UNIVERSIDAD CEU CARDENAL HERRERA

DEPARTAMENTO DE CIENCIAS FÍSICAS, MATEMÁTICAS Y DE LA COMPUTACIÓN

MINERÍA DE DATOS APLICADA AL ANÁLISIS DEL TRATAMIENTO INFORMATIVO DE LA DROGADICCIÓN

MEMORIA DE TRABAJO DE INVESTIGACIÓN

D. PABLO Mª ROMEU GUALLART

DIRIGIDO POR: DR. D. JUAN PARDO ALBIACH

MONCADA, NOVIEMBRE 2010

CEU
Universidad
Cardenal Herrera

ÍNDICE

1.- INTRODUCCIÓN

El tratamiento de las drogodependencias es un tema de gran relevancia en la sociedad, no en vano instituciones públicas y privadas están haciendo un esfuerzo para la sensibilización social hacia la drogodependencia y su prevención. Políticos, fundaciones y medios de comunicación están inmersos en esta ardua tarea de comunicar los riesgos del consumo de drogas, prevenir el consumo y tratar de paliar sus efectos.

Si observamos el problema desde la perspectiva psicológica, nos encontramos con que el ser humano imita aquello que observa, y es por eso que en la sociedad está surgiendo una creciente preocupación por la forma en que se muestran, no sólo las drogas, sino muchos otros temas controvertidos.

Viviendo en una era en la que la información, y por ende la comunicación, son una poderosa herramienta para modelar a la sociedad, la Fundación para el Estudio Prevención y Asistencia a las Drogodependencias de la Generalitat Valenciana y la propia Universidad CEU Cardenal Herrera están colaborando en un proyecto para evaluar cómo se comunican las noticias sobre drogas en nuestro país.

El equipo del proyecto de análisis periodístico dirigido por la Dra. Dña. Pilar Paricio periodístico comenzó a recopilar informaciones sobre drogas aparecidas en diarios de tirada nacional y a realizar algunos estudios sobre los mismos para observar cómo trataban esta información desde el punto de vista periodístico.

En un determinado momento surge en el proyecto la necesidad de realizar un tratamiento estadístico más complejo de los datos. Aparece la necesidad de hallar relaciones ocultas entre la información disponible, que no fueran evidentes a la luz de los primeros resultados estadísticos obtenidos.

En este entorno es en el que aparece el presente proyecto, tratando de aportar nueva información sobre los datos recopilados por el equipo de investigación aplicando técnicas de Minería de Datos sobre una de las muestras recogidas.

En nuestros días, las organizaciones están generando grandes volúmenes de datos como consecuencia de su funcionamiento. Muchos de estos datos terminan, mediante la implantación de Almacenes de Datos, totalizados a modo de, por ejemplo, simple estadística de ventas, o de media salarial, que aportan cierta información. En otras ocasiones, mediante herramientas OLAP (OnLine Analytical Processing) y la generación de hipercubos en el Almacén de Datos, un analista experimentado será capaz de extraer información relevante de estos cubos de datos.

Estos volúmenes de datos contienen en no pocas ocasiones información que podría resultar muy útil a la toma de decisiones, pero que está oculta. Es en este momento en el que surge la necesidad de aplicar técnicas de Minería de Datos.

La Minería de Datos es el núcleo de lo que se conoce como Knowledge Data Discovery –en adelante KDD-. KDD es una metodología que se compone de diferentes fases:

- El proceso de obtención de datos. Este puede ser manual, como en el presente estudio, o automatizado. Esta segunda vertiente abarca el reconocimiento y extracción de patrones en lenguaje natural -Text Mining-, muy extendido en el análisis de textos periodísticos y en particular, en análisis de información online.

- La transformación, limpieza y –en el caso de tener un Almacén de Datos- carga de los datos, para obtener lo que se conoce como vista minable. Esta fase y la anterior se conocen como proceso ETL (Extraction, Transform and Load).

- La selección y aplicación de un modelo de Minería de Datos. Estos pueden ser algoritmos predictivos, reglas de asociación, modelos bayesianos, etc.

- La evaluación y distribución de resultados

El presente estudio se centrará en la transformación y limpieza de datos y en la posterior tarea de Minería de Datos.

La muestra proporcionada por el equipo de investigación se compone de 502 registros de cuatro periódicos diferentes recogidos en un periodo de 6 meses de investigación. Estas muestras contienen más de 50 variables a analizar. Por ejemplo, se sabe que gran parte de estos registros corresponden a artículos donde aparece la droga en un contexto delictivo. El conocimiento que una herramienta de Minería de Datos puede generar sería hallar que en un determinado diario nunca aparece el hachís en un entorno delictivo.

Este ejemplo podría ser obtenido de forma manual por un investigador, de manera fortuita. La Minería de Datos trata de automatizar este procedimiento de "prueba y error" que el investigador debe realizar, de forma que, mediante un proceso semi-dirigido, se puedan hallar estas relaciones, predecir nuevos valores, etc.

En el presente proyecto se pretende ayudar a encontrar aquellas relaciones ocultas en los datos de la muestra que permitan a los investigadores descubrir cómo se realiza el tratamiento de la temática de la drogadicción en la prensa nacional española y dar un soporte más objetivo a sus observaciones.

Por otro lado, se tratarán de confirmar las conclusiones que ya ha publicado el grupo de investigación, e incluso, realizar nuevas aportaciones en aquellos análisis que han ido sugiriendo durante la elaboración de los modelos.

Además, se tratará de estudiar las subpoblaciones de datos, para observar si existen tendencias específicas poblacionales en cada periódico, tema principal de los artículos, droga, etc.

La metodología de trabajo que se sigue durante todo el proyecto se basa en el estándar de Minería de Datos CRISP-DM (Cross-Industry Standard Process for Data Mining). Este estándar está muy orientado a proyectos organizacionales por lo que se ha adaptado a la metodología de un trabajo de investigación.

1.1.- Objetivos

El principal objetivo del presente trabajo de investigación es hallar nueva información estadísticamente relevante de la muestra, que pudiera estar oculta en las relaciones establecidas por los datos. Los objetivos generales serán:

- Realizar una revisión de las técnicas de investigación en el campo de la Minería de Datos, mediante un estudio del estado del arte de la cuestión.
- Además, realizar un estudio del estado del arte de la Minería de Datos exclusivamente en el ámbito de los medios de comunicación.
- Aplicar una metodología de KDD (Knowledge Data Discovery) para la transformación y limpieza de los datos, obteniendo una vista minable, para posteriormente seleccionar y aplicar un modelo de Minería de Datos.
- Comprender el método científico de investigación y aprender la técnica de exposición de resultados.
- Aplicar correctamente la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) para el desarrollo del proyecto de investigación, adaptándola, en su caso, a las peculiaridades de la investigación superior.
- Analizar adecuadamente los resultados y validarlos convenientemente, seleccionando aquellos que resulten relevantes para la investigación.

1.2.- Estructura del trabajo de investigación

El presente estudio se compone de 12 capítulos más los anexos, bibliografía e índices, siendo el primero el presente capítulo de introducción. Los restantes capítulos se estructuran de la siguiente forma:

- En el *capítulo segundo* se tratará de mostrar cuál es el estado del arte de la Minería de Datos, explicando su estrecha relación con los almacenes de datos, el concepto de KDD y una breve clasificación de las diferentes técnicas.

- En el *capítulo tercero* se repasarán los principales algoritmos encuadrados dentro de las diferentes técnicas de Minería de Datos. Entre otros: algoritmos predictivos, de clasificación, de reglas de asociación, modelización estadística paramétrica y no paramétrica.

- El *capítulo cuarto* versa sobre el estado de la cuestión de la Minería de Datos como herramienta de análisis en periodismo. Se revisarán las distintas referencias halladas, así como los dos enfoques que se han observado en estos textos: enfoque de la extracción de información de textos –Text Mining- y el enfoque de análisis periodístico propiamente dicho, donde se aplican los algoritmos de Minería de Datos.

- El *capítulo quinto* introduce la metodología de trabajo CRISP-DM para el desarrollo de proyectos de Minería de Datos.

- En el *capítulo sexto* se abordará el proyecto de investigación del análisis del tratamiento informativo de la drogadicción, mostrando los objetivos del mencionado proyecto, la población de la muestra, las fases del proyecto y las dos metodologías de análisis periodístico empleadas por el equipo de investigación periodística: el análisis de contenido desde la perspectiva del *framing* y el análisis de intensidad formal.

- El *capítulo séptimo* corresponde a la primera fase de la metodología CRISP-DM, con la comprensión del caso de estudio. Se trata de evaluar quién está detrás del estudio, la evaluación de la situación actual, los recursos de los que se dispone y determinar los objetivos de la Minería de Datos.

- El *capítulo octavo* describe el proceso de comprensión de datos del problema, donde se describen lo datos y se exploran para obtener una primera visión de la muestra, así como evaluar la calidad del formato de los mismos.

- El *capítulo noveno* trata de mostrar cómo se han preparado los datos mediante un proceso ETL (Extraction Transform Load) para obtener la vista minable adaptada al algoritmo que posteriormente se utilizará.

- En el *capítulo décimo* se argumenta la selección de la técnica de modelado, se muestran los diferentes parámetros utilizados para la construcción de los distintos modelos, así como las medidas que se utilizarán para evaluarlos.

- En el *capítulo undécimo* se evalúan los resultados y se describen los hallazgos de cada modelo a modo de informe final.

- El *capítulo duodécimo* consta de una primera parte donde se evalúan las principales aportaciones del estudio así como una discusión de las limitaciones del mismo y de las distintas decisiones adoptadas. En una segunda parte, se

proponen nuevas líneas de investigación abiertas durante la realización del presente proyecto.

2.- Estado del arte de la Minería de Datos

2.1.- ¿Qué es la Minería de Datos?

Según [Hernández Orallo, et al. 2004], se llama Minería de Datos a un conjunto de técnicas que van encaminadas a extraer conocimiento a partir de grandes volúmenes de datos.

Hoy en día, las empresas e instituciones almacenan una gran cantidad de datos en bases de datos relacionales y en otros tipos de fuentes, y esta utilización ha aumentado considerablemente durante los últimos años y se prevé que aumentará a un ritmo mayor. A esto han contribuido:

- El desarrollo de las comunicaciones así como la implementación y mejora de las redes informáticas que permiten comunicarse y transmitir información de manera cómoda y fácil. Además, durante los últimos 15 años y gracias al uso de Internet ha aumentado mucho la posibilidad de utilización de fuentes externas.
- Las aplicaciones especializadas de sistemas de información tales como ERPs (Enterprise Resource Planning) o CRMs (Customer Relationship Management).

Muchas veces, las empresas no saben obtener información valiosa de la cantidad ingente de datos que tienen almacenados, aunque el conocimiento que podrían extraer de estos podría ser de gran ayuda en muchas de las áreas y facetas de su negocio.

Tener un aceptable grado de automatización y disponer de almacenes de datos es requisito indispensable ya que, si no se dispone de la infraestructura necesaria para capturar y almacenar convenientemente la información, difícilmente se podrá obtener nada de ella.

Gran parte de la información que se encuentra se corresponde con históricos que ya no sufren variaciones. La información histórica puede ser útil para explicar el pasado y para poder predecir el futuro.

Aquello que se llama el valor añadido de una empresa y que últimamente se denomina con el término inglés *know-how,* se apoya fundamentalmente en el conocimiento de experiencias pasadas y otro tipo de información diversa para poder predecir qué ocurrirá en un futuro. Para obtener conclusiones en una empresa, a menudo, se necesita integrar y analizar información proveniente de diferentes fuentes. La Minería de Datos pretende automatizar estas tareas y realizarlas de forma cuantitativa incluyendo toda la información disponible. En un proceso donde no intervengan este tipo de herramientas, se descartarían muchos datos debido a que a los seres humanos nos resulta imposible obtener conclusiones analizando gran cantidad de datos, o sería

necesario trabajar con resúmenes a partir de esta información totalizada, por lo que se perdería mucha información.

La Gestión del Conocimiento (Knowledge Discovery) abarca todas aquellas tecnologías relativamente nuevas que surgen de la necesidad de procesar, analizar y aprovechar la información "escondida" en grandes volúmenes de datos. La Gestión del Conocimiento requiere una captación, estructuración y transmisión de conocimiento. Permite a los que la usen obtener información útil de la forma más eficiente posible a partir de los datos.

Los procesos de análisis de datos se han realizado en las empresas de forma manual o mediante herramientas de estadística simple. Hasta la fecha, el equipo del proyecto de investigación realizaba estadísticas simples basadas en la observación de parámetros.

Por ejemplo, en los estudios que este grupo viene realizando, como en [Paricio Esteban, et al. 2010] , se detallan estadísticos habituales tales como media, moda, estudio de frecuencias, etc. así como algún estudio de relaciones entre variables, por ejemplo, qué tema de las noticias sobre drogas obtiene una valoración formal más alta.

En este caso, la revisión de los datos es lenta, subjetiva y tiene un alcance muy limitado. Es subjetiva debido a que el grupo de investigación se centra en aquellas relaciones que a priori puede considerar que tienen una relación lógica dentro de la idiosincrasia del propio estudio. Este tipo de estudio obtiene resultados pero es posible que omita relaciones y peculiaridades de los datos ya que sólo realiza análisis desde algunos enfoques y no es exhaustivo. Estas peculiaridades, que podrían resultar interesantes por desviarse de los resultados de otros estudios, serían omitidas si sólo se tiene la intuición del investigador como herramienta principal.

La Minería de Datos es un área dentro de la gestión del conocimiento y se puede definir como un conjunto de metodologías y herramientas que permiten extraer el conocimiento útil para ayuda a la toma de decisiones, comprensión y mejora de procesos o sistemas partiendo de grandes volúmenes de datos.

A continuación se darán varias definiciones de Minería de Datos:

Según [Piatetski-Shapiro, et al. 1991] es el "conjunto de técnicas y herramientas aplicadas al proceso trivial de extraer y presentar el conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con el objeto de predecir de forma automatizada tendencias y comportamientos, y/o descubrir de forma automatizada tendencias y comportamientos y/o descubrir de forma automatizada modelos previamente desconocidos".

Según [Berry, et al. 1997] es la "exploración y análisis, mediante métodos automáticos o semiautomáticos de grandes volúmenes de datos para descubrir reglas o patrones significativos".

Según [Witten, et al. 2000] “la Minería de Datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido desde grandes cantidades de datos almacenados en distintos formatos. Este proceso deberá ser automatizado en mayor o menor medida y deberá generar modelos que ayuden al negocio a tomar decisiones.”

Según [Hand, et al. 2001] es el “análisis de habitualmente grandes series de datos (observaciones) para encontrar relaciones inesperadas y resumir la información de nuevas maneras que sean entendibles y útiles para el propietario de los datos.”

Por lo tanto, y según [Hernández Orallo, et al. 2004], dos son los retos de la Minería de Datos: trabajar con grandes volúmenes de datos y usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. En muchos casos la utilidad del conocimiento minado tiene una relación directa con la comprensibilidad del modelo inferido. Dado que el usuario final en muchas ocasiones no es un experto en técnicas de Minería de Datos, este modelo generado debe ser comprensible. En el caso que nos ocupa, esta premisa es fundamental debido a que el ámbito del grupo de investigación –ciencias sociales- está claramente alejado de las ciencias experimentales.

El objetivo final de la Minería de Datos es transformar datos en conocimiento.

2.2.- ALMACENES DE DATOS

Hace 15 años, el análisis de los datos se realizaba con herramientas de consulta sobre la base de datos operacional, que se basa en un modelo transaccional. Estas herramientas de consulta estaban basadas en lenguajes generalistas de consultas, principalmente SQL. Estas herramientas eran poco flexibles y poco escalables a grandes volúmenes de datos.

Para responder a la necesidad de flexibilizar este tipo de consultas, que tienen un sentido analítico en lugar de operacional se ha creado una nueva tecnología de bases de datos basada en una nueva arquitectura. Estos son los almacenes de datos (datawarehouse). Los almacenes de datos son el repositorio de datos.

Para poder llevar a cabo las técnicas de Minería de Datos de una manera eficiente, se necesita un sistema de adquisición, almacenamiento y manejo de la información eficiente. Por ello se hacen necesarios los almacenes de datos y los sistemas OLAP.

Según [Kimball, et al. 1998] el Data Warehousing es el proceso a través del cual se organiza una gran cantidad de datos heterogéneos y almacenados, de forma que facilite la recuperación de información para llevar a cabo el proceso analítico.

Los almacenes de datos generan bases de datos con una perspectiva histórica, utilizando datos de múltiples fuentes que se fusionan de forma interrelacionada. Estos datos se mantienen estables, sin variar como en los sistemas transaccionales. Los almacenes de datos se alimentan a partir de los datos transaccionales y permiten

realizar consultas operativas de forma que se pueda obtener información para realizar análisis multidimensional tan útil en las empresas hoy en día para los cuadros de mando.

Para comprender cómo se hace la información útil para un analista de negocio, se presenta un ejemplo sencillo de análisis multidimensional en acción.

El ejemplo trata de un mayorista de frutas que compra fruta que proviene de los agricultores y luego transporta y distribuye la fruta en cuatro mercados. Se pretende analizar de las ventas. La Tabla 1 muestra la información de las ventas para el primer y segundo trimestre de 2009.

Tabla 1: Información de ventas del mayorista de frutas del primer semestre de 2009

Trimestre	Ventas
Enero a Marzo	16.000,00 €
Abril a Junio	16.000,00 €
Total	32.000,00 €

Según la Tabla 1, a simple vista parece que el mayorista de frutas tiene el mismo rendimiento en sus ventas en el primer trimestre y en el segundo.

El siguiente paso en el análisis de ventas será seguir analizando las ventas pero desde otra dimensión o perspectiva; por ejemplo, se puede conocer el tipo y el lugar donde la fruta fue vendida. La Tabla 2 muestra esa información.

Tabla 2: División de las ventas del mayorista de frutas por localidades y ventas por productos

Mercado	Ventas	Producto	Ventas
Barcelona	8.000,00 €	Manzanas	8.000,00 €
Madrid	8.000,00 €	Cerezas	8.000,00 €
Sevilla	8.000,00 €	Uvas	8.000,00 €
Valencia	8.000,00 €	Melones	8.000,00 €
Total	32.000,00 €	Total	32.000,00 €

Es destacable que el total de ventas es el mismo 32.000 €, en todas las vistas; esto es un signo de confianza. Esto da la seguridad de que estamos viendo la misma información (las ventas de frutas de la compañía), pero cada vista rompe o totaliza en diferentes categorías. Consideremos por un momento lo que hemos hecho; hemos examinado todas las ventas totalizadas en tres categorías diferentes, tiempo, mercado y producto. Esta categorización es lo que se conoce con el nombre de **dimensiones**.

Basados en los datos presentados en la Tabla 2, no resulta obvio conocer cuál podría ser la siguiente pregunta. En este punto, sabemos que las ventas son idénticas para cada uno de los trimestres, para cada uno de los cuatro productos, y para cada uno de los cuatro mercados. Mejor que analizar la información de ventas en cuatro dimensiones, se puede ver qué sucede cuando se combinan las tres dimensiones existentes para crear una vista multidimensional, como se muestra en la Tabla 3.

Tabla 3: División de las ventas por localidades, trimestres y productos

		Barcelona	Madrid	Sevilla	Valencia	Total
1er Trimestre	Manzanas			2.500,00 €	1.500,00 €	4.000,00 €
	Cerezas			2.000,00 €	2.000,00 €	4.000,00 €
	Uvas	1.000,00 €	3.000,00 €			4.000,00 €
	Melones	2.000,00 €	2.000,00 €			4.000,00 €
	Total	3.000,00 €	5.000,00 €	4.500,00 €	3.500,00 €	16.000,00 €
2º Trimestre	Manzanas	4.000,00 €				4.000,00 €
	Cerezas	1.000,00 €	3.000,00 €			4.000,00 €
	Uvas			1.500,00 €	2.500,00 €	4.000,00 €
	Melones			2.000,00 €	2.000,00 €	4.000,00 €
	Total	5.000,00 €	3.000,00 €	3.500,00 €	4.500,00 €	16.000,00 €
	Totales	8.000,00 €	8.000,00 €	8.000,00 €	8.000,00 €	32.000,00 €

Aquí se observa que información trascendente que estaba oculta por haber analizado las dimensiones de forma separada comienza a aparecer. Por ejemplo, las manzanas y cerezas no se vendieron en Madrid y Barcelona durante el primer trimestre, pero los melones y las uvas sí se vendieron. Sucedió lo contrario durante el segundo trimestre.

Este proceso de interactuar con datos en vistas multidimensionales, es lo que se conoce como “rebanar y dividir” (slice and dice). Esta técnica casi siempre revela nuevas e interesantes informaciones en comparación con los datos aislados en dimensiones sencillas. El análisis multidimensional supone la visualización de los datos simultáneamente en categorías a lo largo de muchas dimensiones, no necesariamente tres dimensiones como el ejemplo anterior.

Gráficamente, el funcionamiento general de los almacenes de datos se puede ver en la Figura 1.

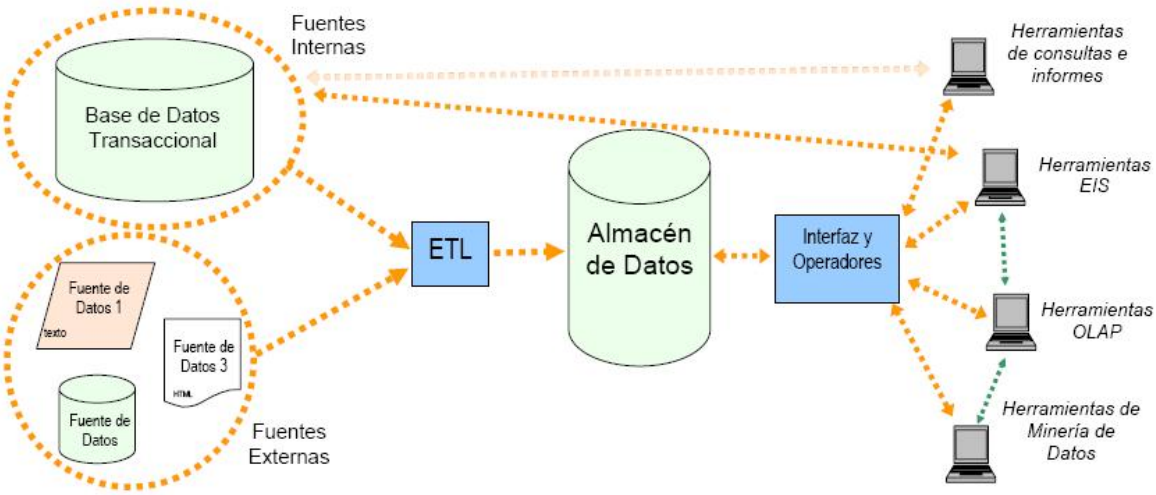

Figura 1: Funcionamiento de un almacén de datos

Tal como se observa en la Figura 1, los almacenes de datos se nutren tanto de fuentes internas de la empresa (las BBDD transaccionales -operacionales- del ERP de la empresa, el CRM, etc.) como de fuentes externas de datos. Un proceso de ejecución periódica se encarga de extraer, transformar y cargar (ETL, Extract, Transform and Load) estos datos y depositarlos en el almacén de datos (DataWarehouse) de manera que estén disponibles para las herramientas de BI encargadas de mostrar los datos y trabajar con ellos. Es importante observar que los datos depositados en el DW no están disponibles en tiempo real para su consulta, solamente lo están una vez el proceso de ETL los ha depositado.

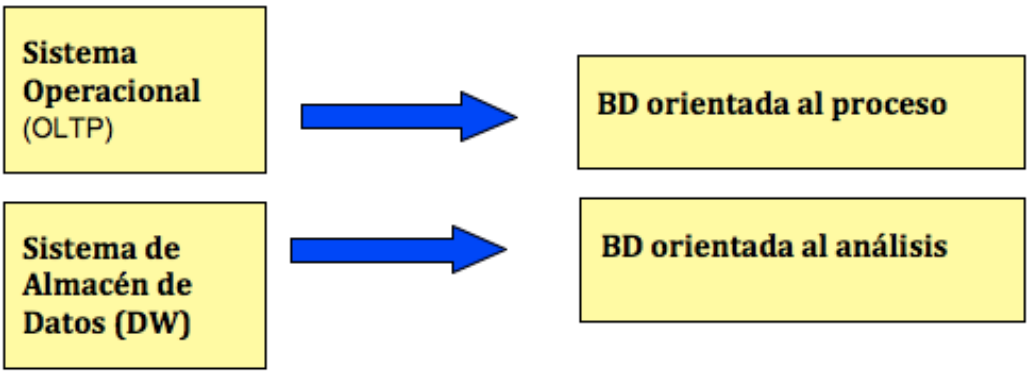

Figura 2 Relación entre tipos de sistemas y tipos de bases de datos utilizadas

En esta **¡Error!No se encuentra el origen de la referencia.** se ve gráficamente que una base de datos de un sistema operacional está orientada al proceso del día a día, mientras que hace falta otros tipo de bases de datos para realizar análisis en sus totalidad del conjunto de datos, y para ello existen los almacenes de datos.

Según [Hernández Orallo, et al. 2004] las principales características de los almacenes de datos son las siguientes:

Están orientados hacia la información relevante de la organización. Se diseñan para consultar eficientemente información relativa a las actividades básicas de la organización (ventas, compras, producción,…) y no para soportar los procesos que se realizan en ella (gestión de pedidos, facturación, control de stocks,…). Tal como se observa en la Figura 3, solo se obtienen los datos de ciertas tablas de la base de datos transaccional, y no de su totalidad. En el almacén de datos estarán únicamente los datos necesarios para el proceso de análisis.

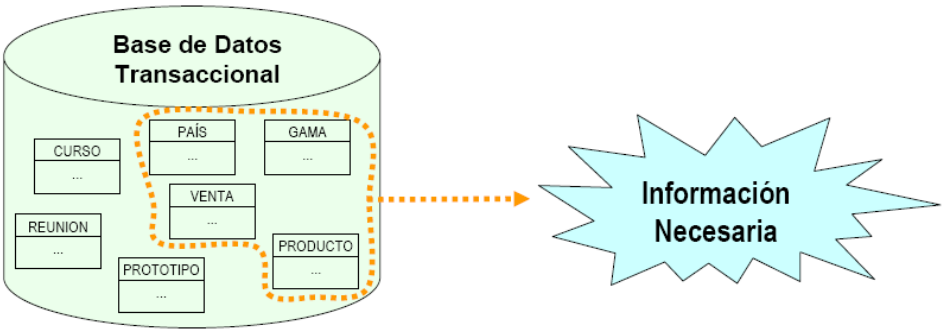

Figura 3: Extracción de información de una base de datos transaccional para su análisis

Están integrados. Integra datos recogidos de diferentes sistemas operacionales de la organización (y/o fuentes externas).

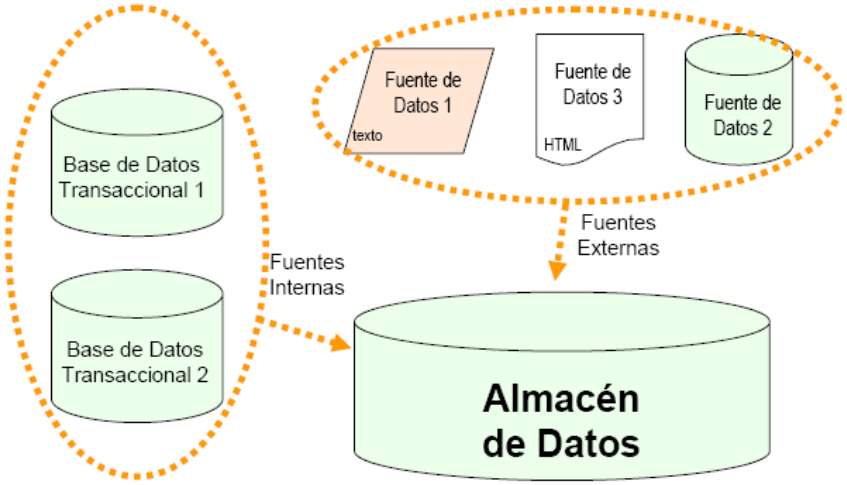

Figura 4: Integración de datos provenientes de diversas fuentes

En la Figura 4 se observa que en muchas ocasiones deben integrarse diversas fuentes internas provenientes de diversas bases de datos transaccionales, que pueden llegar a estar almacenadas en diferentes motores de bases de datos de diferente fabricante, y se integran con fuentes externas que pueden estar en ficheros de texto, ficheros semiestructurados, etc.

Variables en el tiempo. Los datos son relativos a un periodo de tiempo y deben ser incrementados periódicamente, tal como aparece en la Figura 5.

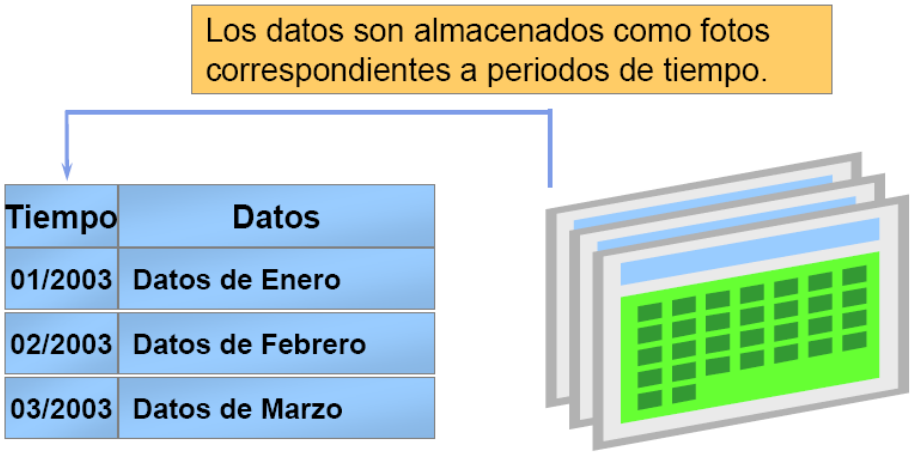

Figura 5: Datos almacenados en un almacén de datos agrupados por periodos de tiempo

No volátiles. Los datos almacenados no son actualizados, sólo incrementados.

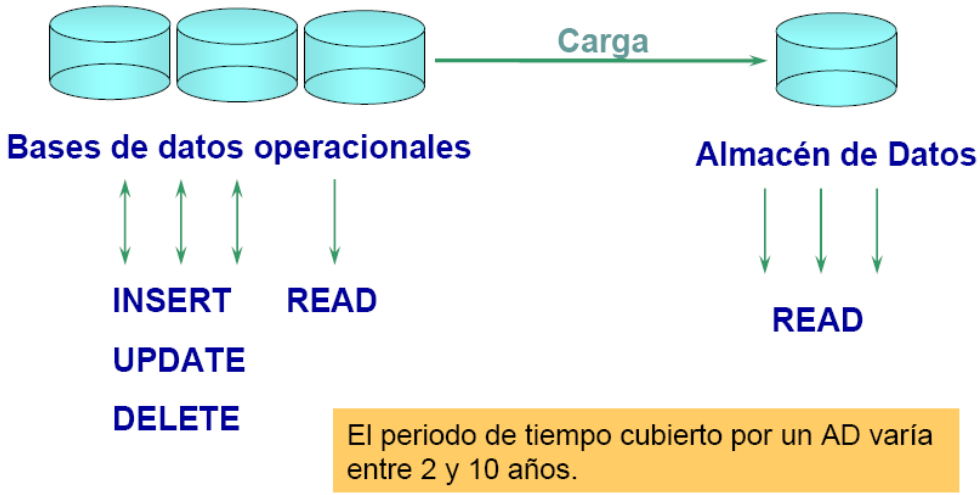

Figura 6: Proceso de carga de un almacén de datos

Tal como muestra la Figura 6, las operaciones habituales en una base de datos operacional son inserciones, borrados, actualizaciones y lecturas, mientras que en un almacén de datos, tras el proceso inicial de carga, se realizan únicamente lecturas.

A continuación se presenta en la Tabla 4, que muestra una comparativa entre los sistemas transaccionales y los almacenes de datos.

Tabla 4: Comparativa entre los sistemas transaccionales y los almacenes de datos

Sistema Transaccional (OLTP)	Datawarehouse (DW)
Almacena Datos actuales	Almacena Datos históricos
Datos dinámicos (actualizables)	Datos estáticos
Elevado número de transacciones	Número bajo de transacciones
Tiempo de respuesta pequeño (segundos)	Tiempo de respuesta variable (secs – horas)
Dedicado al procesamiento de transacciones.	Dedicado al análisis de Datos

Sistema Transaccional (OLTP)	Datawarehouse (DW)
Orientado a los procesos de la organización	Orientado a la información relevante
Soporta decisiones diarias	Soporta decisiones estratégicas
Sirve a muchos usuarios (administrativos)	Sirve a Ejecutivos y Dirección

OLAP (Online Analitical Processing) es la protocolo para poder realizar consultas analíticas, en contraposición a las transaccionales

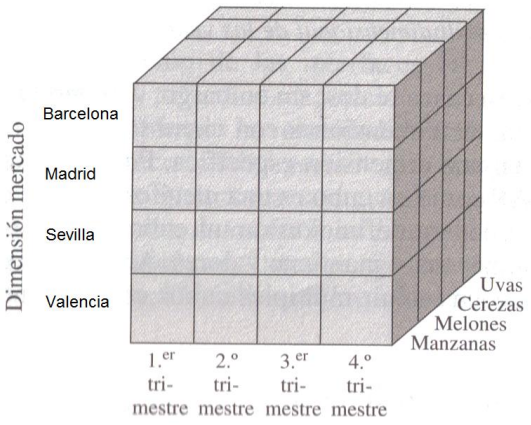

Figura 7: Cubo OLAP del mayorista de frutas

Comúnmente se conoce a los sistemas OLAP como cubos OLAP. Los datos multidimensionales son típicamente visualizados como una estructura de almacenamiento en cubo con un montón de mini-cubos o celdas haciendo el cubo como un todo. La Figura 7 ilustra el cubo que representaría los datos del ejemplo anterior del mayorista de frutas. Un cubo representado en tres dimensiones: Mercados, Tiempo y Productos.

Además de las dimensiones dentro de los sistemas OLAP se encuentran los hechos que son objeto del análisis los cuales representan la información mediante un conjunto de indicadores (medidas). En el ejemplo anterior del mayorista de frutas el hecho a analizar eran las ventas. La medida utilizada era el importe de las mismas expresado en euros (se podría haber cogido como medida la cantidad de fruta vendida expresada en toneladas o kilos).

Otra variable que interviene son los atributos relacionados con la dimensión. En el caso de la dimensión mercado se podría haber visualizado por algún atributo distinto, como

país, provincia del mismo (aunque en el ejemplo coincidiría), etc. que pueden formar una jerarquía. En nuestro ejemplo, podría ser País > Comunidad Autónoma > Provincia…

Así pues tenemos a la hora de analizar un modelo multidimensional los siguientes componentes:

- Dimensiones
- Hechos
- Medidas de los hechos
- Atributos de las dimensiones.

Veamos esto con otro ejemplo. Imaginemos que deseamos analizar las ventas de una cadena de grandes almacenes por Producto, Tiempo y Almacén (tienda). Tal como se ve en la Figura 8, el hecho a analizar son las ventas, que tiene como sus medidas el importe de estas ventas y las unidades vendidas. Como se puede ver, ambos parámetros se refieren a las ventas. Las ventas además pueden seccionarse por las dimensiones que tenemos. Se podrían diseccionar por tiempo, producto o almacén (ubicación). Al diseccionarla por las dimensiones, se pueden tomar diferentes medidas de estas dimensiones, que son los atributos de las dimensiones. Tomando la dimensión tiempo, este tiempo podría ser por años, trimestres, meses, semanas o días. Todas son unidades de tiempo, pero según el nivel de granularidad que se desee obtener al visualizar los datos puede elegir unos atributos u otros.

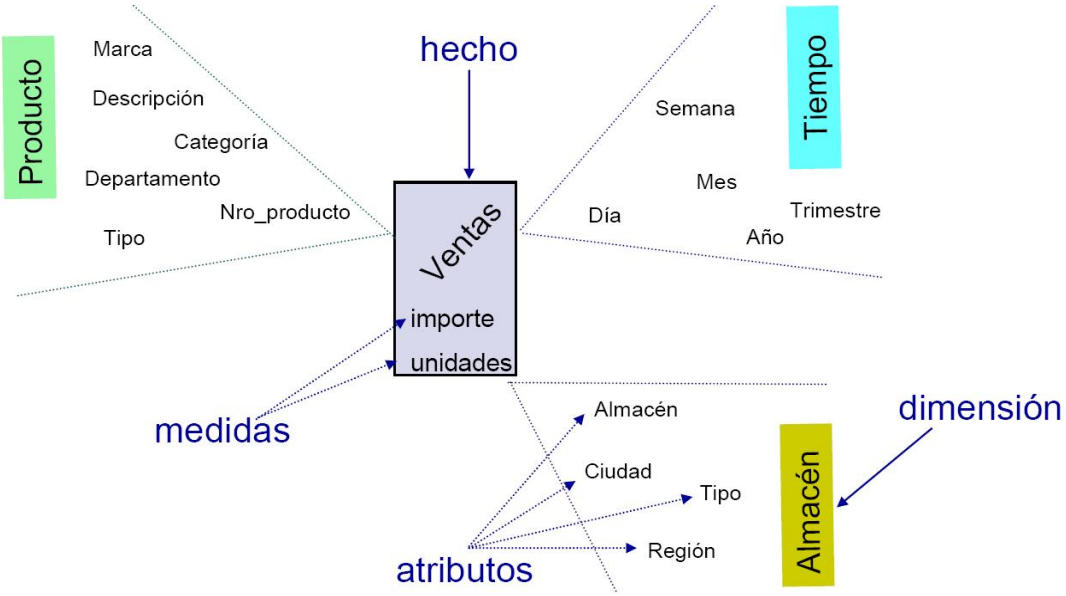

Figura 8: Hechos, medidas, dimensiones y atributos de las dimensiones

Como se ha indicado antes, los atributos pueden organizarse en jerarquías. La jerarquía es una organización en niveles dentro de una dimensión (de sus atributos). La utilización de jerarquías permite ver los datos con mayor o menor nivel de detalle. Así, en el ejemplo de las ventas de frutas, se podría haber definido una jerarquía en la dimensión tiempo que permitiera ver los datos agrupados por trimestres o bien con un mayor detalle a nivel de mes o incluso de semana o día. En la Figura 9 se pueden encontrar ejemplos de jerarquías:

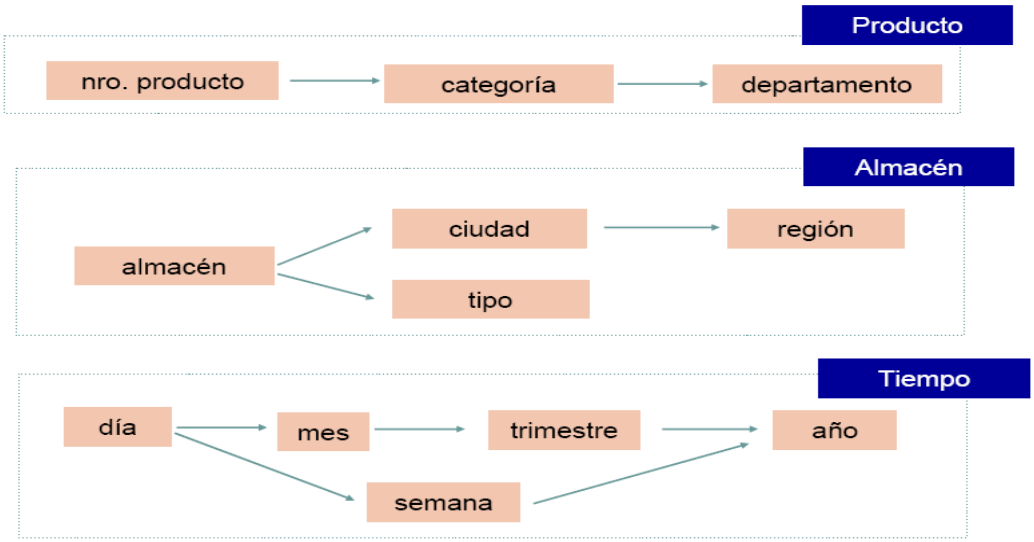

Figura 9: Ejemplos de jerarquías

Por último, es interesante saber distinguir entre los sistemas OLAP y los sistemas orientados a la toma de decisiones, donde estaría englobada la Minería de Datos. La Minería de Datos es el proceso en el que se apoyan las decisiones que buscan, a través de patrones de comportamientos, patrones de información en los datos a partir de los que se podrán obtener las tendencias. Por el contrario, un sistema OLAP ayuda a localizar los resúmenes de la información y es flexible para hacer consultas con mayor o menor nivel de detalle de totalización, así como consultar por diferentes dimensiones.

Mientras que un sistema OLAP nos ayudaría a responder preguntas del tipo: "Compraron más vehículos del modelo X los habitantes del norte de España o del sur en el año 1998?" un sistema DSS (Decision Support System) apoyado en técnicas de Minería de Datos, nos ayudaría a responder preguntas del tipo: *"¿Quiere un modelo que identifique las características predictivas más importantes de las personas que compraron un vehículo de la marca X?".*

2.3.- TIPOS DE DATOS

Tras ver lo que es un almacén de datos se necesita conocer los tipos de datos sobre los que actúa la Minería de Datos. Aunque a los informáticos nos da la sensación de que los datos suelen almacenarse en bases de datos relacionales, por tratarse de la forma más estructurada y más común, existen muchas otras fuentes de datos tratadas por diversas disciplinas de la Minería de Datos.

La fuente de datos con la que habitualmente se enfrenta la Minería de Datos son las bases de datos relacionales. Muchas técnicas de Minería de Datos pueden enfrentarse únicamente con una tabla a la vez por lo que debe realizarse un proceso de desnormalización para reunir todos los datos que nos interesan en una sola tabla y así construir lo que se llama la *vista minable*. También es importante conocer que así como en las bases de datos existen muchos tipos de datos, a nivel de Minería de Datos solo nos interesa distinguir entre dos tipos de datos:

- **Numéricos**: Cualquier tipo de número tanto enteros como reales. Por ejemplo, la edad.
- **Categóricos**: Toman el valor entre un conjunto finito de categorías. Por ejemplo, un valor relacional booleano es un categórico con valores Sí y No. Dentro de los categóricos hay que distinguir entre:
- **ordenados**: Si existe un orden entre las diferentes categorías. Por ejemplo, en nuestro caso: Valoración de la unidad de Análisis Discreta: Baja, Media, Alta
- **no ordenados**: Si no existe un orden entre las diferentes categorías y el conjunto de categorías es una mera enumeración. Por ejemplo: Periódico: El Mundo, ABC, El País, La Razón.

Las bases de datos temporales incluyen atributos donde el tiempo es muy importante. Los atributos son almacenados con la característica del tiempo en que se producen. Este es el caso que nos ocupa, donde se van almacenando las noticias conforme van apareciendo en los periódicos.

Las bases de datos multimedia almacenan imágenes, audio y video. Por ejemplo, el reconocimiento del tema de una imagen o su clasificación por su similitud, podrían ser objeto de esta disciplina de la Minería de Datos.

Las bases de datos documentales contienen documentos de texto, tanto estructurados como semiestructurados y no estructurados. Para poder estudiar estos documentos se utilizan además de las técnicas tradicionales de base de datos relacionales, técnicas específicas para obtener datos a partir de textos tales como las bolsas de palabras.

2.4.- El proceso del KDD (Knowledge Data Discovery)

El KDD es un término muy relacionado con la Minería de Datos y que en muchas ocasiones se confunde con esta, aunque no son lo mismo.

El KDD se refiere al proceso de búsqueda y extracción de conocimiento a partir de las bases de datos, mientras que la Minería de Datos es la parte de este proceso en la que se utilizan las técnicas de inteligencia artificial para obtener un modelo.

Según [Fayyad, et al. 2002] el KDD es "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia (Y sin que sea estrictamente necesario) comprensibles a partir de los datos."

Es muy habitual por tanto confundir la parte (Minería de Datos) con el todo (proceso KDD). Como se ve en la Figura 10, la Minería de Datos forma parte del proceso de KDD, pero este es más amplio y engloba otras tareas además de la Minería de Datos. El KDD forma parte de un área científica más amplia como es el descubrimiento de conocimiento que tiene otras muchas partes dentro de ella diferentes al KDD.

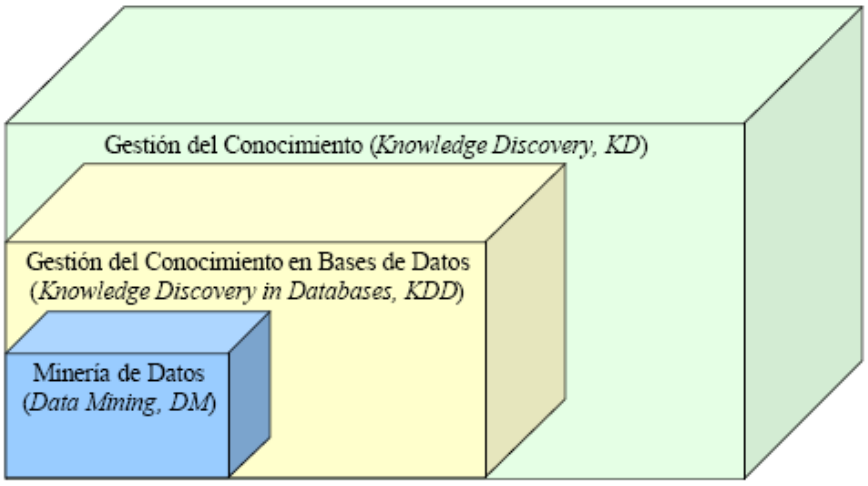

Figura 10: Comparación de los conceptos de Minería de Datos, KDD y Knowledge Discovery

Las fases de un proceso de KDD no están claramente definidas, por lo que a continuación se exponen diversas visiones de autores muy relevantes de las fases del KDD.

Según [Pernía, et al. 2001] las fases del KDD son:

- Exploración del dominio

- Recolección de los datos
- Extracción de patrones en los datos
- Inducir generalizaciones
- Verificación del conocimiento
- Transformación del conocimiento.

[Brachman, et al. 1996] definen las fases así:

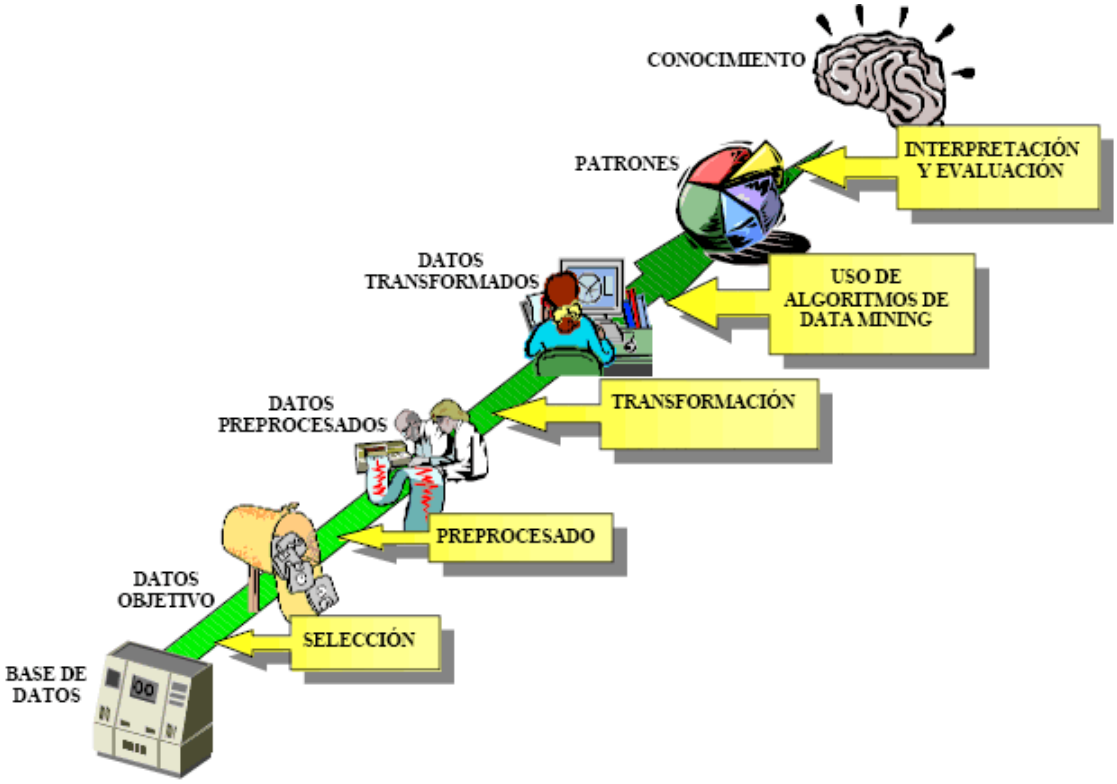

Figura 11: Proceso de KDD según [Brachman, et al. 1996]

En [Hernández Orallo, et al. 2004] se nos exponen las siguientes fases en el proceso de KDD:

- Preparar los datos:
    - Determinar las fuentes de información que pueden ser útiles y dónde conseguirlas.
    - Diseñar el esquema de un almacén de datos (Data Warehouse) que consiga unificar de manera operativa toda la información recogida.

- Implantación del almacén de datos que permita la "navegación" y visualización previa de sus datos, para discernir qué aspectos pueden interesar en el estudio.
- Selección, limpieza y transformación de los datos que se van a analizar. La selección incluye tanto una criba o fusión horizontal (filas) como vertical (atributos).

- Minería de Datos:
  - Seleccionar y aplicar el método de Minería de Datos apropiado.
  - Evaluación/Interpretación/Visualización
  - Evaluación, interpretación, transformación y representación de los patrones extraídos.
  - Difusión y uso del nuevo conocimiento.

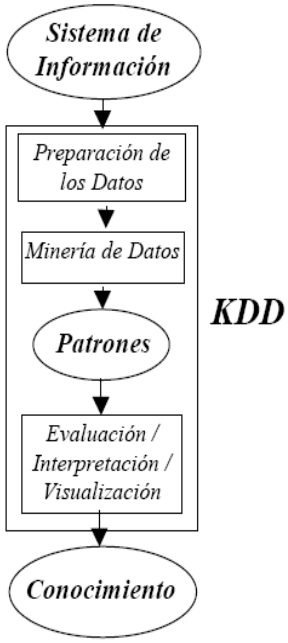

Figura 12: Fases del proceso de KDD

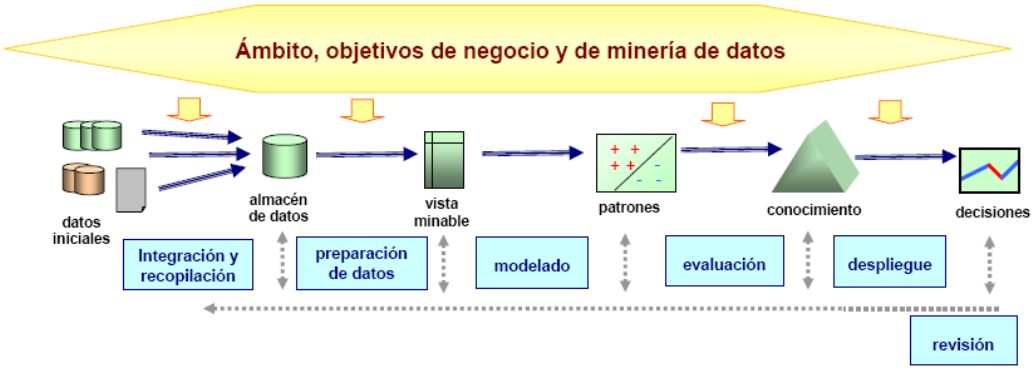

Figura 13: Fases de un proceso de KDD

En la Figura 13, se observan las diferentes fases de un proceso de KDD con los resultados de cada fase. Tras la integración y recopilación de datos se obtiene el almacén de datos. Así mismo tras la fase de preparación de datos se obtiene la vista minable. La fase de modelado parte de la vista minable y genera los patrones o modelos. Por último, tras la evaluación se obtiene el conocimiento del negocio que se aplica en la fase de despliegue con una toma de decisiones.

2.4.1.- FASE DE RECOPILACIÓN E INTEGRACIÓN

La identificación de los datos relevantes para una operación de datamining es una tarea que no puede ser automatizada y que debe ser realizada por el analista. Consiste en crear un conjunto de datos objetivo, seleccionando un conjunto de variables o muestras de datos objetivo. Deben ser seleccionados los datos más relevantes del proceso, así como su disponibilidad. Implica considerar la homogeneidad y variación con el tiempo.

En muchas ocasiones, no solo se tiene que recoger información de distintas fuentes internas (sistemas transaccionales, archivos y almacenes de datos), sino que se debe recurrir a fuentes externas para poder recoger esta información.

Lo ideal es que se recuperaran los datos de un almacén de datos, ya que de otro modo la calidad de los datos puede ser mucho menor. Por ello, es necesaria la siguiente fase.

2.4.2.- FASE DE SELECCIÓN, LIMPIEZA Y TRANSFORMACIÓN

El objetivo de esta fase es crear un conjunto de datos más significativo y manejable en cuanto a tamaño. Según [Pyle. 1999] esta fase puede llevar el 60% del coste del proyecto.

Según [Witten, et al. 2000] los tipos de datos se clasifican en:

- Numéricos o cuantitativos: Son los atributos continuos. Valoración de unidad de Análisis, por ejemplo.
- Nominales o Cualitativos: Son los atributos discretos. Divididos según [Hair, et al. 1999] en:
- Nominales: Asociados con etiquetas o nombres. Por ejemplo, un campo de texto libre como puede ser el nombre del periodista que firma la noticia.
- Ordinales: Nominales ordenados. o la Ubicación de la Noticia dentro del periódico.
- Intervalos: Valores ordenados medidos por intervalos iguales. Valoración de unidad de Análisis discreta, donde cada valor corresponde a un intervalo de datos.
- Ratios: Medidas donde el punto origen está definido en si mismo.

Los atributos deben estar en el tipo más adecuado para los algoritmos que se utilizarán. Por ello, en muchas ocasiones es conveniente la conversión de tipos para acomodarlos al tipo de algoritmo que utilizamos según las necesidades. Es decir, es necesario comprender como trabajan los algoritmos de Minería de Datos que se utilizan para saber cómo preparar los datos.

Muchas veces un atributo numérico puede ser convertido a un atributo ordinal simplemente indicándole al sistema unas reglas que los relacionan. Por ejemplo, la discretización de la Valoración de Unidad de Análisis realizada en nuestro estudio.

Puede ser que para discretizar se utilicen reglas difusas, capaces de tratar las incertidumbres mediante funciones.

Como resultado de esta selección y limpieza se obtiene la vista minable, donde están las diversas fuentes integradas, limpias y los atributos relevantes seleccionados. Muchas de estas operaciones de limpieza requieren el uso de técnicas de Minería de Datos que permiten entre otras cosas:

- Detectar valores anómalos como son los outliers (datos inconsistentes) y eliminarlos.
- Detección de valores faltantes y de registros de baja calidad y su eliminación o relleno.
- Eliminar el ruido.

Dentro de esta fase entran las operaciones de transformación y reducción de datos, ya que en muchas ocasiones estas operaciones marcan el éxito o fracaso del proceso.

Se pretende preparar la información para poder procesarse por los algoritmos de Minería de Datos, y reducir la cantidad de información a tratar en la fase de Minería de Datos.

Se busca:

- Reducir la dimensionalidad: Extraer las características útiles de los datos. Seleccionar los atributos más relevantes, en caso de tener muchas dimensiones.
- Transformar los datos para que su representación sea más intuitiva y manejable. Es decir, generar nuevos atributos que resuman los aspectos más significativos de los anteriores.

Para ello se realizan 3 tareas:

- Reducción de los datos: Consiste en eliminar aquellos datos innecesarios para el proceso de extracción de conocimiento.
- Creación de datos derivados: En ocasiones hay datos que deben combinarse o transformarse para poder obtener información de ellos.
- Transformación de la distribución de los datos: En ocasiones es necesario aplicar transformaciones a los datos como cambiar los ejes de referencia para poder obtener información de los datos.

2.4.3.- Fase de Minería de Datos

Una vez recogidos los datos de interés, se debe decidir qué tipo de patrón o modelo se quiere descubrir. El tipo de conocimiento que se desea extraer va a marcar la técnica de Minería de Datos a utilizar. También, en muchas ocasiones la elección de la técnica depende de los tipos de datos con los que se trabajan y si contienen suficientes ejemplos de cada clase como para obtener un modelo fiable. No todas las técnicas trabajan bien con datos faltantes, con datos escasos o con datos escasos de una de las clases de entrenamiento.

En el caso, el resultado debe ser un conjunto de reglas de amplia cobertura que nos permita obtener conclusiones estadísticas significativas y que nos permita explicar el comportamiento con reglas y a nivel estadístico. En nuestro caso, el algoritmo APRIORI nos ayudará a comprender las distintas relaciones entre los datos.

Las herramientas de Minería de Datos empleadas para extraer el conocimiento se pueden clasificar en dos grandes grupos:

- **Técnicas de verificación**: En las que el sistema se limita a comprobar hipótesis suministradas por el usuario.

- **Métodos de descubrimiento**: En las que se quieren encontrar patrones interesantes de forma automática.

En los métodos predictivos, es importante que se separen los datos de la vista minable en dos partes que no deben ser del mismo tamaño, el conjunto de entrenamiento y el de prueba. El conjunto de entrenamiento que se utiliza en esta fase de Minería de Datos debe ser el más grande. Este conjunto de entrenamiento debe tener suficientes datos como para que la generación del modelo no se sobreajuste a los datos de entrenamiento. El sobreajuste (overfitting) se produce cuando por falta de datos el modelo funciona muy bien para los datos de entrenamiento, pero al cambiar los datos por otros, el modelo tiene muy poca precisión. El sobreajuste produciría reglas muy específicas que únicamente servirían para el conjunto de entrenamiento. El sobreajuste se elimina con el aumento del conjunto de entrenamiento. La otra parte de datos se reservan como conjunto de prueba o validación, y se explicará en la siguiente fase.

Por otra parte, en los métodos de reglas de asociación, y siguiendo a [Hernández Orallo, et al. 2004], la propia regla nos proporciona varios métodos de evaluación objetivos para su validación, tales como soporte, confianza y elevación, que se explicarán en la siguiente fase.

2.4.4.- FASE DE EVALUACIÓN Y VALIDACIÓN

Para seleccionar y validar los modelos predictivos es necesario el uso de criterios de evaluación de hipótesis.

La fase anterior produce una o más hipótesis de modelos. Para seleccionar y validar estos modelos es necesario el uso de criterios de evaluación de hipótesis.

Es en esta fase donde se utiliza el conjunto de prueba o de test. El método de validación más básico es la validación simple, donde se reserva un porcentaje de la vista minable como conjunto de prueba. Este porcentaje suele variar entre el 5% y el 50%. Es deseable que la división de estos datos sea aleatoria.

En caso de tener una cantidad de datos muy moderada y no haber suficientes datos para el entrenamiento, se debe usar la técnica de la validación cruzada. Los datos se dividen aleatoriamente en dos conjuntos iguales y se utiliza el primero para entrenar y el segundo para validar. A continuación se construye el modelo con los datos que se habían utilizado antes para validar y se valida con los datos con los que se ha construido el anterior modelo. Por último se construye un modelo con todos los datos y se calcula un promedio de los ratios de error, así se estima mejor la precisión.

Habitualmente se utiliza una mejora de la validación cruzada que es la validación cruzada de *k*-pliegues. Los datos se dividen aleatoriamente en n grupos, utilizándose el conjunto n de prueba y los k-1 anteriores como conjunto de entrenamiento. Este proceso se repite n veces tomándose cada vez uno de los pliegues como conjunto de

prueba y los k-1 restantes como conjuntos de entrenamiento. Para cada uno de ellos se obtiene un error independiente y por último se combinan estos errores dando un promedio de ellos, obteniéndose el modelo con todos los datos.

Otra técnica cuando el conjunto de entrenamiento es muy escaso es el bootstraping. Esta técnica consiste en construir un modelo con todos los datos iniciales. Entonces se crean conjuntos de datos y se hace un muestreo de los datos originales con reemplazo. Se construye el modelo con cada conjunto y se calcula el error sobre el conjunto de test. El error final estimado para el modelo construido con todos los datos se calcula promediando los errores obtenidos para cada muestra.

Las medidas de evaluación de los modelos dependen de la tarea de Minería de Datos. En el caso de la clasificación, la precisión es el conjunto de pruebas clasificadas correctamente dividido por el número de instancias totales de la prueba. En el caso de que se trabaje con reglas de asociación, se tienen los parámetros de:

- **Cobertura o soporte**: Número de instancias a las que la regla se aplica y predice correctamente.
- **Confianza**: Proporción de instancias que la regla predice correctamente. Es decir, la cobertura dividida por el número de instancias a las que se le puede aplicar la regla.
- **Interés o Elevación**: Es la confianza dividida por el número de instancias del consecuente. Nos permite evaluar si existe una desviación estadísticamente significativa de la regla respecto al conjunto de instancias del consecuente. Es decir, mide cuanto condiciona el antecedente la aparición del consecuente.

La medida tradicional para evaluar un clasificador es el error, que es el porcentaje de instancias mal clasificadas (respecto al conjunto de test). Habitualmente para medir el error se utiliza su inversa que es la precisión (del inglés accuracy).

Por otro lado, tal y como indica [Hernández Orallo, et al. 2004], se debe tener en cuenta en la evaluación subjetiva:

Comprensibilidad: se trata de una medida completamente subjetiva y depende del evaluador, pero en el caso de reglas de asociación se pueden seguir algunas medidas simples, tales como no contemplar reglas con muchos antecedentes (o consecuentes). Las preferencias semánticas del usuario, o un mayor nivel de discretización simplificará la comprensión de las reglas.

