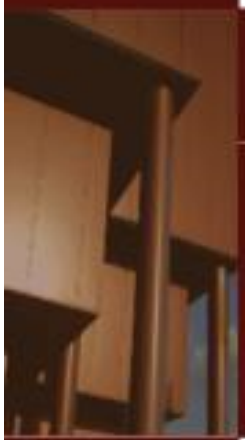
Grado en Ingeniería de Tecnologías Industriales
Trabajo Fin de Grado

Minería de datos. Aplicaciones de técnicas predictivas.

Autor: Diego Morales Cifuentes

Tutor: José Miguel León Blanco



Dep. de Organización Industrial y Gestión de Empresas I
Escuela Técnica Superior de Ingeniería
Universidad de Sevilla

Sevilla, 2016

Grado en Ingeniería de Tecnologías Industriales
Trabajo Fin de Grado

Minería de datos. Aplicaciones de técnicas predictivas.

Autores:
Diego Morales Cifuentes

Tutor:
José Miguel León Blanco
Profesor colaborador

Dep. de Organización Industrial y Gestión de Empresas I
Escuela Técnica Superior de Ingeniería
Universidad de Sevilla
Sevilla, 2016

Trabajo Fin de Grado: Minería de datos. Aplicaciones de técnicas predictivas.

Autor: Diego Morales Cifuentes

Tutor: José Miguel León Blanco

El tribunal nombrado para juzgar el Proyecto arriba indicado, compuesto por los siguientes miembros:

Presidente:

Vocales:

Secretario:

Acuerdan otorgarle la calificación de:

Sevilla, 2016

El Secretario del Tribunal

Resumen

El objetivo marcado de este proyecto, que se ha llevado a cabo en paralelo junto a mi compañero Miguel Novoa, era aplicar la minería de datos a un conjunto de datos que Giahsa, la empresa que se encarga de la canalización del agua en gran parte de la provincia de Huelva, nos proporcionó. En primer lugar, se pretendía detectar los comportamientos anómalos de los clientes para así detectar el fraude y modernizar la forma de detección empleada por la empresa hasta la fecha. En segundo lugar, una vez detectados estos patrones de comportamiento, se pretendía caracterizar a los clientes para agruparlos según sus características, obteniéndose un mejor conocimiento de ellos. Con el uso de ambos métodos se iba a conseguir conocer mejor el perfil de los clientes fraudulentos, lo que supondría un gran avance debido a que si apareciese un nuevo cliente fraudulento, muy probablemente nuestro sistema lo localizaría al ver que cumple con los patrones establecidos con la minería de datos.

Por diversos motivos explicados en nuestro proyecto en profundidad vimos que no íbamos a poder llevar a cabo esto ya que nos faltaban datos, por lo que decidimos emplear otras bases de datos distintas y así poder demostrar la eficacia de la minería de datos cuando se tienen los datos oportunos. Tras realizar una extensa búsqueda por numerosas páginas web, y ver una gran cantidad de bases de datos, se decidió elegir las dos bases de datos nombradas a continuación.

En primer lugar, en mi parte del proyecto se han aplicado técnicas predictivas para detectar a posibles pacientes que pudieran estar padeciendo hipotiroidismo consiguiendo unos resultados asombrosos de un número elevado de acierto.

Y, en segundo lugar, mi compañero Miguel Novoa ha empleado métodos descriptivos para analizar el consumo de alcohol en adolescentes, con el objetivo de ver qué factores influyen más, y poder realizar campañas a grupos de alumnos que compartan características.

Índice

Resumen ..... **i**
Índice ..... **iii**
Índice de ilustraciones ..... **vi**
Índice de tablas ix
1 Minería De Datos **1**
1.1 Definición y origen: KDD **1**
1.2 Etapas del DM **4**
1.3 Ventajas **4**
1.4 Inconvenientes **5**
1.5 Campos de aplicación **5**
1.6 Técnicas **8**
1.7 Tipos de datos **12**
1.8 Softwares de minería de datos 13
1.9 Extensiones del data mining 15
1.9.1 Web mining **15**
1.9.2 Text mining **16**
2 Idea inicial del proyecto **18**
2.1. Introducción 18
2.1.1 Tipos de fraude en la acometida del agua **18**
2.1.2 ¿Por qué aplicar minería de datos al fraude? **19**
2.1.3 Alternativas a la minería de datos **20**
2.1.4 Antecedentes de proyectos similares **20**
2.2 Contacto con Giahsa **22**
2.2.1 Giahsa 22
2.2.2 Datos de Calañas y Manzanilla 23
2.2.3 Preprocesamiento de los datos **24**

2.3 Conversión de los datos de Excel a ".arff" **25**
2.4 Software empleado: Weka **26**
2.4.1 Introducción **26**
2.4.2 Interfaces de Weka **28**
2.5 Resultados de la minería de datos aplicada a la base de datos inicial **38**
2.6. Búsqueda de los nuevos datos 39
2.7. Conclusión **40**
3 Aplicación de técnicas predictivas 42
3.1 Introducción **42**
3.1.1 Breve explicación de la temática **42**
3.1.2 Atributos **43**
3.2 Antecedentes **46**
3.3 Técnicas predictivas empleadas **48**
3.3.1 Naive Bayes **48**
3.3.2 K - Nearest Neighbour (KNN) **49**
3.3.3 Árbol de decisión generado con C4.5 ..... **52**
3.4 Proceso práctico en el software de minería de datos sin preprocesado **53**
3.5 Preprocesado de los datos **55**
3.6 Proceso práctico en el software de minería de datos con preprocesado **56**
3.7 Conclusión **58**
4 Bibliografía **60**

Índice de ilustraciones

Pág.

Ilustración 1. Esquema del proceso KDD.(Han, Kamber, & Pei, 2012) **3**
Ilustración 2. Capas que forman una red neuronal.(Redes de Neuronas Artificiales, 2012) **10**
Ilustración 3. Archivo ".arff". (Elaboración propia, 2016) **26**
Ilustración 4. Logo del software empleado, Weka.(Weka 3: Data Mining Software in Java, s.f.) **27**
Ilustración 5. Interfaces de Weka.(Elaboración propia, 2016) **28**
Ilustración 6. Interfaz*Simple CLI*.(Elaboración propia, 2016) **28**
Ilustración 7. Interfaz*Explorer* con la pestaña *Preprocess*.(Elaboración propia, 2016) **29**
Ilustración 8. Opción de cargar una base de datos desde una dirección URL. (Elaboración propia, 2016) **29**
Ilustración 9. Los 4 modos de entrenamiento de Weka.(Bouckaert, y otros, 2016) **31**
Ilustración 10. Botón *More Options***.**(Elaboración propia, 2016) **32**
Ilustración 11. Icono de Weka que muestra si el software está ejecutando un algoritmo (si gira el Logo) o no.(Elaboración propia, 2016) **32**
Ilustración 12. Resultados de un experimento con un algoritmo Clasificador. (Elaboración propia, 2016) **33**
Ilustración 13. Acceso a opciones adicionales de cada experimento.(Elaboración propia, 2016) **33**
Ilustración 14. Interfaz*Explorer*con la pestaña *Cluster*.(Elaboración propia, 2016) **34**
Ilustración 15. Interfaz*Explorer*con la pestaña *Select Attributes.*(Elaboración propia, 2016) **35**
Ilustración 16. Interfaz *Explorer*con la pestaña *Visualize*.(Elaboración propia, 2016) **36**
Ilustración 17. Interfaz *Experimenter*. (Elaboración propia, 2016) **37**
Ilustración 18. Interfaz *Knowledge Flow*. (Elaboración propia, 2016) **38**
Ilustración 19. Anatomía de la tiroides.(Martínez Fraga, 2012) **41**
Ilustración 20.Síntomas del hipotiroidismo y del hipertiroidismo.(FERTIFARMA, 2016) **42**
Ilustración 21. Glándulas endocrinas del cerebro.(Ropero Lara, 2012) **43**
Ilustración 22. Esquema de las hormonas tiroideas. (Nasser, 2016) **45**
Ilustración 23. Retrato de Thomas Bayes. (Dressler, 2013) **48**
Ilustración 24. Algoritmo KNN. (Elaboración propia, 2016) **49**
Ilustración 25. Ecuación y representación gráfica de la distancia euclídea. (Sancho Caparrini, 2013) **50**

Ilustración 26. Matriz de confusión generada por Weka con el algoritmoNaive Bayes. (Elaboración propia, 2016) **52**
Ilustración 27. Matriz de confusión generada por Weka con el algoritmoKNN con la distancia euclídea.(Elaboración propia, 2016) **52**
Ilustración 28. Matriz de confusión generada por Weka con el algoritmoKNN con la distancia Manhattan.(Elaboración propia, 2016) **52**
Ilustración 29. Matriz de confusión generada por Weka con el algoritmoC4.5 (J48). (Elaboración propia, 2016) **53**
Ilustración 30. Árbol de decisión generado por Weka.(Elaboración propia, 2016) **53**
Ilustración 31. Matriz de confusión generada por Weka con el algoritmoNaive Bayes. (Elaboración propia, 2016) **55**
Ilustración 32. Matriz de confusión generada por Weka con el algoritmoKNN con la distancia euclídea.(Elaboración propia, 2016) **55**
Ilustración 33. Matriz de confusión generada por Weka con el algoritmoKNN con la distancia Manhattan.(Elaboración propia, 2016) **55**
Ilustración 34. Matriz de confusión generada por Weka con el algoritmoC4.5 (J48). (Elaboración propia, 2016) **56**
Ilustración 35. Árbol de decisión generado por Weka.(Elaboración propia, 2016) **56**

Índice de tablas

Pág.

Tabla 1. Clasificación de las técnicas de la minería de datos.(Elaboración propia, 2016) **12**

Tabla 2. Proyectos y técnicas empleadas.(Elaboración propia, 2016) **46**

Tabla 3. Resultados de Weka sin preprocesado. (Elaboración propia, 2016) **57**

Tabla 4. Resultados de Weka con preprocesado. (Elaboración propia, 2016) **58**

Tabla 5. Comparativa de tiempos de ejecución. Tabla en segundos. (Elaboración propia, 2016) **58**

1 Minería De Datos

El fraude es un asunto que ha existido desde casi los inicios de la humanidad, pero desde hace ya muchos años el desarrollo de las tecnologías ha permitido desarrollar muchas más alternativas para llevarlo a cabo. Este proyecto se inicia con la intención de detectar el fraude en el consumo de agua por parte de los consumidores. Por ello se pensó en realizar un estudio a una población definida y comprobar posibles valores anormales que pudieran implicar un caso de fraude. Tras hablar y presentarle la idea a la empresa que se encarga de la gestión y distribución del agua canalizada de la mayoría de los pueblos de la provincia de Huelva, llamada Giahsa, les interesó y nos proporcionaron los datos de un par de municipios para analizarlos.

Giahsa detectó un significante incremento de los casos de fraude en los últimos tiempos, por lo que gestionar esto es un aspecto importante para Giahsa. El procedimiento que sigue esta empresa para la prevención de fraude es muy costoso, lento y aleatorio, por ello, hemos creído que la minería de datos podría ayudar a ver qué casos son más probables de fraude, crear algunos perfiles de casos de fraude que son comunes y así hacer que las inspecciones de contadores sean menos aleatorias, ahorrando tiempo y costes, y ganando en efectividad.

La intención era encontrar patrones de comportamiento de los clientes para detectar operaciones anómalas o sospechosas. Los datos proporcionados por Giahsa necesitarán ser tratados y se le aplicarán unas técnicas para ver si los resultados que obtenemos con los datos proporcionados serán exitosos o no.

Como veremos posteriormente, este objetivo no pudimos cumplirlo por diversas causas que serán explicadas más adelante en el proyecto. Debido a esto, decidimos enfocar nuestro objetivo final a la demostración de la eficacia de la minería de datos y sus técnicas empleando unos datos distintos.

1.1. Definición y origen: KDD

La minería de datos o también llamada “data mining”, DM, es el conjunto de técnicas que permiten explorar de manera automática o semiautomática grandes bases de datos y que se puede clasificar como una de las etapas dentro de un proceso mayor llamado Knowledge Discovery in Databases o KDD. El proceso KDD lo podemos explicar como “el proceso iterativo no trivial de identificar patrones válidos, novedosos y potencialmente útiles y, en última instancia, comprensible a partir de los datos”.

(Herrera Varela, 2006)

Este proceso también podemos verlo nombrado como Data Archeology, DependencyFunctionAnalysis, InformationRecoellect o KnowledgeFishing.

Las etapas del KDD pueden enumerarse de la siguiente manera:

1. Selección de datos: consiste en establecer un objetivo y las herramientas que vamos a necesitar. Es decir, primero debemos tener en cuenta lo que se sabe, lo que se quiere obtener y qué datos vamos a necesitar para conseguir esa información y, de este modo, alcanzar nuestro objetivo.

2. Limpieza de datos: en este segundo paso se limpian los datos, eliminando todos los datos que puedan influir en un análisis inexacto y en resultados incorrectos. Los motivos de esta limpieza son la existencia de datos incompletos, el ruido (valores incorrectos inesperados) y datos inconsistentes.

3. Integración de datos: Combinación de datos de múltiples procedencias.

4. Transformación de los datos: modificación sintáctica sobre los datos sin que esto suponga un cambio para la técnica de minería aplicada. La desventaja fundamental es que se puede disminuir la exactitud del resultado debido a que se pierda alguna información.

5. Reducción de datos: encontrar las características más significativas dependiendo de nuestro objetivo. Podemos emplear métodos de transformación para reducir el número efectivo de variables a ser consideradas o para encontrar otras representaciones de los datos.

6. Minería de datos: búsqueda de patrones que se expresan como un modelo. Se debe especificar un criterio de preferencia para seleccionar un modelo de un conjunto de posibles modelos.

7. Evaluación de los patrones: se evalúan los patrones descubiertos con técnicas que incluyen análisis estadísticos y lenguajes de consultas.

8. Interpretación de resultados: entendimiento de los resultados del análisis y puede llevar alguno de los pasos anteriores.

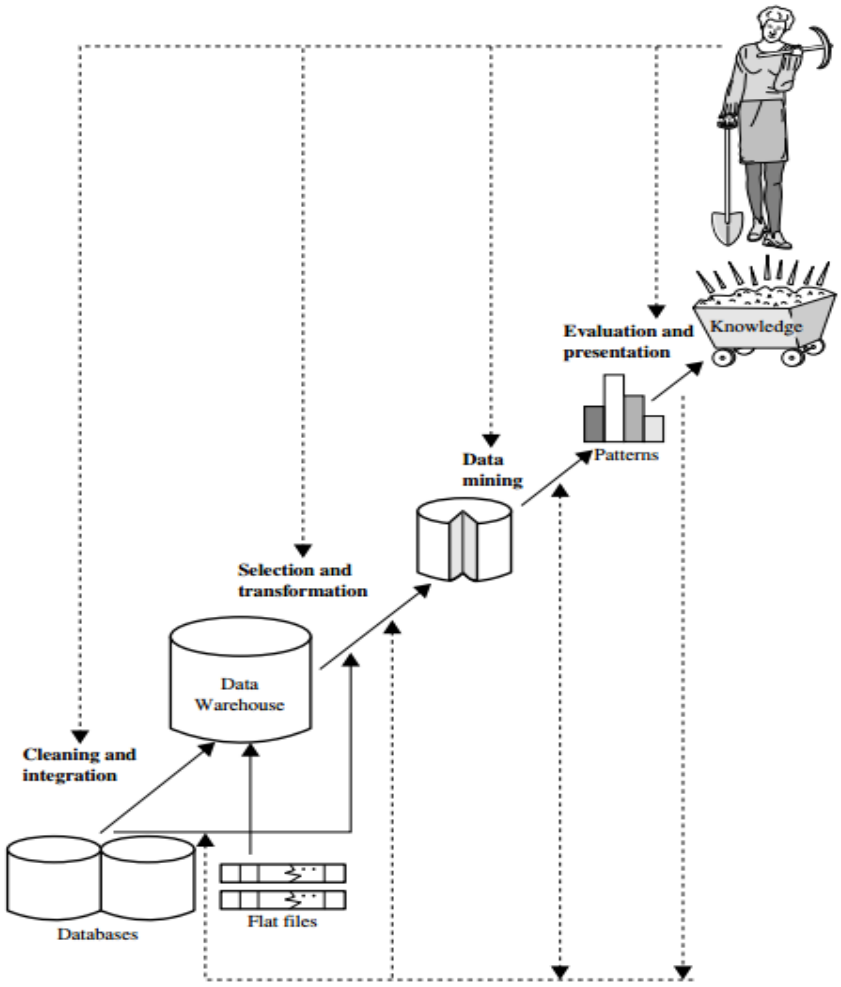


Ilustración 1. Esquema del proceso KDD.(Han, Kamber, & Pei, 2012)

En nuestro trabajo vamos a centrarnos en el sexto paso del KDD como antes hemos dicho, este paso tiene como objetivo comprender el contenido de una base de datos, es decir, esta tecnología pretende encontrar patrones repetitivos, tendencias o reglas para explicar el comportamiento de los datos en un contexto determinado. Estos datos son la materia prima bruta y pasan a ser información cuando el usuario les atribuye un significado especial empleando técnicas de diversas áreas como la Estadística, el Procesamiento Masivo, la Computación Gráfica o la Inteligencia Artificial.

(Marcel, 2014)

Éstas nuevas técnicas han dado lugar a una sustitución paulatina del análisis de datos *dirigido a verificación* por un enfoque de análisis de datos *dirigido al descubrimiento del conocimiento*. La diferencia fundamental entre uno y otro método de análisis se encuentra en que en el último se descubre información sin necesidad de formular previamente una hipótesis.

En la década de los 80’s se produjeron las primeras investigaciones sobre la minería de datos; el avance en la informática, el desarrollo de la inteligencia artificial y el aprendizaje automático ayudaron al impulso de la minería.

Actualmente el valor de la información ha crecido hasta convertirse en un activo estratégico para la competitividad de una empresa, los directivos necesitan obtener una visión más completa y detallada de su negocio, y buscar datos de sus operaciones cotidianas que se salen de los rangos normales, para así poder identificar posibles clientes, puntos de ventas, fraudes y otros muchos aspectos. Es una herramienta que nos ayuda a analizar todos los datos de una empresa.

1.2 Etapas del DM

El DM consta de varias etapas y en cada etapa será necesario el uso de una determinada área antes mencionada. **Las cuatro etapas** más significativas son:

1. Determinar los objetivos→Debido al amplio campo de posibilidades que el DM ofrece hay que establecerse unos objetivos claros y precisos.

2. Pre-procesamiento de los datos→ Es la etapa de mayor peso de las 4, puede consumir en torno al 70% del tiempo/esfuerzo total de cualquier proyecto sobre DM ya que es en la que la base de datos comienza a pasar de materia prima a "información" ya que se agrupan datos, eliminan datos irrelevantes, selecciona lo más importante… Se estudia la calidad de los datos y determinación de las operaciones de minería que se le pueden aplicar.

3. Determinación del modelo→ Para empezar, normalmente se emplean herramientas estadísticas para tantear un poco los datos. Más adelante, para afinar más, se suelen emplear diagramas o gráficas con los que se obtiene una primera aproximación. Según los objetivos establecidos en la primera etapa utilizaremos unos algoritmos u otros.

4. Análisis de resultados→Verificar que los resultados obtenidos son coherentes con la ayuda de las herramientas estadísticas y las gráficas. Y ver si con nuestro trabajo el cliente podrá obtener información antes desconocida.

Cualquier trabajo sobre minería de datos seguirá estas cuatro etapas en este orden, aunque hay que destacar que es un proceso muy iterativo y que será necesario volver a etapas anteriores para así obtener resultados mucho mejores.

1.3 Ventajas

Las principales ventajas que pueden justificar el uso de la minería de datos son:

- La información obtenida ayuda a los usuarios a elegir cursos de acción y a definir estrategias.

- Permite descubrir relaciones que no se conocían anteriormente.

- Puede trabajar siguiendo los mismos criterios con grandes cantidades de información histórica.

- El proceso de búsqueda puede ser realizado por herramientas que automáticamente buscan patrones.

- Enormes bases de datos pueden ser analizadas.

- Podemos verificar si los modelos obtenidos son válidos, gracias a técnicas auxiliares.

- Puede llegar a ahorrar grandes cantidades de dinero a las empresas y abrirles nuevas oportunidades de negocio.

1.4 Inconvenientes

A pesar de las ventajas descritas anteriormente, existen algunos inconvenientes a tener en cuenta cuando nos planteamos el uso de la minería de datos. Estos inconvenientes son los siguientes:

- El tiempo de respuesta es un gran inconveniente, ya que hay veces que es necesario procesar grandes volúmenes de datos lo que implica grandes tiempos de proceso para conseguir un modelo valido y hay problemas que requieren una respuesta en tiempo real.

- El preprocesamiento de datos puede llegar a ser demasiado costoso.

- No está asegurada la obtención de un modelo válido.

1.5 Campos de aplicación

El campo de aplicación del DM es muy amplio. A continuación, se describen varios ejemplos donde se ha visto involucrado el data mining. Se han seleccionado casos de diversos campos y con objetivos muy dispares para así observar con claridad su potencial. Principalmente se han usado árboles y reglas de decisión, reglas de asociación, redes neuronales, redes bayesianas, conjuntos aproximados (rough sets), algoritmos de agrupación (clustering), máquinas de soporte vectorial, algoritmos genéticos y lógica difusa.

- *Análisis de datos financieros*: se emplea tanto en el sector bancario como en el de las finanzas. Se pretende asegurar que es posible practicar análisis sistemáticos en condiciones avanzadas y con un alto grado de fiabilidad. Algunos de los ejemplos más claros dentro de éste ámbito son:

    - Diseñar y construir almacenes de datos para el análisis multidimensional de estos.
    - Predecir el pago de préstamos y analizar las políticas de crédito de los clientes.
    - Clasificar y agrupar a los clientes para crear ofertas personalizadas según las características de cada uno.
    - Detectar el blanqueamiento de dinero y otros fraudes financieros.

- *Industria minorista*: se recogen grandes cantidades de datos provenientes de las ventas, historiales de compra de los clientes o el transporte de mercancías. Los datos recogidos se expanden rápidamente debido al incremento de la facilidad, disponibilidad y popularidad de la web y las transacciones realizadas a través de Internet. Con la minería de datos aplicada a la industria minorista se ayuda a identificar patrones de compra de los clientes y a controlar las tendencias de los

mismos. De este modo, las compañías están en condiciones de ofrecer una mejor calidad de servicio al cliente, aumentando su satisfacción y facilitando su retención. Entre las aplicaciones a las que nos estamos refiriendo podríamos destacar las siguientes:

- Análisis multidimensional de las ventas, los clientes (edad, sexo…), la fecha, el clima y la región.
- Las referencias cruzadas de artículos.
- Análisis de la eficacia de las campañas de ventas.
- La recomendación personalizada de los productos.

- *Industria de las telecomunicaciones*: en el sector de las telecomunicaciones estos datos son especialmente importantes para alcanzar una buena comprensión del negocio. Con la minería de datos y sus aplicaciones específicamente diseñadas para éste área se obtiene una gran ayuda en la identificación de los patrones de telecomunicaciones, facilitando mucho la detección de actividades fraudulentas y posibilitando el hacer un uso óptimo de los recursos con la consiguiente mejora en la calidad de servicio. Entre las más ventajosas están:

  - Análisis multidimensional de datos de telecomunicaciones.
  - Análisis de patrones fraudulentos para adelantarnos a posibles casos.
  - Identificar patrones inusuales, hábitos y tendencias.
  - Asociación multidimensional y análisis de patrones secuenciales.

- *Análisis de datos biológicos*: el campo de la biología es uno de los que más beneficios ha obtenido del avance de la tecnología. La genómica, la proteómica, la genómica funcional y la minería de datos aplicada a la investigación de los seres vivos son algunos ejemplos, sin olvidar la bioinformática. Las aportaciones más importantes de la minería de datos para el análisis de datos biológicos son:

  - Integración semántica de las bases de datos genómicos y proteómicos heterogéneos distribuidos.
  - Alineamiento, indexación, búsqueda de semejanzas y análisis comparativo de múltiples secuencias de nucleótidos.
  - Descubrimiento de patrones y análisis de redes genéticas.
  - Identificación de patrones de proteínas estructurales.

- En la *medicina* para identificar relaciones en el suministro de un fármaco sobre otro fármaco, para relacionar enfermedades y fármaco, para agrupar pacientes…

- En el *sector agropecuario* como instrumento para identificar posibles plagas sobre la fruta, para prever campañas, para analizar proveedores y posibles compradores…

- *Detección de fraudes en tarjetas de crédito*: el Falcon Fraud Manager (FFM) es un sistema inteligente que nació debido a las grandes pérdidas que las instituciones financieras experimentaron en 2001, llegándose a perder más de 2.000 millones de dólares estadounidenses por el fraude con tarjetas de crédito y débito. El FFM examina transacciones, propietarios de tarjetas y datos financieros para detectar y mitigar fraudes. El sistema ha ido evolucionando y se le ha ido incorporando

funcionalidades de análisis en las tarjetas comerciales, de combustibles y de débitos.

