CS 6375.001 MACHINE LEARNING
Project Status Report
November 06, 2016

Eswar Chowdary Ganta (exg151430)
Dileep Gudena (dxg161730)
Divya Reddy Vudem (dxv151430)
Sai Charan Rao Vennamanani (sxv157130)
Santhosh Kamishetty (sxk165130)

In this project, we aim at making a complete analysis of the TITANIC dataset to find what sorts of people were more likely to survive the shipwreck.

The Dataset - TITANIC

The TITANIC dataset is taken from an active Kaggle Competition and the link of which is given below:

https://www.kaggle.com/c/titanic/data?train.csv

- Number of attributes = 11 (including the class attribute)
- Number of instances = 891
- The attributes are the following:
  1. *Survival* – 0 if not survived and 1 if survived
  2. Pclass – Passenger Class (1 - Upper, 2 - Middle and 3 – Lower class)
  3. Name – Name of the passenger
  4. Sex – Gender of the passenger
  5. Age – Age of the passenger
  6. Sibsp – Number of siblings/spouses aboard
  7. Parch – Number of Parents/Children aboard
  8. Ticket – Ticket Number
  9. Fare – Ticket fare
  10. Cabin – Cabin number
  11. Embarked – Port of Embarkation (C, Q and S)

Here is the snapshot of the data:

```
> head(train)
  PassengerId Survived Pclass                                                 Name    Sex Age SibSp Parch
1           1        0      3                              Braund, Mr. Owen Harris   male  22     1     0
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
3           3        1      3                               Heikkinen, Miss. Laina female  26     0     0
4           4        1      1       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
5           5        0      3                             Allen, Mr. William Henry   male  35     0     0
6           6        0      3                                     Moran, Mr. James   male  NA     0     0
            Ticket     Fare Cabin Embarked
1        A/5 21171   7.2500              S
2         PC 17599 71.2833   C85        C
3 STON/O2. 3101282   7.9250              S
4           113803 53.1000  C123        S
5           373450   8.0500              S
6           330877   8.4583              Q
```

Techniques

We planned to apply the following techniques on the data to complete the required analysis –

1. Artificial Neural Network
2. Boosting
3. Random Forest

Experimental Methodology

- We employ the following procedure in our project –
  1. Pre-processing of the dataset
     - This step involves dealing with the NA values,
     - Selecting the attributes that influence the classification by observing the histograms and correlation plots,
     - Scaling the required attributes
  2. On the dataset
     - We perform each of the aforementioned techniques,
     - Also, vary the parameters and find the best one for the technique.
  3. We evaluate the techniques using the following metrics –
     - Accuracy
     - Precision
     - Recall
     - F-measure
  4. We plot the results that aid in comparing the performance of the classifiers.

Programming Language

- We plan to use **R** programming for the project.

Preliminary Results

We now present the result of the work we've done so far.