Interés: según [Freitas. 2002] existen dos tipos de medidas de interés, las objetivas y subjetivas. Las subjetivas se basan en la novedad que aportan los resultados y su relevancia para el sujeto evaluador. Las objetivas son medidas que se obtienen de los propios resultados y tratan de hallar aquellos que presentan una tendencia

estadísticamente significativa. Pueden ser ejemplos de esto la entropía de los datos, distribución, o la elevación estadística mencionada en la fase anterior.

Aplicabilidad: un modelo debe ser aplicable a la predicción o la descripción de los datos para que su evaluación subjetiva sea positiva. Por ejemplo, es poco aplicable una regla cuyo antecedente, en la práctica, no se conozca casi en ningún caso.

2.4.5.- Fase de interpretación y difusión

La obtención de resultados aceptables depende de factores como: la definición de medidas de interés de conocimiento que permitan filtrarlo de forma automática, existencia de técnicas de visualización para facilitar la valoración de los resultados o la búsqueda del conocimiento útil dentro de ellos.

La experiencia en el análisis es un factor determinante. El despliegue del modelo a veces es trivial pero otras veces requiere un proceso de implementación o interpretación:

- El modelo puede requerir implementación.
- El modelo es descriptivo y requiere interpretación.
- El modelo puede tener muchos usuarios y necesita difusión: el modelo puede requerir ser expresado de una manera comprensible. Por ejemplo, como es este caso, solo es válido aquello que es interpretable desde el conjunto de reglas con pocos elementos en la regla.

Los modelos necesitan un mantenimiento:

- Actualización: Un modelo válido puede dejar de ser válido y requerir cambios. Por ejemplo, las empresas evolucionan su negocio.
- Monitorización del modelo para cambiarlo cuando se necesite adaptarlo.

El proceso de KDD necesita en muchas ocasiones realimentarse y reconsiderar decisiones tomadas anteriormente tales como incorporar más datos al análisis o transformación de variables.

2.5.- Taxonomía de las técnicas de Minería de Datos

Las tareas que aborda la Minería de Datos son fundamentalmente de dos tipos:

- **Predictivas**: Donde se pretende predecir el valor que tendrá en un futuro uno o más valores en función de los datos que se disponen hasta el momento.

- **Descriptivas**: Donde se realiza una labor de análisis acerca de los datos que se disponen, intentando describirlos y obtener información de ellos que mediante técnicas estadísticas normales sería complicado.

La clasificación de la Minería de Datos difiere de unos autores a otros:

Según [Joshi. 1997] la Minería de Datos cuenta con los siguientes componentes:

- **Clustering:** Donde se analizan los datos y se generan conjuntos de reglas que agrupan y clasifican los datos futuros.
- **Reglas de asociación:** Reglas que presentan ciertas relaciones entre un grupo de objetos de una base de datos. Un ejemplo de regla de asociación sería: "30% de las transacciones que contienen toallitas de bebé, también contienen pañales; 2% de las transacciones contienen toallitas de bebé". En este caso el 30% es el nivel de confianza de la regla y 2% es la cantidad de casos que respaldan la regla.
- **Análisis de secuencias:** Trata de encontrar patrones que ocurren con una secuencia determinada. Trabaja sobre datos que aparecen en distintas transacciones. "Muchos usuarios que han comprado X luego han comprado Y"
- **Reconocimiento de patrones:** Analiza la asociación de una señal de información de entrada con aquella o aquellas con las que guarda mayor similitud, de entre las catalogadas por el sistema. Se usan para identificar causas de problemas o incidencias y buscar posibles soluciones, siempre y cuando se disponga de la base de información necesaria en la que buscar.
- **Predicción**: Se busca establecer el comportamiento futuro más probable de una variable o un conjunto de variables a partir de la evolución pasada y presente de las mismas o de otras de las que dependen. Las técnicas asociadas a estas herramientas tienen ya un elevado grado de madurez.
- **Simulación**: Comparan la situación actual de una variable y su posible evolución futura.
- **Optimización**: Resuelve el problema de la minimización o maximización de una función que depende de una serie de variables.
- **Clasificación**: Permiten asignar a un elemento la pertenencia a un determinado grupo o clase. Se establece un perfil característico de cada clase y su expresión en términos de un algoritmo o reglas, en función de distintas variables. Se establece también el grado de discriminación o influencia de estas últimas. Con ello es posible clasificar un nuevo elemento una vez conocidos los valores de las variables presentes en él.

[Cabena, et al. 1998] presentan que la Minería de Datos se compone de cuatro grandes operaciones soportadas por algunas técnicas comúnmente usadas:

- Modelización predictiva: Que usa las técnicas de:
    - Clasificación
    - Predicción de valores
- Segmentación de bases de datos: Que usa técnicas de:
    - Clustering poblacional
    - Clustering por redes neuronales
- Análisis de relaciones: Que utiliza las técnicas de:
    - Descubrimiento de asociaciones
    - Descubrimiento de secuencias de patrones
    - Descubrimiento de secuencias temporales similares
- Detección de desviaciones:
    - Técnicas estadísticas
    - Técnicas de visualización

Según [Westphal, et al. 1998] los algoritmos de Minería de Datos pueden ser utilizados para alguna de las siguientes tareas:

- Agrupamiento (clustering) o segmentación: Se busca la identificación de tipologías o grupos en los cuales los elementos guardan similitud entre sí y se diferencian de los otros grupos.
- Asociación: Consiste en establecer las posibles relaciones entre acciones o sucesos aparentemente independientes. Así se puede reconocer cómo la ocurrencia de un determinado suceso puede inducir la aparición de otro u otros.
- Análisis de secuencias: Es un concepto similar al anterior, pero se le añade el factor tiempo.

Se pueden clasificar las técnicas de aprendizaje de la siguiente manera:

- Métodos inductivos: Son aquellos que partiendo de los datos iniciales y del conocimiento generado son capaces de construir modelos que a partir de los datos generen los resultados.

- Técnicas predictivas:
    - **Interpolación:** Genera una función continua sobre varias dimensiones
    - **Predicción secuencial:** En ella, las observaciones están ordenadas secuencialmente y se predice el siguiente valor de la secuencia.
    - **Aprendizaje supervisado:** En éstas técnicas cada observación, compuesta por muchos valores de atributos, incluye un valor de la clase a la que corresponde. Se genera o aprende un clasificador a partir de clases que se proporcionan. Es un caso particular de interpolación en el que la función genera un valor discreto en lugar de continuo.
- Técnicas descriptivas:
    - **Aprendizaje no supervisado**: En ellas, el conjunto de observaciones no tienen clases asociadas. El objetivo es detectar regularidades en los datos de cualquier tipo: agrupaciones de datos *parecidos* o próximos, contornos de delimitación de grupos, asociaciones o valores anómalos.
    - **Métodos abductivos**: Se pretende, partiendo de los valores generados y de las reglas, obtener los datos de origen. El objetivo es explicar la evidencia respecto a los hechos que se han producido, tal cual haría un investigador privado, que a partir de las consecuencias de los hechos y de ciertas reglas de comportamiento es capaz de averiguar los hechos iniciales.

Es importante marcar la diferencia entre lo que es el aprendizaje supervisado y el no supervisado.

- **Aprendizaje supervisado**: El experto define clases y provee ejemplos de cada una. El sistema debe obtener una descripción para cada clase. En algunos casos se puede proveer una sola clase y se daría como resultado que los ejemplos están o no están en la clase.
- **Aprendizaje no supervisado**: El sistema debe agrupar los conceptos. Se reciben los ejemplos y no se definen clases. Se tienen que observar las características y crear grupos en función de los criterios que decida el algoritmo. Se pueden establecer previamente el número de clases.

A partir de [Hernández Orallo, et al. 2004] y de otras fuentes consultadas, se puede establecer esta relación de técnicas en función del objetivo que se pretenda. Esta relación está muy incompleta, pero es lo suficientemente exhaustiva como para dar la sensación de la gran cantidad de técnicas que hay para cubrir cada objetivo.

Si el objetivo de la Minería de Datos es la **Interpolación o la Predicción Secuencial**:

- Si los datos con los que se trabaja son continuos (números reales):
    - Regresión Lineal:
        - Regresión lineal global (clásica).
        - Regresión lineal ponderada localmente.
        - Regresión logística.
        - Análisis de regresión por mínimos cuadrados.
    - Regresión adaptativa (muy usada en compresión de sonido y video):
        - Cadenas de Markov
        - Vector Quantization
        - Algoritmo MARS (Multivariate Adaptive Regression Splines)
    - Regresión No Lineal: Para ello se utilizan técnicas no algebraicas:
        - Redes neuronales
        - Árboles de regresión
        - Máquinas de Vector Soporte (con adaptaciones)
- Si los datos de trabajo son discretos no hay técnicas específicas, aunque se suelen utilizar técnicas de algoritmos genéticos o algoritmos de enumeración refinados.

Si el objetivo es la **clasificación** se encuentran una gran variedad de técnicas (es el campo donde existen más técnicas para poder aplicar).

- Perceptrón y Redes Neuronales Artificiales (Perceptrón multicapa)
- Radial Basis Functions (RBF)
- Máquinas Vector Soporte
- Árboles de Decisión. Su gama es muy amplia, aunque mencionarán algunos muy clásicos como ID3, C4.5 (o C5.0) o CART.
- Clasificadores Bayesianos
- Naive Bayes
- Center Splitting Methods

- Aprendizaje por inducción
- Aprendizaje por ejemplos
- Aprendizaje por observación y descubrimiento
- Algoritmos genéticos
- Métodos Pseudo-relacionales:
    - Supercharging
    - Pick-and-Mix.
- Métodos Relacionales:
    - Programación Lógica Inductiva
    - Programación lógico-funcional inductiva
    - SCIL

Si el objetivo es realizar aprendizaje no supervisado de **segmentación**, es decir, agrupar los elementos en grupos sin conocer previamente las clases, encuentran las siguientes técnicas:

- Jerárquico
- No jerárquico
- K-means
- Redes Neuronales de Kohonen
- Medias Estimadas (Estimated Means)
- Cobweb
- Redes Neuronales
- Árboles de decisión
- AUTOCLASS
- PCA: Principal Component Analysis
- Pairwise hierarchical clustering (clustering jerárquico por pares)
- Técnicas bayesianas

Una secuencia es el conjunto de datos que llegan en el tiempo. Por ejemplo: la secuencia de transacciones bancarias de un cliente, la secuencia de compras de un cliente o la secuencia de las constantes vitales de un paciente en urgencias. Si el objetivo es el análisis por flujos de datos o **análisis de secuencias temporales** (sequence mining).

- Aproximación a matriz de índice bajo
- Projective clustering
- Clustering por dividir y mezclar

Diferente a la anterior es la **minería de reglas de asociación** (association rule mining): Por su origen y principal utilización es la llamada reglas de asociación de cesta de la compra, ya que establece la relación entre los productos vendidos en una tienda. Estudia las relaciones dentro de una transacción que se producen de manera repetitiva. Si el objetivo es este se dispone de los siguientes algoritmos.

- Anti-monotono: Es el método original para la minería de reglas de asociación. Dada una base de datos de transacciones en las que cada transacción consiste en una lista de elementos, consiste en encontrar todos los conjuntos de elementos asociados a través de un proceso de generación de candidatos, recuento del soporte de cada candidato y poda de los candidatos con bajo soporte.
    - A priori
    - CARMA
    - GRI
    - DCP
    - Proyección en árbol
    - FP-Tree
    - GSP
- Basado en prefijo: El problema de los algoritmos anti-monótonos es que generan muchos candidatos sin soporte. Intentan eliminar la fase de generación de candidatos para aumentar la eficiencia, pero necesita encontrar y contar el soporte de los patrones en los datos. Es decir, para cada transacción busca patrones en una transacción y comprueba si tienen suficiente soporte.
    - FP-Growth

- Prefix-Span
- SPADE
- SLPMiner

- Bitmap: representa las transacciones como mapas de bits y las almacena en una estructura de tipo de mapa de bits. Esto puede reducir significativamente el espacio de almacenamiento, reducir los costos de computación, y permite que incluso aplicar algoritmos de compresión de imagen.
  - MAFIA
  - GenMax
  - SPAM

Cada uno de estos algoritmos está orientado a un tipo de escenario tales como grandes datasets de transacciones cortas, pequeñas bases de datos de muchos elementos, etc. Muchos algoritmos de entre los de minería de reglas de asociación están basados en A priori, pero con diversas mejoras.

La **minería de dependencias funcionales** se puede considerar como parte de la minería de reglas de asociación, aunque muchos autores la consideran independiente. Se diferencian de estos primeros en que consideran todos los posibles valores. Por ejemplo serviría para determinar si dado el intervalo de edad, el nivel de ingresos, el código postal, su estado civil, si puedo determinar si el cliente tiene vehículo. Las dependencias funcionales tienen que ver mucho con las reglas de asociación.

Algoritmos:

- Difference Collection
- Logical Multiplication
- Candidate Generation : Trabajan bien con un número limitado de atributos.
- TANE
- Dep-Miner
- FastFDs
- Hypergraph Transversal (interesante en el caso de tener muchos atributos)

A parte de los algoritmos de la fase de Minería de Datos, existen otros algoritmos que ayudan a simplificar ciertos problemas para conseguir preparar los datos de una forma

adecuada para la Minería de Datos. Entre ellos destacan los algoritmos orientados a la selección de atributos. Entre estos se encuentran:

- Algoritmos genéticos: Son los peores ya que tienen un coste computacionalmente alto sin ofrecer ventajas respecto a las aproximaciones envolventes.
- Aproximaciones por filtros: Filtran los atributos irrelevantes. Clasifican por puntuación la relevancia de los atributos, para luego seleccionar los atributos más importantes.
- Aproximaciones envolventes (wrapper): Suelen producir mejores resultados pero su coste computacional es mayor. Utiliza un algoritmo inductivo para establecer una ranking de subconjuntos de característica. Luego se comprueba su precisión. El algoritmo que se use dependerá del método con el que luego se aplicará la Minería de Datos, y ejemplos de estos son:
- Nearest-Neighbour
- Error Nearest-Neighbour
- Selección secuencia hacia atras (Backward Sequential Selection)
- Selección secuencia hacia adelante (Forward Sequential Selection)

Por lo tanto se puede deducir que lo primero que hay que establecer en un proyecto es el objetivo, ya que en función de este habrán habitualmente varias técnicas. Hay que conocer cómo se van a proporcionar los datos al algoritmo de Minería de Datos, ya que esto también condicionará el tipo de técnica que se elige. Además, también hay que conocer ciertas características de los datos, tales como si hay pocos casos, si están muy desequilibrados y si tienen ruido, ya que determinados algoritmos son más convenientes para tratar determinadas situaciones en los datos de entrada. Por eso, para cada área de investigación en la que se trabaje y para cada entorno de trabajo es muy probable que se adapte mejor un algoritmo o una combinación de algoritmos determinada.

3.- TÉCNICAS DE MINERÍA DE DATOS

3.1.- ÁRBOLES DE DECISIÓN

Los árboles de decisión son una de las formas más populares de Minería de Datos porque tienen una representación sencilla de problemas con un número finito (y a ser posible reducido) de clases. Además son modelos comprensibles y proposicionales [Hernández Orallo, et al. 2004]. Se entiende por proposicionales a algoritmos que aprenden modelos sobre una única tabla de datos y que no establecen relaciones entre más de una fila de la tabla a la vez ni sobre más de un atributo a la vez. Las condiciones se expresan sobre el valor de un atributo.

Se observa en la Figura 14 como a partir del valor de la variable X_8, si el valor es menor de 3.2 se continuará la toma de decisiones por la rama izquierda y si es mayor o igual se continuará por la rama de la derecha. A partir de aquí cada rama tiene una variable separadora con un valor de separación, y así sucesivamente formando un árbol.

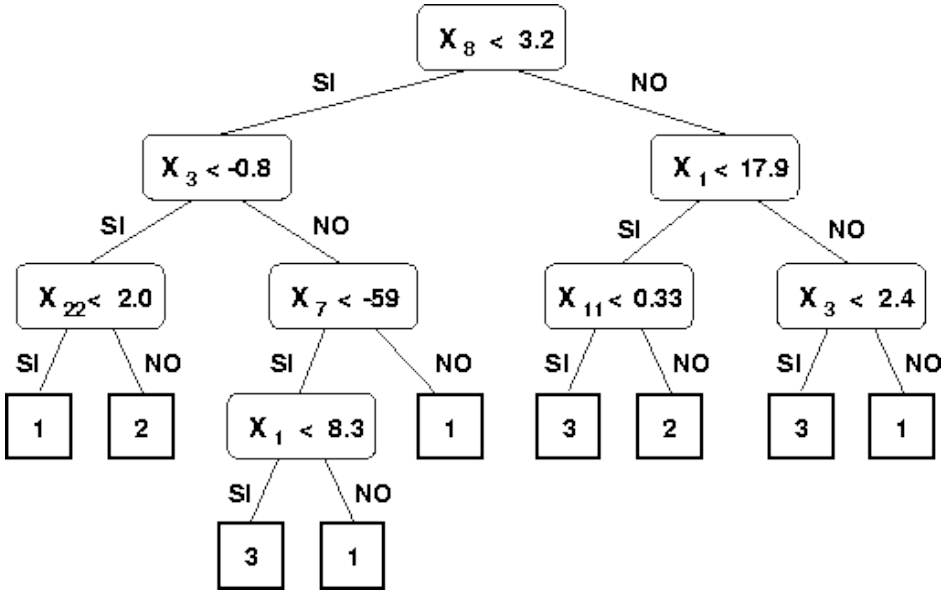

Figura 14: Árbol de decisión

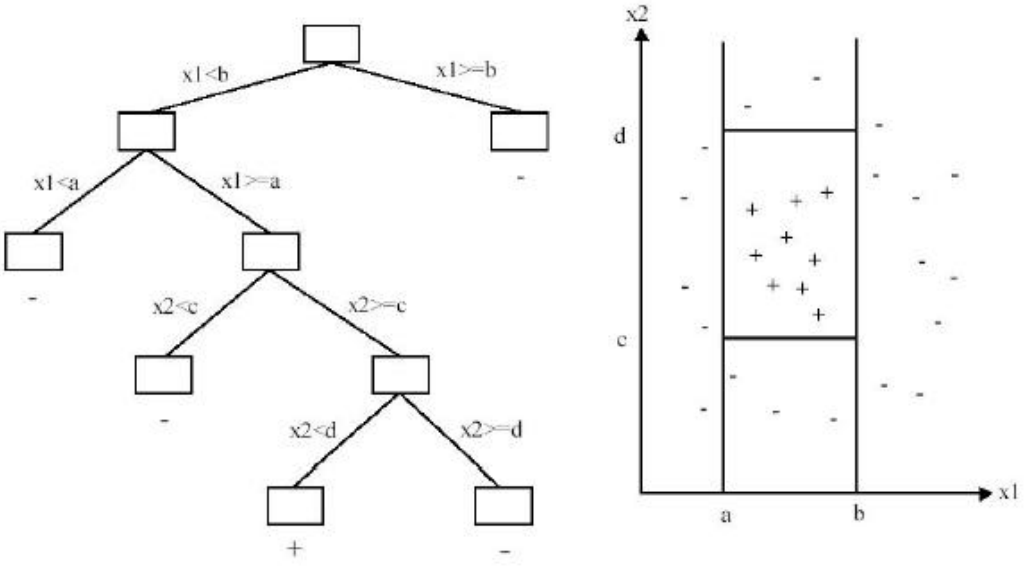

Figura 15: Segmentación en un árbol de decisión por divide y vencerás

En la Figura 15 se observa cómo se representa gráficamente las divisiones en el espacio de estados que produce un árbol de decisión.

Los árboles de decisión son una manera de representación del conocimiento muy antigua y utilizada en muy diversos campos.

La tarea mejor para la que los árboles de decisión se adecuan es la clasificación y se adecuan tanto a valores simbólicos como numéricos.

Los árboles están compuestos de nodos que nunca se unen en bucle cerrado. Se distingue entre nodo raíz (el primero), nodos intermedios y nodos terminales (aquellos que pertenecen al último nivel y a partir del cual ya no se hacen más particiones). Entre ellos se habla de nodo padre al antecesor y de nodos descendientes a los que dependan de él. Cada nodo tiene asociada una regla de modo que se sigue hacia uno u otro descendiente en función de la respuesta a ésta, hasta que se llega a un nodo terminal, donde se efectuará la clasificación.

Se utilizan para realizar la clasificación sistemas por partición. En la fase de entrenamiento, se genera un árbol que será utilizado posteriormente para clasificar cada uno de los casos que se disponen en función de los diferentes valores de sus atributos.

El desarrollo de los árboles de decisión en los algoritmos de partición es el método *divide y vencerás*. Para ello, en cada nodo se va interrogando a cada atributo y según

el valor del mismo, se va particionando el eje espacial que representa, de modo que va generando zonas en el espacio central, tal como ilustra la Figura 15.

Generalmente, en cada nodo se evalúa cada atributo según una constante. Aunque algunos árboles clasifican atributos con valores nominales o comparan dos o más atributos con otro, o utilizan alguna función que incluye uno o varios atributos [Witten, et al. 2000].

Podemos encontrarnos con:

- Árboles que predicen una clase determinada a partir de atributos numéricos, nominales o ambos, llamados árboles clasificadores.
- Árboles que predicen un valor numérico a partir de atributos numéricos a partir de atributos numéricos, nominales o ambos, llamados árboles de regresión.

El algoritmo genérico de un árbol de decisión viene determinado por los siguientes pasos:

- Asignar al nodo raíz todos los datos de entrenamiento (conjunto T)
- Si todos los elementos M nodo raíz pertenecen a la misma clase, finalizar el algoritmo. para ese nodo. (Ya no hace falta seguir partiendo la rama)
- Seleccionar una característica X on clases C_1, C_2, …C_N distintas, creando N nodos T_1, T_2, T_3,…T_N. descendientes del nodo padre T, de forma que T_i sea una partición de T en función de la regla derivada de la clase C_i.
- Para cada T_i hacer $T=T_i$ e ir al paso 2.

Este algoritmo es recursivo, ya que el punto 4 hace una llamada recursiva a ir al punto 2 para cada nodo.

Según [Hernández Orallo, et al. 2004] la técnica de partición realiza lo siguiente:

- Se crea un nodo raíz con S:= todos los ejemplos.
- Si todos los elementos de S son de la misma clase, el subárbol se cierra. Solución encontrada.
- Se elige una condición de partición siguiendo un criterio de partición (split criterion).
- El problema (y S) queda subdivido en dos subárboles (los que cumplen la condición y los que no) y se vuelve a 2 para cada uno de los dos subárboles.

Por ello, uno de los aspectos más importantes es el criterio de partición y los elementos más importantes en de un algoritmo de particionado o técnica de partición son:

- Particiones a considerar
- Criterio de selección de particiones
- Estrategia de poda

Estos aspectos son los que distinguen a los principales algoritmos que se verán más adelante.

Las particiones son un conjunto de condiciones exhaustivas y excluyentes.

El algoritmo de aprendizaje debe perseguir encontrar un buen compromiso entre expresividad y eficiencia. Por este motivo, la mayoría de algoritmos de aprendizaje de árboles de decisión permiten un número muy limitado de particiones.

Tipos de particiones:

- Particiones nominales: Consideran únicamente la partición de los distintos valores posibles del atributo.
- Particiones numéricas: Particionan generalmente con criterios tales como (x<=a, x>a). Muchos algoritmos ordenan todos los posibles valores eliminando repeticiones y después obtienen un valor intermedio entre las particiones y van probando las diferentes particiones. Si hay demasiado valores a probar se elige un subconjunto representativo de estos valores para así ganar en eficiencia.

En cuanto a la expresividad, se conoce como *"expresividad proposicional cuadricular"* aquella de particiones que solo afectan a un atributo y que produce particiones como las de la Figura 15.

Los criterios de selección cuyo objetivo es partir el conjunto de entrenamiento de la manera más disjunta posible, se basan en obtener medidas derivadas de las frecuencias relativas de las clases en cada uno de los hijos de la partición respecto a las frecuencias relativas de las clases en el padre. Por ejemplo, si en un nodo se tienen un 50% de ejemplos de clase A y un 50% de clase B, una partición que de cómo resultado dos nodos donde cada nodo tiene todos los ejemplos de una clase, esa partición es una excelente partición.

Pero, en caso de tener criterios de selección no muy ajustados y seguir el algoritmo anterior indefinidamente, se llega a una solución demasiado ajustada al conjunto de entrenamiento, presentando componentes muy específicos generados por outliers o ruido. Para subsanar este problema se utilizan criterios de parada o podar las ramas de los casos menos numerosos (del inglés pruning).

El primer paso para construir un árbol consistirá en seleccionar el atributo más idóneo. Este será el que genere hijos más puros, es decir aquel que tenga la mejor partición. Esta medida de pureza se llama medida de información o función de impureza. Asociada con un nodo de un árbol, representa la esperada cantidad de información que se debe necesitar para especificar si una nueva serie de atributos puede ser clasificada dentro de una clase u otra.

La bondad de una partición se define cono el decrecimiento de impureza conseguido en ella.

La clasificación por árboles tiene las siguientes ventajas:

- Puede ser aplicado a cualquier tipo de variables de predicción: continuas y categóricas.
- Los resultados son fáciles de entender e interpretar.
- No tiene problema de trabajar con datos perdidos.
- Hace automáticamente selección de variables.
- Es invariante a transformaciones de las variables predictoras.
- Es robusto a la presencia de *"outliers".*
- Es un clasificador no paramétrico, es decir que no requiere suposiciones.
- Toma en cuenta las interacciones que puede existir entre las variables predictoras.
- Es rápido de calcular.

Pero tienen las siguientes desventajas:

- El proceso de selección de variables es sesgado hacia las variables con más valores diferentes.
- Dificultad para elegir el árbol óptimo
- La superficie de predicción no es muy suave, ya que son conjuntos de planos.
- Requiere un gran número de datos para asegurarse que la cantidad de observaciones en los nodos terminales es significativa.
- Ausencia de una función global de las variables y como consecuencia pérdida de la representación geométrica.

Los algoritmos más comunes de árboles de decisión son descritos a continuación.

3.1.1.- CART (CLASSIFICATION AND REGRESSION TREES) O C&RT

[Breiman, et al. 1984] Es un algoritmo que realiza particiones binarias con una estrategia de poda basada en un criterio de coste-complejidad. Las particiones binarias son el resultado de evaluar una condición que tiene dos únicas respuestas. Se define el conjunto estándar de preguntas de la siguiente forma: Cada partición depende de solo un atributo:

- Si X_i es un atributo categórico con valores $\{c_1, c_2,\ldots c_n\}$, Se pregunta si X_i está entre un subconjunto de los valores categóricos, siendo la respuesta sí o no.
- Si X_i es un atributo continuo, Q incluye preguntas del tipo: ¿X_i <=v?, siendo v un valor real cualquiera. Para simplificar, CART toma v como el punto medio entre dos valores consecutivos de X_i .

Cada pregunta da lugar a una partición con una medida de impureza asociada. Esta medida de impureza es el método Gini [Breiman, et al. 1984]

$i_G(t) = \sum_j \sum_k p(j/t)p(k/t) = \sum_{j=1}^{J} p(j/t)(1-p(j/t))$

Si hay dos clases (J=2) presentes entonces:

$i_G(t)=2p(1-p)$

donde p es la proporción de observaciones en la primera clase.

El coeficiente de Gini se puede interpretar como un coeficiente donde se clasifica cada observación de un nodo a una clase j con probabilidad p(j/t) en lugar de clasificar todas las observaciones a la clase con mayor número de observaciones.

Luego se selecciona la mejor puntuación (la mayor reducción de impureza). Así, se obtiene un árbol muy grande (T_{max}) donde los nodos terminales son muy homogéneos o están asociados a pocos patrones. Cuando se ha obtenido el árbol T_{max} se procede a la poda mediante la construcción de una secuencia decreciente y anidada de árboles: $T_{max} = T_0 > T_1 > \ldots > \{t_i\}$ de manera que el árbol $\{t_i\}$ conste de un único nodo.

Cada árbol podado de T_{i+1} se construye seleccionando de entre todos los subárboles de T_i, el subárbol con una menor medida de coste-complejidad asociada. Luego se calcula sobre cada árbol la tasa de error de clasificación para un conjunto de patrones de prueba independiente del conjunto de patrones utilizado en el entrenamiento, y así seleccionar el árbol con menor tasa de error asociada.

3.1.2.- ID3 (INTERACTIVE DICHOMETIZER) O TDIDT (TOP-DOWN INDUCTION OF DECISION TREES)

[Quinlan. 1986] utiliza la entropía como función de impureza. La entropía es la medida de la incertidumbre o aleatoriedad que hay en un sistema, es decir, ante una

determinada situación, la probabilidad de que ocurra cada uno de los posibles resultados.

Construye árboles de decisión a partir de un conjunto de ejemplos que son tuplas compuestas por varios atributos y una única clase. El dominio de cada atributo de estas tuplas está limitado a un conjunto de valores.

Las primeras versiones del ID3 generaban descripciones para dos clases: positiva y negativa. En las versiones posteriores, se eliminó esta restricción, pero se mantuvo la restricción de clases disjuntas. ID3 genera descripciones que clasifican cada uno de los ejemplos del conjunto de entrenamiento.

ID3 es muy utilizado en aplicaciones de dominios médicos, artificiales y en el análisis de ajedrez.

El nivel de precisión en la clasificación es alto. Sin embargo, el sistema no hace uso del conocimiento del dominio. Muchas veces los árboles son demasiado densos, lo que lleva a una difícil interpretación. En estos casos pueden ser transformados en reglas de decisión para hacerlos más comprensibles.

Es capaz de trabajar con atributos discretos y continuos. Si los atributos son discretos, el nodo tendrá tantas ramas como valores posibles tome el atributo.

Como inconvenientes cabría destacar que con atributos numéricos no clasifica de manera muy adecuada. También, tiene el inconveniente de que hay una predisposición a favorecer a los atributos de muchos valores. Por estos motivos Quinlan desarrolló C4.5 como extensión de ID3.

3.1.3.- C4.5 (C5.0)

C4.5 [Quinlan. 1993] y su versión comercial C5.0 son una extensión de ID3 que permite trabajar con valores continuos para los atributos, separando los posibles resultados en dos ramas: una para aquellos $A_i <= N$ y otra para $A_i > N$. C4.5 es capaz de trabajar con ejemplos que contienen valores desconocidos y es tolerante a datos con ruido. Además, es capaz de generar reglas a partir de los árboles.

Este algoritmo genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente. El árbol se construye por primero en profundidad. El algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información.

Para cada atributo discreto, se considera una prueba con n resultados, siendo n el número de valores posibles que puede tomar el atributo. Para cada atributo continuo, se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos.

El criterio ganancia de información de C4.5 ha dado muy buenos resultados. Suele derivar en una preferencia en árboles pequeños (usando el criterio de la Navaja de Occam; el postulado más simple es el que tiene más probabilidades de ser correcto).

C5.0 no es de código abierto y presenta una mejora de rendimiento respecto a C4.5.

Los árboles de Decisión ID3, C4.5 y CART, que son los más conocidos en el mundo de la Minería de Datos, tienen las siguientes ventajas e inconvenientes:

Ventajas:

- Muy fácil de entender y de visualizar el resultado.
- Son robustos al ruido. Existen algoritmos de *post-pruning* para podar hojas poco significativas (que sólo cubren uno o muy pocos ejemplos).

Desventajas:

- Suele ser muy voraces. Al no hacer backtracking tiene el problema de los mínimos locales.
- Si el criterio de partición no está bien elegido, las particiones suelen ser muy ad-hoc y generalizan poco.

3.1.4.- SLIQ (SUPERVISED LEARNING IN QUEST) O QUEST

[Mehta, et al. 1996] fue desarrollado por IBM. Este algoritmo utiliza los árboles de decisión para clasificar grandes cantidades de datos.

Para la construcción del árbol utiliza como función de impureza el índice Gini, como CART. El uso de técnicas de pre-ordenamiento en la etapa de crecimiento del árbol, evita los costos de ordenamiento en cada uno de los nodos. Mantiene una lista ordenada independiente de cada uno de los valores de los atributos continuos y una lista separada de cada una de las clases. Un registro en la lista ordenada de atributos consiste en el valor del atributo y un índice a la clase correspondiente en la lista de clases.

El árbol se construye por el método primero en amplitud. Para cada uno de los atributos, busca en la lista correspondiente y calcula los valores de entropía para cada uno de los nodos de la frontera simultáneamente. A partir de la información obtenida, se particionan los nodos de la frontera, y se expanden para obtener una nueva frontera. El algoritmo necesita que cierta información resida en memoria permanentemente durante la totalidad de la ejecución del mismo. Esta cantidad reservada es proporcional a la cantidad de registros. Por ello, su punto débil reside en la limitación de memoria de la máquina.

SPRINT *(*Scalable PaRallelizable INduction of decision Trees*)* es una mejora del algoritmo que elimina los problemas de memoria del SLIQ.

3.1.5.- BACON

Utiliza algoritmos de análisis de datos para descubrir relaciones matemáticas entre datos numéricos. Ha redescubierto leyes como la ley de Ohm y la ley de Arquímides. Trabaja con datos numéricos.

Sus principales desventajas son:

- No considera el ruido en los datos, ni la inconsistencia o los datos incompletos.
- Considera que todas las variables son relevantes, y explora todas las soluciones posibles utilizando un grafo.

3.1.6.- CHAID

[Kass. 1980] significa "Chi-square automatic interaction detection". Es un derivado del THAID: "A sequential search program for the analysis of nominal scale dependent variables" El criterio para particionar está basado en $\chi 2$.

3.2.- REGLAS DE DECISIÓN

Estos métodos también son proposicionales. Son una generalización de los árboles de decisión. La principal diferencia con los anteriores es la filosofía del algoritmo que utilizan: partición (árboles de decisión) o cobertura (sistemas de reglas). Hay conjuntos de reglas que no se derivan de particiones y son capaces de clasificar la evidencia de una manera conveniente.

Existen métodos que generan conjuntos de reglas que podrían ser en algunos casos contradictorias para algunos ejemplos. Por ello, se da un orden a las reglas, y estas se aplican en ese orden. Estos métodos van generando reglas, una detrás de otra, mientras vayan cubriendo ejemplos de manera consistente y son los llamados métodos por cobertura.

En los algoritmos de aprendizaje de reglas el criterio de selección de condiciones es determinante y debe evaluar muy bien y a priori que particiones son mejores. Este criterio debe incluir la creación de reglas por defecto, los criterios de parada o de evaluación, los tipos de condiciones, etc.

El criterio de selección de las condiciones puede basarse en la pureza (la que elimine más contraejemplos) o en medidas derivadas de la información como Gain, o medidas que ponderen la precisión y el alcance (*precision and recall*).

Algunos algoritmos basados en cobertura son:

3.2.1.- AQ15

[Michalski, et al. 1986] genera reglas de decisión, donde el antecedente es una fórmula lógica.

Una característica particular de este sistema es la inducción constructiva *(constructive induction*), es decir, el uso de conocimientos del dominio para generar nuevos atributos que no están presentes en los datos de entrada.

AQ15 está diseñado para la generación de reglas fuertes, es decir, que para cada clase, se construye una regla que cubre todos los ejemplos positivos y ningún ejemplo negativo. El sistema soluciona el problema de los ejemplos incompletos o inconsistentes mediante un pre o post procesamiento. En el post procesamiento, además, se reduce de forma drástica la cantidad de reglas generadas mediante el truncamiento de reglas, el cual no afecta la precisión de las reglas obtenidas.

3.2.2.- CN2

[Clark, et al. 1989] y [Clark, et al. 1991] es una adaptación del AQ15. La gran desventaja del AQ15 es que elimina los ruidos mediante pre y post procesamiento y no durante la ejecución del algoritmo.

El objetivo del CN2 es incorporar el manejo de datos con ruido al algoritmo en sí. Combina entonces las técnicas de poda utilizadas en el ID3, con las técnicas de reglas condicionales utilizadas en el AQ15. El CN2 genera reglas simples y comprensibles en dominios donde los datos pueden tener ruido. Construye reglas probabilísticas, es decir, el antecedente en cada regla cubre ejemplos positivos de una clase, pero también puede cubrir ejemplos de otra clase en menor número. De esta forma no restringe el espacio de búsqueda únicamente a aquellas reglas inferibles a partir de los ejemplos.

3.2.3.- DBLEARN

[Han, et al. 1994] utiliza conocimientos del dominio para generar descripciones para subconjuntos predefinidos de una base de datos relacional. Su estrategia de búsqueda es *bottom up*. Usa conocimientos del dominio como jerarquías de valores de atributos y también hace uso de álgebra relacional. El conjunto de entrenamiento es una tabla de datos relacional con n-tuplas. Es un algoritmo simple, ya que utiliza solo dos operaciones de generalización para construir los descriptores. La generalización está orientada a los atributos, lo cual limita el conjunto de descriptores que pueden ser construidos. Tiene una complejidad en el tiempo de O(N logN), siendo N la cantidad inicial de tuplas.

3.2.4.- META-DENDRAL

[Lindsay, et al. 1993] representa el conocimiento como la estructura de una molécula (diferente a los anteriores). Busca generar reglas que puedan predecir dónde se romperá la molécula. El sistema ha sido exitoso para encontrar reglas de fragmentación desconocidas hasta el momento.

Sus problemas son:

- Usa una estrategia de búsqueda ineficiente, ya que genera muchas reglas de decisión que luego son eliminadas en la etapa de optimización.
- Es muy difícil encontrar heurísticas que guíen la búsqueda y no existen técnicas explícitas que ayuden a eliminar ruidos o a destacar casos especiales.

3.2.5.- APRENDIZAJE POR INDUCCIÓN

También nos encontramos con las técnicas de **aprendizaje por inducción**. Este es aprendizaje por ejemplos. Se dispone en la base de datos de ejemplos positivos y ejemplos negativos. Los ejemplos positivos generalizan, los negativos previenen que sea excesiva la generalización. Las reglas deben cubrir los ejemplos positivos y no cubrir los negativos. Estos sistemas de reglas de aprendizaje por inducción se pueden dividir en métodos clásicos y el aprendizaje por observación y descubrimiento.

Métodos clásicos:

El **aprendizaje AQ** se basa en la idea de cubrir progresivamente los datos de entrenamiento a medida que se generan reglas de decisión. Su esencia está en la búsqueda de un conjunto de reglas (conjunciones de pares atributo-valor o predicados arbitrarios) que cubran todos los ejemplos positivos y ningún ejemplo negativo. En lugar de dividir los ejemplos en subconjuntos, el aprendizaje AQ generaliza, paso a paso, las descripciones de los ejemplos positivos seleccionados [Michalski, et al. 1997].

El **aprendizaje "divide y vencerás"** particiona el conjunto de ejemplos en subconjuntos sobre los cuales se puede trabajar con mayor facilidad. En la lógica proposicional, por ejemplo, se parte el conjunto de acuerdo a los valores de un atributo en particular, entonces, todos los miembros de un subconjunto tendrán un mismo valor para dicho atributo. Dentro de este tipo de aprendizaje, se encuentra la familia TDIDT (*Top-Down Induction Trees),* que engloba al C45 que ya se ha visto

Aprendizaje por observación y descubrimiento:

El sistema forma teorías o criterios de clasificación en jerarquías taxonómicas, a partir de la inducción realizando tareas de descubrimiento. Es aprendizaje no supervisado y, como tal, permite que el sistema clasifique la información de entrada para formar conceptos. Existen dos formas en las que el sistema interactúa con el entorno:

- La observación pasiva, en la cual el sistema clasifica las observaciones de múltiples puntos del medio
- La observación activa, en la cual el sistema observa el entorno, realiza cambios en el mismo, y luego analiza los resultados.

3.3.- REDES NEURONALES

Se basan en la aproximación mediante algoritmos al funcionamiento de las neuronas humanas. La red de neuronas del cerebro forma un sistema de procesamiento de información masivamente paralelo. Las redes neuronales simulan las propiedades observadas en los sistemas neuronales biológicos que se caracterizan por su generalización y su robustez.

Una red neuronal se compone de unidades llamadas neuronas. Cada neurona recibe una serie de entradas a través de interconexiones y emite una salida. Las redes neuronales pueden ser consideradas como un intento de emular el cerebro humano. En general, las redes neuronales son representaciones de modelos matemáticos donde unidades computacionales son conectadas entre sí por un mecanismo que aprende de la experiencia, es decir de los datos que se han tomado.

Una RNA (Red Neuronal Artificial) se diseña y programa mediante un algoritmo. En las RNA se parte de un conjunto de datos de entrada significativo (conjunto de entrenamiento) y el objetivo es conseguir que la red aprenda automáticamente las propiedades del sistema.

Para ello se dispone de ejemplos positivos y negativos y se debe dividir los datos en tres grupos que no deben ser homogéneos. El más grande es el conjunto de Entrenamiento (del inglés Training Set) que nos sirve para calcular los valores de activación de las neuronas. Para comprobar que estos valores calculados son correctos se dispone del conjunto de Validación (del inglés Validation Set). Por último se utiliza un conjunto de Prueba (del inglés Test Set) para dar los parámetros de fiabilidad de la red neuronal.

La forma como dos neuronas interactúan es modelado por la función de red (función de activación). La neurona recoge las señales por su sinapsis (entradas) sumando todas las influencias excitadoras e inhibidoras. Si las influencias excitadoras positivas dominan, entonces la neurona da una señal positiva y manda este mensaje a otras neuronas por sus sinapsis de salida.

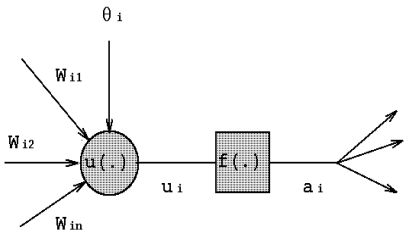

Figura 16: Representación de una neurona artificial

Tal como se puede observar en la Figura 16, las neuronas son representadas por círculos y cajas, mientras que las conexiones entre ellas son representadas por flechas. Los círculos representan variables observadas. El nombre de la variable va adentro del círculo. Las cajas representan valores calculados con una función de uno o más argumentos. El símbolo dentro de la caja indica el tipo de función de activación usada. El punto de partida de la flecha representa el argumento de la función a ser calculada en el punto final de la flecha. Cada flecha tiene usualmente un parámetro correspondiente a ser estimado.

Un perceptrón es un modelo de una neurona. En términos de redes neuronales, un perceptrón calcula una combinación lineal de entradas. Luego, una función de activación, la cual es por lo general no lineal, es aplicada a esta combinación lineal para producir una salida. Es decir, la salida y_j es

$$y_j = f_j\left(\sum_{inputs:i} w_{ij} x_i\right)$$

f_j representa a la función de activación y w_{ij} son los pesos. La red neuronal aprende los pesos de los datos.

Una RNA es un perceptrón multicapa. Esto le permite resolver problemas que no son linealmente separables, lo cual es la principal limitación del perceptrón simple.

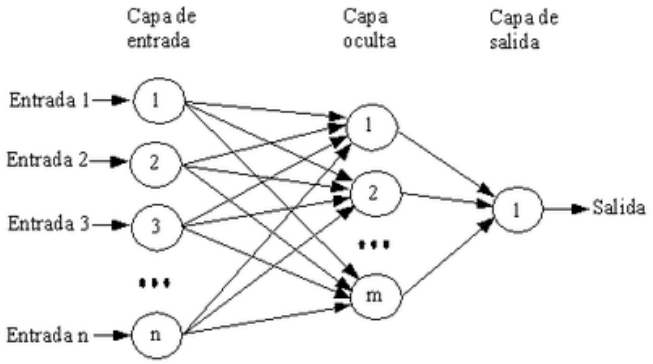

Figura 17: Representación de una red neuronal artificial

En la Figura 17 se observa que las capas en una red neuronal artificial pueden clasificarse en tres tipos:

- Capa de entrada: Constituida por aquellas neuronas que introducen los patrones de entrada en la red. En estas neuronas no se produce procesamiento.
- Capas ocultas: Formada por aquellas neuronas cuyas entradas provienen de capas anteriores y las salidas pasan a neuronas de capas posteriores.
- Capa de salida: Neuronas cuyos valores de salida se corresponden con las salidas de toda la red.

La información se almacena en un conjunto de pesos, no en un programa. Los pesos se deben adaptar cuando se le muestran ejemplos a la red. Las redes son tolerantes a ruido ya que pequeños cambios en la entrada no afecta drásticamente la salida de la red. La red neuronal es capaz de generalizar el conjunto de entrenamiento y así tratar con ejemplos no conocidos. La RNA son buenas para tareas perceptuales y asociaciones que es uno de los grandes problemas de la computación tradicional.

Las ventajas de las redes neuronales son:

- Aprendizaje: tienen la habilidad de aprender mediante una etapa de aprendizaje consistente en proporcionar a la RNA datos como entrada, a la vez que se le indica cual es la salida esperada.
- Auto organización: crea su propia representación de la información en su interior, sin necesidad de una programación explicita. Por ello, junto a la programación evolutiva se le llama computación flexible.
- Tolerancia a fallos: Almacena la información de forma redundante, ésta puede seguir respondiendo aceptablemente aún si se daña parcialmente.
- Flexibilidad: Puede manejar cambios no importantes en la información de entrada, como señales con ruido u otros cambios en la entrada. Por ejemplo, si

la información de entrada es la imagen de un objeto, la respuesta correspondiente es acertada incluso si la imagen tiene parámetros de luz ligeramente distintos o el objeto cambia ligeramente de posición.

- Tiempo real: es paralela, si se implementa con computadoras o en dispositivos electrónicos que utilicen dicha paralelización se pueden obtener respuestas en tiempo real, de la misma manera que el cerebro es capaz de procesar cantidades ingentes de información en paralelo sin un esfuerzo aparente.

3.4.- REDES BAYESIANAS

Las técnicas y métodos bayesianos son técnicas adecuadas para trabajar con incertidumbre. Son métodos utilizados como técnicas tanto descriptivas como predictivas. Como técnicas descriptivas se centran en el descubrimiento de relaciones de independencia y relevancia entre sus variables, y nos permite reflejar muchas relaciones interesantes. Como técnicas predictivas se utilizan los clasificadores bayesianos.

Los métodos bayesianos nos permiten inferir a partir de los datos induciendo modelos probabilísticos que serán usados más tarde para razonar. Nos permite calcular de forma explícita las probabilidades asociadas a cada una de las hipótesis posibles. Su gran problema es el coste computacional, por lo que se simplifican los problemas a partir de premisas. Aún así, se ha demostrado que son buenos clasificadores aún simplificando el problema, tal es el caso de Naive Bayes.

Las redes bayesianas han permitido simplificar el coste computacional , sin perder expresividad.

A partir de unas evidencias (causa) concluimos la probabilidad condicionada de las hipótesis (efecto). Por ello, se utiliza la notación P(h | e), donde h identifica una hipótesis de diagnóstico y e una evidencia. Se debe leer como la probabilidad de h dada la evidencia e.

Tal como se puede ver en la Figura 18, los algoritmos de propagación en redes bayesianas permiten hacer inferencias:

- De tipo abductivo: Por ejemplo “dado que el alumno ha respondido a ciertas preguntas, ¿cuál es la probabilidad de que conozca los conceptos? “
- De tipo predictivo: Por ejemplo, “dado que el alumno conoce ciertos conceptos, ¿cuál es la probabilidad de que responda correctamente a la pregunta?”

Cuando un nodo (o grupo de nodos) se instancia, la información se propaga por la red de forma que se calculan las probabilidades a posteriori de cada uno de los nodos dado el valor que haya tomado el nodo (grupo de nodos) instanciado.

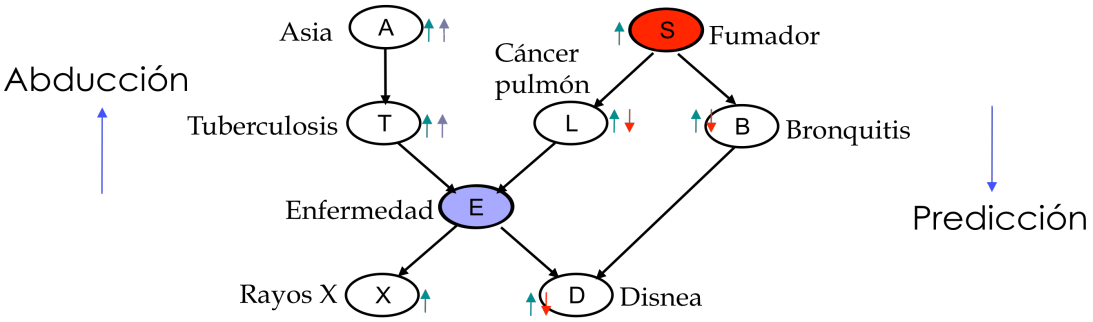

Figura 18: Enfoques abductivo y predictivo de una red bayesiana

Las redes bayesianas representan el conocimiento cualitativo del modelo mediante un grafo dirigido acíclico. Este conocimiento se articula en la definición de relaciones de independencia/dependencia entre las variables que componen el modelo. Estas relaciones abarcan desde una independencia completa hasta una dependencia funcional entre variables que componen del modelo. El uso de la representación gráfica del conocimiento bayesiano hace de esta herramienta un elemento de representación muy interesante. Las redes bayesianas modelan de forma cualitativa el conocimiento y expresan la probabilidad de que se produzcan las relaciones entre las evidencias (variables). Un arco directo entre variables expresa que hay una relación entre ambas. Este arco lleva asociada una probabilidad que es la probabilidad condicionada a que se produzca el nodo origen del arco.

Tal como se observa en la figura anterior, el problema del aprendizaje en redes bayesianas se define como encontrar el grafo dirigido a partir del conjunto de datos. A continuación, se expone un ejemplo que pretende mostrar las relaciones entre los diferentes agentes meteorológicos.

Tabla 5 Frecuencias de Agentes Meteorológicos

Lluvia	Nieve	Granizo	Tormenta	Niebla
5	0	0	0	0
1	0	0	0	0
5	0	0	1	0

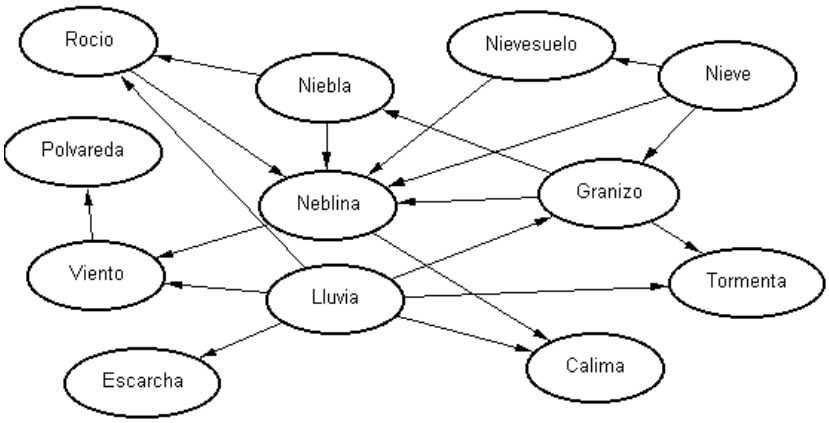

Figura 19 Ejemplo de red Bayesiana entre los agentes meteorológicos

Pero encontrar el conjunto de redes bayesianas con n nodos es super-exponencial, por lo que no es factible. Por ello es necesario la búsqueda heurística para guiar el problema de encontrar la mejor red bayesiana. A esta heurística encargada de calcular la adaptación de los datos a la red bayesiana se le llama medida bayesiana.

Se han planteado diversos algoritmos de búsqueda para la resolución del problema de la estructura de una red bayesiana. Según [Hernández Orallo, et al. 2004] estos algoritmos se pueden dividir en:

- Basados en métodos voraces y locales
- Basados en vecindad local y hill-climbing.

A continuación se muestran algunos clasificadores basados en redes bayesianas:

3.4.1.- CLASIFICADOR NAIVE BAYES (NB)

La simplificación propuesta por Naive Bayes es que las evidencias son independientes unas de otras, es decir, no están condicionadas entre si. Además, supone que las hipótesis son disjuntas, es decir que son excluyentes entre si y que cubren todas las posibles hipótesis. El teorema de Bayes simplificado en Naive Bayes según la h_i aplicado a un ejemplo con n evidencias y k hipótesis se escribe así:

$$P(h_i \mid e_1\, e_2 \ldots e_n) = \frac{P(e_1|h_i) * P(e_2|h_i) \ldots P(e_n|h_i) * P(h_i)}{\sum_{k=1}^{m} P(e_1|h_k) * P(e_2|h_k) \ldots P(e_n|h_k) * P(h_k)}$$

Naive Bayes tiene como problema que si hay pocos datos, es posible que alguna probabilidad condicional sea 0 (porque no ha aparecido un determinado valor de un atributo para una cierta clase).

Para ello se mejoró con el algoritmo Naive Bayes m-estimado. *Se* calcula un m-estimado de la probabilidad (con una fórmula) y el atributo considerado toma un cierto valor, *m es una* constante denominada "tamaño equivalente de muestra".

El algoritmo se quedaría así:

- Se pretende predecir la salida Y de aridad n_Y y valores v_1, v_2, … v_{ny}.
- Hay m atributos de entrada llamados X1, X2, … Xm
- Hay que dividir el dataset n_Y en pequeños datasets llamados DS1, DS2, DS_{ny}. Definir DS_i como los registros en los que $Y=v_i$
- Para cada DS_i, calcular el Estimador de Densidad M_i para modelar la distribución de entrada entre los registros $Y=v_i$.
- M_i se estima $P(X_1, X_2, \ldots X_m \mid Y=v_i)$
- Cuando un nuevo conjunto de entradas $(X_1 = u_1, X_2 = u_2, \ldots. X_m = u_m)$ se reciban para su evaluación se calculará el valor que maximiza la probabilidad de $P(X_1, X_2, \ldots X_m \mid Y=v_i)$.

$$Y^{\text{predicho}} = \underset{v}{\operatorname{argmax}}\, P(X_1 = u_1 \cdots X_m = u_m \mid Y = v)$$

3.4.1.1.- Arboles Bayesianos

[Pearl. 1988] y [Buntine. 1992] consisten en la aplicación de métodos bayesianos a la construcción de árboles de decisión.

3.4.2.- MÁQUINAS DE VECTOR SOPORTE

Se basan en un clasificador lineal muy sencillo (el que maximiza la distancia de los tres ejemplos, que son los vectores soporte), precedido de una transformación de espacio (a través de la función núcleo) para darle potencia expresiva.

El clasificador lineal que se usa, simplemente obtiene la línea (en más dimensiones, el hiperplano) que divide limpiamente las dos clases y además donde los tres ejemplos más próximos a la frontera están lo más distantes posibles.

Son eficientes (incluso para cientos de dimensiones), pues el separador lineal sólo tiene que mirar unos pocos puntos (vectores soporte) y puede descartar muchos que estarán lejos de la frontera.

Si los datos no son separables linealmente se aplica una función núcleo (del alemán "kernel") que suele aumentar el número de dimensiones de tal manera que los datos sean separables.

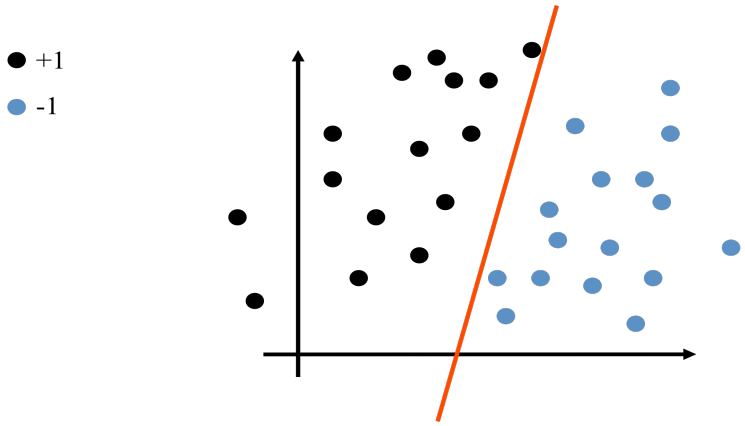

Figura 20: Ejemplo de hiperplano en máquinas de vector soporte

En la Figura 20 se observa como los ejemplos positivos y negativos son separados por un plano.

3.5.- Redes de Kohonen o SOM (Self-Organized Maps)

Son redes neuronales competitivas. Poseen una arquitectura de dos capas, capa de entrada y capa de salida. Tal como se puede ver en la Figura 21, el mapa autoorganizado está formado por una matriz rectangular de neuronas, de modo que las relaciones entre los patrones de entrada son mucho más fácilmente visibles en forma de relaciones de vecindad. Cada neurona sintoniza o aprende por sí misma a reconocer un determinado tipo de patrón de entrada. En el espacio de salida la topología queda preservada respecto a la de entrada, de manera que neuronas próximas en el mapa aprenden a reconocer patrones de entrada similares, cuyas imágenes, por lo tanto, aparecerán cercanas en el mapa creado. Este espacio de salida se representa por una capa discreta de neuronas artificiales o procesadores elementales, generalmente ordenados formando una matriz rectangular.

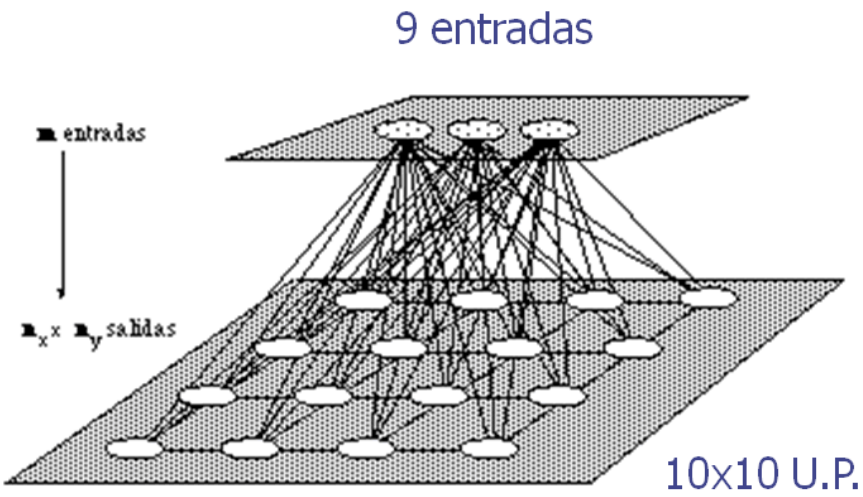

Figura 21: Red de Kohonen de 9 entradas y 10x10 salidas

Existen diversas posibilidades para la información que la salida de una red no-supervisada puede proveer:

- **Análisis de Componentes Principales (PCA)**
- **Familiaridad:** una salida única de tipo real podría indicar el grado de similitud que un patrón de entrada dado tiene con un patrón promedio típico concebido en el pasado por la red. Es decir, la red progresivamente puede aprender qué es lo típico.
- **Determinación de Prototipos:** semejante al caso anterior, pero aquí la red suministra como salida un prototipo o ejemplar representativo de la clase.
- **Agrupamiento:** un conjunto de salidas, de las cuales sólo se activa una a la vez, podrían indicarnos a cuál de diversas categorías pertenece un patrón de entrada dado. Las categorías involucradas han de ser determinadas por la red como consecuencia de las correlaciones en los datos de entrada.
- **Codificación:** la salida es una versión codificada de la entrada, generalmente en menos bits, pero se conserva la mayor cantidad de información relevante posible. Esto es muy empleado como técnica de compresión de información; en este caso, es necesario disponer de una red que ejecute el proceso inverso de descodificación.
- **Correspondencia de Rasgos:** en el caso de que las U.P. de salida posean una disposición geométrica fija (por ejemplo, un retículo cuadrado) y sean activadas una a la vez, entonces el sistema podría producir la correspondencia de los patrones de entrada con cada uno de estos puntos en el retículo. La idea en este

caso es lograr un mapa topográfico de rasgos en el cual patrones con rasgos parecidos siempre activen U.P. de salida vecinas en el retículo.

El objetivo de este tipo de redes es clasificar los patrones de entrada en grupos de características similares, de manera que cada grupo activará siempre las mismas salidas. Cada grupo de entradas queda representado en los pesos de las conexiones de la unidad de salida triunfante. La unidad de salida ganadora para cada grupo de entradas no se conoce a priori, y es necesario averiguarlo entrenando a la red.

Como se muestra en la Figura 22, tras la fase de entrenamiento, ciertas celdas quedarán activadas. En este caso, los ejemplos negativos están numerados en tono oscuro y los positivos en blanco. Las casillas que no tienen números no están activadas.

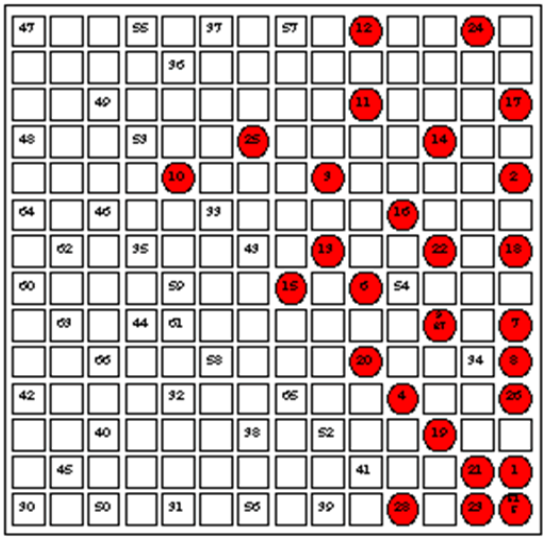

Figura 22: Celdas activadas en una red de Kohonen

Por último, se pretende fijar dos regiones. Se puede elegir por ejemplo marcar tanto la casilla con mayor activación como sus vecinos más cercanos y llegar por medio de esta heurística sencilla a un mapa como el mostrado en la Figura 23.

