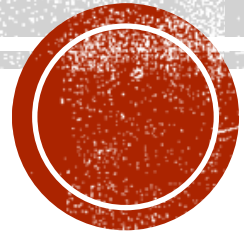
UNIT –I

INTRODUCTION TO ML



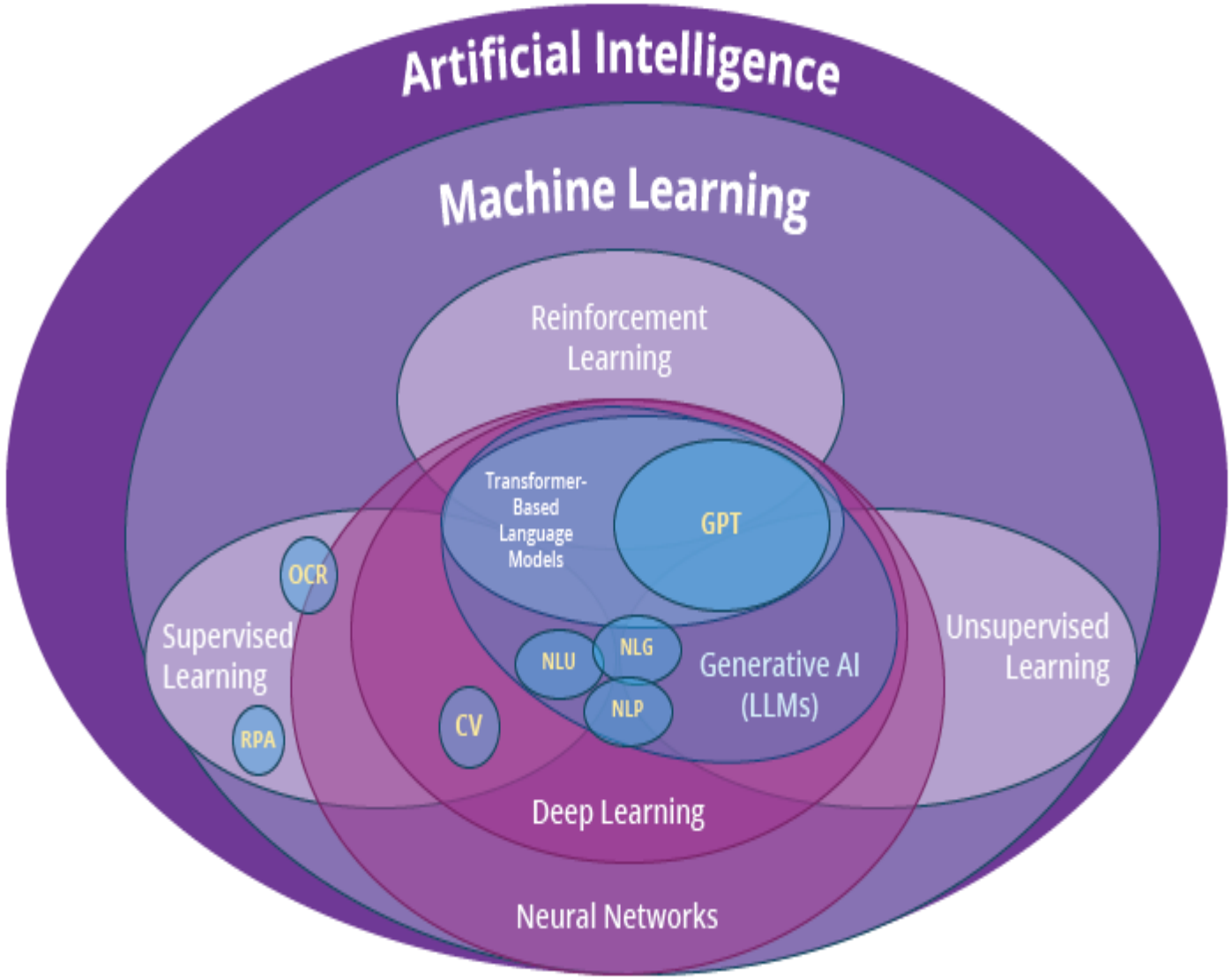
INTRODUCTION TO ML

ML algorithms are trained to find relationships and patterns in data. Using historical data as input, these algorithms can make predictions, classify information, cluster data points, reduce dimensionality and even generate new content.