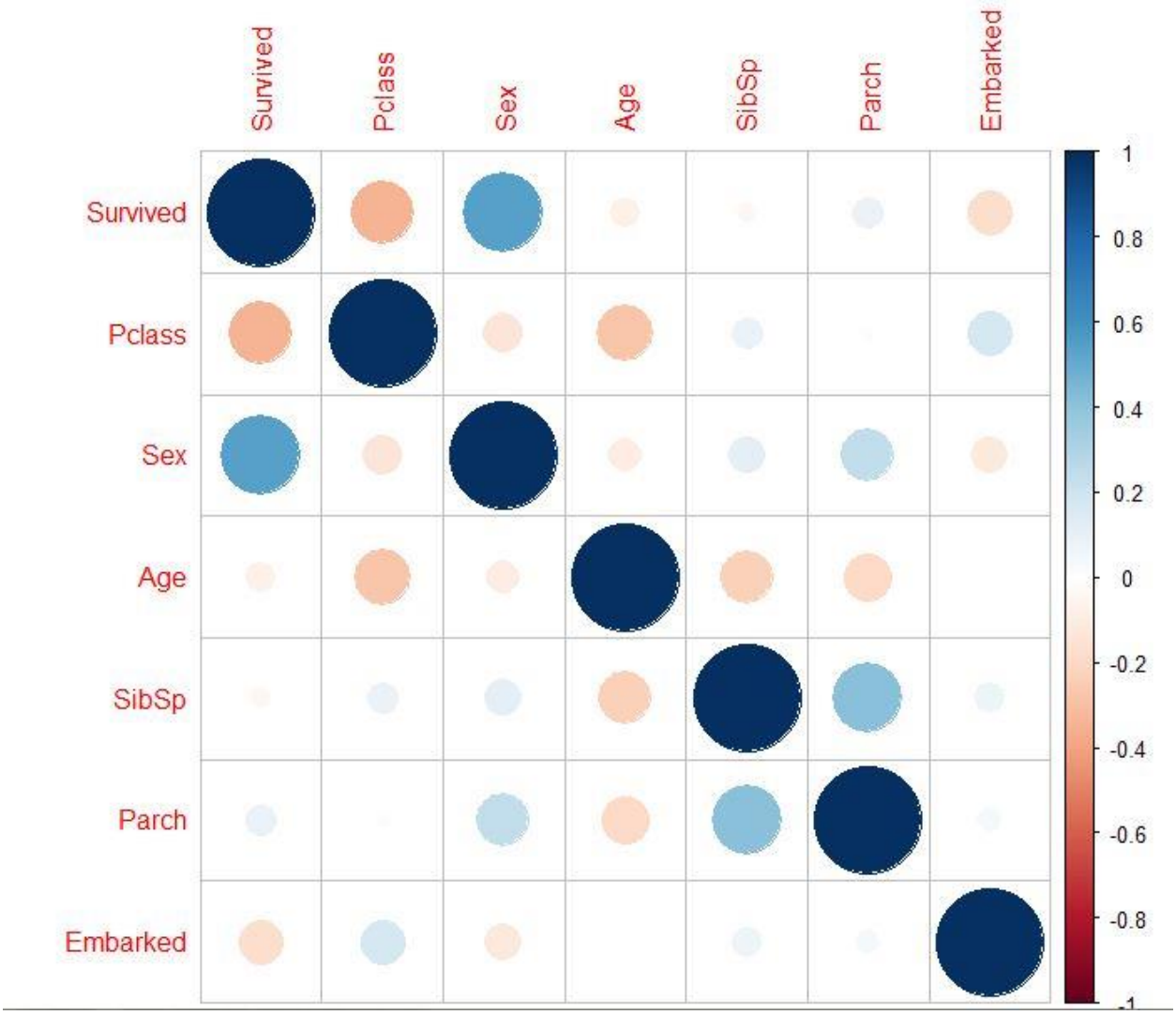
- We removed the following attributes (after initial examination of the dataset) from the dataset as they don't impact the result significantly –
  - Passenger Number (This is just a serial number)
  - Name (Passenger name has nothing to do with his/her survival)
  - Ticket Number
  - Ticket Fare
  - Cabin
- We categorized the AGE attribute into three different intervals as follows:
  - Category_1 – 18 and below (Children)
  - Category_2 – Between 19 and 40

