



U2 MACHINE LEARNING

U2.E1 MACHINE LEARNING OVERVIEW

Artificial Intelligence Technician

March 2021, Version 1



Co-funded by the
Erasmus+ Programme
of the European Union

The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

The student is able to

AIT.U2.E1.PC1	Understand the connection between artificial intelligence and machine learning.
AIT.U2.E1.PC2	Know how to define machine learning.
AIT.U2.E1.PC3	Explain the value proposition of machine learning and understand its particularities and key features.

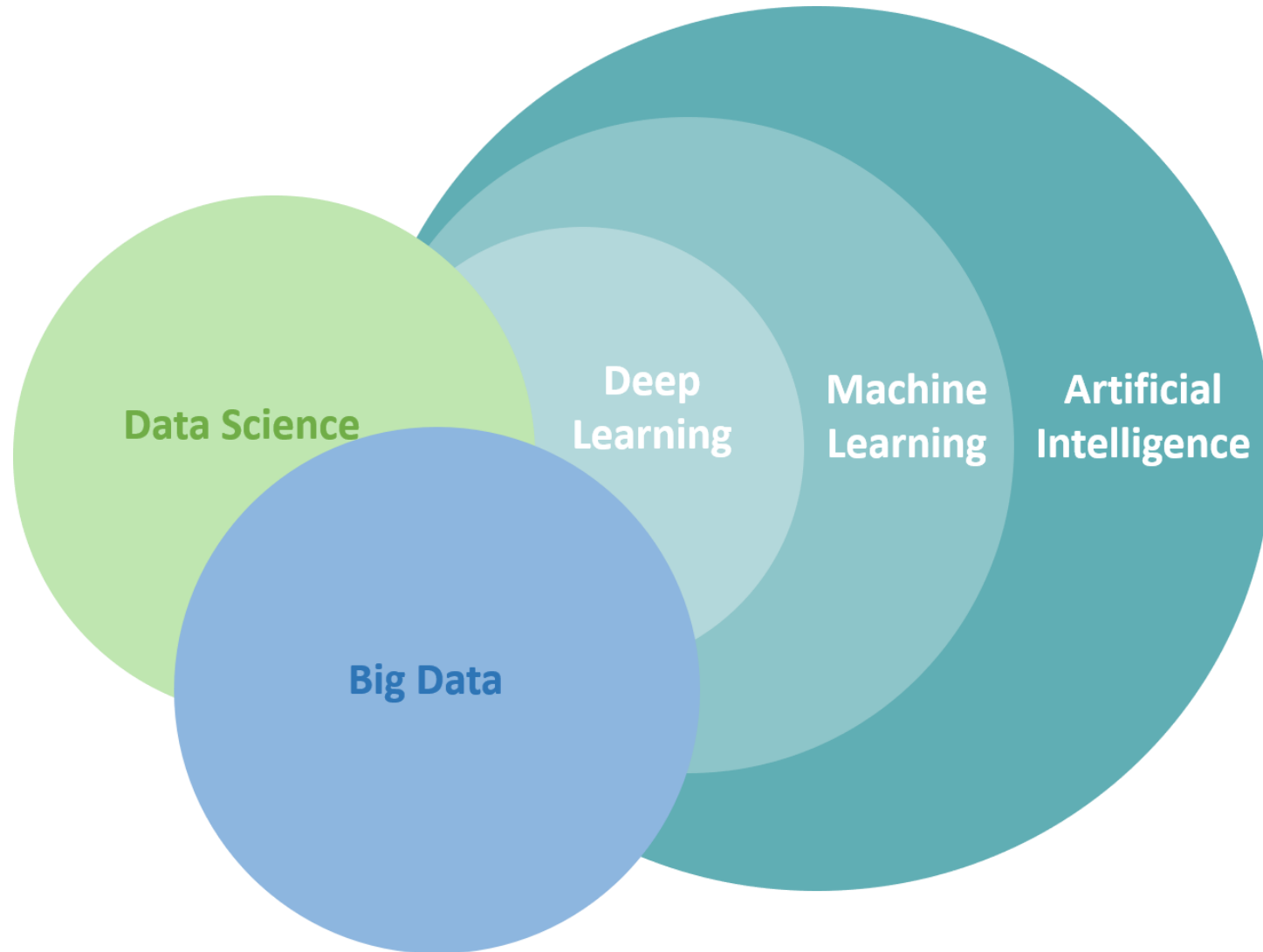
Machine Learning empowers learning systems to act and take data driven decisions to carry out a certain task.

Machine Learning is a subset of Artificial Intelligence that focuses on the development of computer programs (algorithms) that can grant access to data and then use it to learn for themselves.

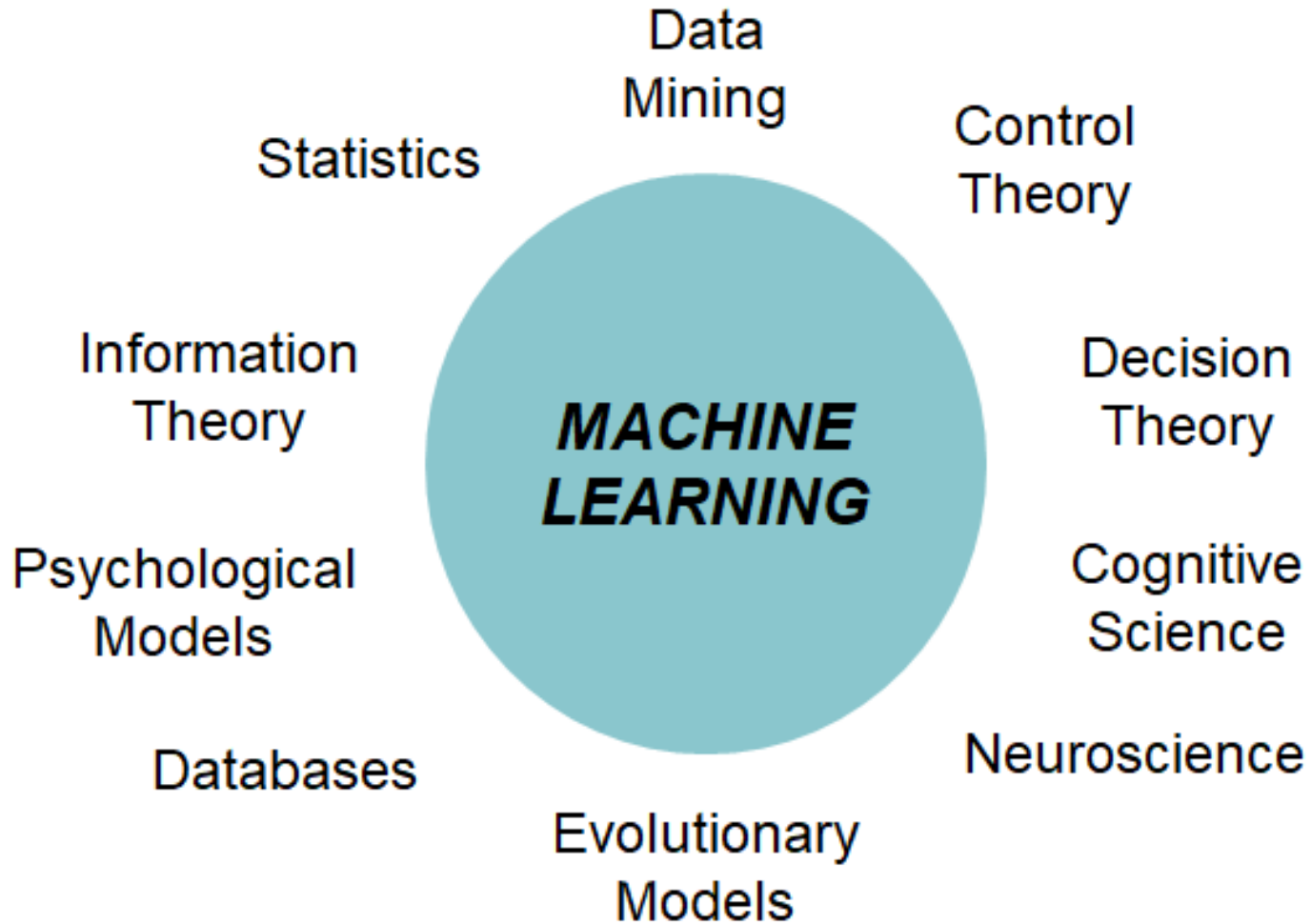
These algorithms can learn and improve over time when exposed to new data, i.e., ML uses statistical methods to enable machines to learn and improve with experience.

Its main goal is to enable computers to learn automatically without human intervention or assistance, as well as, to reduce distance between estimated value and real value (the error).

WHAT IS MACHINE LEARNING?



WHAT IS MACHINE LEARNING?



