Tema 1: Minería de datos y extracción de conocimiento

1.1.Minería de datos y otras disciplinas
1.2.El proceso de extracción de conocimiento desde los datos: Preparación, modelado, evaluación e interpretación.
1.3.Principales problemas/tareas y métodos/técnicas de la minería de datos: extracción de patrones, clustering, clasificación, predicción y asociación.
1.4.Minería de datos, aprendizaje automático y big data

1.1. Minería de datos y otras disciplinas.

Según Witten , Frank y otros (Data Mining) ,la minería de datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos.

Esta definición apela a la extracción (de ahí el nombre de minería) de conocimiento no explícito a partir de los datos. Y esto parece apuntar a las ideas de que "no hay nada nuevo bajo el sol" o a que se trata de "los mismos perros con distintos (o no) collares". La estadística (multivariante) pretende, algo parecido partiendo de los datos como materia prima. ¿Sólo el hecho de que el conocimiento a extraer es "rotundamente" **no explicito** es la relevante diferencia?. En realidad, es más bien la idea de **extracción**. La minería de datos nace, no tanto del desarrollo de técnicas específicas y diferentes de las anteriores , como de la aparición de nuevas necesidades , en la medida en que la omnipresencia de enormes cantidades de almacenes de datos llenos de enormes cantidades de datos convierten a éstos en una potencial materia prima de un "posible" conocimiento "nuevo".

La minería de datos se convierte en una amalgama de procedimientos y métodos que pretenden abordar una serie de problemas de extracción de conocimiento a partir de grandes volúmenes de datos. Muchos de estos procedimientos tienen una base de estadística tradicional, otros participan de métodos de aprendizaje automático, en la línea de algunos desarrollos de la Inteligencia Artificial, otros buscan la obtención de patrones o pautas a partir de la información disponible, etc. En palabras del propio Witten, acaba siendo una disciplina eminentemente **práctica** y empírica que pretende resolver problemas cognoscitivos de una forma **ecléctica** (no comprometida con una u otra metodología)

Para que se comprenda en qué medida lo importante en minería de datos es la resolución de esta clase de problemas de conocimiento citemos aquí algunos ejemplos muy conocidos pero, también, bastante esclarecedores.

Ej.1 El análisis de Créditos Bancarios (Credit Score)

Una entidad financiera quiere contar con un "sistema" para **saber** si un cliente que ha pedido (o va a pedir) un préstamo lo devolverá en tiempo y forma o no. El banco tiene una gran cantidad de información sobre los clientes y pretende aprovecharla para acabar dotándose de una o varias reglas que le permita(n) discernir si un cliente devolverá o no el préstamo. El Análisis Discriminante (técnica tradicional de la estadística multivariante), la obtención de árboles de decisión , el análisis bayesiano, la redes neuronales y algunos otros métodos de muy distintas procedencias metodológicas pueden emplearse para esta **tarea** . Un tarea que podemos catalogar como de **clasificación**. *La minería de datos no suele descartar ninguna técnica por su filiación y acaba tomando partido por una u otra por razones prácticas de eficiencia*.

Ej.2.Análisis de la cesta de la compra.

Es, también, un típico ejemplo de minería de datos. Un supermercado quiere obtener información sobre el comportamiento de compra de sus clientes. A partir de ahí, podrá reubicar productos que se suelen comprar conjuntamente, localizar el emplazamiento óptimo para nuevos productos, realizar campañas de lanzamiento según el perfil de los clientes , etc.. A partir de los datos de un gran número de compras el supermercado podría encontrar, por ejemplo que el 100% de las veces que se compran pañales , se compra también leche, que en el 50 % de la veces que se compran huevos se compran aceite; etc. Estas **asociaciones** altamente frecuentes pueden resultar muy útiles y, quizás también encontrar que otras asociaciones se dan con muy escasa frecuencia para poder descartar ciertas políticas comerciales.

Ej. 3 Determinar las ventas de un producto.

