Artículo Original
Revista Varianza - Año 2022, Nro. 19
ISSN 2789-3529 versión en línea
Páginas: 39-46

Árbol de Decisión en Aprendizaje Automático

Decision Tree in Machine Learning

Fernando Oday Rivero Suguiura[1]
Instituto de Estadística Teórica y Aplicada, Universidad Mayor de San Andrés, La Paz-Bolivia
✉ *friverosuguiura2004@gmail.com*

Artículo recibido: 2022-03-13
Artículo aceptado: 2022-04-01

Resumen

Los árboles de decisión son modelos de clasificación y regresión empleados en el aprendizaje automatizado o también denominado *Machine Learning* de uso en grandes cantidades de datos estadísticos como parte de lo que hoy es la Ciencia de Datos y el *Big Data*.

En este artículo se describe la metodología y se analiza los resultados que arroja el árbol de decisión de clasificación, con uso de variables de la encuesta de hogares, como: género, nivel educativo, ocupación e ingreso en ciudades capitales de Bolivia. Como conclusión, se puede observar que las mujeres están en menor desventaja que los hombres en cuanto a su nivel de formación educativa que influye en la ocupación y el ingreso.

Palabras clave: *Machine Learning*, minería de datos, árbol de decisión, clasificación, entropía, análisis masivo de datos.

Abstract

Decision trees are classification and regression models used in automated learning or also called Machine Learning used in large amounts of statistical data as part of what is now Data Science and Big Data.

This article describes the methodology and analyzes the results of the classification decision tree, using variables from the household survey, such as: gender, educational level, occupation and income in Bolivian capital cities. In conclusion, it can be seen that women are less disadvantaged than men in terms of their level of educational training, which influences occupation and income.

Keywords: Machine Learning, data mining, decision tree, classification, entropy, big data analytics.

1. Introducción

La Minería de Datos comenzó teniendo mucho éxito en los estudios de mercado. Junto con ella se inició la introducción de procesos inductivos basados en los árboles de decisión desarrollados en la Teoría de Decisión. El desarrollo de la informática dio pie a tecnologías especializadas como el aprendizaje automático (*machine learning*) y el reconocimiento de patrones (*pattern recognition*). Un árbol de decisión es una representación de una función multivariada y que fue posible utilizar en la vida práctica a partir del advenimiento de equipos de computación de última generación.

[1] Docente de la carrera de Estadística, Facultad de Ciencias Puras y Naturales de la UMSA. Consultor en muestreo, censos y análisis estadístico en entidades nacionales e internacionales. Magister en Ciencias de la Estadística. *https://orcid.org/0000-0001-9095-7778*

El interés por el uso práctico de los árboles de decisión, tuvo su origen en las necesidades de las ciencias sociales siendo principal el trabajo de Sonquist-Morgan (1964) el software AID (*Automatic Interaction Detection*). Este fue uno de los primeros métodos de ajuste de los datos basados en árboles de clasificación. Con ello los árboles de decisión trascendieron, de ser solo una representación ilustrativa en los cursos de toma de decisiones, para convertirse en una herramienta útil y sencilla de utilizar. Estos avances fueron mejorados por la obra de Breiman-Friedman-Olshen-Stone (1984) llamada (*Classification and Regression Trees*), (Bouza, 2012) - Universidad de La Habana – Cuba, (Santiago, 2012) - Universidad Autónoma de Guerrero – México.

Los datos contienen información oculta potencialmente útil que rara vez se hace explícita o se aprovecha. Con el uso de la minería de datos y las técnicas de aprendizaje automático, se busca determinar patrones de comportamiento en los datos, es decir, se sustenta buscar algún patrón en la información que se encuentre almacenado en las bases de datos, de tal manera que los usuarios puedan solucionar problemas al tratar analíticamente los datos y cuenten con herramientas que les permitan tomar mejores decisiones.

Los algoritmos de aprendizaje basados en árboles decisión se consideran uno de los aventajados y más utilizados métodos de aprendizaje supervisado. Las técnicas basadas en árboles de decisión, potencian los modelos de clasificación y predicción con elevada precisión, confiabilidad, robustez y facilidad de interpretación. A diferencia de los modelos lineales, éstos además consideran posibilidades de inclusión de relaciones no lineales.

