La metodología del *Data Mining*. Una aplicación al consumo de alcohol en adolescentes

The methodology of *Data Mining*. An application to alcohol consumption in teenagers

Elena Gervilla García*, Rafael Jiménez López*, Juan José Montaño Moreno*, Albert Sesé Abad*, Berta Cajal Blasco*, Alfonso Palmer Pol*

*Área de Metodología de las Ciencias del Comportamiento. Departamento de Psicología. Universitat de les Illes Balears.

Enviar correspondencia a:
Alfonso Palmer Pol. E-mail: alfonso.palmer@uib.es

aceptado: marzo 2008
recibido: noviembre 2008

RESUMEN

El presente trabajo pretende principalmente acercar a los investigadores del campo de las drogodependencias una metodología de análisis de datos orientada al descubrimiento de conocimiento en bases de datos (KDD). El KDD es un proceso que consta de una serie de fases, la más característica de las cuales se denomina *Data Mining* (DM), en la que se aplican diferentes técnicas de modelado para detectar patrones y relaciones en los datos. Se analizan los factores comunes y diferenciadores de las técnicas DM más ampliamente utilizadas, desde una visión principalmente metodológica, y ejemplificando su uso con datos provenientes del consumo de alcohol en adolescentes y su posible relación con variables de personalidad (N=7030). Aunque la precisión global obtenida (% de predicciones correctas) es muy similar en los tres modelos analizados, las redes neuronales generan el modelo más preciso (64.1%), seguidas de los árboles de decisión (62.3%) y Naive Bayes (59.9%).

Palabras clave: *Redes Neuronales Artificiales, Árboles de Decisión, Naive Bayes, Reglas de Asociación, alcohol*

ABSTRACT

This paper is aimed mainly at making researchers in the field of drug addictions aware of a methodology of data analysis aimed at knowledge discovery in databases (KDD). KDD is a process consisting of a series of phases, the most characteristic of which is called data mining (DM), whereby different modelling techniques are applied in order to detect patterns and relationships among the data. Common and differentiating factors between the most widely used DM techniques are analysed, mainly from a methodological viewpoint, and their use is exemplified using data related to alcohol consumption in teenagers and its possible relationship with personality variables (N=7030). Although the overall accuracy obtained (% correct predictions) is very similar in the three models analyzed, the Artificial Neural Network (ANN) technique generates the most accurate model (64.1%), followed by Decision Trees (DT) (62.3%) and Naïve Bayes (NB) (59.9%).

Key words: *Artificial Neural Networks, Decision Trees, Naive Bayes, Association Rules, alcohol*

INTRODUCCIÓN

Según Hand, Mannila y Smyth (2001), "Data Mining es el análisis de (a menudo grandes) conjuntos de datos observacionales para encontrar relaciones insospechadas y resumir los datos de nuevas formas que son tanto comprensibles como útiles para el propietario de los datos". De hecho, Data Mining (DM) puede ser visto como el resultado de la evolución natural de la tecnología de la información, debido a la amplia disponibilidad de enormes cantidades de datos y la necesidad inminente de convertir tales datos en información útil y conocimiento (Han y Kamber, 2006).

En este sentido, la idea de extraer información valiosa de los datos no es nueva. La innovación está en los avances tecnológicos proporcionados por diversas disciplinas o campos de conocimiento, como el campo de la informática (procesamiento por ordenador, tecnologías de almacenamiento y bases de datos transaccionales) y el campo de la inteligencia artificial (proveyendo algoritmos computacionales intensivos). Cuando lo combinamos con el conocimiento de la estadística y el campo del análisis de datos, estas nuevas capacidades ofrecen la posibilidad de descubrir patrones y relaciones en los datos que pueden ser usados para hacer predicciones válidas.

Un aspecto relevante e imprescindible asociado a la metodología DM es que forma parte de un proceso denominado descubrimiento de conocimiento en bases de datos «Knowledge Discovery in Databases» (KDD), que explicita los pasos necesarios para reducir riesgos en la búsqueda de modelos de conocimiento al aplicar técnicas de DM. Por ejemplo, los datos requieren un sustancial preprocesamiento para ser modelados (limpieza y preparación de datos) en el proceso KDD (Han y Kamber, 2006; Hand et al., 2001; Hernández, Ramírez y Ferri, 2004).

Dado que la fase de DM es la más característica del proceso KDD, muchas veces se utiliza esta fase para nombrar todo el proceso; en este sentido, algunos autores tratan el DM como un sinónimo para el término KDD (Bigus, 1996; Kantardzic, 2003; Larose, 2005; Two Crows Corporation, 1999).

Las técnicas de DM han aparecido en los campos de las finanzas, el marketing, el comercio, las telecomunicaciones, la manufactura e incluso en la industria de la salud (seguros médicos, diagnóstico y tratamiento de enfermedades), es decir, en sectores que requerían de técnicas avanzadas para la extracción de información útil y rentable dado los grandes volúmenes de datos que manejan habitualmente.

Pocos estudios se han llevado a cabo en el contexto del consumo de drogas, aunque varios trabajos han aplicado técnicas de DM en este campo (Kitsantas, Moore y Sly, 2007; Palmer y Montaño, 1999; Palmer, Montaño y Calafat, 2000). A lo largo de estas líneas, se pretende acercar al investigador del campo de las drogodependencias una visión integradora del uso de esta metodología dentro de un caso práctico: el uso de drogas y su relación con variables de personalidad. Sin embargo, este estudio no pretende

INTRODUCTION

According to Hand, Mannila and Smyth (2001), "Data mining is the analysis of (often large) observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner". In fact, data mining (DM) can be viewed as a result of the natural evolution of information technology, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge (Han and Kamber, 2006).

In this sense, the idea of extracting valuable information from data is not new. The innovation is in the technological advances provided by diverse disciplines or fields of knowledge, such as the field of computerization (computer processing, storage technologies and transactional databases) and the field of artificial intelligence (providing computational intensive algorithms). When combined, along with a knowledge of the statistic and data analysis field, these new capabilities offer the possibility of discovering patterns and relationships in data that may be used to make valid predictions.

A relevant and indispensable aspect associated with DM methodology is that it forms part of a process called Knowledge Discovery in Databases (KDD), which specifies the steps needed to reduce risks in searching for knowledge models when applying DM techniques. For example, data require substantial pre-processing for modelling (cleaning and preparing data) in the KDD process (Han and Kamber, 2006; Hand et al., 2001; Hernández, Ramírez and Ferri, 2004).

Since the DM phase is the most characteristic phase of the KDD process, it is often used to denote the entire process; in this sense, some authors treat the term DM as a synonym for the term KDD (Bigus, 1996; Kantardzic, 2003; Larose, 2005; Two Crows Corporation, 1999).

DM techniques have burst onto the scene in the fields of finance, marketing, commerce, telecommunications, manufacturing and even the health industry (medical insurance and diagnosis and treatment of diseases), i.e., in sectors that require advanced techniques for extracting useful and profitable information from the large volumes of data they customarily handle.

Few studies have been conducted within the context of drug use, although several works have applied DM techniques in this field (Kitsantas, Moore and Sly, 2007; Palmer and Montaño, 1999; Palmer, Montaño and Calafat, 2000). Along these lines, we purport to provide researchers in the field of drug addiction with an integrated view of the use of this methodology within a case study: drug use and its relationship to personality variables. However, this study does not aim to provide substantial information or reach conclusions on the issue, but rather contribute to providing

proporcionar información sustancial o alcanzar conclusiones sobre este tema, sino contribuir a aportar conocimiento metodológico sobre el uso de determinadas técnicas de DM y ejemplificar su uso con la intención de comparar los resultados obtenidos. En concreto, analizamos los factores comunes y diferenciadores de una serie de técnicas DM ampliamente utilizadas: las reglas de asociación, Naive Bayes, las redes neuronales artificiales y los árboles de decisión.

METODOLOGÍA

El KDD (*Knowledge Discovery in Databases*) es un proceso iterativo e interactivo que combina la experiencia en un problema con una variedad de técnicas de análisis de datos tradicionales y tecnologías avanzadas de aprendizaje automático (aprendizaje automático por procedimientos computacionales). El objetivo es descubrir patrones y relaciones en los datos que puedan ser usados para hacer predicciones válidas.

Básicamente, el KDD está compuesto por los pasos de selección de datos (los datos relevantes para el análisis se recuperan de la base de datos), el preprocesamiento de los datos (limpiar y preparar los datos), data mining (construir modelos descriptivos/predictivos) y evaluación del modelo (conseguir los modelos descriptivos/predictivos que mejor solucionen el problema). Mirar Han y Kamber (2006) para una extensa exposición del proceso KDD.

