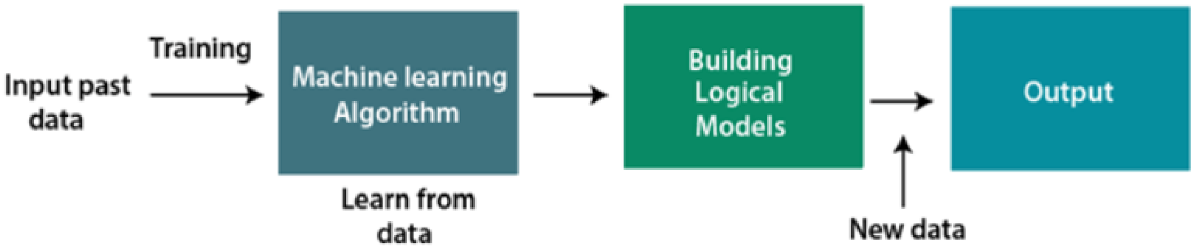
Machine Learning :

- Machine learning allows computers to automatically learn from previous data. For building mathematical models and making predictions based on historical data or information, machine learning employs a variety of algorithms. It is currently being used for a variety of tasks, including speech recognition, email filtering, auto-tagging on Facebook, a recommender system, and image recognition.
- A subset of artificial intelligence known as machine learning focuses primarily on the creation of algorithms that enable a computer to independently learn from data and previous experiences. Arthur Samuel first used the term "machine learning" in 1959.
- Without being explicitly programmed, machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things.
- A machine can learn if it can gain more data to improve its performance.

Working of ML :

- A machine learning system builds prediction models, learns from previous data, and predicts the output of new data whenever it receives it. The amount of data helps to build a better model that accurately predicts the output, which in turn affects the accuracy of the predicted output.



Features :

- Machine learning is much similar to data mining as it also deals with the huge amount of the data.
- It is a data-driven technology.
- It can learn from past data and improve automatically.
- Machine learning uses data to detect various patterns in a given dataset.

Need :

- The need of machine learning is getting increased because it can able to perform tasks that are too complex for human to direct implementation.

- Humans are constrained by our inability to manually access vast amounts of data; as a result, we require computer systems, which is where machine learning comes in to simplify our lives.
- By providing them with a large amount of data and allowing them to automatically explore the data, build models, and predict the required output, we can train machine learning algorithms. **The cost function** can be used to determine the amount of data and the machine learning algorithm's **performance.** We can save both time and money by using machine learning.

- Used in scnarios like : Rapid increment in the production of data, Solving complex problems, which are difficult for a human, Decision making in various sector including finance, Finding hidden patterns and extracting useful information from data.

Classification of machine Learning :

1)Supervised ML : In supervised learning, sample labeled data are provided to the machine learning system for training, and the system then predicts the output based on the training data.

e.g : Spam Filtering ,Forcasting share price ,Disease Prediction (Autism),House price prediction.

There are two categories in SML :
1)Regression (deal with real or continuous values)
Algo : linear regression
2)Classification : classify new data into specific category.(binary(2),Muliclass(3 or more)
Algo : logistic regression,Decision Tree,SVM,KNN,Navie bias classfier,Random Forest

2)Unsupervised ML : Unsupervised learning is a learning method in which a machine learns without any supervision. The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns. In unsupervised learning, we don't have a predetermined result.

1)Association :

[It means finding out association between object things.like suppose in market we look things that are related kept with each other like bread->biscuit->milk .it e.g suppose person buying milk so what is possibility of that person to buy bread ,biscuit].

- It tries to discover some interesting relations or associations between the variables of the dataset. It depends on various rules to find interesting relations between variables in the database
- The association rule learning is the most important approach of machine learning, and it is employed in Market Basket analysisIn market basket analysis, it is an approach used by several big retailers to find the relations between items
- e.g In market basket analysis, customer buying habits are analyzed by finding associations between the different items that customers place in their shopping baskets. By discovering such associations, retailers produce marketing methods by analyzing which elements are frequently purchased by users. This association can lead to increased sales by supporting retailers to do selective marketing and plan for their shelf area .
- e.g :apriori algorithm ,F-P growth algorithm.
-

2)Clustering :

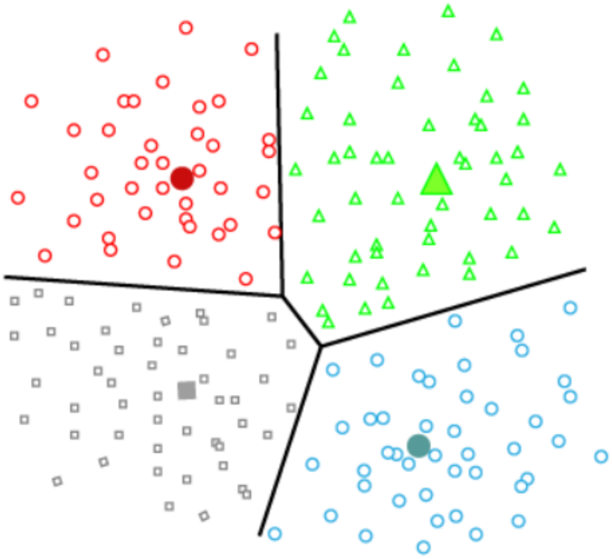
- A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group.
- It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

- After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML system can use this id to simplify the processing of large and complex datasets.
- The clustering technique is commonly used for **statistical data analysis.**
- **Example**: Let's understand the clustering technique with the real-world example of Mall: When we visit any shopping mall, we can observe that the things with similar usage are grouped together. Such as the t-shirts are grouped in one section, and trousers are at other sections, similarly, at vegetable sections, apples, bananas, Mangoes, etc., are grouped in separate sections, so that we can easily find out the things. The clustering technique also works in the same way. Other examples of clustering are grouping documents according to the topic.
- Market Segmentation, Statistical data analysis, Social network analysis ,Image segmentation,Anomaly detection, etc.
- The clustering methods are broadly divided into **Hard clustering** (datapoint belongs to only one group) and **Soft Clustering** (data points can belong to another group also).

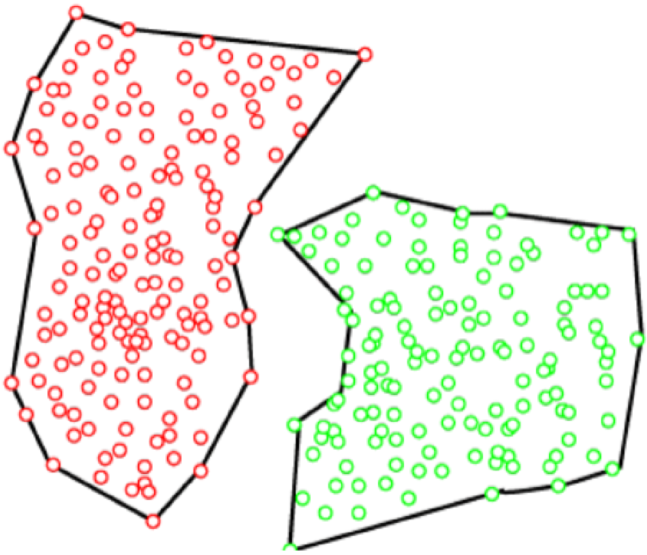
Types :

1)Partioning Clustering : It is a type of clustering that divides the data into non-hierarchical groups. It is also known as the **centroid-based method**. The most common example of partitioning clustering is the K-Means Clustering algorithm.

In this type, the dataset is divided into a set of k groups, where K is used to define the number of pre-defined groups. The cluster center is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid.

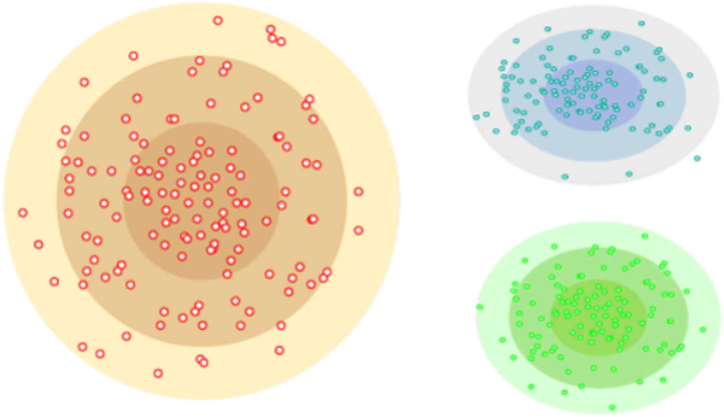


2)Density Based Clustering : The density-based clustering method connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters.

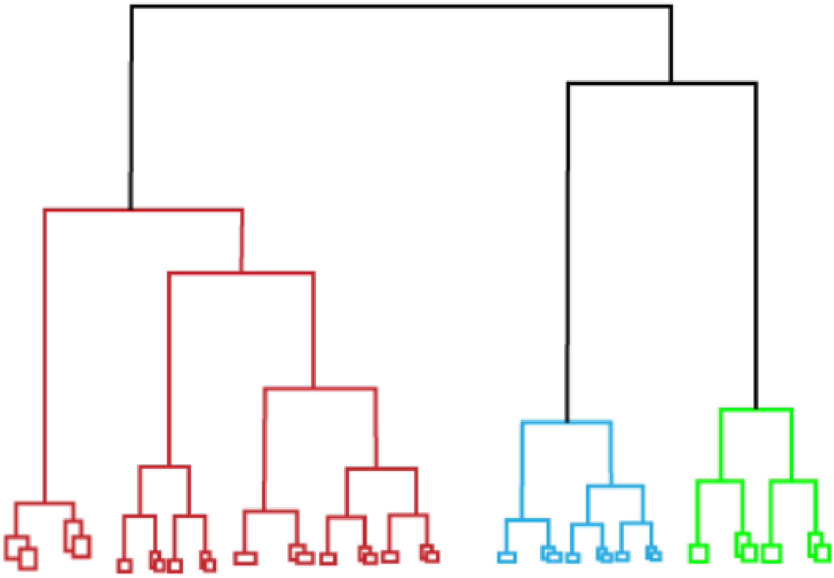


3)Distribution Based Clustering : In the distribution model-based clustering method, the data is divided based on the probability of how a dataset belongs to a particular distribution. The grouping is done by assuming some distributions commonly **Gaussian Distribution**.

