SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with ‘A++’ Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

23AMB201 - MACHINE LEARNING

II YEAR IV SEM

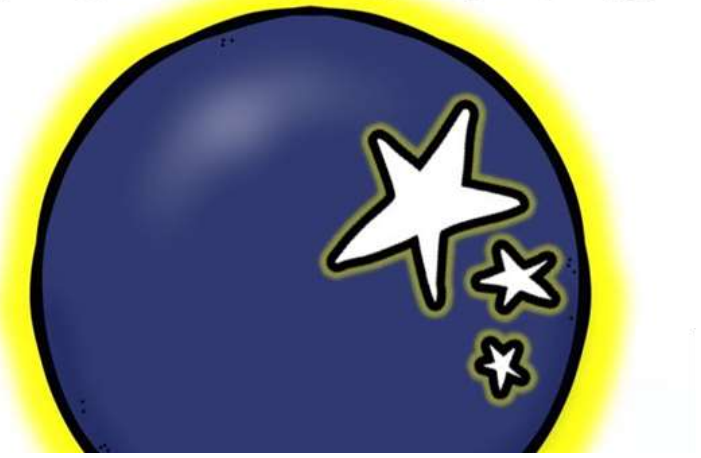
UNIT I – INTRODUCTION

TOPIC 2– Machine Learning process- Preliminaries

Recall

DATA

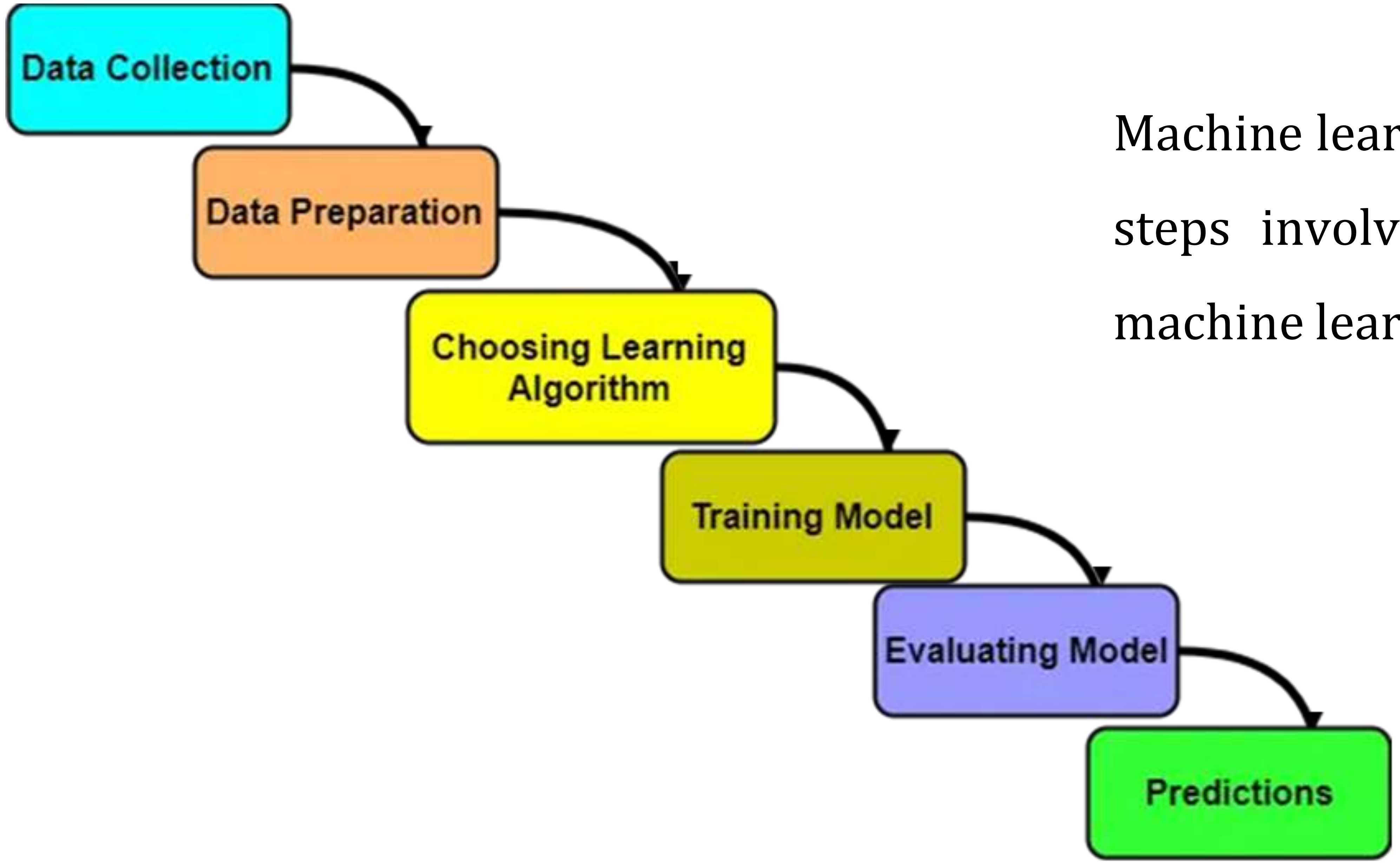
PREDICTIONS



KNOWLEDGE

Types

Machine Learning Process



Machine learning workflow refers to the series of stages or steps involved in the process of building a successful machine learning system.

Data Collection

1. Data is collected from different sources.
2. The type of data collected depends upon the type of desired project.
3. Data may be collected from various sources such as files, databases etc.
4. The quality and quantity of gathered data directly affects the accuracy of the desired system.