Los modelos descriptivos se rigen por un proceso de aprendizaje no supervisado: el objetivo es identificar patrones en los datos sin indicadores externos que guíen al algoritmo (es decir, sin conocer la realidad "a priori"). En este sentido, los modelos descriptivos sirven para explorar las propiedades de los datos examinados.

El «clustering» y las reglas de asociación (RA) son las herramientas más representativas de DM. El objetivo del «clustering» es encontrar grupos que sean muy diferentes unos de otros, pero cuyos miembros sean muy similares entre sí (Ghosh, 2003; Han y Kamber, 2006; Larose, 2005). Las RA son otro instrumento común en el contexto del modelado descriptivo (Agrawal, Imielinski y Swami, 1993); en particular, el análisis de la cesta de la compra es un conocido ejemplo de descubrimiento de asociaciones, donde el objetivo es encontrar reglas sobre los artículos que aparecen juntos en un acontecimiento como una transacción de compra.

Por otro lado, los modelos predictivos requieren de un proceso de aprendizaje supervisado: la técnica supervisa en el modelo en construcción el grado de ajuste a la realidad conocida. En este sentido, dichos modelos pretenden estimar valores futuros o desconocidos de una variable respuesta: cuando es una variable respuesta categórica (para predecir etiquetas de clase), se conoce como un modelo de clasificación; si el valor para ser predicho es numérico (variable respuesta continua) se llama modelo de regresión (según Hand et al., 2001, Hernández et al., 2004

methodological knowledge on the use of certain DM techniques and exemplify its use in order to compare the results obtained. Specifically, we analyse the common and differentiating factors between a series of widely-used DM techniques: association rules, Naive Bayes, artificial neural networks, and decision trees.

METHODOLOGY

KDD (Knowledge Discovery in Databases) is an iterative, interactive process which combines experience in a problem with a variety of traditional data analysis techniques and advanced technologies of machine learning (automatic learning by computational procedures). The aim is to discover patterns and relationships among the data that may be used to make valid predictions.

Basically, KDD is made up of the steps of data selection (data relevant to the analysis task are retrieved from the database), data pre-processing (cleaning and preparing data), data mining (building descriptive/predictive models) and model evaluation (reaching descriptive/predictive models that best solve the problem). See Han and Kamber (2006) for an extensive exposition of the KDD process.

Descriptive models are governed by an unsupervised learning process: the aim is to identify patterns among the data without external indicators guiding the algorithm (that is, without knowing the reality "a priori"). In this sense, descriptive models are useful to explore the properties of the data examined.

Clustering and association rules (AR) are the most representative DM tools. The goal of clustering is to find groups that are very different from each other, but whose members are very similar to each other (Ghosh, 2003; Han and Kamber, 2006; Larose, 2005). AR are another common tool in descriptive modelling context (Agrawal, Imielinski and Swami, 1993); in particular, market basket analysis is a well-known example of association discovery, where the goal is to find rules about items that appear together in an event such as a purchase transaction.

Prediction models, on the other hand, require a supervised learning process: the technique supervises the degree of fit to the known reality in the model under construction. In this sense, these models attempt to estimate the future or unknown values of a response variable: when it is a categorical response variable (in order to predict class labels), it is known as a classification model; if the value to be predicted is numerical (continuous response variable) it is called a regression model (according to Hand et al., 2001, Hernández et al., 2004 or Two Crows Corporation, 1999) or prediction model (according to Han

o Two Crows Corporation, 1999) o modelo de predicción (según Han y Kamber, 2006). En este trabajo, nos referimos al modelo de predicción como un término general, de forma que distinguimos entre modelos de regresión y de clasificación.

Un requisito común a las técnicas de modelado predictivo es la utilización de una muestra de datos (datos test) que es independiente a la empleada en la construcción del modelo (datos de entrenamiento), con la intención de evaluar la capacidad de generalización del modelo (evaluación del modelo). Entre las técnicas de predicción más utilizadas se encuentran el clasificador Naive Bayes (NB), como una moderna técnica estadística, y las Redes Neuronales Artificiales (RNA) y los Árboles de Decisión (AD), métodos que generan modelos de clasificación y regresión a partir de algoritmos de aprendizaje automático (Han y Kamber, 2006; Kantardzic, 2003; Michie, Spiegelhalter y Taylor, 1994; Witten y Frank, 2005; Ye, 2003).

En este sentido, es importante destacar que, por la propia idiosincrasia de los métodos DM, se espera disponer de una amplia base de datos que permita realizar tres grupos (entrenamiento, test y validación) para construir y evaluar los modelos, aspecto que sería problemático con un reducido número de sujetos. En este último caso, cuando se trata de muestras pequeñas, la estadística clásica dispone de técnicas adecuadas para manejar la información; no obstante, la ventaja de DM radica justamente en poder extraer información de grandes volúmenes de información.

A continuación se explican las técnicas DM junto con algunas ventajas e inconvenientes.

Reglas de Asociación

Como ya se ha comentado, los modelos de aprendizaje no supervisado se usan cuando el resultado de interés no es conocido y el sistema debe aprender directamente de los datos. Una de las herramientas más populares incluidas en el aprendizaje no supervisado son las Reglas de Asociación (RA).

Las RA, introducidas por Agrawal et al. (1993), recogen relaciones interesantes entre un gran conjunto de información. Un ejemplo típico de esta aplicación consiste en encontrar asociaciones entre los artículos comprados en los grandes almacenes (análisis de la cesta de la compra). Este tipo de información es muy valiosa para situar estratégicamente los productos en los grandes almacenes o planificar las promociones de determinados artículos, y su uso se ha generalizado a cualquier ámbito en el que se disponga de grandes cantidades de información almacenada.

Las RA están formadas por el antecedente (la primera parte, "si") y el consecuente (la segunda parte, "entonces") y ofrecen la información en forma de declaraciones del tipo "si-entonces" ("Si A, entonces B").

and Kamber, 2006). In this work, we refer to prediction model as a general term, in such a way that we distinguish between classification and regression models.

One requirement common to prediction modelling techniques is the use of a sample of data (test data) that is independent from what is used in constructing the model (training data), in order to evaluate the model's capacity to generalize (model evaluation). Among the most widely-used prediction techniques are the Naive Bayes classifier (NB), as a modern statistical technique, and the Artificial Neural Networks (ANN) and Decision Trees (DT) methods to generate classification and regression models from their machine learning algorithms (Han and Kamber, 2006; Kantardzic, 2003; Michie, Spiegelhalter and Taylor, 1994; Witten and Frank, 2005; Ye, 2003).

In this respect, it is important to emphasize that, for the own idiosyncrasy of DM, one expects to have a wide database that allows to realize three groups (training, test and validation) to construct and to evaluate the models, and this would be problematic with a small number of subjects. In the latter case, when it is a question of small samples, the classic statistics has skills adapted to handle the information; nevertheless, the advantage of DM takes root in being able to extract exactly information of big volumes of information.

Later are explained the technologies DM together with some advantages and disadvantages.

Association rules

As has been commented, unsupervised learning models are used when the result concerned is not known and the system has to learn directly from the data. One of the most popular tools included in unsupervised learning are association rules (AR).

AR, introduced by Agrawal et al. (1993), pick up interesting relationships among a large set of information. A typical example of this application consists of finding associations between the articles purchased in a department store (market basket analysis). This type of information is very valuable in order to strategically place products in the department store or to plan the promotion of certain articles, and their use has become generalized in any field in which there is an availability of great quantities of stored information.

AR are made up of the antecedent (the first part, "if") and the consequent (the second part, "then") and they offer the information in the shape of declarations of the "if-then" type ("If A, then B").

En los últimos años se han desarrollado diversos algoritmos eficaces para extraer RA (Agrawal, Mannila, Srikant, Toivonen y Verkamo, 1996; Hipp, Güntzer y Nakhaeizadeh, 2000). No obstante, el algoritmo clásico para generar RA es el algoritmo Apriori de Agrawal y Srikant (1994). La idea básica de este algoritmo es generar de forma progresiva y recursiva conjuntos de ítems frecuentes que aparecen juntos en la base de datos un porcentaje mínimo de ocasiones.

Apriori recibe este nombre porque reduce el conjunto de ítems frecuentes candidatos descartando, a priori, los conjuntos de k-ítems candidatos que contienen un subconjunto de ítems infrecuente.

El interés de las RA se evalúa mediante una serie de índices que expresan el grado de incertidumbre de las mismas. En concreto, en este trabajo nos centraremos en los siguientes índices: soporte, confianza, «lift» e «hiperlift». Para una revisión detallada de dichos índices, consultar Hahsler, Hornik y Reutterer (2005), en cuyo trabajo presentan un sencillo *framework* probabilístico para el análisis de datos transaccionales, implementando dichos índices en el entorno de programación estadístico R (Ihaka y Gentleman, 1996). Hahsler, Grün y Hornik (2005) dan a conocer en un trabajo posterior el paquete arules para R, que proporciona una infraestructura básica para la extracción de RA y su análisis.