The example of this type is the **Expectation-Maximization Clustering algorithm** that uses Gaussian Mixture Models (GMM).



4)Hirearchical Clustering : Hierarchical clustering can be used as an alternative for the partitioned clustering as there is no requirement of pre-specifying the number of clusters to be created. In this technique, the dataset is divided into clusters to create a tree-like structure, which is also called a **dendrogram**. The observations or any number of clusters can be selected by cutting the tree at the correct level. The most common example of this method is the **Agglomerative Hierarchical algorithm**.



5)Fuzzy Clustering : Fuzzy clustering is a type of soft method in which a data object may belong to more than one group or cluster.

e.g Fuzzy k-means algorithm.

Clustering Algorithms :

1. **K-Means algorithm:** The k-means algorithm is one of the most popular clustering algorithms. It classifies the dataset by dividing the

samples into different clusters of equal variances. The number of clusters must be specified in this algorithm. It is fast with fewer computations required, with the linear complexity of O(n).

2. **Mean-shift algorithm:** Mean-shift algorithm tries to find the dense areas in the smooth density of data points. It is an example of a centroid-based model, that works on updating the candidates for centroid to be the center of the points within a given region.
3. **DBSCAN Algorithm:** It stands for Density-Based Spatial Clustering of Applications with Noise. It is an example of a density-based model similar to the mean-shift, but with some remarkable advantages. In this algorithm, the areas of high density are separated by the areas of low density. Because of this, the clusters can be found in any arbitrary shape.
4. **Expectation-Maximization Clustering using GMM:** This algorithm can be used as an alternative for the k-means algorithm or for those cases where K-means can be failed. In GMM, it is assumed that the data points are Gaussian distributed.
5. **Agglomerative Hierarchical algorithm**: The Agglomerative hierarchical algorithm performs the bottom-up hierarchical clustering. In this, each data point is treated as a single cluster at the outset and then successively merged. The cluster hierarchy can be represented as a tree-structure.
6. **Affinity Propagation:** It is different from other clustering algorithms as it does not require to specify the number of clusters. In this, each data point sends a message between the pair of data points until convergence. It has O(N2T) time complexity, which is the main drawback of this algorithm.

Clustering Applications :

- **In Identification of Cancer Cells:** The clustering algorithms are widely used for the identification of cancerous cells. It divides the cancerous and non-cancerous data sets into different groups.
- **In Search Engines:** Search engines also work on the clustering technique. The search result appears based on the closest object to the search query. It does it by grouping similar data objects in one

group that is far from the other dissimilar objects. The accurate result of a query depends on the quality of the clustering algorithm used.

- **Customer Segmentation:** It is used in market research to segment the customers based on their choice and preferences.
- **In Biology:** It is used in the biology stream to classify different species of plants and animals using the image recognition technique.
- **In Land Use:** The clustering technique is used in identifying the area of similar lands use in the GIS database. This can be very useful to find that for what purpose the particular land should be used, that means for which purpose it is more suitable.

3)Reinforcement ML :

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.

The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

1. Action (A): Moves that the agent makes.
2. State (S): Current situation in the environment.
3. Reward (R): Return sent back from the state to evaluate the last action.
4. Policy: Strategy the agent employs for determining the action based on the current state.
5. Q-value: Often called the action value, it defines the estimation of how good the action taken by the agent is at the state.

Algorithms :

1) Q-learning

Starting with Q-Learning, which is a model-free and off-policy RL algorithm that is based on the Bellman Equation. The algorithm uses a Q-table, which is a lookup table that stores the agent's estimated utility or "quality" of taking a certain action in a given state. The agent updates the Q-values to maximise it in the table through trial and error and, eventually, it converges to the optimal policy.

Algorith not follow single policy like SARSA.

$$Q(s,a) \leftarrow Q(s, a) + \alpha(r + \gamma \max_{a}' (s', a') - Q(s, a))$$

Q(s, a) is the current estimate of the utility of taking action '**a**' in state '**s'**.
α is the learning rate, a value between 0 and 1 that determines the relative weight of the current estimate and the new information.
r is the reward received after taking action 'a' in state 's'.
γ is the discount factor, a value between 0 and 1 that determines the importance of future rewards.
s' is the next state after taking action 'a' in state 's'.
a' is the action selected in state **s'.**

2)Deep Q learning : (neural network is used)

SARSA, or State Action Reward State Action, is similar to Q-Learning but the key difference is that it is an on-policy algorithm, and is often denoted as the 'on-policy Q-learning'. This implies that through this algorithm, Q-value is derived from the action performed by current policy, which is in contrast to Q-learning algorithm that has no constraint over the next action.

The abbreviated name, SARSA, denotes a sequence that the algorithm starts in a state (S), takes action (A), and then the reward is generated (R). This updates the Q-function (value).

$$Q(s,a) \leftarrow Q(s, a) + \alpha(r + \gamma Q(s', a') - Q(s, a))$$

Now, these obtained Q-values are stored in a table and the one with the highest values is chosen by the policy by observing its current state, leading to a new state, and it continues on for next states. Q-values keep getting updated till we find a good policy. But it is constrained by a single policy, thus Q-learning offers more scope for value selection.

Both Q-learning and SARSA are tabular methods and, due to vast memory consumption and failure to visit all states and actions while training, do not scale well for large state and action spaces. This is where neural networks come in.

3)SARSA(state action reward state action)

DQN is an extension of Q-Learning that leverages neural networks to estimate the Q-value function. This enables it to move beyond the limitations of Q-learning, which cannot estimate value for the unseen states. In 2013, DeepMind even applied DQN to the Atari Game.

The neural network is trained based on the Q-learning update equation:

$$Q\_\theta(s, a) \leftarrow Q\_\theta(s, a) + \alpha((r + \gamma \max_{a}' Q_{\boxed{?}a}\square(s', a')) - Q\_\theta(s, a))$$

The standard Q-learning technique finds the optimal values, which are the highest rewards, and then developers decide the optimal function. DQN allows direct approximation of the optimal value using two essential techniques:

1. Experience Relay: To solve the problem of high correlation and less data efficiency, experience relay allows the sample transitions (moves from one state to another by actions) to be stored. This allows detection of trends, which are then randomly selected from the pool to update the knowledge.
2. Target Networks: These help determine if the output reward is already not the best one. This is achieved by going back to the last updated output and considering Q-values as the target.

Applications :

1)Automated Robots: Some restaurants use robots to deliver food to tables. Grocery stores are using robots to identify where shelves are low and order more product. In common settings, automated robots have been used thus far to assemble products; inspect for defects; count, track, and manage inventory; deliver goods; travel long and short distances; input, organize, and report on data; and grasp and handle objects of all different shapes and sizes. As we continue to test robotic abilities, new features are being introduced to expand their potential.

2)Natural Language Processing : Predictive text, text summarization, question answering, and machine translation are all examples of natural language processing (NLP) that uses reinforcement learning.

3)Marketing and Advertisement : Both brands and consumers can use reinforcement learning to their benefit. For brands selling to target audiences, they can use real-time bidding platforms, A/B testing, and automatic ad optimization. This means that they can place a series of advertisements in the marketplace and the host will automatically serve the best-performing ads in the best spots for the lowest prices.

4)Image Processing :

- Robots equipped with visual sensors from to learn their surrounding environment
- Scanners to understand and interpret text
- Image pre-processing and segmentation of medical images, like CT Scans
- Traffic analysis and real-time road processing by video segmentation and frame-by-frame image processing
- CCTV cameras for traffic and crowd analytics.

5)Recommandation :

Recommendation systems also analyze past behaviors to try to predict future ones. So if, for example, a hundred people who bought ski pants then went on to buy ski boots, a company's system learns to send ads for ski boots to anyone who just bought ski pants. If the ads are unsuccessful, they might try to display ads for ski jackets, instead, and see how the results compare.

6)Gaming : RL agents are also used in bug detection and game testing. This is due to its ability to run a large number of iterations without human input, stress testing, and creating situations for potential bugs.