- *En la gestión gubernamental y las ciencias sociales*, permite identificar patrones socioeconómicos, agrupar políticas, analizar el comportamiento de indicadores sociales…

- *Detección de terroristas*: El FBI anunció en julio de 2002 que iban a empezar a introducirse en la gran cantidad de datos comerciales referentes a los hábitos y preferencias de compras de los consumidores, con el objetivo de descubrir potenciales terroristas antes de que ejecuten una acción. Se ha llegado a asegurar que el FBI con esta información uniría todas las bases de datos mediante el número de la Seguridad Social y lograría saber si una persona fuma, que talla y tipo de ropa usa, su registro de arrestos, salario, altura, peso o si tiene abiertas cuentas bancarias, entre otros.

- *Predicción de la audiencia televisiva*: la BBC del Reino Unido emplea la minería de datos para predecir el tamaño de las audiencias televisivas para un determinado programa, así como para definir el mejor horario. El sistema emplea redes neuronales y árboles de decisión aplicados a datos históricos de la cadena.

- En el *deporte* puede emplearse para prevenir lesiones: el AC Milán utiliza un sistema inteligente para prevenir lesiones. El sistema está basado en redes neuronales y optimiza el acondicionamiento de cada atleta. Esto ayudará a la hora de realizar un determinado fichaje o a alertar a los servicios médicos del riesgo que presenta un determinado jugador a las lesiones. El sistema tiene clasificado a los jugadores según rendimiento, alimentación y respuesta a estímulos externos, que se obtienen y analizan cada quince días.

- En la *educación* puede servir de gran ayuda para clasificaciones y diagnósticos de estudiantes, para realizar planes de enseñanza según las capacidades de los estudiantes, descubrimiento de nuevas guías pedagógicas, análisis de profesores…

- *Detección de fraude en el consumo eléctrico y de agua*: esto se consigue estableciendo patrones que responden a formatos de distinta índole, no son sólo numéricos o de fechas. Son patrones complejos como un gran consumo en horas extrañas y esto no se detecta si no utilizamos todo el potencial que nos puede dar el "data mining". Lo óptimo es conseguir automatizar el proceso para así detectar el fraude con facilidad una vez puesta en marcha la maquinaria. Con la aplicación automatizada de algoritmos se detectan con facilidad patrones en los datos que hacen que esta técnica sea mucho más eficiente que el análisis dirigido a la verificación cuando se trabaja con datos procedentes de fuentes de una gran cantidad de datos y de una complejidad elevada. Dichas técnicas, al ser emergentes, se encuentran en constante cambio debido al resultado de la colaboración entre diversos campos de investigación. Una vez alcanzado el objetivo, es recomendable construir modelos predictivos para evitar que se produzca el fraude.

(Molina Félix , 2014)(Blog sobre Bussiness Intelligence,2016)

1.6 Técnicas

Son muchas las técnicas existentes para llevar a cabo una investigación sobre minería de datos. Una clasificación inicial de las técnicas de minería de datos diferencia entre técnicas predictivas, técnicas descriptivas y técnicas auxiliares.

Las técnicas predictivas o de aprendizaje supervisado se basan en el entrenamiento de un modelo o método por medio de diferentes datos para poder predecir una determinada variable partiendo de estos mismos datos. Esta manera de trabajar se desarrolla en dos fases: entrenamiento (construcción de un modelo usando un subconjunto de datos como etiqueta, llamamos etiqueta al atributo del que vamos a predecir su valor) y prueba (prueba del modelo sobre el resto de los datos). Podemos clasificar como técnicas predictivas las siguientes:

- **Regresión lineal:** Método estadístico que nos permite establecer una relación matemática entre un conjunto de variables x_1, $x_2 \ldots x_k$ y una variable dependiente y. Se utiliza en aquellos casos en los que no se puede controlar los valores de las variables independientes. Es la técnica más empleada para comparar datos. Es rápida y eficaz pero no es válida en espacios multidimensionales donde se traten más de dos variables.

- **Análisis de la varianza y la covarianza:** Es una colección de modelos estadísticos, este análisis parte de los conceptos de la regresión lineal y permite eliminar la heterogeneidad causada en la variable de interés por la influencia de una o más variables cuantitativas.

- **Series temporales:** Técnica basada en la sucesión de observaciones de una variable tomada en varios instantes de tiempo. Interesa observar los cambios en esa variable a lo largo del tiempo y poder predecir valores futuros.

- **Métodos bayesianos:** Modelo estadístico en el que las observaciones se emplean para actualizar la probabilidad de que una hipótesis sea cierta o no. Es un método que necesita información anterior para determinar la distribución de probabilidad, además se caracteriza por el uso constante del teorema de Bayes.

- **Algoritmos genéticos:** Se inspiran en la evolución biológica. Se caracterizan por hacer evolucionar una población de datos sometiéndolas a acciones aleatoria semejantes y seleccionar cuáles son los individuos más adaptados y cuáles menos aptos.

- **Análisis discriminante:**: técnica estadística multivariante que ayuda a identificar las características que diferencian a dos o más grupos y a crear una función capaz de distinguir con la mayor precisión posible a los miembros de uno u otro grupo. Es capaz de decirnos que variables permiten diferenciar a los grupos y cuántas de estas variables son necesarias para alcanzar la mejor clasificación posible. Es una técnica muy parecida a la que se va a describir más adelante (Clustering) con la diferencia de que en ésta conocemos el número de datos y los datos que hay en cada grupo.

  Podemos considerarlo como un análisis de regresión donde la variable dependiente es categórica y ésta categoría es la componente diferencial de cada grupo, mientras que las variables independientes son continuas y determinan a que grupos pertenecen los objetos.

- **Análisis de componentes principales (ACP):** técnica que se encarga de sintetizar la información y reduce la dimensión de las observaciones. Ante un banco de datos de muchas variables, se pretende reducirlas perdiendo la menor información posible. El resultado de esta reducción será una combinación lineal de las variables originales y serán independientes entre sí.

  Fases de un ACP:

  1. Análisis de la matriz de correlaciones.
  2. Selección de factores.
  3. Análisis de la matriz factorial.
  4. Interpretación de los factores.
  5. Cálculo de las puntuaciones factoriales.

- **Árboles de decisión:** Son diagramas que representan de forma secuencial condiciones y acciones. Destacan por su sencillez y por poder utilizarse en distintas áreas. Además, cualquier persona que no tenga grandes conocimientos puede entenderlo fácilmente. El objetivo es crear un modelo que predice el valor de una variable de destino en función de diversas variables de entrada. También se puede describir como combinación de técnicas matemáticas, estadísticas y computacionales para ayudar a la descripción y la categorización de un conjunto de datos. El mecanismo es elegir un atributo como raíz y desarrollar el árbol según las variables más significativas.

  Todos los árboles de decisión son similares y están compuestos por los mismos componentes, estos son los cuatro componentes requeridos en cualquier árbol de decisión:

  - Alternativas de decisión en cada punto de decisión.
  - Eventos posibles tras cada alternativa de decisión.
  - Probabilidades de que ocurran cada evento posible.
  - Resultados de las interacciones entre las alternativas de decisión y los eventos.

- **Redes neuronales (RRNN)**: Esta herramienta emplea un conjunto de elementos de procesamiento de información altamente interconectados capaces de aprender con los datos de los que se abastece, tienen un cierto grado de "inteligencia". Las RRNN simulan el comportamiento del sistema nervioso, por lo que reproduce algunas actividades del cerebro. Las características comunes entre las RRNN artificiales y las RRNN biológicas son el paralelismo masivo, la respuesta no lineal de las neuronas frente a la información recibida y el procesamiento de los datos recibidos a través de capas de neuronas. Estas capas de las que hablamos son tres: capa de entrada (recibe la información del exterior), capas ocultas (procesan la información internamente, no tienen ninguna conexión con el exterior) y capa de salida (obtiene la respuesta de la red dada por las capas ocultas y la transfiere al exterior).

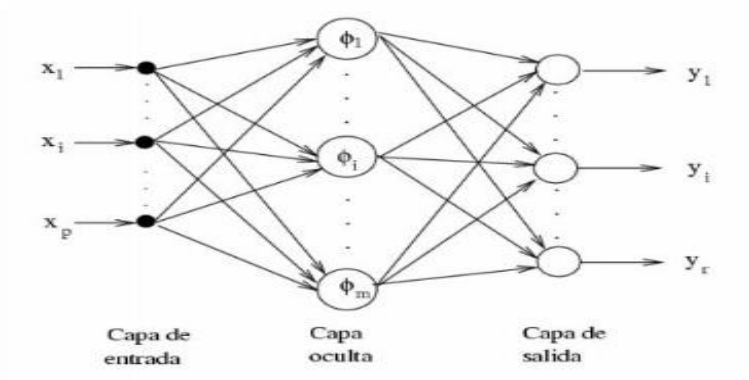


Ilustración 2. Capas que forman una red neuronal.(Redes de Neuronas Artificiales, 2012)

En la imagen anterior podemos ver un esquema de una red neuronal y sus capas. También se dividen las RRNN según el flujo de la información dentro de la misma; puede ser una Red Alimentada hacia delante o "feedforward" (la información va siempre de las primeras capas a las últimas sin opción a que ésta retroceda) o una Red Retroalimentada o "feedback" (aquella en la que la información puede volver a las capas anteriores y reprocesarla).

La respuesta de esta herramienta a los datos introducidos viene dada por tres funciones:

- Función de propagación: Consiste en el sumatorio de cada entrada por el peso de su interconexión.
- Función activación: Puede existir o no, dependiendo de la entrada.
- Función de transferencia: Se aplica al valor devuelto por la función de activación. Se utiliza para acotar la salida de la neurona.

- **Máquinas de soporte vectorial (SVM):** Se usa para problemas supervisados de clasificación, esta técnica está relacionada con la clasificación y la regresión. Se basa en un conjunto de algoritmos los cuales construyen hiperplanos en un espacio de dimensionalidad muy alta, esto permite una separación de clases y una clasificación correcta. Potencialmente es capaz de generar muy buenos modelos predictivos.

Estas tres últimas técnicas son de clasificación que pueden extraer perfiles de comportamientos, cuyo objetivo es construir un modelo que permita clasificar cualquier nuevo dato.

Mientras que las técnicas descriptivas no se asignan ningún papel determinado a las variables. También son llamados métodos simétricos, no supervisados o indirectos. Estos grupos con los que se trabaja no son conocidos con anterioridad, podemos encontrarnos con que las variables estén conectadas entre sí de acuerdo a vínculos desconocidos al principio. Esta opción es la elegida si la aplicación no es lo suficientemente madura como para poder deducir una solución predictiva fiable. No se utilizan datos históricos en esta segunda opción. Las técnicas descriptivas son:

- **Clustering (Análisis de conglomerados**): El funcionamiento de esta técnica se basa en clasificar una muestra de entidades en un número pequeño de grupos de forma que los elementos que formen un mismo grupo sean muy parecidos entre sí y muy distintos del resto de grupos. A diferencia del "Análisis Discriminante" (explicada anteriormente), en el Clustering no se conoce el número y la composición de los grupos. Las distancias más comúnmente utilizadas es la distancia euclidea, manhattan o chebyshev. También se utiliza como paso previo a otras técnicas en la minería de datos. Algunos algoritmos de clustering son:

    - K-means: En este se define el número de clusters que se desean obtener, partir de ahí se forman los centros y se agrupan los datos.
    - X-means: Se elige un límite inferior y otro superior y el algoritmo es capaz de definir el número de grupos óptimos.
    - Cobweb: Realiza agrupaciones instancia a instancia. Va formando un árbol de clasificación.
    - EM: Se puede utilizar para segmentar conjunto de datos, está clasificado como un clustering probabilístico.

- **Asociación:** Se busca encontrar ítems que aparezcan juntos en transacciones de un determinado conjunto de datos. Para encontrar reglas de asociación hay que considerar todas las posibles combinaciones para que haya una consecuencia. Así, se establecen las reglas que indican dependencias entre los ítems de dicho conjunto de datos.

- **Dependencia:**Consiste en buscar un modelo que encuentre dependencias significativas entre el conjunto de datos. Estas dependencias pueden ser usadas para predecir valores futuros.

- **Reducción de la dimensión:** Tiene como objetivo reducir el número de variables aleatorias. Es utilizada cuando se tienen muchas dimensiones (atributos) con respecto al número de instancias, ya que pueden existir muchos grados de libertad.

- **Análisis exploratorio:** La finalidad es conseguir un entendimiento de los datos y de sus relaciones. Permite organizar y preparar los datos, identificar casos atípicos y evaluar datos ausentes.

- **Escalamiento Multidimensional:** Representa en un espacio geométrico de pocas dimensiones las proximidades existentes entre los datos. Puede utilizarse como alternativa o como complemento a otras técnicas.

La técnica de clustering y la de segmentación las incluimos en técnicas descriptivas aunque también son técnicas de clasificación. Es más, son técnicas de clasificación *post hoc* porque realizan la clasificación sin especificación previa de los grupos. El análisis discriminante, los árboles de decisión y las redes neuronales son técnicas de clasificación *ad hoc*, debido a que clasifican las observaciones dentro de grupos previamente definidos.

Tanto las técnicas descriptivas como las técnicas predictivas buscan el descubrimiento del conocimiento de un conjunto de datos. Pero también existen otras técnicas denominadas técnicas auxiliares que son herramientas de apoyo, las cuales están enfocadas más a la verificación. Estas técnicas de verificación o auxiliares son:

- **Proceso Analítico de Transacciones:** También llamado OLAP, su objetivo es agilizar la consulta de grandes cantidades de datos, para ello utiliza estructuras multidimensionales o cubos OLAP (donde los datos son almacenados en un vector multidimensional). Es un proceso distinto a la minería de datos, ya que no busca la creación de patrones a partir de los datos, si no que verifica estos patrones. OLAP y minería de datos son herramientas diferentes pero que se complementan.

- **SQL y herramientas de consulta:** Permiten aplicar el modelo a nuevos datos, obtener un resumen estadístico de los datos, pueden realizar consultas de contenido, de predicción, de detalles y de definición de datos.

- **Reporting:** Herramienta que permite crear, implementar y administrar informes donde se reflejan el análisis de los resultados de la minería de datos.

(Pérez López & Santín González, 2007)

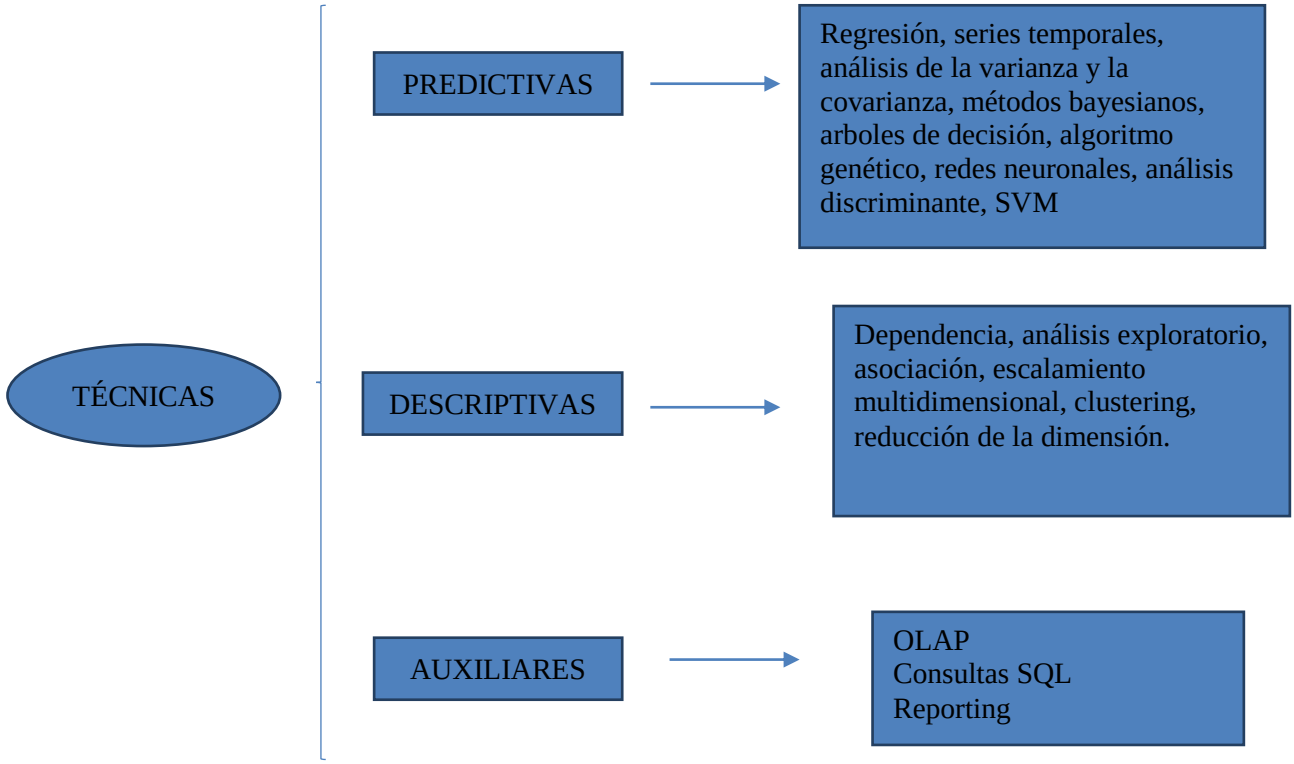


Tabla 1. Clasificación de las técnicas de la minería de datos.(Elaboración propia, 2016)

1.7 Tipos de datos

Un aspecto importante a conocer sobre la minería de datos es saber cuáles son los tipos de datos.“La minería de datos se aplica a todo tipo de datos imaginable: desde datos numéricos a

imágenes de satélite, mamografías, música, archivos de ordenador, imágenes, etc. Podemos decir que "cualquier cosa" constituye un dato. Por tanto, la minería de datos tiene infinitas aplicaciones".

(Krall , 2006)

A pesar de que la minería se puede aplicar a cualquier tipo de dato, es reseñable que hay técnicas de minería de datos que no permiten trabajar con cualquier dato. Los datos están contenidos en bases de datos, las cuales dependiendo de la información que posean pueden ser:

- **Bases de datos relacionales:** Se compone de tablas, consiste en filas y columnas, donde cada columna almacena información sobre un atributo y cada fila contiene una instancia o tupla.

- **Bases de datos transaccionales:** Conjunto de datos que representan transacciones cuyo objetivo es enviar y recibir datos a grandes velocidades. La redundancia de información no es un problema, sin embargo, este tipo de bases de datos son poco comunes.

- **Bases de datos espaciales:** Además de información en alguna de las formas anteriores, también maneja información geográfica como mapas.

- **Bases de datos temporales:**Son aquellas donde los atributos están relacionados con el tiempo.

- **Bases de datos documentales:** Se caracterizan por dar una descripción de los objetos, desde una palabra clave hasta resúmenes.

- **Bases de datos multimedia:** Imágenes, audios y videos son almacenados en este tipo de base de datos.

- **World Wide Web:**"La World Wide Web, WWW, es el repositorio de información más grande y diverso de los existentes en la actualidad".

(Hasperué, 2013)

1.8 Softwares de minería de datos

Identificar los patrones de comportamiento y de relación entre los datos es una función que puede llevar mucho tiempo, pero realizar esto puede ser más fácil si tiene un objetivo establecido y se está trabajando con las variables necesarias.

A continuación, se van a describir las herramientas de uso más comunes y potentes de las que elegiremos una con la que llevar a cabo nuestro proyecto y sobre la que hablaremos más extensamente más adelante para dar una explicación más detallada:

- **Orange:** Software basado en componentes que cuenta con un fácil, potente, rápido y versátil interfaz de programación visual para el análisis exploratorio de datos y visualización. Es un software que puede obtenerse de forma gratuita. Fue desarrollado e implementado por la Universidad de Ljubljana. Cuenta con la licencia GNU o GPL (General Public License), muy empleada para los softwares libres ya que garantiza a los usuarios la libertad para compilar, estudiar, compartir y modificar el software. Las características principales engloban las funcionalidades básicas de un software de este tipo: lectura de datos, generación de tablas de datos, selección de características, entrenamiento de algoritmos y visualización de gráficas. Está escrito en C++ y Python.

  (Orange, s.f.)

- **RapidMiner:**antes llamado YALE (Yet Another Learning Environment). Ofrece más de 500 operadores para los procedimientos de máquina de aprendizaje. Está escrito en un lenguaje de programación Java y proporciona una interfaz gráfica para diseñar y ejecutar flujos de trabajos de análisis. Proporciona esquemas y algoritmos Weka y R scripts. Se distribuye bajo licencia AGPL de código abierto. Este software cuenta con una versión gratuita que se limita a un procesador lógico y a 10.000 filas de datos. Otra versión disponible para estudiantes que dispone gratuitamente está limitada por un procesador lógico pero, sin embargo, el número de filas es ilimitado. Las versiones de pago van desde los 2.500$ al año, con 100.000 filas y dos procesadores lógicos, hasta los 10.000$ al año, con el número de filas y de procesadores lógicos ilimitados.

  (Rapidminer, s.f.)

- **Weka (Waikato Environment for Knowledge Analysis):** Conocido software para máquinas de aprendizaje automático. Desarrollado bajo licencia GPL y de software libre. Programado en Java, contiene una gran cantidad de técnicas para modelado y procesamiento de datos. Tiene una interfaz de usuario muy sencilla y funciona en la mayoría de plataformas. Éste software será explicado más delante en mayor detalle ya que será el empleado para nuestro proyecto.

  (Weka 3: Data Mining Software in Java, s.f)

- **JHepWork:** Diseñado para científicos, ingenieros y estudiantes, de código abierto, para el análisis de datos. Contiene bibliotecas científicas numéricas implementadas en Java. Se basa en Jython (parecido al Phyton) un lenguaje de programación de alto nivel. Cuenta con la licencia GNU también, por lo que su descarga es posible a ningún coste. Este software presenta ciertas restricciones para el uso comercial dado que algunos tipos de archivos tienen la licencia para un uso no comercial. Con este programa nos es posible visualizar gráficas, histogramas, curvas de nivel, redes neuronales…

  (jwork.org, s.f.)

- **Knime:** Plataforma de código abierto con la licencia GNU y de fácil uso, que ofrece a los usuarios la posibilidad de crear de forma visual flujos de datos. Está escrito en Java y basado en Eclipse.Se comenzó a desarrollar en enero de 2004 por un equipo de ingenieros de la Universidad de Costanza como un producto de uso propio. Pero desde 2006 se ha venido empleando en investigaciones farmacéuticas mayoritariamente, además de en otras áreas como en la gestión de relaciones con los clientes (CRM), inteligencia en el negocio y en análisis de datos financieros, llegando en 2012 a tener más de 15.000 usuarios.

  (Knime, s.f.)

- **IBM SPSS Modeler:** Anteriormente llamado Clementine, está diseñada teniendo en cuenta los usuarios empresariales, de manera que no es preciso ser un experto en minería de datos. Es la herramienta más avanzada del mercado, que posee una interfaz simple y sencilla. Se caracteriza por tener una visualización interactiva y por sus numerosas técnicas de modelado. Sus siglas significan *Statistical Product and Service Solutions*, es un software privativo. Que un software es privativo significa que nos e puede acceder a su código fuente de forma libre, éste sólo está a disposición del desarrollador y no es posible modificarlo ni adaptarlo a unas determinadas características libremente. Este programa consta de un módulo base y multitud de módulos anexos que se van implementando y actualizando constantemente y que se pueden adquirir comprándolos por separado. Este software es vendido en dos versiones por IBM:

  - SPSS Modeles Proffessional: empleado para datos estructurados, bases de datos y sistemas de negocio inteligente.

  - SPSS Modeler Premium: incluye todo lo anterior, pero además añade el análisis de textos, el análisis de entidades y el análisis de redes sociales.

  (IBM Analytics, s.f.)

1.9 Extensiones del data mining

La minería de datos engloba tantas herramientas y tanta variedad de campos que no es de extrañar que le hayan surgido variantes muy similares, pero con alguna que otra característica particular. Se pueden diferenciar dos grandes variantes: web mining y text mining:

1.9.1 Web mining

El Web mining consiste en aplicar las técnicas del data mining a documentos y servicios de la web ya que todos los sitios visitados por un usuario en Internet dejan una huella digital que los servidores almacenan en una bitácora de accesos (log). Estos logs son analizados y procesados para buscar patrones que faciliten información significativa como, por ejemplo, cómo es la navegación

de un cliente en el proceso de una compra. Los accesos totales por dominio, horarios de más actividad en la web y la cantidad de visitas diaria, entre otros, son datos que se registran en herramientas estadísticas que ayudan a complementar todo el proceso de análisis de esta extensión. Recientemente ha aparecido un término nuevo como instancia del Web mining, el "multimedia Web mining", un nombre que busca diferenciar los tipos de datos con los que se trabaja en Internet: textos, imagen, vídeo o metadatos.