El soporte de una regla se define como el porcentaje de operaciones en las que aparece un conjunto de ítems. Es la probabilidad de que una operación seleccionada de forma aleatoria de la base de datos contenga todos los ítems del antecedente y el consecuente de la regla (se calcula dividiendo el número de casos que contienen el antecedente y el consecuente entre el número total de casos).

La confianza es la probabilidad condicional de que una operación seleccionada de forma aleatoria incluya todos los ítems en el consecuente si dicha operación incluye todos los ítems en el antecedente (se computa como el número de veces que aparecen juntos el antecedente y el consecuente dividido entre el número de veces que aparece el antecedente).

El «lift» indica qué probabilidad existe de encontrar el consecuente limitando la búsqueda a aquellos conjuntos de ítems donde el antecedente está presente (se calcula dividiendo la confianza de antecedente y consecuente entre el soporte del consecuente).

Finalmente, el «hyperlift» es una adaptación del índice anterior y resulta una medida más robusta para sucesos con baja prevalencia (véase Hahsler, Hornik y Reutterer, 2005, p. 10). Está basado en la idea que, bajo independencia, el soporte de las transacciones que contienen todos los ítems en la regla sigue una distribución hipergeométrica con los parámetros dados por el soporte del antecedente y el consecuente.

In recent years, different efficient algorithms have been developed in order to extract AR (Agrawal, Mannila, Srikant, Toivonen and Verkamo, 1996; Hipp, Güntzer and Nakhaeizadeh, 2000). Nevertheless, the classical algorithm to generate AR is the Apriori algorithm by Agrawal and Srikant (1994). The basic idea of this algorithm is to generate frequent itemsets that appear together in the database on a minimal percentage of occasions.

Apriori receives this name because it reduces the frequent itemset candidates by ruling out, a priori, the sets of k-item candidates that contain a subset of infrequent items.

The interest of AR is evaluated by means of a series of indices that express the degree of uncertainty of the same. In particular, in this study we will focus on the following indices: support, confidence, lift and hyper-lift. For a detailed review of these indices, consult Hahsler, Hornik and Reutterer (2005), where they present a simple probabilistic framework for the analysis of transactional data, by implementing these indices in the statistical R programming environment (Ihaka and Gentleman, 1996). Hahsler, Grün and Hornik (2005) in a later paper publish the arules package for R, which provides a basic infrastructure for the extraction of AR and their analysis.

The support of a rule is defined as the percentage of operations in which an itemset appears. It is the probability that an operation selected from the database at random contains all the items of the antecedent and the consequent of the rule (it is calculated by dividing the number of cases containing the antecedent and the consequent by the total number of cases).

The confidence is the conditional probability that an operation selected at random includes all the items in the consequent if this operation includes all the items in the antecedent (it is computed as the number of times the antecedent and the consequent appear together divided by the number of times the antecedent appears).

The lift indicates what probability exists of finding the consequent by limiting the search to the itemsets where the antecedent is present (it is calculated by dividing the antecedent and consequent confidence by the support of the consequent).

Finally, the hyper-lift is an adaptation of the previous index and is a more robust measure for low prevalence events (see Hahsler, Hornik and Reutterer, 2005, p. 10). It is based on the idea that, under independence, the support of the transactions that contain all the items in the rule follows a hyper-geometric distribution with the parameters given by the support of the antecedent and the consequent.

Naive Bayes

Naive Bayes (NB) es una de las técnicas de clasificación más ampliamente usada, debido a su proceso computacionalmente simple incluso para un conjunto de datos de entrenamiento grande. Está basado en el teorema de Bayes, que puede predecir la probabilidad de que un caso dado pertenezca a una clase determinada. Su simplicidad computacional se debe a la suposición conocida como *independencia condicional de clase* (supone que el efecto de un valor de atributo sobre una clase dada es independiente de los valores de los otros atributos), y en este sentido es considerado "ingenuo" (Han y Kamber, 2006).

Los estudios que comparan algoritmos de clasificación (p.ej., Michie et al., 1994) a menudo han mostrado que NB es comparable en el funcionamiento con RNA y clasificadores AD, y de hecho supera a estos sofisticados clasificadores si los atributos son condicionalmente independientes dada la clase. De hecho, este mejor rendimiento parece mantenerse en muchas bases de datos médicas, según la experiencia de estos autores. En este sentido, según Witten y Frank (2005), la regla es "intentar siempre las cosas sencillas primero", ya que los investigadores de forma repetitiva, después de mucho esfuerzo, han obtenido buenos resultados usando métodos de estudio sofisticados sólo para descubrir años más tarde que métodos simples como NB lo hacen igual de bien, o incluso mejor.

Una solución para eliminar la influencia de un valor de atributo con frecuencia nula en una clase (probabilidades que el cero sostenga un veto sobre los otros), es seguir la técnica del *estimador de Laplace*: añadir 1 a cada frecuencia de cada variable. Esta estrategia asegura que un valor de atributo que ocurre cero veces en el conjunto de datos de entrenamiento reciba una probabilidad que no es nula, aunque muy pequeña.

La atracción del clasificador NB reside en su simplicidad, eficiencia computacional y buen rendimiento en clasificación. No obstante, presenta tres importantes inconvenientes (Shmueli, Patel y Bruce, 2008): en primer lugar, requiere un gran número de casos para obtener buenos resultados; en segundo lugar, si una categoría de predicción no está presente en los datos de entrenamiento, la técnica asume que un nuevo caso con esa categoría en el predictor tiene probabilidad cero; finalmente, aunque se obtiene buen rendimiento si el objetivo es clasificación u ordenación de los casos de acuerdo a su probabilidad de pertenecer a una clase determinada, este método ofrece resultados muy sesgados cuando el objetivo es estimar la probabilidad de pertenencia a una clase.

Para explorar la técnica NB con más profundidad, recomendamos consultar Larose (2006), cuyo trabajo también presenta un ejemplo de análisis que usa el NB en el programa de código abierto Weka (Witten et al., 1999; Witten y Frank, 2005).

Naive Bayes

Naive Bayes (NB) is one of the most widely used classification techniques, because of its computationally simple process even for a large training dataset. It is based on Bayes' theorem, which can predict the probability that a given case belongs to a particular class. Its computational simplicity is due to the assumption known as *class conditional independence* (suppose that the effect of an attribute value on a given class is independent of the values of the other attributes), and in this sense is considered "naive" (Han and Kamber, 2006).

Studies comparing classification algorithms (e.g., Michie et al., 1994) have often shown that NB is comparable in performance whith ANN and DT classifiers, and indeed outperforms these sophisticated classifiers if the attributes are conditionally independent given the class. In fact, this outperformance seems to hold true for many medical datasets, according to these authors' experience. In this sense, according to Witten and Frank (2005), the moral is, "always try the simple things first", since repeatedly people have eventually, after an extended struggle, obtained good results using sophisticated learning methods only to discover years later that simple methods such as NB do just as well, or even better.

A solution to eliminate the influence of an attribute value with null frequency in a class (probabilities that are zero hold a veto over the other ones), is to follow the *Laplace estimator* technique: adding 1 to each frequency value of each variable. This strategy ensures that an attribute value that occurs zero times in the training dataset receives a probability which is nonzero, albeit very small.

The NB classifier's beauty is in its simplicity, computational efficiency, and good classification performance. However, there are three main shortcomings (Shmueli, Patel and Bruce, 2008): first, it requires a very large number of records to obtain good results; second, where a predictor category is not present in the training data, NB assumes that a new record with that category of the predictor has zero probability; finally, albeit good performance is obtained when the goal is classification or ranking of records according to their probability of belonging to a certain class, this method provides very biased results when the goal is actually to estimate the probability of class membership.

To explore the NB technique in further depth, we recommend consulting Larose (2006), whose work also presents an example of analysis using the NB in the Weka open source software (Witten et al., 1999; Witten and Frank, 2005).

Redes neuronales artificiales

Las redes neuronales artificiales (RNA) son sistemas de procesamiento de datos cuya estructura y operación están inspirados en las redes neuronales biológicas. Las RNA se desarrollaron basándose en las siguientes directrices:

El procesamiento de información ocurre en elementos sencillos llamados neuronas.

Las neuronas transmiten señales mediante conexiones establecidas.

Cada conexión (enlace de comunicación) tiene un peso asociado.

