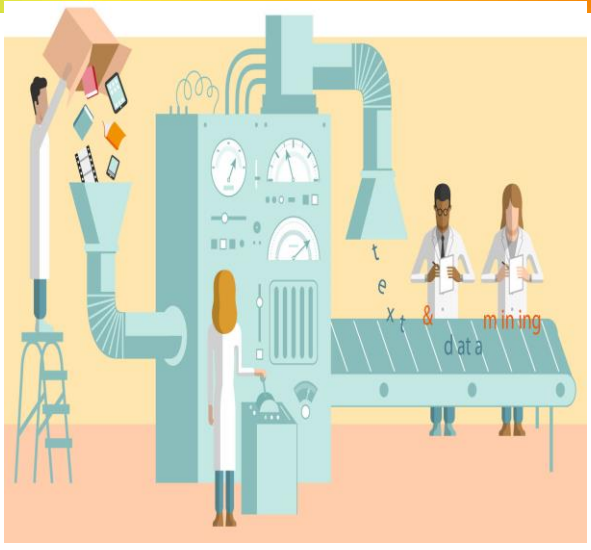
IPN-ESCOM

Presentación del curso de *Data Mining* (Minería de datos)

Profesora:
Dra. Fabiola Ocampo Botello

2023-2. Grupo: 3CV15

Fuente de la imagen:
https://www.inteldig.com/2018/12/las-mejores-herramientas-mineria-criptomonedas-linux/



Fuente de la imagen: Rudolf Mumenthaler
https://ruedimumenthaler.ch/2015/06/09/trend-und-herausforderung-text-and-data-mining/

¿Qué es la minería de datos?

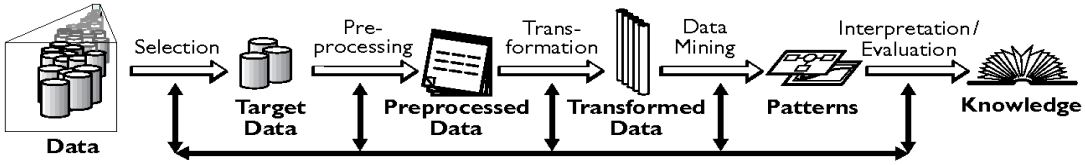
La minería de datos es un campo multidisciplinario que integra trabajo de diversas áreas como tecnología de bases de datos, aprendizaje automático (machine learning), estadística, reconocimiento de patrones, recuperación de información, redes neuronales, sistemas basados en conocimiento, inteligencia artificial, cómputo de alto rendimiento y visualización de datos (Sahu, Shrma, & Gondhalakar, 2011).

Fuente de la imagen:
https://www.duperrin.com/english/wp-content/uploads/2017/06/data-driven.jpg

La minería de datos es un proceso para extraer información y conocimiento implícitos, potencialmente útiles y que son desconocidos por las personas, los cuales se encuentran en datos masivos, incompletos, difusos y aleatorios (Sahu, Shrma, & Gondhalakar, 2011).

La minería de datos ha sido comúnmente definida como encontrar información en una base de datos, ha sido llamada análisis de datos exploratorio, descubrimiento conducido por datos y aprendizaje deductivo (Dunham, M. H., 2002).

Extracción de conocimiento en bases de datos (*Knowledge Discovery in Databases, KDD*)



Fuente de la imagen: Exponentis
http://exponentis.es/el-proceso-kdd-para-extraer-conocimiento-util-de-volumenes-de-datos

KDD es el proceso de extraer información y la minería de datos es el uso de algoritmos para extraer esa información y forma parte del KDD.

Es el proceso no trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información. No es un proceso automático, es un proceso iterativo que, exhaustivamente explora volúmenes muy grandes de datos para determinar relaciones (U Fayyad et al 1996, citado en Joyanes, 2019:232).

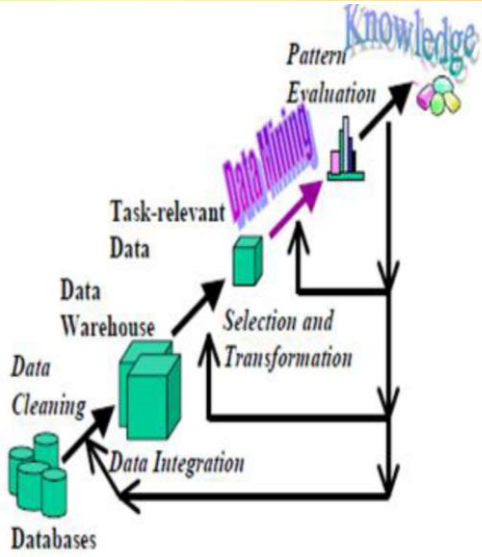


Imagen tomada de Sahu, Shrma, & Gondhalakar (2011).

