Six Weeks Summer Training Report

On

Machine Learning Advanced Certification Training

Submitted by

Hemchand Chandravanshi

Registration no. – 12008563

Course code - CSE443

Program – B TECH CSE

School of Computer Science and Engineering

Lovely Professional University, Phagwara, Punjab

Declaration

I hereby declare that I have completed my six-week summer training program at **SimpliLearn** from **10th June, 2022** to **10th July,2022.** I have declared that I have worked with full dedication during these six weeks of training and my learning outcomes fulfil the requirements of training for the award of degree of Bachelor of Technology in CSE.

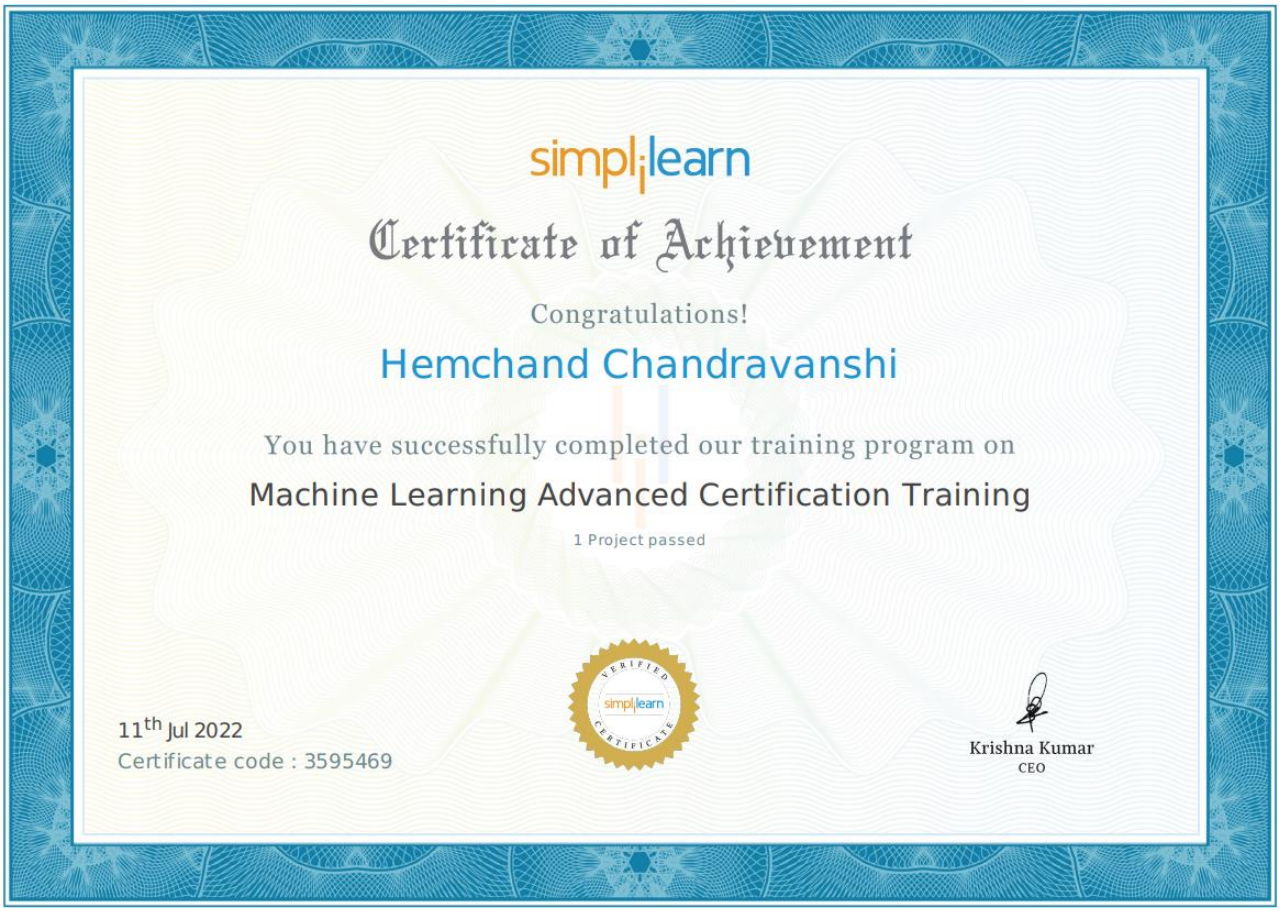
Acknowledgement

The success and final outcome of learning Machine Learning is required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my course and few of the projects. All that I have done is only due to such supervision and assistance and I would no forget to thank them.

I respect and thank **Simplilearn**, for providing me an opportunity to do the course and project work and giving me all support and guidance, which made me complete the course duly.

I am thankful to and fortunate enough to get constant encouragement support and guidance from all Teaching staffs of Simplilearn which helped us in successfully completing my course and project work.

Completion Certificate



Link - https://certificates.simplicdn.net/share/3595469.pdf

The above link is to download the certificate

Contents

1. Introduction……06
   1.1. Artificial Intelligence……06
   1.2. Data Economy …… 06
   1.3. Machine Learning……07
   1.4. Relation between Artificial intelligence, Machine Learning and Data Science……07

2. Application of Machine Learning ……08

3. Technology Learnt……09

   3.1. Machine Learning Algorithm……09

   3.2. Techniques of Machine Learning……10

   3.3. Data Preprocessing……11

   3.4. Supervised Learning……12

      3.4.1. Supervised Learning Classification……13

         3.4.1.1 Regression……13
            3.4.1.1.1 Linear Regression……14
            3.4.1.1.2 Multiple Linear Regression……14
            3.4.1.1.3 Polynomial Regression……15
            3.4.1.1.4 Decision Tree Regression……16
            3.4.1.1.5 Random Forest Regression……16
         3.4.1.2 Classification……17
            3.4.1.2.1 Linear Models……17
            3.4.1.2.2 Nonlinear Models……19

   3.5. Unsupervised Learning ……21
      3.5.1 Clustering……21
         3.5.1.1 Clustering Algorithms…… ……21
         3.5.1.2 K- means Clustering……22
   3.6. Feature Engineering……23
      3.6.1. Feature Scaling ……24

      3.6.2. Datasets……26

      3.6.3. Dimensionality Reduction with Principal Component Analysis……26

4. Reason for choosing Machine Learning……27

5. Learning Outcome……28

6. Project Description ……29

7. Gantt Chart……30

8. References……30

1. Introduction:

Definition:

1.1 Artificial Intelligence:

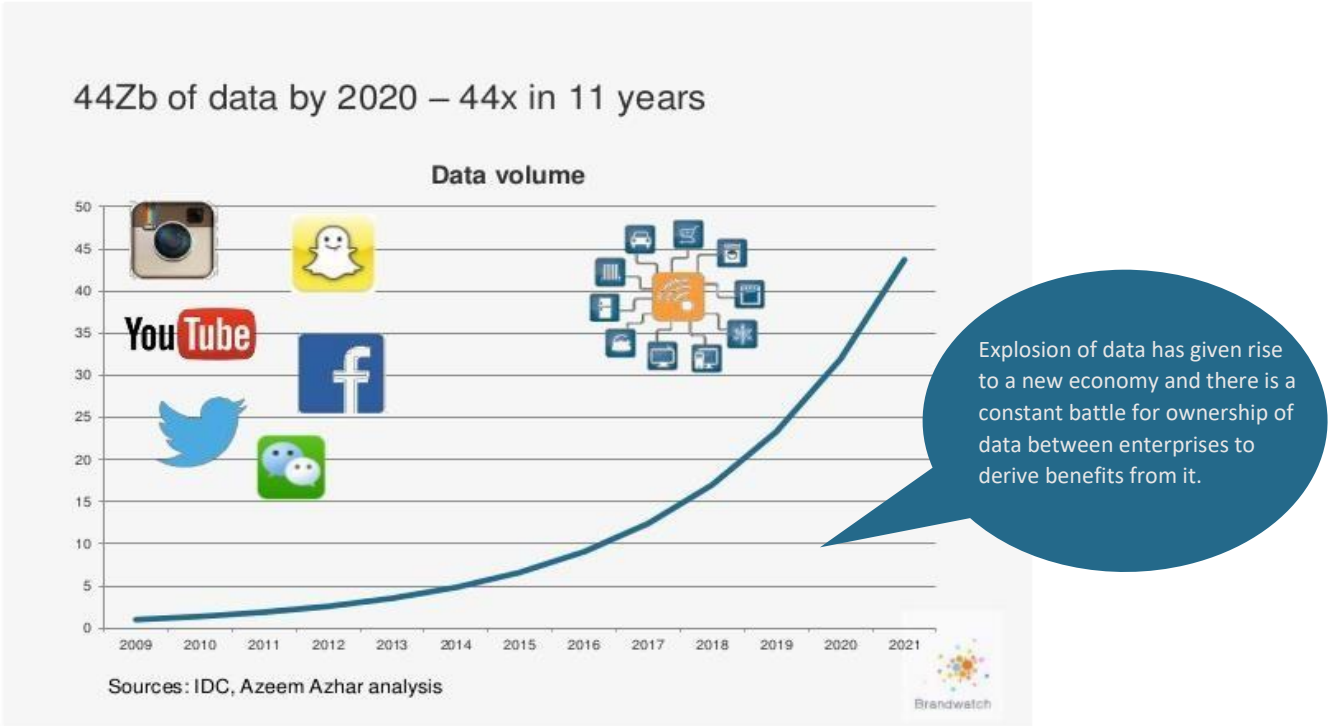
Artificial Intelligence refers to intelligence displayed by machines that simulates human and animal intelligence.

Artificial intelligence (AI) is **the ability of machines to replicate or enhance human intellect, such as reasoning and learning from experience**. Artificial intelligence has been used in computer programs for years, but it is now applied to many other products and services.

1.2 Data Economy:

World is witnessing real time flow of all types structured and unstructured data from social media, communication, transportation, sensors, and devices.

International Data Corporation (IDC) forecasts that 180 zettabytes of data will be generated by 2025.



This explosion of data has given rise to a new economy known as the **Data Economy.**

Data is the new oil that is precious but useful only when cleaned and processed.

There is a constant battle for ownership of data between enterprises to derive benefits from it.