- ✓ AI is defined as acquisition of knowledge intelligence
- ✓ The aim is to increase the chance of success
- ✓ It works as a computer program that does smart work
- ✓ The goal is to simulate natural intelligence to solve complex problems

- ✓ ML is defined as the acquisition of knowledge or skill
- ✓ The aim is to increase accuracy
- ✓ It is a simple machine that takes data and learns from it
- ✓ The goal is to learn from data a specific task to maximize machine performance on this task

- ✓ AI is decision making
- ✓ It leads to the development of a system that mimics human behavior to respond to circumstances
- ✓ AI pursues the optimal solution
- ✓ Leads to intelligence or wisdom

- ✓ ML enables systems to learn new things from data
- ✓ It is involved in the creation of self-learning algorithms
- ✓ ML will only come up with a solution that is either optimal or not
- ✓ Leads to knowledge

WHY DO WE NEED MACHINE LEARNING?

*The work that
we need to do is
increasing
day-to-day*

*Data-based
decisions
increasingly make a
difference between
companies*

*Saves the
manpower of the
organization and
also increases the
productivity*

- **No human experts**
 - industrial/manufacturing control
 - mass spectrometer analysis, drug design, astronomic discovery, etc.

- **Black-box human expertise**
 - face/handwriting/speech recognition
 - emotion and activity recognition
 - driving a car, flying a plane

WHY DO WE NEED MACHINE LEARNING?

- **Rapidly changing phenomena**
 - credit scoring, financial modeling
 - medical diagnosis, fraud detection

- **Need for customization/personalization**
 - personalized feeds, news reader, etc.
 - products/services recommendation

WHY DO WE NEED MACHINE LEARNING?



input
data



analyze data



find
patterns



prediction



improvement



HEALTHCARE



ENTERTAINMENT



RETAIL



TRANSPORT



SOCIAL MEDIA



EDUCATION

WHAT KIND OF PROBLEMS CAN BE TACKLED USING MACHINE LEARNING?

- **Text or document classification**, which includes problems such as assigning a topic to a text or a document, spam detection, and determining automatically if the content of a web page is inappropriate or too explicit.
- **Natural language processing (NLP)**, which includes part-of-speech tagging, named-entity recognition, context-free parsing, or dependency parsing.
- **Computer vision applications**, which includes object recognition, object identification, face detection, Optical Character Recognition (OCR), content-based image retrieval, or pose estimation.

WHAT KIND OF PROBLEMS CAN BE TACKLED USING MACHINE LEARNING?

- **Speech processing applications**, which includes speech recognition, speech synthesis, speaker verification, speaker identification, as well as sub-problems such as language modeling and acoustic modeling.
- **Computational biology applications**, which includes protein function prediction, identification of key sites, or the analysis of gene and protein networks.
- Other applications such as fraud detection, network intrusion, learning to play games, unassisted vehicle control such as robots or cars, medical diagnosis, design of recommendation systems, search engines, or information extraction systems.

WHAT KIND OF PROBLEMS CAN BE TACKLED USING MACHINE LEARNING?

This list is by no means comprehensive!



Most prediction problems found in practice can be cast as learning problems and the practical application area of machine learning keeps expanding.

- **This trend is accelerating:**

- Improved machine learning algorithms;
- Improved data capture, networking and faster computers;
- Software too complex to write by hand;
- New sensors / IO devices;
- Demand for self-customization to user/environment;
- It turns out to be difficult to extract knowledge from human experts → *failure of expert systems in the 1980's.*

■ Type of Feedback

- supervised (labeled examples)
- unsupervised (unlabeled examples)
- semi-supervised (labeled and unlabeled examples)
- reinforcement (reward)

■ Representation

- attribute-based (feature vector)
- relational (first-order logic)

■ Use of Knowledge

- empirical (knowledge-free)
- analytical (knowledge-guided)

Supervised
Learning

Unsupervised
Learning

Semi-
Supervised
Learning

Reinforced
Learning

- **Supervised (inductive) learning**

Training data includes desired outputs

- **Unsupervised learning**

Training data does not include desired outputs

- **Semi-supervised learning**

Training data includes a few desired outputs

- **Reinforcement learning**

Rewards from sequence of actions

Examples: Items or instances of data used for learning and testing.

Features: The set of attributes also referred to as variables, often represented as a vector, associated to an example. In datasets, features appear as columns.

Labels: Values or categories assigned to examples. In classification problems, examples are assigned specific categories, while in regression, items are assigned real-valued labels.

Hyperparameters (Fitting Parameters): Free parameters that are not determined by the learning algorithm, but rather specified as inputs to the learning algorithm. The optimization of these parameters can be made through greedy search, gradient descent, linear programming and many variations.

Training sample: Examples used to train a learning algorithm.

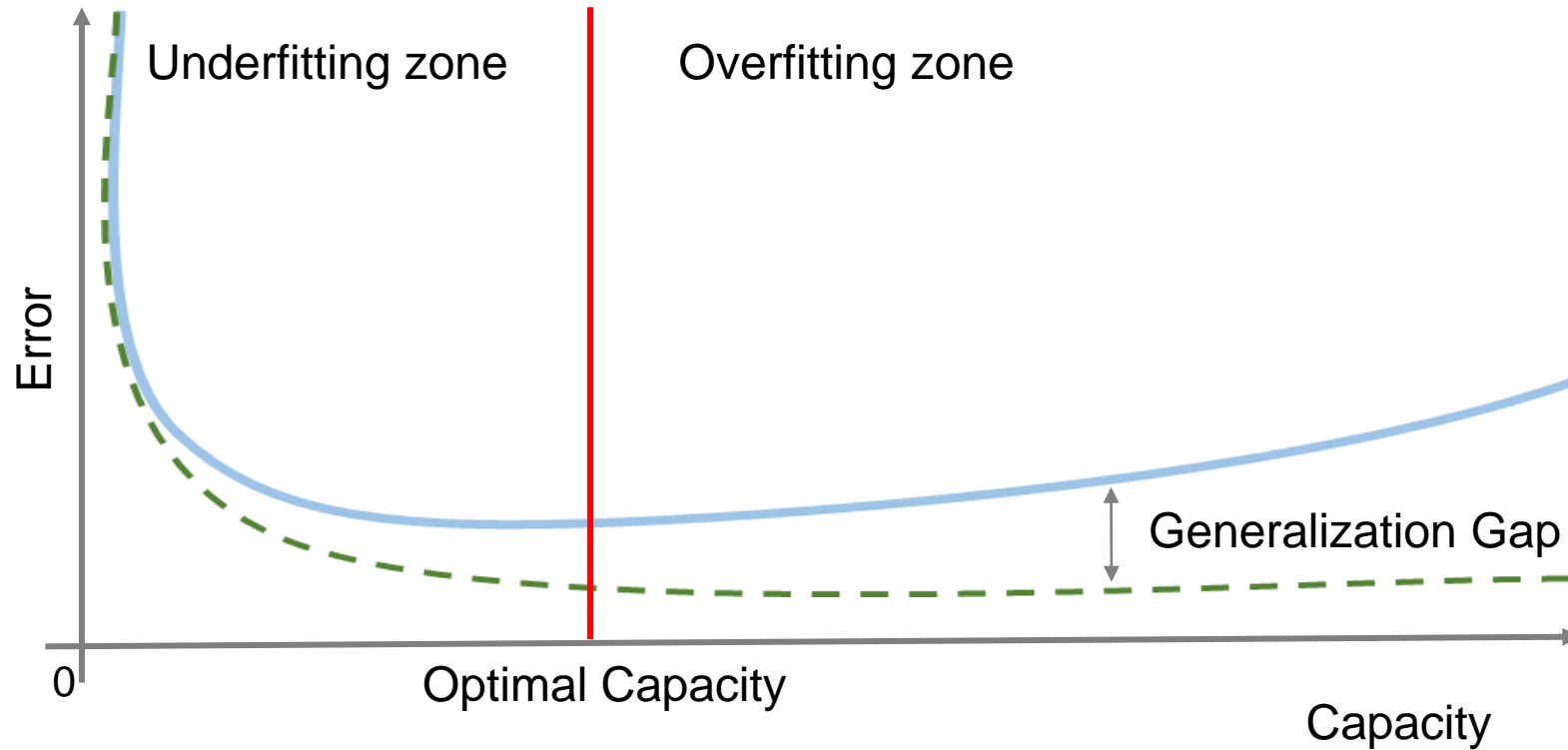
Validation sample: Examples used to tune the parameters of a learning algorithm when working with labeled data. The validation sample is used to select appropriate values for the hyperparameters.

Test sample: Examples used to evaluate the performance of a learning algorithm. The test sample is separate from the training and validation data and is not made available in the learning stage.

