SUPERVISED LEARNING

Consider we have a problem to find out whether a penguin is male or female with the given data.

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	Male
1	Adelie	Biscoe	39.5	17.4	186.0	3800.0	Female
2	Gentoo	Torgersen	46.1	13.2	211.0	4900.0	Female
3	Gentoo	Biscoe	50.0	16.3	230.0	5700.0	Male
4	Chinstrap	Dream	48.5	17.9	192.0	3500.0	Female
5	Chinstrap	Dream	50.0	19.5	196.0	3900.0	Male

In traditional approach, we need to use certain rules like these

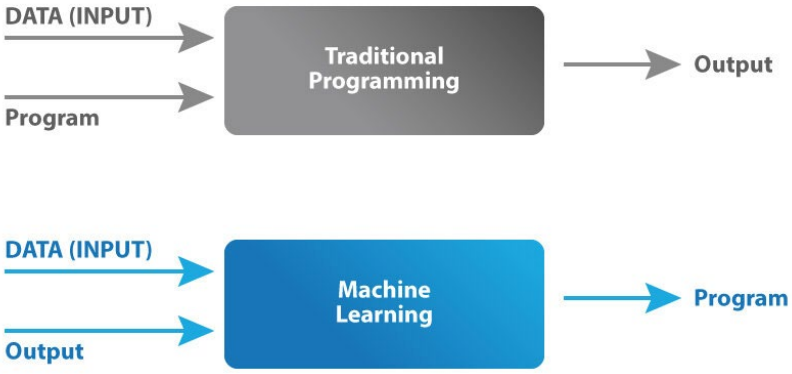
- If the species is **Adelie and the island is Biscoe**, classify as **male**
- If the species is **Chinstrap and the bill length is greater than 47mm** classify as **female**
- If the species is **Gentoo and the body mass is greater than 4800g**, classify as **female,** etc.

But this method has the limitations especially when the amount of data increases

- May struggle with complexity as the number of rules increases.

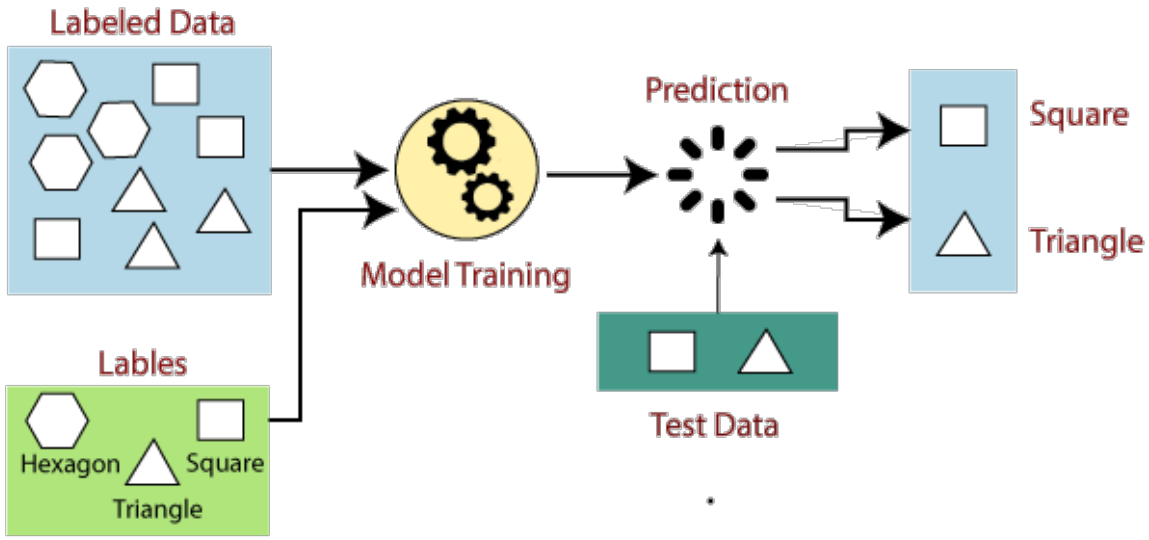
- Requires manual updates and modifications for new information or changing requirements.
- Relies on human expertise and predefined rules.

To overcomes this, we can use the ML (Machine Learning) techniques. ML can be explained as **automating and improving the learning process of computers based on their experiences** without being actually programmed i.e., without any human assistance.



For this particular problem we can go through one subset of ML called **Supervised Learning in which machines are trained using well "labelled" training data** (here “male” and “female”), and on basis of that data, machines predict the output.

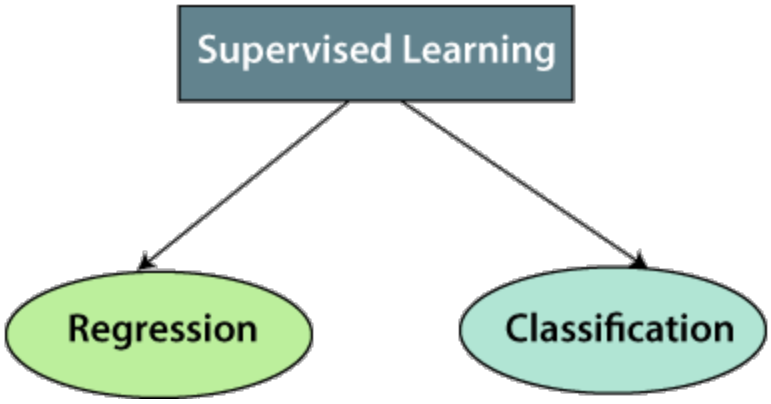
Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**



Supervised Learning has the super powers of

- Adaptability: **models can adapt to new data and changing conditions without requiring manual updates**.
- Automation: automates the learning process by automatically discovering patterns and extracting relevant features from the data. This reduces the need for manual feature engineering and rule definition, saving time and effort.
- Handling Complexity: **algorithms can capture intricate patterns and interactions** among features that may be challenging or impractical to define explicitly in traditional programming.
- Scalability: **models can efficiently handle large amounts of data**, making them scalable for big data challenges.
- Generalization: **They learn from training examples and apply that knowledge to make predictions or decisions on new, unseen data**, improving their ability to handle real-world scenarios.
- Improved Accuracy: models, when trained and tuned properly, **can achieve high levels of accuracy and predictive performance.**

Let's dive into Supervised learning more, it can be classified as



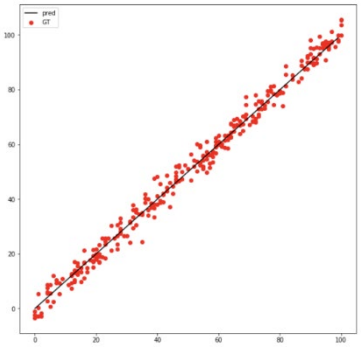
1. Regression

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables.

Applications of Regression includes:

1. **Economic Forecasting:**
   Regression models are used to forecast economic variables such as GDP growth, stock prices, inflation rates, and interest rates.

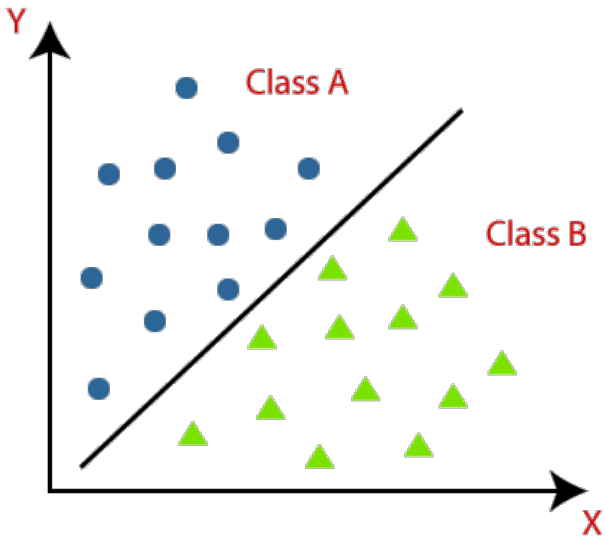
2. **Marketing and Sales Analysis:**
   Regression analysis is utilized to understand the relationship between marketing efforts (advertising, pricing, promotions) and sales. It helps businesses optimize marketing strategies and allocate resources effectively.



3. **Demand and Price Analysis:**
   Regression models are used to analyze the demand for products or services based on factors such as price, income levels, demographics, and market conditions. This information assists in pricing decisions and demand forecasting.
4. **Risk Assessment and Insurance**:
   Regression analysis is used in actuarial science to assess risk factors for insurance claims. It helps insurance companies estimate the probability of claims and determine appropriate premium rates.
5. **Environmental Studies:**
   Regression models are applied to analyze the impact of environmental factors on various phenomena, such as climate change, air quality, water pollution, and species distribution. It helps in understanding and predicting environmental changes.
6. **Real Estate and Housing Market Analysis:**
   Regression models are used to study the factors affecting housing prices, rental rates, and property valuation. It helps buyers, sellers, and real estate professionals make informed decisions.

2. Classification

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, **Yes or No, 0 or 1, Spam or Not Spam, cat or dog, male or female,** etc. Classes can be called as targets/labels or categories.



