MACHINE LEARNING

With content adapted from the thesis PhD “Desenvolvimento de Modelos Analíticos de Apoio à Gestão de Instituições do Ensino Superior, com Recurso a Data Mining”, of Maria Martins, 2020

Artificial Intelligence

- In an attempt to deal with, more and more, complex problems, it has long time been tried to incorporate into computer programs and other computer systems, comprehension and learning skills. Skills that are usually associated with human beings
  - *This remains, therefore, the main goal of Artificial Intelligence (IA)*
- By other hand, the machine learning (ML) has always presented itself as one of the main concepts and challenges of AI
  - *but it has been most in the present times that it has been imposed as an important area, of increasing interest, both in terms of research and in the field of its applicability*
- The main focus of this Curricular Unit is precisely to this important sub-area of AI
  - *The machine learning*

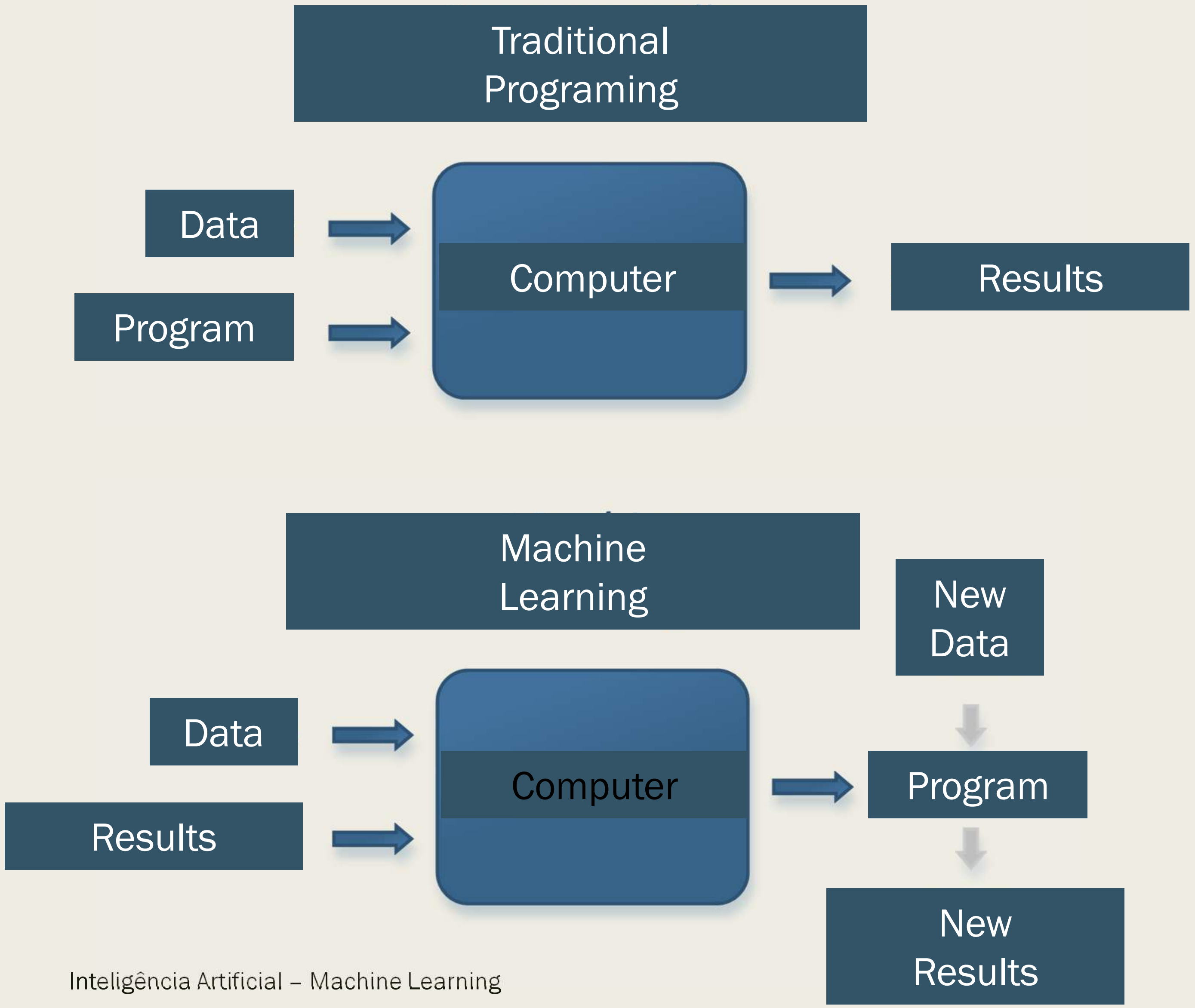
Machine Learning

- Ml's main objective is to give computers the ability to learn without being explicitly programmed
  - *In ML, algorithms we try to develop algorithmics with the ability to automatically adjust himself (improving their performance) from datasets that we provided to them*
  - *Learning is achieved through models with tuning parameters that are automatically adjusted according to different performance criteria*
- ML can then be understood as being the set of algorithms and techniques used in the development of systems that have the ability to learn from data,
  - *automatically being able to make predictions or recognize patterns in datasets that are provided to them*