Cada neurona aplica una función de activación (usualmente no lineal) a la entrada total recibida de las neuronas conectadas (suma de entradas ponderadas por los pesos de conexión), obteniendo así un valor de salida que actuará como valor de entrada que se transmitirá al resto de la red (Figura 1).

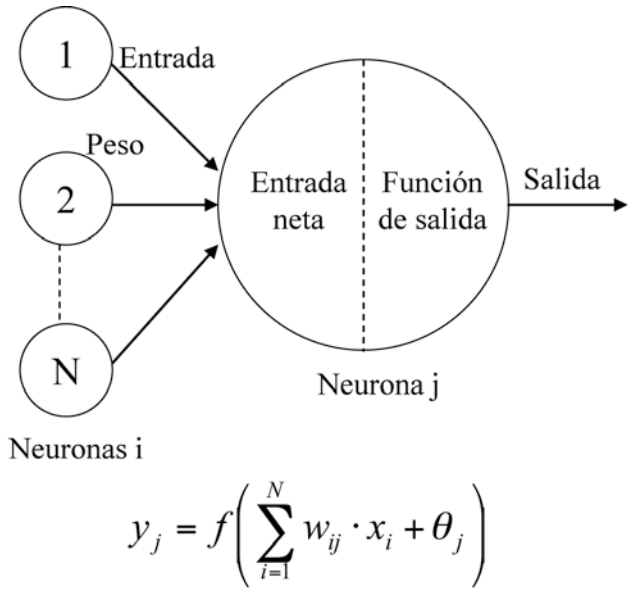

Figura 1. Funcionamiento general de una neurona artificial y su representación matemática (adaptado de Palmer y Montaño, 1999)

Hay una amplia selección de modelos de RNA. La combinación de la topología (el número de neuronas y capas ocultas, y cómo están conectadas), el paradigma de aprendizaje y el algoritmo de aprendizaje definen un modelo de RNA (Bigus, 1996).

Una RNA de tipo perceptrón multicapa parte de una capa de entrada, en la que cada nodo o neurona se corresponde con una variable predictora. Estas neuronas de entrada se conectan con cada una de las neuronas que forman la capa oculta de la red (número de nodos determinado por el usuario). Los nodos de la capa oculta se conectan a su vez con las neuronas de otra capa oculta. La capa de salida se compone de una (predicción binaria) o más neuronas de

Artificial neural networks

Artificial neural networks (ANN) are data processing systems whose structure and operation are inspired by biological neural networks. ANN were developed on the basis of the following guidelines:

Data processing occurs in simple elements called neurons.

Neurons transmit signals along established connections.

Each connection (communication link) has an associated weight.

Each neuron applies an activation function (usually nonlinear) to the total input received from the linked neurons (the sum of weighed inputs times connection weights), thus obtaining an output value that will act as the input value to be transmitted to the rest of the network (Figure 1).

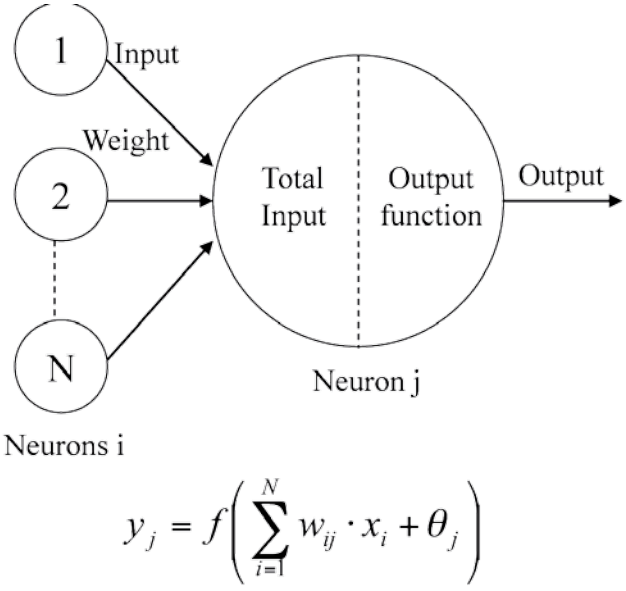

Figure 1. Generic working of an artificial neuron and its output mathematical representation (adapted from Palmer and Montaño, 1999)

There are a wide selection of ANN models. The combination of topology (the number of neurons and hidden layers, and how they are connected), the learning paradigm and the learning algorithm define an ANN model (Bigus, 1996).

Multi-layer perceptron ANN are based on an input layer in which each node or neuron corresponds to a prediction variable. These input neurons are connected to each one of the neurons that form the network's hidden layer (number of nodes determined by the user). The nodes in the hidden layer are in turn connected to the neurons of another hidden layer or an output layer. The output layer is made up of one

salida. En este tipo de arquitectura, la información siempre se transmite desde la capa de entrada hacia la capa de salida.

La popularidad del perceptrón multicapa se debe principalmente a que es capaz de actuar como un aproximador universal de funciones. Más concretamente, una red «backpropagation» que contiene al menos una capa oculta con suficientes unidades no lineales puede aprender cualquier tipo de función o relación continua entre un grupo de variables de entrada (discretas y/o continuas) y una variable de salida (discreta o continua). Esta propiedad convierte a las redes perceptrón multicapa en herramientas de propósito general, flexibles y no lineales. Además, otra ventaja fundamental de los modelos neuronales en general es que normalmente no imponen ningún tipo de restricción respecto a los datos de partida (tipo de relación funcional entre variables), ni suelen partir de supuestos concretos (como el tipo de distribución que siguen los datos).

Un tercer conjunto de datos independiente (conjunto de validación) se emplea para evitar el sobreajuste del modelo durante el proceso de aprendizaje de la red, debido a un número de parámetros o pesos excesivo en relación al problema (Hastie, Tibshirani, y Friedman, 2001, p. 356)

A pesar de las ventajas expuestas acerca de la técnica, por contrapartida, una de las críticas más importantes que se han lanzado contra el uso de las RNA se centra en el hecho de que el conocimiento de los pesos de la red no ayuda en general a la interpretación del proceso subyacente que genera la predicción de un determinado valor de salida. Pese a ello, esta percepción acerca de las RNA como una compleja "caja negra" no es del todo cierta. En este sentido, han surgido diferentes intentos por interpretar los pesos de la red neuronal, de los cuales el más ampliamente utilizado es el denominado análisis de sensibilidad, implementado en programas RNA como el recientemente presentado por Palmer, Fernández y Montaño (2001), bajo el nombre de Sensitivity Neural Network 1.0.

La ventaja más prominente de las RNA es su buen funcionamiento predictivo. Las RNA tienen una alta tolerancia al ruido en los datos y la habilidad de capturar relaciones muy complicadas entre los predictores y la respuesta; no obstante, su extremada flexibilidad reside en disponer de datos de entrenamiento suficientes y requieren más tiempo para su ejecución que otras técnicas (Shmueli, Patel y Bruce, 2008).

Árboles de decisión

Los árboles de decisión (AD) permiten representar de forma gráfica una serie de reglas sobre la decisión que se debe tomar en la asignación de un valor de salida a un determinado registro. Se componen de nodos (variables de entrada), ramas (grupos de registros en las variables de entrada) y hojas o nodos hoja (valores de la variable de salida).

La construcción de un AD se basa en el principio de "divide y vencerás": a través de un algoritmo de aprendizaje

(binary prediction) or more output neurons. Data are always transmitted from the input layer to the output layer in this type of architecture.

The popularity of the multi-layer perceptron is mainly thanks to its capacity to act as a universal function approximator. More specifically, a back-propagation network containing at least one hidden layer with sufficient nonlinear units can learn any type of function or continuous relationship among a group of (discrete and/or continuous) input variables and an (discrete or continuous) output variable. This property turns multi-layer perceptron networks into flexible, nonlinear all-purpose tools. Furthermore, another fundamental advantage of neural models in general is that they do not normally impose any type of restriction on data (type of functional relationship between variables), neither are they usually based on specific assumptions (such as the type of distribution the data follow).

A third set of independent data (validation dataset) is used in ANN to prevent the model overfitting during the network learning process, caused by an excessive number of parameters or weights in relationship to the problem (Hastie, Tibshirani and Friedman, 2001, p. 356).

In contrast to the advantages of the technique described, one of the most important criticisms launched against the use of ANN focuses on the fact that knowledge of network weights does not generally help interpret the underlying process that generates the prediction of a certain output value. Despite this, the perception of ANN as a complex "black box" is not absolutely accurate. Different attempts have been made to interpret neural network weights, of which the more widely used is known as *sensitivity analysis*, implemented in ANN programs such as the one recently presented by Palmer, Fernandez and Montaño (2001) as *Sensitivity Neural Network 1.0.*

The most prominent advantage of ANN is their good predictive performance. They are known to have high tolerance to noisy data and the ability to capture highly complicated relationships between the predictors and a response; however, their extreme flexibility relies heavily on having sufficient data for training purposes and they require a longer runtime than other techniques (Shmueli, Patel and Bruce, 2008).

