**REPORT ON INTERNSHIP ASSIGNMENT**

**for**

**Summer Internship at IIT Bombay,2019**

**made by**

**Mitali Sheth**

**Chapter 1:EXPERIMENT**

The problem statement that I have chosen for the assignment is coding up a decision tree learning algorithm with the given specifications. I have tried to incorporate as much of the specifications as possible within the stipulated time. I have validated the algorithm with the Mushroom dataset as given in the specifications.

With the given details of the specifications, I have listed here the specifications that I have used.

1. IMPURITY MEASURE:
   1. entropy impurity
   2. gini impurity
2. STOPPING CRITERIA:
   1. check label: to check if only one label is present in the subtree
   2. Max\_depth: to check for maximum number of levels that can be created
3. BINARY SPLITS: Incorporated
4. EVALUATION:
   1. Using Confusion Matrix and F1 score
   2. Used Mushroom Dataset
5. MISSING ATTRIBUTE:
   1. Restored the missing values with the mode of the entire column(Implemented)
   2. Ignore and remove them from dataset
   3. Pass the rows to test case and check its behaviour
6. REPORTING NUMBERS:
   1. Calculated as much as possible within the stipulated time and tabulated them in results section as below.

Chapter 2: EVALUATION:

The evaluation for this algorithm has been done using both Titanic Dataset initially followed by Mushroom Dataset.  
Initially I had used the cross validation method for accuracy and later incorporated the confusion matrix.

Chapter 3:OBSERVATIONS:

|  |  |  |
| --- | --- | --- |
| **Condition** | **Accuracy** | **F1 Score** |
| train data for gini impurity | 1.0 | 1.0 |
| test data with gini impurity | 0.9519 | 0.9517 |
| training data with entropy impurity | 1.0 | 1.0 |
| test data with entropy impurity | 0.8801 | 0.8836 |
| train data for entropy impurity with check label stopping criteria only | 1.0 | 1.0 |
| test data for entropy impurity with check label stopping criteria only | 0.9369 | 0.9370 |
| train data for gini impurity with check label stopping criteria only | 1.0 | 1.0 |
| test data for gini impurity with check label stopping criteria only | 0.9556 | 0.9562 |