PATTERN RECOGNITION AND MACHINE LEARNING

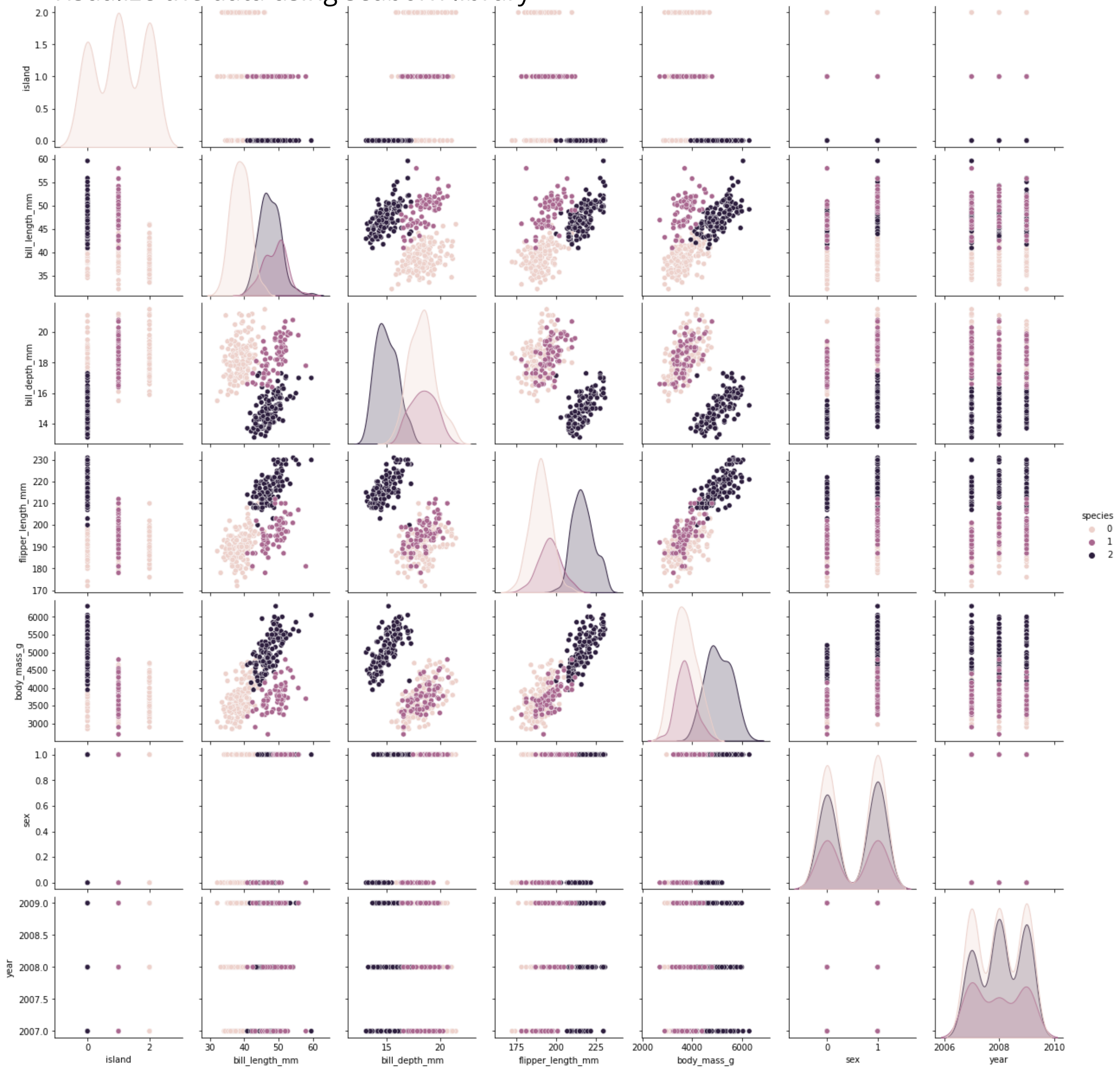
LAB 02

LAB REPORT

QUESTION 1.

1)

- Preprocess the data by replacing NAN values in int/float columns by respective column means
- Visualize the data using seaborn library



2)

- Implementing Gini index as the cost function.
- Gini index of an array arr = 1 - Summation(prob(arr[i])^2)

3)

- Use label encoder from sklearn to encode respective categorical features.
- Spit the data in x_train, x_test, y_train, y_test using train_test_split from model_selection in sklearn.
- Get an optimum threshold for each feature in x_train and edit each feature to contain just 2 values: 0 and 1 according to the optimal threshold found.

4) AND 5)

- Get the best attribute to split on with respect to the gini index of all features.
- This will happen after finding the gini index of all features.
- Create 2 classes for representing Leaf and DecisionNode nodes.
- Design them with appropriate constructors and attributes.
- Build the actual tree with respect to the x_train dataframe using a recursive BuildTree function.
- Print the tree with yet another recursive function just to have a look at the tree.

6)

- Design a function Classify to classify a given input to predict its output.
- This will work recursively by comparing input feature data to threshold feature data.

7)

- Classify each of the rows in x_test and compare the prediction with y_test to find its accuracy.
- The accuracy turns out to be 94.23 percent which is fascinating since only 2 children are being made of each node and all the variety in all features is reduced to just 2.