A1

ML vs Traditional programming

- ML changes the programming paradigm we were used to



Diapositivo 4

A1 Autor; 17/09/2020

ML types

- ML, is typically divided into 3 large sub-areas:
  1. *Supervised learning: algorithms that learns from a set of examples that include their response (also called labeled examples), and then generalize their behavior to all possible inputs – these techniques will essentially be the target of our study.*
  2. *Unsupervised learning: algorithms that learns from unlabeled examples, trying to classify these same examples according to similarity criteria – we will study only one example of these technique.*
  3. *Reinforcement learning: algorithms that learns from the feedback they receive about the quality of the solutions they are producing (performance is being improved by iteratively exploring the solution space) – technique not addressed in our study.*
- ML techniques are commonly used to extract knowledge in large databases, integrated into a more comprehensive process known as Knowledge Discovery in Databases (KDD) and involving various processing phases.

Diapositivo 5

A2 Autor; 17/09/2020

A3

The KDD process

- The constant evolution of computer systems has led to organizations collecting, processing, and depositing large volumes of information from their daily activities
- This large and diverse volume of data, coming from multiple sources and available in heterogeneous configurations, makes it almost impossible to analyze
  - *the proportion of data that is analyzed compared to those that are generated will be very small;*
  - *information, potentially useful, present in that large volumes of data, are not properly used.*
- It was due to the need to design and develop a methodology that would allow the extraction of useful knowledge from large volumes of data that emerged the **KDD (Knowledge Discovery in Database)**
- Second Fayyad et al.[1], considered the main precursors of this area of research,
  - *"KDD is the iterative and iterative non-trivial, of valid, original, and potentially useful identification of understandable patterns in the data".*
  - *“Involving a large interdisciplinary relationship, KDD's unifying goal is to extract knowledge of high-level data from low-level data in the context of large data sets”.*

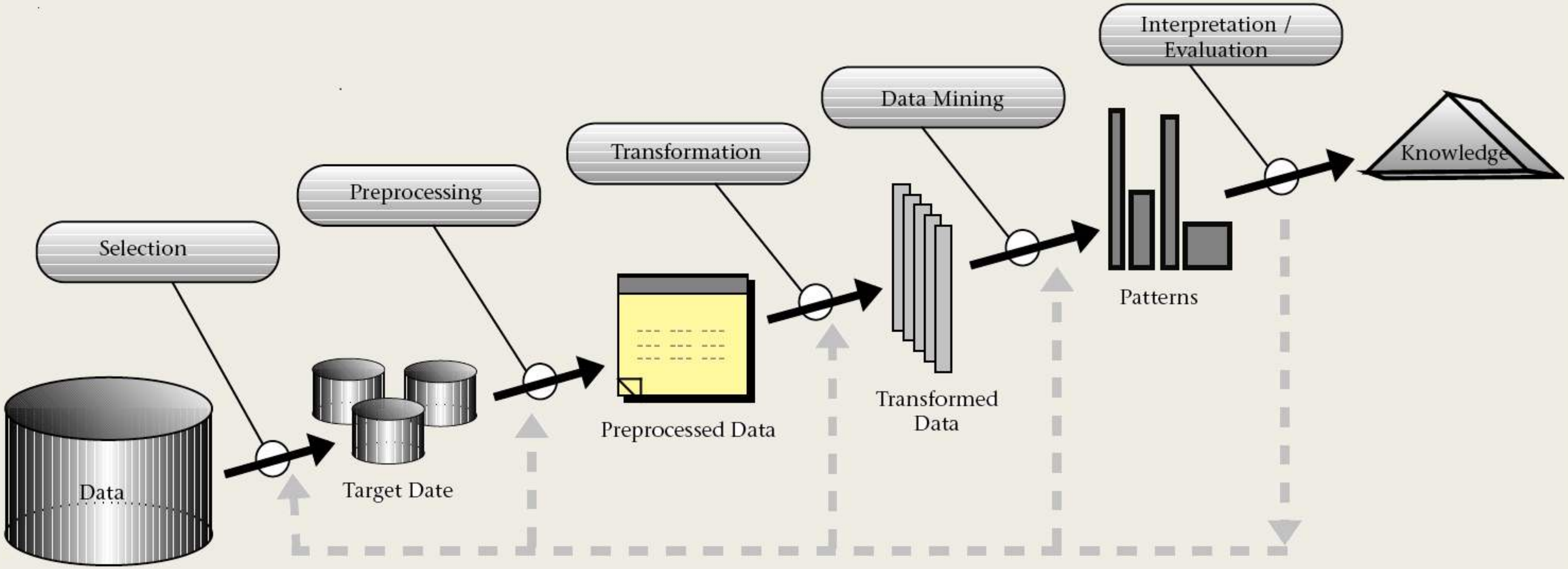
[1] *Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. AI magazine, 17(3):37, 1996*

