IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Passenger data analysis of Titanic using machine learning approach in the context of chances of surviving the disaster

To cite this article: Md Arfinul Haque *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1065** 012042

Passenger data analysis of Titanic using machine learning approach in the context of chances of surviving the disaster

Md Arfinul Haque[1], Shivaprasad G[2,4] and Guruprasad G[3]

[1]PG Student, Department of CSE, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India

[2]Department of CSE, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India

[3]Department of CSE, Yenopaya Institute of Technology, Mangalore, India

[4]shiva.prasad@manipal.edu

Abstract. Titanic disaster occurred about 100 years back but still it attracts the researchers to understand and study that how some passengers survived and others perished. In this work, the characteristics of the passengers will be identified and the relationship of survival chance from the disaster is found. Feature engineering techniques will be performed, where the alphabetic values will be changed to numeric values, the family size will be calculated. Also, we will extract the title from the name, and deck label from ticket number. Classification is done using Decision tree machine learning classification algorithm using two classes which are survived and not survived. R programming has been used for its implementation. Clustering is performed using KMeans machine learning algorithm. Its implementation has been done using Python programming.

1. Introduction

One the most infamous Shipwrecks in history is the sinking of the RMS Titanic. It happened on April 15, 1912. When an iceberg collided with the titanic, it sank and 1502 passengers died from the total of 2224 passengers as well as the crew people got killed. The International community got shocked by this sensational tragedy and because of this, now there are better safety regulations for the ships.

Based on data set obtained from Kaggle, the purpose is about survival prediction of the titanic [4]. We will split the historical data set that we obtained from kaggle into two groups – train set and test set. They are the two different files in csv format. The train data will be an input for the training model. This data will be used to build the model for the generations of predictions for the test data. There can be many flaws in the train set because of which training of the model on the basis of such data is not possible or can leads to an incorrect output. It is possible that in the csv file of the train set, there will be few blank cells where no values is mentioned. For an example, if the age of a person is not mentioned then this a flaw in the train data set. Hence, it has to be filled with appropriately calculated age value before training phase. This is feature engineering that we do to make the train data set perfect for giving it as an input for training the model. The titles for detecting the crew members and other professional among the passengers also has to be grabbed. Such titles can be like Dona, Lady, Capt, Sir, Doc etc.

Male and female will be identified using those titles of the names. The titles belonging to a female and male professionals will help to classify the female and male people, then all such titles which were in huge number will be shorted down into titles like Master, Miss, Mr, Mrs and Rare Title. The surname from the passenger names has to be grabbed for detecting the members in the family. This can be done by breaking the passenger names into some meaningful variables. The unique surnames will be grabbed to analyze the families. After applying further such feature engineering on the train data, the training model will be applied using classification algorithm. Decision tree machine learning algorithm will be used for predicting the survival status of each passenger in the test data set. Decision tree is used to train the machine about the survival of the people in titanic. Two classes will be used for predicting the survival status [3]. At every level in the decision tree the probability of the survival will be identified. Every node in the decision tree can be identified using the node number and at every node number the number of observations can be observed. The information at every node in the tree will be explained using the parameters like predicted class, expected loss, P(node), class counts, probabilities. These parameters will be used to identify the information of people whose family size is zero or was travelling alone. For family, at every node the information will be explained using complexity parameter, predicted class, expected loss, P(node), class counts, probabilities, left son and right son. Further the primary splits and surrogate splits will be displayed using age, passenger class and family size. Overall summary of decision tree classification will be examined using conditional probability, number of splits, relative error and expected error. Accuracy of the predicted test data will be calculated with respect to the survived information available in the train data set.

Similarly, there are modifications that needs to be performed on training data set. The prediction that whether or not the titanic passengers survived the sinking will be found for each passenger in the test set. Different machine learning techniques will be applied on the data set available. By the results obtained from those techniques we will be able to understand the statistics more clearly. Algorithms used includes feature engineering, decision tree visualization, random forest etc. [1].

2. Problem Definition

We have to perform the complete analysis about what kind or sort of titanic passengers likely to survive in the titanic disaster. The type of passengers include those who travelled alone and passengers with their family. Passengers type can be analyzed by family size, family size has to be categorized into three categories like singleton, small and large family. Family size as 1 will belong to singleton, from 2 to 4 family size as small family and family size greater than 4 as a large family. Depending on three family categories, the survival rate is small or large can be identified. It has been illustrated in the Fig.1 for understanding the family size categorization. This will help in training the data about survival and then same can be predicted in test data.