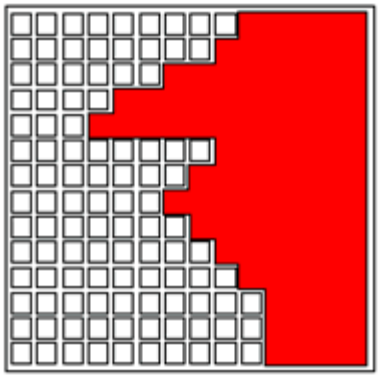

Figura 23: Regiones resultantes en una red de Kohonen

3.6.- PCA (PRINCIPAL COMPONENT ANALYSIS)

El propósito del análisis de componentes principales es transformar una matriz X de p variables en otra matriz Y de variables virtuales incorreladas ordenadas de mayor a menor varianza. En otras palabras, se pretende reducir la dimensionalidad de un dataset encontrando un conjunto de variables, más pequeñas que el original que mantenga casi toda la información de los datos originales.

La selección de ejes factoriales se lleva a cabo en orden de relevancia en términos de aportación de información (autovalores), de manera que cada eje que se determina debe aportar cada vez menor información. En la Figura 24 se observa que el primer autovalor de la nube de puntos corresponde al eje que maximiza la inercia de los puntos. Los siguientes autovalores corresponderán a ejes que irán aportando cada vez menor información.

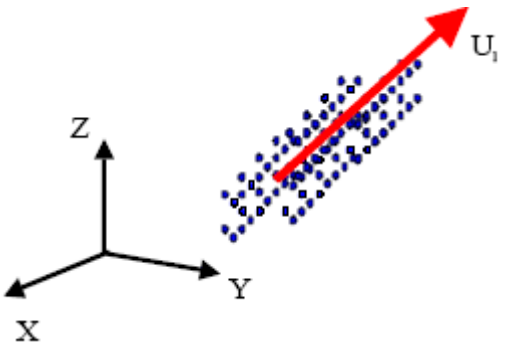

Figura 24. Primer autovalor obtenido de la nube de puntos.

Lo que se busca, por lo tanto, es seleccionar los suficientes autovalores de forma que la pérdida de información resultante no exceda del 10%.

3.7.- REGLAS DE ASOCIACIÓN

La minería de reglas de asociación (association rule mining), también llamada reglas de asociación de cesta de la compra por iniciarse para descubrir la relación entre productos vendidos en una tienda, trabaja con las asociaciones intratransaccionales (asociaciones entre elementos de la misma transacción) y no impone orden en los elementos que constituyen la transacción. Tratan principalmente atributos nominales. Además de la cesta de la compra son muy utilizadas en análisis de textos, búsquedas de patrones en páginas web, etc.

En las reglas de asociación se trabaja con dos medidas de calidad, la cobertura (del inglés support) y la confianza (del inglés confidence). La cobertura o soporte es el número de instancias que la regla predice correctamente. La confianza o precisión mide el porcentaje de veces que la regla se cumple cuando se puede aplicar.

Se podría considerar una regla de asociación de la siguiente manera

Si yogur griego Danoni y cervezas El Murcielago ENTONCES pañales Dodit Soporte=10%, Confianza=70%

Esta regla nos indica que había un 10% de casos donde aparecía una compra de yogur griego Danoni y de cerveza El Murcielago, y en 70% de ellas se había comprado además pañales Dodit.

[Ceglar, et al. 2006] hace una comparativa de los distintos algoritmos que han ido surgiendo en los últimos años. Se van a describir los principales algoritmos citados, agrupados por características de funcionamiento. Esto es: algoritmos anti-monotono, basados en prefijo, basados en bitmap.

3.7.1.- ALGORITMOS ANTI-MONOTONO

La característica principal de estos algoritmos es la forma de hallar los candidatos. Tratan de hallar aquellos candidatos con soporte superior a un mínimo dado, y a partir de esos candidatos generan reglas, calculando el soporte de éstas, realizando una poda de candidatos con bajo soporte y repitiendo el proceso.

El más característico y más conocido es el algoritmo “Apriori” que se describe a continuación.

3.7.1.1.- Apriori

El algoritmo Apriori fue propuesto por [Agrawal, et al. 1994]. Dada una base de datos de transacciones donde cada transacción consiste en una lista de los elementos que contiene, Apriori encuentra todos los conjuntos de elementos soportados utilizando un proceso de generación de candidatos, recuento del soporte y poda de estos candidatos.

Para ello hay que tener en cuenta las siguientes reglas para la generación de candidatos:

- **Principio de cierre hacia abajo (downward closure principle):** Antes de generar candidatos de longitud n se comprueba si los candidatos de longitud n-1 que son subelementos de este tienen soporte. Ya que si no tiene soporte un subelemento de este, no puede tener soporte el conjunto de elementos de longitud n, ya que el primero contiene al segundo y por tanto es más restrictivo. Es decir, el antecedente de la regla debe cumplir un soporte mínimo para generar otras reglas.

- Si hay dos o más elementos con soporte que comparten el mismo subelemento de longitud n-2, pueden ser combinados para generar un candidatos de longitud n. Por ejemplo, los conjuntos de elementos (0,1,2,3) y (0,1,2,4) comparten el prefijo (0,1,2) y pueden ser combinados formando el elemento (0,1,2,3,4). Es decir, si $A \rightarrow B$ y $C \rightarrow B$, entonces $A \cap C \rightarrow B$.

El algoritmo quedaría de la siguiente forma:

Repetir:

- Generación de candidatos: A partir de los conjuntos de elementos de longitud n-1, se generan nuevos conjunto de elementos candidatos de longitud n.

- Poda:. Se hace un recuento del soporte y todos aquellos candidatos que no tienen soporte se podan.

- Hasta que no se puedan generar nuevos candidatos.

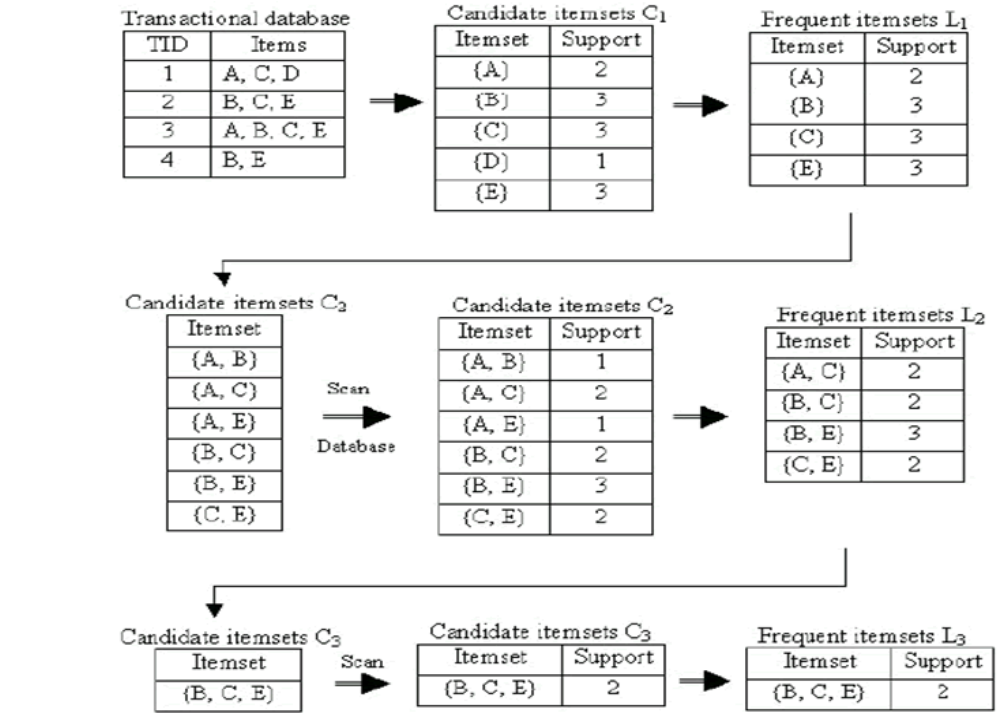

Figura 25: Ejemplo de conjuntos de elementos candidatos y frecuentes generados por Apriori

En el ejemplo de la Figura 25 se parte de las transacciones indicadas en el primer bloque y en la primera fase se calcula el soporte de cada elemento. En el tercer paso nos hemos quedado con los elementos con un soporte de 2 como mínimo. En el siguiente paso se combinan los elementos de longitud 1 formando pares de elementos y se vuelve a calcular su soporte, eliminando las que tengan soporte menor de 2. Luego se combinan en la siguiente fase las que tienen un elemento en común formando elementos de longitud 3, y se vuelve a calcular el soporte, y así sucesivamente.

Podemos indicar que sus principales ventajas son:

- Es un algoritmo muy sencillo de entender.
- Al tratarse de un algoritmo anti-monotono –el principio de downward-closure mencionado anteriormente-, se puede garantizar que si un item escogido es frecuente, sus subconjuntos también lo son, permitiendo realizar una búsqueda eficiente.

Como principales desventajas, citar:

- Es necesario decrementar mucho los valores mínimos de soporte y confianza para hallar relaciones. Esto provoca la aparición de gran cantidad de reglas.

- Es necesario trabajar con atributos categóricos, ya sean conjuntos o binarios. No se puede trabajar con atributos numéricos y deben ser discretizados.
- Se realizan múltiples recorridos de toda la base de datos para poder calcular el soporte y confianza de cada iteración.

3.7.1.2.- General Rule Induction - GRI

El algoritmo GRI propuesto por [Smyth, et al. 2002] es capaz de tomar valores numéricos y categóricos, siendo menos restrictivo que Apriori en este sentido. En cualquier caso, los consecuentes de las reglas inducidas deben ser categóricos. Utiliza una medida de interés J para determinar cómo de interesante es una regla y si puede ser interesante tratar de hallar más antecedentes a la regla, esto es, profundizar en ella o bien podarla. Esta medida se calcula como:

$$J(A \rightarrow B) = P(AB) \log(P(B|A) / P(B)) + P(A\neg B) \log(P(\neg B|A)/P(\neg B))$$

3.7.1.3.- CARMA

El algoritmo CARMA, propuesto por [Hidber. 1999], funciona básicamente igual que Apriori si tuviéramos todos los campos como entrada y salida. Además, permite ajustar el soporte de las reglas de forma continua, pero necesita un identificador por cada fila de datos.

Su principal ventaja radica en que se puede ir variando el soporte de las reglas a medida que se realiza el análisis. Es decir, los valores iniciales para la ejecución del algoritmo pueden ir cambiando a medida que éste se va procesando.

3.7.1.4.- DHP

El algoritmo DHP fue propuesto por [Park, et al. 1995] para mejorar el recuento de soporte, mediante lo que denomina *Direct Hashing and Pruning*. Se trata de usar una tabla de dispersión (*hashing table*) que almacena la frecuencia de los diferentes candidatos de longitud k+1. Así, no se almacenan aquellos candidatos con soporte inferior al mínimo. Después, el algoritmo va podando las ramas que ya no se usan en la construcción de posteriores reglas.

[Orlando, et al. 2001] realizó una mejora denominada DCP (Direct Count and Prune) que trata de optimizar el descubrimiento de candidatos incorporando técnicas de poda más eficientes e introduciendo recuento directo en la validación de candidatos. Mantiene una matriz m^2 para punteros a los candidatos y su soporte en vez de estructuras de árbol.

3.7.1.5.- Proyección en árbol

[Agarwal, et al. 2001] introducen el algoritmo de Proyección en Árbol. La principal innovación es trabajar con árboles lexiocográficos –las ramas siguen un órden- que necesitan bastante menos memoria que los árboles de dispersión. El soporte de los candidatos se cuenta proyectando las transacciones en los nodos del arbol, lo que mejora el rendimiento de recuento de transacciones con candidatos frecuentes.

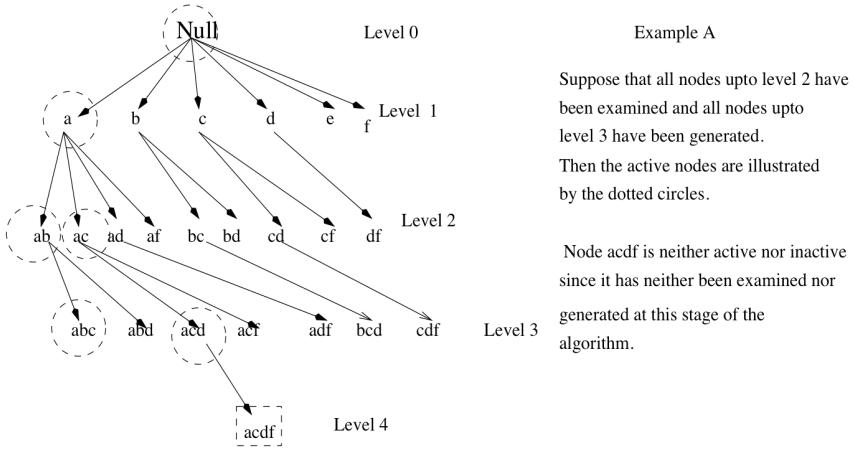

FIG. 1. The lexicographic tree

Figura 26 Arbol lexicográfico del algoritmo Tree Projection

En la Figura 26 podemos apreciar cómo se insertan nodos de forma ordenada. En estos nodos se proyectarían las transacciones de forma que se tendría un recuento de soporte. Las hojas son nodos terminales, los nodos en un rectángulo punteado aún no han sido evaluados, los nodos en un círculo punteado están activos y pueden generar más hijos.

3.7.2.- ALGORITMOS BASADOS EN PREFIJO

Los algoritmos basados en prefijo surgen para solventar el problema de la generación de muchos candidatos sin soporte en los algoritmos anti-monótonos. En cada iteración, los algoritmos anti-monotono tienen que calcular todos los soportes de todos los candidatos, resultando muchos de ellos con un soporte inferior al mínimo requerido. Esto provoca un gran sobrecoste. Los algoritmos basados en prefijo tratan de eliminar este sobrecoste. Entre ellos cabe destacar:

3.7.2.1.- FP-Growth

Propuesto por [Han, et al. 2004] trata de minimizar el coste del algoritmo al generar candidatos. El algoritmo básicamente consiste en ordenar por frecuencias los datos y construir un FP-Tree (Frequent Patern Tree) de prefijos. Una vez hecho esto se

proyectan las posibles combinaciones en otros árboles, creando árboles más pequeños con las ocurrencias de que se deseen evaluar.

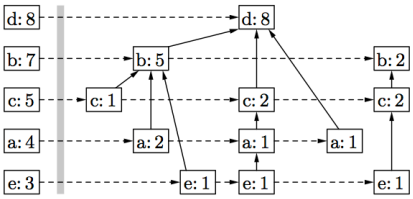

Figura 27 FP-Tree inicial en [Borgelt. 2005]

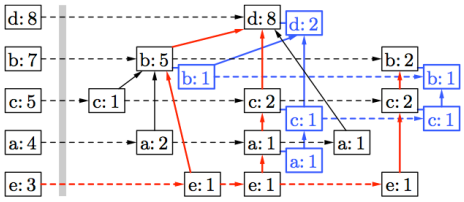

Figura 28 FP-Tree proyectado en [Borgelt. 2005]

3.7.2.2.- Prefix-Span

[Mortazavi-Asl, et al. 2004] proponen este algoritmo que usa una estrategia de "divide y vencerás", así como la proyección de prefijos frecuentes. El algoritmo proyecta recursivamente la base de datos y encuentra sufijos frecuentes. Una vez hecho esto vuelve a proyectar con ellos.

3.7.2.3.- SPADE

Este algoritmo propuesto por [Zaki. 2001] se basa en una lista de identificadores para cada uno de los valores posibles. Todas las ocurrencias de un valor aparecen como un objeto asociado a su identificador. Así, las ocurrencias de n valores pueden hallarse intersectando los objetos de n identificadores.

3.7.2.4.- SLPMiner

Propuesto por [Seno, et al. 2002] usa como idea principal decrementar el soporte exigido a los ítems frecuentes de los prefijos. Teniendo en cuenta que los prefijos de n candidatos siempre son un subconjunto de los de n-1 candidatos, estos prefijos siempre tendrán un soporte menor o igual a los prefijos de longitud n-1. Esto permite no podar ramas de n atributos en las que el soporte inicial no se cumple, cosa habitual a medida que las reglas crecen en antecedentes.

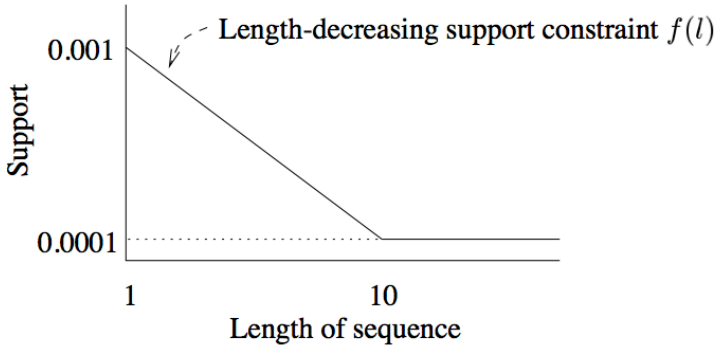

Figura 29 Decrecimiento del soporte en SLPMiner

3.7.3.- ALGORITMOS BASADOS EN BITMAP

Los algoritmos basados en bitmap representan las transacciones como mapas de bits y las almacena en una estructura de tipo de mapa de bits. Esto permite una reducción considerable del almacenamiento, costes de computación y permite incluso aplicar algoritmos de compresión de imágenes. Entre ellos destacan los siguientes.

3.7.3.1.- MAFIA

[Burdick, et al. 2005] proponen este algoritmo que se basa en generar un arbol lexicográfico en profundidad e ir podando las ramas que tengan soporte inferior al establecido como mínimo. Además, el almacenamiento de la información se realiza mediante mapas de bits verticales de cada transacción. Para calcular el soporte conjunto de dos elementos basta con aplicar la función AND a sus correspondientes bitmaps y sumar el bitmap resultante. El resultado es un conjunto de los patrones con mayor frecuencia en el conjunto de datos. Funciona especialmente bien con grandes cantidades de datos.

3.7.3.2.- GenMax

En [Gouda, et al. 2002] se propone este algoritmo que tiene algunas mejoras respecto a MAFIA haciendo uso de backtracking para determinar qué subespacios son más interesantes –tienen mayor probabilidad de contener un patrón de máxima frecuencia- y varias técnicas eficientes de podado.

3.7.3.3.- SPAM

[Ayres, et al. 2002] plantea este algoritmo centrado en secuencias. La peculiaridad de este algoritmo es que incluye en un árbol lexicográfico no sólo los ítems sino las secuencias en orden de aparición.

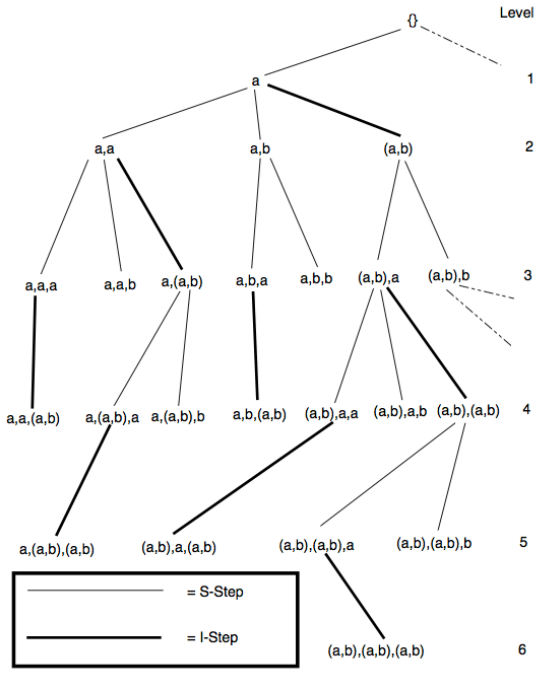

Figura 30 Árbol lexicográfico mostrando pasos Item y Secuencia en algoritmo SPAM

3.8.- ALGORITMOS GENÉTICOS

Son una clase particular de algoritmos evolutivos que se utilizan para la resolución de problemas de búsqueda y optimización. Se basan en técnicas inspiradas en la evolución biológica. Se aplican sobre una población representada de forma abstracta como *cromosomas*, que son la codificación de soluciones candidatas a un problema. La evolución comienza desde una población aleatoria y en cada generación, la *selección natural* elegirá que individuos son más aptos, modificándolos y mutándolos para la siguiente generación.

Los algoritmos genéticos no tienen un marco teórico genérico para aplicarlos a todos los problemas, sino que se adaptan específicamente a los problemas que van a resolver.

Para resolver un problema usando algoritmos genéticos se necesita poder:

- Representar soluciones: Tradicionalmente se representan mediante una cadena de bits. Los *cromosomas* son vectores numéricos que representan el rango de variación.
- Medir la calidad de cada solución con respecto al problema a resolver: Gracias a esta medida de calidad se sabe como seleccionar los candidatos mejores para la nueva generación. Se usa una función de selección.

Los cromosomas se emparejan y combinan, eliminándose los elementos del nuevo conjunto que tengan la calidad más pobre. El resultado se muta con una probabilidad dada, produciéndose esta mutación de manera aleatoria. Esta mutación sirve para evitar el estancamiento en máximos locales.

En general los algoritmos genéticos tienen el siguiente funcionamiento:

- Se crea una población inicial generando individuos aleatoriamente.
- Se repite hasta que se alcance el individuo óptimo o el número máximo de generaciones:
- *Asignar* un valor de supervivencia a cada miembro de la población.
- *Seleccionar* a un conjunto de individuos que actuarán como padres usando como criterio su probabilidad de supervivencia, por la función de calidad.
- *Emparejar* un grupo de padres para crear descendencia.
- *Combinar* la descendencia con la población actual para crear la nueva población.

Como ventajas en el uso de algoritmos genéticos tenemos:

- Es poco sensible a los mínimos locales, lo cual les confiere robustez.
- No dependen de las condiciones iniciales, debido a que se usa búsqueda estocástica y ésta hace al principio un gran número de intentos aleatorios.
- El tiempo de convergencia de los AG es predecible por la naturaleza paralela de la búsqueda estocástica.
- Funciona de forma paralela, por lo que pueden usarse en sistemas distribuidos para mejorar más la velocidad de búsqueda.

3.9.- MODELIZACIÓN ESTADÍSTICA PARAMÉTRICA

Trata tanto problemas con variables numéricas (regresión) como en problemas de variables categóricas, dando lugar a la clasificación supervisada. La regresión y la clasificación son dos de las tareas más habituales en Minería de Datos. Son los modelos que dependen de un número finito de parámetros. Este es el nombre para definir en estadística las técnicas de clasificación supervisada.

3.9.1.- ANÁLISIS DISCRIMINANTE

El análisis discriminante es un tipo de modelización estadística paramétrica. Su objetivo es encontrar reglas de asignación de individuos a una de las clases de una clasificación preestablecida.

Dado que las variables no siempre van a permitir separar las clases, se utiliza un criterio natural como maximizar la probabilidad de acierto y minimizar la de mala clasificación. Cada región queda asignada a partir de funciones monótonas de la probabilidad a posteriori que se llaman funciones discriminantes. Un ejemplo de función discriminante es el discriminante de Fischer que consiste en la búsqueda de la combinación lineal que maximiza la varianza de la separación entre las dos clases, respecto a la varianza dentro de las clases. La calidad de las funciones discriminantes depende de muy diversos factores tales aunque la precisión es la mejor medida de discriminación.

3.10.- MODELIZACIÓN ESTADÍSTICA NO PARAMÉTRICA

Es la que hace uso de métodos núcleo (kernel) para poder construir modelos más flexibles que en la modelización paramétrica, al hacer ajustes de manera local (ajustes a los diferentes puntos y no de manera conjunta al problema). Los modelos paramétricos presentan el problema de tener una estructura muy rígida para ajustarse s los problemas, por ello surge la estadística no paramétrica.

Según [Hernández Orallo, et al. 2004] el conjunto de técnicas no paramétricas es muy amplio, destacando de entre ellas:

- Ajuste local de modelos paramétricos: Se trata de varios o infinitos ajustes paramétricos tomando en consideración únicamente los datos cercanos al punto de estimación de la función.
- Métodos basados en series ortogonales de funciones: Se elige una base ortonormal del espacio vectorial de funciones y se estiman los coeficientes del desarrollo en esa base de la función de regresión. Los ajustes por series de Fourier y mediante wavelets son los dos enfoques más utilizados.
- Suavizado mediante splines: Busca la función que minimiza la suma de los cuadrados de los errores más un término que minimiza la falta de suavidad de las funciones.
- Técnicas de aprendizaje supervisado: Las rede neuronales, K-NN y los árboles de regresión.

3.10.1.- DEA (DATA ENVELOPMENT ANALYSIS)

Un modelo no paramétrico es DEA, muy utilizado en investigación operativa y economía para estimar las fronteras de producción. Se usa para medir empíricamente la eficiencia en la producción para las unidades de toma de decisiones (DMUs = Decission Making Units).

DEA asigna pesos a las entradas y salidas de una DMU que le produce una mayor eficiencia. Así, les asigna un peso de importancia relativa a las variables de entrada y

salida que refleja la importancia en que tienen que ser consideradas en una DMU. Al mismo tiempo, DEA deja las otras DMUs los mismos pesos y compara las eficiencias resultantes con aquella DMU que tiene el foco. Si la DMU con el foco parece ser tan buena como las otras, recibe una puntuación de máxima eficiencia, pero si alguna otra DMU parece mejor que la del foco, los pesos se calculan para ser más favorables a la del foco. Luego recibirá una puntuación de eficiencia menor que el máximo.

CEU
Universidad
Cardenal Herrera

4.- Minería de Datos en el Análisis Periodístico

4.1.- El Análisis de Datos en Periodismo

En cualquier ámbito del conocimiento, el análisis de los datos es una necesidad para el estudio de los hechos observados. En las ciencias sociales como el periodismo, esta tarea es más ardua si cabe, debido a la subjetividad y creatividad de los datos tratados.

El periodismo es un gran usuario y consumidor de datos: recopila grandes cantidades de información relacionada, la procesa de muchas formas -periodismo informativo, de opinión, gráfico, etc.- y la almacena en sistemas documentales. La aplicación de técnicas de Data Mining en periodismo puede obtener un valor agregado de la información que de otra forma pasaría desapercibida.

Una búsqueda sobre los principales ámbitos del análisis de datos en periodismo puede llevar a resultados diversos: desde la aplicación de técnicas de Minería de Datos para optimizar las ventas de periódicos, pasando por la minería de textos –Text Mining-, la extracción automatizada de opiniones o intencionalidad, la clasificación automática de textos, hasta el análisis de datos estadísticos.

En el ámbito del análisis de datos en periodismo, [Walter Lima. 2008] distingue dos vertientes bien diferenciadas: Text Mining y Data Mining. Estos dos enfoques pueden ser implementados de forma independiente, si bien ambos son complementarios y pueden hallarse juntos en un mismo estudio.

A continuación se describen ambos enfoques y posteriormente se describirán algunas referencias notables sobre la temática. De entre ellos, es destacable el caso [Colle. 2002] por ser una referencia en español que trata específicamente la temática del presente estudio y que se analizará más detenidamente.

4.2.- El Enfoque del Análisis de Contenido: Text Mining

En la aplicación de algoritmos de Minería de Datos es necesario proporcionar como entrada datos estructurados y preparados para un algoritmo concreto. Por ello, se utilizan procesos ETL –Extracción, Transformación y Carga- para preparar los datos para el modelado concreto con un algoritmo.

Sin embargo, tal y como argumenta [Weiss. 2005], la información que se proporciona habitualmente para la Minería de Datos dista mucho de ser un conjunto estructurado de datos, entendido como conjuntos de datos categóricos, o datos numéricos.

En el ámbito del presente estudio, puede comprenderse con facilidad la problemática de dicha situación. El análisis de textos periodísticos tiene como datos textos escritos por seres humanos, que no pueden ser tratados directamente por los algoritmos de

Minería de Datos. Se hace entonces necesario el uso de técnicas de extracción de la información de los textos a estudiar. Estas técnicas se conocen como Text Mining.

Según [Francis. 2006], se entiende por Text Mining:

> *"Text mining refers to a collection of methods used to find patterns and create intelligence from unstructured text data."*

Es decir, las técnicas de Text Mining tratan de obtener información de datos no estructurados en forma de textos escritos.

4.3.- Minería de Datos sobre Información Periodística

Las necesidades de análisis en el periodismo como ciencia social han llevado a diversos autores a tratar de desarrollar una metodología de análisis para los textos periodísticos. Algunas de éstas se describirán más adelante en el apartado *6.2.-* El Análisis Periodístico en Prensa Escrita. Estas metodologías proveen al analista de datos formalmente estructurados que sí pueden ser interpretados por algoritmos de Minería de Datos.

Habitualmente es posible encontrar textos que realizan una aproximación estadística al problema, entre ellos los del grupo de investigación periodística del presente proyecto -[Rodríguez Luque, et al. 2011], [Paricio Esteban, et al. 2010], [Paricio, et al. 2010], [Núñez-Romero Olmo. 2009], [Rodríguez Luque. 2009]-. Estas aproximaciones estadísticas son la antesala de la Minería de Datos en el análisis de la información periodística.

Como ya se ha comentado, es necesario obtener una representación formal y fuertemente estructurada de los datos para poder realizar el análisis de datos mediante algoritmos de Data Mining. Una vez obtenida esta representación, y dependiendo de las necesidades del análisis, se pueden aplicar las técnicas de Minería de Datos pertinentes. Los problemas que afrontan los textos suelen ser:

- Clasificación de textos: tratar de clasificar un texto por su contenido en algún tipo de categoría predefinida. Para ello se suelen usar algoritmos de redes neuronales, clustering, modelos bayesianos, etc.

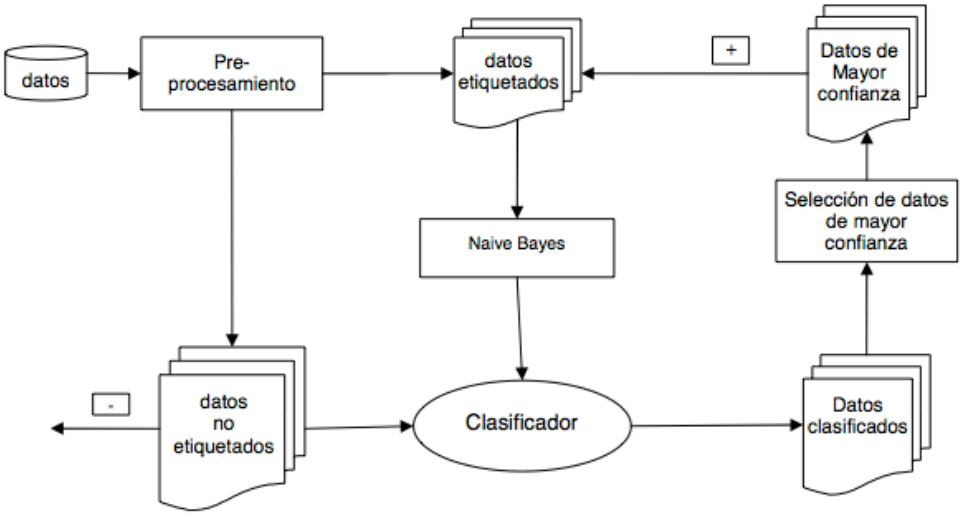

Figura 31 Método semi-automatizado de clasificación de textos Naive-Bayes en [Araujo Arredondo. 2009]

- Predicciones: algunos textos tratan de predecir la mejor ubicación de publicidad en periódicos, o las ventas de los mismos. Para ello se usan algoritmos predictivos y árboles de decisión.

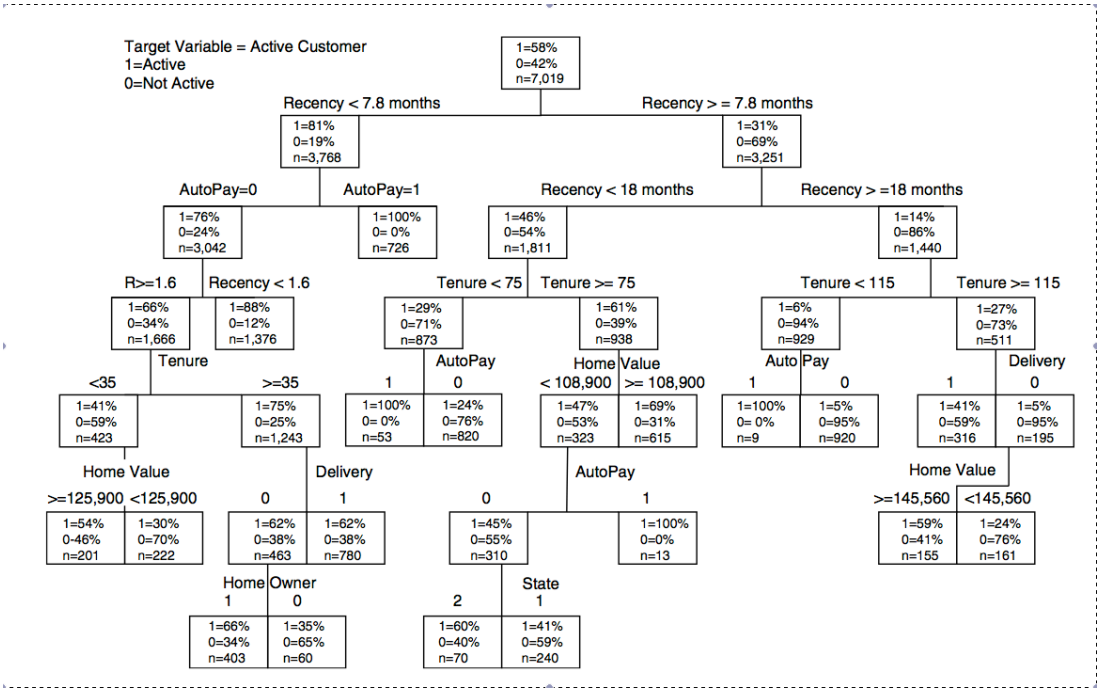

Figura 32 Árbol de Decisión en [Gunnarsson, et al. 2007]

- Extracción de información novedosa: otros estudios tratan de hallar relaciones entre variables que puedan resultar extrañas o novedosas respecto al conocimiento previo del problema. Para ello se utilizan técnicas de visualización gráfica y reglas de asociación.

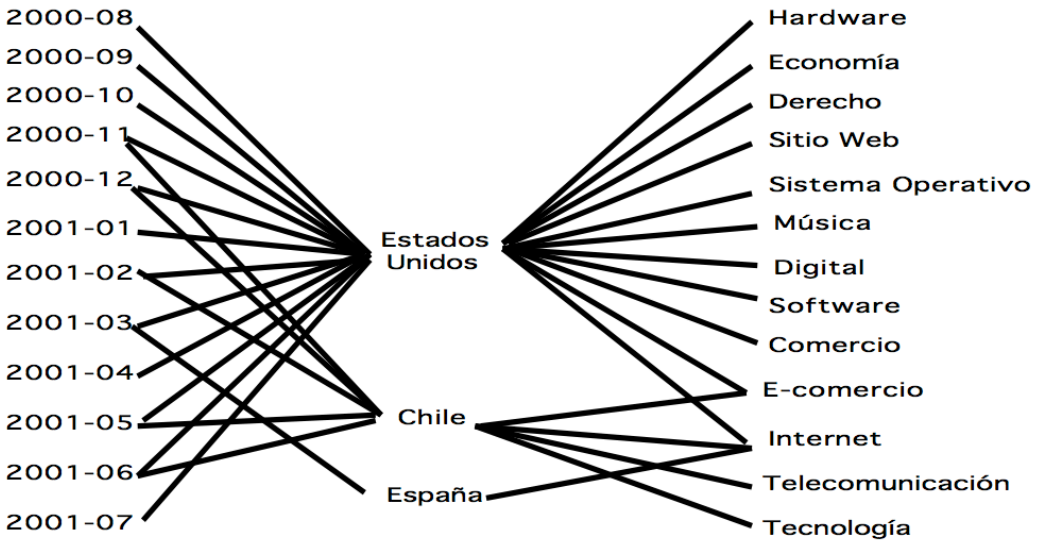

Figura 33 Análisis Gráfico de la tríada fecha-lugar-tema en [Colle. 2002]

4.4.- REFERENCIAS EN TEXT MINING Y DATA MINING SOBRE ANÁLISIS PERIODÍSTICO

Sin ánimo de ser exhaustivo, este apartado trata de mostrar algunas referencias de Text Mining y Data Mining relacionadas con el análisis periodístico y el periodismo en general, para poder proporcionar una visión general del estado de la materia de la Minería de Datos en el ámbito del periodismo.

[Alexa. 1997] trata de recapitular las metodologías de trabajo para la extracción automatizada de la información de textos de ciencias sociales, en el momento de la publicación de su artículo.

[Hong, et al. 2002] utiliza mapas cognitivos y redes neuronales para predecir los tipos de interés basándose en conocimiento previo, una base de datos financiera y el conocimiento extraído de las noticias financieras en Internet. Destaca en este estudio el uso de diferentes fuentes de información, así como el análisis de fuentes directas –noticias financieras en Internet-.

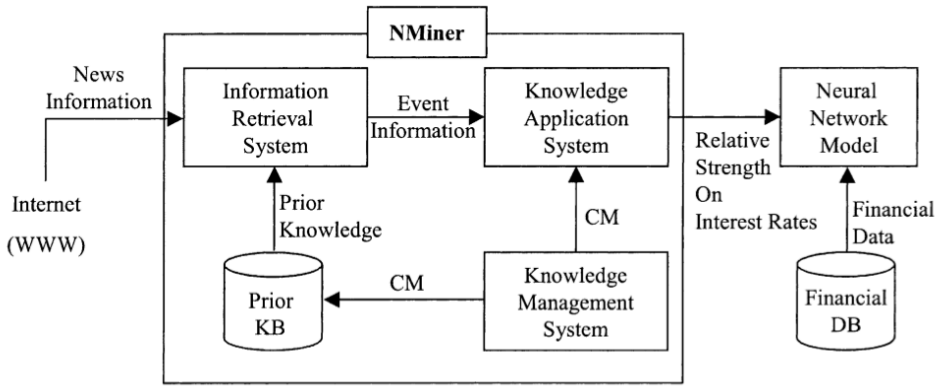

Figura 34 Esquema del modelo de Minería de Datos utilizado en [Hong, et al. 2002]

[Chung, et al. 2003] proponen usar clustering para la agrupación de noticias aparecidas en Internet por su temática. Se centra en el hecho de que las noticias habitualmente tienen precedentes y sucesores relacionados (por Ej. el rescate de un menor secuestrado probablemente estará precedido por la noticia de su secuestro). Además, proponen una forma de categorizar las noticias generando una ontología propia de cada tema, para evitar confusiones con términos idénticos de otras temáticas.

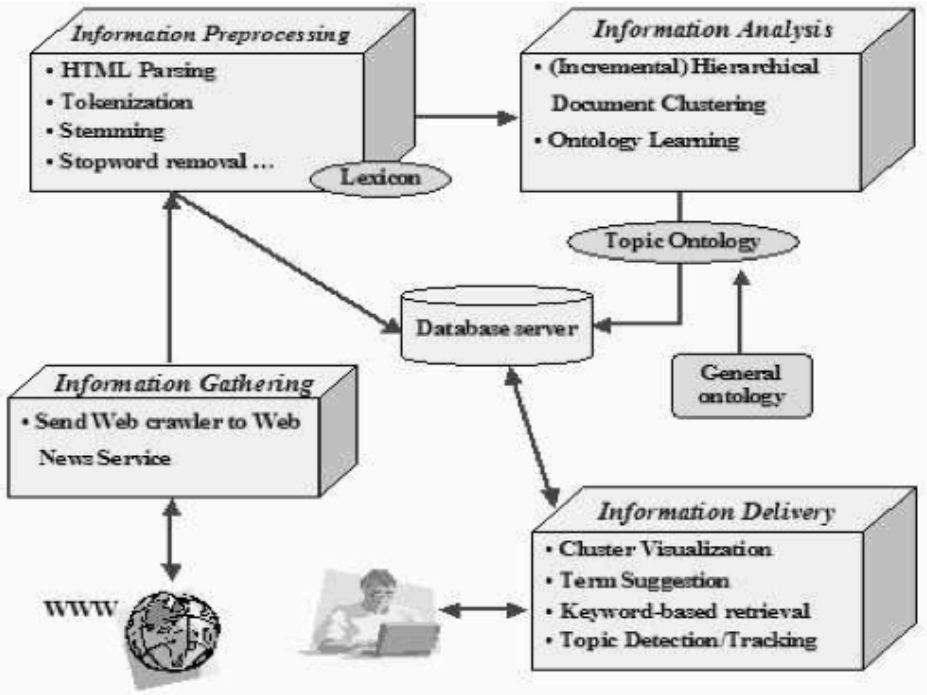

Figura 36 Modelo propuesto en [Chung, et al. 2003]

En [Clifton. 2004] se propone la técnica de Text Mining TopCat, que automatiza la búsqueda de información relevante dentro de textos, clasificándolos así por su temática. Utiliza para ello análisis de frecuencias y clustering. En el caso de estudio mostrado, TopCat identifica patrones en una base de datos de 60.000 registros de noticias.

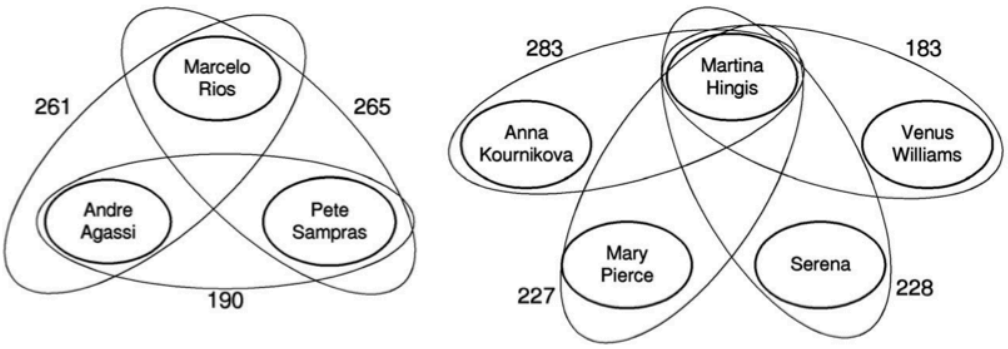

Figura 35 Agrupaciones de la categoría "Tenis" en [Clifton. 2004]

[Bordignon, et al. 2004] es una referencia en español que usa un clasificador Naive Bayes con un corpus de noticias extraído de la USENET. El principal objetivo del estudio es observar si existen diferencias en el comportamiento de este clasificador comparando los resultados que se obtienen utilizando como datos noticias en español y en inglés.

[Liang, et al. 2005] es un estudio centrado en la predicción de valores bursátiles en EEUU y Europa mediante redes neuronales utilizando noticias del mercado de valores. El estudio contiene una parte sobre procesamiento del lenguaje natural y cómo cuantificar la aparición de expresiones, palabras, etc. en una noticia para la predicción objetivo.

En [Norvag. 2005] los autores se centran en la extracción de información de sitios de noticias Web –periódicos digitales- y la problemática que supone la no estandarización de la estructura de los mismos a la hora de extraer información relevante de las noticias. Los autores profundizan características propias de los diarios digitales, tales como hipervínculos, *frame*s, imágenes, HTML, etc. para extraer información útil que les permita discriminar contenidos. Utilizan algoritmos de clusterización para agrupar noticias por temáticas.

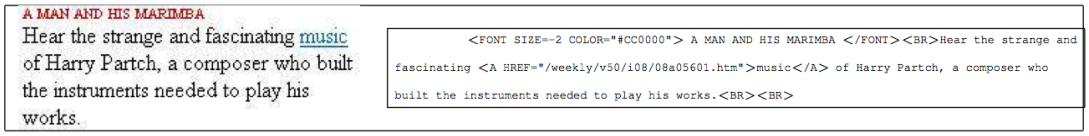

Figura 37 Uso de html para encontrar el cuerpo de una noticia en [Norvag. 2005]

[Gunnarsson, et al. 2007] es un caso de estudio en el que se aplican técnicas de Minería de Datos a los datos de un periódico del medio oeste de EEUU para predecir qué clientes pueden cambiar de periódico y prevenirlo. En la Figura 32 se mostró un árbol de decisión mostrado en este caso.

En [Walter Lima. 2008] se discuten los retos de representar en una estructura formal el conocimiento que un periodista utiliza para establecer la importancia de una noticia. Para ello, se introducen conceptos de Text Mining y además, se propone la Minería de Datos como una metodología empírica para obtener evidencias no triviales en las valoraciones de las noticias.

[Reilly. 2009] presentan 3 casos de estudio en el que se analiza la automatización del análisis de textos periodísticos y otras fuentes documentales accesibles en Internet, con la finalidad de ser clasificados por documentalistas. El estudio está centrado en textos políticos y económicos.

[Araujo Arredondo. 2009] es una referencia en español que se centra en la clasificación de textos de opinión utilizando un método semi-supervisado centrado en un algoritmo Naive-Bayes. Se puede ver el modelo propuesto en la Figura 31.

4.4.1.- Un caso de estudio de Data Mining en Análisis Periodístico

En el apartado anterior se han mostrado varios artículos centrados en el Text Mining y la clasificación de textos. Algunos de ellos trataban de predecir la fidelidad de los lectores de un periódico o el valor de las acciones en base a las noticias de los medios digitales. [Colle. 2002] es un libro sobre el uso de herramientas de Minería de Datos en español en el que el autor se apoya en un caso de estudio para mostrar su aplicación. Se ha seleccionado este caso debido a la similitud con el presente estudio, ya que, al contrario que la mayoría de textos consultados, no se centra en la extracción de la información de textos periodísticos, sino en el análisis de datos obtenidos de estos textos.

En el mencionado texto el autor se centra en un caso de estudio de 1.766 noticias sobre tecnologías digitales de comunicación aparecidas en medios periodísticos. Su modelo de datos se compone de: Implicados, Lugar, Fecha, Fuente y Tema.

En el texto, se realiza el proceso de la Minería de Datos en varias fases. En la primera, el autor revisa la consistencia de los datos, agrupa el atributo Implicados en clases para reducir su dispersión y prepara los datos para obtener la vista minable.

En una segunda fase realiza un análisis de frecuencias y un estudio de predictibilidad de las variables utilizando el coeficiente lambda de Goodman y Kruskal.

En una tercera fase se realiza un análisis de concurrencias internas, es decir, entre los valores de una variable.

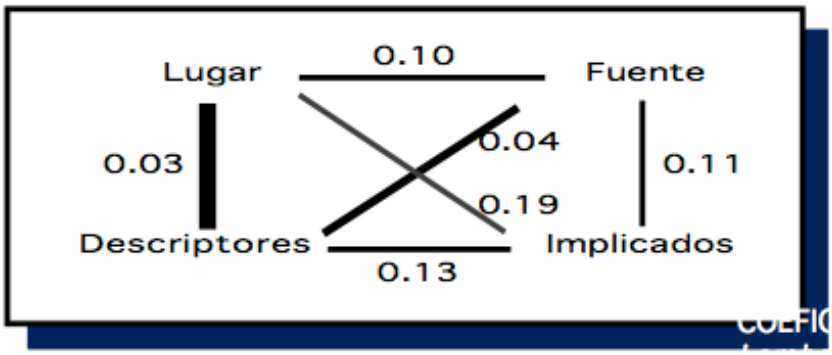

Figura 38 Predictibilidad de las variables en [Colle. 2002]

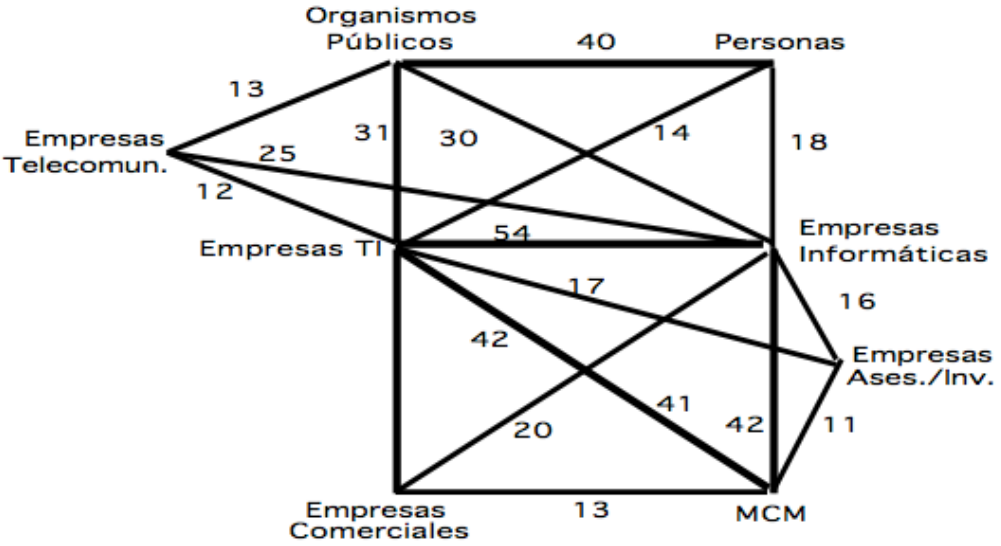

Figura 40 Coocurrencias entre valores de "Implicados" en [Colle. 2002]

En una cuarta fase el autor analiza las relaciones entre distintas variables, dos a dos.

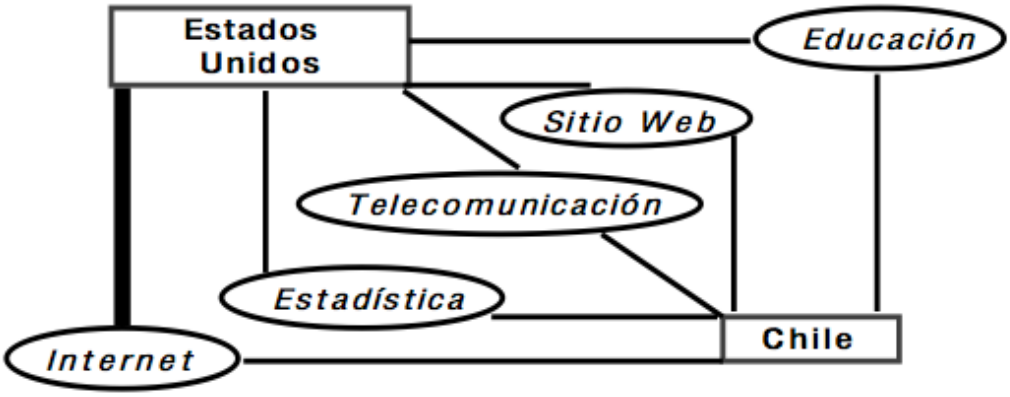

Figura 39 Coocurrencias lugar y tema en [Colle. 2002]

En una quinta fase el autor trata de hallar relaciones entre tríadas de variables. Un ejemplo de este análisis se puede observar en la Figura 33.

Por último, el texto analiza los datos mediante representaciones gráficas multidimensionales.

Como conclusión el autor indica, entre otras cosas, la dificultad que ha supuesto la evaluación de los resultados, debido a la gran cantidad de datos que las herramientas que utiliza han generado. Es decir, que no se han aplicado medidas de interés para evaluar los resultados y descartar aquellos fácilmente predecibles o no interesantes.

Si bien en el estudio se generan gran cantidad de resultados con concurrencias de datos, el autor no realiza ningún modelado de reglas de asociación, que hubiesen permitido establecer confianzas y soportes a sus resultados, dando una medida objetiva de evaluación de sus observaciones.

5.- CRISP-DM

CRISP-DM es el estándar de facto de ciclo de vida de un proyecto de KDD. La metodología de CRISP-DM está descrita en términos de un modelo de proceso jerárquico, consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo más general a lo más específico): fase, tarea genérica, tarea especializada, e instancia de procesos. Lo que aquí se presenta en un resumen de la metodología extraído a partir de [Chapman, et al. 2000].

La metodología CRISP-DM fue creado entre 1996 y 1999 por:

- Daimler-Benz con experiencia en proyectos de Minería de Datos industrial y comercial.
- SPSS que había lanzado la primera herramienta de trabajo comercial de Minería de Datos, Clementine.
- NCR, que quería integrar su herramienta de almacenes de datos, Teradata, con utilidades de Minería de Datos.

En el nivel superior, el proceso de Minería de Datos es organizado en un número de fases; cada fase consiste de varias tareas genéricas de segundo nivel. Este segundo nivel lo llaman genérico porque está destinado a ser bastante general para cubrir todas las situaciones posibles de Minería de Datos. Las tareas genéricas están destinadas a ser tan completas y estables como sea posible. El tercer nivel, el nivel de tarea especializada, es el lugar para describir cómo las acciones en las tareas genéricas deberían ser realizadas en ciertas situaciones específicas. Por ejemplo, en el segundo nivel podría haber una tarea genérica llamada limpieza de datos. El tercer nivel describe como esta tarea se diferencia en situaciones diferentes, como la limpieza de valores numéricos contra la limpieza de valores categóricos, o si el tipo de problema es agrupamiento o el modelado predictivo. Muchas de las tareas pueden ser realizadas en un orden diferente, y esto a menudo implicará volver a hacer tareas anteriores repetidamente y repetir ciertas acciones. El cuarto nivel, la instancia de proceso, es un registro de las acciones, decisiones, y de los resultados de una Minería de Datos real contratada.

Horizontalmente, la metodología de CRISP-DM se distingue entre el modelo de referencia y la guía de usuario. El modelo de referencia presenta una descripción rápida de fases, las tareas, y sus salidas, y describen que hacer en el proyecto de Minería de Datos. La guía de usuario da consejos más detallados e insinuaciones para cada fase y cada tarea dentro de una fase, y representa como realizar un proyecto de Minería de Datos.

El contexto de Minería de Datos traza un mapa entre lo genérico y el nivel especializado en CRISP-DM. Actualmente, se distingue entre cuatro dimensiones diferentes de contextos de Minería de Datos [Chapman, et al. 2000]:

- El dominio de aplicación es el área específica en la que el proyecto de Minería de Datos toma lugar los tipos de problemas de Minería de Datos describen la(s) clase(s) específica(s) de objetivo(s).
- El aspecto técnico cubre cuestiones específicas en Minería de Datos que describen diferentes dificultades que por lo general ocurren durante la Minería de Datos.
- La herramienta y las especificaciones de dimensión técnica en la que las herramienta(s) de Minería de Datos y/o técnicas son aplicadas durante el proyecto de Minería de Datos.

Distinguimos entre dos tipos diferentes de estrategias:

- Estrategia para el presente: Si sólo aplicamos el modelo de proceso genérico para realizar un proyecto de minería simple, e intentar pasar de tareas genéricas y sus descripciones al proyecto específico como requerido, hablamos sobre una estrategia solo para (probablemente) un solo uso.
- Estrategia para el futuro: Si sistemáticamente especializamos el modelo de proceso genérico según un contexto predefinido.

El modelo de proceso proporciona una descripción del ciclo de vida del proyecto de Minería de Datos. Este contiene las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. El ciclo de vida del proyecto de Minería de Datos consiste en seis fases.

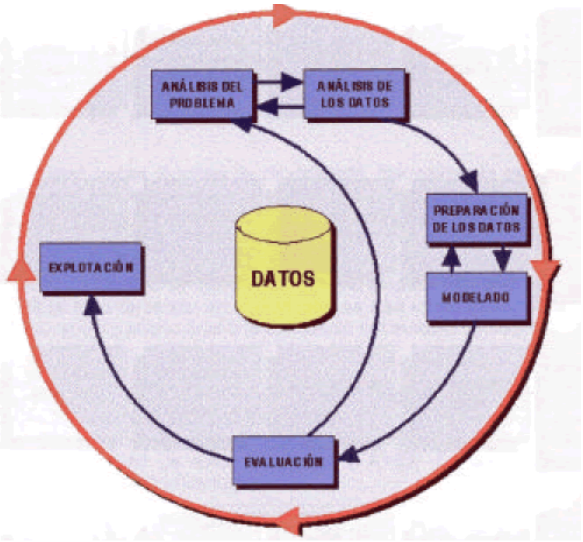

Figura 41: Ciclo de vida de CRISP-DM

El movimiento hacia adelante y hacia atrás entre fases diferentes es siempre necesario. El resultado de cada fase determina que la fase, o la tarea particular de una fase, tienen que ser realizados después. Las flechas indican las más importantes y frecuentes dependencias entre fases.

Las fases son las siguientes:

- **Comprensión del negocio:** Comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego se convierte este conocimiento de los datos en la definición de un problema de Minería de Datos y en un plan preliminar diseñado para alcanzar los objetivos.

- **Comprensión de los datos:** Comienza con la colección de datos inicial y continua con las actividades que le permiten familiarizar primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

- **Preparación de datos:** Cubre todas las actividades necesarias para construir el conjunto de datos final que constituirán la vista minable. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

- **Modelado:** Se seleccionan varias técnicas de modelado y se aplican, y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de Minería de Datos. Algunas técnicas tienen

requerimientos específicos sobre la forma de datos, por lo que volver a la fase de preparación de datos puede ser necesario.

- **Evaluación:** Se comprueba la calidad del modelo.
- **Desarrollo**: La creación del modelo no es generalmente el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento ganado tendrá que ser organizado y presentado en el modo en el que el cliente pueda usarlo. La fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de Minería de Datos a través de la empresa.

5.1.- COMPRENSIÓN DEL NEGOCIO

5.1.1.- DETERMINACIÓN DE OBJETIVOS DE NEGOCIO

Consiste en entender, desde una perspectiva de negocio, lo que el cliente realmente quiere lograr. El objetivo es mostrar factores importantes.

Se debe registrar la información que se conoce sobre la situación de negocio de la organización al principio del proyecto. Hay que describir el objetivo primario del cliente, desde una perspectiva de negocio. Además de los objetivos del negocio primario, se describen otras preguntas de negocio típicamente relacionadas con lo que al cliente le gustaría gestionar. Por ejemplo, el objetivo primario de negocio podría ser mantener a los clientes actuales por predicción en los que han tenido tendencia a pasar a ser clientes de la competencia. Los ejemplos de preguntas relacionadas de negocio son "¿Cómo el uso del canal principal afecta si los clientes se quedan o se van? " o "¿Bajar los honorarios considerablemente reducirá que el número de los clientes más importante se vayan?"

Hay que describir los criterios para llevar el proyecto a buen término desde el punto de vista del negocio. Esto podría ser bastante específico y capaz de ser medido objetivamente, por ejemplo, la reducción de clientes se reduce a un cierto nivel o valor, o esto podría ser general y subjetivo, como "dar ideas útiles en las relaciones".

5.1.2.- EVALUACIÓN DE LA SITUACIÓN

Esta tarea implica la investigación más detallada sobre todos los recursos, restricciones, presunciones, y otros factores que deberían considerarse en la determinación del objetivo de análisis de datos y el plan de proyecto.

Para esto hay que listar los recursos disponibles para el proyecto, incluyendo el personal y sus perfiles, datos, recursos hardware, y software (herramientas de Minería de Datos y otro software relevante).

Hay que listar todos los requerimientos del proyecto, incluyendo cómo se termina, la comprensibilidad y calidad de los resultados, y la seguridad, así como las cuestiones legales. Como parte de esta salida, hay que asegurarse de la disponibilidad de los datos, ya que en ocasiones los datos no pueden estar accesibles legalmente o solo bajo determinadas circunstancias.

Se deben también listar las presunciones hechas por el proyecto. Estas pueden ser presunciones sobre los datos que pueden ser verificadas durante la Minería de Datos, pero también puede incluir presunciones no comprobables sobre el negocio relacionado con el proyecto.

Adicionalmente, se deben listar las restricciones sobre el proyecto. Estas pueden ser restricciones sobre la disponibilidad de recursos, pero puede también incluir coacciones tecnológicas como el tamaño de conjunto de datos que es práctico para usar en el modelado. En nuestro caso, el tamaño de la muestra condicionará la capacidad del algoritmo de trabajar con la muestra al completo.

Se debe, por último, listar los riesgos o los acontecimientos que podrían retrasar el proyecto o hacer que fracase. Listar los planes de contingencia correspondientes y que acción será tomada si estos riesgos o acontecimientos ocurren.

Es muy interesante hacer un glosario de terminología relevante al proyecto. Esto puede incluir dos partes:

- Un glosario de terminología relevante del negocio, que forma la parte de la comprensión del negocio disponible al proyecto.
- Un glosario de terminología de Minería de Datos, ilustrada con ejemplos relevantes al problema del negocio en cuestión.

Por último, hay que construir un análisis coste-beneficio para el proyecto, que compare los gastos del proyecto con los beneficios potenciales al negocio si tiene éxito.

5.1.3.- Determinación de los objetivos de la Minería de Datos

En este punto se trata de traducir los objetivos de negocio en objetivos del proyecto en términos técnicos. Por ejemplo, para el objetivo de negocio podría ser “Aumentar catálogos de ventas a clientes existentes”, un objetivo de Minería de Datos podrían ser “Predecir cuantos artículos un cliente comprará, obteniendo datos de sus compras de tres años pasados, información demográfica (edad, sueldo, ciudad, etc.), y el precio del artículo.”

Hay que definir los criterios de un resultado con éxito para el proyecto en términos técnicos.

5.1.4.- PRODUCIR EL PLAN DEL PROYECTO

En esta tarea se debe describir el plan intencionado para alcanzar los objetivos de Minería de Datos y así alcanzar los objetivos de negocio. El plan debería especificar los pasos para ser realizados durante el resto del proyecto. Se debe para ello, Listar las etapas a ser ejecutadas en el proyecto, juntos con su duración, recursos requeridos, entradas, salidas, y dependencias. Es también importante, analizar dependencias entre la planificación de tiempo y los riesgos.

Por último en esta fase se deben evaluar inicialmente las herramientas y técnicas a utilizar.

5.2.- COMPRENSIÓN DE DATOS

5.2.1.- RECOLECCIÓN DE LOS DATOS INICIALES

Esta recolección inicial incluye cargar los datos si fuera necesario para la comprensión de los mismos. Se debe listar el conjunto de datos adquiridos, juntos con su lugar de captura, los métodos usados para adquirirlos, y algunos de los problemas encontrados.

5.2.2.- DESCRIBIR LOS DATOS

En esta tarea se debe describir los datos que han sido adquiridos, incluyendo el formato de los datos, la cantidad de datos, los identificadores de los campos, y cualquier otro rasgo superficial que ha sido descubierto. Se evaluará si los datos adquiridos satisfacen las exigencias relevantes.