Normalmente, el Web mining se clasifica en tres grandes grupos de extracción de conocimiento:

1. Web content mining (minería de contenido web). Es la parte encargada de la extracción de conocimiento del contenido de documentos o sus descripciones. Otras tareas de esta parte serían la localización de patrones en el texto de documentos, el descubrimiento del recurso basado en conceptos de indexación o la tecnología que se basa en agentes.

2. Web structure mining (minería de estructura web). Es la minería que se encarga de la estructura de la web, es decir, identifica la relación entre páginas que están vinculadas por un enlace o por información.

3. Web usage mining (minería de uso web). Basado en la extracción de modelos interesantes usando los logs de los accesos a la web.

Para que se entienda mejor el tipo de resultados que se pueden obtener de esta herramienta se va a mostrar un ejemplo práctico y, así, conseguir que no queden dudas sobre lo que nos ofrece esta variante de la minería de datos: si el sistema detecta que un alto porcentaje de los clientes que hacen una compra online en */adquisición/productoA.html* también compraron en */adquisición/productoB.html* semanas posteriores. Esto nos está marcando la opción de ofrecer un pack que incluya ambos productos ahorrando así los gastos de envío del segundo producto.

1.9.2 Text mining

Es sabido que un elevado porcentaje de la información de las compañías está almacenada en forma de documentos. He aquí donde aparece el text mining o minería de texto. El text mining incluye técnicas como la categorización de textos, el procesamiento de lenguaje natural, la extracción y recuperación de la información o el aprendizaje automático. No se debe confundir el text mining con la recuperación de la información (Information Retrieval o IR). Ésta última consiste en la recuperación automática de documentos destacados mediante indexación de textos, clasificación, categorización… Normalmente utiliza palabras claves cuando se busca una información importante en un determinado texto. En cambio, el text mining se encarga de examinar una colección de documentos y descubrir información no contenida en ningún documento concreto de la colección.

Una aplicación muy conocida es narrada en Hearst (1999). Donde se relata cómo mediante cadenas de implicaciones causales dentro de la literatura médica pueden llevarnos a hipótesis para enfermedades poco frecuentes. Este hallazgo tiene su importancia debido a que los expertos sólo pueden leer una pequeña parte de lo que se publica en su campo obviando los avances que se pueden estar dando en otros campos. En este caso, se investigó sobre la migraña y se extrajeron varias piezas de evidencia a partir de artículos de la literatura biomédica. Algunas de las evidencias detectadas fueron:

- El estrés está ligado a la migraña.
- El estrés puede producir pérdida de magnesio.
- Los bloqueadores de canales de calcio previenen la migraña.
- El magnesio es un bloqueador del canal de calcio.

Estas claves dejan ver que la deficiencia de magnesio puede representar un importante papel en algunos tipos de migraña, una teoría que no existía en la literatura y que, con ésta investigación, se encontró.

Nuestra capacidad para almacenar datos ha aumentado en los últimos tiempos a velocidad de vértigo. Sin embargo, nuestra capacidad para procesarlos y utilizarlos no ha ido a la par. Por este motivo, la minería de datos y sus alternativas se presenta como una gran herramienta de apoyo para explorar, analizar, comprender y aplicar el conocimiento obtenido usando grandes volúmenes de datos.

En el ámbito comercial resulta muy interesante encontrar patrones ocultos de consumo de los clientes, predecir el comportamiento futuro de los clientes basándose en datos históricos de clientes que presentaron una misma forma de proceder.

(Hearst, 1999)

2 Idea inicial del proyecto

Como se ha dicho anteriormente, el fraude ha existido siempre, por lo que las empresas buscan modernizar sus métodos de detección de este tipo de acciones para conseguir automatizar esto dentro de las posibilidades que poseen. Es aquí donde mi compañero y yo vimos la posibilidad de poder ayudar a la empresa Giahsa facilitándole este trabajo.

Este trabajo es un muy duro ya que implica tener muchos datos y mucha información que sea útil y haga que los algoritmos empleados logren sacar patrones de comportamiento que vislumbren los posibles casos de fraude. Giahsa mostraba mucho interés en este caso debido que nota un gran aumento de este tipo de acciones en su facturación y buscan pararlo de la forma más eficaz y rápida posible. Tras ponernos en contacto con la empresa, esta nos proporcionó un libro de Excel en el que venía toda la facturación de 2013, 2014 y 2015 junto a sus atributos (mes, año, puerta, piso, número, calle, facturación, identificador de usuario…). Tras un tiempo trabajando en el proyecto, nos dimos cuenta de que si no teníamos los datos de otros clientes fraudulentos que ya hubiesen sido detectados, no podíamos detectar a los nuevos porque no contábamos con unos patrones que introducirle al algoritmo. Debido a esto, nos volvimos a poner en contacto con la empresa y nos proporcionó otro libro de Excel con los clientes que tenían almacenados por haber cometido ya alguna irregularidad. Lógicamente, los nombres de los consumidores eran confidenciales, por lo que la empresa nos proporcionó un identificador para cada cliente; es así como diferenciábamos a cada uno.

Con todo esto, nos pusimos en marcha para ver que patrones éramos capaces de sacar para detectar operaciones anómalas o sospechosas de serlo.

2.1. Introducción

2.1.1 Tipos de fraude en la acometida del agua

Antes de analizar los métodos más utilizados por los usuarios para trucar las acometidas de agua, es necesario conocer que es una acometida y como son las estructuras de los contadores. Una acometida es un conjunto de tuberías y otros elementos que unen las conducciones de abastecimiento viarias con las instalaciones del inmueble. Debe estar dimensionada según las características de la instalación del inmueble, además debe tener una válvula de registro en el

acerado, en acceso público y lo más cerca posible de la parcela, la cual delimita el límite de competencias a efectos del mantenimiento y de la conservación.

Un contador es un aparato de gran precisión que permite contabilizar la cantidad de agua que pasa a través de él. El contador es propiedad de la compañía suministradora, quienes lo instalan y lo reponen. Los contadores se encuentran a la entrada de cada vivienda, siempre con acceso desde la vía pública, además están verificados y precintados para garantizar que la medida sea precisa. Los contadores están compuestos por un conjunto de relojes que indican los registros de volumen. En la mayoría de modelos podemos apreciar un indicador de consumo total, unos relojes que indican los metros cúbicos consumidos y un indicador de movimiento que registra el paso del agua. Cada contador cuenta con un número de identificación que proporciona el año de fabricación, el número de serie e información sobre la marca.

Giahsa utiliza dos tipos de contadores:

- Contador único: Cuando en el inmueble solo existe una vivienda y en suministros provisionales para obras.

- Batería de contadores divisionarios: Cuando exista más de una vivienda, es obligatorio un aparato de medida para cada una de ellas y los necesarios para los servicios comunes.

(Giahsa, 2012)

Además, esta empresa utiliza contadores con distintos calibres en función de la demanda de cada usuario. Estos calibres pueden ser de 13 mm, 15 mm y 20 mm.

Una vez explicado esto, ahora sí podemos analizar los métodos más utilizados por los usuarios para estafar a las compañías de agua. Un puente contador, un enganche justo antes del instrumento de control de abastecimiento, tomas clandestinas o acometidas históricas sin registrar son formas de engañar por parte del consumidor.

El método de fraude más común es el de la instalación de un puente paralelo al contador con un trozo de tubo y la colocación de una válvula de cierre y otra de retorno. También hay muchos usuarios que quitan el precinto y colocan el contador al revés para que cuente hacia atrás el consumo.

Abundan los intentos de trucar el funcionamiento, como perforar la base del instrumento de medida e introducir una pequeña barra para que corra más lento. Otra forma es poner un alambre de acero clavado en el ventilador del tubo que se enlaza con la tubería general. Algún usuario ha introducido una radiografía en el contador hasta que toque la rueda que contabiliza el consumo y hace que esta se mueva más lento. Es muy popular la utilización de imanes para ralentizar la velocidad de giro de la rueda.

2.1.2 ¿Por qué aplicar minería de datos al fraude?

Las herramientas de minería de datos son muy buenas para su clasificación, para tratar de entender por qué un grupo de personas es diferente de otro. Es una buena herramienta para poder trabajar con la cantidad de registros que se tienen de los usuarios. Además, si se disponen de usuarios que ya han sido identificados anteriormente como fraudulentos, a partir de una técnica

predictiva se puede comparar con el resto de usuarios y comprobar que usuarios se asemejan a las características del usuario fraudulento.

2.1.3 Alternativas a la minería de datos

A parte de la minería de datos hay otras técnicas que permiten identificar posibles casos de fraude como:

- Modelos de probabilidad, los cuales son una clase de modelos que la gente utilizaba en la antigüedad cuando los datos no estaban disponibles en abundancia. Con la cantidad de datos que hay en la actualidad, estos modelos son más difíciles de utilizar, aunque se pueden combinar con la minería de datos.

- Personal cualificado y autorizado visita e inspecciona las instalaciones, observando si existe alguna anormalidad. Normalmente estas inspecciones periódicas se realizan por áreas y en cada área se inspecciona al azar algunos suministros, excepto en algunos casos donde se intuyen una alta probabilidad de fraude y se les realiza la inspección.

2.1.4 Antecedentes de proyectos similares

Antes de poner en marcha nuestro proyecto hemos creído conveniente consultar otros proyectos de temática parecida y ver que técnicas han empleado otras personas para hacernos una ligera idea de cómo afrontar nuestra tarea de minería de datos.

En Internet hemos encontrado infinidad de trabajos relacionados con la minería de datos, nosotros hemos elegido algunos sobre los que hablar aquí basándonos en las técnicas empleadas y la temática a tratar intentado que fueran lo mas similares a nuestro proyecto.

A. Minería de datos para la predicción de fraudes en tarjetas de crédito.

Como ya explicamos anteriormente en la parte de los “Campos de Aplicación”, la minería de datos se ha utilizado para la detección de fraudes en tarjetas de crédito y uno de los trabajos encontrados tratan este campo. Con el propósito de descubrir transacciones sospechosas, el autor de este trabajo decide hacer uso de los algoritmos de árboles de decisión (J48 exactamente) y reglas de asociación. Como podemos notar, la temática es muy parecida a la nuestra ya que ambas pretenden detectar posibles clientes fraudulentos a través de revisar el historial de estos y hallar patrones ocultos que todos los clientes de este tipo cumplan. Cabe destacar que el autor ha utilizado los árboles de decisión como técnica, pero indica que los algoritmos que se suelen usar para la detección de fraude son los propios árboles de decisión, las redes neuronales y los análisis bayesianos. A continuación, vamos a hablar un poco de lo que se encarga cada técnica empleada:

- Discretización: transformación de valores continuos en discretos.

- Normalización: preprocesamiento necesario en los datos en el que los valores estarán entre 0 y 1.

- Árboles de decisión: esta técnica es muy útil cuando unas distintas situaciones suceden de forma sucesiva y así se detectan los posibles comportamientos anómalos. El empleado por el autor de este trabajo, el J48, es uno de los más populares de WEKA.

- Reglas de asociación: agrupa según los atributos tengan información en común, por ello es necesario discretizar los datos anteriormente.

(Wanumen Silvaz, 2010)

B. Aplicación de Minería de Datos para la Detección de Anomalías: Un Caso de Estudio.

Otro proyecto del que nos pareció interesante hacer mención es uno de la Universidad de La Frontera (Chile) en el que el objetivo es que la empresa Aguas Araucanía S.A., que presta servicios de agua potable y de alcantarillado a la ciudad de Lautaro, detecte comportamientos anómalos en su red de abastecimiento. En este otro caso, las técnicas que se han utilizado son el clustering y la metodología CRISP-DM. Con estos avances la empresa fue capaz de reducir mucho el tiempo gastado en detectar los posibles fraudes. En este proyecto pudimos comprobar que para que la minería de datos sea una técnica satisfactoria es necesario emplear diferentes técnicas para que sea posible una óptima solución del problema. Los métodos clustering utilizados son:

- K-means: se encarga de dividir un conjunto de datos en un conjunto de K grupos pretendiendo optimizar el criterio de particionamiento elegido.

- COBWEB: es un algoritmo en el que se van agrupando las instancias una a una. Se basa en un árbol de clasificación, donde cada segmento está representado por las hojas y el conjunto de datos de entrada están englobados por el nodo raíz. No se recomienda esta técnica para la detección de registros fraudulentos ya que tiende a agrupar la mayoría de los registros en un solo segmento.

- Expectation-Maximization (EM): se utiliza para fragmentar conjunto de datos basándose en la obtención de la función de densidad de probabilidades a la que pertenecen los datos. Ésta técnica obtuvo resultados más valiosos que el K-means.

(Cravero Leal & Sepúlveda Cuevas , 2009)

C. Minería de datos aplicada a la detección de clientes con alta probabilidad de fraudes en sistemas de distribución.

Conforme encontramos más proyectos de temas parecidos íbamos descubriendo que las técnicas empleadas comenzaban a repetirse como hemos podido comprobar en este otro trabajo de la Universidad Tecnológica de Pereira de Colombia. En este otro, el objetivo era aplicar la minería de datos para detectar a clientes con alta probabilidad de fraude en sistemas de distribución. Aquí se vuelven a repetir las técnicas de aprendizaje no supervisado descritas en el caso anterior: K-means, EM Y COBWEB. La novedad de este proyecto es el uso de la máquina de soporte vectorial (SVM), una herramienta muy poderosa para clasificar datos y para localizar aquellos que pueden ser determinantes a la hora de llegar a la meta propuesta.

(Ríos Villegas& Uribe Aguirre, 2013)

D. Modelo de detección de fraude basado en el descubrimiento simbólico de reglas de clasificación extraídas de una red neuronal.

Este otro proyecto fue desarrollado en la Universidad Nacional de Colombia y el objetivo es el de crear e implementar una herramienta de colaboración que ayudara a los expertos en negocios a examinar y verificar con facilidad los resultados obtenidos para ayudar a la toma de decisiones pudiendo recurrir rápidamente a las decisiones tomadas con anterioridad comprobando la respuesta que se obtuvo según la decisión y comprobando que situaciones reflejaban un posible escenario fraudulento. Se emplearon dos técnicas ya nombradas en los artículos anteriores: los árboles de decisión y las redes neuronales. A través de la minería de datos puede desarrollarse una red neuronal de la que se extraen reglas de clasificación una vez esta red ha sido entrenada.

(Santamaría Ruíz, 2010)

E. Estrategia inteligente para la detección eficiente de clientes residenciales con condiciones fraudulentas de las empresas de servicio eléctrico.

Aquí también podemos encontrarnos con un proyecto análogo al nuestro en el que en lugar de buscar posibles fraudes en el consumo de agua los buscan en el consumo eléctrico. Este proyecto fue llevado a cabo en Venezuela en la Universidad Nacional Experimental Politécnica "Antonio José de Sucre". En este caso, como en el anterior se emplean las redes neuronales para detectar el fraude y se centra en la aplicación de esta técnica a los usuarios particulares (sector residencial), ya que el número de estos es mucho mayor que el de industrias y comercios.

(Lima & Vásquez, 2013)

2.2 Contacto con Giahsa

2.2.1 Giahsa

Giahsa es la empresa que nos ha facilitado los datos para llevar a cabo nuestra investigación para detectar el fraude en el consumo de agua. Ésta es una empresa onubense nacida por la necesidad de modernizar las infraestructuras de abastecimiento, saneamiento y depuración de la costa de esta provincia debido a que las redes existentes hasta entonces no eran capaces de soportar la alta demanda de un turismo cada vez mayor con los años. Las bases de la empresa se pusieron en 1989 cuando aunaron fuerzas los ayuntamientos de Lepe, Aljaraque, Ayamonte, Cartaya, Isla Cristina, Moguer, Punta Umbría y San Juan del Puerto para formar la Mancomunidad de Aguas Costa de Huelva y, junto a ésta, de su empresa pública de gestión, Giahsa.

La meta de Giahsa es la gestión de los servicios públicos adjudicados por la Mancomunidad y la gestión técnica de los mismos con eficacia y eficiencia en la administración de los recursos comprometidos. Equilibrar la calidad del servicio con el menor coste posible es la base fundamental de su actuación, como forma para lograr un sistema tarifario lógico y comparable para el nivel de servicios demandados por la Mancomunidad.

Podríamos estructurar la oferta de servicios de Giahsa de la siguiente manera:

- Gestión del ciclo integral del agua (abastecimiento, saneamiento y depuración).
- Recogida y tratamiento de Residuos Sólidos Urbanos (RSU).

- Otros servicios como proyectos y obras, telecontrol de instalaciones, procedimiento integrado de gestión, energías alternativas y cooperación al desarrollo.
- Nuevos servicios como establecer un plan de eficiencia energética.

La infraestructura se puede detallar de la siguiente forma:

- Abastecimiento: 361 instalaciones, de las cuales 18 son ETAPs (Estación de Tratamiento de Agua Potable), y una red de 1.830.122 metros.Saneamiento: 46 EDARs (Estación de Depuración de Aguas Residuales) y 1.354.985 metros de red.
- RSU: tres puntos limpios y una PSEL (Planta de Selección de Envases Ligeros). 20 almacenes, 181 vehículos en flota, 361 proveedores y 553 empleados.

(Giahsa, s.f.) (¿Qué es Giahsa?, 2012)

2.2.2 Datos de Calañas y Manzanilla

Giahsa nos ha proporcionado los datos de dos pueblos; uno de ellos es Calañas, municipio de la provincia de Huelva, situado en el pleno corazón de la comarca del Andévalo, se caracteriza por su terreno montañoso y pedregoso. El otro pueblo del que hemos recibido los datos de facturación es Manzanilla, éste junto a otros pueblos determinan el denominado Condado de Huelva. Este Condado se justifica por las circunstancias geográficas, orográficas y económicas. Manzanilla se encuentra a 54 kilómetros de la capital de la provincia, Huelva, y es conocido por el buen vino que se produce y las diferentes Rutas de Vino que pueden realizarse por el Condado.

Debido a la confidencialidad no tenemos datos personales, pero si un identificador de concesión, el cual utilizaremos para distinguir cada usuario.

Todos estos datos pertenecen al área de abastecimiento de agua, las lecturas pertenecen a los años 2013, 2014 y 2015, la lectura de contadores se realiza cada dos meses, mientras que la factura se emite cada mes. Por lo que un mes habrá factura estimada y otro mes se factura sobre el consumo real. Se puede identificar en los datos si una factura es real, ya que en este caso la clave de facturación es 1, mientras que el resto de números de la clave de facturación son lecturas estimadas por distintas causas.

Además, también contamos con el tipo de abonado, que puede ser AC o BJ, fecha de alta del abonado, la calle de cada registro.

Giahsa también nos ha proporcionado el tipo de calibre de cada usuario que puede ser de 13 mm, 15 mm y 20 mm, aunque en los datos hay algún otro tipo de calibre, pero son excepcionales. El factorN nos indica el número de viviendas que comparten una acometida, por ejemplo, N=2 implica que la acometida da agua a dos viviendas, mientras que el código de actividad dice el tipo de edifico.

El tipo de actividad indica, si es de uso doméstico, comercial o municipal. La empresa clasifica el tipo de suministro dependiendo del uso del agua que se haga:

- Doméstico: Se aplica a locales destinados a viviendas, siempre que en dicho local no se realicen actividades industriales, comerciales o profesionales. Tampoco se incluyen en este grupo cocheras que sean independientes de la vivienda.

- Usos comerciales: Son aquellos suministros en el que el agua es un elemento no básico en una actividad profesional.

- Usos industriales: Se caracteriza por el uso básico y directo del agua en la actividad industrial o comercial.

- Centros oficiales: Suministros que sirven a centros y dependencias del Estado.

- Otros usuarios: Aquellos que no puedan ser clasificados en alguno de los grupos anteriores.

Y por ultimo contamos con los consumos en metros cúbicos de los usuarios, el número de empadronados en cada registro e información de que usuarios han sido fraudulentos y en qué mes se detectó dicho fraude.

2.2.3 Preprocesamiento de los datos

Los datos que Giahsa nos proporcionó venían en dos libros de Excel distintos por lo que tuvimos que llevar a cabo una unificación de los mismos en un solo libro. Por una parte, en un libro venía toda la facturación de los habitantes de los pueblos a tratar en el que cada registro tenía los siguientes atributos: POLIZA (identificador distinto para cada usuario), CONCEPTO_FACT (en todos los registros lo declara como B1), DESC_CONCEP_FACT (en todos los registros lo declara como ABASTECIMIENTO), AÑO_PERIOD_FACT, MES_PERIOD_FACT, ST_POLIZA (hay dos posibles opciones: AC o BJ), CONCE_UBIC, CONCESIÓN, FECHA_ALT_ABON, MUNICIPIO (CALAÑAS o MANZANILLA), COD_POBLACIÓN (siendo 64 para Manzanilla y 75, 75B, 75C, 75L Y 75P para Calañas), POBLACIÓN (incluye ambos pueblos más la aldeas que también pertenecen a éstos), CALLE, NUM, KM (todos los registros tienen un cero), BLOQUE (columna vacía), PORTAL (columna vacía), ESCALERA (columna vacía), PISO, LOCAL (información redundante en COD_ACTIV2), PUERTA (información redundante en COD_ACTIV2), CALIBRE (13, 15, 20, 25, 30 ó 40), FACTOR_N (1, 2 ó 6), M3_FACTURADO (consumo de cada póliza en el periodo marcado), MENSUAL_BIMENSUAL (todo marcado con M), CLAVE_FACT (1, 2, 3, 4, 5 ó 6), MES_FACTURA (del 1 al 12), AÑO_FACTURA (todos de 2015), COD_ACTIV2 (información muy útil que esta repetida en PUERTA y LOCAL), TIPO_ACTIV (indica si la actividad desarrollada en el lugar donde trabaja el contador es doméstica, comercial, industrial…).

Y por otro lado teníamos otro libro de Excel donde venían los diferentes casos de fraude con los siguientes atributos: ACTA_FR (número de póliza del cliente), FEC_ALT_FR (fecha de alta del fraude), STATUS_FR (puede ser PP, RE ó IN), DESC_FRAUDE_LECTOR (puede ser ABONADO ó NO ABONADO), COD_POBL_AS (código del municipio), CALLE, NUM, KM, BLQ, PORTAL, ESCALERA, PISO, LOCAL, PUERTA, ACOMETIDA, UBICACIÓN, SATUS_UBIC.

Como necesitábamos unificar estos dos ficheros Excel, hemos decidido hacerlo del siguiente modo: en el fichero que teníamos con todos los consumos hemos añadido una columna nueva (un nuevo atributo) llamada Fraude y hemos ido mirando que casos estaban en el libro de Excel de los fraudes y cuáles no, para así tener en un solo Excel señalado con exactitud que cliente ha realizado fraude. Por lo tanto, esta columna estará formada por SI o NO en todas sus filas.

Antes de empezar a pasar los datos a un archivo ".arff", ha sido necesario aligerar un poco la base de datos ya que había varios atributos que no aportaban nada a la investigación. Estos atributos que hemos considerado inútiles para nuestro objetivo los hemos eliminado del Excel ya unificado, consiguiendo una reducción de la dimensión y mayor velocidad de procesado. La explicación de por qué se eliminan estos atributos se encuentra en el siguiente apartado.

A continuación, se va a explicar que tratamiento hemos tenido que darle al Excel con el que hemos trabajado para poder introducirle los datos al Weka.

2.3 Conversión de los datos de Excel a ".arff"

Para poder analizar el conjunto de datos y poder aplicarle técnicas de minería de datos utilizando el software elegido, Weka, es necesario crear un archivo ".arff". En este caso, la empresa nos ha suministrado los datos en un archivo Excel, por lo que a partir de este Excel hemos tenido que transformarlo a un archivo ".arff".

Lo primero que debemos hacer en el propio Excel es guardar dicho archivo como ".csv" (delimitado por comas). Posteriormente, se abre este archivo con el bloc de notas o notepad++ y se modifica el archivo dándole una estructura ".arff". Para ello, debemos introducir en la cabecera "@Relation" seguido del nombre del archivo, en cada una de las siguientes líneas escribiremos "@Attribute" seguido del nombre de cada atributo. En cada línea sólo se podrá colocar un único atributo, por lo que tendremos tantas filas con esta estructura como atributos tenga nuestra base de datos. Siguiendo el nombre de cada atributo escribiremos el tipo de archivo, normalmente este será *numeric* o *real* (para datos numéricos reales), *string*(texto), *integer*(expresa números enteros), *date*(expresa fechas) o una serie de datos entre {}, por ejemplo {1,2}. Cuando terminemos con todos los atributos y en la línea anterior a los datos ponemos "@Data". Todas las filas de datos deberán tener el mismo número de columnas, el cual tiene que coincidir con el número de atributos declarados anteriormente. Si no se dispone de algún dato se colocará "?". Para los decimales se utiliza el punto y los datos de tipo *string* deberán estar entre comillas simples. Si queremos poner un comentario utilizaremos %, que indicará que desde ese símbolo hasta el final de la línea es todo un comentario.

