Algorithm Vs Model

Algorithm:

- a set of rules that specifies how to perform a particular task or solve a problem.
- derived by statisticians and mathematicians for a particular task.
- used to process data, learn patterns, and make predictions.
- **Examples**: linear regression, decision trees, neural networks, and k-means clustering.

Model:

- the result of training an algorithm on data,

  Model = Data + Algorithm
- represents the patterns, relationships, and parameters learned by the algorithm during the training process.
- used to make predictions/decisions on new, unseen data.
- **For example**, a linear regression model learns the coefficients that define the linear relationship between input features and the target variable.
- contains 4 steps: Data preprocessing, Feature engineering, Data management and performance measurement.

- **Algorithms** in ML were derived many years ago. Only when they were implemented in the form of a code in a computer, the algorithms' utility increased to a very great extent since the computers can handle high computation very easily.

- **Example**:

  y = w0+w1x, You might be knowing that this is an equation of a line, where
  w0 corresponds to the y-intercept and
  w1 corresponds to slope of the line.

This is nothing but the equation of **linear regression** with one variable. Similarly every algorithm has some mathematical form underneath it, which when implemented in a machine developed to form a ML algorithm.

Now, coming to defining a model. In the above equation, you cannot find y if you don't know w0 and w1. So how to find it? Suppose you are given a set of sample data, say 2 values of x and y, then certainly you can find the slope by slope-point form. Again let's take the 2 points be: **(x1, y1) = (1, 1)** and **(x2, y2) = (2, 2)**

Now by slope-point form we can find w1 for which the formula is: **w1 = y2 −y1 / x2 − x1**

So, **w1 = 1.** Now To find **w0**, use the equation of the line **y = w0+w1x**.

Substitute one of the points into this equation and solve for **w0**. Let's use **(x1, y1) = (1, 1)**:

$$1 = w0 + 1*1$$

$$w0=0$$

By all this calculation, we have an equation,

y=0+(1)x, which is a model.

So we can now say that **a model is an equation which is formed by finding out the parameters (w0,w1) in the equation of the algorithm**. And you create a model using some data, in this case, the two points which we helped us calculate w0,w1 . This is called training a model.

Now we can find any value of **y** given a new value of **x**. This is how prediction takes place using algorithms.

ML Techniques

What is Supervised Machine Learning?

- the algorithm is provided an input dataset, and is rewarded or optimized to meet a set of specific outputs.

- **For example**,
  - deployed in image recognition, using a classification technique.

  - used in predicting demographics such as population growth or health metrics, utilizing a regression technique.

What is Unsupervised Machine Learning?

- the algorithm is provided an input dataset, but not rewarded or optimized to specific outputs, and instead trained to group objects by common characteristics.

- **For example**, recommendation engines on online stores rely on unsupervised ml, specifically a clustering technique.

What is Reinforcement Learning?

- the algorithm is made to train itself using many trial and error experiments.

- happens when the algorithm interacts continually with the environment, rather than relying on training data.

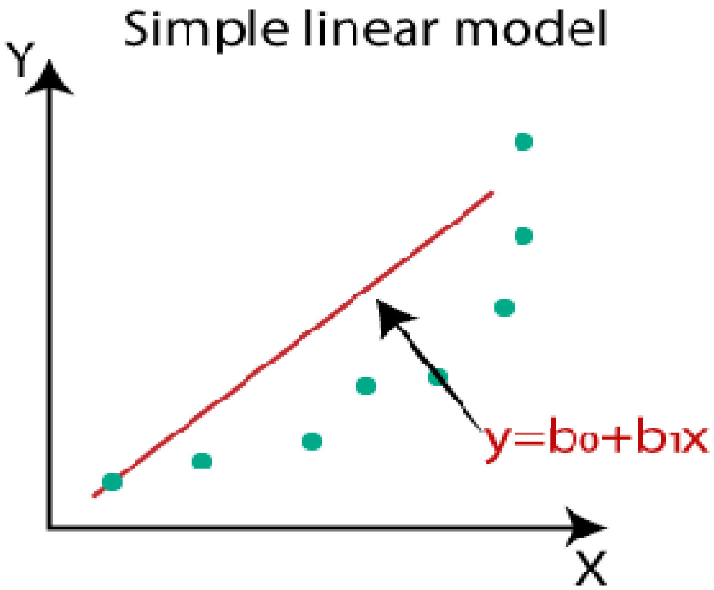
- **For example,** autonomous driving.

