STATISTICAL LEARNING PROJECT

UNSUPERVISED LEARNING

PRINCIPAL COMPONENT ANALYSIS APPLIED TO THE TITANIC CASE

Report by

Amina Khalfa

ID 941835

DSE MSc Student

Titanic case

Table of Contents

ALGORITHMS

PRINCIPAL COMPONENT ANALYSIS: PCA (principal component analysis) is a method for emphasizing variance and highlighting significant patterns in a dataset. It's frequently used to make data exploration and visualization simple. It is mostly used to visualize patterns in multivariate data. Its goal is to depict data points' relative locations in fewer dimensions while maintaining as much information as feasible, as well as to investigate correlations between dependent variables. It's a hypothesis-generating approach that's meant to depict patterns rather than test formal statistical hypotheses in general. PCA was designed to study continuous variables, but it may also be used to examine ordinal and presence–absence data. The smaller of the number of original variables or the number of observations minus one equals the number of different primary components. This transformation is specified so that the first principal component has the greatest possible variance, and each subsequent component has the greatest possible variance while remaining orthogonal to the previous components. The generated vectors form an orthogonal uncorrelated basis set.

CLUSTERING: One of the most basic and often used unsupervised machine learning techniques is K-means clustering. A cluster is a collection of data points that have been grouped together due to particular commonalities. We'll use the number k to represent the number of centroids required in the dataset. A centroid is a fictional or actual place that represents the cluster's center. By lowering the in-cluster sum of squares, each data point is assigned to one of the clusters. To put it another way, the K-means algorithm finds k centroids and then assigns each data point to the closest cluster while keeping the centroids as small as feasible. The average of the data, or determining the centroid, is what the 'means' in K-means refers to. The K-means technique in data mining starts with a first group of randomly picked centroids, which serve as the starting points for each cluster, and then performs iterative (repetitive) computations to optimize the centroids' placements. It stops forming and optimizing clusters when either: the centroids have stabilized — their values haven't changed because clustering was successful; or the centroids have stabilized — their values haven't changed because clustering was successful. The specified number of iterations was completed.

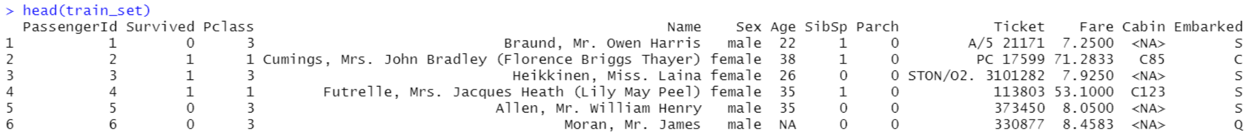
PROBLEM DEFINATION AND DATA PREPARATION

PROBLEM DEFINATION

It can be seen from previous analysis that the Titanic data by Cabin is an attempt to locate lucky ship placements. However, the majority of cabin data is missing, which will have an impact on the results. This report aims to address the issue of missing cabin number data. First, unsupervised learning technique Principal Components Analysis (PCA) is used to determine whether cabin number is important. Missing cabin number is associated to PClass to some extent (0.738), according to missing data study. Finally, missing data is imputed using Multiple Imputation (MI). Then another technique clustering is used to group the survived and not survived people.

DATA PREPROCESSING

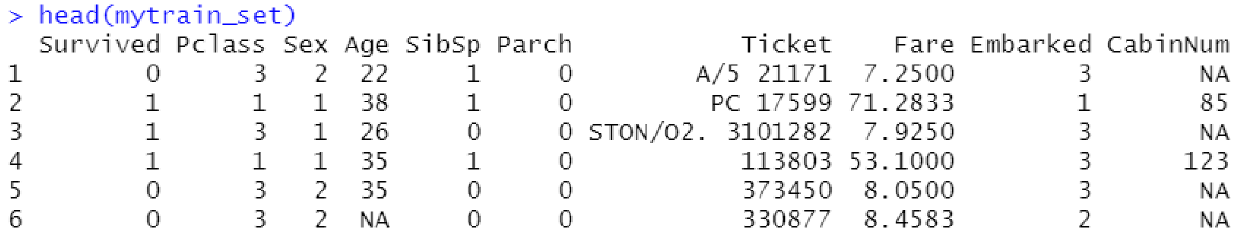
Data is loaded and prepared in this section. Cabin number is taken from the Cabin variable and added to the data set as a variable. Factor class is created from the variables Sex and Embarked. The following analysis will ignore variables such as PassengerId, Name, and so on.



