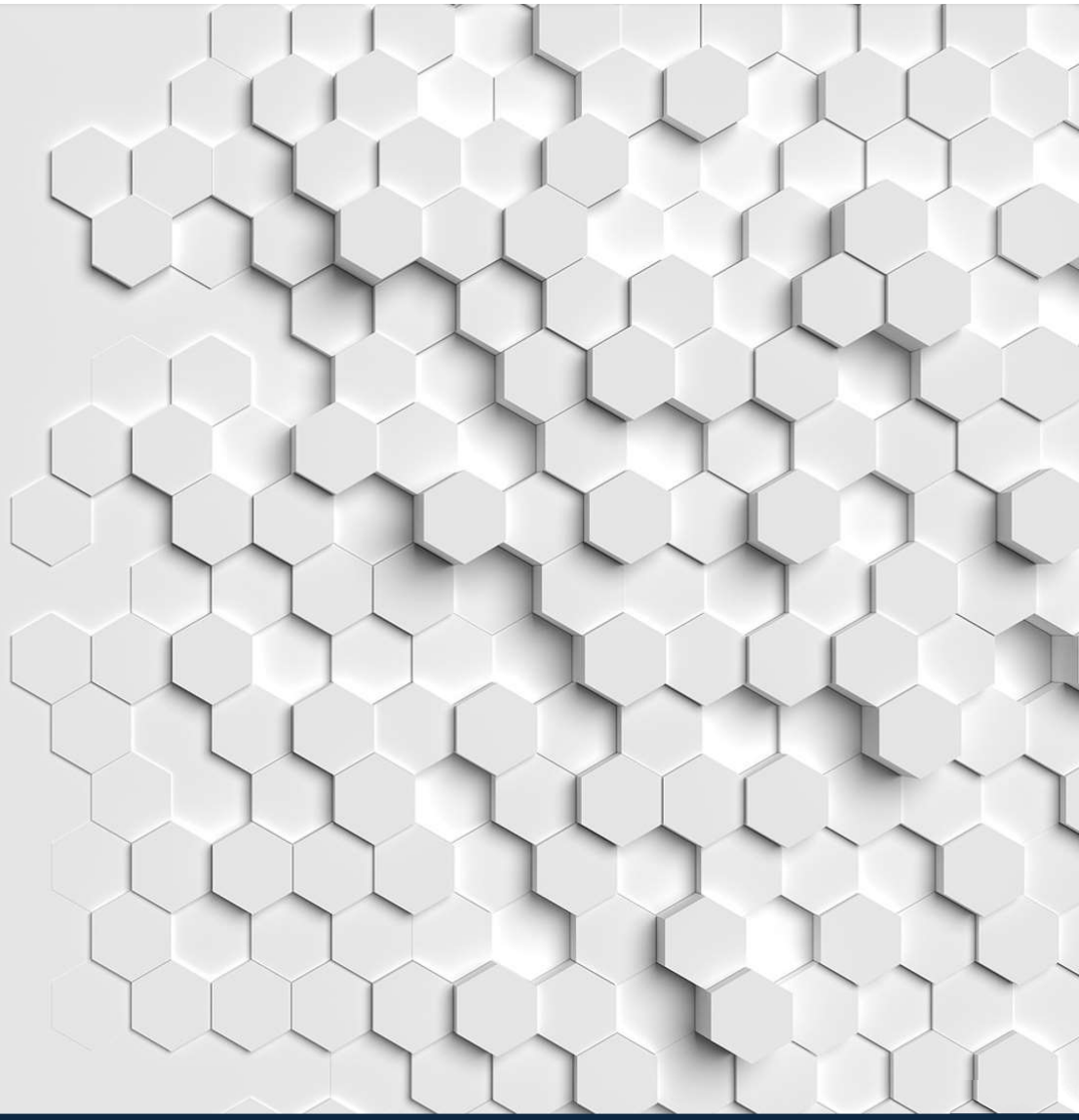
Statistical Computing

Regression Tree & KNN| August , 2022

Emmanuelle Rodrigues Nunes