5.2.3.- EXPLORAR LOS DATOS

Se revisan los datos para ver si se necesitan realizar agregaciones, donde están las claves, las propiedades de las subpoblaciones significativas, y análisis estadísticos simples. Este análisis puede dirigir los objetivos de Minería de Datos, contribuir o refinar la descripción de datos e informes de calidad, y alimentar en la transformación y otros pasos de preparación de datos necesarios para análisis futuros.

En este paso, se describen los resultados de esta tarea, incluyendo primeras conclusiones o hipótesis iniciales y su impacto sobre el resto del proyecto. Si es apropiado, se incluyen gráficos para indicar las características de datos que sugieren más examen de subconjuntos de datos interesantes.

5.2.4.- VERIFICAR LA CALIDAD DE LOS DATOS

Se debe verificar la calidad de los datos, realizando preguntas como: ¿Los datos están completos? ¿Son correctos, o estos contienen errores y, si hay errores, son frecuentes? ¿Hay valores omitidos en los datos?

Se debe obtener un informe de calidad de datos y si existen problemas de calidad en estos, que es lo más común, listar las posibles soluciones. Las soluciones a los problemas de calidad de datos generalmente dependen tanto del conocimiento de los datos y como del negocio.

5.3.- PREPARACIÓN DE DATOS

Durante la fase de preparación de datos, se deben generar el conjunto de datos, que será usada para modelar o para el trabajo principal de análisis del proyecto, es decir, la vista minable y describirlos.

5.3.1.- SELECCIÓN DE DATOS

Hay que decidir qué datos serán usados para el análisis. Los criterios incluyen la importancia de los objetivos de la Minería de Datos, la calidad, y las restricciones técnicas como límites sobre el volumen de datos o los tipos de datos según la técnica que será usada posteriormente. La selección de datos cubre la selección de atributos (columnas) así como la selección de registros (filas) en una tabla. Se debe listar los datos que deben ser incluidos/excluidos y los motivos para tomar estas decisiones.

5.3.2.- LIMPIEZA DE DATOS

Conseguir que la calidad de los datos alcance el nivel requerido por las técnicas de análisis seleccionadas con las que luego se realizarán la fase de Minería de Datos. Esto puede implicar la selección de subconjuntos de datos, la inserción de valores por defecto adecuados, o la estimación de datos faltantes mediante modelado.

Es importante haber descrito las decisiones y acciones tomadas para dirigir los problemas de calidad de datos informados durante la tarea de Verificación de Calidad de Datos de los Datos de la fase de Comprensión de Datos. Las transformaciones de los datos para una apropiada limpieza y el posible impacto en el análisis de resultados deberían ser considerados.

5.3.3.- CONSTRUIR DATOS

Se debe construir las operaciones de preparación de datos tales como la producción de atributos derivados o el ingreso de nuevos registros, o la transformación de valores para atributos existentes.

Los atributos derivados son los atributos nuevos que son construidos a partir de uno o más atributos existentes en el mismo registro. Ejemplo: área = longitud * anchura.

Puede ser necesaria también la creación de registros completamente nuevos. Ejemplo: Crear registros para los clientes que no hicieron compras durante el año pasado. No había ninguna razón de tener tales registros en los datos brutos, pero para el objetivo

del modelado esto podría tener sentido para representar explícitamente el hecho que ciertos clientes no hayan hecho compras.

5.3.4.- INTEGRAR DATOS

Incluye los métodos por los que la información es combinada de múltiples tablas o registros para crear nuevos registros o valores.

Los datos combinados también incluyen las agregaciones. La agregación se refiere a operaciones en la que son calculados nuevos valores de información resumida a partir de múltiples registros y/o tablas. Por ejemplo, convirtiendo una tabla de compra de clientes donde hay un registro para cada compra en una tabla nueva donde hay un registro para cada cliente, con campos tales como el número de compras, el promedio de la cantidad de compra, el porcentaje de ordenes cobradas por tarjeta de crédito, el porcentaje de artículos bajo promoción, etc.

5.3.5.- FORMATEAR DATOS

Estas transformaciones se refieren a modificaciones principalmente sintácticas hechas a los datos sin cambiar estos su significado, ya que pueden ser requeridos así por la herramienta de modelado.

Algunas herramientas tienen requerimientos sobre el orden de los atributos, tales como el primer campo que es un único identificador para cada registro o el último campo es el campo resultado que el modelo debe predecir. Por ello, podría ser importante cambiar el orden de los registros en el conjunto de datos. Además, hay cambios sintácticos hechos para satisfacer las exigencias de la herramienta de modelado específica.

5.4.- MODELADO

5.4.1.- SELECCIÓN DE LA TÉCNICA DE MODELADO

Lo primero debe ser seleccionar la técnica de modelado específico, por ejemplo, un árbol decisión construido con C4.5, o la generación de una red neuronal por Retro-Propagación. Si múltiples técnicas son aplicadas, se realizan esta tarea separadamente para cada técnica.

Muchas técnicas de modelado hacen presunciones específicas sobre los datos. Por ejemplo, que todos los atributos tengan distribuciones uniformes, no encontrar valores no permitidos, el atributo de clase debe ser simbólico, etc.

5.4.2.- GENERACIÓN DE LA PRUEBA DE DISEÑO

Antes de que en realidad se construya un modelo, hay que generar un procedimiento o el mecanismo para probar la calidad y validez del modelo. Por ejemplo, en tareas de

Minería de Datos supervisadas como la clasificación, esto es común usar tasas de errores como medida de calidad para modelos de Minería de Datos. Por lo tanto, típicamente se separa el conjunto de datos en datos de entrenamiento y datos de prueba, se construye el modelo sobre el conjunto de series, y se estima su calidad sobre el conjunto de prueba separado.

Describir el plan destinado al entrenamiento, la prueba, y la evaluación de los modelos. Un componente primario del plan es determinar como dividir un conjunto de datos disponible en datos de entrenamiento, datos de prueba, y conjunto de datos de validación.

5.4.3.- CONSTRUCCIÓN DEL MODELO

Ejecutar la herramienta de modelado sobre el conjunto de datos preparados para crear uno o más modelos.

Con cualquier herramienta de modelado, hay a menudo un gran número de parámetros que pueden ser ajustados. Listar los parámetros y sus valores escogidos, también redactar el razonamiento para elegir los parámetros de ajuste.

Estos son los modelos reales producidos por la herramienta de modelado. Hay que describir los modelos obtenidos, informar sobre la interpretación de los modelos y documentar cualquier dificultad encontrada con sus significados.

5.4.4.- EVALUACIÓN DEL MODELO

El ingeniero de Minería de Datos interpreta los modelos según su conocimiento de dominio, los criterios de éxitos de Minería de Datos, y el diseño de prueba deseado. El ingeniero de Minería de Datos juzga el éxito de la aplicación de modelado y descubre técnicas; se pone en contacto con analistas de negocio y expertos en el dominio para hablar de los resultados de la Minería de Datos en el contexto de negocio. En esta tarea sólo se considera modelos, mientras que la fase de evaluación también toma en cuenta los resultados que fueron producidos durante el proyecto.

Se intenta clasificar los modelos según los criterios de evaluación teniendo en cuenta objetivos del negocio y criterios de éxito del negocio. En los grandes proyectos de Minería de Datos, se aplica una sola técnica más de una vez, o genera resultados de Minería de Datos con varias técnicas diferentes. En esta tarea, también se comparan todos los resultados según los criterios de evaluación.

Según la evaluación del modelo, hay que revisar los parámetros de ajuste y repetir la construcción y evaluación del modelo hasta que se encuentre con el modelo adecuado. Hay que documentar las revisiones y las evaluaciones.

Los pasos de esta evaluación tratan con factores como la exactitud y la generalidad del modelo. Este paso evalúa el grado al que el modelo responde (encuentra) los objetivos de negocio y procura determinar si hay alguna decisión de negocio por el que este modelo es deficiente. Otra opción de evaluación es probar el/los modelo/s sobre aplicaciones de prueba en la aplicación real, si el tiempo y las restricciones de presupuesto lo permiten.

5.5.- Evaluación

5.5.1.- Evaluación de los resultados

Esta es una evaluación de los resultados de la Minería de Datos en lo que concierne a criterios de éxito de negocio, incluyendo una declaración final en cuanto si el proyecto ya encuentra los objetivos iniciales de negocio. Después de esta evaluación los modelos generados que cumplen los criterios seleccionados son los modelos aprobados. Los modelos resultantes deben ser satisfactorios para las necesidades de negocio.

5.5.2.- Revisión

Hay que hacer una revisión más cuidadosa de los compromisos de la Minería de Datos para determinar si hay cualquier factor importante o tarea que de algún modo haya sido pasada por alto. Esta revisión también cubre cuestiones de calidad como: ¿Construimos correctamente el modelo? ¿Usamos sólo los atributos que nos permitieron usar y que están disponibles para análisis futuros? Por ello, se debe resumir la revisión de proceso y destacar las actividades que han sido omitidas y/o aquellas que deberían ser repetidas.

5.5.3.- Determinar los próximos pasos

Se debe decidir si hay que terminar este proyecto y tomar medidas sobre el desarrollo si es apropiado, tanto iniciar más iteraciones, o comenzar nuevos proyectos de Minería de Datos. Para ello, se debe listar las acciones futuras potenciales, con los motivos a favor y en contra de cada opción y describir la decisión en cuanto a cómo proceder, junto con el razonamiento.

5.6.- Desarrollo

5.6.1.- Desarrollo del plan

De acuerdo al desarrollo de los resultados de Minería de Datos en el negocio, esta tarea toma los resultados de la evaluación y determina una estrategia para el despliegue. Si un procedimiento general ha sido identificado como capaz de crear modelos relevantes, este procedimiento se documenta aquí para el desarrollo posterior.

Se debe resumir la estrategia de desarrollo, incluyendo los pasos necesarios y cómo realizarlos.

5.6.2.- PLANEAR LA SUPERVISIÓN Y EL MANTENIMIENTO

La supervisión y el mantenimiento son cuestiones importantes si los resultados de Minería de Datos son parte del trabajo cotidiano. La preparación cuidadosa de una estrategia de mantenimiento ayuda evitar largos periodos innecesarios de uso incorrecto de resultados de Minería de Datos. Para supervisar el desarrollo de los resultados de la Minería de Datos, el proyecto necesita un plan detallado del proceso de supervisión. Este plan tiene en cuenta el tipo específico de desarrollo. Por ello, hay que resumir la estrategia de supervisión y mantenimiento incluyendo los pasos necesarios y cómo realizarlos.

5.6.3.- INFORME DEFINITIVO DE PRODUCTO

Este informe puede ser sólo un resumen del proyecto y sus experiencias (si estas aún no han sido documentadas como una actividad en curso) una presentación final y exhaustiva de los resultados de Minería de Datos.

Esto incluye todo el desarrollo anterior, el resumen y la organización de los resultados. Además, habrá una reunión en la conclusión del proyecto en el que los resultados son presentados oralmente.

5.6.4.- REVISIÓN DEL PROYECTO

Evaluar todo aquello realizado correctamente, así como lo incorrecto. Especialmente anotar aquello susceptible de ser mejorado.

6.- Estudio del Caso del Análisis del Tratamiento Informativo de la Drogadicción

6.1.- Descripción del Proyecto

6.1.1.- Introducción

El presente proyecto se basa en los datos obtenidos por otro proyecto de investigación denominado "Análisis y diseño de campañas y programas de sensibilización y prevención de las drogodependencias en los medios de comunicación" que está llevando a cabo el equipo de la Dra Dña. Pilar Paricio Esteban en la Universidad CEU Cardenal Herrera. A continuación se tratará de dar una descripción del proyecto así como su metodología de trabajo, fases, población de muestra, etc. para encuadrar el caso de estudio del presente trabajo de investigación.

6.1.2.- El proyecto "Análisis y diseño de campañas y programas de sensibilización y prevención de las drogodependencias en los medios de comunicación"

Que las instituciones responsables de la prevención y la asistencia en las drogodependencias lleven a cabo una acertada política comunicativa es un factor esencial para logar la mayor eficacia de sus mensajes en las audiencias clave, en particular, en los grupos que se inician en las drogas y en los segmentos de principales consumidores, además de en la sociedad en su conjunto. Para alcanzar esos objetivos, las instituciones desarrollan su comunicación principalmente a través de los medios, ya sea mediante campañas publicitarias o bien mediante las Relaciones con dichos medios.

Con el objetivo de completar la metodología utilizada en el Proyecto "La eficacia de la comunicación institucional en la prevención y asistencia a la drogadicción: Análisis de campañas y del tratamiento informativo de las drogas en el periodo Enero-Junio de 2008" presentado en la I Convocatoria de FEPAD (Fundación de la Comunitat Valenciana para el Estudio, Prevención y Asistencia a las Drogodependencias), el presente Proyecto se orienta al análisis de las campañas realizadas y el tratamiento informativo recibido por la drogadicción en los medios más afines al público objetivo durante el periodo de Enero-Junio de 2009. El estudio planteado daría continuidad al anterior Proyecto, donde se realizó un pilotaje con la metodología aplicada que se ha redefinido para el nuevo Proyecto, y, además, permitirá realizar una comparativa en periodos diferentes descubriendo la evolución en los modelos comunicativos de la prevención de las drogodependencias en los medios. En este sentido, los resultados del nuevo proyecto permitirían implementar recomendaciones que ayuden a mejorar la eficacia de las campañas de prevención de la drogadicción y de la comunicación

institucional de los organismos nacionales y autonómicos que trabajan en la prevención y asistencia a las drogodependencias.

El Proyecto incorpora también una fase de diseño de campañas publicitarias (taller de investigación y diseño de campañas con aplicación en medios audiovisuales y en medios no convencionales) con el objetivo de trabajar con el público diana, que puede aportar interesantes conclusiones a la hora de mejorar los códigos para conseguir una mayor penetración del mensaje en dicho público objetivo y, por consiguiente, aumentar el impacto a nivel cuantitativo y cualitativo.

En la sociedad de la comunicación y en la era de la información, los medios actúan como instituciones mediadoras entre la población y la realidad y, en muchos casos, como prescriptores. Existen diferentes investigaciones sobre el papel socializador de los medios y sus efectos sobre la persona, en particular, sobre los niños, pre-adolescentes y adolescentes. Según los principales teóricos de la comunicación social y según los principales estudiosos de los efectos cognitivos de la comunicación de masas, los mensajes transmitidos a través de los medios no sólo informan, sino que guían nuestras experiencias influyendo en la sociedad y conformando opinión. En este sentido, los medios, al llamar la atención sobre ciertos temas en lugar de otros, o al destacar unos contenidos temáticos frente a otros tratados con un encuadre menor, orientan nuestra atención, conforman la agenda de temas predominantes que reclaman dicha atención y jerarquizan la relevancia de dichos temas.

Una parte más de la comunicación de masas es la publicidad, que es una comunicación pagada efectuada a través de los medios masivos cuyos objetivos principales son informar y persuadir al público al que se dirige. En este sentido, la publicidad forma parte de la vida cotidiana tanto como los programas o artículos periodísticos junto a los que figura en cualquier medio. En nuestro sistema, la Administración es un usuario más de la Publicidad, porque pone en marcha campañas institucionales de corte social que tienen por objeto modificar actitudes o comportamientos negativos para el individuo o la sociedad. Las organizaciones públicas centran los objetivos de sus campañas en informar, favorecer el conocimiento de las leyes, modificar comportamientos, hábitos y costumbres, crear, mantener o mejorar la imagen... De todos estos objetivos, en las campañas publicitarias sobre la drogadicción, el fin principal sería modificar hábitos y costumbres y, en definitiva, romper con un modelo cultural que provoca la muerte. En este caso, la instrumentalización de la publicidad adquiere su más noble cometido.

De todo lo anterior podemos concluir que los medios de comunicación y la publicidad son elementos clave que pueden incidir determinantemente en la eficacia de los programas de prevención de las drogodependencias. En este sentido, quedaría justificado el interés de una investigación orientada a evaluar la eficacia de la comunicación institucional dirigida a medios y las campañas publicitarias en un periodo representativo cuyos resultados pueden aportar datos de gran interés para los organismos que trabajan en la prevención y asistencia de las drogodependencias.

El equipo de trabajo ya dispone de bibliografía publicada sobre la temática en cuestión como [Paricio Esteban, et al. 2010] y [Paricio Esteban, et al. 2002]

El proyecto[1] tiene continuidad en el presente, utilizando los datos obtenidos en el periodo enero 2009 – junio 2009 y otros que se recopilaron posteriormente. Se están realizando otro tipo de análisis, aparte del aspecto institucional.

6.1.3.- OBJETIVOS GENERALES DEL PROYECTO

Los objetivos del Proyecto son los siguientes:

- Analizar el tratamiento publicitario e informativo de las drogodependencias en los medios de comunicación en el periodo descrito.
- Implementar recomendaciones que permitan mejorar la eficacia de las campañas de prevención de la drogadicción y de la comunicación institucional de los organismos nacionales y autonómicos que trabajan en la prevención y asistencia a las drogodependencias.

6.1.4.- OBJETIVOS ESPECÍFICOS

Los objetivos específicos son los siguientes:

- Estudiar las campañas publicitarias de prevención y asistencia a las drogodependencias en el ámbito local (Comunidad Valenciana) y nacional con el fin de medir su eficacia realizando un análisis comparativo y aplicando la metodología perfilada a partir del pilotaje realizado en el Proyecto presentado en la I Convocatoria de FEPAD.
- Estudio del tratamiento informativo en los medios de comunicación de las campañas e instituciones que trabajan en la prevención y asistencia de las drogodependencias a partir de un análisis de las noticias que informan y tratan los distintos aspectos de la drogadicción realizando un análisis comparativo y aplicando la metodología perfilada a partir del pilotaje realizado en el Proyecto presentado en la I Convocatoria de FEPAD.
- Analizar la repercusión en los medios de comunicación de las campañas de prevención de las drogodependencias en el periodo estudiado.
- Analizar el tratamiento informativo que los medios de comunicación propician para las noticias relacionadas con las drogodependencias.
- Realizar un análisis comparativo frente a los resultados obtenidos en el Proyecto presentado en la I Convocatoria de FEPAD.

[1] La descripción del proyecto, así como una enumeración de referencias externas, que dan soporte al proyecto de investigación pueden obtenerse del informe interno [Paricio Estéban. 2009]

- Concluir el tipo de tratamiento informativo propiciado por los medios de comunicación social sobre las drogodependencias, sus causas y sus consecuencias, así como de las actuaciones desarrolladas en materia de asistencia y prevención. Dicho tratamiento será determinante dado el peso de la interpretación de la realidad social ofrecida por los medios de comunicación en la configuración de la agenda pública y, por tanto en la opinión pública.

- Establecer un contacto sistematizado entre profesionales de los medios de comunicación, investigadores en drogodependencias y personal de la Dirección General de Drogodependencias.

- Verificar el grado de sensibilización de los medios de comunicación en la prevención de las drogodependencias.

- Trabajar con el público diana, que puede aportar interesantes conclusiones a la hora de mejorar los códigos para conseguir una mayor penetración del mensaje en dicho público objetivo y, por consiguiente, aumentar el impacto a nivel cuantitativo y cualitativo a partir del diseño de campañas publicitarias (taller de investigación y diseño de campañas con aplicación en medios audiovisuales y medios no convencionales).

- Formación de investigadores noveles.

6.1.5.- POBLACIÓN DE MUESTRA

Hay varias muestras en el proyecto original. Nos centraremos en noticias sobre los diferentes aspectos de la drogadicción en el periodo Enero-Junio 2009 en periódicos de información general nacionales con edición Comunidad Valenciana:

- EL País
- El Mundo
- ABC
- La Razón

6.1.6.- FASES DEL PROYECTO

- Fase 0. Elaboración de guía para instituciones sobre revistas para adolescentes.
    - Temporización: Abril-Septiembre 2009
    - Tareas:

- Resumir resultados del Proyecto financiado por I Convocatoria FEPAD
- Reunir datos descriptivos sobre publicaciones para adolescentes
- Redacción de la guía
- Edición y elaboración de la guía
- Envío a las instituciones

- Fase 1. Investigación del tratamiento informativo de las drogas en prensa dirigida a adolescentes
  - Temporizació: Marzo-Septiembre 2009
  - Tareas:
    - Fijación del Corpus de investigación
    - Búsqueda hemerográfica
    - Análisis de contenido de las informaciones sobre drogas
    - Resultados y conclusiones
- Fase 2. Investigación del tratamiento informativo de las drogas en prensa de información general.
  - Temporización: Enero-Diciembre 2009
  - Tareas:
    - Fijación del Corpus de investigación
    - Búsqueda hemerográfica
    - Análisis de contenido de las informaciones sobre drogas
    - Resultados y conclusiones
- Fase 3. Análisis de campañas publicitarias sobre drogas
  - Temporización: Enero-Junio 2009
  - Tareas:
    - Fijación corpus investigación

- Búsqueda documental: Recopilación de campañas Análisis de contenido de las campañas
- Resultados y conclusiones

- Fase 4. Investigación y diseño de campañas con público diana (Taller de publicidad)
  - Temporización: Febrero-octubre 2009
    - Tareas:
    - Análisis de los mensajes publicitario de prevención de drogas
    - Propuestas de nuevos hilos argumentales, mensajes y formatos
    - Realización de spots y acciones
- Fase 5. Artículos para revistas
  - Temporización: Enero-Marzo 2010
  - Tareas:
    - Redacción y envío de artículos según normas de extensión y presentación fijados por las Revistas
- Fase 6. II Jornadas sobre campañas y comunicación institucional para la prevención de la drogadicción
  - Temporización: Marzo 2010
- Fase 7. Publicación
  - Temporización: Marzo 2010
  - Tareas:
    - Estructuración de la publicación
    - Redacción propia
    - Solicitud de textos de las ponencias de las Jornadas
    - Integración de textos
    - Revisión
    - Edición

- Fase 8. Talleres con periodistas
    - Temporización: Abril 2009-Marzo 2010
    - Tareas:
        - Reuniones con miembros de FEPAD para preparar encuentros con periodistas
        - Resumir resultados del Proyecto financiado por I Convocatoria FEPAD
        - Elaboración de contenidos de talleres
        - Contacto y convocatoria de periodistas al taller
        - Organización de rueda de prensa de presentación de nuevos datos sobre drogodependencias a periodistas

6.1.7.- METODOLOGÍA EMPLEADA

Se realizó una búsqueda documental y bibliográfica sobre la población de muestra anteriormente descrita.

Metodología de investigación aplicada en la Fase 1: Análisis de contenido de tipo categorial y del discurso de las campañas publicitarias para la prevención de las drogodependencias.

Metodología de investigación aplicada en las Fases 2 y 3: Metodología de medición de la exposición al mensaje: Análisis de contenido de las informaciones sobre el tema de tipo categorial y evaluativo de la intensidad (encuadre). Siguiendo la clasificación de [Bardin. 1986], serán de aplicación dos tipos de análisis de contenido: el de tipo categorial y el evaluativo ampliado desde la perspectiva del *framing* y de la hemerografía comparativa francesa. Como hemos reseñado anteriormente, el análisis categorial comprenderá varias variables de estudio:

Análisis de la fuente de la información (donde se analizará la fuente manifiesta y la citada), que nos permitirá valorar la eficacia de las instituciones y organizaciones promotoras de las campañas como fuente fiable para los medios de comunicación, así como descubrir las principales fuentes consultadas por los medios en el tratamiento de este tipo de noticias. Por otra parte, aportará datos sustanciales sobre los periodistas que intervienen en el proceso de elaboración de las noticias sobre el tema.

Análisis temático donde las categorías se confeccionan “ad hoc” en función de las variables de análisis mencionadas anteriormente.

El análisis evaluativo del encuadre nos permitirá verificar el contexto en el que los medios enmarcan las informaciones sobre drogodependencias (el judicial, sanitario...), así como la intensidad con la que los medios tratan las informaciones de drogas, de lo que se puede concluir la importancia concedida a las informaciones sobre el tema y su grado de sensibilidad con el tema.

La recogida de informaciones objeto de estudio se realizará mediante exploración diaria de las publicaciones analizadas y en colaboración con FEPAD, que realiza ya un vaciado de informaciones de prensa relacionadas con las drogodependencias.

6.2.- El Análisis Periodístico en Prensa Escrita

A continuación se explicará brevemente el análisis periodístico que se realiza sobre prensa escrita y que se ha realizado en el proyecto descrito.

Cuando se habla de análisis periodístico, es conveniente remarcar que este análisis se encuadra en la prensa escrita, debido a que, pese a la existencia de otros medios de comunicación, la prensa escrita tiene un formato muy conocido y poco variable, que permite realizar afirmaciones sobre un escrito periodístico en función de valores perfectamente medibles. No así otros medios como el audiovisual, o por ejemplo, los medios digitales.

Es de destacar que, pese a que los diarios de los que se recoge la muestra tienen una versión digital, esta versión no siempre corresponde a la versión impresa, y además, no se puede aplicar el mismo tipo de análisis al medio digital por sus peculiaridades técnicas. Esto se tratará más adelante en la descripción de la recolección de los datos iniciales.

El análisis periodístico es una disciplina que tiene varias décadas de existencia. Básicamente, se pueden identificar dos vertientes del análisis periodístico: El Análisis de Contenido desde la perspectiva del *Framing*, cuyos máximos exponentes son [Bardin. 1986, Scheufele. 1999] y el Análisis de la Intensidad Formal, con su principal postulador, [Kayser. 1974].

6.2.1.- El Análisis de Contenido desde la perspectiva del *Framing*

Siguiendo la tesis de [Rodríguez Luque. 2009] sobre el *Framing* o Encuadre, se trata de una teoría del análisis de contenido que proviene de la sociología interpretativa. Aunque el concepto de *frame* surge en el contexto de la sociología interpretativa, sus orígenes inmediatos proceden de la psicología de la mano de [Goffman, et al. 1974] que, al recogerlo, añade los matices sociológicos que permitirán que se aplique posteriormente al estudio de los medios de comunicación [Sádaba. 2001, Reese, et al. 2001].

Así, [Sádaba. 2001] indica sobre [Goffman, et al. 1974]: "el *frame* para Goffman es tanto un marco como un esquema. Un marco que designa el contexto de la realidad y un esquema o estructura mental que incorpora los datos externos objetivos".

Para [Entman. 1993] que es uno de los teóricos más citados en estudios de comunicación, enmarcar- *to frame*- se define del siguiente modo:

"To frame is to select some aspect of a perceived reality and make them more salient in a communicating text in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and for treatment recommendation for the item described."

Es decir, podemos entender el concepto del *framing* como el enfoque que sobre una misma realidad se le da a una noticia.

Hay que aclarar que, pese a que existe una técnica objetiva definida, el análisis del *framing* puede parecer subjetivo comparado con el análisis de intensidad formal. El análisis del *framing* se realiza sobre elementos objetivos que aparecen en el texto, tratando de una evaluar el enfoque dado por el autor al escribir el texto periodístico. Por tanto, se valoran parámetros subjetivos mediante técnicas y datos objetivos. Se tratará de observar cómo percibe la realidad objetiva la persona que escribe el texto periodístico y aquello que quiere transmitir.

En el presente estudio las variables del *framing* a estudiar han sido codificadas por el equipo de investigación como ausente, presente o destacado en el tema principal y la entradilla.

6.2.2.- El Análisis de Intensidad Formal

[Canga Larequi, et al. 2010] nos remite a la tesis de [Núñez-Romero Olmo. 2009] para un análisis exhaustivo de la metodología. Para ello, [Núñez-Romero Olmo. 2009] nos introduce en el concepto de la hemerografía de la mano de [Casasús. 1985]: *"El objeto de la hemerografía es el examen, estudio y descripción totales de los periódicos diarios."*

La hemerografía analítica se subdivide en tres: la hemerografía registral –la que se ocupa de la identificación de periódicos–, la hemerografía comparada –la que estudia la evolución de los medios a través del tiempo, ya sea examinando su manifestación diaria o su manifestación en determinadas fechas separadas por unos lapsus determinados–, y la hemerografía estructural –la que propone un tratamiento de la espacialidad del medio basado en las técnicas de confección y compaginación de los diarios–. Y es este último en el que [Kayser. 1974, Casasús. 1985, Martinez Albertos. 1984, Nelkin. 1987] se vuelven referencias imprescindibles para el presente estudio.

El objeto que se persigue es el del análisis de la estructura del periódico y de su presentación, es decir: revelar lo que un periódico ha querido comunicar a sus lectores y presumir la influencia que una lectura normal del periódico ha ejercido en el lector.

Entrarían dentro de este análisis mediciones como el emplazamiento dentro del periódico –portada, página par/impar, etc.-, la superficie del titular, jerarquía del titular, etc.

En el presente estudio han sido codificadas estas variables de forma numérica y se ha obtenido un valor global, ponderado al peso específico de cada una de las variables, para permitir analizar el peso global que se le ha querido asignar al texto estudiado dentro de un periódico concreto.

7.- Comprensión del Caso de Estudio

Se pretende en este capítulo determinar cuáles son los objetivos del caso de estudio desde el punto de vista periodístico, cual es la situación actual del mismo viendo cómo resuelven el análisis de datos el grupo de investigación y por último, ver cuáles son los objetivos de la Minería de Datos, determinando cuáles son los criterios de éxito.

7.1.- Determinación de los Objetivos del Caso de Estudio

A continuación se tratará de explicar quiénes son los agentes implicados en el grupo de Investigación y cuáles son los antecedentes del análisis periodístico sobre la temática de la drogadicción.

7.1.1.- Fundación de la Comunitat Valenciana para el Estudio, Prevención y Asistencia a las Drogodependencias

La Conselleria de Sanitat de la Generalitat Valenciana, a través de la Dirección General de Drogodependencias, viene desarrollando una extensa política de actuaciones para conseguir la máxima eficacia en la lucha contra las drogodependencias.

En la búsqueda de soluciones, y dentro del marco operativo definido en el Decreto Legislativo 1/2003 de 1 de abril, por el que se aprueba el Texto Refundido de la Ley de Drogodependencias y otros Trastornos Adictivos, el Consell de la Generalitat aprobó la constitución de la Fundación de la Comunitat Valenciana para el Estudio, Prevención y Asistencia a las Drogodependencias (en adelante, FEPAD).

La FEPAD, se constituye con carácter preventivo y científico con la finalidad de aportar, en el ámbito de la Comunitat Valenciana, un mayor conocimiento y comprensión sobre las conductas adictivas, fortaleciendo la labor investigadora en todas sus vertientes como fuente inequívoca del saber y base de la eficacia de posteriores actuaciones.

La Fundación dirige su acción a:

- Fomentar líneas de estudio e investigación que aporten un mayor conocimiento sobre las conductas adictivas.
- Desarrollar acciones de prevención dirigidas a evitar, reducir y controlar los trastornos adictivos.
- Ofrecer a los profesionales del campo de las drogodependencias un espacio de formación en el que intercambiar sus experiencias y actualizar sus conocimientos.
- Colaborar activamente con otras entidades nacionales e internacionales y contribuir en proyecto de cooperación al desarrollo.

Encuadrado dentro de sus objetivos, la FAD firmó un convenio especifico de colaboración entre la propia Fundación y la Universidad CEU Cardenal Herrera del que es fruto el proyecto de investigación periodística.

7.1.2.- UNIVERSIDAD CEU CARDENAL HERRERA

La Universidad CEU Cardenal Herrera pertenece a la Fundación CEU San Pablo, institución benéfico-docente sin ánimo de lucro y con más de 70 años de experiencia en el campo de la enseñanza. Los promotores de la Fundación CEU San Pablo pertenecen a la Asociación Católica de Propagandistas, fundada por el padre jesuita Ángel Ayala y cuyo primer presidente sería el abogado Ángel Herrera-Oria.

En la mencionada universidad, se constituyó el grupo de investigación dirigido por la doctora Pilar Paricio Esteban y compuesto por Mª José Rabadán, Cristina Rodríguez Luque y Francisco Núñez-Romero Olmo.

7.1.3.- ENCUADRE DEL PROBLEMA

Los estudios que el mencionado grupo viene realizando se centran principalmente en la evaluación estadística de los datos incluidos en la muestra. Así, se puede observar que los principales informes y comunicaciones se centran en un aspecto de los datos y omiten su relación con el resto.

Existen ya varias referencias propias del grupo sobre la temática de drogas frente a periodismo. Por citar alguna de ellas: [Núñez-Romero Olmo, et al. 2010], [Paricio Esteban, et al. 2010], [Paricio Esteban, et al. 2010], [Paricio Esteban, et al. 2002], [Paricio, et al. 2010], etc. Estas publicaciones se han llevado a cabo realizando un análisis estadístico básico de los datos recogidos, así como gráficos a modo de ejemplo.

En el ámbito de la temática que nos ocupa, esto es, Drogas vs. Periodismo, se pueden hallar referencias de otros autores que tratan de una forma similar la drogadicción en los textos periodísticos y medios de comunicación en general: [Berrio.], [FERNÁNDEZ-CID.], [Nebreda, et al. 1987], [Rekalde, et al. 2002], [Vega Fuente. 1995], etc.

Los análisis se realizan mediante las dos metodologías explicadas previamente en el apartado “6.2.- El Análisis Periodístico en Prensa Escrita.”

Los distintos estudios se centran en un aspecto concreto del análisis, tratando de mostrar qué conclusiones se pueden obtener del análisis de datos. Unos se centran en el *framing*, otros en la imagen de las instituciones que aparecen en la noticia, otros en la valoración formal que realizan los periódicos sobre las noticias relacionadas con la drogadicción, otros en la relación de las fuentes con otros aspectos de la noticia, etc.

Tanto los textos del grupo de investigación como los externos se centran en un aspecto concreto del análisis periodístico.

7.2.- Evaluación de la Situación

En el presente apartado se trata de explicar de qué recursos se dispone para realizar el proyecto de investigación, así como de dar una visión global de cómo está tratando el grupo de investigadores el problema del análisis periodístico de los datos de los que disponen.

7.2.1.- Recursos Disponibles

Los recursos inicialmente disponibles son los siguientes:

- Materiales:
    - SPSS Clementine versión 11.1. Herramienta de IBM para la Minería de Datos.
    - SPSS Statistics versión 18. Herramienta de IBM para el análisis estadístico de datos
- Humanos:
    - Francisco Núñez-Romero Olmo, como asesor en el ámbito del análisis periodístico de los datos y enlace con el grupo de investigación.
    - Juan Pardo Albiach, como director del presente proyecto.
    - Pablo M Romeu Guallart como investigador principal.
- Datos de trabajo:
    - Se dispone de la serie de datos comprendida entre enero y junio de 2009 de la muestra.
    - Se dispone de los documentos publicados por los diferentes integrantes del equipo de investigación.

7.2.2.- Descripción de la Situación Actual

Pese a que existen varias muestras delimitadas por el rango temporal, inicialmente se nos proporcionó una muestra de datos que incluía los recogidos entre enero y junio de 2009. El tamaño de esta muestra es de 502 registros.

Pese a que el objetivo inicial del proyecto trataba de hallar la visión que los medios escritos dan de las instituciones, especialmente de la FEPAD, los datos recogidos están siendo utilizados para multitud de estudios de varias perspectivas.

Mediante una exploración de los artículos realizados por el grupo de investigación periodística, se observan dos formas de tratar la información, que se describen a continuación utilizando dos de los textos.

En [Paricio Esteban, et al. 2010] se analizan 423 del total de textos recopilados en la muestra. Inicialmente se comprueba la fiabilidad de la muestra mediante un índice Kappa de Cohen. Posteriormente se realiza un análisis de frecuencias por periódico, género periodístico, secciones, temas tratados, encuadre, etc. y se obtienen conclusiones de ellas.

Por otro lado en [Rodríguez Luque, et al. 2011], se realiza un análisis de las 502 referencias de las que consta la base de datos, realizando los siguientes estudios:

- Las frecuencias conjuntas de valoración de la unidad de análisis y fuentes del documento.
- Se realiza un análisis segmentado por periódicos de lo anterior.
- Se realiza un análisis de frecuencias conjuntas de fuentes y enfoques (*Framing*). También se analiza la ausencia de enfoques y fuentes.
- Se realiza un análisis de frecuencias conjuntas de Fuentes y Tema Principal
- Se realiza un último análisis confrontando Fuentes y sustancias.

En este capítulo de libro, el análisis es más profundo que en el anterior texto, mostrando confianzas y soportes en las relaciones estudiadas.

Todos los estudios del grupo de investigación trantan los datos de esta manera, no confrontando más de dos variables por estudio. Sólo de manera excepcional se realiza una segmentación por alguna de las variables, introduciendo un tercer nivel de estudio.

En los estudios del equipo de investigación periodística se realizan estadísticas simples, análisis de confianzas y correlaciones de variables, y en algún caso, estudios de subpoblaciones de la muestra como en el caso de la segmentación por periódicos de [Rodríguez Luque, et al. 2011].

7.3.- Determinación de los Objetivos de la Minería de Datos

7.3.1.- Objetivo general

Como objetivos generales de la aplicación de la Minería de Datos al caso de estudio se plantean los siguientes:

- Dar soporte mediante técnicas de Minería de Datos a las conclusiones del grupo de investigación.

- Tratar de hallar nuevas relaciones entre los distintas variables.

7.3.2.- OBJETIVOS ESPECÍFICOS

Los objetivos específicos de la Minería de Datos, y por tanto, criterios de éxito de su aplicación serán:

- Comprobar o refutar de forma motivada los resultados del grupo de investigación.
- Realizar un estudio estadístico genérico de los datos.
- Estudiar qué variables influyen en el valor final de cada una de las variables.
- Hallar nuevas relaciones en la muestra, con relevancia para el grupo de investigación periodística.
- Hallar nuevas relaciones en la muestra, atendiendo a la ausencia de variables.
- Hallar nuevas relaciones en la muestra, atendiendo a subpoblaciones.

8.- Comprensión de los Datos del Problema

8.1.- Recolección de los datos iniciales

En la muestra utilizada se recogieron las reseñas de los textos de forma manual, por el personal del grupo de investigación periodística. Se obtenían los ejemplares de los cuatro periódicos analizados de forma diaria, y se buscaban página por página cualquier texto que pudiera ser relevante para la investigación. En el presente, las muestras se están obteniendo mediante la aplicación "MyNews", que permite buscar textos en la edición escrita de distintos periódicos.

Figura 43 Buscador MyNews

Figura 42 Vista de una información sobre drogas en MyNews

Como ya se ha anticipado en el apartado *6.2.-* El Análisis Periodístico en Prensa Escrita*,* el tipo de análisis planteado por el grupo de investigación periodística no se puede realizar en la edición digital de los periódicos, por varios motivos:

- El formato de las versiones escritas de los diarios siguen un patrón medible y poco cambiante. No así los medios digitales.
- Por este motivo, no existe, hasta la fecha, una metodología de análisis morfológico ampliamente aceptada para publicaciones periodísticas principalmente escritas en Internet.
- Es difícil relacionar la motivación del lector de una noticia con el formato de aparición de la noticia en medios digitales, debido a que muchos de los lectores pueden llegar a ella por medio de agregadores de noticias o hipervínculos.

Para la codificación de los datos, se utilizó un manual de codificación que se incluye en el anexo *13.1.-* Anexo: Libro de instrucciones para base de datos de análisis de prensa.

Los datos incluidos en esta muestra se han recogido en un fichero Excel, existiendo otras muestras de datos recogidas en otros ficheros y bases de datos.

8.2.- DESCRIPCIÓN DE LOS DATOS

Las variables estudiadas, tipos y valores son los siguientes, así como el tratamiento que se le dará en el presente proyecto son los siguientes:

8.2.1.- SECCIÓN DE IDENTIFICACIÓN

Se presenta en este apartado la descripción de datos de la sección de identificación.

Tabla 6 Descripción de Datos – Sección Identificación

Nombre	Valores	Tipo	Tratamiento
Id	1..n	Entero	Omitir
Periódico	El Mundo, La Razón…	Texto	Conjunto
Edición	Nacional, Comunidad Valenciana	Texto	Conjunto
Año	2009	Entero	Conjunto Ordenado
Mes	Enero,Febrero…	Texto	Conjunto Ordenado
Dia	1..31	Entero	Conjunto Ordenado

Nombre	Valores	Tipo	Tratamiento
Género	Artículo, Reportaje…	Texto	Conjunto
Sección	Cultura, Sucesos…	Texto	Conjunto
Sección Otros	Crónica	Texto	Omitir
Página	1..n	Entero	Omitir
Extensión	1..n	Entero	Conjunto Ordenado
Forma de Aparición	Dedicación Principal/Incrustación más imagen/Referencia más imagen/ Incrustación/ Referencia / Sin Referencia	Texto	Conjunto Ordenado

8.2.2.- SECCIÓN DE FORMA

Se presenta en este apartado la descripción de datos de la sección de forma.

Tabla 7 Descripción de Datos – Sección de Forma

Nombre	Valores	Tipo	Tratamiento
Ubicación en el Periódico	Portada/Contraportada/ Página Impar Apertura Sección / Página Impar / Página Par Apertura Sección /Página Par	Texto	Conjunto Ordenado
Valor Ubicación	2,5,10,20,30	Entero	Rango
Jerarquía de la página	Portada Noticia Principal/ Portada Noticia no Principal / Contraportada Noticia Principal / Contraportada Noticia no Principal / Apertura de Página / No Apertura de Página	Texto	Conjunto Ordenado
Valor Jerarquía	1,2,5,10	Entero	Rango
Altura del Titular	Una Línea, Dos Líneas, Tres Líneas, Más de Tres Líneas	Texto	Conjunto Ordenado
Valor Altura Titular	1,2,3,5	Entero	Rango
Ancho del Titular	Menos de 1/4 de página / Entre 1/4 de página y 1/2 de página / Entre 1/2 de página y 3/4 de página / Entre 3/4 de página y 1 toda la página	Texto	Conjunto Ordenado

Nombre	Valores	Tipo	Tratamiento
Valor Ancho Titular	1,2,5,10,12	Entero	Rango
Superficie del Titular	Entre 1 y 4 módulos / Entre 5 y 9 módulos / Más de 9 Módulos	Texto	Conjunto Ordenado
Valor Superficie Titular	1,2,5	Entero	Rango
Cuerpo del Titular	Pequeño / Mediano / Grande	Texto	Conjunto Ordenado
Valor Cuerpo Titular	1,2,5	Entero	Rango
Jerarquía del Titular	Jerarquía 1 en la página / Jerarquía 1 compartida en la página / Jerarquía 2 en la página / Jerarquía 2 compartida en la página / Otros	Texto	Conjunto Ordenado
Valor Jerarquía Titular	0,2,5,7,10	Entero	Rango
Nº Imágenes	Ninguna, 1, 2, 3, más de 3	Texto	Conjunto Ordenado
Valor Nº Imágenes	0, 2, 4, 7, 10	Entero	Rango
Ilustración prioritaria	SI/NO	Booleano	Conjunto Ordenado
Valor Ilustración prioritaria	0,5	Entero	Rango
Tipografía o página especial	SI/NO	Booleano	Conjunto Ordenado
Valor Tipografía o página especial	0,5	Entero	Rango
Valoración Unidad de Análisis	0..100	Entero	Rango

8.2.3.- SECCIÓN DE CONTENIDO

Se presenta en este apartado la descripción de datos de la sección de contenido.

Tabla 8 Descripción de Datos – Sección de Contenido

Nombre	Valores	Tipo	Tratamiento
Análisis del Tono	Ambivalente / Neutro / Predominio de Frases Positivas / Predominio de Frases Negativas	Texto	Conjunto Ordenado
Tema Principal	11. Tráfico de drogas, alijos importantes 12. Tráfico de drogas, menudeo 13. Producción de sustancias 14. Tráfico de drogas, en general 21. Consecuencias relacionadas con la conducción 22. Consecuencias: conflictos/delitos 221. Celebrities 222. Violencia contra la mujer 23. Consecuencias sobre la salud física 24. Consecuencias sobre la salud psíquica 31. Datos sobre consumo de drogas 32. Datos sobre trastornos adictivos comportamentales 33. Dopaje 34. Datos sobre consumo en general 41. Prevención 42. Institucional 43. Presentación de estudios y resultados de investigaciones 44. Famosos y prevención 51. Ocio 52. Otros	Texto	Conjunto
Tema Secundario	Mismos Valores que Tema Principal	Texto	Omitir

8.2.3.1.- Subsección de Sustancias

Se presenta en este apartado la descripción de datos de la subsección de sustancias, dentro de la sección de contenido.

Tabla 9 Descripción de Datos - Subsección de Sustancias

Nombre	Valores	Tipo	Tratamiento
Tabaco	0.Ausente/1.Presente	Texto	Conjunto
Alcohol	0.Ausente/1.Presente	Texto	Conjunto

Nombre	Valores	Tipo	Tratamiento
Cannabis	0.Ausente/1.Presente	Texto	Conjunto
Hachis	0.Ausente/1.Presente	Texto	Conjunto
Marihuana	0.Ausente/1.Presente	Texto	Conjunto
Cocaína	0.Ausente/1.Presente	Texto	Conjunto
Heroína	0.Ausente/1.Presente	Texto	Conjunto
Crack/Cocaína Base	0.Ausente/1.Presente	Texto	Conjunto
Cristal	0.Ausente/1.Presente	Texto	Conjunto
Éxtasis o MDMA	0.Ausente/1.Presente	Texto	Conjunto
Anabonizantes	0.Ausente/1.Presente	Texto	Conjunto
Psicofármacos	0.Ausente/1.Presente	Texto	Conjunto
Drogas en General	0.Ausente/1.Presente	Texto	Conjunto
Otra	0.Ausente/1.Presente	Texto	Conjunto
Otra Droga	"*"	Texto	Omitir

8.2.3.2.- Subsección de Encuadre o *Frame*

Se presenta en este apartado la descripción de datos de la subsección de Encuadre, dentro de la sección de contenido.

Tabla 10 Descripción de Datos - Subsección de Encuadre

Nombre	Valores	Tipo	Tratamiento
Nueva Investigación	0 Ausente/1 Presente /2 Destacado en el tema principal, en la entradilla	Texto	Conjunto
Contexto General Científico/Médico	0 Ausente/1 Presente /2 Destacado en el tema principal, en la entradilla	Texto	Conjunto
Ética/Moralidad	0 Ausente/1 Presente /2 Destacado en el tema principal, en la entradilla	Texto	Conjunto

Nombre	Valores	Tipo	Tratamiento
Estrategia Política	0 Ausente/1 Presente /2 Destacado en el tema principal, en la entradilla	Texto	Conjunto
Política/Legislación	0 Ausente/1 Presente /2 Destacado en el tema principal, en la entradilla	Texto	Conjunto
Mercado/Empresa	0 Ausente/1 Presente /2 Destacado en el tema principal, en la entradilla	Texto	Conjunto
Epidemiología	0 Ausente/1 Presente /2 Destacado en el tema principal, en la entradilla	Texto	Conjunto
Opinión Pública	0 Ausente/1 Presente /2 Destacado en el tema principal, en la entradilla	Texto	Conjunto
Opinión no Experta	0 Ausente/1 Presente /2 Destacado en el tema principal, en la entradilla	Texto	Conjunto
Personalización Anecdótica	0 Ausente/1 Presente /2 Destacado en el tema principal, en la entradilla	Texto	Conjunto
Delito	0 Ausente/1 Presente /2 Destacado en el tema principal, en la entradilla	Texto	Conjunto

8.2.3.3.- Subsección de Fuentes

Se presenta en este apartado la descripción de datos de la subsección de fuentes, dentro de la sección de contenido.

Tabla 11 Descripción de Datos - Subsección de Fuentes

Nombre	Valores	Tipo	Tratamiento
Fuente Manifiesta	1. Nombre propio 2. Redacción 3. Agencia 4. Corresponsal 5. Enviado especial 6. No figura 7. Otros	Texto	Conjunto
Firma Periodista	"*"	Texto	Libre

Nombre	Valores	Tipo	Tratamiento
Análisis Cuantitativo de Fuentes	0..n	Entero	Rango
Políticos	0..n	Entero	Rango
Fuerzas y Cuerpos de Seguridad del Estado	0..n	Entero	Rango
Tribunal	0..n	Entero	Rango
Ciencia	0..n	Entero	Rango
Académicos	0..n	Entero	Rango
Privada	0..n	Entero	Rango
Psicosanitarios	0..n	Entero	Rango
No Expertos	0..n	Entero	Rango
Nombre de Persona	"*"	Texto	Libre
Afiliación institucional	0. Ausente / 1. Presente	Texto	Conjunto
Nombre de la Institución	"*"	Texto	Libre
Cargo	0. Ausente / 1. Presente	Texto	Conjunto
Nombre del Cargo	"*"	Texto	Libre

8.3.- EXPLORACIÓN DE LOS DATOS

En el presente capítulo mostraremos algunos datos estadísticos de cada uno de los campos a estudiar. La muestra tiene un total de 502 registros.

En primer lugar, se presentan en el anexo *13.2.1.-* Anexo: Estadísticos de Datos Numéricos algunos estadísticos de los campos numéricos. De entre los datos analizados destacar:

- Aparece un valor anómalo de Extensión de 0 palabras.
- La media de valoración de unidad de análisis es 29,388 sobre 100.
- El máximo de fuentes es de 8, siendo la media de 1,209.

- Los campos Fuerzas, Tribunal, Privada y Psico-Sanitarios aparecen como numérico, cuando deberían de ser categóricos. Además es destacable que cuando no aparecen no se codifican como el valor 0, sino que el valor queda en blanco, lo que provoca un valor nulo.

A continuación se presentan algunas propiedades estadísticas sobre los campos con valores discretos, obtenidas de los datos del *13.2.2.*-Anexo: Estadísticos de Datos Discretos*:*

- Las variables en las que aparecen textos libres tienen gran cantidad de valores nulos. Estas variables son Sección Otros (139 valores válidos), Otra Droga (33), Nombre de la persona (46), Nombre de la Institución (99), Nombre del Cargo (37), Firma del Periodista (333).

- Muchas de las variables binarias (Presente/Ausente) sólo se codifican cuando están presentes, con lo que existen gran cantidad de valores nulos en estas también. Es el caso de Tema Principal (439 valores válidos), Tema Secundario (202), Políticos (86), Ciencia (18), Académicos (19), Afiliación Institucional (99), Cargo (43).

Se presenta en el anexo *13.2.3.*-Anexo: Frecuencias y Procentajes de Valores un estudio de frecuencias absolutas y porcentajes de cada campo[2]. Se han limpiado los datos y se han añadido estadísticas de campos que se han creado *ad hoc* para el presente estudio de datos. Para más información ver el apartado *9.1.*- Selección y Transformación de Datos*.*

Se observan las siguientes características destacables de los datos:

- Sección de Identificación:
    - La edición nacional corresponde por completo al diario ABC.
    - El periódico “El País” tiene una línea editorial muy distante a los otros tres periódicos elegidos para la muestra.
    - Existe una ligera tendencia a publicar más textos en domingo, ya que hay un 17,73% en la muestra -89 casos- cuando la probabilidad aleatoria es de 14,28%.
    - Tan sólo hay 25 textos con una extensión larga -4,98%-.

[2] Se omiten los campos nombre de la institución, nombre de la persona, nombre del cargo, otra droga para mejorar la legibilidad, ya que son campos de texto libre con muy poca frecuencia y que posteriormente serán omitidos. Se omite también la valoración de la unidad de análisis, por ser un campo numérico que se discretizará posteriormente.

- El género es principalmente la noticia -363 casos y 72,31%- seguido de lejos por el artículo -36 y 7,17%- y Otros -34 y 6,77%- y por Reportaje -31 y 6,18%-. El resto de valores no alcanzan el 3% de los casos.

- Respecto a la sección, la mayoría de casos se encuentran en secciones específicas de cada periódico, marcado como "Otros", con 135 casos y 26,89% de casos, seguido por Sociedad -83 casos y 16,53%- y Nacional/Política/España -75 casos y 14,94%-. Es destacable la sección de suplemento de salud con 42 casos y 8,37%.

- La forma de aparición "Dedicación Principal" predomina en 411 casos, lo que representa un 81,87% de la muestra. El resto de valores quedan todos por debajo del 6% de los casos.

- Sección de Forma:

  - La ubicación más habitual es Página par normal, con 216 casos y 43,03% de los casos, seguida de Página impar normal, con 170 casos. Destacar que los textos sólo obtuvieron 29 portadas -5,78% de la muestra-. Sumando los casos de páginas par por una parte, y páginas impar por otra, se obtiene que los textos aparecen en su mayoría en páginas pares -261 casos frente a 210-.

  - Respecto a la valoración de unidad de análisis discreta, los grupos están balanceados con cierta tendencia a encuadrar casos en valores superiores. Ver el apartado *9.1.-* Selección y Transformación de Datos para más información sobre cómo está creado este campo.

  - A la mayoría de los textos le acompaña una imagen -234 casos y 46,61%- o sin imagen -222 casos y 44,22%-.

- Sección de Contenido:

  - La categoría de tema principal Ocio tiene 20 ocurrencias -3,98%- y la categoría Otros tan sólo 8 casos con 1,59% del total.

  - El tráfico de drogas es categoría de tema principal más abundante, con el 32,07% de casos y 161 ocurrencias.

  - Subsección de Sustancias:

    - Los Anabonizantes tienen un presencia muy baja. Sólo el 1% con 5 casos.

    - El Cannabis tiene una presencia muy baja con 5,98% de los casos y 30 ocurrencias.

- La cocaína está fuertemente presente con 126 casos -25,10%-.
- Las drogas de diseño están prácticamente ausentes en la muestra. El Crack/Cocaína base y el Cristal aparecen en muy pocas ocasiones, tan sólo hay 4 casos -0,8%- y 1 caso -0,2%-. El éxtasis/MDMA por su parte, tan sólo en 9 ocasiones con un 1,79%.
- En 192 ocasiones -38,25%- se habla de drogas en general.
- El hachís aparece en 40 textos, con un 7,97% del total. La marihuana por su parte, aparece en 35 casos, con un 6,97%.
- La heroína aparece el 6,77% de los casos, con sólo 34 casos.
- El Tabaco aparece presente en 113 casos con un 22,51%. El alcohol aparece en 103 textos con un 20,52% sobre la muestra.

o Subsección de Encuadre o *Frame:*

- Los textos se encuadran en un contexto general científico/médico tan sólo el 6'57% de las ocasiones. La muestra contiene sólo 33 ocurrencias.
- El Delito es, con mucho, la forma de encuadre más habitual con 256 ocasiones -51%-.
- El enfoque ético o de moralidad está presente en sólo 32 casos -6,37%-.

o Subsección de Fuentes:

- La fuente ciencia tiene una ocurrencia muy baja. Tan sólo 18 casos con un 3,59%.
- La fuente Fuerzas aparece en 116 ocasiones -23,11%- de los casos.
- La fuente tribunal está presente en pocas ocasiones, con 31 casos y 6,18% sobre el total.
- La opinión no experta aparece en sólo 18 ocasiones y 3,59% de la muestra.
- El Cargo aparece en el 7,17% de los textos, 36 veces en valor absoluto.

- La fuente manifiesta es el Nombre Propio en 304 ocasiones -60,56%-. Inmediatamente después está el valor “No Figura” con 114 casos y 22,71%. El resto de valores no alcanzan el 5% de los casos.

8.4.- CALIDAD DE LOS DATOS

La muestra de datos está bastante depurada, ya que se sigue un método de introducción de datos estándar entre los diferentes codificadores. Aún así, hay datos que no se codifican si no aparecen, lo que da lugar a valores nulos. Otros contienen espacios en blanco, otros se tratan como texto cuando deberían ser numéricos, etc.

Se han hallado las siguientes peculiaridades acerca de la calidad de los datos:

- En el número de páginas faltan 2 registros.
- Hay extensiones con tamaño 0, lo cual no puede ser correcto.
- Los campos de fuentes deberían de ser numéricos, pero sólo lo son Fuerzas, Tribunal, Privada, Psico-sanitarios y No expertos son numéricos.
- Falta el número de imágenes en 1 registro.
- Faltan temas principales. Esto ocurre cuando el tema principal no son las drogas.
- Faltan temas secundarios que no se codifican si no hay temas secundarios.
- Faltan entradas en otra droga que no se codifica si no aparece una droga que no tiene campo asignado.
- No se codifican si no aparecen, firma periodista, Nombre de la persona, Afiliación institucional, Nombre de la institución, Cargo, Nombre del Cargo.
- Faltan por codificar muchos campos fuentes.
- Los campos de texto libre están codificados de forma que varios valores pueden referirse a una misma entidad, por ejemplo, “Oficina de Drogas y Delitos de la ONU” y “Oficina de la ONU contra la Droga y el delito”.
- Debido a diferentes transformaciones, existen espacios perdidos antes o después de los valores de los campos que pueden hacer que dos valores idénticos se evalúen como diferentes, por ejemplo:
    - “1. Presente”
    - “ 1. Presente”

9.- PREPARACIÓN DE LOS DATOS PARA EL PROCESO DE MODELADO

En el presente capítulo se explicará la preparación de los datos que se ha propuesto para obtener una vista minable. Esta preparación tratará de:

- Eliminar valores anómalos, inconsistencias, valores ausentes, etc.
- Seleccionar datos a tratar
- Modificar valores para su mejor tratamiento en función del algoritmo seleccionado o de una mayor representatividad.

9.1.- SELECCIÓN Y TRANSFORMACIÓN DE DATOS

El primer paso que se va a tener en cuenta es la selección de campos a utilizar en el proceso de Minería de Datos. No es necesario integrar datos de diferentes fuentes, por tanto, habrá sólo un proceso de trasformación, pues todos los datos provienen de las mismas fuentes.

Los campos que se descartarán[3] son los siguientes:

- Id: se trata de un identificador único. No aporta ninguna información
- Edición: Hay sólo un diario que tiene edición "Nacional" por lo que este campo no aporta información.
- Año: Todos los registros corresponden a 2009. Este campo no aporta información.
- Sección Otros: Se trata de un campo de Texto Libre. Esto permite errores por parte del codificador, como hemos señalado anteriormente. Además, la baja frecuencia de cada valor hace que la información aportada sea muy baja.
- Página: el número de página dentro de un periódico no aporta información.
- Los campos de la sección Forma se descartarán, salvo la Ubicación en el Periódico, ya que el campo "Valoración de la unidad de Análisis" es un campo calculado con los valores de todos ellos. Se trabajará con este campo.
- El análisis del tono: se descartará inicialmente, pues no es relevante para los objetivos actuales del equipo de investigación periodística.
- El tema secundario: se descartará, ya que tiene multitud de campos vacíos y sólo aparece cuando el tema principal no es de drogas.

[3] Se decartarán también los campos numéricos originales una vez hayan sido derivados en campos de datos discretos. Ver más adelante en este mismo apartado la construcción de campos derivados.

- Otra droga: se descartará por ser campo de Texto Libre, por los mismos motivos que Sección Otros.
- Firma del periodista: se descartará por ser campo de Texto Libre, por los mismos motivos que Sección Otros.
- Nombre de Persona: se descartará por ser campo de Texto Libre, por los mismos motivos que Sección Otros.
- Nombre de la Institución: se descartará por ser campo de Texto Libre, por los mismos motivos que Sección Otros.
- Nombre del Cargo: se descartará por ser campo de Texto Libre, por los mismos motivos que Sección Otros.

Las tareas de limpieza de datos a realizar serán las siguientes:

- Conversión de Tipos:
    - Los campos de fuentes deberían de ser numéricos, pero sólo lo son Fuerzas, Tribunal, Privada, Psico-sanitarios y No expertos son numéricos.
- Modificar datos erróneos:
    - Falta el número de imágenes en 1 registro. Poner a 0
- Añadir datos faltantes:
    - Insertar nuevo tema principal: "0. No Relacionado con Drogas".
    - Insertar el valor 0 en los campos de fuentes vacíos.
- Trimming:
    - Eliminar espacios perdidos antes y después de cadenas.

Una vez seleccionados y limpiados los datos, se procederá a crear nuevos campos derivados para permitir al algoritmo seleccionado[4] trabajar con los datos de la muestra, o bien se agruparán valores de la muestra para aumentar su representatividad.

Para ello se crearán los siguientes campos discretos:

- Cantidad de fuentes se discretizará de la siguiente forma, según criterio del equipo de investigación periodística:

[4] Para más información ver el apartado *10.1.*- Selección de la técnica de modelado

- 0 → “Pocas Fuentes”
- 1 → “Suficientes Fuentes”
- >1 → “Muchas Fuentes”

- Categoría de Tema Principal se agrupará siguiendo la descripción de niveles del documento de codificación del grupo de investigación y las sugerencias del equipo de investigación. Esto permitirá obtener más representatividad en esta variable. Así pues, queda de la siguiente forma:
  - 0. No Relacionado con Drogas → Corresponde a campos vacíos
  - 1. Trafico de Drogas →
    - 11. Tráfico de drogas, alijos importantes
    - 12. Tráfico de drogas, menudeo
    - 13. Producción de sustancias
    - 14. Tráfico de drogas, en general
  - 2. Consecuencias →
    - 21. Consecuencias relacionadas con la conducción
    - 22. Consecuencias: conflictos/delitos
    - 221. Celebrities
    - 222. Violencia contra la mujer
    - 23. Consecuencias sobre la salud física
    - 24. Consecuencias sobre la salud psíquica
  - 3. Consumo →
    - 31. Datos sobre consumo de drogas
    - 32. Datos sobre trastornos adictivos comportamentales
    - 33. Dopaje
    - 34. Datos sobre consumo en general
  - 4. Estudios y Prevención →

        - 41. Prevención
        - 42. Institucional
        - 43. Presentación de estudios y resultados de investigaciones
    - 5. Ocio → 51. Ocio
    - 6. Otros → 52. Otros

- Las fuentes se discretizarán de forma que tomarán los siguientes valores:
    - 0 → "0. Ausente"
    - >1 → "1. Presente"
- El día de la semana pasará a ser un campo binario, que indicará si el día es domingo o bien, no lo es. Esta transformación fue sugerida por el equipo de investigación periodístico debido a que es representativo el que la noticia aparezca en domingo o no. Para ello se deberá realizar un cálculo.
- La extensión del artículo se discretizará según criterio del equipo de investigación periodística en los siguientes intervalos:
    - <80 → 1. Corto
    - <80 y <180 → 2. Normal
    - >180 → 3. Largo
- La Valoración de la Unidad de Análisis se discretizará mediante la aplicación del criterio de los cuantiles, es decir, se crearán 3 grupos en los que se tratará de ubicar el mismo número de casos, pasándose los casos frontera al grupo superior. Así pues, la valoración de la unidad de análisis queda:
    - Alto >= 36
    - Medio >= 17
    - Bajo < 17
- Se modificarán ciertos valores para permitir una mayor representatividad semántica de la información. Se eliminarán los valores "2. Destacado" pasando a ser "1. Presente", ya que un valor destacado implica un valor presente.
- El mes se ha transformado en numérico para poder calcular el día de la semana, y para poder generar un grupo ordenado. En la muestra los meses comprenden un intervalo desde Enero (Mes 1) a Junio (Mes 6).