Applications of ML:

- Image Recognition
- Speech Recognition
- Recommender Systems
- Fraud Detection
- Medical Diagnosis
- Stock Market Trading

Broad Fields of AI



STATISTICAL LEARNING

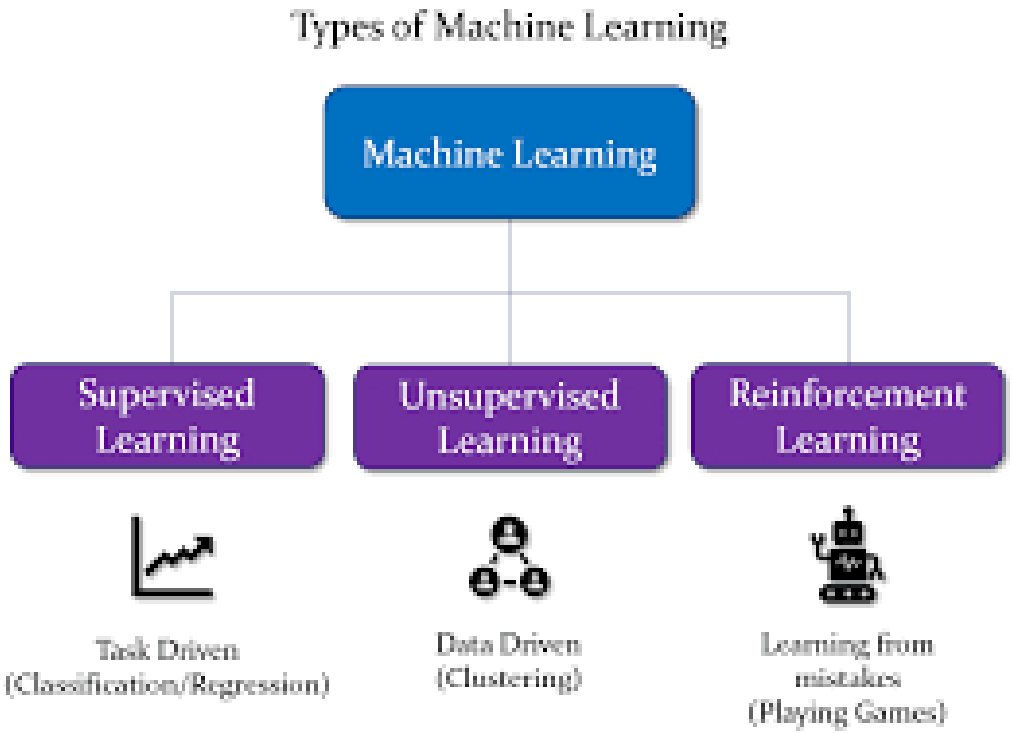
Statistical learning theory is a framework for machine learning that draws from statistics and functional analysis. It deals with finding a predictive function based on the data presented. The main idea in statistical learning theory is to build a model that can draw conclusions from data and make predictions.