Decision trees

Decision trees (DT) allow the graphic representation of a series of rules on the decisions to be made in assigning output value to a certain entry. They are made up of nodes (input variables), branches (groups of entries in the input variables) and leaves or leaf nodes (output variable values).

Constructing a DT is based on the principle of "divide and conquer": successive splitting of the multivariate space is achieved through a supervised learning algorithm for the

supervisado se realizan divisiones sucesivas del espacio multivariable para maximizar la distancia entre grupos en cada división (esto es, realizar particiones que discriminen). El proceso de división finaliza cuando todos los registros de una rama tienen el mismo valor en la variable de salida (nodo hoja puro), dando lugar al modelo completo (máxima especificidad). Cuanto más abajo están las variables de entrada en el árbol, menos importantes son en la clasificación de salida (y menos generalización permiten, debido a la disminución del número de entradas en las ramas descendientes).

Para evitar el sobreajuste del modelo, se puede realizar una poda del árbol que elimine las ramas con pocos registros o poco significativas. En consecuencia, si partimos del modelo completo, tras la poda del árbol éste ganará en capacidad de generalización (evaluada con datos de test), a costa de reducir el grado de pureza de sus hojas (Hernández et al., 2004; Larose, 2005).

purpose of maximising the distance between groups in each splitting (i.e., to create partitions that are discriminatory). The splitting process ends when the value of the output variable is the same for all entries in a given branch (pure leaf node), resulting in the complete model (maximum specificity). The farther down the input variables are on the tree, the less important they are in output classification (and the less they allow for generalizability, due to the decrease of the number of entries in the descending branches).

To avoid overfitting of the model, it is possible to carry out some pruning on the tree so as to eliminate the branches with few or not very significant entries. As a result, if we start from the whole model, after pruning the tree it will gain in generalization capacity (evaluated with test data), at the expense of reducing the degree of purity of its leaves (Hernández et al., 2004; Larose, 2005).

Tabla 1. Comparativa entre algoritmos de aprendizaje para árboles de decisión
Table 1. Comparative between learning algorithms for decision trees

ALGORITMOS *ALGORITHMS*	Variables de entrada *Input variables*	Variable de salida *Output variable*	Tipo de predicción *Type of prediction*	Ramas por división *Splitting branches*	Criterio de división *Splitting criterion*
CHAID	categórica / numérica *categorical / numerical*	categórica / numérica *categorical / numerical*	clasificación / regresión *classification / regression*	≥*2*	*Ji-cuadrado / F* *Chi-square / F*
CART	categórica / numérica *categorical / numerical*	categórica / numérica *categorical / numerical*	clasificación / regresión *classification / regression*	=2	GINI / Desviación cuadrática mínima *GINI / Least squared deviation*
C4.5 / C5.0	categórica / numérica *categorical / numerical*	categórica *categorical*	clasificación *classification*	≥2	Ganancia proporcional *Gain ratio*

Existen diversos algoritmos de aprendizaje diseñados para la obtención de modelos AD (véase Tabla 1). Destacan el algoritmo CHAID (Chi-squared Automatic Interaction Detection), implementado por Kass (1980), el algoritmo CART (Classification And Regression Trees) desarrollado por Breiman, Friedman, Losen y Stone (1984), el algoritmo ID3 (Iterative Dichotomiser 3) de Quinlan (1986), y sus posteriores evoluciones C4.5 (Quinlan, 1993) y C5.0 (Quinlan, 1997).

El algoritmo de aprendizaje determina los siguientes aspectos:

Compatibilidad específica con el tipo de variables: naturaleza de las variables de entrada y la variable de salida.

Procedimiento de evaluación de la distancia entre grupos en cada división: criterio de división.

Puede imponer restricciones en el número de ramas en que se divide cada nodo.

There are different learning algorithms designed to obtain DT models (see Table 1). 'The most outstanding ones are the CHAID (Chi-squared Automatic Interaction Detection) algorithm implemented by Kass (1980), the CART (Classification And Regression Trees) algorithm developed by Breiman, Friedman, Losen and Stone (1984), the ID3 (Iterative Dichotomiser 3) algorithm by Quinlan (1986), and its later evolutions C4.5 (Quinlan, 1993) and C5.0 (Quinlan, 1997).

The learning algorithm determines the following aspects:

Specific compatibility with the type of variables: nature of the input and output variables.

Evaluation procedure of distance between groups in each splitting: splitting criterion.

It can impose restrictions on the number of branches each node can be split into.

Parámetros de poda [prepoda / postpoda]: nº mínimo de registros por nodo o rama, valor crítico del criterio de división, diferencia de rendimiento entre el árbol ampliado y el reducido. La prepoda implica utilizar criterios de parada durante la construcción del árbol, mientras que la postpoda aplica los parámetros de poda al árbol completo.

Una de las ventajas más sobresalientes de los AD es su carácter descriptivo, que permite entender e interpretar fácilmente las decisiones tomadas por el modelo, ya que tenemos acceso a las reglas que se utilizan en la tarea predictiva (aspecto no contemplado en otras técnicas, como las RNA). De hecho, es posible derivar fácilmente reglas de decisión (para cada rama terminal) siguiendo las rutas marcadas en la estructura del árbol que llevan a un determinado nodo hoja (la decisión del modelo).

Por otro lado, las reglas de decisión proporcionadas por un modelo de árbol tienen poder predictivo (no sólo descriptivo) desde el momento en que se evalúa su precisión a partir de unos datos independientes (datos test) a los utilizados en la construcción del modelo (datos de entrenamiento); esta particularidad no queda recogida en los modelos obtenidos bajo aprendizaje no supervisado, como es el caso de la técnica AR.

Otro rasgo atractivo de los AD es que son intrínsecamente robustos a los valores perdidos. No obstante, los AD presentan algunas debilidades (Shmueli, Patel y Bruce, 2008): son sensibles a pequeños cambios en los datos y, a diferencia de los modelos que asumen una relación particular entre la respuesta y la predicción (p. ej., una relación lineal como en una regresión lineal), los AD son no lineales y no paramétricos. Esto permite una amplia gama de relaciones entre los predictores y la respuesta, pero puede ser una debilidad: dado que las particiones se realizan sobre predictores únicos más que sobre combinaciones de predictores, el AD probablemente omite relaciones entre predictores, en particular estructuras lineales como las de los modelos de regresión lineal o logística.

Pruning parameters [pre-pruning/post-pruning]: minimum number of entries per node or branch, critical value of the splitting criterion, performance difference between the amplified and reduced tree. Pre-pruning implies using halting criteria during the construction of the tree, whereas post-pruning applies the pruning parameters to the whole tree.

One of the most outstanding advantages of DT is their descriptive character, which allows the model's decisions to be easily understood and interpreted, since we have access to the rules used in the prediction task (an aspect not included in other techniques, such as ANN). In fact, decision rules (for each terminal branch) can be easily derived by following the routes marked in the tree structure that lead to a certain leaf node (the model's decision).

On the other hand, the decision rules provided by a tree model have a predictive power (not only descriptive) from the time in which their accuracy is evaluated from data (test dataset) that are independent from the data used in constructing the model (training dataset); this peculiarity is not included in the models obtained under unsupervised learning, as is the case of the AR technique.

Another appealing feature of DT is that they are intrinsically robust to outliers. However, DT present some weakness (Shmueli, Patel and Bruce, 2008): they are sensitive to slight changes in the data and, unlike models that assume a particular relationship between the response and prediction (e.g., a linear relationship such as in linear regression), DT are nonlinear and nonparametric. This allows for a wide range of relationships between the predictors and the response, but this can also be a weakness: since the splits are done on single predictors rather than on combinations of predictors, the DT is likely to miss relationships between predictors, in particular linear structures like those in linear or logistic regression models.