kaggle

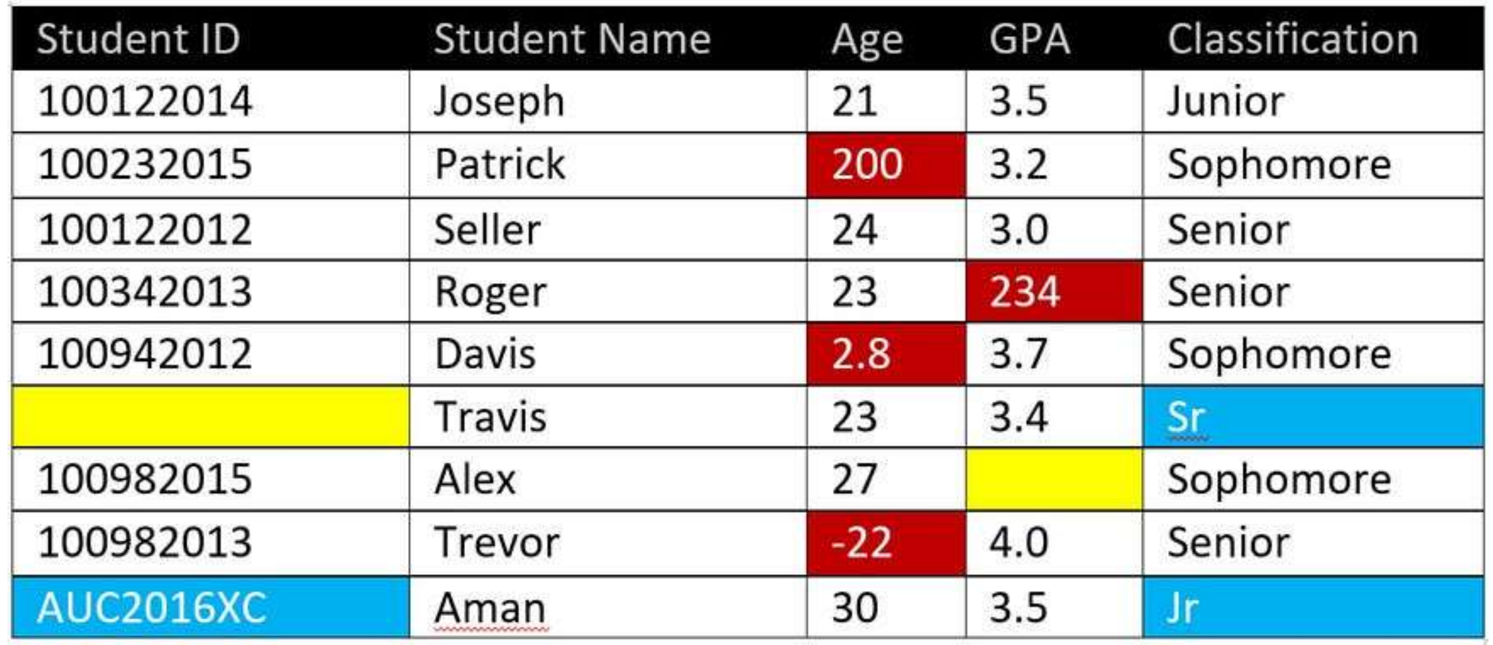
MAVEN ANALYTICS

UC Irvine

tableau SOFTWARE

Data Preparation/Cleaning

1. Data preparation is done to clean the raw data.
2. Data collected from the real world is transformed to a clean dataset.
3. Raw data may contain missing values, inconsistent values, duplicate instances etc.
4. So, raw data cannot be directly used for building a model.



Student ID	Student Name	Age	GPA	Classification
100122014	Joseph	21	3.5	Junior
100232015	Patrick	200	3.2	Sophomore
100122012	Seller	24	3.0	Senior
100342013	Roger	23	234	Senior
100942012	Davis	2.8	3.7	Sophomore
	Travis	23	3.4	Sr
100982015	Alex	27		Sophomore
100982013	Trevor	-22	4.0	Senior
AUC2016XC	Aman	30	3.5	Jr

