Beer & Diaper: An Impossible Correlation

Summer Internship

2019

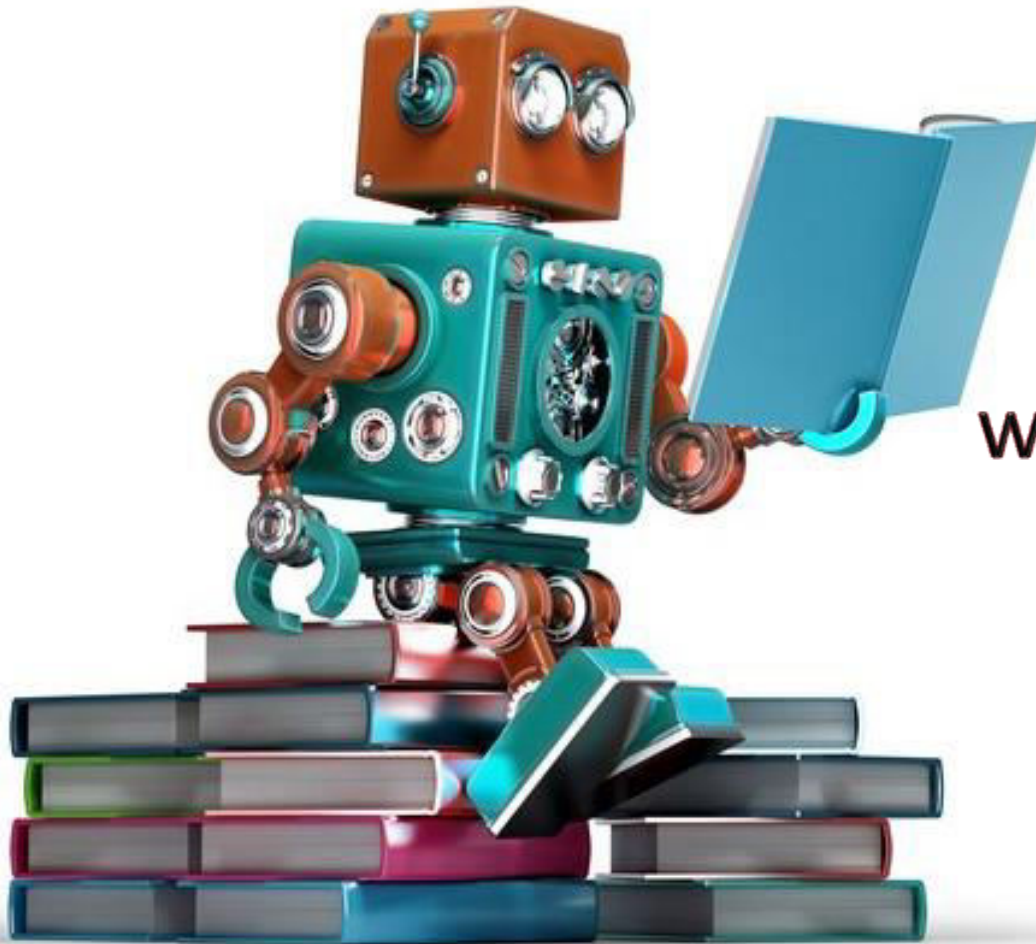**01 Company**
Cognifront

02 Domain
Machine Learning

03 Time Span
6th June – 21st June

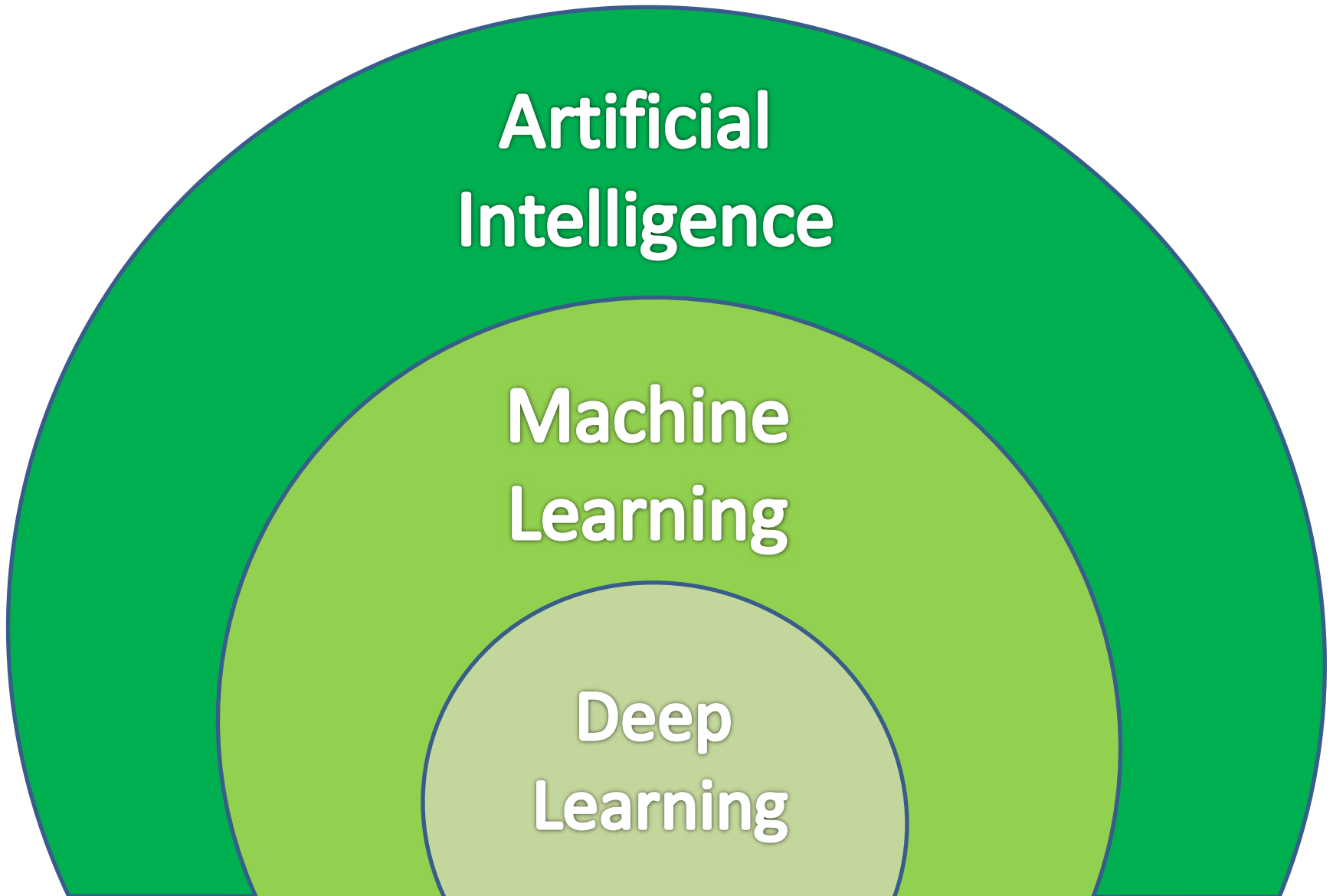# Why do we need Machine Learning ?

