TASK 2

1) According to you, why do overfitting and underfitting occur, and how resolve them? What is the difference between them?

A machine learning algorithm is said to be underfitting when it cannot capture the trend of the data. It performs well on training data but poorly on the testing data. It reduces the accuracy of the machine learning model. It happens when the training data is small or insufficient. It can be reduced by using more training data and reducing the noise in the input training data. Increasing the number of epochs, number of features also helps in getting better results.

A machine learning algorithm is said to be overfitting when the model cannot make proper predictions, accurate predictions on the testing data. When the training data is huge, it learns from the noise also. This results in the overfitting of data. By reducing the model complexity, we can reduce overfitting

### 2) What kind of pattern did you analyze in the Train and Test score while running the code of overfitting?

We can see in the observation that the when the number of tree\_depths is 30, the testing accuracy is increased from 76.7 to 90.8

When the number of tree\_depths is increased to 50, the testing accuracy is increased from 76.7 to 92.1

It can be observed that as the number of tree\_depths is increased, accuracy is being increased which can be a good or a bad sign depending on the application we are using the model for.

### 3) What is cross-validation, and what did you analyze in a different type of validation that you performed?

Cross-Validation gives the ability to estimate performance of a model on unseen data which is not used during the training step.

Other than K-fold, Repeated K-Fold and Leave One Out Validations I have tried Shuffle and Split, Leave One Group First, Leave P Groups Out Cross Validations.

K-Fold : Input samples are divided into k groups,. 1 group is left for validation. Here we have done 2 and 4 fold cross validation on a dataset with 4 samples.

Repeated K-Fold : We execute the k-fold n times instead of n-1.

Leave One Out : We have n training and testing sets. N-1 samples are trained.

Shuffle and Split: It returns stratified splits. Preserves the same percentage for each target class as complete set.

Leave One Group First: holds samples according to third part provided integer array groups.

Leave P Groups Out: similar to leave one group. But removes samples related to P groups. All combinations of p groups are left out.

### 4) Explain the analysis from generated ROC and validation curve and what they represent?

ROC is known as Receiver Operating Characteristic curve

In the code we are testing it against the true fpositive rate and false positive rate at various threshold levels. WHerther true positivity rate can represent factors like sensitivity and probability.False positive rate is to estimate the probability of errors.

Validation curve: is generally used to represent the performance of a machine learning model.