LAB Assignment 2
CSL 2050

By :- Akshat Jain B21CS005

1. In this we Performed regression using a Decision tree.
   1.1. Since the data doesn't have any missing values we just split into training,validation and test data with required 70:20:10 ratio.
   1.2. Here we have ran the code for each hyperparameter varying them iteratively and then Checked the best values for those hyperparameters by minimizing the mean squared error. We vary parameters like random_state, min_samples_leaf, max_depth, min_samples_split, max_leaf_nodes and find from the MSE plots that varying each of these has different effects on the value of MSE. Especially in the case of max_leaf_nodes and max_depth we see that the MSE reaches its lowest point and then gently increases and Plateus
   1.3. I have Done the stated questions and plotted the decision tree for viewing with the optimal hyperparameters found in above Question.
   1.4. Here we see that Absolute_error or the L1 Criterion works better Since it does not take into account the Blowing up of error rate due to squaring it.
2. Classification
   2.1. Here we have Made and plotted the decision tree. In the plot it shows that what split was made at which depth. We have also plotted the Decision Boundary in this case.
   2.2. Here we have removed the Specified data points and then constructed the Decsion tree. We have also plotted the decision boundary in this case.
   2.3. Here I have removed the constraints of Max_depth. After this we find that as opposed to the simple decision boundary in question 2.1 we have a more complicated and fitted Decision

Boundary. In the First Case there are some points classified incorrectly. But in this case the no. of these points is much lower.

2.4. Here I had first used numpy.random to create an array of values between 0 and 1 after this I have scaled them up for 0 to . I have then taken these values and created the X1 and X2 features of the dataset. The Decision Boundary for this is a simple line at x1 = 2.5. After rotating them Clockwise 45 degree we Find that the new decision boundary is very different. It also becomes a little rotated and Warped.

2.5. Decision Tree Classifiers Have a habit of Making Decision Boundaries which are not smooth Curves but are instead Clunky looking. We also see that changing the MAx_depth has a great impact on the Bias and Variance of a decision tree.

2.6. Here we have plotted the predictions against the actual Values at each Max_depth. We find that as we increase the max_depth the closer it reaches to the actual Predictions. In this case the black line is the actual prediction.

2.7. In this Case we observe that both the cases have different plots to each other however we see that the plot for Min_samples_leaf = 1 has a very close shape to the actual results as compared to the min_samples_leaf = 10

3. Gini Index is the cost Function.

3.1. We have performed the above tasks and plotted the graphs of species vs. sex, species vs island, species vs year etc. in order to visualise the dataset

3.2. I have implemented The Cost function as gini value or gini index. Here we see the gini indexes of various categorical columns.

3.3. I have then implemented the Cont_to_categorical Function which iteratively looks for the best split on the basis of gini indexes and then gives us the best values for split.

3.4. For questions 3.4,3.5,3.6 we have made a class called DecisionTreeClassifier to complete these tasks here we have different Functions like best split, Fit, Predict, Best_gini etc.

These helper functions along with our build tree function help us to create this decision tree classifier.

3.5. See above

3.6. See Above

3.7. We then create a Function to see What our Accuracy is in this case and whether we get the correct classifications for our testing database