A partir de la información de las ventas de distintos productos durante un suficiente periodo de tiempo puede determinarse los niveles de ventas del próximo mes de cada producto y organizar la gestión de stocks de una manera más eficiente. En definitiva se trata de un problema de **predicción**. Para la solución de este tipo de problemas existen técnicas tradicionales estadísticas como las series temporales y los modelos de regresión mútiple, pero otras muchos procedimientos pueden también emplearse; algunos estadísticos (Modelos lineales generalizados, por ejemplo) y otros no estadísticos (redes neuronales, por ejemplo)

Ej. 4 Determinar grupos diferenciados de empleados (o de clientes) .

Una gran compañía puede desear tener catalogados a sus empleados según grupos o categorías de comportamiento similar según algunas características. El objetivo de ello puede ser llevar a cabo una política de personal más adecuada y eficiente y el resultado final podría resultar algo así como:

Los empleados pueden catalogarse como pertenecientes a uno de estos grupos:

Grupo 1: Sin hijos, vivienda de alquiler, poco sindicados, muchas bajas

Grupo 2: Sin hijos, con coche muy sindicados con pocas bajas, mayoritariamente mujeres y viven en casa alquilada

Grupo 3: Casados y/o con hijos con coche propietarios de vivienda ,poco sindicados y mayoritariamente hombres

Relaciones de la minería de datos con otras disciplinas

La minería de datos es un campo multidsiciplinar que se ha desarrollado en paralelo o como prolongación de otras disciplinas y tecnologías. De forma que buena parte de los avances en minería de datos se nutren de mejoras y avances en estas otras disciplinas por lo que no debe perderse de vista su , a veces muy estrecha, relación.

Del tratamiento de las bases de datos se heredan conceptos y técnicas de procesamiento en línea y algunas herramientas de gestión de bases de datos.

De la recuperación de la información, se recogen técnicas de recuperación de la información desde datos textuales o desde internet o de búsquedas en contextos de información masiva.

La estadística ha aportado muchos conceptos, algoritmos, técnicas y métodos de análisis: reducción de datos, análisis uni y multidimensional, regresión, muestro,,

validación cruzada, modelización paramétrica y no paramétrica, técnicas bayesianas, análisis mulivariante (Cluster, discriminante, factorial, correspondencias) . Algunos paquetes estadísticos se comercializan como herramientas de minería de datos o incluyen módulos con esa denominación (SAS, SPSS, R)

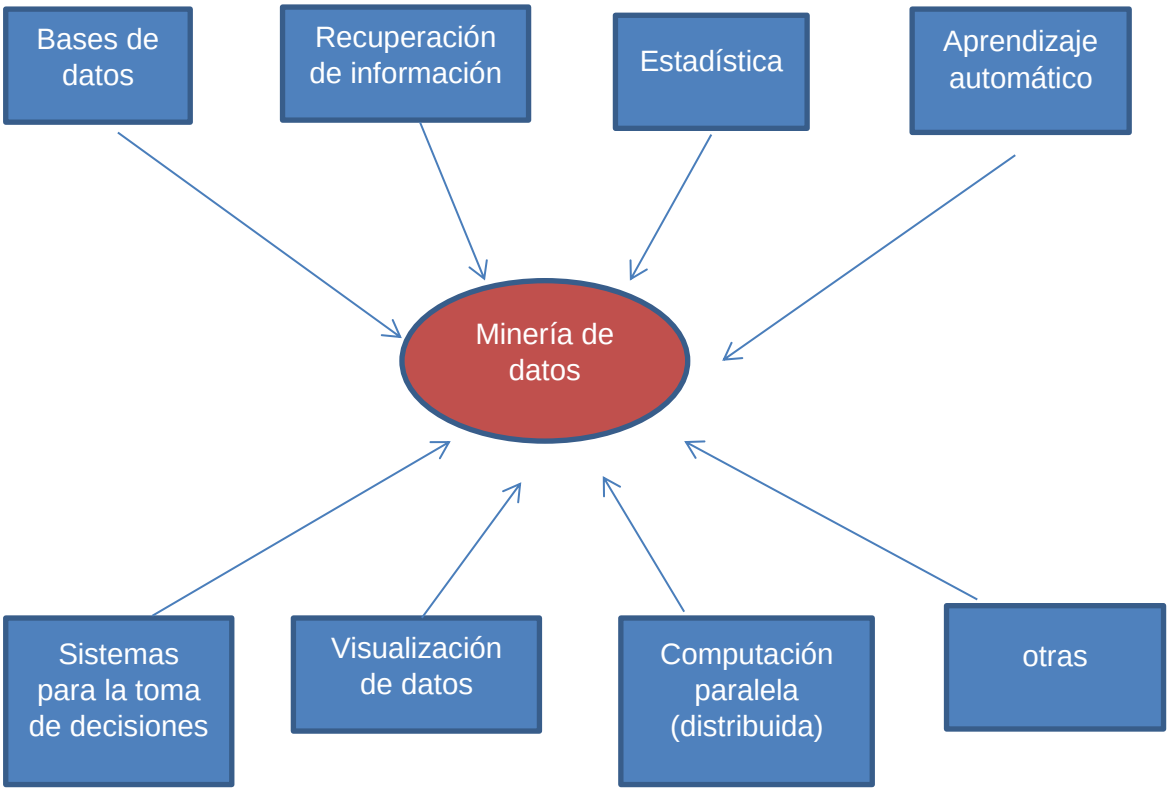
El aprendizaje automático , campo de la Inteligencia Artificial, que desarrolla algoritmos capaces de aprender, constituye , junto con la estadística, el corazón metodológico de la minería de datos. Comparten el principio básico de aprender (generar/generalizar) un modelo a partir de ejemplos para la resolución de problemas y, a menudo, el procedimiento y la técnica para llevarlo a cabo.