El proceso iterativo consiste de los siguientes pasos:

- **Limpieza de datos**. Se eliminan los datos con ruido o sucios y los datos irrelevantes que se encuentran en la colección de datos.
- **Integración de datos**. Se integran múltiples fuentes de datos comúnmente heterogéneas en una fuente de datos en común.
- **Selección de datos**. Se seleccionan de la fuente de datos, los datos relevantes para el análisis.
- **Transformación de datos**. También conocida como consolidación de los datos. Los datos seleccionados se transforman a formas apropiadas para realizar el proceso de la minería.
- **Minería de datos**. Es el paso crucial en el cual se aplican técnicas inteligentes para la extracción de patrones de relación potencialmente útiles.
- **Evaluación de los patrones**. Se identifican patrones estrictamente interesantes que representen conocimiento sustentados en las medidas proporcionadas.
- **Representación del conocimiento**. El conocimiento descubierto es presentado al usuario de manera visual, las técnicas de representación visual permite a los usuarios interpretar y entender los resultados de la minería de datos.

Ejemplo de problemas de Minería de datos

Las compañías de tarjetas de crédito deben determinar si autorizan las compras que se realizan con tarjeta de crédito. Supongamos que con base en información histórica sobre compras, cada compra se coloca en una de cuatro clases:
(1) autorizar,
(2) solicitar una identificación adicional antes de la autorización,
(3) no autorizar, y
(4) no autorizar y contactar a la policía.

Fuente de la imagen: https://www.profe-de-espanol.de/2015/06/23/de-compras-consumo-y-costumbres/

Las funciones de minería de datos en este ejemplo son necesarias, debido dos aspectos: Primero se deben examinar los datos históricos para determinar cómo los datos encajan en las cuatro clases.
Posteriormente el problema es aplicar este modelo a cada nueva compra.
Si bien la segunda parte puede expresarse como una simple consulta de base de datos, la primera parte no puede ser (Ejemplo adaptado de Dunham, M. H., 2002).

Una determinada cadena de tiendas departamentales crea catálogos especiales dirigidos a varios grupos demográficos basados en atributos como ingresos, ubicación y características físicas de los clientes potenciales (edad, altura, peso, etc.). Para determinar los correos de destino de los diversos catálogos y ayudar en la creación de catálogos nuevos y más específicos, la empresa realiza una agrupación de clientes potenciales en función de los valores de atributo determinados.

Fuente de la imagen: Running Life
http://www.runninglife.com.mx/2018/03/27/asics-mexico-estrena-nueva-tienda-en-cdmx-conocela/

Los resultados del ejercicio de agrupación en clústeres son luego utilizados por la gerencia para crear catálogos especiales y distribuirlos a la población objetivo correcta según el clúster de ese catálogo. (Agrupamiento). (Ejemplo adaptado de Dunham, M. H., 2002).

Fuente de la imagen:
https://agendainmobiliariatv.blogspot.com/2015/12/3-consejos-para-usar-la-banca-online.html

Un banco por Internet desea obtener reglas para predecir qué personas de las que solicitan un crédito no lo devuelven. La entidad bancaria cuenta con los datos correspondientes a los créditos concedidos con anterioridad a sus clientes (cuantía del crédito, duración en años...) y otros datos personales como el salario del cliente, si posee casa propia, etc. (Ejemplo tomado de Hernández, Ramírez y Ferri, 2004).

A partir de éstos, las técnicas de minería de datos podrían sintetizar algunas reglas, como por ejemplo:

SI Cuentas_Morosas > 0
ENTONCES Devuelve_Credito = NO
SI Cuentas_Morosas = 0 **Y** ((Salario > 2,500) **O** (D_Credito > 10))
ENTONCES Devuelve_Credito = SI

Un profesor desea alcanzar cierto nivel de ahorros antes de su jubilación. Periódicamente, él predice cuáles serán sus ahorros según su valor actual y varios valores del pasado, por lo que utiliza una fórmula de regresión lineal simple para predecir este valor ajustando el comportamiento pasado a una función lineal y entonces usa esta función para predecir los valores de esos puntos en el futuro.