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

Figure 1: Dataset Overview

Now there is need to convert data character variables in integer format which gives the final look of data given below

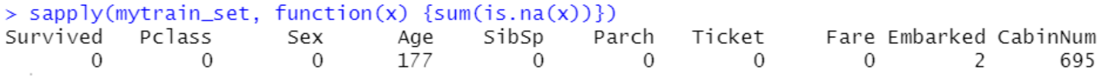


```
> head(mytrain_set)
  Survived Pclass Sex Age SibSp Parch           Ticket    Fare Embarked CabinNum
1        0      3   2  22     1     0        A/5 21171  7.2500        3       NA
2        1      1   1  38     1     0         PC 17599 71.2833        1       85
3        1      3   1  26     0     0 STON/O2. 3101282  7.9250        3       NA
4        1      1   1  35     1     0           113803 53.1000        3      123
5        0      3   2  35     0     0           373450  8.0500        3       NA
6        0      3   2  NA     0     0           330877  8.4583        2       NA
```

Figure 2: Transformed Data

MISSING DATA ANALYSIS

Now we will finally check how much missing values are there in this dataset. So, here we made a new function in sapply() to get the number of missing values in each variable. Therefore, below figure shows that age has 177 missing values while Embarked has 2 and Cabin number has 695 missing values which we try to impute in next sections.



```
> sapply(mytrain_set, function(x) {sum(is.na(x))})
Survived   Pclass      Sex      Age    SibSp    Parch   Ticket     Fare Embarked CabinNum 
       0        0        0      177        0        0        0        0        2      695 
```

Figure 3: Missing Values

The total number of cases in the data collection is tallied. That suggests that one or more missing values can be found in 80% of observations. The following graph depicts the missing data pattern.

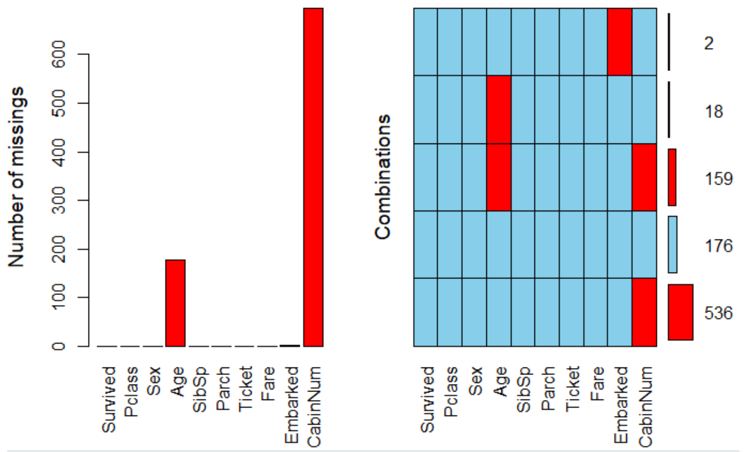
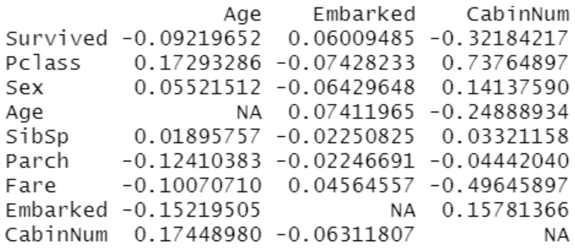