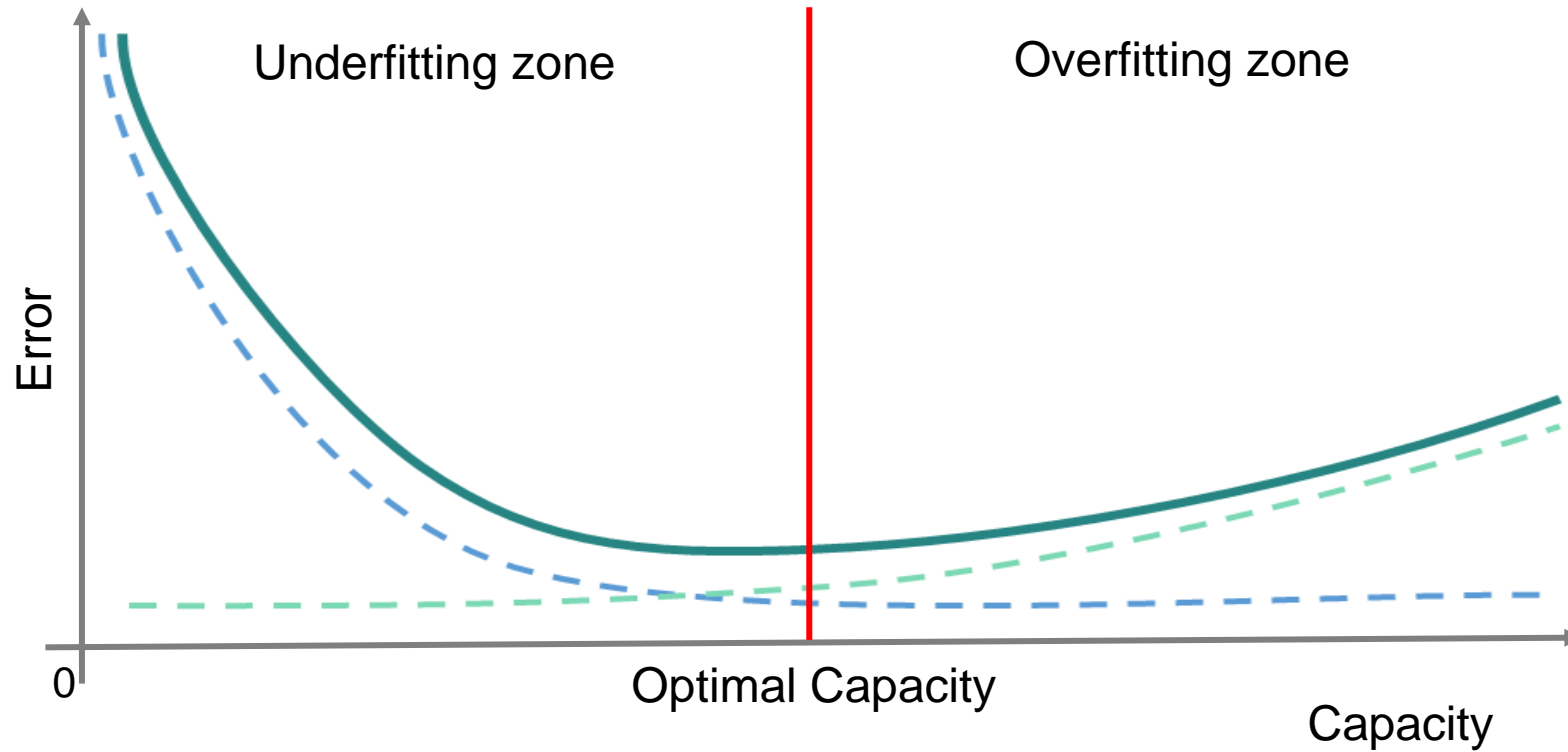
Loss Function: Evaluation of the models through metrics such as accuracy, precision, and squared error.

Deployment/Experimentation Cycle: Dealing with mistakes, adaption over time and maintaining balance.

➔ Generalization and Capacity



➔ Bias and Variance



Both Data Mining (DM) and Machine Learning (ML) are rooted in Data Science. The limits between the two concepts are often blurred, but there are a few differences between them.

ML is the process of discovering algorithms that can learn from and make predictions on the data. It is the design, study and development of algorithms that allow machines to learn without human intervention.

DM is a process for extracting useful information from a large amount of data. It is used to discover new, reliable and useful patterns in the data, to find meaning and information relevant to the company or to the person who needs it. It is used by humans.

These concepts are often intersected or confused and individuals erroneously use these two terms interchangeably.

WHY DO HUMANS CONFUSE THESE TERMS?



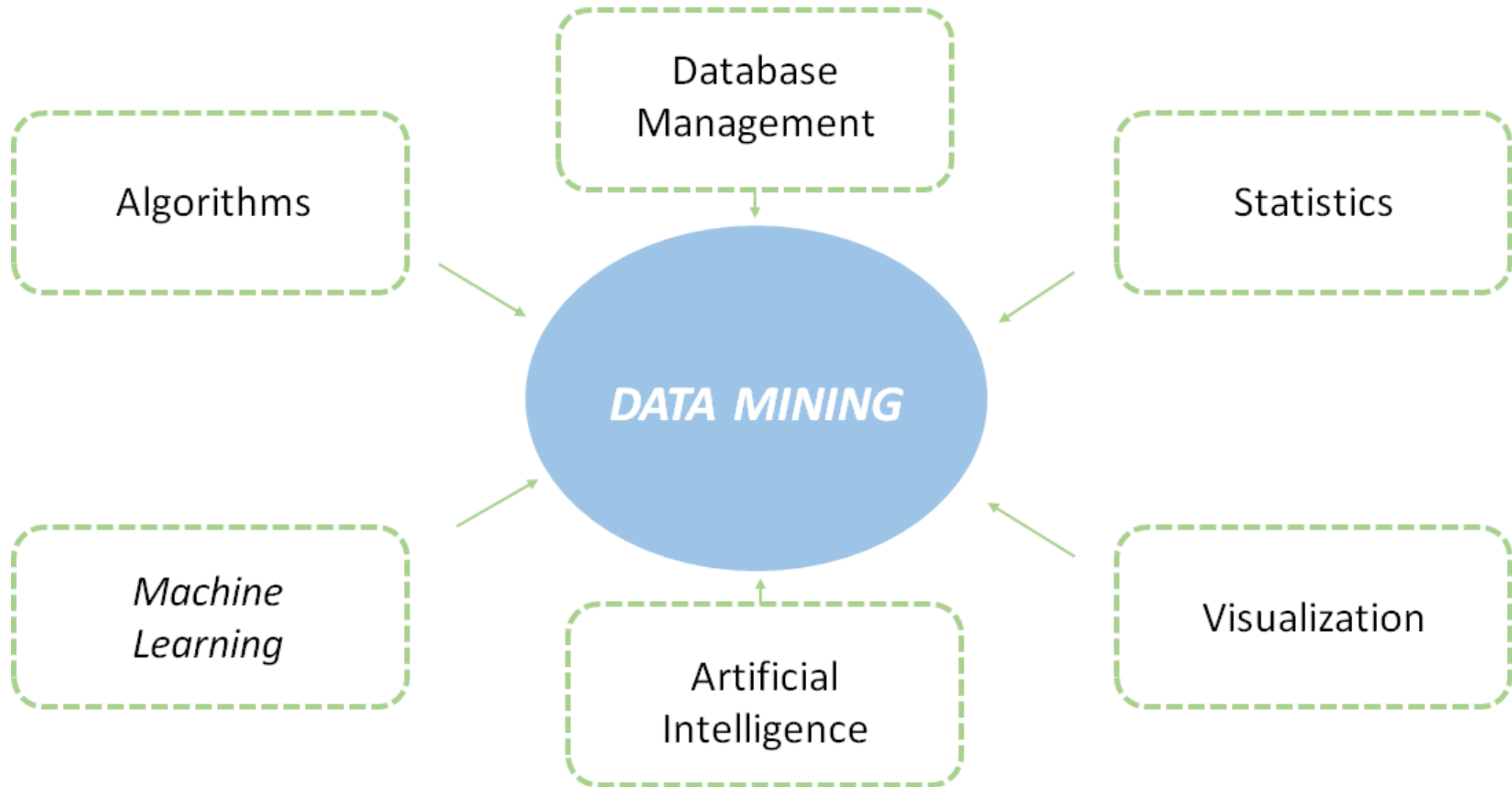
Machine Learning and Data Mining are both analytics processes.

