Artículo Original / Original Research

Ciencias Matemáticas, Vol. 38, No. 2, 2024, Pag. 29-39
Recibido (*Received*): 10-04-2025, Revisado (*Reviewed*): 14-05-2025
Aceptado (*Accepted*): 21-05-2025, En línea (*Online*): 26-07-2025

Modelos y técnicas aplicables al análisis predictivo
Models and techniques for predictive analysis

Carmen Elena Viada González[1], Carlos Narcizo Bouza Herrera[2], Sira María Allende Alonso[3], Gemayqzel Bouza Allende[4], Aliuska Frias Blanco[5], Lazara García Fernández[6], Martha María Fors López[7]*

Resumen El análisis predictivo es un área de la minería de datos que consiste en la extracción de información existente en los datos y su utilización para predecir tendencias y patrones de comportamiento pasado, presente o futuro. Se describen los modelos aplicables al análisis predictivo: modelos predictivos, modelos descriptivos, modelos de decisión, modelos de conjuntos y modelos de elevación y el proceso de validación de modelos. Por otra parte, se describen las técnicas aplicables al análisis predictivo: modelo de regresión lineal, análisis de supervivencia, los árboles de clasificación y las curvas de regresión adaptativa multivariable. Ademas, las técnicas de aprendizaje computacional como redes neuronales, máquinas de vectores de soporte, *naïve bayes* y $k-$vecinos más cercanos. Finalmente, se describen las herramientas computacionales con las que se cuanta como *SPPS* y *SAS*, que son software propietarios, y *R* y *Weka*, que son de código abiertos y libres.

Palabras Clave: árboles de decisión, modelos predictivos, redes neuronales, regresión lineal.

Abstract *Predictive analytics is an area of data mining that involves extracting existing information form data and using it to predict trends and patterns of past, present, and future behavior. The models applicable to predictive analysis are described: predictive models, descriptive models, decision models, set models, and elevation models and the model validation process. On the other hand, the techniques applicable to predictive analysis are described: linear regression model, survival analysis, classification trees, and multivariate adaptive regression curves. In addition, computational learning techniques such as neural networks, support vector machines, naïve bayes and $k-$nearest neighbors. Finally, the computational tools are described, such as SPSS and SAS, which are property software, and R and Weka, which are open source and free.*

Keywords: *decision trees, predictive models, neural networks, linear regression.*

Mathematics Subject Classification: 62-XX, 68T05, 62H30, 90B50, 62P30.

[1]Dirección de Investigaciones Clínicas, Centro de Inmunología Molecular, La Habana, Cuba. Email: carmen@cim.sld.cu.
[2]Departamento Matemática Aplicada, Facultad de Matemática y Computación, Universidad de La Habana, La Habana, Cuba. Email: bouza@matcom.uh.cu.
[3]Departamento Matemática Aplicada, Facultad de Matemática y Computación, Universidad de La Habana, La Habana, Cuba. Email: sira@matcom.uh.cu.
[4]Departamento Matemática Aplicada, Facultad de Matemática y Computación, Universidad de La Habana, La Habana, Cuba. Email: gema@matcom.uh.cu.
[5]Dirección de Investigaciones Clínicas, Centro de Inmunología Molecular, La Habana, Cuba. Email: aliuska@cim.sld.cu.
[6]Dirección de Investigaciones Clínicas, Centro de Inmunología Molecular, La Habana, Cuba. Email: lazarag@cim.sld.cu.
[7]Universidad de las Américas, Quito, Ecuador. Email: forsmarthamaria@gmail.com.
***Autor para Correspondencia (*Corresponding Author*)**

Editado por (*Edited by*): Damian Valdés Santiago, Facultad de Matemática y Computación, Universidad de La Habana, La Habana, Cuba.
Maquetado por (*Layout by*): Paulo Enrique Lantigua Cuervo, Instituto de Criptografía, Universidad de La Habana, La Habana, Cuba.

Citar como: Viada González, C.E., Bouza Herrera, C.N., Allende Alonso, S.M., Bouza Allende, G., Frias Blanco, A., García Fernández, L., & Fors López, M.M. (2024). Modelos y técnicas aplicables al análisis predictivo. *Ciencias Matemáticas*, 38(2), 29–40. DOI: https://doi.org/10.5281/zenodo.16455547. Recuperado a partir de https://revistas.uh.cu/rcm/article/view/11157.

1. Introducción

El análisis predictivo es un área de la minería de datos que consiste en la extracción de información existente en los datos y su utilización para predecir tendencias y patrones de comportamiento, pudiendo aplicarse sobre cualquier evento desconocido, ya sea en el pasado, presente o futuro. El análisis

predictivo se fundamenta en la identificación de relaciones entre variables en eventos pasados, para luego explotar dichas relaciones y predecir posibles resultados en futuras situaciones. Ahora bien, hay que tener en cuenta que la precisión de los resultados obtenidos depende mucho de cómo se ha realizado el análisis de los datos, así como de la calidad de las suposiciones [15].

En un principio puede parecer que el análisis predictivo es lo mismo que hacer un pronóstico (que hace predicciones a un nivel macroscópico), pero se trata de algo completamente distinto. Mientras que un pronóstico puede predecir cuántos helados se van a vender el mes que viene, el análisis predictivo puede indicar qué individuos es más probable que se coman un helado. Esta información, si se utiliza de la forma correcta, supone un cambio radical, ya que permite orientar los esfuerzos para ser más productivos en la consecución de los objetivos.

Para llevar a cabo el análisis predictivo es indispensable disponer de una considerable cantidad de datos, tanto actuales como pasados, para poder establecer patrones de comportamiento y así inducir conocimiento. Por ejemplo, en el caso comentado en el párrafo anterior, si se cruzan datos acerca de la temperatura registrada, la época del año y si es fin de semana o festivo se puede inferir el perfil de persona que comerá helado [6].

Este proceso se realiza gracias al aprendizaje computacional. Las computadoras pueden "aprender"de manera autónoma y de esta forma desarrollar nuevo conocimiento y capacidades, para ello basta con proporcionarles el más potente y gran recurso de la sociedad moderna: los datos [12].

2. Materiales y Métodos

El aprendizaje computacional es parte fundamental en un proceso de análisis predictivo. El aprendizaje computacional proporciona las técnicas de análisis de datos mediante las cuales se pueden descubrir relaciones entre variables que en un principio pueden parecer insignificantes, pero que tras la aplicación de estas técnicas puede descubrirse la trascendencia de las mismas.