Diapositivo 6

A3 Autor; 17/09/2020

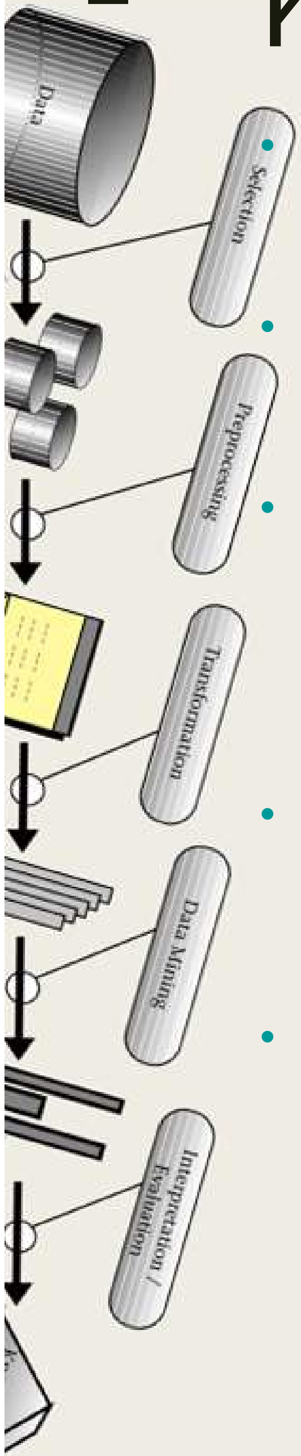
The life cycle of KDD

- The life cycle that characterises KDD comprises a sequence of phases and tasks associated with each other
  - *even if they have a trendily sequential order, one can go back to any of the phases, depending on the quality of the results that are obtained*



(Fayyad et al.)

KDD Phases



- Selection

  this phase Includes the selection of the correct data to be analysed, the most relevant attributes and the most appropriate period

- Preprocessing

  properly prepare the data by removing the noise and outliers (atypical values) present in the data and deciding what to do with missing values

- Transformation

  after the selection and preprocessing, the data are converted to the format that best appropriates the subsequent application of data mining algorithms, applying to them normalisation, aggregation or discretisation transformations, with the idea of reducing the adequate number of variables, looking for invariant representations for the data or other characteristics valuable for knowledge

- Data Mining (Data Prospecting)

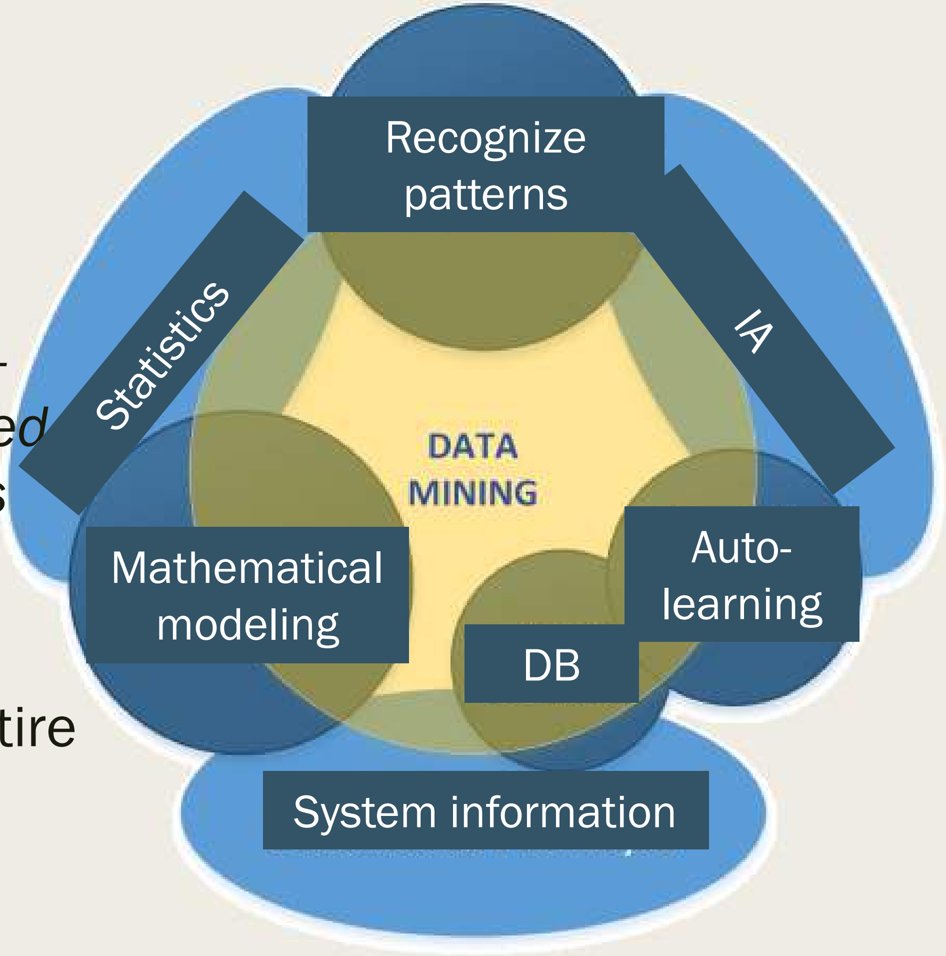
  application of robust methods and algorithms of automatic analysis (usually machine learning algorithms) that serve the purpose of prospecting — the discovery of patterns and relationships that exist in the data