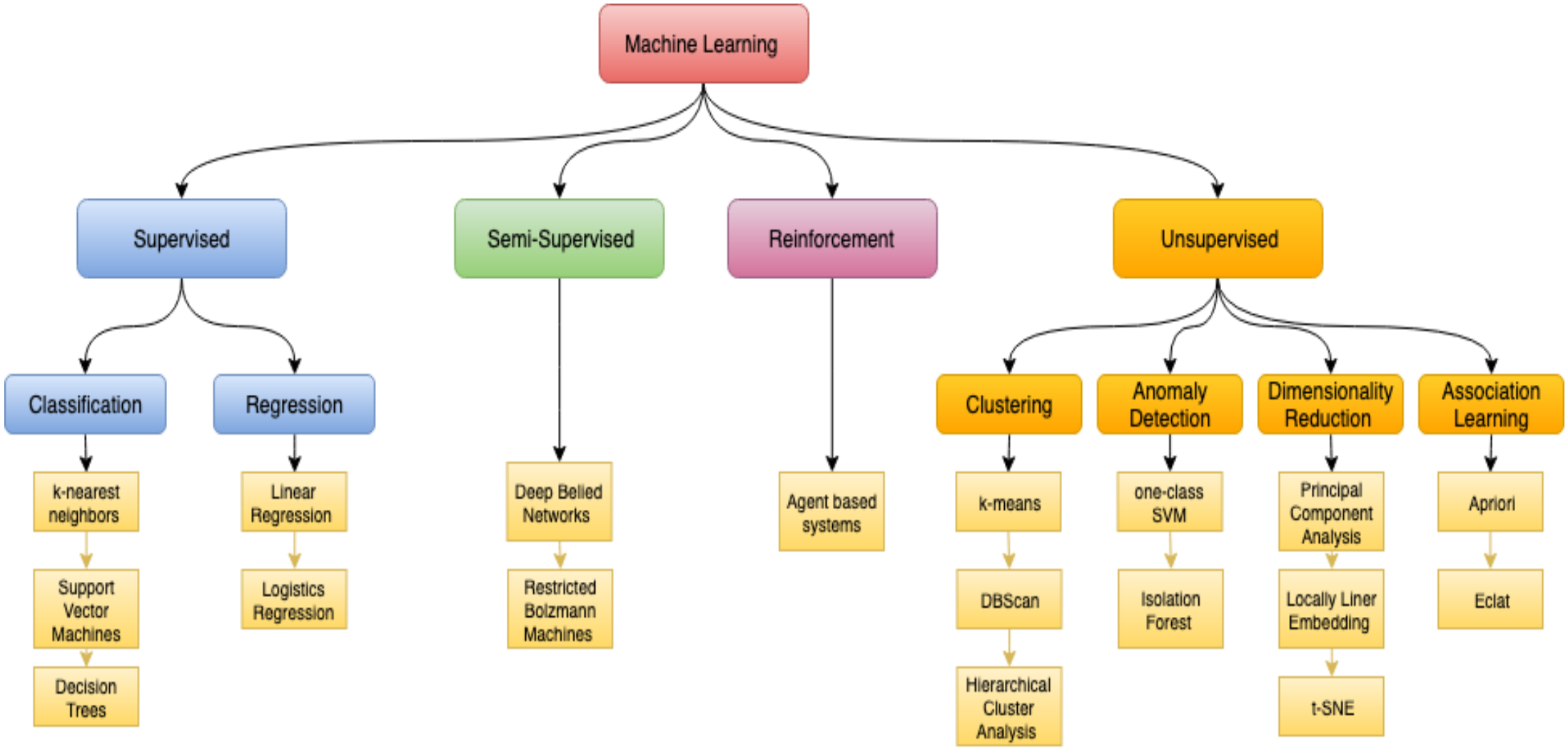
Types of Data in Statistical Learning:

- With statistical learning theory, there are two main types of data:
- Dependent Variable — a variable (y) whose values depend on the values of other variables (a dependent variable is sometimes also referred to as a target variable)
- Independent Variables — a variable (x) whose value does not depend on the values of other variables (independent variables are sometimes also referred to as predictor variables, input variables, explanatory variables, or features)

TYPES OF MACHINE LEARNING:

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning.




