Report on Offline 1: Topic: Implementation And Performance Evaluation Of Decision Tree

Prepared By Md. Salman Estyak

Student No, 1105094

Performance measures Table:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Accuracy | TPR | TNR | PPV | NPV | FPR | FNR | FDR | F1 Score |
| 96.99 | 97.14 | 96.93 | 94.44 | 98.95 | 2.04 | 2.85 | 5.55 | 95.77 |
| 96.61 | 95.24 | 97.04 | 93.75 | 98.99 | 1.97 | 3.17 | 6.25 | 95.24 |
| 94.98 | 91.58 | 96.23 | 92.45 | 98.94 | 2.73 | 2.8 | 7.54 | 94.68 |
| 94.17 | 90.806 | 95.81 | 92.94 | 98.28 | 3.35 | 3.44 | 7.05 | 94.61 |
| 92.78 | 85.77 | 96.53 | 94.31 | 96.98 | 2.77 | 5.60 | 5.68 | 94.08 |

Table: 5 Set of Performance factors for 5 fold cross-validation

Average accuracy = (96.99+ 96.61+ 94.98 + 94.17 + 92.78)/5 %

= 95.106 %

Q1. *Why are you using cross validation? Do the dataset justify it?*

One of the main reasons for using 5 fold cross validation is to avoid overfitting of data. As we are testing and training the decision tree using different segments from the data, we can get a measure of its performance on average.

Another reason is to use more of data as test data set, when the dataset is not large enough to partition it into two suitable sets(training data, testing data), without losing significant modelling or testing capability.

In this assignment, we are given a dataset of size 670, however as each feature has 10 distinct values, with 8 such features, total number of unique instances becomes as large as 10^8. As we have a small data set, using cross validation is justified.

Q2. *Besides accuracy, which of the criteria mentioned above should be used in cross  
validation for the given data set? Explain.*

F1Score can be used as a performance measure as it is a tradeoff of two useful measures, i) recall and ii) precision.