Los problemas de clasificación generalmente son aquellos en los que se intentan predecir los valores de una variable dependiente categórica a partir de una o más variables predictoras continuas, discretas o categóricas, por ejemplo, será importante clasificar a las personas por la variable sexo como dependiente y a partir de ella, considerar algunas variables independientes predictivas como el nivel de educación, si trabaja o no y cuánto es su ingreso, etc., que es el ejemplo que se estudia más adelante en este artículo. (Breiman, 1984), (Friedman, 1984).

2. Metodología

2.1 Terminología del árbol de decisión

Un árbol de decisión contiene los siguientes elementos:

- **Nodo de decisión.** Aquel cuando un subnodo se divide en subnodos adicionales indicando la decisión que se tomará ante la disyuntiva.
- **Nodo de probabilidad u hoja.** Presenta múltiples resultados inciertos. Describe las probabilidades de acierto de cada una de las opciones que se plantea ante una disyuntiva.
- **Nodo terminal.** Indica el resultado definitivo como última opción en una ruta de decisión.
- **División**. Proceso de división de un nodo en dos o más subnodos.
- **Flecha**. Nexo que unen los nodos entre sí.
- **Vector**. Representa la opción por la que se opta entre todas las posibilidades que define el nodo.
- **Poda**. Esta se da cuando se reduce el tamaño de los árboles de decisión eliminando nodos opuestos a la división.
- **Rama o subárbol.** Una subsección del árbol de decisión se denomina rama o subárbol.
- **Etiqueta.** Permite la unión de nodos y flechas que denominan las acciones que se llevan a cabo.

2.2. Descripción del árbol de decisión

Un árbol de decisión en *Machine Learning* es una estructura de árbol similar a un diagrama de flujo, ver Figura 1, donde un nodo interno representa una característica (o atributo), la rama es una regla de decisión y cada nodo u hoja simboliza el resultado y el nodo superior se conoce como el nodo raíz. Este tipo de árbol se conoce como árbol de clasificación, donde cada ramificación contiene un conjunto de atributos o reglas de clasificación asociadas a una etiqueta de clase específica que se halla al final de la ramificación.

El árbol de decisión clasificatorio aprende a particionar en función del valor del atributo, éste se divide de manera recursiva buscando un mínimo local de acuerdo a medidas como la entropía en base a un conjunto de datos masivo de entrenamiento.

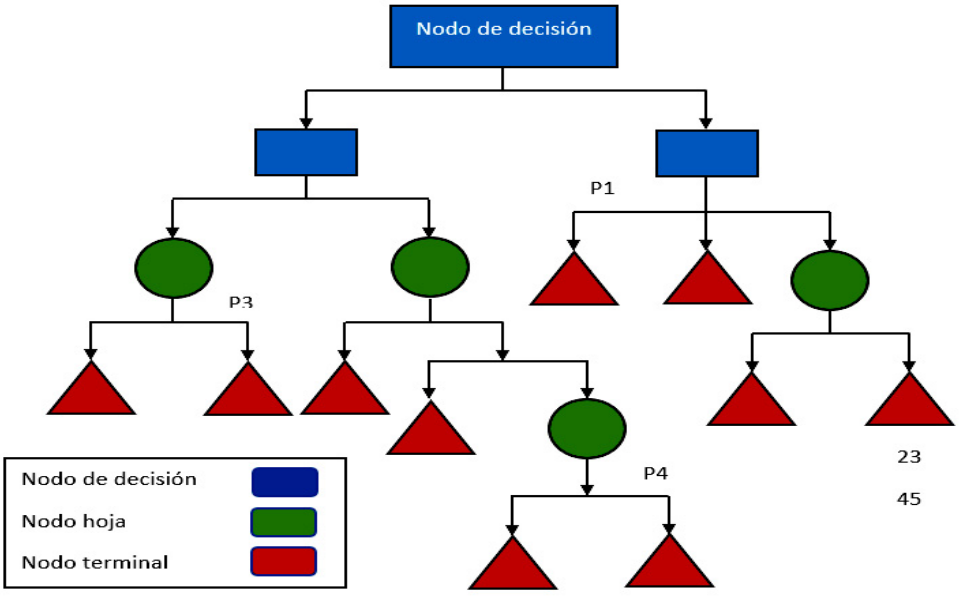
Las ventajas de utilidad de los árboles de decisión en el aprendizaje automático son:

- El costo del uso del árbol para predecir los datos disminuye con cada punto de datos adicional.
- Funciona para los datos numéricos o categóricos.
- Puede modelar problemas con múltiples resultados.
- Usa un modelo de caja blanca (lo que hace que los resultados sean fáciles de explicar).
- La fiabilidad de un árbol se puede cuantificar y poner a prueba.
- Tiende a ser preciso independientemente de si viola las suposiciones de los datos de origen.

