**Readme**

**Files Used:-**

1. Q1.py
2. Q2.py
3. Q3.py

There are three questions in the assignment, so there are three code files, one for each question.

**About q1.py :**

1. The file q1.py contains the four functions which are as follows :

* Q1\_a() - For solving 1(a) part.
* Q1\_b() - For solving 1(b) part.
* Q1\_c() - For solving 1(c) part.
* Q1\_d() - For solving 1(d) part.

1. Q1\_a() – This function loads the dataset\_1.mat using the loadmat() and then plot the 10 samples for each class in the dataset\_1
2. Q2\_b() – This function loads the dataset\_2.mat and then creates the scatter plot of the data using the scatterplot() function of the seaborn library.
3. Q2\_c() – This function loads the dataset\_1.mat.The loaded dataset is then reduced in the 2 Dimensional Space using the tSNE algorithm. Finally, the scatter plot is created for reduced dimensions.
4. Q2\_d() - This function loads the dataset\_1.mat.The loaded dataset is then reduced in the 3 Dimensional Space using the tSNE algorithm. Finally, the scatter plot is created for reduced dimensions.

**About q2.py :**

1. The file q2.py contains the following functions which are as follows :

* Load\_Dataset()
* Spilt\_DataSet(Data,r)
* accuracy(actual,predicted)
* OptimalDepth(Train,Test,Hyper)
* TrainingVsValidation(Train,Test,Hyper)
* TrainVsValid(Train,Test,Hyper)

1. Load\_Dataset() – This function loads the dataset\_2.It transforms into a data frame and finally retuns the data frame.
2. Spilt\_DataSet(Data,r) – This function receives two arguments :

* Data: Dataframe
* r – split size

The Data is splitting in the two data frames as Train and Test with the size ration (1-r):r.

The splitting of the Data takes place randomly using the sample() function of the pandas library.

These data frames are returned by the method.

1. accuracy(actual, predicted): This function receives the two lists named actual and predicted. It returned the scalar value based on the corresponding matching of values in the two lists. This scalar value is the accuracy value.
2. OptimalDepth(Train, Test, Hyper): This function takes the three parameters as Train, Test, and the list of Maximum Depth Values for the Decision Tree and find the optimal depth for which testing accuracy is high. For accuracy calculation, it uses the above-defined accuracy() function.
3. TrainingVsValidation(Train, Test, Hyper): This function plots the Train/Testing accuracy curve with the accuracy calculated from the above-defined accuracy() function.
4. TrainVsValid(Train, Test, Hyper): This function plots the Train/Testing accuracy curve with the accuracy calculated from the sklearn accuracy function.

**About q3.py :**

1. The file q3.py contains the following functions which are as follows :

* Spilt\_DataSet(Data,r)
* accuracy(actual,predicted)
* Q3\_a()
* Q3\_b()
* GetData()
* Q3\_c(Train,Test,Stumps,Max)
* Q3\_d()

1. Spilt\_DataSet(Data,r) , accuracy(actual,predicted) have the same functionality as above.
2. GetData(): Load the CSV data and then split the data into Train and Test and return them.
3. Q3\_a(): This function is used to find the best criteria for the decision tree among Gini and entropy-based on testing accuracy on the Testing Data. The accuracy is calculated from the above-defined accuracy function.
4. Q3\_b(): This function is used to find the optimal max depth parameter for the decision tree and plot the training and testing curve.
5. Q3\_c(Train, Test, Stumps, Max): This function takes the training data, testing data, Number of stumps, and Max(maximum depth). Here , the number of stumps = 100 and Max = 3, it training the 100 decision trees each of depth 3, then use the maximum voting rule to predict the class labels of test data instances.
6. Q4\_d(): This function is used to find the optimal Number of Stumps and Max Depth using the Grid Search and find the maximum testing accuracy.