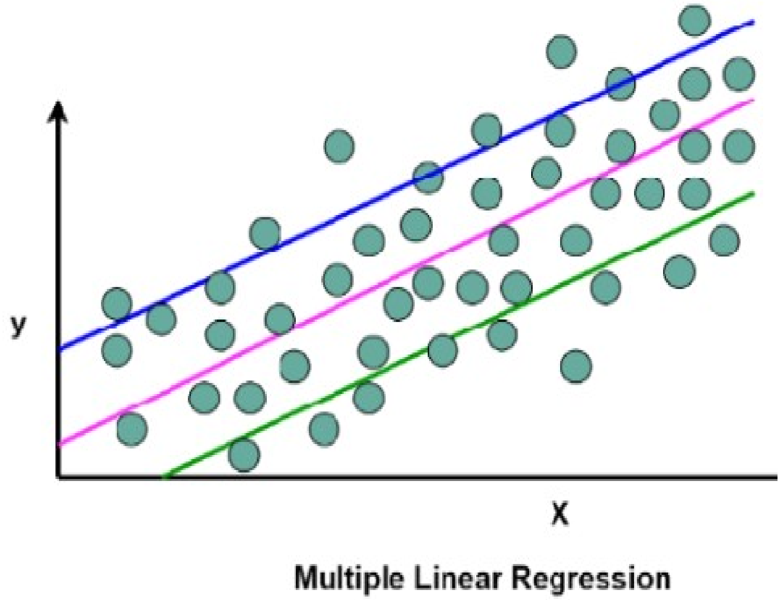
ML Models

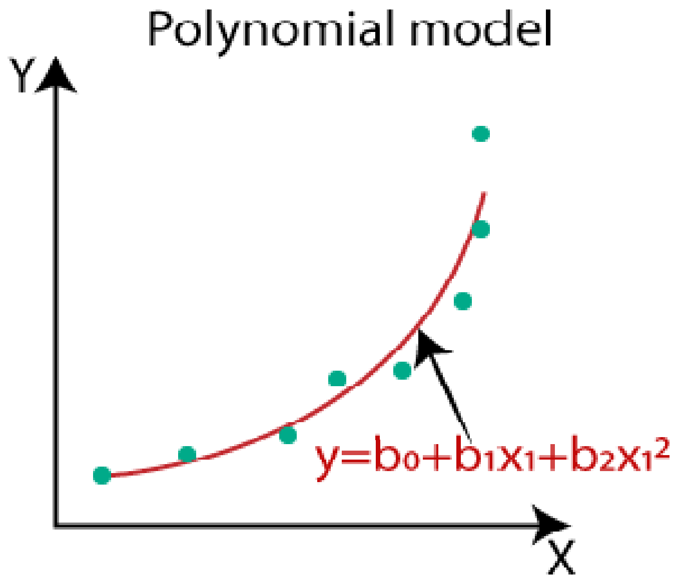
Supervised Machine Learning

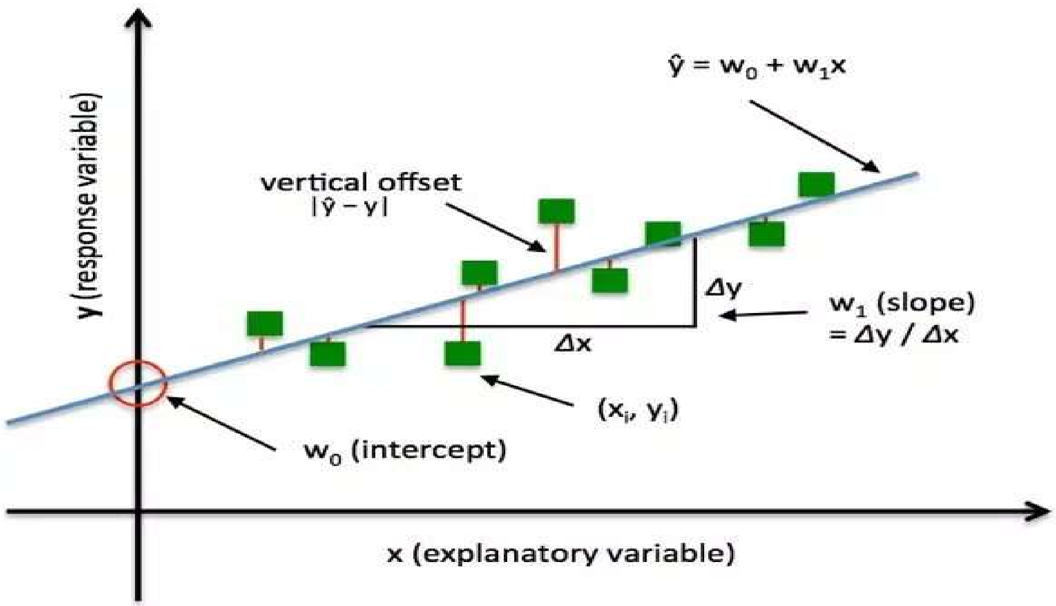
1. **Linear Regression:** is used to identify relationships between the variable of interest and the inputs, and predict its values based on the values of the input variables.