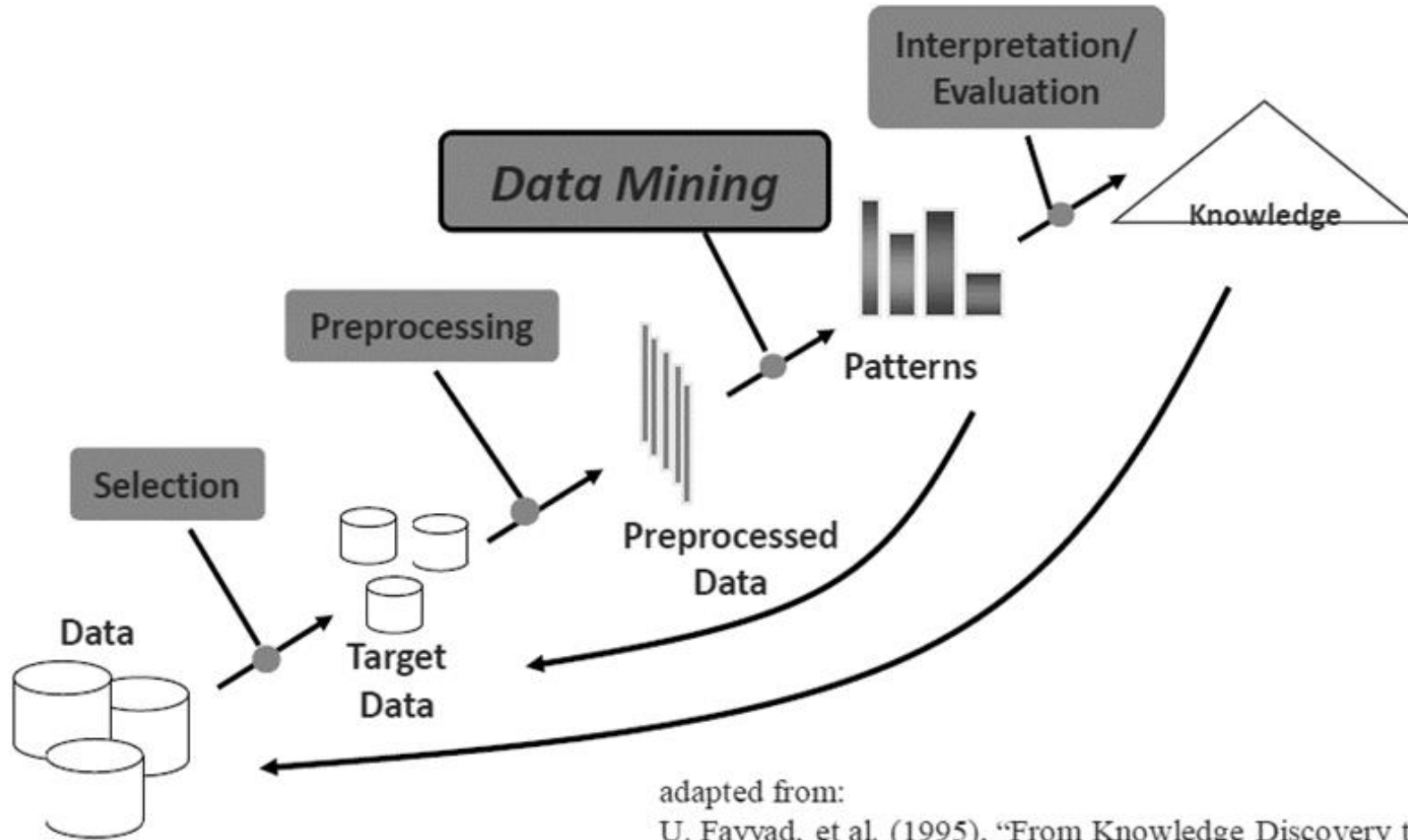


Machine Learning is sometimes used as a means for carrying out useful Data Mining activities.



Both aim to learn from data in order to improve decision-making.





adapted from:

U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

DIFFERENCES



DATE OF INVENTION

DM predates ML by two decades.



PURPOSE

DM is designed to extract rules from a large amount of data, while ML teaches a computer how to learn and understand information to perform complex tasks.

DIFFERENCES



HUMAN FACTOR

DM relies on human intervention and is ultimately designed for human use.

Whereas ML teaches itself and not depends on human influence or actions.



GROWTH ABILITY

DM cannot learn or adapt as it follows pre-set rules and is static by nature, while ML adjusts algorithms as circumstances occur.

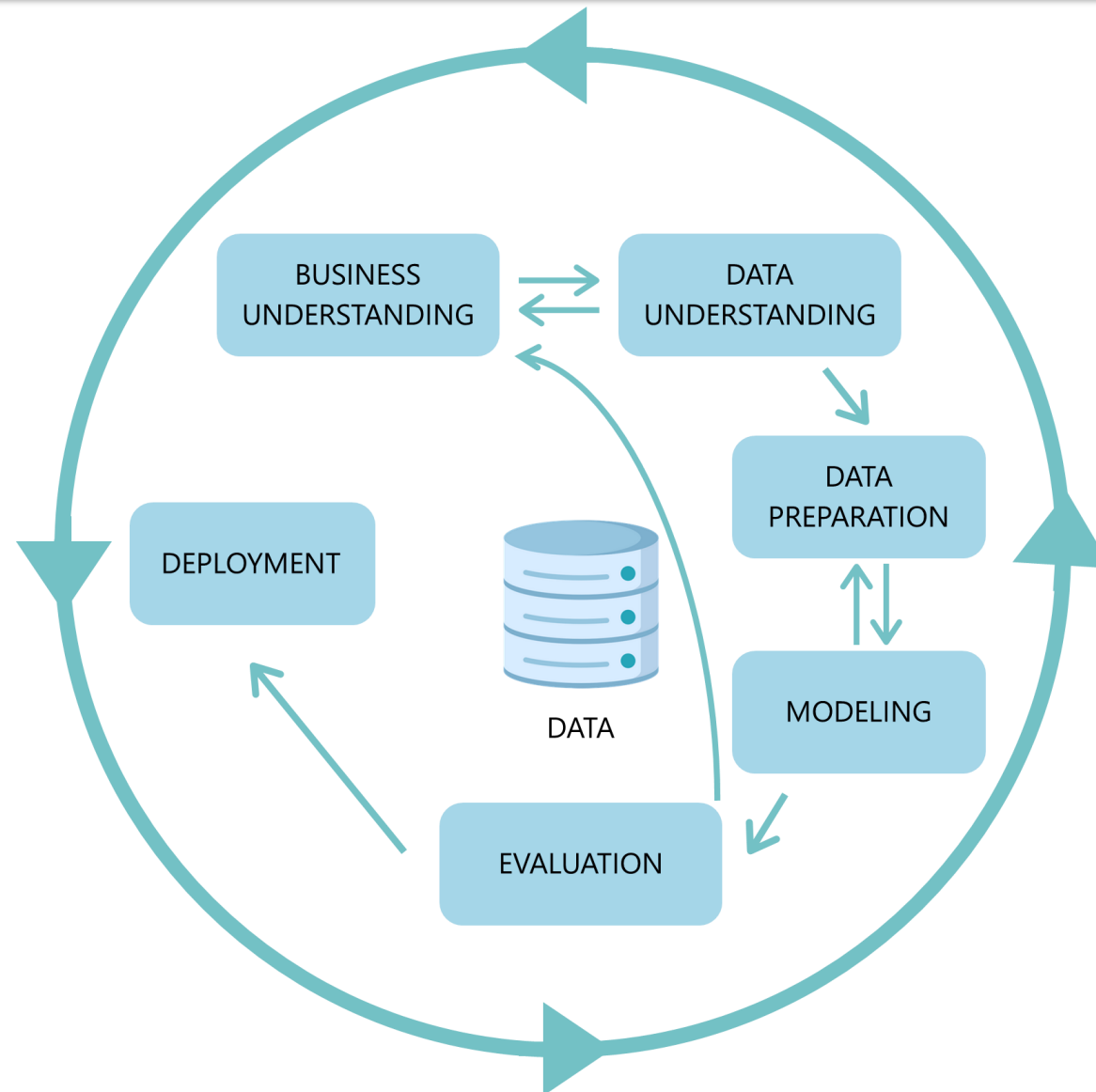
DIFFERENCES



USE OF DATA

Data mining is used on an existing dataset, such as a data warehouse, to discover patterns. Machine learning, on the other hand, is trained on a training data set, which teaches the computer how to make sense of the data, and then makes predictions about new data sets.

CRISP-DM LIFECYCLE



Business Understanding focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

1. Definition of the objectives in business terminology.
2. Definition of the objectives in technical terms.
3. Design a preliminary research plan.

1. Definition of the objectives in business terminology

- Understand client's needs and expectations;
- Uncover important factors (constraints, competing objectives);
- Identify the business units impacted by the project;
- Define business success criteria;
- Describe the problem in general terms regarding business questions and expected benefits.

2. Definition of the objectives in technical terms

- Identify knowledge sources and types;
- Identify software and hardware available;
- Describe relevant background;
- Translate the business questions into Data Mining goals;
- Specify the Data Mining problem type (classification, regression, clustering, etc.);
- Specify performance criteria for model assessment.