Las desventajas que presenta son:

- La incorporación de datos categóricos con múltiples niveles provoca que los resultados se inclinen a favor de los atributos con mayoría de niveles.
- Los cálculos pueden volverse complejos al afrentarse con la falta de certezas y numerosos resultados relacionados.

Figura 1.
Estructura de un árbol de clasificación



Fuente: Elaboración propia

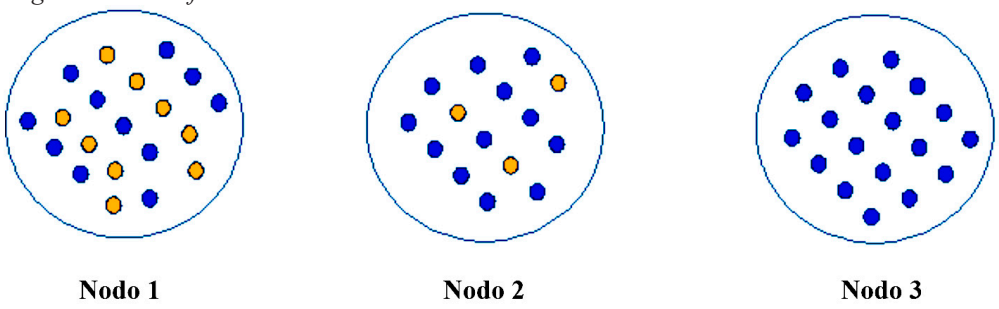
2.3. Medidas de selección

La medida de selección de atributos o características es una heurística para seleccionar el criterio que divida los datos de la mejor manera posible. También se conoce como reglas de división porque ayudan a determinar puntos de interrupción en un nodo dado. Los utilizados, son los siguientes:

Ganancia de información

La ganancia de información, es una propiedad estadística que mide homogeneidad, qué tan bien un atributo se asemeja en el entrenamiento de patrones a otro atributo según la clasificación que se busca. De acuerdo a la Figura 2, se puede observar que el nodo 3 describe muy fácilmente la información pues contiene datos con valores similares (llamado nodo puro), mientras que el nodo 2 todavía requiere regular información de entrenamiento para alcanzar la descripción del nodo 3. A su vez el nodo 1 necesitará mayor información para llegar a los anteriores, a este último se le denomina nodo impuro.

Figura 2.
Nodos con ganancia de información



Fuente: Elaboración propia.

La expresión para determinar la ganancia de información por nodo, es:

$$Ganancia(p, X) = Entropía - \sum_{i=1}^{k} (P_j \text{ x } Entropía)$$

Donde P_j es el conjunto de valores posibles para el atributo X, donde el valor de entropía se define luego.

Entropía

La entropía es una medida de la homogeneidad de los datos aleatorios presentes en un nodo de un árbol de clasificación. Si los datos son similares, el valor de entropía es pequeño o nulo.

De cada nodo se puede medir el valor de entropía para su respectivo control. La expresión siguiente mide entropía:

$$Entropía = -\sum_{i=1}^{k} P_i \log_2 P_i$$

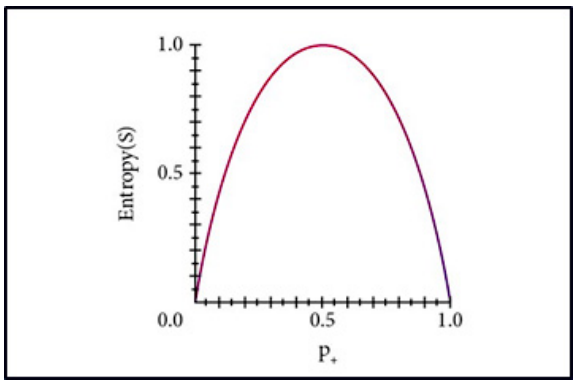
Por ejemplo, si se calcula la entropía de los nodos 1 y 3, se tiene:

$$Entropía_1 = -(11/20)\log_2(11/20)-(9/20)\log_2(9/20)=0{,}993$$
$$Entropía_3 = -(20/20)\log_2(20/20)-(0/20)\log_2(0/20)=0$$