Figure 4: Missing Values Patterns

The missing data is mostly found in the Cabin and Age variables. Other variables, with the exception of Embarked, contain no missing data. Is there a link between Cabin Variable and Age? There are 159 observations missing when the two are combined, which is a significant quantity when compared to the sample size.



```
                 Age    Embarked    CabinNum
Survived -0.09219652  0.06009485 -0.32184217
Pclass    0.17293286 -0.07428233  0.73764897
Sex       0.05521512 -0.06429648  0.14137590
Age               NA  0.07411965 -0.24888934
SibSp     0.01895757 -0.02250825  0.03321158
Parch    -0.12410383 -0.02246691 -0.04442040
Fare     -0.10070710  0.04564557 -0.49645897
Embarked -0.15219505          NA  0.15781366
CabinNum  0.17448980 -0.06311807          NA
```

Figure 5: Relationship of CabinNum with other variables

Those missing data, however, are unlikely to be missing together, according to the preceding correlation efficiency matrix.

With missing data and additional factors, the correlation coefficient of variables Age, Embarked, and CabinNum is calculated. There is no way to count a ticket as a character. PClass (r = 0.73) and Missing CabinNum are tightly connected. CabinNum is more likely to be absent if PClass is large. Furthermore, there is no evident relationship between Age and the other factors. CabinNum info that is missing might come from MAR (Missing at Random). And embarked might be MCAR (Missing Completely at Random) or MNAR (Missing Not at Random).

RESEARCH

PRINCIPAL COMPONENT ANALYSIS

The number of principle components is determined by the correlation coefficient between variables using a scree plot and parallel analysis. Missing data is omitted in this phase. Because character classes cannot be counted, Variable Ticket is also removed.

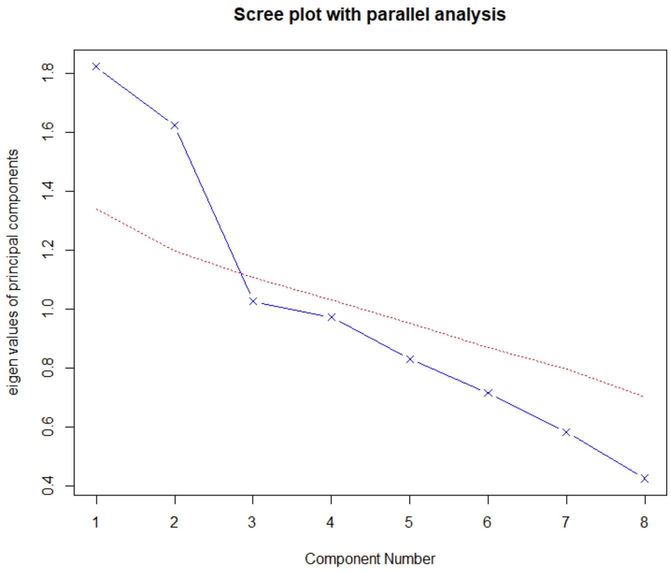
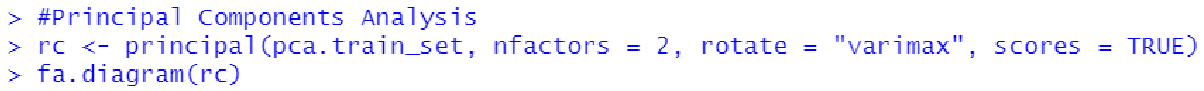


