1.Difference between decision tree and random forest.

The basic idea behind a decision tree is to build a “tree” using a set of predictor variables that predicts the value of some response variable using decision rules.

An extension of the decision tree is a model known as a random forest, which is essentially a collection of decision trees.

Here are the steps we use to build a random forest model:

1. Take bootstrapped samples from the original dataset.

2. For each bootstrapped sample, build a decision tree using a random subset of the predictor variables.

3. Average the predictions of each tree to come up with a final model.

2.List down the advantages and dis-advantages of random forest.

Advantage:

1. Random Forest algorithm is less prone to overfitting than Decision Tree and other algorithms
2. Random Forest algorithm outputs the importance of features which is a very useful

Disadvantage

1. Random Forest algorithm may change considerably by a small change in the data.
2. Random Forest algorithm computations may go far more complex compared to other algorithms.

3.What are the other parameters to assess the classifier.

ROC AUC

F-Beta Score

4. Perform EDA & all different classifier/regression for the case study (given in the session) and evaluate it based on the performance

EDA in Python uses data visualization to draw meaningful patterns and insights. It also involves the preparation of data sets for analysis by removing irregularities in the data.

Based on the results of EDA, companies also make business decisions, which can have repercussions later.

- If EDA is not done properly then it can hamper the further steps in the machine learning model building process.
- If done well, it may improve the efficacy of everything we do next.

In this article we'll see about the following topics:

1. Data Sourcing
2. Data Cleaning
3. Univariate analysis
4. Bivariate analysis
5. Multivariate analysis

1. Data Sourcing

Data Sourcing is the process of finding and loading the data into our system. Broadly there are two ways in which we can find data.

1. Private Data
2. Public Data

Private Data

As the name suggests, private data is given by private organizations. There are some security and privacy concerns attached to it. This type of data is used for mainly organizations internal analysis.

Public Data

This type of Data is available to everyone. We can find this in government websites and public organizations etc. Anyone can access this data, we do not need any special permissions or approval.

We can get public data on the following sites.

- https://data.gov
- https://data.gov.uk
- https://data.gov.in
- https://www.kaggle.com/
- https://archive.ics.uci.edu/ml/index.php
- https://github.com/awesomedata/awesome-public-datasets

The very first step of EDA is Data Sourcing, we have seen how we can access data and load into our system. Now, the next step is how to clean the data.

2. Data Cleaning

After completing the Data Sourcing, the next step in the process of EDA is **Data Cleaning**. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

Irregularities are of different types of data.

- Missing Values
- Incorrect Format
- Incorrect Headers
- Anomalies/Outliers