3. Design a preliminary research plan

- Define an initial process plan;
- Discuss its feasibility with involved personnel and stakeholders;
- Estimate efforts and resources;
- Identify challenges and critical steps.

Data Understanding begins with the initial data collection and proceeds with activities aimed at getting acquainted with the data, identifying problems with the quality of the data, discovering initial insights from the data or detecting interesting subsets to form hypotheses for hidden information.

1. Data acquisition.

2. Data analysis and exploration:

- Understand the meaning of each attribute and its value in terms of business goal;
- Analyse attribute types and ranges;
- Compute basic statistics, such as distribution, average/mode and standard deviation, for each attribute;
- Review the dataset's variability and assess the need to cover more cases;
- Analyse properties of attributes and relations between them;
- Identify data inconsistencies, duplicated instances, missing values and outliers.

Data Preparation may take longer than the Data Mining process itself. The importance of data preparation is based on three aspects:

1. Real world data may be incomplete (missing values), noisy (outlier) and inconsistent (female, woman, F, W);
2. High-performance mining systems require quality data;
3. Quality data is a prerequisite for the production of effective models and quality standards.

DATA INTEGRATION

Integration of multiple databases or files.

DATA CLEANING

Removal of duplicates, treatment of missing values, treatment of outliers, resolution of inconsistencies, etc.

DATA TRANSFORMATION

Create attributes, rename attributes, convert data types, change data format, normalize data, etc.

DATA REDUCTION

Feature Selection, Discretization, etc.

DATA SAMPLING

Oversampling, Undersampling

Missing Attributes

Recover Missing Values

Please contact the participants and ask them to fill in the missing values

Remove Missing Values

Delete instances that contain missing values

*** If the sample is large enough, it is likely to be able to remove the data without significant loss of statistical power.*

Impute Missing Values

Replace missing values with alternative values

Outliers

Maintain Outliers

In some cases, outliers do not originate from data errors and correspond to natural aberrant values

Remove Outliers

Remove instances that contain outliers

*** If the sample is large enough, it is likely to be able to remove the data without significant loss of statistical power.*

Replace Outliers

Replace the outliers with the highest or second lowest value in the observations, except for the outliers.

Normalization

When there are attributes with disparate value ranges or at different scales, attributes with values at a higher scale may unrealistically overshadow a significant or equally important attribute (but at a lower scale). Thus, attributes are normalized to transform all attributes on the same scale.

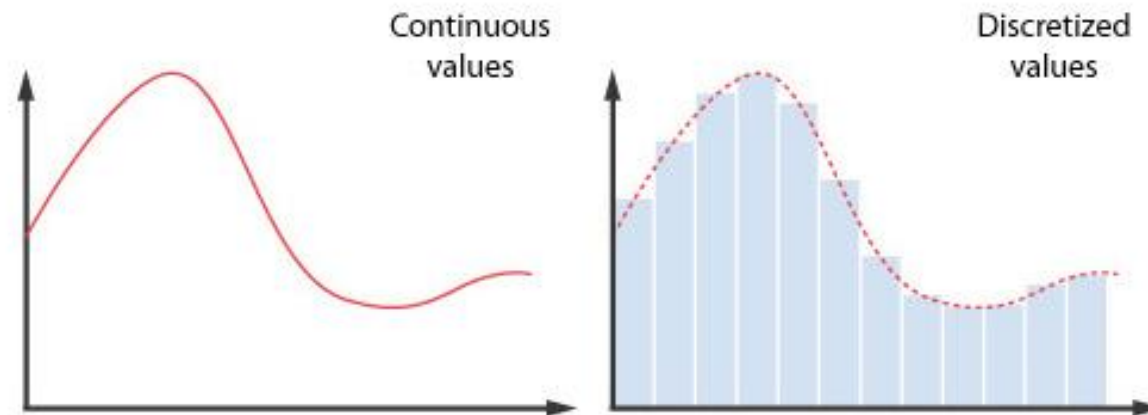
Data normalization allows a new scale to be assigned to an attribute so that the values of that attribute can fall in a new scale in a specific range from 0 to 1 for example.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \rightarrow \text{Min-Max Normalization}$$

➔ Discretization/Binning

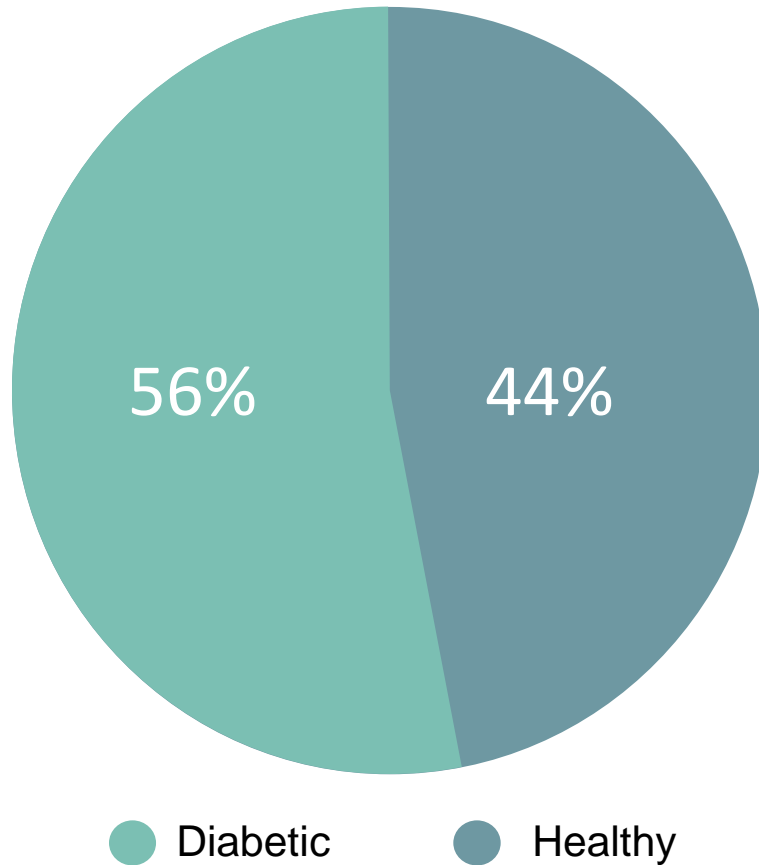
The objective of discretization is to transform a continuous attribute into a discrete attribute. In several Data Mining algorithms, it is necessary to use discretized data since these algorithms can only handle discrete attributes.

Discretization reduces the impact that small fluctuations in the data have on the model, often small fluctuations are just noise. Each "bin" soothes the fluctuations/noise.

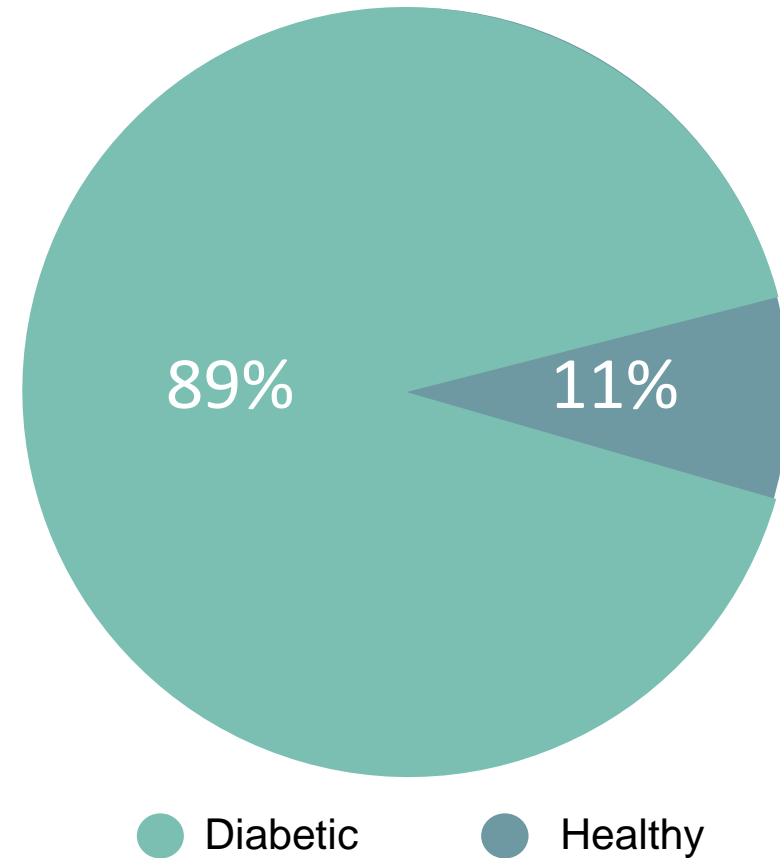


➔ Data Sampling

Balanced Dataset



Unbalanced Dataset



Data Sampling

In these cases, the algorithm receives significantly more examples from a class, which leads it to be skewed to that specific class. Due to the disparity of classes, the algorithm is then prone to categorize instances into the majority class and does not learn what makes the other class "different", nor does it understand the underlying patterns that allow classes to be distinguished.