Los sistemas para la toma de decisiones, son sistemas automatizados que asisten a los directivos en la toma de decisiones. Algunas herramientas como el análisis ROC o los árboles de decisión provienen de éste área.

De la visualización de datos se heredan distintas técnicas que permiten “ver” patrones extraídos como solución de conocimiento que en su expresión matemática o textual resultan complejos de comprender.

La computación paralela o distribuida permite una capacidad de ejecución de algoritmos de minería de datos y/o de representación y almacenamiento de los datos que hacen que las aplicaciones de minería de datos pueden alcanzar una potencia ciertamente “abrumadora” en la actualidad y, aún más, en próximo futuro.

Otras disciplinas como el análisis del lenguaje natural, el proceso de imágenes o señales o el grafismo automático aportan también una gran cantidad de técnicas de tratamiento cuando los datos a tratar son textuales, gráficos o complejos.



1.2.El proceso de extracción de conocimiento desde los datos: Preparación , modelado, evaluación e interpretación.

Si la clave de la minería de datos es la extracción de conocimiento (no explícito) desde los (masivos) datos. Debemos atender al "proceso" por el que se lleva a cabo esta extracción.

Este proceso es lo que se conoce, también como "proceso de descubrimiento del conocimiento en bases de datos":KDD, por sus siglas en inglés: Knowledge Discovery in Databases .

Aunque a veces se confunde la totalidad del proceso KDD con el Data Mining (o mimería de datos) son , en realidad, términos diferentes. El proceso KDD sería un conjunto de fases o etapas que buscan la extracción del conocimiento y que :

1.- partiría de la recopilación de la los datos, su selección y limpieza, esto es: la **preparación.**

2.- aplicaría una serie de técnicas y métodos de extracción de conocimiento en lo que podríamos llamar **modelado** o modelización y que constituiría la minería de datos , propiamente dicha Un modelo, en este contexto sería un representación simbólica y resumida de los datos que permitiría extraer conclusiones de los mismos.

3.- y terminaría con la **interpretación** de los resultados y la **evaluación** del (de los) modelos utilizados

1.3.Principales problemas/tareas y métodos/técnicas de la minería de datos: extracción de patrones, clustering, clasificación, predicción y asociación.