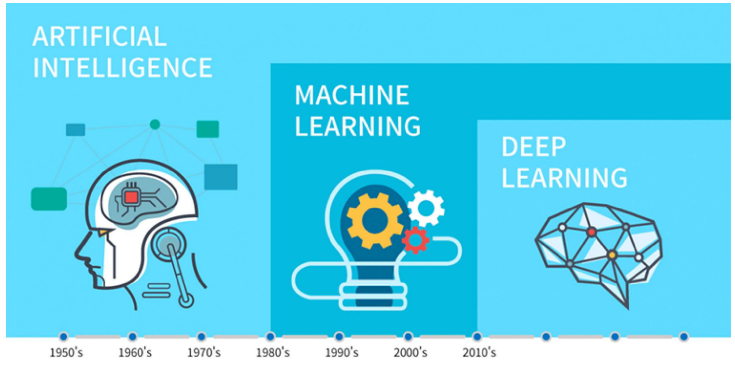


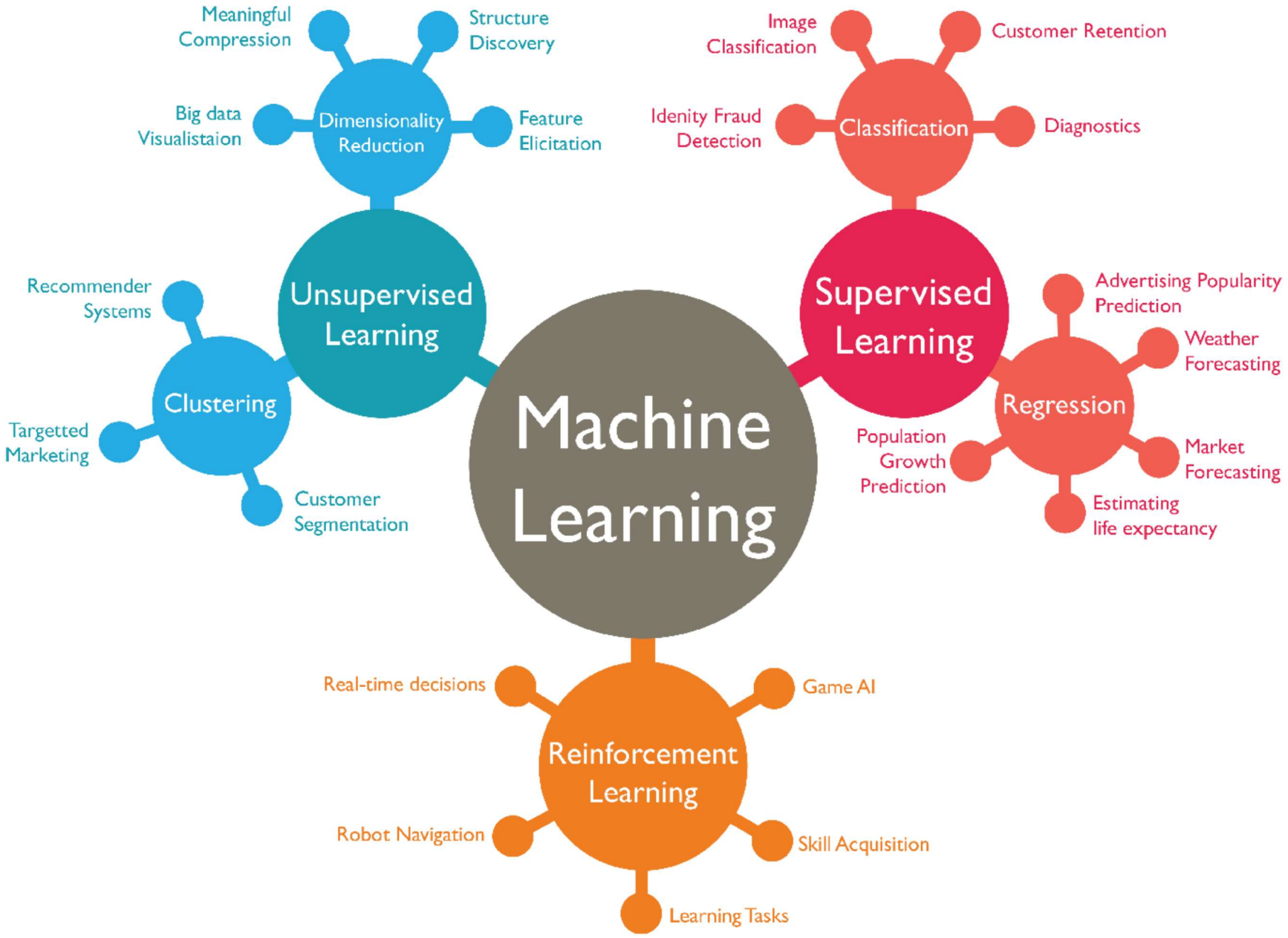
CIDRA-AIPD
African Institute for Professional Development