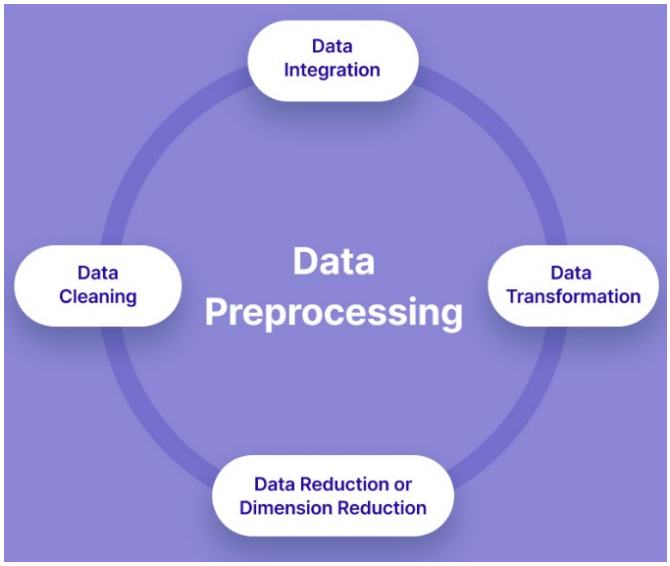
Applications of Classification:

- Email Spam Detection
- Disease Diagnosis: By analyzing patient data such as symptoms, medical history, and test results, the models can classify patients into different disease categories, helping doctors make accurate diagnoses.
- Credit Risk Assessment: By considering factors such as credit history, income, and loan application details, models can classify applicants into high or low credit risk categories.
- Image and Object Recognition: Used in computer vision to recognize and classify objects or patterns within images.
- Sentiment Analysis: Determine the sentiment or emotion expressed in textual data, such as social media posts, customer reviews, or survey responses.
- Fraud Detection: By analyzing patterns and anomalies in transaction data, models can classify transactions as either legitimate or fraudulent.
- Document Classification: Classification models are employed to automatically categorize documents into specific topics or classes. This aids in organizing and retrieving information, as well as filtering content based on user preferences.

So far, we got an overview of Supervised Learning method. Now let's get into a step-by-step process of how supervised machine learning is implemented

1. Data Collection: Gather a dataset that contains paired examples of input data and their corresponding output labels. The quality and size of the dataset are crucial for the success of the model. This step depends on the specific problem you are working on (like the penguin data set given earlier).

2. Data Preprocessing: Clean the data by handling missing values, normalizing or scaling numerical features, and encoding categorical variables. This step is essential to ensure the data is in a suitable format for the learning algorithm.

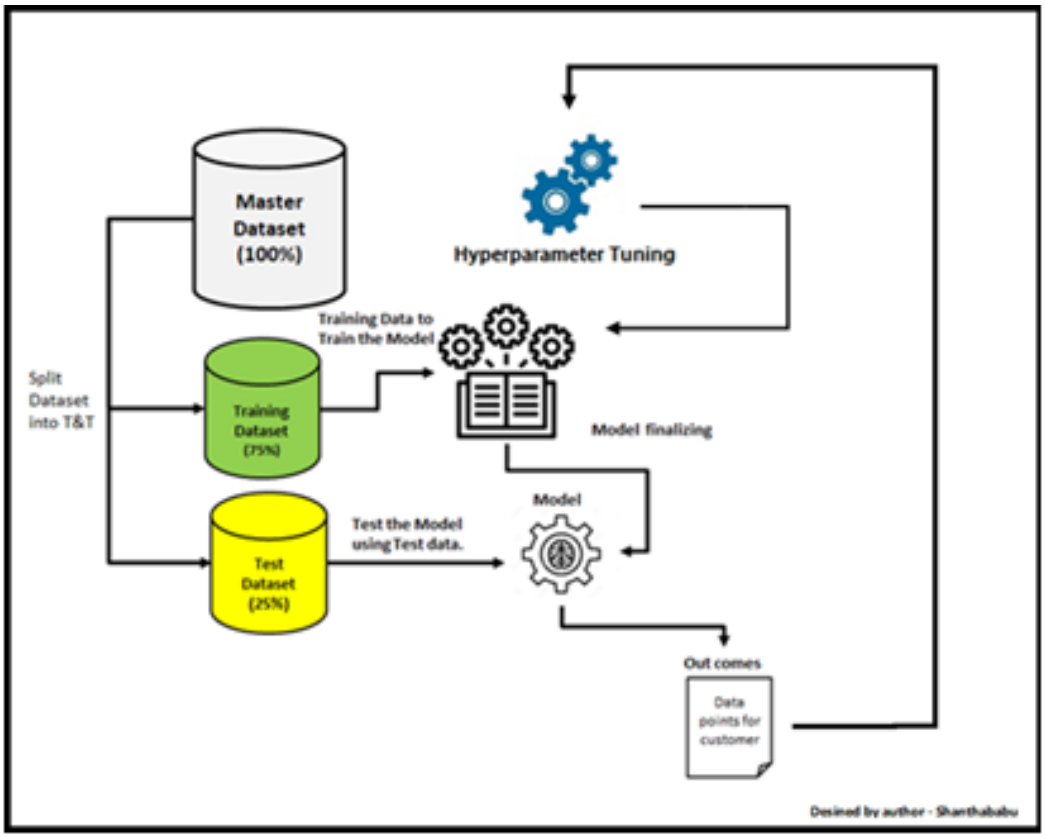


3. Splitting the Dataset: Divide the dataset into two or three subsets: the training set, the validation set, and the test set. The training set is used to train the model, the validation set is used to tune hyperparameters and avoid overfitting, while the test set is used to evaluate the model's performance on unseen data.

4. Selecting a Model: Choose a suitable supervised learning algorithm for your task, such as linear regression, logistic regression, decision trees, support vector machines, or neural networks. The choice of the model depends on the nature of the data and the problem you are trying to solve.

5. Model Training: Feed the training data into the selected model, and it will learn to make predictions based on the input features and their associated labels. The model will adjust its internal parameters during the training process to minimize the prediction errors.

6. Hyperparameter Tuning: Adjust the hyperparameters of the model to optimize its performance on the validation set. Hyperparameters are settings that are not learned during training and need to be set beforehand.

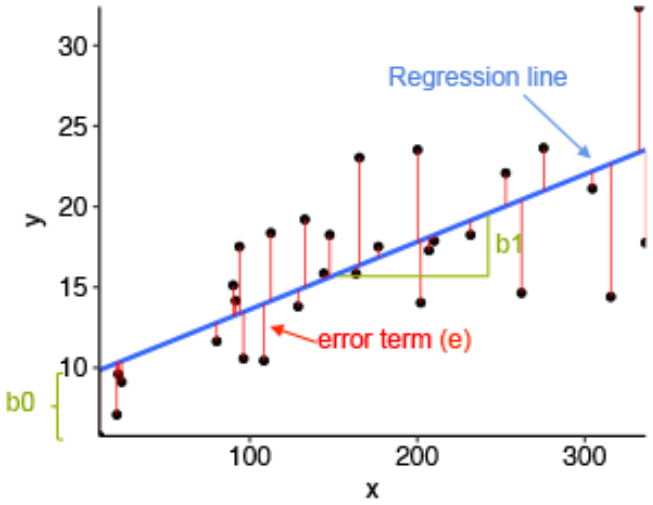


7. Model Evaluation: After training and tuning, evaluate the model's performance on the test set to estimate how well it will generalize to new, unseen data. Common evaluation metrics include accuracy, precision, recall, F1-score, and mean squared error, among others.

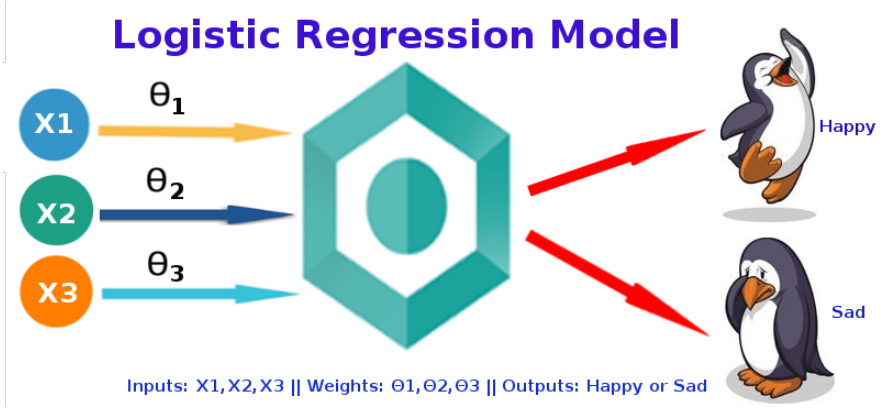
8. Model Deployment: If the model's performance is satisfactory, it can be deployed in a real-world application to make predictions on new incoming data.