CASO PRÁCTICO

Una vez expuesta la base metodológica de las técnicas DM objeto de este artículo, en este apartado pretendemos proporcionar elementos comparativos de la información proporcionada por dichas técnicas en un contexto aplicado: el consumo de alcohol en adolescentes y su relación con determinadas variables de personalidad. Con ello se pretende aportar una visión más integradora, si cabe, de la metodología DM, puesto que se aportan parámetros de evaluación comunes para comparar los resultados obtenidos, en el contexto indicado, a partir de los modelos generados. No obstante, como se ha indicado anteriormente, no se pretende extraer conclusiones sustantivas en el campo de las adicciones a partir de dicho ejemplo. En este sentido, la intención es dar a conocer a los investigadores de dicho campo una serie de herramientas metodológicas (desde una visión comparativa) que permiten detectar patrones

A CASE STUDY

Having described the methodological basis of the DM techniques covered in this article, this section endeavours to furnish comparative elements of the information provided by these techniques within an applied context: alcohol consumption in teenagers and its relationship to personality variables. With this instance, we hope to offer a more integrated vision of DM methodology, since we provide common evaluation parameters to compare the results obtained from the models generated. Nevertheless, as indicated above, we do not aim to reach substantial conclusions in the field of addictions from this instance. In this sense, the intention is to provide researchers in this field with a series of methodological tools (from a comparative view) which will enable them to detect knowledge patterns in a practically automated way. These patterns can be used

de conocimiento de una manera casi-automatizada. Estos patrones pueden ser utilizados para generar modelos descriptivos de la realidad que aporten mayor información sobre las variables asociadas a un fenómeno de drogodependencia concreto; y además, permite a los investigadores poder generar modelos predictivos que faciliten la toma de decisiones en el campo aplicado.

La muestra se compuso de 7030 adolescentes entre 14 y 18 años de edad, 53% de los cuales consumían alcohol, en oposición al 47% que no lo hacían. Se proporcionó información sobre variables de personalidad focalizadas en los constructos de autoestima, impulsividad, conducta antisocial y búsqueda de sensaciones (un total de 20 variables). Nos planteamos extraer modelos de conocimiento que dieran cuenta de la posible relación entre estas variables de personalidad y la variable de consumo de alcohol.

En primer lugar, nos planteamos realizar una búsqueda de RA desde una perspectiva puramente descriptiva. En este sentido, desde el contexto del aprendizaje no supervisado, aplicamos el algoritmo Apriori a los 7030 casos del estudio con la finalidad de extraer RA fuertes. En concreto, el sistema generó 205312 reglas con una confianza igual o superior a 0.7. De éstas, 32 conducen al consumo de alcohol, es decir, aparece el consumo de alcohol en el consecuente de la regla. En la Tabla 2 se presenta un ejemplo de las RA generadas:

to generate descriptive models of reality which provide more information about the variable associated with a specific drug addiction phenomenon; what is more, it enables researchers to generate predictive models that will facilitate decision making in the applied field.

The data sample was made up of 7030 teenagers between 14 and 18 years of age, 53% of whom consumed alcohol, as opposed to 47% who did not. Information was provided on personality variables that focused on the constructs of self-esteem, impulsiveness, sensation-seeking and anti-social conduct (a total of 20 variables). We considered extracting knowledge models that give an account of the possible relationship between these personality variables and the alcohol consumption variable.

First, we decided to carry out an AR search from a purely descriptive point of view. In this sense, from the unsupervised learning context, we applied the Apriori algorithm to the 7030 case studies with the aim of extracting strong AR. Specifically, the system generated 205312 rules with a confidence equal to or greater than 0.7. Of these, 32 led to alcohol consumption, that is, alcohol consumption appears in the consequent of the rule. Table 2 shows an example of the AR generated:

Tabla 2. Reglas de asociación con medidas de interés (información generada con el package arules integrado en el programa de libre distribución R, versión 2.2.1)

Table 2. Association rules with interesting measurements (information generated with the arules package integrated in the freely distributed R programme, version 2.2.1)

Reglas de asociación *Association rules*	Soporte *Support*	Confianza *Confidence*	«Lift» *Lift*	«Hyperlift» *Hyperlift*
Si le gustan las experiencias nuevas/excitantes y hace cosas ilegales, entonces consume alcohol *If he/she likes new/exciting experiences and does illegal things, then he/she consumes alcohol*	0.1044	0.8101	1.5248	1.4709
Si hace cosas ilegales, entonces consume alcohol *If he/she does illegal things, then he/she consumes alcohol*	0.1098	0.7991	1.5041	1.4511
Si hace cosas impulsivamente, le gustan las experiencias nuevas/excitantes y el desenfreno/la desinhibición, entonces consume alcohol *If he/she acts impulsively, likes new/exciting experiences and licentiousness/a lack of inhibition, then he/she consumes alcohol*	0.1098	0.7394	1.3918	1.3449
Si no piensa que es un fracaso como persona, hace cosas impulsivamente y le gusta el desenfreno/la desinhibición, entonces consume alcohol *If he/she doesn't think he/she is a failure as a person, he/she acts impulsively and likes licentiousness/a lack of inhibition, then he/she consumes alcohol*	0.1022	0.7344	1.3823	1.3339
Si hace cosas impulsivamente y le gusta el desenfreno/la desinhibición, entonces consume alcohol *If he/she acts impulsively and likes licentiousness/a lack of inhibition, then he/she consumes alcohol*	0.1152	0.7290	1.3722	1.3278

Desde la perspectiva del modelado predictivo, se proporciona información del rendimiento de los modelos generados con las técnicas NB, RNA y AD (Tabla 3). En concreto, se trata de un sencillo problema de clasificación, en el cual la variable de salida "consumo de alcohol" (variable predicha) es dicotómica (consume/no consume), al igual que las variables de entrada (variables predictoras).

From the perspective of predictive modelling, information on the models' performance generated with the NB, ANN and DT techniques is provided (Table 3). In particular, this is a simple problem of classification, in which the "alcohol consumption" output variable (predicted variable) is dichotomous (consumes/does not consume), as are the input variables (predictive variables).

Tabla 3. Matriz de confusión y rendimiento del modelo con datos de test
Table 3. Confusion matrix and model performance with test data

		Categoría actual *Actual category*		
Naive Bayes		No consumo *No consumption*	Consumo *Consumption*	Total
Categoría predicha *Predicted category*	No consumo *No consumption*	**688**	556	1244
	Consumo *Consumption*	404	**743**	1147
NB	Total	1092	1299	**2391**
	Precisión ***Accuracy***	63.0%	57.2%	**59.9%**
Red Neuronal Artificial		No consumo *No consumption*	Consumo *Consumption*	Total
Categoría predicha *Predicted category*	No consumo *No consumption*	**691**	433	1124
	Consumo *Consumption*	425	**841**	1266
RNA *ANN*	Total	1116	1274	2390
	Precisión ***Accuracy***	61.9%	66.0%	**64.1%**
Árbol de decisión		No consumo *No consumption*	Consumo *Consumption*	Total
Categoría predicha *Predicted category*	No consumo *No consumption*	**719**	493	1212
	Consumo *Consumption*	408	**771**	1179
AD *DT*	Total	1127	1264	2391
	Precisión	63.8%	60.1%	**62.3%**

La tabla 3 ofrece información sobre los casos clasificados correctamente (precisión) con datos de test, es decir, en una muestra de datos independientes a la empleada en la construcción del modelo con cada técnica particular. En concreto, se ha utilizado un 34% del conjunto total de datos como muestra de test. Las técnicas NB y AD se usaron en el resto de casos para construir el modelo (66%), mientras que en el caso particular de las RNA se ha empleado un 50% como datos de entrenamiento, y un 16% para validación.

A partir de la matriz de confusión de cada modelo (Tabla 3) es fácil obtener información, no únicamente sobre la precisión global de dicho modelo (% de predicciones

Table 3 provides information on correctly classified cases (accuracy) with test data, i.e., in a sample of data independent from the data used to construct the model with each particular technique. Specifically, 34% of the total dataset was used as a test sample. The NB and DT techniques were used in the remaining cases to construct the model (66%), whereas 50% were used as training data and 16% for validation in the particular case of ANN.

It is easy to obtain information, not only on the overall accuracy of a model (% correct predictions), but also on the specific accuracy of each real classification category from each model's confusion matrix (Table 3). That is to say, we observe that the NB and DT models make better

correctas), sino también sobre la precisión particular para cada categoría de clasificación real. En concreto, podemos observar que los modelos NB y AD predicen mejor para la categoría "no consume" que para la categoría "consume", mientras que en el modelo RNA ocurre lo contrario. En cualquier caso, en términos generales, la precisión global obtenida es bastante similar en los tres modelos; de hecho, si aplicamos una técnica estadística clásica como la regresión logística, se obtiene una precisión similar (63.2%) aunque está por debajo de la técnica RNA, que es la que genera el modelo más preciso.

Dicha baja precisión indica que las variables de entrada del modelo tienen un poder predictivo bajo sobre la variable salida. En otras palabras, estas variables en conjunto discriminan muy poco en el proceso de clasificación. Esta baja discriminación de las variables de personalidad analizadas puede ser debida a la edad de los sujetos de la muestra, ya que la personalidad por lo general no está bien definida en los adolescentes si los comparamos con los adultos (Caspi, Roberts y Shiner, 2005; MacDonald, 2005).