Missing Data

Inconsistent Data

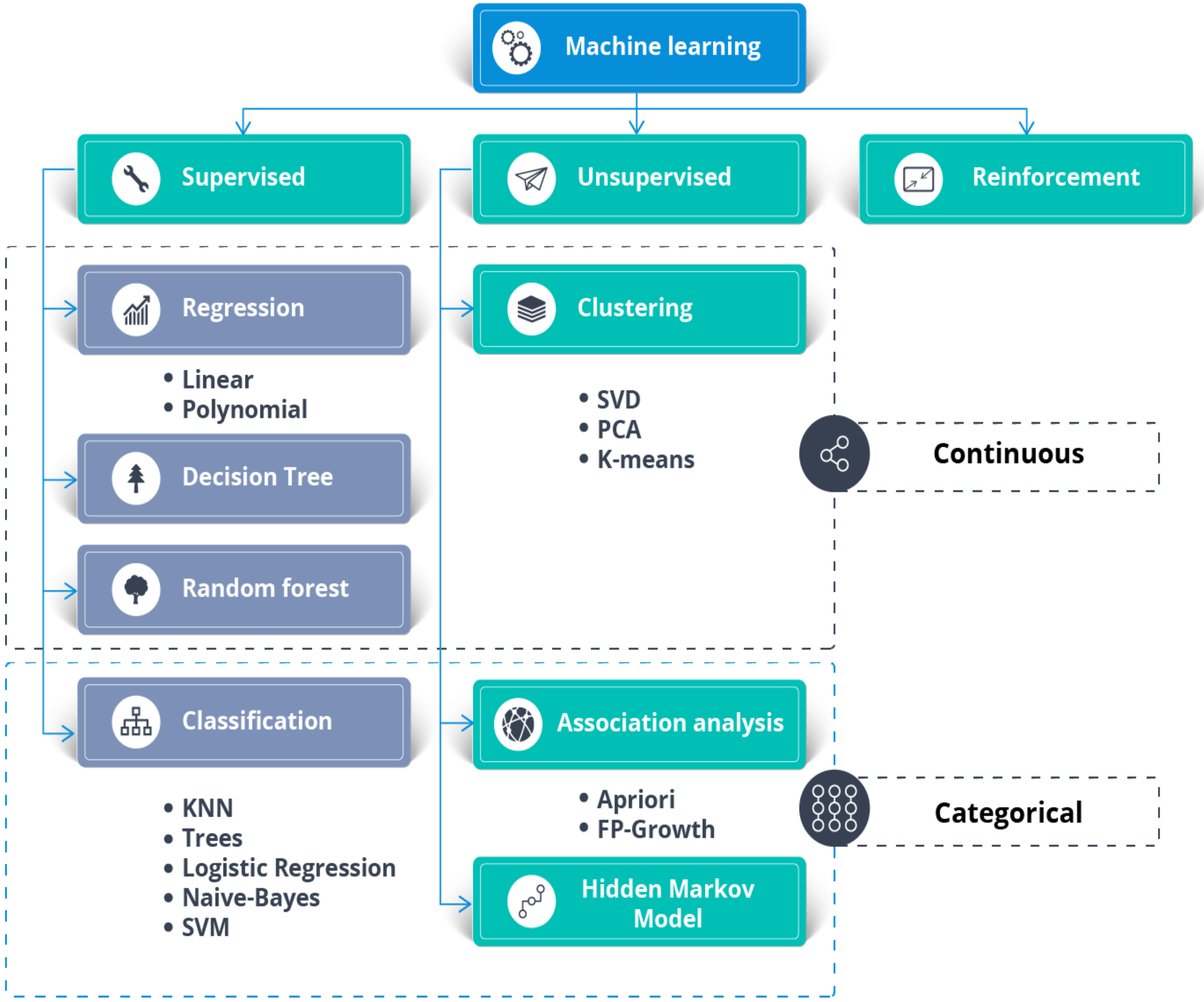
Noisy Data

Methods to remove noise

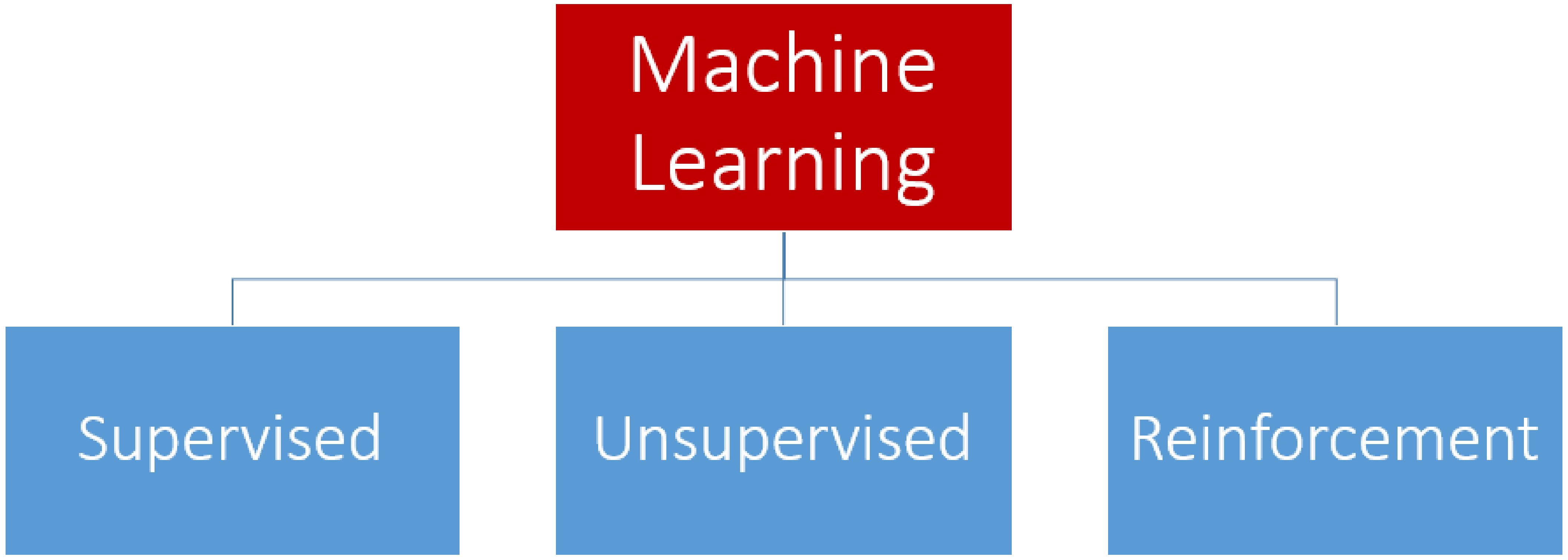
1. Ignoring the missing values
2. Removing instances having missing values from the dataset.
3. Estimating the missing values of instances using mean, median mode.
4. Removing duplicate instances from the dataset.
5. Normalizing the data in the dataset.