Machine Learning

Machine Learning - Concepts

- Machine Learning (ML) is a subset of Artificial Intelligence (AI)
- Algorithms that can improve automatically through experience and by the use of data without being explicit programmed, reason why we say that the algorithms learn.
- With ML algorithms we can build a model to make predictions or decisions.
- Machine learning algorithms are used many different applications, for example:
  - Medicine
  - Email filtering
  - Speech recognition
  - Computer vision




Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

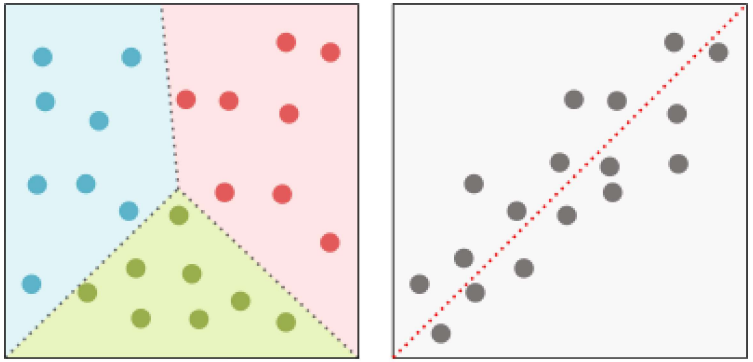
Supervised Learning

Supervised Learning

Supervised learning is where you have input variables (X) and an output variable (y) and you use an algorithm to learn the mapping function from the input to the output.

$$y = f(X)$$

- It is the most common type of Machine Learning problem
- It is called **supervised** because we have the label that tell us the correct information, and we are going to be corrected if we predict wrong.
- Supervised learning can be grouped into two problems:
    - **Regression:** The output variable is a real number, for example, weight
    - **Classification:** The output variable is a category, for example, disease and no disease



Classification

Classification

Classification is a type of **supervised** learning where we categorise data into classes. There are many different algorithms that can help us solve this kind of problems.

Classification requires a training dataset with many examples of inputs and outputs from which to learn. It can be categorised in two types of problems:

- **Binary classification:** The outcome has only **two** labels, for example, disease and not disease.
  - Some popular algorithms are: Logistic Regression, Decision Tree, K-Nearest Neighbour (KNN)
- **Multi-label classification:** The outcome has multiple labels, for example, dog, cat, bird and other.
  - Some popular algorithms: KNN, Decision Tree, Random Forest, Naive Bayes

To evaluate the model performance we can make use of ROC, confusion matrix, etc. We need to be aware of **class imbalance** problems.

Decision Tree (CART)

Decision Tree

There are various algorithms that can grow a tree.

- Differences:
  - Possible structure of the tree (e.g. number of splits per node)
  - Criteria how to find the splits
  - Criteria to stop splitting
  - How to estimate the simple models within the leaf nodes.

The **Classification and Regression Trees (CART)** algorithm is probably the most popular algorithm for tree induction.

- We will focus on CART, but the interpretation is similar for most other tree types.

> Note: Decision Trees can be used for both Regression and Classification problems

Theory

The processes behind *classification* and *regression* in tree analysis is very similar, but we need to first distinguish the two.