Now you have an idea of where to start and how to pass through, but before proceeding we need to chose the best model to train our model also (it actually depends on the problems we are trying to solve). Here are some common ones to choose from,

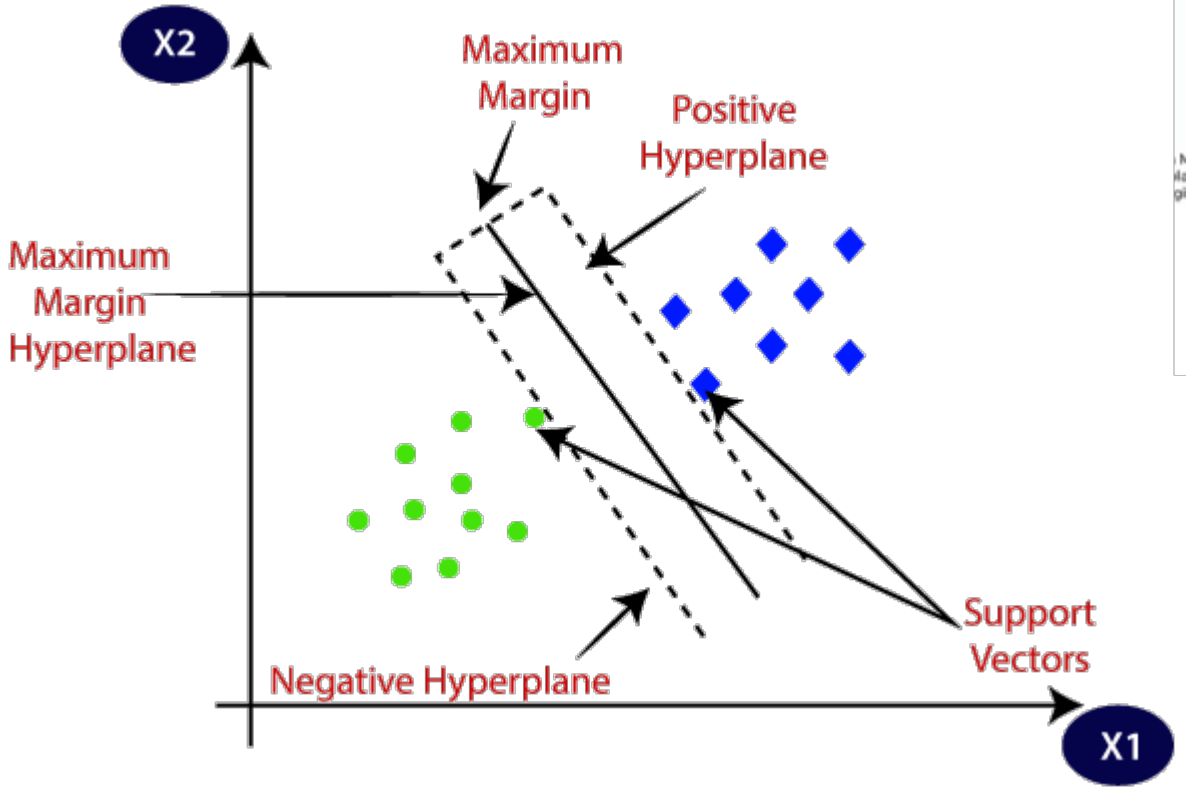
1. Linear Regression: Linear regression is a simple and widely used method for predicting a continuous output variable (target) based on one or more input features by fitting a linear equation to the observed data.



2. Logistic Regression: Logistic regression is used for binary classification problems, where the output variable takes one of two possible classes. It models the probability of the target belonging to a particular class using a logistic function.

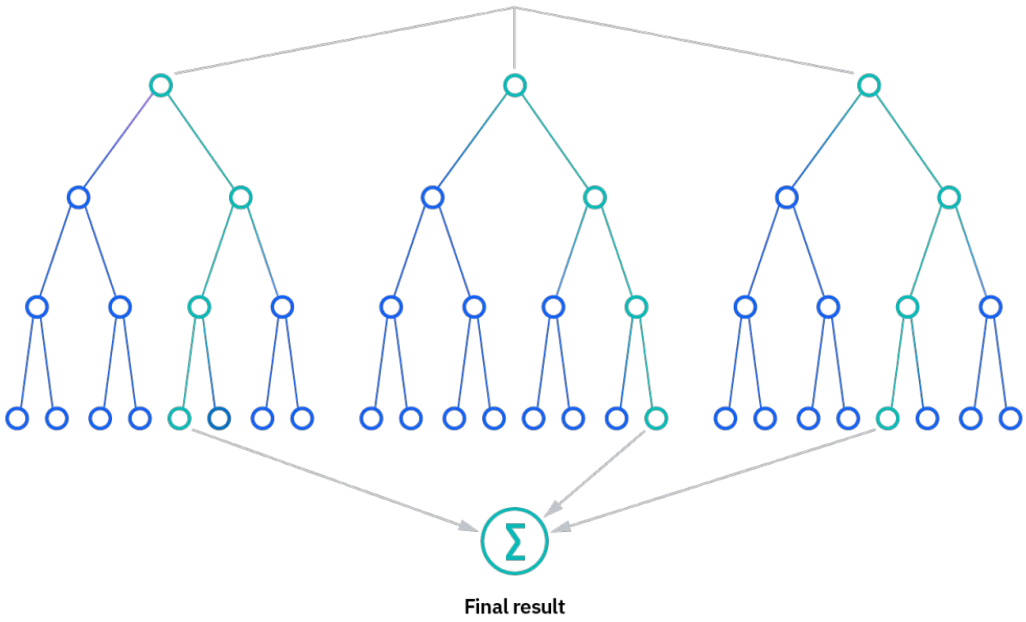


3. Support Vector Machines (SVM): SVM is a powerful and widely used algorithm for classification and regression tasks. However, primarily, it is used for Classification problems in Machine Learning. It finds an optimal hyperplane that best separates different classes in the data. SVM is known for its ability to handle high-dimensional data and works well for both linearly and non-linearly separable datasets.



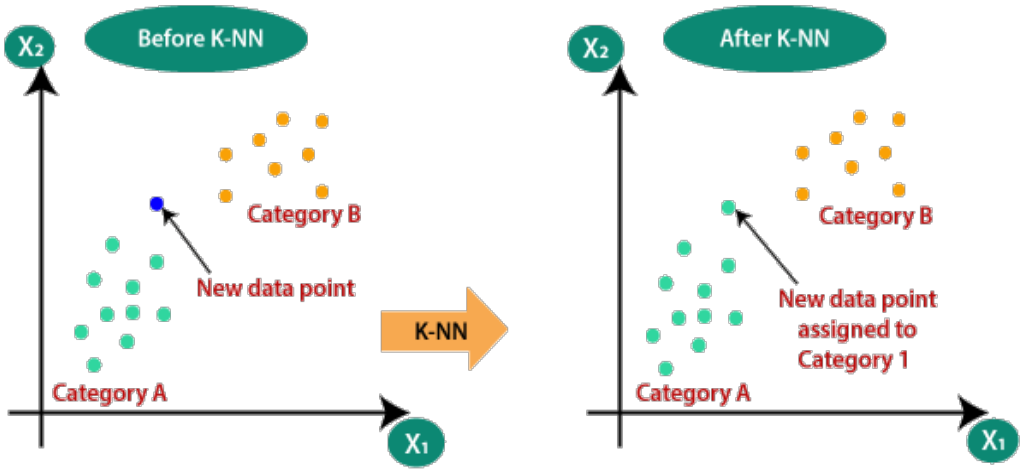
4. Decision Trees: Decision trees create a tree-like model of decisions and their possible consequences. Each internal node represents a test on a feature, each branch corresponds to the outcome of the test, and each leaf node represents a class label or numerical value
.

5. Random Forest: Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the mode (classification) or mean prediction (regression) of the individual trees.



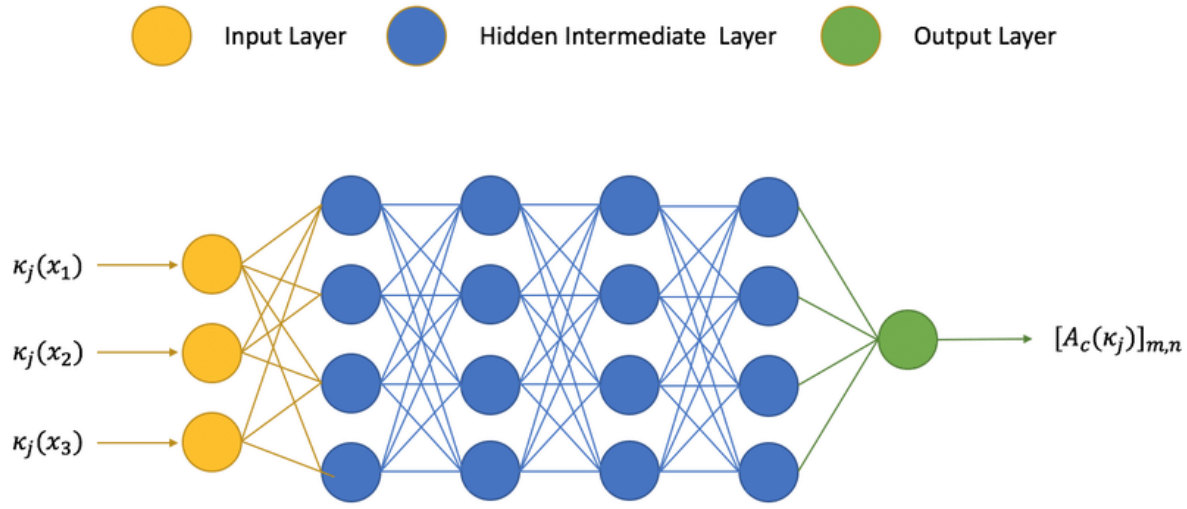
6. Naive Bayes: Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with the assumption of independence between features. It is particularly useful for high-dimensional datasets.

7. K-Nearest Neighbors (KNN): KNN is a simple and effective algorithm for classification and regression tasks. It assigns a class label or numerical value based on the majority class (or average value) among its k-nearest neighbors in the feature space.



8. Gradient Boosting Machines (GBM): GBM is an ensemble learning technique that builds multiple weak learners (often decision trees) sequentially. Each new tree corrects the errors of its predecessor, leading to improved predictions.

9. Neural Networks (Deep Learning): Neural networks are a class of algorithms inspired by the biological neural networks in the human brain. Deep learning, a subset of neural networks, uses multiple layers of interconnected neurons to learn hierarchical representations of data.



While Going through the above contents, have you noticed some notations and terms that aren't familiar, Let's see a little about those and some related things too,

Normalization

Normalization is one of the most frequently used data preparation techniques, which helps us to change the values of numeric columns in the dataset to use a common scale.

Mathematically, we can calculate normalization with the below formula:

$$Xn = \frac{X - Xmin}{Xmax - Xmin}$$

X_n = Value of Normalization

X_{max} = Maximum value of a feature

X_{min} = Minimum value of a feature

Normalization techniques in Machine Learning

- Min-Max Scaling

- Standardization scaling

Feature scaling

Feature scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.