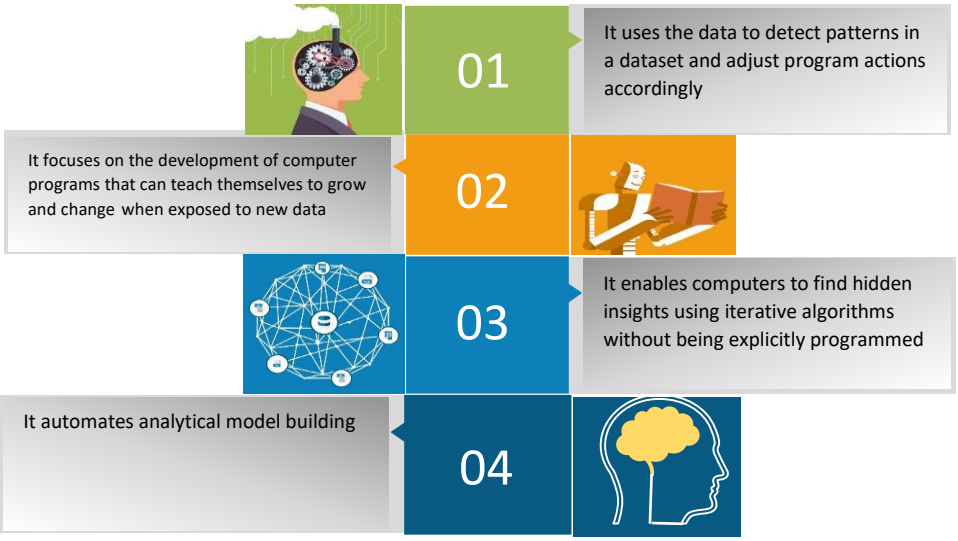
1.3 Machine Learning:

The term machine learning was coined in **1959** by Arthur Samuel, an IBM employee and pioneer in the field of computer gaming and artificial intelligence. Also the synonym self-teaching computers were used in this time period.

The capability of Artificial Intelligence systems to learn by extracting patterns from data is known as Machine Learning.

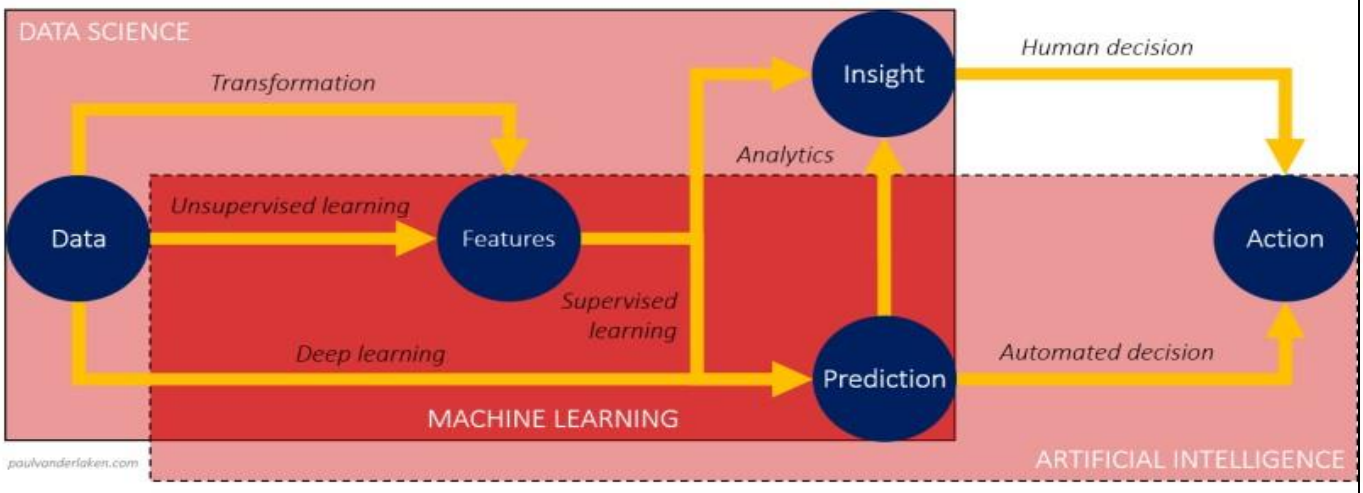
Machine Learning is an approach or subset of Artificial Intelligence that is based on the idea that machines can be given access to data along with the ability to learn from it.

1.3.1Features of Machine Learning:

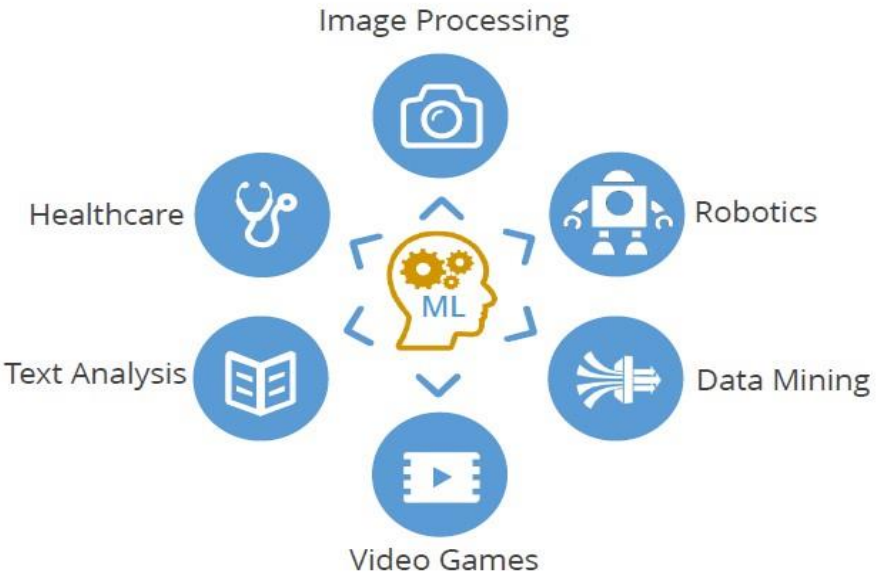


1.4 Relationship between Artificial Intelligence, Machine Learning, and Data Science:

Even though the terms data science, machine learning, and artificial intelligence (AI) fall in the same domain and are connected to each other, they have their specific applications and meaning.