La fase fundamental del proceso de KDD es, la minería de datos, propiamente dicha.(Hasta el punto, que puede confundirse con el proceso completo).Es en esta fase cuando hay que determinar la **tarea** de minería de datos que habrá que llevar a cabo, el modelo a desarrollar y la elección del o de los algoritmos concretos para desarrollar el modelo y aplicarlo a los datos ya previamente preparados. Por ejemplo, podemos desear llevar a cabo un tarea de **clasificación** de clientes de un banco entre los que devolverán el préstamo y los que no (ej. 1 del apartado anterior). Quizás queramos utilizar un árbol de decisión porque queremos obtener un **modelo** que describa el conocimiento generado en forma de reglas y, dentro de los distintos algoritmos de árboles podemos querer usar un algoritmo C5 o un CART.

La cuestión es que vemos que los problemas de conocimiento a los que nos enfrentamos los podemos estructurar en tareas a realizar (clasificación, agrupación o clustering, predicción y asociación, son las más habituales). Y, por otro lado, estas tareas se llevan a cabo utilizando distintos métodos o técnicas (técnicas estadísticas multivariantes, regresión, métodos bayesianos , árboles de decisión, redes neuronales, etc.).

1.3.1 Tareas de la minería de datos.

Cada tarea es un tipo de problema a ser resuelto que podrá considerarse *descriptiva* como el clustering o la obtención de reglas de asociación, ya que no concluyen conocimiento extra-muestral, o bien *predictivas* como la clasificación o la regresión que sí extrapolan los resultados para situaciones extra-muestrales.

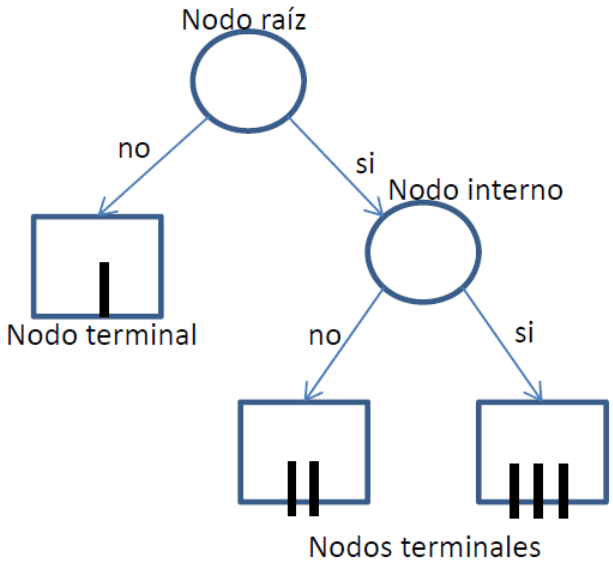
La **clasificación** es quizá la tarea más utilizada. En ella cada **instancia** o registro de la base de datos (cada individuo) pertenece a una clase lo que si indica por el valor de un atributo que juega el papel especial de clasificación. El resto de **atributos** (a

menudo en minería de datos se utiliza el nombre de atributo tanto para los atributos propiamente dichos [nominales, binarios, u ordinales] como para las variables numéricas, así lo hace el entorno WEKA, p. ejemplo), se utilizarán para predecir la clase de las futuras instancias (individuos).
El objetivo es predecir la clase de nuevas instancias cuya adscripción se desconoce a partir de la información de los demás rasgos o atributos. Las técnicas apropiadas de minería de datos proveerán un modelo que utilizando la información muestral suministrada sea capaz de aventurar un valor para el atributo "clase" de la nueva instancia maximizando algún criterio de precisión.
La **regresión** (en un sentido algo más amplio que el habitual en Estadística) es una tarea predictiva que persigue *aprender* una función real que asigna a cada instancia un valor real. La principal diferencia con la clasificación es que la predicción es numérica.
El **agrupamiento** o **clustering** es una tarea descriptiva que pretende obtener "grupos naturales" a partir de los datos. A diferencia de la clasificación en la que los datos muestrales están etiquetados en clases, aquí se pretende encontrar una "variable de etiquetado" que aún no existe. A menudo el problema del clustering se relaciona con la segmentación (partición de una población en segmentos o grupos).
La **reglas de asociación** (su obtención, más bien) es una tarea descriptiva que pretende identificar relaciones no explicitas (aún) entre atributos categóricos) pretendería obtener una versión cualitativa del análisis de las correlaciones. Como ocurre con las correlaciones las asociaciones no implican necesariamente una relación causa- efecto sino sólo una frecuente (o muy frecuente) concomitancia.
Un caso particular de reglas de asociación son las llamadas reglas de asociación secuenciales que introducen la relación entre acciones que se dan en momentos de tiempo subsecuentes.

