Co-funded by the European Union

Specialized and updated training on supporting advance technologies for early childhood education and care professionals and graduates

MÓDULO IV.1

Técnicas de Observación y Evaluación a partir de Recursos Inteligentes: Introducción a la Minería de Datos

Docentes

Dr. Álvar Arnaiz González
Dr. José Francisco Díez Pastor
Dra. Sandra Rodríguez Arribas
Departamento de Ingeniería Informática
Universidad de Burgos

e-EarlyCare-T

“Specialized and updated training on supporting advance technologies for early childhood education and care professionals and graduates”, e-EarlyCare-T, reference 2021-1-ES01-KA220-SCH-000032661, is co-financed by the European Union's Erasmus+ programme, line KA220 Strategic Partnerships Scholar associations. The content of the publication is the sole responsibility of the authors. Neither the European Commission nor the Spanish Service for the Internationalization of Education (SEPIE) is responsible for the use that may be made of the information disseminated herein.”

Índice de contenidos

I. Introducción

Vivimos en la sociedad de la información y la comunicación, la tecnología que empleamos en el siglo XXI lleva asociada la recopilación y almacenamiento de grandes cantidades de datos. La **Minería de Datos** o ***Data Mining*** (**DM**) permite encontrar información contenida en los datos que no siempre resulta aparente, ya que, dado el gigantesco volumen de datos existentes, gran parte de ese volumen nunca será analizado.

II. Objetivos

1. Conocer conceptos clave relacionados con la **Minería de Datos**

2. Conocer y aplicar técnicas sencillas de **Minería de Datos** en el campo de la atención temprana.

III. Contenidos específicos del tema

3.1. Minería de Datos.

La **minería de datos** también conocida como ***Data Mining*** **(DM)** en inglés, es el proceso de búsqueda y análisis en grandes bases de datos para encontrar información útil que sirva para la toma de decisiones.

Existen numerosas técnicas de **DM** que emplean el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque el volumen de datos que hay que analizar es demasiado grande.

En la actualidad en campo de la **minería de datos** se emplea continuamente para el análisis de grandes cantidades de datos en diversos campos de conocimiento como la educación, la economía, los negocios, el medio ambiente...

3.1.1 Conceptos básicos en Minería de Datos

Antes de conocer el proceso que se realiza y los tipos de algoritmos que se utilizan en el **DM** es importante aclarar algunos conceptos básicos que aparecen con frecuencia en la bibliografía asociados a la **Minería de Datos**.

Data set o Conjunto de datos

Es una colección grande de datos generalmente organizados en filas y columnas que contienen variables y atributos. Cada uno de estos valores se conoce con el nombre de dato. El conjunto de datos también puede consistir en una colección de documentos o de archivos.

Clases o etiquetas

En el campo de la **minería de datos**, una clase es el atributo discreto cuyo valor se desea predecir en función de los valores de otros atributos. También se conoce como etiqueta.

Instancia

Una instancia es cada uno de los datos de los que se disponen para hacer un análisis. Cada instancia, a su vez, está compuesta de características que la describen. Por ejemplo, en una hoja de cálculo, las instancias serían las filas y las características la información almacenada en las columnas.

Algoritmo

En informática un algoritmo es un conjunto de instrucciones definidas, ordenadas y acotadas para resolver un problema, realizar un cálculo o desarrollar una tarea. En otras palabras, es un procedimiento paso a paso para obtener un resultado.

3.1.2 Proceso de Aplicación de técnicas de Minería de Datos

El proceso consta de cuatro fases principales que se enumeran a continuación:

1. **Definición del problema**: es la primera fase en la que se traduce un problema específico en un problema de **minería de datos** en el que se plantean los objetivos del análisis y las preguntas de investigación.
2. **Preparación y recopilación de datos**: es la fase más extensa del proceso ya que la calidad de los datos es uno de los retos más importantes en la **minería de datos**. Los datos brutos deben ser identificados, limpiados y almacenados en un formato preestablecido.
3. **Modelado y evaluación**: en este paso se seleccionan y aplican diferentes técnicas de modelado de datos (algoritmos) y después se establecen los parámetros y valores óptimos de dichas técnicas.
4. **Despliegue**: es la última fase en la que se organizan y presentan los resultados de la **minería de datos** mediante gráficos e informes.