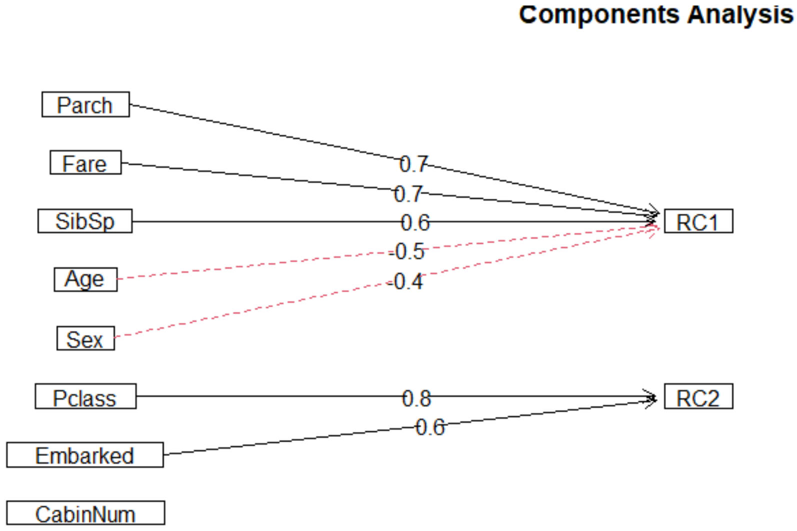
Figure 6: PCA Scree Plot

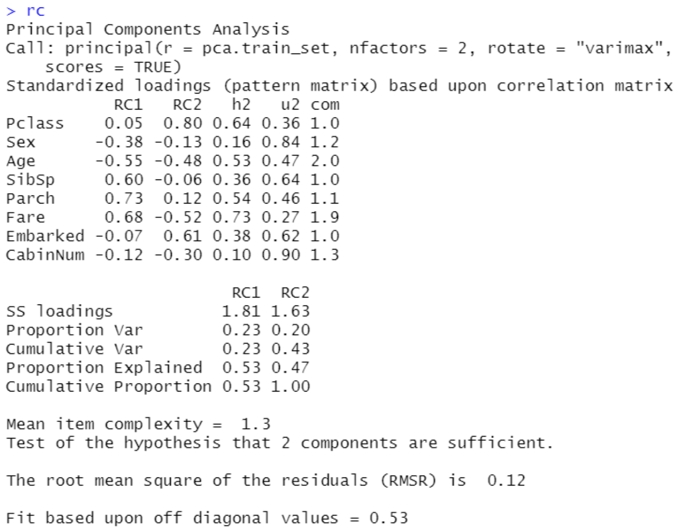
Parallel analysis reveals that the number of components = 2 based on the Kaiser-Harris criteria (i.e., the eigenvalue-greater-than-one rule).

The following is the PCA (Principal Components Analysis) code and result.



```
> #Principal Components Analysis
> rc <- principal(pca.train_set, nfactors = 2, rotate = "varimax", scores = TRUE)
> fa.diagram(rc)
```



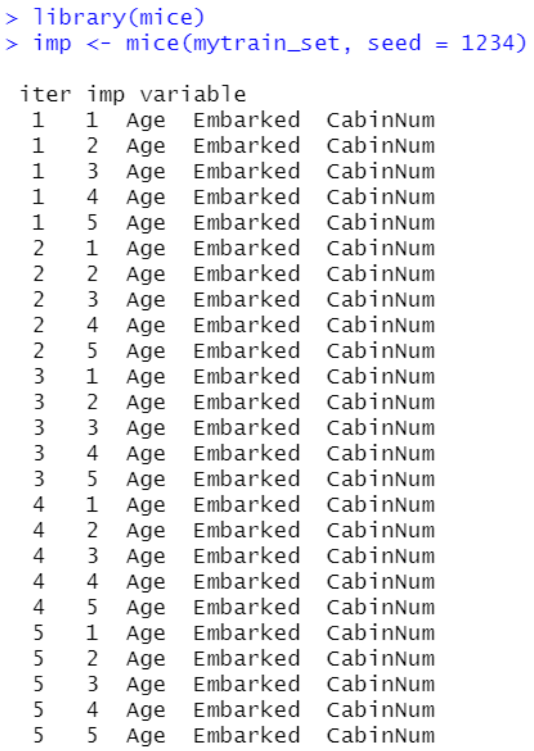


```
> rc
Principal Components Analysis
Call: principal(r = pca.train_set, nfactors = 2, rotate = "varimax",
    scores = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
           RC1   RC2   h2   u2 com
Pclass    0.05  0.80 0.64 0.36 1.0
Sex      -0.38 -0.13 0.16 0.84 1.2
Age      -0.55 -0.48 0.53 0.47 2.0
SibSp     0.60 -0.06 0.36 0.64 1.0
Parch     0.73  0.12 0.54 0.46 1.1
Fare      0.68 -0.52 0.73 0.27 1.9
Embarked -0.07  0.61 0.38 0.62 1.0
CabinNum -0.12 -0.30 0.10 0.90 1.3

                       RC1  RC2
SS loadings           1.81 1.63
Proportion Var        0.23 0.20
Cumulative Var        0.23 0.43
Proportion Explained  0.53 0.47
Cumulative Proportion 0.53 1.00

Mean item complexity =  1.3
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.12

Fit based upon off diagonal values = 0.53
```