9.2.- TRANSFORMACIÓN ETL EN SPSS PARA OBTENCIÓN DE VISTA MINABLE

En el presente apartado se ilustrará cómo se realizan las transformaciones ETL (Extraction Transform y Load) en la aplicación SPSS Clementine para obtener la vista minable.

Clementine permite mediante un flujo de trabajo realizar transformaciones sobre los datos. Los flujos se componen de nodos y los nodos, a su vez, se pueden conectar con flechas a modo de diagrama dirigido. Hay varios tipos de nodos:

- Orígenes: nodos de origen de datos. Permiten especificar la fuente –o fuentes- de los datos.
- Operaciones con registros: Permiten seleccionar, filtrar o muestrear registros entre otras cosas.
- Operaciones con campos: Permiten realizar modificaciones en los campos, así como filtrarlos. También pueden crear campos derivados, reclasificar los valores, etc.
- Gráficos: Permiten realizar gráficos con los datos.
- Modelado: Contiene nodos con los diferentes algoritmos de modelado de Minería de Datos.
- Resultado: Permiten Obtener resultados, evaluar modelos, evaluar datos, visualizar datos de forma tabular, etc.
- Exportar: Permiten exportar los resultados.

En este caso se tratará de realizar un proceso de limpieza y transformación de los datos para conseguir una vista minable que pueda procesar el algoritmo escogido.

Para ello, se comenzó creando una nueva ruta cuyo aspecto final será el siguiente:

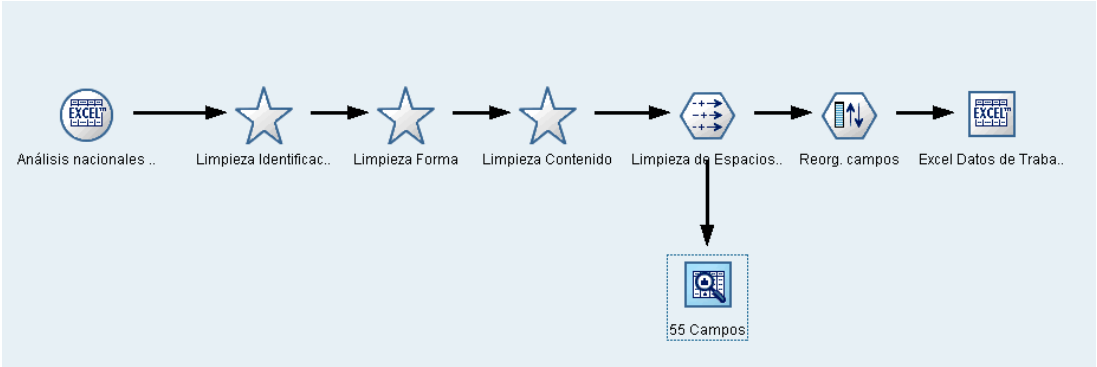

Figura 44 Ruta de Limpieza y Transformación Clementine

En la figura anterior se pueden observar varios nodos en forma de estrella, que son nodos especiales llamados "supernodos" que representan una agrupación de nodos. Estos supernodos permiten que no aparezcan multitud de nodos en la ruta principal, para facilitar la comprensión de la misma y el trabajo.

En la siguiente figura se puede observar el fichero de rutas y cómo cuelgan los supernodos a modo de árbol.

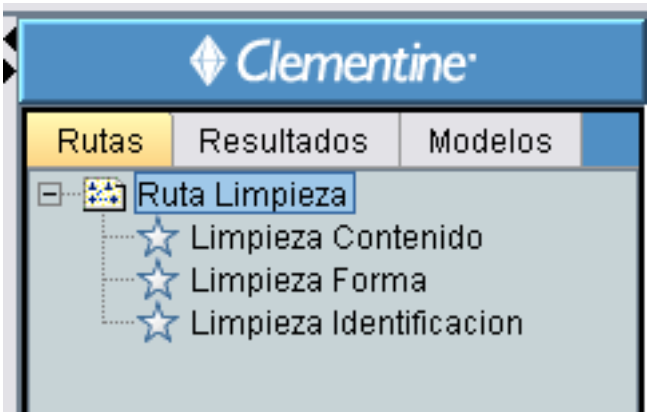

Figura 45 Rutas y Supernodos

En la ruta anterior se tienen los nodos que describimos a continuación.

Un Nodo de Origen Excel apuntando a los datos del fichero “Análisis Nacionales FEPAD 2.xls”. Este nodo –y todos los demás- se puede configurar mediante pantallas como la siguiente:

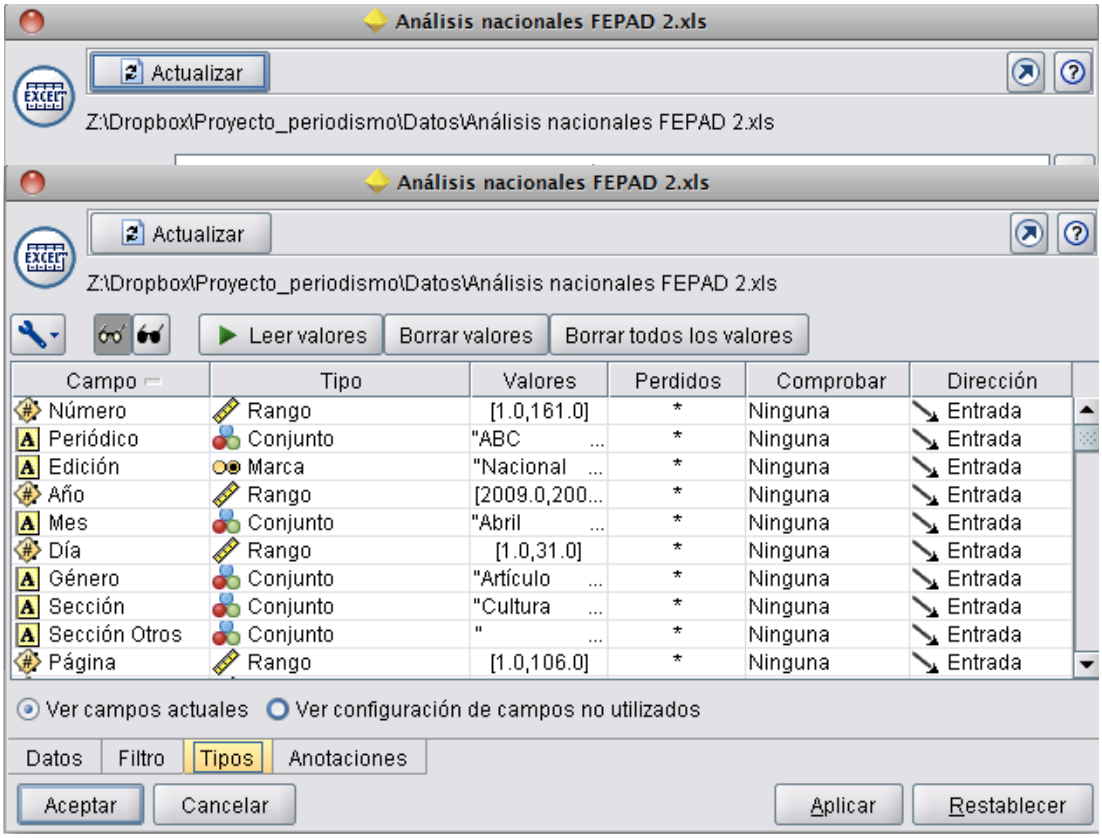

Figura 47 Tipos de datos del nodo origen Excel

Se puede observar en esta pantalla que el programa permite seleccionar los datos a importar así como configurar varias opciones. Es destacable que en las pestañas inferiores permite filtrar los datos a importar por campos, y sobretodo, permite asignar tipos de datos a los datos importados.

En la figura anterior puede observarse cómo trata los valores Clementine. Llama rango a los valores numéricos, Conjunto a los discretos, Marca a los booleanos, etc. Además, se tienen las siguientes columnas:

- La columna perdidos permite establecer qué hacer con los datos faltantes.
- La columna comprobar permite realizar comprobaciones sobre los datos y correcciones.

- La columna Dirección permite especificar si el campo se tomará como entrada, salida, ambas, o nada en el momento de modelar con un algoritmo concreto de Minería de Datos.

Los supernodos de la Figura 44 son los que realizan la limpieza y transformación de datos. Cada uno se centra en un grupo de campos de la muestra. Esto es: Identificación, Forma y Contenido. Se explicarán con detalle más adelante.

Inmediatamente después hay un nodo para eliminar espacios sobrantes en los distintos campos de texto.

Después hay un nodo terminal de auditoría de datos que permite evaluar si todas las transformaciones han sido correctas.

Tras el nodo de *trimming* hay un nodo de reorganización de campos que permite cambiar el orden de aparición. No es necesario, pero se ha incluido por comodidad, para facilitar posteriormente el trabajo.

Al final hay un nodo terminal de exportación, que escribe los datos limpios y transformados en un nuevo fichero llamado “Datos de Trabajo.xls”.

Se describen ahora los supernodos de limpieza y transformación.

El primer nodo, que trata el grupo de campos de identificación, tiene el siguiente aspecto:

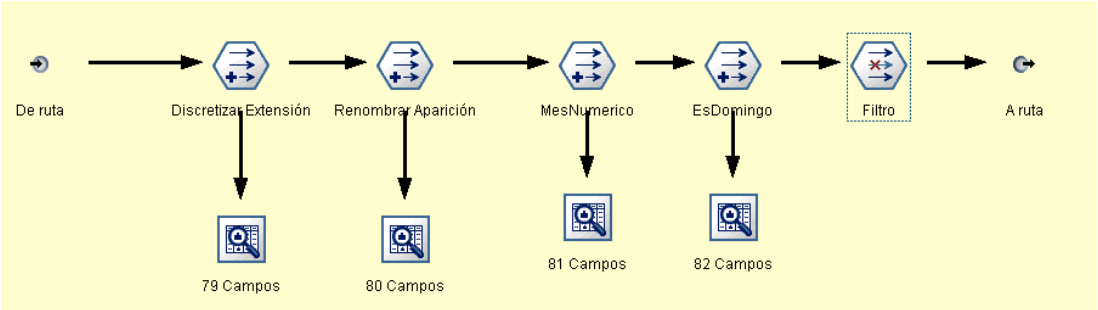

Figura 48 Ruta ETL Identificación

En primer lugar se discretiza la extensión con los parámetros antes descritos, luego se renombra la Forma de Aparición para generar campo derivado que sea un Conjunto Ordenado, inmediatamente después se calcula el mes numérico y después se crea el campo derivado EsDomingo. Para finalizar se filtran los campos que ya no se necesitan: Mes, Día, Extensión, Forma de Aparición, etc.

El segundo supernodo, que trata el grupo de campos de forma, tiene el siguiente aspecto:

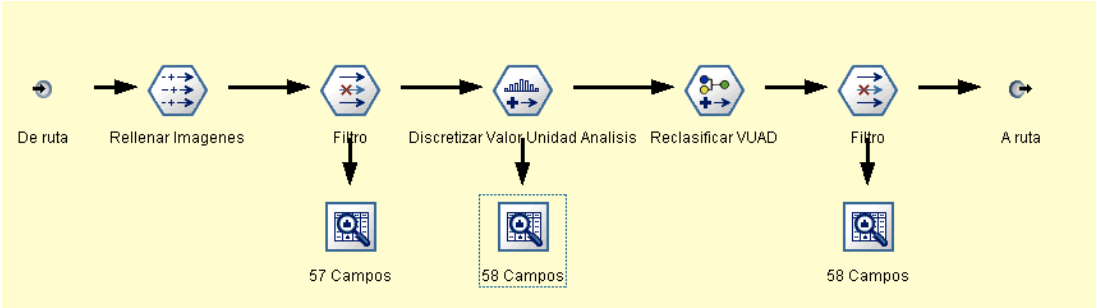

Figura 49 Ruta ETL Forma

El primer nodo que se puede observar en la Figura 49 es un nodo de relleno de datos, que añade el valor “Ninguna” a los valores vacíos del campo “Nº de imágenes”. Posteriormente se filtran todos los campos de la unidad de análisis con la salvedad de las imágenes y la valoración de la unidad de análisis.

Inmediatamente después se generan los tres cuantiles del campo de valoración de la unidad de análisis con los siguientes parámetros:

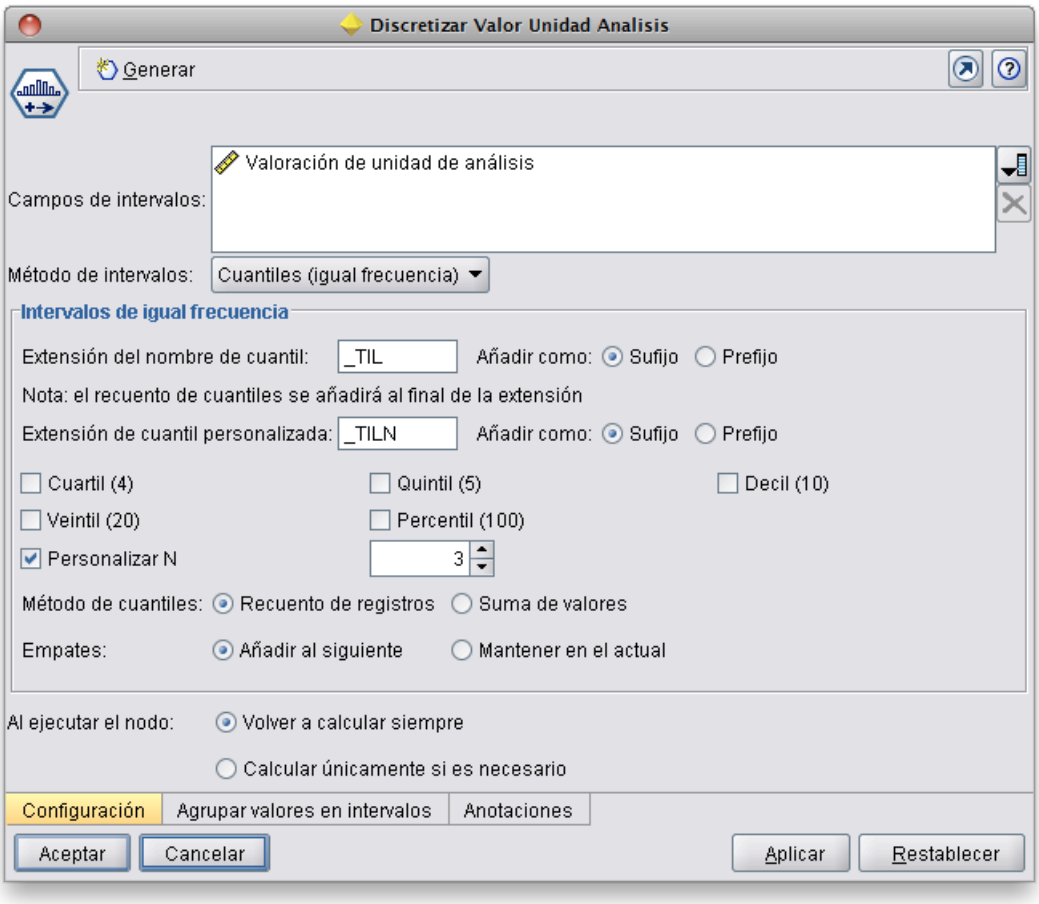

Figura 50 Discretización de la Valoración de la Unidad de Análisis

Inmediatamente después se tiene un nodo de reclasificación, que crea el campo Valoración de Unidad de Análisis Discreto con los valores del anterior nodo, siendo valores posibles las cadenas “Alto”,”Medio y “Bajo”.

El tercer supernodo, que realizar la transformación ETL de los campos de contenido, tiene el siguiente aspecto:

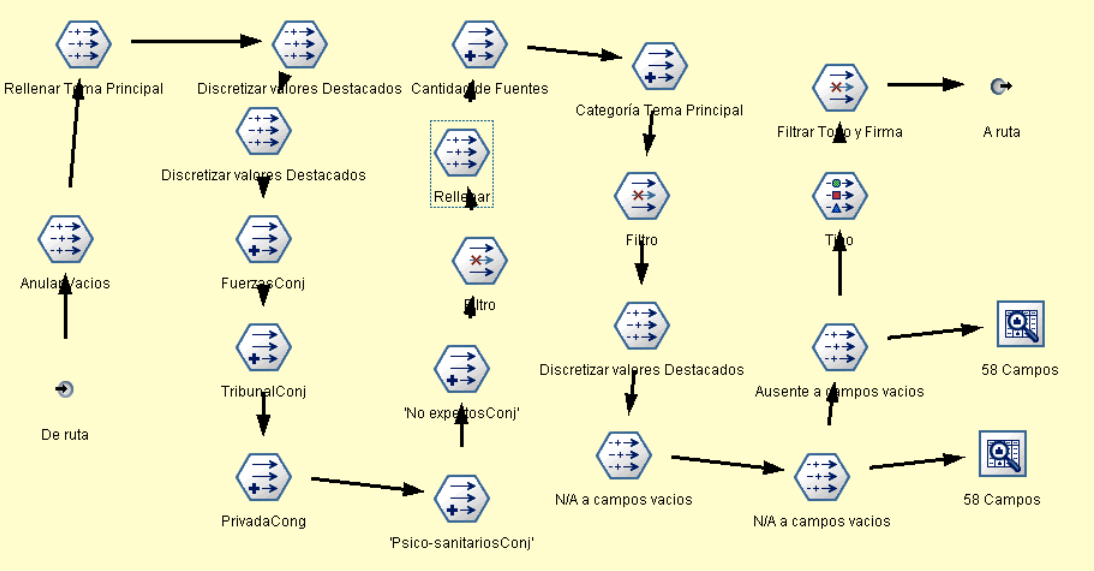

Figura 51 Ruta ETL de Contenido

El primer nodo que se observa es “Anular Vacios”. Este nodo trata de poner valores nulos en aquellos campos que están vacíos para después poder rellenarlos con valores “N/A”, “0. Ausente”, etc. según corresponda.

Después de este nodo, se tiene un nodo que rellena con “00. No relacionado con drogas” los valores anulados del campo Tema Principal.

Le sigue un nodo que permite cambiar los valores destacados del *Frame* por “0. Presente”. El siguiente sustituye los valores nulos por “0. Ausente”.

Los siguientes nodos hasta “No ExpertosConj” crean campos derivados con dos valores “0. Ausente/1. Presente” teniendo en cuenta los distintos campos de fuentes. Después, en el nodo “Filtro” se filtran los campos numéricos antiguos y se renombran los nuevos.

En el siguiente nodo “Rellenar” se rellenan con “0. Presente/1. Ausente” los campos de fuentes políticos, ciencia y académicos.

En el siguiente nodo “Cantidad de Fuentes” se genera un campo derivado con el número de fuentes discretizado.

En el siguiente nodo se genera el campo derivado “Categoría de Tema Principal”. En el siguiente nodo se filtra el tema principal.

En el siguiente nodo se trata de eliminar espacios perdidos en cadenas.

En los dos nodos “N/A a campos vacios” se inserta la cadena “N/A” en los campos de texto que no contengan valor alguno. En el siguiente nodo se codifica como “0. Ausente” los valores de los campos booleanos que no contengan ningún valor.

En el siguiente campo se preparan los campos como entradas para los modelos de Minería de Datos y finalmente se filtran los campos sobrantes Tono y Firma.

Una vez llegados a este punto se tiene la Vista Minable preparada para ser utilizada. En el presente caso, se insertó la vista en un nuevo fichero excel para poder facilitar su importación y, sobre todo, evitar tener que realizar la computación de toda la ruta de limpieza en cada transformación.

10.- MODELADO

En el presente capítulo se describen las técnicas de modelado consideradas, la constricción del modelo y la evaluación del mismo. Posteriormente, en el siguiente capítulo, se evaluarán los resultados.

10.1.- SELECCIÓN DE LA TÉCNICA DE MODELADO

En la selección de la técnica de modelado a aplicar han pesado diferentes factores. Por un lado la disponibilidad técnica de los diferentes modelos en la herramienta y por otro las necesidades del estudio. La selección se realizó teniendo en cuenta las características descritas en la clasificación realizada en el apartado *3.-* Técnicas de Minería de Datos*.*

Como indica [Hernández Orallo, et al. 2004], una de las principales características que se deben tener en cuenta en la selección de la técnica a utilizar es la adecuación al objetivo de la Minería de Datos. Por tanto, siguiendo los objetivos que se propusieron en el apartado *7.3.2.-* Objetivos Específicos se realizaron las siguientes selecciones previas de algoritmos en función de los objetivos descritos.

El objetivo “Realizar un estudio estadístico genérico de los datos” ya se ha realizado en apartados previos.

Para el objetivo “Estudiar qué variables influyen en el valor final de cada una de las variables.” se tienen modelos en la herramienta que se utliizarán para este fin. Se describirán más adelante.

Para los objetivos “Comprobar o refutar de forma motivada los resultados del grupo de investigación”, “Hallar nuevas relaciones en la muestra, con relevancia para el grupo de investigación periodística”, “Hallar nuevas relaciones en la muestra, atendiendo a la ausencia de variables”, “Hallar nuevas relaciones en la muestra, atendiendo a subpoblaciones”, se evaluaron dos posibilidades: Algoritmos predictivos y algoritmos de reglas de asociación.

La comprensión de los resultados por parte del equipo de investigación es una de las variables que se han evaluado para la selección de la técnica de modelado. Se ha de tener en cuenta que los resultados deben ser entendibles y fácilmente explicables. En este sentido los algoritmos predictivos pueden realizar una buena predicción de datos con una muestra grande, pero algunos de ellos pueden resultar difíciles de comprender. Por otro lado, los modelos de reglas de asociación son fáciles de entender en general, siempre que no se tenga una cantidad de antecedentes excesiva.

Además, se tuvo en cuenta que muchos de los datos tienen el formato típico de problema de “cesta de la compra” que son bien evaluados con algoritmos de reglas de asociación.

Algunos de los objetivos a cumplir implican descubrir nuevas relaciones en la muestra, que resulten significativas para los integrantes del grupo de investigación periodística. Esto nos lleva de nuevo a optar por los algoritmos de reglas de asociación.

Una vez establecida la conveniencia de trabajar con los algoritmos de reglas de asociación, se selecciona de entre ellos el algoritmo Apriori, que es, de los descritos, el que se tiene disponible en la herramienta SPSS Clementine. El algoritmo GRI, propuesto por [Smyth, et al. 2002], pese a que es capaz de tomar valores numéricos como antecedentes no es adecuado ya que escala peor que Apriori con el tamaño de los datos, debido a que tiene que calcular una medida de interés J-Measure basada en las probabilidades de las variables. Además, utiliza esta medida para podar los resultados antes de presentarlos, con lo que es difícil controlar su exhaustividad.

El algoritmo CARMA, propuesto por [Hidber. 1999], funciona básicamente igual que Apriori teniendo todos los campos como entrada y salida. Además, permite ajustar el soporte de las reglas –la frecuencia en que antecedente y consecuente aparecen juntos-, y necesita un identificador por cada fila de datos. Apriori, en este sentido, permite evaluar el soporte del antecedente y generar una regla aunque la regla en sí tenga bajo soporte.

La aparición de algunos campos numéricos hará necesaria su discretización, ya que Apriori no puede trabajar con variables no discretas.

Por otro lado, el tamaño de la muestra y la cantidad de campos hacen que la aplicación directa de un algoritmo Apriori puede conllevar dificultades debido a la necesidad de generación de todos los candidatos posibles. Esta generación no escala bien con el tamaño de la muestra y la cantidad de variables a utilizar. En el apartado siguiente se expone la metodología que se ha seguido para evitar este problema con el algoritmo.

10.2.- CONSTRUCCIÓN DEL MODELO

En el apartado anterior se ha seleccionado el algoritmo Apriori para el modelado de datos. [Hipp, et al. 2000] realizan una revisión de los algoritmos de reglas de asociación e indican que Apriori no escala bien para una cantidad de datos como los de la muestra. Además si lo que se necesita es obtener datos con soporte y/o confianza bajos, el rendimiento del algoritmo empeora considerablemente, puesto que no es capaz de podar muchas reglas de bajo soporte. La opción que se ha seguido realizar evaluaciones de datos segmentadas por grupos de campos, cruzando valores, estudiando subpoblaciones, etc.

Para la construcción de los modelos se acordó con el equipo de investigación periodística exigir un mínimo de ***10 elementos de soporte en cualquier caso***, sin un mínimo de soporte relativo o confianza mínimo, en principio.

El algoritmo Apriori que implementa Clementine permite realizar podas de reglas habituales utilizando medidas de interés. No se van a utilizar debido a que se desean obtener tanto relaciones habituales como relaciones no habituales. Sin embargo, sí que se van a obtener los valores para identificar qué reglas son interesantes y cuáles no.

De la misma forma el algoritmo Apriori permite utilizar sólo los valores positivos de los campos binarios o bien, ambos valores. Por lógica booleana, el segundo análisis contiene al primero, pero también genera una cantidad muy superior de resultados a filtrar. Por ese motivo, se han realizado ambos análisis en cada uno de los modelos construidos. En el análisis con marcas positivas, se han reducido los requisitos de soporte y confianza, que han tenido que ser elevados en los análisis con todos los valores de las marcas para evitar obtener grandes conjuntos de reglas.

Según [Hernández Orallo, et al. 2004] aumentar el número de antecedentes en las reglas dificulta la comprensión y proporciona un bajo aporte de información novedosa. El número de antecedentes se ha establecido en 1 para poder observar las relaciones inmediatas de los elementos de cada análisis.

Se describen a continuación los modelos construidos. Se numerarán los modelos para su posterior identificación en la descripción de resultados.

Los análisis generales de los datos realizados son los siguientes:

1. Ruta General: Incluyendo todos los campos. Permitirá una visión general.
2. Ruta Drogas: Sólo los campos de sustancias. Permitirá saber cuándo aparecen conjuntamente las drogas.
3. Ruta *Frame*: Sólo los campos del Encuadre. Permitirá ver la concurrencia de encuadres.
4. Ruta Fuentes: Sólo los campos de Fuentes. Permitirá observar la concurrencia de las fuentes.

Los análisis individuales realizados se han centrado en obtener subpoblaciones de cada uno de los valores de los campos que se estudian. Por ejemplo, en el caso de la Categoría del Tema Principal, en el primer análisis, se ha filtrado la muestra para dejar sólo aquellos registros que tengan el valor “Tráfico de Drogas”. Los análisis realizados han sido los siguientes:

5. Análisis Individual de la Categoría del Tema Principal.
6. Análisis Individual por Cantidad de Fuentes.
7. Análisis Individual de cada Fuente

8. Análisis Individual de la Forma de Aparición
9. Análisis Individual de EsDomingo
10. Análisis Individual de Drogas
11. Análisis Individual de Extensión
12. Análisis Individual de *Frame*
13. Análisis Individual de Fuente Manifiesta
14. Análisis Individual de Género Periodístico
15. Análisis Individual de Ilustración: Se han considerado ilustradas aquellos textos con número de imágenes mayor que 0.
16. Análisis Individual por Meses
17. Análisis Individual por Periódicos
18. Análisis Individual por Sección
19. Análisis Individual por Ubicación
20. Análisis Individual por Valoración de Unidad de Análisis

El conocimiento del problema es un factor a tener en cuenta a la hora de abordar el análisis en Minería de Datos, por ello se solicitó al equipo de investigación periodística que sugiriera posibles análisis.

Además de los mencionados, el equipo de investigación periodística sugirió los análisis cruzados mostrados en la siguiente tabla:

Tabla 12 Análisis sugeridos por el Equipo de Investigación Periodística

	Valoración formal	Tema principal	Frame	Tipo de fuente	Droga
21. Valoración formal		x	x	x	x
22. Tema principal			x	x	x
23. Frame				x	x
24. Tipo de fuente presente					x

Por otro lado, se realizaron análisis de los resultados obtenidos por el grupo de investigación, centrándose en las siguientes observaciones:

25. La presencia de fuentes psico-sanitarias aumenta el valor de la unidad de análisis.
26. Tener imágenes en un texto aumenta la probabilidad de que exista una fuente de tipo tribunal.
27. Las drogas blandas no son portada.
28. La valoración de la unidad de análisis está relacionada con las fuentes.
29. Que la noticia verse sobre el tema "Celebrities" implica una mayor valoración de la unidad de análisis.

Además de lo anterior, durante la construcción de los modelos se sugirió hacer una análisis agrupado de drogas, de forma que se crearan cuatro categorías: Drogas Duras, Drogas Blandas, Drogas Legales y Drogas de Diseño. Se realizó un análisis de esta nueva agrupación de drogas:

30. Análisis Agrupado de Drogas

Por otro lado, se realizaron análisis para observar qué campos influyen en el valor final de una variable. Para ello se utilizó el nodo Selector de Características de Clementine que calcula la dependencia entre variables categóricas basándose en la distribución Chi-Cuadrado de Pearson. Para ello se realizaron los siguientes análisis:

31. Análisis de Campos Influyentes en Drogas
32. Análisis de Campos Influyentes en *Frame*
33. Análisis de Campos Influyentes en Fuentes
34. Análisis de Campos Influyentes en Otros: Incluye la categoría de tema principal, forma de aparición, valoración de unidad de análisis, número de imágenes, ubicación en el periódico, mes y EsDomingo.

10.3.- EVALUACIÓN DEL MODELO

La evaluación de un modelo predictivo supone la generación de un modelo con un conjunto de datos y su evaluación con otro diferente, para calcular su fiabilidad. [Hernández Orallo, et al. 2004] indica que la evaluación de reglas de asociación se puede realizar con su soporte y confianza.

Aún así, existen otras medidas de evaluación de reglas de decisión que se pueden clasificar en objetivas y subjetivas. Las evaluaciones subjetivas son observaciones realizadas por personas con conocimiento del problema, que miden diversos parámetros pero no responden a un criterio objetivo medible. Por el contrario, las medidas objetivas responden a estos criterios objetivos medibles.

Algunas medidas subjetivas son las siguientes:

- Comprensibilidad: se trata de una medida completamente subjetiva, puesto que lo que es comprensible para un usuario puede no serlo para otro. Así, en el caso de reglas de asociación el tener pocos antecedentes puede ayudar a su fácil comprensión. También hay que tener en cuenta en este sentido cuestiones semánticas y de presentación de resultados. Las discretizaciones suelen facilitar también la comprensión del modelo aunque nos hacen perder detalle. Por ejemplo, "Salario=alto" es más fácilmente comprensible que "Salario>50.000".

- Aplicabilidad: las reglas con poca aplicación práctica tienen poco interés para el observador. Si una regla predice un consecuente con un antecedente que en la realidad sucede en muy raras ocasiones, su aplicabilidad es muy pobre

- Novedad: los algoritmos de reglas de asociación a menudo se centran en reglas obvias y bien conocidas, que no aportan nada al observador. Por ello, es importante hallar reglas no obvias que aporten novedades.

[Geng, et al. 2006] realiza un estudio sobre las medidas de interés objetivas que se han propuesto en los últimos años. Siendo la regla a evaluar A → B se pueden definir algunas de ellas son:

- Soporte y confianza mencionados antes.

- Cobertura: Mide la probabilidad de ocurrencia de A.

- Prevalencia: mide la probabilidad de ocurrencia de B.

- Retorno: mide la probabilidad de B → A.

- Elevación o Interés: Mide cómo condiciona A a la aparición de B dividiendo la confianza por la probabilidad de B. Las medidas que se alejan de 1 implican una probabilidad condicionada mayor –o menor-.

En la Tabla 13 se pueden observar cómo se calculan estas medidas:

Tabla 13 Medidas de Interés basadas en Probabilidades

Medida	Cálculo
Soporte	P(AB)
Confianza/Precisión	P(B|A)
Cobertura	P(A)
Prevalencia	P(B)
Retorno	P(A|B)
Interés/Elevación	P(B|A)/P(B) o bien P(AB)/P(A)P(B)

En el presente estudio se han tenido en cuenta ambos tipos de evaluación para obtener resultados. En cuanto a las medidas subjetivas se han realizado reuniones periódicas con el equipo de investigación periodística en las que se ha ido obteniendo un *feedback* de los resultados obtenidos desde el punto de vista del usuario final.

En cuanto a las medidas de interés, se han utilizado tres: soporte, confianza y elevación. El uso de las medidas de elevación han sido determinantes para hallar reglas cuya precisión fuera estadísticamente relevante.

11.- Evaluación y Desarrollo del Informe Final

11.1.- Evaluación de los resultados

Teniendo en cuenta que se solicitó por parte del equipo de proyecto un soporte mínimo de 10 casos, se trabajó con un soporte cercano al 5% sobre la muestra global. La obtención de resultados fue, por tanto, una tarea laboriosa debido a la gran cantidad de los mismos que APRIORI obtiene cuando se utilizan niveles de soporte y confianza bajos. Además, no sólo la presencia de ciertos atributos binarios podía ser de interés, sino también la ausencia, con lo que el algoritmo APRIORI se preparó de forma que pudiese evaluar ausencias significativas. Esto agravó el problema de la cantidad de resultados.

Como ya se ha comentado, se han utilizado varias medidas de interés para seleccionar y filtrar resultados, entre ellas, la más importante ha sido la elevación, que ha permitido hallar qué reglas permiten condicionar la aparición del consecuente. Obteniendo valores de elevación por encima de 1, indica que la regla predice el consecuente con una confianza superior a si sólo se tiene en cuenta la probabilidad del consecuente. Si la elevación es inferior a 1, el razonamiento es el inverso.

11.2.- Resultados

Se presentan a continuación los principales resultados de los distintos análisis realizados. Se utilizará la misma numeración que en el apartado *10.2.-* Construcción del modelo. En los análisis generales se mostrarán prácticamente todos los resultados para ayudar a la comprensión del problema. En los análisis específicos se presentarán las reglas más relevantes halladas, y se omitirán resultados ya hallados, obvios o poco relevantes.

En los análisis generales se ordenarán los resultados por soporte, exigiendo un mínimo de un 2% -10,02 casos-, para poder observar las características más relevantes de la muestra. En el resto se ordenarán por elevación de forma que se pueda observar qué reglas aportan mayor información.

Hay que tener en cuenta que una elevación por debajo de 1 implica que el antecedente es un mal predictor para el consecuente, o lo que es lo mismo, que dado el antecedente, la confianza de ocurrencia del consecuente es inferior a la que se daría sin el antecedente. Por tanto, con elevación menor que 1, se debe entender que la regla indica que el antecedente condiciona reduciendo el soporte del consecuente.

En el análisis de resultados se mostrarán pares (soporte,confianza) de las reglas. En caso de que un antecedente obtenga más de un consecuente, en el segundo y sucesivos tan sólo se mostrará la confianza, ya que el soporte es el mismo.

11.2.1.- ANÁLISIS GENERALES

1. Ruta General[5]: La confianza mínima se estableció al 80%. Si no se indica lo contrario, la elevación es superior a 1. Los resultados más interesantes han sido:

    a. Si no aparecen fuentes privadas no aparece afiliación institucional (80,47% de soporte, 99,50% de confianza).

    b. Si aparece Delito, no aparece Tabaco (50,99%, 97,26%). Tampoco la Afiliación Institucional (91,79%) ni fuentes privadas (91,79%). No aparece la Epidemiología (97,65%). Tampoco aparece Personalización Anecdótica (92,18%)

    c. La no aparición de delito implica que no se hable de Cocaína (49%, 85,77%)

    d. Cuando aparece Drogas en General, no aparecen ni Tabaco (38,24%, 93,75%) ni Alcohol (89,06%).

    e. Cuando la Valoración de Unidad de Análisis es alta, aparece como fuente manifiesta el Nombre Propio (35,84%, 85,55%).

    f. Cuando la Categoría de Tema Principal es Tráfico de Drogas, no aparecen el Tabaco (32,07%, 98,75%), ni Alcohol (99,37%), ni fuente Privada (93,78%), ni Afiliación Institucional (93,16%), ni Epidemiología (100%), ni Personalización Anecdótica (94,40%), ni Política/Legislación (97,51%), ni fuentes No Expertas (95,03%).

    g. Cuando aparecen las Fuerzas y Cuerpos de Seguridad del Estado, no aparecen ni el Tabaco (23,10%, 97,41%), ni el Hachís (80,17%). Este último aparece como menos habitual que en la muestra general, con una elevación de (0,87).

    h. Cuando aparece el Tabaco no aparecen ni Delito (22,50%, 93,80%), ni Drogas en General (89,38%), ni Cocaína (90,26%). Tampoco son frecuentes ni Psico-Sanitarios (82,30%), ni un contexto Científico/Médico (82,30%), pero aquí la elevación nos indica que "Tabaco" es un mal predictor (0,88).

    i. Cuando aparece la Categoría de Tema Principal Estudios y Prevención, no aparecen ni Cocaína (20,71%, 86,53%) ni Personalización Anecdótica (96,15%).

[5] Los resultados detallados se incluyen en los anexos

j. Cuando aparece el Alcohol no aparece la Estrategia Política (20,51%, 96,11%).

k. Cuando aparece una Fuente Privada, no aparecen las Fuerzas y Cuerpos de Seguridad del Estado (19,52%, 95,91%).

l. Cuando la Categoría de Tema Principal es Consecuencias, no aparece ni Cocaína (18,12%, 87,81%), ni la Estrategia Política (95,60%).

m. Cuando aparece Personalización Anecdótica no aparece la Cocaína (17,33%, 83,90%) ni la Marihuana (82,75%), pero este último con una elevación de 0,88.

n. Cuando aparece Estrategia Política, no aparece ni el Alcohol (16,53%, 95,18%) ni Personalización Anecdótica (96,38%).

o. Cuando la sección es Sociedad, se le da Dedicación Principal (16,35%, 95,18%).

p. En Nacional/Política/España no aparece la Cocaína (14,94%, 85,33%).

q. Cuando aparece Política/Legislación, no aparece Cocaína (13,14%, 92,42%), Estrategia Política (100%), ni la Ética/Moralidad (83,33%). Este último con una elevación de 0,89.

r. Cuando la Categoría de Tema Principal es Consumo, no aparece Delito (10,95%, 89,09%) y aparece con Dedicación Principal (92,72%).

s. Cuando aparece Hachís, se encuadra en Tráfico de drogas (7,96%, 82,5%), se asocia con Delito (92,5%), y no aparece con Alcohol (97,5%) ni Tabaco (95%), ni aparecen Políticos (95%),

t. Cuando aparece Psico-Sanitarios, no aparece Cocaína (7,17%, 86,11%).

u. Cuando aparece la Marihuana, no aparece el Tabaco (6,97%, 97,14%), no aparece en Domingo (94,28%), no aparece el Alcohol (88,57%) ni Ética/Moralidad (80%). Este último con una elevación de 0,85.

v. Si la heroína está presente, no aparece Política/Legislación (6,77%, 97,05%).

w. Cuando hay un contexto general científico/médico no aparece Personalización Anecdótica (6,57%,100%), Estrategia Política (100%), ni Cocaína (84,84%),.

x. Cuando aparece Ética/Moralidad no aparece Delito (6,37%, 93,75%), Cocaína (90,62%), Tabaco (87,5%) ni Mercado/Empresa (100%).

y. Cuando aparece Tribunal, no aparece Alcohol (6,17%, 96,77%), Tabaco (93,54%), y la Extensión es corta (80,64%).

z. Cuando aparece Cannabis, el texto es Corto (5,97%, 80%) y no aparecen las Fuerzas y Cuerpos de Seguridad del Estado (86,66%).

aa. Los textos que aparecen en Portada, suelen aparecer de lunes a sábado (5,77%, 96,55%), no se habla de Política/Legislación (96,55%), aparece con una imagen (89,65%), y no aparece Cannabis (82,75%), pero este último con una elevación de 0,88.

bb. Si la extensión es Largo, no aparece Cocaína (4,98%,88%), Alcohol (92%), Fuerzas y Cuerpos de Seguridad del Estado (88%), fuente Privada (92%) ni Mercado/Empresa (80%), pero este último con una elevación de 0,89.

cc. Cuando la Categoría de Tema Principal es Ocio, no aparece Cannabis (3,98%, 80%) pero con elevación 0,85.

dd. Cuando Académicos está presente, la extensión es corta (3,78%, 89,47%).

ee. Cuando Ciencia está presente, Cocaína está ausente (3,58%, 83,33%).

ff. Cuando la Fuente Manifiesta es Enviado Especial, el Periódico es El Mundo (2,39%, 83,33%)

2. Ruta Drogas[6]: La confianza mínima se estableció al 50%. Si no se indica lo contrario, la elevación es superior a 1. Los resultados más interesantes, no mencionados previamente, han sido:

    a. Cuando se habla de Alcohol, no aparece Drogas en General (20,51%, 79,61%)

[6] A partir de el análisis de Ruta Drogas, los resultados presentados en los anexos son resúmentes de los datos obtenidos.

b. Cuando se habla de Marihuana, no se habla de Drogas en General (6,97%, 71,42%)

c. Le aparición de Heroína condiciona la aparición de Cocaína (6,77%, 52,94%).

d. La aparición de Cannabis condiciona la aparición de Cocaína (5,97%, 66,66%). Cuando aparece el Cannabis, no aparece Tabaco (60%) ni Alcohol (53,33%) con una elevación de 0,77 y 0,67 respectivamente.

3. Ruta *Frame*: La confianza mínima se estableció al 50%. Si no se indica lo contrario, la elevación es superior a 1. Los resultados más interesantes, no mencionados previamente, han sido:

    a. La presencia de Ética/Moralidad implica una ausencia de Personalización Anecdótica (6,37%, 65,62%) y ausencia de Política/Legislación (65,62%) con una elevación de 0,79 y 0,75 respectivamente.

4. Ruta Fuentes: La confianza mínima se estableció al 50%. Si no se indica lo contrario, la elevación es superior a 1. No se han hallado resultados interesantes no mencionados previamente.

11.2.2.- ANÁLISIS INDIVIDUALES

En esta categoría, al tratarse de subpoblaciones, se indicarán el número de ocurrencias, soporte y confianza, como una terna entre paréntesis.

5. Análisis Individual de la Categoría del Tema Principal

    a. Subpoblación Tráfico de Drogas:

        i. La presencia de Políticos condiciona la ausencia de Cocaína (22 casos, 13,66% de soporte, 81,81% de confianza)

    b. Subpoblación Consecuencias:

        i. Un contexto general Científico/Médico condiciona la aparición de Tabaco (10,10,98%,70%)

        ii. En el periódico ABC, la Marihuana está Ausente (16,17,58%, 68,75%) pero con elevación 0,78.

        iii. La ubicación en página impar con apertura de sección corresponde al periódico La Razón (14,15,38%, 64,28%)

iv. Cuando se trata de un contexto Científico/Médico, el diario es El País (10, 10,98%, 60%)

c. Subpoblación Consumo

i. El Alcohol condiciona la presencia de la Cocaína (16, 29,09%, 68,75%)

ii. A la inversa obtenemos que la presencia de la Cocaína implica presencia de Alcohol (17, 30,09%, 64,70%)

iii. Si el periódico es La Razón, el Alcohol está presente (12, 21,81%, 66,67%)

d. Subpoblación Estudios y Prevención:

i. La presencia de Ética/Moralidad implica una fuente Privada (15,14,42%, 73,33%)

ii. La presencia de Delito implica presencia de Estrategia Política (16,15,38%, 62,5) y de Políticos (16,15,38%, 62,5).

6. Análisis Individual por Cantidad de Fuentes

a. Subpoblación ninguna:

i. Si la valoración de la unidad de análisis es alta, entonces el Tabaco está presente (31, 22,79%, 61,29%) y Personalización Anecdótica está presente (31, 22,79%, 64,51%).

ii. Si la categoría del tema principal es Estudios y Prevención, el periódico es El País (14, 10,29%, 64,28%) y el Tabaco está presente (14, 10,29%, 71,42%).

iii. Cuando la categoría del tema principal no está relacionada con Drogas, la valoración de la unidad de análisis es alta (24, 17,64%, 66,66%)

b. Subpoblación muchas

i. Si el Delito está presente, aparece Estrategia Política (20, 33,89%, 60%)

ii. Si la categoría de tema principal es Tráfico de Drogas, la estrategia política está presente (13, 22,03%, 69,23%)

7. Análisis Individual de cada Fuente:
    a. Subpoblación fuente Privada:
        i. El Cannabis se asocia con la Cocaína (15, 15,36%, 73,33%)
    b. Subpoblación Fuerzas y Cuerpos de Seguridad del Estado:
        i. La presencia de Alcohol implica categoría de tema principal Consecuencias (16, 13,79%, 62,5%)
    c. Subpoblación fuentes Políticos:
        i. Si el tabaco está presente, drogas en general está ausente (21, 24,41%, 85,71%)
        ii. Con Política/Legislación presente, Estrategia Política está ausente (12, 13,95%, 100%)
        iii. Cuando categoría de tema principal es Consecuencias, Alcohol está presente (11, 12,79%, 63,63%).
        iv. Si aparece el Cannabis lo hace junto con Cocaína (11, 12,70%, 90,90%)
8. Análisis Individual de la Forma de Aparición: No se obtuvo ningún resultado interesante.
9. Análisis Individual de EsDomingo
    a. Subpoblación Presente:
        i. Si aparece Hachís, aparecen Fuerzas y Cuerpos de Seguridad del Estado (11, 12,35%, 63,63%) y el tema principal es Tráfico de Drogas (11, 13,35%, 81,81%)
        ii. Si aparece el tema Estudios y Prevención, aparece el Tabaco (17, 19,01%, 76,47%)
        iii. Si aparece Estrategia Política, el periódico es El País (13, 14,60%, 61,53%)
10. Análisis Individual de Drogas
    a. Subpoblación Alcohol
        i. Cuando aparece Política/Legislación, aparece Mercado/Empresa (17, 16,50%, 23,52%) con elevación 4,03.

ii. Cuando aparece Personalización Anecdótica aparece Ética/Moralidad (19, 18,44%, 21,05%) con elevación 2,4.

b. Subpoblación Tabaco:

i. Si el periódico es La Razón, Epidemiología está ausente (26, 23%, 61%) con elevación 0,79

11. Análisis Individual de Extensión: no se hallaron relaciones interesantes.

12. Análisis Individual de *Frame*

a. Subpoblación Nueva Investigación:

i. Si aparece Contexto Científico/Médico, aparece el Tabaco (15, 24,59%, 66,67%)

b. Subpoblación Contexto Científico/Médico:

i. Si aparece en Página Impar Normal, es el Periódico El Mundo (15, 24,59%, 66,66%)

c. Subpoblación Ética-Moralidad:

i. Si el Periódico es ABC, se ubica la información en Página impar normal (11, 34,37%, 63,63%)

d. Subpoblación Estrategia Política

i. Si aparece el Tabaco, la categoría de tema principal es Estudios y Prevención (16, 19,27%, 75%), la ubicación es Página Impar Normal (16, 19,27%, 62,5%), no aparece el delito (16, 19,27%, 93,75%) y no se habla de drogas en general (16, 19,27%, 100%).

e. Subpoblación Política y Legislación:

i. Si está presente Ética/Moralidad, la extensión es Normal (11, 16,66%, 63,63%)

f. Subpoblación Delito:

i. Si el tema es Estudios y Prevención, el periódico es El País (16, 6,25%, 62,5%), aparece estrategia política (16, 6,25% 62,5%) y Políticos (16, 6,25%, 62,5%)

ii. Si aparece el Alcohol, el tema principal se encuadra en Consecuencias (26, 14,06%, 75%)

iii. Si la ubicación es Portada, el periódico es El Mundo (16, 6,25%, 75%)

13. Análisis Individual de Fuente Manifiesta: No se obtuvo nada interesante

14. Análisis Individual de Género Periodístico

    a. Subpoblación Artículo:

        i. Si es domingo, se ubica en página par normal (11, 30,55%, 63,63%)

        ii. Si es página par normal, el periódico es El País (12, 33,33%, 66,66%)

    b. Subpoblación Noticia

        i. Si es página impar apertura de sección, el periódico es La Razón.

15. Análisis Individual de Ilustración

    a. Subpoblación Presente (1 o más ilustraciones)

        i. Si la ubicación es portada, el periódico es El Mundo (29, 10,35%, 62,06%) y drogas en general está presente (29, 10,35%, 65,51%)

16. Análisis Individual por Meses: No se obtuvo información relevante de este análisis.

17. Análisis Individual por Periódicos:

    a. Subpoblación de “El Mundo”

        i. Si aparece Tribunal, aparece Delito (11, 6,74%, 63,63%)

        ii. La aparición de Psico-Sanitarios, provoca una valoración de unidad de análisis alta (12, 7,36%, 75%)

        iii. La aparición de Políticos provoca una valoración de unidad de análisis alta (28, 17,17%, 67,85%)

        iv. La Personalización Anecdótica provoca una valoración de unidad de análisis alta (35, 21,47%, 71,42%)

v. La aparición de la Heroína provoca una valoración de unidad de análisis alta (12, 7,36, 66,66%)

vi. La aparición de fuentes no expertas obtienen una valoración de unidad de análisis alta (33, 20,24%, 63,63%)

vii. La aparición del alcohol está asociada a la no aparición de una fuente de tipo tribunal (26, 15,95%, 100%)

viii. Cuando el tema principal es Tráfico de Drogas la heroína no aparece (58, 35,58%, 94,82%) y la Marihuana tampoco (58, 35,58%, 96,55%) ni el hachís (58, 35,58%, 91,37%)

b. Subpoblación "ABC":

i. Cuando aparece Política/Legislación, aparece Ética/Moralidad (10, 12,65%, 60%)

ii. Cuando aparece Ética/Moralidad, no aparece Delito (11, 13,92%, 100%)

iii. Cuando aparece Hachís, aparece Delito (10, 12,65%, 90%)

iv. Cuando el tema principal es Tráfico de Drogas, no aparece Marihuana (30, 37,97%, 90%) ni Cannabis (30, 37,97%, 93,33%).

c. Subpoblación "El País":

i. Cuando aparece el Hachís, aparece Tráfico de Drogas (16,10,59%, 75%) y Delito (16, 10,59%, 93,75%)

d. Subpoblación "La Razón"

i. Cuando aparece el tema Consecuencias, aparece Alcohol (21, 19,26%, 71,42%)

ii. Cuando aparece epidemiología, la valoración de unidad de análisis es alta (19, 17,43%, 68,42%)

iii. Cuando aparecen fuentes Privadas, la valoración de unidad de análisis es alta (15, 13,76%, 66,66%)

18. Análisis Individual por Sección: No se obtuvo nada concluyente de este análisis.

19. Análisis Individual por Ubicación: No se obtuvo nada concluyente de este análisis.

20. Análisis Individual por Valoración de Unidad de Análisis

    a. Subpoblación Alto:

        i. Existe una ausencia generalizada de Ética/Moralidad (180, 100%, 90,55%)

11.2.3.- ANÁLISIS SUGERIDOS

En este apartado se codifican de nuevo soporte y confianza. Si hay más de un consecuente el segundo y sucesivos sólo codifican la confianza.

21. Valoración formal vs. Tema Principal y *Frame* y Fuente y Droga

    a. Resultados que condicionan una Valoración de la Unidad de Análisis alta :

        i. Académicos (3,78%, 68,42%)

        ii. Psico-Sanitarios (7,17%, 66,66%)

        iii. Enviado Especial (2,39%, 66,66%)

        iv. Heroína (6,77%, 58,82%)

        v. Personalización Anecdótica (17,33%, 58,62%)

        vi. Ciencia (3,58%, 55,55%)

        vii. No Expertos (14,74%, 54,05%)

        viii. Ética/Moralidad (6,37%, 53,12%)

        ix. Tribunal (6,17%, 51,61%)

        x. Fuente Manifiesta Nombre Propio (60,55%, 50,65%)

        xi. Fuente Manifiesta Corresponsal (3,58%, 50%)

        xii. Políticos (17,13%, 50%)

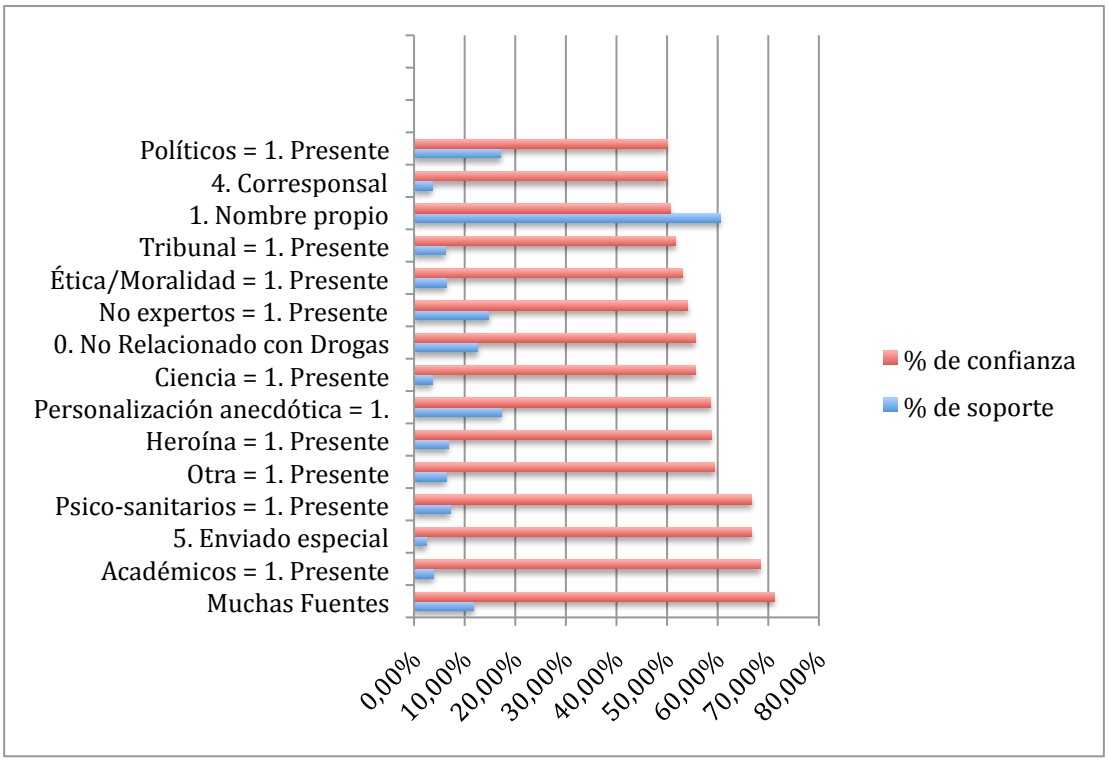

Figura 52 Campos más influyentes en la Valoración de Unidad de Análisis

b. Resultados que condicionan una Valoración de la Unidad de Análisis baja:

   i. Fuente manifiesta es Otros (3,78%, 78,94%)

   ii. fuente manifiesta es Agencia (4,78%, 78,94%)

   iii. No figura la fuente manifiesta (22,70%, 54,38%)

22. Tema Principal vs. *Frame* y Fuente y Droga

   a. Antecedentes que condicionan Tráfico de Drogas:

      i. Hachís (7,96%, 82,5%)

      ii. Fuerzas (23,10%, 68,96%)

      iii. Fuente Manifiesta Enviado Especial (2,39%, 66,66%)

      iv. Cocaína (23,09%, 59,52%)

v. Delito (50,99%, 59,375%)

vi. Tribunal (6,17%, 54,83%)

vii. Agencia (4,78%, 50%)

b. Antecedentes que condicionan Consumo:

i. Epidemiología (12,54%, 52,38%)

23. *Frame* vs. Fuente y Droga

a. Antecedentes que condicionan Nueva Investigación:

i. Ciencia (3,58%, 66,66%)

ii. Académicos (3,78%, 52,63%)

b. Antecedentes que condicionan Epidemiología:

i. Cannabis (5,97%, 53,33%)

ii. Académicos(3,78%, 52,63%)

c. Antecedentes que condicionan Estrategia política:

i. Políticos (17,13%, 53,48%)

ii. Enviado Especial (2,39%, 50%)

d. Antecedentes que condicionan Delito:

i. Enviado Especial (2,39%, 91,66%)

ii. Hachis (7,96%, 92,5%)

iii. Fuerzas (23,10%, 91,37%)

iv. Tribunal (6,17%, 83,87%)

v. Cocaína (25,09%, 72,22%)

vi. Agencia (4,78%, 70,83%)

vii. Tabaco ausente (77,49%, 64,01%)

e. Antecedentes que condicionan la ausencia de Delito

i. Ciencia (3,58%, 94,44%)

ii. Tabaco (22,50%, 93,80%)

iii. Alcohol (20,51%, 65,04%)

iv. Cannabis (5,97%, 63,33%)

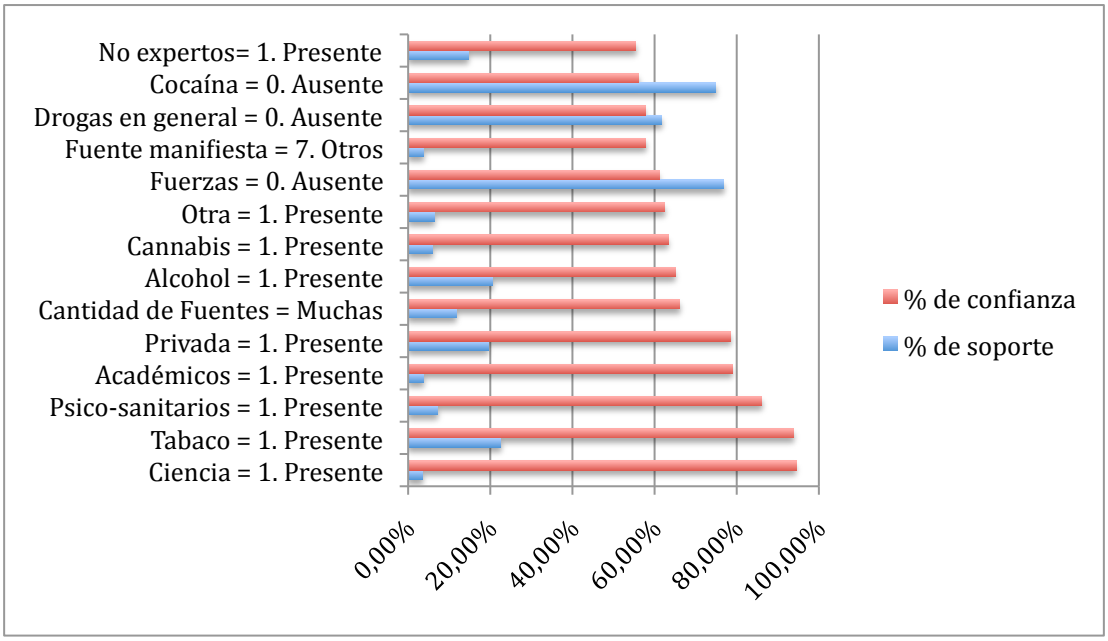

Figura 53 Antecedentes que condicionan la Ausencia de Delito

24. Fuente vs. Droga

    a. La presencia de Psico-Sanitarios condiciona la aparición de Tabaco (7,17%, 55,55%)

    b. Antecedentes que condicionan la aparición de Drogas en General:

        i. Redacción (2,19%, 72,72%)

        ii. Enviado Especial (2,39%, 58,33%)

        iii. Políticos (17,13%, 58,13%)

        iv. Corresponsal (3,58%, 55,55%)

25. La presencia de fuentes psico-sanitarias aumenta el valor de la unidad de análisis.

a. Resultado: La aparición de fuentes psico-sanitarias parece estar relacionada con una valoración alta. (36 ocurrencias, 7,17%, 66,66%) con 1,85 de elevación.

26. Tener imágenes en un texto aumenta la probabilidad de que exista una fuente de tipo tribunal.

    a. Resultado: En este caso se debe tener en cuenta que pese a la baja confianza de la regla que predice la aparición de Tribunal (6,57% y 15,15% de confianza), la elevación -2,45- indica que es un predictor muy fuerte. Además el resto de reglas nos sugieren que la mayoría de casos se ubican en un contexto con 2 imágenes.

        i. Ninguna → Tribunal ausente (44,22% y 95,04%)

        ii. 1 imagen → Tribunal ausente (46,61% y 93,58%)

        iii. 2 imágenes → Tribunal ausente (6,57% y 84,84%)

        iv. 2 imágenes → Tribunal presente (6,57% y 15,15%)

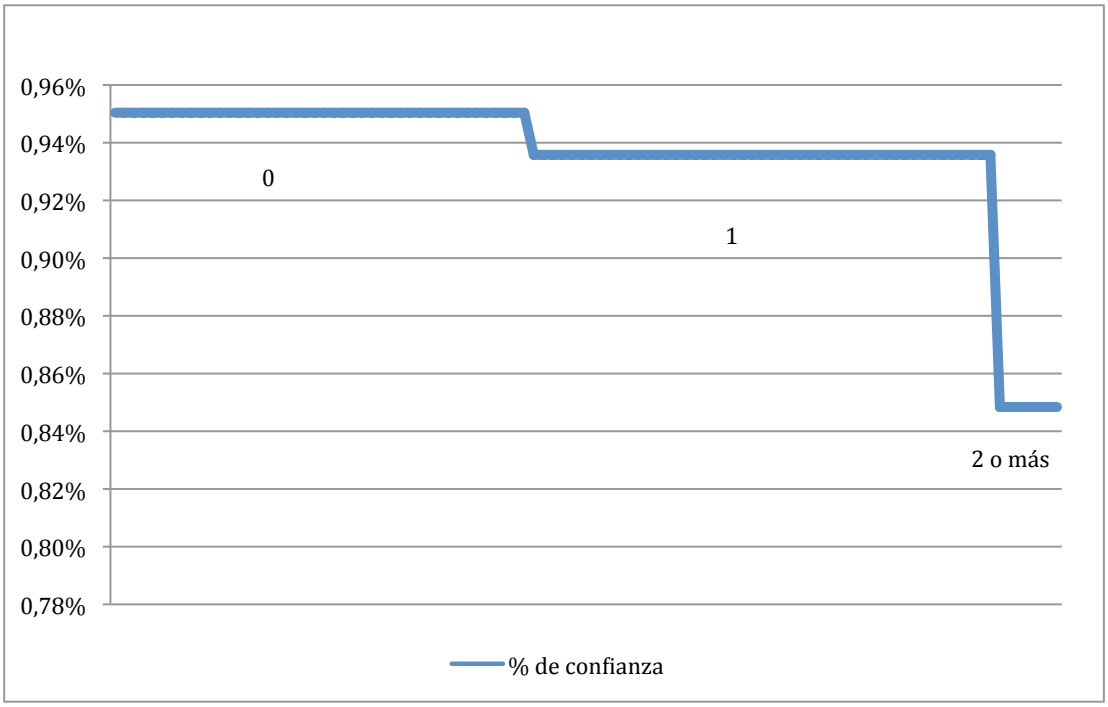

Figura 54 Confianza de ausencia de Tribunal por nº de Imágenes

27. Las drogas blandas –achís, cannabis, marihuana- no son portada.

    a. Resultado: De la misma forma que la anterior, el Cannabis, pese a su poca incidencia en la muestra, aparece en portada con más frecuencia de lo que debería para su poca incidencia.

        i. Portada → Cannabis (6%, 17%) con una elevación 3.