Ver Figura 1.

Figura 1. Proceso de aplicación de técnicas de **Minería de datos**. Fuente: Elaboración propia

Es importante señalar que todo proceso de **minería de datos** es un proceso iterativo, lo que significa que el proceso no se detiene cuando se despliega una solución concreta. Puede ser sólo una nueva entrada para otro proceso de **minería de datos** (Rodríguez-Arribas, 2021). Es decir, en numerosas ocasiones la aplicación de técnicas de **DM** requiere de varias iteraciones y del empleo de algoritmos diferentes para poder extraer los resultados finales de la investigación que estamos realizando.

3.2. Tipos de Aprendizaje en Minería de Datos

Existen numerosas clasificaciones de los algoritmos que se emplean en el mundo de la **Minería de Datos**, pero es fundamental entender que hay dos enfoques básicos: el aprendizaje supervisado y el aprendizaje no supervisado. La principal diferencia es que en el aprendizaje supervisado existe una clase que se emplea para obtener una función que permite asociar nuevos datos con la clase correspondiente. Sin embargo, en el aprendizaje no supervisado no existe ninguna clase, en este caso los algoritmos tratan de descubrir patrones ocultos en los datos sin intervención humana en forma de etiquetas asociadas a los datos (Chapelle, Schölkopf y Zien, 2006).

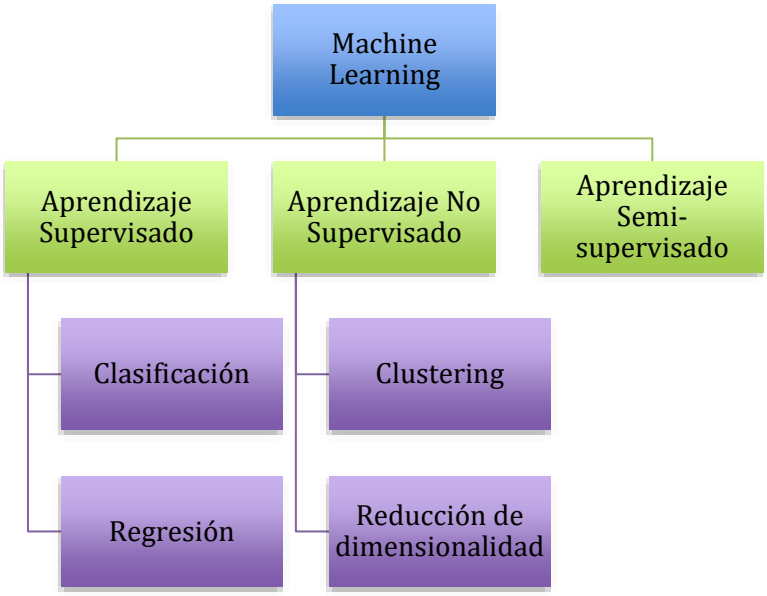

Figura 2. Métodos de **Minería de datos**. Fuente: elaboración propia

Cuando tenemos que decidir qué algoritmo se empleará para realizar el análisis de los datos es importante tener en cuenta que tipo de aprendizaje se está utilizando, es decir, si se está hablando de aprendizaje supervisado o no supervisado (García, Luengo y Herrera, 2015). De acuerdo con el tipo de aprendizaje utilizado se emplearán diferentes técnicas y algoritmos como puede observarse en la Figura 2.

3.2.1 Aprendizaje Supervisado

Una de las modalidades de aprendizaje del ***Machine Learning***, como se ha comentado anteriormente, es la de aprendizaje supervisado.

El objetivo fundamental del aprendizaje supervisado es la creación de un modelo que sea capaz de predecir valores correspondientes a objetos de entrada después de haberse familiarizado con una serie de ejemplos, los datos de entrenamiento.