Figure 7: Components Analysis

MISSING VALUES IMPUTATION

In this section, package mice are used to fill in blanks in lacking data. The method mice () impute the source data set and returns entire data sets in a single object. For repeatable research, the random seed is set to 1234.



```
> library(mice)
> imp <- mice(mytrain_set, seed = 1234)

 iter imp variable
  1   1  Age  Embarked  CabinNum
  1   2  Age  Embarked  CabinNum
  1   3  Age  Embarked  CabinNum
  1   4  Age  Embarked  CabinNum
  1   5  Age  Embarked  CabinNum
  2   1  Age  Embarked  CabinNum
  2   2  Age  Embarked  CabinNum
  2   3  Age  Embarked  CabinNum
  2   4  Age  Embarked  CabinNum
  2   5  Age  Embarked  CabinNum
  3   1  Age  Embarked  CabinNum
  3   2  Age  Embarked  CabinNum
  3   3  Age  Embarked  CabinNum
  3   4  Age  Embarked  CabinNum
  3   5  Age  Embarked  CabinNum
  4   1  Age  Embarked  CabinNum
  4   2  Age  Embarked  CabinNum
  4   3  Age  Embarked  CabinNum
  4   4  Age  Embarked  CabinNum
  4   5  Age  Embarked  CabinNum
  5   1  Age  Embarked  CabinNum
  5   2  Age  Embarked  CabinNum
  5   3  Age  Embarked  CabinNum
  5   4  Age  Embarked  CabinNum
  5   5  Age  Embarked  CabinNum
```

Figure 8: Imputation of Missing values

K-MEANS CLUSTERING

Clustering is a term that refers to a collection of approaches for identifying subgroups of observations in a dataset. We would like to have observations, in the same group, to be similar and observations in other groups to be distinct when we cluster them.

When we clused, it is advisable to use the least number of cluster possible, according to experts. Cluster partitioning approaches, such as k-means clustering, work by defining clusters so that total intra-cluster variance (also known as total within-cluster variation) is minimized. As a result, we can apply the procedure below to find the best clusters:

- Calculate different values of k using a clustering technique (e.g., k-means clustering). For example, if you change k, you might get a different result.
- Calculate the total within-cluster sum of squares for each k. (WSS)
- Make a curve graph.

In most cases, the location of a bend (knee) in the plot is used to determine the proper number of clusters.