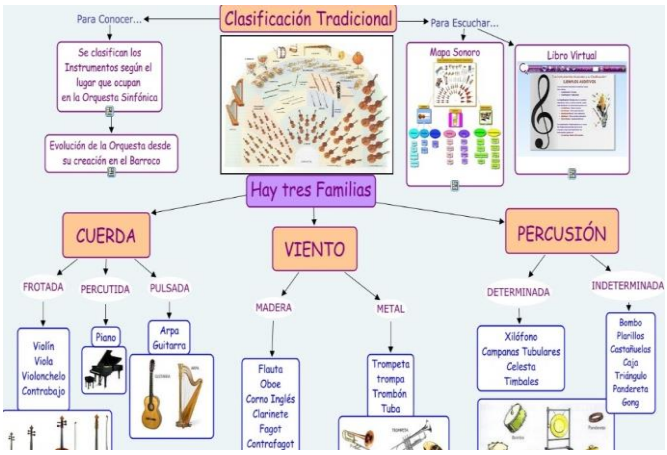
Con base en estos valores, el profesor puede tomar decisiones respecto a sus inversiones. Regresión lineal. (Ejemplo adaptado de Dunham, M. H., 2002).

Fuente de la imagen: El rincón del emprendedro
https://rincondelemprendedor.es/diferencias-entre-las-inversiones-rentables-e-inversiones-seguras/

Tareas y Técnicas de minería de datos

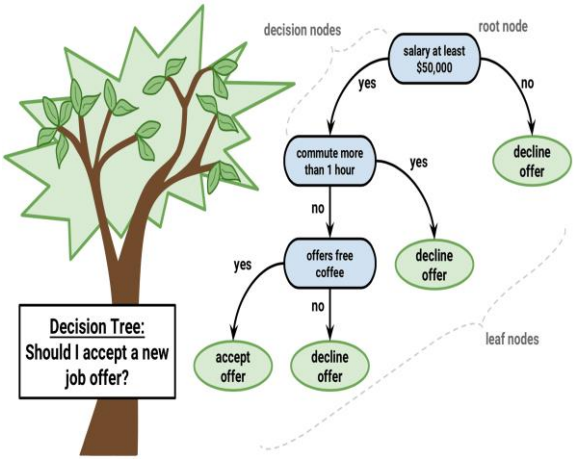
Clasificación

Han, Kamber & Pei (2012) establecen que la clasificación es una forma de analizar datos para generar modelos que describen importantes clases de datos. Estos modelos se llaman clasificadores, permiten predecir etiquetas de clases categóricas (discretas, desordenadas).



Fuente de la imagen: Nuestra cita musical
https://nuestracitamusical.blogspot.com/2017/02/clasificacion-tradicional-de-los.html

Árboles de decisión



Fuente de la imagen
https://eduladder.com/viewquestions/25662/Explain-Decision-Tree-algorithm-in-detail

Han, Kamber & Pei (2012) establecen que Un árbol de decisión es una estructura de árbol similar a un diagrama de flujo, donde cada nodo interno (nodo no hoja) denota una prueba en un atributo, cada rama representa un resultado de la prueba y cada nodo hoja (o nodo terminal) tiene una etiqueta de clase.

Reglas de asociación

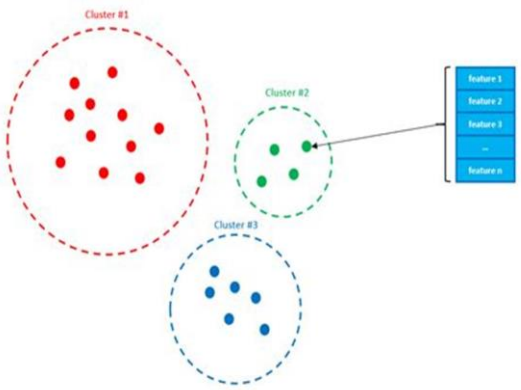
Larose & Larose (2015) establecen que El análisis de afinidad se refiere al estudio o características que "van juntas". Los métodos para el análisis de afinidad son conocidos como "análisis de la canasta de mercado", la cual busca descubrir asociaciones entre los atributos, es decir, busca descubrir reglas para cuantificar la relación entre dos o más atributos.
Las reglas de asociación son de la forma:

Si el *antecedente* entonces el *consecuente*.

