CS 6375

ASSIGNMENT \_\_\_ 5\_\_\_

Names of students in your group:

Mansi Kukreja(mxk174330)

Deepan Verma(dxv160430)

Number of free late days used: \_\_\_\_\_\_\_\_\_0\_\_\_\_\_\_\_\_\_   
Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

**Dataset Source:**

The dataset for the following assignment has been chosen from the following source listed below:

<https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center>

**Dataset Details:**

The details of the dataset are as follow:

**Dataset Characteristics:** Multivariate

**Attribute Characteristics:** Real

**Associated Tasks:** Classification

**Number of Instances:** 748

**Number of Attributes:** 5

**Missing Values:** N/A

**Area:** Business

**Date Donated:** 2008-10-03

**Summary of the Dataset:**

**Number of Instances: 748**

**Number of Attributes: 5**

**Missing Attribute Values: NONE**

**Number of Cross-Validation performed: 10**

**Attribute Information:**

1. **RECENCY:** Months since last donation
2. **FREQUENCY:** Total number of donation
3. **MONETARY:** Total blood donated in c.c.
4. **TIME:** Months since last donation
5. **CLASS:** A binary variable i.e. 0 or 1 where 0 represents for not donating blood and 1 stands for donation of blood.

**Preprocessing the Dataset:**

1. **Checking the attributes and predicted values for null values:**

apply(dataset, MARGIN = 2, FUN = function(x) sum(is.na(x)))

Recency Frequency Monetary Time Class

0 0 0 0 0

1. **Scaling the dataset:**

The entire dataset has been scaled i.e. all the values have been scaled between 0 and 1.

##CALCULATING THE MAXIMUM AND MINIMUM VALUES FROM THE DATASET AND STORING THEM IN THE RESPECTIVE VARIABLES AND USING THEM FOR SCALING PURPOSES

maxValues = apply(dataset, MARGIN = 2, max)

minValues = apply(dataset, MARGIN = 2, min)

##SCALING THE DATASET AND FORMING A DATA FRAME WHICH IS STORED AS SCALED SET

scaledDataset = as.data.frame(scale(dataset, center = minValues, scale = maxValues-minValues))

1. **Looking for categorical attributes and making them categorical if the attribute isn’t categorical**

scaledDataset$class <- factor(scaledDataset$class)

1. **Finding the correlations**

##OBTAINING THE CORRELATION MATRIX AND STORING IT IN THE CORRELATION MATRIX

correlationMatrix<-cor(scaledDataset[,1:5])

##PRINTING THE CORRELATION MATRIX

print(correlationMatrix)

**Since Frequency and Monetary have the same values we have removed Monetary from the dataset for the analysis.**

1. **Evaluation Metric Used**

We used area under ROC curve as an evaluation metric.

1. **Results**

|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | **Best Parameters**  **Used** | **Accuracy** | **Evaluation Metric**  **(Area under ROC curve)** |
| **Decision Tree** | method="class”, parms=list(split="information") | 0.7135 | 0.5 |
| **Perceptron** | hidden = 0, threshold = 0.1, err.fct = "sse", linear. output = FALSE, act.fct = "logistic" | 0.7027 | 0.5 |
| **Neural Net** | hidden = c (2,2), err.fct = "ce”, rep = 5, threshold = 0.01, linear.output = FALSE, learningrate = 0.5,act.fct = "logistic" | 0.7027 | 0.5 |
| **Deep Learning** | numepochs = 100, momentum = 0.5, activationfun = "sigm”, learningrate\_scale=1 | 0.8513 | 0.5 |
| **SVM** | kernel = "linear", cost = 100, gamma = 1 | 0.7027 | 0.5 |
| **Naive Bayes** | laplace = 3 | 0.697 | 0.491 |
| **Logistic Regression** | family="binomial" | 0.7081 | 0.5 |
| **K-nearest Neighbors** | cl=train\_data$class,k=4,prob = TRUE,use.all = TRUE | 1 | 1 |
| **Bagging** | ns=275, nbagg=500, control=rpart. control(minsplit=5, cp=0, xval=0,maxsurrogate=0) | 0.6945 | 0.5 |
| **Random Forest** | importance = TRUE, mtry = 2, ntree = 500 | 0.7027 | 0.5 |
| **Adaboost** | mfinal = 10, control = rpart.control(maxdepth = 1) | 0.6945 | 0.491 |
| **Gradient Boosting** | max.depth=2,nrounds = 2 | 0.6459 | 0.49 |