**REPORT – ASSIGNMENT - 5**

Shown are the analysis made for every classifier predicting different data.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Number of Instances | Number of Attributes | Number of Cross Validation folds |
|  |  |  |  |
| processed.cleveland.data.csv | 303 | 14 | 10 |
| pima-indians-diabetes.data.csv | 768 | 9 | 10 |
| transfusion.data.csv | 748 | 5 | 10 |
| car.data.csv | 1728 | 7 | 10 |
| bupa.data.csv | 345 | 7 | 10 |

**Accuracies of classifiers**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | KNN | Random Forests | Bagging | AdaBoost | Gradient Boosting |
|  |  |  |  |  |  |
| processed.cleveland.data.csv | 57.94727 | 58.06748 | 56.52966 | 55.7281 | 59.05821 |
| pima-indians-diabetes.data.csv | 74.60868 | 77.61278 | 75.78606 | 73.30998 | 76.82502 |
| transfusion.data.csv | 77.94234 | 75.53514 | 74.18739 | 72.59459 | 78.48468 |
| car.data.csv | 93.74943 | 98.66714 | 98.84358 | 98.95786 | 97.91537 |
| bupa.data.csv | 63.15126 | 74.78151 | 69.2437 | 67.83193 | 72.18487 |

The datasets, R script has been zipped in this folder and they are used.

These can be inferred from the given analysis with the best classifier which can be used with it:

For, processed.cleveland.data.csv – **Gradient Boosting**

For, pima-indians-diabetes.data.csv – **Random Forests**

For, transfusion.data.csv – **KNN/Gradient Boosting**

For, car.data.csv – **AdaBoost**

For, bupa.data.csv – **Random Forests**

From the experimental results, we can see that the execution time for ada boosting is high since it involves complex operations, and the performance is good. Gradient boosting for the given datasets perform well.

Random forests performs better when the variance of the model is low.

Bagging performs well if we don’t have enough number of samples and when we are trying to reduce the bias.

When we already have a weak classifier in order to improve the performance of this classifier, boosting is preferred and it is usually effective with weak classifiers.

KNN performs well when the value of k is optimal i.e. when we keep increasing the k value, it begins to over fit at some point. One important thing to note is KNN is a lazy classifier.

**Concluding that, we have to choose a classifier by considering the dataset in which we are operating. Every classifier has its own advantages and disadvantages in which the estimate should be made based on the data samples we get.**