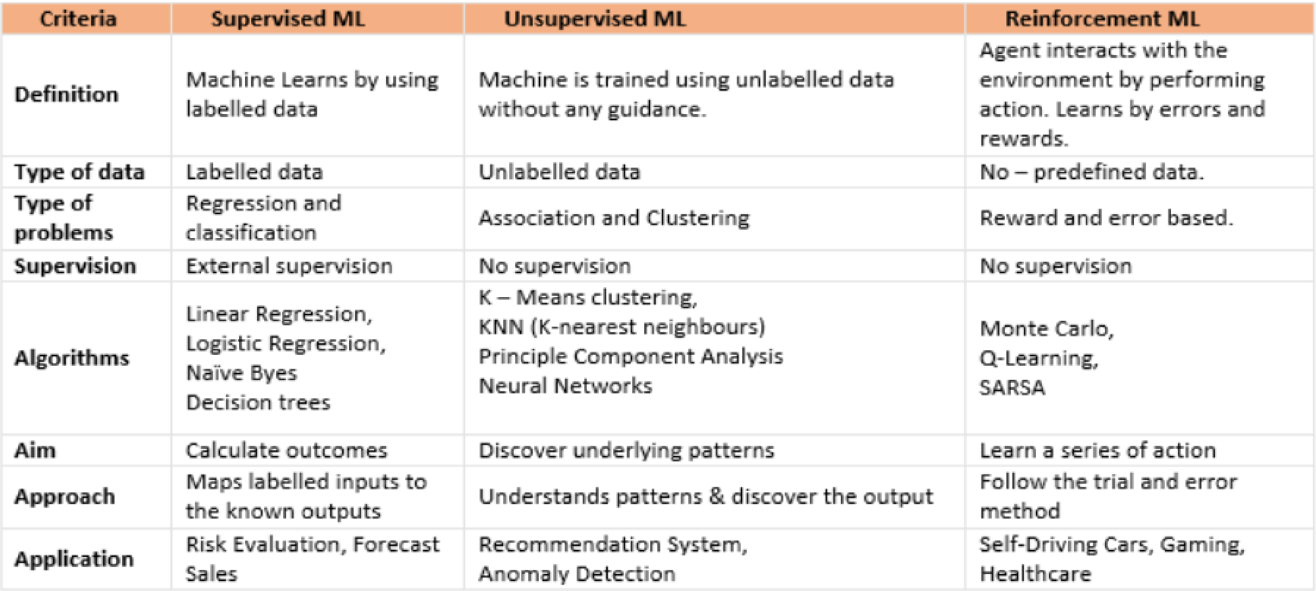
e.g Chess ,Tic Tac Toe

7)Healthcare :

Healthcare employs machine learning and artificial intelligence in much of its work, and RL is no exception. It has been used in automated medical diagnosis, resource scheduling, drug discovery and development, and health management.

8)Self Driving Cars :

he algorithms learn to recognize pedestrians,roads, traffic, detect street signs in the environment and act accordingly.identify what should be speed on busy streets ,in traffic.



Criteria	Supervised ML	Unsupervised ML	Reinforcement ML
Definition	Machine Learns by using labelled data	Machine is trained using unlabelled data without any guidance.	Agent interacts with the environment by performing action. Learns by errors and rewards.
Type of data	Labelled data	Unlabelled data	No – predefined data.
Type of problems	Regression and classification	Association and Clustering	Reward and error based.
Supervision	External supervision	No supervision	No supervision
Algorithms	Linear Regression, Logistic Regression, Naïve Byes Decision trees	K – Means clustering, KNN (K-nearest neighbours) Principle Component Analysis Neural Networks	Monte Carlo, Q-Learning, SARSA
Aim	Calculate outcomes	Discover underlying patterns	Learn a series of action
Approach	Maps labelled inputs to the known outputs	Understands patterns & discover the output	Follow the trial and error method
Application	Risk Evaluation, Forecast Sales	Recommendation System, Anomaly Detection	Self-Driving Cars, Gaming, Healthcare

- **Batch and Epoch :**

In Machine Learning, whenever you want to train a model with some data, then **Epoch** refers to one complete pass of the training dataset through the algorithm. Moreover, it takes a few epochs while training a machine learning model, but, in this scenario, you will face an issue while feeding a bunch of training data in the model. This issue happens due to limitations of computer storage. To overcome this issue, we have to break the training data into small batches according to the computer memory or storage capacity. Then only we can train a machine learning model by feeding these batches without any hassle. This process is called batch in machine learning, and ***further, when all batches are fed exactly once to train the model, then this entire procedure is known as Epoch in Machine Learning***. In this article, **''Epoch in Machine Learning''** we will briefly discuss the Epoch, batch, and sample, etc. So let's start with the definition of the Epoch in Machine Learning

2)ML vs DL

3)Cross validation

4)Dimentionality reduction

5)Principle Compomnent Analysis.

6)Regularization

7)Feature Engineering.

8)Stacking

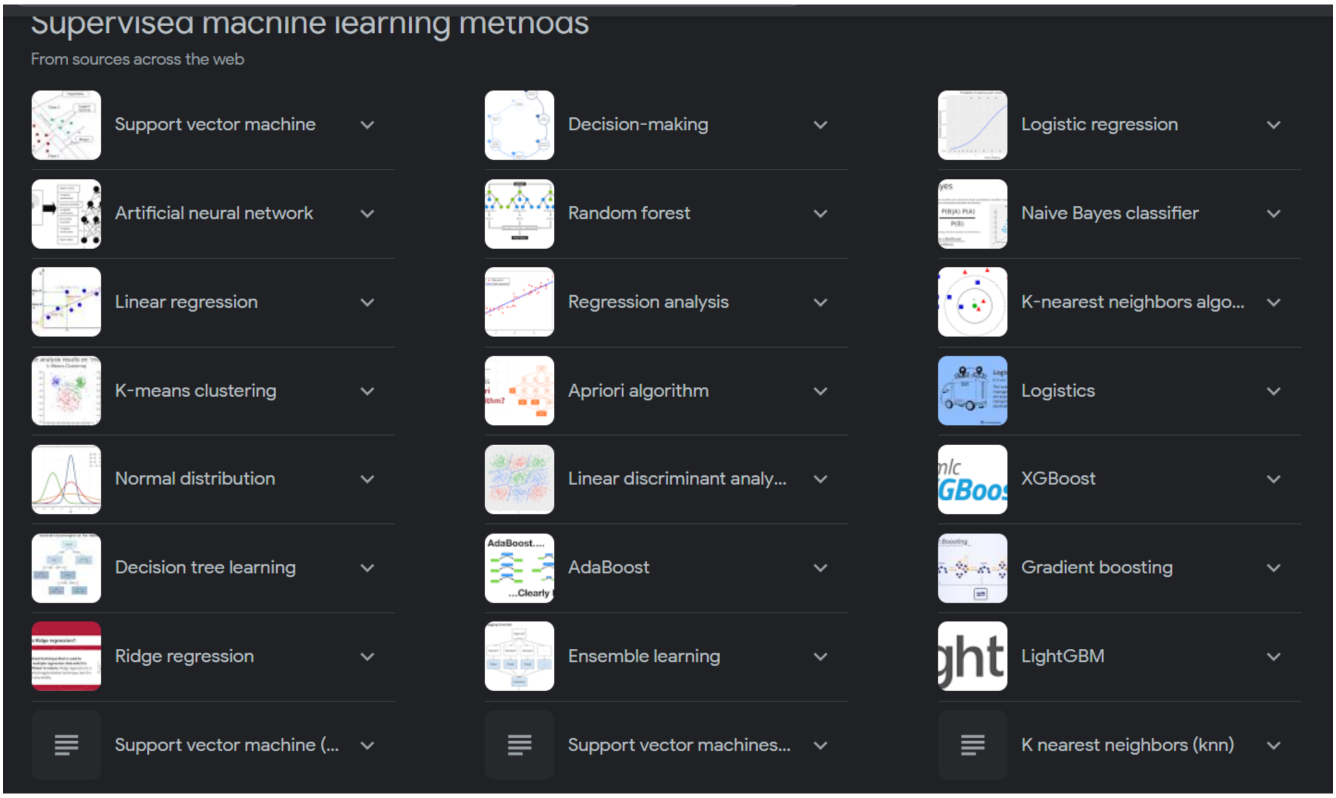
9)Genetic Algorithms in ML

10)Auto ML

11)Challenges and importance of ML

12)Parameters Vs Hyperparameters

13)



Supervised machine learning methods

From sources across the web

- Support vector machine
- Artificial neural network
- Linear regression
- K-means clustering
- Normal distribution
- Decision tree learning
- Ridge regression
- Support vector machine (...
- Decision-making
- Random forest
- Regression analysis
- Apriori algorithm
- Linear discriminant analy...
- AdaBoost
- Ensemble learning
- Support vector machines...
- Logistic regression
- Naive Bayes classifier
- K-nearest neighbors algo...
- Logistics
- XGBoost
- Gradient boosting
- LightGBM
- K nearest neighbors (knn)

Supervised Machine Learning Algorithms :

Regression :

1)Linear Regression : Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the **dependent variable**. The variable you are using to predict the other variable's value is called the **independent variable.**

Linear regression is commonly used in many fields, including economics, finance, and social sciences, to analyze and predict trends in data. It can also be extended to multiple linear regression, where there are **multiple independent variables,** and logistic regression, which is used for binary classification problems.

a)Simple Linear Regression : In a simple linear regression, there is **one independent** variable and **one dependent variable**. The model estimates the slope and intercept of the line of best fit, which represents the relationship between the variables. The **slope** represents the **change in the dependent variable** for each unit change in the independent variable, while **the intercept represents the predicted value** of the **dependent variable when the independent variable is zero.**

e.g : Salary Prediction Dependant variable : Salary /Independent Variable : Year of experience

Line of Regression : The line is determined by minimizing the sum of the squared differences between the predicted values and the actual values.

Best Fit: The best fit line is a line that fits the given scatter plot in the best way. Mathematically, the best fit line is obtained by minimizing the Residual Sum of Squares(RSS).