- Category_3 – 41 and above

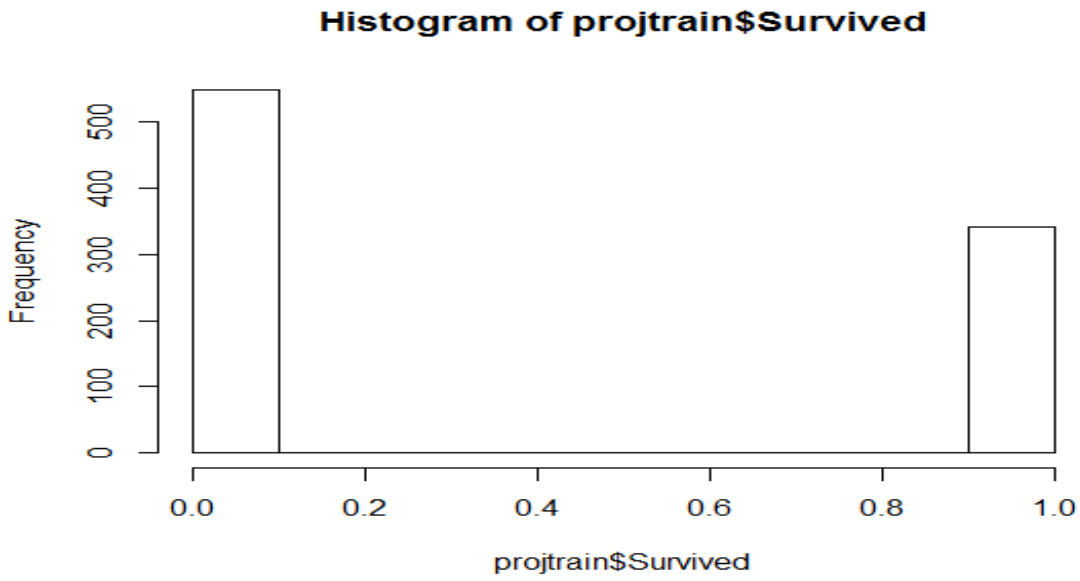
- The attribute SEX is labelled as follows:
  MALE – 1
  FEMALE – 2

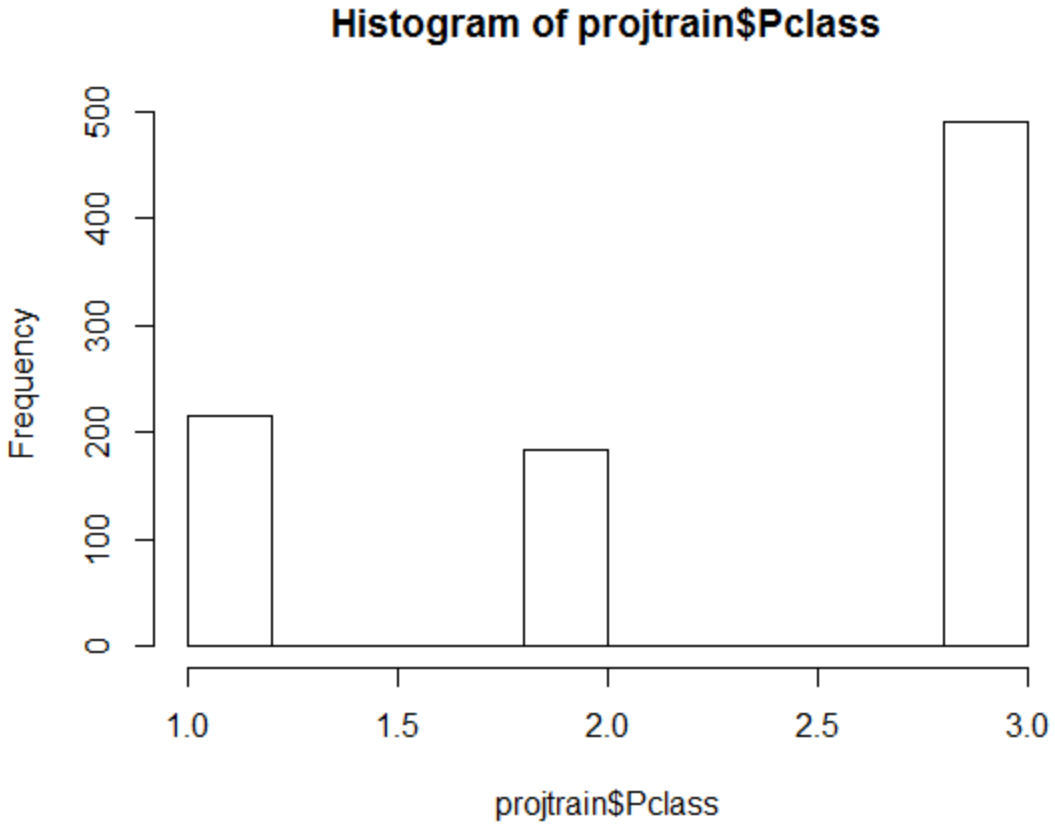
- We plotted the CORRPLOT which aids in identifying the correlation of the attributes with the class attribute.