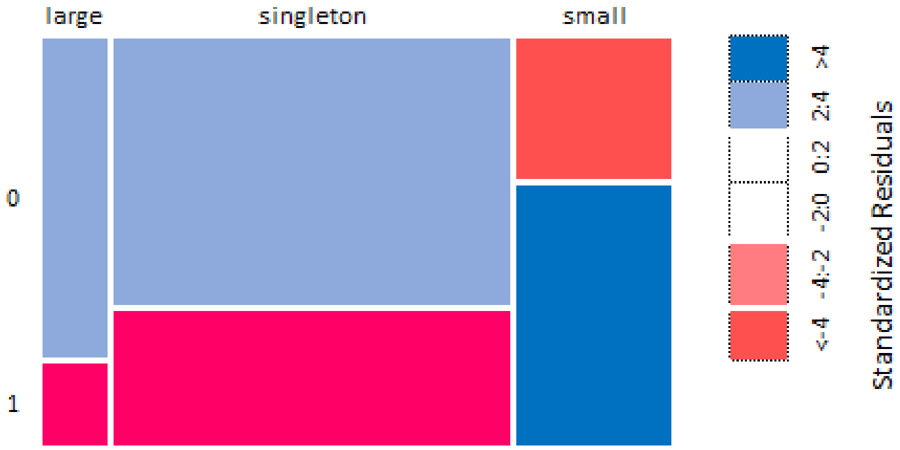
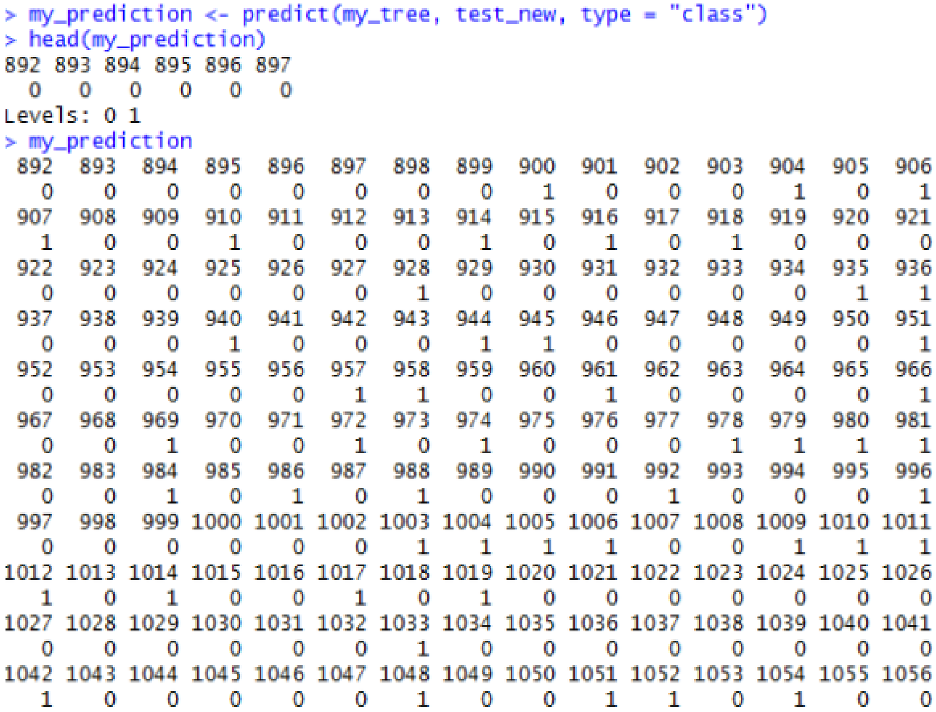


Fig. 1. Family Size Survival

Similarly, the crew members also have to be classified by their working departments. Finally, the output will be the survival status as 1 and not survived as 0. Survival status has to be understood using classification algorithm and using also using a clustering algorithm. For classification and clustering, decision tree and KMeans algorithm has to be implemented. Accuracy has to be checked for the machine learning algorithms implemented [5]. Expected accuracy is greater than 80 percent but for future work more than the resulted accuracy rate has to be implemented. This is possible by modifying and feature engineering the train data more accurately. Therefore, for this, we will be applying the tools of machine learning for prediction of passengers survived the tragedy. In the decision tree, there will be 2 classes i.e. male and female then depending on the survival status 0 and 1, the decision tree has to be grown. For the survival status of male and female respectively, further survival rate will be calculated. Same way the decision tree will grow depth wise. Each node of decision tree will indicate either 1 or 0 which are the status of survived and not survived passengers.