Una vez se han establecido correlaciones entre variables entra en juego la labor del ser humano, que consiste en saber interpretar las mismas y hacer las suposiciones apropiadas.

Si bien establecer correlaciones entre variables puede proporcionar información muy valiosa, hay que saber interpretar las mismas del modo correcto para no llegar a conclusiones erróneas. La correlación no implica causalidad. El descubrimiento de una relación entre A y B no implica que una cause la otra.

Una vez definidas las suposiciones correctas hay que tratar de aprovechar las mismas, que se utilizarán para realizar predicciones.

Tras identificar las correlaciones entre variables mediante técnicas de aprendizaje computacional y establecer las suposiciones correctas, se identifican patrones de comportamiento que permiten crear un modelo predictivo.

Este modelo predictivo se podrá utilizar para predecir qué probabilidades hay de que una persona – en función de los datos que se disponga de la misma – reaccione de una manera determinada (si comprará un producto, si cambiará de voto, si contratará un servicio, etc.). Una vez introducidos los datos de la persona y se aplique el modelo predictivo se obtendrá una calificación que indicará la probabilidad de que se produzca la situación estudiada por el modelo.

2.1 Modelos aplicables en el análisis predictivo

Generalmente, se usa el término análisis predictivo cuando en realidad se está hablando del modelado predictivo, que realiza calificaciones mediante modelos predictivos y pronósticos. Sin embargo, cada vez se está utilizando más el término para referirse a todo lo relacionado con la disciplina analítica, como el modelado descriptivo o el modelado decisivo. Estas disciplinas implican un riguroso análisis de datos y son ampliamente utilizadas en negocios mecanismo de ayuda a la toma de decisiones.

Un modelo predictivo es un mecanismo que predice el comportamiento de un individuo. Utiliza las características del individuo como entrada y proporciona una calificación predictiva como salida. Cuanto más elevada es la calificación, más alta es la probabilidad de que el individuo exhiba el comportamiento predicho [10].

2.1.1 Modelos predictivos

Los modelos predictivos son modelos de la relación entre el rendimiento específico de una unidad en una muestra y uno o más atributos o características de esta. El objeto del modelo es evaluar la probabilidad de que una unidad similar en una muestra diferente exhiba un comportamiento específico. Esta categoría abarca modelos que se encuentran en muchas áreas, como el *marketing*, donde buscan patrones de datos ocultos para responder preguntas sobre el desempeño del cliente, como los modelos de detección de fraude.

Los modelos predictivos a menudo ejecutan cálculos durante las transacciones en curso, por ejemplo, para evaluar el riesgo o la oportunidad de un cliente o transacción en particular, de forma que aporte conocimiento a la hora de tomar una decisión. Con los avances en la velocidad de computación, los sistemas de modelado de agentes individuales han sido capaces de simular el comportamiento humano o reacciones ante estímulos o escenarios específicos.

El análisis predictivo construye un modelo estadístico que utiliza los datos existentes para predecir datos de los cuales no se dispone. Como ejemplo del análisis predictivo se incluyen las líneas de tendencia o la puntuación de la influencia [6].

Para la creación del modelo predictivo se utilizan unidades de muestra disponibles con atributos conocidos y un comportamiento conocido, a este conjunto de datos se le denomina conjunto de entrenamiento. Por otro lado, se utilizará una serie de unidades de otra muestra con atributos similares, pero de las cuales no se conoce su comportamiento, a este conjunto de datos se le denomina conjunto de prueba [9].

2.1.2 Modelos descriptivos

Los modelos descriptivos cuantifican las relaciones entre los datos de manera que son utilizadas a menudo para clasificar clientes o contactos en grupos. A diferencia de los modelos predictivos que se centran en predecir el comportamiento de un cliente en particular, los modelos descriptivos identifican diferentes relaciones entre los clientes y los productos.

La analítica descriptiva proporciona resúmenes simples sobre la muestra y sobre las observaciones que se han hecho. Estos resúmenes pueden constituir la base de la descripción inicial de los datos como parte de un análisis estadístico más amplio, o pueden ser suficientes en sí mismos para una investigación en particular.

Los modelos descriptivos no clasifican ni ordenan a los clientes por su probabilidad de realizar una acción particular, de la forma en la que lo hacen los modelos predictivos. Sin embargo, los modelos descriptivos pueden ser utilizados, por ejemplo, para asignar categorías a los clientes según su preferencia en productos o su franja de edad. Estos modelos descriptivos pueden ser utilizados para desarrollar nuevos modelos que pueden imitar un gran volumen de agentes individuales y hacer predicciones. Entre los modelos descriptivos se pueden citar los modelos de simulación, la teoría de colas o las técnicas de previsión.

El análisis descriptivo calcula estadísticas descriptivas para resumir los datos. La mayoría de los análisis sociales pertenecen a esta categoría.

2.1.3 Modelos de decisión

Los modelos de decisión describen la relación entre todos los elementos de una decisión – los datos conocidos (incluyendo los resultados de los modelos predictivos), la decisión y el pronóstico de los resultados de una decisión – con la intención de predecir los resultados de una decisión en la que se involucran gran cantidad de variables.

Estos modelos pueden ser utilizados en la optimización o maximización de determinados resultados mientras minimizan otros. Los modelos de decisión se utilizan, en general, para el desarrollo de la decisión lógica o conjunto de reglas de negocio que deberían producir el resultado deseado para cada cliente o circunstancia.

Los modelos de decisión se usan para modelar una decisión que se toma una vez, así como para modelar un enfoque de toma de decisiones repetible que se utilizará una y otra vez. Como ejemplos de este tipo de modelo cabe destacar los árboles de decisión, el análisis Pareto, el análisis SWOT o el análisis de la matriz de decisiones [10].

2.1.4 Modelos de conjuntos

Los modelos de conjuntos (*ensemble models*) consisten en la aplicación del modelado predictivo para combinar dos o más modelos y luego sintetizar los resultados en una sola puntuación o propagación para mejorar la precisión. Al aplicar un solo modelo basado en una muestra de datos puede tener sesgos, una alta variabilidad o inexactitudes absolutas que afectan la confianza de sus hallazgos analíticos. El uso de técnicas de modelado específicas puede presentar inconvenientes similares. Al combinar diferentes modelos o analizar múltiples muestras, se pueden reducir los efectos de esas limitaciones.

Este sistema de modelado considera las predicciones de ambos modelos caso por caso. En ciertos casos, puede dar más credibilidad a un modelo sobre otro, o al revés. Al hacerlo, el modelo de conjuntos se capacita para predecir qué casos son puntos débiles para cada modelo que lo compone. Puede haber muchos casos en los que los dos modelos están de acuerdo, pero cuando hay desacuerdo, el trabajo conjunto de los modelos ofrece la oportunidad de mejorar el rendimiento.