- **Human Expertise does not exist.**

  eg. **Navigating on Mars**

- **Solution Changes in Time.**

  eg. **Routing on Computer Networks**

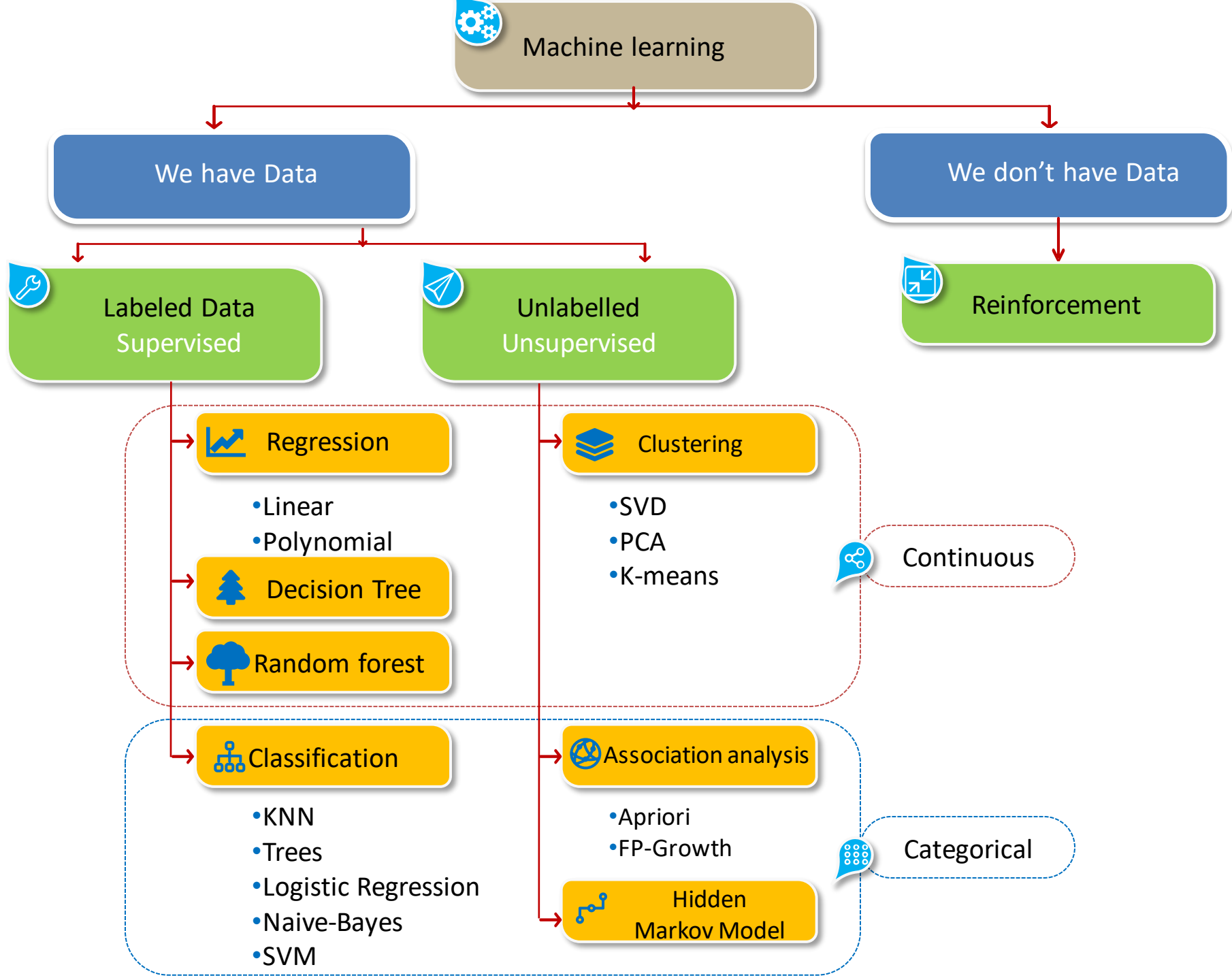- **Improving Sales.**

  eg. **Market Basket Analysis**

Machine Learning ?