3. Objective

In the test data, all the attributes of train data will be mentioned except the survived column. So, the objective is to get the survival status of each passenger in the test data. The same expected result is illustrated in the Fig.2 and also the accuracy value in this prediction when compared with the survived status of train data set.



```
> my_prediction <- predict(my_tree, test_new, type = "class")
> head(my_prediction)
892 893 894 895 896 897 
  0   0   0   0   0   0 
Levels: 0 1
> my_prediction
 892  893  894  895  896  897  898  899  900  901  902  903  904  905  906 
   0    0    0    0    0    0    0    0    1    0    0    0    1    0    1 
 907  908  909  910  911  912  913  914  915  916  917  918  919  920  921 
   1    0    0    1    0    0    0    1    0    1    0    1    0    0    0 
 922  923  924  925  926  927  928  929  930  931  932  933  934  935  936 
   0    0    0    0    0    0    1    0    0    0    0    0    0    1    1 
 937  938  939  940  941  942  943  944  945  946  947  948  949  950  951 
   0    0    0    1    0    0    0    1    1    0    0    0    0    0    1 
 952  953  954  955  956  957  958  959  960  961  962  963  964  965  966 
   0    0    0    0    0    1    1    0    0    1    0    0    0    0    1 
 967  968  969  970  971  972  973  974  975  976  977  978  979  980  981 
   0    0    1    0    0    1    0    1    0    0    0    1    1    1    1 
 982  983  984  985  986  987  988  989  990  991  992  993  994  995  996 
   0    0    1    0    1    0    1    0    0    0    1    0    0    0    1 
 997  998  999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 
   0    0    0    0    0    0    1    1    1    1    0    0    1    1    1 
1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 
   1    0    1    0    0    1    0    1    0    0    0    0    0    0    0 
1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 1040 1041 
   0    0    0    0    0    0    1    0    0    0    0    0    0    0    0 
1042 1043 1044 1045 1046 1047 1048 1049 1050 1051 1052 1053 1054 1055 1056 
   1    0    0    0    0    0    1    0    0    1    1    0    1    0    0 
```

Fig. 2. Survival status in test data

In order to get the expected result, there is a need to apply 3 techniques which are feature engineering technique, classification algorithm and clustering algorithm.

Feature Engineering [7][8][9][10][11] is required for modifying the train data set because there can be some empty spaces, title for some passengers will be missing, department for some crew members may not be mentioned. Unique surname has to be extracted then calculation of family size will be required for analyzing the families through unique surnames. Visualization of relationship between the family size and survival is required. Hence for all such errors this technique has be applied before applying the machine learning techniques for training the model. The best way for doing this using R programming for data extraction and analyzing using graphs.

Classification Technique [12][13] is used to determine that for which class or category a given observation will belong to. There are many algorithms related to classification but for this work, the decision tree algorithm is employed. All nodes in the tree will indicate either the passenger survived or not survived using 0 and 1 respectively. At the first level, the relation between the child nodes and the root node will indicate the gender of passenger. Further, the relation will be based on age, Passenger class and family size. The implementation will be done using the library called "rpart" in R programming. Accuracy for the resulted survival status in the test data has to be calculated by comparing with the survival status of passengers in train data.

Clustering Technique [14][15] is used to group a set of observations in such a way that observation of similar types belongs to a group. There can be many groups defined while clustering. This technique has also to be implemented for examining passenger's survival in titanic disaster. This has been illustrated using KMeans algorithm in Python as illustrated in the Fig. 2.

4. Background

This research work is part of machine learning and data mining domain. A computer can get an ability to learn without being explicitly programmed by machine learning. Algorithms with machine learning techniques can be constructed in such a way that a computer can learn from and also make predictions on data available. Computational statistics is closely related or we can say overlapped in machine learning. Computers can do predictions on data available by machine learning techniques. Even machine learning is very closely related to mathematical optimization by which many methods, theory and applications gets developed.

5. Implementation

For implementation of the task to be performed, R and Python has been used. There are different R libraries involved which can be understood by the table (Table 1).

Feature engineering technique where Passenger names has to be broken down into some meaningful variables which will be used for predictions or for creating new variables. As passenger title is mentioned in passenger name, we can use a new variable as surname to represent the families. After that all the unique surnames of all the passengers in the train data has to be grabbed to analyze the families. The family size will be calculated using this result of unique name extraction. The family size calculation will be done using the formula as:

$$\text{Family Size} = \text{Parents Children} + \text{Siblings Spouses} + 1$$

Further, the missing value of passengers age has to be calculated for filling such missed age values of passengers. After applying all such feature engineering methods, while analyzing the data, it can be noticed that the death of passengers who travelled alone was more compared to the survival of singleton passengers. As illustrated in Fig.3, death was less than survival for passengers whose family size was 2. Similar condition was observed with the passengers with family size of 3 and 4. This indicates that survival status for small sized family was more compared to singleton and large family size. Further, after analyzing the data, the training model will be generated.