- **Classification:** For a response variable which has *classes*, we want to organize the dataset into groups by the response variable.
- **Regression:** When our response variable is instead numeric or continuous we wish to use the data to predict the outcome, and will use regression trees in this situation.

Essentially, a classification tree splits the data based on homogeneity by categorizing the data based on similarity, filtering out the "noise" and making the data "pure", hence the concept of **purity criterion**.

When the response variable does not have classes, a regression model is fit to each of the independent variables, isolating these variables as nodes where their inclusion decreases error.

<!--

--> <!-- -->

Theory

CART takes a feature and determines which cut-off point minimizes:

- The variance of Y for a regression task
  - The variance tells us how much the y values in a node are spread around their mean value
- The Gini index of the class distribution of Y for classification tasks
  - The Gini index tells us how "impure" a node is, e.g. if all classes have the same frequency, the node is impure, if only one class is present, it is maximally pure.

Variance and Gini index are minimized when the data points in the nodes have very similar values for Y. As a consequence, the best cut-off point makes the two resulting subsets as different as possible with respect to the target outcome.

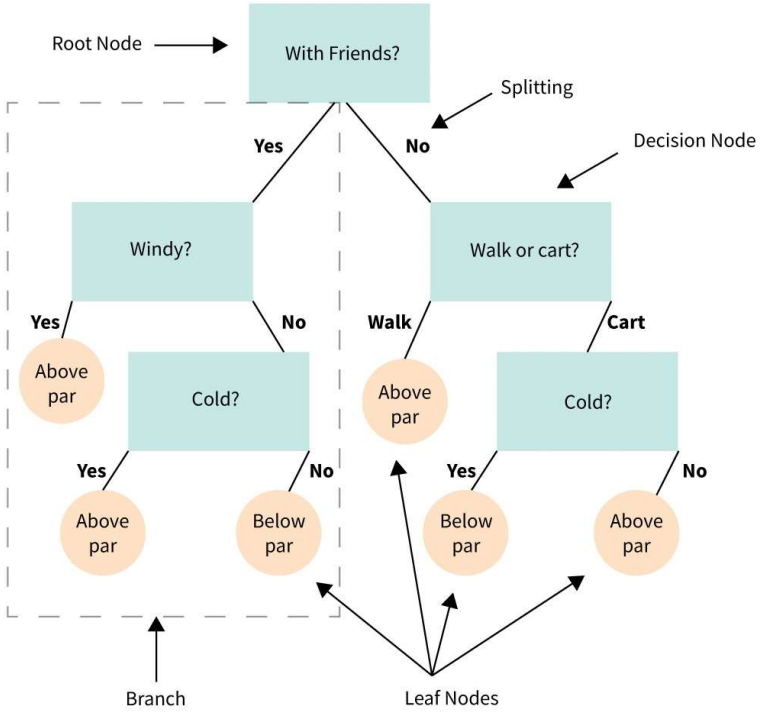
For categorical features, the algorithm tries to create subsets by trying different groupings of categories. After the best cutoff per feature has been determined, the algorithm selects the feature for splitting that would result in the best partition in terms of the variance or Gini index and adds this split to the tree. The algorithm continues this search-and-split recursively in both new nodes until a stop criterion is reached. Possible criteria are: A minimum number of instances that have to be in a node before the split, or the minimum number of instances that have to be in a terminal node.

Interpretation

Let's first define some keys terms:

- **Root node:** The base of the decision tree.
- **Splitting:** The process of dividing a node into multiple sub-nodes.
- **Decision node:** When a sub-node is further split into additional sub-nodes.
- **Leaf node:** When a sub-node does not further split into additional sub-nodes; represents possible outcomes.
- **Pruning:** The process of removing sub-nodes of a decision tree.
- **Branch:** A subsection of the decision tree consisting of multiple nodes.

Starting from the root node, you go to the next nodes and the edges tell you which subsets you are looking at. Once you reach the leaf node, the node tells you the predicted outcome. All the edges are connected by 'AND', so we are conditioning on the variable.



For example, if you are with friends *AND* it is windy, *THEN* it is above par.

References

Majid (2013)