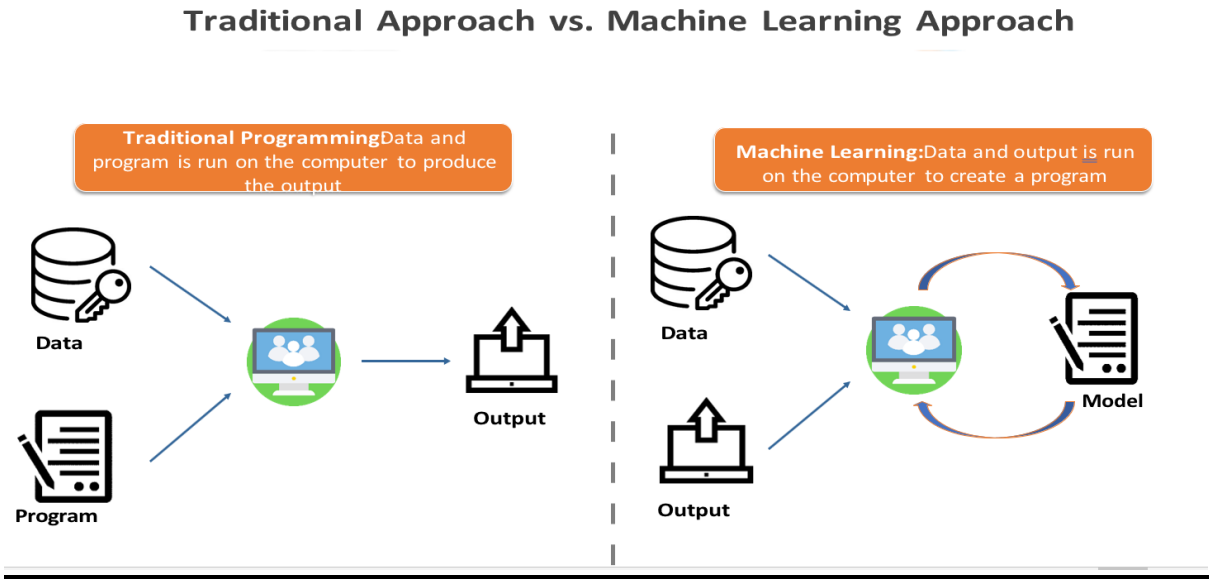
2. Applications of Machine Learning

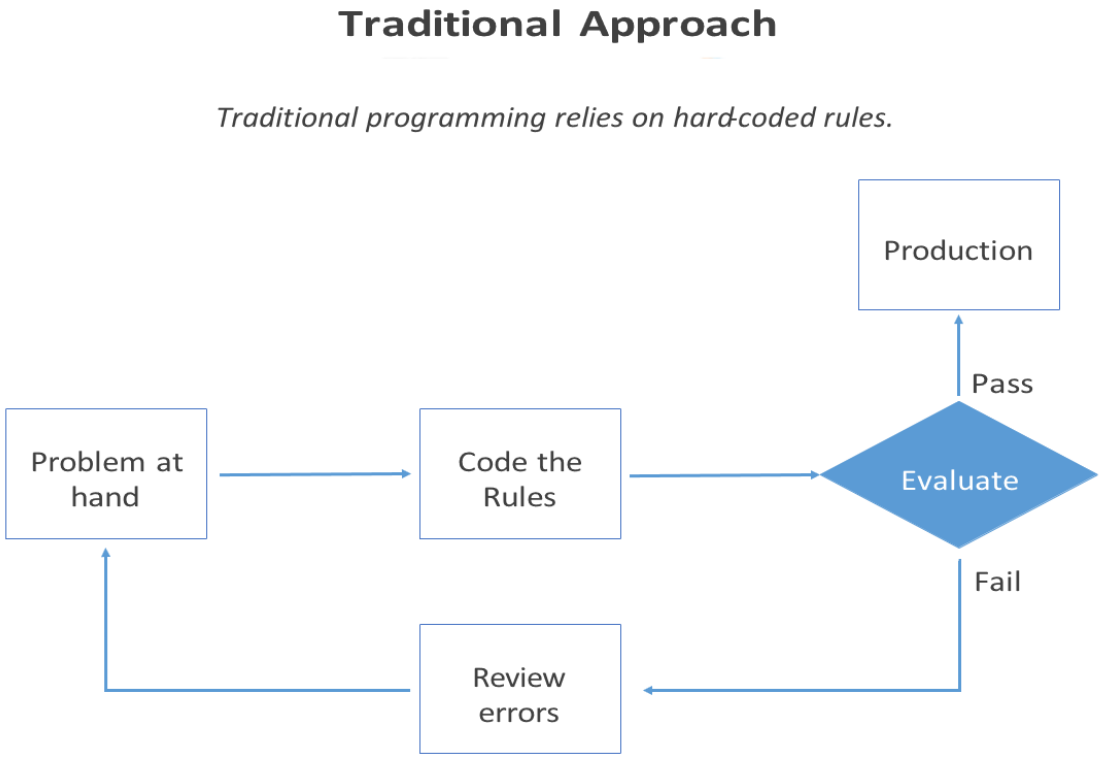


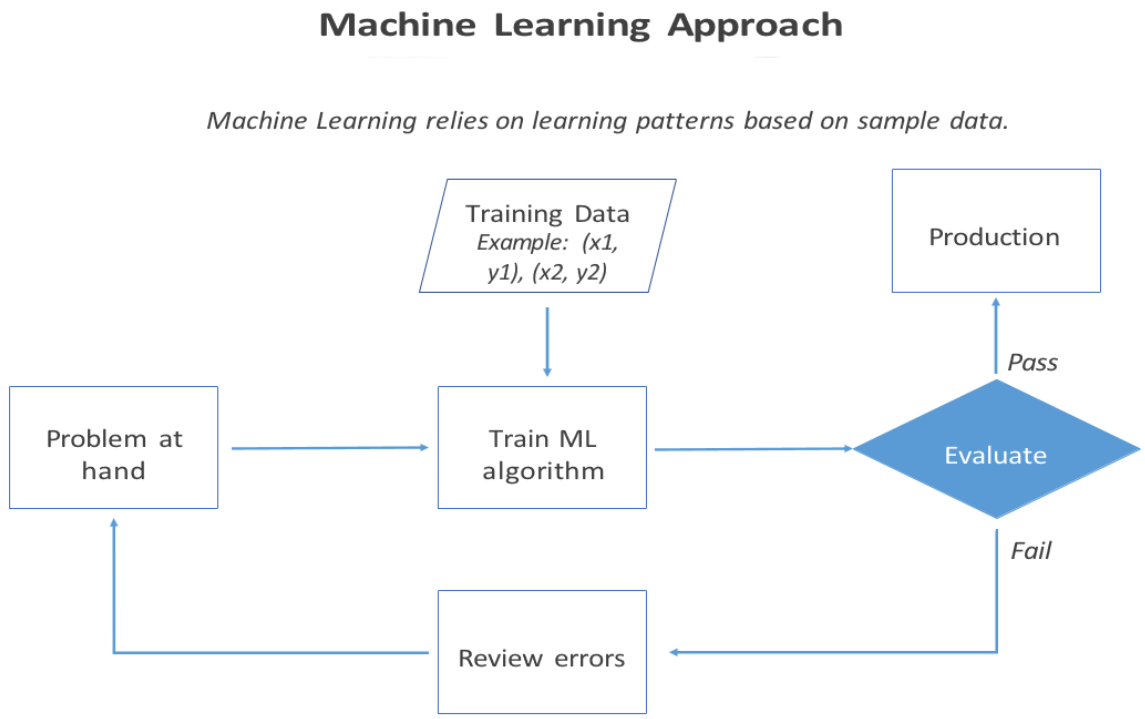
- **Image Processing**
  - Optical Character Recognition (OCR)
  - Self-driving cars
  - Image tagging and recognition
- **Robotics**
  - Industrial robotics
  - Human simulation
- **Data Mining**
  - Association rules
  - Anomaly detection
  - Grouping and Predictions
- **Video games**
  - Pokémon
  - PUBG
- **Text Analysis**
  - Spam Filtering
  - Information Extraction
  - Sentiment Analysis
- **Healthcare**
  - Emergency Room & Surgery
  - Research
  - Medical Imaging & Diagnostics

3. Technology Learnt

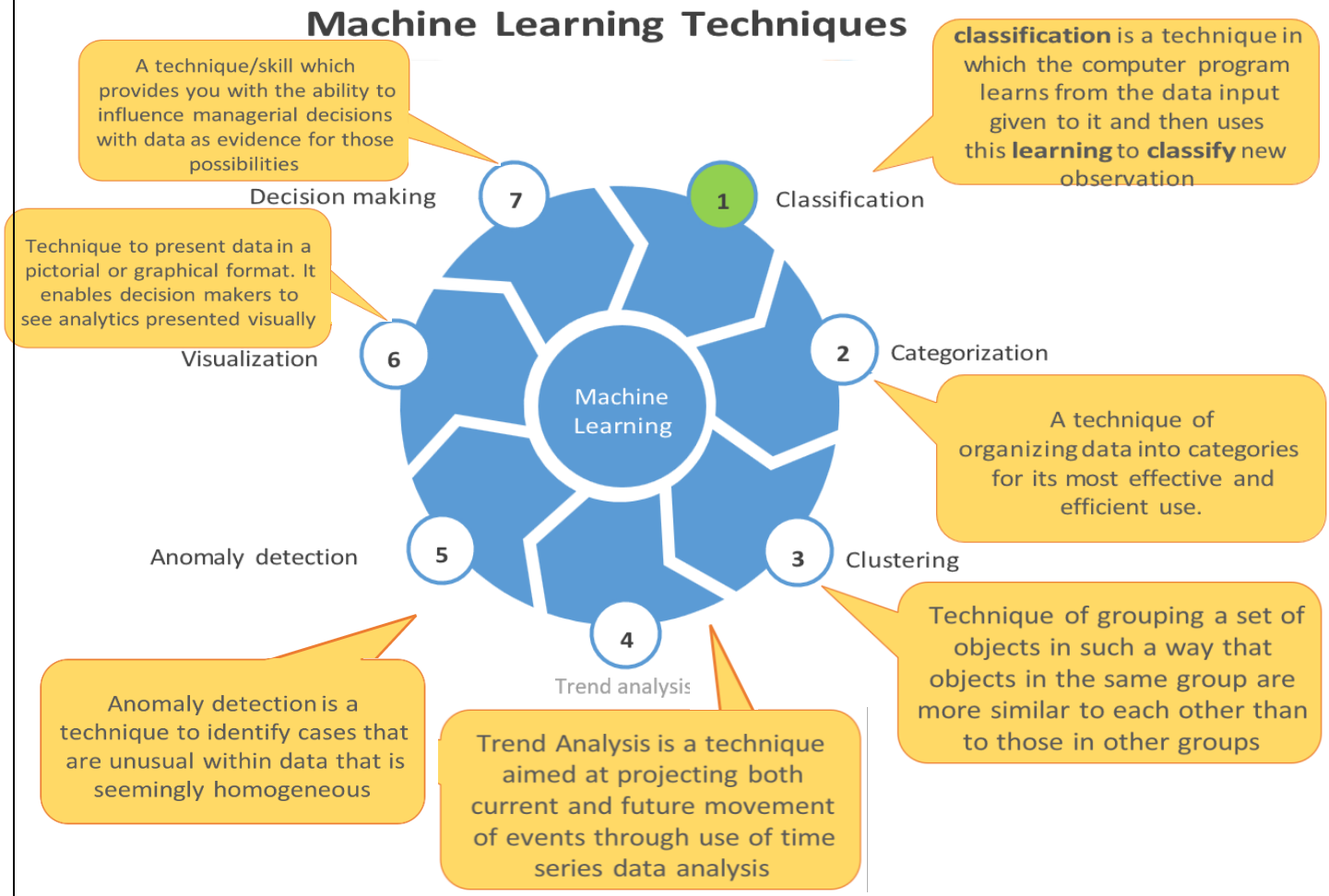
3.1 Machine Learning Algorithms







3.2 Techniques of Machine Learning



3.3 Data Preprocessing

3.3.1. Data Preparation

- **Data Preparation Process**

  Machine Learning depends largely on test data.

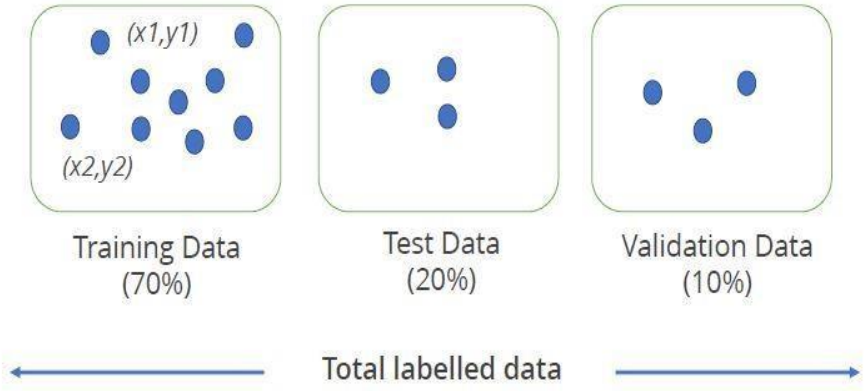
  Data preparation involves data selection, filtering, transformation, etc.

  Data preparation is a crucial step to make it suitable for ML.

  A large amount of data is generally required for the most common forms of ML.

- **Types of Data**

  Labelled Data or Training Data



  Unlabeled Data

  Test Data

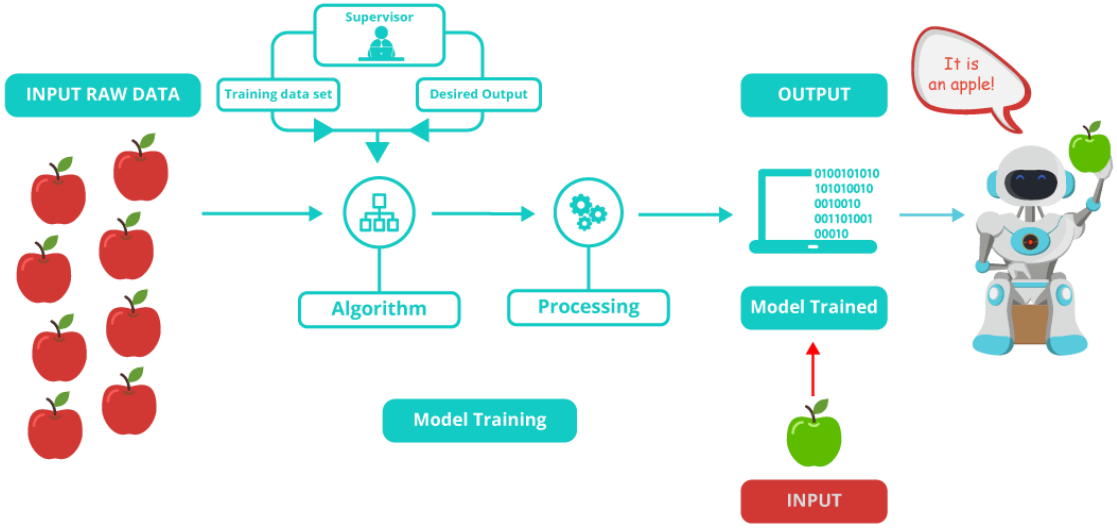
  Validation Data

<u>Machine Learning can learn from labelled data (known as supervised learning) or unlabeled data (known as unsupervised learning).</u>

3.4 Supervised Learning

✞ Define Supervised Learning

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.