Se concluye que la entropía es cero si todos los elementos del grupo pertenecen a la misma clase, y la entropía es uno cuando la muestra contiene el mismo número de observaciones positivas y negativas. Si la muestra contiene un número desigual de observaciones positiva y

negativas, la entropía está entre 0 y 1. La Figura 3, presenta la forma de la función de entropía en relación con una clasificación booleana.

Figura 3.
Función entropía



Fuente: Elaboración propia

Gini

Gini es una medida de impureza de las observaciones contenidas en un determinado nodo. Si el índice de Gini vale 0, significa que el nodo es totalmente puro. Se determina de la siguiente manera:

$$Gini = \sum_{i=1}^{k} P_i(1 - P_i)$$

Si se determina el índice de Gini para los nodos 1 y 3 anteriores, se tiene:

$$Gini_1 = (11/20)(9/20)+(9/20)(11/20) = 0{,}495$$

$$Gini_3 = (20/20)(0/20)+(0/20)(20/20) = 0$$

Luego el nodo 3 es puro y el nodo 1 es impuro.

3. Resultados

3.1. Aplicación de un árbol de clasificación

Para el entrenamiento del presente árbol de clasificación, se han empleado las siguientes variables

Variable dependiente categórica:

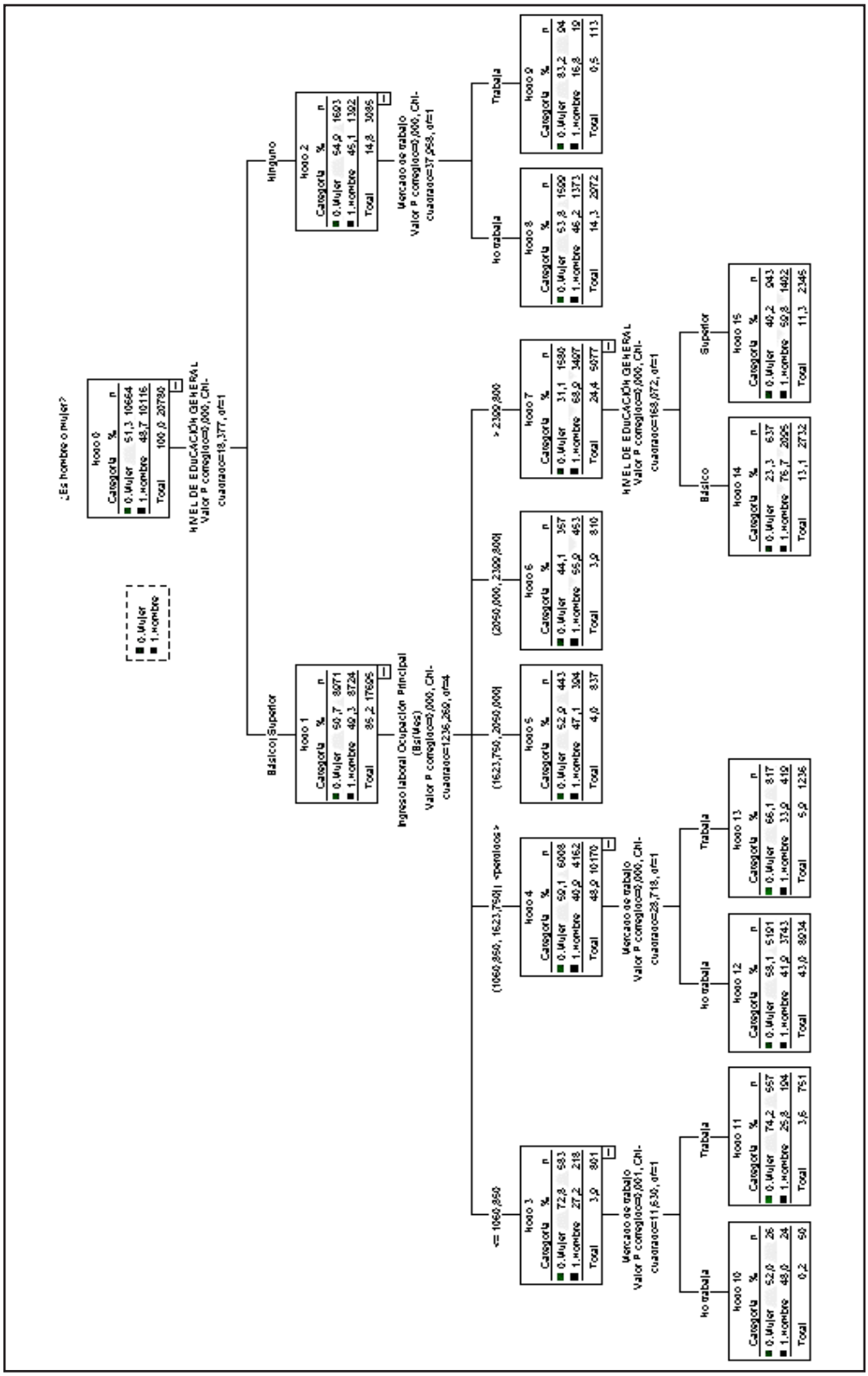
Sexo (masculino - femenino)