Esta técnica consta de dos pasos fundamentales:

1. Una fase de entrenamiento donde se utiliza un conjunto de datos etiquetados, que contienen los datos de entrada y los resultados deseados para esos datos de entrenamiento con un algoritmo que permita deducir una función a partir de los datos que le estamos proporcionando al algoritmo.

2. La fase de prueba, en donde se utiliza la función obtenida en el paso anterior para generar nuevas predicciones con nuevos conjuntos de datos (ver Figura 3).

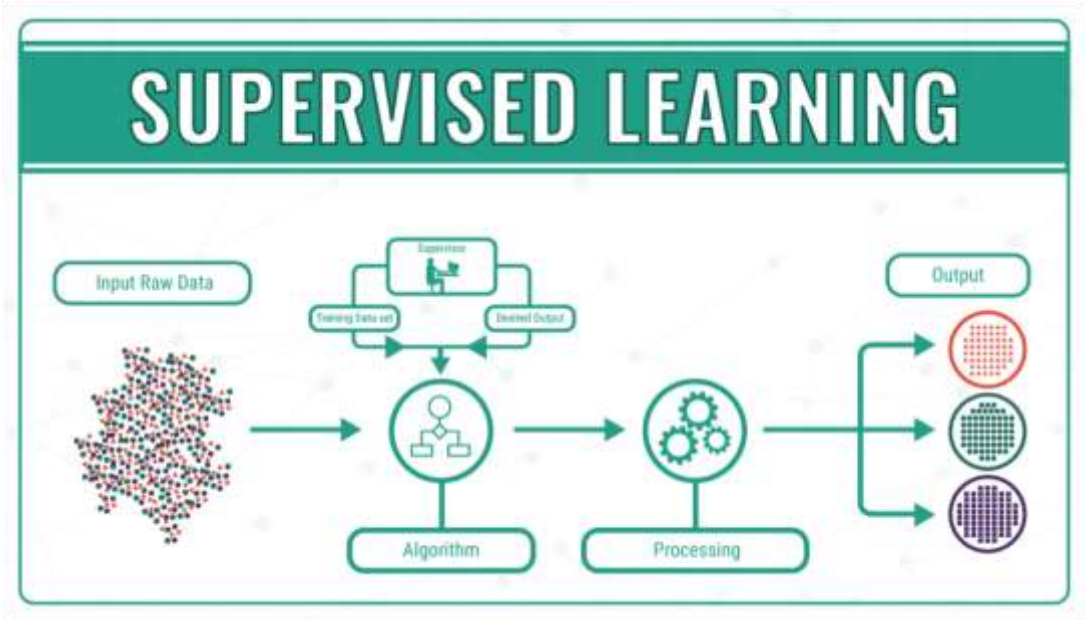

Figura 3. Proceso de funcionamiento del aprendizaje supervisado. Fuente: ExperiencIA Oracle.

El proceso es conocido como aprendizaje supervisado, pues al conocer las respuestas de cada ejemplo del conjunto de entrenamiento, es posible corregir la función generada por el algoritmo. Se supervisa el entrenamiento del algoritmo mediante la corrección de parámetros del mismo, según sean los resultados que se obtienen de forma iterativa.

3.2.2 Aprendizaje No Supervisado

Este tipo de aprendizaje es el otro enfoque básico del ***Machine Learning*** (**ML**). El aprendizaje no supervisado tiene datos sin etiquetar que el algoritmo tiene que intentar entender por sí mismo.

El objetivo de este tipo de aprendizaje es dejar que la máquina aprenda sin ayuda o indicaciones de los científicos de datos, es decir, sin supervisión y sin un conjunto de datos de entrenamiento. Además, la propia máquina realizará ajustes en los resultados y agrupaciones cuando haya resultados más adecuados, permitiendo que la máquina comprenda los datos y los procese de la mejor manera (ver Figura 4).