Considerando una medida de soporte (*support*) y una medida de confianza (*confidence*).

Fuente de la imagen: Sin referencia

Agrupamiento



Fuente de la imagen: https://datascience.stackexchange.com/questions/33937/how-to-solve-online-clustering-problem

El agrupamiento (*Clustering*) es similar a la clasificación en que los datos están agrupados. Sin embargo, a diferencia de la clasificación, los grupos no están predefinidos. En cambio, el agrupamiento se logra al encontrar similitudes entre los datos de acuerdo con las características encontradas en los datos reales. (Dunham, 2002).

TABLE 5.1: Sample Data for Example 5.1

Income	Age	Children	Marital Status	Education
$25,000	35	3	Single	High school
$15,000	25	1	Married	High school
$20,000	40	0	Single	High school
$30,000	20	0	Divorced	High school
$20,000	25	3	Divorced	College
$70,000	60	0	Married	College
$90,000	30	0	Married	Graduate school
$200,000	45	5	Married	Graduate school
$100,000	50	2	Divorced	College

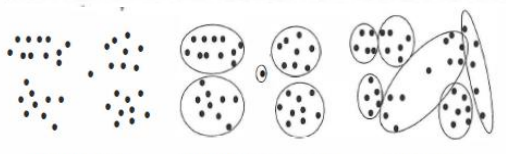


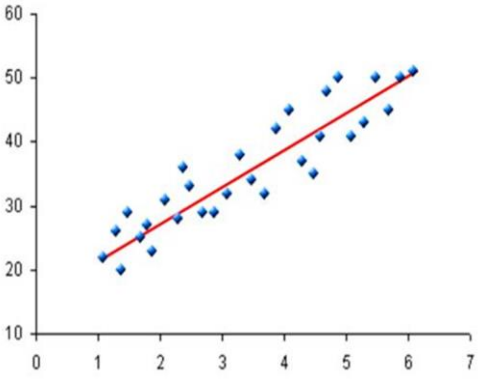
FIGURE 5.1: Different clustering attributes.

<u>Ejemplo y figuras tomados de Dunham (2002:125-126):</u>

Una empresa internacional de catálogos en línea desea agrupar a sus clientes en función de características comunes. La dirección de la empresa no tiene etiquetas predefinidas para estos grupos. Según el resultado de la agrupación, dirigirán campañas de marketing y publicidad a los diferentes grupos. La información que tienen sobre los clientes incluye ingresos, edad, número de hijos, estado civil y educación.

Regresión lineal

Carollo (2012) establece que "El objetivo de un modelo de regresión es tratar de explicar la relación que existe entre una variable dependiente (variable respuesta) Y y un conjunto de variables independientes (variables explicativas) X1,..., Xn. En un modelo de regresión lineal simple tratamos de explicar la relación que existe entre la variable respuesta Y y una única variable explicativa X."



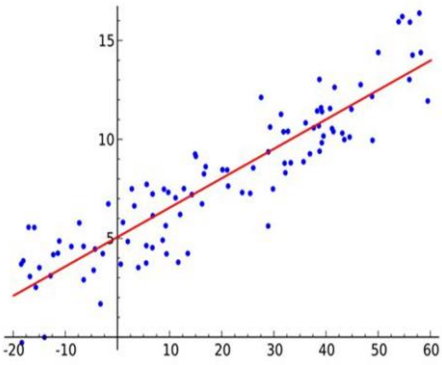
Fuente de la imagen: aprendiendocalidadyadr.com
https://aprendiendocalidadyadr.com/diagrama-de-dispersion/

Se tienen los datos de 10 pizzerías (Pizzerías "Polito") ubicadas cerca de los campus universitarios. Tanto la cantidad de alumnos y las ganancias se expresan en miles, como se muestra en la siguiente tabla.
Ejemplo de regresión lineal adaptado de Anderson, Sweeney & Williams (2008).

