Co-funded by the European Union


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

# MODULE IV.1

## Observation and Evaluation Techniques from Intelligent Resources: Introduction to Data Mining

Teachers
Dr. Álvar Arnaiz González
Dr. Jose Francisco Díez Pastor
Dra. Sandra Rodríguez Arribas
Department of Computer Engineering
University of Burgos

e-EarlyCare-T

## Tabla de contenido

"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."

# I. Introduction

We live in the information and communication society, the technology we use in the twenty-first century is associated with the collection and storage of large amounts of data. **Data Mining** (**DM**) allows you to find information contained in the data that is not always apparent, since, given the gigantic volume of existing data, much of that volume will never be analysed.

# II. Objectives

1. Know key concepts related to **Data Mining**

2. Know and apply simple **Data Mining** techniques in the field of early care.

# III. Topic-specific contents

### 3.1. Data Mining.

**Data Mining** is the process of searching and analyzing large databases to find useful information that is useful for decision making.

There are numerous **DM** techniques that employ mathematical analysis to deduce the patterns and trends that exist in the data. Typically, these patterns cannot be detected by traditional data exploration because the relationships are too complex or because the volume of data to be analyzed is too large.

Currently in the field of **Data Mining** it is continuously used for the analysis of large amounts of data in various fields of knowledge such as education, economics, business, the environment.

### 3.1.1 Basic concepts in Data Mining

Before knowing the process that is carried out and the types of algorithms that are used in the **DM** it is important to clarify some basic concepts that appear frequently in the bibliography associated with **Data Mining**.

**Data set**

It is a large collection of data usually organized into rows and columns containing variables and attributes. Each of these values is known by the data name. The dataset can also consist of a collection of documents or files.

**Classes or tags**

In the field of **Data Mining**, a class is the discrete attribute whose value you want to predict based on the values of other attributes. It is also known as a label.

**Instance**

An instance is each of the data that is available for analysis. Each instance, in turn, is composed of features that describe it. For example, in a spreadsheet, the instances would be the rows and the features the information stored in the columns.

**Algorithm**

In computer science, an algorithm is a set of defined, ordered, and bounded instructions to solve a problem, perform a calculation or develop a task. In other words, it is a step-by-step procedure to get a result.

### 3.1.2. Process of Application of Data Mining Techniques

The process consists of four main phases listed below:

1. **Problem definition**: this is the first phase in which a specific problem is translated into a **data mining** problem in which the objectives of the analysis and research questions are raised.
2. **Data preparation and collection**: It is the most extensive phase of the process since data quality is one of the most important challenges in **data mining**. Raw data must be identified, cleaned, and stored in a preset format.

3. **Modeling and evaluation**: in this step, different data modeling techniques (algorithms) are selected and applied and then the optimal parameters and values of these techniques are established.
4. **Deployment**: this is the last phase in which the results of **data mining** are organized and presented using graphs and reports. See Figure 1.