Learning Algorithms

1. The best performing learning algorithm is researched.
2. It depends upon the type of problem that needs to solved and the type of data we have.
3. If the problem is to classify and the data is labeled, classification algorithms are used.
4. If the problem is to perform a regression task and the data is labeled, regression algorithms are used.
5. If the problem is to create clusters and the data is unlabeled, clustering algorithms are used.



Learning Algorithms: Use case



Machine Learning

Supervised

Unsupervised

Reinforcement

(Learning by examples)

Classification

Disease detection
Email spam detection
Image classification
Bank loan prediction

Regression

House price prediction
Stock price prediction

(Learning by observation)

Clustering

Search engines
Face recognition
Targetted marketing
Recommender system

Association Rule Mining

Market basket analysis
Medical diagnosis
Census data

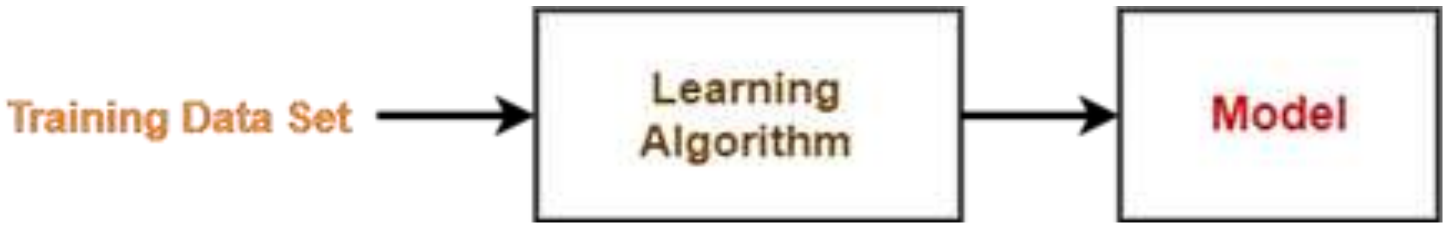
(Learning from mistakes)

Example

Self-driving car
Gaming AI
Robot navigation
Inventory management
Finance sector

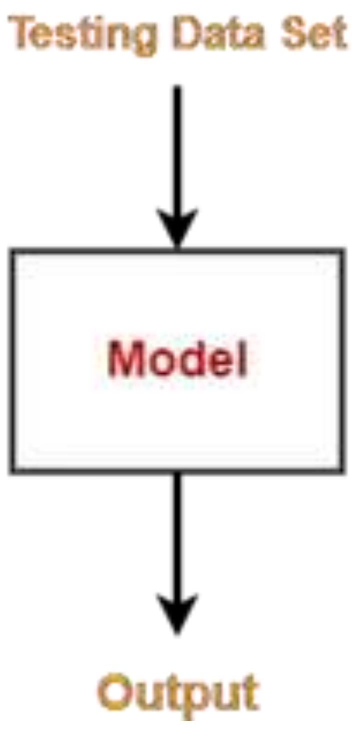
Training Model

1. The model is trained to improve its ability.
2. The dataset is divided into training dataset and testing dataset. (Training and Testing split is order of 80/20 or 70/30)
3. It also depends upon the size of the dataset.
4. Training dataset is used for Learning purpose.
5. Testing dataset is used for the Evaluating purpose.
6. Training dataset is fed to the learning algorithm.
7. The learning algorithm finds a mapping between the input and the output and generates the model.



Evaluating Model & Prediction

1. It allows to test the model against data that has never been used before for training.
2. Metrics such as accuracy, precision, recall etc are used to test the performance.
3. If the model does not perform well, the model is re-built using different hyper parameters.
4. The accuracy may be further improved by tuning the hyper parameters.



Testing Data Set → Model → Output

The built system is finally used to do something useful in the real world.

References

1. Sebastian Raschka , Yuxi (Hayden) Liu Machine Learning with PyTorch and Scikit-Learn: Developmachine learning and deep learning models with Python Packt Publishing Limited (23 December 2022).

2. Aurélien Géron "Hands-On Machine Learning with Scikit-Learn and TensorFlow" Publisher(s): O'Reilly Media, Inc 2017.

3. https://www.gatevidyalay.com/machine-learning-workflow-process-steps/