Classifiers generated from unbalanced datasets have high false negative rates for the less common classes.

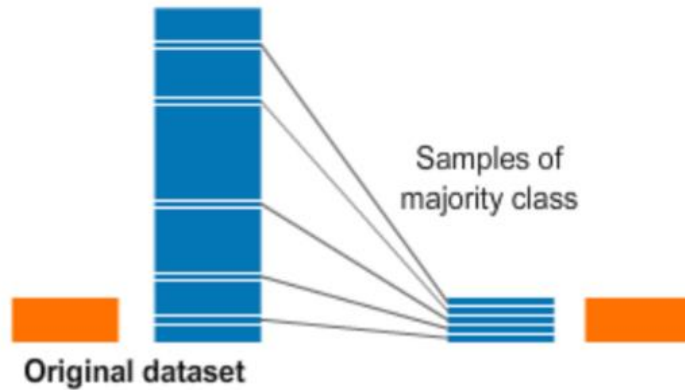
As there are few instances of the minority class, the associated error is reduced, giving at the same time the false sense that we are building a highly accurate model. Both the inability to predict rare events, i.e., the minority class, and the misleading accuracy decrease the performance of the prediction models built.

➡ Data Sampling

Alteration of the class distributions in the data set, with the aim of reducing the imbalance and obtaining better classifiers than those obtained from the original distribution.

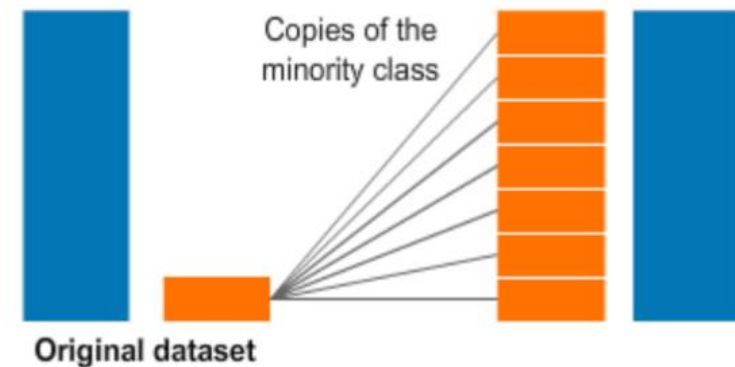
Undersampling

Removal of cases from the majority class



Oversampling

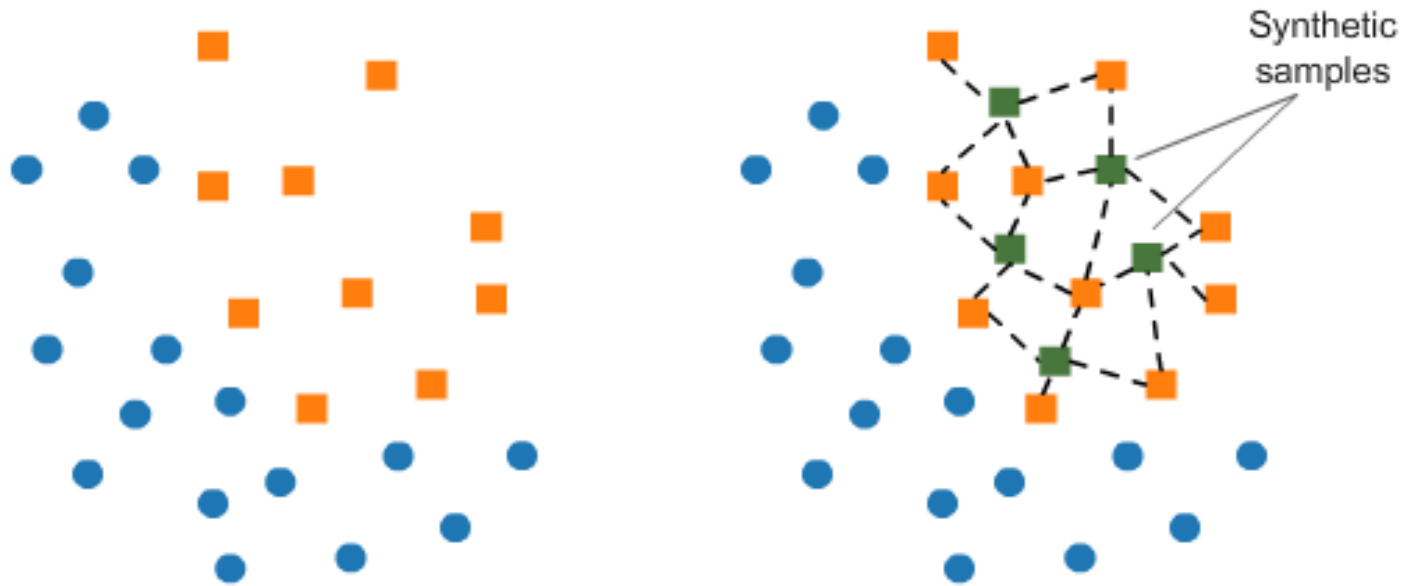
Replication of cases from the minority class



*** substantial loss of statistical power may occur*

➡ Data Sampling – **SMOTE** (Synthetic Minority Over-sampling Technique)

It is an oversampling technique based on the k nearest neighbor, judged by the Euclidean distance between data points in the feature space.



At the **Modeling** stage, algorithms are used to determine patterns in the data previously processed. As a result, several modeling techniques are selected and applied, and their parameters are calibrated to the optimum values. In this way:

1. Based on the defined objectives, modeling techniques should be selected for the previously prepared data set.
2. Some scenarios should be defined to test and verify the quality and validity of the model.
3. Finally, the models should be executed in the prepared data set.

SUPERVISED LEARNING

Classification or Regression

UNSUPERVISED LEARNING

Dimensionality Reduction or Clustering

SEMI-SUPERVISED LEARNING

Predictions in the medical field (tests and diagnostics are expensive and time consuming and only part of the population has them)

REINFORCEMENT LEARNING

Gaming, Finance Sector, Manufacturing, Inventory Management, Robot Navigation

1. Evaluation of the results achieved:

- Understand the results and verify their impact on the data mining objective initially defined;
- Verify the result against existing literature in order to see whether innovative and useful discoveries have been made;
- Draw relevant conclusions from the results achieved;
- Analyze whether there are new objectives that can be addressed in the future.

2. Review the data mining process to identify possible failures, neglected factors, changes in steps or unexpected options.

3. Refine the process and analyze the implementation potential.

SPLIT VALIDATION



K-FOLD CROSS VALIDATION



➔ The **Confusion Matrix** is a table with four different combinations of predicted and actual values.



		ACTUAL	
		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

➔ **Accuracy** measures the ability of the model to capture true positive as positive and true negative as negative. It can be a useful measure if there is the same number of samples per class, but if, on the contrary, the set of samples is unbalanced, the accuracy is not an adequate measure.



PREDICTED

Positives (1)

Negatives (0)

ACTUAL

Positives (1)

Negatives (0)

		Positives (1)	Negatives (0)
Positives (1)	TP	FP	
Negatives (0)	FN	TN	

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

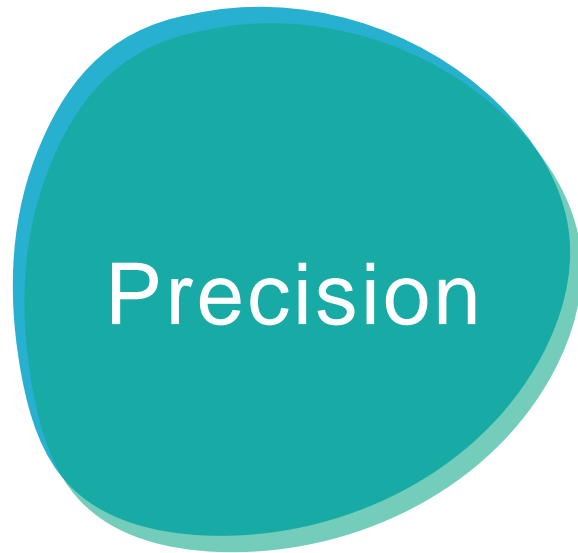
➔ **Classification Error** measures the number of instances incorrectly classified by the model, that is, the number of False Positives, also known as Type I error, and the number of False Negatives, also known as Type II error.