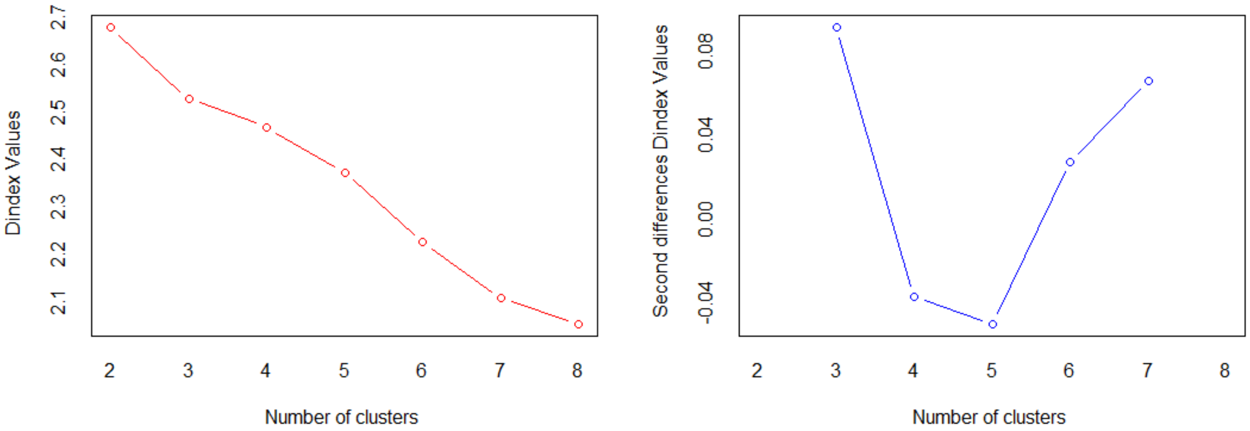


Figure 9: Optimal Number of Clusters

Below figure shows the clusters. We can see that our groups resulted in two cluster sizes of 30 and 20 if we print the data. For each observation, we also obtain the cluster assignment (i.e., Survived was assigned to cluster 1, Not Survived was assigned to cluster 1, etc.).

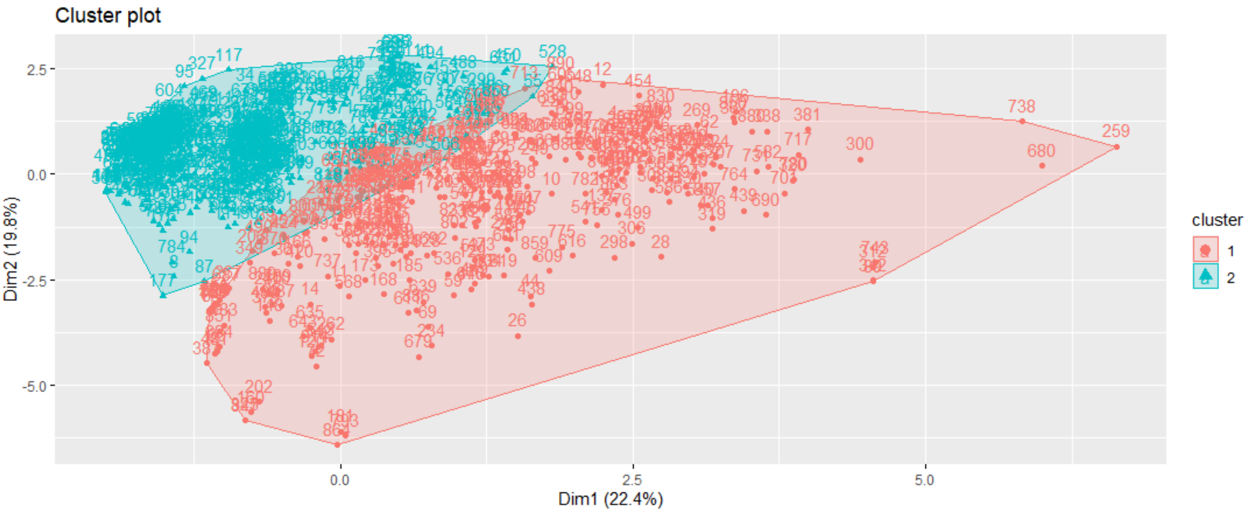
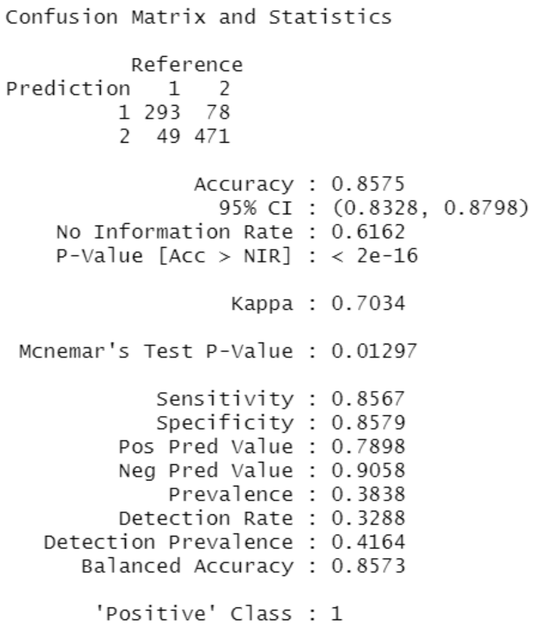


