Machine Learning Assignment 1

Grp no: 35
Grp members:
Roopak Priydarshi (20CS30042)
Saras Umakant Pantulwar (20CS30046)

Question 1: Decision Tree

1. In this we first did the data analysis of the given data and found out following:
    a. There were no NULL entries
    b. The data was a mixture of continuous and discrete features
    c. The result was a discrete feature (“Response”)
    d. The shape of the dataset was (131689, 12)

2. We then encoded the discrete values i.e Vehicle_Age, Gender, Vehicle_Damage
3. We used a Node class to store the nodes of the tree
4. To calculate entropy we have used calculate_entropy using the formula:
    entropy = -p1 * np.log2(p1) - (1 - p1) * np.log2(1 - p1)
5. Then we calculated information_gain using the formula:
    Information_gain = entropy of parent node - entropy of current node

    Also, we used two different functions for calculating information gain for continuous and discrete features according to the conditions required
6. So to build the tree we call build_tree with required parameter which in turn calls recursive function build_node. So by this, we get a list of nodes in the tree and as the node class has listed as an attribute in it which has nodes (list of children) as its elements, so we can traverse the tree.
7. For traversing we call the recursive function traverse tree by sending the root node to it. For any non-leaf node, it checks the condition on input and then moves to the right child if the input is >= the condition, else if goes to the left node. Hence in this way, we can traverse to the bottom of the tree.

8. In the train_test_split function we split the dataset into 80% train, and 20% test parts, ten times randomly, and then we print max_accuracy and depth of the tree with max accuracy. And then we prune this tree and print it. Note that we have used functions like accuracy to calculate accuracy, print_tree to print_tree, and prune_tree to prune the tree.

Question 2: Bayesian (Naïve Bayes) Classifier

1. In this we first did the data analysis of the given data and found out following:

    e. There were no NULL entries
    f. The data was a mixture of continuous and discrete features
    g. The result was a discrete feature (“Response”)
    h. The shape of the dataset was (131689, 12)

2. As this dataset contains discrete and continuous values so we applied multinomial Naive Bayes by discretizing the continuous values in bins using the inbuild pandas function.
3. Three bin values gave an accuracy of 77.3%, but binary binning (2 bins) gave an accuracy of 83%.
4. Observation: Almost 87% of the
 y values in the dataset are zero so simple probability-based classifier would have given better accuracy for random test sets.

```
Fold   1: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.226
Fold   2: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.583
Fold   3: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.119
Fold   4: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.135
Fold   5: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.097
Fold   6: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.742
Fold   7: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.324
Fold   8: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 84.015
Fold   9: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.226
```

```
Fold  10: [Training,Test] Split Distribution: [118521, 13168],
Accuracy: 83.278

10-fold-Cross-Validation accuracy: 83.374 +/- 0.291

Fold   1: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.643
Fold   2: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.226
Fold   3: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 82.945
Fold   4: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.530
Fold   5: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 82.778
Fold   6: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.788
Fold   7: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.651
Fold   8: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 84.023
Fold   9: [Training,Test] Split Distribution: [118520, 13169],
Accuracy: 83.598
Fold  10: [Training,Test] Split Distribution: [118521, 13168],
Accuracy: 82.655

10-fold-Cross-Validation accuracy: 83.384 +/- 0.435
```