Un ejemplo de modelado de conjuntos es el modelo de bosque aleatorio (*random forest*). Este modelo combina árboles de decisión que pueden analizar diferentes datos de muestra, evaluar diferentes factores o variables comunes de peso de manera diferente. Los resultados de los diversos árboles de decisión se convierten entonces en un promedio simple o agregados a través de una ponderación adicional.

2.1.5 Modelos de elevación

Los modelos de elevación (modelo *uplift*), también conocido como modelo incremental o de red, es una técnica de modelado predictivo que modela el impacto incremental producido por el tratamiento (como una acción de *marketing*) sobre el comportamiento de un sujeto. Este modelo predictivo predice la influencia de un tratamiento en el comportamiento de un individuo.

El modelo *uplift* utiliza un control científico aleatorio para medir no sólo la eficacia de una acción de *marketing*, sino también para construir un modelo predictivo para la respuesta incremental a la acción de *marketing*. Se trata de una técnica de minería de datos que se ha aplicado principalmente en las industrias de servicios financieros, telecomunicaciones y comercio minorista para la retención de clientes y ventas cruzadas.

Este modelo trata de identificar el *uplift* de una acción, por ejemplo, de una campaña de *marketing*. El *uplift* se define como la diferencia de la tasa de respuesta entre un grupo, tratado mediante una campaña de *marketing*, y otro grupo aleatorio de control. Ambos grupos tendrán las mismas características salvo que el grupo de control no recibirá el tratamiento.

2.1.6 Validación de los modelos

Una vez se ha creado un modelo es necesario comprobar que este funciona de manera correcta, pues este es el aspecto más importante de los modelos predictivos, su validación. Dado que es relativamente fácil crear un modelo, se hace muy importante la validación, ya que es la única forma de saber si el modelo funciona.

Una manera muy extendida para comprobar el modelo consiste en dividir el conjunto de datos del que se dispone en dos. Por un lado, se dispone de un conjunto de datos sobre el cual se desarrollará el modelo, este conjunto abarcará dos terceras partes de la muestra y se denomina *training set* (conjunto de entrenamiento). Por otro lado, la tercera parte sobrante se utilizará para validar el modelo y se denomina *test*

set (conjunto de prueba).

2.2 Técnicas aplicables al análisis predictivo

Los enfoques y técnicas utilizados para realizar el análisis predictivo pueden agruparse de una manera muy general en técnicas de regresión y técnicas de aprendizaje computacional.

Técnicas de regresión

2.2.1 Modelo de regresión lineal

El modelo de regresión lineal analiza la relación existente entre la variable dependiente o de respuesta y un conjunto de variables independientes o predictoras. Esta relación se expresa como una ecuación que predice la variable de respuesta como una función lineal de los parámetros. Estos parámetros se ajustan para que la medida de ajuste sea óptima. Gran parte del esfuerzo en la adaptación del modelo se centra en minimizar el error, así como en asegurarse que está distribuido de forma aleatoria respecto a las predicciones del modelo.

La fórmula de la regresión lineal se expresa matemáticamente como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

donde:

- Y es la variable dependiente que queremos predecir,
- β_0 es el intercepto con el eje Y,
- $\beta_1, \cdots, \beta_p$ representan los coeficientes,
- $X_1, \cdots, X_p$ representan las variables independientes, y
- ε es el término de error.

Para utilizar la regresión lineal, primero es necesario tener datos de X (variables independientes) y Y (variable dependiente). Luego, se pueden estimar los coeficientes β_0 y β_1 que mejor ajustan los datos mediante métodos de optimización, como el método de mínimos cuadrados. Una vez que se obtienen los coeficientes, se utiliza la fórmula de regresión para predecir Y para nuevos valores de X.

El objetivo de la regresión es seleccionar los parámetros del modelo que minimizan la suma de los errores al cuadrado. Esto se conoce como estimación de mínimos cuadrados ordinarios y se logran mejores estimaciones lineales no sesgadas de los parámetros si y sólo si se satisfacen las suposiciones de Gauss-Markov.

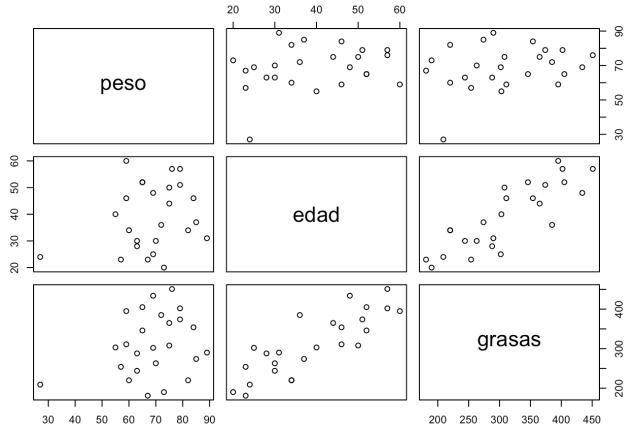
Una vez que se ha estimado el modelo, es necesario saber si las variables predictoras pertenecen al mismo. Para ello, se comprueba la significación estadística de los coeficientes del modelo que pueden medirse utilizando el estadístico t. Esto equivale a probar si el coeficiente es significativamente diferente de cero.

Ejemplo 1 *Los datos del fichero corresponden a tres variables medidas en 25 individuos: edad, peso y cantidad de grasas en sangre. Para leer el fichero de datos y saber los nombres de las variables:*

```
grasas <- read.table('http://verso.mat.uam.es/~
  joser.berrendero/datos/EdadPesoGrasas.txt',
  header = TRUE)
```

Para determinar las relaciones existentes entre cada par de variables se representa una matriz de diagramas de dispersión. Al parecer existe una relación lineal bastante clara entre la edad y las grasas, pero no entre los otros dos pares de variables. Por otra parte, el fichero contiene un dato atípico.

```
pairs(grasas)
```



Para cuantificar el grado de relación lineal, se calcula la matriz de coeficientes de correlación:

```
cor(grasas)

##              peso      edad    grasas
## peso    1.0000000 0.2400133 0.2652935
## edad    0.2400133 1.0000000 0.8373534
## grasas  0.2652935 0.8373534 1.0000000
```

Este ejemplo ha sido tomado de https://rpubs.com/joser/RegresionSimple.