Figure 10: Cluster Visualization

The confusion matrices below reveal that k-means has an accuracy of 85.75 percent, with 78 false positives (FP) and 49 false negatives (FN) (FN).



```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 293  78
         2  49 471

               Accuracy : 0.8575
                 95% CI : (0.8328, 0.8798)
    No Information Rate : 0.6162
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7034

 Mcnemar's Test P-Value : 0.01297

            Sensitivity : 0.8567
            Specificity : 0.8579
         Pos Pred Value : 0.7898
         Neg Pred Value : 0.9058
             Prevalence : 0.3838
         Detection Rate : 0.3288
   Detection Prevalence : 0.4164
      Balanced Accuracy : 0.8573

       'Positive' Class : 1
```

Figure 11: K-Means Confusion Matrix

The trade-off between sensitivity (or TPR) and specificity (1 – FPR) is depicted by the ROC curve. Classifiers with curves that are closer to the top-left corner perform better. Here we got a good AUC of 0.41 showing that unsupervised learning is performing better, by a small variation.

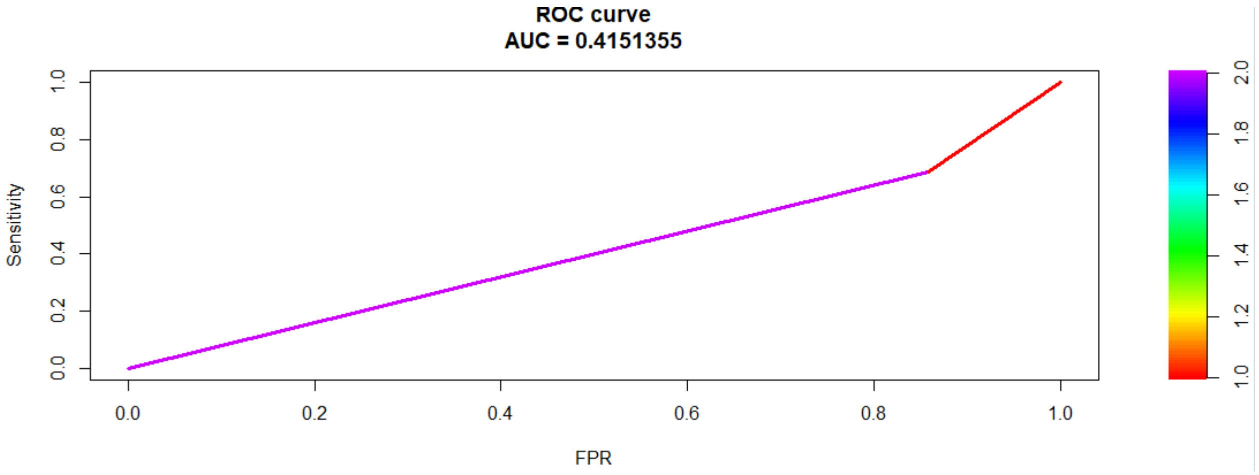


Figure 12: K-Means ROC Curve

CONCLUSION

This report is based on the implementation of unsupervised learning techniques for dimensional reduction. Furthermore, we have the imputation of missing data using PCA and them predicting the survival ratio using clustering algorithm. So, first factor analysis is putted into action to check the components. Then, missing values are imputed with the help of this PCA process and finally, the process starts on this complete data. Regarding clustering, first there is a need to find optimal number of clusters which leads to the use of Elbow method with in sum of squares method. This method gives 3 optimal number of clusters but 3rd number shows that some values were missing which we imputed later. After that, the k-means model is implemented using 2 clusters which are shown in given figure. Then, this model is evaluated with the help of confusion matrix which shows the 85% accuracy of model, with the ROC curve showing how good of a fit is the model. In the end, the AUC shows a percent about 0.42, that considering everything is a little variation, still including the fact that is done, in unsupervised learning, it is considered a good result.

LIST OF FIGURES