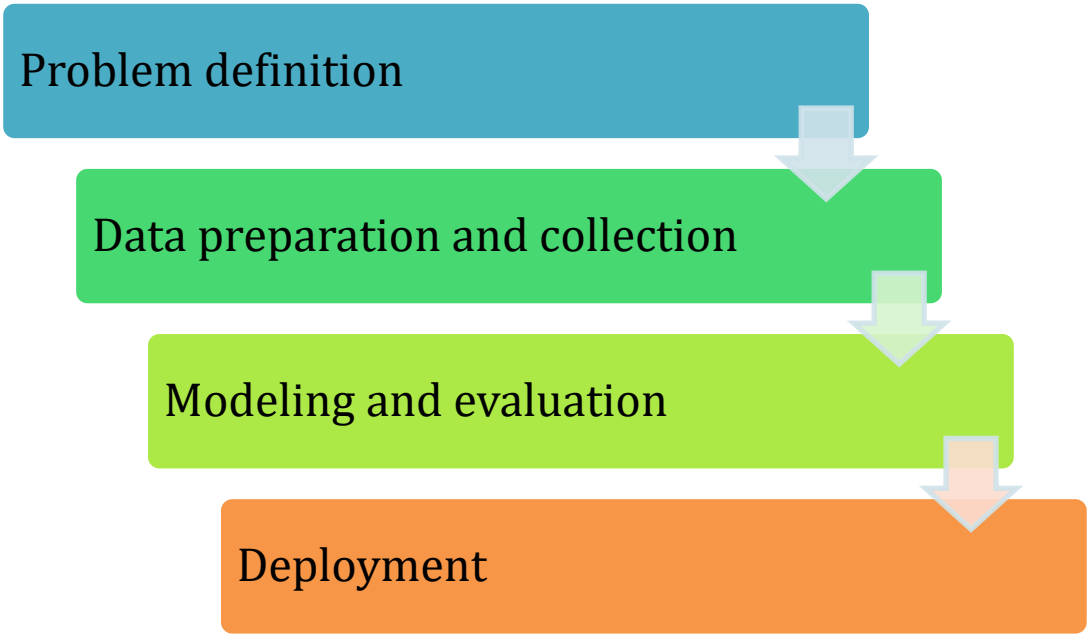


**Figure 1.** Process of application of **data mining** techniques. Source: Own elaboration

It is important to note that every **data mining** process is an iterative process, which means that the process does not stop when a particular solution is deployed. It may be just a new entry for another **data mining** process (Rodríguez-Arribas, 2021). That is, on many occasions the application of **DM** techniques requires several iterations and the use of different algorithms to be able to extract the final results of the research we are doing.

### 3.2 Types of Learning in Data Mining

There are numerous classifications of the algorithms used in the world of **Data Mining**, but it is essential to understand that there are two basic approaches: supervised learning and unsupervised learning. The main difference is that in supervised learning there is a class that is used to obtain a function that allows you to associate new data with the corresponding class. However, in unsupervised learning there is no class, in this case algorithms try to discover hidden patterns in the data

without human intervention in the form of tags associated with the data. (Chapelle, Schölkopf y Zien, 2006).

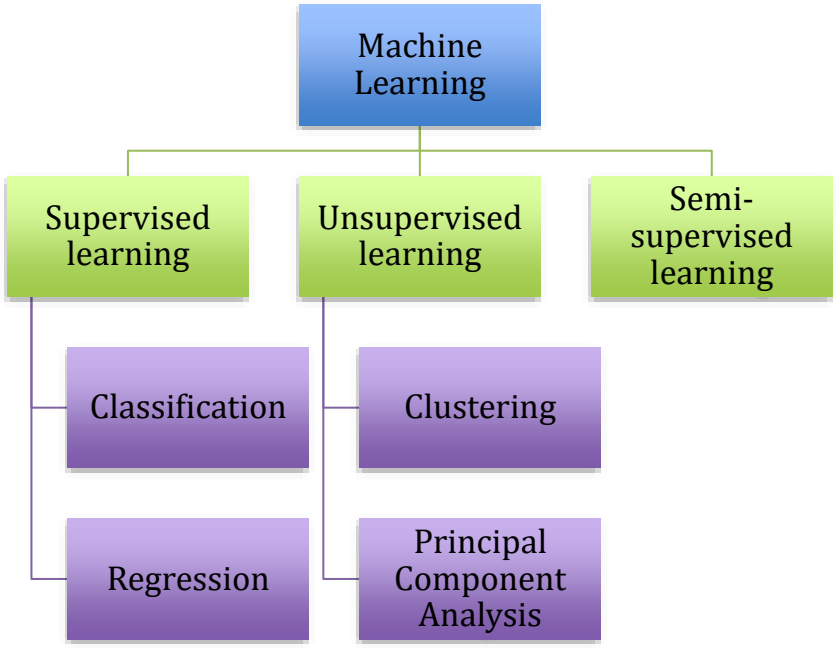


**Figure 2.** Data Mining Methods. Source: own elaboration.

When we must decide which algorithm will be used to perform data analysis, it is important to take into account what type of learning is being used, that is, if we are talking about supervised or unsupervised learning (García, Luengo and Herrera, 2015). According to the type of learning used, different techniques and algorithms will be used as can be seen in the previous image.

### 3.2.1. Supervised Learning

One of the learning modalities of **Machine Learning**, as mentioned above, is supervised learning.

The fundamental objective of supervised learning is the creation of a model that is able to predict values corresponding to input objects after having become familiar with a series of examples, training data.