Machine Learning is a field of **Computer Science** which gives computer the ability to learn **without being explicitly programmed**.

Relation Between AI, ML & Deep Learning

Artificial Intelligence

Machine Learning

Deep Learning

Machine learning

We have Data

We don't have Data

Labeled Data
Supervised

Unlabelled
Unsupervised

Reinforcement

Regression
•Linear
•Polynomial

Decision Tree

Random forest

Classification
•KNN
•Trees
•Logistic Regression
•Naive-Bayes
•SVM

Clustering
•SVD
•PCA
•K-means

Association analysis
•Apriori
•FP-Growth

Hidden
Markov Model

Continuous

Categorical

Linear Regression

- Linear Regression is a linear approach to modeling the relationship between a dependent and one or more independent variable.

- Types :
  - Simple Linear Regression.
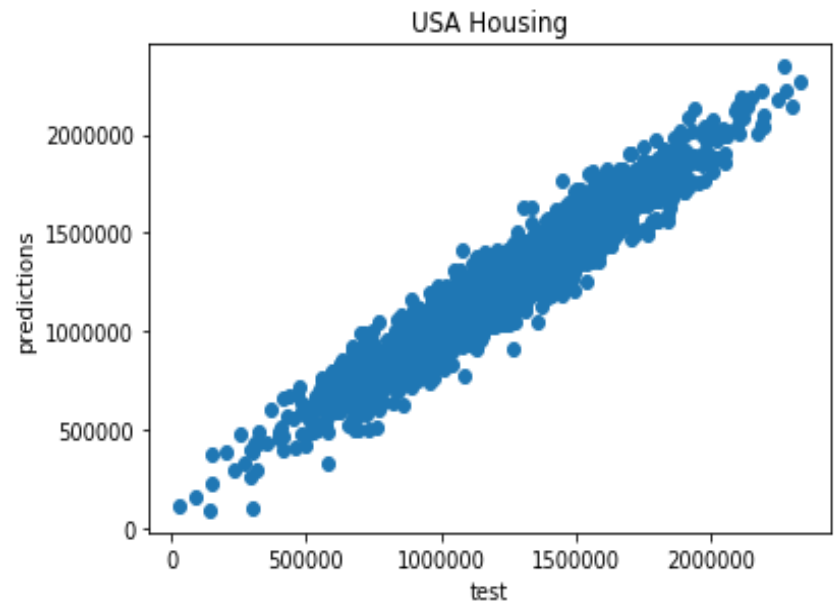 - Multiple Linear Regression.

