ASSIGNMENT 3

In this assignment you will compare the performance of the classifiers that you have learned so far in the course.

Datasets:

UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html) is one of the most popular sources of datasets. You can filter them for classification tasks and also choose other characteristics such as attribute type and application areas e.g. Life Sciences, CS/Engineering, etc. For this assignment, you will choose any 5 datasets that are interesting to you. You obviously want to ensure that they are for classification tasks.

Here are some interesting datasets that you might want to consider (you are free to choose others)

- Census income dataset: http://archive.ics.uci.edu/ml/datasets/Adult
- Chemical origin of wines: http://archive.ics.uci.edu/ml/datasets/Wine
- Car evaluation dataset: http://archive.ics.uci.edu/ml/datasets/Car+Evaluation
- Heart disease dataset: http://archive.ics.uci.edu/ml/datasets/Heart+Disease
- Internet advertisements dataset: http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements

Classifiers to implement:

So far, we have studied the following classifiers:

- Decision Trees
- Perceptron (Single Linear Classifier)
- Neural Net
- Support Vector Machines
- Naïve Bayes Classifiers

You will test the accuracy of classification of the above classifiers. For this assignment, you are free to use any programming language (including R). Be sure to mention the details and also how to compile your code in the README file.

You can use any R package, but be sure to mention them in the README file and also have a statement at the top of your code, such as:

require(packagename)

Training and Testing Methodology

You will need to split the dataset into training and test partitions. You can use an 80/20 or 90/10 percent split for training and testing respectively. These splits should be performed using random sampling. R has a function called sample that you will find useful.

You will build your model using the training dataset and test it on the test part and report testing accuracy. For each dataset, you should run the classifier twice on different samples (the ratio of training to test can remain the same).

Comparing Classifier Performance:

The real aim of this assignment is to compare the performance of the classifiers in terms of their accuracy. So, for each sample you should run all of the classification algorithms and compare their accuracy.

The pseudo-code for the whole question can be summarized as:

```
D = List of datasets
A = List of algorithms

for d in  datasets D
        for numSamples in 1:2
                // create training data
                training = sample(d, 80)
                // create test data
                test = d – training
                for a in algorithms A
                        model <- createModel(a, train)
                        accuracy <- findAccuracy(model, test)
                        write accuracy to file
                next a
        next numSamples
next d
```

Reporting your results:

You should report your results as follows:

Dataset	Number of total instances	Number of attributes	Percent split	Decision Tree Accuracy	…. (Other methods)	Naïve Bayes Accuracy
D1 -1	1000	10	80/20	85%		90%
D1 -2	1000	10	80/20	..	..	..
..						

Analysis:

Write a short paragraph explaining your results. Which method performs best and why do you think it performs the best? Which method is worst and why? Any other analysis that you wish to report