Los AD no sólo proporcionan información general sobre la precisión de sus predicciones, sino que además permiten entender e interpretar fácilmente las decisiones tomadas por dichos modelos para un caso particular, ya que tenemos acceso a las reglas que se utilizan en la tarea predictiva. En concreto, podemos extraer reglas que tengan un cierto grado de soporte y confianza en su decisión. Se muestran a modo de ejemplo algunas de las reglas de decisión ligadas al modelo AD (Figura 2) cuyo rendimiento aparece en la Tabla 3 (entre paréntesis se indica la confianza de la regla, es decir, la proporción de entradas para las que la regla de decisión es correcta –son correctamente clasificados).

predictions for the "does not consume" category than for the "consumes" category, whereas the reverse is true in the ANN model. In any case, the overall accuracy obtained is very similar in the three models; in fact, if we apply a classical statistical technique such as logistic regression, a similar accuracy is obtained (63.2%) although it is below the ANN technique, which is the one that generates the most accurate model.

Such low accuracy indicates that the model's input variables have a low predictive power over the output variable. In other words, these variables as a whole discriminate very little in the classification process. This low discrimination of the personality variables analysed may be due to the age of the subjects in the sample, since personality is not usually very well defined in teenagers as compared to adults (Caspi, Roberts and Shiner, 2005; MacDonald, 2005).

Not only do DT provide general information on the accuracy of their predictions, but they also allow decisions made by these models in a particular case to be easily understood and interpreted, since we have access to the rules used in the prediction task. Specifically, rules can be extracted that have a certain degree of support and confidence in their decision. Shown as an example are several of the decision rules linked to the DT model (Figure 2) whose performance is noted in Table 3 (the confidence of the rule is indicated in brackets, i.e., the proportion of entries for which the rule's decision is correct – are correctly classified):

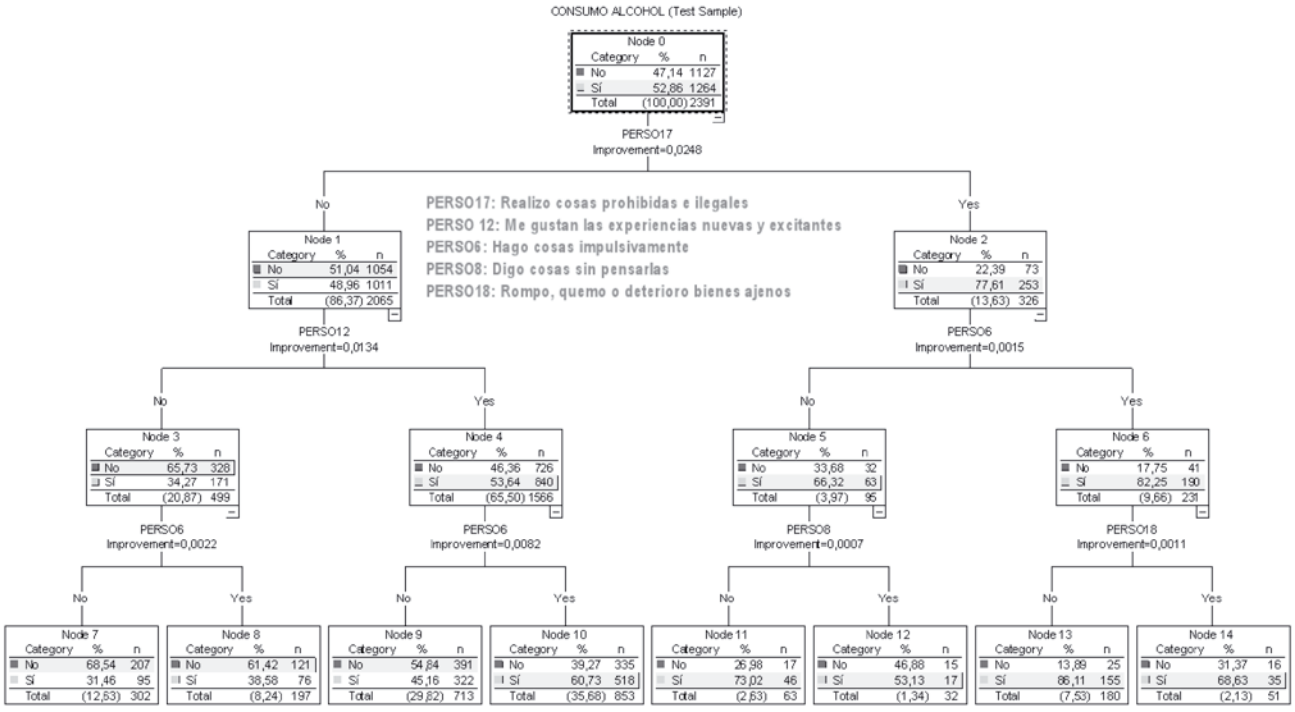

Figura 2. Árbol de clasificación generado por el algoritmo CART (Breiman et al., 1984)

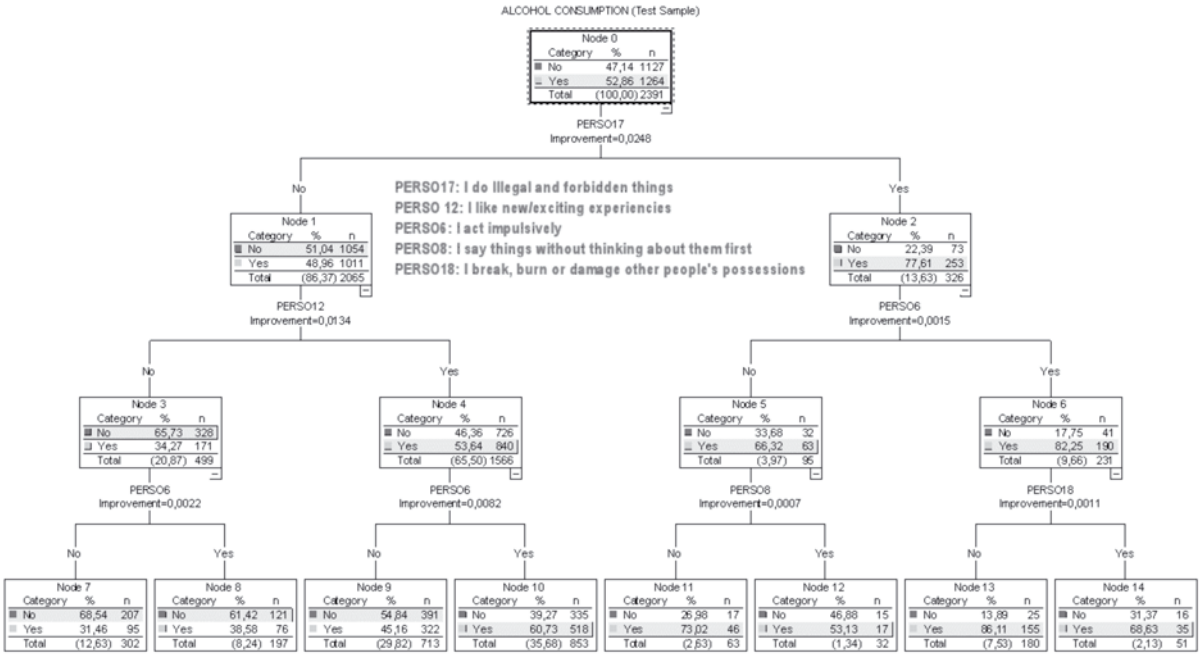

Figure 2. Classification tree generated by the CART algorithm (Breiman et al., 1984))

SI (Hago cosas prohibidas e ilegales = "No") y (Me gustan las experiencias nuevas y excitantes = "No") y (Actúo impulsivamente = "No") ENTONCES (consumo de alcohol = "No") (confianza: 0.6854) [Nodo 7, Figura 2]

Si (Hago cosas prohibidas e ilegales = "Si") y (Actúo impulsivamente = "Si") ENTONCES (consumo de alcohol = "Si") (confianza: 0.8225) [Nodo 6, Figura 2]

Si (Hago cosas prohibidas e ilegales = "Si") y (Actúo impulsivamente = "Si") y (Rompo, quemo o deterioro propiedades de otras personas = "No") ENTONCES (consumo de alcohol = "Si") (confianza: 0.8611) [Nodo 13, Figura 2]

IF (I do illegal and forbidden things = "No") and (I like new/exciting experiences = "No") and (I act impulsively = "No") THEN (alcohol consumption = "No") (confidence: 0.6854) [Node 7, Figure 2]

IF (I do illegal and forbidden things = "Yes") and (I act impulsively = "Yes") THEN (alcohol consumption = "Yes") (confidence: 0.8225) [Node 6, Figure 2]

IF (I do illegal and forbidden things = "Yes") and (I act impulsively = "Yes") and (I break, burn or damage other people's possessions= "No") THEN (alcohol consumption = "Yes") (confidence: 0.8611) [Node 13, Figure 2]

DISCUSIÓN

La revisión de la literatura sugiere que las técnicas DM aún no tienen un papel importante en el contexto de las drogodependencias, a pesar de su utilidad demostrada en otros ámbitos de conocimiento. En este sentido, nuestra intención ha sido acercar a sus investigadores una visión integradora de esta metodología, entendida como un proceso de extracción de conocimiento (KDD) que focaliza su atención en distintas fases, tal como se ha comentado anteriormente, siendo la fase DM la más destacable; además, se ha proporcionado literatura básica para que los investigadores puedan profundizar en dicha metodología.