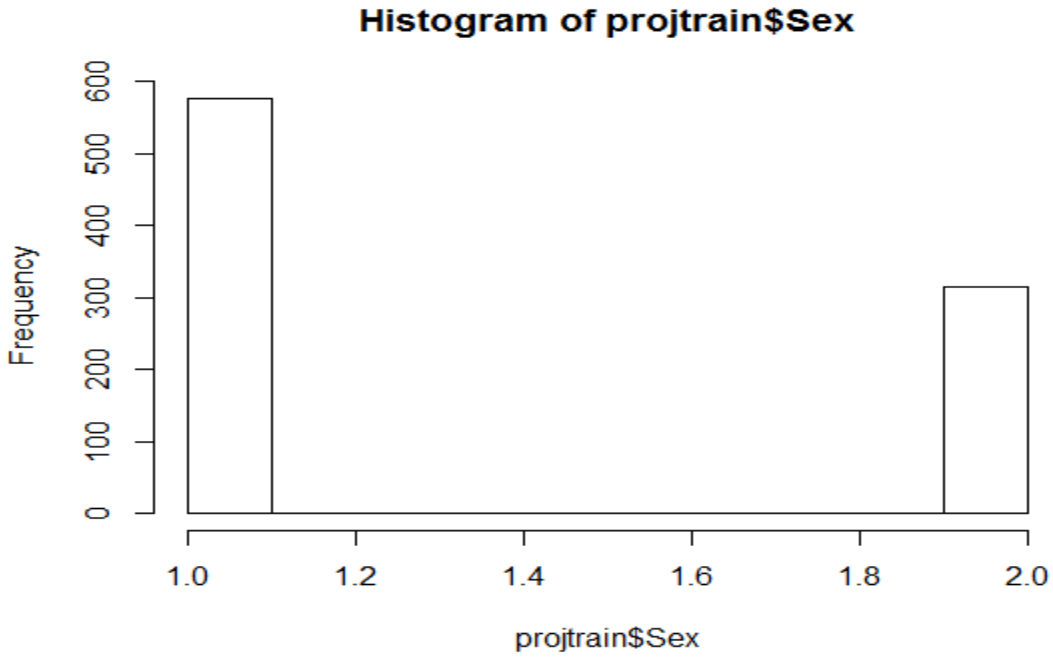
  The plot is as follows:

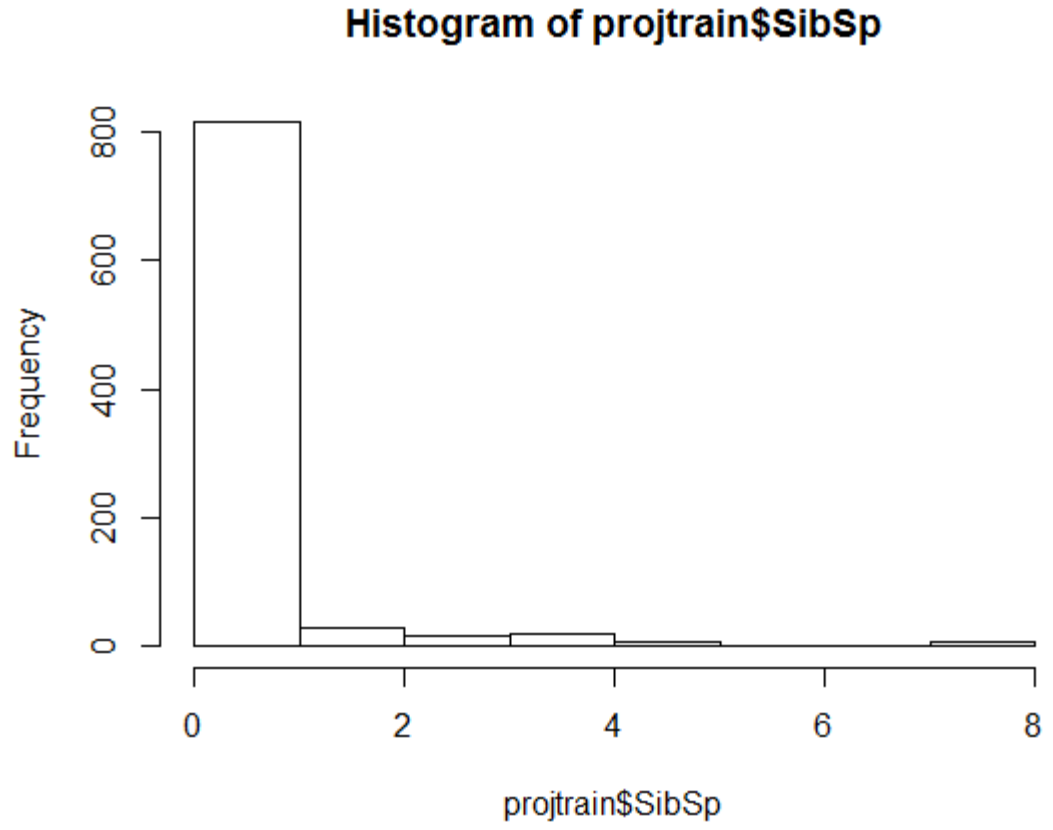


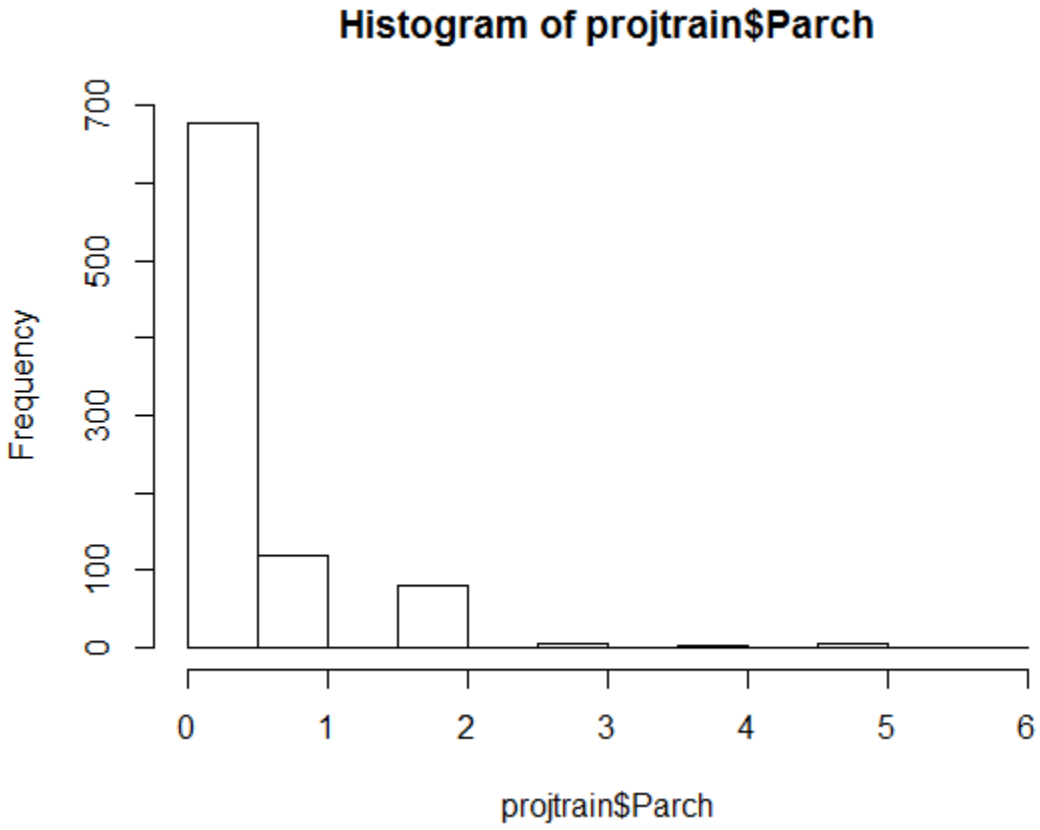
The histograms of the data are as follows:

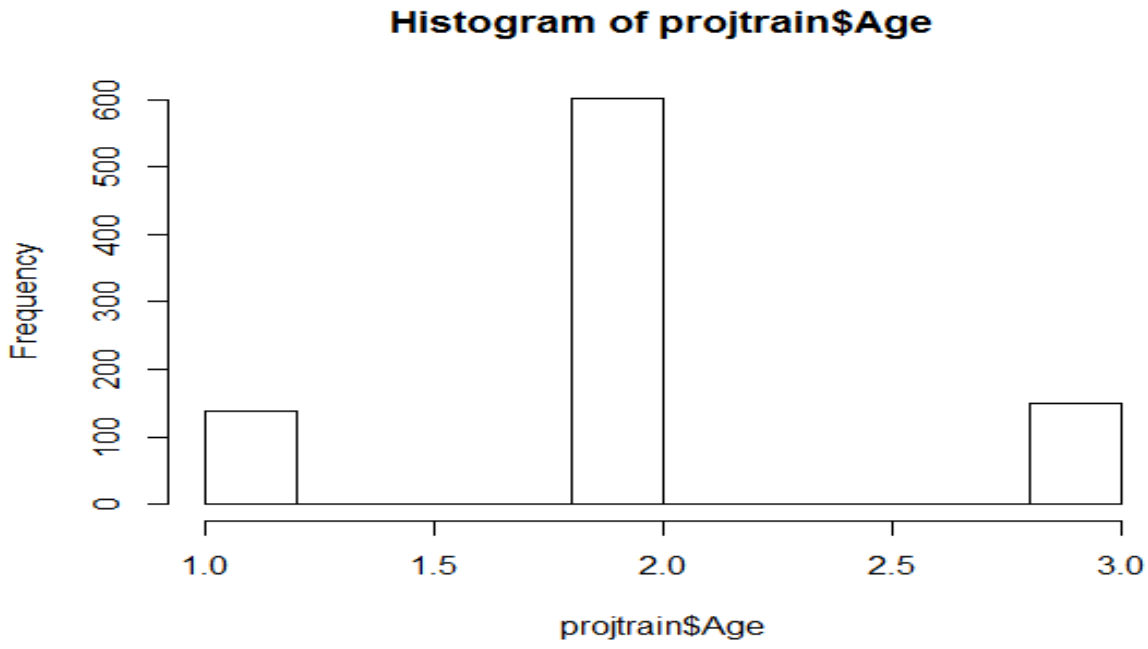





Histogram of projtrain$Sex
Frequency
0
100
200
300
400
500
600
1.0
1.2
1.4
1.6
1.8
2.0
projtrain$Sex


Histogram of projtrain$SibSp
Frequency
0
200
400
600
800
0
2
4
6
8
projtrain$SibSp


Histogram of projtrain$Parch
Frequency
0
100
300
500
700
0
1
2
3
4
5
6
projtrain$Parch


Histogram of projtrain$Age
Frequency
0
100
200
300
400
500
600
1.0
1.5
2.0
2.5
3.0
projtrain$Age

R – CODE

```
#Read the dataset
myData <-
read.csv("D:/Academics/Fall_2016/ML/Assignments/Project/train.csv")
View(myData)

#Pre-processing
projtrain <- myData
#Removing the attributes that are not required
projtrain <- projtrain[-c(1,4,9,10,11)]
View(projtrain)
head(projtrain)
summary(projtrain)

projtrain$Embarked <- as.numeric(projtrain$Embarked)
projtrain$Embarked
ageavg<-mean(na.omit(projtrain$Age))
ageavg
#Replacing the NAs in the age attribute with the average age value
projtrain[is.na(projtrain)]<-ageavg

projtrain[,3] = ifelse(projtrain[,3]=="male",1, 2)
projtrain[,4] = ifelse(projtrain[,4]<=18,1,
ifelse(projtrain[,4]<=40,2,3))

#Finding the correlation plot
library(corrplot)
p<-cor(projtrain)
corrplot(p,method = "circle")

#Plotting the Histograms
hist(projtrain$Survived)
hist(projtrain$Pclass)
hist(projtrain$Sex)
hist(projtrain$Age)
hist(projtrain$SibSp)
hist(projtrain$Parch)
hist(projtrain$embarked)
```