In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal).

✞ Supervised Learning Flow

Data Preparation

- Clean data
- Label data (x, y)
- Feature Engineering
- Reserve 80% of data for Training (Train_X) and 20% for Evaluation (Train_E)

Training Step

- Design algorithmic logic
- Train the model with Train X
- Derive the relationship between x and y, that is, y = f(x)

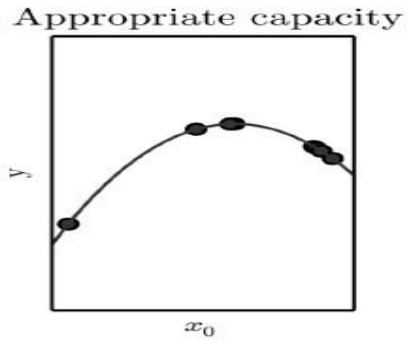
Evaluation or Test Step

- Evaluate or test with Train E
- If accuracy score is high, you have the final learned algorithm y = f(x)
  If accuracy score is low, go back to training step

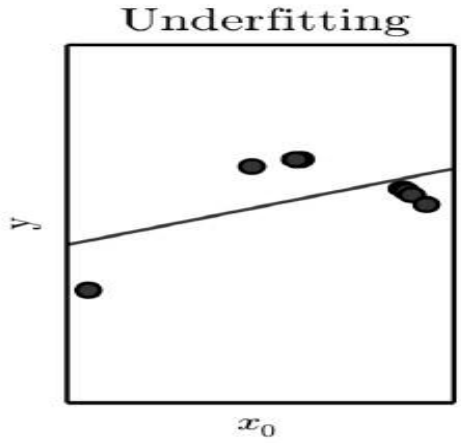
✞ Testing the Algorithms

Once the algorithm is trained, test it with test data (a set of data instances that do not appear in the training set).

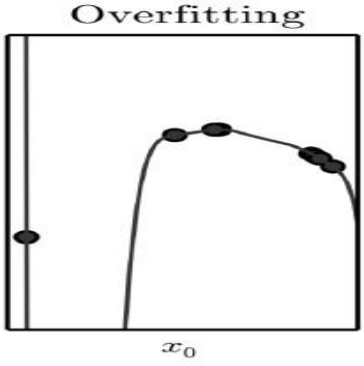
A well-trained algorithm can predict well for new test data.



If the learning is poor, we have an underfitted situation. The algorithm will not work well on test data. Retraining may be needed to find a better fit.



If learning on training data is too intensive, it may lead to overfitting–a situation where the algorithm is not able to handle new testing data that it has not seen before. The technique to keep data generic is called **regularization.**



✞ **Examples of Supervised Learning**

Voice Assistants
Gmail Filters
Weather Apps

3.4.1 Supervised learning – Classification

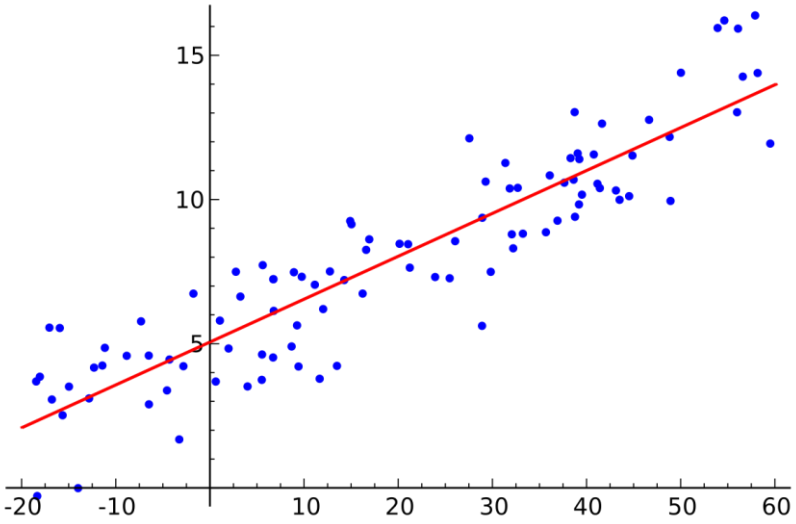
3.4.1.1 Regression

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables.

It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').

More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

3.4.1.1.1 Linear Regression



- Linear regression is a linear approach for modeling the relationship between a scalar dependent variable y and an independent variable x.

$$\hat{y} = \boldsymbol{w}^\top \boldsymbol{x}$$

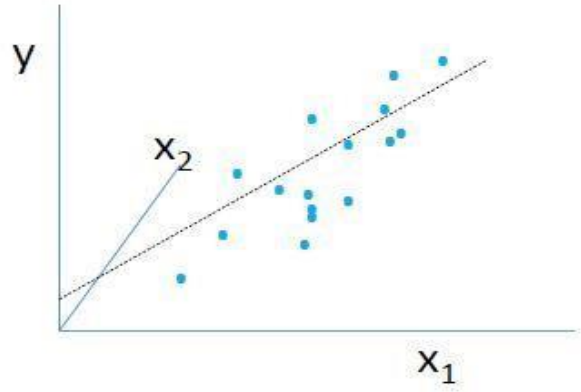
- where x, y, w are vectors of real numbers and w is a vector of weight parameters.
- The equation is also written as:

$$\mathbf{y = wx + b}$$

- where b is the bias or the value of output for zero input

3.4.1.1.2 Multiple Linear Regression

It is a statistical technique used to predict the outcome of a response variable through several explanatory variables and model the relationships between them.

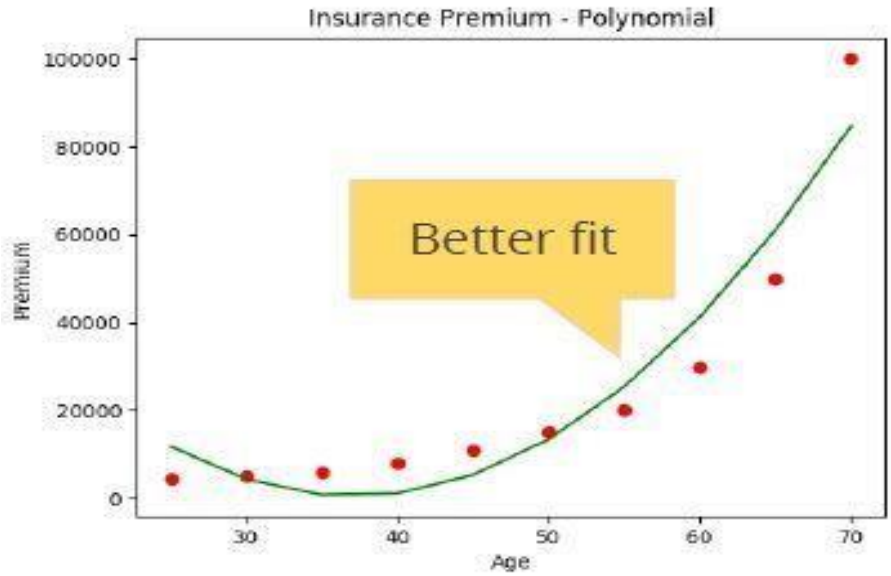


The graph shows dependent variable y plotted against two independent variables x_1 and x_2. It is shown in 3D. More independent variables (if involved) will increase the dimensions further.

- It represents line fitment between multiple inputs and one output, typically:

$$y = w_1x_1 + w_2x_2 + b$$

3.4.1.1.3 Polynomial Regression



- Polynomial regression is applied when data is not formed in a straight line.
- It is used to fit a linear model to non-linear data by creating new features from powers of non-linear features.

Example: Quadratic features

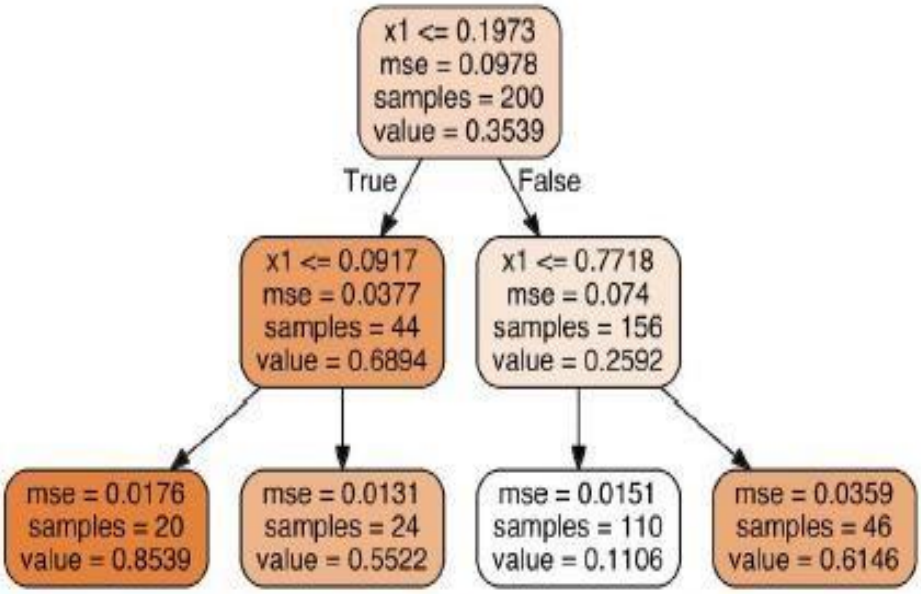
$$x_2' = x_2^2$$

$$y = w_1x_1 + w_2x_2^2 + 6$$

$$= w_1x_1 + w_2x_2' + 6$$

3.4.1.1.4 Decision Tree Regression

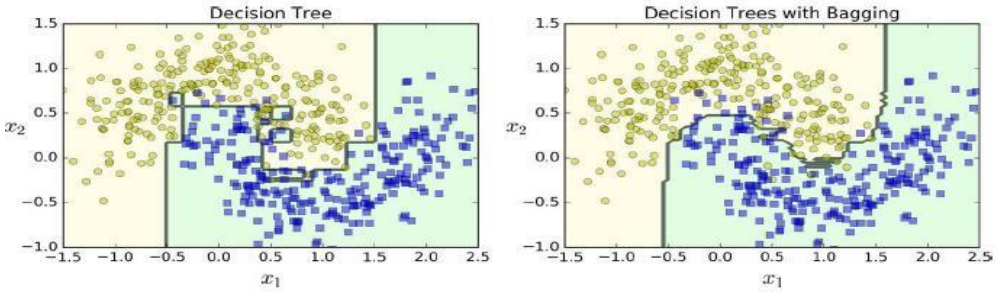
- A decision tree is a graphical representation of all the possible solutions to a decision based on a few conditions. o Decision Trees are non-parametric models, which means that the number of parameters is not determined prior to training. Such models will normally overfit data.
- In contrast, a parametric model (such as a linear model) has a predetermined number of parameters, thereby reducing its degrees of freedom. This in turn prevents overfitting.