		ACTUAL	
		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

$$\text{Classification Error} = \frac{FP + FN}{TP + FP + FN + TN}$$

➔ **Precision** measures the accuracy of the model against the predicted positives and determines how many of them are actually positive. Precision is a good measure if the cost of False Positives is high (e.g.: SPAM detection).



$$Precision = \frac{TP}{TP + FP}$$

		ACTUAL	
		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

➔ **Recall** also called **Sensitivity** or **True Positive Rate** calculates how many of the true positives the model captures as being positive. Recall should be the metric to be use when there is a high cost associated with false negatives (e.g. medical diagnosis).



$$Recall = \frac{TP}{TP + FN}$$

		ACTUAL	
		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

➔ The **F1 score** is adequate when it is necessary to find a balance between Precision and Recall and when there is an uneven distribution of the class.



PREDICTED

Positives (1)

Negatives (0)

ACTUAL

Positives (1)

Negatives (0)

	ACTUAL	
	Positives (1)	Negatives (0)
Positives (1)	TP	FP
Negatives (0)	FN	TN

$$F1\ score = 2 * \frac{precision * recall}{precision + recall}$$

➔ **Specificity** or **True Negative Rate** calculates how many of the true negatives the model captures as being negative. Consider the example of a medical examination to diagnose a disease, the Specificity relates to the ability of the test to correctly reject healthy patients. A test with a higher Specificity has a lower error rate of Type I.



$$\text{Specificity} = \frac{TN}{FP + TN}$$

		ACTUAL	
		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

➔ **Fall-out** or **False Positive Rate** calculates how many false positives the model was unable to capture as being negative.



PREDICTED

Positives (1)

Negatives (0)

ACTUAL

Positives (1)

Negatives (0)

		Positives (1)	Negatives (0)
Positives (1)	TP	FP	
Negatives (0)	FN	TN	

$$FPR = \frac{FP}{FP + TN} = 1 - \text{specificity}$$

➔ **K statistic** is a measure of the reliability among evaluators and the discrepancy between them, taking into account the possibility that the agreement may occur by chance.



PREDICTED

Positives (1)

Negatives (0)

ACTUAL

Positives (1)

Negatives (0)

		Positives (1)	Negatives (0)
PREDICTED	Positives (1)	TP	FP
	Negatives (0)	FN	TN

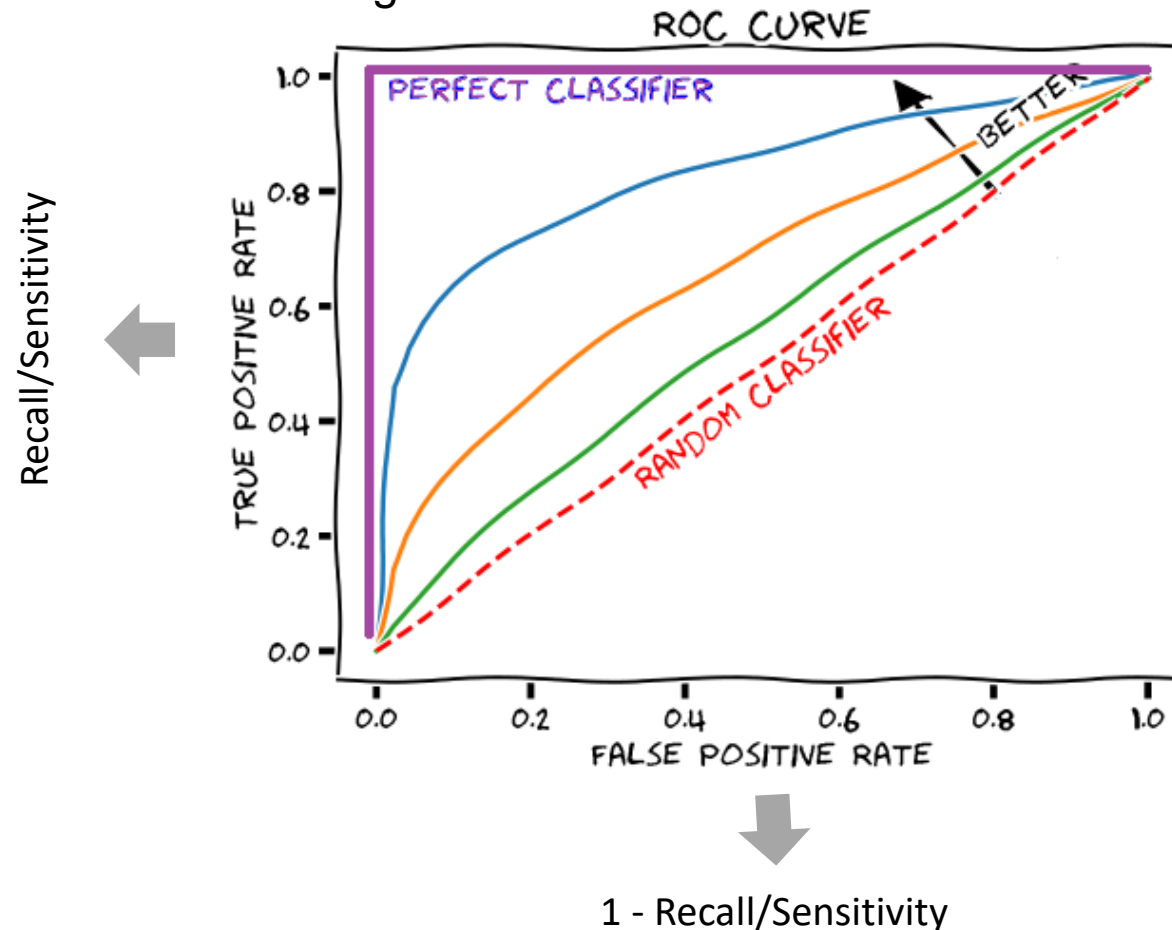
$$K \text{ statistic} = \frac{\text{accuracy} - p_e}{1 - p_e}$$

$$p_e = p_{yes} + p_{no}$$

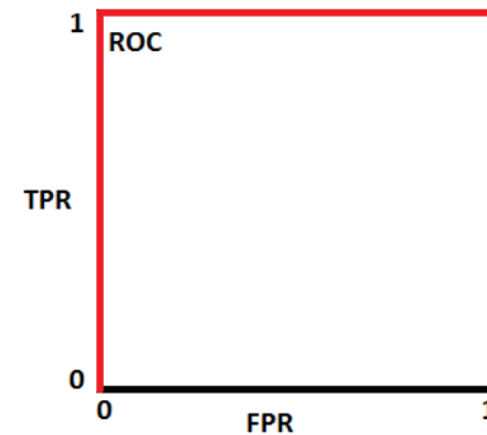
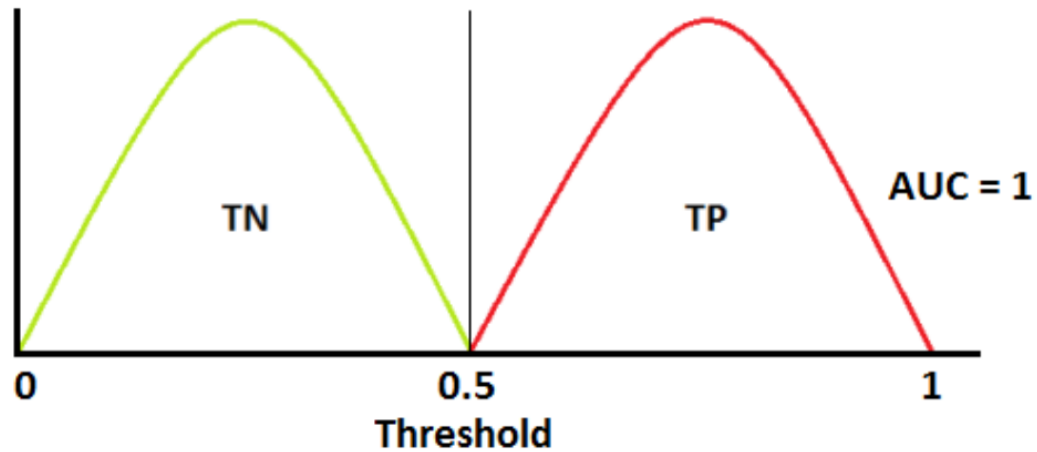
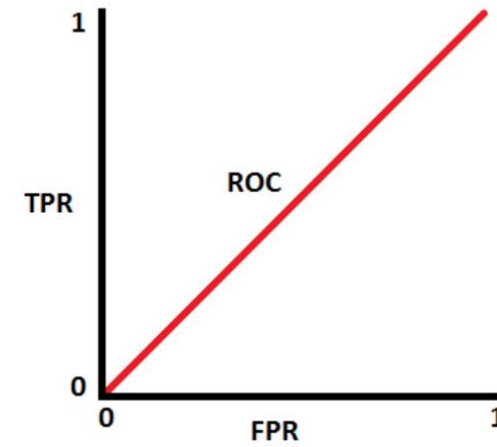
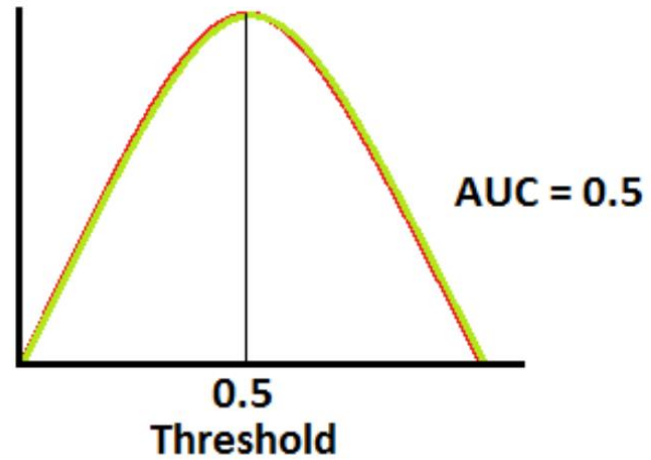
$$p_{yes} = \frac{TP + FP}{TP + FP + FN + TN} * \frac{TP + FN}{TP + FP + FN + TN}$$

$$p_{no} = \frac{TN + FN}{TP + FP + FN + TN} * \frac{TN + FP}{TP + FP + FN + TN}$$