Loss function

A mathematical function that quantifies the difference between the predicted output and the actual output (label). The model aims to minimize this loss during training. A loss function is **for a single training example**

Cost function

Cost function is a measure of how wrong the model is in estimating the relationship between X(input) and Y(output) Parameter. A cost function, on the other hand, is the **average loss over the entire training dataset**, it is an important parameter that determines how well a machine learning model performs for a given dataset.

Gradient descent

It is an optimization algorithm used in machine learning to minimize the cost function by iteratively adjusting parameters in the direction of the negative gradient, aiming to find the optimal set of parameters.

Precision and Recall in Machine Learning

Precision and recall are performance metrics used for pattern recognition and classification in machine learning. Some of the models in machine learning require more precision and some model requires more recall.

Precision is defined as the ratio of correctly classified positive samples (True Positive) to a total number of classified positive samples (either correctly or incorrectly).

The recall is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples. The recall measures the model's ability to detect positi ve samples.

Some new terms we came across are

- Dependent Variable: These are the output of the process, which we want to predict or understand. It is also called target variable(y).

- Independent Variable: The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a features(x).

- Outliers: Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.

- Multicollinearity: If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity.

- Training Set: A set of observations used to generate machine learning models.

- Validation Set: A set of observations used to tune the hyperparameters

- Test Set: A set of observations used at the end of model training and validation to assess the predictive power of your model.

- Model: The algorithm or mathematical function that learns from the training data and makes predictions on new, unseen data.

- Hypothesis function: In the context of supervised learning, this refers to the function learned by the model that maps the input features to the predicted output.
- Generalization: The ability of a supervised learning model to perform well on new, unseen data, beyond the data it was trained on.

- Underfitting and Overfitting: If our algorithm works well with the training dataset but not well with test dataset, then such problem is called Overfitting. And if our

algorithm does not perform well even with training dataset, then such problem is called underfitting.

- Confusion matrix: The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known.

- ROC curve: An ROC curve, or receiver operating characteristic curve, is like a graph that shows how well a classification model performs. It helps us see how the model makes decisions at different levels of certainty.

Now you got a solid understanding about Supervised Learning and the terminologies related to it, Let's see some of its pros and cons also

Pros of Supervised Machine Learning

- You will have **an exact idea about the classes** in the training data.
- Supervised learning is **a simple process** for you to understand. In the case of unsupervised learning, we don't easily understand what is happening inside the machine, how it is learning, etc.
- You can **find out exactly how many classes** are there before giving the data for training.
- It is **possible to be very specific** about the definition of the classes, that is, you **can train the classifier in a way which has a perfect decision boundary** to distinguish different classes accurately.
- After the entire training is completed, you **don't necessarily need to keep the training data in your memory**. Instead, you **can keep the decision boundary as a mathematical formula**.
- Supervised learning can be very **helpful in classification problems.**
- Another typical task of supervised machine learning is **to predict a numerical target value** from some given data and labels.

- Supervised learning in Machine Learning allows you to **collect data or produce a data output from the previous experience**
- Helps you to **optimize performance criteria** using experience
- Supervised machine learning helps you to **solve various types of real-world computation problems.**
- An example of linear regression is **easy to understand and fairly straightforward**. It can also be **normalized to avoid overfitting.** Moreover, by

using stochastic gradient descent, linear models can be updated easily with new data.

- The use of well-known and labelled input data makes supervised learning produce **a far more accurate and reliable than unsupervised learning.** With the access to labels, it can use to improve its performance on some tasks.
- Efficient in finding solutions to **several linear and non-linear problems such as classification, robotics, prediction and factory control**.
- Able to **solve complex problem by having hidden neuron layer**.

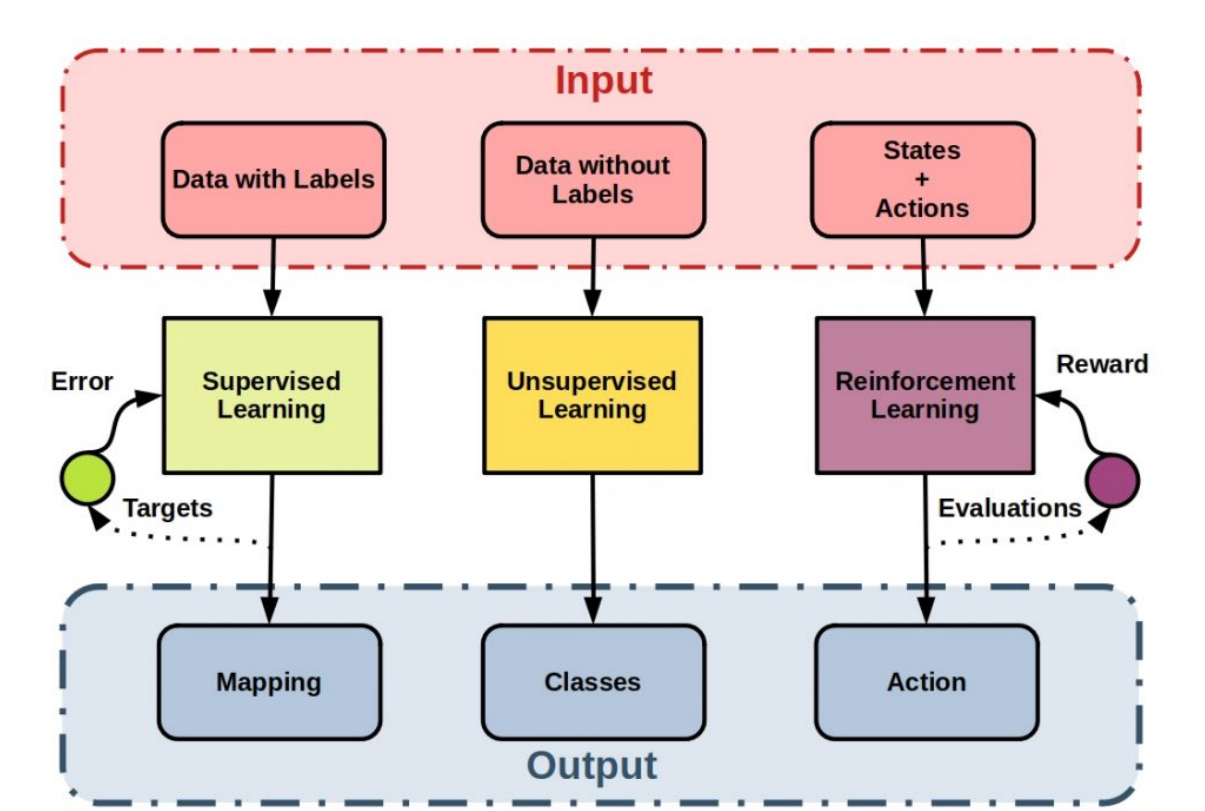
Cons of Supervised Machine Learning

- Supervised learning is limited in a variety of sense so that it **can't handle some of the complex tasks** in machine learning.
- Supervised learning **cannot give you unknown information** from the training data like unsupervised learning do.
- It **cannot cluster or classify data by discovering its features on its own**, unlike unsupervised learning.
- In the case of classification, **if we give an input that is not from any of the classes in the training data, then the output may be a wrong class label**. For example, let's say you trained an image classifier with cats and dogs data. Then if you give the image of a giraffe, the output may be either cat or dog, which is not correct.
- Similarly, let's say **your training set does not include some examples that you want to have in a class.** Then, when you use those examples after training, you **might not get the correct class label as the output.**
- While you are training the classifier, you **need to select a lot of good examples from each class.** Otherwise, the accuracy of your model will be very less. This is difficult when you deal with a large amount of training data.
- Usually, training needs **a lot of computation time**, so do the classification, especially if the data set is very large. This will test your machine's efficiency and your patience as well.
- We **cannot always give lots of information with supervision.** A lot of the time, the machine needs to learn by itself from the training data.
- **Takes a long time** for the algorithm to compute by training because supervised learning can grow in complexity. Therefore, it is not giving result in real time since majority of world's data is unlabeled, the performance is quite limited.
- Performs poorly when there are non-linear relationships. One of supervised learning method like linear regression not flexible to apprehend more complex structure. It takes a lot of computation time and also difficult to append the right polynomials or interaction terms.

- It's **not cost efficient** if the data keeps growing that adds to the uncertainty of data labelling to predefine outputs. Example, It is costly to manually label an image dataset, and the most high quality image dataset has only one thousand labels.
- Classifying big data can be a real challenge.
- **Decision boundary might be overtrained** if your training set which doesn't have examples that you want to have in a class
- Data preparation and pre-processing is always a challenge.
- Anyone can **overfit** supervised algorithms easily.

Having a better understanding of what makes supervised learning differ from other machine learning techniques is also important to know

	SUPERVISED LEARNING	UNSUPERVISED LEARNING
Input Data	Uses Known and Labelled Data as input	Uses Unknown Data as input
Computational Complexity	Less Computational Complexity	More Computational Complex
Real Time	Uses off-line analysis	Uses Real Time Analysis of Data
Number of Classes	Number of Classes are known	Number of Classes are not known
Accuracy of Results	Accurate and Reliable Results	Moderate Accurate and Reliable Results
Output data	Desired output is given.	Desired output is not given.
Model	In supervised learning it is not possible to learn larger and more complex models than with supervised learning	In unsupervised learning it is possible to learn larger and more complex models than with unsupervised learning
Training data	In supervised learning training data is used to infer model	In unsupervised learning training data is not used.
Another name	Supervised learning is also called classification.	Unsupervised learning is also called clustering.
Test of model	We can test our model.	We cannot test our model.
Example	Optical Character Recognition	Find a face in an image.