- Interpretation/Evaluation

  interpretation of the discovered patterns and the evaluation of their usefulness for the intended application — it is in this final phase that the discovered knowledge is consolidated or not, and the need for new iterations is assessed, returning, if necessary, to any of the previous phases, for further improvements

The multidisciplinarity of Data Mining

- It is through the KDD process that raw data is transformed into valuable and understandable knowledge,
  - *this information, "hidden" in large-volume databases, is then retrieved through computational algorithms related to the Data Mining (DM) phase.*
- DM is the most critical phase of the entire KDD process
  - *Like The KDD itself, DM is also an interdisciplinary area*



(adaptada de Turban et al.[1])

[1] *Efraim Turban, Ramesh Sharda, and Dursun Delen. Decision support and business intelli- gence systems. Pearson Education India, ninth edition, 2011*

Data Mining with Machine Learning

- Even if we are talk about a methodology that involves interdisciplinary knowledge, Data Mining is closely linked to Machine Learning:
  - *"DM is a step in the global process of knowledge discovery, in which, by extracting, through the application of machine learning algorithms a set of patterns and relationships of the data, which allow to reveal, automatically or semi-automatically, implicit information that is present in large databases", Fayyad et al.*
  - *"DM is an interdisciplinary field that brings together techniques of machine learning, pattern recognition, statistics, databases and visualization, in order to be able to extract useful information from large databases", Cabena et al.*[1]
  - *"DM is a powerful artificial intelligence tool, capable of discovering useful information by analyzing data from many backgrounds or dimensions, categorizing this information and summarizing the relationships identified in databases", Algarni.*[2]

[1] Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi. Discovering data mining: from concept to implementation. Prentice-Hall, Inc., 1998

[2] Abdulmohsen Algarni. Data mining in education. International Journal of Advanced Computer Science and Applications, 7:456–461, 2016

Data Mining Models

- In essence, data mining aims to create analytical or computational models from large data sets.
- Depending on the domain of the problem they represent, the models can be categorised into two large groups:
  - *descriptive models*
    - the main objective is to find frequent patterns that can explain or generalise the intrinsic structure of the data, including its relationships.

