LAB 5-Report

Name:Devyani Gorkar
Roll no.:B20ME027

Task1:

Main data frame is df.X is the dataframe of all columns except drugs.Using OneHotencoder all nominal features are encoded and using ordinal encoder all ordinal features are encoded.Y is multidimensional column consisting of Target values.
A function called ‘decision_tree’ is created which takes 5 inputs X_train,y_train,X_test,y_testand criterion,where criterion is used to measure the quality of split.
When the ratio is 70:30-

- The criterion used to calculate the quality of the split is Gini index.
- On calling the function’decision_tree’ and on training the classifier using X_train we can get the model accuracy on the training set and predict the values using X_test.
- The model accuracy obtained after training the dataset is 1.This is because we trained the model using X_train and y_train. The decision tree had gained knowledge about that data and therefore on giving the same data it gives correct results each time.
- Since the training data gives zero error so it is possible that the model overfits training data .But since the maximum depth of the tree is equal to 4 which is not a large value and also the accuracy of test data is good, therefore we can conclude that the model does not overfit the data.
- After calculating the confusion matrix of y_test we can see that almost all non diagonal elements are zero which means that there are very less incorrect predictions therefore the model is very accurate on test data.

When the ratio is 80:20 and 90:10-

- In this case we use entropy/gini as our criterion to calculate the quality of the split.
- On calling the function’decision_tree’ and on training the classifier using X_train2/X_train3 we can get the model accuracy on the training set and predict the values using X_test2/X_train3.
- On calculating model accuracy we get it as 1.This is because we trained the data using X_train2/X_train3,y_train2/y_train3 ,decision tree and gained knowledge about that data and therefore on giving the same data it gives correct results each time.
- The training data gives zero error so it is possible that the model overfits training data .But since the maximum depth of the tree is equal to 4 which is not a large value and also the accuracy of test data is good, therefore we can conclude that the model does not overfit the data
- The test accuracy in this case is 1 because we are using maximum data to train the model and our test data is very less.Also we may say that the complexity of the model is very less hence it becomes easier to give precise and accurate results.

Task 2:

Main data frame is df.X is the dataframe of all columns except Compressive strength,which is target value of the dataset.There are no columns with null or missing values.

Regression using MSE:

- In this case we use Mean squared error as our cost function which is to be minimized by decision tree regressor.
- On predicting the values for X_test after training the model we get the test accuracy as 0.9953.
- The max depth of the decision tree is 19 which indicates that the model is complex but still gives good results on the test data.

Regression using MAE:

- In this case we use Mean absolute error as our cost function.
- On predicting the values for X_test after training the model we get the test accuracy as 0.92387 which is less than the earlier one.This shows that when we use MSE as our cost function we get better accuracy because MAE consists of modulus which does not have differentiability.