- **max_depth** –limit the maximum depth of the tree
- **min_samples_split** –the minimum number of samples a node must have before it can be split
- **min_samples_leaf** –the minimum number of samples a leaf node must have o **min_weight_fraction_leaf** –same as min_samples_leaf but expressed as a fraction of total instances
- **max_leaf_nodes** –maximum number of leaf nodes
- **max_features** –maximum number of features that are evaluated for splitting at each node

3.4.1.1.5 Random Forest Regression

- Ensemble Learning uses the same algorithm multiple times or a group of different algorithms together to improve the prediction of a model.

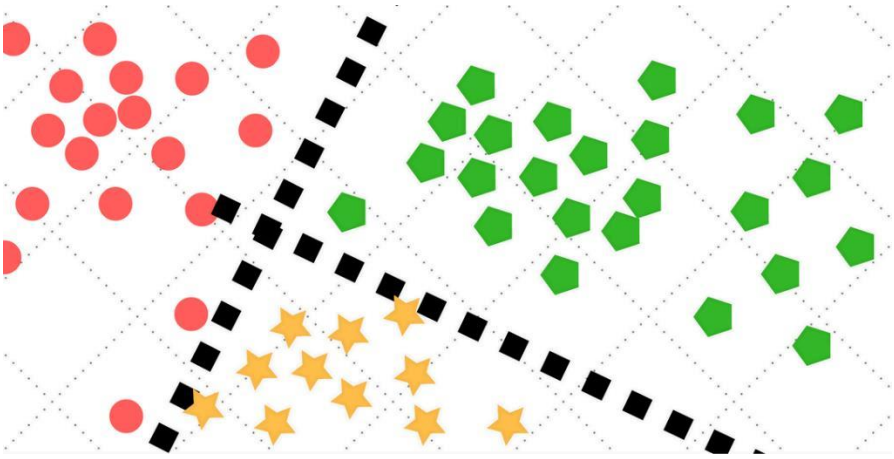


A single decision tree vs. a bagging ensemble of 500 trees

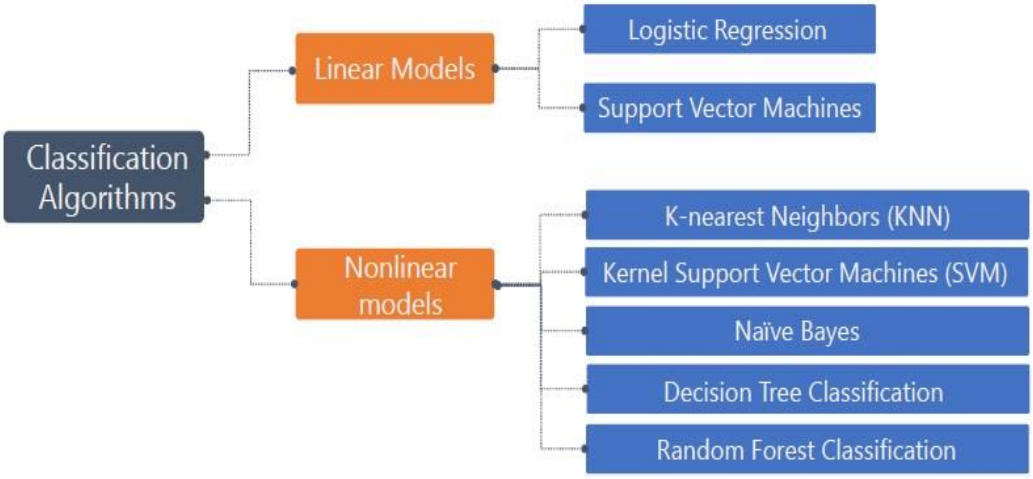
- Random Forests use an ensemble of decision trees to perform regression tasks.

3.4.1.2 Classification

- It specifies the class to which data elements belong to.
- It predicts a class for an input variable.
- It is best used when the output has f inite and discreet values.



- There are 2 types of classification, **binomial and multi -class** .

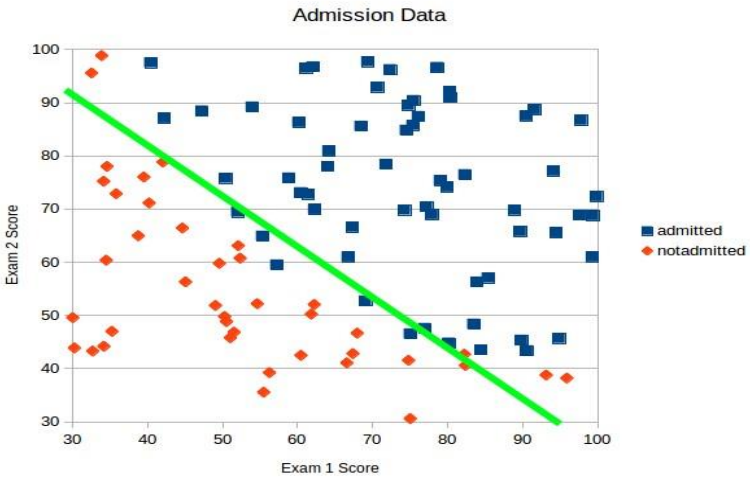


3.4.1.2.1 Linear Models

3.4.1.2.1.1 Logistic Regression

This method is widely used for binary classification problems. It can also be extended to multi-class classification problems.

A binary dependent variable can have only **two values**, like 0 or 1, win or lose, pass or fail, healthy or sick, etc.

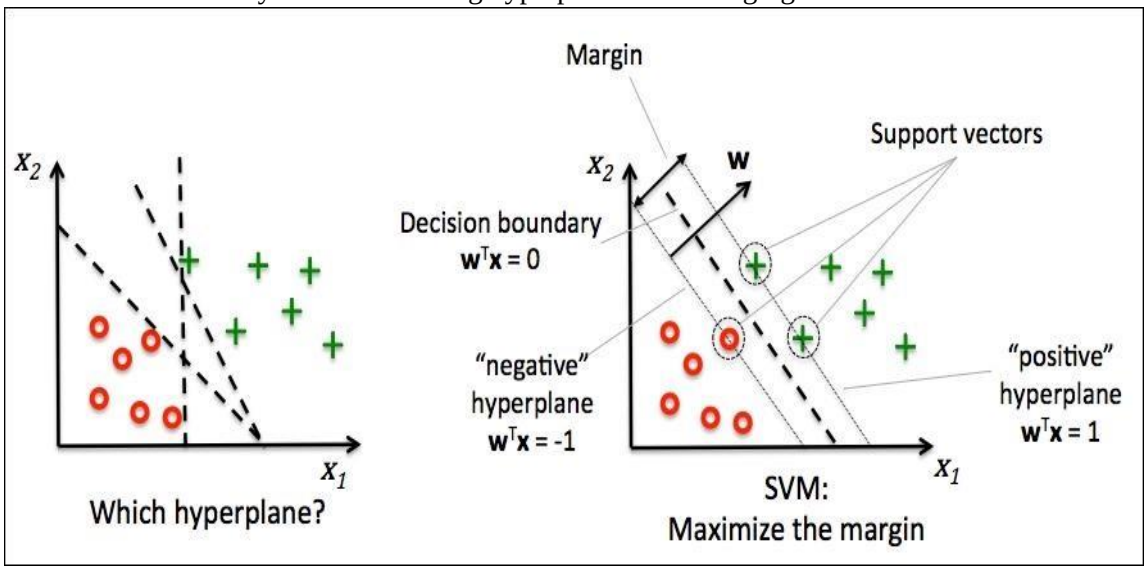


The probability in the logistic regression is often represented by the **Sigmoid function** (also called the **logistic function** or the **S-curve**)

$$S(t) = \frac{1}{1 + e^{-t}}$$

3.4.1.2.1.2 Support Vector machines

- SVMs are very versatile and are also capable of performing linear or nonlinear classification, regression, and outlier detection.
- They involve detecting hyperplanes which segregate data into classes.

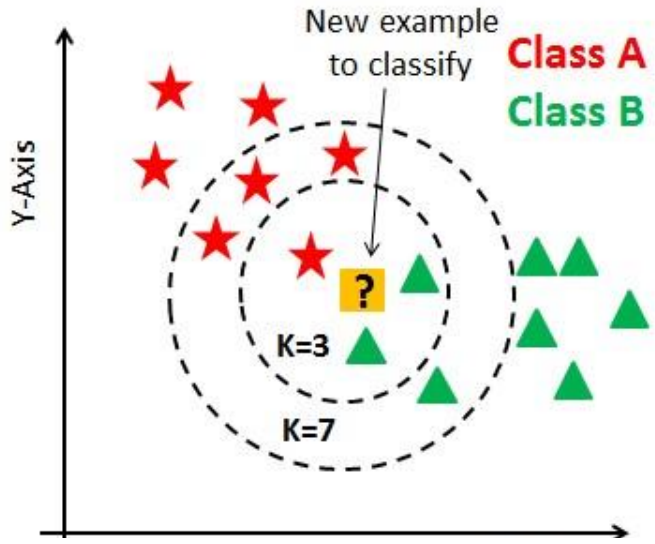


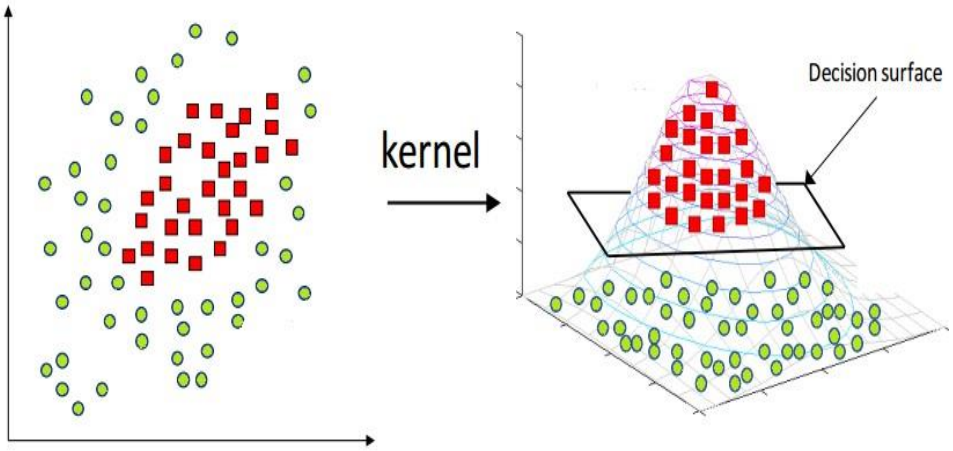
- The optimization objective is to find "**maximum margin hyperplane**" that is farthest from the closest points in the two classes (these points are called support vectors).