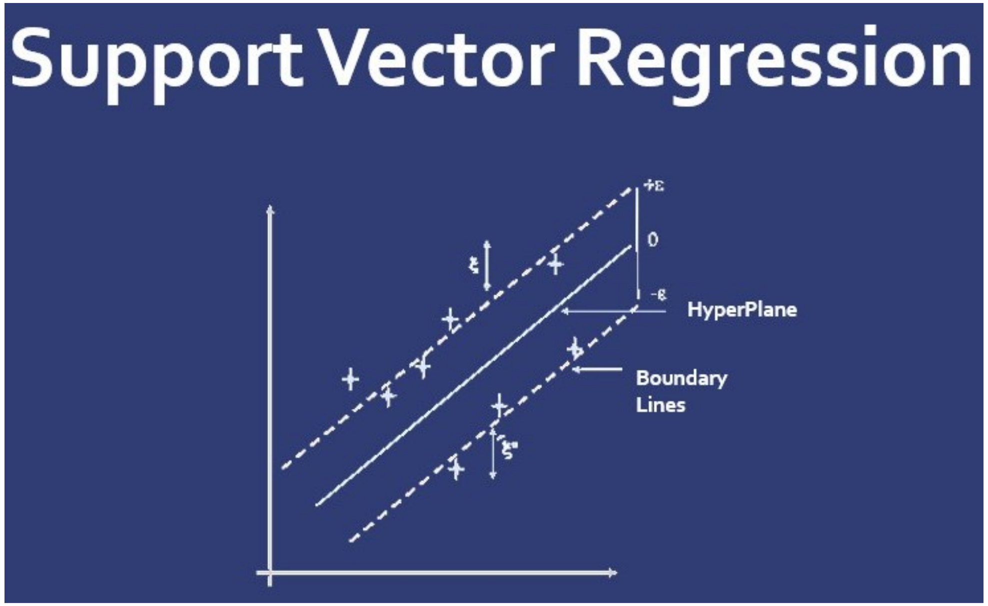
Regression Type	Description	Example
Simple Linear Regression	Models the relationship between two variables using a straight line ($y = \beta 0 + \beta 1x$).	Predicting a person's weight based on their height.
Multiple Linear Regression	Extends simple linear regression to model the relationship between one dependent variable and multiple independent variables ($y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta nxn$).	Predicting house prices based on square footage, number of bedrooms, and location.
Polynomial Regression	Models the relationship between variables as an nth-degree polynomial ($y = \beta 0 + \beta 1x + \beta 2x^2 + ... + \beta nx^n$).	Predicting a car's fuel efficiency based on engine size, where the relationship is nonlinear.
Support Vector Regression	• Uses support vector machines to predict values by finding a function that approximates the data within a certain margin of tolerance (ε) while maximizing the margin between support vectors. or • capable of handling non-linear relationships and outliers.	Predicting stock prices with a focus on capturing complex, non-linear patterns.