Comparing with some other techniques,

Semi-Supervised Learning:

- Combination of Supervised and Unsupervised: Semi-supervised learning is a hybrid approach that leverages both labeled and unlabeled data. It uses a small amount of labeled data along with a larger pool of unlabeled data to train the model.
- Advantages: Semi-supervised learning can be beneficial when acquiring labeled data is costly or time-consuming, as it allows the model to benefit from the additional information in the unlabeled data.
- Use Cases: Semi-supervised learning is useful in scenarios where obtaining labeled data is challenging but there is an abundance of unlabeled data, such as in certain natural language processing tasks or image recognition tasks.

Reinforcement Learning:

- Feedback Mechanism: While supervised learning relies on labeled data, reinforcement learning operates in an interactive setting where the algorithm learns from feedback in the form of rewards or penalties based on its actions in an environment.

- Goal-Oriented: The objective in reinforcement learning is to learn a policy that maximizes cumulative rewards over time, enabling the agent to make decisions to achieve a specific goal.
- Use Cases: Reinforcement learning is well-suited for scenarios where the model needs to learn by trial and error, such as game playing, robotic control, and optimization problems.

Transfer Learning:

- Reusability of Knowledge: Transfer learning is a technique that allows a model trained on one task to be reused or adapted for a related or different task.
- Knowledge Transfer: In supervised learning, the model is trained for a specific task with labeled data. In transfer learning, the pre-trained model's knowledge is used as a starting point for a new task, potentially requiring less data and computation.
- Use Cases: Transfer learning is commonly used in situations where the target task has limited data, as it can leverage the knowledge gained from larger, related datasets.

In summary, supervised learning is a specific type of machine learning that requires labeled data for training. Other techniques like unsupervised learning, semi-supervised learning, reinforcement learning, and transfer learning address different learning scenarios and data availability, making the machine learning field diverse and adaptable to various real-world problems

What will be the experts going to ask you about Supervised Learning, Let's see some different aspects of it through some interview questions

1)Give a real-life example of *Supervised Learning.*

Answer

- You get a bunch of photos **with information about what is on them** and then you train a model to recognize new photos.

- You have a bunch of molecules and **information about which are drugs** and you train a model to answer whether a new molecule is also a drug.

- Based on past information about spams, filtering out a new incoming email into **Inbox** (normal) or **Junk folder** (Spam)

- Cortana or any speech automated system in your mobile phone trains your voice and then starts working based on this training.

- Train your handwriting to OCR system and once trained, it will be able to convert your hand-writing images into text (till some accuracy obviously)

2) What is *Bias* in Machine Learning?

Answer

In supervised machine learning an algorithm learns a model from training data. The goal of any supervised machine learning algorithm is to best estimate the mapping function (f) for the output variable (Y) given the input data (X). The mapping function is often called the target function because it is the function that a given supervised machine learning algorithm aims to approximate.

Bias are **the simplifying assumptions** made by a model to make the target function easier to learn. Generally, linear algorithms have a high bias making them fast to learn and easier to understand but generally less flexible.

Examples of **low-bias** machine learning algorithms include: Decision Trees, k-Nearest Neighbors and Support Vector Machines.

Examples of **high-bias** machine learning algorithms include: Linear Regression, Linear Discriminant Analysis and Logistic Regression.

3) What is *Overfitting* in Machine Learning?

Answer

Overfitting refers to a model that models the training data too well. It happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the model's ability to generalize.

4) What is *Underfitting* in Machine Learning?

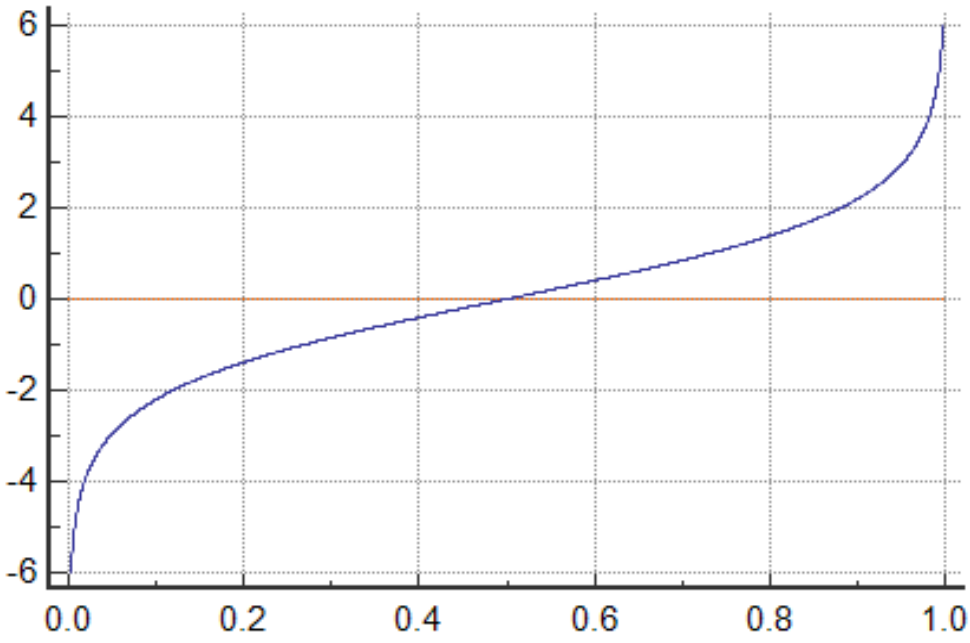
Answer

Underfitting refers to a model that can neither model the training data nor generalizes to new data. An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.
Underfitting is often not discussed as it is easy to detect given a good performance metric. The remedy is to move on and try alternate machine learning algorithms.

5) How do you use a supervised *Logistic Regression* for Classification?

Answer

Logistic regression is a statistical model that utilizes ***logit*** function to model classification problems. It is a regression analysis to conduct when the dependent variable is *binary*. The **logit** function is shown below:



- Looking at the logit function, the next question that comes to mind is *how to fit that graph/equation*. The fitting of the logistic regression is done using the *maximum likelihood* function.
- In a supervised logistic regression, **features** are mapped onto the **output**. The output is usually a categorical value (which means that it is mapped with one-hot vectors or binary numbers).
- Since the **logit** function always outputs a value between 0 and 1, it gives the **probability of the outcome**.

6) What are some challenges faced when using a *Supervised Regression Model*?

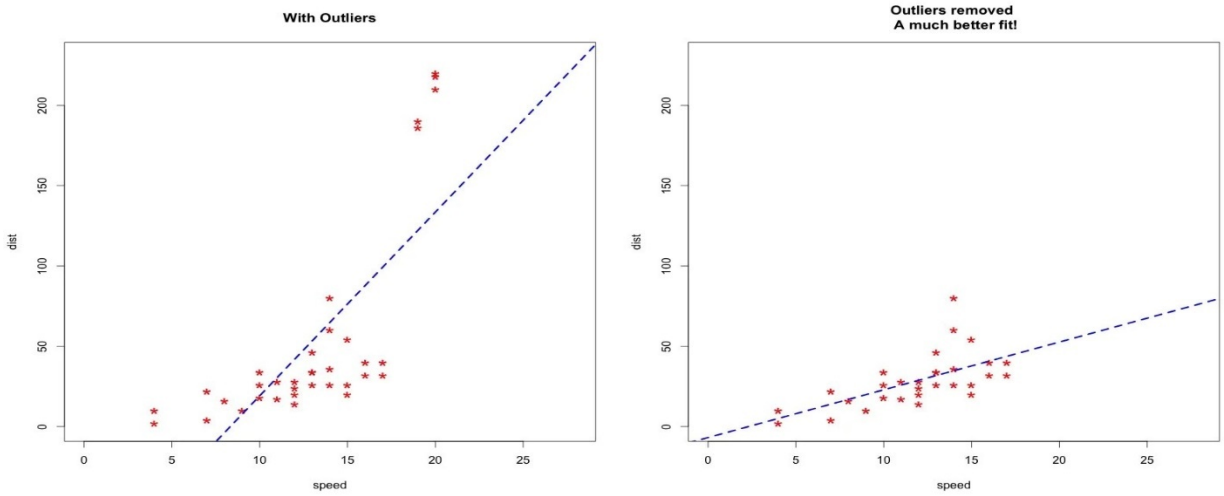
Answer

Some challenges faced when using a supervised regression model are:

- **Nonlinearities:** Real-world data points are more complex and do not follow a linear relationship. Sometimes a non-linear model is better at fitting the dataset. So, it is a challenge to find the perfect equation for the dataset.

- **Multicollinearity:** Multicollinearity is a phenomenon where one *predictor variable* in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. If there is a problem of multicollinearity then even the slightest change in the independent variable causes the output to change erratically.