Variables independientes:

Nivel educativo(ninguno, básico y superior)
Si la persona trabaja o no
Ingreso mensual en Bs.

La información estadística corresponde a la encuesta de hogares del año 2018 de personas que habitan las ciudades capitales y El Alto. La muestra alcanza un total de 20.780 registros disponibles en la base de datos inferida a la población total. El resultado del árbol de clasificación se presenta en la Figura 4.

Figura 4.
Árbol de clasificación por sexo



Fuente: Elaboración propia - Programa SPSS

Haciendo un análisis del árbol de clasificación de la Figura 4, se tiene que: el nodo raíz, indica que el 51,3% de la población es mujer y el 48,7% es hombre con un total de 20.780 registros. La población se divide en dos grupos por el nivel educativo y se puede ver que el 50,7% de mujeres y 49,3% de hombres tienen un nivel educativo entre básico y superior. Por otro lado, se observa que el 54,9% de mujeres y el 45,1% de hombres no tienen nivel educativo, sin embargo, esta última categoría, es de solo el 14,8% en el global. Pasando a la tercera fila del árbol de clasificación, de los que tienen un nivel educativo básico y superior y cuentan con un ingreso mensual en bolivianos menor a 1.061 Bs, el 72,8% son mujeres y el 27,2% son hombres, pero apenas representan el 3,9% del total de la población. De los que tienen ingreso entre 1.061 y 1.624 Bs., el 59,1% son mujeres y el 40,9% son hombres de un total de 48,9% poblacional. Asimismo, de los que tienen ingreso mensual entre 1.624 y 2.050 Bs., el 52,9% son mujeres y el 47,1% son hombres de un total poblacional de apenas 4%.

El nodo 6 de la misma fila indica que de los que tienen un nivel educativo básico y superior y tienen ingreso mensual en bolivianos entre 2.050 y 2.400 Bs., aproximadamente, el 44,1% son mujeres y el 55,9% son hombres, representando solo el 3,9% del total de la población. De los que tienen ingresos por encima de los 2.400 Bs. y que representa el 24,4% de la población, el 31,1% son mujeres y el 68,9% son hombres.

Si se observa la parte derecha del árbol de clasificación en el nodo 8 con el 14,3% del total de la población de los que no trabajan, el 53,8% son mujeres y el resto hombres.

El nodo 12 con un 43% de la población, indica que de entre los que tienen un ingreso entre 1.061 y 1.624 Bs. mensual, no trabajan y de ellos el 58,1% son mujeres y el 41,9% son hombres. El nodo 14 describe, con una población del 13,1% son los que tienen un ingreso mensual por encima de los 2.400 Bs. y tienen sorpresivamente un nivel educativo básico, de ellos, el 23,3% son mujeres y el 76,7% son hombres. Lo mismo pasa en el nodo 15, que con una población de 11,3% son los que tienen un ingreso mensual por encima de los 2.400 Bs. y cuentan con un nivel educativo superior, de ellos, el 40,2% son mujeres y el 59,8% son hombres.

