Machine Learning (Supervised machine learning)

1. Artificial Neural Network (ANN):

Regression:

- It is used to model the relationship between dependent and independent variables.
- It is a method of predicting the continuous quantity (Numerical type)
- **Multicollinearity**: When dependent variable are co-related with each other or when Independent variable sharing non-linear relationship between each other i.e one variable is affected by affecting the other variable then it is called Multicollinearity.

Following are the type of regression model:

a. Linear regression:

if the dependent variable is continuous in nature then we go for the linear regression.

- **Simple linear regression:**

If we have only one independent variable then it is called simple linear regression.

- **Multilevel linear regression:**

if we have multiple independent variables then we go for the multilevel linear regression

b. Polynomial regression:

If independent variable varies exponentially or non-linearly then we go for the polynomial regression.
Ex- Salary of employee based on there position.

c. Support vector regression (SVR):

- Instant of having simple line as a linear regression, SVR contains a tube, having width w+ , w-
- Any data points fall inside the tube will discard the error.

i.e A margin of error we are allowing to our model to have and not care about error inside the tube.

d. Decision Tree:

- Generally we have many data points available in a plane. So by applying ML algorithm, we used to consider all the data points available in the plane, then we used to find their features.
- However using DT we split the data point into many subset and find the features of every segment of data so that we can get better accuracy.
- In DT we need not to apply feature scaling and also we need not to split the dataset into training & Testing
- DT model is clearly not the best model to use on single features dataset.
- It is more adapted to the dataset with many features or high dimensional dataset.

e. Random Forest

- It is based on concepts of ensemble learning. that follow the bagging technique, which is a process of combining multiple classifier to solve a complex problem and improve the performance of the model.
- The base estimator in the random forest are decision tree.
- Instead of depending on one decision tree, RF takes the prediction from each tree and based on majority votes of prediction, it predict the final output.

  i.e Higher DT --> Higher accuracy and prevent from overfitting

Classification:

- It is used to predict the classes of input dataset / category.

a. Logistic regression:

- Logistic regression is a linear classification model, which simply convert prediction into probability using Sigmoid function (that ranges from 0 to 1)
- Sigmoid function is used for binary classification.

b. K- Nearest Neighbour (KNN):

- KNN is one of the simplest ML algorithm based on supervised learning technique.
- If we provide a new data point to the ML model then with the help of KNN we can collect K nearest data point w.r.to new data point and assign the categories to the new data points based on majority of nearest data point categories.

c. Support vector machine (SVM)

d. Decision Tree

# Entropy, Information Gain and GINI entropy

e. Random Forest:

- Random forest reduce the high variance or RF minimize the variance of model.
- In RF, DT are build using bagging technique
- Data are randomly chosen and build a separate Decision tree.

How Random forest model works

- Data are randomly chosen and build a separate decision tree.
- At each node in the decision tree, only a random set of features are considered to decide the best split.
- A decision tree model is fitted on each of the subset.
- Finally prediction is calculated by averaging the predictions from all decision trees.

Key concepts:

❖ Activation Function

- AF is a non-linear transformation. It basically decide whether a neuron should be activated or not. (Neuron is activated means transferring the signal and helping us to classify the output)
  - a. **Sigmoid :** Used in o/p layer for Binary classification
  - b. **Softmax :** Used in o/p layer for Multi classification
  - c. **Relu :** Used in hidden layers (Used for regression)

❖ Loss function / Cost function

For Classification:

a. Binary cross entropy
b. Categorical cross entropy
c. Sparse categorical cross entropy

For Regression:

a. Mean absolute error (MAE)
b. Mean square error (MSE)
c. Root mean square error (RMSE)

❖ Underfitting / Overfitting

- **Underfitting :** if we use very simple model then it is not able to recognize complex features that are present in the dataset.
- To overcome with underfitting we can use complex model (By adding more number of hidden layers and more number of neurons) **Or** we can overcome by training the model longer