Simple linear model
Y
X
y=b0+b1x


y
x
Multiple Linear Regression


Polynomial model
Y
X
y=b0+b1x1+b2x1²

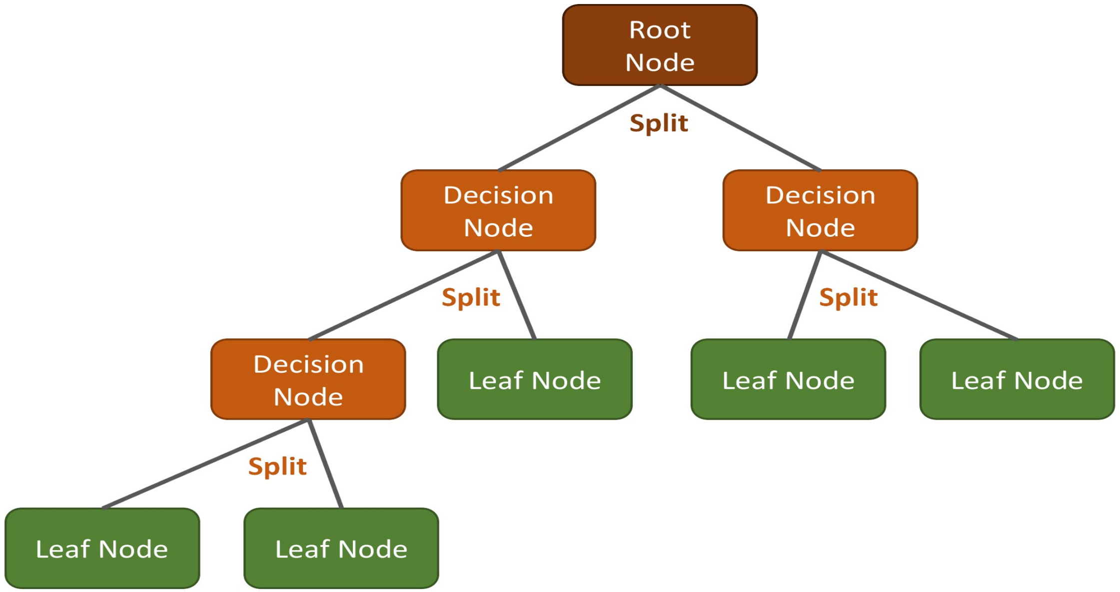

ŷ = w0 + w1x
vertical offset
|ŷ − y|
Δy
Δx
w1 (slope)
= Δy / Δx
(xi, yi)
w0 (intercept)
y (response variable)
x (explanatory variable)

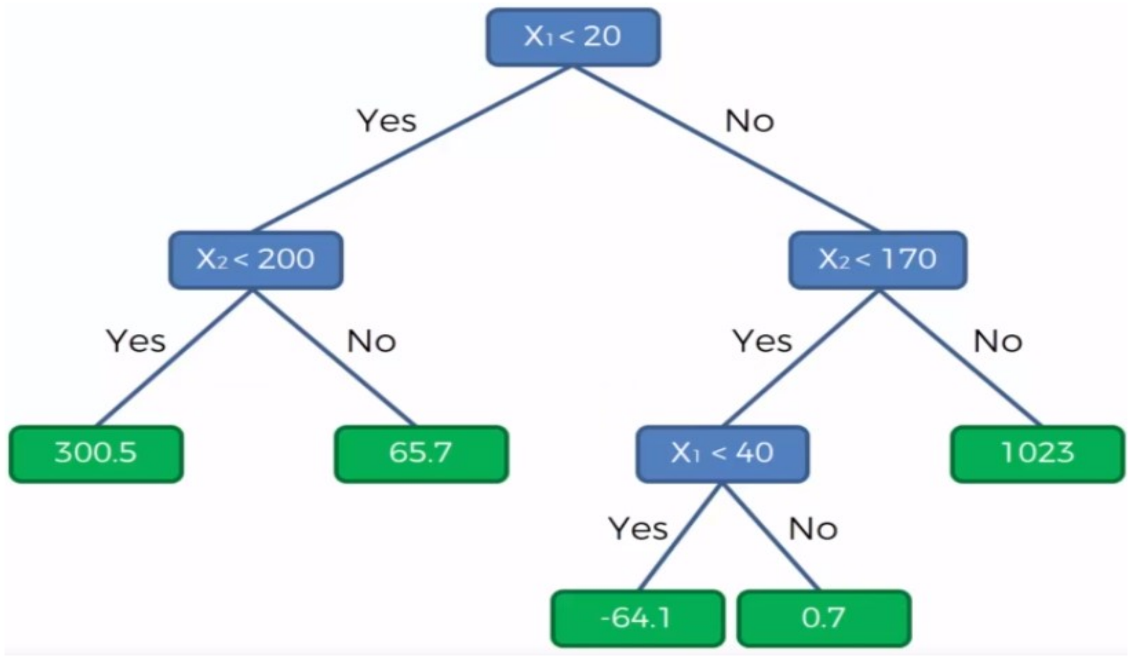

Support Vector Regression
+ε
0
-ε
HyperPlane
Boundary
Lines

Supervised Machine Learning

2. Decision Trees:

- a predictive approach to determine what class an object belongs to.
- a tree-like flow chart where the class of an object is determined step-by-step using certain known conditions.
- Used for tasks where interpretability is crucial, such as risk assessment, decision-making processes.



