ASSIGNMENT 4

In this assignment, you will compare the performance of various classifiers that we have learned so far. You will write code in R that takes a dataset URL and properties from the command line and analyzes it using various classifiers.

Below is the list of the classifiers that you should implement:

1. Decision Tree
2. Support Vector Machines
3. Naïve Bayesian
4. kNN
5. Logistic Regression
6. Neural Network
7. Bagging
8. Random Forest
9. Boosting

In this assignment, you will read the path of the dataset via the command line. You will also read a Boolean flag indicating whether the dataset contains a header, and the position of the class attribute column from the command line.

To help you get started, I have created a template of the assignment in file assignment4.R. It shows you how to read the parameters from the command line. A shell script that contains the list of datasets and various parameters is also attached (testScript.sh). Your code will be tested using this script.

A list of datasets used is also included at the end of this file. For each dataset, you will **create ten different samples each having a ratio of 90:10 for training and testing**. Each classifier will be trained on the training part of the samples and tested on the test part. You will print the output of each classifier on screen as well as submit a report that contains summary of the output.

It is for you to find the best parameters for each classifier. You should experiment with different parameters until you find the best set.

Your code will be run from the UNIX shell script as:

./testScript.sh
**Note: The two files should be in the same directory and the shell script should be executable on your system. You can do that by chmod u+x testScript.sh

If you don't have UNIX on your computer, you can use the UTD Linux cluster as follows:
ssh <YourNetID>@csgrads1.utdallas.edu

** Also be sure to load all packages that you use. For example, if you use the package rpart, you should include following at the top:
library(rpart)

Pseudocode for the assignment is below. Please follow it carefully.

PseudoCode:

Read the dataset using the command line
Load in memory

Repeat 10 times
    Create random training and test sample using a ratio of 90:10
    for each of the classifiers
        Build a model using the training data
        Predict on the test data
        Calculate accuracy
        Output accuracy on screen
    end for
end Repeat

Format of report:

In the report, you have to build tables of output like:

Dataset1:

		Accuracy					
Method	**Best Parameters**	**Sample1**	**Sample2**	**…**	**..**	**Sample10**	**Average of 10 samples**
Decision Tree							
SVM							
…							
…							
Boosting							

Similarly for dataset 2, 3, 4, and 5.

You will be using the following datasets:

1. Credit default dataset that is available at:

https://gist.github.com/Bart6114/8675941

You can download a csv of this dataset from:
http://www.utdallas.edu/~axn112530/cs6375/creditset.csv

Class variable is attribute: **default10yr (Position 6 in the columns)**

2. Graduate school admission dataset that can be obtained from:

http://www.ats.ucla.edu/stat/data/binary.csv

Class variable is attribute: **admit (Position 1 in the columns)**

3. Wisconsin Prognostic Breast Cancer (WPBC) dataset:

Description is here:
https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wpbc.names

Data is here:
http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wpbc.data

Note: Read the field names in the description file. The class attribute is: **Outcome (Position 2 in the columns).** Also note that the data has no header field.

4. Wisconsin Diagnostic Breast Cancer (WDBC) dataset:
Description is here:
https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names

Data is here:
http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data

Note: Read the field names in the description file. The class attribute is: **Diagnosis (Position 2 in the columns).** Also note that the data has no header field.

5. Ionosphere dataset:
Description is here:
https://archive.ics.uci.edu/ml/machine-learning-databases/ionosphere/ionosphere.names

Data is here:
http://archive.ics.uci.edu/ml/machine-learning-databases/ionosphere/ionosphere.data

Note: Read the field names in the description file. The class attribute has **Position 35 in the columns.** Also note that the data has no header field.