Hemos analizado los factores comunes y diferenciadores de una serie de técnicas ampliamente utilizadas para la extracción de modelos de conocimiento. Hemos intentado mantener un equilibrio entre la información necesaria para introducir al lector en aspectos básicos de dichas técnicas

DISCUSSION

A review of the literature suggests that DM techniques have not yet begun to play a significant role within the context of drug addiction, in spite of their proven utility in other fields of knowledge. Therefore, our aim has been to provide its researchers with an integrated vision of this methodology, understood as a knowledge mining process (KDD) which focuses on different phases, as mentioned above, in which the DM phase is the most outstanding; furthermore, basic literature has been provided so that researchers can examine this methodology in further depth.

We have analysed the common and differentiating factors in a series of widely-used techniques for extracting knowledge models. We have been careful to maintain a balance between the information needed to introduce readers to the basic features of these techniques and the

y la inclusión de referencias útiles para profundizar en las mismas.

Como se ha indicado, la metodología KDD necesita preprocesar los datos antes de aplicar una determinada técnica DM, y a su vez incide en otro aspecto clave común a todas las técnicas de propósito predictivo, la evaluación del rendimiento de los modelos generados. De hecho, se ha proporcionado información al respecto para cada una de las técnicas revisadas, destacando la visión integradora de dicha metodología al emplear unos parámetros de evaluación comunes; esto es, el uso de la información proporcionada por las matrices de confusión para comparar el rendimiento de modelos que han sido generados con diferentes técnicas.

Ciertamente, este estudio destaca la vertiente aplicada de las técnicas presentadas. En este sentido, hemos acudido al campo del uso de drogas. De hecho, en trabajos previos (Palmer y Montaño, 1999; Palmer et al., 2000) ya se ha utilizado una de las técnicas clásicas de DM, las RNA, para estudiar la relación entre el consumo de drogas y variables conductuales. Por otro lado, Kitsantas et al. (2007) usaron los árboles de clasificación (un tipo de AD) para perfilar la conducta de fumar de los adolescentes. La novedad de nuestro artículo, respecto a los trabajos citados, reside en la contextualización de éstas y otras técnicas de DM en una metodología unificada (llamada KDD), y la intención de informar, desde una visión más metodológica que aplicada, de los factores comunes y diferenciadores de las técnicas DM más ampliamente usadas para obtener modelos de conocimiento.

inclusion of useful references for an examination of them in further depth.

As indicated, the KDD methodology states the need to pre-process data before applying a DM technique, and also affects another key aspect that all prediction techniques have in common: the evaluation of the generated model's performance. In fact, information in this regard was provided for each of the supervised learning techniques reviewed, emphasising the integrated vision of this methodology when using common evaluation parameters; i.e., the use of information provided by confusion matrices to compare the performance of models generated by these techniques.

Certainly, this study highlights the applied aspect of the techniques presented. To do so, we have turned to a context within the field of drug use. In fact, previous studies (Palmer and Montaño, 1999; Palmer et al., 2000) have already used one of the classic DM techniques, ANN, to study the relationship between drug use and behavioural variables. On the other hand, Kitsantas et al. (2007) used classification trees (a type of DT) to profile adolescent smoking behaviour. The novelty of our article, with respect to the aforementioned studies, lies in the contextualisation of these and other DM techniques within a unifying methodology (called KDD), and the intention to inform, from a more methodological rather than applied view, of the common and differentiating factors of the DM techniques most widely-used in obtaining knowledge models.

REFERENCIAS / REFERENCES

Agrawal, R. y Srikant, R. (1994). Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Databases, 487-499.

Agrawal, R., Imielinski, T. y Swami, A. (1993). Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data, 207-216.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. y Verkamo, A. I. (1996). Fast Discovery of Association Rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth y R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining (pp. 307-328). AAAI/MIT Press.

Bigus, J.P. (1996). Data mining with neural networks: solving business problems from application development to decision support. New York: McGraw-Hill.

Breiman, L., Friedman, J. H., Olshen, R. A. y Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Caspi, A., Roberts, B. W. y Shiner, R. L. (2005). Personality development: stability and change. Annual Review of Psychology, 56, 453-484.

Ghosh, J. (2003). Scalable Clustering. In N. Ye (Ed.), The Handbook of Data Mining (pp. 247-277). Mahwah, NJ: Lawrence Erlbaum Associates.

Hahsler, M., Grün, B. y Hornik, K. (2005). Arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. Journal of Statistical Software, 14, 1-25.

Hahsler, M., Hornik, K. y Reutterer, T. (2005). Implications of probabilistic data modeling for rule mining. Report 14, Research Report Series, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Austria.

Han, J. y Kamber, M. (2006). Data Mining: Concepts and Techniques (2nd. ed.). San Francisco: Morgan Kaufmann.

Hand, D., Mannila, H. y Smyth, P. (2001). Principles of Data Mining. Cambridge, MA: The MIT Press.

Hastie, T., Tibshirani, R. y Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.

Hernández, J., Ramírez, M. J. y Ferri, C. (2004). Introducción a la Minería de Datos [Introduction to Data Mining]. Madrid: Pearson Educación, S.A.

Hipp, J., Güntzer, U. y Nakhaeizadeh, G. (2000). Algorithms for Association Rule Mining – A general survey and comparison. SIGKDD Explorations, 2, 58-64.

Ihaka, R. y Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics, 5, 299-314.

Kantardzic, M. (2003). Data Mining: Concepts, Models, Methods, and Algorithms. New York: Wiley.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 29, 119-127.

Kitsantas, P., Moore, T. W. y Sly, D. F. (2007). Using classification trees to profile adolescent smoking behaviors. Addictive Behaviors, 32, 9-23.

Larose, D. T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. Hoboken, NJ: Wiley.

Larose, D. T. (2006). Data Mining Methods and Models. Hoboken, NJ: Wiley.

MacDonald, K. (2005). Personality, Evolution, and Development. In R. Burgess and K. MacDonald (Eds.), Evolutionary Perspectives on Human Development (pp. 207-242). Thousand Oaks, CA: Sage.

Michie, D., Spiegelhalter, D. J. y Taylor C. C. (Eds.) (1994). Machine Learning, Neural and Statistical Classification. New York: Ellis Horwood Ltd.

Palmer, A. y Montaño, J. J. (1999). ¿Qué son las redes neuronales artificiales? Aplicaciones realizadas en el ámbito de las adicciones [What are artificial neural networks? Applications in the field of addictions]. Adicciones, 11, 243-255.

Palmer, A., Fernández, C. y Montaño, J. J (2001). Sensitivity Neural Network 1.0 [Computer program]. Available at mailto:alfonso.palmer@uib.es

Palmer, A., Montaño, J. J. y Calafat, A. (2000). Predicción del consumo de éxtasis a partir de redes neuronales artificiales [Ecstasy consumption prediction on the basis of artificial neural networks]. Adicciones, 12, 29-41.

Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1, 81-106.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann.

Quinlan, J. R. (1997). C5.0 Data Mining Tool. RuleQuest Research, http://www.rulequest.com.

Shmueli, G., Patel, N. R. y Bruce, P. C. (2007). Data Mining for Business Intelligence. Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner. New Jersey: John Wiley & Sons, Inc.

Two Crows Corporation (1999). Introduction to Data Mining and Knowledge Discovery (3th. ed.). Maryland: Two Crows Corporation.

Witten, I. H. y Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques (2nd. ed.). San Francisco: Morgan Kaufmann.

Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, G. y Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with Java implementations. In N. Kasabov and K. Ko (Ed.), Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems (pp. 192-196). Dunedin, New Zealand.

Ye, N. (Ed.) (2003). The Handbook of Data Mining. Mahwah, NJ: Lawrence Erlbaum Associates.