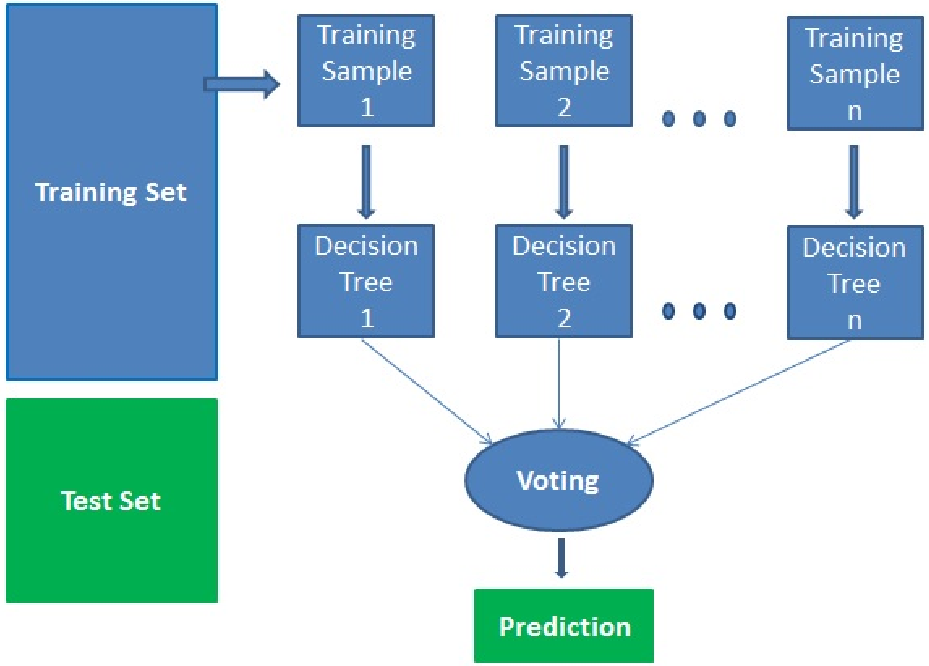


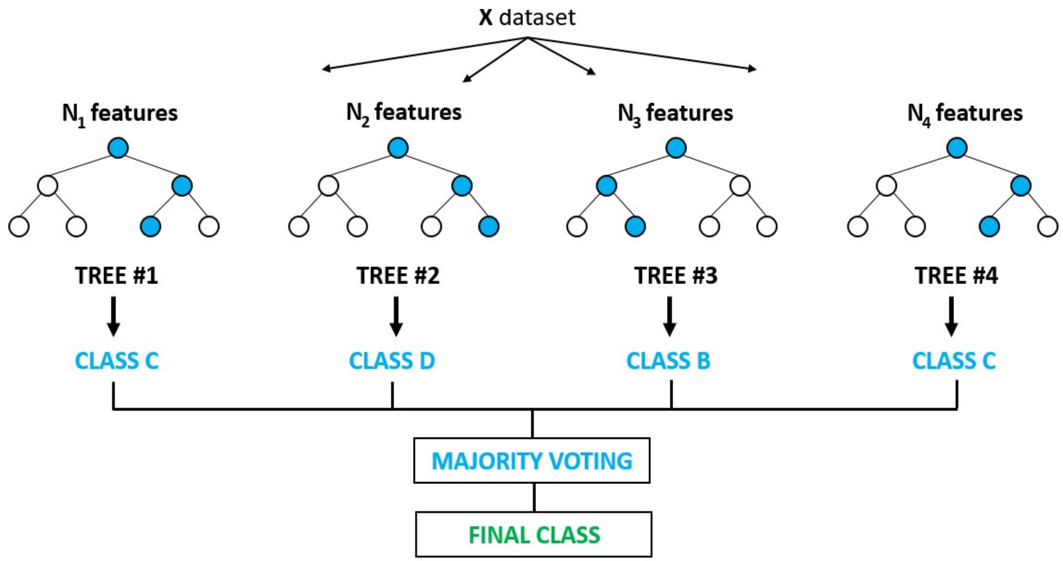
Supervised Machine Learning

3. Random Forest:

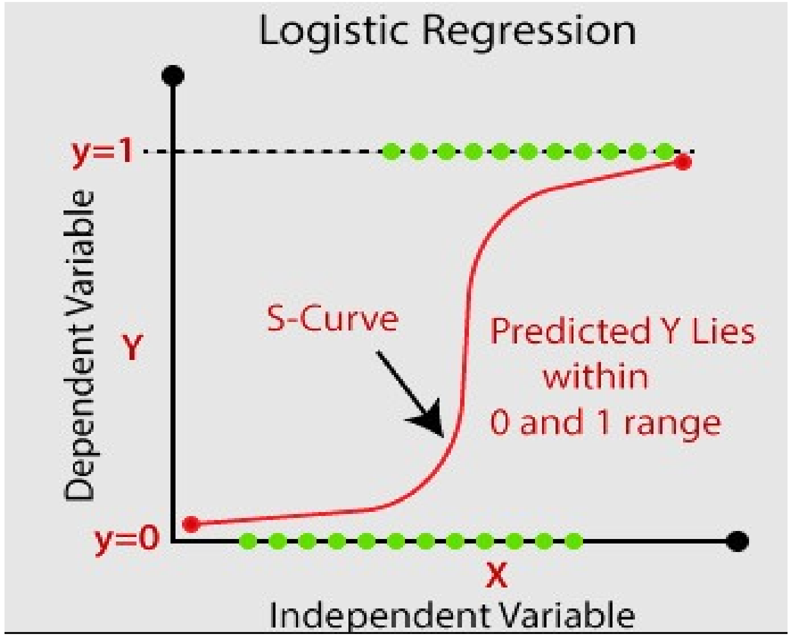
- collection of decision trees from random subsets of the data, resulting in a combination of trees that may be more accurate in prediction than a single decision tree.