Acto seguido se guarda el archivo como ANSI y con la extensión ".arff". Una vez hecho todo esto, ya es posible abrir este archivo con nuestro software Weka y proceder a su análisis. Debemos tener cuidado ya que al abrirlo en Weka es posible que nos de error, para ello debemos ir línea a línea viendo los posibles fallos por los que Weka no puede leer el archivo. Son muy comunes los ";" ente los datos, es necesario cambiarlos por",". También hay que tener cuidado con los espacios entre los datos, dado que debemos eliminarlos o no será posible cargar el archivo.

En el caso de los datos de consumo proporcionados por GIAHSA tuvimos diversos problemas para crear dicho archivo ".arff" debido al desorden de los datos, a la gran cantidad de datos erróneos y a que había datos faltantes que eran necesarios para el análisis. Estos datos los buscamos y comparamos con otros registros, ya que pertenecían al mismo usuario. A continuación, se verán más detallados los cambios que hubo que realizar para crear el archivo ".arff":

Muchos de los atributos que se eliminaron para realizar el archivo arff son: CONCEPTO_FACT, DESC_CONCEP_FACT, POLIZA, CONCESION, ST_UBICACION, POBLACION, KM, BLQ, PORTAL, ESCALERA, PISO, LOCAL, PUERTA, INFORME, FECHA_FACTURA, AÑO_FACTURA, MES_FACTURA, MENSUAL_BIMENSUAL.

Estos atributos se han eliminado por falta de información, por incoherencias de datos, por ser atributos no necesarios para el estudio y por proporcionar información redundante.Además de estos cambios se realizaron todos los pasos explicados y detallados anteriormente.

Había instancias con meses repetidos, para un mismo usuario, donde existían consumos diferentes. También, se apreciaban mismos usuarios con calibres distintos, cosa que es imposible. En un gran número de instancias se mezclaban tres atributos, es decir, la información de un atributo estaba en el siguiente, la de este último en el siguiente y así repetidamente. Había una gran cantidad de datos faltantes, por ejemplo, en la columna de los metros cúbicos gastados había muchos meses donde no aparecía cantidad alguna.

```
@RELATION student

@ATTRIBUTE school {GP,MS}
@ATTRIBUTE sex {F,M}
@ATTRIBUTE age numeric
@ATTRIBUTE address {U,R}
@ATTRIBUTE famsize {GT3,LE3}
@ATTRIBUTE Pstatus {A,T}
@ATTRIBUTE Medu numeric
@ATTRIBUTE Fedu numeric
@ATTRIBUTE Mjob {at_home,teacher,services,health,other}
@ATTRIBUTE Fjob {at_home,teacher,services,health,other}
@ATTRIBUTE reason {reputation,course,home,other}
@ATTRIBUTE guardian {father,mother,other}
@ATTRIBUTE traveltime numeric
@ATTRIBUTE studytime numeric
@ATTRIBUTE failures {0,1,2,3}
@ATTRIBUTE schoolsup {yes,no}
@ATTRIBUTE famsup {yes,no}
@ATTRIBUTE paid {yes,no}
@ATTRIBUTE activities {yes,no}
@ATTRIBUTE nursery {yes,no}
@ATTRIBUTE higher {yes,no}
@ATTRIBUTE internet {yes,no}
@ATTRIBUTE romantic {yes,no}
@ATTRIBUTE famrel numeric
@ATTRIBUTE freetime numeric
@ATTRIBUTE goout numeric
@ATTRIBUTE Dalc numeric
@ATTRIBUTE Walc numeric
@ATTRIBUTE health numeric
@ATTRIBUTE absences numeric
@ATTRIBUTE G1 numeric
@ATTRIBUTE G2 numeric
@ATTRIBUTE G3 numeric
@ATTRIBUTE subject {por,mat}

@DATA
GP,F,18,U,GT3,A,4,4,at_home,teacher,course,mother,2,2,0,yes,no,no,no,yes,yes,no,no,4,3,4,1,1,3,4,0,11,11,por
GP,F,17,U,GT3,T,1,1,at_home,other,course,father,1,2,0,no,yes,no,no,no,yes,yes,no,5,3,3,1,1,3,2,9,11,11,por
```

Ilustración 3. Archivo ".arff". (Elaboración propia, 2016)

Se intentó realizar la conversión automática de la hoja Excel a un archivo .arff mediante alguna página web, pero siempre daba error, ya que, como se ha comentado antes el archivo tenía demasiado datos faltantes y errores.

Una vez hemos explicado todo el tratamiento que le hemos dado a los datos para poder aplicarle las diferentes técnicas con el software elegido, vamos a pasar a explicar en detalle cómo funciona el software Weka viendo las distintas interfaces que maneja, aunque nos centraremos más en la interfaz con la que hemos trabajado más en nuestro proyecto.

2.4 Software empleado: Weka

2.4.1 Introducción

WEKA (Waikato Environment for Knowledge Analysis) es un potente programa de aprendizaje automático utilizado para la minería de datos, que fue desarrollado por la universidad de Waikato en Nueva Zelanda en 1993. En sus orígenes fue desarrollado para analizar los datos originados por la agricultura y se programó en lenguaje C, aunque desde 1997 está programado en Java.

Ilustración 4. Logo del software empleado, Weka. (Weka 3: Data Mining Software in Java, s.f.)

Es de código abierto publicado bajo la Licencia Pública General de GNU. Software que contiene herramientas para el procesamiento de datos, clasificación, regresión, clustering, reglas de asociación y la visualización. Es muy recomendable para la experimentación y la investigación en el reconocimiento de patrones ocultos. Este programa tiene acceso a 4 interfaces:

- Simple CLI: Permite acceder a todas las funciones de WEKA desde la línea de comandos.

- Experimenter: Permite aplicar experimentos a gran escala.

- KnowledgeFlow: Genera proyectos mediante la generación de flujo de información.

- Explorer: Tiene acceso a todas las funciones de una manera muy sencilla, además es la más usada y la que hemos empleado nosotros más para nuestra investigación. Esta interfaz presenta distintos paneles:

  - Preprocess: Tiene opciones para importar datos y para procesarlos gracias a los filtros.
  - Classify: Permite aplicar técnicas de clasificación y regresión. También posibilita estimar la exactitud del modelo predictivo resultante.
  - Associate: Proporciona acceso a las reglas de asociación.
  - Cluster: Acceso a los algoritmos de clustering.
  - Selectedattributes: Permite identificar que atributos son mas predictivos.
  - Visualize: Matriz de puntos, los cuales pueden seleccionarse y analizarse con mayor detalle.

De entre sus muchas ventajas cabe destacar que es de fácil acceso para los usuarios, tiene una interfaz de usuario muy sencilla, gracias a los varios tipos de gráficos que contiene permite una mejor comprensión de los datos y ofrece una amplia gama de técnicas para modelado y procesamiento de datos. Sin embargo, no incluye algoritmos para modelar secuencias. El formato que lee WEKA por defecto es ".arff", aunque también admite archivos CSV y archivos codificados según el formato C4.5. WEKA, además de permitir abrir eso tipos de archivos que tengamos guardados en el ordenador, nos permite obtener los datos desde una dirección Url o desde una base de datos. La descarga de este programa se puede realizar desde http://www.cs.waikato.ac.nz/ml/weka/downloading.html, donde se pueden encontrar diferentes opciones para descargar el archivo ejecutable. Las versiones que hemos utilizado nosotros son Weka 3.8 y Weka 3.6.1.

Una vez descargamos y abrimos el programa nos encontramos con la posibilidad de elegir entre uno de los cuatros interfaces diferentes, anteriormente mencionados.

Cabe destacar que Weka está en continuo desarrollo y cada interfaz evoluciona por separado. Además, como se puede apreciar en la imagen de abajo en la version 3.8 hay una nueva interfaz llamada Workbench, la cual combina todo en uno.

Ilustración 5. Interfaces de Weka. (Elaboración propia, 2016)

2.4.2 Interfaces de Weka

- *SIMPLE CLI*

Abreviación de Simple Client. Proporciona una consola para introducir comandos. Permite realizar cualquier operación de forma directa, pero es muy complicada de manejar, debido a que es necesario tener un conocimiento extenso del software. En la actualidad, Weka posee más interfaces por lo que su utilidad se ha quedado reducida a ayudar a la fase de pruebas. A continuación se puede visualizar la pantalla que se nos muestra si seleccionamos la interfaz SIMPLE CLI.

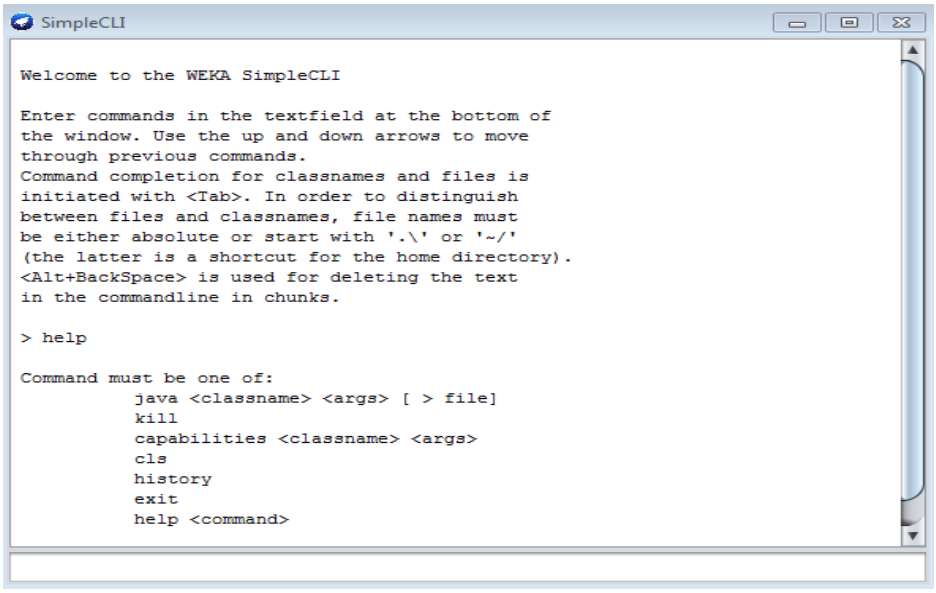


Ilustración 6. Interfaz *Simple CLI*. (Elaboración propia, 2016)

- *EXPLORER*

Es el interfaz más usado, ya que permite visualizar y aplicar multitud de algoritmos a un conjunto de datos. Es el interfaz que se ha empleado en este proyecto tanto para las técnicas descriptivas como para las predictivas, por lo que será el interfaz en el que profundizaremos más y sobre el que entraremos a detallar más a fondo. Cada una de las operaciones de minería de datos está representada por una pestaña en la parte superior.

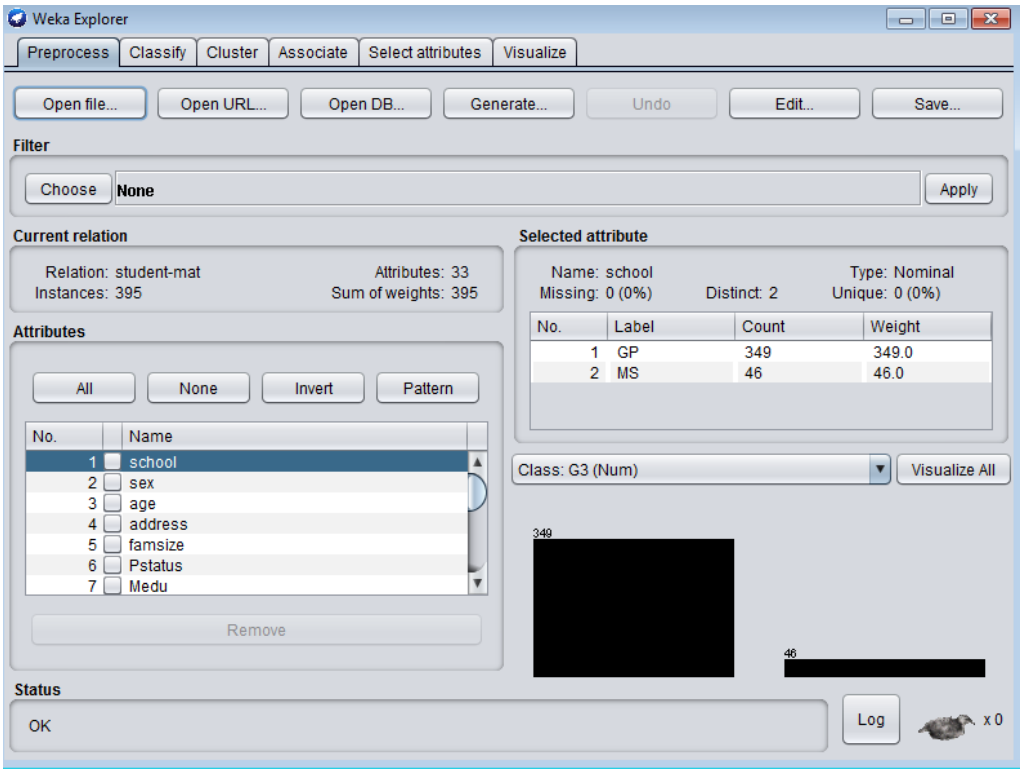


Ilustración 7. Interfaz *Explorer* con la pestaña *Preprocess*.(Elaboración propia, 2016)

➢ <u>Preprocesado de datos y filtros</u>

Este programa soporta diversas fuentes para poder leer los datos:

I. Open file: Al abrir esta pestaña el formato por defecto es el ".arff", aunque no es el único. También podemos utilizar CSV o C4.5 entre otros.

II. Open URL: A partir de este botón se podrá introducir una dirección, en la que se encuentra el fichero con los datos para el análisis. En

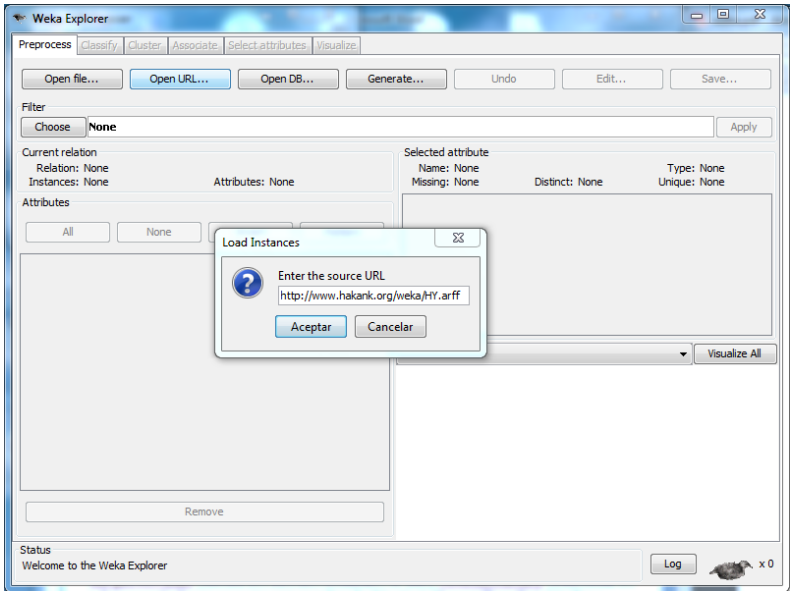


Ilustración 8. Opción de cargar una base de datos desde una dirección URL. (Elaboración propia, 2016)

laimagen se puede ver la forma de cargar los datos de esta forma.

III. Open DB: En esta opción se puede obtener los datos desde una base de datos.

Una vez ya tenemos los datos podemos aplicarles filtros. Los botones *undo* y *save* son para deshacer cambios y guardar los nuevos datos transformados. En la parte izquierda de la imagen anterior se pueden observar los atributos, si seleccionamos un atributo se pueden apreciar estadísticas sobre este como media aritmética, tipo, número de instancias distintas, etc. En el cuadro inferior derecho se aprecia una representación gráfica del atributo. En la pestaña *Visualize All* se abre una ventana que muestra todas las graficas de los atributos.

Si pulsamos *Choose* se despliega un árbol con los distintos filtros, estos se pueden aplicar a atributos o a instancias.

Para atributos podemos encontrar: *Add* (añade un atributo más), *AddExpression* (agrega una función al final del atributo), *AddNoise* (añade ruido a un atributo), *ClusterMembership*(da la probabilidad de que cada atributo este clasificado en una clase u otra), *Copy* (copia un conjunto de atributos), *Discretize* (discretiza un conjunto de valores en rangos), *FirstOrder* (realiza una transformación de los datos obteniéndose la diferencia entre pares consecutivos de datos), *MakeIndicator* (reemplaza un atributo nominal por uno booleano), *MergeTwoValues* (fusiona dos atributos nominales en uno), *NominalToBinary* (transforma los valores nominales en un vector con coordenadas binarias), *NumericTransform* (similar a *AddExpression*), *Obfuscate* (útil para compartir una base de datos pero no se quiere compartir cierta información), *PKIDiscretize* (discretiza atributos numéricos), *RandomProjection* (reduce la dimensionalidad de los datos), *Remove* (borra un conjunto de atributos), *RemoveType* (elimina un conjunto de atributos de un tipo), *RemoveUseless* (elimina atributos que oscilan menos de un nivel de variación), *ReplaceMissingValues*(reemplaza valores indefinidos), *Standarize* (estandariza datos númericos), *StringToNominal* (convierte un atributo tipo string en nominal), *SwapValues* (intercambia valores de dos atributos que son nominales), *TimeSeriesDelta* (asume que las instancias forman una serie temporal, reemplaza los valores por la diferencia entre el valor actual y el pronosticado para dicha instancia).

Para instancias podemos encontrar: *NonSparseToSparse* (transforma una muestra de modo completo a modo abreviado), *Randomize* (modifica el orden de las instancias), *RemoveFolds* (elimina conjunto de datos), *RemoveMisclassified* (aplica un método de clasificación a las muestras y elimina las que queden mal clasificadas), *RemovePercentage* (suprime un porcentaje de muestras), *RemoveRange* (elimina un rango de instancias), *RemoveWithValues* (elimina instancias según una restricción), *Resample* (obtiene un subconjunto del conjunto inicial),*SparseToNonSparse* (transforma una muestra a modo completo).

- Clasificación

El modo clasificación se encuentra en la parte superior, la segunda pestaña. En esta podemos clasificar mediante varios métodos los datos.

Lo primero que se debe hacer para aplicar una clasificación es elegir el clasificador y configurarlo, para ello presionar el botón *Choose*. Elegimos el clasificador y lo configuramos pudiendo variar las características que Weka trae por defecto. Es aquí donde elegimos si vamos a emplear el algoritmo C4.5 (J48 en Weka), el KNN (IBK en Weka) o si el Naive Bayes, por ejemplo.

Posteriormente se configura el modo de entrenamiento. Weka nos proporciona 4 modos:

a. *Use training set*: Entrenará con todos los datos. Normalmente no da unos resultados muy óptimos cuando contamos con gran cantidad de datos debido a los

tiempos de proceso. Por lo que es necesario ser cuidadoso con esto o emplear los modos de entrenamiento explicados a continuación, evitando el uso de todos los datos para la fase de entrenamiento.

b. *Supplied test set* : Aplica el clasificador a un fichero con datos distintos a los de entrenamiento.

c. *Cross-Validation*: Realiza una validación cruzada de K hojas. Es decir, divide los datos en K partes y por cada parte se construye un clasificador con las K-1 partes restantes y se prueba con esa. Por defecto, Weka emplea una K igual a 10 pero este dato podemos alterarnos a nuestro gusto.

d. *Percentage split* : Se define un porcentaje con el que se construirá el clasificador y se probará con el resto. Es igual que el caso anterior, *Cross-Validation*, pero en lugar de seleccionar el número de hojas se selecciona un porcentaje. Por defecto, el porcentaje seleccionado para el entrenamiento es el 66%, pero también es posible cambiarlo a nuestro gusto. Cabe señalar que esta opción desordena aleatoriamente el conjunto inicial y después es cuando lleva a cabo la división de los datos en 2 partes: una para entrenamiento y otra para el test. Por este motivo, si practicáramos con este clasificador dos veces, obtendríamos dos resultados ligeramente distintos debido a la desordenación previa llevada a cabo por el clasificador. En el botón *More options…* podemos cambiar las opciones por defecto de Weka y conseguir que se mantenga el orden de los datos siempre, evitando así que los resultados vayan variando a no ser que cambiemos el porcentaje.

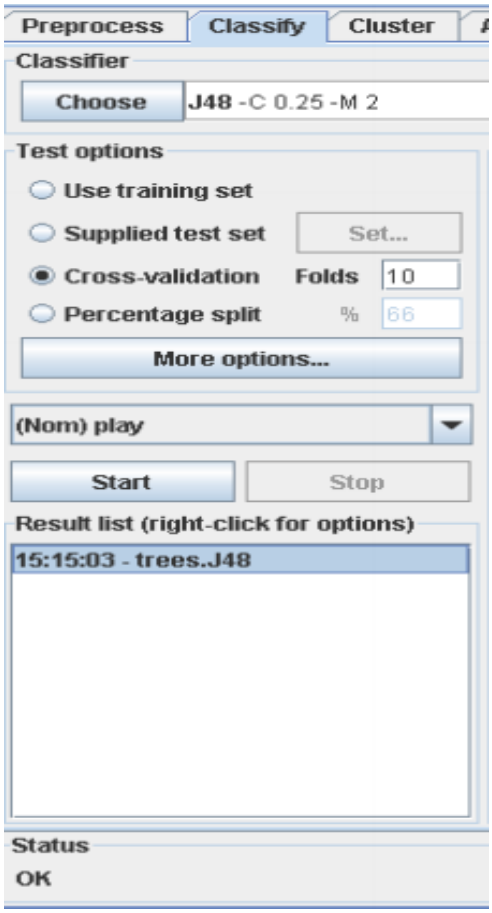


Ilustración 9. Los 4 modos de entrenamiento de Weka.(Bouckaert, y otros, 2016).

Después de elegir el método de entrenamiento, se puede seleccionar algunas opciones más en *More Options:*

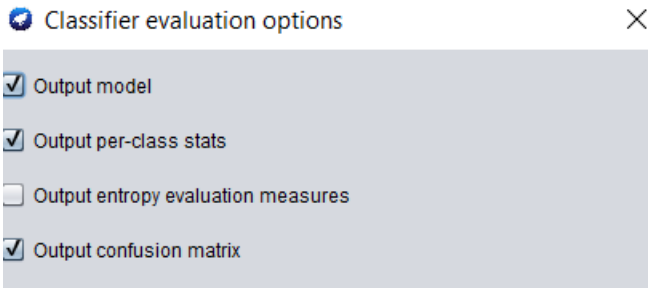


Ilustración 10. Botón *More Options***.** (Elaboración propia, 2016)

a. *Output Model*: Si se activa mostrara el modelo que ha construido.
b. *Output per-class stats*: Muestra estadísticas referentes a cada clase.
c. *Output entropy evaluation mesures*: Informa de las mediciones de la entropía en la clasificación.
d. *Output confusion matrix*: Presenta la matriz de confusión del clasificador. Donde las columnas son las categorías clasificadas por el clasificador y las filas las categorías reales. Es decir, en la diagonal principal estarán los elementos acertados y el resto serán erróneos.

Debajo de *More Options* hay un menú desplegable que permite seleccionar un atributo de nuestra muestra. Este será el resultado real de la clasificación y suele ser el último atributo de nuestra base de datos.

Ilustración 11. Icono de Weka que muestra si el software está ejecutando un algoritmo (si gira el Logo) o no. (Elaboración propia, 2016)

Ya tenemos todo listo para implementar un método de clasificación, solo queda pulsar el botón *Start*. Cuando se pulsa *Start* el icono de Weka que está en la esquina inferior derecha empieza a girar y cuando éste pare es cuando Weka ha terminado de estudiar los datos y ya ofrece el resultado en la pantalla *Classifier output*. A continuación se puede apreciar el resultado de aplicar un clasificador aleatorio, hemos elegido el de Naive Bayes aleatoriamente.