Representación de la recta de mínimos cuadrados

El comando básico es `lm` (linear models) donde el primer argumento de este comando es una fórmula $y \sim x$ en la que se especifica cuál es la variable respuesta o dependiente (y) y cuál es la variable regresora o independiente (x). Mediante el comando `summary` se obtiene un resumen de los principales resultados:

```
regresion <- lm(grasas ~ edad, data = grasas)
summary(regresion)

## Call:
## lm(formula = grasas ~ edad, data = grasas)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.478 -26.816  -3.854  28.315  90.881
## Coefficients:
##             Estimate Std. Error t value Pr(>|t
    |)
```

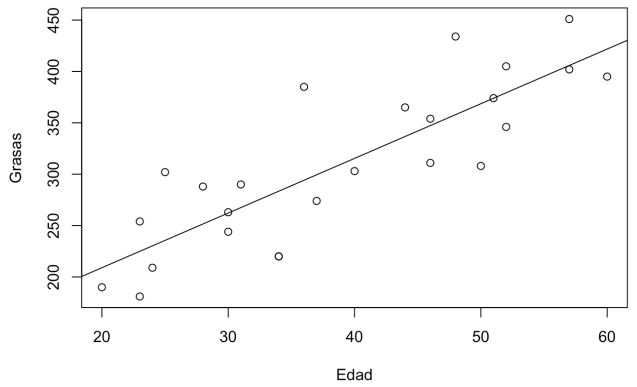
```
## (Intercept) 102.5751    29.6376   3.461
    0.00212 **
## edad          5.3207     0.7243   7.346 1.79e
    -07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
    0.05 '.' 0.1 ' ' 1
## Residual standard error: 43.46 on 23 degrees of
     freedom
## Multiple R-squared:  0.7012, Adjusted R-squared
    :  0.6882
## F-statistic: 53.96 on 1 and 23 DF,  p-value:
    1.794e-07
```

La columna `Estimate` de la tabla `Coefficients` de la salida anterior define la relación entre grasa en sangre y peso.

El coeficiente de determinación (es decir, el coeficiente de correlación al cuadrado) mide la bondad del ajuste de la recta a los datos. A partir de la salida anterior, vemos que su valor en este caso es `Multiple R-squared: 0.701`.

Para graficar la recta se utiliza `plot`:

```
plot(grasas$edad, grasas$grasas, xlab='Edad', ylab
    ='Grasas')
abline(regresion)
```



Cálculo de predicciones

Para predecir sobre nuevos valores basta agregar nuevos datos, sea directamente o en un nuevo archivo.

```
nuevas.edades <- data.frame(edad = seq(30, 50))
predict(regresion, nuevas.edades)

##        1        2        3        4        5
         6
## 262.1954 267.5161 272.8368 278.1575 283.4781
    288.7988
##        9       10       11       12       13
        14
## 304.7608 310.0815 315.4022 320.7229 326.0435
    331.3642
##       17       18       19       20       21
## 347.3263 352.6469 357.9676 363.2883 368.6090
```

2.2.2 Análisis de supervivencia o duración

El análisis de supervivencia es otro nombre para el análisis del tiempo hasta el evento. Estas técnicas se desarrollaron principalmente en las ciencias médicas y biológicas, pero también se usan ampliamente en las ciencias sociales como la economía, así como en la ingeniería (fiabilidad y análisis del tiempo de falla).

La censura y la no-normalidad, que son características de los datos de supervivencia, generan dificultad al intentar analizar los datos usando modelos estadísticos convencionales como la regresión lineal múltiple. La distribución normal, que es una distribución simétrica, toma tanto valores positivos como negativos, pero la duración por su propia naturaleza no puede ser negativa y, por lo tanto, no se puede asumir la normalidad cuando se trata de datos de duración/supervivencia. Por lo tanto, la suposición de normalidad de los modelos de regresión es violada.

El supuesto es que si los datos no fueron censurados sería representativo de la población de interés. En el análisis de supervivencia, las observaciones censuradas surgen cuando la variable dependiente de interés representa el tiempo hasta un evento terminal, y la duración del estudio es limitada en el tiempo.

Un concepto importante en el análisis de supervivencia es la tasa de riesgo, definida como la probabilidad de que el evento ocurra en el tiempo t condicional a sobrevivir hasta el tiempo t. Otro concepto relacionado con la tasa de riesgo es la función de supervivencia que puede definirse como la probabilidad de sobrevivir al tiempo t.

La mayoría de los modelos intentan modelar la tasa de riesgo eligiendo la distribución subyacente dependiendo de la forma de la función de riesgo. Una distribución cuya función de riesgo se inclina hacia arriba se dice que tiene una dependencia de duración positiva, un riesgo decreciente muestra una dependencia de duración negativa mientras que un riesgo constante es un proceso sin memoria usualmente caracterizada por la distribución exponencial.

Función de supervivencia Es la probabilidad de que el evento de interés suceda después del tiempo t.

$$S(t) = P(T > t).$$

Estimador de Kaplan-Meier

$$\hat{S}(t) = \prod_{i:Y_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

donde:

- Y_i: es el tiempo al evento,
- d_i: número de muertes en Y_i,
- n_i: número de individuos a riesgo justo antes de Y_i, y
- $\frac{d_i}{n_i}$: riesgo en Y_i.

2.2.3 Árboles de clasificación y regresión

El análisis discriminante óptico jerárquico (*hierarchial optimal discriminat analysis*, HODA) es una generalización del análisis discriminante óptimo que puede ser utilizado para identificar el modelo estadístico que tiene la máxima precisión para predecir el valor de una variable categórica dependiente para un conjunto de datos que consiste en variables categóricas y variables continuas.

La salida de HODA es un árbol que combina variables categóricas y puntos de corte para variables continuas que proporciona máxima precisión predictiva y una evaluación de potencial generalización cruzada del modelo estadístico.

El análisis discriminante óptimo es una alternativa al análisis de varianza (ANOVA) y al análisis de regresión, que intentan expresar una variable dependiente como una combinación lineal de otras características o medidas. Sin embargo, ANOVA y el análisis de regresión dan una variable dependiente que es una variable numérica, mientras que el análisis discriminante óptimo jerárquico da una variable dependiente que es una variable de clase [8].

Los árboles de clasificación y regresión (*classification and regression trees*, CART) son una técnica de aprendizaje de árboles de decisión no paramétrica que produce árboles de clasificación o regresión, dependiendo de si la variable dependiente es categórica o numérica, respectivamente [14].