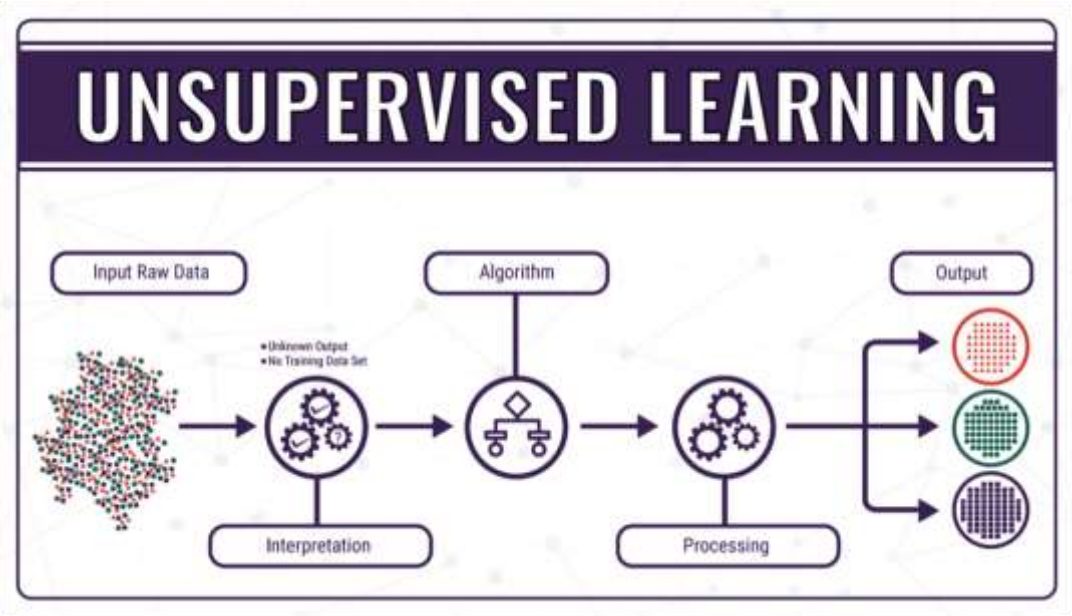

Figura 4. Proceso de funcionamiento del aprendizaje no supervisado. Fuente: ExperiencIA Oracle.

El aprendizaje no supervisado se utiliza para explorar datos desconocidos y sin etiquetar. Puede revelar patrones que podrían haberse pasado por alto o examinar grandes conjuntos de datos que serían demasiado para que los abordara una sola persona.

3.2.3 Aprendizaje Semi-Supervisado

Actualmente se están realizando numerosas investigaciones con métodos de aprendizaje semi-supervisado. Estas técnicas de aprendizaje automático utilizan datos de entrenamiento tanto etiquetados como no etiquetados: normalmente una pequeña

cantidad de datos etiquetados junto a una gran cantidad de datos no etiquetados (Zhu y Goldberg, 2009). Es decir, buscan mejorar los modelos de predicciones que se obtienen al utilizar exclusivamente datos etiquetados explorando la información estructural que contienen los datos no etiquetados.

Podemos decir el aprendizaje semi-supervisado trata de combinar los dos enfoques tradicionales de la **minería de datos** (aprendizaje supervisado y aprendizaje no supervisado) para quedarse con lo mejor de cada uno de ellos.

3.3. Algoritmos de Clasificación

Los algoritmos de clasificación son aquellos que utilizamos cuando el resultado esperado es una etiqueta discreta. Es decir, son útiles cuando la respuesta a la pregunta de investigación se encuentra dentro de un conjunto finito de resultados posibles.

Estos algoritmos trabajan generalmente sobre la información entregada por un conjunto de muestras, patrones, ejemplos o prototipos de entrenamiento que son tomados como representantes de las clases, y los mismos conservan una etiqueta de clase correcta. A este conjunto de prototipos correctamente etiquetados se les llama conjunto de entrenamiento, y es el conocimiento disponible para la clasificación de nuevas muestras. El objetivo de la clasificación supervisada es determinar, según lo que se tenga conocimiento, cual es la clase a la que debería concernir una nueva muestra, teniendo en cuenta la información que se pueda extraer (ver Figura 5).