Cost Function : Used to measure perfiormance of linear regression **. MSE (Mean Squared Error) :** squared difference between actual and predicted values**.**

Evaluation of Linear Regression : Strength of model can be assessed using

a)R-Squared : It always ranges between 0 & 1 . Overall, the higher the value of R-squared, the better the model fits the data.R2 = 1 – (RSS[actual-predicted]) higher r-square indicate model is good fit and lowerb r-square indicate not good fit

b)RMSE : R-squared is a better measure than RSME. Because the value of Root Mean Squared Error depends on the units of the variables.

Bias : It is the error due to the model's inability to represent the true relationship between input and output accurately. When a model has poor performance both on the training and testing data means high bias because of the simple model, indicating underfitting.

Variance : error due to model sensitivity.**High variance** occurs when a model learns the training data's noise and random fluctuations rather than the underlying pattern. As a result, the model performs well on the training data but poorly on the testing data, indicating overfitting.[determine how model react when cange in input data]

Bias Variance Tradeoff : The aim of any supervised machine learning algorithm is to achieve low bias and low variance as it is more robust. So that the algorithm should achieve better performance. here is an inverse relationship between bias and variance,

- An increase in bias will decrease the variance.

- An increase in the variance will decrease the bias.

There is a trade-off that plays between these two concepts and the algorithms must find a balance between bias and variance.

Standard Deviation(Gap between Values) : Standard deviation is a number that describes how spread out the values are.

A low standard deviation means that most of the numbers are close to the mean (average) value.

A high standard deviation means that the values are spread out over a wider range

Model parameters : *are configuration variables that are internal to the model, and a model learns them on its own. For example, W Weights or Coefficients of independent variables in the Linear regression model. or Weights or Coefficients of independent variables in SVM, weight, and biases of a neural network, cluster centroid in clustering. Some key points for model parameters are as follows:*

- They are used by the model for making predictions.
- They are learned by the model from the data itself
- These are usually not set manually.
- These are the part of the model and key to a machine learning Algorithm.

"***Hyperparameters :*** *are defined as the parameters that are explicitly defined by the user to control the learning process." These are external to the model, and their values cannot be changed during the training process.*

- The k in kNN or K-Nearest Neighbour algorithm
- Learning rate for training a neural network
- Train-test split ratio
- Batch Size
- Number of Epochs
- Branches in Decision Tree
- Number of clusters in Clustering Algorithm

a)Hyperparameter for Optimization : The process of selecting the best hyperparameters to use is known as hyperparameter tuning, and the tuning process is also known as hyperparameter optimization. E.g learing rate ,batch size

b) Hyperparameter for Specific Models : Hyperparameters that are involved in the structure of the model are known as hyperparameters for specific models

e.g : No of hidden unit ,number of layers.

Learning rate : The learning rate is a tuning parameter in an optimization algorithm that determines the step size at each iteration while moving toward a minimum of a loss function.

Batch Size: To enhance the speed of the learning process, the training set is divided into different subsets, which are known as a batch.

Number of Epochs: An epoch can be defined as the complete cycle for training the machine learning mode.

Mean ,Mode ,Median :

Mean : Avarage of values
Mode : frequent value.
Median : sorted middle value (if count is odd take middle value as median else middle two valpue and avarge it)

b)Multiple Linear Regression : MLR examines how multiple independent variables are related to one dependent variable.

E,g Car price prediction using various independent variables like fueltype ,aspiration, doornumber ,carbody ,drivewheel ,enginelocation , wheelbase ,carlength , carwidth , carheight , enginesize,fuelsystem, boreratio ,stroke, compressionratio, horsepower ,peakrpm citympg, highwaympg,price.

Consideration For Multiple Linear regression :

Good Fitting : Low bias ,low variance .

Multicollinearity: It is the phenomenon where a model with several independent variables, may have some variables interrelated.

Feature Selection: With more variables present, selecting the optimal set of predictors from the pool of given features (many of which might be redundant) becomes an important task for building a relevant and better model.

Underfitting : 1)It is inability of machine learning model to learn from training data effectively that result poor performance on training and testing data. underfit model's are inaccurate, especially when applied to new, unseen examples.

Reasons of underfitting :

1)The model is too simple, So it may be not capable to represent the complexities in the data.
2)Size of training data is not enough.

Techniques Used :

1) Increase model complexity.
2)Increase no of features ,performing feature engineeeruing {as there are various features but some feature affect most on the result extracting such features called feature engineering }
3)Remove noise and increase duration of traing to get better result.
OverFitting : A `statistical model` is said to be overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise.

Reasons :
1)High Variance and Low bias
2)Large amount of data & start learning from noise as well
3)Complex model

Techniques used :

1)Reduce model complexity
2)Ridge and Lasso regiularization (improve performance)
3)Dropout from neural network (drop node that are not useful)

Gradient Descent for Liear Regression : Gradient Descent is one of the optimization algorithms that optimize the cost function(objective function) to reach the optimal minimal solution. To find the optimum solution we need to reduce the cost function(MSE) for all data points.

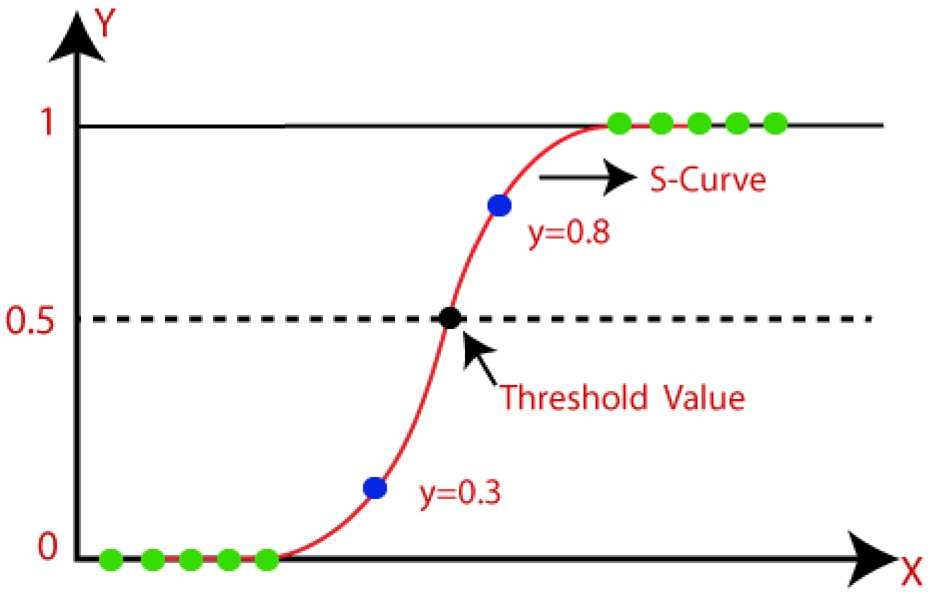
Ridge and Lasso(Least Absolute Shrinkage and Selection Operator) : Both used for dealing with overfitting
Diff:

- Ridge: It includes all (or none) of the features in the model. Thus, the major advantage of ridge regression is coefficient shrinkage and reducing model complexity.
- Lasso: Along with shrinking coefficients, the lasso also performs feature selection. (Remember the ‘selection‘ in the lasso full-form?) As we observed earlier, some of the coefficients become exactly zero, which is equivalent to the particular feature being excluded from the model.

Classification :

1)Logistic Regression :

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
- Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).



- **Logistic Function (Sigmoid Function):**

1) It maps any real value into another value within a range of 0 and 1.
2) The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
3) In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

- **Types :**

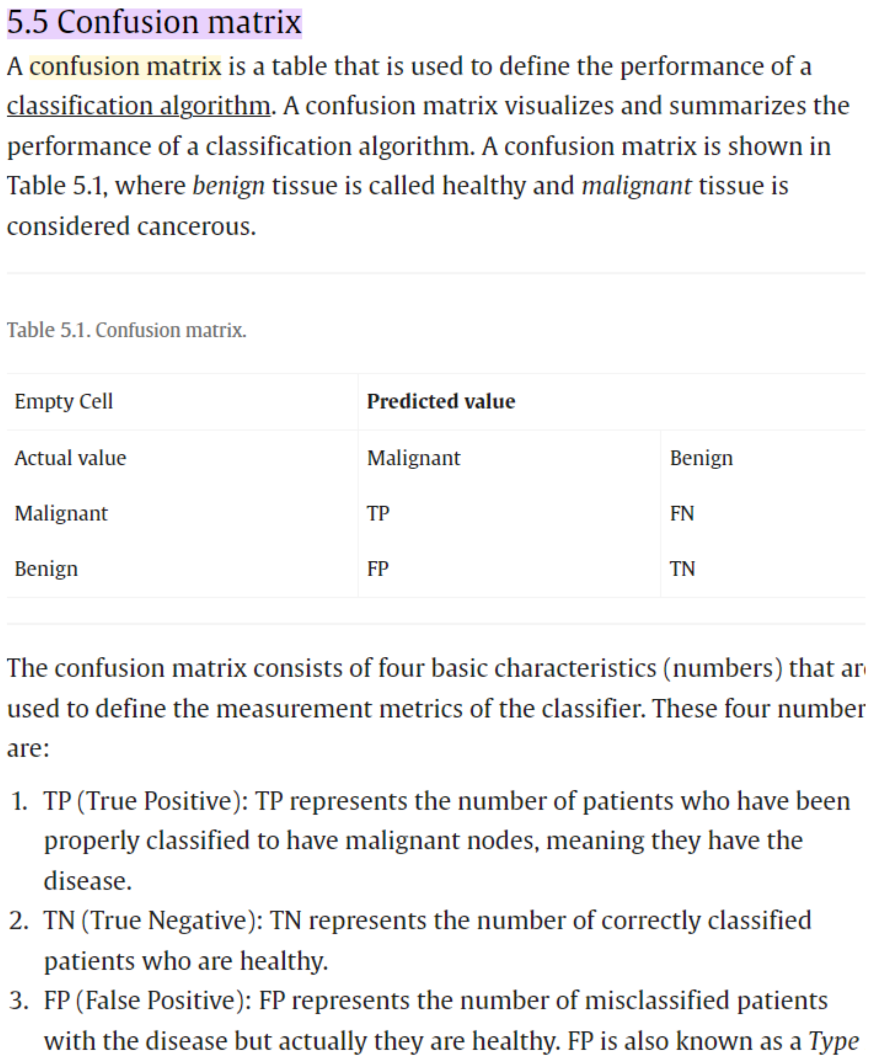
1)Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.