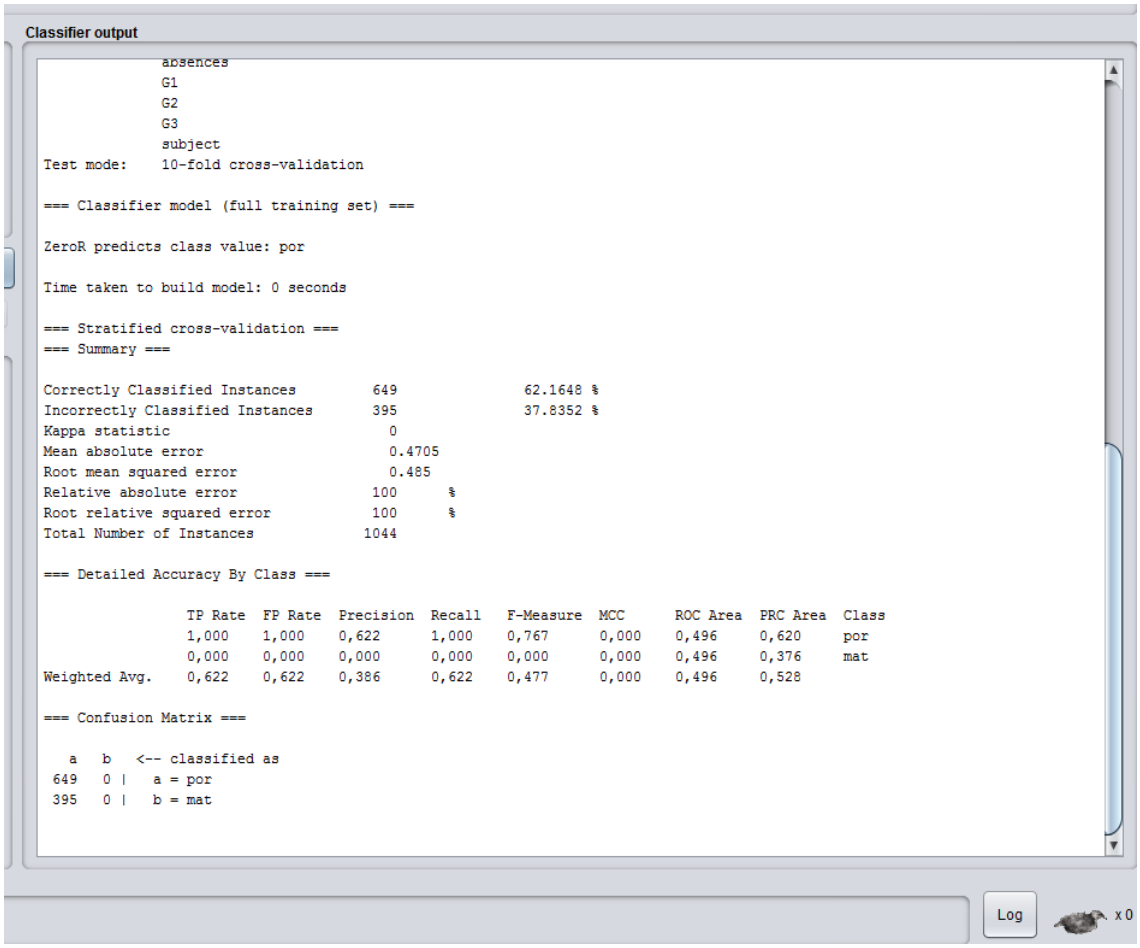


Ilustración 12. Resultados de un experimento con un algoritmo Clasificador. (Elaboración propia, 2016)

En la esquina inferior izquierda se encuentra la lista de resultados en la que podemos consultar todos los experimentos que hayamos realizado. Pulsando con el botón secundario del ratón sobre los diferentes experimentos realizados podremos tener acceso a opciones adicionales, algunas de ellas específicas para cada algoritmo empleado y, por esto, no siempre todas las opciones están disponibles. En la imagen de la derecha podemos ver las opciones adicionales, las que están en color gris son aquellas que no están disponibles para el clasificador sobre el que hemos dado con el ratón.

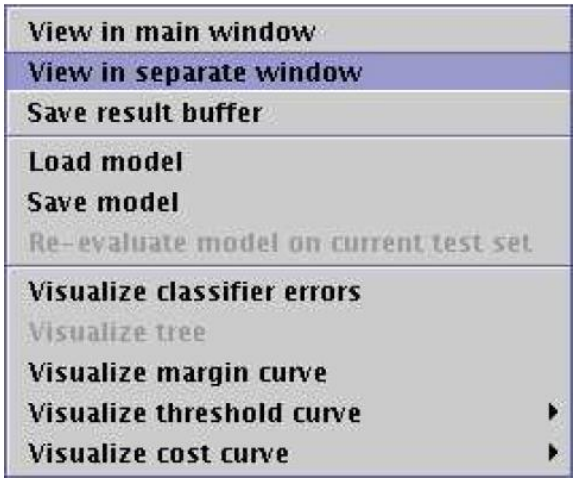


Ilustración 13. Acceso a opciones adicionales de cada experimento. (Elaboración propia, 2016)

- Clustering

Una vez descrita la segunda pestaña, pasamos a la tercera que es llamada *Cluster*, esta pestaña se encarga del *clustering* de información, como su propio nombre indica. El funcionamiento es prácticamente igual que el de la pestaña de clasificación: se elige un método de *clustering*, se elige que forma de entrenamiento quiero emplear y dándole al botón *Start* el software empieza a trabajar y muestra los resultados en la pantalla llamada *Clusterer output* en este caso.

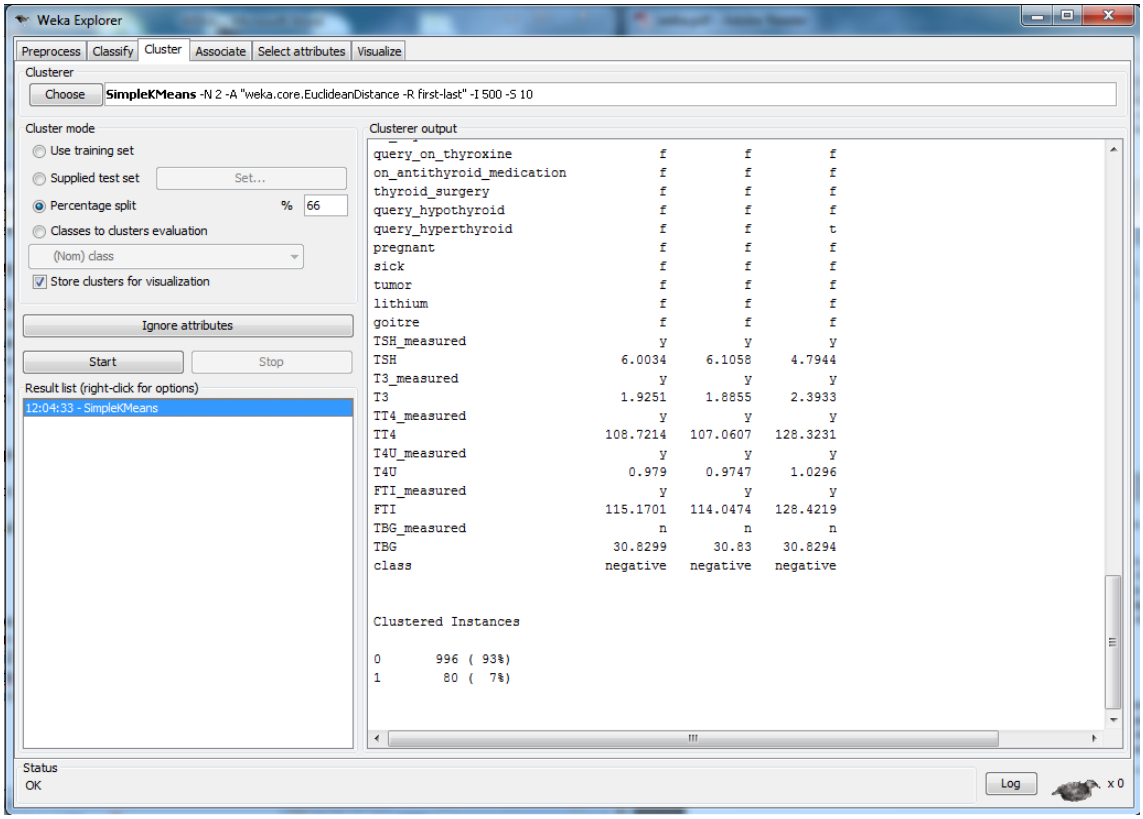


Ilustración 14. Interfaz *Explorer*con la pestaña *Cluster*. (Elaboración propia, 2016)

Una novedad reseñable de esta pestaña, con respecto a la anteriormente descrita, es la posibilidad de ver gráficamente la asignación en clusters. Esto se obtiene manteniendo activa la opción *Store clusters for visualization* antes de iniciar el experimento. Una vez tengamos los resultados del experimento, pulsamos con el botón derecho sobre el algoritmo en cuestión en la lista de resultados y marcamos *Visualize cluster assignments*.

- Búsqueda de asociaciones

La cuarta pestaña, *Associate*, nos permite asociar datos y poco más. Los métodos que se aplican en esta pestaña sólo son válidos cuando estamos trabajando con datos nominales. Ésta es la opción más sencilla y fácil de manejar ya que no tiene muchas variantes que añadir; sólo es necesario decidir que método emplear y configurarlo.

- Selección de atributos

En esta pestaña es posible ver que atributos son los que tienen mayor importancia a la hora de determinar si los datos son de una clase o de otra. Se puede emplear en la fase del preprocesado de los datos para ver que atributos podemos eliminar de nuestra base de datos y, así, aligerar un poco la cantidad que le introducimos al software. Con esta descarga de datos conseguimos disminuir los tiempos de funcionamiento del software, punto muy importante cuando tratamos con bases de datos muy densas.

En esta pestaña sí hay varias cosas a seleccionar para poder llevar a cabo la selección de atributos. Para empezar, lo primero que hay que seleccionar es un método de evaluación de atributos, *Attribute evaluator*. Éste se encargará de evaluar cada uno de los atributos a los casos a los que haga frente dándole un peso a cada atributo. El proceso de selección es igual que como

dijimos anteriormente; seleccionamos el método que queramos emplear y, si lo vemos necesario, alteramos sus propiedades pulsando sobre el nombre del mismo.

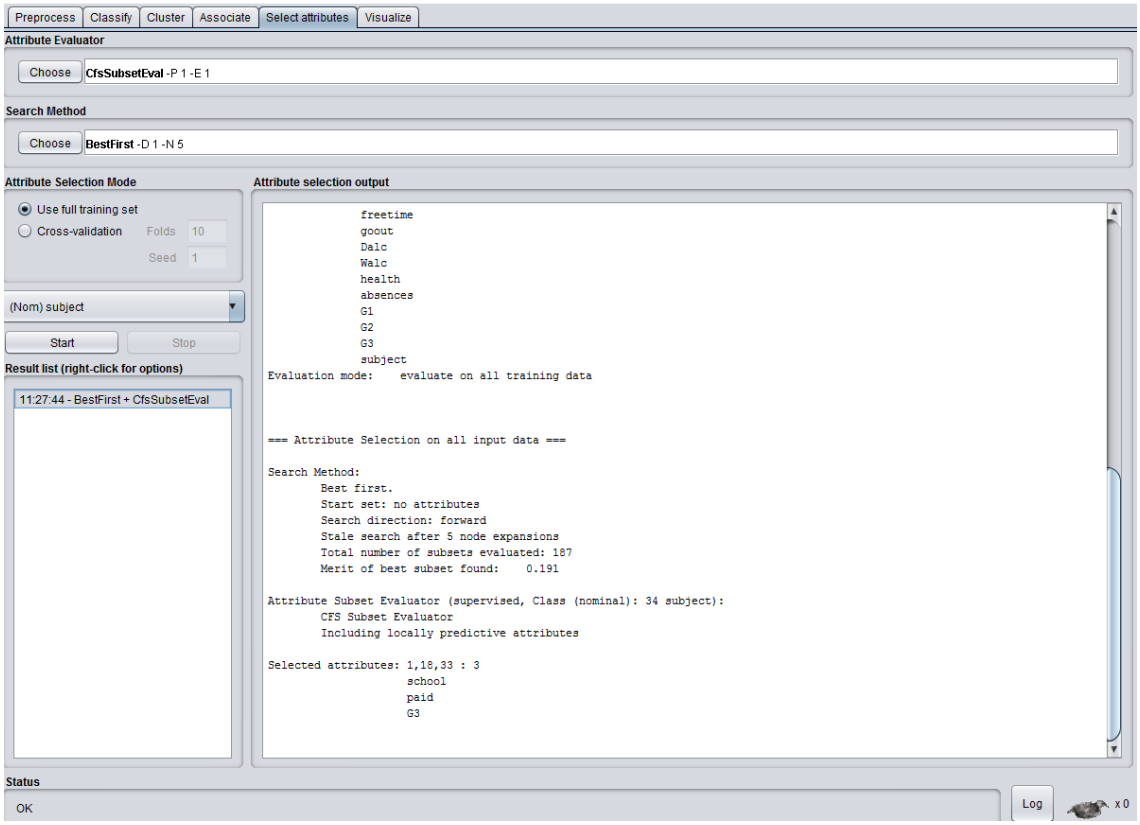


Ilustración 15. Interfaz *Explorer*con la pestaña *Select Attributes*. (Elaboración propia, 2016)

La siguiente decisión a tomar será la de elegir el método de búsqueda, con el fin de producir el espacio de pruebas. Una vez seleccionadas las dos opciones descritas sólo faltaría el método de prueba, el atributo sobre el que se va a realizar el estudio e iniciar la exploración pulsando el botón *Start*.

- Visualización de datos

Ésta última pestaña del *Explorer* es el modo *Visualize*, éste modo representa gráficamente como se distribuyen todos los atributos con los que estamos trabajando. Las gráficas mostradas en esta ventana son en dos dimensiones, se representan en los ejes todos los atributos de dos en dos permitiéndonos ver correlaciones y asociaciones visualmente.

Si pinchamos sobre cualquier gráfica de las que nos aparecen en la pantalla, se nos abrirá otra ventana con más nivel de detalle sobra la gráfica seleccionada. Este modo de visualización nos ofrece 3 diferentes opciones que se manejan mediante una barra deslizante. Las posibles opciones son:

- Plotsize: determina el tamaño del lateral de cada gráfica en píxeles, puede variar de 50 a 500.

- Pointsize: determina el tamaño del punto de las gráficas expresado en píxeles también, va desde uno hasta diez.

- Jitter: esta opción añade un ruido al azar con el objetivo de separar las muestras que están muy juntas. En algunas ocasiones se busca conseguir esto para que nos sea posible diferenciar los diferentes puntos que hay en un área si estos están demasiado concentrados.

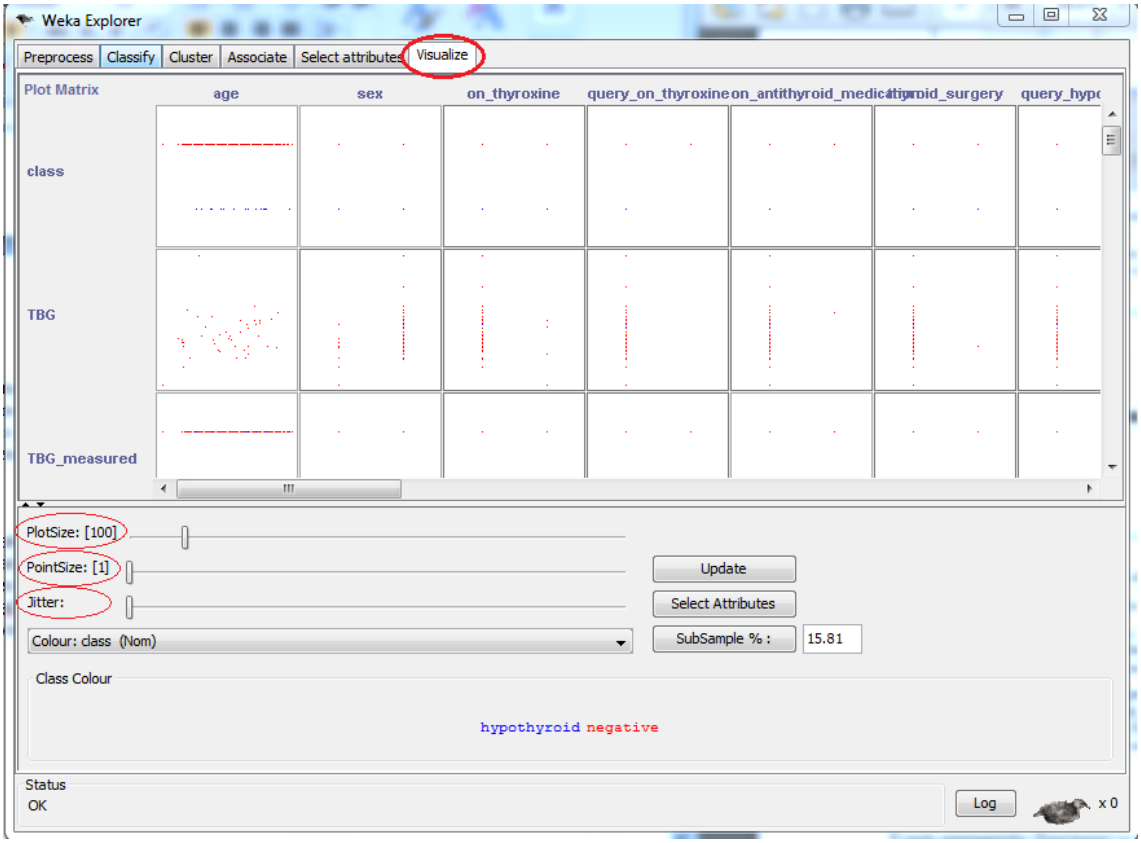


Ilustración 16. Interfaz *Explorer*con la pestaña *Visualize*. (Elaboración propia, 2016)

Una vez establecidas las diferentes características de este modo, debemos pulsar en el botón *Update* para que se queden grabadas y se actualicen las gráficas. En esta pestaña hay dos botones más que pueden ser útiles: el Select Attributes y el SubSample %. El primero de ellos nos permite seleccionar los diferentes atributos que queremos representar en las gráficas. El último botón nos da la opción de elegir qué porcentaje de muestras se va a representar.

- *EXPERIMENTER*

Este modo es útil para aplicar uno o varios métodos de clasificación sobre grandes bases de datos y, luego, realizar contrastes estadísticos entre ellos. Esta interfaz cuenta con tres pestañas que vamos a comentar a continuación.

1. **Setup.**

La primera pestaña de esta interfaz es la de *Setup*, esta pestaña se encarga de la configuración del *Experimenter*. Por defecto, viene configurado en modo simple y este modo nos da la capacidad de definir infinidad de cosas. Lo primero que debemos definir es el fichero configuración que contendrá todos los ajustes, ficheros involucrados, notas, etc. También hay que elegir donde se van a almacenar los resultados, si queremos almacenarlos, y en que formato, Weka nos ofrece tres opciones: archivo ".arff", fichero CSV o en una base de datos. Lo siguiente que debemos definir es el tipo de validación (que no difiere mucho del caso del *Explorer*): validación cruzada, entrenamiento con un porcentaje y tomando a la población al azar y entrenamiento con un

porcentaje pero tomando a la población de forma ordenada. Llegados a este punto sólo nos queda declarar que archivos queremos que sean parte de nuestra investigación y cuántas repeticiones del experimento queremos realizar.

Con lo explicado anteriormente ya tendríamos configurado un modo simple. Si por el contrario quisiéramos configurar un modo *Advanced* sería necesario llevar a cabo todas las configuraciones anteriores con algo más de concreción ya que la gran diferencia entre el modo simple y el avanzado es que, el segundo, está enfocado a tareas específicas.

2. Run

Con todas las características del experimento declaradas en la pestaña de configuración, *Setup*, pasamos a esta pestaña que sirve para comenzar el experimento. Pulsando el botón *Start* comienza Weka a ejecutar el experimento, experimento que podemos parar en cualquier momento pulsando *Stop*. Cabe destacar que un experimento que paramos en mitad del proceso no se puede volver a iniciar en el punto en el que lo hemos parado.

3. Analyse

Nos encontramos en la tercera, y última, pestaña del *Experimenter* que sirve para ver los resultados de nuestro experimento, contrastarlo estadísticamente, etc.

En este modo podemos definir el origen de los datos de los resultados y el test que queremos llevar a cabo. Cuando tengamos ambas cosas configuradas debemos emplear el botón *Perform test*, botón que realiza el análisis t de Student. El resultado de este será reflejado en al cuadro llamado *Test Output*. Weka nos permite guardar los resultados en un fichero de texto mediante el botón *Save Output*.

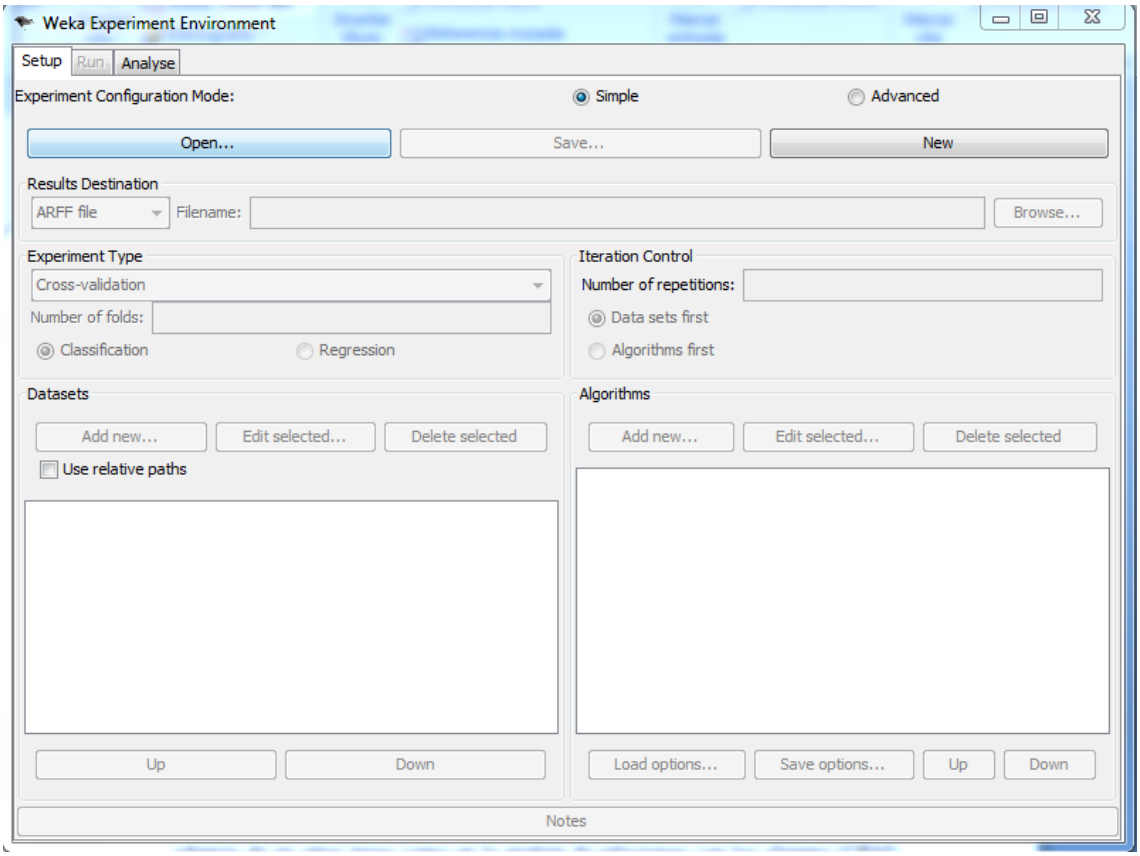


Ilustración 17. Interfaz *Experimenter*. (Elaboración propia, 2016)

- *KNOWKEDGE FLOW*

Esta interfaz es la más cuidada y la que mejor muestra cómo funciona nuestro programa internamente. Tiene un funcionamiento gráfico y se basa en la colocación en el panel de trabajo de elementos base, de manera que creemos una ruta que explique nuestro experimento. Esta interfaz aún está en progreso por lo que algunas funcionalidades del *Explorer* no están disponibles aún aquí. Por otro lado, hay algunas funciones que pueden llevarse a cabo en el *Knowledge Flow* pero no en el *Explorer*.