y

dependent variable

minimize errors

x

independent variable
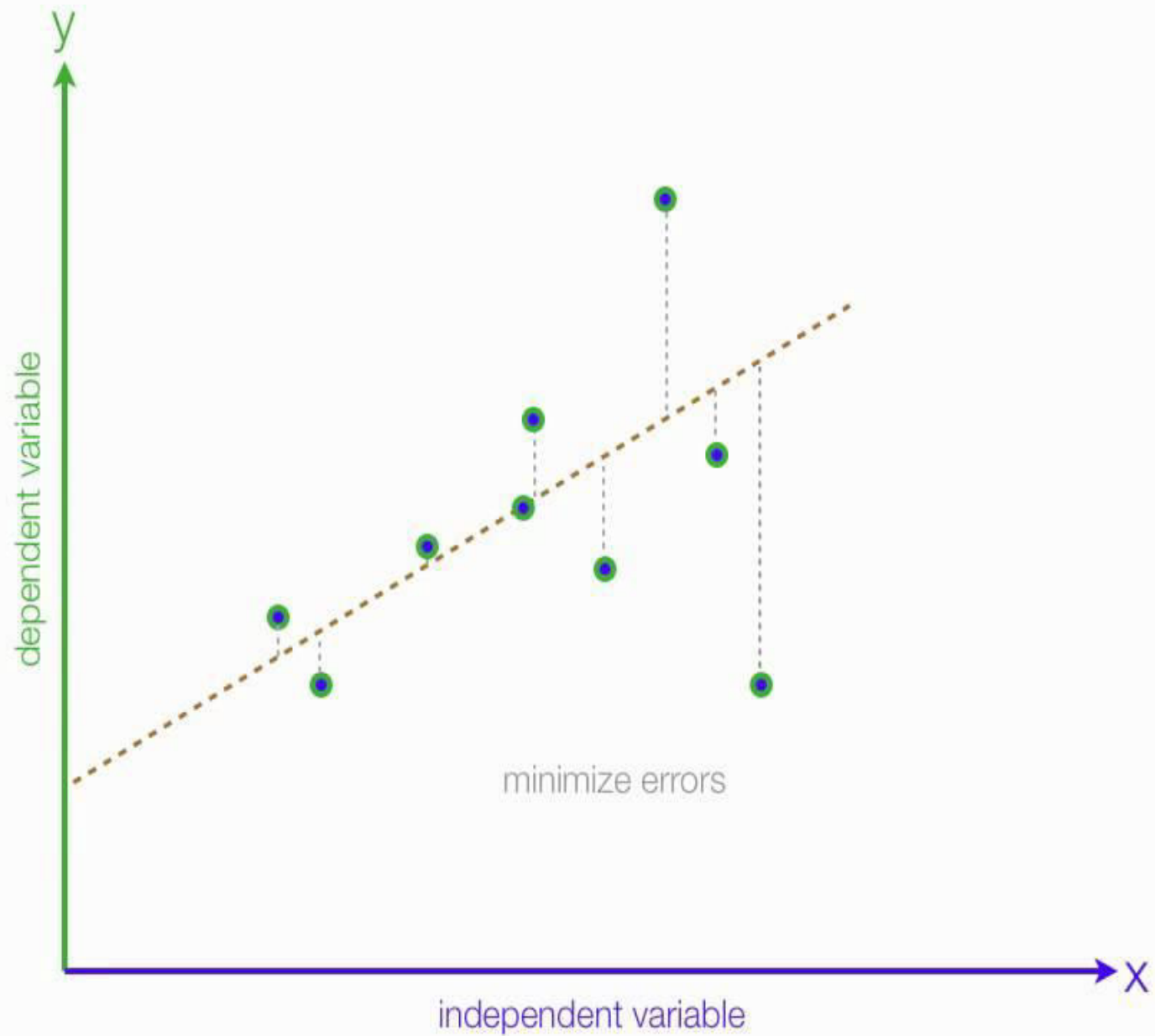
Case Study

Simple Linear Regression

Multiple Linear Regression

Salary Prediction on basis of work Experience

USA Housing price prediction

Accuracy : 90%

Accuracy : 93%

Linear Regression

Y=1

Y-Axis

Y=0

X-Axis

Logistic Regression

Y=1

Y-Axis

Y=0

X-Axis
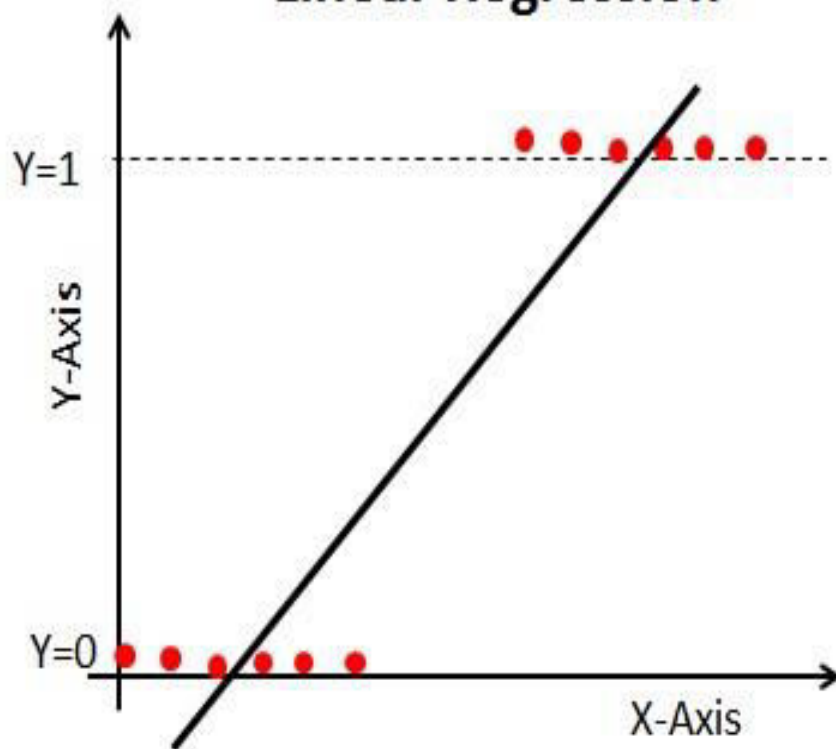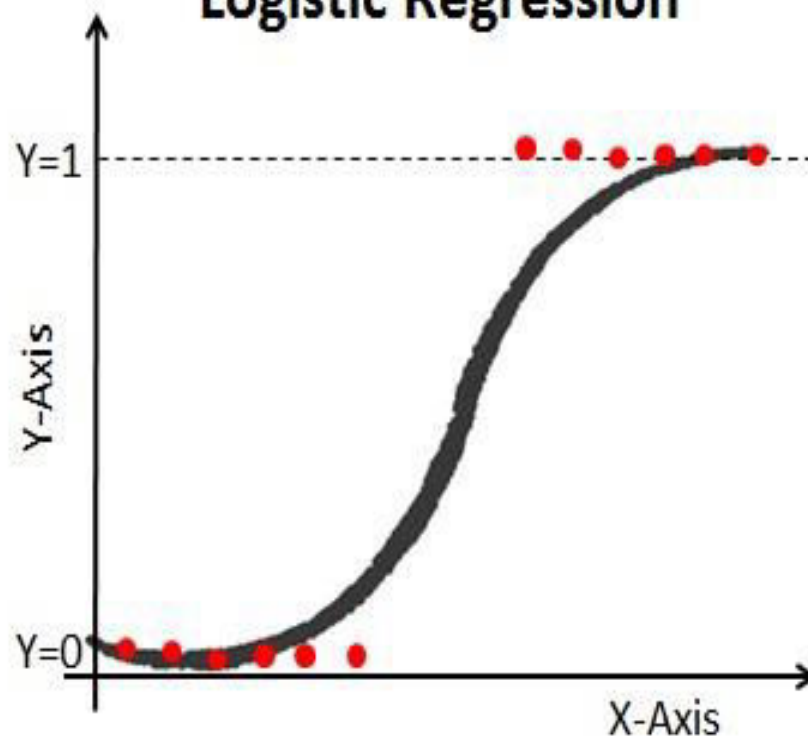
Logistic Regression

- It is an Classification Algorithm.