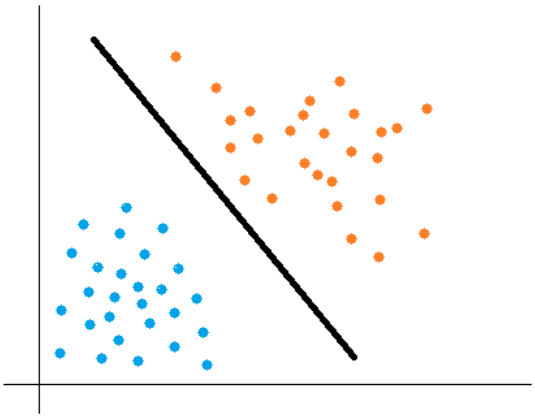

Figura 5. Algoritmo de clasificación. Fuente: elaboración propia

La clasificación es muy similar al proceso de aprendizaje de las personas, ya que poseemos la capacidad de clasificar alimentos, libros, animales, planetas, es decir, todo lo que nos rodea.

3.4. Algoritmos de *Clustering*

Los algoritmos de agrupamiento o de *clustering* se encargan de agrupar los objetos de un conjunto de datos en función de sus similitudes. De este modo los objetos que están dentro de un clúster o grupo tienen más similitudes entre ellos que diferencias (ver Figura 6).

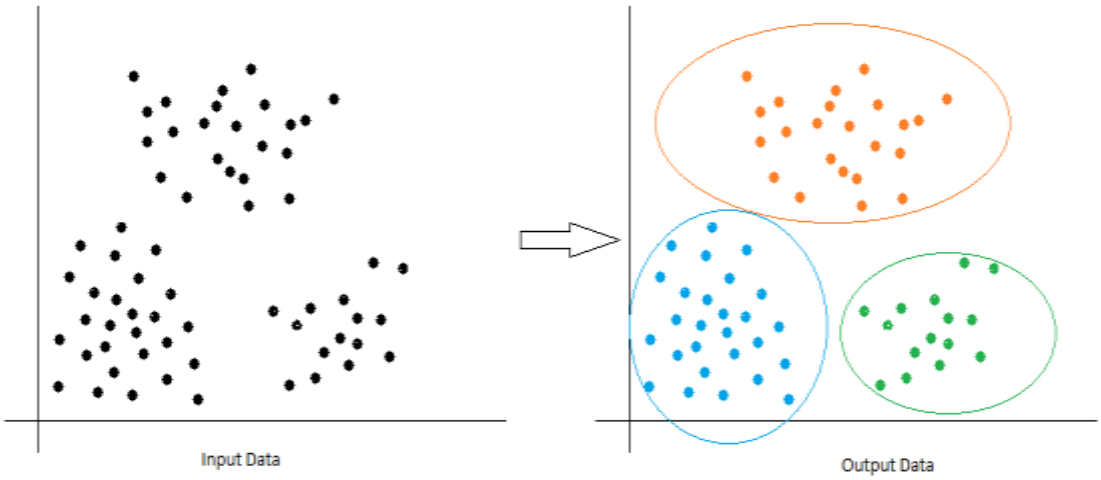

Figura 6. Algoritmo de *clustering*. Fuente: elaboración propia.

Estos algoritmos trabajan con datos no etiquetados por lo que es el propio algoritmo el que analiza los datos para encontrar el número de agrupamientos óptimo para el conjunto de datos de entrada ya que no disponemos de conocimientos previos sobre las características de los datos y sus clases.

Los agrupamientos que realizan los algoritmos pueden ser de dos tipos:

- **Clúster duro**: cada dato pertenece exclusivamente a un grupo
- **Clúster blando (difuso):** los datos pueden pertenecer a varios grupos en distintos grados, es decir, un mismo dato puede tener un grado de pertenencia del 60% al grupo 1 y del 40% en el grupo 2.

3.5. Algoritmos de Regresión

Los algoritmos de regresión es un subcampo del aprendizaje supervisado cuyo objetivo es establecer un método para la relación entre un cierto número de características y una variable objetivo continua.