Fuente de la imagen: La Cocinika de Ana
https://ana-lacocinikadeana.blogspot.com/2012/10/dominos-pizza.html

No	NoEstud x	Ventas y
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Ejemplo de Regresión lineal



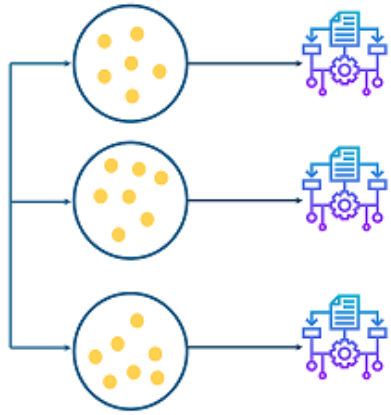
Fuente de la imagen: Wikipedia
https://en.wikipedia.org/wiki/Linear_regression

El vicepresidente de una compañía química y de fabricación de fibras cree que las ganancias anuales de la empresa dependen de la cantidad gastada en investigación y Desarrollo (ID), pero el nuevo presidente de la compañía no está de acuerdo, por lo que ha solicitado una ecuación para pronosticar los beneficios anuales derivados de la cantidad presupuestada para ID.

Ejemplo tomado de Levín et al (2004:510).

Bagging

El término bagging deriva del mecanismo denominado *bootstrap aggregation*, mecanismo que genera subconjuntos de entrenamiento seleccionando aleatoriamente y con reemplazamiento. Dado que hay un conjunto de clasificadores, la predicción de nuevos ejemplos se efectúa por votación mayoritaria. (Hernández, Ramírez y Ferri, 2004).

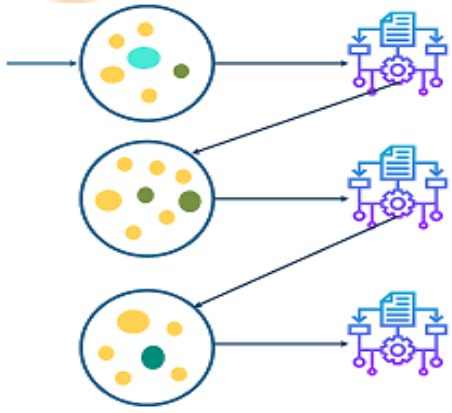


Fuente de la imagen: Machine Learning para todos
https://machinelearningparatodos.com/cual-es-la-diferencia-entre-los-metodos-de-bagging-y-los-de-boosting/

Boosting

Hernández, Ramírez y Ferri (2004) establecen que:

- La estrategia de *boosting* construye los nuevos modelos tratando de corregir los errores cometidos previamente.
- Existen muchas variantes del algoritmo básico de *boosting*, siendo probablemente AdaBoost una de las versiones originales y todavía más populares.
- A diferencia del algoritmo *Bagging*, este algoritmo no siempre realiza las k iteraciones requeridas por el usuario, dado que considera un criterio de parada de acuerdo con el error e.

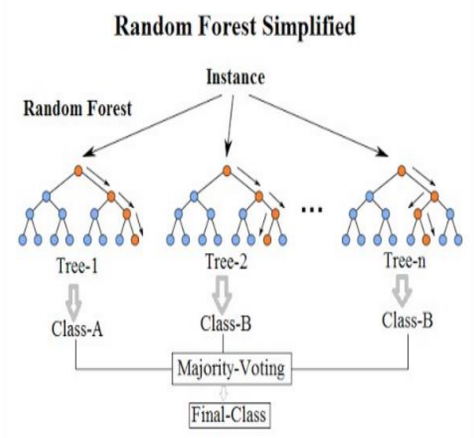


https://machinelearningparatodos.com/cual-es-la-diferencia-entre-los-metodos-de-bagging-y-los-de-boosting/

Random Forest

Imagine que cada uno de los clasificadores del conjunto es un clasificador de árbol de decisión, de modo que la colección de clasificadores es un "bosque". Los árboles de decisión individuales se generan utilizando una selección aleatoria de atributos en cada nodo para determinar la división (Han, Kamber, & Pei, 2012:383).

Un conjunto de bosque aleatorio utiliza una gran cantidad de árboles de decisión individuales sin podar que se crean aleatorizando la división en cada nodo del árbol de decisión (Rokach, L. & Maimon, O., 2015).



Fuente de la imagen: Wikipedia
https://en.wikipedia.org/wiki/Random_forest

Herramientas de minería de datos

Joyanes (2019:250-255) presenta un conjunto de herramientas para la minería de datos, las cuales clasifica como herramientas de código abierto y herramientas comerciales propietarias.