Los árboles de decisión están formados por una colección de reglas basadas en variables en el conjunto de datos de modelado:

- Las reglas basadas en valores de variables se seleccionan para obtener la mejor división para diferenciar observaciones basadas en la variable dependiente.
- Una vez que se selecciona una regla y divide un nodo en dos, se aplica el mismo proceso a cada nodo secundario, es decir, es un procedimiento recursivo.
- La división se detiene cuando CART detecta que no se pueden realizar más ganancias o se cumplen algunas reglas de parada preestablecidas. Cada rama del árbol finaliza en un nodo terminal. Cada observación cae en un nodo terminal, y cada nodo terminal es definido de manera única por un conjunto de reglas.

Los árboles de decisión se utilizan para:

- Clasificación:
  - Binaria: fraude/no fraude, morosidad, *spam* en correos.
  - Multiclase: niveles de satisfacción (completamente, bastante, poco satisfecho, totalmente insatisfecho).
- Regresión:
  - Pagos de compañías de seguros.
  - Gasto de compras por clientes.

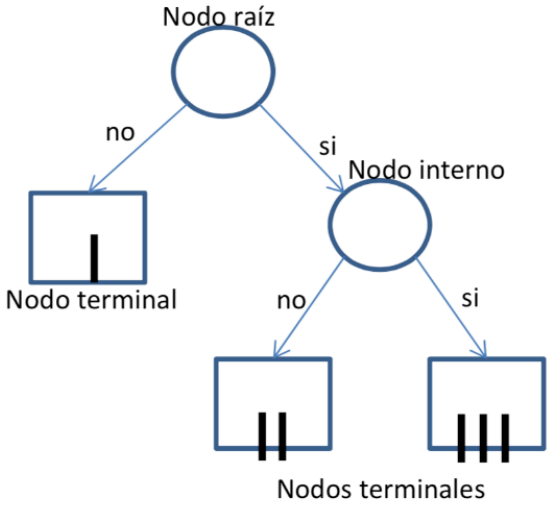


Figura 1. Elementos de un árbol de decisión [*Elements of a decision tree*].

Todo árbol tiene como mínimo un nodo raíz, un nodo interno y nodos terminales (Figura 1). El nodo raíz es particionado en nodos internos (hijos o ramas). Se busca la homogeneidad de los nodos terminales, esto es que las variables resultantes en cada nodo terminal sean homogéneas. Lo que está definido por una noción de impureza que puede estar determinado a su vez por tres funciones diferentes: mínimo error o error de Baye, entropía e índice de Gini.

Los nodos terminales son aquellos que ya no pueden ser divididos en base a las variables determinantes. Los árboles de decisión se generan a partir de algoritmos de segmentación recursiva, que determinan el mecanismo de segmentación y el criterio de parada definiendo el nodo terminal, esta determinado por tres procedimientos: *Chi-Square Automatic Interaction Detector* (CHAID), creado por IBM e incluido en el software SPSS Modeler; *Quick Unbiased Efficient Statistical Tree* (QUEST) y CART.

CART

Este algoritmo fue desarrollado en la Universidad de Berkeley y consiste en tres pasos:

1. Construcción del árbol.
2. Elección del tamaño correcto.
3. Clasificación de nuevos datos a partir del árbol ya construido.

El algoritmo trabaja con variables de todo tipo. No necesita discretizar las variables explicativas continuas. El corte en cada nodo viene dado por reglas de tipo binario. Se pueden formular como preguntas: ¿Es $X < a$? ¿Pertenece X a un subconjunto E de estados? Da lugar a estructuras de árbol de mayor profundidad.

Ejemplo 2 *Para este ejemplo se utilizó la información en los siguientes enlaces:*

- *Dataset:* https://www.kaggle.com/janiobachmann/bank-marketing-dataset
- https://rpubs.com/Peters64s/451160
- https://rpubs.com/jboscomendoza/arboles_decision_clasificacion
- https://rpubs.com/amaurandi/ejemploTree
- https://rpubs.com/Cristina_Gil/arboles_ensemble

```
library(tidyverse)
library(rpart)
library(rpart.plot)
library(caret)

# Datos
download.file("https://archive.ics.uci.edu/ml/
    machine-learning-databases/wine/wine.data", "
    wine.data")

# Información
download.file("https://archive.ics.uci.edu/ml/
    machine-learning-databases/wine/wine.names", "
    wine.names")
vino <- read.table("wine.data", sep = ",", header
    = FALSE)

# Copiar archivo de nombres de variables de vinos
file.copy(from = "wine.names", to = "wine_names.
    txt")
file.show("wine_names.txt")

# Obtener los nombres de las columnas
nombres <- readLines("wine_names.txt")[58:70] %>%
    gsub("[[cntrl:]].*\\)", "", .) %>% trimws()
    %>% tolower() %>% gsub(" |/", "_", .) %>%

# Agregar el nombre "tipo", para la primera
    columna con los tipos de vino
c("tipo", .)
names(vino) <- nombres

# Cambiar el tipo de dato de la columna tipo a
    factor usando la función mutate_a t() de dplyr
    , para poder hacer clasificaciones
vino <- vino %>%
mutate_at("tipo", factor)
head(vino)

## tipo alcohol malic_acid ash alcalinity_of_ash
    magnesium total_phenols
## 1 1 14.23 1.71 2.43 15.6 127 2.80
## 2 1 13.20 1.78 2.14 11.2 100 2.65
## 3 1 13.16 2.36 2.67 18.6 101 2.80
## 4 1 14.37 1.95 2.50 16.8 113 3.85
## 5 1 13.24 2.59 2.87 21.0 118 2.80
## 6 1 14.20 1.76 2.45 15.2 112 3.27

## flavanoids nonflavanoid_phenols proanthocyanins
     color_intensity hue
## 1 3.06 0.28 2.29 5.64 1.04
## 2 2.76 0.26 1.28 4.38 1.05
## 3 3.24 0.30 2.81 5.68 1.03
## 4 3.49 0.24 2.18 7.80 0.86
## 5 2.69 0.39 1.82 4.32 1.04
## 6 3.39 0.34 1.97 6.75 1.05

## od280_od315_of_diluted_wines proline
## 1 3.92 1065
## 2 3.40 1050
## 3 3.17 1185
## 4 3.45 1480
## 5 2.93 735
## 6 2.85 1450
```

Creando conjuntos de entrenamiento y prueba

```
# Usar la función sample_frac() de dplyr para
    obtener un subconjunto de los datos, que
    consiste en 70% del total de ellos. Usar set.
    seed() para que este ejemplo sea reproducible.
set.seed(1649)
vino_entrenamiento <- sample_frac(vino, .7)

# Con setdiff() de dplyr, se obtiene el
    subconjunto de datos complementario al de
    entrenamiento para el conjunto de prueba, esto
     es, el 30% restante.
vino_prueba <- setdiff(vino, vino_entrenamiento)
```