This technique consists of two fundamental steps:

1. A training phase where a set of labeled data is used, which contain the input data and the desired results for that training data with an algorithm that allows to deduce a function from the data that we are providing to the algorithm

2. The test phase, where the function obtained in the previous step is used to generate new predictions with new data sets.

See Figure 3.

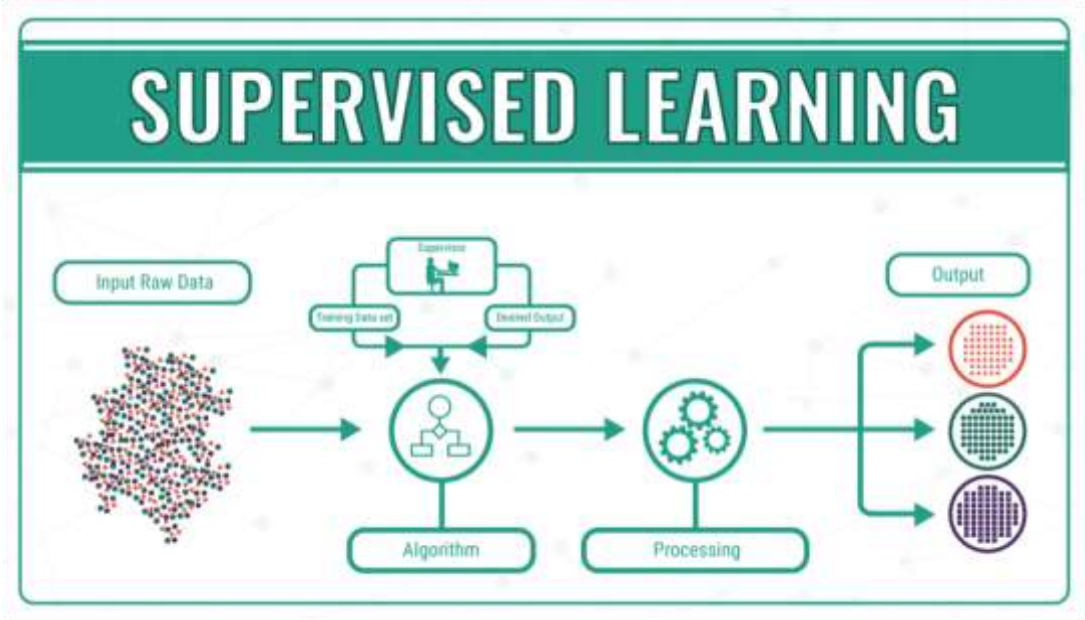


**Figure 3.** Process of operation of supervised learning. Source: ExperiencIA Oracle.

The process is known as supervised learning, since by knowing the responses of each example of the training set, it is possible to correct the function generated by the algorithm. The training of the algorithm is supervised by correcting its parameters, depending on the results obtained iteratively.

### 3.2.2. Unsupervised Learning

This type of learning is the other basic approach to **Machine Learning** (**ML**). Unsupervised learning has unlabeled data that the algorithm must try to understand for itself.

The goal of this type of learning is to let the machine learn without help or directions from data scientists, that is, without supervision and without a training

dataset. In addition, the machine itself will adjust the results and groupings when there are more suitable results, allowing the machine to understand the data and process it in the best way (see Figure 4).