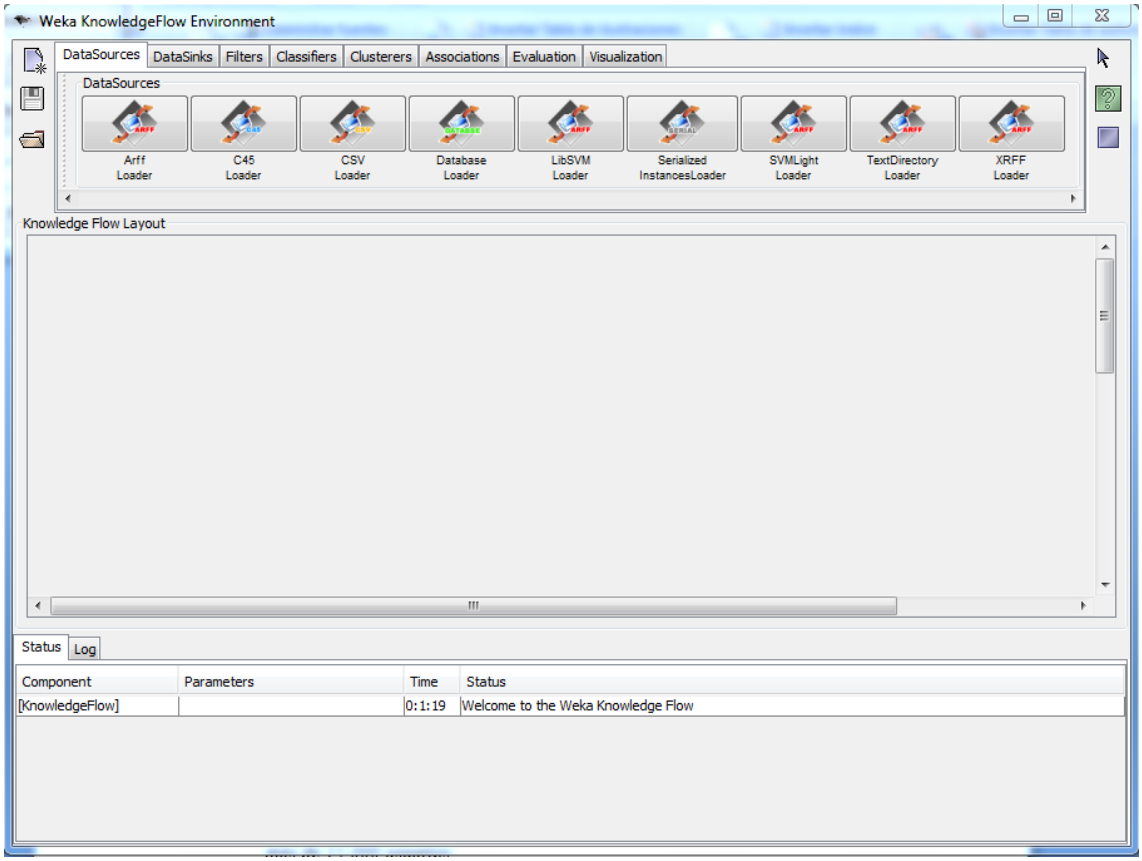


Ilustración 18. Interfaz *Knowledge Flow*. (Elaboración propia, 2016)

2.5 Resultados dela minería de datos aplicada a la base de datos inicial

Una vez realizado todo el preprocesado anterior de los datos y tras investigar en profundidad el software sobre el que íbamos a tratar hemos pasado a comprobar si con éste software y con los datos que Giahsa nos proporcionó, seríamos capaces de conseguir predecir los casos de fraude, idea inicial con la que comenzamos este proyecto.

Al iniciar este proceso, nos fijamos en que contábamos con más de 70.000 instancias, y que esto era una muy buena opción para aplicar minería de datos ya que teníamos muchísima información almacenada. Sin embargo, cuando nos pusimos en contacto con la empresa por segunda vez para que nos proporcionara los casos de fraude de esos clientes, nos encontramos con muy pocos casos de clientes fraudulentos en comparación con el número de instancias. Para ser más concreto, contábamos con más de 70.000 instancias, como hemos comentado, y con menos de 200 casos de fraude. Esta cifra representa menos de un 0,3% de casos fraudulentos dentro del total.

¿Qué pasa entonces ante este problema? Ante este problema nos hemos encontrado con el inconveniente de que no nos es posible entrenar al algoritmo de forma efectiva ya que no tiene los

suficientes datos para obtener patrones que se cumplan en la mayoría de los casos. Al entrenar al algoritmo y luego testearlo, hemos obtenido unos resultados desastrosos con un porcentaje de acierto muy bajo ya que lo que hace el algoritmo es clasificar como fraudulento el 99,5% de los casos. Con algunos parámetros del KNN se ha conseguido que clasificara algunos registros, aunque veíamos que con esta opción fallaba más que si no hiciera nada, ya que obtenía más fraudes que los que teníamos realmente.

Cuando hicimos la prueba para las técnicas descriptivas pues nos daba lo normal, dos grupos perfectamente definidos (los dos pueblos con los que contábamos). Tocando alguna que otra característica conseguimos obtener otras divisiones más pero poco relevantes.

Como curiosidad, hemos encontrado un proyecto similar al nuestro, donde la mitad de los clientes eran fraudulentos, 3.628 de 7.256. Este proyecto desarrolla un modelo de detección de fraude en clientes de una empresa de agua en Santiago de Chile, en él se obtuvieron buenos resultados ya que había miles de datos de fraude donde encontrar un patrón. Mientras que, como ya se mencionó antes, nosotros solo contábamos con 200 casos de 70.000 registros.

Como contraste con el proyecto mencionado anteriormente, una empresa sevillana pretendía conseguir detectar los posibles casos de fraude entre sus clientes. Esta empresa invirtió gran cantidad de dinero de cara a conseguir capturar a estos clientes obteniendo beneficios a largo plazo. Sin embargo, este proyecto no llegó a buen puerto ya que no se obtuvieron los resultados esperados debido a la falta de datos y de claridad de estos, por lo que todo el dinero invertido por la empresa fue derrochado para un fin que no se pudo conseguir.

2.6. Búsqueda de los nuevos datos

Como alternativa a estos malos resultado no esperados por nosotros, hemos creído oportuno buscar otras dos bases de datos con las que seamos capaces de demostrar que la minería de datos es una herramienta muy efectiva y válida, pero es necesario tener los datos suficientes y adecuados para el fin propuesto.

Estos dos conjuntos de datos no fueron los primeros ni los únicos vistos, ya que, hasta la elección de estas dos bases de datos, se miraron en profundidad otros muchos conjuntos de datos y páginas web. De entre los conjuntos de datos analizados, se vieron datos de distintos ámbitos como:

- Clasificación de vinos: esta base de datos estaba muy bien y fue seria candidata a tratar con técnicas descriptivas. Clasificaba los vinos según su pH, alcohol, sulfatos, densidad o acidez, entre otros.

- Valorar si una persona ingresa más de 50.000 dólares: esta base de datos no sabíamos muy claro si trataba sobre este tema, pero por sus características es lo que dedujimos. La descartamos debido a esta incertidumbre que teníamos.

- Clasificación de hojas de habas:esta base de datos contaba con 36 atributos y tenía una estructura perfecta. El inconveniente en este conjunto de datos ha sido que solo contaba con 800 instancias y el atributo llamado a ser clase era muy amplio, por lo que se desechó.

- Seguro de vehículos: esta base de datos contaba con 86 atributos y parecía muy buena ya que era bastante gruesa. El inconveniente que le vimos es que no tenía un atributo

muy claro sobre el que basar la investigación, es decir, no veíamos un buen candidato a clase.

- Evaluación de los coches de un concesionario: esta mostraba los coches de un concesionario junto a seis características (atributos).

- Juego asiático: esta base de datos nos resultó extraña cuando la vimos por su forma, nos documentamos acerca de ella y vimos que se basaba en un juego asiático que consiste en eliminar al Rey del tablero de ajedrez con una Torre. Esta base de datos representa las soluciones óptimas del juego (menor número de movimientos) en todas las posibles posiciones del tablero. Fue descartada porque no tiene mucha aplicación en la minería de datos, más bien es para entrenar algoritmos de inteligencia artificial.

- Multisensor de gas: esta estaba basada en un dispositivo multisensor de gas colocado en el campo de una ciudad italiana con mucha contaminación. Contaba con 15 atributos y parecía buena pero la descartamos ya que el Excel en el que venían los datos traía muchas erratas y era difícil de conocer el significado de algunos atributos.

Estas son algunas de las vistas, finalmente la elección de los dos conjuntos analizados en el proyecto fue debido a que el conjunto de datos tenía que tener un mínimo de instancias y atributos, a que los campos que se eligieron eran más interesantes de investigar que otros vistos, a que al analizar por encima los datos son las dos donde se encontraron unos objetivos interesantes y claros de alcanzar, mientras que en el resto eran objetivos más flojos o, simplemente, en esos conjuntos era difícil encontrar un fin hacia donde enfocar nuestro análisis. Para valorar la validez de todas las bases de datos que hemos manejado, hemos tenido en cuenta los siguientes aspectos de cada una:

- Temática.
- Atributos analizados.
- Número de instancias.
- Número de atributos.
- Objetivo que podíamos marcarnos.

Teniendo en cuenta estos aspectos anteriores llegamos a la selección de nuestras dos bases de datos. Una de ellas trata sobre la predicción de casos de hipotiroidismo y otra sobre la clasificación de alumnos con el objetivo de darles charlas sobre el consumo de alcohol, teniendo en cuenta los factores que influyen más en el alcoholismo. Ambos conjuntos de datos se han sacado mediante un enlace de la web oficial de Weka, más exactamente de la Universidad de California, a través de un repositorio donde se pueden encontrar infinidad de bases de datos.

(Lichman, 2013)

2.7. Conclusión

Una vez hemos comentado todo lo anterior, podemos ver que la minería de datos es muy útil, pero es necesario contar con los datos correctos o podemos llegar a resultados incongruentes. En el caso de las técnicas predictivas, el problema ha sido que al contar con menos de un 0,3% de casos de fraude dentro del total de los datos, el software cataloga todo como no fraudulento porque no es capaz de sacar patrones de comportamiento debido a los escasos datos de fraude con los que

lo hemos alimentado. Esto da un resultado incongruente porque catalogándolo todo como "No fraudulento" acierta con una eficacia de más del 99,7% debido a que los datos están así distribuidos, pero ese resultado no es real. Aplicándole otras técnicas como el KNN y jugando con el parámetro K, se consiguió que clasificara algunos casos, pero igualmente eran resultados bastante malos porque cometía más errores que si no hiciera nada, catalogando como "Fraudulentos" más casos de los que en realidad se tenían.

En el caso de las técnicas descriptivas con las que trabajaba mi compañero, se observó como la clasificación que llegaba a cabo el software era la más lógica: dividía a la población en dos grandes grupos que se correspondían con los dos pueblos (Calañas y Manzanilla).

Debido a esto, decidimos llevar a cabo la búsqueda de dos bases de datos alternativas para poder demostrar que cuando se cuenta con los datos y la información válida, la minería de datos consigue sacar los patrones de comportamiento necesarios obteniendo unos resultados espectaculares.

En la siguiente sección, se podrá ver el análisis de la base de datos dedicada a aplicarle algoritmos predictivos, cuyo objetivo final será detectar a los posibles pacientes que padezcan hipotiroidismo. Mi compañero Miguel Novoa, se encargará de comentar el conjunto de datos que se ha empleado para aplicarle métodos descriptivos a unos datos enfocados obtener los factores que influyen más en el consumo de alcohol de los jóvenes para orientar las charlas de concienciación según el público al que tengan que ir dirigidas.

3 Aplicación de técnicas predictivas

Ante la imposibilidad de llevar a cabo la idea inicial que teníamos en el proyecto de detectar a los posibles clientes fraudulentos aplicándole algoritmos predictivos a los datos proporcionados por Giahsa hemos decidido tomar como la alternativa la búsqueda de alguna base de datos por Internet en la que la aplicación de este tipo de técnicas nos dieran unos resultados más coherentes y así mostrar el aprendizaje adquirido en este campo.

Tras documentarnos con varias bases de datos sobre temas muy diversos hemos decidido que la que mejor se adaptaba al objetivo propuesto era una base de datos basada en la detección del hipotiroidismo. Aunque se trata de una base de datos de muchas menos instancias que la que teníamos proporcionada por Giahsa, aún con esto podemos ver a simple vista que es una base de datos mejor distribuida que la anterior ya que mientras antes teníamos unos 70 casos de fraude para 70.000 instancias ahora tenemos 151 dentro del total que lo conforman 3.163. Como se puede ver en un primer vistazo, es un porcentaje mucho mayor, lo que facilita que el resultado sea más exacto ya que los algoritmos consiguen sacar unos patrones fiables y reales de los casos en los que se detecta la enfermedad.

3.1 Introducción

3.1.1 Breve explicación de la temática

Ya que hemos cambiado de ámbito dentro del proyecto, se va proceder a explicar un poco la base de datos elegida, en que se basa y de que atributos está compuesta. Para empezar, se va a dar una breve explicación de qué es el hipotiroidismo; dicha enfermedad es una afección que se produce cuando la glándula tiroides no es capaz de producir suficientes hormonas tiroideas. La tiroides es encargada de generar hormonas que controlan la forma en la que cada célula emplea la energía.

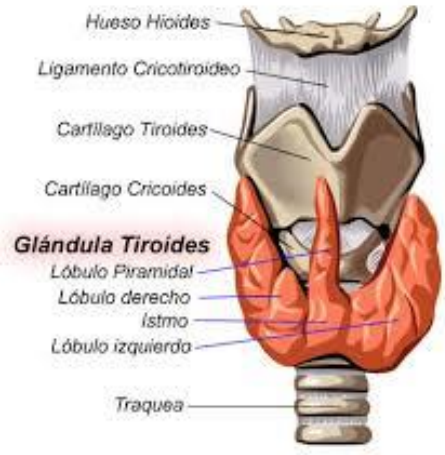


Ilustración 19. Anatomía de la tiroides.(Martínez Fraga, 2012)

Hay multitud de causas que provocan la aparición del hipotiroidismo, pero la principal es la tiroiditis: la hinchazón e inflamación de la glándula tiroides que daña las células generadas por ésta. Otras causas que originan esta enfermedad son: embarazo (también llamado tiroiditis posparto), infecciones virales, infecciones respiratorias, algunos medicamentos, anomalías congénitas, terapias de radiación en el cuello o cerebro para tratar cánceres, la extirpación quirúrgica de parte o de toda la glándula tiroidea…

Entre los síntomas relacionados con el hipotiroidismo podemos diferenciar entre dos grupos: los síntomas iniciales y los síntomas tardíos si no se trata la enfermedad cuando aparecen los iniciales. Algunos de los síntomas iniciales más comunes son: estreñimiento, fatiga, sentirse lento y pesado, menstruación abundante e irregular, dolor muscular o articular o un aumento de peso repentino. Si una vez se aprecian estos síntomas tratamos la enfermedad evitaremos que aparezcan los síntomas siguientes (los tardíos): disminuye el sentido del gusto y el olfato, ronquera, se hinchan la cara, las manos y los pies o baja frecuencia cardíaca.

Ilustración 20.Síntomas del hipotiroidismo y del hipertiroidismo. (Fertifarma, 2016)

En la imagen lateral podemos ver muchos de los síntomas de los que hemos hablado anteriormente comparados con otra enfermedad relacionada con la glándula tiroides también, el hipertiroidismo.

3.1.2 Atributos

Una vez se ha explicado qué es el hipotiroidismo, sus causas y sus síntomas; se va a pasar a comentar los atributos de los que se compone esta base de datos y a comentar porque se han tenido en cuenta en cada caso.

La base de datos con la que hemos pasado a trabajar tiene un total de 26 atributos:

1. Edad (número real)
2. Sexo (Hombre o Mujer)

Dos atributos importantes ya que se sabe por la experiencia que el hipotiroidismo es más propenso a aparecer en personas mayores de 50 años y más aún en las mujeres que en los hombres.

3. Tratamiento en tiroxina (Sí o No).
4. Si se ha visitado recientemente a un médico para evaluar sus niveles de tiroxina (Sí o No).

Estos son otros dos atributos que van de la mano. La tiroxina es, de todas las hormonas que segrega la glándula tiroides, la más importante ya que actúa sobre todas las células del metabolismo. Esta hormona ayuda a regular el sistema suprarrenal, influye en el crecimiento normal y el desarrollo de la persona, ayuda a mantener un peso saludable, y ayuda a mantener equilibrado el estado de ánimo.

5. Saber si el paciente ha estado recientemente o está en tratamiento con medicamentos antitiroideos (Sí o No).

Es otro factor a tener en cuenta y, por lo tanto, lo hemos considerado un atributo válido, ya que puede aportarnos información necesaria. Los antitiroideos se emplean como apoyo al diagnóstico en anomalías en la glándula tiroidea. Muchas veces es solicitada para detectar las causas del incremento de tamaño de la glándula de la que hablamos (bocio) o la causa de segregación de elevadas (hipertiroidismo) o insuficientes (hipotiroidismo) hormonas tiroideas.

6. Cirugía en la glándula tiroides (Sí o No).

Otro atributo que se ha tenido en cuenta es saber si el paciente ha sido sometido con anterioridad a cirugía en la glándula tiroides o no. Como se ha comentado antes, se dijo que una de las causas por las que podía aparecer este problema era por la extirpación de parte o la totalidad de la glándula tiroides.

7. Consulta para evaluar el hipotiroidismo (Sí o No).
8. Consulta para evaluar el hipertiroidismo (Sí o No).

Estos dos atributos son muy similares entre sí y evalúan si el paciente ha asistido a un médico recientemente para ver los niveles de hormonas que genera su glándula tiroides y ver si están en el nivel correcto.

9. Embarazo (Sí o No).
10. Enfermo (Sí o No).
11. Tumor (Sí o No).
12. Medicamentos con Litio (Sí o No).
13. Bocio: incremento del tamaño de la glándula tiroidea (Si o No).

Estos 4 atributos también tienen que ver con el historial médico del paciente y son de respuesta “Sí” o “No”. Embarazo, si está pasando o ha pasado recientemente una enfermedad, si ha sufrido algún tumor, si ha tratado con medicamentos que pueden facilitar la aparición del hipotiroidismo; en este caso se evalúa el trato con el litio, presente en muchos medicamentos y, finalmente, si sufre de bocio, lo que hemos dicho anteriormente que es un incremento del tamaño de la glándula tiroides. Como podemos observar hasta este punto, todos los atributos que se tienen en cuenta han sido nombrados en la parte de las causas de la aparición del hipotiroidismo, con esto se puede observar que los atributos elegidos no están seleccionados al azar, si no que cada uno de ellos está por una importancia palpable en el desarrollo de la enfermedad.

A continuación, se van a enumerar 12 atributos que se pueden emparejar de dos en dos: si un determinado valor se ha medido o no y, en caso de haberlo medido, su resultado.

14. Medida del TSH, Thyroid Stimulating Hormone u Hormona Estimulante de la Tiroides (Sí o No).
15. TSH (número real).

La determinación de TSH es el parámetro más sensible para diagnosticar hipotiroidismo. Un número elevado de TSH indica una insuficiencia en la función del tiroides. Esta hormona es producida por

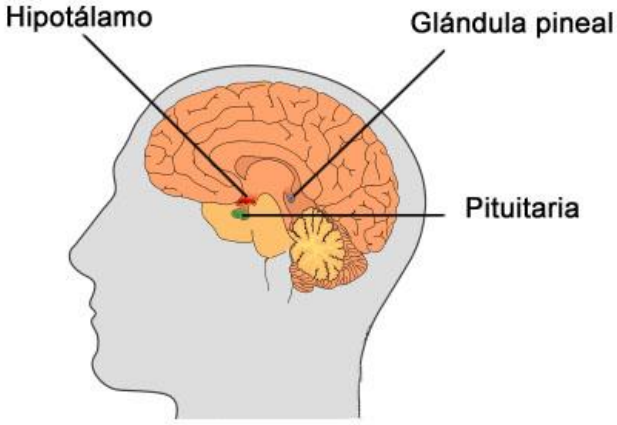


Ilustración 21. Glándulas endocrinas del cerebro.(Ropero Lara, 2012)

la hipófisis o glándula pituitaria, una glándula que se localiza en la base del cerebro y que se encarga de regular la actividad de otras glándulas de controlar determinadas funciones del cuerpo. "Los valores normales de TSH pueden fluctuar de 0.4 a 4.0 mlU/L (miliunidades internacionales por litro)".

(MedlinePlus, 2016.)

Los valores pueden variar a lo largo del mismo día y es aconsejable hacerse la prueba a primera hora del día.

16. Medida de T3, es otra hormona segregada por la glándula tiroides y se conoce como triyodotironina (Sí o No).
17. T3 (número real).

La T3 es una hormona de la tiroides como se ha dicho anteriormente y que tiene gran importancia en el control del metabolismo. Unos valores normales de T3 estarían entre 100 y 200 ng/dL (nanogramos por decilitro).

18. Medida de T4, otra de las hormonas segregadas por la glándula tiroides y conocida como tiroxina (Sí o No).
19. T4 (número real).

Anteriormente ya comentamos la función de la tiroxina, cuando explicábamos los atributos tres y cuatro. Los valores esperados de esta medición van desde 4.5 a 11.2 mcg/dL (microgramos por decilitro).

(MedlinePlus, 2014)

20. Medida de T4U (Sí o No).
21. T4U (número real).
22. Medida de FTI, Free Thyroxine Index o Índice de Tiroxina Libre.
23. FTI

La tiroxina libre (T4) puede medirse de dos formas: directamente (FT4) o mediante el FTI. El índice de tiroxina libre indica la cantidad de tiroxina libre comparándolo con la tiroxina atada. Con este parámetro se pueden emplear para hacer un seguimiento del tratamiento que se está llevando a cabo y ver si está dando el resultado esperado o no. Unos valores adecuados del FTI serían aquellos que se comprendieran entre 0.4 y 2.5 ng/dL (nanogramos por decilitro).

24. Medida del TBG, Globulina Fijadora de Tiroxina (Sí o No).
25. TBG (número real).

La Globulina Fijadora de Tiroxina es una proteína que se une en la circulación sanguínea a las hormonas mencionadas anteriormente: T3 y T4. La TBG es producida en el hígado. Los valores normales de esta prueba dependen de la forma en que se lleve a cabo la medición: si se utiliza electroforesis, los valores pueden moverse de 10 mg/100mL a 24 mg/100mL; si, por otro lado, se emplea radioinmunoanálisis, los valores pueden oscilar de 1.3 a 2.0 mg/100mL. En nuestro caso, se ha empleado la primera forma explicada ya que los datos observados en la base de datos así lo confirman.

(MedlinePlus,2014)

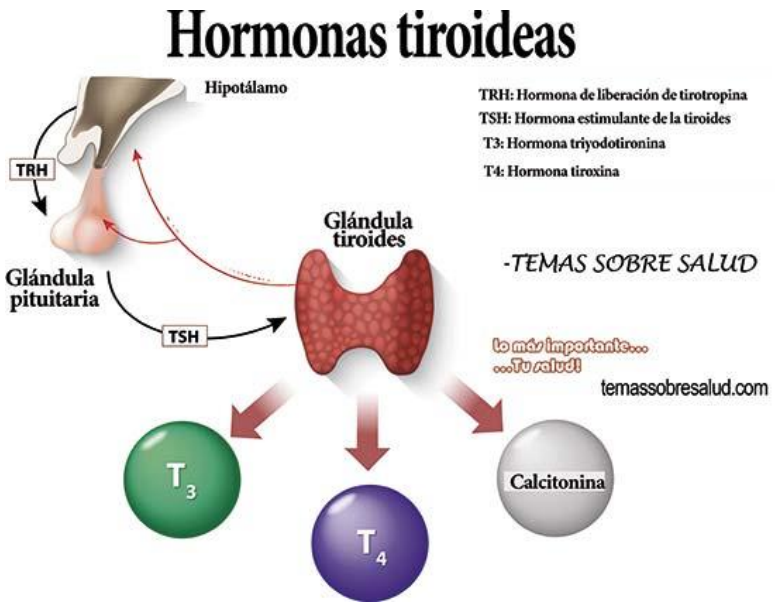


Ilustración 22. Esquema de las hormonas tiroideas. (Nasser, 2016)

26. Clase (hipotiroidismo positivo ó hipotiroidismo negativo).

Este último atributo es en función del que se estudiará y se entrenará a los algoritmos para que estos decidan si un paciente presenta síntomas que puedan ayudarnos a detectar esta enfermedad en sus inicios, evitando así males mayores.

3.2. Antecedentes

Para afinar más en las técnicas que debíamos emplear para este proyecto, se decidió investigar que técnicas habían sido empleadas en otros proyectos similares en el campo de la salud. En estos estudios no solo se tratan temas de tiroides, hay estudios de todo tipo en la detección de cualquier enfermedad.

Los proyectos que hemos empleado para documentarnos en este trabajo nuestro son los siguientes:

1. *Decision Support in Heart Disease Prediction System using Naive Bayes:* este proyecto fue publicado por el *Indian Journal of Computer Science and Engineering* en 2011 y el objetivo del proyecto era el de predecir que pacientes tienen más probabilidad de sufrir una enfermedad cardíaca. En este experimento la técnica empleada fue Naive Bayes (Bayes ingenuo) obteniéndose unos resultados muy buenos.

   (G.Subbalakshmi, K. Ramesh, & M. Chinna Rao, 2011)

2. *Dengue Disease Prediction Using Weka Data Mining Tool:* este proyecto fue llevado a cabo por la Universidad Central Jamia Millia Islamia de Nueva Delhi, India. Este proyecto pretendía detectar a los pacientes potenciales de sufrir una enfermedad mortal

como lo es el dengue. Este proyecto empleó cinco técnicas distintas: Naive Bayes, árboles de decisión (C4.5, Random tree, REP tree) y SMO. Como podemos leer en el documento, los mejores resultados se obtuvieron con el Naive Bayes y el algoritmo C4.5. Además de esto, observamos que el software empleado fue el Weka.

(Ara Shakil, Anis, & Alam, 2015)

3. *Classification of Multi-dimensional Thyroid Dataset Using Data Mining Techniques: Comparison Study:* este otro proyecto llevado a cabo por la Universidad de Ingeniería de Tiruchirappalli, una de las ciudades más grandes de lal estado de Tamil Nadú en la India. Este proyecto también trata sobre las múltiples dimensiones de la tiroides y emplea las siguientes técnicas:Naive Bayes, KNN, árboles de decisión y CN2. En este trabajo también se trabaja primero con el grueso de los datos, llevando a cabo una reducción de la dimensionalidad más tarde. Tras esto, se ve que no hay diferencias significativas entre un caso y otro.