1.3.2.Técnicas de minería de datos

Las técnicas utilizadas en minería de datos provienen de distintos paradigmas metodológicos : técnicas de inferencia estadística, árboles de decisión, redes neuronales, inducción de reglas, aprendizaje basado en instancias, algoritmos genéticos, aprendizaje bayesiano, programación lógica inductiva y estimación núcleo, entre otros. Cada uno de estos métodos suele disponer, a menudo, de distintos algoritmos y variantes y suele verse afectado por restricciones de aplicación que hacen que su mayor o menor efectividad ante unas y otras tareas dependa frecuentemente del dominio de aplicación, esto es, del caso concreto de problema a acometer con nuestra tarea de minería de datos. Por ejemplo para clasificar los clientes de un cierto banco en "devolvedores o no de un préstamo" puede funcionar muy bien un análisis discriminante y muy mal un árbol de decisión J-48 y, en cambio, para clasificar los conductores asegurados en una compañía entre los que tienen siniestros y los que no puede ocurrir al revés. Analizar las condiciones de aplicación y conocer las restricciones de aplicabilidad de cada técnica es, sin duda , importante, pero también lo es ensayar distintas estrategias y evaluar la eficiencia entre distintas alternativas.
Entre las técnicas estadísticas tradicionales la regresión y la correlación lineales y no lineales, la estimación kernel, y los modelos lineales generalizados puede ser empleados en tareas de predicción. El análisis discriminante es una técnica estadística multivariante muy útil (concebida , de hecho para ello) para la clasificación. Existen también técnicas estadísticas de análisis clúster jerárquico y no jerárquico que pueden resolver el problema (la tarea) del agrupamiento.
Técnicas derivadas del teorema de Bayes pueden usarse para generar reglas de clasificación.

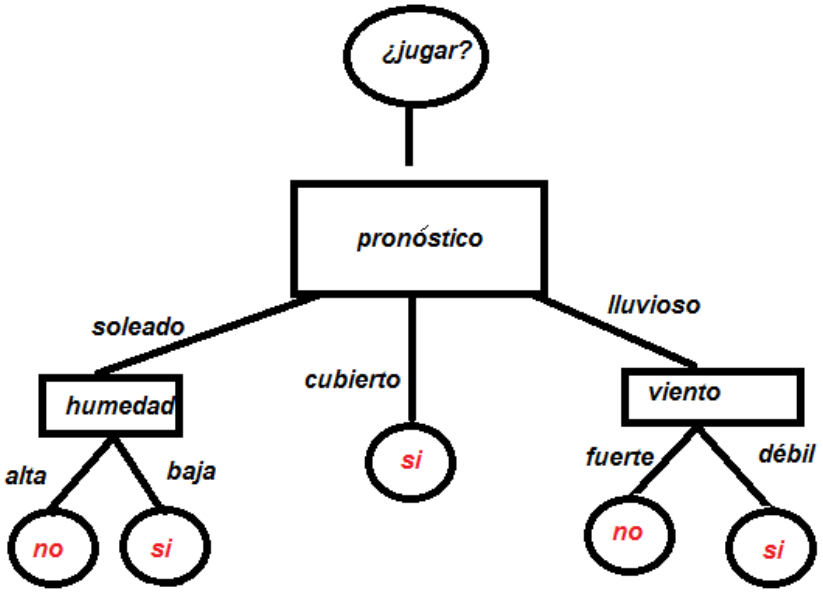
A caballo entre la estadística tradicional y los desarrollos específicos de algoritmos orientados a la clasificación encontramos muchos algoritmos de construcción de árboles de decisión (y de regresión, si pretenden predecir variables numéricas)
Un árbol de decisión propone como solución de un problema de clasificación (o de regresión) un conjunto de bifurcaciones anidadas en forma de árbol que va dividiendo paulatinamente los individuos con el objetivo de que al llegar a una rama terminal del árbol encontremos sólo individuos (instancias) perteneciente a una única clase tal como se muestra en este esquema.