Machine Learning
Supervised
Semi-Supervised
Reinforcement
Unsupervised
Classification
Regression
k-nearest neighbors
Support Vector Machines
Decision Trees
Linear Regression
Logistics Regression
Deep Belied Networks
Restricted Bolzmann Machines
Agent based systems
Clustering
Anomaly Detection
Dimensionality Reduction
Association Learning
k-means
DBScan
Hierarchical Cluster Analysis
one-class SVM
Isolation Forest
Principal Component Analysis
Locally Liner Embedding
t-SNE
Apriori
Eclat

REINFORCEMENT LEARNING:

Reinforcement learning involves programming an algorithm with a distinct goal and a set of rules to follow in achieving that goal. The algorithm seeks positive rewards for performing actions that move it closer to its goal and avoids punishments for performing actions that move it further from the goal.

Reinforcement learning is often used for tasks such as the following:

- Helping robots learn to perform tasks in the physical world.
- Teaching bots to play video games.
- Helping enterprises plan allocation of resources.

SUPERVISED LEARNING:

supplies algorithms with labeled training data and defines which variables the algorithm should assess for correlations. Both the input and output of the algorithm are specified. Initially, most ML algorithms used supervised learning, but unsupervised approaches are gaining popularity.

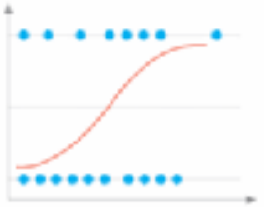
Supervised learning algorithms are used for numerous tasks, including the following:

- **Binary classification.** This divides data into two categories.
- **Multiclass classification.** This chooses among more than two categories.
- **Ensemble modeling.** This combines the predictions of multiple ML models to produce a more accurate prediction.
- **Regression modeling.** This predicts continuous values based on relationships within data.

5 types of regression

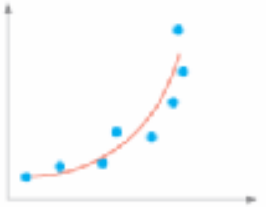
Linear regression

Predicts a continuous output by modeling a straight-line relationship between input features and target variables, such as estimating the impact of price changes on demand.



Logistic regression

Models the probability of binary outcomes, such as predicting customer churn; commonly used in classification tasks.

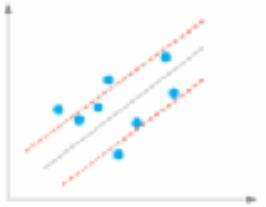


Polynomial regression

Captures nonlinear relationships, such as estimating the impact of ad spending on sales, by fitting a polynomial curve to data points.

Time series regression

Predicts future values in a time-dependent data set; often employed to forecast future values based on past observations, as seen in stock market analysis.



Support vector regression

Approximates a continuous function by identifying a hyperplane that best represents the data's structure; valuable in various applications, including financial market prediction.

UNSUPERVISED LEARNING:

Unsupervised learning doesn't require labeled data. Instead, these algorithms analyze unlabeled data to identify patterns and group data points into subsets using techniques such as gradient descent. Most types of deep learning, including neural networks, are unsupervised algorithms.

Unsupervised learning is effective for various tasks, including the following:

- Splitting the data set into groups based on similarity using clustering algorithms.
- Identifying unusual data points in a data set using anomaly detection algorithms.
- Discovering sets of items in a data set that frequently occur together using association rule mining.
- Decreasing the number of variables in a data set using dimensionality reduction techniques.

SEMI-SUPERVISED LEARNING:

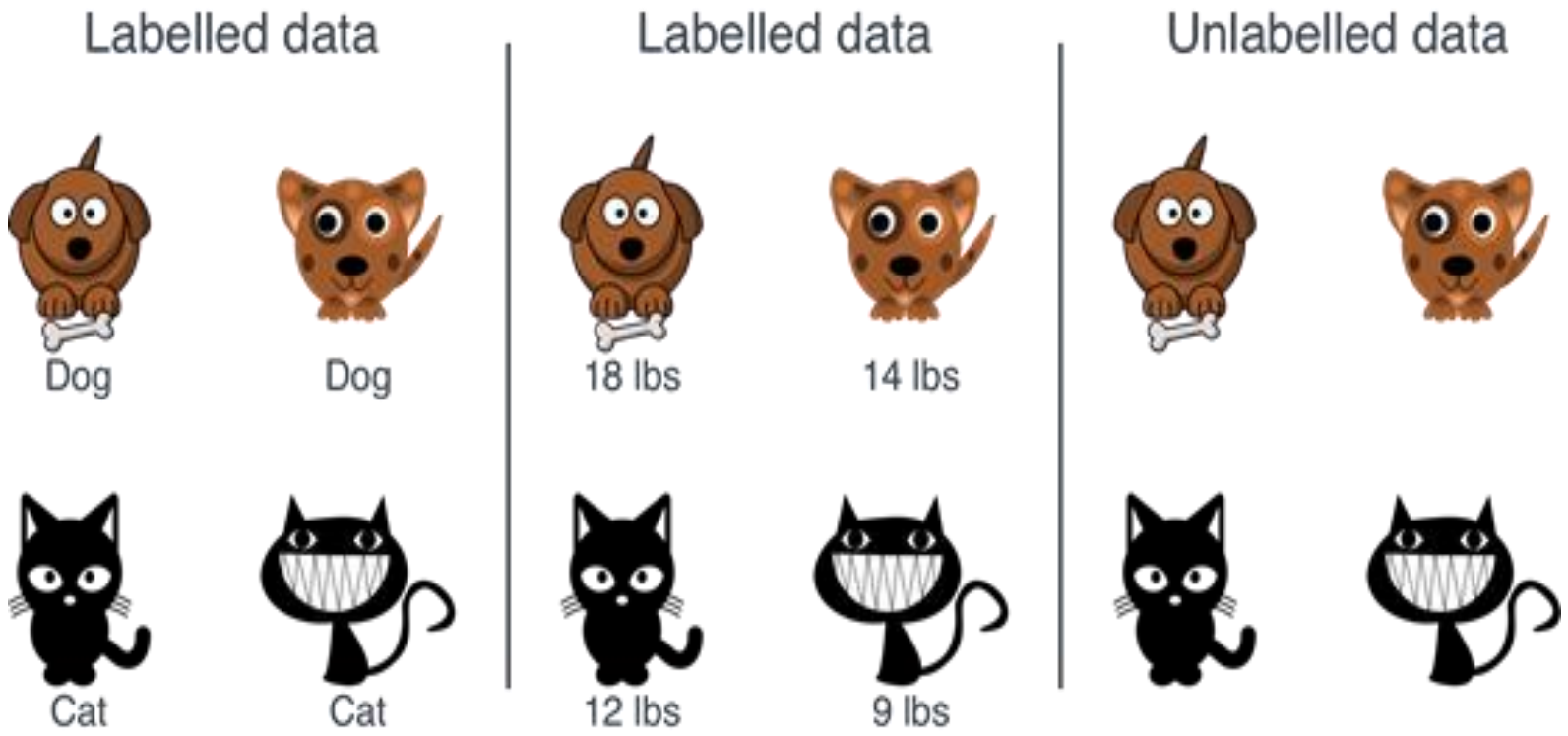
Semi-supervised learning provides an algorithm with only a small amount of labeled training data. From this data, the algorithm learns the dimensions of the data set, which it can then apply to new, unlabeled data. Note, however, that providing too little training data can lead to overfitting, where the model simply memorizes the training data rather than truly learning the underlying patterns.

Although algorithms typically perform better when they train on labeled data sets, labeling can be time-consuming and expensive. Semisupervised learning combines elements of supervised learning and unsupervised learning, striking a balance between the former's superior performance and the latter's efficiency.