28. La valoración de la unidad de análisis está relacionada con la fuente manifiesta.

    a. Resultado: La fuente manifiesta, pese a tener una confianza cercana al 50%, parece ser un buen predictor para la valoración de la unidad de análisis según la elevación. Ver el análisis 21.- para más detalles.

29. Que la noticia verse sobre el tema "Celebrities" implica una mayor valoración de la unidad de análisis.

    a. Resultado: Se comprobó la hipótesis analizando la subpoblación. Se observa que los resultados donde aparecen "Celebrities" se agrupan en la parte media-superior de la valoración.

        i. Celebrities → Valoración alto (30 ocurrencias, 100%, 43%)

        ii. Celebrities → Valoración medio (30, 100%, 40%)

        iii. Celebrities → Valoración bajo (30, 100%, 17%)

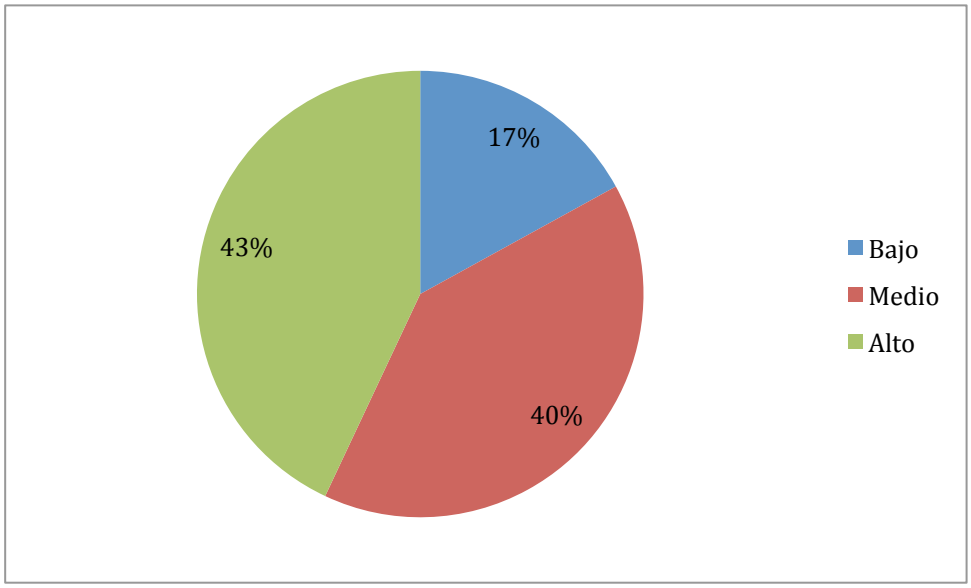

Figura 55 Distribución de aparición de Celebrities en cada Valor de la Unidad de Análisis

11.2.4.- ANÁLISIS AGRUPADOS

30. Análisis Agrupado de Drogas

a. Análisis Drogas como Consecuente:

i. Antecedentes que condicionan la presencia de Drogas legales:

1. Científico/Médico (6,57%, 78,78%)
2. Opinión pública (7,17%, 72,22%)
3. Psico-Sanitarios (7,17%, 72,22%)
4. Epidemiología (12,54%, 66,66%)
5. Nueva investigación (12,15%, 63,93%)
6. Política/Legislación (13,14%, 63,63%)
7. Estudios y Prevención (20,71%, 63,46%)

ii. Antecedentes que condicionan la ausencia de drogas legales:

1. Estrategia Política (16,53%, 77,10%)
2. Portada (5,77%, 72,41%)
3. ABC (15,73%, 70,88%)
4. El Mundo (32,47%, 66,87%)

iii. Antecedentes que condicionan la ausencia de drogas blandas:

1. Estudios y Prevención (20,71%, 87,5)
2. El Mundo (32,47%, 87,11%)

iv. Antecedentes que condicionan la ausencia de drogas duras:

1. Política/Legislación (13,14%, 90,90%)

2. Ética/Moralidad (6,37%, 87,5%)
3. Estudios y Prevención (20,71%, 85,57%)
4. Psico-Sanitarios (7,17%, 83,33%)
5. Contexto General Científico/Médico (6,57%, 81,81%)
6. El País (30,07%, 79,47%)

b. Análisis de subpoblación con Drogas Duras Ausente:
   i. Si la ubicación es portada, el periódico es El Mundo (18 casos, 5%, 77,77%)
   ii. Si la extensión es largo, el periódico es El País (22 casos, 6,11%, 63,63%)

c. Análisis de subpoblación con Drogas Legales Presente
   i. Si aparece Mercado/Empresa entonces el periódico es El País (26 casos, 13,54%, 69,23%)
   ii. Si aparece Ética/Moralidad entonces la valoración de la unidad de análisis es alta (12 casos, 6,25%, 75%)
   iii. Si aparece Psico-Sanitarios, la valoración de la unidad de análisis es alta (26 casos, 13,54%, 65,38%)

d. Análisis de subpoblación con Drogas Legales Ausente
   i. Si aparece Psico-Sanitarios aparece Epidemiología (10 casos, 3,22%, 70%)

11.2.5.- Análisis de Campos Influyentes

Se presentan los campos más influyentes en la aparición de un valor concreto, por orden de influencia.

31. Análisis de Campos Influyentes en Drogas
    a. Cocaína: Crack/Cocaína Base, Epidemiología, Éxtasis/MDMA y Cannabis.
    b. Heroína: Crack/Cocaína Base, Cocaína. Epidemiología, Éxtasis/MDMA y Cannabis.

c. Crack: Heroína, Cannabis, Nº de imágenes, Cocaína, Epidemiología.

d. Cristal: Contexto Científico/Médico, Nueva Investigación, Privada, Afiliación Institucional.

e. Éxtasis: Epidemiología, Opinión Pública, Alcohol, Cocaína, Cannabis, Heroína, Nueva Investigación, Académicos, Forma de Aparición, Valoración de Unidad de Análisis, Categoría Tema Principal, Otra.

f. Drogas en General: Tabaco,Estrategia política, Delito, Categoría Tema Principal, Alcohol, Políticos, Valoración de Unidad de análisis, Política/Legislación, Hachís, Contexto general científico/médico, Opinión pública, Psico-sanitarios, Otra, Cannabis y Epidemiología.

g. Otra: Anabolizantes, Categoría Tema Principal, Epidemiología, Ciencia, Académicos, Psico-sanitarios, Cannabis, Psicofármacos, Opinión pública, Cantidad de Fuentes, Cocaína, Drogas en general, Afiliación institucional, Valoración de Unidad de análisis, y Éxtasis/MDMA.

h. Psicofármacos: Ciencia, Epidemiología, Otra, Delito, Psico-sanitarios, imágenes.

i. Marihuana: Personalización anecdótica, Ética/Moralidad, No expertos, Tabaco.

j. Hachis: Categoría Tema Principal, Delito, Fuerzas, Alcohol, Forma de Aparición, Tabaco, Cannabis, Drogas en general, Tribunal, imágenes y Políticos.

k. Cannabis: Epidemiología, Cocaína, Afiliación institucional, Privada, Cargo, Categoría Tema Principal, Crack/Cocaína base, Alcohol, Éxtasis/MDMA, Otra, Heroína, Políticos, Cantidad de Fuentes, Hachís, Tabaco, Drogas en general y Psico-sanitarios.

l. Alcohol: Categoría Tema Principal, Epidemiología, Nueva investigación, Éxtasis/MDMA, Drogas en general, Estrategia política, Cannabis, Delito, Cantidad de Fuentes, Hachís, Opinión pública, Psico-sanitarios, Ciencia, Tribunal, Académicos, Contexto general científico/médico, Valoración de Unidad de análisis, Afiliación institucional, No expertos y Fuerzas.

m. Tabaco: Delito, Categoría Tema Principal, Drogas en general, Fuerzas, Contexto general científico/médico, Psico-sanitarios, Cocaína, Política/Legislación, Epidemiología, imágenes, Mercado/Empresa, Marihuana, Opinión pública, Hachís, Afiliación institucional, Privada, Personalización anecdótica, Cannabis, Valoración de Unidad de análisis, Tribunal y Académicos

32. Análisis de Campos Influyentes en *Frame*

a. Nueva Investigación: Ciencia, Contexto general científico/médico, Delito, Privada, Categoría Tema Principal, Académicos, Afiliación institucional, Alcohol, Epidemiología, Cantidad de Fuentes, Opinión pública, Psico-sanitarios, Éxtasis/MDMA, Fuerzas, Crystal/Cristal, Estrategia política, Tribunal, Personalización anecdótica, Mercado/Empresa y MesNumerico.

b. Contexto General Científico/Médico: Psico-sanitarios, Nueva investigación, Delito, Ciencia, Tabaco, Categoría Tema Principal, C Crystal/Cristal, Académicos, Fuerzas, Cantidad de Fuentes, Personalización anecdótica, Estrategia política, 'No expertos', Drogas en general, Alcohol, MesNumerico, Privada, Valoración de Unidad de análisis y Afiliación institucional.

c. Ética/Moralidad: Cargo, Privada, Afiliación institucional, Delito, Categoría Tema Principal, Política/Legislación, Marihuana, Cantidad de Fuentes, Personalización anecdótica, Cocaína y Mercado/Empresa.

d. Estrategia Política: Políticos, Drogas en general, Categoría Tema Principal, Política/Legislación, Alcohol, Personalización anecdótica, Mercado/Empresa, Contexto general científico/médico, Nueva investigación, Cantidad de Fuentes, MesNumerico y Epidemiología.

e. Política/Legislación: Categoría Tema Principal, Delito, Estrategia política, Tabaco, Ética/Moralidad, Cocaína, No expertos, Opinión no experta, Drogas en general, Privada, Afiliación institucional, Fuerzas e Imágenes.

f. Mercado/Empresa: Privada, Afiliación institucional, Categoría Tema Principal, Estrategia política, Tabaco, Personalización anecdótica, Valoración de Unidad de análisis, Cargo, Ética/Moralidad y Nueva investigación.

g. Epidemiología: Categoría Tema Principal, Delito, Cannabis, Alcohol, Psico-sanitarios, Otra, Académicos, Éxtasis/MDMA, Nueva

investigación, Afiliación institucional, Fuerzas, Privada, Tabaco, Heroína, Cantidad de Fuentes, Valoración de Unidad de análisis, Cocaína, Psicofármacos, Ciencia, Opinión pública, Crack/Cocaína base, Políticos, Drogas en general y Estrategia política.

h. Opinión Pública: Delito, Categoría Tema Principal, Éxtasis/MDMA, Nueva investigación, Tabaco, Alcohol, Cantidad de Fuentes, Otra, Privada, Fuerzas, Afiliación institucional, Ciencia, Drogas en general y Epidemiología.

i. Opinión no Experta: No expertos, MesNumerico, Política/Legislación, Categoría Tema Principal y Delito.

j. Personalización Anecdótica: No expertos, Categoría Tema Principal, Delito, Imágenes, Cantidad de Fuentes, Marihuana, Estrategia política, Políticos, Valoración de Unidad de análisis, Mercado/Empresa, Fuerzas, Contexto general científico/médico, Ética/Moralidad, Tabaco, Cocaína y Nueva investigación.

k. Delito: Categoría Tema Principal, Tabaco, Fuerzas, Epidemiología, Afiliación institucional, Privada, Nueva investigación, Personalización anecdótica, Política/Legislación, Contexto general científico/médico, Cocaína, Hachís, Opinión pública, Ética/Moralidad, Drogas en general, Psico-sanitarios, Ciencia, Tribunal, Alcohol, Imágenes, Cargo, Académicos, Psicofármacos, Cantidad de Fuentes y Opinión no experta.

33. Análisis de Campos Influyentes en Fuentes

a. Análisis Cuantitativo de Fuentes: Privada, Afiliación institucional, Fuerzas, No expertos, Psico-sanitarios, Políticos, Cargo, Personalización anecdótica, Académicos, Nueva investigación, Tribunal, Ciencia, Categoría Tema Principal, Imágenes, Epidemiología, Valoración de Unidad de análisis, Alcohol, Contexto general científico/médico, Ética/Moralidad, Opinión pública, Cannabis, Otra, Estrategia política y Delito.

b. Políticos: Estrategia política, Cantidad de Fuentes, Drogas en general, No expertos, Personalización anecdótica, Cannabis, Epidemiología, Hachís y Valoración de Unidad de análisis.

c. Fuerzas y Cuerpos de Seguridad del Estado: Categoría Tema Principal, Delito, Cantidad de Fuentes, Tabaco, Cocaína, Hachís, Afiliación institucional, Privada, Epidemiología, Contexto general científico/médico, Personalización anecdótica, Nueva investigación,

Opinión pública, Cargo, Ciencia, Política/Legislación, Psico-sanitarios, Alcohol y Anabolizantes.

d. Tribunal: Cantidad de Fuentes, Delito, Alcohol, Hachís, Cocaína, Tabaco, Nueva investigación y Categoría Tema Principal.

e. Ciencia: Nueva investigación, Contexto general científico/médico, Académicos, Psicofármacos, Delito, Otra, Cantidad de Fuentes, Psico-sanitarios, Epidemiología, Alcohol, Opinión pública, Fuerzas y Categoría Tema Principal.

f. Académicos: Nueva investigación, Ciencia, Epidemiología, Psico-sanitarios, Cantidad de Fuentes, Categoría Tema Principal Valoración de Unidad de análisis, Otra, Contexto general científico/médico, Éxtasis/MDMA, Delito, Alcohol, Imágenes, Tabaco y MesNumerico.

g. Privada: Afiliación institucional, Cargo, Cantidad de Fuentes, Categoría Tema Principal, Delito, Ética/Moralidad, Nueva investigación, Fuerzas, Cannabis, Mercado/Empresa, Epidemiología, Opinión pública, Valoración de Unidad de análisis, Tabaco, Política/Legislación, Psico-sanitarios, Contexto general científico/médico y Crystal/Cristal.

h. Psicosanitarios: Contexto general científico/médico, Cantidad de Fuentes, Epidemiología, Tabaco, Delito, Categoría Tema Principal, Académicos, Valoración de Unidad de análisis, Nueva investigación, Ciencia, Otra, Alcohol, Psicofármacos, Drogas en general, Fuerzas, Privada, Cannabis, Imágenes y Cocaína.

i. No Expertos: Personalización anecdótica, Cantidad de Fuentes, Opinión no experta, Políticos, Política/Legislación, Categoría Tema Principal, Imágenes, Marihuana, Contexto general científico/médico y Alcohol.

34. Análisis de Campos Influyentes en Otros:

a. Categoría de tema principal: Delito, Epidemiología, Fuerzas, Tabaco, Alcohol, Cocaína, Afiliación institucional, Política/Legislación, Personalización anecdótica, Privada, Hachís, Otra, Nueva investigación, Cargo, Opinión pública, Ética/Moralidad, Psico-sanitarios, Contexto general científico/médico, Drogas en general, Cannabis, Académicos, Estrategia política, Mercado/Empresa, Cantidad de Fuentes, 'No expertos', Imágenes,

Valoración de Unidad de análisis, Ciencia, Opinión no experta, Tribunal y Éxtasis/MDMA.

b. Forma de aparición: Hachís, EsDomingo y Éxtasis/MDMA.

c. Valoración de unidad de análisis: Imágenes, Académicos, Psico-sanitarios, Drogas en general, Epidemiología, Personalización anecdótica, Cantidad de Fuentes, Categoría Tema Principal, Cargo, Privada, Mercado/Empresa, Afiliación institucional, Alcohol, Tabaco, Éxtasis/MDMA, MesNumerico, Otra, Contexto general científico/médico y Políticos.

d. Número de imágenes: Valoración de Unidad de análisis, Personalización anecdótica, Cantidad de Fuentes, Tabaco, Delito, No expertos, Categoría Tema Principal, Cocaína, Crack/Cocaína base, Académicos, Hachís, Psico-sanitarios, Política/Legislación y Psicofármacos.

e. EsDomingo: Forma de Aparición.

12.- CONCLUSIONES

12.1.- PRINCIPALES APORTACIONES Y DISCUSIÓN

En el presente estudio se ha tratado de analizar en profundidad una muestra de datos de drogas en el ámbito de la prensa escrita española. Durante sus primeras etapas, se ha realizado un análisis del estado del arte de la Minería de Datos y se han revisado las diferentes técnicas así como sus algoritmos más representativos.

Posteriormente se ha realizado un estudio bibliográfico de las referencias relativas a la Minería de Datos en el análisis periodístico, explicando los dos principales enfoques hallados en los textos: el Text Mining y el Data Mining. Posteriormente se centró el análisis en algunos estudios y especialmente en el libro de [Colle. 2002] que es una de las pocas referencias de análisis periodístico halladas.

Más adelante se describió CRISP-DM, como una metodología formal de trabajo en proyectos de Data Mining.

En el siguiente punto se abordó el análisis del caso de estudio desde la perspectiva del análisis periodístico del tratamiento informativo de la drogadicción, describiendo el proyecto del que proviene la muestra analizad y los objetivos específicos del proyecto. Además, se abordó desde la perspectiva del periodismo, una aproximación al análisis de contenido desde la perspectiva del *framing* y al análisis de intensidad formal.

Posteriormente se realizó una evaluación del caso de estudio para tratar de definir los objetivos de la Minería de Datos, así como un estudio inicial de los datos de la muestra, para tratar de describir cada uno de los campos, las secciones, sus valores y anomalías.

En una etapa posterior se trató de realizar el proceso ETL –Extracción, Transformación y Carga- de los datos, para obtener la vista minable. Este proceso estuvo condicionado por la selección del algoritmo APRIORI, que se discutió en el punto *10.1.*- Selección de la técnica de modelado.

En la etapa de modelado se describieron los modelos construidos y se discutió la forma de evaluación objetiva y subjetiva a realizar. Finalmente, se presentaron los resultados.

De entre estos resultados, cabe destacar:

- Cuando aparece Drogas en General, no aparecen ni Tabaco (38,24%, 93,75%) ni Alcohol (89,06%).
- Cuando aparece la Categoría de Tema Principal Estudios y Prevención, no aparece Cocaína (20,71%, 86,53%).

- Cuando aparece el Alcohol no aparece la Estrategia Política (20,51%, 96,11%).
- Cuando la Categoría de Tema Principal es Consecuencias, no aparece Cocaína (18,12%, 87,81%).
- Cuando la sección es Sociedad, se le da Dedicación Principal (16,35%, 95,18%).
- Cuando aparece Política/Legislación, no aparece Cocaína (13,14%, 92,42%).
- Si la heroína está presente, no aparece Política/Legislación (6,77%, 97,05%).
- Cuando aparece Tribunal, no aparece Alcohol (6,17%, 96,77%), Tabaco (93,54%), y la Extensión es corta (80,64%).
- Cuando la Categoría de Tema Principal es Ocio, no aparece Cannabis (3,98%, 80%) pero con elevación 0,85. Es decir, en la temática Ocio aparece Cannabis con una frecuencia superior que en la muestra en general.
- Cuando se habla de Marihuana, no se habla de Drogas en General (6,97%, 71,42%)
- En la subpoblación de Tráfico de Drogas, la presencia de Políticos condiciona la ausencia de Cocaína (22 casos, 13,66% de soporte, 81,81% de confianza).
- En la subpoblación de Consecuencias, si el periódico es ABC, la Marihuana está Ausente (16,17,58%, 68,75%) pero con elevación 0,78. Si aparece un contexto Científico/Médico, el diario es El País (10, 10,98%, 60%).
- En la subpoblación de Consumo, si el periódico es La Razón, el Alcohol está presente (12, 21,81%, 66,67%).
- En la subpoblación EsDomingo, si aparece Estrategia Política, el periódico es El País (13, 14,60%, 61,53%).
- En la subpoblación Alcohol, cuando aparece Política/Legislación, aparece Mercado/Empresa (17, 16,50%, 23,52%) con elevación muy alta (4,03).
- En la subpoblación de encuadre Delito, si la ubicación es Portada, el periódico es El Mundo (16, 6,25%, 75%).
- Las Celebrities aumentan la Valoración de Unidad de Análisis.
- La aparición de fuente “Tribunal” aumentan el número de imágenes que acompañan al texto.

- Si los periódicos son ABC (15,73%, 70,88%) o El Mundo (32,47%, 66,87%) Drogas legales está ausente.

- Si aparece Política/Legislación (13,14%, 90,90%) o Ética/Moralidad (6,37%, 87,5%) o Estudios y Prevención (20,71%, 85,57%) o Psico-Sanitarios (7,17%, 83,33%) o Contexto General Científico/Médico (6,57%, 81,81%) o El País (30,07%, 79,47%) drogas duras está ausente.

- En la subpoblación de Drogas Legales, si aparece Ética/Moralidad, la Valoración de Unidad de Análisis es Alta (12 caso, 6,25%, 75%).

Hay que mencionar que los resultados han estado condicionados en todo momento por la elección del algoritmo disponible en la herramienta. Sería interesante comprobar si se obtendrían los mismos resultados en caso de haber seleccionado otro algoritmo, o bien, haber utilizado otros parámetros de entrada para APRIORI. Además de APRIORI, existen gran cantidad de algoritmos de reglas de asociación en la bibliografía, tales como CARMA, GRI, FP-Growth, etc. Sería interesante comprobar si la selección de otro algoritmo obtiene resultados de forma más eficaz.

Por otro lado, debido a que APRIORI no escala bien, por la generación de candidatos previa que debe realizar y sus múltiples combinaciones con el resto de elementos, se han tratado de reducir al máximo los resultados, utilizando para ello unos niveles mínimos de soporte y precisión, un máximo de un antecedente por regla, así como una medida de interés –la elevación- para la selección de resultados. Mediante unas pruebas previas para evaluar el escalado de APRIORI, estableciendo todos las variables como antecedente y consecuente, un 2% mínimo de soporte, un 50% de confianza y dos antecedentes por regla, se obtuvieron más de 250.000 reglas por evaluar. Es posible que existan otros parámetros que obtengan resultados más interesantes. También es posible que utilizar otra medida de interés permita obtener resultados más eficientemente en este caso.

Por otro lado, debido a esta poda de resultados objetiva, realizada por medio de subpoblaciones y la medida de interés elevación, y a la poda subjetiva realizada por el equipo de investigación periodística, es posible que muchos resultados que pudieran resultar interesantes al equipo de investigación periodística, no aparezcan en este estudio. Por ejemplo, no se han obtenido reglas con 3 o más antecedentes que es posible que existan en la muestra.

Además, al necesitar datos categóricos, los valores numéricos han sido discretizados obligando a perder precisión en el análisis. Del mismo modo, algunas variables de tipo conjunto han sido categorizadas para evitar la dispersión de valores y permitir obtener conclusiones con un soporte mínimo. Es posible que se hayan perdido relaciones interesantes por esta pérdida de precisión.

Acerca de la muestra, tal y como se indicó en el apartado *8.3.-* Exploración de los datos*,* se debe tener en cuenta que está restringida a 6 meses en 4 periódicos de línea editorial diferente. Estos se seleccionaron por su difusión en el ámbito nacional, es decir, por tirada. De la selección de periódicos se observa que uno de ellos, "El País", tiene una línea editorial diferenciada del resto. Hay otras peculiaridades de la muestra que se deben tener en cuenta, como que se trata de periódicos impresos de tirada nacional, la recopilación se ha realizado de forma manual, las drogas de diseño apenas aparecen, etc. Es posible que estas características hayan provocado un sesgo en los datos, y por tanto, en los resultados. Sería interesante trabajar con otras muestras para observar si se obtienen los mismos resultados.

Otro de los problemas que ha afrontado el proyecto es la falta de referencias anteriores al mismo, hasta donde sabemos. Pese a la existencia de numerosos estudios de Text Mining y reconocimiento de patrones en lenguaje natural, existe una carencia de casos de estudio de Minería de Datos en análisis periodístico, y el caso hallado de [Colle. 2002] no realiza ninguno de los dos análisis llevados a cabo en este estudio.

12.2.- FUTURAS LÍNEAS DE INVESTIGACIÓN

Una vez vistos los resultados de los diferentes modelos y las limitaciones que se han presentado en el presente estudio, podría ser interesante tratar de hallar predictores a las diferentes variables del estudio. De este modo, sería posible evaluar futuros valores de las variables y observar si se corresponden con los modelos generados.

Sería interesante estudiar otras muestras temporales para observar si se obtienen los mismos resultados que en la presente muestra. Así mismo, podría ampliarse el alcance a más periódicos nacionales y/o internacionales.

En el presente estudio se ha aplicado la metodología KDD, centrándose en la fase de Minería de Datos, debido a que los datos ya estaban extraídos de la información original por parte de un observador humano. Otra línea abierta sería el estudio de la información original en la fase de extracción de KDD, mediante la aplicación del Text Mining a los textos y observar si se obtienen los mismos datos de un proceso automatizado que de un observador humano, como en la muestra presente.

En el caso del presente estudio, al aplicar el algoritmo APRIORI, ha sido necesario discretizar todas las variables implicadas. Además, se han agrupado muchos valores de forma que su frecuencia aumentara y resultara suficientemente representativa para que no fueran podados por el soporte mínimo aplicado al algoritmo. Sería interesante que aplicar un algoritmo fuera capaz de trabajar con valores numéricos, de forma que no se perdiera precisión en la discretización y agrupación de variables. Un buen candidato para este trabajo sería el algoritmo GRI, que admite valores numéricos y utiliza una medida de interés propia (J-mensure) para evaluar las reglas que son interesantes.

La selección del algoritmo ha condicionado todo el proceso de trabajo. Una posible investigación futura sería aplicar un algoritmo que minimice los resultados, pero que sea lo más exhaustivo posible para tratar de hallar el máximo de resultados interesantes para un observador subjetivo. Sin embargo, esta exhaustividad exige una gran cantidad de combinaciones a analizar. Se plantea por tanto el problema de hallar medidas que permitan predecir qué reglas van a generar otras reglas interesantes y cuáles no lo van a hacer.

Las medidas de interés han sido otra de las vías de investigación abiertas por el presente estudio. Si bien la elevación ha sido una medida heurística útil para el proceso de selección de resultados, no es apropiada para podar ramas. Por otro lado, el algoritmo APRIORI utiliza el soporte como medida de poda, ya que en los algoritmos anti-monotonos el soporte de reglas posteriores es predecible. Sin embargo, esta forma de podar resultados puede dejar de evaluar reglas interesantes por falta de soporte mínimo. Se podría tratar de hallar una medida que minimice la cantidad de resultados, y que permita hallar aquellos más representativos para el problema.

13.- ANEXOS

13.1.- ANEXO: LIBRO DE INSTRUCCIONES PARA BASE DE DATOS DE ANÁLISIS DE PRENSA

La unidad de análisis es la unidad redaccional entendida ésta como el texto periodístico, el texto periodístico más su acompañamiento gráfico, o el elemento gráfico (cualquiera de las tres opciones es válida) que, conformando una unidad informativa, traten como tema principal o secundario, las drogas en cualquiera de sus versiones: sustancias adictivas, trastornos de comportamiento, consecuencias del consumo, datos de consumo, investigaciones, campañas de prevención, incautaciones, etc.

Variables:

Id

Es un número de identificación que se genera de forma automática.

Periódico

Nombre del periódico. Elegir de la lista.

Edición

Edición del ejemplar. Elegir de la lista.

Fecha

Fecha de publicación del periódico. Elegir de la lista.

Género

Género periodístico de la unidad redaccional. Elegir de la lista. Si no se sabe qué género es, hacerlo constar de alguna manera para que lo revisemos.

Sección

Sección del periódico donde se encuadra la unidad redaccional. Elegir de la lista.

Sección Otros

Si la sección es “Otros”, escribir aquí el nombre de la sección.

Página

Nº de página donde comienza la unidad redaccional en número.

Extensión

Cálculo aproximado de la extensión de la unidad redaccional el número de líneas suponiendo que una columna completa tiene 104 líneas. En este cálculo hay que tener en cuenta el titular, el texto y los elementos gráficos. Ejemplo: una unidad redaccional que ocupe media columna (titular incluido) tendrá una extensión de 50. Otro ejemplo: una unidad redaccional que ocupe media página (titular y foto incluidos) tendrá una extensión de 50x5 columnas= 250.

Forma de aparición

Elegir de la lista según estos criterios:

Dedicación principal: el tema relacionado con la droga aparece en el titular.

Incrustación más imagen: el tema relacionado con la droga aparece destacado en el texto y en una imagen.

Referencia más imagen: el tema relacionado con la droga aparece poco destacado en el texto y en una imagen.

Incrustación: el tema relacionado con la droga aparece destacado en el texto, pero no aparece en las imágenes.

Referencia: el tema relacionado con la droga aparece poco destacado en el texto.

Sin referencia: el tema relacionado con la droga aparece muy poco destacado en el texto, pero no tiene ninguna importancia. Ejemplo: algunas veces se menciona la palabra droga o alguna droga, pero esa es la única referencia que hay en el texto al tema que nos ocupa.

Ubicación en el periódico

Elegir de la lista.

Jerarquía en la página

Elegir de la lista.

Altura de titular

Es el número de líneas que tiene sólo el título, sin contar antetítulo ni subtítulo.

Ancho de titular

Es el número de columnas que ocupa todo el titular, incluido antetítulo y subtítulo.

Superficie de titular

Es el número de módulos aproximados que ocupa todo el titular. Un módulo es cada uno de los cuadrados imaginarios en los que se divide una columna. Para la medición tomaremos que una columna son 8 módulos. Ejemplo: un titular que vaya a una sola columna de ancho no tendrá más de un módulo. Otro ejemplo: un titular que vaya a cuatro columnas de ancho tendrá, probablemente, entre 4 y 8 módulos dependiendo del número de líneas que tenga y de si tiene o no antetítulo y/o subtítulos.

Cuerpo de titular

Elegir de la lista. Normalmente: noticia grande, titular grande; noticia mediana, titular mediano; noticia pequeña o en una columna, titular pequeño.

Jerarquía de titular

Ver si es el más importante en la página, el segundo en importancia o el tercero o el cuarto. Escoger de la lista.

Nº de imágenes

Ilustración prioritaria

Escoger de la lista si la imagen/imágenes de la unidad redaccional son las principales de la página.

Tipografía o página especial

Escoger de la lista si el tipo de letra o la página donde va la unida redaccional son especiales, distintas a lo que normalmente hace el periódico.

Análisis del tono

Escoger de la lista según la impresión general de la lectura del texto de la unidad redaccional conforme al tratamiento que da a las drogas.

Tema principal

Escoger de la lista. No puede quedar en blanco. Esta variable debe recoger la motivación principal del trabajo periodístico. Normalmente el tema principal queda reflejado en el Titular, ante título y subtítulo así como en la entradilla. Si no existe entradilla resaltada tipográficamente contará a efectos de este análisis el primer párrafo del texto. Ahora se detallan los temas y cómo están clasificados:

- Tráfico de drogas, alijos importantes. Se indican cantidades y se hace expreso la cantidad y su importancia
- Tráfico de drogas, menudeo. Venta en puertas de los colegios/institutos, discotecas, gramos, papelinas

- Producción de sustancias. Elaboración/Crianza sustancias, naturalmente ello conlleva tráfico, pero como objetivo destaca la fabricación
- Tráfico de drogas, en general. Cuando no se especifican las cantidades, porque no es relevante para la noticia, incluidos los anabolizantes, por ejemplo, la detención de un narco, conflictividad en una zona/barrio por el tráfico de drogas.
- Consecuencias relacionadas con la conducción
- Consecuencias: conflictos/delitos. No relacionados directamente con el tráfico de drogas, ni con las celebrities. También pobreza como consecuencia de las adicciones, juego, heroína…
- Celebrities . Relacionado con personas famosas adictas a las drogas.
- Violencia contra la mujer
- Consecuencias sobre la salud física. Cualquier texto centrado en efectos sobre enfermedades y muerte, no datos
- Consecuencias sobre la salud psíquica. Centrado en trastornos mentales y adición
- Sobre consumo de sustancias. Hace referencias a datos de consumo, se habla de prevalencias del consumo de sustancias, de personas que reciben tratamiento por consumo, datos sobre mortalidad atribuible a las sustancias.
- Sobre trastornos adictivos comportamentales. Incluye los relacionados con las nuevas tecnologías
- Dopaje. Dopaje deportivo, porque no suelen aparecer los relacionados con la vida cotidiana, aunque puede darse el caso, consumo medicamentos para mejorar el rendimiento académico
- Sobre consumo en general. Se habla indistintamente de consumo de sustancias, de nuevas tecnologías, sin perder de vista que estamos en el bloque de datos. Fundamentalmente se trata de dar datos sobre consumo pero no se puede incluir en ninguna de las tres categorías anteriores.
- Prevención. Programas, actividades de este tipo.
- Institucional. Presentación de datos acerca de instituciones/asociaciones, fundaciones… donde la noticia es la propia institución, por ejemplo la presentación de memorias; suelen hacerlo también algunas UCA. Presentación de campañas publicitarias, empresas del sector, así como sus problemas económicos y laborales. También se incluye aquí la política relacionada con las adicciones (Tanto la normativa como las reclamaciones de la oposición)
- Presentación de estudios y resultados de investigaciones. Descubrimientos…
- Famosos y prevención. Actividades de prevención realizadas por famosos.
- Ocio. Ocio y drogas, ocio y adicción a los juegos multimedia, a las loterías… sin hablar de prevención, de consecuencias o de datos de consumo. Ej. Botellón, fiestas rave, qué hacen los jóvenes en el tiempo libre…
- Otros

Tema secundario

Escoger de la lista según los mismos criterios

Drogas: ahora vienen una serie de sustancias en las que hay que decir se aparecen o no en la unidad redaccional que se está analizando. Tabaco; Alcohol; Cannabis; Hachís; Marihuana; Cocaína; Heroína; Crack o Cocaína base; Cristal; Éxtasis o MDMA; Anabolizantes; Psicofármacos; Drogas en general; Otra

Si aparece otra droga que no está en la lista, se marca que sí en otra y se escribe en texto qué droga es en la variable Otra droga.

Dispositivos narrativos y de encuadre. Los siguientes encuadres o dispositivos narrativos los codificaremos como ausentes o sustantivamente presentes en todo el texto. Codificamos la ausencia y la presencia sustantiva. Los encuadres que aparecen de manera insustancial no se consideran. Hay que tener en cuenta que la información contextual o de situación se considera explicación para ayudar a los lectores a entender el texto y no debe codificarse como encuadre. Se considera entradilla el texto resaltado tipográficamente como tal, en caso de que no esté resaltado, codificaremos como entradilla el primer párrafo y si el texto es de un solo párrafo, se contarán las tres primeras líneas.

0 Ausente 1 Presente 2 Destacado en el tema principal, en la entradilla

El encuadre Destacado el tema principal, en la entradilla excluye el encuadre Presente. Sólo se puede elegir uno de los tres en cada encuadre. Se explican a continuación los encuadres.

Nueva investigación. Se centra en la comunicación de una nueva investigación, anuncio de un descubrimiento, nueva aplicación médica o científica, anuncio de resultados de pruebas clínicas (Ej. Estudios públicos, artículos de publicaciones científicas, comunicaciones o ponencias, contenidos científicos transmitidos mediante rueda de prensa)

Contexto general científico médico. Conocimientos generales científicos o médicos sobre el tema. (Ej. Descripción de resultados previos, recapitulación de resultados , hallazgos, descripción de aplicaciones o usos potenciales, tratamientos médicos).

Ética moralidad. Centrado en los aspectos éticos o morales de las investigaciones, anuncios de informes de comités éticos, centrados en una perspectiva religiosa o de valores, énfasis en los bioéticos.

Estrategia política. Centrado en la estrategia política, acción política o deliberación de cifras políticas, administraciones presidenciales, miembros de l congreso u otros funcionarios elegidos bien federales o del estado, o agencias gubernamentales, y la presión de grupos de interés. El enfoque aquí no es sólo específico de acción política, sino más bien mantener, ganar o perder apoyo político y de integrantes.

Política Legislación. Centrado en normas de regulación sobre el tráfico de drogas o consumo.

Mercado Empresa. Centrado en el valor del mercado, el crecimiento del sector o de una empresa, su situación en el mercado, la reacción de los inversores, desarrollo de productos en el mercado, implicaciones para la economía doméstica o competitividad

global, mercado clandestino tales como locales de venta de plantas de cannabis o setas alucinógenas.

Epidemiología. Centrada en los resultados de las últimas encuestas, estadísticas que muestran situaciones de consumo entre diversos colectivos de población.

Opinión pública. Centrada en los resultados de las últimas encuestas, estadísticas que muestran el punto de vista del público, referencias generales a apoyo/rechazo público a las drogas o a la batalla en la opinión pública, así como en las acciones de diversos actores sociales o grupos de presión implicados en el debate.

Opinión no experta. Centrado en la reacción o la opinión del hombre de la calle o alguien no experto, no paciente, líder de una comunidad local, sin relación política con la investigación.

Personalización anecdótica. Centrado en pacientes, familiares o amigos de enfermos que están recibiendo tratamiento relacionado, sufren enfermedades relacionadas con células madre, sufren enfermedades o afecciones relacionadas con ello o pueden beneficiarse de las investigaciones. Se centra en la narrativa personal o testimonial.

*Delito.*Centrado en operaciones policiales o judiciales, incautaciones de estupefacientes, desarticulaciones de redes de narcotráfico, detenciones, procesamientos judiciales, juicios, sentencias todo ello en relación con el tráfico de drogas o con delitos cometidos que tienen como causa o detonante el consumo de estupefacientes o drogas, o el consumo de alcohol.

Fuente manifiesta

Autor del texto. Escoger de la lista.

Firma del periodista.

Si firma un periodista, escribir su nombre tal y como aparece.

Nº de fuentes

Se cuenta el número total de fuentes personales o documentales que aparecen en la unidad redaccional. Estas fuentes se desglosan a continuación. En cada una de ellas sólo hay que contar el número de fuentes distintas que aparecen de cada tipo de manera que la suma de ellas debe ser el valor que se coloque en la variable Nº de fuentes. Cada tipo de fuente se explica a continuación.

Políticos: Incluimos a todas las personas con un cargo oficial (local, autonómico, nacional o internacional y a todos los miembros del sistema político: presidente, ministros, secretarios de estado, delegados autonómicos, locales o provinciales de distintos departamentos relacionados, Ciencia y Tecnología, Sanidad, así como a las

figuras equivalentes en los partidos políticos de la oposición, así como personas con cargos públicos. Se incluyen tantos como aparezcan en el texto.

Fuerzas y Cuerpos de Seguridad del Estado: Personas implicadas en operaciones policiales de narcotráfico, incautación de drogas o estupefacientes, ministerio y ministro del Interior , policía nacional, autonómica y local y guardia civil, servicios de vigilancia de costas, y cualquier cuerpo que pueda estar relacionado con la seguridad....

Tribunal. Tribunales y personal del mundo jurídico: Profesionales de la Justicia, Tribunal Superior de Justicia, Audiencia Nacional, Tribunales de ámbito autonómico y local, así como abogados, magistrados, letrados, etc. que intervengan en las informaciones

Científicos: Investigadores, científicos, biólogos, etc. Se sugiere que participan en el proceso aplicando el método científico. Alguien que trabaja en el departamento de informática científica no contará a no ser que se sugiera que no aplica el método.

Académicos: Asociado con una universidad, escuela o instituto, pero no descrito como científico o investigador. Incluye a profesores, decanos, directores y otros cargos facultativos.

Privada. Asociaciones privadas, profesionales no médicos: Colectivos privados no médicos que ayudan a la rehabilitación de drogodependientes tales como ONG, asociaciones de profesionales, fundaciones, especializadas en la atención a los adictos a los estupefacientes.

Psicosanitarios. Psicólogos y personal de asistencia sanitaria: Aquí se incluyen aquellos profesionales médicos, psicólogos, personal de Unidades de Conductas adictivas, asistentes sociales, etc. Que se dedican a la asistencia directa a los adictos

No expertos: Miembro del público o ciudadano en general. Puede estar asociado con algún grupo ciudadano, pero no es experto en ninguna de las categorías que hemos designado.

Cuando haya fuentes privadas, completar las siguientes variables:

Nombre de la persona. Se escribe el nombre.

Afiliación institucional. Si aparece o no aparece

Nombre de la institución. Se escribe el nombre.

Cargo. Si aparece o no aparece

Nombre del cargo. Se escribe el nombre.

13.2.- ANEXO: ESTADÍSTICOS DE DATOS

13.2.1.- ANEXO: ESTADÍSTICOS DE DATOS NUMÉRICOS

Campo	Mín	Máx	Media	Desv. típica	Válidos
Número	1	161	67,574	42,792	502
Año	2009	2009	2009	0	502
Día	1	31	15,863	8,792	502
Página	1	106	24,754	18,937	500
Extensión	0	444	64,102	71,227	502
Valor Ubicación en el periódico	2	30	4,596	6,771	502
Valor Jerarquía en la página	1	10	1,761	1,328	502
Valor Altura de titulares	1	5	3,598	1,426	502
Valor Ancho de titulares	1	15	4,351	4,086	502
Valor Superficie de titulares	1	5	1,797	1,236	502
Valor Cuerpo de titulares	1	5	3,323	1,834	502
Valor Jerarquía de titulares	0	10	5,765	4,339	502
Valor Nº de imágenes de la unidad de análisis	0	10	1,43	1,749	502
Valor Ilustración prioritaria	0	5	2,171	2,481	502
Valor Tipografía o página especial	0	5	0,598	1,624	502
Valoración de unidad de análisis	7	92	29,388	16,86	502
Análisis cuantitativo de fuentes	0	8	1,209	1,257	502
Fuerzas	1	2	1,078	0,269	116
Tribunal	1	2	1,097	0,301	31
Privada	1	4	1,337	0,688	98
Psico-sanitarios	1	4	1,639	0,899	36

Campo	Mín	Máx	Media	Desv. típica	Válidos
No expertos	1	6	1,419	0,876	74

13.2.2.- ANEXO: ESTADÍSTICOS DE DATOS DISCRETOS

Campo	Únicos	Válidos
Periódico	4	502
Edición	2	502
Mes	6	502
Género	9	502
Sección	10	502
Sección Otros	47	139
Forma de aparición	6	502
Ubicación en el periódico	6	502
Jerarquía en la página	5	502
Altura de titulares	4	502
Ancho de titulares	6	502
Superficie de titulares	3	502
Cuerpo de titulares	3	502
Jerarquía de titulares	5	502
Nº de imágenes de la unidad de análisis	5	501
Ilustración prioritaria	2	502
Tipografía o página especial	2	502
Análisis del tono	4	502
Tema principal	21	439

Campo	Únicos	Válidos
Tema secundario	20	202
Tabaco	2	502
Alcohol	2	502
Cannabis	2	502
Hachís	2	502
Marihuana	2	502
Cocaína	2	502
Heroína	2	502
Crack/Cocaína base	2	502
Crystal/Cristal	2	502
Éxtasis/MDMA	2	502
Anabolizantes	2	502
Psicofármacos	2	502
Drogas en general	2	502
Otra	2	502
Otra droga	31	33
Nueva investigación	3	502
Contexto general científico/médico	3	502
Ética/Moralidad	3	502
Estrategia política	3	502
Política/Legislación	3	502
Mercado/Empresa	3	502
Epidemiología	3	502

Campo	Únicos	Válidos
Opinión pública	3	502
Opinión no experta	3	502
Personalización anecdótica	3	502
Delito	3	502
Fuente manifiesta	7	502
Firma del periodista	232	333
Políticos	3	86
Ciencia	3	18
Académicos	4	19
Nombre de la persona	37	46
Afiliación institucional	2	99
Nombre de la institución	79	99
Cargo	3	43
Nombre del cargo	21	37

13.2.3.- ANEXO: FRECUENCIAS Y PROCENTAJES DE VALORES

Campo	Valor	Frecuencia	Porcentaje
'No expertos'	0. Ausente	428	85,26
'No expertos'	1. Presente	74	14,74
'Psico-sanitarios'	0. Ausente	466	92,83
'Psico-sanitarios'	1. Presente	36	7,17
Académicos	0. Ausente	483	96,22
Académicos	1. Presente	19	3,78

Campo	Valor	Frecuencia	Porcentaje
Afiliación institucional	0. Ausente	403	80,28
Afiliación institucional	1. Presente	99	19,72
Alcohol	0. Ausente	399	79,48
Alcohol	1. Presente	103	20,52
Anabolizantes	0. Ausente	497	99,00
Anabolizantes	1. Presente	5	1,00
Año	2009	502	100,00
Cannabis	0. Ausente	472	94,02
Cannabis	1. Presente	30	5,98
Cantidad de Fuentes	Muchas Fuentes	59	11,75
Cantidad de Fuentes	Pocas fuentes	136	27,09
Cantidad de Fuentes	Suficientes fuentes	307	61,16
Cargo	0. Ausente	466	92,83
Cargo	1. Presente	36	7,17
Categoría Tema Principal	0. No Relacionado con Drogas	63	12,55
Categoría Tema Principal	1. Trafico de Drogas	161	32,07
Categoría Tema Principal	2. Consecuencias	91	18,13
Categoría Tema Principal	3. Consumo	55	10,96
Categoría Tema Principal	4. Estudios y Prevención	104	20,72
Categoría Tema Principal	5. Ocio	20	3,98
Categoría Tema Principal	6. Otros	8	1,59
Ciencia	0. Ausente	484	96,41
Ciencia	1. Presente	18	3,59

Campo	Valor	Frecuencia	Porcentaje
Cocaína	0. Ausente	376	74,90
Cocaína	1. Presente	126	25,10
Contexto general científico/médico	0. Ausente	469	93,43
Contexto general científico/médico	1. Presente	33	6,57
Crack/Cocaína base	0. Ausente	498	99,20
Crack/Cocaína base	1. Presente	4	0,80
Crystal/Cristal	0. Ausente	501	99,80
Crystal/Cristal	1. Presente	1	0,20
Delito	0. Ausente	246	49,00
Delito	1. Presente	256	51,00
Drogas en general	0. Ausente	310	61,75
Drogas en general	1. Presente	192	38,25
Edición	Comunidad Valenciana	423	84,26
Edición	Nacional	79	15,74
Epidemiología	0. Ausente	439	87,45
Epidemiología	1. Presente	63	12,55
EsDomingo	Falso	413	82,27
EsDomingo	Verdadero	89	17,73
Estrategia política	0. Ausente	419	83,47
Estrategia política	1. Presente	83	16,53
Ética/Moralidad	0. Ausente	470	93,63
Ética/Moralidad	1. Presente	32	6,37
Éxtasis/MDMA	0. Ausente	493	98,21

Campo	Valor	Frecuencia	Porcentaje
Éxtasis/MDMA	1. Presente	9	1,79
Extensión	'3. Largo"	25	4,98
Extensión	1. Corto	348	69,32
Extensión	2. Normal	129	25,70
Forma de Aparición	1. Dedicación Principal	411	81,87
Forma de Aparición	2. Incrustación más imagen	20	3,98
Forma de Aparición	3. Referencia más imagen	15	2,99
Forma de Aparición	4. Incrustación	11	2,19
Forma de Aparición	5. Referencia	30	5,98
Forma de Aparición	6. Sin Referencia	15	2,99
Fuente manifiesta	1. Nombre propio	304	60,56
Fuente manifiesta	2. Redacción	11	2,19
Fuente manifiesta	3. Agencia	24	4,78
Fuente manifiesta	4. Corresponsal	18	3,59
Fuente manifiesta	5. Enviado especial	12	2,39
Fuente manifiesta	6. No figura	114	22,71
Fuente manifiesta	7. Otros	19	3,78
Fuerzas	0. Ausente	386	76,89
Fuerzas	1. Presente	116	23,11
Género	Artículo	36	7,17
Género	Columna/Tribuna	6	1,20
Género	Crítica	5	1,00
Género	Crónica	9	1,79

Campo	Valor	Frecuencia	Porcentaje
Género	Editorial	5	1,00
Género	Entrevista	13	2,59
Género	Noticia	363	72,31
Género	Otros	34	6,77
Género	Reportaje	31	6,18
Hachís	0. Ausente	462	92,03
Hachís	1. Presente	40	7,97
Heroína	0. Ausente	468	93,23
Heroína	1. Presente	34	6,77
Marihuana	0. Ausente	467	93,03
Marihuana	1. Presente	35	6,97
Mercado/Empresa	0. Ausente	449	89,44
Mercado/Empresa	1. Presente	53	10,56
MesNumerico	1	95	18,92
MesNumerico	2	84	16,73
MesNumerico	3	87	17,33
MesNumerico	4	83	16,53
MesNumerico	5	76	15,14
MesNumerico	6	77	15,34
Nº de imágenes de la unidad de análisis	1.0000000000000000e+000	234	46,61
Nº de imágenes de la unidad de análisis	2.0000000000000000e+000	33	6,57
Nº de imágenes de la unidad de análisis	3.0000000000000000e+000	4	0,80
Nº de imágenes de la unidad de análisis	Más de tres	9	1,79

Campo	Valor	Frecuencia	Porcentaje
Nº de imágenes de la unidad de análisis	Ninguna	222	44,22
Nueva investigación	0. Ausente	441	87,85
Nueva investigación	1. Presente	61	12,15
Opinión no experta	0. Ausente	484	96,41
Opinión no experta	1. Presente	18	3,59
Opinión pública	0. Ausente	466	92,83
Opinión pública	1. Presente	36	7,17
Otra	0. Ausente	470	93,63
Otra	1. Presente	32	6,37
Periódico	ABC	79	15,74
Periódico	El Mundo	163	32,47
Periódico	El País	151	30,08
Periódico	La Razón	109	21,71
Personalización anecdótica	0. Ausente	415	82,67
Personalización anecdótica	1. Presente	87	17,33
Política/Legislación	0. Ausente	436	86,85
Política/Legislación	1. Presente	66	13,15
Políticos	0. Ausente	416	82,87
Políticos	1. Presente	86	17,13
Privada	0. Ausente	404	80,48
Privada	1. Presente	98	19,52
Psicofármacos	0. Ausente	496	98,80
Psicofármacos	1. Presente	6	1,20

Campo	Valor	Frecuencia	Porcentaje
Sección	Cultura	29	5,78
Sección	Internacional/Mundo	54	10,76
Sección	Nacional/Política/España	75	14,94
Sección	Ocio	1	0,20
Sección	Opinión	23	4,58
Sección	Otros	135	26,89
Sección	Secciones regionales	14	2,79
Sección	Sociedad	83	16,53
Sección	Sucesos	46	9,16
Sección	Suplemento de salud	42	8,37
Tabaco	0. Ausente	389	77,49
Tabaco	1. Presente	113	22,51
Tribunal	0. Ausente	471	93,82
Tribunal	1. Presente	31	6,18
Ubicación en el periódico	Contraportada	2	0,40
Ubicación en el periódico	Página impar apertura de sección	40	7,97
Ubicación en el periódico	Página impar normal	170	33,86
Ubicación en el periódico	Página par apertura de sección	45	8,96
Ubicación en el periódico	Página par normal	216	43,03
Ubicación en el periódico	Portada	29	5,78
Valoración de Unidad de análisis Discreta	Alto	180	35,86
Valoración de Unidad de análisis Discreta	Bajo	149	29,68
Valoración de Unidad de análisis Discreta	Medio	173	34,46

13.3.- ANEXO: ANÁLISIS GENERALES

13.3.1.- ANEXO: RUTA GENERAL

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Crystal/Cristal = 0. Ausente		100	99,80079681	1
Crack/Cocaína base = 0. Ausente		100	99,20318725	1
Anabolizantes = 0. Ausente		100	99,00398406	1
Psicofármacos = 0. Ausente		100	98,80478088	1
Éxtasis/MDMA = 0. Ausente		100	98,20717131	1
Opinión no experta = 0. Ausente		100	96,41434263	1
Ciencia = 0. Ausente		100	96,41434263	1
Académicos = 0. Ausente		100	96,21513944	1
Cannabis = 0. Ausente		100	94,02390438	1
Tribunal = 0. Ausente		100	93,8247012	1
Otra = 0. Ausente		100	93,62549801	1
Ética/Moralidad = 0. Ausente		100	93,62549801	1
Otra droga = N/A		100	93,42629482	1
Contexto general científico/médico = 0. Ausente		100	93,42629482	1
Heroína = 0. Ausente		100	93,22709163	1
Marihuana = 0. Ausente		100	93,02788845	1
Opinión pública = 0. Ausente		100	92,82868526	1
Cargo = 0. Ausente		100	92,82868526	1
'Psico-sanitarios' = 0. Ausente		100	92,82868526	1
Nombre del cargo = N/A		100	92,62948207	1
Hachís = 0. Ausente		100	92,03187251	1
Nombre de la persona = N/A		100	90,83665339	1
Mercado/Empresa = 0. Ausente		100	89,44223108	1
Nueva investigación = 0. Ausente		100	87,84860558	1
Epidemiología = 0. Ausente		100	87,4501992	1
Política/Legislación = 0. Ausente		100	86,85258964	1
'No expertos' = 0. Ausente		100	85,25896414	1
Edición = Comunidad Valenciana		100	84,26294821	1
Estrategia política = 0. Ausente		100	83,46613546	1
Políticos = 0. Ausente		100	82,8685259	1

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Personalización anecdótica = 0. Ausente		100	82,66932271	1
EsDomingo = Falso		100	82,27091633	1
Forma de Aparición = 1. Dedicación Principal		100	81,87250996	1
Privada = 0. Ausente		100	80,47808765	1
Afiliación institucional = 0. Ausente		100	80,27888446	1
Nombre de la institución = N/A		100	80,27888446	1
Afiliación institucional = 0. Ausente	Privada = 0. Ausente	80,47808765	99,5049505	1,239490947
Nombre de la institución = N/A	Privada = 0. Ausente	80,47808765	99,5049505	1,239490947
Nombre de la institución = N/A	Afiliación institucional = 0. Ausente	80,27888446	100	1,245657568
Afiliación institucional = 0. Ausente	Nombre de la institución = N/A	80,27888446	100	1,245657568
Privada = 0. Ausente	Afiliación institucional = 0. Ausente	80,27888446	99,75186104	1,239490947
Privada = 0. Ausente	Nombre de la institución = N/A	80,27888446	99,75186104	1,239490947
Tabaco = 0. Ausente	Delito = 1. Presente	50,99601594	97,265625	1,255201639
Afiliación institucional = 0. Ausente	Delito = 1. Presente	50,99601594	91,796875	1,143474721
Nombre de la institución = N/A	Delito = 1. Presente	50,99601594	91,796875	1,143474721
Privada = 0. Ausente	Delito = 1. Presente	50,99601594	91,796875	1,140644338
Epidemiología = 0. Ausente	Delito = 1. Presente	50,99601594	97,65625	1,116707005
Personalización anecdótica = 0. Ausente	Delito = 1. Presente	50,99601594	92,1875	1,115135542
Fuerzas = 0. Ausente	Delito = 0. Ausente	49,00398406	95,93495935	1,247651544
Cocaína = 0. Ausente	Delito = 0. Ausente	49,00398406	85,77235772	1,145152223
Personalización anecdótica = 0. Ausente	Nº de imágenes de la unidad de análisis = Ninguna	44,22310757	93,69369369	1,133355042
Tabaco = 0. Ausente	Drogas en general = 1. Presente	38,24701195	93,75	1,209832905
Alcohol = 0. Ausente	Drogas en general = 1. Presente	38,24701195	89,0625	1,120535714
Fuente manifiesta = 1. Nombre propio	Valoración de Unidad de análisis Discreta = Alto	35,85657371	85,55555556	1,412792398
Género = Noticia	Ubicación en el periódico = Página impar normal	33,86454183	82,35294118	1,138875385
Edición = Comunidad Valenciana	Periódico = El Mundo	32,47011952	100	1,186761229
Delito = 1. Presente	Categoría Tema Principal = 1. Trafico de Drogas	32,07171315	94,40993789	1,851319876
Tabaco = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	32,07171315	98,75776398	1,27445752
Alcohol = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	32,07171315	99,37888199	1,250330796
Privada = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	32,07171315	93,78881988	1,165395732

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Afiliación institucional = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	32,07171315	93,16770186	1,160550529
Nombre de la institución = N/A	Categoría Tema Principal = 1. Trafico de Drogas	32,07171315	93,16770186	1,160550529
Epidemiología = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	32,07171315	100	1,143507973
Personalización anecdótica = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	32,07171315	94,40993789	1,142019008
Política/Legislación = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	32,07171315	97,51552795	1,122770528
'No expertos' = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	32,07171315	95,0310559	1,11461659
Edición = Comunidad Valenciana	Periódico = El País	30,07968127	100	1,186761229
Nº de imágenes de la unidad de análisis = Ninguna	Valoración de Unidad de análisis Discreta = Bajo	29,6812749	80,53691275	1,821150009
Personalización anecdótica = 0. Ausente	Valoración de Unidad de análisis Discreta = Bajo	29,6812749	95,30201342	1,152809897
Fuerzas = 0. Ausente	Cantidad de Fuentes = Pocas fuentes	27,09163347	100	1,300518135
Afiliación institucional = 0. Ausente	Cantidad de Fuentes = Pocas fuentes	27,09163347	100	1,245657568
Nombre de la institución = N/A	Cantidad de Fuentes = Pocas fuentes	27,09163347	100	1,245657568
Privada = 0. Ausente	Cantidad de Fuentes = Pocas fuentes	27,09163347	100	1,242574257
Políticos = 0. Ausente	Cantidad de Fuentes = Pocas fuentes	27,09163347	100	1,206730769
'No expertos' = 0. Ausente	Cantidad de Fuentes = Pocas fuentes	27,09163347	100	1,172897196
Forma de Aparición = 1. Dedicación Principal	Extensión = 2. Normal	25,69721116	98,4496124	1,202474585
Tabaco = 0. Ausente	Cocaína = 1. Presente	25,09960159	91,26984127	1,177826743
Cannabis = 0. Ausente	Cocaína = 1. Presente	25,09960159	84,12698413	0,894740382
Delito = 1. Presente	Fuerzas = 1. Presente	23,10756972	91,37931034	1,791891164
Cantidad de Fuentes = Suficientes fuentes	Fuerzas = 1. Presente	23,10756972	86,20689655	1,409637201
Tabaco = 0. Ausente	Fuerzas = 1. Presente	23,10756972	97,4137931	1,257113731
Afiliación institucional = 0. Ausente	Fuerzas = 1. Presente	23,10756972	96,55172414	1,202703859
Nombre de la institución = N/A	Fuerzas = 1. Presente	23,10756972	96,55172414	1,202703859
Privada = 0. Ausente	Fuerzas = 1. Presente	23,10756972	96,55172414	1,199726869
Epidemiología = 0. Ausente	Fuerzas = 1. Presente	23,10756972	98,27586207	1,123792318
Personalización anecdótica = 0. Ausente	Fuerzas = 1. Presente	23,10756972	92,24137931	1,115787287
Hachís = 0. Ausente	Fuerzas = 1. Presente	23,10756972	80,17241379	0,871137483
Delito = 0. Ausente	Tabaco = 1. Presente	22,50996016	93,80530973	1,914238434
Drogas en general = 0. Ausente	Tabaco = 1. Presente	22,50996016	89,38053097	1,447387953
Fuerzas = 0. Ausente	Tabaco = 1. Presente	22,50996016	97,34513274	1,265991105
Cocaína = 0. Ausente	Tabaco = 1. Presente	22,50996016	90,26548673	1,205140275