En cada bifurcación (nodo) se evalúa una determinada variable (atributo) predictora y según su resultado se sigue una u otra dirección. La dicotomía (puede haber más de dos ramas también) planteada en cada nodo es la clave del proceso y ella obedecerá al criterio utilizado en cada algoritmo según alguna función que evalúe la eficiencia clasificadora.
Por ilustrar la cuestión supongamos un conjunto de datos sobre las posibilidades de jugar un partido de tenis según las condiciones meteorológicas.

#instancia	pronostico	humedad	viiento	jugar
1	soleado	alta	débil	no
2	cubierto	alta	débil	si
3	lluvioso	alta	débil	si
4	lluvioso	normal	fuerte	no
5	soleado	normal	débil	si
….				

A partir de esta información un árbol de decisión podría dar como resultado una solución como:



Los árboles de decisión pueden considerarse una forma de aprendizaje de reglas , ya que cada rama de árbol puede interpretarse como un regla ,los nodos internos van definiendo los términos de la conjunción del antecedente de la regla y la clase final asignada sería el consecuente.
Así se seguimos , por ejemplo, la rama más “extrema de la derecha”, jugar-pronóstico-lluvioso-viento-debil-si” ,tendríamos la regla:
Si l pronóstico = lluvioso **y** viento = débil **entonces** jugar= sí.
Si consideramos todas la ramas que terminan en un nodeo final “sí” podemos construir una regla general añadiendo la regla por defecto alternativa “en otro caso”. Veamos:

Si pronóstico= soleado **y** humedad= normal **entonces** jugar =sí
Si pronostico= cubierto **entonces** jugar= sí
Si pronostico = lluvioso **y** viento=débil **entonces** jugar= sí
En otro caso jugar= no

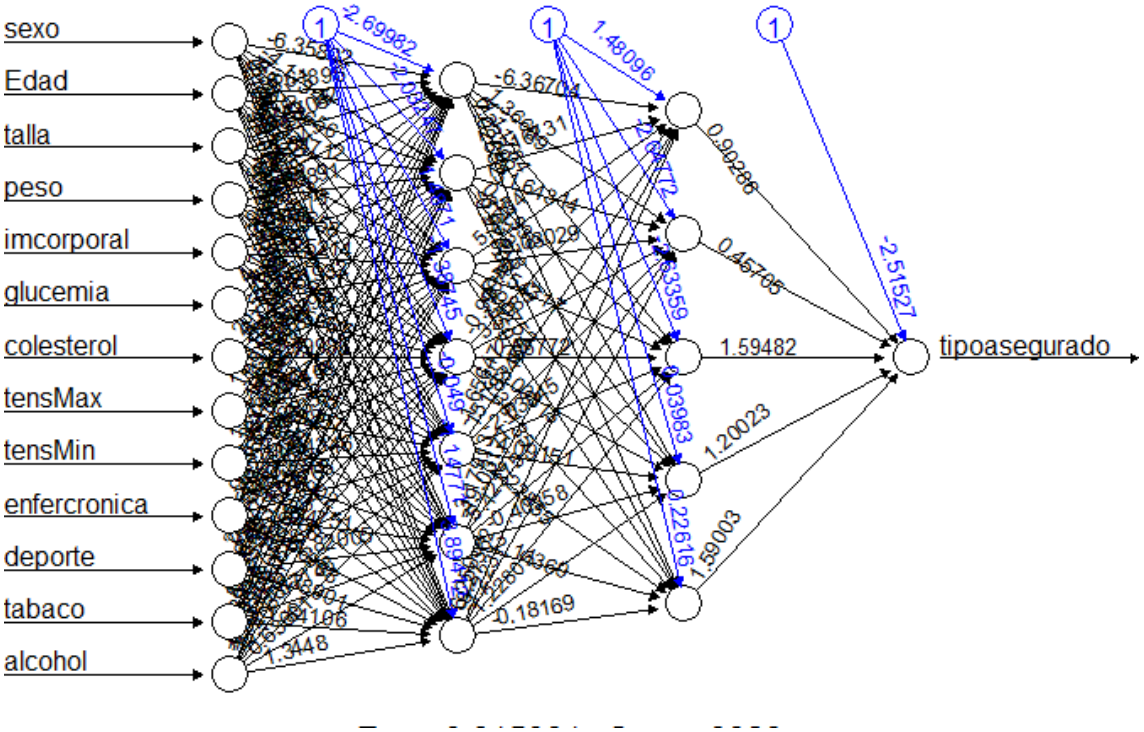
Aunque los árboles de decisión pueden producir reglas los **métodos de inducción de reglas**, en realidad, deben considerarse como técnicas diferentes ya que incluyen otras posibilidades de estructura además de las que generan los árboles:

Las reglas se consideran, en principio independientes y no tienen por qué formar un árbol .
Las reglas generadas no tienen por qué cubrir todas las posibilidades existentes (los árboles sí lo hacen).
La reglas pueden estar en conflictos en sus predicciones en cuyo caso es necesario tener un criterio de prelación o preferencia en su utilización (habitualmente se asigna a cada regla un valor de confianza basado en la frecuencia de éxito de la regla en el conjunto muestral de datos).

Las redes neuronales constituyen un paradigma computacional muy potente capaz de modelar problemas complejos. Una red neuronal puede verse como un grafo dirigido con muchos *nodos* (elementos de proceso) y *arcos* (interconexiones).Suelen contar con una *capa* de (nodos de) *entrada* y una de *salida* , pudiendo haber una o varias capas *ocultas*. Las conexiones entre nodos están ponderadas por un peso numérico

que se va viendo modificado en el proceso de aprendizaje hasta alcanzar el funcionamiento deseado en el que la red procesará las entradas para obtener la salida.
En algunos tipos de red (de aprendizaje supervisado) las entradas son las variables predictoras y las salidas la variable a predecir o a utilizar como clasificador . En otros tipos de red (de aprendizaje no supervisado) las salidas constituyen una malla en la que las distintas instancias puedan ser asignada según su comportamiento situándose cerca las de comportamiento similar y lejos las de comportamiento más diferente.
Las redes supervisadas suelen ser útiles en las tareas de clasificación y regresión y las no supervisadas en las de clustering.

El gráfico de abajo muestra un red MLP (multilayer perceptron) de aprendizaje supervisado utilizada para clasificar clientes de una compañía de seguros de salud. La red (ya está entrenada con un conjunto de datos, se uso para ello el paquete neuralnet de R) cuenta con una capa de entrada con un nodo para cada variable clasificadora , una capa de salida para el atributo de clase “tipoasegurado” y dos capas intermedias ocultas de “proceso ciego”.



La redes neuronales tienen un gran capacidad de generalización para relaciones altamente no lineales entre las variables de entrada y salida pero entre sus inconvenientes están que requieren gran número de datos para un adecuado entrenamiento y que el modelo aprendido suele ser de muy difícil interpretación. Fijémonos en el caso del gráfico resulta imposible una interpretación general del clasificador obtenido y es altamente complejo discernir si un atributo (fumar o no, por ejemplo) es más o menos relevante que otro (practicar deporte, pongamos por caso) .

En el aprendizaje basado en instancias o casos , éstos son almacenados en memoria y conforme se van introduciendo nuevos casos se intenta relacionar los nuevos casos con las ya almacenados. Se comparan los nuevos y los almacenados según alguna métrica de forma que si resultan cercano tienden a agruparse juntos y si no es así no.