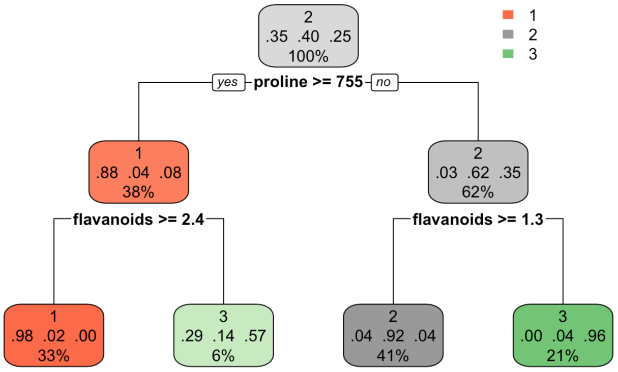
Entrenamiento `rpart` es la función que pide una fórmula para especificar la variable objetivo de la clasificación. La fórmula que se usará es tipo $\sim$, la cual expresa que se desea clasificar `tipo` usando a todas las demás variables como predictoras:

```
arbol_1 <- rpart(formula = tipo ~ ., data = vino_
    entrenamiento)
arbol_1

## n= 125
##
## node), split, n, loss, yval, (yprob)
## * denotes terminal node
##
## 1) root 125 75 2 (0.35200000 0.40000000
    0.24800000)
## 2) proline>=755 48 6 1 (0.87500000 0.04166667
    0.08333333)
## 4) flavanoids>=2.35 41 1 1 (0.97560976
    0.02439024 0.00000000) *
## 5) flavanoids< 2.35 7 3 3 (0.28571429
    0.14285714 0.57142857) *
## 3) proline< 755 77 29 2 (0.02597403 0.62337662
    0.35064935)
## 6) flavanoids>=1.265 51 4 2 (0.03921569
    0.92156863 0.03921569) *
## 7) flavanoids< 1.265 26 1 3 (0.00000000
    0.03846154 0.96153846) *

rpart.plot(arbol_1)
```

Dentro del rectángulo de cada nodo se muestra qué proporción de casos pertenecen a cada categoría y la proporción del total de datos que han sido agrupados allí. Por ejemplo, el rectángulo en el extremo inferior izquierdo de la gráfica tiene 92 % de casos en el tipo 1, y 4 % en los tipos 2 y 3, que representan 39 % de todos los datos.



2.2.4 Curvas de regresión adaptativa multivariable

Las curvas de regresión adaptativa multivariable (*multivariate adaptative regression splines*, MARS) son una técnica no paramétrica que construye modelos flexibles al ajustar regresiones lineales por piezas. Un concepto importante asociado con curvas de regresión es el de un nudo. Un nudo es el punto donde un modelo de regresión local da paso a otro y, por lo tanto, es el punto de intersección entre dos curvas.

En las curvas de regresión adaptativa multivariante, las funciones de base son la herramienta utilizada para generalizar la búsqueda de nudos. Las funciones básicas son un conjunto de funciones utilizadas para representar la información contenida en una o más variables.

El modelo MARS casi siempre crea las funciones de base en parejas. La curva de regresión adaptativa multivariable es un modelo que primero realiza un sobreajuste y luego hace una poda para obtener un modelo óptimo. El algoritmo es computacionalmente muy intensivo y en la práctica se requiere especificar un límite superior en el número de funciones de base.

Cabe destacar que estos no son los únicos modelos, existen otros modelos de regresión como los modelos de elección discreta, de regresión logística, de regresión logística multinomial, modelos *probit*, o los modelos de series temporales. Más detalles pueden encontrase en [13].

Técnicas de aprendizaje computacional

2.2.5 Redes neuronales

Las redes neuronales son técnicas de modelado no lineal sofisticadas que son capaces de modelar funciones complejas. Pueden aplicarse a problemas de predicción, clasificación o control en un amplio espectro de campos como las finanzas, la psicología cognitiva/neurociencia, la medicina, la ingeniería y la física.

Las redes neuronales se utilizan cuando no se conoce la naturaleza exacta de la relación entre los valores de entrada y de salida. Una característica clave de las redes neuronales es que aprenden la relación entre los valores de entrada y salida a través del entrenamiento.

Existen tres tipos de entrenamiento en redes neuronales utilizadas por diferentes redes, el aprendizaje por refuerzo, el supervisado y no supervisado, siendo el supervisado el más común [3].

2.2.6 Máquinas de vectores de soporte

Las máquinas de vectores de soporte (SVM, por sus siglas en ingl es) se usan para detectar y explotar patrones complejos de datos agrupando, ordenando y clasificando los datos. Son máquinas de aprendizaje que se utilizan para realizar clasificaciones binarias y estimaciones de regresión. Usualmente usan métodos basados en *kernel* para aplicar técnicas de clasificación lineal a problemas de clasificación no lineal. Hay una serie de tipos de SVM tales como lineal, polinomial, sigmoidal, etc. [7].

2.2.7 Naïve Bayes

El clasificador bayesiano ingenuo se basa en la regla de probabilidad condicional de Bayes, que se utiliza para la tarea de clasificación. El clasificador bayesiano asume que los predictores son estadísticamente independientes, lo que hace que sea una herramienta de clasificación eficaz que sea fácil de interpretar. Se emplea mejor cuando se enfrenta al problema de la "maldición de la dimensionalidad", es decir, cuando el número de predicciones es muy alto [2].

2.2.8 $k-$vecinos más cercanos

El algoritmo $l-$vecinos más cercanos ($k-$*nearest neighbors*, $k-$NN) pertenece a la clase de métodos estadísticos de reconocimiento de patrones. El método no impone *a priori* ninguna suposición sobre la distribución de la que se extrae la muestra de modelado. Se trata de un conjunto de entrenamiento con valores positivos y negativos. Una nueva muestra se clasifica calculando la distancia al vecino más cercano del conjunto de entrenamiento. El signo de ese punto determinará la clasificación de la muestra. En el clasificador $k-$vecino más cercano, se consideran los k puntos más cercanos y se utiliza el signo de la mayoría para clasificar la muestra.

El rendimiento del algoritmo $k-$NN está influenciado por tres factores principales:

- La medida de distancia utilizada para localizar a los vecinos más cercanos.
- La regla de decisión usada para derivar una clasificación de los $k-$vecinos más cercanos.
- el número de vecinos utilizados para clasificar la nueva muestra.