- Preferred for tasks requiring high accuracy and robustness, such as image classification.



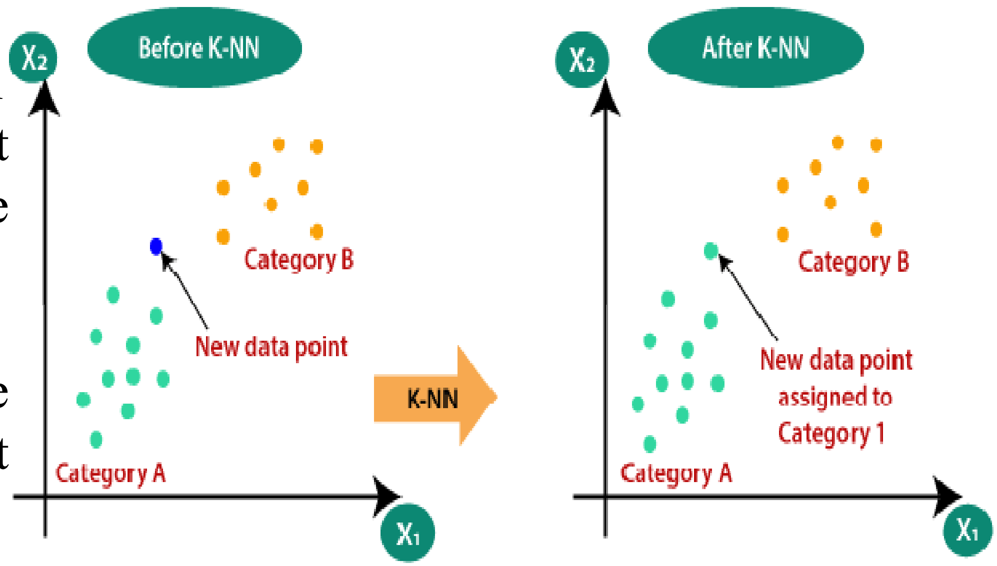


Supervised Machine Learning



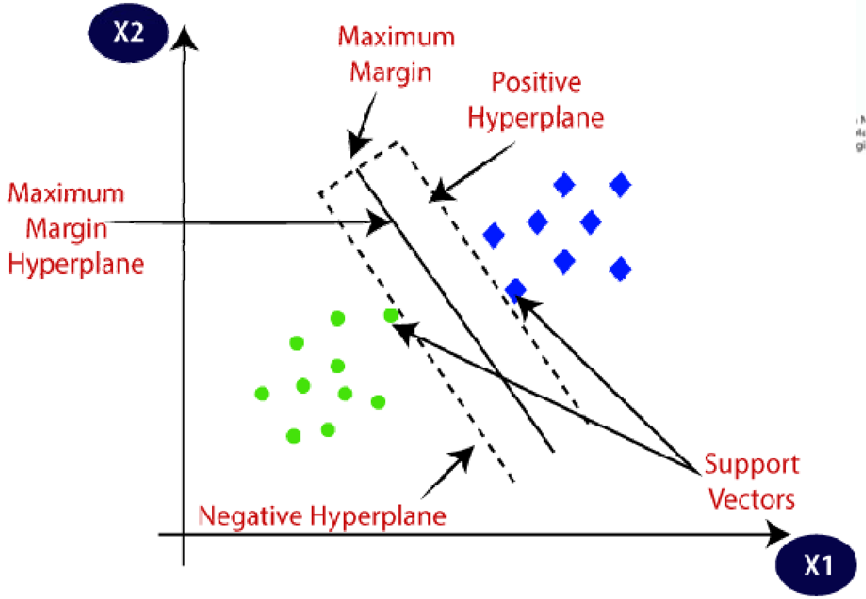
4. **Logistic Regression:** Logistic Regression is used to determine if an input belongs to a certain group or not

5. **k-Nearest Neighbors:**
   - It involves grouping the closest objects in a dataset and finding the most frequent or average characteristics among the objects.
   - To find the optimal K value, use cross-validation to evaluate model performance across different K values, and select the K that minimizes the error rate or maximizes accuracy.
   - Common methods include plotting the error rate versus K and choosing the K with the lowest error.