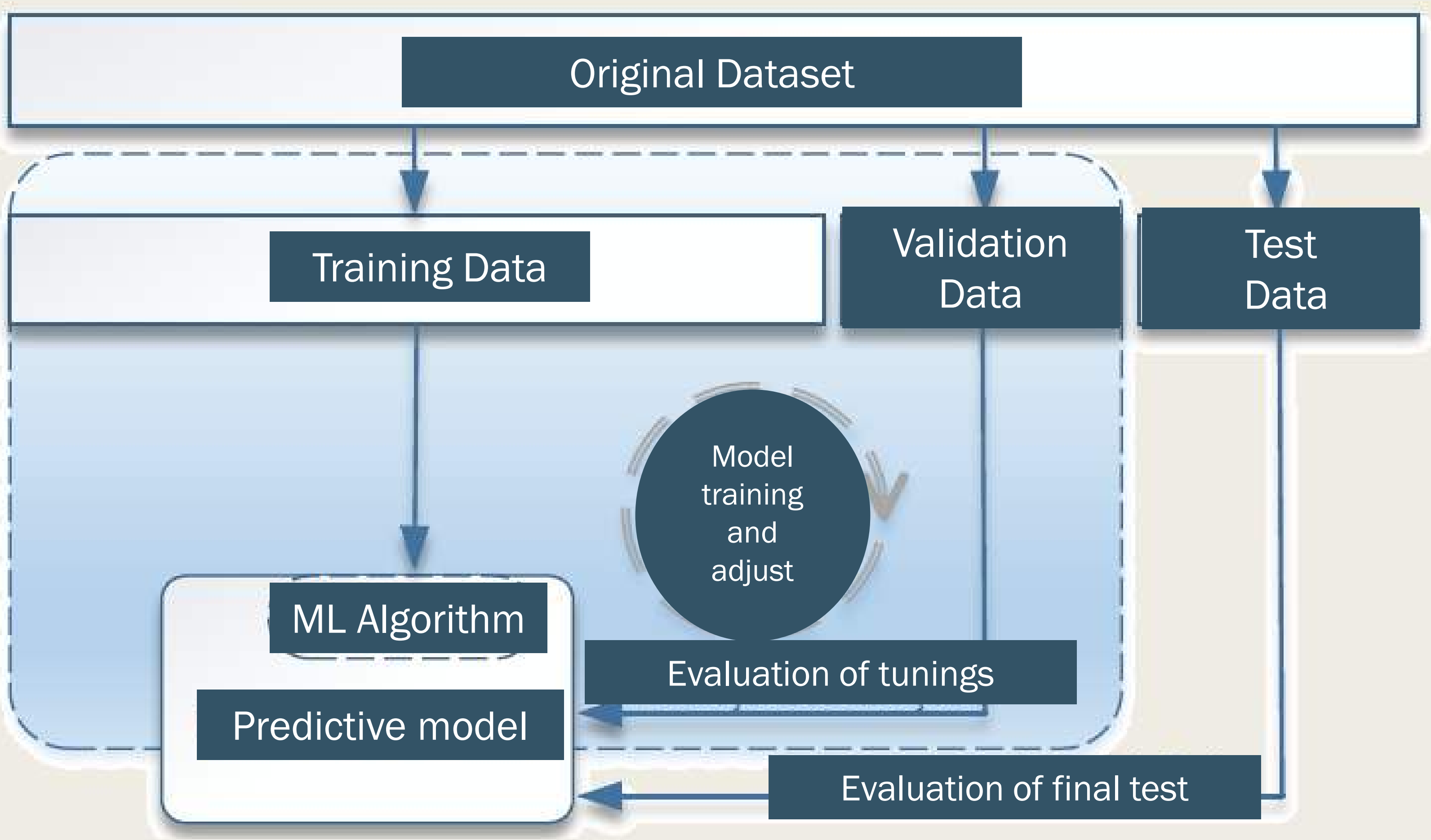
      (patterns can be, for example, presence of anomalies, trends, groupings between objects, associations, and correlations between variables)
  - *predictive models*
    - the main objective is to estimate values, unknown or future, of one or more variables that we consider with interest from some combination of other characteristics present in the data.

      (therefore, they allow for future circumstances, such as specific trends and certain behaviours)
      - *the variable for the prediction is designated target variable, response variable, or dependent variable (VD),*
      - *the attributes used for the prediction are designated independent, predictive, or explanatory variables.*

Predictive model

- Predictive models may also provide some description of the problem and descriptive models reveal predictions of upcoming events
    - *The distinction between one and the other is very little*

      (usualy, models whose representation is not easily interpretable will be used for prediction; on the other hand, predictive models whose representation arises in the form of rules, or in some interpretable structure can also be used to describe the data)
- The following diagram, illustrative the entire process of inducing a predictive model



Induction of the predictive model

- For the model to be induced,
  - *the initial dataset is typically divided into three subsets: training, validation, and the testing;*
  - *the training set is processed by a Machine Learning algorithm, serving the second subset to adjust the hyperparameters responsible for small adjustments (re-configuring) on the algorithm, in a normally interactive process in which the model is evaluated with validation data;*
  - *finally, finished the training and adjust of the model, the data from the test subset are showing for the first time (and last) to the model founded, in order to evaluate its true predictive capacity and in particular its generalization capacity*
- If the model shows high accuracy in the data set left for testing, it is assumed that the model has effectively good generalization capacity, that is, it is expected to perform well with future and unknown data
  - *on the other hand, if the model performs poorly with the test set data, then the model we create will not be suitable for forecasting*

    (given this second situation, it is usually to return to the pre-processing phase to improve the data, or simply change to another learning algorithm)

Better understanding the types of learning

- As mentioned already, depending on the kind of information that is made available to the model, it will be based on a Supervised or Unsupervised Learning algorithm
  - *In supervised learning, the model is trained with data that include the response variable, trying to understand (learn) the relationship between this variable and the other variables (predictive variables)*
  - *In unsupervised learning, dataset instances are only characterised by input attributes (predictive variables), and there is no information about the value of the response variable associated with each example*
    - the learning of the model is carried out by discovering similarities in the data, forming groups of data with similar characteristics
- In any learning, whether or not supervised, one of the main objectives of data mining is the creation of models with generalisation capability
  - *it is a concept that reflects the ability of a model to accurately predict new examples not yet observed after being constructed only based on a learning data set*
  - *it is considered to be* overfitting *when the model adjusts (becomes too addictive) to the training data, thus compromising its generalisation capacity*
    - and underfitting when not even with the training data, the model achieves a good performance