Se trata de algoritmos que establecen una recta para proporcionar la tendencia de un conjunto de datos, es decir, el fin de estos algoritmos es relacionar un número de características y una variable objetivo continua (ver Figura 7).

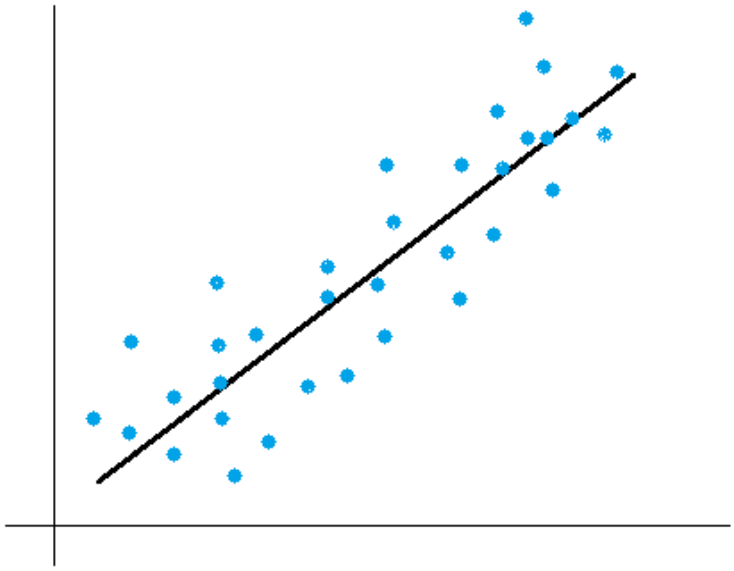

Figura 7. Algoritmo de regresión. Fuente: elaboración propia.

Esta técnica es útil para predecir resultados que son valores continuos, eso significa que la respuesta a la pregunta de investigación se presenta mediante una cantidad que puede determinarse de manera flexible en función de las entradas del modelo en lugar de limitarse a un conjunto de etiquetas finito como en el caso de la clasificación.

3.6. KNIME

KNIME es una aplicación de código abierto (*open source*) que permite aplicar a nuestros propios conjuntos de datos o a conjuntos de datos de ejemplo:

- Métodos estadísticos
- Algoritmos de **minería de datos** o aprendizaje automático.

- Técnicas de visualización.

Al tratarse de un software de código abierto tiene muchas ventajas, su código pertenece a la comunidad de usuarios y desarrolladores, lo que garantiza que siempre será una herramienta libre y gratuita. En contraposición el software privado pertenece en exclusiva a una empresa y esta empresa puede permitir su uso gratuito, pero también cobrar un precio elevado o exigir el pago de una subscripción mensual.

Se trata de una herramienta diseñada para ser sencilla de usar. El concepto más importante en el uso de la herramienta es el de *workflow* (en castellano, flujo de trabajo). Un flujo de trabajo es una secuencia de pasos configurada por el usuario. Formalmente es un conjunto de nodos unidos entre sí con flechas. Un nodo encapsula distintos trabajos que se pueden realizar con los datos, existen nodos para muchas tareas. Un flujo de trabajo podría tener un nodo para cargar un conjunto de datos a partir de un fichero de Excel, a continuación, un nodo para seleccionar atributos (columnas) de dicho conjunto de datos y por último otro nodo para visualizar estadísticas de los atributos seleccionados (ver Figura 8).