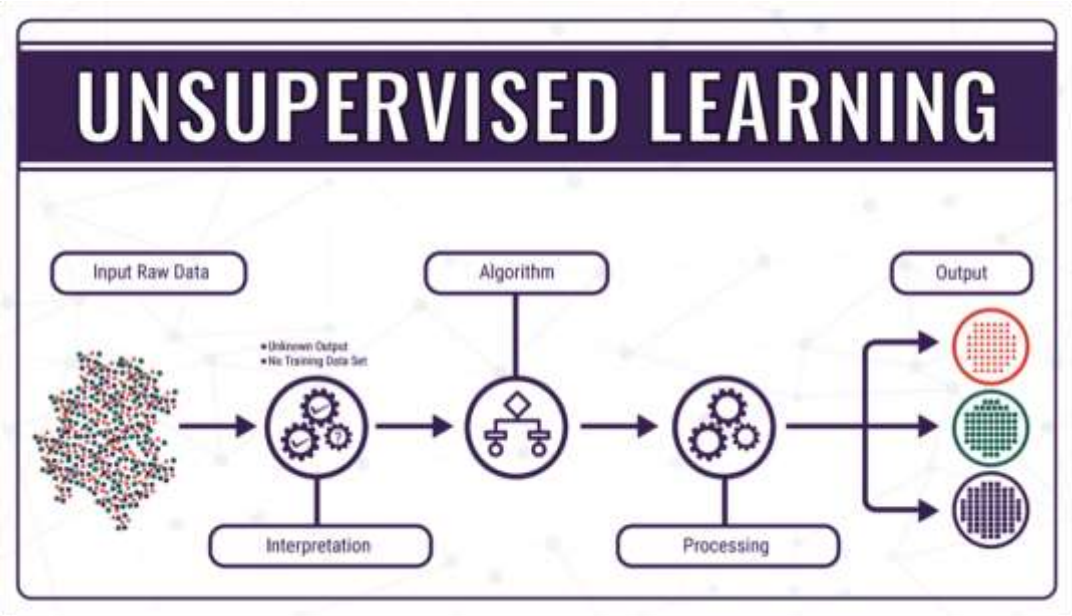


**Figure 4.** Process of operation of unsupervised learning. Source: ExperiencIA Oracle

Unsupervised learning is used to explore unknown and unlabeled data. It can reveal patterns that might have been overlooked or examine large data sets that would be too much for a single person to address.

### 3.2.3 Semi-Supervised Learning

Numerous investigations are currently being conducted with semi-supervised learning methods. These **Machine Learning** techniques use both labeled and unlabeled training data: typically, a small amount of labeled data alongside a large amount of unlabeled data (Zhu and Goldberg, 2009). That is, they seek to improve the prediction models that are obtained by using exclusively labeled data by exploring the structural information contained in the unlabeled data.

We can say semi-supervised learning tries to combine the two traditional approaches of data mining (supervised learning and unsupervised learning) to keep the best of each of them.

### 3.3. Classification Algorithms.

Classification algorithms are those we use when the expected result is a discrete label. That is, they are useful when the answer to the research question lies within a finite set of possible outcomes.

These algorithms generally work on the information delivered by a set of samples, patterns, examples, or training prototypes that are taken as representatives of the classes, and they retain a correct class label. This set of correctly labeled prototypes is called a training set, and it is the knowledge available for the classification of new samples. The objective of supervised classification is to determine, according to what is known, which class a new sample should concern, considering the information that can be extracted (see Figure 5).

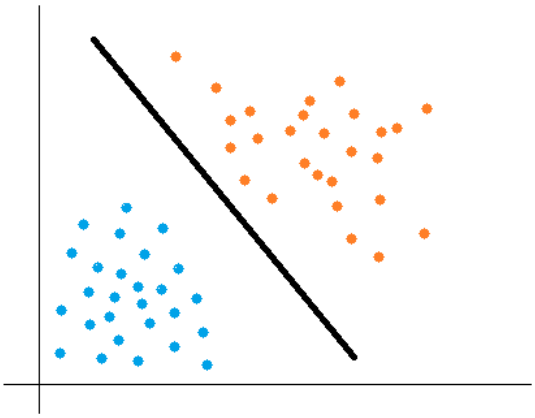

**Figure 5.** Classification algorithm. Source: own elaboration

Classification is very similar to the learning process of people, since we possess the ability to classify food, books, animals, planets, that is, everything around us.

### 3.4. Clustering algorithms

**Clustering** algorithms are responsible for grouping the objects in a dataset according to their similarities. In this way the objects that are within a cluster or group have more similarities between them than differences (see Figure 6).