- Logistic Regression is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

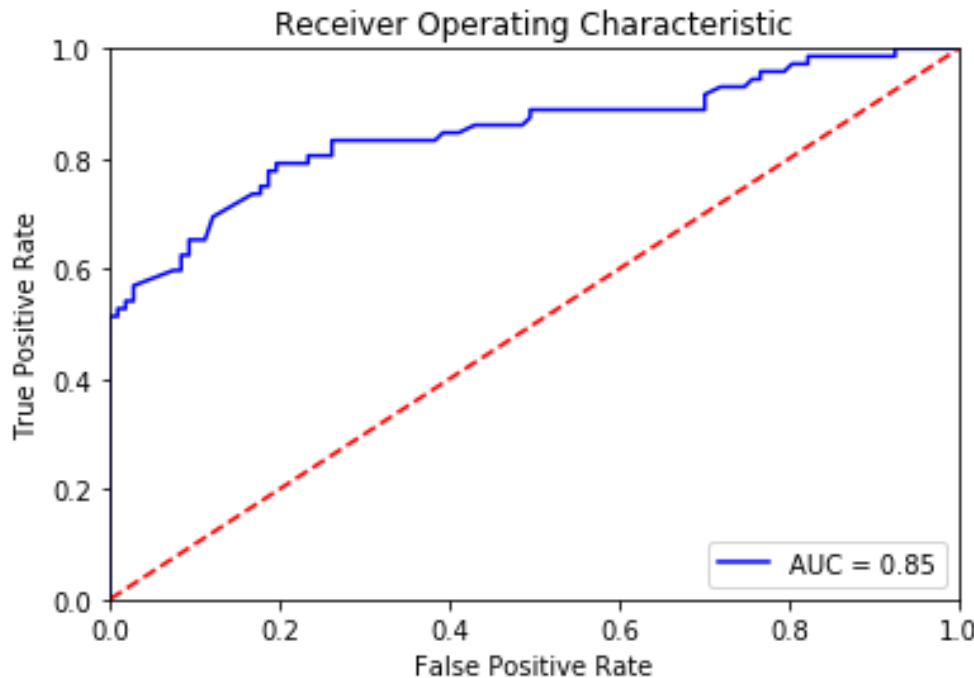- Uses Sigmoid Function for finding probability.

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

Case Study

- Logistic Regression

- Titanic Survival Classification

Accuracy : 82%

- Framingham Heart Study

Accuracy : 84%
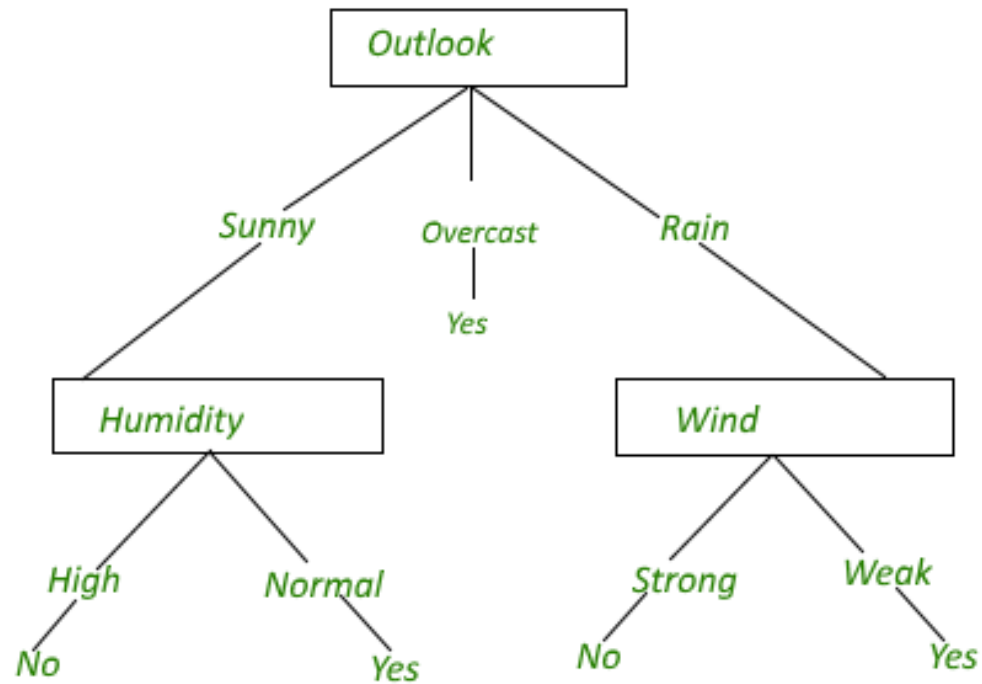
Decision Tree Classifier

- A **decision tree** is a decision support tool that uses a tree-like graph or model of decisions.

- A decision tree is a flowchart-like structure in which -

  - Internal Node : Test.

  - Branch : Outcome of the test.

  - Leaf Node : Class label.

Player will play Tennis or not ?

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Decision Tree

Decision Tree for *PlayTennis*

- Outlook
  - Sunny → Humidity
    - High → No
    - Normal → Yes
  - Overcast → Yes
  - Rain → Wind
    - Strong → No
    - Weak → Yes