Classification vs Regression

- The methods of predictions can be classified essentially into two different typologies: classification and regression.
  - *The main difference between them lies in the typology of the target variable for the prediction:*
    - in the classification the target variable is categorical,
    - in the regression is numerical, assuming continuous values.
- Regression Methods
  - *The models aim to predict the future values, or unknown ones, of one or more continuous numerical variables, starting from other attributes present in the dataset.*
- Classification Methods
  - *Classify is the action of categorizing a particular object/instance according to its own characteristics*
    - The idea is that the generated model, assign to each of one objects/instances the predefined categories (labels), depending on their explanatory variables

Evaluation metrics

Whatever the predictive model, it is always necessary to evaluate how accurate its performance is

Some of the most commonly used predictive performance assessment metrics

- *in regression problems are:*
  - Mean quadratic error (MSE) or root of that value (RMSE)
  - Coefficient of determination (R2)
- *in classification problems are:*
  - Confusion Matrix
  - A diverse set of measures that are obtained from the values that integrates the Confusion Matrix
    - *True positive rates, true negatives, false positives, false negatives, accuracy...*
  - Receiver Operating Characteristic (ROC) curve and the area below the ROC curve, called AUC value

Metrics for regression

- *R2 (coefficient of determination)*

  Measures the correlation between the observed (real) values and the predicted (estimated) values - ranges from 0 to 1

- MSE (mean square error)

  Measuares the predicts deviation from the effective value, calculating the average of the squares of the distances between the knowed values y_i and the predicted values $\hat{f}(x_i)$ by the model:

$$\mathtt{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

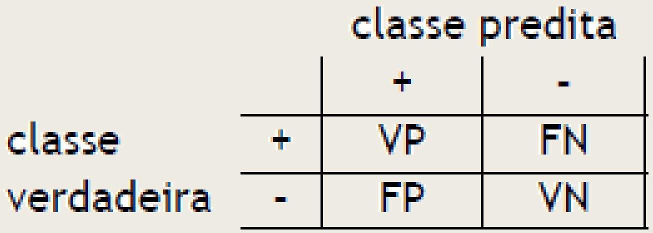
- RMSE (*root mean square error*)

  square root of the MSE (error measure more used than the previous one, since it gives a more direct indication of the true deviation)

Confusion Matrix - to evaluate classifiers

This is a table that illustrates the number of correct and incorrect predictions in each class (category), positive (+) or negative (-)

- For a given dataset, the array rows represent the true classes (categories), and the columns represent the classes predicted by the classifier



		classe predita	
		+	-
classe verdadeira	+	VP	FN
	-	FP	VN

- True Positives (VP): Represents the number of positive predictions that are correct;
- False Positives (FP): Represents the number of positive predictions that are incorrect;
- False Negatives (FN): Is the number of negative predictions that are incorrect;
- True Negatives (VN): is the number of negative predictions that are correct.

- Therefore, each element m_{ij} of a confusion matrix presents the number of examples from the class i classified as belonging to that class j.
- For k classes the confusion matrix would have the dimension k × k.
- The diagonal displays the cases correctly classified by the model, while the other elements correspond to the errors made in its predictions

Other metrics for classification

From the values of the Confusion Matrix, several other measures are considered

- False Negative rate (TFN)

  represents the error rate in the positive class. It is a measure of the proportion of positive class examples incorrectly classified by the predictor

$$\mathsf{TFN} = \frac{\mathsf{FN}}{\mathsf{VP} + \mathsf{FN}}$$

- False Positive rate (TFP)

  represents the error rate in the negative class. It is a measure of the proportion of negative class examples incorrectly classified by the predictor

$$\mathsf{TFP} = \frac{\mathsf{FP}}{\mathsf{FP} + \mathsf{VN}}$$

- Error rate

  represents the percentage of incorrect classifications of total n examples, regardless of the direction of the error

$$\mathsf{erro} = \frac{\mathsf{FP} + \mathsf{FN}}{n}$$

Other metrics for classification

- Hit or accuracy rate

  represents the percentage of correct classifications of total n examples, regardless of the direction of the hit. It is calculated by the sum of the values of the main diagonal of the matrix, divided by the sum of the values of all elements on the matrix (n)

$$\text{acurácia} = \frac{\text{VP} + \text{VN}}{n}$$

- Precision

  represents the rate of success among the examples classified by the predictor as positive

$$\text{precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

- True Positiverate (DVT) or Sensitivity (Recall)

  represents the proportion of positive examples that were correctly classified by the predictor

$$\text{sensibilidade} = \text{recall} = \text{TVP} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

Other metrics for ranking

- Rate of True Negative (TVN) or Specificity

  corresponds to the rate of correct answer in the negative class, that is, the proportion of negative examples that were correctly classified by the predictor. The complement corresponds to the False Positive rate (TFP)

$$\text{especificidade} = 1 - \text{TFP} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