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Edición = Comunidad Valenciana	Tabaco = 1. Presente	22,50996016	93,80530973	1,113245047
'Psico-sanitarios' = 0. Ausente	Tabaco = 1. Presente	22,50996016	82,30088496	0,886588932
Contexto general científico/médico = 0. Ausente	Tabaco = 1. Presente	22,50996016	82,30088496	0,880917788
Edición = Comunidad Valenciana	Periódico = La Razón	21,71314741	100	1,186761229
Delito = 0. Ausente	Categoría Tema Principal = 4. Estudios y Prevención	20,71713147	84,61538462	1,72670419
Fuerzas = 0. Ausente	Categoría Tema Principal = 4. Estudios y Prevención	20,71713147	96,15384615	1,250498206
Personalización anecdótica = 0. Ausente	Categoría Tema Principal = 4. Estudios y Prevención	20,71713147	96,15384615	1,163113994
Cocaína = 0. Ausente	Categoría Tema Principal = 4. Estudios y Prevención	20,71713147	86,53846154	1,155380524
Nombre de la persona = N/A	Categoría Tema Principal = 4. Estudios y Prevención	20,71713147	81,73076923	0,899755398
Estrategia política = 0. Ausente	Alcohol = 1. Presente	20,51792829	96,11650485	1,151562898
Privada = 1. Presente	Afiliación institucional = 1. Presente	19,72111554	97,97979798	5,018965162
Fuerzas = 0. Ausente	Afiliación institucional = 1. Presente	19,72111554	95,95959596	1,247971947
Cannabis = 0. Ausente	Afiliación institucional = 1. Presente	19,72111554	83,83838384	0,891670947
Afiliación institucional = 1. Presente	Privada = 1. Presente	19,52191235	98,97959184	5,018965162
Fuerzas = 0. Ausente	Privada = 1. Presente	19,52191235	95,91836735	1,247435762
Cocaína = 0. Ausente	Categoría Tema Principal = 2. Consecuencias	18,12749004	87,91208791	1,173719897
Estrategia política = 0. Ausente	Categoría Tema Principal = 2. Consecuencias	18,12749004	95,6043956	1,145427365
Fuerzas = 0. Ausente	Personalización anecdótica = 1. Presente	17,33067729	89,65517241	1,165981776
Estrategia política = 0. Ausente	Personalización anecdótica = 1. Presente	17,33067729	96,55172414	1,15677722
Políticos = 0. Ausente	Personalización anecdótica = 1. Presente	17,33067729	95,40229885	1,151248895
Cocaína = 0. Ausente	Personalización anecdótica = 1. Presente	17,33067729	83,90804598	1,120261678
Marihuana = 0. Ausente	Personalización anecdótica = 1. Presente	17,33067729	82,75862069	0,889610869
Personalización anecdótica = 0. Ausente	Políticos = 1. Presente	17,1314741	95,34883721	1,153376296
'No expertos' = 0. Ausente	Políticos = 1. Presente	17,1314741	97,6744186	1,145620517
Alcohol = 0. Ausente	Estrategia política = 1. Presente	16,53386454	95,18072289	1,197511852
Personalización anecdótica = 0. Ausente	Estrategia política = 1. Presente	16,53386454	96,38554217	1,165916679
Forma de Aparición = 1. Dedicación Principal	Sección = Sociedad	16,53386454	95,18072289	1,162548002
Política/Legislación = 0. Ausente	Estrategia política = 1. Presente	16,53386454	100	1,151376147
Género = Noticia	Sección = Sociedad	16,53386454	83,13253012	1,149656477

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Edición = Nacional	Periódico = ABC	15,73705179	100	6,35443038
Periódico = ABC	Edición = Nacional	15,73705179	100	6,35443038
Tabaco = 0. Ausente	Periódico = ABC	15,73705179	91,13924051	1,176141356
Tabaco = 0. Ausente	Edición = Nacional	15,73705179	91,13924051	1,176141356
Género = Noticia	Sección = Nacional/Política/España	14,94023904	97,33333333	1,346042241
Cocaína = 0. Ausente	Sección = Nacional/Política/España	14,94023904	85,33333333	1,13929078
Políticos = 0. Ausente	'No expertos' = 1. Presente	14,74103586	97,2972973	1,174116424
Delito = 0. Ausente	Política/Legislación = 1. Presente	13,14741036	81,81818182	1,66962306
Cocaína = 0. Ausente	Política/Legislación = 1. Presente	13,14741036	92,42424242	1,23396196
Estrategia política = 0. Ausente	Política/Legislación = 1. Presente	13,14741036	100	1,198090692
Fuerzas = 0. Ausente	Política/Legislación = 1. Presente	13,14741036	87,87878788	1,142879573
Ética/Moralidad = 0. Ausente	Política/Legislación = 1. Presente	13,14741036	83,33333333	0,890070922
Delito = 0. Ausente	Epidemiología = 1. Presente	12,5498008	90,47619048	1,846302749
Fuente manifiesta = 1. Nombre propio	Categoría Tema Principal = 0. No Relacionado con Drogas	12,5498008	82,53968254	1,36299081
Cocaína = 0. Ausente	Categoría Tema Principal = 0. No Relacionado con Drogas	12,5498008	95,23809524	1,271529889
Fuerzas = 0. Ausente	Epidemiología = 1. Presente	12,5498008	96,82539683	1,259231845
Afiliación institucional = 0. Ausente	Categoría Tema Principal = 0. No Relacionado con Drogas	12,5498008	92,06349206	1,146795856
Nombre de la institución = N/A	Categoría Tema Principal = 0. No Relacionado con Drogas	12,5498008	92,06349206	1,146795856
Epidemiología = 0. Ausente	Categoría Tema Principal = 0. No Relacionado con Drogas	12,5498008	98,41269841	1,125357052
Privada = 0. Ausente	Categoría Tema Principal = 0. No Relacionado con Drogas	12,5498008	90,47619048	1,124233852
Nombre de la persona = N/A	Epidemiología = 1. Presente	12,5498008	80,95238095	0,891186299
Heroína = 0. Ausente	Epidemiología = 1. Presente	12,5498008	82,53968254	0,885361552
Académicos = 0. Ausente	Epidemiología = 1. Presente	12,5498008	84,12698413	0,874363272
Delito = 0. Ausente	Nueva investigación = 1. Presente	12,15139442	85,24590164	1,739570838
Fuerzas = 0. Ausente	Nueva investigación = 1. Presente	12,15139442	91,80327869	1,193918288
Estrategia política = 0. Ausente	Nueva investigación = 1. Presente	12,15139442	95,08196721	1,139168199
Nombre de la persona = N/A	Nueva investigación = 1. Presente	12,15139442	80,32786885	0,884311188
Opinión pública = 0. Ausente	Nueva investigación = 1. Presente	12,15139442	81,96721311	0,882994442
'Psico-sanitarios' = 0.	Nueva investigación = 1.	12,15139442	81,96721311	0,882994442

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Ausente	Presente			
Académicos = 0. Ausente	Nueva investigación = 1. Presente	12,15139442	83,60655738	0,868954282
Ciencia = 0. Ausente	Nueva investigación = 1. Presente	12,15139442	80,32786885	0,833152689
Género = Noticia	Cantidad de Fuentes = Muchas Fuentes	11,75298805	83,05084746	1,148526871
Académicos = 0. Ausente	Cantidad de Fuentes = Muchas Fuentes	11,75298805	86,44067797	0,898410359
Opinión pública = 0. Ausente	Cantidad de Fuentes = Muchas Fuentes	11,75298805	83,05084746	0,894667928
Delito = 0. Ausente	Categoría Tema Principal = 3. Consumo	10,9561753	89,09090909	1,818033999
Fuerzas = 0. Ausente	Categoría Tema Principal = 3. Consumo	10,9561753	96,36363636	1,253226566
Forma de Aparición = 1. Dedicación Principal	Categoría Tema Principal = 3. Consumo	10,9561753	92,72727273	1,132581287
'Psico-sanitarios' = 0. Ausente	Categoría Tema Principal = 3. Consumo	10,9561753	81,81818182	0,881388997
Otra = 0. Ausente	Categoría Tema Principal = 3. Consumo	10,9561753	81,81818182	0,873887814
Cannabis = 0. Ausente	Categoría Tema Principal = 3. Consumo	10,9561753	81,81818182	0,8701849
Otra droga = N/A	Categoría Tema Principal = 3. Consumo	10,9561753	80	0,856289979
Género = Noticia	Sección = Internacional/Mundo	10,75697211	85,18518519	1,178043057
Alcohol = 0. Ausente	Sección = Internacional/Mundo	10,75697211	90,74074074	1,141650422
Personalización anecdótica = 0. Ausente	Mercado/Empresa = 1. Presente	10,55776892	98,11320755	1,186815185
Forma de Aparición = 1. Dedicación Principal	Mercado/Empresa = 1. Presente	10,55776892	92,45283019	1,129229215
Alcohol = 0. Ausente	Mercado/Empresa = 1. Presente	10,55776892	88,67924528	1,115713813
Nombre del cargo = N/A	Mercado/Empresa = 1. Presente	10,55776892	83,01886792	0,896246703
Género = Noticia	Sección = Sucesos	9,163346614	97,82608696	1,35285663
Extensión = 1. Corto	Sección = Sucesos	9,163346614	82,60869565	1,191654173
Edición = Comunidad Valenciana	Sección = Sucesos	9,163346614	100	1,186761229
Forma de Aparición = 1. Dedicación Principal	Sección = Sucesos	9,163346614	93,47826087	1,141753941
Alcohol = 0. Ausente	Ubicación en el periódico = Página par apertura de sección	8,964143426	88,88888889	1,118351434
Edición = Comunidad Valenciana	Sección = Suplemento de salud	8,366533865	100	1,186761229
Otra = 0. Ausente	Sección = Suplemento de salud	8,366533865	83,33333333	0,890070922
Categoría Tema Principal = 1. Trafico de Drogas	Hachís = 1. Presente	7,96812749	82,5	2,572360248
Delito = 1. Presente	Hachís = 1. Presente	7,96812749	92,5	1,813867188
Drogas en general = 0. Ausente	Hachís = 1. Presente	7,96812749	80	1,295483871

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Alcohol = 0. Ausente	Hachís = 1. Presente	7,96812749	97,5	1,226691729
Tabaco = 0. Ausente	Hachís = 1. Presente	7,96812749	95	1,22596401
Estrategia política = 0. Ausente	Ubicación en el periódico = Página impar apertura de sección	7,96812749	100	1,198090692
Políticos = 0. Ausente	Ubicación en el periódico = Página impar apertura de sección	7,96812749	95	1,146394231
Políticos = 0. Ausente	Hachís = 1. Presente	7,96812749	95	1,146394231
Afiliación institucional = 0. Ausente	Hachís = 1. Presente	7,96812749	90	1,121091811
Nombre de la institución = N/A	Hachís = 1. Presente	7,96812749	90	1,121091811
Personalización anecdótica = 0. Ausente	Hachís = 1. Presente	7,96812749	92,5	1,118915663
Privada = 0. Ausente	Hachís = 1. Presente	7,96812749	90	1,118316832
Afiliación institucional = 1. Presente	Cargo = 1. Presente	7,171314741	100	5,070707071
Privada = 1. Presente	Cargo = 1. Presente	7,171314741	97,22222222	4,98015873
Delito = 0. Ausente	Opinión pública = 1. Presente	7,171314741	91,66666667	1,870596206
Delito = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	86,11111111	1,757226739
Fuente manifiesta = 1. Nombre propio	Cargo = 1. Presente	7,171314741	91,66666667	1,51370614
Fuente manifiesta = 1. Nombre propio	'Psico-sanitarios' = 1. Presente	7,171314741	80,55555556	1,330226608
Drogas en general = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	80,55555556	1,304480287
Drogas en general = 0. Ausente	Opinión pública = 1. Presente	7,171314741	80,55555556	1,304480287
Fuerzas = 0. Ausente	Opinión pública = 1. Presente	7,171314741	94,44444444	1,228267127
Fuerzas = 0. Ausente	Cargo = 1. Presente	7,171314741	94,44444444	1,228267127
Fuerzas = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	91,66666667	1,192141623
Cocaína = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	88,88888889	1,186761229
Cocaína = 0. Ausente	Opinión pública = 1. Presente	7,171314741	86,11111111	1,149674941
Estrategia política = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	94,44444444	1,131530098
Ciencia = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	86,11111111	0,893135904
Otra = 0. Ausente	Opinión pública = 1. Presente	7,171314741	83,33333333	0,890070922
Nombre de la persona = N/A	'Psico-sanitarios' = 1. Presente	7,171314741	80,55555556	0,886817739
Académicos = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	83,33333333	0,866114562
Otra droga = N/A	'Psico-sanitarios' = 1. Presente	7,171314741	80,55555556	0,862236437
Otra droga = N/A	Opinión pública = 1. Presente	7,171314741	80,55555556	0,862236437

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Otra = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	80,55555556	0,860401891
Tabaco = 0. Ausente	Marihuana = 1. Presente	6,972111554	97,14285714	1,253617334
EsDomingo = Falso	Marihuana = 1. Presente	6,972111554	94,28571429	1,146039433
Alcohol = 0. Ausente	Marihuana = 1. Presente	6,972111554	88,57142857	1,114357322
Ética/Moralidad = 0. Ausente	Marihuana = 1. Presente	6,972111554	80	0,854468085
Extensión = 1. Corto	Género = Otros	6,772908367	97,05882353	1,40010142
Fuente manifiesta = 1. Nombre propio	Heroína = 1. Presente	6,772908367	82,35294118	1,359907121
Estrategia política = 0. Ausente	Género = Otros	6,772908367	100	1,198090692
Políticos = 0. Ausente	Género = Otros	6,772908367	94,11764706	1,135746606
Política/Legislación = 0. Ausente	Heroína = 1. Presente	6,772908367	97,05882353	1,117512142
Forma de Aparición = 1. Dedicación Principal	Género = Otros	6,772908367	91,17647059	1,113639616
Cannabis = 0. Ausente	Heroína = 1. Presente	6,772908367	82,35294118	0,875872383
Valoración de Unidad de análisis Discreta = Alto	Nº de imágenes de la unidad de análisis = 2.0000000000000000e+000	6,573705179	90,90909091	2,535353535
Delito = 0. Ausente	Contexto general científico/médico = 1. Presente	6,573705179	96,96969697	1,978812515
Fuente manifiesta = 1. Nombre propio	Nº de imágenes de la unidad de análisis = 2.0000000000000000e+000	6,573705179	81,81818182	1,351076555
Drogas en general = 0. Ausente	Contexto general científico/médico = 1. Presente	6,573705179	81,81818182	1,324926686
Fuerzas = 0. Ausente	Contexto general científico/médico = 1. Presente	6,573705179	100	1,300518135
Personalización anecdótica = 0. Ausente	Contexto general científico/médico = 1. Presente	6,573705179	100	1,209638554
Estrategia política = 0. Ausente	Contexto general científico/médico = 1. Presente	6,573705179	100	1,198090692
'No expertos' = 0. Ausente	Contexto general científico/médico = 1. Presente	6,573705179	100	1,172897196
Cocaína = 0. Ausente	Contexto general científico/médico = 1. Presente	6,573705179	84,84848485	1,132817537
Académicos = 0. Ausente	Contexto general científico/médico = 1. Presente	6,573705179	84,84848485	0,881862099
Delito = 0. Ausente	Ética/Moralidad = 1. Presente	6,374501992	93,75	1,913109756
Drogas en general = 0. Ausente	Otra = 1. Presente	6,374501992	81,25	1,315725806
Cocaína = 0. Ausente	Ética/Moralidad = 1. Presente	6,374501992	90,625	1,20994016
Fuerzas = 0. Ausente	Ética/Moralidad = 1.	6,374501992	90,625	1,17859456

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
	Presente			
Forma de Aparición = 1. Dedicación Principal	Otra = 1. Presente	6,374501992	93,75	1,145072993
Tabaco = 0. Ausente	Ética/Moralidad = 1. Presente	6,374501992	87,5	1,129177378
Mercado/Empresa = 0. Ausente	Ética/Moralidad = 1. Presente	6,374501992	100	1,118040089
Política/Legislación = 0. Ausente	Otra = 1. Presente	6,374501992	96,875	1,115395642
Anabolizantes = 0. Ausente	Otra = 1. Presente	6,374501992	87,5	0,883802817
Nombre del cargo = N/A	Otra = 1. Presente	6,374501992	81,25	0,877150538
Académicos = 0. Ausente	Otra = 1. Presente	6,374501992	84,375	0,876940994
Opinión pública = 0. Ausente	Otra = 1. Presente	6,374501992	81,25	0,87526824
Ciencia = 0. Ausente	Otra = 1. Presente	6,374501992	84,375	0,875129132
Cannabis = 0. Ausente	Otra = 1. Presente	6,374501992	81,25	0,864141949
Delito = 1. Presente	Tribunal = 1. Presente	6,175298805	83,87096774	1,644657258
Alcohol = 0. Ausente	Tribunal = 1. Presente	6,175298805	96,77419355	1,217560029
Tabaco = 0. Ausente	Tribunal = 1. Presente	6,175298805	93,5483871	1,207231114
Extensión = 1. Corto	Tribunal = 1. Presente	6,175298805	80,64516129	1,163329626
Nueva investigación = 0. Ausente	Tribunal = 1. Presente	6,175298805	100	1,138321995
Otra = 0. Ausente	Género = Reportaje	6,175298805	83,87096774	0,895813315
Ética/Moralidad = 0. Ausente	Género = Reportaje	6,175298805	83,87096774	0,895813315
Hachís = 0. Ausente	Tribunal = 1. Presente	6,175298805	80,64516129	0,876274263
Extensión = 1. Corto	Forma de Aparición = 5. Referencia	5,976095618	100	1,442528736
Drogas en general = 0. Ausente	Cannabis = 1. Presente	5,976095618	80	1,295483871
Cocaína = 0. Ausente	Forma de Aparición = 5. Referencia	5,976095618	86,66666667	1,157092199
Extensión = 1. Corto	Cannabis = 1. Presente	5,976095618	80	1,154022989
Alcohol = 0. Ausente	Forma de Aparición = 5. Referencia	5,976095618	90	1,132330827
Fuerzas = 0. Ausente	Forma de Aparición = 5. Referencia	5,976095618	86,66666667	1,127115717
Fuerzas = 0. Ausente	Cannabis = 1. Presente	5,976095618	86,66666667	1,127115717
'Psico-sanitarios' = 0. Ausente	Cannabis = 1. Presente	5,976095618	83,33333333	0,897711016
Hachís = 0. Ausente	Cannabis = 1. Presente	5,976095618	80	0,869264069
Heroína = 0. Ausente	Cannabis = 1. Presente	5,976095618	80	0,858119658
Otra = 0. Ausente	Cannabis = 1. Presente	5,976095618	80	0,854468085
Valoración de Unidad de análisis Discreta = Alto	Ubicación en el periódico = Portada	5,77689243	100	2,788888889
Nº de imágenes de la unidad de análisis = 1.0000000000000000e+000	Ubicación en el periódico = Portada	5,77689243	89,65517241	1,923371648
Extensión = 1. Corto	Ubicación en el periódico = Portada	5,77689243	93,10344828	1,343043995

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Extensión = 1. Corto	Sección = Cultura	5,77689243	82,75862069	1,193816885
Edición = Comunidad Valenciana	Ubicación en el periódico = Portada	5,77689243	100	1,186761229
EsDomingo = Falso	Ubicación en el periódico = Portada	5,77689243	96,55172414	1,1735827
Política/Legislación = 0. Ausente	Ubicación en el periódico = Portada	5,77689243	96,55172414	1,111673521
Cannabis = 0. Ausente	Ubicación en el periódico = Portada	5,77689243	82,75862069	0,880187025
Forma de Aparición = 1. Dedicación Principal	Extensión = '3. Largo''	4,980079681	100	1,221411192
Cocaína = 0. Ausente	Extensión = '3. Largo''	4,980079681	88	1,174893617
Alcohol = 0. Ausente	Extensión = '3. Largo''	4,980079681	92	1,157493734
Fuerzas = 0. Ausente	Extensión = '3. Largo''	4,980079681	88	1,144455959
Privada = 0. Ausente	Extensión = '3. Largo''	4,980079681	92	1,143168317
Edición = Comunidad Valenciana	Extensión = '3. Largo''	4,980079681	96	1,13929078
Mercado/Empresa = 0. Ausente	Extensión = '3. Largo''	4,980079681	80	0,894432071
Nº de imágenes de la unidad de análisis = Ninguna	Fuente manifiesta = 3. Agencia	4,780876494	87,5	1,978603604
'No expertos' = 0. Ausente	Fuente manifiesta = 3. Agencia	4,780876494	100	1,172897196
Personalización anecdótica = 0. Ausente	Fuente manifiesta = 3. Agencia	4,780876494	95,83333333	1,159236948
Estrategia política = 0. Ausente	Fuente manifiesta = 3. Agencia	4,780876494	95,83333333	1,148170247
Afiliación institucional = 0. Ausente	Fuente manifiesta = 3. Agencia	4,780876494	91,66666667	1,141852771
Nombre de la institución = N/A	Fuente manifiesta = 3. Agencia	4,780876494	91,66666667	1,141852771
Privada = 0. Ausente	Fuente manifiesta = 3. Agencia	4,780876494	91,66666667	1,139026403
Tabaco = 0. Ausente	Fuente manifiesta = 3. Agencia	4,780876494	87,5	1,129177378
Extensión = 1. Corto	Sección = Opinión	4,581673307	82,60869565	1,191654173
Edición = Comunidad Valenciana	Sección = Opinión	4,581673307	95,65217391	1,135162915
Delito = 0. Ausente	Categoría Tema Principal = 5. Ocio	3,984063745	80	1,632520325
Drogas en general = 0. Ausente	Categoría Tema Principal = 5. Ocio	3,984063745	90	1,457419355
Extensión = 1. Corto	Forma de Aparición = 2. Incrustación más imagen	3,984063745	95	1,370402299
Cantidad de Fuentes = Suficientes fuentes	Forma de Aparición = 2. Incrustación más imagen	3,984063745	80	1,308143322
Tabaco = 0. Ausente	Categoría Tema Principal = 5. Ocio	3,984063745	95	1,22596401
Estrategia política = 0. Ausente	Categoría Tema Principal = 5. Ocio	3,984063745	100	1,198090692
Edición = Comunidad Valenciana	Forma de Aparición = 2. Incrustación más imagen	3,984063745	100	1,186761229
Políticos = 0. Ausente	Categoría Tema Principal = 5. Ocio	3,984063745	95	1,146394231

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Afiliación institucional = 0. Ausente	Categoría Tema Principal = 5. Ocio	3,984063745	90	1,121091811
Nombre de la institución = N/A	Categoría Tema Principal = 5. Ocio	3,984063745	90	1,121091811
Privada = 0. Ausente	Categoría Tema Principal = 5. Ocio	3,984063745	90	1,118316832
Mercado/Empresa = 0. Ausente	Forma de Aparición = 2. Incrustación más imagen	3,984063745	100	1,118040089
Opinión no experta = 0. Ausente	Categoría Tema Principal = 5. Ocio	3,984063745	85	0,88161157
'Psico-sanitarios' = 0. Ausente	Categoría Tema Principal = 5. Ocio	3,984063745	80	0,861802575
Cannabis = 0. Ausente	Categoría Tema Principal = 5. Ocio	3,984063745	80	0,850847458
Extensión = 1. Corto	Académicos = 1. Presente	3,784860558	89,47368421	1,290683606
Fuerzas = 0. Ausente	Fuente manifiesta = 7. Otros	3,784860558	94,73684211	1,232069812
Personalización anecdótica = 0. Ausente	Fuente manifiesta = 7. Otros	3,784860558	100	1,209638554
Cocaína = 0. Ausente	Fuente manifiesta = 7. Otros	3,784860558	89,47368421	1,194568869
Fuerzas = 0. Ausente	Académicos = 1. Presente	3,784860558	89,47368421	1,163621489
Personalización anecdótica = 0. Ausente	Académicos = 1. Presente	3,784860558	94,73684211	1,145973367
Políticos = 0. Ausente	Fuente manifiesta = 7. Otros	3,784860558	94,73684211	1,143218623
Nueva investigación = 0. Ausente	Fuente manifiesta = 7. Otros	3,784860558	100	1,138321995
Alcohol = 0. Ausente	Fuente manifiesta = 7. Otros	3,784860558	89,47368421	1,125709009
Cannabis = 0. Ausente	Académicos = 1. Presente	3,784860558	84,21052632	0,895628903
Delito = 0. Ausente	Ciencia = 1. Presente	3,585657371	94,44444444	1,927280939
Fuerzas = 0. Ausente	Ciencia = 1. Presente	3,585657371	100	1,300518135
Fuerzas = 0. Ausente	Opinión no experta = 1. Presente	3,585657371	94,44444444	1,228267127
Cocaína = 0. Ausente	Opinión no experta = 1. Presente	3,585657371	88,88888889	1,186761229
Género = Noticia	Ciencia = 1. Presente	3,585657371	83,33333333	1,152433425
Política/Legislación = 0. Ausente	Fuente manifiesta = 4. Corresponsal	3,585657371	100	1,151376147
Política/Legislación = 0. Ausente	Ciencia = 1. Presente	3,585657371	100	1,151376147
Tabaco = 0. Ausente	Fuente manifiesta = 4. Corresponsal	3,585657371	88,88888889	1,147100828
Epidemiología = 0. Ausente	Opinión no experta = 1. Presente	3,585657371	100	1,143507973
Epidemiología = 0. Ausente	Fuente manifiesta = 4. Corresponsal	3,585657371	100	1,143507973
Personalización anecdótica = 0. Ausente	Opinión no experta = 1. Presente	3,585657371	94,44444444	1,142436412
Personalización anecdótica = 0. Ausente	Ciencia = 1. Presente	3,585657371	94,44444444	1,142436412
Políticos = 0. Ausente	Ciencia = 1. Presente	3,585657371	94,44444444	1,139690171
Estrategia política = 0.	Ciencia = 1. Presente	3,585657371	94,44444444	1,131530098

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Ausente				
Cocaína = 0. Ausente	Fuente manifiesta = 4. Corresponsal	3,585657371	83,33333333	1,112588652
Cocaína = 0. Ausente	Ciencia = 1. Presente	3,585657371	83,33333333	1,112588652
Psicofármacos = 0. Ausente	Ciencia = 1. Presente	3,585657371	88,88888889	0,899641577
Extensión = 1. Corto	Forma de Aparición = 6. Sin Referencia	2,988047809	100	1,442528736
Extensión = 1. Corto	Forma de Aparición = 3. Referencia más imagen	2,988047809	93,33333333	1,346360153
Drogas en general = 0. Ausente	Forma de Aparición = 3. Referencia más imagen	2,988047809	80	1,295483871
EsDomingo = Falso	Forma de Aparición = 6. Sin Referencia	2,988047809	100	1,215496368
Edición = Comunidad Valenciana	Forma de Aparición = 6. Sin Referencia	2,988047809	100	1,186761229
Edición = Comunidad Valenciana	Forma de Aparición = 3. Referencia más imagen	2,988047809	100	1,186761229
Tabaco = 0. Ausente	Forma de Aparición = 3. Referencia más imagen	2,988047809	86,66666667	1,118423308
Mercado/Empresa = 0. Ausente	Forma de Aparición = 6. Sin Referencia	2,988047809	100	1,118040089
Ciencia = 0. Ausente	Forma de Aparición = 6. Sin Referencia	2,988047809	86,66666667	0,898898072
Éxtasis/MDMA = 0. Ausente	Forma de Aparición = 3. Referencia más imagen	2,988047809	86,66666667	0,882488168
Opinión pública = 0. Ausente	Forma de Aparición = 3. Referencia más imagen	2,988047809	80	0,861802575
Tribunal = 0. Ausente	Forma de Aparición = 6. Sin Referencia	2,988047809	80	0,852653928
Periódico = La Razón	Sección = Secciones regionales	2,788844622	92,85714286	4,276539974
Forma de Aparición = 1. Dedicación Principal	Sección = Secciones regionales	2,788844622	100	1,221411192
Políticos = 0. Ausente	Sección = Secciones regionales	2,788844622	100	1,206730769
Estrategia política = 0. Ausente	Sección = Secciones regionales	2,788844622	100	1,198090692
Edición = Comunidad Valenciana	Sección = Secciones regionales	2,788844622	100	1,186761229
Género = Noticia	Sección = Secciones regionales	2,788844622	85,71428571	1,185360094
Política/Legislación = 0. Ausente	Sección = Secciones regionales	2,788844622	100	1,151376147
Cocaína = 0. Ausente	Sección = Secciones regionales	2,788844622	85,71428571	1,1443769
Ciencia = 0. Ausente	Sección = Secciones regionales	2,788844622	85,71428571	0,889020071
Cocaína = 0. Ausente	Género = Entrevista	2,589641434	100	1,335106383
Epidemiología = 0. Ausente	Género = Entrevista	2,589641434	100	1,143507973
Nueva investigación = 0. Ausente	Género = Entrevista	2,589641434	100	1,138321995
Periódico = El Mundo	Fuente manifiesta = 5. Enviado especial	2,390438247	83,33333333	2,566462168
Delito = 1. Presente	Fuente manifiesta = 5. Enviado especial	2,390438247	91,66666667	1,797526042

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Género = Noticia	Fuente manifiesta = 5. Enviado especial	2,390438247	91,66666667	1,267676768
Tabaco = 0. Ausente	Fuente manifiesta = 5. Enviado especial	2,390438247	91,66666667	1,182947729
Política/Legislación = 0. Ausente	Fuente manifiesta = 5. Enviado especial	2,390438247	100	1,151376147
Epidemiología = 0. Ausente	Fuente manifiesta = 5. Enviado especial	2,390438247	100	1,143507973
Nueva investigación = 0. Ausente	Fuente manifiesta = 5. Enviado especial	2,390438247	100	1,138321995
'Psico-sanitarios' = 0. Ausente	Fuente manifiesta = 5. Enviado especial	2,390438247	83,33333333	0,897711016
Marihuana = 0. Ausente	Fuente manifiesta = 5. Enviado especial	2,390438247	83,33333333	0,895788722
Tribunal = 0. Ausente	Fuente manifiesta = 5. Enviado especial	2,390438247	83,33333333	0,888181175
Extensión = 1. Corto	Forma de Aparición = 4. Incrustación	2,19123506	100	1,442528736
Drogas en general = 0. Ausente	Forma de Aparición = 4. Incrustación	2,19123506	81,81818182	1,324926686
Alcohol = 0. Ausente	Fuente manifiesta = 2. Redacción	2,19123506	100	1,258145363
Cocaína = 0. Ausente	Fuente manifiesta = 2. Redacción	2,19123506	90,90909091	1,213733075
Personalización anecdótica = 0. Ausente	Fuente manifiesta = 2. Redacción	2,19123506	100	1,209638554
Personalización anecdótica = 0. Ausente	Forma de Aparición = 4. Incrustación	2,19123506	100	1,209638554
Fuerzas = 0. Ausente	Forma de Aparición = 4. Incrustación	2,19123506	90,90909091	1,182289213
'No expertos' = 0. Ausente	Forma de Aparición = 4. Incrustación	2,19123506	100	1,172897196
Política/Legislación = 0. Ausente	Fuente manifiesta = 2. Redacción	2,19123506	100	1,151376147
Alcohol = 0. Ausente	Forma de Aparición = 4. Incrustación	2,19123506	90,90909091	1,143768512
Epidemiología = 0. Ausente	Fuente manifiesta = 2. Redacción	2,19123506	100	1,143507973
Mercado/Empresa = 0. Ausente	Forma de Aparición = 4. Incrustación	2,19123506	100	1,118040089
'Psico-sanitarios' = 0. Ausente	Fuente manifiesta = 2. Redacción	2,19123506	81,81818182	0,881388997
Ética/Moralidad = 0. Ausente	Fuente manifiesta = 2. Redacción	2,19123506	81,81818182	0,873887814
Ética/Moralidad = 0. Ausente	Forma de Aparición = 4. Incrustación	2,19123506	81,81818182	0,873887814

13.3.2.- ANEXO: RUTA DROGAS

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Drogas en general = 0. Ausente		100	61,75298805	1
Cocaína = 0. Ausente		100	74,90039841	1
Tabaco = 0. Ausente		100	77,49003984	1

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Alcohol = 0. Ausente		100	79,48207171	1
Hachís = 0. Ausente		100	92,03187251	1
Marihuana = 0. Ausente		100	93,02788845	1
Heroína = 0. Ausente		100	93,22709163	1
Cannabis = 0. Ausente		100	94,02390438	1
Éxtasis/MDMA = 0. Ausente		100	98,20717131	1
Psicofármacos = 0. Ausente		100	98,80478088	1
Anabolizantes = 0. Ausente		100	99,00398406	1
Crack/Cocaína base = 0. Ausente		100	99,20318725	1
Crystal/Cristal = 0. Ausente		100	99,80079681	1
Drogas en general = 0. Ausente	Tabaco = 0. Ausente	77,49003984	53,72750643	0,870038975
Tabaco = 0. Ausente	Drogas en general = 0. Ausente	61,75298805	67,41935484	0,870038975
Tabaco = 0. Ausente	Drogas en general = 1. Presente	38,24701195	93,75	1,209832905
Alcohol = 0. Ausente	Drogas en general = 1. Presente	38,24701195	89,0625	1,120535714
Tabaco = 0. Ausente	Cocaína = 1. Presente	25,09960159	91,26984127	1,177826743
Cannabis = 0. Ausente	Cocaína = 1. Presente	25,09960159	84,12698413	0,894740382
Drogas en general = 0. Ausente	Tabaco = 1. Presente	22,50996016	89,38053097	1,447387953
Cocaína = 0. Ausente	Tabaco = 1. Presente	22,50996016	90,26548673	1,205140275
Drogas en general = 0. Ausente	Alcohol = 1. Presente	20,51792829	79,61165049	1,289195114
Drogas en general = 0. Ausente	Hachís = 1. Presente	7,96812749	80	1,295483871
Alcohol = 0. Ausente	Hachís = 1. Presente	7,96812749	97,5	1,226691729
Tabaco = 0. Ausente	Hachís = 1. Presente	7,96812749	95	1,22596401
Tabaco = 0. Ausente	Marihuana = 1. Presente	6,972111554	97,14285714	1,253617334
Drogas en general = 0. Ausente	Marihuana = 1. Presente	6,972111554	71,42857C143	1,156682028
Alcohol = 0. Ausente	Marihuana = 1. Presente	6,972111554	88,57142857	1,114357322
Cocaína = 1. Presente	Heroína = 1. Presente	6,772908367	52,94117647	2,109243697
Cannabis = 0. Ausente	Heroína = 1. Presente	6,772908367	82,35294118	0,875872383
Cocaína = 1. Presente	Cannabis = 1. Presente	5,976095618	66,66666667	2,656084656
Drogas en general = 0. Ausente	Cannabis = 1. Presente	5,976095618	80	1,295483871
Hachís = 0. Ausente	Cannabis = 1. Presente	5,976095618	80	0,869264069
Heroína = 0. Ausente	Cannabis = 1. Presente	5,976095618	80	0,858119658

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Tabaco = 0. Ausente	Cannabis = 1. Presente	5,976095618	60	0,774293059
Alcohol = 0. Ausente	Cannabis = 1. Presente	5,976095618	53,33333333	0,67101086

13.3.3.- ANEXO: RUTA *FRAME*

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Opinión no experta = 0. Ausente		100	96,41434263	1
Ética/Moralidad = 0. Ausente		100	93,62549801	1
Contexto general científico/médico = 0. Ausente		100	93,42629482	1
Opinión pública = 0. Ausente		100	92,82868526	1
Mercado/Empresa = 0. Ausente		100	89,44223108	1
Nueva investigación = 0. Ausente		100	87,84860558	1
Epidemiología = 0. Ausente		100	87,4501992	1
Política/Legislación = 0. Ausente		100	86,85258964	1
Estrategia política = 0. Ausente		100	83,46613546	1
Personalización anecdótica = 0. Ausente		100	82,66932271	1
Delito = 1. Presente		100	50,99601594	1
Delito = 1. Presente	Epidemiología = 0. Ausente	87,4501992	56,9476082	1,116707005
Delito = 1. Presente	Personalización anecdótica = 0. Ausente	82,66932271	56,86746988	1,115135542
Epidemiología = 0. Ausente	Delito = 1. Presente	50,99601594	97,65625	1,116707005
Personalización anecdótica = 0. Ausente	Delito = 1. Presente	50,99601594	92,1875	1,115135542
Nueva investigación = 0. Ausente	Delito = 0. Ausente	49,00398406	78,86178862	0,897701086
Política/Legislación = 0. Ausente	Delito = 0. Ausente	49,00398406	78,04878049	0,898635041
Epidemiología = 0. Ausente	Delito = 0. Ausente	49,00398406	76,82926829	0,878548808
Personalización anecdótica = 0. Ausente	Delito = 0. Ausente	49,00398406	72,76422764	0,880184151
Estrategia política = 0. Ausente	Personalización anecdótica = 1. Presente	17,33067729	96,55172414	1,15677722
Delito = 0. Ausente	Personalización anecdótica = 1. Presente	17,33067729	77,01149425	1,571535371
Política/Legislación = 0. Ausente	Estrategia política = 1. Presente	16,53386454	100	1,151376147

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Personalización anecdótica = 0. Ausente	Estrategia política = 1. Presente	16,53386454	96,38554217	1,165916679
Mercado/Empresa = 0. Ausente	Estrategia política = 1. Presente	16,53386454	78,31325301	0,875573564
Estrategia política = 0. Ausente	Política/Legislación = 1. Presente	13,14741036	100	1,198090692
Ética/Moralidad = 0. Ausente	Política/Legislación = 1. Presente	13,14741036	83,33333333	0,890070922
Delito = 0. Ausente	Política/Legislación = 1. Presente	13,14741036	81,81818182	1,66962306
Delito = 0. Ausente	Epidemiología = 1. Presente	12,5498008	90,47619048	1,846302749
Nueva investigación = 0. Ausente	Epidemiología = 1. Presente	12,5498008	69,84126984	0,795018537
Estrategia política = 0. Ausente	Nueva investigación = 1. Presente	12,15139442	95,08196721	1,139168199
Delito = 0. Ausente	Nueva investigación = 1. Presente	12,15139442	85,24590164	1,739570838
Opinión pública = 0. Ausente	Nueva investigación = 1. Presente	12,15139442	81,96721311	0,882994442
Contexto general científico/médico = 0. Ausente	Nueva investigación = 1. Presente	12,15139442	75,40983607	0,807158586
Epidemiología = 0. Ausente	Nueva investigación = 1. Presente	12,15139442	68,85245902	0,787333358
Personalización anecdótica = 0. Ausente	Mercado/Empresa = 1. Presente	10,55776892	98,11320755	1,186815185
Estrategia política = 0. Ausente	Mercado/Empresa = 1. Presente	10,55776892	66,03773585	0,791191966
Delito = 0. Ausente	Mercado/Empresa = 1. Presente	10,55776892	60,37735849	1,232090811
Delito = 0. Ausente	Opinión pública = 1. Presente	7,171314741	91,66666667	1,870596206
Política/Legislación = 0. Ausente	Opinión pública = 1. Presente	7,171314741	77,77777778	0,895514781
Epidemiología = 0. Ausente	Opinión pública = 1. Presente	7,171314741	75	0,857630979
Nueva investigación = 0. Ausente	Opinión pública = 1. Presente	7,171314741	69,44444444	0,790501386
Personalización anecdótica = 0. Ausente	Contexto general científico/médico = 1. Presente	6,573705179	100	1,209638554
Estrategia política = 0. Ausente	Contexto general científico/médico = 1. Presente	6,573705179	100	1,198090692
Delito = 0. Ausente	Contexto general científico/médico = 1. Presente	6,573705179	96,96969697	1,978812515
Nueva investigación = 0. Ausente	Contexto general científico/médico = 1. Presente	6,573705179	54,54545455	0,620902907
Mercado/Empresa = 0. Ausente	Ética/Moralidad = 1. Presente	6,374501992	100	1,118040089
Delito = 0. Ausente	Ética/Moralidad = 1. Presente	6,374501992	93,75	1,913109756
Personalización anecdótica = 0. Ausente	Ética/Moralidad = 1. Presente	6,374501992	65,625	0,793825301

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Política/Legislación = 0. Ausente	Ética/Moralidad = 1. Presente	6,374501992	65,625	0,755590596
Epidemiología = 0. Ausente	Opinión no experta = 1. Presente	3,585657371	100	1,143507973
Personalización anecdótica = 0. Ausente	Opinión no experta = 1. Presente	3,585657371	94,44444444	1,142436412
Delito = 0. Ausente	Opinión no experta = 1. Presente	3,585657371	72,22222222	1,473803071
Estrategia política = 0. Ausente	Opinión no experta = 1. Presente	3,585657371	66,66666667	0,798727128
Política/Legislación = 0. Ausente	Opinión no experta = 1. Presente	3,585657371	66,66666667	0,767584098

13.3.4.- ANEXO: RUTA FUENTES

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Fuerzas = 0. Ausente	Ciencia = 1. Presente	3,585657371	100	1,300518135
'No expertos' = 0. Ausente	Políticos = 1. Presente	17,1314741	97,6744186	1,145620517
Políticos = 0. Ausente	'No expertos' = 1. Presente	14,74103586	97,2972973	1,174116424
Privada = 0. Ausente	Fuerzas = 1. Presente	23,10756972	96,55172414	1,199726869
Ciencia = 0. Ausente		100	96,41434263	1
Académicos = 0. Ausente		100	96,21513944	1
Fuerzas = 0. Ausente	Privada = 1. Presente	19,52191235	95,91836735	1,247435762
Políticos = 0. Ausente	Ciencia = 1. Presente	3,585657371	94,44444444	1,139690171
Tribunal = 0. Ausente		100	93,8247012	1
'Psico-sanitarios' = 0. Ausente		100	92,82868526	1
Fuerzas = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	91,66666667	1,192141623
Fuerzas = 0. Ausente	Académicos = 1. Presente	3,784860558	89,47368421	1,163621489
Ciencia = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	86,11111111	0,893135904
'No expertos' = 0. Ausente		100	85,25896414	1
Académicos = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	83,33333333	0,866114562
Políticos = 0. Ausente		100	82,8685259	1
Privada = 0. Ausente		100	80,47808765	1
Fuerzas = 0. Ausente		100	76,89243028	1
'No expertos' = 0. Ausente	Tribunal = 1. Presente	6,175298805	74,19354839	0,870214049
Ciencia = 0. Ausente	Académicos = 1. Presente	3,784860558	73,68421053	0,764245324
'Psico-sanitarios' = 0. Ausente	Ciencia = 1. Presente	3,585657371	72,22222222	0,778016214
Académicos = 0. Ausente	Ciencia = 1. Presente	3,585657371	72,22222222	0,75063262
'Psico-sanitarios' = 0. Ausente	Académicos = 1. Presente	3,784860558	68,42105263	0,737067992

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Privada = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	66,66666667	0,828382838

13.4.- ANEXO: ANÁLISIS INDIVIDUALES

13.4.1.- ANEXO: CATEGORÍA TEMA PRINCIPAL

13.4.1.1.- Tráfico de Drogas

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Cocaína = 0. Ausente	Políticos = 1. Presente	22	13,66459627	81,81818182	1,531712474
Estrategia política = 0. Ausente	Sección = Nacional/Política/España	27	16,77018634	62,96296296	0,77977208

13.4.1.2.- Consecuencias

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Ética/Moralidad = 0. Ausente	Marihuana = 1. Presente	11	12,08791209	72,72727273	0,760710554
Tabaco = 1. Presente	Contexto general científico/médico = 1. Presente	10	10,98901099	70	4,246666667
Marihuana = 0. Ausente	Periódico = ABC	16	17,58241758	68,75	0,78203125
'Psico-sanitarios' = 0. Ausente	Tabaco = 1. Presente	15	16,48351648	66,66666667	0,71372549
Periódico = La Razón	Ubicación en el periódico = Página impar apertura de sección	14	15,38461538	64,28571429	2,785714286
Periódico = El País	Contexto general científico/médico = 1. Presente	10	10,98901099	60	2,1
'Psico-sanitarios' = 0. Ausente	Contexto general científico/médico = 1. Presente	10	10,98901099	60	0,642352941

13.4.1.3.- Consumo

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Cocaína = 1. Presente	Alcohol = 1. Presente	16	29,09090909	68,75	2,224264706
Alcohol = 1. Presente	Periódico = La Razón	12	21,81818182	66,66666667	2,291666667
Alcohol = 1. Presente	Cocaína = 1. Presente	17	30,90909091	64,70588235	2,224264706

13.4.1.4.- Estudios y Prevención

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Privada = 1. Presente	Ética/Moralidad = 1. Presente	15	14,42307692	73,33333333	2,061261261
Afiliación institucional = 1. Presente	Ética/Moralidad = 1. Presente	15	14,42307692	73,33333333	2,061261261
Estrategia política = 1. Presente	Delito = 1. Presente	16	15,38461538	62,5	2,096774194
Políticos = 1. Presente	Delito = 1. Presente	16	15,38461538	62,5	2,407407407

13.4.2.- ANEXO: ANÁLISIS INDIVIDUAL DE CANTIDAD DE FUENTES

13.4.2.1.- Ninguna

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Tabaco = 1. Presente	Valoración de Unidad de análisis Discreta = Alto	31	22,79411765	61,29032258	2,38156682
Periódico = El País	Categoría Tema Principal = 4. Estudios y Prevención	14	10,29411765	64,28571429	2,428571429
Personalización anecdótica = 1. Presente	Valoración de Unidad de análisis Discreta = Alto	31	22,79411765	64,51612903	2,506912442
Tabaco = 1. Presente	Categoría Tema Principal = 0. No Relacionado con Drogas	24	17,64705882	66,66666667	2,59047619
Tabaco = 1. Presente	Categoría Tema Principal = 4. Estudios y Prevención	14	10,29411765	71,42857143	2,775510204
Valoración de Unidad de análisis Discreta = Alto	Categoría Tema Principal = 0. No Relacionado con Drogas	24	17,64705882	66,66666667	2,924731183

13.4.2.2.- Muchas

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Estrategia política = 1. Presente	Delito = 1. Presente	20	33,89830508	60	2,082352941
Tabaco = 1. Presente	'Psico-sanitarios' = 1. Presente	16	27,11864407	62,5	2,169117647

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Estrategia política = 1. Presente	Categoría Tema Principal = 1. Trafico de Drogas	13	22,03389831	69,23076923	2,402714932
Personalización anecdótica = 1. Presente	Categoría Tema Principal = 3. Consumo	12	20,33898305	75	2,458333333

13.4.3.- ANEXO: ANÁLISIS INDIVIDUAL DE FUENTES

13.4.3.1.- Privada

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Tabaco = 1. Presente	'Psico-sanitarios' = 1. Presente	12	12,24489796	66,66666667	2,107526882
Tabaco = 1. Presente	Contexto general científico/médico = 1. Presente	11	11,2244898	81,81818182	2,586510264
Cocaína = 1. Presente	Cannabis = 1. Presente	15	15,30612245	73,33333333	3,593333333

13.4.3.2.- Fuerzas y Cuerpos de Seguridad del Estado

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Cantidad de Fuentes = Muchas Fuentes	Políticos = 1. Presente	13	11,20689655	61,53846154	4,461538462
Categoría Tema Principal = 2. Consecuencias	Alcohol = 1. Presente	16	13,79310345	62,5	4,53125

13.4.3.3.- Políticos

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Drogas en general = 0. Ausente	Tabaco = 1. Presente	21	24,41860465	85,71428571	2,047619048
Estrategia política = 0. Ausente	Política/Legislación = 1. Presente	12	13,95348837	100	2,15
Cantidad de Fuentes = Muchas Fuentes	Fuerzas = 1. Presente	13	15,11627907	61,53846154	2,646153846
Tabaco = 1. Presente	Política/Legislación = 1. Presente	12	13,95348837	66,66666667	2,73015873
Alcohol = 1. Presente	Categoría Tema Principal = 2. Consecuencias	11	12,79069767	63,63636364	2,736363636
Cocaína = 1. Presente	Cannabis = 1. Presente	11	12,79069767	90,90909091	4,343434343

13.4.4.- ANEXO: ANÁLISIS INDIVIDUAL DE ESDOMINGO

13.4.4.1.- Presente

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Fuerzas = 1. Presente	Hachís = 1. Presente	11	12,35955056	63,63636364	2,980861244
Tabaco = 1. Presente	Categoría Tema Principal = 4. Estudios y Prevención	17	19,1011236	76,47058824	2,768888889
Categoría Tema Principal = 1. Trafico de Drogas	Hachís = 1. Presente	11	12,35955056	81,81818182	2,767942584
Tabaco = 1. Presente	Mercado/Empresa = 1. Presente	9	10,11235955	77,77777778	2,722352941
Valoración de Unidad de análisis Discreta = Alto	Personalización anecdótica = 1. Presente	20	22	70	2,397306397
Periódico = El País	Estrategia política = 1. Presente	13	14,60674157	61,53846154	2,282051282

13.4.5.- ANEXO: ANÁLISIS INDIVIDUAL DE DROGAS

13.4.5.1.- Alcohol

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Mercado/Empresa = 1. Presente	Política/Legislación = 1. Presente	17	16,50485437	23,52941176	4,039215686
Ética/Moralidad = 1. Presente	Personalización anecdótica = 1. Presente	19	18,44660194	21,05263158	2,409356725

13.4.5.2.- Tabaco

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Epidemiología = 0. Ausente	Periódico = La Razón	26	23,00884956	61,53846154	0,799292661
Delito = 0. Ausente	Drogas en general = 1. Presente	12	10,61946903	75	0,799528302
Delito = 0. Ausente	Cannabis = 1. Presente	12	10,61946903	75	0,799528302
Alcohol = 1. Presente	Cannabis = 1. Presente	12	10,61946903	75	3,53125
Epidemiología = 1. Presente	Cannabis = 1. Presente	12	10,61946903	83,33333333	3,621794872
Cocaína = 1. Presente	Cannabis = 1. Presente	12	10,61946903	83,33333333	8,560606061

13.4.6.- ANEXO: ANÁLISIS INDIVIDUAL DE *FRAME*

13.4.6.1.- Nueva Investigación

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Tabaco = 1. Presente	Contexto general científico/médico = 1. Presente	15	24,59016393	66,66666667	2,259259259

13.4.6.2.- Contexto Científico/Médico

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Periódico = El Mundo	Ubicación en el periódico = Página impar normal	11	33,33333333	63,63636364	2,333333333

13.4.6.3.- Ética/Moralidad

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Ubicación en el periódico = Página impar normal	Periódico = ABC	11	34,375	63,63636364	2,036363636

13.4.6.4.- Estrategia Política

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Categoría Tema Principal = 4. Estudios y Prevención	Tabaco = 1. Presente	16	19,27710843	75	2,008064516
Ubicación en el periódico = Página impar normal	Tabaco = 1. Presente	16	19,27710843	62,5	2,075
Delito = 0. Ausente	Tabaco = 1. Presente	16	19,27710843	93,75	2,161458333
Drogas en general = 0. Ausente	Tabaco = 1. Presente	16	19,27710843	100	2,862068966

13.4.6.5.- Política/Legislación

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Extensión = 2. Normal	Ética/Moralidad = 1. Presente	11	16,66666667	63,63636364	2,210526316

13.4.6.6.- Delito

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Periódico = El País	Categoría Tema Principal = 4. Estudios y Prevención	16	6,25	62,5	2,222222222
Periódico = El Mundo	Ubicación en el periódico = Portada	16	6,25	75	2,37037037
Estrategia política = 1. Presente	Categoría Tema Principal = 4. Estudios y Prevención	16	6,25	62,5	3,404255319

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Políticos = 1. Presente	Categoría Tema Principal = 4. Estudios y Prevención	16	6,25	62,5	3,80952381
Políticos = 1. Presente	Cantidad de Fuentes = Muchas Fuentes	20	7,8125	65	3,961904762
Categoría Tema Principal = 2. Consecuencias	Alcohol = 1. Presente	36	14,0625	75	4,085106383

13.4.7.- ANEXO: ANÁLISIS INDIVIDUAL DE FUENTE MANIFIESTA

13.4.7.1.- Nombre Propio

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Categoría Tema Principal = 1. Trafico de Drogas	Fuerzas = 1. Presente	72	23,68421053	69,44444444	2,454780362
Epidemiología = 1. Presente	Categoría Tema Principal = 3. Consumo	35	11,51315789	65,71428571	4,250455927

13.4.8.- ANEXO: ANÁLISIS INDIVIDUAL DE GÉNERO PERIODÍSTICO

13.4.8.1.- Artículo

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Ubicación en el periódico = Página par normal	EsDomingo = Verdadero	11	30,55555556	63,63636364	1,909090909
Periódico = El País	Ubicación en el periódico = Página par normal	12	33,33333333	66,66666667	1,6

13.4.8.2.- Noticia

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Epidemiología = 1. Presente	Cannabis = 1. Presente	21	5,785123967	61,9047619	4,781155015
Periódico = La Razón	Ubicación en el periódico = Página impar apertura de sección	21	5,785123967	61,9047619	2,956766917

13.4.9.- ANEXO: ANÁLISIS INDIVIDUAL DE ILUSTRACIÓN

13.4.9.1.- Presente

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Periódico = El Mundo	Ubicación en el periódico = Portada	29	10,35714286	62,06896552	1,791681479
Drogas en general = 1. Presente	Ubicación en el periódico = Portada	29	10,35714286	65,51724138	1,581450654

13.4.10.- Análisis Individual por Periódicos

13.4.10.1.- El Mundo

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Cocaína = 1. Presente	Tribunal = 1. Presente	11	6,748466258	63,63636364	2,160984848
Valoración de Unidad de análisis Discreta = Alto	'Psico-sanitarios' = 1. Presente	12	7,36196319	75	1,746428571
Valoración de Unidad de análisis Discreta = Alto	Personalización anecdótica = 1. Presente	35	21,47239264	71,42857143	1,663265306
Valoración de Unidad de análisis Discreta = Alto	Políticos = 1. Presente	28	17,17791411	67,85714286	1,580102041
Valoración de Unidad de análisis Discreta = Alto	Heroína = 1. Presente	12	7,36196319	66,66666667	1,552380952
Valoración de Unidad de análisis Discreta = Alto	'No expertos' = 1. Presente	33	20,24539877	63,63636364	1,481818182
Tribunal = 0. Ausente	Alcohol = 1. Presente	26	15,95092025	100	1,072368421
Heroína = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	58	35,58282209	94,82758621	1,023635533
Marihuana = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	58	35,58282209	96,55172414	1,015350389
Hachís = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	58	35,58282209	91,37931034	0,95479664

13.4.10.2.- ABC

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Ética/Moralidad = 1. Presente	Política/Legislación = 1. Presente	10	12,65822785	60	4,309090909
Delito = 0. Ausente	Ética/Moralidad = 1. Presente	11	13,92405063	100	2,393939394
Delito = 1. Presente	Hachís = 1. Presente	10	12,65822785	90	1,545652174
Marihuana = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	30	37,97468354	90	1,06119403
Cannabis = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	30	37,97468354	93,33333333	1,010045662

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Hachís = 0. Ausente	Categoría Tema Principal = 1. Trafico de Drogas	30	37,97468354	70	0,801449275

13.4.10.3.- El País

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Categoría Tema Principal = 1. Trafico de Drogas	Hachís = 1. Presente	16	10,59602649	75	2,903846154
Delito = 1. Presente	Hachís = 1. Presente	16	10,59602649	93,75	1,966145833

13.4.10.4.- La Razón

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Alcohol = 1. Presente	Categoría Tema Principal = 2. Consecuencias	21	19,26605505	71,42857143	2,162698413
Valoración de Unidad de análisis Discreta = Alto	Epidemiología = 1. Presente	19	17,43119266	68,42105263	1,818998716
Valoración de Unidad de análisis Discreta = Alto	Privada = 1. Presente	15	13,76146789	66,66666667	1,772357724

13.4.11.- ANEXO: ANÁLISIS INDIVIDUAL DE VALORACIÓN DE UNIDAD DE ANÁLISIS

13.4.11.1.- Alto

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Ética/Moralidad = 0. Ausente		180	100	90,55555556	1

13.5.- ANEXO: ANÁLISIS SUGERIDOS

13.5.1.- VALORACIÓN FORMAL VS. TEMA PRINCIPAL, *FRAME*, FUENTE Y DROGA

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Valoración de Unidad de análisis Discreta = Bajo	Fuente manifiesta = 7. Otros	3,784860558	78,94736842	2,659837513
Valoración de Unidad de análisis Discreta = Bajo	Fuente manifiesta = 3. Agencia	4,780876494	62,5	2,105704698
Valoración de Unidad de análisis Discreta = Alto	Cantidad de Fuentes = Muchas Fuentes	11,75298805	71,18644068	1,985310734
Valoración de Unidad de análisis Discreta = Alto	Académicos = 1. Presente	3,784860558	68,42105263	1,908187135

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Valoración de Unidad de análisis Discreta = Alto	Fuente manifiesta = 5. Enviado especial	2,390438247	66,66666667	1,859259259
Valoración de Unidad de análisis Discreta = Alto	'Psico-sanitarios' = 1. Presente	7,171314741	66,66666667	1,859259259
Valoración de Unidad de análisis Discreta = Bajo	Fuente manifiesta = 6. No figura	22,70916335	54,38596491	1,832332509
Valoración de Unidad de análisis Discreta = Alto	Otra = 1. Presente	6,374501992	59,375	1,655902778
Valoración de Unidad de análisis Discreta = Alto	Heroína = 1. Presente	6,772908367	58,82352941	1,640522876
Valoración de Unidad de análisis Discreta = Alto	Personalización anecdótica = 1. Presente	17,33067729	58,62068966	1,6348659
Valoración de Unidad de análisis Discreta = Alto	Ciencia = 1. Presente	3,585657371	55,55555556	1,549382716
Valoración de Unidad de análisis Discreta = Alto	Categoría Tema Principal = 0. No Relacionado con Drogas	12,5498008	55,55555556	1,549382716
Valoración de Unidad de análisis Discreta = Alto	'No expertos' = 1. Presente	14,74103586	54,05405405	1,507507508
Valoración de Unidad de análisis Discreta = Alto	Ética/Moralidad = 1. Presente	6,374501992	53,125	1,481597222
Valoración de Unidad de análisis Discreta = Alto	Tribunal = 1. Presente	6,175298805	51,61290323	1,439426523
Valoración de Unidad de análisis Discreta = Alto	Fuente manifiesta = 1. Nombre propio	60,55776892	50,65789474	1,412792398
Valoración de Unidad de análisis Discreta = Alto	Fuente manifiesta = 4. Corresponsal	3,585657371	50	1,394444444
Valoración de Unidad de análisis Discreta = Alto	Políticos = 1. Presente	17,1314741	50	1,394444444

13.5.2.- TEMA PRINCIPAL VS. *FRAME*, FUENTE Y DROGA

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Categoría Tema Principal = 3. Consumo	Epidemiología = 1. Presente	12,5498008	52,38095238	4,780952381
Categoría Tema Principal = 1. Trafico de Drogas	Hachís = 1. Presente	7,96812749	82,5	2,572360248
Categoría Tema Principal = 1. Trafico de Drogas	Fuerzas = 1. Presente	23,10756972	68,96551724	2,150353395
Categoría Tema Principal = 1. Trafico de Drogas	Fuente manifiesta = 5. Enviado especial	2,390438247	66,66666667	2,078674948
Categoría Tema Principal = 1. Trafico de Drogas	Cocaína = 1. Presente	25,09960159	59,52380952	1,855959775
Categoría Tema Principal = 1. Trafico de Drogas	Delito = 1. Presente	50,99601594	59,375	1,851319876
Categoría Tema Principal = 1. Trafico de Drogas	Tribunal = 1. Presente	6,175298805	54,83870968	1,70987778
Categoría Tema Principal = 1. Trafico de Drogas	Fuente manifiesta = 3. Agencia	4,780876494	50	1,559006211

13.5.3.- FRAME VS. FUENTE Y DROGA

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Nueva investigación = 1. Presente	Ciencia = 1. Presente	3,585657371	66,66666667	5,486338798
Nueva investigación = 1. Presente	Académicos = 1. Presente	3,784860558	52,63157895	4,331320104
Epidemiología = 1. Presente	Cannabis = 1. Presente	5,976095618	53,33333333	4,24973545
Epidemiología = 1. Presente	Académicos = 1. Presente	3,784860558	52,63157895	4,193817878
Política/Legislación = 1. Presente	Fuente manifiesta = 7. Otros	3,784860558	52,63157895	4,003189793
Estrategia política = 1. Presente	Políticos = 1. Presente	17,1314741	53,48837209	3,235079854
Estrategia política = 1. Presente	Fuente manifiesta = 5. Enviado especial	2,390438247	50	3,024096386
Delito = 0. Ausente	Ciencia = 1. Presente	3,585657371	94,44444444	1,927280939
Delito = 0. Ausente	Tabaco = 1. Presente	22,50996016	93,80530973	1,914238434
Delito = 1. Presente	Hachís = 1. Presente	7,96812749	92,5	1,813867188
Delito = 1. Presente	Fuente manifiesta = 5. Enviado especial	2,390438247	91,66666667	1,797526042
Delito = 1. Presente	Fuerzas = 1. Presente	23,10756972	91,37931034	1,791891164
Delito = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	86,11111111	1,757226739
Delito = 1. Presente	Tribunal = 1. Presente	6,175298805	83,87096774	1,644657258
Delito = 0. Ausente	Académicos = 1. Presente	3,784860558	78,94736842	1,611039795
Delito = 0. Ausente	Privada = 1. Presente	19,52191235	78,57142857	1,603368177
Delito = 1. Presente	Cocaína = 1. Presente	25,09960159	72,22222222	1,416232639
Delito = 1. Presente	Fuente manifiesta = 3. Agencia	4,780876494	70,83333333	1,388997396
Delito = 0. Ausente	Cantidad de Fuentes = Muchas Fuentes	11,75298805	66,10169492	1,348904506
Delito = 0. Ausente	Alcohol = 1. Presente	20,51792829	65,04854369	1,327413371
Delito = 0. Ausente	Cannabis = 1. Presente	5,976095618	63,33333333	1,292411924
Delito = 1. Presente	Drogas en general = 1. Presente	38,24701195	65,10416667	1,276652018
Delito = 0. Ausente	Otra = 1. Presente	6,374501992	62,5	1,275406504
Delito = 1. Presente	Tabaco = 0. Ausente	77,49003984	64,01028278	1,255201639
Delito = 0. Ausente	Fuerzas = 0. Ausente	76,89243028	61,13989637	1,247651544
Personalización anecdótica = 0. Ausente	Fuente manifiesta = 2. Redacción	2,19123506	100	1,209638554
Personalización anecdótica = 0. Ausente	Fuente manifiesta = 7. Otros	3,784860558	100	1,209638554
Delito = 0. Ausente	Fuente manifiesta = 7. Otros	3,784860558	57,89473684	1,181429183
Delito = 0. Ausente	Drogas en general = 0. Ausente	61,75298805	57,74193548	1,178311041
Personalización anecdótica = 0. Ausente	Fuente manifiesta = 3. Agencia	4,780876494	95,83333333	1,159236948
Personalización anecdótica = 0. Ausente	Políticos = 1. Presente	17,1314741	95,34883721	1,153376296
Estrategia política = 0. Ausente	Alcohol = 1. Presente	20,51792829	96,11650485	1,151562898
Política/Legislación = 0. Ausente	Fuente manifiesta = 2. Redacción	2,19123506	100	1,151376147
Política/Legislación = 0. Ausente	Fuente manifiesta = 5. Enviado especial	2,390438247	100	1,151376147