3.4.1.2.2 Nonlinear Models

3.4.1.2.2.1 K-Nearest Neighbors (KNN)

- K-nearest Neighbors algorithm is used to assign a data point to clusters based on similarity measurement.





A new input point is classified in the category such that it has the **greatest number of neighbors** from that category.

3.4.1.2.2.2. Kernel Support Vector Machines (SVM)

Kernel SVMs are used for classification of nonlinear data. In the chart, nonlinear data is projected into a higher dimensional space via a mapping function where it becomes linearly separable. A reverse projection of the higher dimension back to original feature space takes it back to nonlinear shape.

3.4.1.2.2.3. Naïve Bayes

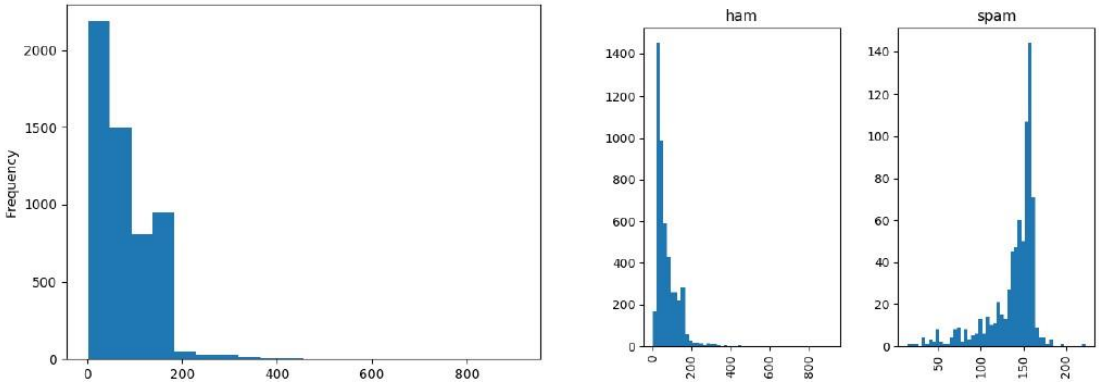
- According to Bayes model, the conditional probability P(Y|X) can be calculated as:

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

- This means you have to estimate a very large number of P(X|Y) probabilities for a relatively small vector space X.

Naïve Bayes Classifier for SMS Spam Detection

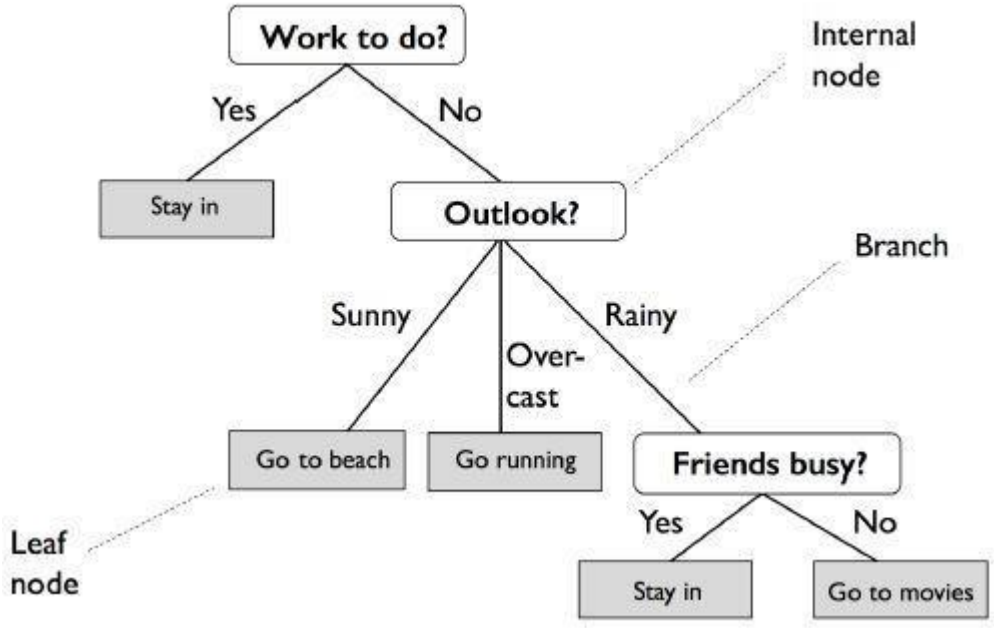
The message lengths and their frequency (in the training data set) are as shown below:



3.4.1.2.2.4. Decision Tree Classification

The advantage of decision trees is that they require very little data preparation. They do not require feature scaling or centering at all.

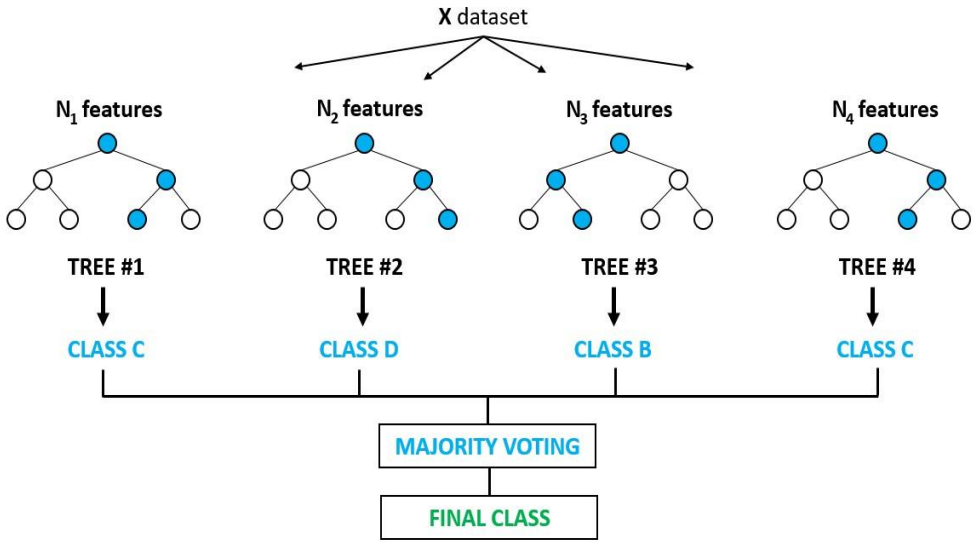
They are also the fundamental components of Random Forests, one of the most powerful ML algorithms.



Start at the tree root and split the data on the feature using the decision algorithm, resulting in the **largest information gain** (IG).

3.4.1.2.2.5. Random Forest Classification

- Random decision forests correct for decision trees' habit of overfitting to their training set.

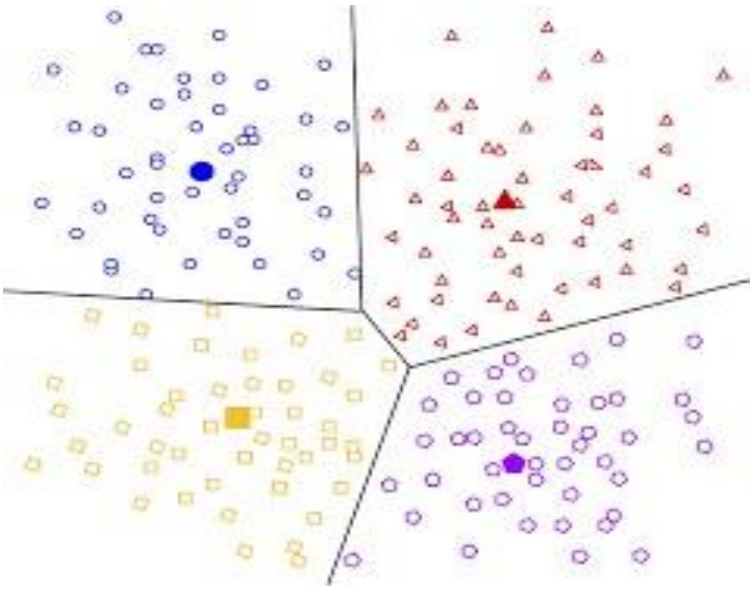


- Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

3.5. Unsupervised learning

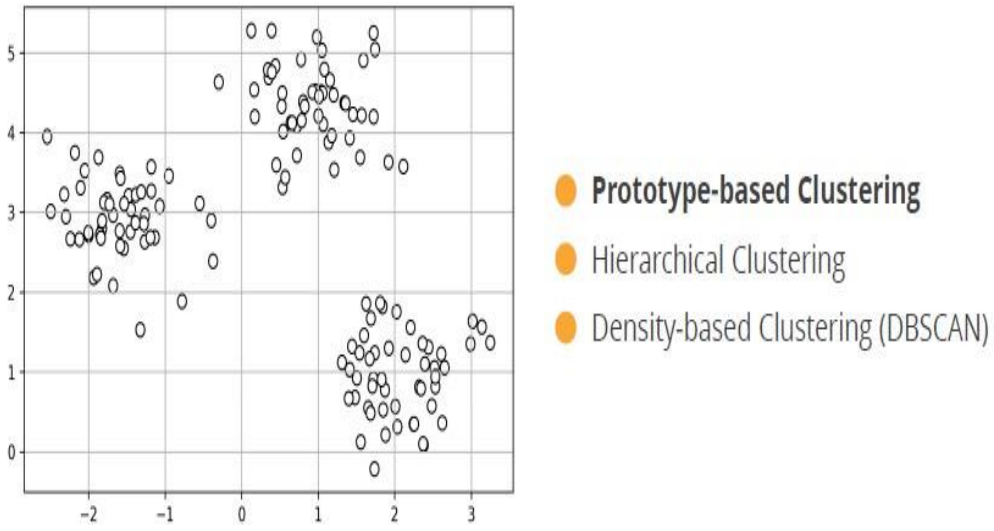
3.5.1 Clustering

3.5.1.1 Clustering Algorithms



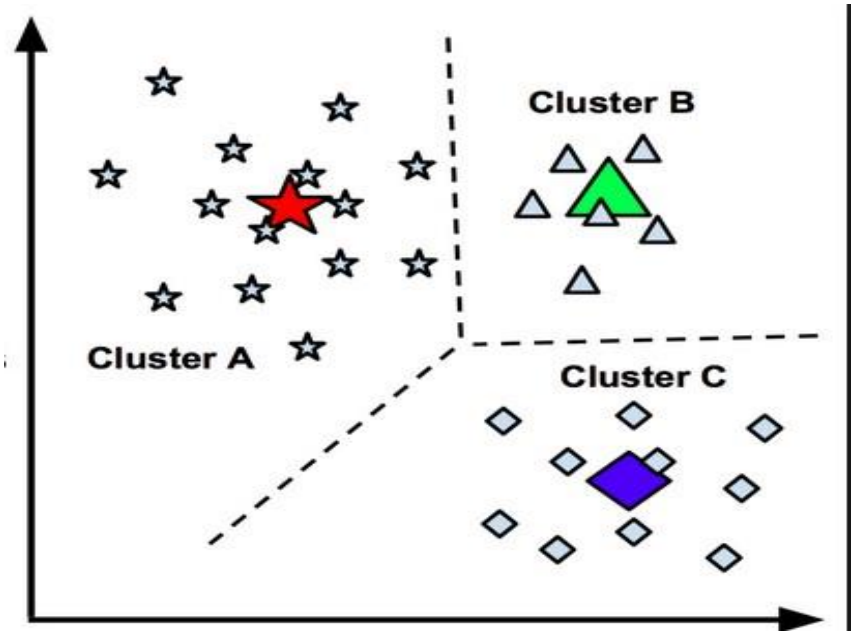
✞ **Clustering means**

✓ Clustering is a Machine Learning technique that involves the grouping of data points.