Semi-supervised learning can be used in the following areas, among others:

- **Machine translation.** Algorithms can learn to translate language based on less than a full dictionary of words.
- **Fraud detection.** Algorithms can learn to identify cases of fraud with only a few positive examples.
- **Labeling data.** Algorithms trained on small data sets can learn to automatically apply data labels to larger sets.

LABELLED DATA Vs UNLABELLED DATA



TYPES OF ML MODELS:

Models are the central concept in machine learning as they are what one learns from data in order to solve a given task. There is a huge variety of machine learning models available. This is particularly due to the omnipresence of tasks that machine learning aims to solve. The 3 most common groups of models that we see are :

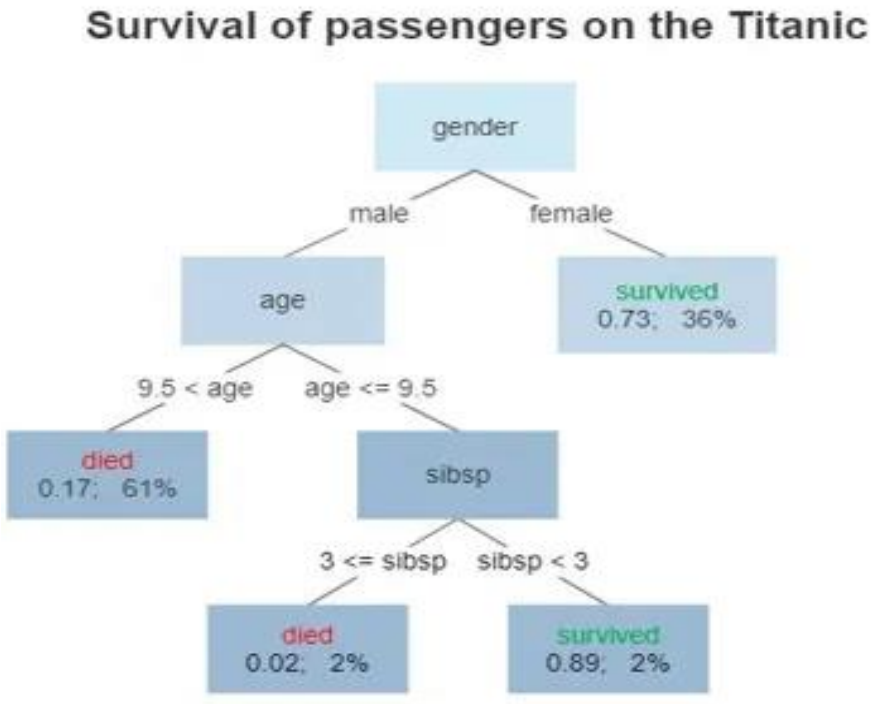
- ***Logical***,
- ***Geometric***
- ***Probabilistic*** Models.

LOGICAL MODEL:

"Logical" because models of this kind can easily be translated into ***rules*** that humans can understand, such as ., *if lottery = 1 then class = Y = spam.* Such rules are easily arranged in a tree structure, which we refer to as a ***feature tree***.

Feature trees whose leaves are labelled with classes are commonly called **decision trees***.*

A simple logical model is shown below:



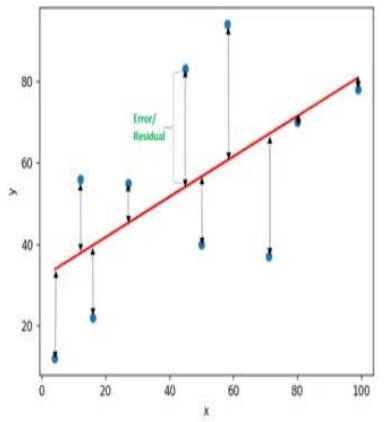
GEOMETRIC MODEL:

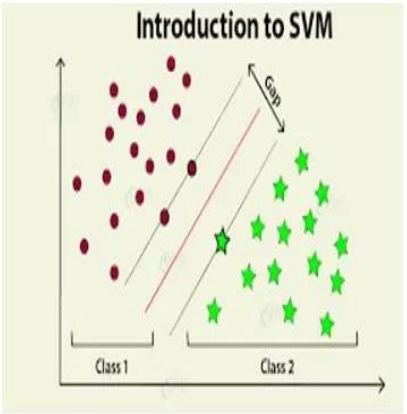
Geometric models/feature learning is a technique of combining machine learning and computer vision to solve visual tasks. These models define similarity by considering the geometry of the instance space. Here, features could be described as points in two dimensions (x- and y-axis) or a three dimensional space (x, y, and z).

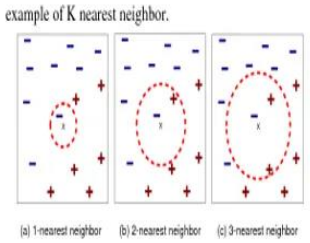
*An **instance space** is the set of all possible or describable instances, whether they are present in our data set or not. Usually this set has some geometric structure. For instance, if all features are numerical, then we can use each feature as a coordinate in a Cartesian coordinate system.*

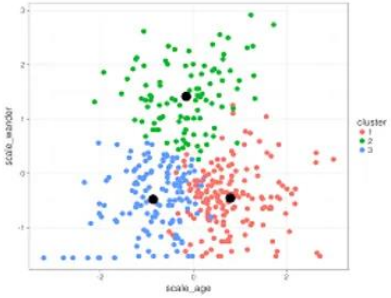
- Geometric models are basically of two types :-

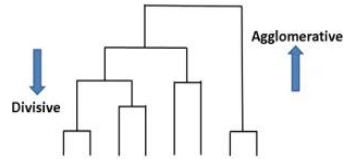
- A geometric model that is constructed directly in instance space, using geometric concepts such as lines or planes are used to segment the instance space known as Linear Models.
- A geometric model that uses distance as a metric to represent the similarity between the instances is known as Distance based Models. The distance metrics commonly used are, Euclidean, Minkowski, Manhattan, and Mahalanobis.











PROBABILISTIC MODEL:

A probability model/method is based on the theory of probability, or the fact that randomness play a role in predicting future events.

Naïve Bayes is an example of Probabilistic models, which follows **Bayes theorem**.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- P(Y/X) = Posterior probability (probability of hypothesis is true given the evidence)
- P(X/Y) = Likelihood ratio (probability of seeing the evidence if the hypothesis is true)
- P(Y) = Class Prior probability (probability of hypothesis is true, before any evidence is present)
- P(X) = Predictor Prior probability (probability of observing the evidence)

MODEL EVALUATION:

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses.

Why is Evaluation necessary for a successful model?

Evaluation is necessary for ensuring that machine learning models are reliable, generalizable, and capable of making accurate predictions on new, unseen data, which is crucial for their successful deployment in real-world applications. Overfitting and underfitting are the two biggest causes of poor performance of machine learning algorithms.

- ***Overfitting:*** *Occurs when the model is* ***so closely*** *aligned to the training data that it does not know how to respond to new data.*
- ***Underfitting****: Occurs when the model* ***cannot adequately capture*** *the underlying structure of the data.*
- ***Right Fit:*** *Occurs when both the training data error and the test data are minimal*

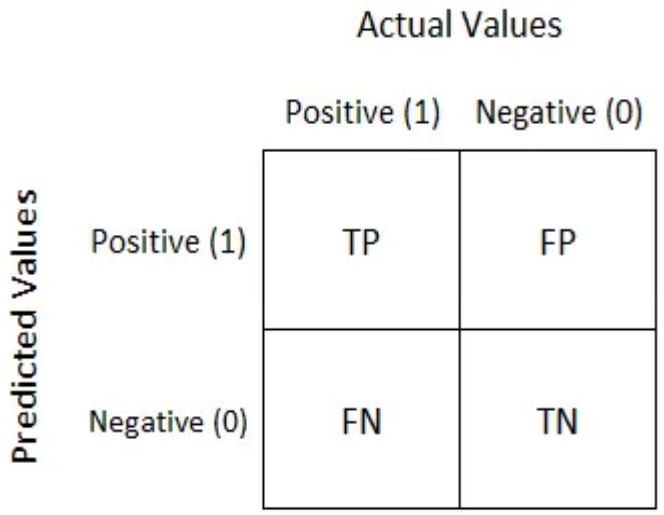
Error	Overfitting	Right Fit	Underfitting
Training	Low	Low	High
Test	High	Low	High