2)Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep”.

3)Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

E.g : we have customer data such as gender ,age and salary and we have to check weather that person will buy item or not

- **Classification Algorithms :**
  - Support Vector Machine
  - Naïve bias
  - K-nearest neighbour
  - Decision Tree
  - Random Forest
- **Performance of classification Algorithms :**
  - A confusion matrix is a table that is used to define the performance of a classification algorithm.



5.5 Confusion matrix

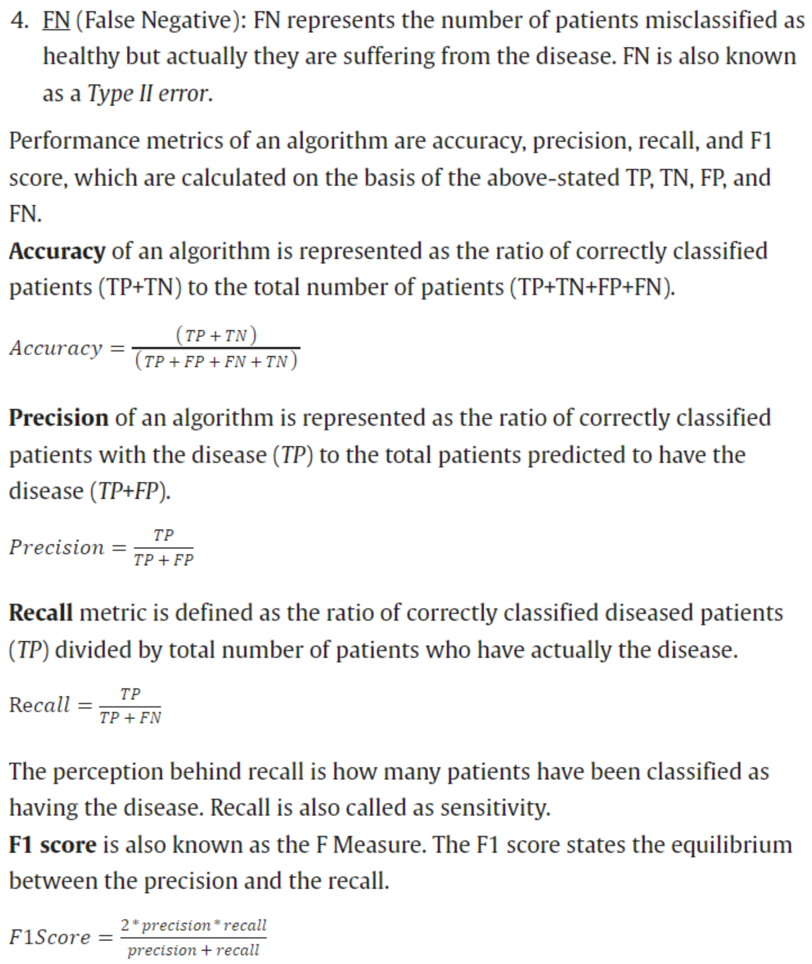
A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm. A confusion matrix is shown in Table 5.1, where *benign* tissue is called healthy and *malignant* tissue is considered cancerous.

Table 5.1. Confusion matrix.

Empty Cell	**Predicted value**	
Actual value	Malignant	Benign
Malignant	TP	FN
Benign	FP	TN

The confusion matrix consists of four basic characteristics (numbers) that ar used to define the measurement metrics of the classifier. These four number are:

1. TP (True Positive): TP represents the number of patients who have been properly classified to have malignant nodes, meaning they have the disease.
2. TN (True Negative): TN represents the number of correctly classified patients who are healthy.
3. FP (False Positive): FP represents the number of misclassified patients with the disease but actually they are healthy. FP is also known as a *Type*



4. FN (False Negative): FN represents the number of patients misclassified as healthy but actually they are suffering from the disease. FN is also known as a *Type II error*.

Performance metrics of an algorithm are accuracy, precision, recall, and F1 score, which are calculated on the basis of the above-stated TP, TN, FP, and FN.

Accuracy of an algorithm is represented as the ratio of correctly classified patients (TP+TN) to the total number of patients (TP+TN+FP+FN).

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

Precision of an algorithm is represented as the ratio of correctly classified patients with the disease (*TP*) to the total patients predicted to have the disease (*TP*+*FP*).

$$Precision = \frac{TP}{TP + FP}$$

Recall metric is defined as the ratio of correctly classified diseased patients (*TP*) divided by total number of patients who have actually the disease.

$$\text{Recall} = \frac{TP}{TP + FN}$$

The perception behind recall is how many patients have been classified as having the disease. Recall is also called as sensitivity.

F1 score is also known as the F Measure. The F1 score states the equilibrium between the precision and the recall.

$$F1Score = \frac{2 * precision * recall}{precision + recall}$$

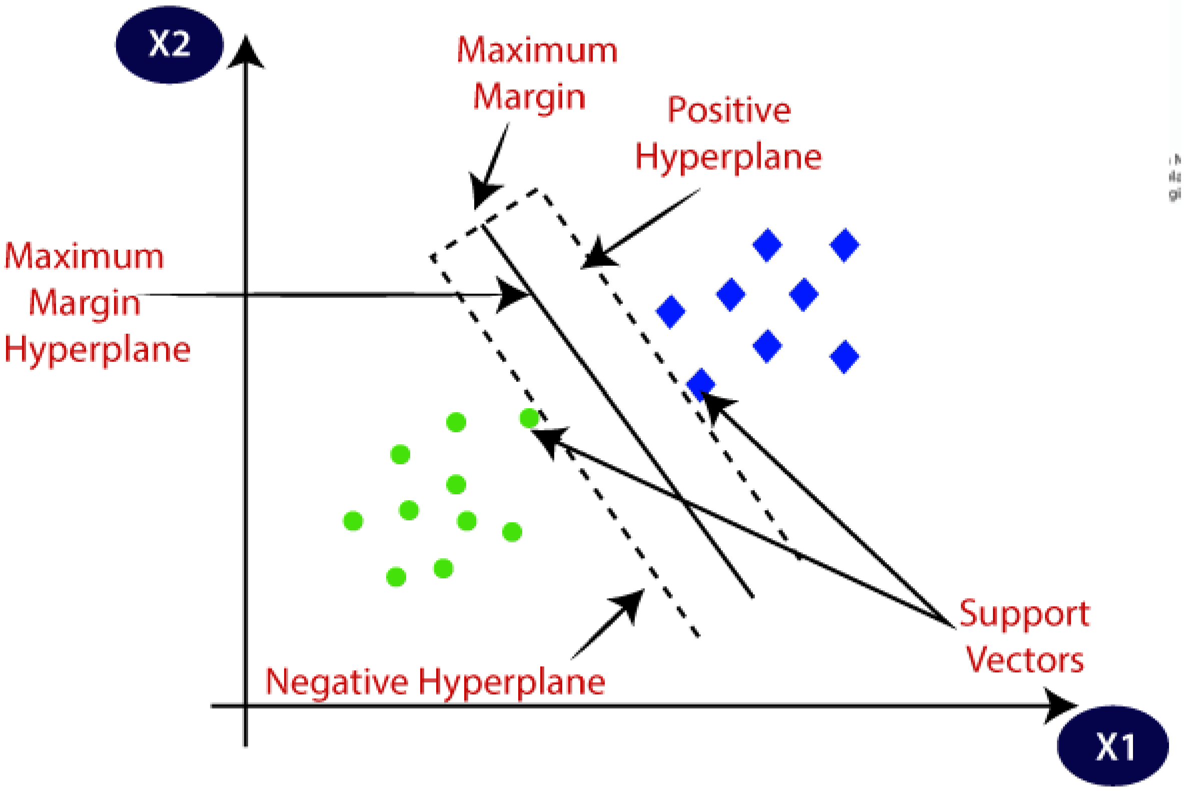
- Value close to 1 have best F1 score.

Support Vector Machine :

1)Supervised Machine Learning Technique used for classification

2)goal is create best line that segregate n-dimensional space into classes so that we can easily put new data point into specific category.

3) SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.



Example : Suppose we see a strange cat that also has some features of dogs, so if we want a model that can accurately identify whether it is a cat or dog, so such a model can be created by using the SVM algorithm. We will first train our model with lots of images of cats and dogs so that it can learn about different features of cats and dogs, and then we test it with this strange creature. So as support vector creates a decision boundary between these two data (cat and dog) and choose extreme cases (support vectors), it will see the extreme case of cat and dog. On the basis of the support vectors, it will classify it as a cat.

Application : Face detection, image classification, text categorization.

- **Hyperplane:** There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

  The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

  We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

- **Support vectors :** The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector.