- **Outliers:** Outliers can change and make huge impact on the machine learning model. This happens because the regression model tries to fit the outliers into the model as well. The problem of outliers is shown in the figure below:

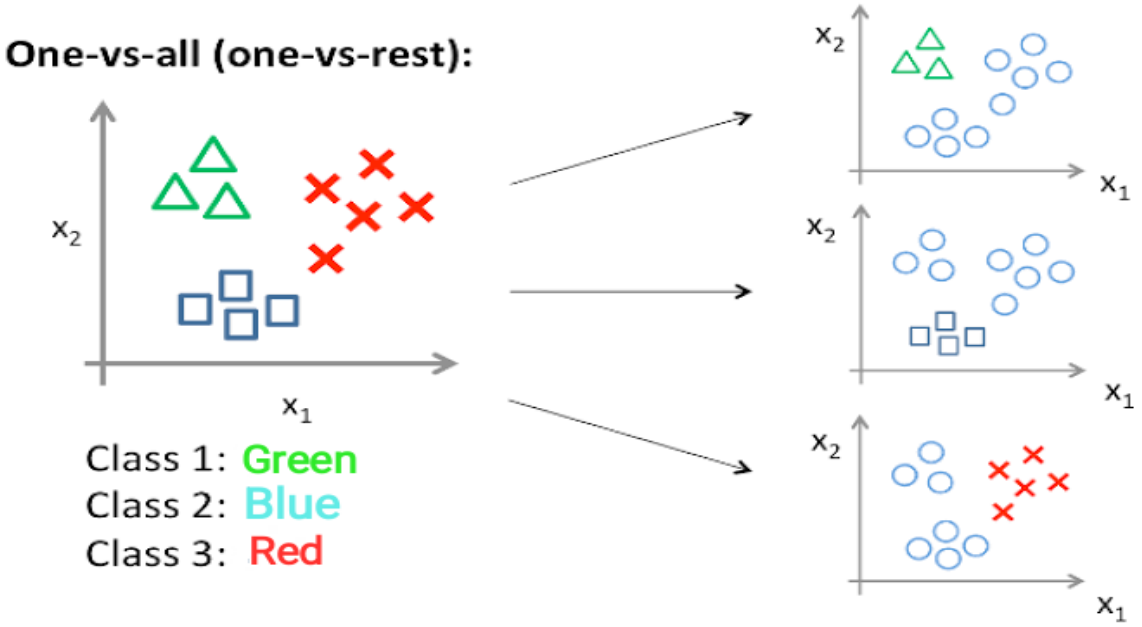


7) What's the difference between *One-vs-Rest* and *One-vs-One*?

Answer

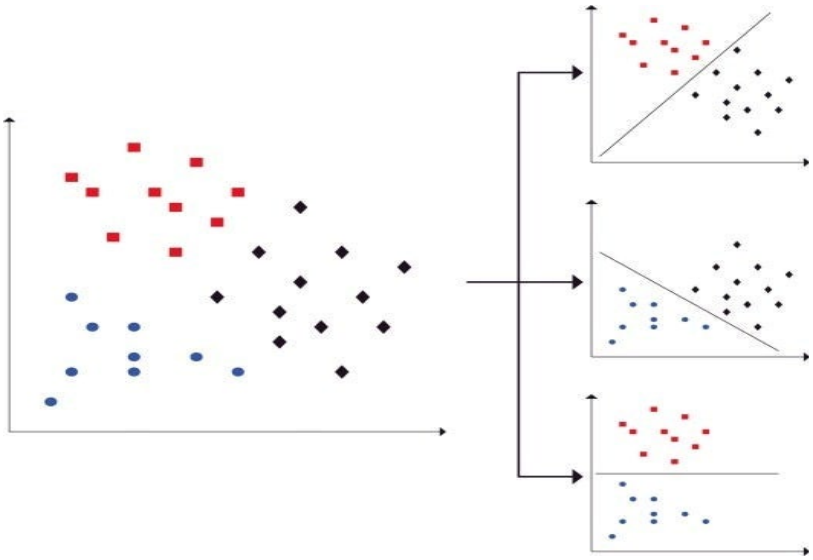
Both **One-vs-Rest** and **One-vs-One** are two techniques for splitting the *multi-class dataset* into *multiple binary classification* problems. It allows us to categorize the *test data* into multiple class labels present in *trained data* as a model *prediction*. One key difference in both techniques is the number of classifiers that are created in each one.

In **One-vs-Rest**, for the N-class instances dataset, we have to generate the N binary classifier models. The predictions are then made using the model that is the most confident. For example, given a multi-class classification problem with examples for each class 'Red', 'Blue' and 'Green', the way to divide into three binary classifications could be as follows:



- Classifier 1:- [Green] vs [Red, Blue].
- Classifier 2:- [Blue] vs [Green, Red].
- Classifier 3:- [Red] vs [Blue, Green].

In **One-vs-One**, we split the primary dataset into one dataset for each class opposite to every other class, so for the N-class instances dataset, we have to generate N*(N-1)/2 binary classifier models. Taking the above example, we have a classification problem having three types: 'Green', 'Blue', and 'Red' (N=3). So we divide this problem into N* (N-1)/2 = 3 binary classifier problems:



- Classifier 1: [Green] vs. [Blue].
- Classifier 2: [Green] vs. [Red].

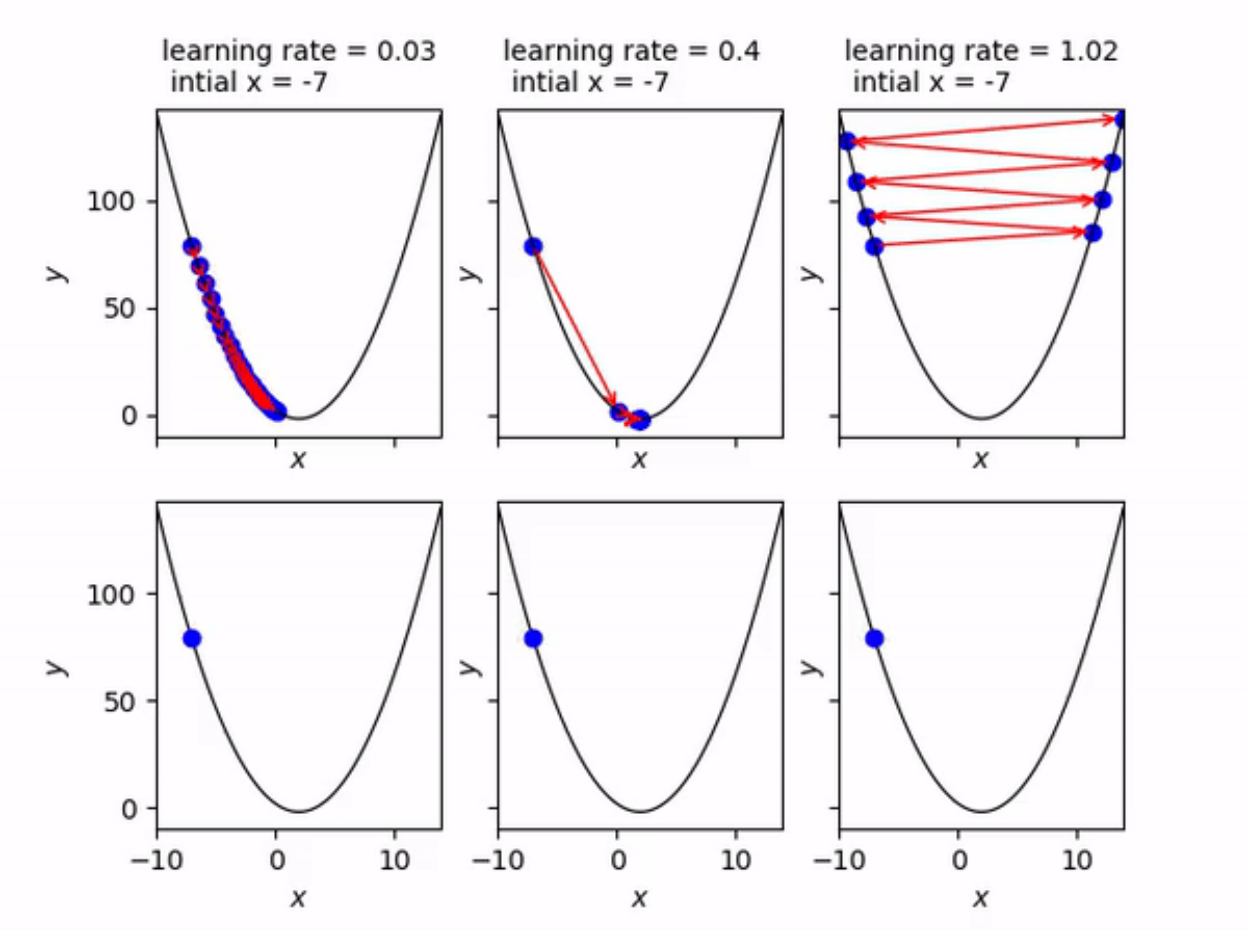
- Classifier 3: [Blue] vs. [Red].

Each binary classifier predicts one class label. When we input the test data to the classifier, then the model with the majority counts is concluded as a result.

8) Provide an intuitive explanation of the *Learning Rate*?