Se puede demostrar que, a diferencia de otros métodos, este método es universal y asintóticamente convergente, es decir, a medida que el tamaño del conjunto de entrenamiento aumenta, si las observaciones son independientes e idénticamente distribuidas, independientemente de la distribución a partir de la cual se dibuja la muestra, la clase predicha convergerá a la asignación de clase que minimiza el error de clasificación errónea.

Al igual que en apartado anterior, estos nos son las únicas técnicas de aprendizaje computacional, existen otras como la función de base radial, el perceptrón multicapa o modelado predictivo geoespacial [5].

2.3 Ventajas del análisis predictivo

A continuación se mencionan los beneficios del análisis predictivo:

- **Mejora en la toma de decisiones**: La medida en que el análisis predictivo puede mejorar el proceso de toma de decisiones de una organización está directamente relacionada con la cantidad de datos a los que tiene acceso la organización.

  El análisis predictivo permite a las empresas identificar patrones en el comportamiento de compra de los clientes, determinar que prácticas ayudan o dificultan las ganancias y decidir que acciones tomar para mejorar el negocio.

- **Aumenta la eficiencia**: Muchas industrias dependen de mantenimiento predictivo para mantener el funcionamiento de los equipos y reducir las interrupciones en la cadena de suministro de una empresa.

  El análisis predictivo no solo aumenta la eficiencia al prevenir el mal funcionamiento de los equipos, sino que también permite identificar nuevas formas de agilizar las transacciones comerciales, reducir los desperdicios innecesarios y permitir que las empresas se adapten a las tendencias más rápidamente.

- **Mejora la gestión de riesgos**: Cada industria implica una cierta cantidad de riesgo en sus operaciones diarias. La forma en que las empresas equilibran estos riesgos puede determinar su éxito o su fracaso. Una gestión eficaz de los riesgos permite a las empresas crecer y expandirse en la dirección correcta.

  El análisis predictivo permite estudiar cantidades masivas de datos para anticipar y prevenir fraudes, detectar vulnerabilidades y evitar pérdidas financieras importantes.

- **Aumenta las ventas**: Al estudiar los patrones de comportamiento humano, el análisis predictivo puede ayudar a las empresas a aumentar sus ganancias. Esta forma de análisis de datos permite a las organizaciones realizar un seguimiento de los clientes individuales y crear estrategias de marketing personalizadas adaptadas a los intereses de cada persona.

  Las organizaciones pueden distinguir las campañas de marketing exitosas de las que no lo son para crear las condiciones necesarias para que un cliente desee un producto.

- **Proporciona inteligencia competitiva**: La información que brindan las métricas predictivas puede ser el arma secreta de una empresa. Cualquier empresa compite con docenas de otras para ofrecer a los clientes el mismo producto, por lo que tener una ventaja sobre los competidores puede significar la diferencia entre obtener o perder ganancias.

- **Mejora la gestión de la cadena de suministro**: El análisis predictivo agiliza la gestión de la cadena de suministro al hacer un seguimiento de cómo se utilizan los recursos y predecir cuándo es necesario reponerlos. Este tipo de análisis puede identificar patrones en los que se envían recursos para que el reabastecimiento sea un proceso más automatizado.

2.4 Desventajas del análisis predictivo

Si bien las herramientas de análisis predictivo pueden ser útiles en el arsenal de una empresa, existen algunas desventajas que los líderes de la organización deben tener en cuenta:

- **No se puede predecir todo el comportamiento humano**: Es cierto que el análisis predictivo puede predecir de forma precisa y fiable el comportamiento humano, y que esta herramienta puede suponer un cambio radical para muchas empresas. Sin embargo, es importante reconocer que no todo el comportamiento humano se puede predecir.

  La crisis financiera mundial es un ejemplo de la incapacidad de la tecnología para prever todos los resultados probables.

- **Los conjuntos de datos deben actualizarse constantemente**: El éxito del análisis predictivo depende de la actualización constante de la información. En este sentido, el tiempo es un factor importante en la precisión de una predicción. Los datos de un año anterior pueden estar demasiado desactualizados para predecir tendencias y patrones en el mercado global actual, lo que podría causar importantes pérdidas financieras para una organización.

- **Debe tener objetivos claros**: Antes de que una organización invierta en análisis predictivos, es fundamental tener claros los objetivos que se deben alcanzar o los problemas que se deben resolver. De lo contrario, se puede perder tiempo y dinero valioso extrayendo datos que no tienen correlaciones reales que comprender.

  Cuando comienzan a trabajar con el análisis predictivo, algunas empresas creen que extraer datos (todos los datos) les permitirá obtener información valiosa que les permitirá cambiar la forma de operar de la empresa. Este tipo de pensamiento generalizado puede ser peligroso, ya que el verdadero valor del análisis predictivo proviene de preguntas bien pensadas sobre problemas conocidos.

- **Datos incompletos**: Las organizaciones que utilizan el análisis predictivo parten del supuesto de que hay

suficientes datos disponibles para generar información útil. Pero, ¿qué sucede cuando un conjunto de datos no está completo? Un conjunto de datos incompletos distorsionará la información, lo que puede aumentar los riesgos de una empresa.

- **Algunos datos pueden ser inexactos**: Las empresas que se basan en datos recopilados mediante encuestas saben que no todos los clientes brindan información honesta o precisa. Los datos inexactos no ocurren porque las personas sean deshonestas, sino que pueden estar más influenciados por reservas personales. En cualquier caso, los datos inexactos solo brindarán información sesgada.

Conclusiones

El análisis predictivo ha dejado de estar reservado a grandes corporaciones, gobiernos o universidades y se ha generalizado como una herramienta más de la inteligencia de negocios (*busines intelligence*) a disposición de todo tipo de empresas y organizaciones.

El requerimiento fundamental para realizar análisis predictivo es la existencia de un conjunto, lo suficientemente amplio de datos, como para permitir detectar en ellos patrones que permitan formular reglas capaces de anticipar previsiones [11].

La capacidad de almacenar y gestionar conjuntos de datos masivos ha crecido de manera exponencial en los últimos años, al tiempo que ha aparecido una cultura empresarial y gubernamental que apuesta por la recolección de datos de manera sistematizada, con la confianza de que en algún momento podrá extraerse de los mismos información relevante.