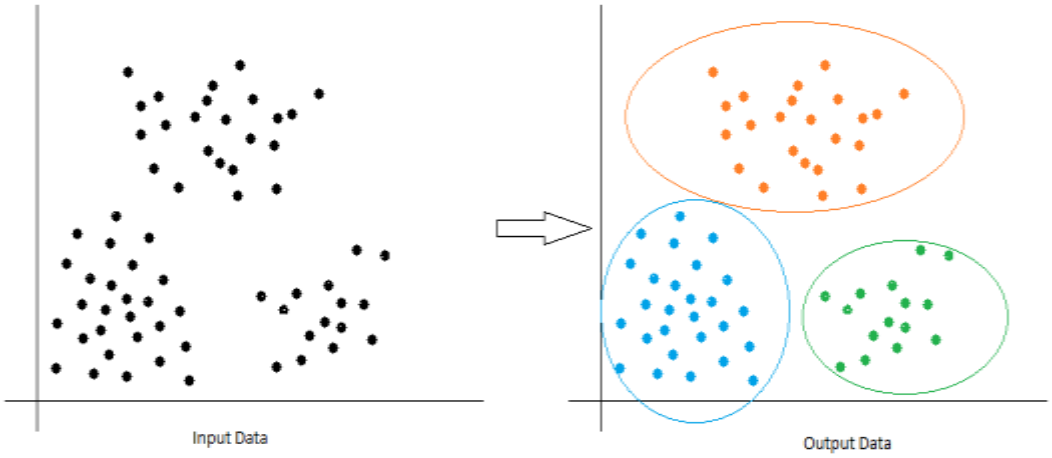


**Figure 6.** *Clustering* algorithm. Source: own elaboration

These algorithms work with unlabeled data, so it is the algorithm itself that analyzes the data to find the optimal number of groupings for the input data set since we do not have prior knowledge about the characteristics of the data and its classes.

The groupings performed by the algorithms can be of two types:

1. **Hard cluster**: each piece of data belongs exclusively to a group.
2. **Soft (diffuse) cluster:** the data can belong to several groups in different degrees, that is, the same data can have a degree of belonging of 60% to group 1 and 40% in group 2.

### 3.5. Regression Algorithms

Regression algorithms are a subfield of supervised learning whose goal is to establish a method for the relationship between a certain number of characteristics and a continuous target variable.

These are algorithms that establish a line to provide the trend of a set of data, that is, the purpose of these algorithms is to relate a number of characteristics and a continuous objective variable (see Figure 7).

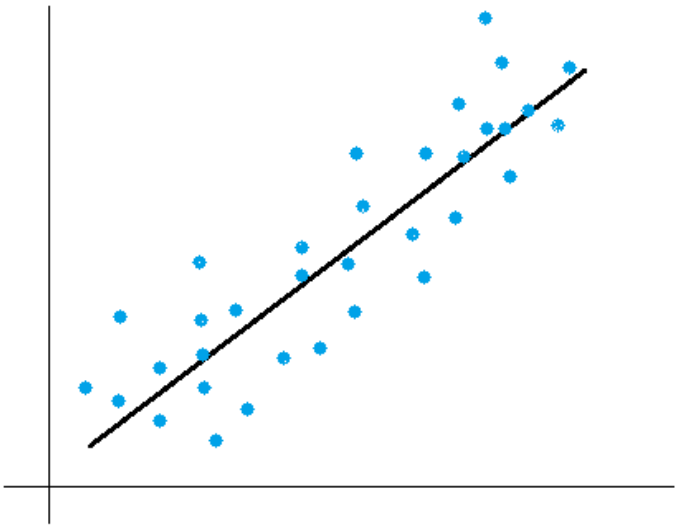

**Figure 7.** Regression algorithm. Source: own elaboration

This technique is useful for predicting outcomes that are continuous values, that means that the answer to the research question is presented by a quantity that can be flexibly determined based on model inputs rather than being limited to a finite set of labels as in the case of classification.

### 3.6. KNIME.

KNIME is an open-source application that allows us to apply to our own datasets or to sample datasets:

1. Statistical methods
2. **Data mining** algorithms or **Machine Learning**.
3. Visualization techniques.

Being an open-source software has many advantages, its code belongs to the community of users and developers, which guarantees that it will always be a free and free tool. In contrast, private software belongs exclusively to a company and this company can allow its free use, but also charge a high price or demand the payment of a monthly subscription.

It is a tool designed to be simple to use. The most important concept in the use of the tool is that of *workflow* (in Spanish, workflow). A workflow is a sequence of steps configured by the user. Formally it is a set of nodes joined together with

arrows. A node encapsulates different jobs that can be done with the data, there are nodes for many tasks. A workflow might have a node to load a dataset from an Excel file, then a node to select attributes (columns) from that dataset, and then another node to display statistics for the selected attributes (see Figure 8).