Los método de análisis clúster de K-vecinos más próximos, K-medias y otros métodos no jerárquicos están entre los más frecuentemente utilizados.
Su campo de aplicación fundamental es en tareas de clustering y son especialmente útiles cuando los datos no son de naturaleza “estándar” (datos textuales, o multimedia, por ejemplo.) aunque suele requerir un proceso de **preparación** apropiado para convertir éstos datos en atributos manejables métricamente (binarios , habitualmente)
Los algoritmos genéticos o evolutivos buscan soluciones al problema, generando muchas posibles soluciones y haciéndolas evolucionar (reproducirse y mutar) simulando el proceso evolutivo biológico de forma que conforme vayan mostrando un “mejor comportamiento” tenderán a “sobrevivir” y se extinguirán si no es así. Pueden usar en distintas tareas de minería de datos y combinarse con otras técnicas a tales efectos.

1.4.Minería de datos, aprendizaje automático y big data

Como ya ha quedado de manifiesto la minería de datos pretende generar un conocimiento no explícito a partir de conjuntos de datos generalmente extensos. En la medida en que la interacción social en el mundo actual genera un enorme cantidad información sobre conjuntos individuos de distintas categorías una de las aplicaciones posibles de la minería de datos, es precisamente la aplicación de procesos de generación de conocimiento a estos enormes conjuntos de datos en los que se conoce ya con el anglicismo de **big data**. La minería de datos con datos ubicuos y masivos presenta problemas de todo tipo , tanto técnicos, como sociales, éticos, jurídicos y políticos. Obviamente este no es el espacio para dar cuenta de ellos pero sí es preciso en este tema introductorio hacer esta pequeña reseña.

Otro campo con el que se relaciona (y, a veces se confunde) la minería de datos es el aprendizaje automático o machine learning (en inglés). El aprendizaje automático es un paradigma de investigación de la inteligencia artificial que busca procedimientos en los que sea posible el aprendizaje autónomo de ciertos modelos más o menos complejos de asignación de valores a ciertos atributos. En buena media, y ya lo hemos visto de pasada, algo más arriba, buena parte de la generación de conocimiento en las tareas de minería de datos tienen que ver con que, a partir de los datos se “aprenda” un patrón de respuesta o lo que hemos llamado un modelo de conocimiento. Así, ya hemos mencionado (para las redes neurales) los aprendizajes supervisados y no supervisados, sobre los que volveremos más adelante.

De hecho, la interrelación es tan estrecha ,que, a menudo, las distintas técnicas usadas en minería de datos, con independencia de su origen metodológico, (Estadística, Lógica, teoría de Grafos, etc.) , y no sólo las procedentes de la Inteligencia artificial , tienden a considerarse como fruto de cierto “aprendizaje” que se considera , según el caso, **supervisado** o **no supervisado**.

Como ya se ha mencionado someramente, el aprendizaje supervisado obtiene la generalización de un patrón aprendido de respuesta (ante futuras entradas) a partir de la presentación al algoritmo de una gran cantidad de ejemplos (ejemplares, técnicamente hablando) de entradas y salidas. (valores de las variables de entrada y valores de las variables de salida que tienen cada uno de los individuos de la base de datos [cada instancia de la muestra]) . La idea es que algoritmo acabe encontrando

un modelo de asignación de salidas a las entradas que sea capaz de dar cuenta acertadamente de la información ya registrada para poder acertar en el futuro ante nuevos datos.

En cambio, en el aprendizaje no supervisado no se presentan pares de entradas y salidas sino sólo la información de entrada que de alguna manera el algoritmo debe agrupar según sus “parecidos razonables”.

La relación casi directa de estos dos tipos de aprendizaje con algunas de las típicas tareas de la minería de datos es casi natural. Clasificación, predicción, regresión, tienen que ver con aprendizaje supervisado. Agrupamiento, generación de reglas de asociación/correlación tiene que ver con aprendizaje no supervisado.

El entorno WEKA , en su “pestaña” de “preparación de datos” que llama pestaña preproceso se dispone de muchos “filtros” (es el nombre que usa WEKA) o procedimientos de filtrado, manipulación, transformación etc. de variables(**atributos**) e individuos (**instancias** o casos) que cataloga entre filtros supervisados y filtros no supervisados. La diferencia está, aproximadamente en lo comentado más arriba si bien en ocasiones es difícil de discernir si determinada transformación de las variables o determinado filtrado de casos debe considerarse supervisado o no supervisado.