Las operadoras telefónicas almacenan el geoposicionamiento de los usuarios, los bancos almacenan millones de transacciones con tarjetas de crédito, Google permite almacenar en gigabytes de correos electrónicos, al tiempo que en las redes sociales se invita a los usuarios a compartir opiniones, fotos y vídeos.

Todos esos elementos constituyen, en última instancia, datos y del análisis de los datos emergen pautas de comportamiento susceptibles de ser utilizadas en la planificación del transporte o en el *marketing* personalizado entre un sinnúmero de posibles aplicaciones [4].

Esa posibilidad real de almacenar y procesar datos, unida a la cultura de conservarlos, requiere de un tercer elemento: las herramientas capaces de encontrar patrones que permitan formular reglas [1].

Desde la década de los sesenta existen en el mercado potentes herramientas capaces de realizar complejos análisis estadísticos (*SPSS*, *SAP Business Suite* o *SAS Software Package*). Esta oferta de herramientas especializadas se ha visto complementada por herramientas de software libre entre las que destacan las analizadas en este trabajo (*R* y *Weka*).

El análisis de *R* y *Weka* se ha realizado creando con ambas herramientas modelos predictivos. En los dos casos se han realizado árboles de decisión y modelos de agrupamientos con los algoritmos propios de cada una de ellas. En ambos casos se ha finalizado creando un modelo de reglas de asociación utilizando el algoritmo *a priori*.

El resultado de la comparación puede resumirse afirmando que ambas herramientas cumplen con las prestaciones exigibles y que la curva de aprendizaje de *R*, más dura y pronunciada, es compensada por su mayor potencia y flexibilidad.

Suplementos

Este artículo no contiene información suplementaria.

Conflictos de interés

Se declara que no existen conflictos de interés. No existen subvenciones involucradas en este trabajo.

Contribución de autoría

Conceptualización C.E.V.G., C.N.B.H.

Curación de datos A.F.B., L.G.F.

Análisis formal C.E.V.G., M.M.F.L.

Investigación S.M.A.A., G.B.A.

Metodología C.E.V.G., M.M.F.L.

Administración de proyecto S.M.A.A., G.B.A.

Software A.F.B., L.G.F.

Supervisión C.E.V.G., M.M.F.L.

Validación S.M.A.A., G.B.A.

Visualización C.E.V.G.

Redacción: preparación del borrador original C.E.V.G., M.M.F.L.

Redacción: revisión y edición S.M.A.A., G.B.A.

Referencias

[1] Anónimo: *Análisis Predictivo para comprender el consumo de agua y mejorar la administración de recursos*. Be Smart Company. http://www.besmart.company/blog/analisispredictivo-para-comprender-el-consumo-de-agua-y-mejorar-la-administracion-de-recursos.

[2] Barranco Frangoso, R.: *¿Qué es Big Data?* IBM Developer Works, 2012. https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/.

[3] Cía, J.F.: *El ranking de las mejores soluciones de análisis predictivo para empresas*. BBVA Open4u. https://bbvaopen4u.com/es/actualidad/el-ranking-de-lasmejores-soluciones-de-analisis-predictivo-para-empresas.

[4] Hacienda Administraciones Públicas, Ministerio de: *Plan de Transformación Digital de la Administración General del Estado y sus Organismos Públicos (Estrategia TIC 2015-2020)*. Informe técnico, Gobierno de España, 2015. https://administracionelectronica.gob.es/pae_Home/pae_Estrategias/Plan_Digitalizacion_AAPP/planes-anteriores.html.

[5] Mathew, G.: *Five Ways Data Analytics Will Shape Business, Sports And Politics In 2016*. Forbes, 2016. http://www.forbes.com/sites/valleyvoices/2016/01/20/five-ways-data-analytics-will-shape-business-sports-and-politics-in-2016.

[6] Mayor, E.: *Learning Predictive Analytics with R*. Birmingham: Pack Publishing, 2015, ISBN 978-1-78213-935-2. https://www.researchgate.net/publication/308117456_Learning_predictive_analytics_with_R.

[7] Merino, P.P.: *Los datos, el nuevo petróleo del siglo XXI*. Ecommerce News, 2016. http://ecommerce-news.es/actualidad/los-datos-nuevo-petroleo-del-sigloxxi-41824.html.

[8] Molina Félix, L.C. y R. Sangüesa i Solé: *Reglas de asociación. Data mining*, 2022. https://openaccess.uoc.edu/bitstream/10609/138187/25/Data%20mining_Módulo%206_Reglas%20de%20asociación.pdf.

[9] Nyce, C.: *Predictive Analytics White Paper*. Technical report, 2007. https://www.the-digital-insurer.com/wp-content/uploads/2013/12/78-Predictive-Modeling-White-Paper.pdf.

[10] Regan, P.J. and S. Holtzman: *R&D Decision Advisor: An interactive approach to normative decision system model construction*. European Journal of Operational Research, 84(1):116–133, 1995. https://doi.org/10.1016/0377-2217(94)00321-3.

[11] Shimada, T. y F. López: *Analítica Predictiva: cómo convertir la información en ventaja competitiva*. Illatam, 2014. http://www.il-latam.com/wp-content/uploads/2018/08/articulo-revista-109-como-convertir-la-informacion-en-ventaja-competitiva.pdf.

[12] Siegel, E.: *Predictive Analytics: The power to predict who will click, buy, lie, or die*. Wiley, 2013, ISBN 978-1-119-14567-7. https://www.wiley.com/en-ie/Predictive+Analytics%3A+The+Power+to+Predict+Who+Will+Click%2C+Buy%2C+Lie%2C+or+Die%2C+Revised+and+Updated-p-9781119145677.

[13] Solé, R. Sangüesa i: *Agregación (clustering). Data mining*. https://openaccess.uoc.edu/bitstream/10609/138187/24/Data%20mining_Módulo%205_Agregación%20(clustering).pdf.

[14] Solé, R. Sangüesa i: *Clasificación: árboles de decisión. Data mining*. https://openaccess.uoc.edu/bitstream/10609/138187/22/Data%20mining_Módulo%203_Clasificación%2C%20árboles%20de%20decisión.pdf.

[15] Strickland, J.: *Predictive Analytics using R*. Lulu.com, 2014, ISBN 978-1312841017. https://books.google.com.cu/books?hl=en&lr=&id=7l0sCQAAQBAJ&oi=fnd&pg=PR5&dq=Strickland,+Predictive+Analytics+using+R&ots=n2sinD3xvY&sig=TcPDI04TL8tq078_4ZVeiEnZ5I4&redir_esc=y#v=onepage&q=Strickland%2C%20Predictive%20Analytics%20using%20R&f=false.