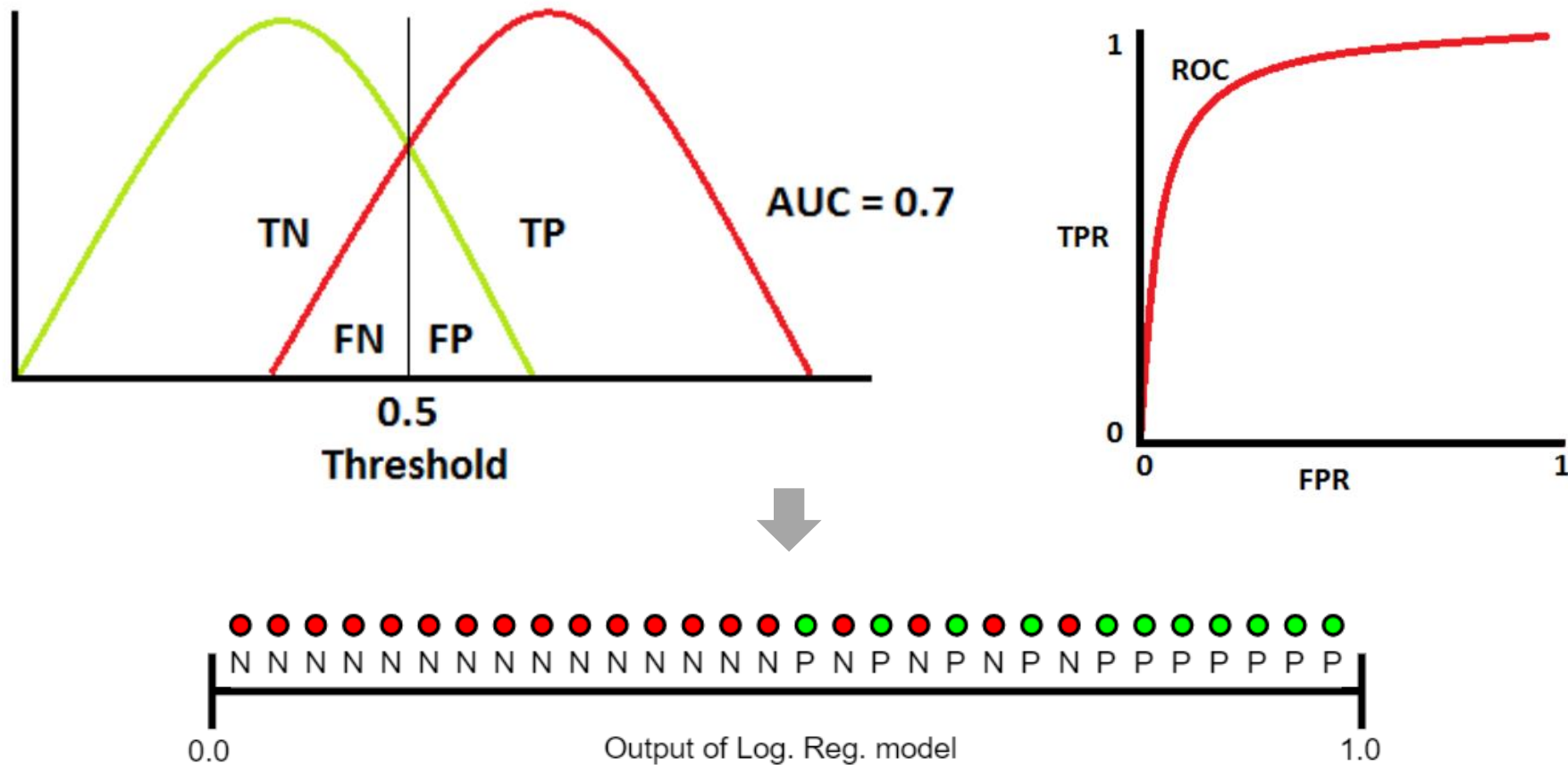
➔ **Receiver Operating Characteristic (ROC)** is a probability curve, and **Area Under the Curve (AUC)** is a separability measure that informs the ability of the model to distinguish classes. The higher the AUC, the better the model predicts 0s as being 0s and 1s as being 1s.



➔ Receiver Operating Characteristic (ROC)



➔ Receiver Operating Characteristic (ROC)



Deployment concerns the tactics to organize, present, and deploy the results of evaluation. Deployment can be as simple as generating a report or as complex as implementing a repeatable data mining process.

1. Implementation of the final models in a real environment.
2. Monitoring and maintenance of the Data Mining models.

- Machine Learning is a branch of Artificial Intelligence;
- Machine Learning is based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention;
- There are four types of Machine Learning models, namely Supervised, Unsupervised, Semi-supervised and Reinforcement Learning models.
- Machine learning and Data mining are rooted in Data Science but are different;
- Machine Learning is sometimes used as a means for carrying out useful Data Mining activities;
- CRISP-DM is a popular methodology used for increasing the success of a DM project and is composed by six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment;

- Business Understanding focuses on the definition of the project objective from a business perspective, then converting it into a DM problem definition;
- Data Understanding involves acquisition, analysis and exploration of data;
- Data Preparation involves data integration, cleaning, transformation, reduction, and sampling;
- Modeling consists in the application of different ML algorithms;
- Evaluation regards the assessment of the quality of the results obtained by the models and the verification of their impact on the DM objective initially defined. There are several evaluation metrics available to assess the models such as Recall, Accuracy, and AUC;
- Deployment concerns the implementation, monitoring and maintenance of the final models.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Zhang, X. D. (2020). Machine learning. In *A Matrix Algebra Approach to Artificial Intelligence* (pp. 223-440). Springer, Singapore. https://doi.org/10.1007/978-981-15-2770-8_6.

Sammut, C., & Webb, G. I. (Eds.). (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

Bonaccorso, G. (2017). *Machine learning algorithms*. Packt Publishing Ltd.

Langley, P. (1996). *Elements of machine learning*. Morgan Kaufmann.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Machine learning basics. *Deep learning*, 1, 98-164.



Diana Ferreira

- PhD student
in Biomedical Engineering
- Research Collaborator of the
Algoritmi Research Center

 [0000-0003-2326-2153](https://orcid.org/0000-0003-2326-2153)



Regina Sousa


- PhD student
in Biomedical Engineering
- Research Collaborator of the
Algoritmi Research Center

 [0000-0002-2988-196X](https://orcid.org/0000-0002-2988-196X)



José Machado

- Associate Professor with
Habilitation at the University of
Minho
- Integrated Researcher
of the Algoritmi Research Center

 [0000-0003-4121-6169](https://orcid.org/0000-0003-4121-6169)



António Abelha

- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

 [0000-0001-6457-0756](https://orcid.org/0000-0001-6457-0756)



Victor Alves

- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

 [0000-0003-1819-7051](https://orcid.org/0000-0003-1819-7051)

This Training Material has been certified according to the rules of **ECQA – European Certification and Qualification Association**.

The Training Material was developed within the international job role committee “**Artificial Intelligence Technician**”:

UMINHO – University of Minho (<https://www.uminho.pt/PT>)

The development of the training material was partly funded by the EU under Blueprint Project DRIVES.



Thank you for your attention

DRIVES project is project under **The Blueprint for Sectoral Cooperation on Skills in Automotive Sector**, as part of New Skills Agenda.

The aim of the Blueprint is **to support an overall sectoral strategy and to develop concrete actions to address short and medium term skills needs.**

Follow DRIVES project at:



More information at:

www.project-drives.eu



Co-funded by the
Erasmus+ Programme
of the European Union

The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.