DATA MINING

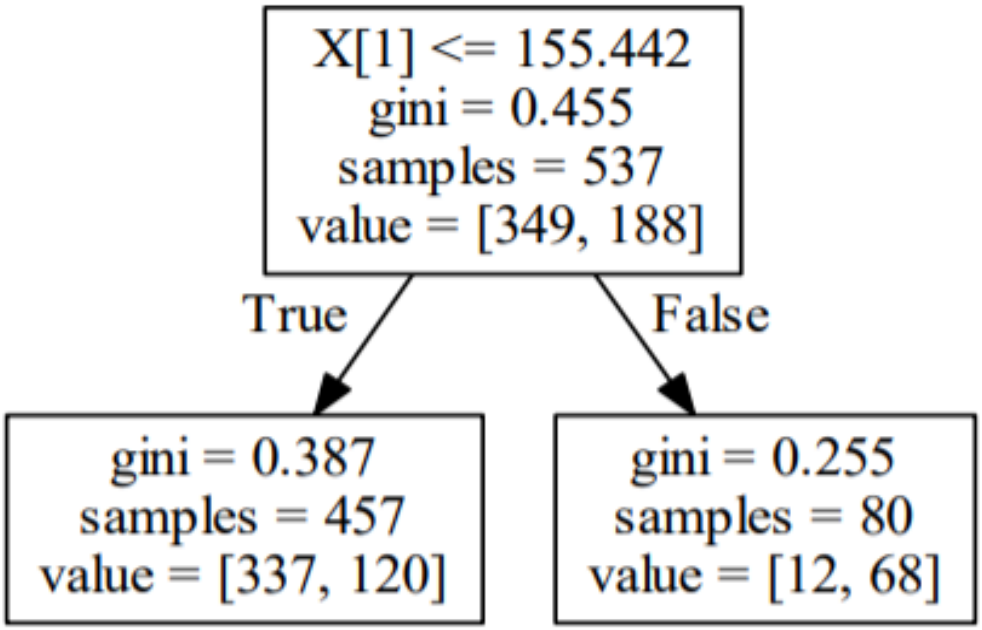
MINI PROJECT 2

Members:

Harshil Shah	1213348082
Jay Bhanushali	1213436781
Ronak Vikam	1213203327
Viren Bhanushali	1213232850

Q.1) **Decision Tree**

a) A decision Tree model was built for given dataset. We have used the Grid Search Method Was used to find out the optimal values of the parameters and then use that model to predict. The train – split used for the model building process was 70-30.



b) Identify the model parameters for your selected model

The model parameters that are considered are in this problem are:

i)	Criterion: The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain.
ii)	splitter: The strategy used to choose the split at each node. Supported strategies are “best” to choose the best split and “random” to choose the best random split.
iii)	max_depth: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
iv)	min_samples_split: The minimum number of samples required to split an internal node: If into, then consider min_samples_split as the minimum number. If float, then min_samples_split is a fraction and ceil (min_samples_split * n_samples) are the minimum number of samples for each split.

c) Provide an accuracy evaluation of your model and describe clearly the method you used to evaluate the accuracy.

Best Parameters are {'criterion': 'gini', 'max_depth': 1, 'min_samples_split': 2, 'splitter': 'random'}
Testing Accuracy is 0.718614718615

As it can be seen we have grid search method to evaluate the accuracy, what it basically does is calculates the accuracy considering all the possible values of various parameters used and gives us the model and parameters with the best accuracy.
The testing: training data is taken as 70:30 and '. score' method is used to calculate the accuracy over the testing data.
It can be inferred that the best model after building a decision tree classifier is able to have an accuracy of 0.718 which means that it positively classifies 71.86% of the data correctly.

It can also be seen from the confusion matrix that type 0 is being classified more accurately than type 1 hence to improve the model we might have to concentrate on type 0.

Q.2) **Support Vector Machine**

a) Provide a graphical display of your final tree
Ans) Since it is a SVM model there is no decision tree to display graphically.

b) Identify the model parameters for your selected model
Ans) The parameters used to control the complexity and accuracy of the model are given below:

i) C: Penalty parameter C of the error term.

ii) Gamma**:** Kernel coefficient for 'rbf', 'poly' and 'sigmoid'.

iii) Kernel**:** Specifies the kernel type to be used in the algorithm. It must be one of 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed' or a callable. If none is given, 'rbf' will be used. If a callable is given it is used to pre-compute the kernel matrix from data matrices; that matrix should be an array of shape.

c) Provide an accuracy evaluation of your model and describe clearly the method you used to evaluate the accuracy.
Ans)