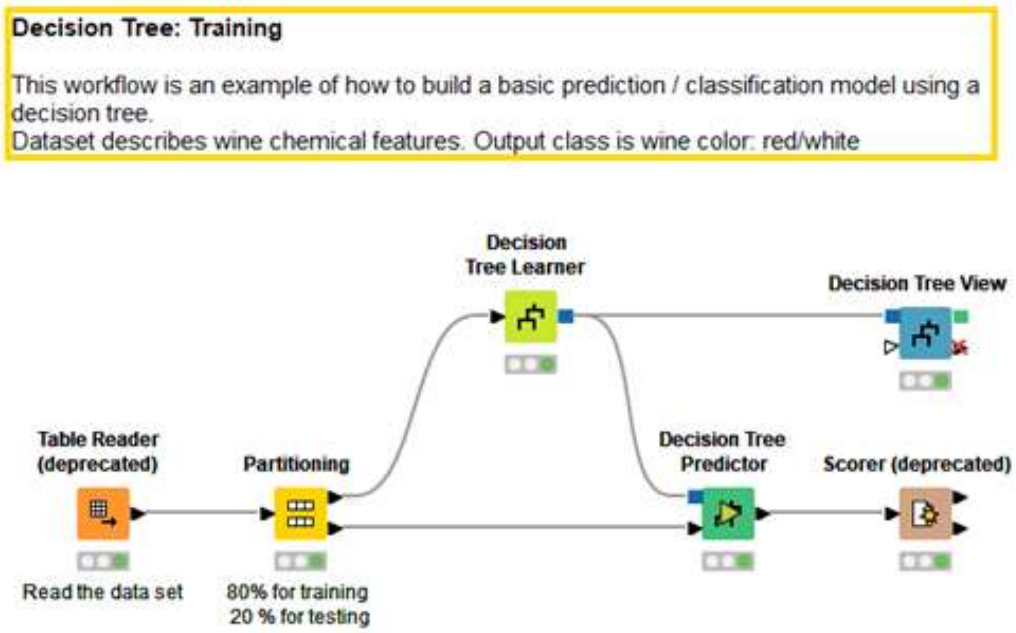

Figura 8. Ejemplo de flujo de trabajo en KNIME

Las características fundamentales por las que KNIME es una herramienta sencilla de usar son las siguientes:

1. Es una herramienta de "Programación visual". El análisis de los datos se puede realizar de forma intuitiva configurando el proceso simplemente haciendo *clics* con el ratón. Se colocan los "nodos" que necesitemos, sin necesidad de conocer su nombre o como se configuran, puesto que en todo momento disponemos de ayudas.
2. Existen nodos para aplicar cualquier procedimiento o técnica que se desee, además al tratarse de una herramienta de código abierto, los propios usuarios pueden crear sus propios nodos. Existen nodos para:

    a. Cargar datos desde ficheros o bases de datos.
    b. Crear, modificar o eliminar filas o columnas del conjunto de datos con el que estemos trabajando.
    c. Calcular estadísticas: medias, percentiles, correlaciones etc.
    d. Combinar datos de fuentes de datos distintas.
    e. Construir y evaluar modelos de aprendizaje automático como: clasificación, regresión o *clustering*.
    f. Visualizar los datos usando gráficos de barras, tarta, dispersión y también otros tipos de gráficos más avanzados.
    g. Generación de informes.

3.6.1. Instalación.

KNIME es una aplicación Java, lo que significa que será necesario tener instalada la máquina virtual de Java antes de poder instalar y ejecutar el programa.

Para instalar el software deberemos ir a https://www.knime.com/downloads, una vez allí descargaremos "*KNIME Analytics Platform*" eligiendo la versión que corresponda para el ordenador personal del que dispongamos: Mac, Windows 32 bits (ordenadores antiguos), Windows 64 (ordenadores modernos) o Linux.

3.6.2. El *Workspace*.

El *Workspace* (espacio de trabajo), es la carpeta o directorio de nuestro ordenador donde están almacenados todos los proyectos realizados con KNIME. Será necesario

elegir un *Workspace* antes de arrancar el programa (también se puede dejar la carpeta que aparece por defecto al realizar la instalación).

3.6.3. Ejemplos de uso.

En el material adicional están disponibles ejemplos donde se repasan algunos conceptos clave de KNIME, aunque dichos conceptos se aprenden mucho mejor si el estudiante los realiza en su propio ordenador mientras sigue las diapositivas.

Resumen

En este tema IV.1 se han abordado conceptos básicos relacionados con **la Minería de Datos**, así como algunas técnicas sencillas de **Minería de datos** para aplicar a investigaciones en el campo de la atención temprana.