(Senthilkumar, Sheelarani , & Paulraj, 2015)

	G.Subbalakshmi, K. Ramesh, & M. Chinna Rao, 2011	Ara Shakil, Anis, & Alam, 2015	Senthilkumar, Sheelarani , & Paulraj, 2015
Naive Bayes, Red bayesiana	**X**	**X**	**X**
KNN			**X**
Árbol de Decisión (C4.5, REP tree, Random tree, CART, Decision Stump		**X**	**X**
Red Neuronal			
SMO		**X**	
CN2			**X**

Tabla 2. Proyectos y técnicas empleadas.(Elaboración propia, 2016)

A partir de la tabla anterior podemos ver cuáles son las técnicas que más se emplean en los proyectos similares al nuestro. Vemos que los árboles de decisión y el Naive Bayes son los más utilizados obteniéndose, además, unos resultados excelentes. Ya que hemos visto que Naive Bayes se emplea en todos, hemos buscado los motivos por los que ocurre esto. Las causas encontradas para justificar el uso de este algoritmo son las siguientes: es muy útil cuando trabajamos con gran cantidad de datos, cuando los atributos son independientes entre ellos y cuando queremos comparar la eficiencia de este algoritmo con la de otra técnica. Con los otros cuatro algoritmos vemos que empatan a la hora de ver cuánto se usan, por lo que se ha decidido que emplearemos el KNN ya que es una técnica muy fiable y fácil de entender para una persona que no conozca esto algoritmos.

3.3. Técnicas predictivas empleadas

3.3.1 Naive Bayes

Para explicar a fondo este teorema debemos ir hacia atrás hasta llegar al Teorema de Bayes y a la Regla de Laplace. Comenzaremos con la Regla de Laplace, esta norma dice que si realizamos un experimento aleatorio en el que hay N sucesos elementales equiprobables (todos igualmente probables), la probabilidad de que ocurra un suceso se calcula como el cociente del número de casos favorables entre el número de resultados posibles. Tomando como caso favorable el caso A, en la imagen siguiente vemos la ecuación:

$$P(A) = \frac{número\ de\ casos\ favorables\ a\ A}{número\ de\ casos\ posibles}$$

Ahora vamos a pasar a explicar el Teorema de Bayes, teorema en el que se basa el algoritmo de Naive Bayes. Este teorema, que fue planteado por Thomas Bayes en 1763, es de increíble relevancia ya que vincula la probabilidad de A dado B y la de B dado A (condicionalidad de A y B). Con la probabilidad condicionada se busca obtener la probabilidad de que ocurra un suceso A, sabiendo que ya ha sucedido un suceso B. Y la fórmula que define esta probabilidad condicionada es el cociente de la probabilidad de que ocurran ambos eventos (intersección de A y B) entre la probabilidad de que ocurra el que ya sabemos que ha ocurrido, B.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

A partir de la Regla de Laplace y la de la probabilidad condicionada se puede explicar el Teorema de Bayes y ver de dónde sale dicha norma. A continuación, vamos a ver el desarrollo de lo explicado hasta llegar al Teorema de Bayes alcanzado por el matemático inglés, Thomas Bayes.

$$P(A \cap B) = P(A|B)P(B)$$

$$P(B \cap A) = P(B|A)P(A)$$

Como la intersección de A y B tiene que ser igual a la intersección de B y A, podemos igualar las dos ecuaciones anteriores, quedándonos:

$$P(A|B)P(B) = P(B|A)P(A)$$

Despejando en la ecuación anterior llegamos al Teorema de Bayes:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Con esta ecuación que hemos alcanzado combinando las 3 reglas anteriores es con la que trabaja el teorema de Naive Bayes o de Bayes Ingenuo. Este teorema es un clasificador muy fácil de construir que con pocos parámetros introducidos puede constituir una herramienta muy potente para aplicarlos a grandes bases de datos. A pesar de su simplicidad, el clasificador NB obtiene resultados sorprendentemente buenos y es utilizado en infinidad de proyectos de minería de datos.

Ilustración 23. Retrato de Thomas Bayes. (Dressler, 2013)

Este algoritmo asume que la presencia o ausencia de una determinada propiedad no tiene relación alguna con la presencia o ausencia de cualquier otra propiedad. Es esto lo que le da el adjetivo "Naive" (ingenuo) al algoritmo. Esta independencia entre las variables no es siempre cierta, sin embargo, el método ha sido fructífero debido a que la información relevante se encuentra en las magnitudes relativas entre las cantidades y no en los valores de las probabilidades en sí. Son modelos fácilmente entrenables para aplicarle las técnicas descriptivas con las que se le va a tratar. Se considera un algoritmo fácilmente entrenable ya que no necesita de gran cantidad de datos para su entrenamiento.

Lo que hace tan bueno este modelo es su simplicidad y lo fácil que es de explicar y de entender para cualquier público, ya sea un público especializado en la estadística, la minería de datos o cualquier otro tipo de ámbito.

3.3.2 K - Nearest Neighbour (KNN)

El algoritmo KNN (K vecinos más cercanos) es otro de los algoritmos más simples de entender pero trabaja espectacularmente bien en la práctica.

"El KNN es un algoritmo *de aprendizaje vago no parametrizado"*que clasifica a los diferente sujetos comparándolos con los K vecinos más cercanos según la distancia que se tome como medida de uso.

(Thirumuruganathan, 2010)

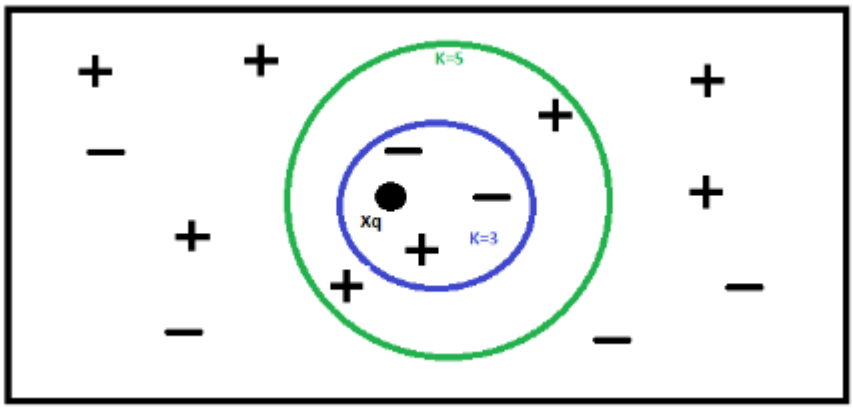


Ilustración 24. Algoritmo KNN. (Elaboración propia, 2016)

Vamos a proceder a explicar que significa que un algoritmo es *vago no parametrizado*. Cuando decimos que un algoritmo es *no parametrizado*, queremos decir que no se hace ningún supuesto sobre la distribución de los datos con los que se trabaja. Este término hace que sea un algoritmo muy útil, ya que, en los casos prácticos reales, pocas veces se obedecen a las suposiciones teóricas típicas enunciadas.

También hemos dicho que se trata de un algoritmo *vago*, con esto nos referimos a que éste no suele necesitar entrenamiento de los datos para sacar un patrón. En otras palabras, no existe una fase de entrenamiento explícita o, si la hay, es mínima. Esta falta de generalización nos indica que todos los datos de entrenamiento son necesarios durante la fase de testeo.

Se puede apreciar claramente que la fase de entrenamiento es nula o mínima pero que, sin embargo, la fase de testeo será una fase costosa. El coste es acentuado tanto en tiempo como en memoria.

KNN tiene varias suposiciones que se deben comentar. Con KNN se asume que los datos están en un espacio medible; los datos pueden ser escalares o, incluso, vectores multidimensionales. Se debe tener siempre una referencia para medir las distancias entre las diferentes instancias con las que contamos y así saber con cuáles de su alrededor hay que compararlas. La distancia más usada es la euclídea, por eso es la que se ha empleado en nuestro proyecto. Además, para comprobar que según la distancia empleada varían los resultados, también hemos utilizado la distancia Manhattan. Más adelante explicaremos como se miden dichas distancias para cada caso.

El número K es el número con el que nosotros debemos de jugar hasta ver para que K se obtienen los mejores resultados. Este número decide cuántos vecinos influyen a la hora de clasificar a los nuevos parámetros para testarlos. Se suelen probar números impares para evitar posibles empates, es decir, si tomamos una K=2 dentro de un experimento de clasificación para "positivo" o "negativo", estamos diciendo que vamos a comparar a nuestro sujeto con los dos vecinos más cercanos y puede darse el caso en el que uno de ellos sea "positivo" y el otro "negativo", lo que llevaría a un empate. Para evitar esto siempre daremos valores impares eliminando así esta incertidumbre.

A continuación, vamos a pasar a detallar las dos distancias con las que se ha trabajado en este proyecto.

3.3.2.1 Distancia euclídea.

La distancia euclídea es la distancia que hay entre dos puntos del espacio medida con una regla. Se deduce a partir del Teorema de Pitágoras.

Si nos encontramos en un espacio bidimensional, la distancia euclídea entre dos puntos P_1 (X_1,Y_1) y P_2 (X_2,Y_2) vendría definida por la siguiente ecuación:

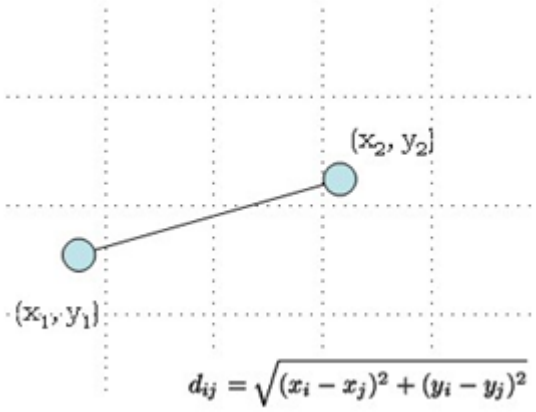


Ilustración 25. Ecuación y representación gráfica de la distancia euclídea. (Sancho Caparrini, 2013)

Esta ecuación anterior la podemos generalizar para el caso de un espacio n-dimensional, quedando de la siguiente manera:

$$d_E(P,Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

3.3.2.2 Distancia Manhattan

Esta distancia fue considerada por Hermann Minkowski en el siglo XIX, se trata de una nueva métrica en la que la distancia entre dos puntos del espacio se calcula como la suma de las diferencias absolutas de sus coordenadas. Es una peculiaridad de la distancia de Minkowski, esta distancia de Minkowski se define mediante la siguiente ecuación:

$$d_{m_q}(i,j) = (\sum_{k=1}^{p} |X_{ik} - X_{jk}|^q)^{1/q} \quad , q > 0$$

La distancia de Manhattan sale del caso en el que q=1 en la ecuación anterior, quedando la expresión de la manera que vemos a continuación:

$$d_{m_1}(i,j) = \sum_{k=1}^{p} |X_{ik} - X_{jk}|$$

3.3.3 Árbol de decisión generado con C4.5

C4.5 es un algoritmo que se emplea para la generación de árboles de decisión. "Quinlan (1993) propone una mejora del algoritmo ID3, al que denomina C4.5", el algoritmo ID3 también fue desarrollado por éste ingeniero australiano. Los árboles de decisión generados con C4.5 se utilizan para clasificación.

(Larranaga, Inza, & Moujahid)

Las principales características del algoritmo son las descritas a continuación:

- Es posible trabajar con valores continuos en los atributos, ya que separa los posibles resultados en 2 ramas por encima y por debajo de X, siendo X un umbral previamente seleccionado.
- Los árboles son menos densos debido a que cada hoja se encarga de una distribución de clases, no de una clase en concreto.
- Entrena los datos para generar el árbol utilizando el método "divide y vencerás".
- Basado en el criterio de proporción de ganancia.
- Es recursivo.

La forma de trabajar de este algoritmo considera todas las posibles opciones en las que se pueda dividir el conjunto de datos y se queda con la opción en la que haya generado la mayor ganancia de información. Para los atributos discretos se hace una prueba con n resultados, siendo n la cantidad de valores que puede tomar el atributo en cuestión. Si trabajamos con atributos continuos, se realiza una prueba binaria (1,0) sobre cada valor que el atributo toma en los datos. En cada nodo, se debe decidir que patrón se elige para dividir los datos. Se pueden tomar tres posibles patrones para el C4.5:

1. Un resultado y una rama para cada valor que pueda tomar la variable (discreta). Esta opción es conocida como la prueba estándar.
2. Una prueba un poco más compleja, que también trata con variables discretas, es aquella en la que los valores posibles son asignados a un número de grupos que varía con un resultado posible para cada grupo, y no de para cada valor como anteriormente.Si se trabaja con variables continuas, se realiza una prueba binaria con dos rangos de resultados, un rango menor o igual que X, A<=X, y otro por encima de X, A>X; siendo X un valor límite que se debe determinar.

Otra restricción que se suele añadir es la de que, para cualquier división, al menos dos de los subconjuntos Ci contengan un mínimo número de casos razonables, evitando así infinidad de subdivisiones que no aporten ninguna información al árbol y que lo único que consiguen es ensuciar el árbol con más ramas inservibles.

3.4. Proceso práctico en el software de minería de datos sin preprocesado

Ahora vamos a pasar a comentar los resultados proporcionados por el software de minería de datos empleado. Como vamos a observar a continuación con la explicación de los resultados vamos a ver que da una información buenísima y una precisión muy alta de acierto a la hora de predecir si una persona está sufriendo o va a sufrir hipotiroidismo.

Aplicando el algoritmo Naive Bayes podemos ver que clasifica bien en un 97,91% de los casos, lo que deja un error del 2,09%. La imagen que vemos a la derecha refleja la matriz de confusión generada por el propio Weka. En esta matriz podemos ver cuántos casos ha catalogado correctamente como positivos y cuáles negativos y en cuántos casos se ha cometido error. En este caso en concreto, podemos ver que 117 casos los ha catalogado como positivo correctamente mientras que otros 32 casos los ha catalogado como positivos siendo negativos en realidad. Por otro lado, vemos que 2980 casos negativos han sido acertados correctamente, dando otros 34 casos como negativos cuando éstos sí que eran positivos. Como breve resumen, decir que los casos en los que el algoritmo acierta se reflejan en la diagonal principal de la matriz, quedando los errores en la diagonal secundaria. El tiempo de ejecución de este algoritmo fue en torno a 0.23 segundos.

```
   a    b   <-- classified as
 117   34 |    a = hypothyroid
  32 2980 |    b = negative
```

Ilustración 26. Matriz de confusión generada por Weka con el algoritmoNaive Bayes. (Elaboración propia, 2016)

KNN (llamado IBK en Weka) con la distancia euclídea da el mejor resultado utilizando un número K igual a tres. El porcentaje de acierto es de 97,38%, dejando un error del 2,62%. Observando la matriz de confusión vemos como se han generado estos porcentajes. Ha clasificado bien 81 casos de hipotiroidismo mientras que ha dado 13 falsos positivos. Por otro lado, ha acertado al asignar 2.999 casos como negativos, clasificando erróneamente otros 70 como negativos (siendo positivos). El tiempo de ejecución de este algoritmo se encuentra en torno a 0.12 segundos.

```
   a    b   <-- classified as
  81   70 |    a = hypothyroid
  13 2999 |    b = negative
```

Ilustración 27. Matriz de confusión generada por Weka con el algoritmoKNN con la distancia euclídea. (Elaboración propia, 2016)

KNN (IBK) con la distancia de Manhattan da un resultado mejor que con la distancia euclídea, aunque cabe destacar que en este caso da el mejor resultado con una K distinta a la elegida en la opción anterior. Para esta cuestión hemos visto que el mejor porcentaje de acierto se encontraba con una K igual a 5, dando un éxito del 97,57%. En la matriz de confusión podemos ver cómo este algoritmo ha errado en 77 casos, acertando en 3086 tal y como vemos sumando las cantidades de los términos que se encuentran en la diagonal principal. El tiempo de ejecución de este algoritmo fue en torno a 0.10 segundos.

```
   a    b   <-- classified as
  81   70 |    a = hypothyroid
   7 3005 |    b = negative
```

Ilustración 28. Matriz de confusión generada por Weka con el algoritmoKNN con la distancia Manhattan. (Elaboración propia, 2016)

Finalmente, con el árbol de decisión generado por el algoritmo C4.5 (llamado J48 en Weka), el algoritmo más sofisticado de los que hemos empleado, podemos ver que se consiguen los mejores resultados, aún sabiendo que los tres casos anteriores también nos han dado resultados muy buenos. Aquí podemos ver la matriz de confusión de este algoritmo en el que se ha dado un acierto del 99,24% y, por lo tanto, un error muy pequeño. Este algoritmo solo se confundiría en el 0,76% de las ocasiones, lo que sería en 24 casos solo de confusión, la suma de 11 más 13. Con este algoritmo C4.5, el Weka ha generado un árbol de decisión de siete hojas y tiene un tamaño de 13, llevándole esto un tiempo de 0.23 segundos. A continuación se puede ver la representación tanto codificada como gráfica del árbol en sí. En primer lugar se va a mostrar la representación gráfica:

```
  a    b   <-- classified as
138   13 |    a = hypothyroid
 11 3001 |    b = negative
```

Ilustración 29. Matriz de confusión generada por Weka con el algoritmoC4.5 (J48). (Elaboración propia, 2016)

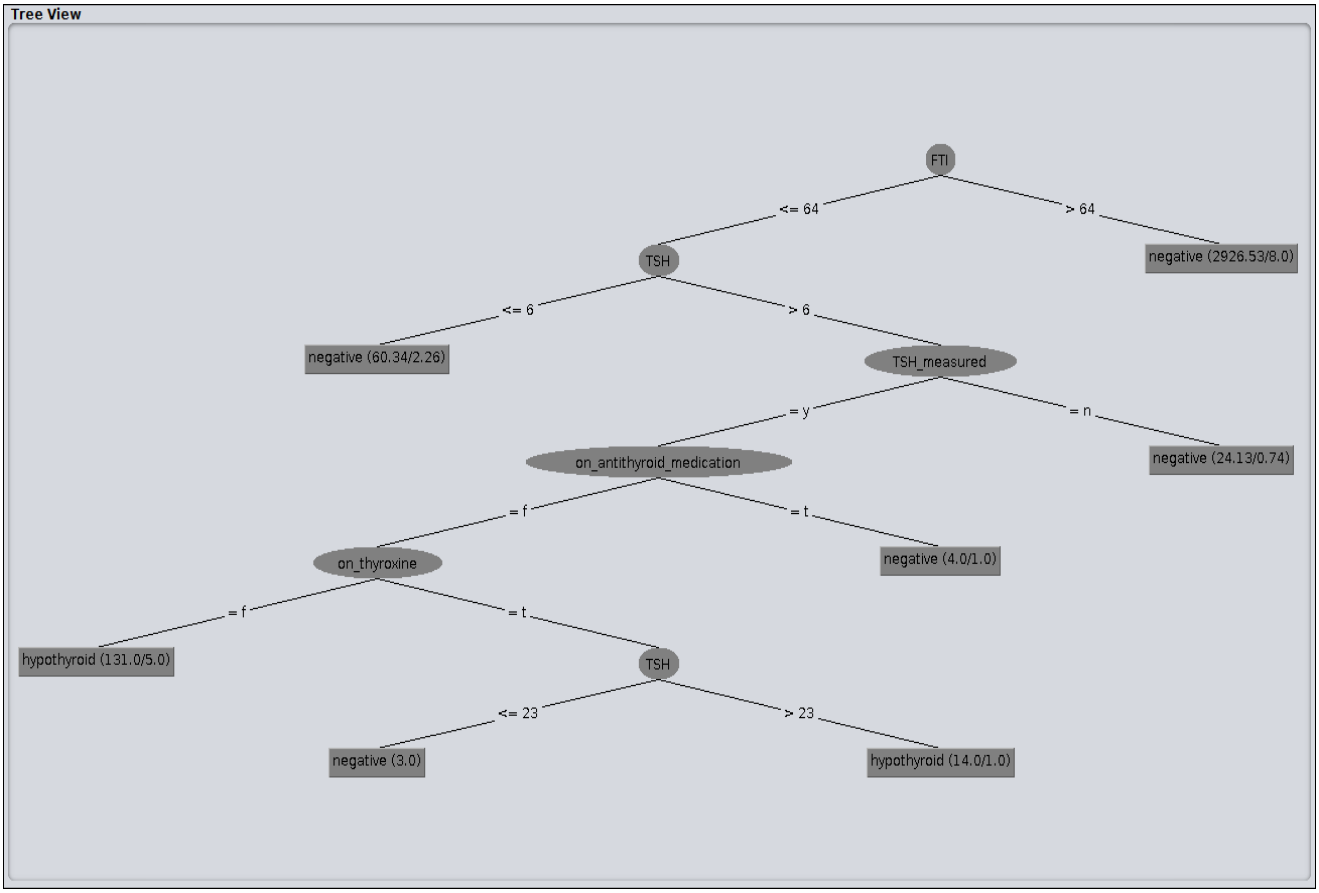


Ilustración 30. Árbol de decisión generado por Weka.(Elaboración propia, 2016)

Una vez vemos el árbol representado, ahora nos disponemos a plasmar el árbol codificado:

```
FTI <= 64
|   TSH <= 6: negative (60.34/2.26)
|   TSH > 6
|   |   TSH_measured = y
|   |   |   on_antithyroid_medication = f
|   |   |   |   on_thyroxine = f: hypothyroid (131.0/5.0)
|   |   |   |   on_thyroxine = t
|   |   |   |   |   TSH <= 23: negative (3.0)
```

```
|  |  |  |  |  TSH > 23: hypothyroid (14.0/1.0)
|  |  |  on_antithyroid_medication = t: negative (4.0/1.0)
|  |  TSH_measured = n: negative (24.13/0.74)
FTI > 64: negative (2926.53/8.0)
```

Tras haber visto los resultados obtenidos con los distintos métodos, podemos decir que contamos con unos datos iniciales muy precisos y correctos, ya que todos los algoritmos dan grandes probabilidades de éxito. Lógicamente, podemos ver a simple vista que el mejor clasificador para predecir a los posibles pacientes es el árbol de decisión, ya que alcanza una eficacia del 99%.

3.5. Preprocesado de los datos

Una vez realizado el análisis de los datos con el software, nos hemos planteado la posibilidad de omitir los atributos que no fueran determinantes a la hora de generar los patrones de decisión.

Esta tarea es bueno realizarla ya que cuando se trabajan con bases de datos muy densas hay que intentar minimizarlas todo lo que se pueda dentro de las posibilidades. ¿Por qué es aconsejable minimizar los datos al máximo? La razón por la que se minimizan los datos es por el tiempo que puede llevar el análisis de los datos por parte del software cuando se tienen millones de instancias con su infinidad de atributos, y un coste en tiempo siempre implica un coste económico.

En el caso de nuestra base de datos, no se trata de una base de datos tan grande como para que sea estrictamente necesario aplicarle un preprocesado, sin embargo, hemos considerado conveniente hacerlo para mostrar un ejemplo de cómo se realiza esto y para ver los resultados que da una vez se ha eliminado esta información que hemos considerado irrelevante. Los resultados normalmente, tras un preprocesado, suelen ser un poco peores ya que algo de información siempre estamos obviando. Aún con este empeoramiento (siempre que no sea muy acentuado), es preferible minimizar los datos perdiendo algo de efectividad debido al gran ahorro de tiempo que puede suponer como hemos descrito anteriormente.

Para llevar a cabo el preprocesado hemos empleado la pestaña que el Weka tiene destinada a esta función, "Select attributes". En esta pestaña tenemos que seleccionar dos cosas: el evaluador del atributo y el método de búsqueda. En el primero de los dos hemos elegido "CfsSubsetEval", un evaluador que devuelve el número de veces que un atributo ha sido seleccionado en la validación cruzada. Este evaluador junto con el método de búsqueda seleccionado ("BestFirst"), selecciona los atributos que aportan más información aplicándole cambios a cada uno de los atributos y va midiendo el impacto que éstos tienen en relación con la clase seleccionada. El orden de los atributos escogidos es aquellos que tienen un mayor impacto respecto a la clase, teniendo éstos una relación directa o inversamente proporcional con la clase. Este software, basándose en las normas descritas anteriormente, ha seleccionado como atributos más relevantes y determinantes a la hora de clasificar las instancias los seis siguientes más la clase que, lógicamente, hay que tenerla en cuenta para poder llevar a cabo la clasificación:

1. Si ha sufrido cirugía en la glándula tiroides o no.
2. Si ha sufrido tumor o no.
3. Niveles de TSH medidos.
4. Niveles de T3 medidos.
5. Niveles de FTI medidos.
6. Si se ha medido los niveles de TBG, Globulina Fijadora de Tiroxina.