Supervised Machine Learning

6. **SVM:** Support Vector Machines create coordinates for each object in an n-dimensional space and uses a hyperplane to group objects by common features



7. **Naive Bayes:** algorithm that assumes independence among variables and uses probability to classify objects based on features

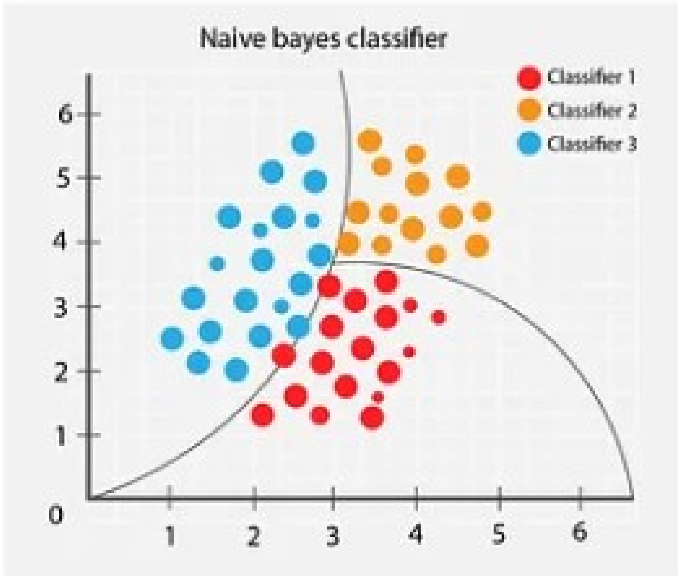
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$

where A and B are events and $P(B) \neq 0$.

- $P(A \mid B)$ is a conditional probability: the likelihood of event A occurring given that B is true.
- $P(B \mid A)$ is also a conditional probability: the likelihood of event B occurring given that A is true.
- $P(A)$ and $P(B)$ are the probabilities of observing A and B independently of each other; this is known as the marginal probability.



Supervised Machine Learning

8. Boosting algorithms:

- Boosting algorithms, such as Gradient Boosting Machine, XGBoost, AdaBoost, CatBoost and LightGBM, use ensemble learning.
- They combine the predictions from multiple algorithms while taking into account the error from the previous algorithm.
- Boosting algorithms combine multiple weak learners in a sequential method, which iteratively improves observations.
- This approach helps to reduce high bias that is common in machine learning models.

What is a Classifier in Machine Learning?

- A classifier is a machine learning algorithm that assigns an object as a member of a category or group.

- **For example**, classifiers are used to detect if an email is spam, or if a transaction is fraudulent.

Decision Tree Vs Decision Tree Classifier

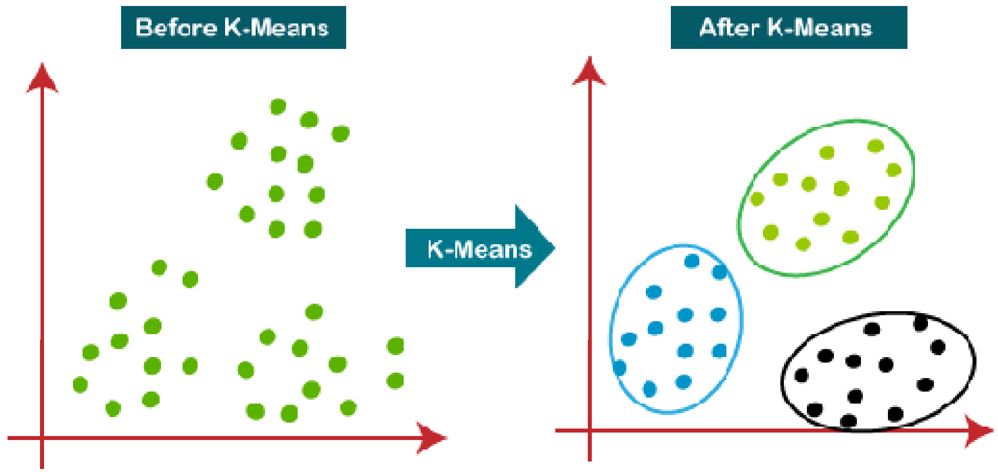
Feature	Decision Tree (General Concept)	Decision Tree Classifier
Purpose	Can be used for both classification and regression tasks.	Specifically used for classification tasks.
Output	Can be a continuous value (regression) or a class label (classification).	Outputs a class label (e.g., "spam" or "not spam").
Examples	Predicting house prices (regression), classifying emails (classification)	Classifying emails as "spam" or "not spam", determining if a patient has a disease.
Components	Root node, internal nodes, branches, and leaf nodes.	Same components, but the leaf nodes represent class labels.
Splitting Criteria	Splits data based on minimizing error in regression or impurity in classification.	Splits data based on minimizing classification error.