4. Discusión

Es interesante contar con un método de clasificación como el de árboles de decisión multivariante, que incluye en el análisis, muchas observaciones o datos masivos, y que va entrenando el algoritmo hasta lograr una confiable homogeneidad de observaciones en diferentes grupos excluyentes, dando lugar a una interpretación de los datos y análisis de forma sencilla. Si se compara con el método de análisis multivariante de conglomerados no jerárquico, siempre existirá en la clasificación, observaciones que pueden pertenecer a otros grupos o conglomerados de acuerdo al manejo de distancia entre objetos u observaciones. Luego, de acuerdo a lo discutido, será muy importante realizar un análisis y comparación de clasificación de observaciones mediante otras técnicas multivariantes exploratorias con el que arroja el método de árbol de decisión.

5. Conclusión

Existen otros métodos de clasificación de observaciones multivariantes que emplea la ciencia de los datos dentro del *Machine Learning* o la Minería de Datos, sin embargo, incluido este último método como es el de árbol de clasificación, no siempre son perfectos y deben ser utilizados de acuerdo a la información disponible, además, requieren siempre del cuidado de la inclusión de la información masiva que por manejar demasiada información, puede contener datos incoherentes o no válidos que

distorsionarían el análisis. Lo mismo ocurriría con la inclusión en demasía de variables y datos que pueden ser atípicos y/o faltantes.

En cuanto a los resultados obtenidos del árbol de decisión, se observa que el género mujer está en desventaja en relación a los hombres en cuanto a su nivel de formación educativa influyente en la ocupación y el ingreso.

Referencias Bibliográficas

Bouza, C., Santiago A. *Classification and Regresssion Trees.* Universidad de La Habana, Cuba y Universidad Autónoma de Guerrero, México.

Breiman, L., Friedman, J.H., Olsen, R. A. & Stone, C. J. (1984). *Classification and Regresssion Trees.* Wadsworth, N. York.

Cox, D. R., Snell, E. J. (1989). *The Analysis of Binary Data*. Chapman and Hall. London.

De la Fuente, F, S. (2011), *Análisis de conglomerados.* Universidad Autónoma de Madrid (UAM).

Diaz, L. (2012). *Análisis Estadístico de Datos Multivariantes.* Universidad Nacional de Colombia.

Francois, H. (2013). *Análisis de Datos con R*. *Colombia.* Escuela Colombiana de Ingeniería.

Hair, J, F., Anderson, R, E., Tatham, R, L. & Black, W,C. (1999). *Análisis Multivariante*. Madrid, España. Editorial Prentice Hall.

Hosmer, D. W. & Lemeshow, S. (1989). *Applied Logistic Regression.* Wiley Interscience. New York.

Aldás, J. & Uriel, E. (2017). *Análisis Multivatiante Aplicado con R.* Alfacentauro.

Johnson, R. A. *Applied Multivariate Statistical Analysis.* Prentice Hall. New Jersey.

López, A. (2012). *Análisis Multivariante para la Inteligencia de Mercados.* Tecnológico de Monterrey.

Ruiz, M. (2013). *Árboles de Decisión y ELECTRA I*. Biblioteca Universitaria.

Mitchell, T. M. (1997). *Machine Learning. Texto: McGraw-Hill, Cap. III. Web: http://www.cs.cmu.edu/~tom/mlbook.html*

Muller, A. & Guido, S. (2016). *Introduction Machine Learning with Python*. Jupyter

Loyola, O. (2012). *Inducción de Árboles de Decisión*. Academia Española.

Theobald, O. (2018). *Machine Learning for Absolute Beginners.* Andriy Burkov

Peña, D. (2002). *Análisis de Datos Multivariantes*. McGraw Hill.

Pérez, C. (2004). *Técnicas de Análisis Multivariante de Datos.* Pearson Educación S.A., España.

Poole, D. (1998). *Al Computational Intelligence. A Logical Approach.* Oxford University Press.

Uriel, E. & Aldás, J. (2005). *Análisis Multivariante Aplicado*. Madrid, España. Thomson.