- **Prototype Based Clustering**
  - Prototype-based clustering assumes that most data is located near prototypes; example: centroids (average) or medoid (most frequently occurring point)
  - K-means, a Prototype-based method, is the most popular method for clustering that involves:
    - Training data that gets assigned to matching cluster based on similarity
    - Iterative process to get data points in the best clusters possible

3.5.1.2 K-means Clustering

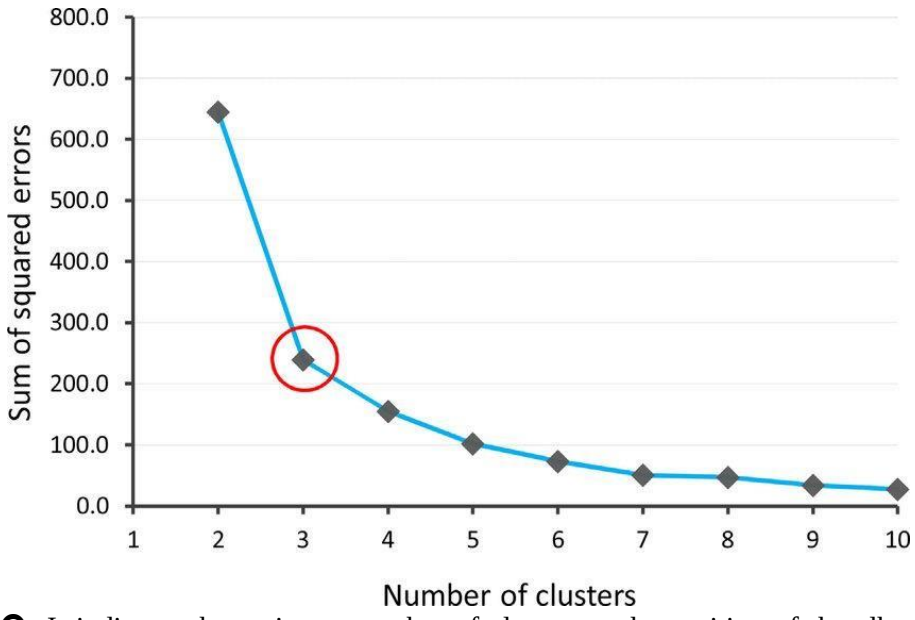


- **K-means Clustering Algorithm**
  - Step 1: randomly pick k centroids
  - Step 2: assign each point to the nearest centroid
  - Step 3: move each centroid to the center of the respective cluster
  - Step 4: calculate the distance of the centroids from each point again

Step 5: move points across clusters and re-calculate the distance from the centroid

Step 6: keep moving the points across clusters until the Euclidean distance is minimized

- **Elbow Method**
  - One could plot the Distortion against the number of clusters K. Intuitively, if K increases, distortion should decrease. This is because the samples will be close to their assigned centroids. This plot is called the Elbow method.



  - It indicates the optimum number of clusters at the position of the elbow, the point where distortion begins to increase most rapidly.

- **Euclidian Distance**

✓ K-means is based on finding points close to cluster centroids. The distance between two points x and y can be measured by the squared Euclidean distance between them in an m-dimensional space.

Here, *j* refers to *j*-th dimension (or *j*-th feature) of the data point.

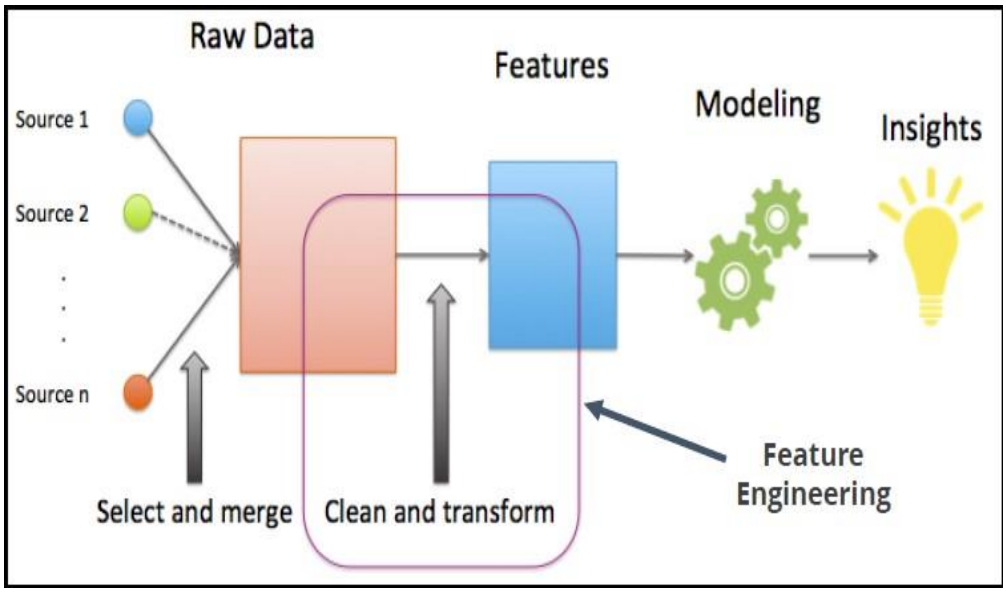
$$d(\boldsymbol{x}, \boldsymbol{y})^2 = \sum_{j=1}^{m} \left(x_j - y_j\right)^2 = \left\|\boldsymbol{x} - \boldsymbol{y}\right\|_2^2$$

- **Examples of K-means Clustering**
  - Grouping articles (example: Google news)
  - Grouping customers who share similar interests
  - Classifying high risk and low risk patients from a patient pool

3.6. Feature Engineering

- **Define Feature Engineering**

The transformation stage in the data preparation process includes an important step known as Feature Engineering.



Feature Engineering refers to selecting and extracting right features from the data that are relevant to the task and model in consideration.

- **Aspects of Feature Engineering**

Feature Selection

Most useful and relevant features are selected from the available data

Feature Addition

New features are created by gathering new data

Feature Extraction

Existing features are combined to develop more useful ones

Feature Filtering

Filter out irrelevant features to make the modelling step easy

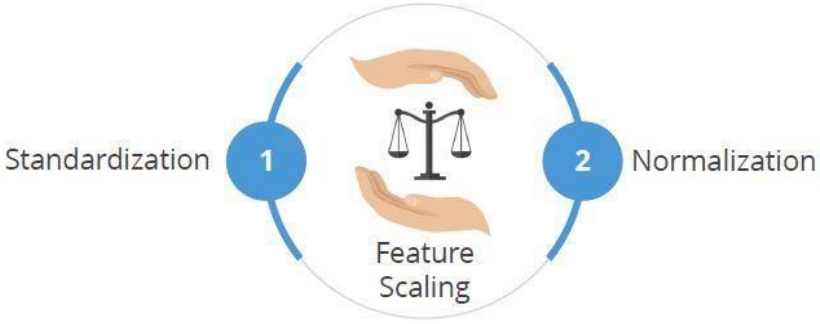
3.6.1 Feature Scaling

- **Define Feature Scaling**

Feature scaling is an important step in the data transformation stage of data preparation process.

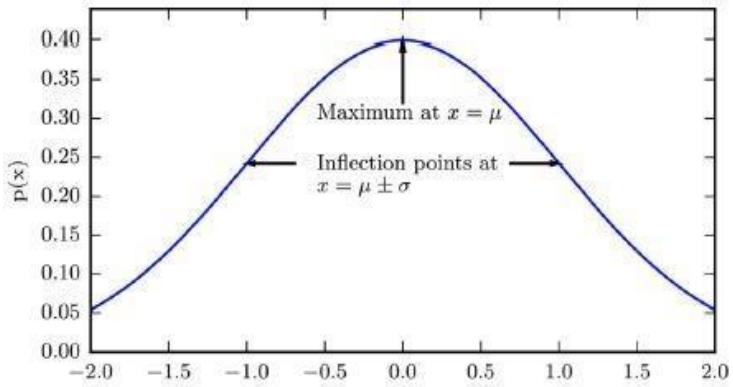
Feature Scaling is a method used in Machine Learning for standardization of independent variables of data features.

✞ Techniques of Feature Scaling



Standardization

- ✦ Standardization is a popular feature scaling method, which gives data the property of a standard normal distribution (also known as Gaussian distribution).
- ✦ All features are standardized on the normal distribution (a mathematical model).
- ✦ The mean of each feature is centered at zero, and the feature column has a standard deviation of one.



$$x'_j = \frac{x_j - \mu_j}{\sigma_j}$$

The ML library scikit-learn implements a class for standardization called StandardScaler, as demonstrated here:

```
>>> from sklearn.preprocessing import StandardScaler
>>> stdsc = StandardScaler()
>>> X_train_std = stdsc.fit_transform(X_train)
>>> X_test_std = stdsc.transform(X_test)
```

✓ **Normalization**

- In most cases, normalization refers to rescaling of data features between 0 and 1, which is a special case of Min-Max scaling.

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{\min}}{x_{\max} - x_{\min}}$$

- In the given equation, subtract the min value for each feature from each feature instance and divide by the spread between max and min.
- In effect, it measures the relative percentage of distance of each instance from the min value for that feature.