Case Study

- Decision Tree Classifier

- Diabetes Prediction
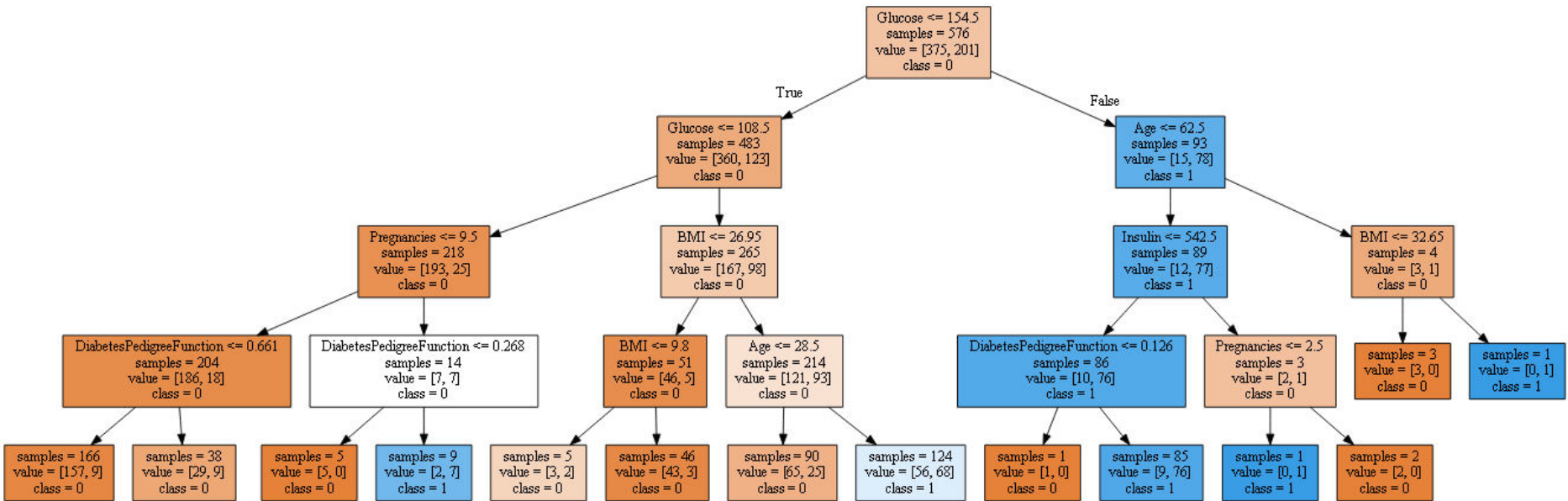
Accuracy : 70%

Note : Accuracy can be increased up to 80% if SVM is used instead of Decision Tree classifier

Random Forest Classifier

- **Random forests** method operates by constructing a multiple decision trees at training time and **outputting the mean prediction** (regression) of the individual trees.

- It uses multiple **Decision Trees** as well as **Bagging Method**, hence very time consuming (depending on processing power)

Detecting whether an URL is spam or not
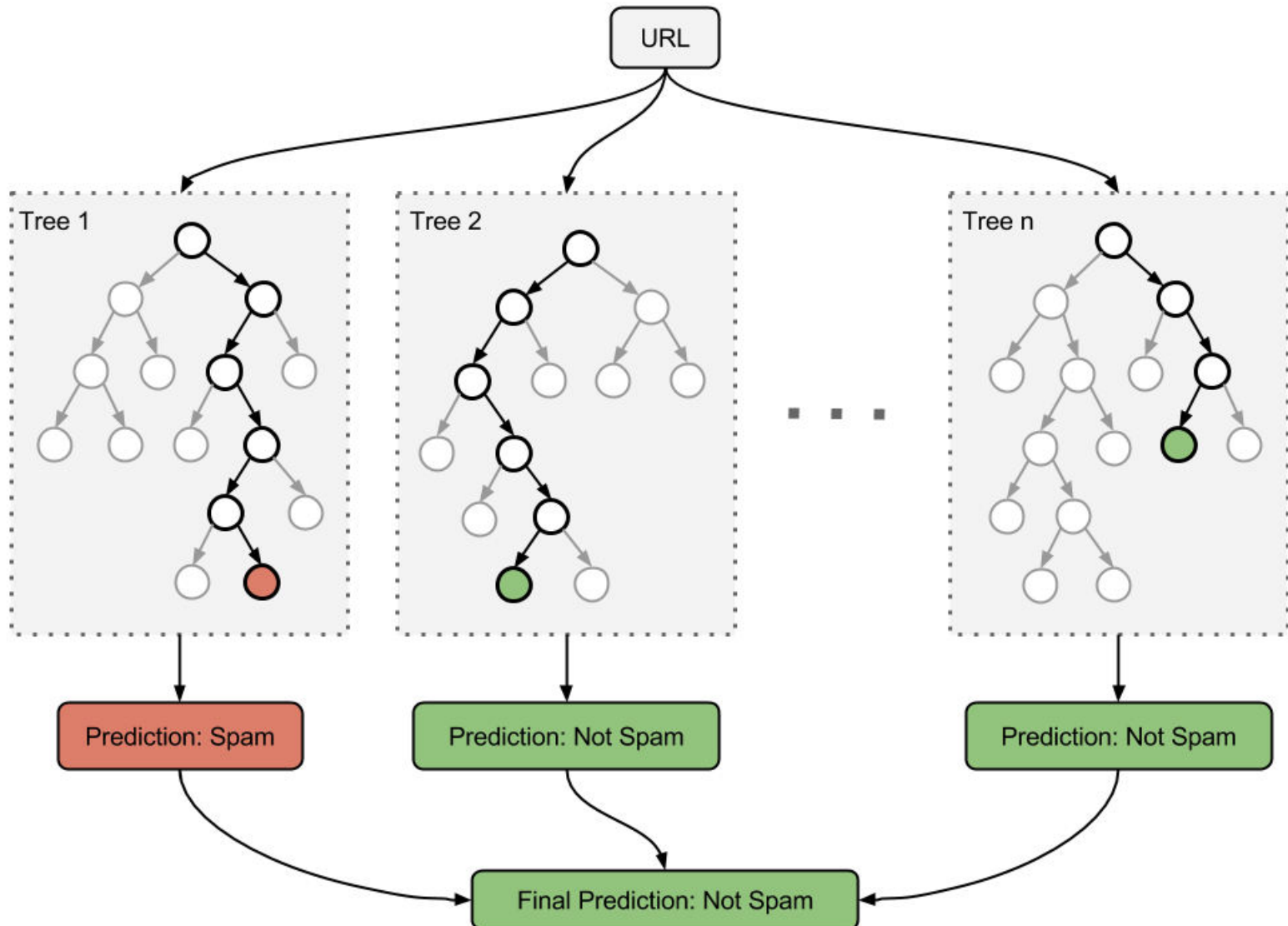
Case Study

