**REPORT – ASSIGNMENT - 3**

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

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Number of Instances | Number of Attributes | Percent Split |
|  |  |  |  |
| processed.cleveland.data.csv | 303 | 14 | 80/20 |
| pima-indians-diabetes.data.csv | 768 | 9 | 80/20 |
| transfusion.data.csv | 748 | 5 | 80/20 |
| car.data.csv | 1728 | 7 | 80/20 |
| bupa.data.csv | 345 | 7 | 80/20 |

**RUN – 1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Decision Tree | Support Vector Machines | Naïve Bayes | Neural Networks | Perceptron |
|  |  |  |  |  |  |
| processed.cleveland.data.csv | 62.29508 | 62.29508 | 57.37705 | 57.37705 | 62.29508 |
| pima-indians-diabetes.data.csv | 81.81818 | 65.58442 | 74.67532 | 68.18182 | 77.92208 |
| transfusion.data.csv | 81.33333 | 76.00000 | 81.33333 | 80.66667 | 80.66667 |
| car.data.csv | 93.93064 | 86.70520 | 86.99422 | 98.26590 | 93.35260 |
| bupa.data.csv | 72.46377 | 52.17391 | 57.97101 | 52.17391 | 68.11594 |

**RUN – 2**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Decision Tree | Support Vector Machines | Naïve Bayes | Neural Networks | Perceptron |
|  |  |  |  |  |  |
| processed.cleveland.data.csv | 60.65574 | 54.09836 | 52.45902 | 54.09836 | 54.09836 |
| pima-indians-diabetes.data.csv | 75.32468 | 65.58442 | 75.97403 | 77.92208 | 79.87013 |
| transfusion.data.csv | 75.33333 | 77.33333 | 76.66667 | 78.00000 | 78.00000 |
| car.data.csv | 94.21965 | 88.43931 | 83.81503 | 96.24277 | 91.61850 |
| bupa.data.csv | 63.76812 | 53.62319 | 57.97101 | 73.91304 | 52.17391 |

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 – **Decision Tree**

For, pima-indians-diabetes.data.csv - **Decision Tree**

For, transfusion.data.csv – **Naïve Bayes / Support vector Machines**

For, car.data.csv – **Neural Networks / Support Vector Machines**

For, bupa.data.csv – **Neural Networks**

For data which are linearly separable, Decision Trees or Neural Networks / Perceptrons will perform the best. For data which cannot be separated linearly, Neural Networks or Support Vector Machines will give us the maximum accuracy. In general Support Vector Machines is the best classifier, but it’s very complicated one. However, Neural Networks is also very efficient when they are dealing with complex separations. In the other hand, decision trees are easy to implement and debug. But they tend to over fit the data which might lead to decline in accuracy. Naïve Bayes is also a good classifier but it’s generally suited for smaller datasets. It doesn’t over fit data in any case.

For non-linear datasets, Perceptrons are very poor as there is no hidden layer. Neural Networks are better. However, as complexity increases we can use complex kernel functions in Support Vector Machines so that we can classify as much as possible.

**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. Few of them are listed in the above two paragraphs in which the estimate had been made based on the datasamples.**