- F-measure

  is the harmonic mean of accuracy and sensitivity. It is a single measure that values the mistakes made in any way (FP and FN)

$$\text{Fmedida} = \frac{(w+1) \times \text{precisao} \times \text{sensibilidade}}{w \times \text{precisao} + \text{sensibilidade}}$$

- F1

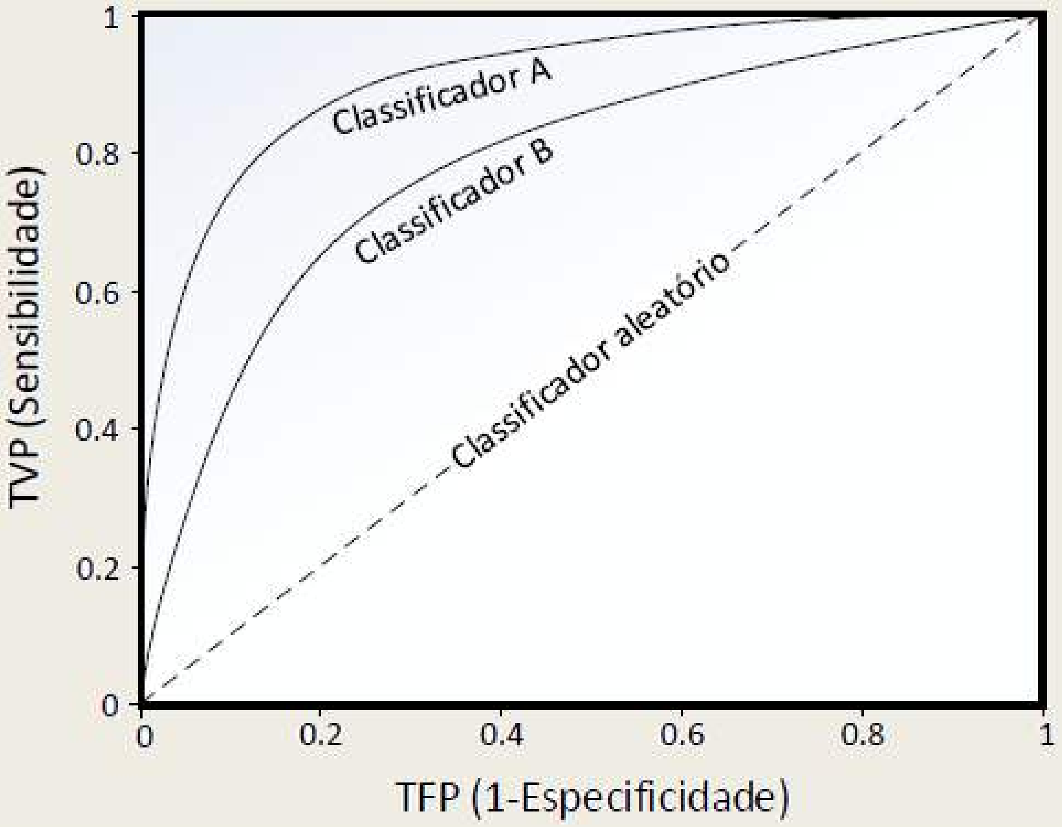
  If errors are valued equally, the F-measured harmonic mean is no longer simple (w = 1), assuming the form it normally called F1

$$\text{F1} = \frac{2 \times \text{precisao} \times \text{sensibilidade}}{\text{precisao} + \text{sensibilidade}}$$

ROC Curves

- Yet another way of evaluating algorithms is through Receiver Operating Characteristic (ROC) curves
- A ROC curve is a graph that illustrates the performance of a binary classification model through the variation of the threshold of discrimination between positive and negative elements

- One classifier is considered better than another, if its curve in the ROC space is positioned above and to the left of the curve corresponding to the second classifier
- When comparing two or more curves, if they do not intersect, the one closest to the point (0.1) corresponds to the best performance.
  - *In case intersections occur, each algorithm has a region with better performance.*



AUC Value

- It is also usually to measure the performance of a classifier in terms of a single measure extracted from its ROC curve:

  *the **area below the ROC curve,** called AUC (Area Under Curve).*

- The AUC measure, produces values between 0 and 1.
  - *Values closer to 1 are considered better, i.e., the larger the area under the ROC curve, the higher the accuracy of the algorithm.*

In conclusion...

- Each of the metrics presented quantitatively evaluates (with the exception of the ROC curve, which is qualitative) a model, providing information about the effectiveness of the learning method.
  - *None of them, completely replace all the others, so it is common to use several at the same time*

  For example, the error and the accuracy are simple measures that, in the face of unbalanced problems, do not allow to evidence the difference between false positives and false negatives, leading to illusory results.

  In this case, when the number of examples of each class is very different, it is recommended to use metrics that emphasize both error measurement, FP and FN, combining them, such as f-measure and the AUC value itself.