Answer

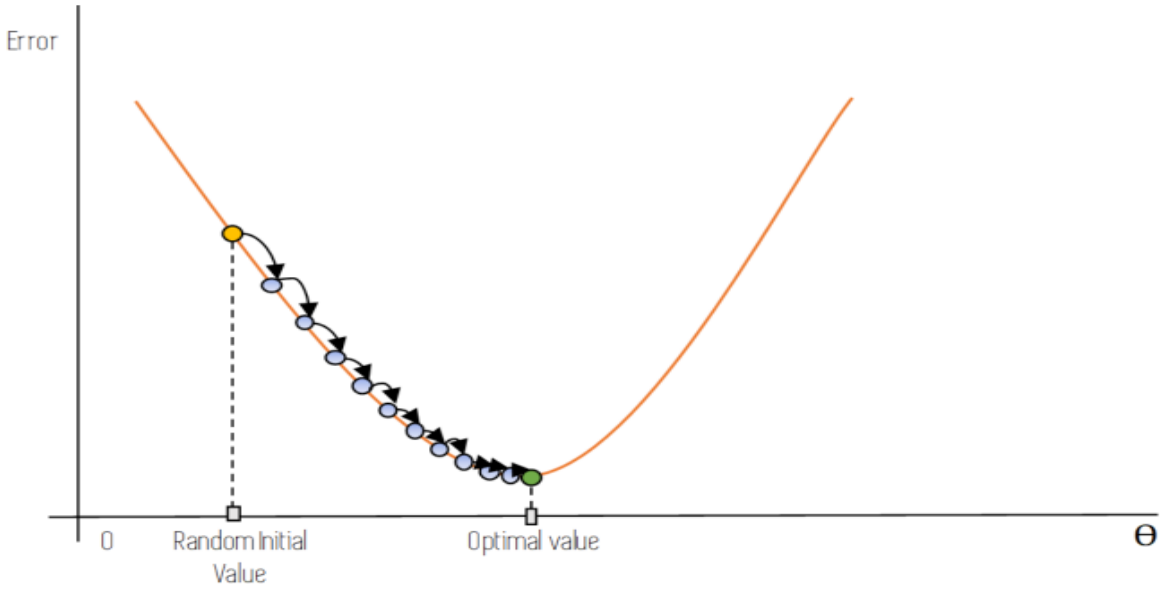
The **Learning Rate** is a **hyper-parameter** that can determine the *speed* or step size at each *iteration* while moving towards a minimal point in **Gradient Descent**. This value should not be too *small* or *too high* because if it's too small then it takes too much time to *converge* and if it's too large then the step size will increase and it moves quickly and never reach global minima *point* even after repeated iterations.



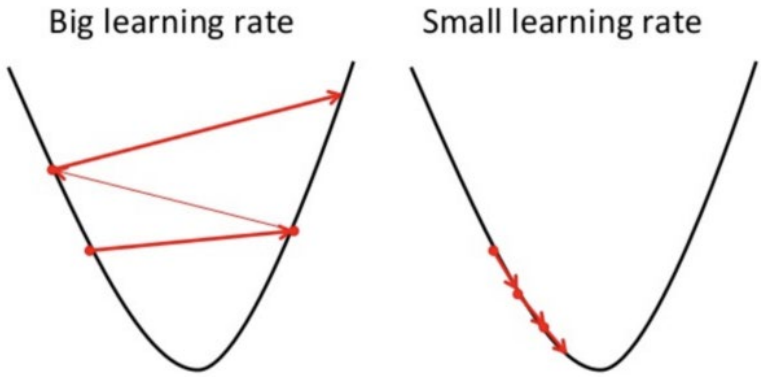
9) Explain the intuition behind Gradient Descent algorithm

Answer

Gradient descent is an optimization algorithm that's used when training a machine learning model and is based on a **convex** function and tweaks its parameters iteratively to minimize a given function to its local minimum (that is, *slope = 0*).
For a start, we have to select a random *bias* and *weights*, and then *iterate* over the slope function to get a slope of 0.



The way we change update the value of the bias and weights is through a variable called the learning rate. We have to be wise on the learning rate because choosing:
A small leaning rate may lead to the model to take some time to learn
A large learning rate will make the model converge as our pointer will shoot and we'll not be able to get to minima.



10) What is the coefficient of determination (R-squared)?

Answer

The coefficient of determination, often referred to as R-squared, is a statistical measure that represents the proportion of the variance in the dependent variable that can be explained by the independent variables in a regression model. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data.

11) What is heteroscedasticity in regression analysis?

Answer

Heteroscedasticity in regression analysis refers to a violation of the assumption that the variability of the residuals is constant across all levels of the independent variables. It occurs when the spread or dispersion of the residuals changes systematically as the values of the independent variables change, indicating unequal variances in the data.

12) What is the purpose of residual analysis in regression?

Answer

The purpose of residual analysis in regression is to assess the adequacy of the regression model. Residuals are the differences between the observed values and the predicted values. By examining the pattern and distribution of residuals, one can identify potential model violations, such as heteroscedasticity, outliers, or nonlinear relationships, and make appropriate model adjustments.

13) What is the difference between an outlier and an influential point in regression?

Answer

An outlier is an observation that deviates significantly from the overall pattern of the data. It has an extreme value and can have a disproportionate impact on the regression model's fit. In contrast, an influential point has a substantial effect on the estimated regression coefficients and can significantly alter the model's results when removed.

14) What is the difference between an influential variable and a confounding variable in regression?

Answer

An **influential variable**, also known as an independent variable or predictor variable, directly affects the dependent variable and is essential for understanding the relationship between the variables in the regression model. On the other hand, a **confounding variable** is a variable that is related to both the independent and dependent variables, creating a spurious association between them if not properly accounted for.

15) What is the difference between mean absolute error (MAE) and mean squared error (MSE)?

Answer

Mean Absolute Error (MAE) and Mean Squared Error (MSE) are both measures of the accuracy of a regression model's predictions, but they differ in how they capture the errors.

MAE: MAE is the average absolute difference between the predicted and actual values. It is calculated as:

MAE = (1/n) * Σ|Y_actual - Y_predicted|

MSE: MSE is the average squared difference between the predicted and actual values. It amplifies larger errors more than MAE. It is calculated as:

MSE = (1/n) * Σ(Y_actual - Y_predicted)^2

In summary, MAE measures the average magnitude of errors, while MSE gives more weight to larger errors due to squaring.

16) What is the difference between L1 regularization (Lasso) and L2 regularization (Ridge)?

Answer

L1 Regularization (Lasso):

- It adds the absolute value of the coefficients as a penalty term to the loss function.
- It tends to produce sparse models by driving some coefficients to exactly zero, effectively performing feature selection.
- It is useful when the dataset has many irrelevant or redundant features.

L2 Regularization (Ridge):

- It adds the squared value of the coefficients as a penalty term to the loss function.
- It encourages smaller but non-zero coefficients, reducing the impact of less influential features.

- It is beneficial when all features are potentially relevant and should be considered in the model.

- It is less likely to result in exactly zero coefficients and tends to distribute the impact across all variables.

17) What is the purpose of interaction terms in regression analysis?

Answer

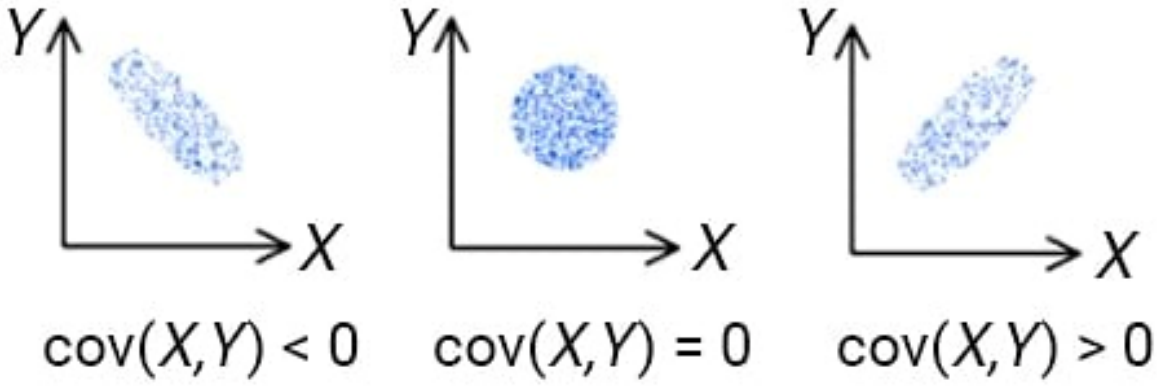
The purpose of interaction terms in regression analysis is to capture the combined effect of two or more independent variables on the dependent variable. By including interaction terms, the model can account for the possibility that the relationship between the variables is not simply additive but may exhibit synergistic or conditional effects.

18) What's the difference between *Covariance* and *Correlation*?

Answer

Covariance measures whether a **variation** in one *variable* results in a variation in *another variable*, and deals with the linear relationship of only 2 variables in the dataset. Its value can take range from -∞ to +∞. Simply speaking **Covariance** indicates the direction of the linear relationship between variables.

Correlation measures how strongly two or more variables are **related** to each other. Its values are between -1 to 1. **Correlation** measures both the strength and direction of the linear relationship between two variables. Correlation is a function of the covariance.



19) What is a neural network regressor?

Answer

A neural network regressor is a type of neural network model used for regression tasks, where the goal is to predict continuous numerical values. It takes input data, passes it

through multiple layers of interconnected neurons, and produces an output value that represents the predicted numerical value.

20) Can you choose a *classifier* based on the *size of the training set*?

Answer

If the *availability of data is a constraint*, i.e. if the training data is smaller or if the dataset has a fewer number of observations and a higher number of features, we can choose algorithms with *high bias/low variance* like ***Naïve Bayes*** and ***Linear SVM***.

If the training data is *sufficiently large* and the number of observations is higher as compared to the number of features, one can go for *low bias/high variance* algorithms like ***K-Nearest Neighbors***, ***Decision trees***, ***Random forests***, and ***kernel SVM***.

21) Could you *convert* Regression into Classification and vice versa?

Answer

Yes, depending on the problem and the context is possible to convert a *regression* to a *classification* problem and vice versa.

To convert Regression into Classification we perform ***discretization***: a process through which we can transform *continuous variables* into an *ordered relationship* (called *ordinal*). For example, amounts in a continuous range between $0 and $100 could be converted into 2 buckets:

Class 0: $0 to $49.