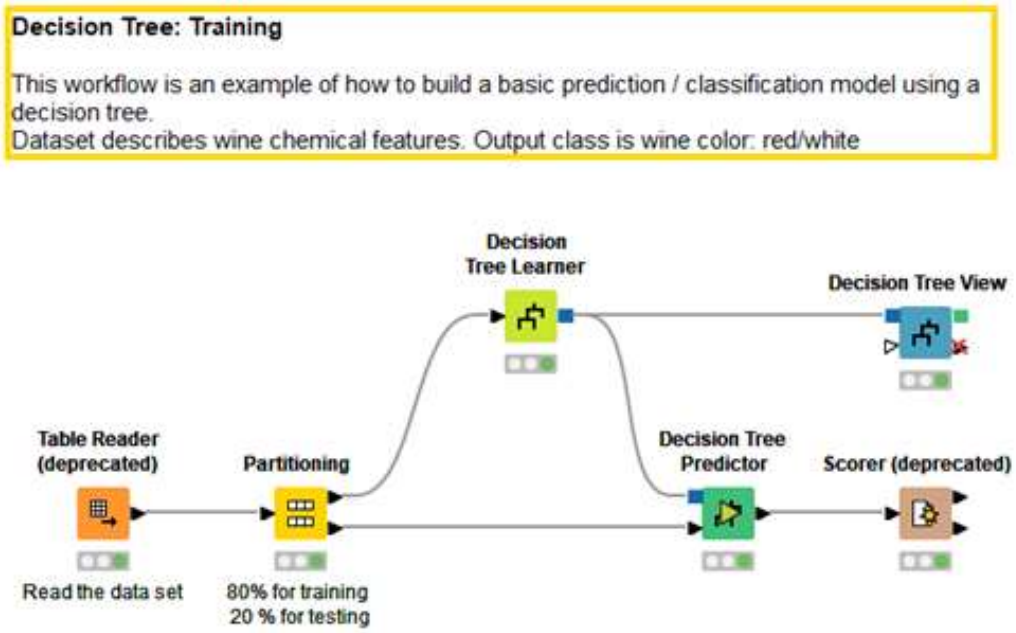


**Figure 8.** Example of a workflow in KNIME

The fundamental features why KNIME is an easy-to-use tool are the following:

1. It is a "Visual Programming" tool. Data analysis can be done intuitively by setting up the process simply by *clicking* the mouse. The "nodes" that we need are placed, without the need to know their name or how they are configured, since at all times we have help.
2. There are nodes to apply any procedure or technique you want, in addition to being an open-source tool, users themselves can create their own nodes. There are nodes for:

    a. Load data from files or databases.
    b. Create, modify, or delete rows or columns from the dataset we are working with.
    c. Calculate statistics means, percentiles, correlations etc.
    d. Combine data from different data sources.
    e. Build and evaluate **Machine Learning** models such as: classification, regression, or **clustering**.

f. Visualize the data using bar charts, pie charts, scatter charts, and also other more advanced chart types.
g. Generation of reports.

### 3.6.1. Installation.

KNIME is a Java application, which means that you will need to have the Java virtual machine installed before you can install and run the program.

To install the software, we must go to https://www.knime.com/downloads, once there we will download "KNIME Analytics Platform" choosing the corresponding version for the personal computer we have: Mac, Windows 32 bits (old computers), Windows 64 (modern computers) or Linux.

### 3.6.2. The Workspace

The workspace is the folder or directory of our computer where all the projects carried out with KNIME are stored. It will be necessary to choose a workspace before starting the program (you can also leave the folder that appears by default when installing).

### 3.6.3. Examples of use.

Examples are available in the additional material where some key KNIME concepts are reviewed, although these concepts are much better learned if the student performs them on their own computer while following the slides.

## Summary

In this IV.1-unit, basic concepts related to Data Mining have been addressed, as well as some simple data mining techniques to apply to research in the field of early care.

## Glossary

**Clustering**: is a data mining technique, which is generally used with unlabeled data, which allows data to be grouped according to their similarities or differences.

**DM**: **Data Mining**, is a set of techniques and technologies that allow you to explore large databases, with the aim of finding repetitive patterns that explain the behavior of this data and that these can be used to draw conclusions.

**ML**: **Machine Learning**, is a discipline in the field of Artificial Intelligence that gives machines the ability to "learn", from the analysis of data tries to identify patterns and support decision making.

## Bibliography

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., 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.

### Basic Bibliography Module

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

## Resources

### Software

KNIME

https://www.knime.com/downloads