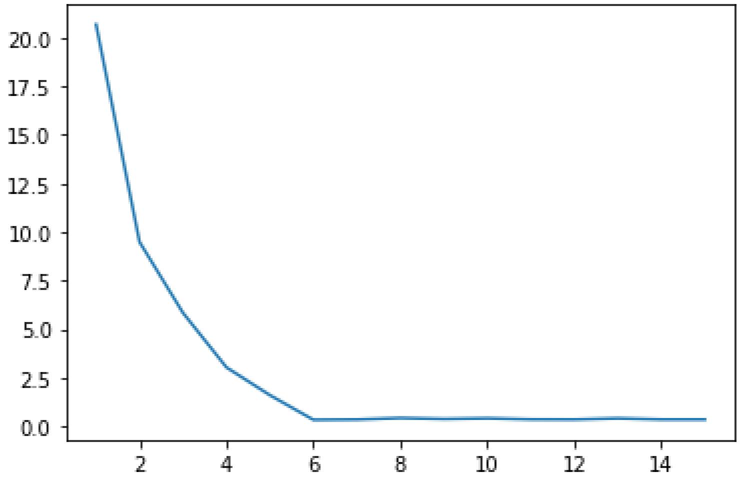
QUESTION 2.

1)

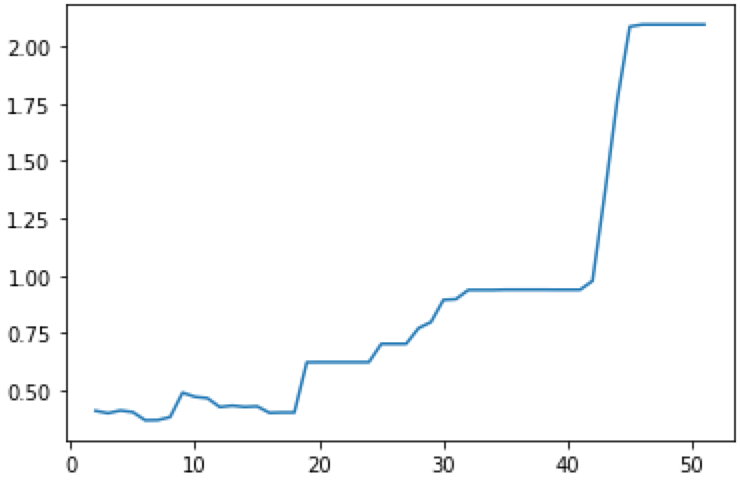
- Preprocess the data to replace all NAN values with the column mean values
- Split the data into x_train, x_test, x_validation, y_train, y_test, y_validation using train_test_split from model_selection in sklearn.

2)

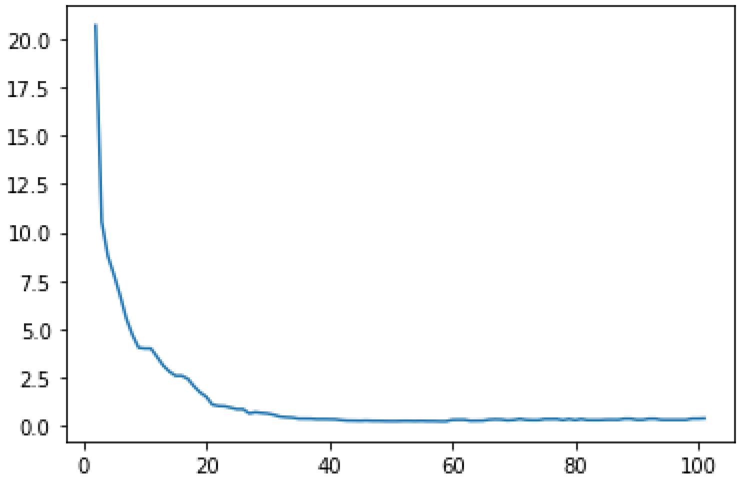
- Find appropriate values for hyperparameters to best train the data.
- This is done by calculating MSE between predictions made by the model and given validation data.
- GRAPHS: MSE vs HYPERPARAMETER VALUE
- Max Depth, min MSE at max_depth = 6



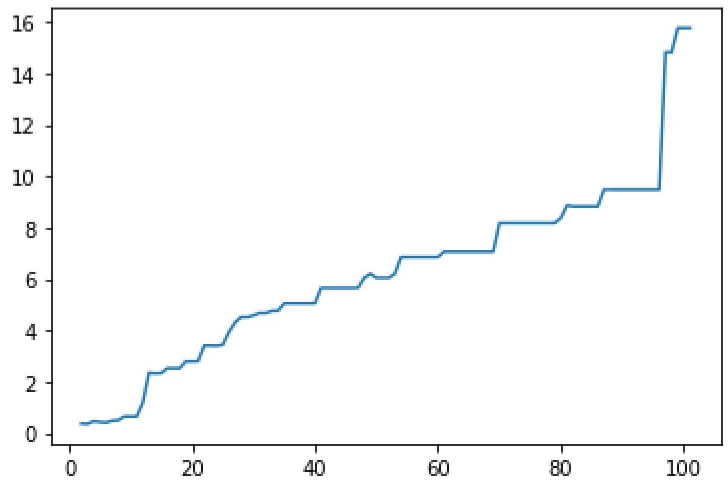
- Minimum Samples Split, min MSE at 6



- Max Leaf Node, min MSE at 58



- Min Samples Leaf, min MSE at 3



- Hence the optimal values for all the hyper parameters are obtained.

3)

- Train the model using k fold from sklearn.model_selection taking all the optimal hyperparameters into consideration.
- The accuracy turns out to be 96.43 percent.
- Plot the decision tree.