Glosario

***Clustering*:** es una técnica de minería de datos, que se emplea generalmente con datos no etiquetados, que permite agrupar datos en función de sus similitudes o diferencias.

DM: *Data Mining* o Minería de datos, es un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, con el objetivo de encontrar patrones repetitivos que expliquen el comportamiento de estos datos y, que estos, puedan ser utilizados para extraer conclusiones.

ML: *Machine Learning*, es una disciplina del campo de la Inteligencia Artificial que dota a las máquinas de la capacidad de "aprender", a partir del análisis de datos trata de identificar patrones y apoyar en la toma decisiones.

Bibliografía

Bogarín, A., Romero, C., & Cerezo, R. (2016). Aplicando minería de datos para descubrir rutas de aprendizaje frecuentes en Moodle. *Revista de Educación Mediática y TIC*, 5(1), 73-92

Chapelle, O., Schölkopf, B. y Zien, A. (2006). *Semi-Supervised Learning: Adaptive computation and machine learning*. MIT Press

Cunningham, P., Cord, M., & Delany, S. J. (2008). *Supervised learning. In Machine learning techniques for multimedia* (pp. 21-49). Springer, Berlin, Heidelberg.

Peterson, P. L., Baker, E., & McGaw, B. (2010). *International encyclopedia of education*. Elsevier Ltd

Rodríguez-Arribas, S. (2021). *Minería de datos aplicada al procesamiento automático en el análisis del proceso de enseñanza-aprendizaje* [Tesis doctoral, Universidad de Burgos]. Repositorio académico de la Universidad de Burgos https://riubu.ubu.es/handle/10259/6704

Romero, C., Cerezo, R., Bogarín, A., Sánchez-Santillán, M. (2016). Educational Process Mining: A tutorial and case study using Moodle data sets. En S. Elatia, D. Ipperciel., & O.R. Zaïane (Eds.), Data Mining and Learning Analytics (pp. 3-28). New Jersey: Wiley Online Library. doi: 10.1002/9781118998205.ch

Sáiz-Manzanares, M.C., Escolar-Llamazares, M.C., Rodríguez-Media, J. (2019). *Investigación cualitativa. Aplicación de métodos mixtos y de técnicas de minería de datos*. Burgos: Servicio de Publicaciones de la UBU. ISBN: 978-84-16283-79-8.

Sáiz-Manzanares, M.C., Marticorena, R., Arnaiz, Á., Díez-Pastor, J.F., y García-Osorio, C.I. (2020). Measuring the functional abilities of children aged 3-6 years old with observational methods and computer tools. *Journal of Visualized Experiments*, e60247, 1-17. https://doi.org/10.3791/60247

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1), 1-130.

Bibliografía básica Módulo

García, S., Luengo, J., y Herrera, F. (2015). *Data Preprocessing in Data Mining* / by Salvador García, Julián Luengo, Francisco Herrera. Springer

Sáiz-Manzanares, M.C., Marticorena, R., y Arnaiz-Gonzalez, Á. (2022). Improvements for therapeutic intervention from the use of web applications and machine learning techniques in different affectations in children aged 0-6 years. *Int. J. Environ. Res. Public Health*, 19, 6558. https://doi.org/10.3390/ijerph19116558

Sáiz-Manzanares, M.C., Marticorena, R., & Arnaiz, Á. (2020). Evaluation of Functional Abilities in 0–6 Year Olds: An Analysis with the eEarlyCare Computer Application. (2020). I*nt. J. Environ. Res. Public Health*, 17(9), 3315, 1-17 https://doi.org/10.3390/ijerph17093315

Sáiz-Manzanares, M.C., Marticorena, R., Arnaiz-González, Á., Díez-Pastor, J.F., & Rodríguez-Arribas, S. (2019, March). Computer application for the registration and automation of the correction of a functional skills detection scale in Early Care. 13th International Technology, Education and Development Conference Proceedings of INTED2019 Conference 11th-13th (5322-5328). IATED: Valencia. doi: 10.21125/inted.2019.1320

Recursos

Software

KNIME

https://www.knime.com/downloads