Types :

- **Linear SVM :** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

- **Non Linear SVM :** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

K-Nearest Neighbour : -

- Supervised Machine learning algorithm used to solve classification as well as regression tasks.
- Find similarity between new cases and available cases and place new data into similar category.
- non parametric algorithm means does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

K-nearest Neighbour : The algorithms performance can be sensitive the value of k. When k is small, the model is sensitive to noise and is prone to overfitting whereas large values of k can lead to underfitting, especially if there is class inbalance.When the number of classes is 2, k should be an odd number to prevent any 'tied votes' when making predictions.The value of k should be bigger than the number of classes for similar reasons.

Steps :

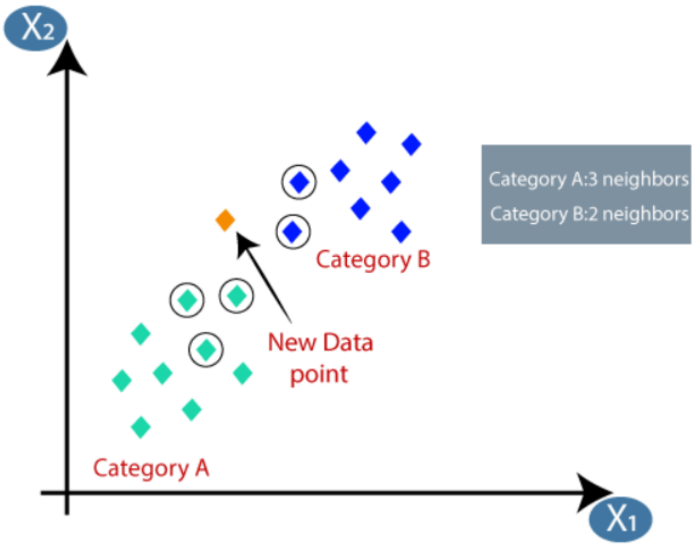
The K-NN working can be explained on the basis of the below algorithm:

- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.

- **Step-6:** Our model is ready.

Euclidean Distance between A_1 and B_2 = $\sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$

- By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A a two nearest neighbors in category B. Consider the below image:



How to select the value of K in the K-NN Algorithm?

Below are some points to remember while selecting the value of K in the K-NN algorithm:

- There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
- A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
- Large values for K are good, but it may find some difficulties.

Advantages of KNN Algorithm:

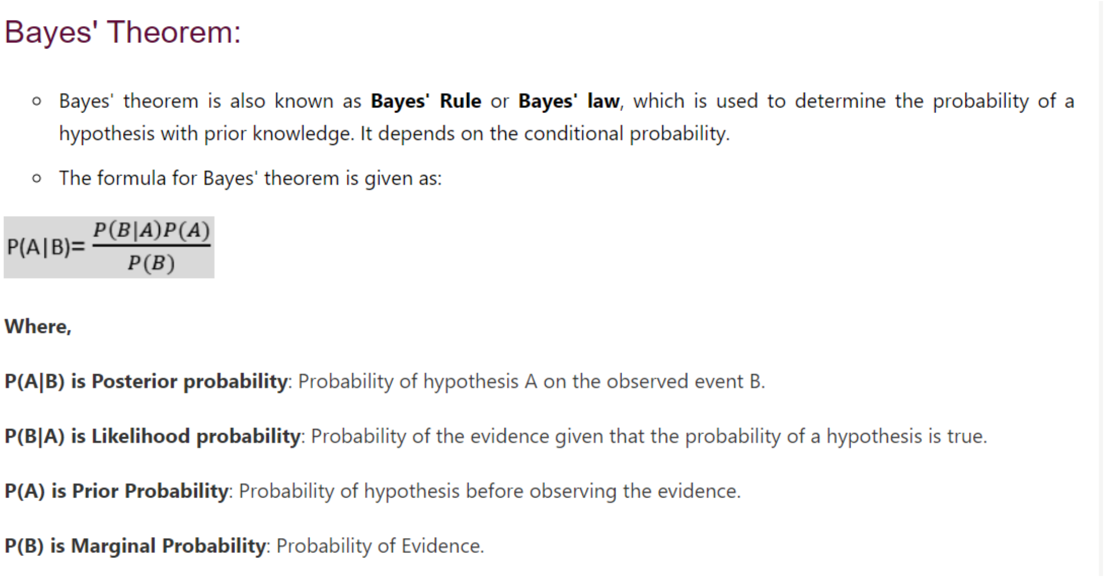
- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

Naïve Bias Algorithm :

- Used in text classification.
- Based on Bayes theorem ,finds probability.
- Most effective help to build fast ML models.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.
- It assume all features independent so cannot learn relationship between features.



Bayes' Theorem:

- Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- The formula for Bayes' theorem is given as:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

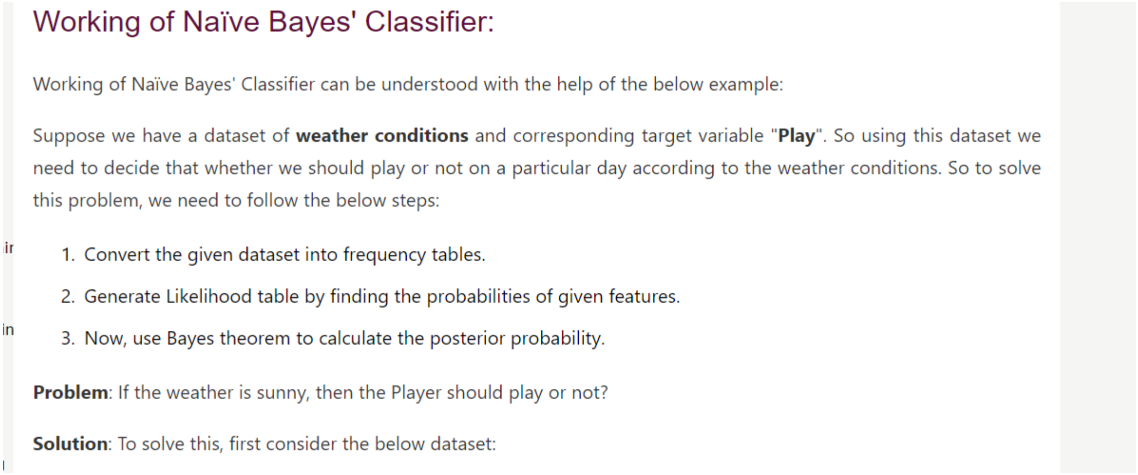
Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.



Working of Naïve Bayes' Classifier:

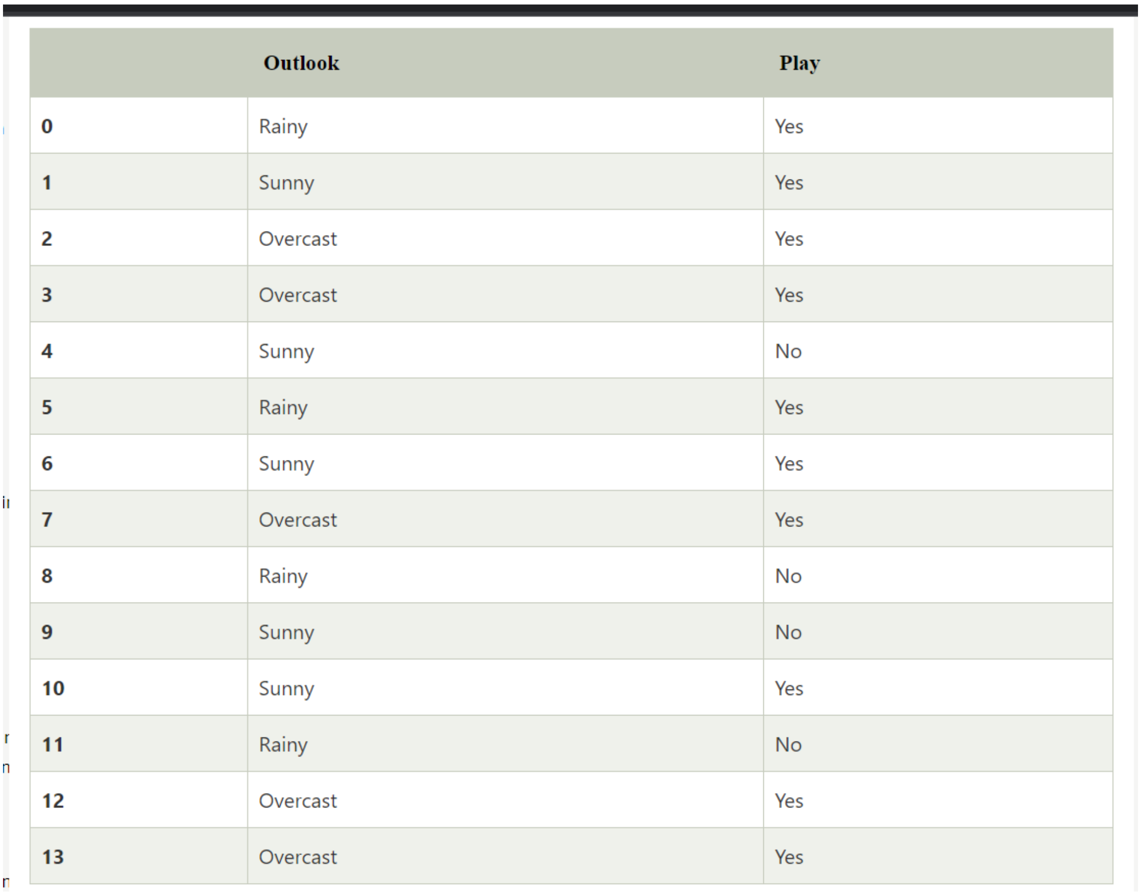
Working of Naïve Bayes' Classifier can be understood with the help of the below example:

Suppose we have a dataset of **weather conditions** and corresponding target variable "**Play**". So using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions. So to solve this problem, we need to follow the below steps:

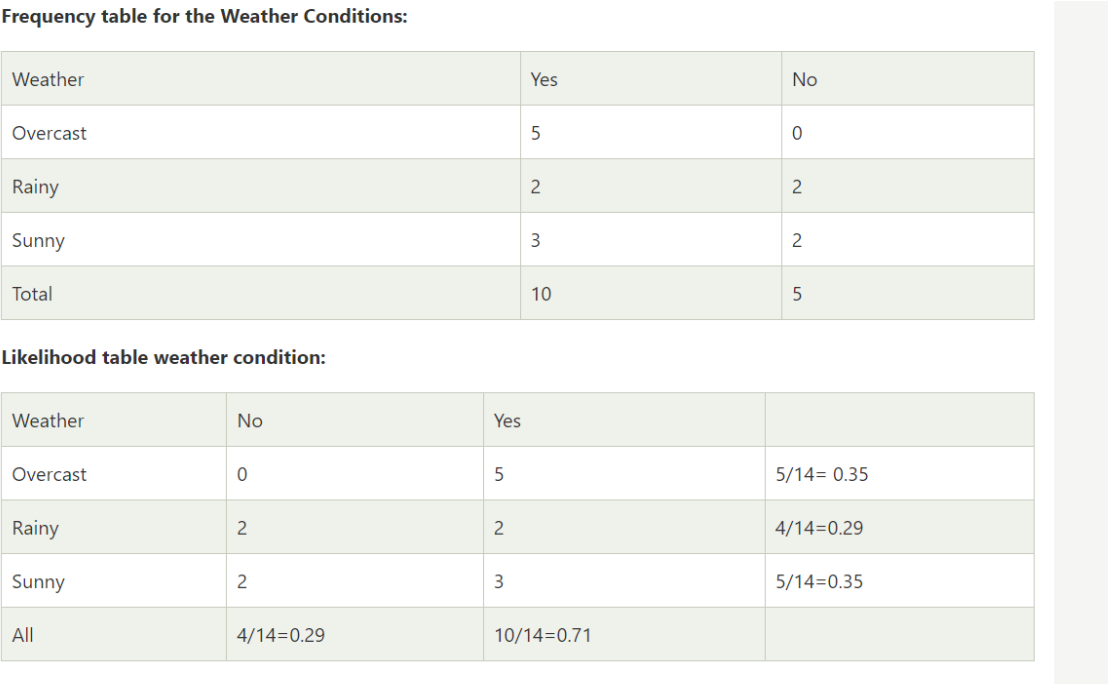
1. Convert the given dataset into frequency tables.
2. Generate Likelihood table by finding the probabilities of given features.
3. Now, use Bayes theorem to calculate the posterior probability.

Problem: If the weather is sunny, then the Player should play or not?

Solution: To solve this, first consider the below dataset:



	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

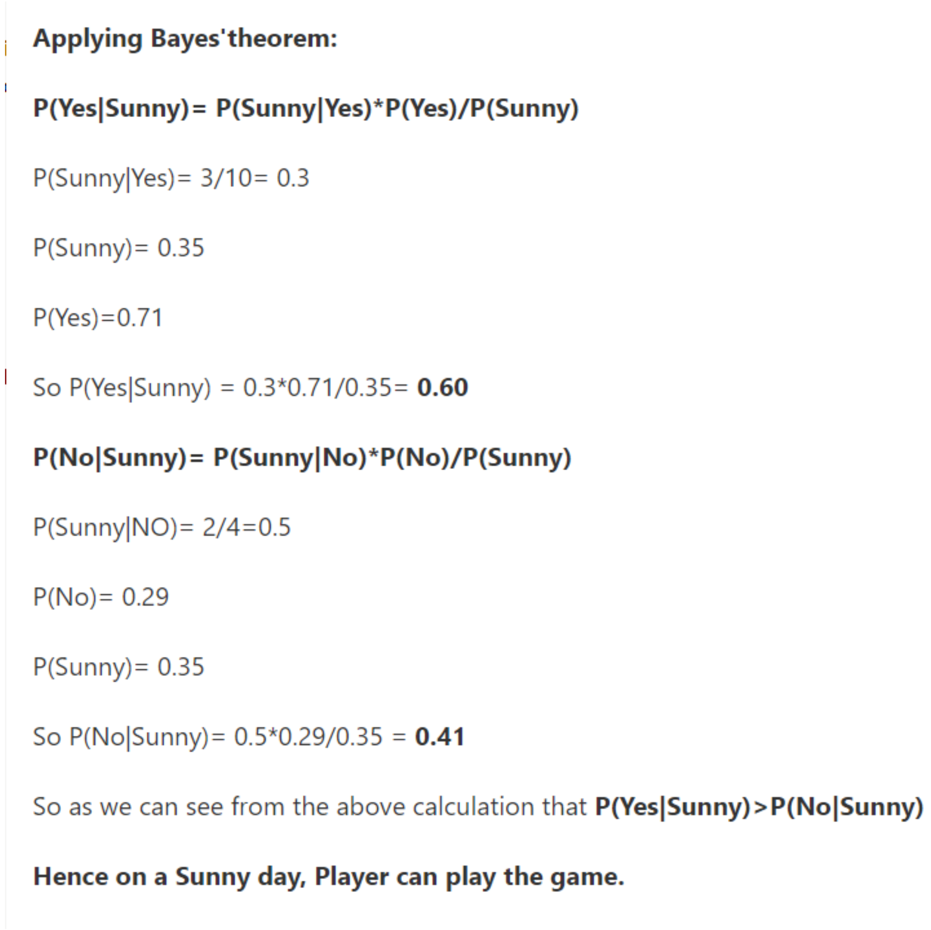


Frequency table for the Weather Conditions:

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	5

Likelihood table weather condition:

Weather	No	Yes	
Overcast	0	5	5/14= 0.35
Rainy	2	2	4/14=0.29
Sunny	2	3	5/14=0.35
All	4/14=0.29	10/14=0.71	



Applying Bayes'theorem:

P(Yes|Sunny)= P(Sunny|Yes)*P(Yes)/P(Sunny)

P(Sunny|Yes)= 3/10= 0.3

P(Sunny)= 0.35

P(Yes)=0.71

So P(Yes|Sunny) = 0.3*0.71/0.35= **0.60**

P(No|Sunny)= P(Sunny|No)*P(No)/P(Sunny)

P(Sunny|NO)= 2/4=0.5

P(No)= 0.29

P(Sunny)= 0.35

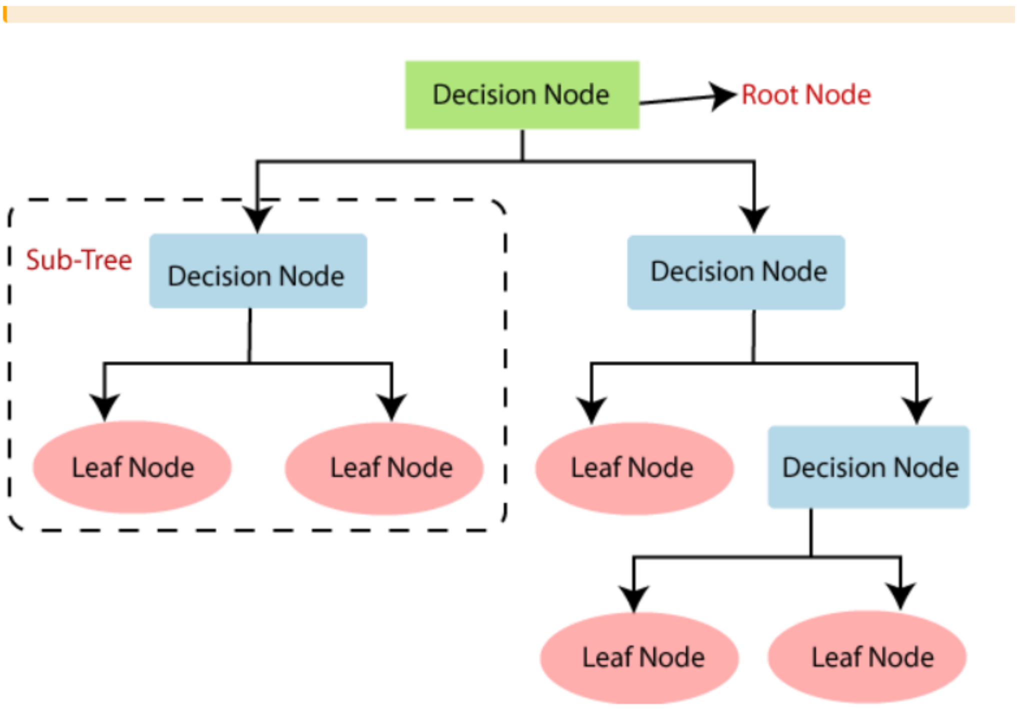
So P(No|Sunny)= 0.5*0.29/0.35 = **0.41**

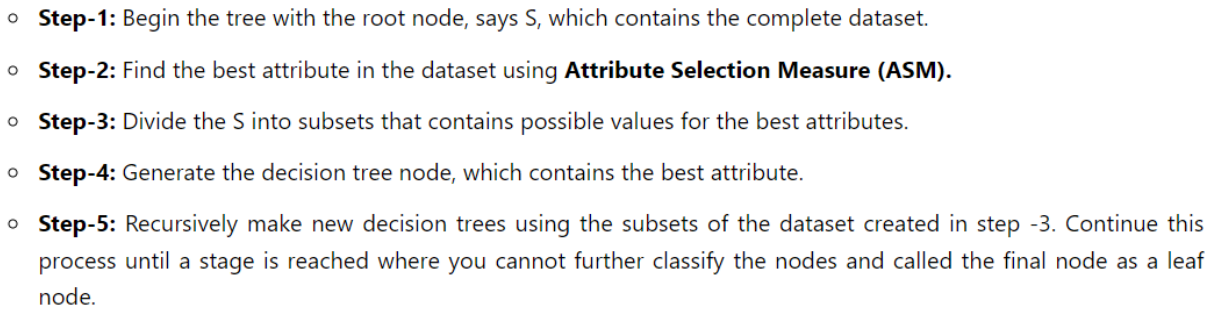
So as we can see from the above calculation that **P(Yes|Sunny)>P(No|Sunny)**

Hence on a Sunny day, Player can play the game.