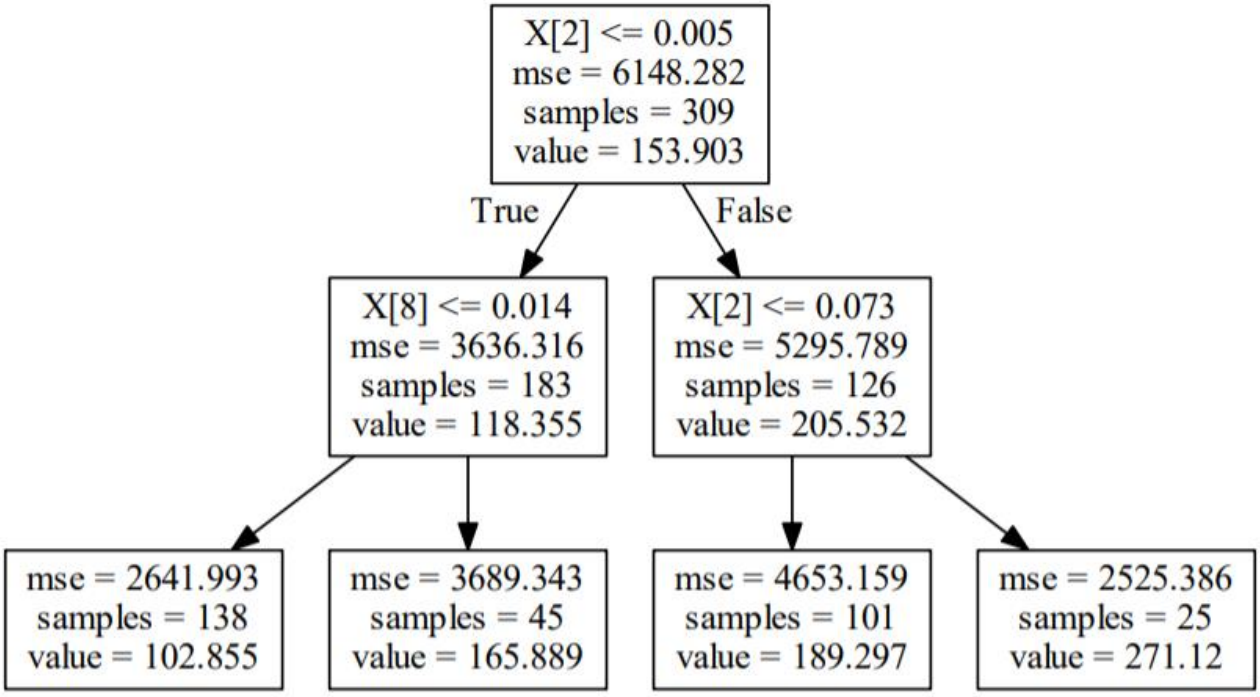
Best Parameters are {'C': 10000.0, 'gamma': 1.0000000000000001e-05, 'kernel': 'linear'} Testing Accuracy is 0.705627705628
As we can see the best values of the parameters that we can get is shown above. It Is evident that the testing accuracy that we got is 0.7056 which says that 70.56 % of the data is classified accurately.

The method that we used to evaluate the accuracy was the grid search method, what is basically does is calculates the accuracy considering all the possible values of various parameters used and gives us the model and parameters with the best accuracy. The testing: training data is taken as
70:30 and. score method is used to calculate the accuracy over the testing data.

Q.3) **Decision Regression Tree Model**

a) Provide a graphical display of your final tree

A decision Tree Regression model was built for given dataset. We have used the Grid Search Method Was used to find out the optimal values of the parameters and then use that model to predict. The train – split used for the model building process was 70-30.



b) Identify the model parameters for your selected model.

Ans) Following parameters are used to control the complexity and accuracy of the model.

i)	criterion: The function to measure the quality of a split. Supported criteria are “friedman_mse” for the mean squared error with improvement score by Friedman, “mse” for mean squared error, and “mae” for the mean absolute error. The default value of “friedman_mse” is generally the best as it can provide a better approximation in some cases.
ii)	Max_depth: maximum depth of the individual regression estimators. The maximum depth limits the number of nodes in the tree. Tune this parameter for best performance; the best value depends on the interaction of the input variables.
iii)	Min_samples_split: The minimum number of samples required to split an internal node
iv)	splitter: The strategy used to choose the split at each node. Supported strategies are “best” to choose the best split and “random” to choose the best random split.

c)Provide an accuracy evaluation of your model and describe clearly the method you used to evaluate the accuracy.

Ans)

Best Parameters are {'criterion': 'mse', 'max_depth': 2, 'min_samples_split': 2, 'splitter': 'best'}
Testing Accuracy is 0.355419437342.

we have grid search method to evaluate the accuracy, what it basically does is calculates the accuracy considering all the possible values of various parameters used and gives us the model and parameters with the best accuracy. The testing: training data is taken as 70:30 and. score method is used to calculate the accuracy over the testing data.