- **Overfitting :** Overfitting most likely occur if we have very small data set
- If we have small data set, then our model quickly recognized features & pattern in the dataset can highly overfitted (i.e model gives high accuracy in training dataset, but low accuracy in test dataset)

- **How to overcome with overfitting:**

1. We can increase the size of training dataset

- But if we have limited amount of data / data is too expensive then **how will increase the training dataset**.
- By using **data augmentation** (i.e Increase the training dataset by rotating, cropping in the case of Image dataset)

2. With the help of **Regularization** , Here we need not to increase training dataset to overcome with overfitting.

Regularization :

- Regularization is a technique which is used for smoothing the complicated curve to fit well in both training and testing.
- There are mainly two type of Regularization :
  a. **L2 regularization and**
  b. **Dropout regularization** : (Most commonly used technique)
- This two technique perform well to overcome with overfitting problem and to achieve better accuracy in both training as well as testing.

Bias/ Variance

Bias: Bias in data tell us about the inconsistency in data.

- Bias is an error due to over simplified assumption in learning algorithm.
- It can lead the model to underfitting.

Bias = expected or avg prediction of model - Correct value which we are trying to predict

Variance: Variance is an error due to too much complexity in the learning algorithm

- It can lead the model to overfitting

Hypothesis test :

P-value : P value is used to make a decision about a hypothesis test.

- P value is the minimum significant level at which we can reject the null hypothesis.
- Lower the P value --> more likely to reject the null hypothesis.

Outliers detection technique:

- Outliers are data points that doesn't belongs to a certain population. **or**
- Outliers is an abnormal observation that lies far away from other values.
- Following technique are used to detect the outliers:

a. Standard deviation:

- If we have any data points that are more than 3 times the standard deviation, then those points are very likely to anomalous or outliers.

b. Box plot:

- Box plot are graphical representation of numerical data through their quantiles.
- It is very simple but effective way to visualize outliers.
- Any data points that show above or below the whiskers can be considered outliers or anomalous.

c. Scatter plot :

- Scatter plot is a plot or mathematical diagram using Cartesian co-ordinate to display values for typically two variables for a set of data.
- The point which are very far away from the general spread of data and have a very few neighbours are consider are consider as outliers.

Following are the approach to handle the outliers:

- Drop the outliers
- Assign new value (mean ,mode, median)
- If % wise number of outliers is less, but when we see the numbers that are several, then in that case dropping them might cause a big loss. In that case we should group them and run our analysis separately on them.

❖ Encoding Technique

- In many practical data activities, the data set contain categorical variables. These variables are typically store as text values.
- Since ML is based on mathematical equation, it would caused a problem when we keep categorical variables. That's why we need to encode them
- Following are the common encoding technique

a. Label encoder:

- In label encoding we map each category to a number or label.
- Mostly for binary categorical data we use label encoding.

b. One-Hot encoder:

- In this method we map each category to a vector that contains 1 and 0 denoting the presence of feature or not.

❖ Feature scaling

a. Standardisation

b. Normalisation

❖ Optimizer Function

- It is used to reduce the loss value. In order to reduce the loss value, we basically have to use back propagation algorithm, that means we basically have to update the waits.
- Once we get the updated waits, again forward propagation will go ahead.

a. Gradient Descent (GD) / Stochastic Gradient Descent (SGD) / Mini-Batch SGD

- **Gradient Descent:** it takes all the records and then doing the forward propagation, computes the loss then do the backward propagation and finally update the weight. (But required huge resources)
- **SGD**: it does the training w.r.to 1 record at a time, do the forward propagation and backward propagation. (But training is very much slow in SGD)
- **Mini-Batch SGD**: Here we define some batch size based on that we do the forward propagation and backward propagation. (But it has noise due to that convergence will take time)
- **Exponentially moving weight average**: Avg is calculated at each step as we encounter new points and it calculated in such a way that we gave higher weight to new data points, lower weight to older data points.

b. SGD with momentum:

- With the help of SGD with momentum, noisy data in SGD is reduced
- By applying the moving weighted average we basically **smoothing the curve**

c. Adaptive Gradient (AdaGrad)

- We change the learning rate as number of epoch increased.

d. RMS propagation / Ada delta

- Changing the learning rate in efficient manner, such that α(t) doesn't goes high

e. Adam optimizer

- It is combination of SGD with momentum and RMS prop

❖ Ensemble method

- Decision tree have been around for a long time and also known to suffer from bias and variance.
- Simple tree --> Large Bias , Complex tree --> Large Variance
- Ensemble learning: Which combines several decision trees to produce better predictive performance than utilizing a single decision tree.
- The main principle behind the ensemble model is that a group of weak learners comes together to form a strong learner.
- There are two technique to perform ensemble decision tree:

a. Bagging

- Bagging is used when our goal is to reduced the variance of a decision tree.
- Here idea is to create several subset of data from the training sample chosen randomly.
- Now each collection of subset data is used to train their decision trees.
- As a result we end up with an ensemble of different models.
- Average of all the predictions from different trees are used which is more robust than single decision tree.
- Most widely algorithm used for bagging technique is **Random Forest**

b. Boosting

- Boosting is another ensemble technique to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple model to the data and then analysing the data for error. **Or**
- We fit consecutive trees (random sample), and at every step the goal is to solve for net error from the prior tree.
- Following are the different type of boosting algorithm:

a. Adaptive Boost (Ada Boost):

- Ada boost is a sequential learning process,one trees is dependent on previous trees.
- Initially, initial weight is assigned to all the record
- All the record are equally important for the model.
- In the **next iteration**: if any record is miss classified by the previous model then weight for that record updated with higher weight and normalized the remaining weight with lesser weight.
- In Ada boost technique more importance is given to the previous miss classified record.
- That's why it's name is Adaptive boosting, because it adopt the previous model.

b. Gradient Boost

- Learning happen by optimizing the loss function
- It iteratively corrects the mistakes of the weak classifier and improves accuracy by combining weak learners and gives the better accuracy in most of the case.
- But it has space and time complexity

c. Extreme Gradient Boost (XG Boost)

- XG boost is one of the most popular algorithm for the data analysis.
- Speed and Performance of XG boost is quite better than other algorithm.
- Speed is high due to Parallelization, Cache optimization.
- Performance is better due to Regularization.
- XG boost also take care of missing value treatment.
- New prediction= Previous prediction + learning rate * output

❖ K- fold cross validation

❖ Performance Metrics

For Classification:

a. Confusion Matrix

- Confusion matrix is a table which is used for summarizing the performance of a classification algorithm.

b. Recall / Precision / F1 Score

Recall =Correctly classified positive examples/Total number of positive examples= TP/(TP+TN)
Precision=Correctly classified positive /Total number of predicted positive = TP/(TP+FP)

F1 Score: Which uses harmonic mean in place of arithmetic mean as it punished the extreme values more

- F1 score will always be nearer to the smaller value of Precision or Recall.
- F1 score = (2* Recall * Precision)/ (Recall + Precision)

For Regression:

a. R square (R2) , Adjusted R2

- R2 is a statical measure of how close the data are fitted regression line.
- It is also known as the coefficient of determination or the coefficient of multiple determination for multiple regression.

b. Mean square error (MSE) , RMSE, Mean absolute error (MAE)

Difference between Linear regression and Logistic regression

Linear Regression	Logistic Regression
1. Linear regression is used when dependent variable is continuous and the nature of the regression line is linear	Logistic regression is used when the dependent variable is binary
2. Required to establish the linear relationship among dependent and independent variables.	It is not necessary for Logistic regression
3. Independent variable can be correlated with each other	Variable must not be correlated with each other

2. Recurrent Neural Network (RNN) : For Time series data analysis

- RNN suitable where data in a particular time series is important.
- **Main Goal :** Predicting the future (Forecasting) and assign the categories

Algorithm:

a. Long-Sort term memory (LSTM)

Continue……