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Política/Legislación = 0. Ausente	Fuente manifiesta = 4. Corresponsal	3,585657371	100	1,151376147
Política/Legislación = 0. Ausente	Ciencia = 1. Presente	3,585657371	100	1,151376147
Estrategia política = 0. Ausente	Fuente manifiesta = 3. Agencia	4,780876494	95,83333333	1,148170247
Personalización anecdótica = 0. Ausente	Académicos = 1. Presente	3,784860558	94,73684211	1,145973367
Delito = 0. Ausente	Cocaína = 0. Ausente	74,90039841	56,11702128	1,145152223
Epidemiología = 0. Ausente	Fuente manifiesta = 2. Redacción	2,19123506	100	1,143507973
Epidemiología = 0. Ausente	Fuente manifiesta = 5. Enviado especial	2,390438247	100	1,143507973
Epidemiología = 0. Ausente	Fuente manifiesta = 4. Corresponsal	3,585657371	100	1,143507973
Personalización anecdótica = 0. Ausente	Ciencia = 1. Presente	3,585657371	94,44444444	1,142436412
Delito = 1. Presente	Privada = 0. Ausente	80,47808765	58,16831683	1,140644338
Nueva investigación = 0. Ausente	Fuente manifiesta = 5. Enviado especial	2,390438247	100	1,138321995
Nueva investigación = 0. Ausente	Fuente manifiesta = 7. Otros	3,784860558	100	1,138321995
Nueva investigación = 0. Ausente	Tribunal = 1. Presente	6,175298805	100	1,138321995
Estrategia política = 0. Ausente	Ciencia = 1. Presente	3,585657371	94,44444444	1,131530098
Estrategia política = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	94,44444444	1,131530098
Delito = 0. Ausente	'No expertos' = 1. Presente	14,74103586	55,40540541	1,130630631
Epidemiología = 0. Ausente	Fuerzas = 1. Presente	23,10756972	98,27586207	1,123792318
Personalización anecdótica = 0. Ausente	Hachís = 1. Presente	7,96812749	92,5	1,118915663
Política/Legislación = 0. Ausente	Heroína = 1. Presente	6,772908367	97,05882353	1,117512142
Personalización anecdótica = 0. Ausente	Fuerzas = 1. Presente	23,10756972	92,24137931	1,115787287
Política/Legislación = 0. Ausente	Otra = 1. Presente	6,374501992	96,875	1,115395642

13.5.4.- FUENTE VS DROGA

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Tabaco = 1. Presente	'Psico-sanitarios' = 1. Presente	7,171314741	55,55555556	2,468043265
Drogas en general = 1. Presente	Fuente manifiesta = 2. Redacción	2,19123506	72,72727273	1,901515152
Drogas en general = 1. Presente	Fuente manifiesta = 5. Enviado especial	2,390438247	58,33333333	1,525173611
Drogas en general = 1. Presente	Políticos = 1. Presente	17,1314741	58,13953488	1,520106589
Drogas en general = 1. Presente	Fuente manifiesta = 4. Corresponsal	3,585657371	55,55555556	1,452546296

Consecuente	Antecedente	% de soporte	% de confianza	Elevación
Drogas en general = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	80,55555556	1,304480287
Drogas en general = 0. Ausente	Fuente manifiesta = 3. Agencia	4,780876494	79,16666667	1,281989247
Alcohol = 0. Ausente	Fuente manifiesta = 2. Redacción	2,19123506	100	1,258145363
Tabaco = 0. Ausente	Fuerzas = 1. Presente	23,10756972	97,4137931	1,257113731
Alcohol = 0. Ausente	Tribunal = 1. Presente	6,175298805	96,77419355	1,217560029
Cocaína = 0. Ausente	Fuente manifiesta = 2. Redacción	2,19123506	90,90909091	1,213733075
Tabaco = 0. Ausente	Tribunal = 1. Presente	6,175298805	93,5483871	1,207231114
Cocaína = 0. Ausente	Fuente manifiesta = 7. Otros	3,784860558	89,47368421	1,194568869
Drogas en general = 0. Ausente	Académicos = 1. Presente	3,784860558	73,68421053	1,193208829
Cocaína = 0. Ausente	'Psico-sanitarios' = 1. Presente	7,171314741	88,88888889	1,186761229
Tabaco = 0. Ausente	Fuente manifiesta = 5. Enviado especial	2,390438247	91,66666667	1,182947729
Drogas en general = 0. Ausente	Ciencia = 1. Presente	3,585657371	72,22222222	1,16953405
Tabaco = 0. Ausente	Fuente manifiesta = 4. Corresponsal	3,585657371	88,88888889	1,147100828
Drogas en general = 0. Ausente	Fuente manifiesta = 6. No figura	22,70916335	70,1754386	1,13638936
Tabaco = 0. Ausente	Fuente manifiesta = 3. Agencia	4,780876494	87,5	1,129177378
Alcohol = 0. Ausente	Fuente manifiesta = 7. Otros	3,784860558	89,47368421	1,125709009
Cocaína = 0. Ausente	Ciencia = 1. Presente	3,585657371	83,33333333	1,112588652
Cocaína = 0. Ausente	Fuente manifiesta = 4. Corresponsal	3,585657371	83,33333333	1,112588652

13.6.- ANEXO: ANÁLISIS AGRUPADOS

13.6.1.- ANÁLISIS AGRUPADO DE DROGAS

13.6.1.1.- Drogas como consecuente

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Drogas Legales = 1. Presente	Contexto general científico/médico = 1. Presente	33	6,573705179	78,78787879	2,059974747
Drogas Legales = 1. Presente	Opinión pública = 1. Presente	36	7,171314741	72,22222222	1,888310185
Drogas Legales = 1. Presente	'Psico-sanitarios' = 1. Presente	36	7,171314741	72,22222222	1,888310185
Drogas Legales = 1. Presente	Epidemiología = 1. Presente	63	12,5498008	66,66666667	1,743055556
Drogas Legales = 1. Presente	Nueva investigación = 1. Presente	61	12,15139442	63,93442623	1,671618852
Drogas Legales = 1. Presente	Política/Legislación = 1. Presente	66	13,14741036	63,63636364	1,663825758

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Drogas Legales = 1. Presente	Categoría Tema Principal = 4. Estudios y Prevención	104	20,71713147	63,46153846	1,659254808
Drogas Duras = 0. Ausente	Política/Legislación = 1. Presente	66	13,14741036	90,90909091	1,267676768
Drogas Duras = 0. Ausente	Categoría Tema Principal = 0. No Relacionado con Drogas	63	12,5498008	90,47619048	1,261640212
Drogas Legales = 0. Ausente	Estrategia política = 1. Presente	83	16,53386454	77,10843373	1,248659153
Drogas Duras = 0. Ausente	Ética/Moralidad = 1. Presente	32	6,374501992	87,5	1,220138889
Drogas Duras = 0. Ausente	Categoría Tema Principal = 4. Estudios y Prevención	104	20,71713147	85,57692308	1,19332265
Drogas Legales = 0. Ausente	Ubicación en el periódico = Portada	29	5,77689243	72,4137931	1,172636263
Drogas Duras = 0. Ausente	'Psico-sanitarios' = 1. Presente	36	7,171314741	83,33333333	1,162037037
Drogas Legales = 0. Ausente	Periódico = ABC	79	15,73705179	70,88607595	1,147897101
Drogas Duras = 0. Ausente	Contexto general científico/médico = 1. Presente	33	6,573705179	81,81818182	1,140909091
Drogas Duras = 0. Ausente	Periódico = El País	151	30,07968127	79,47019868	1,10816777
Drogas Legales = 0. Ausente	Periódico = El Mundo	163	32,47011952	66,87116564	1,082881457
Drogas Blandas = 0. Ausente	Categoría Tema Principal = 4. Estudios y Prevención	104	20,71713147	87,5	1,07396088
Drogas Blandas = 0. Ausente	Periódico = El Mundo	163	32,47011952	87,11656442	1,069254654

13.6.1.2.- Drogas Duras ausente

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Periódico = El Mundo	Ubicación en el periódico = Portada	18	5	77,77777778	2,568807339
Periódico = El País	Extensión = '3. Largo''	22	6,111111111	63,63636364	1,909090909

13.6.1.3.- Drogas Legales presente

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Periódico = El País	Mercado/Empresa = 1. Presente	26	13,54166667	69,23076923	2,179066835
Valoración de Unidad de análisis Discreta = Alto	Ética/Moralidad = 1. Presente	12	6,25	75	1,714285714
Valoración de Unidad de análisis Discreta = Alto	'Psico-sanitarios' = 1. Presente	26	13,54166667	65,38461538	1,494505495

13.6.1.4.- Drogas Legales Ausente

Consecuente	Antecedente	Ocurrencias	% de soporte	% de confianza	Elevación
Epidemiología = 1. Presente	'Psico-sanitarios' = 1. Presente	10	3,225806452	70	10,33333333
Periódico = El País	Extensión = '3. Largo''	16	5,161290323	62,5	2,152777778
Periódico = El Mundo	Ubicación en el periódico = Portada	21	6,774193548	61,9047619	1,760594146

13.7.- Análisis de Campos Influyentes

13.7.1.- Drogas

13.7.1.1.- Cocaina

Rango	Campo	Tipo	Importancia	Valor
1	Crack/Cocaína base	flag	0	0.999999950227037
2	Epidemiología	flag	0	0.9996937680316695
3	Éxtasis/MDMA	flag	0	0.9986248816565281
4	Cannabis	flag	0	0.9970543919121049
5	Política/Legislación	flag	1	0.9318463897769075

13.7.1.2.- Heroína

Rango	Campo	Tipo	Importancia	Valor
1	Crack/Cocaína base	flag	0	0.999999950227037
2	Cocaína	flag	0	0.9998945997326015
3	Epidemiología	flag	0	0.9996937680316695
4	Éxtasis/MDMA	flag	0	0.9986248816565281
5	Cannabis	flag	0	0.9970543919121049
6	Política/Legislación	flag	1	0.9318463897769075

13.7.1.3.- Crack

Rango	Campo	Tipo	Importancia	Valor
1	Heroína	flag	0	0.999999950227037
2	Cannabis	flag	0	0.9998079908888758
3	Nº de imágenes de la unidad de análisis	set	0	0.9874721678986765
4	Cocaína	flag	0	0.9791659025393573
5	Epidemiología	flag	0	0.9767913513920718

13.7.1.4.- Cristal

Rango	Campo	Tipo	Importancia	Valor
1	Contexto general científico/médico	flag	0	0.9998391282322531

Rango	Campo	Tipo	Importancia	Valor
2	Nueva investigación	flag	0	0.9928859235135
3	Privada	flag	0	0.9578875102349219
4	Afiliación institucional	flag	0	0.9565763403765961

13.7.1.5.- Éxtasis

Rango	Campo	Tipo	Importancia	Valor
1	Epidemiología	flag	0	0.9999992393254697
2	Opinión pública	flag	0	0.9999877593647015
3	Alcohol	flag	0	0.9999823234667342
4	Cocaína	flag	0	0.9997648768281203
5	Cannabis	flag	0	0.9995237442350323
6	Heroína	flag	0	0.9986248816565281
7	Nueva investigación	flag	0	0.9972294776473174
8	Académicos	flag	0	0.9965536414190317
9	Forma de Aparición	orderedSet	0	0.9901205889273322
10	Valoración de Unidad de análisis Discreta	set	0	0.9766772183883381
11	Categoría Tema Principal	set	0	0.9599122616599837
12	Otra	flag	0	0.9504461814744894
13	Drogas en general	flag	1	0.9090311824883914

13.7.1.6.- Drogas en General

Rango	Campo	Tipo	Importancia	Valor
1	Tabaco	flag	0	0.9999999999933437
2	Estrategia política	flag	0	0.999999962407937
3	Delito	flag	0	0.9999993519397775
4	Categoría Tema Principal	set	0	0.9999752763332395
5	Alcohol	flag	0	0.9999712523999245
6	Políticos	flag	0	0.9999695149853642
7	Valoración de Unidad de análisis Discreta	set	0	0.9992003958951069
8	Política/Legislación	flag	0	0.9879957857692367
9	Hachís	flag	0	0.9866869404750581
10	Contexto general científico/médico	flag	0	0.9858641890257429
11	Opinión pública	flag	0	0.9840189558832787
12	'Psico-sanitarios'	flag	0	0.9840189558832787
13	Otra	flag	0	0.9809936847717992
14	Cannabis	flag	0	0.9660637412913259
15	Epidemiología	flag	0	0.9508206622867879
16	Anabolizantes	flag	1	0.9230385420597661
17	'No expertos'	flag	1	0.9172456667702767
18	Éxtasis/MDMA	flag	1	0.9090311824883914

13.7.1.7.- Otra

Rango	Campo	Tipo	Importancia	Valor
1	Anabolizantes	flag	0	0.9999999999873666
2	Categoría Tema Principal	set	0	0.9999999838846187
3	Epidemiología	flag	0	0.9999999632928365
4	Ciencia	flag	0	0.9998466121366523
5	Académicos	flag	0	0.9997136422414261
6	'Psico-sanitarios'	flag	0	0.9991367736211401
7	Cannabis	flag	0	0.9983698520967129
8	Psicofármacos	flag	0	0.9934592614996197
9	Opinión pública	flag	0	0.9912992640156206
10	Cantidad de Fuentes	orderedSet	0	0.9898204546018777
11	Cocaína	flag	0	0.9880874844991985
12	Drogas en general	flag	0	0.9809936847717992
13	Afiliación institucional	flag	0	0.9686892099750036
14	Valoración de Unidad de análisis Discreta	set	0	0.9674712045079453
15	Éxtasis/MDMA	flag	0	0.9504461814744894
16	Cargo	flag	1	0.9445719868614185
17	Nueva investigación	flag	1	0.9181237517309231
18	Política/Legislación	flag	1	0.9170744858954581
19	Privada	flag	1	0.9163405529398762

13.7.1.8.- Psicofármacos

Rango	Campo	Tipo	Importancia	Valor
1	Ciencia	flag	0	0.9999194023277662
2	Epidemiología	flag	0	0.9946595769072154
3	Otra	flag	0	0.9934592614996197
4	Delito	flag	0	0.9880578225881932
5	'Psico-sanitarios'	flag	0	0.9875356867840172
6	Nº de imágenes de la unidad de análisis	set	0	0.950408838140222
7	Académicos	flag	1	0.9037829739733458

13.7.1.9.- Marihuana

Rango	Campo	Tipo	Importancia	Valor
1	Personalización anecdótica	flag	0	0.9999647431841404
2	Ética/Moralidad	flag	0	0.9993762050306841
3	'No expertos'	flag	0	0.9961138477506994
4	Tabaco	flag	0	0.99610203569509
5	EsDomingo	flag	1	0.9463492611044984
6	Epidemiología	flag	1	0.9272850806032382
7	'Psico-sanitarios'	flag	1	0.9117772731759578

13.7.1.10.- Hachís

Rango	Campo	Tipo	Importancia	Valor
1	Categoría Tema Principal	set	0	0.999999999562722
2	Delito	flag	0	0.9999999558820882
3	Fuerzas	flag	0	0.9999999251160749
4	Alcohol	flag	0	0.9967334118968498
5	Forma de Aparición	orderedSet	0	0.9960865494692241
6	Tabaco	flag	0	0.9942899877421457
7	Cannabis	flag	0	0.9879177365335937
8	Drogas en general	flag	0	0.9866869404750581
9	Tribunal	flag	0	0.9843501793525344
10	Nº de imágenes de la unidad de análisis	set	0	0.9718528801589782
11	Políticos	flag	0	0.9662183977353093
12	MesNumerico	range	1	0.9492463936335519
13	Nueva investigación	flag	1	0.9485214040244446
14	Opinión pública	flag	1	0.9331074455097725
15	Contexto general científico/médico	flag	1	0.9196683246945058
16	Personalización anecdótica	flag	1	0.913147340843523
17	EsDomingo	flag	1	0.9083329735005313

13.7.1.11.- Cannabis

Rango	Campo	Tipo	Importancia	Valor
1	Epidemiología	flag	0	0.9999999999964464
2	Cocaína	flag	0	0.9999999387834706
3	Afiliación institucional	flag	0	0.9999981731377694
4	Privada	flag	0	0.9999859701621194
5	Cargo	flag	0	0.999980283323536
6	Categoría Tema Principal	set	0	0.9999637756093478
7	Crack/Cocaína base	flag	0	0.9998079908888758
8	Alcohol	flag	0	0.9997453675189013
9	Éxtasis/MDMA	flag	0	0.9995237442350323
10	Otra	flag	0	0.9983698520967129
11	Heroína	flag	0	0.9970543919121049
12	Políticos	flag	0	0.9965955905281462
13	Cantidad de Fuentes	orderedSet	0	0.9913627863799233
14	Hachís	flag	0	0.9879177365335937
15	Tabaco	flag	0	0.9819939991922357
16	Drogas en general	flag	0	0.9660637412913259
17	'Psico-sanitarios'	flag	0	0.9623639780994336
18	Nueva investigación	flag	1	0.9467891589719334
19	Académicos	flag	1	0.9341869015012044

Rango	Campo	Tipo	Importancia	Valor
20	Valoración de Unidad de análisis Discreta	set	1	0.9070151597920257

13.7.1.12.- Alcohol

Rango	Campo	Tipo	Importancia	Valor
1	Categoría Tema Principal	set	0	1.0
2	Epidemiología	flag	0	0.9999999998025477
3	Nueva investigación	flag	0	0.9999990394655641
4	Éxtasis/MDMA	flag	0	0.9999823234667342
5	Drogas en general	flag	0	0.9999712523999245
6	Estrategia política	flag	0	0.9998940500114334
7	Cannabis	flag	0	0.9997453675189013
8	Delito	flag	0	0.9997414703789047
9	Cantidad de Fuentes	orderedSet	0	0.9973027870508631
10	Hachís	flag	0	0.9967334118968498
11	Opinión pública	flag	0	0.9953881060577382
12	'Psico-sanitarios'	flag	0	0.9953881060577382
13	Ciencia	flag	0	0.9895334209089329
14	Tribunal	flag	0	0.986157559935027
15	Académicos	flag	0	0.9824742217059371
16	Contexto general científico/médico	flag	0	0.9803004864033218
17	Valoración de Unidad de análisis Discreta	set	0	0.9774343957538606
18	Afiliación institucional	flag	0	0.9672604994854518
19	'No expertos'	flag	0	0.9664256849965506
20	Fuerzas	flag	0	0.959179263972609
21	Privada	flag	1	0.945375191829447
22	Mercado/Empresa	flag	1	0.920424143291205

13.7.1.13.- Tabaco

Rango	Campo	Tipo	Importancia	Valor
1	Delito	flag	0	1.0
2	Categoría Tema Principal	set	0	1.0
3	Drogas en general	flag	0	0.9999999999933437
4	Fuerzas	flag	0	0.9999999953526512
5	Contexto general científico/médico	flag	0	0.9999999407881223
6	'Psico-sanitarios'	flag	0	0.9999991664770128
7	Cocaína	flag	0	0.9999812543946826
8	Política/Legislación	flag	0	0.9998771267050535
9	Epidemiología	flag	0	0.9998624307799981
10	Nº de imágenes de la unidad de análisis	set	0	0.9992403623688458

Rango	Campo	Tipo	Importancia	Valor
11	Mercado/Empresa	flag	0	0.9983900177433517
12	Marihuana	flag	0	0.99610203569509
13	Opinión pública	flag	0	0.9957153984077463
14	Hachís	flag	0	0.9942899877421457
15	Afiliación institucional	flag	0	0.9909266691870018
16	Privada	flag	0	0.9840644385534434
17	Personalización anecdótica	flag	0	0.9825074720606847
18	Cannabis	flag	0	0.9819939991922357
19	Valoración de Unidad de análisis Discreta	set	0	0.9771804391743518
20	Tribunal	flag	0	0.9729020833523698
21	Académicos	flag	0	0.9629266519033004

13.7.2.- ANÁLISIS DE CAMPOS INFLUYENTES EN *FRAME*

13.7.2.1.- Nueva Investigación

Rango	Campo	Tipo	Importancia	Valor
1	Ciencia	flag	0	0.9999999999994383
2	Contexto general científico/médico	flag	0	0.9999999986215135
3	Delito	flag	0	0.999999998470491
4	Privada	flag	0	0.9999999707463118
5	Categoría Tema Principal	set	0	0.9999999663649846
6	Académicos	flag	0	0.9999999632415847
7	Afiliación institucional	flag	0	0.9999999581563097
8	Alcohol	flag	0	0.9999990394655641
9	Epidemiología	flag	0	0.9999971029249478
10	Cantidad de Fuentes	orderedSet	0	0.9999324500100766
11	Opinión pública	flag	0	0.999548320338264
12	'Psico-sanitarios'	flag	0	0.999548320338264
13	Éxtasis/MDMA	flag	0	0.9972294776473174
14	Fuerzas	flag	0	0.996798357220331
15	Crystal/Cristal	flag	0	0.9928859235135
16	Estrategia política	flag	0	0.9908282821605606
17	Tribunal	flag	0	0.9674673415113946
18	Personalización anecdótica	flag	0	0.9556571132046235
19	Mercado/Empresa	flag	0	0.951602471968607
20	MesNumerico	range	0	0.9509146173322969
21	Hachís	flag	1	0.9485214040244446
22	Cannabis	flag	1	0.9467891589719334
23	Cargo	flag	1	0.9450826628063008
24	Forma de Aparición	orderedSet	1	0.9302537675054509

Rango	Campo	Tipo	Importancia	Valor
25	Otra	Flag	1	0.9181237517309231

13.7.2.2.- Contexto General Científico/Médico

Rango	Campo	Tipo	Importancia	Valor
1	'Psico-sanitarios'	flag	0	0.9999999999823627
2	Nueva investigación	flag	0	0.9999999986215135
3	Delito	flag	0	0.9999999881975272
4	Ciencia	flag	0	0.9999999824128845
5	Tabaco	flag	0	0.9999999407881223
6	Categoría Tema Principal	set	0	0.99998108461687
7	Crystal/Cristal	flag	0	0.9998391282322531
8	Académicos	flag	0	0.9995999367249068
9	Fuerzas	flag	0	0.9988781937437771
10	Cantidad de Fuentes	orderedSet	0	0.9949736536650172
11	Personalización anecdótica	flag	0	0.9934951550731342
12	Estrategia política	flag	0	0.991835136923623
13	'No expertos'	flag	0	0.9865357701703956
14	Drogas en general	flag	0	0.9858641890257429
15	Alcohol	flag	0	0.9803004864033218
16	MesNumerico	range	0	0.9755273103001172
17	Privada	flag	0	0.9616325882806854
18	Valoración de Unidad de análisis Discreta	set	0	0.9612920219353115
19	Afiliación institucional	flag	0	0.9579699445531442
20	Hachís	flag	1	0.9196683246945058

13.7.2.3.- Ética/Moralidad

Rango	Campo	Tipo	Importancia	Valor
1	Cargo	flag	0	0.9999999999999655
2	Privada	flag	0	0.9999999997688319
3	Afiliación institucional	flag	0	0.9999999996732345
4	Delito	flag	0	0.9999998331946558
5	Categoría Tema Principal	set	0	0.999990479729702
6	Política/Legislación	flag	0	0.9997598530038161
7	Marihuana	flag	0	0.9993762050306841
8	Cantidad de Fuentes	orderedSet	0	0.994749263563708
9	Personalización anecdótica	flag	0	0.9915255086843366
10	Cocaína	flag	0	0.966012982779988
11	Mercado/Empresa	flag	0	0.9554200480996974
12	Fuerzas	flag	1	0.9431719191587324
13	'No expertos'	flag	1	0.9093146848139245

13.7.2.4.- Estrategia Política:

Rango	Campo	Tipo	Importancia	Valor
1	Políticos	flag	0	1.0
2	Drogas en general	flag	0	0.999999962407937
3	Categoría Tema Principal	set	0	0.9999425269910038
4	Política/Legislación	flag	0	0.9998954701949451
5	Alcohol	flag	0	0.9998940500114334
6	Personalización anecdótica	flag	0	0.9996980085444979
7	Mercado/Empresa	flag	0	0.9996955120123768
8	Contexto general científico/médico	flag	0	0.991835136923623
9	Nueva investigación	flag	0	0.9908282821605606
10	Cantidad de Fuentes	orderedSet	0	0.9871052217122112
11	MesNumerico	range	0	0.963142164011905
12	Epidemiología	flag	0	0.9505071295126049
13	Opinión no experta	flag	1	0.9492956326072273
14	'Psico-sanitarios'	flag	1	0.9342848931746135
15	'No expertos'	flag	1	0.9239637397510032

13.7.2.5.- Político/Legislación

Rango	Campo	Tipo	Importancia	Valor
1	Categoría Tema Principal	set	0	0.999999999999884
2	Delito	flag	0	0.9999999894810039
3	Estrategia política	flag	0	0.9998954701949451
4	Tabaco	flag	0	0.9998771267050535
5	Ética/Moralidad	flag	0	0.9997598530038161
6	Cocaína	flag	0	0.9995735937785033
7	'No expertos'	flag	0	0.9994477106338024
8	Opinión no experta	flag	0	0.9901510078295548
9	Drogas en general	flag	0	0.9879957857692367
10	Privada	flag	0	0.9822633285299087
11	Afiliación institucional	flag	0	0.979569529357949
12	Fuerzas	flag	0	0.9769157190631331
13	Nº de imágenes de la unidad de análisis	set	0	0.9565195298904822
14	Heroína	flag	1	0.9318463897769075
15	Otra	flag	1	0.9170744858954581
16	Ciencia	flag	1	0.9072569626063557
17	Opinión pública	flag	1	0.9055515885644192
18	Valoración de Unidad de análisis Discreta	set	1	0.9028190057396636

13.7.2.6.- Mercado/Empresa:

Rango	Campo	Tipo	Importancia	Valor
1	Privada	flag	0	0.9999804656456204
2	Afiliación institucional	flag	0	0.9998819927185645
3	Categoría Tema Principal	set	0	0.999832564476585
4	Estrategia política	flag	0	0.9996955120123768
5	Tabaco	flag	0	0.9983900177433517
6	Personalización anecdótica	flag	0	0.9983152517261086
7	Valoración de Unidad de análisis Discreta	set	0	0.9862162249367564
8	Cargo	flag	0	0.981913131930645
9	Ética/Moralidad	flag	0	0.9554200480996974
10	Nueva investigación	flag	0	0.951602471968607
11	Alcohol	flag	1	0.920424143291205
12	Delito	flag	1	0.9201138005115863

13.7.2.7.- Epidemiología

Rango	Campo	Tipo	Importancia	Valor
1	Categoría Tema Principal	set	0	1.0
2	Delito	flag	0	0.9999999999980977
3	Cannabis	flag	0	0.9999999999964464
4	Alcohol	flag	0	0.9999999998025477
5	'Psico-sanitarios'	flag	0	0.9999999979720312
6	Otra	flag	0	0.9999999632928365
7	Académicos	flag	0	0.9999999240711192
8	Éxtasis/MDMA	flag	0	0.9999992393254697
9	Nueva investigación	flag	0	0.9999971029249478
10	Afiliación institucional	flag	0	0.9999793820466086
11	Fuerzas	flag	0	0.9999402209435361
12	Privada	flag	0	0.999930274163827
13	Tabaco	flag	0	0.9998624307799981
14	Heroína	flag	0	0.9996937680316695
15	Cantidad de Fuentes	orderedSet	0	0.9995732978028532
16	Valoración de Unidad de análisis Discreta	set	0	0.9991791471185751
17	Cocaína	flag	0	0.9984513886056086
18	Psicofármacos	flag	0	0.9946595769072154
19	Ciencia	flag	0	0.9932864371971136
20	Opinión pública	flag	0	0.980736220864259
21	Crack/Cocaína base	flag	0	0.9767913513920718
22	Políticos	flag	0	0.9735457906246402
23	Drogas en general	flag	0	0.9508206622867879
24	Estrategia política	flag	0	0.9505071295126049

Rango	Campo	Tipo	Importancia	Valor
25	Cargo	flag	1	0.9309695577301862
26	Marihuana	flag	1	0.9272850806032382

13.7.2.8.- Opinión Pública

Rango	Campo	Tipo	Importancia	Valor
1	Delito	flag	0	0.9999998931362026
2	Categoría Tema Principal	set	0	0.9999962856898486
3	Éxtasis/MDMA	flag	0	0.9999877593647015
4	Nueva investigación	flag	0	0.999548320338264
5	Tabaco	flag	0	0.9957153984077463
6	Alcohol	flag	0	0.9953881060577382
7	Cantidad de Fuentes	orderedSet	0	0.9939879076158795
8	Otra	flag	0	0.9912992640156206
9	Privada	flag	0	0.9908491628895454
10	Fuerzas	flag	0	0.9904883100557133
11	Afiliación institucional	flag	0	0.9896886025418525
12	Ciencia	flag	0	0.9882813545548038
13	Drogas en general	flag	0	0.9840189558832787
14	Epidemiología	flag	0	0.980736220864259
15	Hachís	flag	1	0.9331074455097725
16	Política/Legislación	flag	1	0.9055515885644192

13.7.2.9.- Opinión no experta

Rango	Campo	Tipo	Importancia	Valor
1	'No expertos'	flag	0	0.9999827149588698
2	MesNumerico	range	0	0.9919803850972939
3	Política/Legislación	flag	0	0.9901510078295548
4	Categoría Tema Principal	set	0	0.964598633225114
5	Delito	flag	0	0.9552323744990806
6	Estrategia política	flag	1	0.9492956326072273
7	Fuerzas	flag	1	0.9280090322358908

13.7.2.10.- Personalización Anecdótica

Rango	Campo	Tipo	Importancia	Valor
1	'No expertos'	flag	0	0.9999999999999872
2	Categoría Tema Principal	set	0	0.9999999999997312
3	Delito	flag	0	0.9999999909421307
4	Nº de imágenes de la unidad de análisis	set	0	0.999999957619005
5	Cantidad de Fuentes	orderedSet	0	0.9999852025493511
6	Marihuana	flag	0	0.9999647431841404
7	Estrategia política	flag	0	0.9996980085444979

Rango	Campo	Tipo	Importancia	Valor
8	Políticos	flag	0	0.9993563828605072
9	Valoración de Unidad de análisis Discreta	set	0	0.9989458190988516
10	Mercado/Empresa	flag	0	0.9983152517261086
11	Fuerzas	flag	0	0.9981041395990283
12	Contexto general científico/médico	flag	0	0.9934951550731342
13	Ética/Moralidad	flag	0	0.9915255086843366
14	Tabaco	flag	0	0.9825074720606847
15	Cocaína	flag	0	0.9669269822243998
16	Nueva investigación	flag	0	0.9556571132046235
17	Hachís	flag	1	0.913147340843523
18	MesNumerico	range	1	0.9111722446749874

13.7.2.11.- Delito

Rango	Campo	Tipo	Importancia	Valor
1	Categoría Tema Principal	set	0	1.0
2	Tabaco	flag	0	1.0
3	Fuerzas	flag	0	1.0
4	Epidemiología	flag	0	0.9999999999980977
5	Afiliación institucional	flag	0	0.9999999999631677
6	Privada	flag	0	0.9999999999328365
7	Nueva investigación	flag	0	0.999999998470491
8	Personalización anecdótica	flag	0	0.9999999909421307
9	Política/Legislación	flag	0	0.9999999894810039
10	Contexto general científico/médico	flag	0	0.9999999881975272
11	Cocaína	flag	0	0.9999999635438214
12	Hachís	flag	0	0.9999999558820882
13	Opinión pública	flag	0	0.9999998931362026
14	Ética/Moralidad	flag	0	0.9999998331946558
15	Drogas en general	flag	0	0.9999993519397775
16	'Psico-sanitarios'	flag	0	0.999996209883824
17	Ciencia	flag	0	0.9999141928395245
18	Tribunal	flag	0	0.999843236089435
19	Alcohol	flag	0	0.9997414703789047
20	Nº de imágenes de la unidad de análisis	set	0	0.9989261305509013
21	Cargo	flag	0	0.9987980612845924
22	Académicos	flag	0	0.9922271339909762
23	Psicofármacos	flag	0	0.9880578225881932
24	Cantidad de Fuentes	orderedSet	0	0.980055643767246
25	Opinión no experta	flag	0	0.9552323744990806
26	Mercado/Empresa	flag	1	0.9201138005115863

13.7.3.- Análisis de Campos Influyentes en Fuentes

13.7.3.1.- Análisis Cuantitativo de Fuentes

Rango	Campo	Tipo	Importancia	Valor
1	Privada	flag	0	0.999999999999998
2	Afiliación institucional	flag	0	0.9999999999999978
3	Fuerzas	flag	0	0.999999999999556
4	'No expertos'	flag	0	0.9999999999991879
5	'Psico-sanitarios'	flag	0	0.9999999998958313
6	Políticos	flag	0	0.9999999997100489
7	Cargo	flag	0	0.9999999980230477
8	Personalización anecdótica	flag	0	0.9999852025493511
9	Académicos	flag	0	0.9999705153317332
10	Nueva investigación	flag	0	0.9999324500100766
11	Tribunal	flag	0	0.9998679676497493
12	Ciencia	flag	0	0.9997699887486947
13	Categoría Tema Principal	set	0	0.9996902317650563
14	Nº de imágenes de la unidad de análisis	set	0	0.9996102053027042
15	Epidemiología	flag	0	0.9995732978028532
16	Valoración de Unidad de análisis Discreta	set	0	0.9984333562851232
17	Alcohol	flag	0	0.9973027870508631
18	Contexto general científico/médico	flag	0	0.9949736536650172
19	Ética/Moralidad	flag	0	0.994749263563708
20	Opinión pública	flag	0	0.9939879076158795
21	Cannabis	flag	0	0.9913627863799233
22	Otra	flag	0	0.9898204546018777
23	Estrategia política	flag	0	0.9871052217122112
24	Delito	flag	0	0.980055643767246

13.7.3.2.- Políticos

Rango	Campo	Tipo	Importancia	Valor
1	Estrategia política	flag	0	1.0
2	Cantidad de Fuentes	orderedSet	0	0.9999999997100489
3	Drogas en general	flag	0	0.9999695149853642
4	'No expertos'	flag	0	0.9996398115168268
5	Personalización anecdótica	flag	0	0.9993563828605072
6	Cannabis	flag	0	0.9965955905281462
7	Epidemiología	flag	0	0.9735457906246402
8	Hachís	flag	0	0.9662183977353093
9	Valoración de Unidad de análisis Discreta	set	0	0.9564484862764281

Rango	Campo	Tipo	Importancia	Valor
10	Fuerzas	flag	1	0.946555506975698

13.7.3.3.- Fuerzas y Cuerpos de Seguridad del Estado

Rango	Campo	Tipo	Importancia	Valor
1	Categoría Tema Principal	set	0	1.0
2	Delito	flag	0	1.0
3	Cantidad de Fuentes	orderedSet	0	0.999999999999556
4	Tabaco	flag	0	0.9999999953526512
5	Cocaína	flag	0	0.9999999945460496
6	Hachís	flag	0	0.9999999251160749
7	Afiliación institucional	flag	0	0.9999994919366836
8	Privada	flag	0	0.9999993668643762
9	Epidemiología	flag	0	0.9999402209435361
10	Contexto general científico/médico	flag	0	0.9988781937437771
11	Personalización anecdótica	flag	0	0.9981041395990283
12	Nueva investigación	flag	0	0.996798357220331
13	Opinión pública	flag	0	0.9904883100557133
14	Cargo	flag	0	0.9904883100557133
15	Ciencia	flag	0	0.9821468291015294
16	Política/Legislación	flag	0	0.9769157190631331
17	'Psico-sanitarios'	flag	0	0.9709429107000578
18	Alcohol	flag	0	0.9591792639726089
19	Anabolizantes	flag	0	0.9508025027820637
20	Políticos	flag	1	0.946555506975698
21	Ética/Moralidad	flag	1	0.9431719191587324
22	Opinión no experta	flag	1	0.9280090322358908

13.7.3.4.- Tribunal

Rango	Campo	Tipo	Importancia	Valor
1	Cantidad de Fuentes	orderedSet	0	0.9998679676497493
2	Delito	flag	0	0.999843236089435
3	Alcohol	flag	0	0.986157559935027
4	Hachís	flag	0	0.9843501793525344
5	Cocaína	flag	0	0.9743811184375246
6	Tabaco	flag	0	0.9729020833523698
7	Nueva investigación	flag	0	0.9674673415113946
8	Categoría Tema Principal	set	0	0.9642000807263083
9	'No expertos'	flag	1	0.9272089314636089

13.7.3.5.- Ciencia

Rango	Campo	Tipo	Importancia	Valor
1	Nueva investigación	flag	0	0.9999999999994383
2	Contexto general científico/médico	flag	0	0.9999999824128845
3	Académicos	flag	0	0.9999999444382374
4	Psicofármacos	flag	0	0.9999194023277662
5	Delito	flag	0	0.9999141928395245
6	Otra	flag	0	0.9998466121366523
7	Cantidad de Fuentes	orderedSet	0	0.9997699887486947
8	'Psico-sanitarios'	flag	0	0.9994412139599762
9	Epidemiología	flag	0	0.9932864371971136
10	Alcohol	flag	0	0.9895334209089329
11	Opinión pública	flag	0	0.9882813545548038
12	Fuerzas	flag	0	0.9821468291015294
13	Categoría Tema Principal	set	0	0.9689736580991546
14	Política/Legislación	flag	1	0.9072569626063557

13.7.3.6.- Académicos

Rango	Campo	Tipo	Importancia	Valor
1	Nueva investigación	flag	0	0.9999999632415847
2	Ciencia	flag	0	0.9999999444382374
3	Epidemiología	flag	0	0.9999999240711192
4	'Psico-sanitarios'	flag	0	0.999973750906737
5	Cantidad de Fuentes	orderedSet	0	0.9999705153317332
6	Categoría Tema Principal	set	0	0.999959990585708
7	Valoración de Unidad de análisis Discreta	set	0	0.9999559123758466
8	Otra	flag	0	0.9997136422414261
9	Contexto general científico/médico	flag	0	0.9995999367249068
10	Éxtasis/MDMA	flag	0	0.9965536414190317
11	Delito	flag	0	0.9922271339909762
12	Alcohol	flag	0	0.9824742217059371
13	Nº de imágenes de la unidad de análisis	set	0	0.9759801909737407
14	Tabaco	flag	0	0.9629266519033004
15	MesNumerico	range	0	0.9558152657706275
16	Cannabis	flag	1	0.9341869015012044
17	Psicofármacos	flag	1	0.9037829739733458

13.7.3.7.- Privada

Rango	Campo	Tipo	Importancia	Valor
1	Afiliación institucional	flag	0	1.0
2	Cargo	flag	0	1.0

Rango	Campo	Tipo	Importancia	Valor
3	Cantidad de Fuentes	orderedSet	0	0.999999999999998
4	Categoría Tema Principal	set	0	0.9999999999996136
5	Delito	flag	0	0.9999999999328365
6	Ética/Moralidad	flag	0	0.9999999997688319
7	Nueva investigación	flag	0	0.9999999707463118
8	Fuerzas	flag	0	0.9999993668643762
9	Cannabis	flag	0	0.9999859701621194
10	Mercado/Empresa	flag	0	0.9999804656456204
11	Epidemiología	flag	0	0.999930274163827
12	Opinión pública	flag	0	0.9908491628895454
13	Valoración de Unidad de análisis Discreta	set	0	0.9883430044553155
14	Tabaco	flag	0	0.9840644385534434
15	Política/Legislación	flag	0	0.9822633285299087
16	'Psico-sanitarios'	flag	0	0.9699889055494891
17	Contexto general científico/médico	flag	0	0.9616325882806854
18	Crystal/Cristal	flag	0	0.9578875102349217
19	Alcohol	flag	1	0.945375191829447
20	Otra	flag	1	0.9163405529398762

13.7.3.8.- Psico-Sanitarios

Rango	Campo	Tipo	Importancia	Valor
1	Contexto general científico/médico	flag	0	0.9999999999823627
2	Cantidad de Fuentes	orderedSet	0	0.9999999998958313
3	Epidemiología	flag	0	0.9999999979720312
4	Tabaco	flag	0	0.9999991664770128
5	Delito	flag	0	0.999996209883824
6	Categoría Tema Principal	set	0	0.9999898856663585
7	Académicos	flag	0	0.999973750906737
8	Valoración de Unidad de análisis Discreta	set	0	0.9997937704641732
9	Nueva investigación	flag	0	0.999548320338264
10	Ciencia	flag	0	0.9994412139599762
11	Otra	flag	0	0.9991367736211401
12	Alcohol	flag	0	0.9953881060577382
13	Psicofármacos	flag	0	0.9875356867840172
14	Drogas en general	flag	0	0.9840189558832787
15	Fuerzas	flag	0	0.9709429107000578
16	Privada	flag	0	0.9699889055494891
17	Cannabis	flag	0	0.9623639780994336
18	Nº de imágenes de la unidad de análisis	set	0	0.9599553037743737
19	Cocaína	flag	0	0.955475418005051

Rango	Campo	Tipo	Importancia	Valor
20	Estrategia política	flag	1	0.9342848931746135
21	Marihuana	flag	1	0.9117772731759578
22	Afiliación institucional	flag	1	0.9100581768449596

13.7.3.9.- No Expertos

Rango	Campo	Tipo	Importancia	Valor
1	Personalización anecdótica	flag	0	0.9999999999999872
2	Cantidad de Fuentes	orderedSet	0	0.9999999999991879
3	Opinión no experta	flag	0	0.9999827149588698
4	Políticos	flag	0	0.9996398115168268
5	Política/Legislación	flag	0	0.9994477106338024
6	Categoría Tema Principal	set	0	0.9993260630349089
7	Nº de imágenes de la unidad de análisis	set	0	0.9976040407451525
8	Marihuana	flag	0	0.9961138477506994
9	Contexto general científico/médico	flag	0	0.9865357701703956
10	Alcohol	flag	0	0.9664256849965506
11	Tribunal	flag	1	0.9272089314636089
12	Estrategia política	flag	1	0.9239637397510032
13	Drogas en general	flag	1	0.9172456667702769
14	Ética/Moralidad	flag	1	0.9093146848139245

13.7.4.- Análisis de Campos Influyentes en Otros Campos

13.7.4.1.- Categoría de Tema Principal

Rango	Campo	Tipo	Importancia	Valor
1	Delito	flag	0	1.0
2	Epidemiología	flag	0	1.0
3	Fuerzas	flag	0	1.0
4	Tabaco	flag	0	1.0
5	Alcohol	flag	0	1.0
6	Cocaína	flag	0	0.9999999999999097
7	Afiliación institucional	flag	0	0.9999999999998996
8	Política/Legislación	flag	0	0.999999999999884
9	Personalización anecdótica	flag	0	0.9999999999997312
10	Privada	flag	0	0.9999999999996136
11	Hachís	flag	0	0.999999999562722
12	Otra	flag	0	0.9999999838846187
13	Nueva investigación	flag	0	0.9999999663649846
14	Cargo	flag	0	0.9999999279406314
15	Opinión pública	flag	0	0.9999962856898486

Rango	Campo	Tipo	Importancia	Valor
16	Ética/Moralidad	flag	0	0.999990479729702
17	'Psico-sanitarios'	flag	0	0.9999898856663585
18	Contexto general científico/médico	flag	0	0.99998108461687
19	Drogas en general	flag	0	0.9999752763332395
20	Cannabis	flag	0	0.9999637756093478
21	Académicos	flag	0	0.999959990585708
22	Estrategia política	flag	0	0.9999425269910038
23	Mercado/Empresa	flag	0	0.999832564476585
24	Cantidad de Fuentes	orderedSet	0	0.9996902317650563
25	'No expertos'	flag	0	0.9993260630349089
26	Nº de imágenes de la unidad de análisis	set	0	0.9948005519282928
27	Valoración de Unidad de análisis Discreta	set	0	0.9930831396182129
28	Ciencia	flag	0	0.9689736580991546
29	Opinión no experta	flag	0	0.964598633225114
30	Tribunal	flag	0	0.9642000807263083
31	Éxtasis/MDMA	flag	0	0.9599122616599838

13.7.4.2.- Forma de Aparición

Rango	Campo	Tipo	Importancia	Valor
1	Hachís	flag	0	0.9960865494692241
2	EsDomingo	flag	0	0.9946362529648998
3	Éxtasis/MDMA	flag	0	0.9901205889273322
4	Nueva investigación	flag	1	0.9302537675054509

13.7.4.3.- Valoración de la Unidad de Análisis

Rango	Campo	Tipo	Importancia	Valor
1	Nº de imágenes de la unidad de análisis	set	0	1.0
2	Académicos	flag	0	0.9999559123758466
3	'Psico-sanitarios'	flag	0	0.9997937704641732
4	Drogas en general	flag	0	0.9992003958951069
5	Epidemiología	flag	0	0.9991791471185751
6	Personalización anecdótica	flag	0	0.9989458190988516
7	Cantidad de Fuentes	orderedSet	0	0.9984333562851232
8	Categoría Tema Principal	set	0	0.9930831396182129
9	Cargo	flag	0	0.9888945653978238
10	Privada	flag	0	0.9883430044553155
11	Mercado/Empresa	flag	0	0.9862162249367564
12	Afiliación institucional	flag	0	0.9855585858187628
13	Alcohol	flag	0	0.9774343957538606
14	Tabaco	flag	0	0.9771804391743518

Rango	Campo	Tipo	Importancia	Valor
15	Éxtasis/MDMA	flag	0	0.9766772183883381
16	MesNumerico	range	0	0.9692409377410837
17	Otra	flag	0	0.9674712045079453
18	Contexto general científico/médico	flag	0	0.9612920219353115
19	Políticos	flag	0	0.9564484862764281
20	Cannabis	flag	1	0.9070151597920257
21	Política/Legislación	flag	1	0.9028190057396636

13.7.4.4.- Imágenes

Rango	Campo	Tipo	Importancia	Valor
1	Valoración de Unidad de análisis Discreta	set	0	1.0
2	Personalización anecdótica	flag	0	0.999999957619005
3	Cantidad de Fuentes	orderedSet	0	0.9996102053027042
4	Tabaco	flag	0	0.9992403623688458
5	Delito	flag	0	0.9989261305509013
6	'No expertos'	flag	0	0.9976040407451525
7	Categoría Tema Principal	set	0	0.9948005519282928
8	Cocaína	flag	0	0.9880346723098237
9	Crack/Cocaína base	flag	0	0.9874721678986765
10	Académicos	flag	0	0.9759801909737407
11	Hachís	flag	0	0.9718528801589782
12	'Psico-sanitarios'	flag	0	0.9599553037743737
13	Política/Legislación	flag	0	0.9565195298904822
14	Psicofármacos	flag	0	0.950408838140222

13.7.4.5.- EsDomingo

Rango	Campo	Tipo	Importancia	Valor
1	Forma de Aparición	orderedSet	0	0.9946362529648998
2	Marihuana	flag	1	0.9463492611044984
3	Hachís	flag	1	0.9083329735005313

14.- BIBLIOGRAFÍA

AGARWAL, R.C., AGGARWAL, C.C. AND PRASAD, V. 2001. A tree projection algorithm for generation of frequent item sets. *Journal of parallel and Distributed Computing* 61, 350-371.

AGRAWAL, R. Y SRIKANT, R. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases,* Anonymous Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 487-499.

ALEXA, M. 1997. Computer assisted text analysis methodology in the social sciences. ZUMA, .

ARAUJO ARREDONDO, N.P. 2009. Método Semisupervisado para la Clasificación Automática de Textos de Opinión.

AYRES, J., FLANNICK, J., GEHRKE, J. Y YIU, T. 2002. Sequential pattern mining using a bitmap representation. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining,* Anonymous ACM, , 429-435.

BARDIN, L. 1986. El análisis de contenido. Ediciones Akal, Madrid.

BERRIO, J.L. Medios de comunicación y drogas.

BERRY, M.J.A. Y LINOFF, G. 1997. Data Mining Techniques for Marqueting Sales and Customer Support. Wiley, .

BORDIGNON, F., PERI, J., TOLOSA, G., VILLA, D. Y PAOLETTI, L. 2004. Experimentos en clasificación automática de noticias en español utilizando el modelo bayesiano.

BORGELT, C. 2005. An Implementation of the FP-growth Algorithm. In *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations,* Anonymous ACM, , 1-5.

BRACHMAN, R. Y ANAND, T. 1996. The process of Knowledge Discovery in Databases: A human centered approach. Advances in Knowledge Discovery and Data Mining. AAAII MIT Press, .

BREIMAN, L., FRIEDMAN, J.H., OLSEN, R.A. Y STONE, C.J. 1984. Classification and Regression Trees. Wadsworth and Brooks/Cole, Monterey.

BUNTINE, W. 1992. Learning classification trees. Springer, .

BURDICK, D., CALIMLIM, M., FLANNICK, J., GEHRKE, J. Y YIU, T. 2005. Mafia: A maximal frequent itemset algorithm. *Knowledge and Data Engineering, IEEE Transactions on* 17, 1490-1504.

CABENA, P., HADJINIAN, P., STADLER, R., VERHEES, J., ZANASI, A., INTERNATIONAL BUSINESS MACHINES CORPORATION (SAN JOSE, CALIFORNIA). AND INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION (SAN

JOSE, CALIFORNIA). 1998. Discovering data mining: from concept to implementation. Prentice Hall, .

CANGA LAREQUI, J., COCA GARCÍA, C., PEÑA FERNÁNDEZ, S. Y PÉREZ DASILVA, J. 2010. Terrorismo y política dominan las portadas de la prensa vasca. Análisis de contenido y superficie de las primeras páginas de los diarios autonómicos. *Revista Latina de Comunicación Social* 65, 28 de febrero de 2010-61; 70.

CASASÚS, J.M. 1985. Ideología y análisis de medios de comunicación.

CEGLAR, A. Y RODDICK, J.F. 2006. Association mining. *ACM Computing Surveys (CSUR)* 38, 5.

CHAPMAN, P., CLINTON, J., KERBER, R., KHABAZA, T., REINARTZ, T., SHEARER, C. AND WIRTH, R. 2000. CRISP-DM 1.0: Step-by-step data mining guide. SPSS, .

CHUNG, S. Y MCLEOD, D. 2003. Dynamic topic mining from news stream data. *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE* 653-670.

CLARK, P. Y BOSWELL, R. 1991. Rule induction with CN2: Some recent improvements. In *Machine learning—EWSL-91,* Anonymous Springer, , 151-163.

CLARK, P. Y NIBLETT, T. 1989. The CN2 induction algorithm. *Machine Learning* 3, 261-283.

CLIFTON, C. 2004. TopCat: data mining for topic identification in a text corpus. *Knowledge and Data Engineering, IEEE Transactions on* 949.

COLLE, R. 2002. Explotar la información noticiosa: Data mining aplicado a la documentación periodística. Universidad Complutense, Departamento de Biblioteconomía y Documentación, .

ENTMAN, R.M. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication* 43, 51-58.

FAYYAD, U.M., GRINGSTEIN, G.G. Y WIERSE, A. 2002. Information and Visualization in Data Mining and Knowledge Diskovery. Morgan Kauffmann, .

FERNÁNDEZ-CID, M. OTROS (1996): Tratamiento periodístico de las drogas y las drogodependencias. *Madrid, Coordinadora de ONG's que intervienen en Drogodependencias* .

FRANCIS, L.A. 2006. Taming Text: An Introduction to Text Mining. In *Casualty Actuarial Society Forum, Winter,* Anonymous Citeseer, .

FREITAS, A.A. 2002. Data mining and knowledge discovery with evolutionary algorithms. Springer Verlag, .

GENG, L. Y HAMILTON, H.J. 2006. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)* 38, 9.

GOFFMAN, E. Y BERGER, B.M. 1974. Frame analysis: An essay on the organization of experience.

GOUDA, K. Y ZAKI, M.J. 2002. Efficiently mining maximal frequent itemsets. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on,* Anonymous IEEE, , 163-170.

GUNNARSSON, C.L., WALKER, M.M., WALATKA, V. Y SWANN, K. 2007. Lessons learned: A case study using data mining in the newspaper industry. *Journal of Database Marketing & Customer Strategy Management* 14, 271-280.

HAIR, J.F., ANDERSON, R.E., TATHAM, R.L. Y BLACK, W.C. 1999. Análisis multivariante. Prentice Hall, .

HAN, J., FU, Y., HUANG, Y., CAI, Y. Y CERCONE, N. 1994. DBLearn: A system prototype for knowledge discovery in relational databases. In *Proceedings of the 1994 ACM SIGMOD international conference on Management of data,* Anonymous ACM, , 516.

HAN, J., PEI, J., YIN, Y. Y MAO, R. 2004. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery* 8, 53-87.

HAND, D., MANNILA, H. Y SMYTH, P. 2001. Principles of Data Mining. A Bradford Book. MIT Press, London.

HERNÁNDEZ ORALLO, J., RAMÍREZ QUINTANA, M.J. Y FERRI RAMÍREZ, C. 2004. Introducción a la minería de datos. Pearson Education, Valencia.

HIDBER, C. 1999. Online association rule mining. *ACM SIGMOD Record* 28, 145-156.

HIPP, J., GUNTZER, U. Y NAKHAEIZADEH, G. 2000. Algorithms for association rule mining — a general survey and comparison. *SIGKDD Explor.Newsl.* 2, 58-64.

HONG, T. Y HAN, I. 2002. Knowledge-based data mining of news information on the Internet using cognitive maps and neural networks. *Expert Systems with Applications* 23, 1-8.

JOSHI, K.P. 1997. Analysis of data mining algorithms. *University of Minnesota.Retrieved July* 25, 2005.

KASS, G.V. 1980. An exploratory technique for investigating large quantities of categorical data. *Applied statistics* 29, 119-127.

KAYSER, J. 1974. El diario francés. ATE, .

KIMBALL, R., REEVES, L., ROSS, M. Y THORNTHWAITE, W. 1998. The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses John Wiley & Sons, New York.

LIANG, X. Y RONG-CHANG CHEN. 2005. Mining stock news in cyberworld based on natural language processing and neural networks. 8-8.

LINDSAY, R.K., BUCHANAN, B.G., FEIGENBAUM, E.A. Y LEDERBERG, J. 1993. DENDRAL: a case study of the first expert system for scientific hypothesis formation. *Artificial Intelligence* 61, 209-261.

MARTINEZ ALBERTOS, J.L. 1984. Curso general de redacción periodística.

MEHTA, M., AGRAWAL, R. Y RISSANEN, J. 1996. SLIQ: A fast scalable classifier for data mining. *Advances in Database Technology—EDBT'96* 18-32.

MICHALSKI, R.S., AMAREL, S., LENAT, D.B., MICHIE, D. Y WINSTON, P.H. 1986. Machine learning: Challenges of the eighties. *Machine Learning: An Artificial Intelligence Approach* 2, 27-41.

MICHALSKI, R.S., BRATKO, I. Y KUBAT, M. 1997. Machine Learning, Data Mining and Knowledge Discovery: Methods and Applications. John Wiley and Sons, .

MORTAZAVI-ASL, B., WANG, J., PINTO, H. Y CHEN, Q. 2004. Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering* 16, .

NEBREDA, B.G., SENDRA, J.M., ALBERT, A.P. Y ESPAÑA. DELEGACIÓN DEL GOBIERNO PARA EL PLAN NACIONAL SOBRE DROGAS. 1987. La imagen de la droga en la prensa española. Ministerio de Sanidad y Consumo, Delegación del Gobierno para el Plan Nacional sobre Drogas, Publicaciones, Documentación y Biblioteca, .

NELKIN, D. 1987. Selling science. How the press covers science and technology.

NORVAG, K. 2005. News Item Extraction for Text Mining in Web Newspapers. 195.

NÚÑEZ-ROMERO OLMO, F. 2009. La formación de las secciones de deportes en los diarios de información general españoles antes de 1936. Análisis hemerográfico estructural comparado.

NÚÑEZ-ROMERO OLMO, F., PARICIO ESTEBAN, P. Y RABADÁN ZARAGOZÁ, M.J. 2010. Puesta en página de las drogas en la prensa de información general española. In *XIII Jornadas de Fotografía, Edición y Diseño,* Universidad CEU San Pablo. Madrid, 3 y 4 de marzo de 2010, Anonymous .

ORLANDO, S., PALMERINI, P. Y PEREGO, R. 2001. *The DCP algorithm for Frequent Set Counting* .

PARICIO ESTEBAN, M.P. Y RABADÁN ZARAGOZÁ, M.J. 2010. Comunicación y prevención de las drogodependencias. In *Campañas y comunicación institucional para la prevención de la drogadicción,* M.P. PARICIO ESTEBAN, Ed. Erasmus, Barcelona, 37-60.

PARICIO ESTÉBAN, P. 2009. Informe Interno del Proyecto: "Análisis y diseño de campañas y programas de sensibilización y prevención de las drogodependencias en los medios de comunicación".

PARICIO ESTEBAN, P., NÚÑEZ-ROMERO OLMO, F. Y SANFELIU AGUILAR, P. 2010. Tratamiento informativo de las drogas en las revistas para adolescentes 2008-2009. In *V Congreso Internacional Prensa y Periodismo Especializado. Historia y Realidad Actual,* Guadalajara. España, 6-7 de mayo de 2010, Anonymous .

PARICIO ESTEBAN, P., NÚÑEZ-ROMERO OLMO, F., RODRÍGUEZ LUQUE, C. Y RABADÁN ZARAGOZÁ, M.J. 2010. La falta de perspectiva social en las informaciones sobre drogas de los diarios generalistas nacionales en España en 2009. In *XII Congreso de la Sociedad Española de Periodística,* Universidad CEU Cardenal Herrera. Valencia, 21 y 22 de mayo de 2010, Anonymous .

PARICIO ESTEBAN, P., SANFELIU AGUILAR, P. Y SANFELIU MONTORO, A. 2002. Las campañas de comunicación y publicitarias sobre sida y drogas. *Revista española de drogodependencias* 27, 489-513.

PARICIO, P. Y SANFELIU, P. 2010. El tratamiento informativo de la drogadicción y su prevención en la prensa dirigida a adolescentes. In *Campañas y comunicación institucional para la prevención de la drogadicción,* P. PARICIO ESTEBAN, Ed. Erasmus Ediciones, Villafranca del Penadés, 109-134.

PARK, J.S., CHEN, M.S. Y YU, P.S. 1995. An effective hash-based algorithm for mining association rules. *ACM SIGMOD Record* 24, 175-186.

PEARL, J. 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, .

PERNÍA, A., MARTINEZ DE PISÓN, F.J., ORDIERES, J.B., CASTEJÓN, M. Y DE COS, F.J. 2001. Gestión del Conocimiento y Minería de datos. In *XVII Congreso Nacional de Ingeniería de Proyectos,* Murcia, Anonymous .

PIATETSKI-SHAPIRO, G. Y FRAWLEY, W.J. 1991. Knowledge Discovery in Databases. AAAI/MIT Press, .

PYLE, D. 1999. Data Preparation for Datamining. Morgan Kaufmann Publishers, San Francisco, California.

QUINLAN, J.R. 1986. Induction of decision trees. *Machine Learning* 1, 81-106.

QUINLAN, J.R. 1993. C4. 5: programs for machine learning. Morgan Kaufmann, .

REESE, S.D., GANDY, O.H. Y GRANT, A.E. 2001. Framing public life: Perspectives on media and our understanding of the social world. Lawrence Erlbaum, .

REILLY, B.F. 2009. When Machines Do Research: Automated Analysis of News and Other Primary Source Texts. *Journal of Library Administration* 49, 507-517.

REKALDE, A. Y ROMANÍ, O. 2002. Los medios de comunicación social ante el fenómeno de las drogas: un análisis crítico.

RODRÍGUEZ LUQUE, C. 2009. El tratamiento periodístico de las "células madre" desde la perspectiva del "framing". El País y ABC (1996-2006).

RODRÍGUEZ LUQUE, C. Y NÚÑEZ-ROMERO OLMO, F. 2011. Tratamiento periodístico de las drogas en los medios de información general nacionales. El caso español. *El Mundo, ABC, El País, La Razón* (Enero-Junio 2009) - Capítulo de libro.

SÁDABA, M.T. 2001. Origen, aplicación y límites de la teoría del encuadre (framing) en comunicación. *Comunicación y sociedad* 14, 143-175.

SCHEUFELE, D.A. 1999. Framing as a theory of media effects. *Journal of communication* 49, 103-122.

SENO, M. Y KARYPIS, G. 2002. Slpminer: An algorithm for finding frequent sequential patterns using length-decreasing support constraint.

SMYTH, P. Y GOODMAN, R.M. 2002. An information theoretic approach to rule induction from databases. *Knowledge and Data Engineering, IEEE Transactions on* 4, 301-316.

VEGA FUENTE, A. 1995. Los medios de comunicación social y las drogas: entre la publicidad y el control social. *Revista española de drogodependencias* 20, 99-111.

WALTER LIMA, J. 2008. Texts and data mining and their possibilities applied to the process of news production. *BRAZILIAN JOURNALISM RESEARCH* 4, .

WEISS, S.M. 2005. Text mining: predictive methods for analyzing unstructured information. Springer-Verlag New York Inc, .

WESTPHAL, C. Y BLAXTON, T. 1998. Data mining solutions: methods and tools for solving real-world problems. John Wiley & Sons, Inc. New York, NY, USA, .

WITTEN, I.H. Y FRANK, E. 2000. Data Mining Practical Machine Learning Tools and Techniques with Java Implementation. Morgan Kaufmann Publishers, .

ZAKI, M.J. 2001. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning* 42, 31-60.

15.- ÍNDICE DE TABLAS

16.- ÍNDICE DE FIGURAS