There are many metrics like Accuracy, Precision, Recall, F1 score, Area under Curve, Confusion Matrix, and Mean Square Error. Cross Validation is one technique that is followed during the training phase and it is a model evaluation technique as well.

Accuracy

Accuracy is defined as the ratio of the number of correct predictions to the total number of predictions. This is the most fundamental metric used to evaluate the model. The formula is given by:

```
Accuracy = (TP+TN)/(TP+TN+FP+FN)
```



True Positives: It is also known as TP. It is the output in which the actual and the predicted values are YES.

True Negatives: It is also known as TN. It is the output in which the actual and the predicted values are NO.

False Positives: It is also known as FP. It is the output in which the actual value is NO but the predicted value is YES.

False Negatives: It is also known as FN. It is the output in which the actual value is YES but the predicted value is NO.

Precision and Recall

- **Precision** is the ratio of true positives to the summation of true positives and false positives. It basically analyses the positive predictions.

```
Precision = TP/(TP+FP)
```

The drawback of Precision is that it does not consider the True Negatives and False Negatives.

- **Recall** is the ratio of true positives to the summation of true positives and false negatives. It basically analyses the number of correct positive samples.

```
Recall = TP/(TP+FN)
```

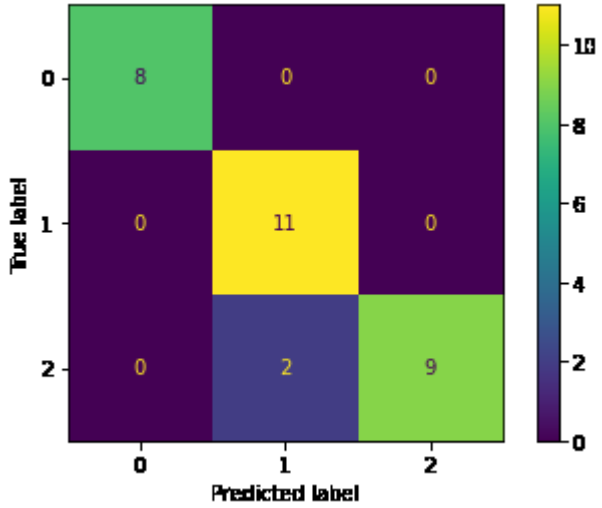
- The drawback of Recall is that often it leads to a higher false positive rate.

F1 score

- The F1 score is the harmonic mean of precision and recall. It is seen that during the precision-recall trade-off if we increase the precision, recall decreases and vice versa. The goal of the F1 score is to combine precision and recall.

```
F1 score = (2×Precision×Recall)/(Precision+Recall)
```

Confusion Matrix- A confusion matrix is an N x N matrix where N is the number of target classes. It represents the number of actual outputs and the predicted outputs.



AUC-ROC Curve

AUC (Area Under Curve) is an evaluation metric that is used to analyze the classification model at different threshold values. The Receiver Operating Characteristic(ROC) curve is a probabilistic curve used to highlight the model's performance. The curve has two parameters:

•TPR: It stands for True positive rate. It basically follows the formula of Recall.

•FPR: It stands for False Positive rate. It is defined as the ratio of False positives to the summation of false positives and True negatives.

This curve is useful as it helps us to determine the model's capacity to distinguish between different classes.