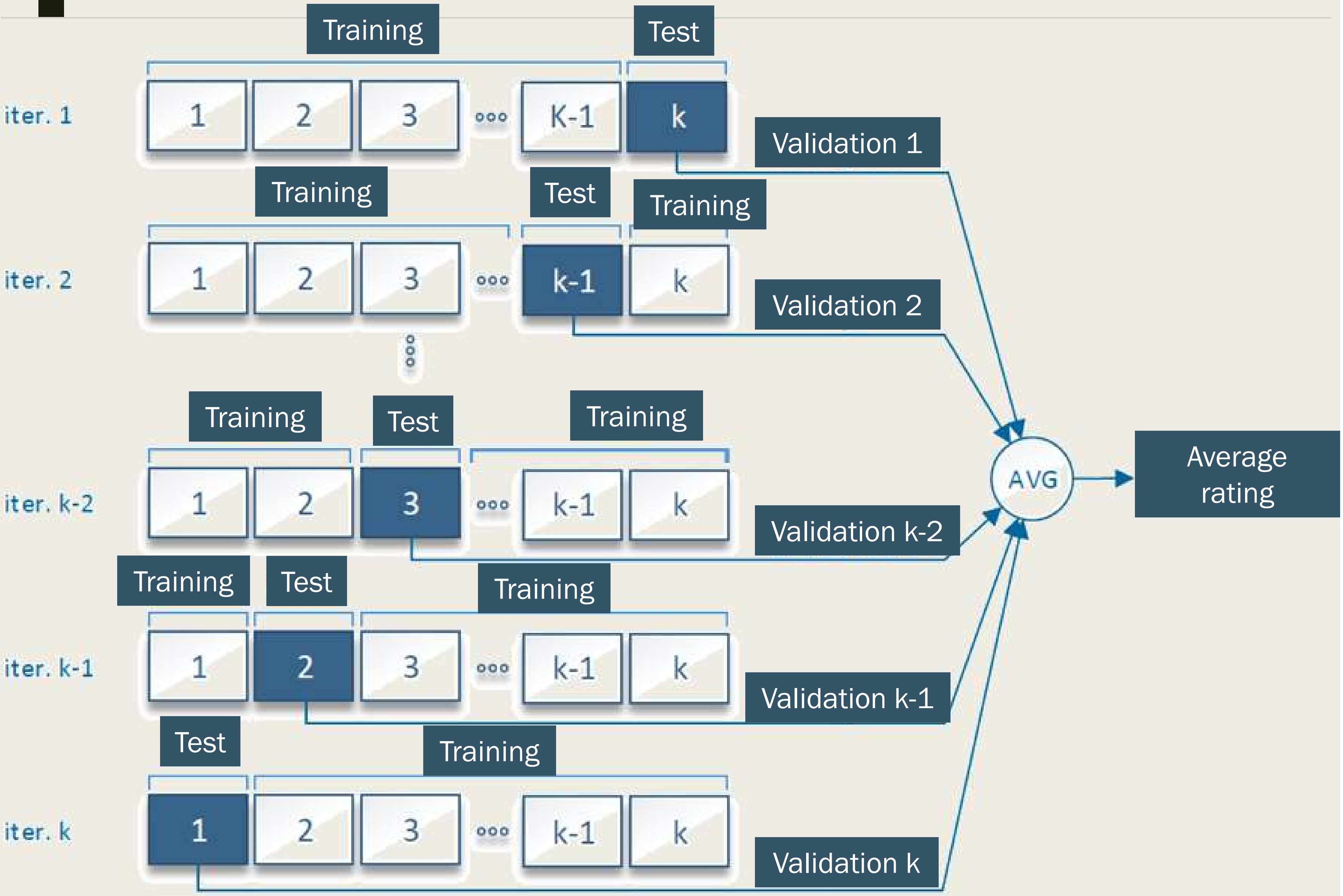
- AUC value is one of the most appreciated metrics used by data mining analysts, in part because it handles relatively precisely with unbalanced datasets.

Cross-Validation K-folds

The data mining literature suggests that the previous metrics are calculated by the cross-validation method.

- Cross-validation is a widely used technique when we want to obtain more stable and reliable predictive performance results,
  - *especially indicated for the adjust phase of the models (tuning)*
- In the K-folds cross-validation method, the initial dataset is divided into k partitions (subsets), approximately with the exact sizes
  - *Then we take each of the k partitions:*
    - and We use this partition for testing and all the rest for training
  - *The final performance of the model is obtained by the average of the observed performances on each subset of the test,*
    - thus achieving an estimate of performance that is considered more consistent
- One of the most commonly used values for the number of partitions is K=10.
- This technique is inconvenient, about simple partitioning, to increase computational effort

Cross-Validation K-folds



solve exercise #18

from the book of exercises