3.6. Proceso práctico en el software de minería de datos con preprocesado

Una vez hemos realizado el preprocesado de los datos, vamos a pasar a comentar los resultados que nos ha dado el software aplicándole, a la base de datos reducida, las mismas técnicas que a la base de datos sin el preprocesado. Vamos a observar si los resultados son los esperados o si, por el contrario, vemos algo que salga de la normalidad.

Vamos a empezar comentando el Naive Bayes. En esta ocasión, el Naive Bayes nos da una efectividad del 97,72% y, por consiguiente, sólo un 2,28% de error. Anteriormente, nos dio una efectividad del 97,91%, prácticamente nos ha dado lo mismo por lo que se encuentra dentro de los baremos esperados. Incluso ha empeorado un poco, consecuencia de lo que comentábamos antes cuando decíamos que algo de información siempre se pierde por lo que los resultados pueden empeorar un poco. Cierto es que este 0,2% que se ha perdido en este caso no lo podríamos considerar un empeoramiento ya que es ínfimo. También podemos ver que el tiempo de ejecución ha disminuido hasta los 0.13 segundos tal y como esperábamos.

```
  a    b   <-- classified as
104   47 |    a = hypothyroid
 25 2987 |    b = negative
```

Ilustración 31. Matriz de confusión generada por Weka con el algoritmoNaive Bayes. (Elaboración propia, 2016)

Con el KNN vamos a ver algo que sí puede ser significativo, ya que además de variar los porcentajes de acierto y error como es lógico, también va a variar el número de instancias cercanas con las que se comparará cada una (el número K). Para el caso en el que hemos empleado la distancia euclídea, tras múltiples iteraciones hemos visto que los mejores resultados han salido cuando la K es igual a 21 (antes era K=3). Ahora el porcentaje de acierto ha sido de 95,83%, empeorando el caso sin preprocesado en un 2% aproximadamente. Comparando ambas matrices de confusiones podemos ver reflejados estos porcentajes viendo el número de aciertos y fallos en cada caso. También podemos ver en este caso como el tiempo de ejecución a disminuido de 0.12 segundos a 0.08 segundos.Vemos que estos cambios también se encuentran dentro de los resultados esperados que podíamos esperar.

```
  a    b   <-- classified as
 96   55 |    a = hypothyroid
 77 2935 |    b = negative
```

Ilustración 32. Matriz de confusión generada por Weka con el algoritmoKNN con la distancia euclídea. (Elaboración propia, 2016)

Aplicando el KNN con la distancia de Manhattan alcanza su mayor eficiencia cuando se compara con los 53 vecinos más cercanos al caso estudiado. En esta ocasión, ha variado de cinco a 53. Aquí los porcentajes de éxito no difieren mucho tampoco, pasa del 97,57% anterior a un 97,06% ahora, apenas varía un 0,5%; cantidad irrelevante a la hora de analizarlo pero que podemos ver que siguen estando dentro de los resultados esperados y que daríamos como buenos. El tiempo de ejecución también ha pasado a ser en torno a 0.08 segundos

```
  a    b   <-- classified as
 78   73 |    a = hypothyroid
 20 2992 |    b = negative
```

Ilustración 33. Matriz de confusión generada por Weka con el algoritmoKNN con la distancia Manhattan. (Elaboración propia, 2016)

En el último caso, empleando el algoritmo C4.5 para generar un árbol de decisión, vemos que el resultado que nos da el software, una vez le hemos realizado el preprocesado, sigue dentro de los parámetros esperados. La efectividad de este algoritmo apenas baja un 0,4%; de un 99,24% a un 98,89% ahora. Haciendo una comparativa entre las dos matrices de confusión podemos ver que si antes sólo se equivocaba al catalogar 24 casos, ahora lo hace en 35 de estos. Este incremento de 11 unidades no supone prácticamente nada ya que estamos hablando de que tenemos más de 3.000 instancias. El tiempo de ejecución ha variado de 0.23 segundos a 0.17 segundos, disminución que entra dentro de lo que esperábamos. Para este algoritmo también podemos comparar tanto el código que el Weka ha generado como la representación gráfica. En ambos casos observaremos que en los generados tras el preprocesado se nota una gran simplificación de los mismos. A continuación, se puede ver el árbol generado esta vez, en el que el número de hojas baja de siete a cuatro y su tamaño decrece de 13 a siete, lo que hace una representación menos frondosa y más clara del caso de estudio.

```
  a    b   <-- classified as
138   13 |    a = hypothyroid
 22 2990 |    b = negative
```

Ilustración 34. Matriz de confusión generada por Weka con el algoritmoC4.5 (J48). (Elaboración propia, 2016)

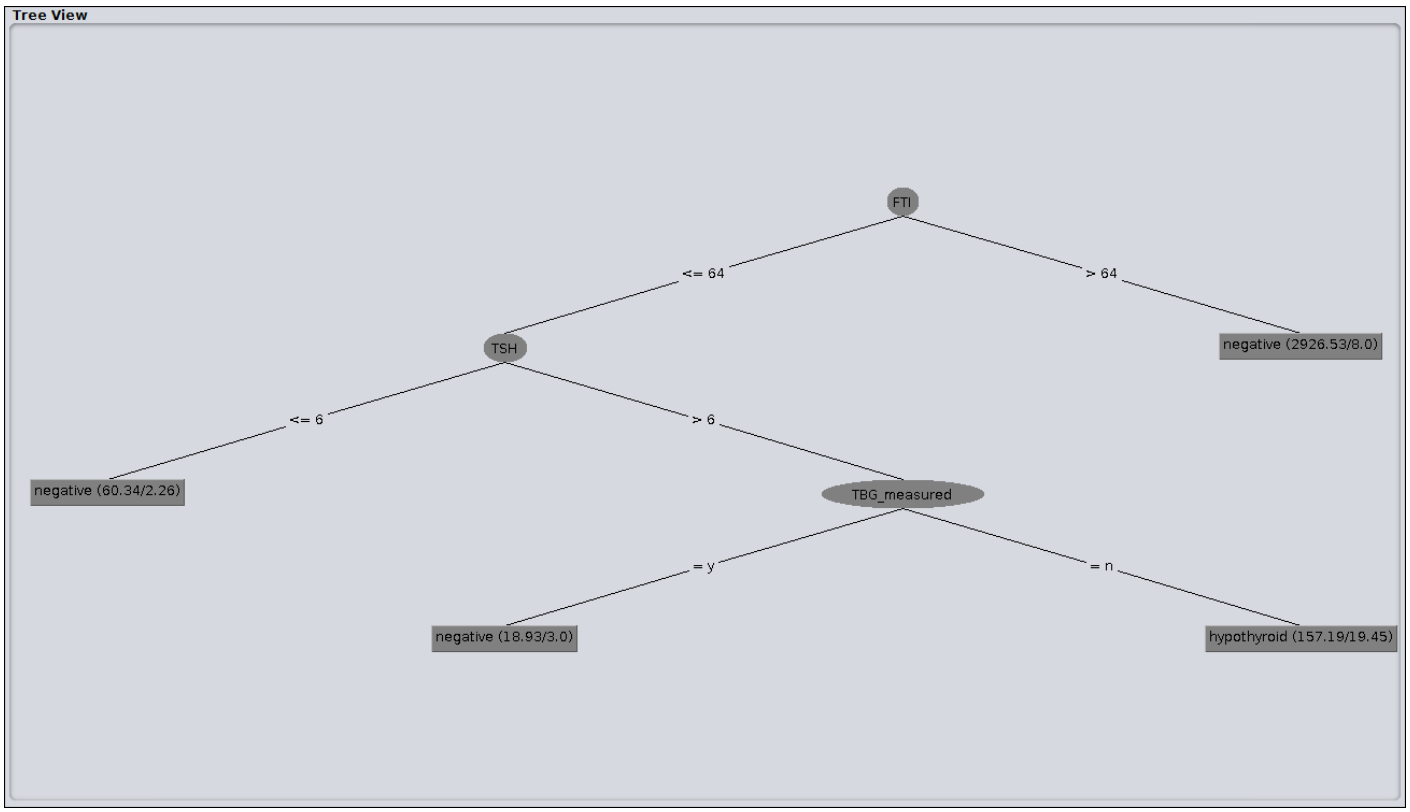


Ilustración 35. Árbol de decisión generado por Weka. (Elaboración propia, 2016)

Una vez vista la representación gráfica se va a mostrar el código en el que también podemos apreciar la disminución de la complejidad ya que es muy parecido y simple pero con menos líneas de código.

```
FTI <= 64
|   TSH <= 6: negative (60.34/2.26)
|   TSH > 6
|   |   TBG_measured = y: negative (18.93/3.0)
```

```
|  |  TBG_measured = n: hypothyroid (157.19/19.45)
FTI > 64: negative (2926.53/8.0)
```

3.7. Conclusión

Como comentarios finales a los resultados obtenidos hay varias cosas a mencionar una vez que hemos visto los exitosos resultados que hemos hallado empleando tres técnicas distintas.

En primer lugar, se van a comentar los resultados hallados con el experimento realizado antes de aplicarle el preprocesado a los datos. En este caso, vemos claramente que el algoritmo que proporciona los mejores resultados es el algoritmo C4.5, llamado J48 en Weka. Tanto el KNN (llamado IBK en Weka), con la distancia euclídea y la distancia Manhattan, como el Naive Bayes proporcionan unos resultados muy buenos y que podrían sernos útiles para un posible estudio más a fondo ya que alcanzan el 97% de acierto. Sin embargo, no son comparables con el árbol de decisión creado con el algoritmo C4.5 que vemos que llega al 99% de acierto.

	Naive Bayes	KNN con distancia euclídea (K=3)	KNN con distancia Manhattan (K=5)	Árbol de decisión (C4.5)
Acierto	97,91%	97,38%	97,57%	99,24%
Error	2,09%	2,62%	2,43%	0,76%

Tabla 3. Resultados de Weka sin preprocesado. (Elaboración propia, 2016)

Una vez se había realizado el análisis de todas las instancias con sus respectivos atributos, pensamos que sería interesante ver que atributos eran los más decisivos a la hora de decidir si un paciente presenta hipotiroidismo o no. En este caso, podríamos haber elegido entre varios algoritmos. Nosotros hemos seleccionado el algoritmo que selecciona los atributos según el número de veces que haya sido seleccionado un atributo en la validación cruzada, lo que hace que sean los atributos más decisivos para la clase elegida (si padece hipotiroidismo o no).

Antes de comenzar el análisis de los datos, debemos decir que los resultados que se esperan, en estos casos en los que ya tenemos unos resultados sin preprocesado anterior, suelen ser peores que los que se obtienen sin el preprocesado. Esto se debe a que al eliminar ciertos atributos estamos eliminando información útil, no determinante, pero sí útil. Sin embargo, esta pérdida de eficacia no es un gran contratiempo (siempre que hablemos de una pérdida liviana) porque al eliminar atributos estamos consiguiendo aligerar nuestra base de datos, lo que repercute en un ahorro de tiempo que es de agradecer cuando trabajamos con millares de instancias.

Pasando a comentar los resultados que hemos obtenido con el preprocesado, lo primero que podemos observar es que ha sucedido lo que esperábamos. El porcentaje de acierto ha disminuido en todos los casos de estudio, cumpliéndose lo que ya adelantábamos antes. Seguimos viendo que el algoritmo que nos ha dado los mejores resultados es el árbol de decisión generado con C4.5, ya que sigue rozando el 99% de acierto (98,89% exactamente). Vemos que se ha perdido menos de un 0,5%, lo que podemos dar por una pérdida asumible. Con las otras técnicas empleadas también vemos que la pérdida es muy leve ya que Naive Bayes y el KNN con la distancia Manhattan siguen sin bajar del 97%, y el KNN con la distancia euclídea es el algoritmo que más nota este preprocesamiento bajando su porcentaje de acierto hasta un 95,83%.

	Naive Bayes	KNN con distancia euclídea (K=21)	KNN con distancia Manhattan (K=53)	Árbol de decisión (C4.5)
Acierto	97,72%	95,83%	97,06%	98,89%
Error	2,28%	4,17%	2,94%	1,11%

Tabla 4. Resultados de Weka con preprocesado. (Elaboración propia, 2016)

Una vez hemos comentado los cambios de eficacia de los algoritmos antes y después de llevar al cabo el preprocesado, ahora vamos a pasar a mostrar en una tabla los cambios de tiempos de ejecución donde vamos a poder ver claramente como estos tiempos han disminuido al eliminar la información no tan importante para la predicción de la enfermedad.

	Naive Bayes	KNN con distancia euclídea	KNN con distancia Manhattan	Árbol de decisión (C4.5)
Sin preprocesado	0,23	0,12	0,1	0,23
Con preprocesado	0,13	0,08	0,08	0,17

Tabla 5. Comparativa de tiempos de ejecución. Tabla en segundos. (Elaboración propia, 2016)

Una vez comentados todos los resultados, podemos decir con total seguridad que la minería de datos es una herramienta muy fiable a la hora de detectar el hipotiroidismo siempre que contemos con los datos oportunos y que es conveniente aplicarle un preprocesado previo para así aligerar la base de datos y disminuir el tiempo de ejecución de los cálculos por el software.

4 Bibliografía

(s.f.). Recuperado el 20 de Julio de 2016, de http://www.sc.ehu.es/towsoesj/Biologia%20Educacion/Web%20HORMONAS/HORMONAS/tiroxina.htm

(s.f.). Recuperado el 20 de Julio de 2016, de http://hipertiroidismo.org/tiroxina/

Cárdenas-Montes, M. (s.f.). *Medidas de Distancia.* Obtenido de http://wwwae.ciemat.es/~cardenas/docs/lessons/MedidasdeDistancia.pdf

Fehrman, E., Muhammad, A., Mirkes, E., Egan, V., & Gorban, A. (s.f.). *The Five Factor Model of personality and evaluation of drug consumption risk.*Nottingham, Leicester. Obtenido de https://arxiv.org/ftp/arxiv/papers/1506/1506.06297.pdf

¿Qué es Giahsa? (2012). *Revista aguas*. Obtenido de http://www.revistaaguas.es/que-es-giahsa/
(2009). *Algoritmo "Naive Bayes"*. Obtenido de http://algoritmosmineriadatos.blogspot.com.es/2009/12/algoritmo-naive-bayes.html

Bouckaert, R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2016). *WEKA Manual for Version 3-8-0.* MANUAL, University of Waikato, New Zealand.

Córdoba Fallas, L. (s.f.). Obtenido de http://cor-mineriadedatos.blogspot.com.es/2011/06/weka.html

Cravero Leal, A., & Sepúlveda Cuevas , S. (2009). Aplicación de Minería de Datos para la Detección de Anomalías: Un Caso de Estudio. *WORKSHOP INTERNACIONAL EIG2009*. Obtenido de http://ceur-ws.org/Vol-558/Art_8.pdf

Dressler, M. (Julio de 2013). Thomas Bayes y las sutilezas de la estadística. *Investigación y ciencia*. Obtenido de http://www.investigacionyciencia.es/revistas/investigacion-y-ciencia/numero/442/thomas-bayes-y-las-sutilezas-de-la-estadstica-11212

Félix, L. C. (2002). Data mining: torturando a los datos hasta que confiesen. *UOC*. Obtenido de http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.pdf

Galofré Ferrater, J. (s.f.). *Clínica Universidad de Navarra*. Obtenido de http://www.cun.es/enfermedades-tratamientos/enfermedades/hipotiroidismo

García Morate, D. (s.f.). *MANUAL DE WEKA.* Manual. Obtenido de http://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/InteligenciaDeNegocio/weka.pdf

Giahsa. (s.f.). Recuperado el 5 de Julio de 2016, de http://www.giahsa.com/wps/portal/giahsa/Conoce-Giahsa/Quienes%20Somos/Historia/!ut/p/z1/rZZfe5owFMa_Sm-8jEkggbA7t7UoU9vibGtufAIGZJM_Ata1n35hPs-mpcJ8CneB9_1xcnJOEsjhE-SJeI5CUUZpIjZqvODGcjjEo6FN0NjWZwQN7q07615zEUMYPr4RuKamBJOvD0w3kU0MyM_7Z65Z89d-UPnRmWeA4AP

Giahsa. (2012). *REGLAMENTO DEL SUMINISTRO DOMICILIARIO DE AGUA.* Obtenido de https://www.giahsa.com/wps/wcm/connect/a2fbba79-e070-4ffb-982c-3f47fef4c5a2/01_Reglamento%2Bde%2BSuministro%2BGIAHSA.pdf?MOD=AJPERES

Grané, A. (s.f.). *Distancias estadísticas y Escalado Multidimensional (Análisis de Coordenadas Principales).* Universidad Carlos III de Madrid, Departamento de Estadística, Madrid. Obtenido de http://halweb.uc3m.es/esp/Personal/personas/agrane/ficheros_docencia/MULTIVARIANT/slides_Coorp_reducido.pdf

Hall, M., & Reutemann, P. (2008). *WEKA KnowledgeFlow Tutorial*. MANUAL, University of Waikato. Obtenido de http://software.ucv.ro/~eganea/AIR/KnowledgeFlowTutorial-3-5-8.pdf

Han, J., Kamber, M., & Pei, J. (2012). *DATA MINING. Concepts and Techniques*.ELSEVIER.

Hasperué, W. (2013). *Extracción de Conocimiento en Grandes Bases de Datos Utilizando Estrategias Adaptativas.* Obtenido de http://sedici.unlp.edu.ar/bitstream/handle/10915/35555/Documento_completo.pdf?sequence=1

Hearst, M. (1999). *Untangling Text Data Mining*.Obtenido de http://people.ischool.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html

IMSS. (2015). *síntomas del hipotiroidismo.* Obtenido de Cintalapanecos.com

Krall , C. (2006). *Minería de datos (data mining). Qué es y para qué sirve. (2ª parte).* Obtenido de http://www.aprenderaprogramar.com/index.php?option=com_attachments&task=download&id=203

LAB TEST ONLINE. (s.f.). Obtenido de http://www.labtestsonline.es/tests/thyroidantibodies.html?tab=3

Larranaga, P., Inza, I., & Moujahid, A. (s.f.). *Árboles de Clasificación.* Universidad del PaÍs Vasco. Obtenido de http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t10arboles.pdf

Las pruebas de la hormona tiroidea. (12 de Abril de 2013). Obtenido de http://listaexamenesmedicos.blogspot.com.es/2013/04/las-pruebas-de-la-hormona-tiroidea.html

Lichman, M. (2013). *UCI Machine Learning Repository*. Recuperado el 2016, de http://archive.ics.uci.edu/ml

Lima, L., & Vásquez, C. (2013). ESTRATEGIA INTELIGENTE PARA LA DETECCIÓN EFICIENTE DE CLIENTES RESIDENCIALES CON CONDICIONES FRAUDULENTAS DE LAS EMPRESAS DE SERVICIO ELÉCTRICO. *Dialnet*. Obtenido de https://dialnet.unirioja.es/servlet/articulo?codigo=4777894

MedlinePlus. (2016). Recuperado el 21 de Julio de 2016, de https://medlineplus.gov/spanish/ency/article/003684.htm

MedlinePlus. (2014). Recuperado el 21 de Julio de 2016, de https://medlineplus.gov/spanish/ency/article/003374.htm

MedlinePlus. (2014). Recuperado el 21 de Julio de 2016, de https://medlineplus.gov/spanish/ency/article/003517.htm

Mesa, F., Raineri, A., Maturana, S., & Kaempffer, A. (2009). Fraudes a los sistemas de salud en Chile: un modelo para su detección. *Panam Salud Publica*. Obtenido de http://www.scielosp.org/pdf/rpsp/v25n1/09.pdf

Mitchell, T. (2016). *GENERATIVE AND DISCRIMINATIVE CLASSIFIERS: NAIVE BAYES AND LOGISTIC REGRESSION*.Obtenido de https://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

Moreno García, M., Miguel Quintales, L., García Peñalvo, F., & Polo Martín , M. (s.f.). *APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EN LA CONSTRUCCIÓN Y VALIDACIÓN DE MODELOS PREDICTIVOS Y ASOCIATIVOS A PARTIR DE ESPECIFICACIONES DE REQUISITOS DE SOFTWARE.* Universidad de Salamanca. , Departamento de Informática y Automática , Salamanca. Obtenido de http://ceur-ws.org/Vol-84/paper4.pdf

My Weka page. (26 de Julio de 2016). Obtenido de http://www.hakank.org/weka/

Nasser, J. (2016). *¿Cuáles Son Los Niveles Normales De La Hormona TSH?*

Pérez López, C., & Santín González, D. (2007). *Minería de datos: técnicas y herramientas.* Thomson-Paraninfo.

Pérez Marqués, M. (2014). *Minería de datos a través de ejemplos.* Madrid.
(2012). *Redes de Neuronas Artificiales.* UC3M. Obtenido de http://www.lab.inf.uc3m.es/~a0080630/redes-de-neuronas/

Rios Villegas, A., & Uribe Aguirre, K. (2013). *MINERIA DE DATOS APLICADA A LA DETECCIÓN DE CLIENTES CON ALTA PROBABILIDAD DE FRAUDES EN SISTEMAS DE DISTRIBUCIÓN.* Obtenido de http://recursosbiblioteca.utp.edu.co/dspace/bitstream/handle/11059/3856/006312R586.pdf;jsessionid=6C34F2EF442102C4D5B5EB00B61A19CE?sequence=1

Sancho Caparrini, F. (2013). *Mapas semánticos: clasificación y representación.* Obtenido de http://www.cs.us.es/~fsancho/?e=44

Santamaría Ruíz, W. (2010). *MODELO DE DETECCION DE FRAUDE BASADO EN EL DESCUBRIMIENTO SIMBOLICO DE REGLAS DE CLASIFICACIÓN EXTRAÍDAS DE UNA RED NEURONAL*. UNIVERSIDAD NACIONAL DE COLOMBIA, DEPARTAMENTO DE INGENIER´IA DE SISTEMAS E INDUSTRIAL, Bogotá. Obtenido de https://core.ac.uk/download/files/334/11053314.pdf

(s.f.). *Técnicas de Análisis de Datos en WEKA.* Obtenido de http://isa.umh.es/asignaturas/crss/turorialWEKA.pdf

Thirumuruganathan, S. (2010). A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm. Obtenido de https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/

Wanumen Silvaz, L. (2010). *Minería de datos para la predicción de fraudes en tarjetas de crédito.* Obtenido de http://revistas.udistrital.edu.co/ojs/index.php/vinculos/article/viewFile/4162/5825

Weka 3: Data Mining Software in Java. (s.f.). Recuperado el 9 de Julio de 2016, de http://www.cs.waikato.ac.nz/ml/weka/

Ara Shakil, K., Anis, S., & Alam, M. (2015). *DENGUE DISEASE PREDICTION USING WEKA DATA MINING TOOL.*Obtenido de https://arxiv.org/ftp/arxiv/papers/1502/1502.05167.pdf

Fertifarma. (2016). Como Afecta El Hipotiroidismo A La Fertilidad. *FERTIFARMA*. Obtenido de http://www.fertifarma.com/articulos/como-afecta-el-hipotiroidismo-a-la-fertilidad.php

G.Subbalakshmi, K. Ramesh, & M. Chinna Rao. (2011). *Decision Support in Heart Disease Prediction System using Naive Bayes* (Vol. 2). Obtenido de http://www.ijcse.com/docs/IJCSE11-02-02-56.pdf

Martínez Fraga, J. (2012). *Anatomía y Fisiología.* Obtenido de http://www.elmodernoprometeo.es/Sitio_web/Anatomia_files/endocrino.pdf

Ropero Lara, A. (2012). *HORMONAS*. Recuperado el 28 de JULIO de 2016, de http://las-hormonas.blogspot.com.es/2012/10/las-hormonas-del-cerebro.html

Senthilkumar, D., Sheelarani , N., & Paulraj, S. (2015). Classification of Multi-dimensional Thyroid Dataset Using Data Mining Techniques: Comparison Study. *Advances in Natural and Applied Sciences*. Obtenido de http://www.aensiweb.com/old/anas/2015/Special1%20IC/24-28.pdf

Blog sobre Bussiness Intelligence. (s.f.). Minería de datos: aplicaciones más populares a día de hoy. *Blog sobre Bussiness Intelligence*. Recuperado el 17 de Junio de 2016, de http://www.lantares.com/blog/mineria-de-datos-aplicaciones-que-ya-son-una-realidad

Herrera Varela, R. (2006). minería de datos y descubrimiento de conocimiento en bases de datos aplicados al ámbito bibliotecario. (Primera parte). *Forinf@ Online, 33*.

IBM. (s.f.). Obtenido de http://www.ibm.com/analytics/us/en/technology/spss/

jwork.org. (s.f.). Obtenido de http://jwork.org/main/

knime. (s.f.). Obtenido de http://www.knime.org/

Marcel. (2014). *datamashup.info*. Recuperado el 15 de Junio de 2016, de http://www.datamashup.info/what-is-data-mining-video/

Molina Félix , L. (2014). *Data mining: torturando a los datos hasta que confiesen.* Obtenido de http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html

orange. (s.f.). Obtenido de http://orange.biolab.si/

Rapidminer. (s.f.). Obtenido de https://rapidminer.com/