Herramientas de código abierto	Herramientas comerciales propietarias
• WEKA. • KNIME. • RapidMiner. • KEEL (*Knowledge Extraction for Evolution*). • Orange. • Lenguaje R. • NLTK.	• IBM SPSS Modeler (Clementine). • SAS Enterprise Miner. • Oracle Data Mining. • SAP Business Object. • Microsoft SQL Server Data Mining.

Tipos de Actividades a desarrollar en el curso

Tareas
Prácticas
Lecturas
Exámenes
Exposiciones
Ejercicios de clase
Proyectos parciales
Proyecto final

Individual
En equipo
Grupal

Referencias Bibliográficas

Anderson, Sweeney & Williams. (2008). Estadística para administración y economía, 10ª edición. Cengage Learning.

Bennet, Briggs & Triola (2011). Razonamiento estadístico. Pearson. México.

Carollo Limeres, M. Carmen. (2012). Regresión lineal simple. Apuntes del departamento de estadística e investigación operativa . Disponible en: http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140116_Regr_%20simple_2011_12.pdf

Dunham, M. H. (2002). Data mining: introductory and advanced topics. Prentice Hall.

Han, Jiawei; Kamber, Micheline & Pei, Jian. (2012). *Data Mining: concepts and techniques*. Third edition. Morgan Kaufman Series.

Hernández Orallo, José; Ramírez Quintana, Mª José y Ferri Ramírez, César. (2004). Introducción a la Minería de datos. Editorial Pearson.

Joyanes Aguilar, Luis. (2019). Inteligencia de negocios y analítica de datos. Una visión global de Business inteligence& Analytics. Alfaomega.

Larose, T. Daniel & Larose, D. Chantal. (2015). Data Mining and Predictive Analytics. Second Edition. Wiley.

Levín, Rubín, Balderas, Del Valle y Gómez. (2004). Estadística para administración y economía. Séptima Edición. Prentice-Hall.

Rokach, L. & Maimon, O. (2015). Data Mining with decision trees. Theory and Applications. Second Edition. World Scientific Publishing Co. Pte. Ltd.

Sahu, Hemlata; Shrma, Shalini; Gondhalakar, Seema. (2011). A Brief Overview on Data Mining Survey. International Journal of Computer Technology and Electronics Engineering (IJCTEE). Vol.1.

Bibliografía sugerida para el curso

Anderson, Sweeney & Williams. (2008). Estadística para administración y economía, 10ª edición. Cengage Learning.

Bennet, Briggs & Triola (2011). Razonamiento estadístico. Pearson. México.

Tan, Pang-Ning, Steinbach, Michael & Kumar, Vipin. (2014). Introduction to data mining. Pearson.

Dunham, M. H. (2002). Data mining: introductory and advanced topics. Prentice Hall.

Han, Jiawei; Kamber, Micheline & Pei, Jian. (2012). *Data Mining: concepts and techniques*. Third edition. Morgan Kaufman Series.

Hernández Orallo, José; Ramírez Quintana, Mª José y Ferri Ramírez, César. (2004). Introducción a la Minería de datos. Editorial Pearson.

Joyanes Aguilar, Luis. (2019). Inteligencia de negocios y analítica de datos. Una visión global de Business inteligence& Analytics. Alfaomega.

Larose, T. Daniel & Larose, D. Chantal. (2015). Data Mining and Predictive Analytics. Second Edition. Wiley.

Levín, Rubín, Balderas, Del Valle y Gómez. (2004). Estadística para administración y economía. Séptima Edición. Prentice-Hall.

Maimon, O. & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook. Second Edition. Springer.

Rokach, L. & Maimon, O. (2015). Data Mining with decision trees. Theory and Applications. Second Edition. World Scientific Publishing Co. Pte. Ltd.

Sahu, Hemlata; Shrma, Shalini; Gondhalakar, Seema. (2011). A Brief Overview on Data Mining Survey. International Journal of Computer Technology and Electronics Engineering (IJCTEE). Vol.1.

Tan, Pang-Ning, Steinbach, Michael & Kumar, Vipin. (2014). Introduction to data mining. Pearson.

Fuente de la imagen: La casa de Raul
https://claseraul.es/bienvenidos/

Bienvenidos al curso de Data Mining (Minería de datos)