Decision Tree :

- It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- The decisions or the test are performed on the basis of features of the given dataset.
- In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.





- **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
- **Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**
- **Step-3:** Divide the S into subsets that contains possible values for the best attributes.
- **Step-4:** Generate the decision tree node, which contains the best attribute.
- **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as Attribute selection measure or ASM. By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:

- **Information Gain :** it is change in entropy. It calculates how much information a feature provides us about a class. According to the value of information gain, we split the node and build the decision tree. A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first.
- Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)

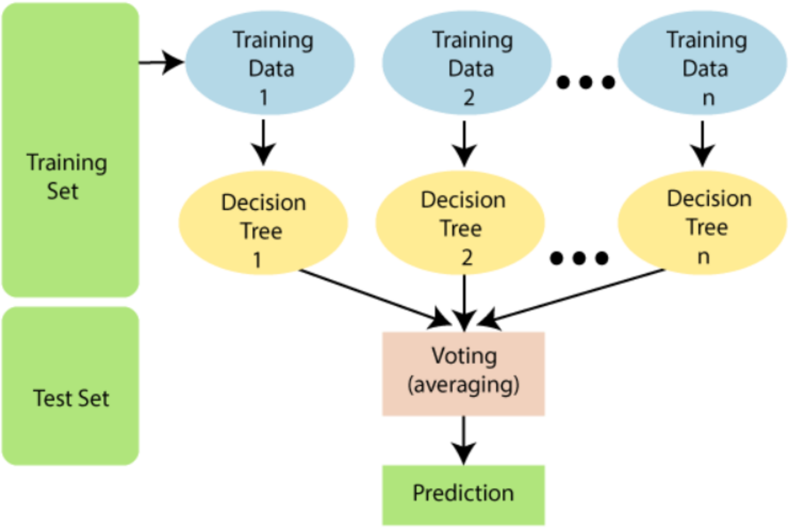
- **Entropy:** Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:
- Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no) Where,
- S= Total number of samples
- P(yes)= probability of yes
- P(no)= probability of no

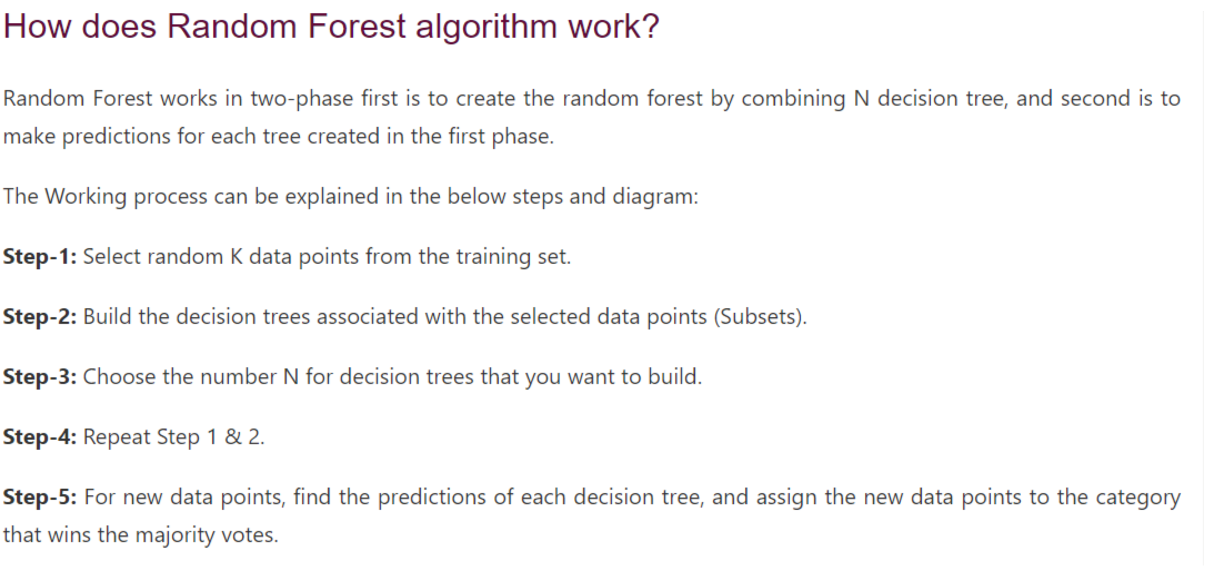
- **Gini Index :** Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.

Gini Index= $1- \sum_j P_j^2$

Random Forest Algorithm :

- It is based on the concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model*
- Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
- Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.





How does Random Forest algorithm work?

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

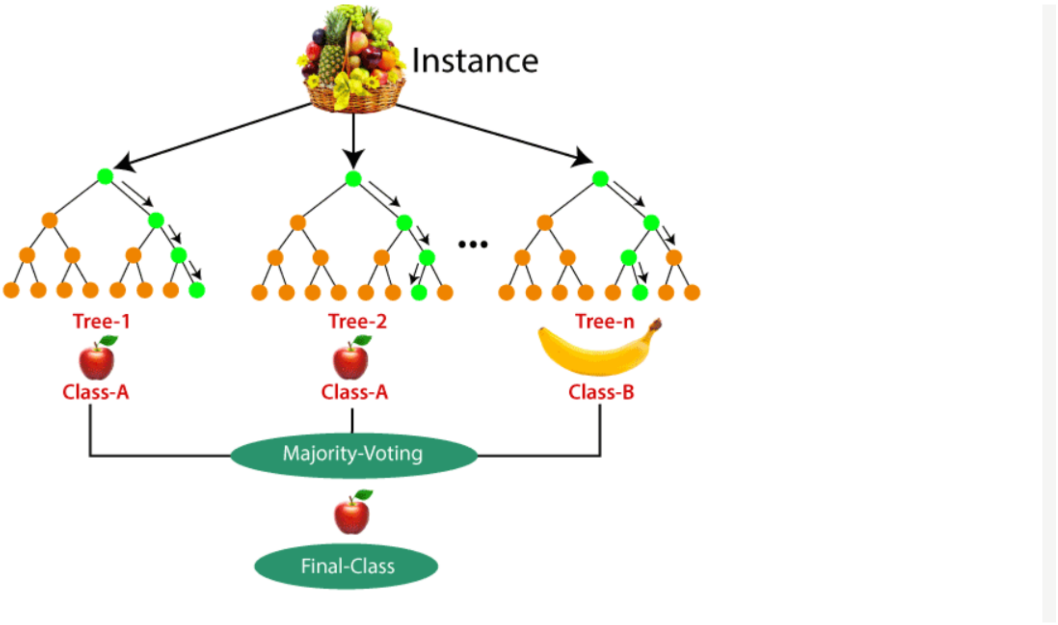
Step-1: Select random K data points from the training set.

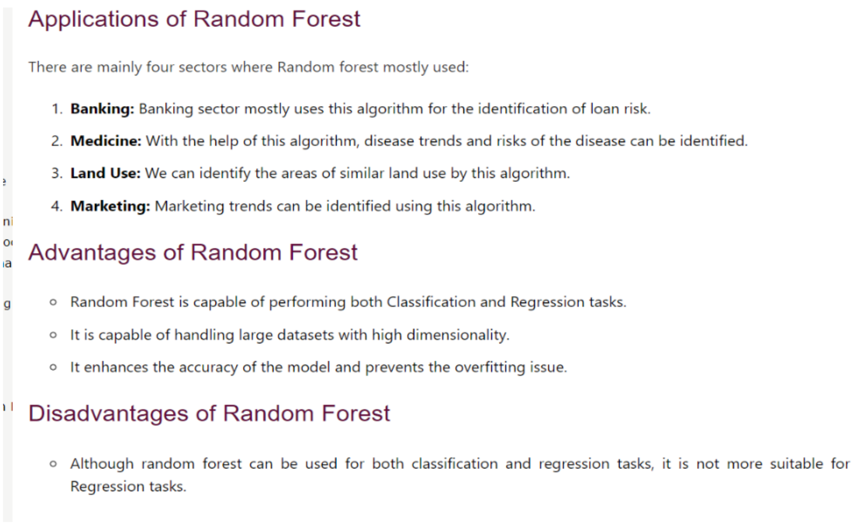
Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.





Applications of Random Forest

There are mainly four sectors where Random forest mostly used:

1. **Banking:** Banking sector mostly uses this algorithm for the identification of loan risk.
2. **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.
3. **Land Use:** We can identify the areas of similar land use by this algorithm.
4. **Marketing:** Marketing trends can be identified using this algorithm.

Advantages of Random Forest

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

Disadvantages of Random Forest

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.