The ML library scikit-learn has a MinMaxScaler class for normalization. It implements normalization as explained on the previous slide.

```
>>> from sklearn.preprocessing import MinMaxScaler
>>> mms = MinMaxScaler()
>>> X_train_norm = mms.fit_transform(X_train)
>>> X_test_norm = mms.transform(X_test)
```

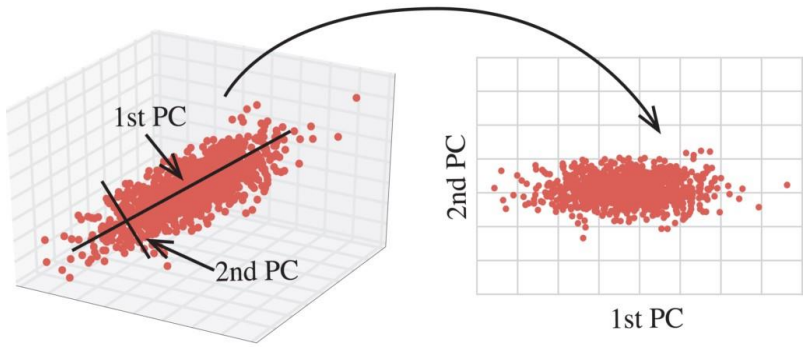
3.6.2. Datasets

- Machine Learning problems often need training or testing datasets.
- A dataset is a large repository of structured data.
- In many cases, it has input and output labels that assist in Supervised Learning.

3.6.3. Dimensionality Reduction with Principal Component Analysis

✞ **Define Dimensionality Reduction**

Dimensionality reduction involves transformation of data to new dimensions in a way that facilitates discarding of some dimensions without losing any key information.



✞ **Define Principal Component Analysis (PCA)**

Principal component analysis (PCA) is a technique for dimensionality reduction that helps in arriving at better visualization models.

1. Let $\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$.
2. Replace each $x^{(i)}$ with $x^{(i)} - \mu$.
3. Let $\sigma_j^2 = \frac{1}{m}\sum_i (x_j^{(i)})^2$
4. Replace each $x_j^{(i)}$ with $x_j^{(i)}/\sigma_j$.

✞ **Applications of PCA**

Noise reduction
Compression
Preprocess

4. Reason for choosing Machine Learning

➢ Learning machine learning brings in better career opportunities

Machine learning is the shining star of the moment.

Every industry looking to apply AI in their domain, studying machine learning opens world of opportunities to develop cutting edge machine learning applications in various verticals – such as cyber security, image recognition, medicine, or face recognition.

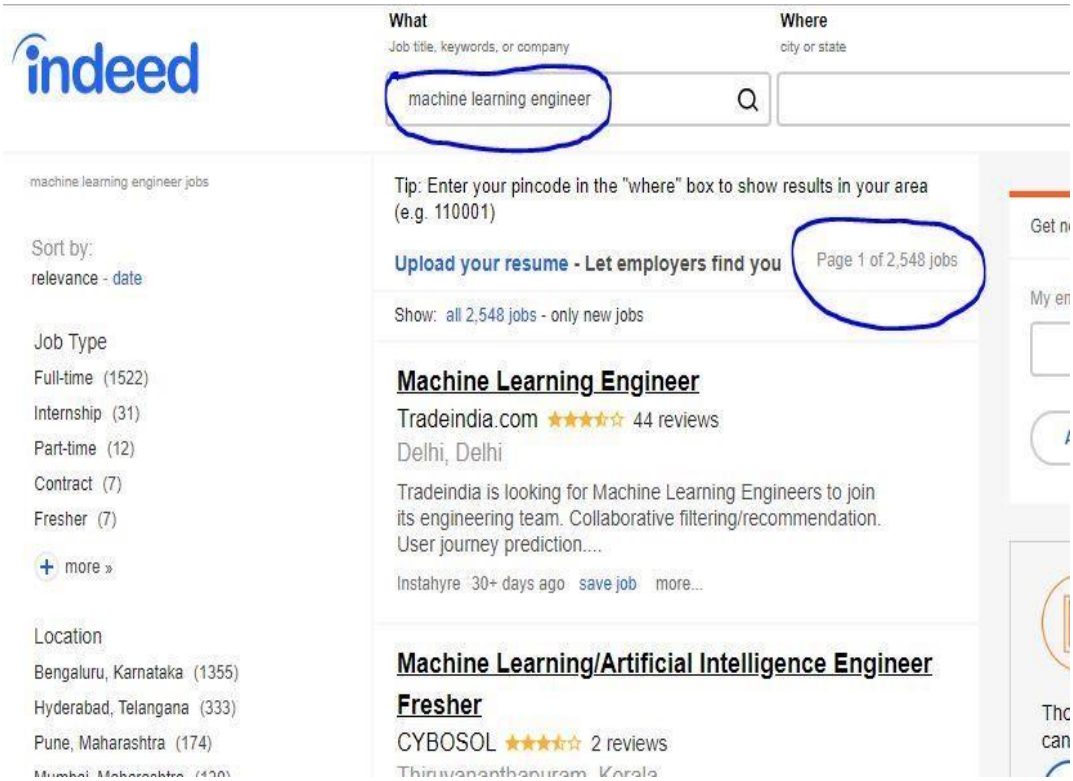
Several machine learning companies on the verge of hiring skilled ML engineers, it is becoming the brain behind business intelligence.

➢ Machine Learning Jobs on the rise

The major hiring is happening in all top tech companies in search of those special kind of people (machine learning engineers) who can build a hammer (machine learning algorithms).

The job market for machine learning engineers is not just hot but it's sizzling.

Machine Learning Jobs on Indeed.com - 2,500+(India) & 12,000+(US)



5. Learning Outcome

- Have a good understanding of the fundamental issues and challenges of machine learning: data, model selection, model complexity, etc.
- Have an understanding of the strengths and weaknesses of many popular machine learning approaches.
- Appreciate the underlying mathematical relationships within and across Machine Learning algorithms and the paradigms of supervised and un-supervised learning.
- Be able to design and implement various machine learning algorithms in a range of real-world applications.
- Ability to integrate machine learning libraries and mathematical and statistical tools with modern technologies
- Ability to understand and apply scaling up machine learning techniques and associated computing techniques and technologies.

6. Project Description

Mercedes-Benz Greener Manufacturing

DESCRIPTION

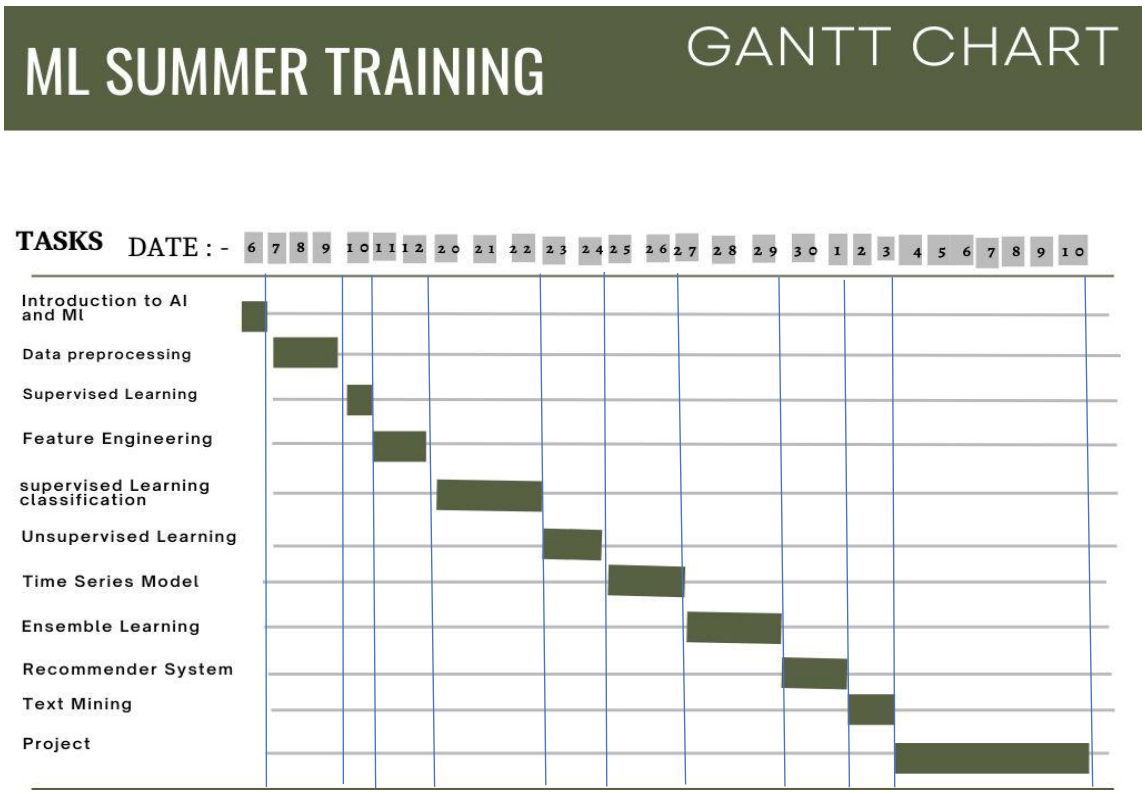
Reduce the time a Mercedes-Benz spends on the test bench.

Problem Statement Scenario: Since the first automobile, the Benz Patent Motor Car in 1886, Mercedes-Benz has stood for important automotive innovations. These include the passenger safety cell with a crumple zone, the airbag, and intelligent assistance systems. Mercedes-Benz applies for nearly 2000 patents per year, making the brand the European leader among premium carmakers. Mercedes-Benz is the leader in the premium car industry. With a huge selection of features and options, customers can choose the customized Mercedes-Benz of their dreams.

To ensure the safety and reliability of every unique car configuration before they hit the road, the company's engineers have developed a robust testing system. As one of the world's biggest manufacturers of premium cars, safety and efficiency are paramount on Mercedes-Benz's production lines. However, optimizing the speed of their testing system for many possible feature combinations is complex and time-consuming without a powerful algorithmic approach.

You are required to reduce the time that cars spend on the test bench. Others will work with a dataset representing different permutations of features in a Mercedes-Benz car to predict the time it takes to pass testing. Optimal algorithms will contribute to faster testing, resulting in lower carbon dioxide emissions without reducing Mercedes-Benz's standards.

7. Gantt Chart



8. References

https://www.simplilearn.com/

https://www.wikipedia.org/

https://towardsdatascience.com/

https://www.expertsystem.com/

https://www.coursera.org/

https://www.edureka.co/

https://www.forbes.com/

https://medium.com/

https://www.google.com/