Random Forest vs Random Forest classifier

Feature	Random Forest (General Concept)	Random Forest Classifier
Purpose	Can be used for both regression and classification tasks, as well as other tasks like feature selection.	Specifically designed for classification problems, where the goal is to predict a discrete label or class.
Output	Depending on the task: predicted values (regression) or classes (classification).	A predicted class label, typically based on the majority vote of the individual trees.
Algorithm	Combines multiple decision trees; the final output depends on the task (e.g., averaging for regression, majority vote for classification).	Uses multiple decision trees where each tree votes on the class; the most voted class is the final output.
Examples	Can be used for predicting house prices (regression) or identifying if an email is spam (classification).	Used specifically for classification tasks like predicting if a patient has a disease based on medical records.
Evaluation Metrics	Mean Squared Error (MSE) for regression, Accuracy, Precision, Recall for classification.	Accuracy, Precision, Recall, F1-score, ROC-AUC, etc., for evaluating classification performance.

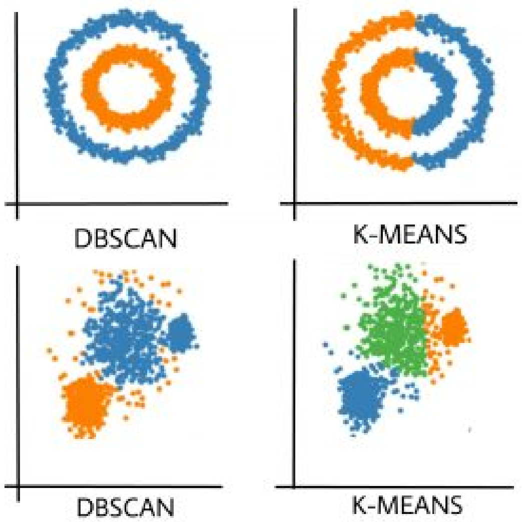
Unsupervised Machine Learning

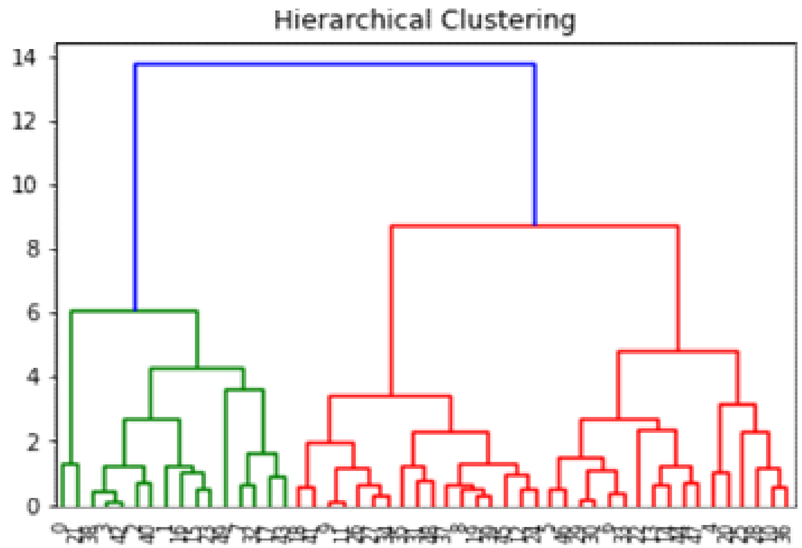
1. **K-Means:** finds similarities between objects and groups them into K different clusters.



2. **Hierarchical Clustering:** builds a tree of nested clusters without having to specify the number of clusters.

3. **Density-Based Spatial Clustering of Applications with Noise (DBSCAN):**
   - finds core samples in regions of high density and expands clusters from them.
   - good for data which contains clusters of similar density.





Accuracy and Loss Metrics:

- **Mean Absolute Error (MAE):** Measures the average magnitude of errors in a set of predictions, without considering their direction.
  - Formula: $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- **Mean Squared Error (MSE):** Measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.
  - Formula: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- **Root Mean Squared Error (RMSE):** The square root of the average of squared differences between predicted and actual values. It gives a sense of the average error magnitude.
  - Formula: $RMSE = \sqrt{MSE}$
- **MAE, MSE, RMSE:** Lower values indicate better model performance.

Accuracy and Loss Metrics:

- **R-squared (R^2):** Indicates how well the independent variables explain the variance in the dependent variable.
  - Formula: $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$, where SS_{res} is the sum of squares of residuals and SS_{tot} is the total sum of squares.

- **Adjusted R-squared (for Multiple Linear Regression):** Adjusts the R2 value for the number of predictors in the model, providing a more accurate measure when multiple variables are used.

- **R2 and Adjusted R2**: Higher values indicate better model performance.

What is Time Series Machine Learning?

- A time-series machine learning model is one in which one of the independent variables is a successive length of time minutes, days, years etc.), and has a bearing on the dependent or predicted variable.

- Time series machine learning models are used to predict time-bound events.

- **for example** - the weather in a future week, expected number of customers in a future month, revenue guidance for a future year, and so on.