- Random Forest Classifier
- Predicting whether Mushroom is Edible or Poisonous.
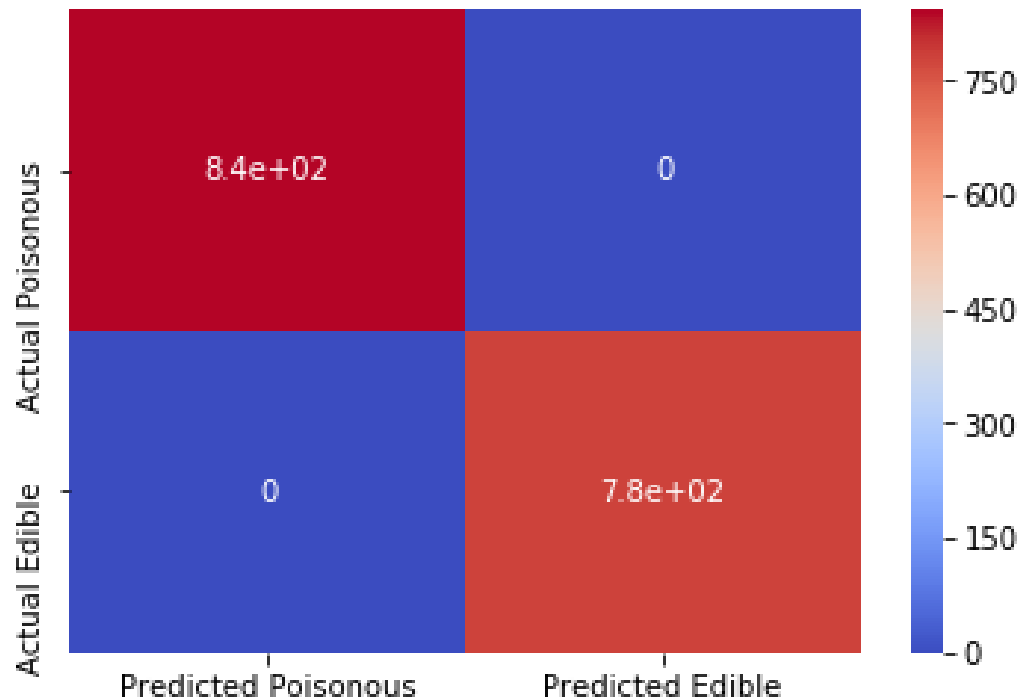
Accuracy : 100%

Hierarchical clustering

- hierarchical clustering is a method of **cluster analysis** which seeks to build a hierarchy of clusters.

- Strategies for hierarchical clustering generally fall into two types:

  - Agglomerative (Bottom – Up Approach)
  - Divisive (Top - Down)

Case Study

- Hierarchical Clustering

- Clustering of customers on the basis of Annual Income and Spending Score
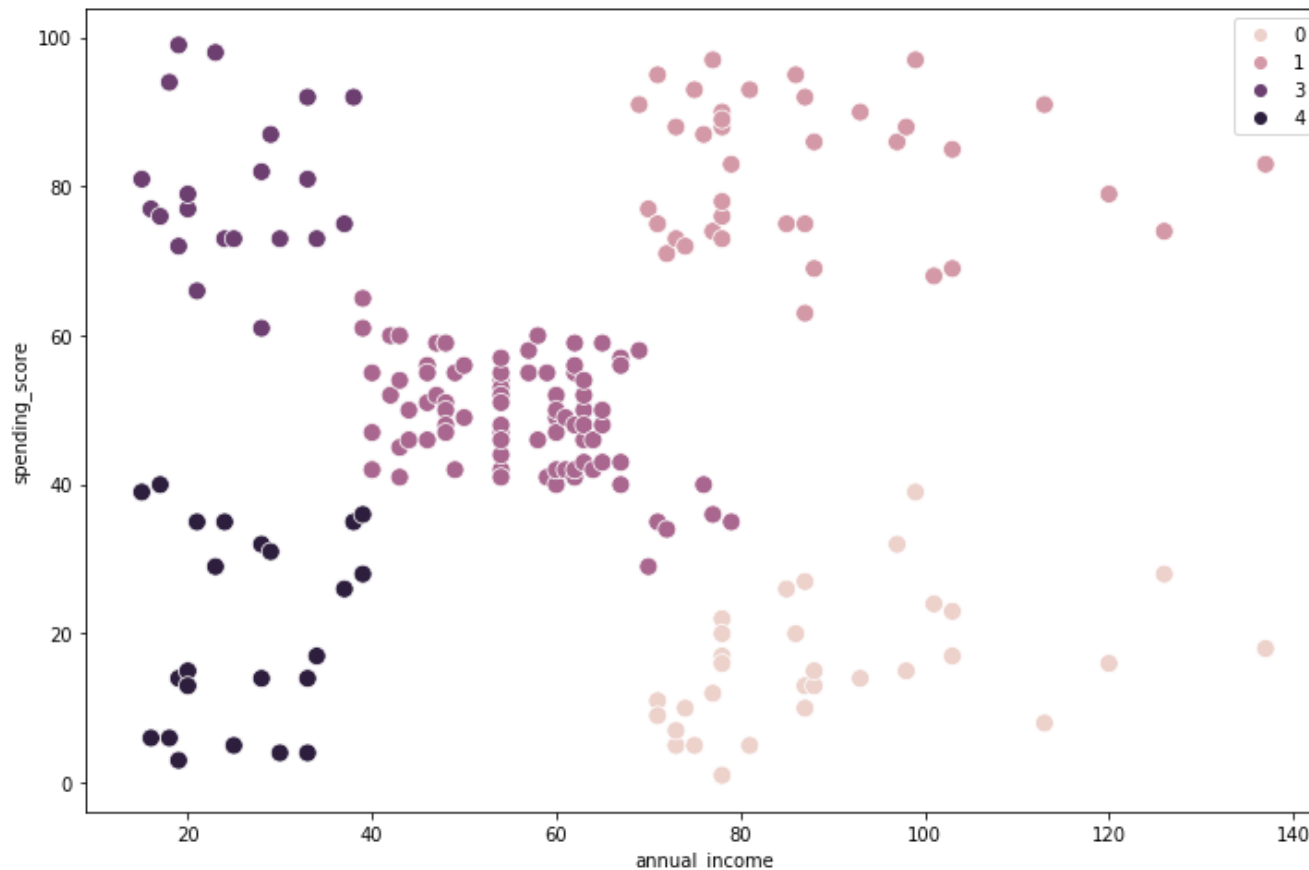
K means clustering