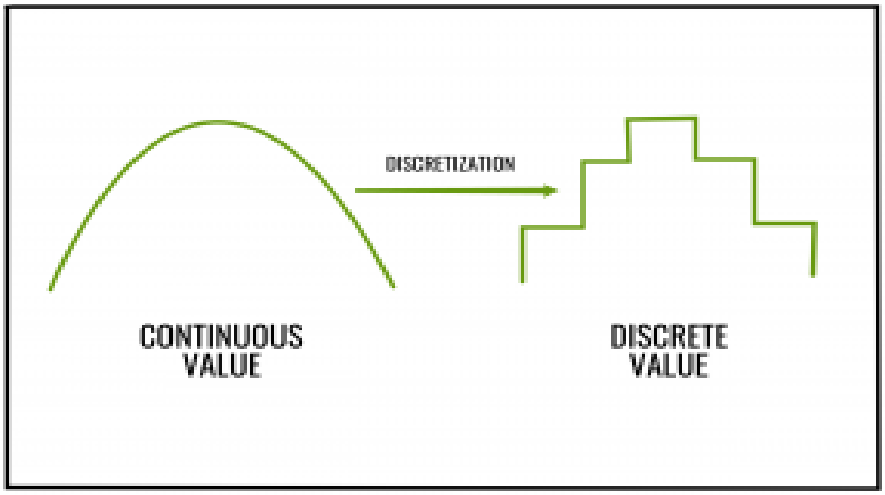
Class 1: $50 to $100.

To convert Classification into Regression, a *label* can be converted into a *continuous range*, i.e. we perform the inverse operation described above:

$0 to $49 for Class 1.

$50 to $100 for Class 2.

Here, we must be sure that the class labels in the classification problem do have a *natural ordinal relationship*. If not, the conversion from classification to regression may result in surprising or poor performance as the model may learn a false or non-existent mapping from inputs to the continuous output range.



22) Name some *classification metrics* and when would you use each one.

Answer

- **Accuracy**: is the proportion of *true positives* among the total number of cases examined. Is suited for classification problems that are well balanced and not skewed.

- **Precision**: is the proportion of *true positive* values between *all positive values*. Is better to use when we want to be very sure of our prediction.

- **Recall**: measures the proportion of *true positives* correctly classified. This is a good metric to use when we want to capture as many positives as possible. For example: If we are building a system to predict if a person has cancer or not, we want to capture the disease even if we are not very sure.

- **F1-score**: it is a number between 0 and 1 and is the *harmonic mean* of *precision* and *recall*. This metric maintains a balance between these two, so if the precision is low, the F1 is low, and if the recall is low again the F1 score is low. It is suited for imbalanced class distributions problems.

23) What is cross-validation and why is it important in classification?

Answer

Cross-validation is a technique used to assess the performance and generalization ability of a classification model. It involves partitioning the data into multiple subsets (folds), training the model on some folds, and evaluating it on the remaining fold. It helps to

estimate how well the model will perform on unseen data and helps in hyperparameter tuning and model selection.

24) How can you handle imbalanced classes in classification?

Answer

Imbalanced classes occur when one class has significantly more instances than the others. Some techniques to handle imbalanced classes include:

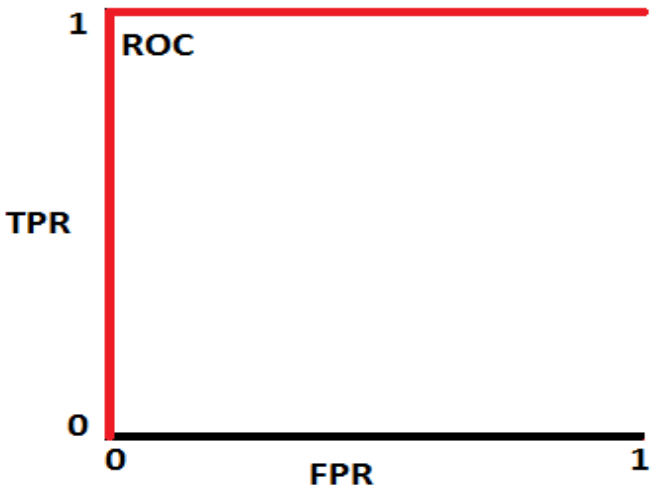
- Resampling the data (under sampling the majority class or oversampling the minority class)
- Using appropriate performance metrics like F1 score or area under the ROC curve (AUC-ROC)
- Using ensemble methods like Random Forests or boosting algorithms that can handle class imbalances

25) How is *AUC - ROC* curve used in classification problems?

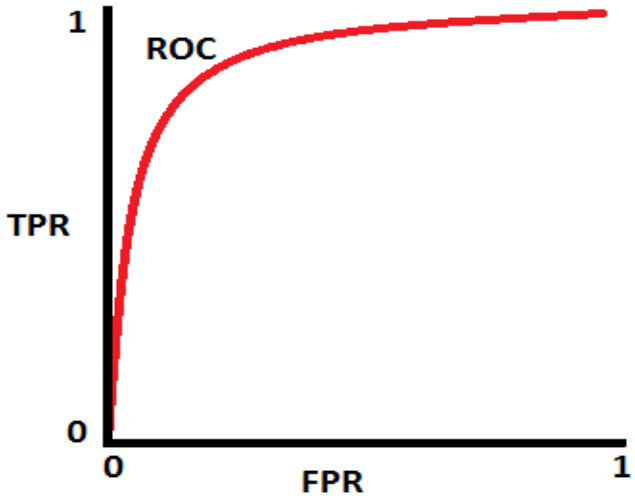
Answer

A **ROC** *(receiver operating characteristic curve)* is used to measure the performance of a classification model at various classification thresholds. This lets us essentially separates the *signal* from the *noise* using the **Area Under the Curve** (AUC) as the measure of the ability of a classifier to distinguish between classes. Its values are between 0 and 1 and are used as a summary of the *ROC curve*.

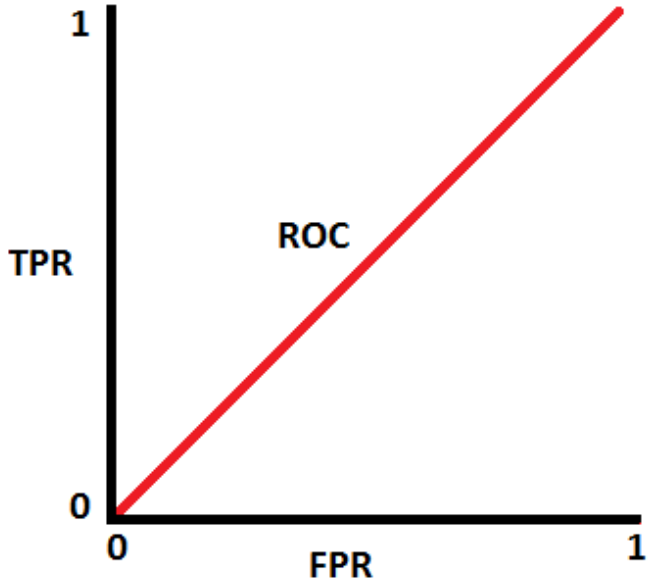
- If AUC = 1 the classifier can perfectly distinguish between all the *Positive* and *Negative* class points correctly.



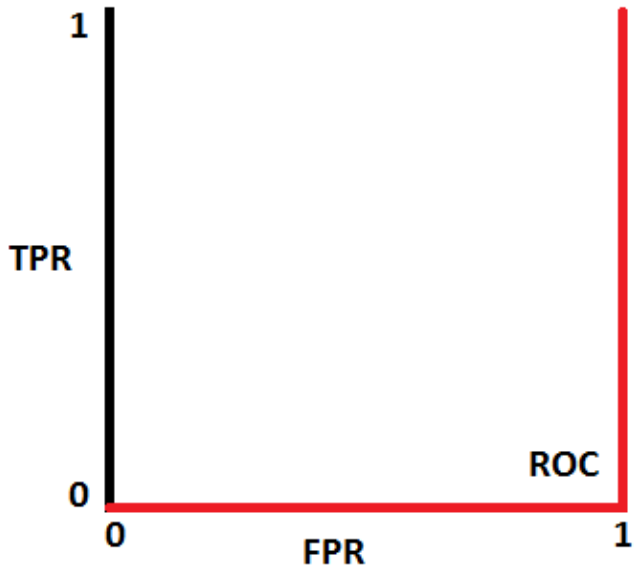
- If 0.5 < AUC < 1 there is a high chance that the classifier will be able to distinguish the *positive class values* from the *negative class values*. This is so because the classifier can detect more numbers of *True positives* and *True negatives* than *False negatives* and *False positives*.



- If AUC = 0.5, then the classifier is not able to distinguish between *Positive* and *Negative* class points. Meaning either the classifier is predicting a random class or a constant class for all the data points.



- If AUC = 0 then the classifier would be predicting all *Negatives* as *Positives*, and all *Positives* as *Negatives*.



In summary, the higher the *AUC value* for a classifier, the better its ability to distinguish between *positive* and *negative* classes.

Here is some YouTube channels for self-study.

Reference	Channel Name	Channel Link
Machine Learning Full Course	Edureka	https://www.youtube.com/watch?v=N5fSpaaxoZc
Machine Learning - Malayalam & Deep Learning	Sivahari Nandakumar	https://youtu.be/fqsOWEBu99U
Complete Roadmap To Follow To Prepare Machine Learning	Krish Naik	https://youtu.be/VOpETRQGXy0
Python Machine Learning Tutorial (Data Science)- Simple Steps	Programming with Mosh	https://youtu.be/7eh4d6sabA0
Simulating the Evolution of Aggression	Primer	https://www.youtube.com/watch?v=YNMkADpvO4w
Math vs animation	Alan Becker	https://youtu.be/B1J6Ou4q8vE
Supervised Learning in Machine learning|| Explanation in Malayalam	brAIn Tek	https://www.youtube.com/watch?v=c0u5TSd8Nx4
Welcome to machine learning	Python Data Structures	https://youtu.be/-xxkz4IvF0Y