- k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Case Study

- ## K means Clustering
- Customer clustering based on brand loyalty score and price sensitivity score
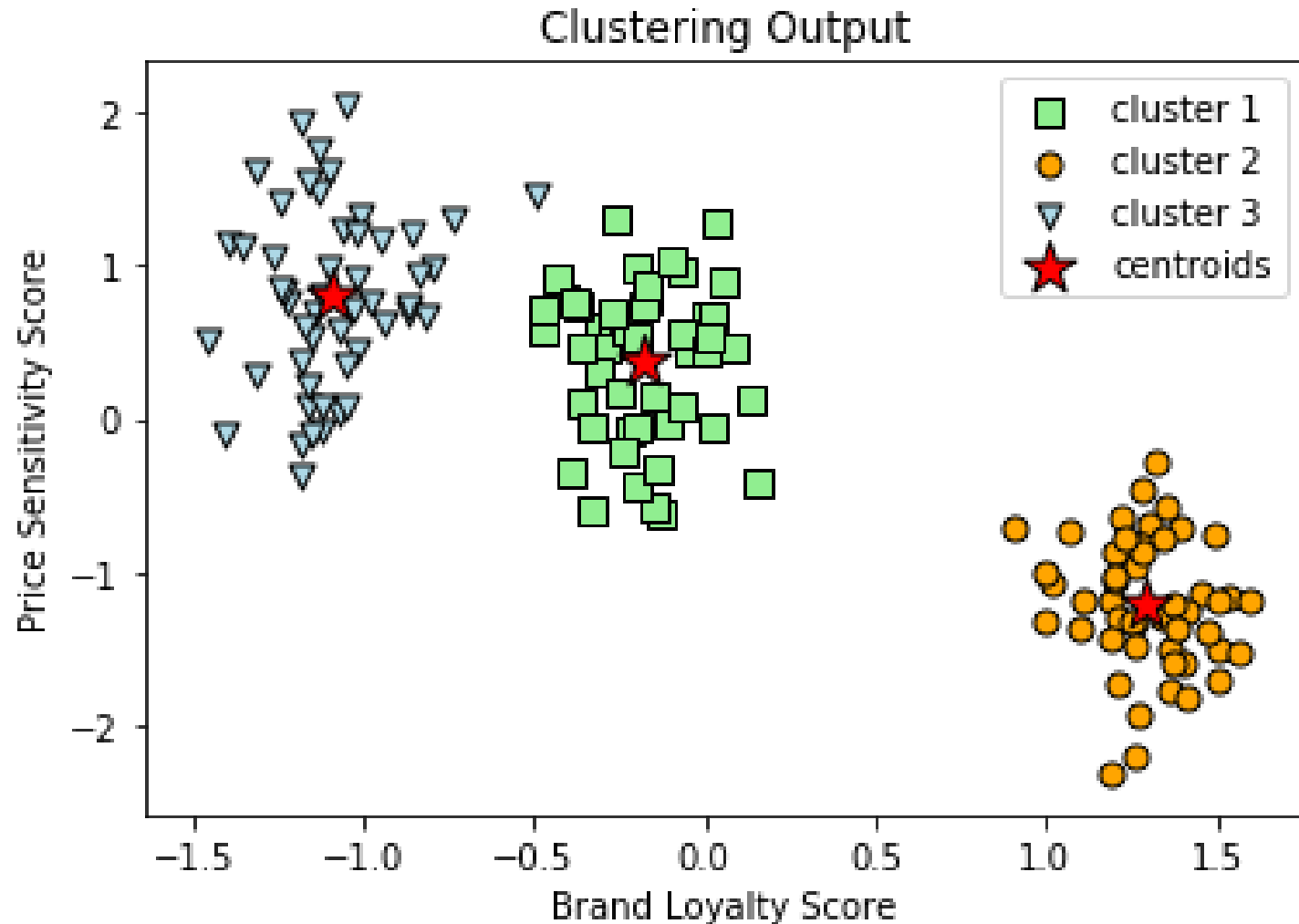
Day Wise Summary

Date	Syllabus Covered
6th June	Inauguration, Introduction to Machine Learning
7th June	Data types & Output Formatting in Python
8th June	Flow Control and Matrices
10th June	Set Operations and Functions
11th June	Modules, File Operations, NumPy Library and Operations
12th June	Pandas Library and operations
13th June	Operations on Dataframes
14th June	MatPlotLib, Seaborne, EDA
15th June	Statistics and Linear Regression
17th June	Logistic Regression
18th June	Decision Tree Classifier
19th June	Random Forest & Clustering
20th June	ANN, Certificate Distribution