Table 1. Different R Libraries Involved

Purpose	R Library Involved
Visualization	library('ggplot2')
	library('ggthemes')
	library('scales')
Data manipulation	library('dplyr')
Imputation	library('mice')
Decision tree	library(rpart)

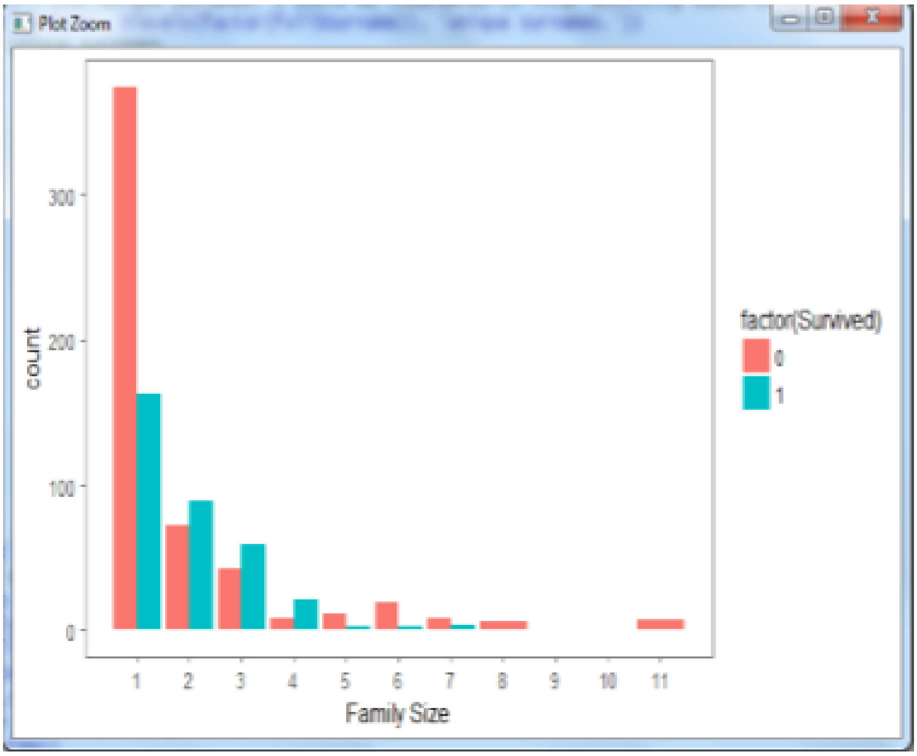
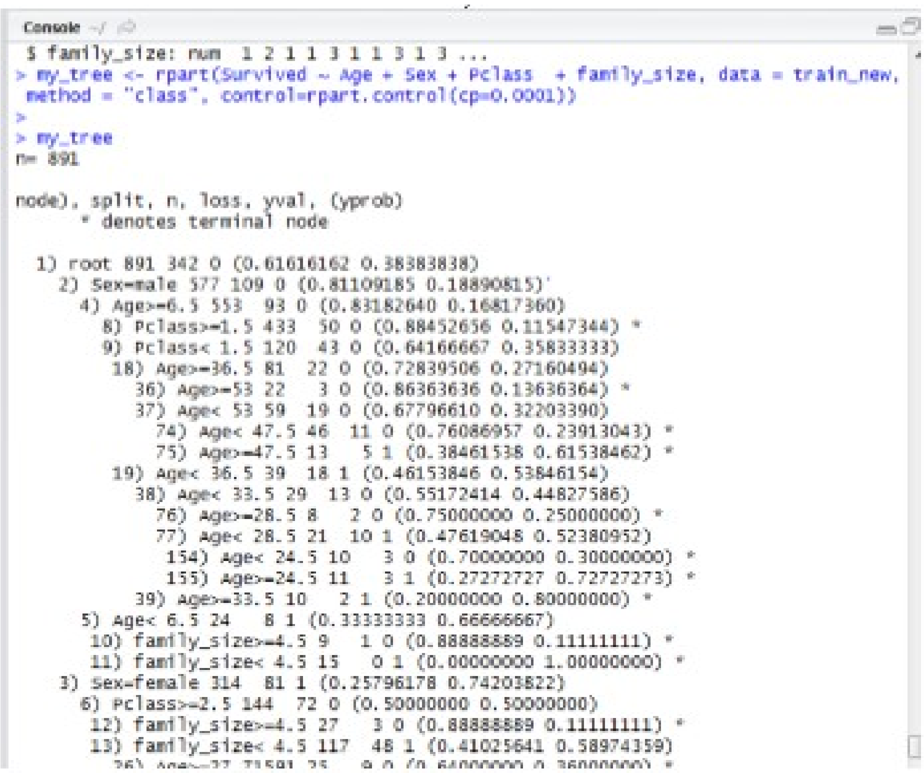


Fig. 3. Family size & Survival

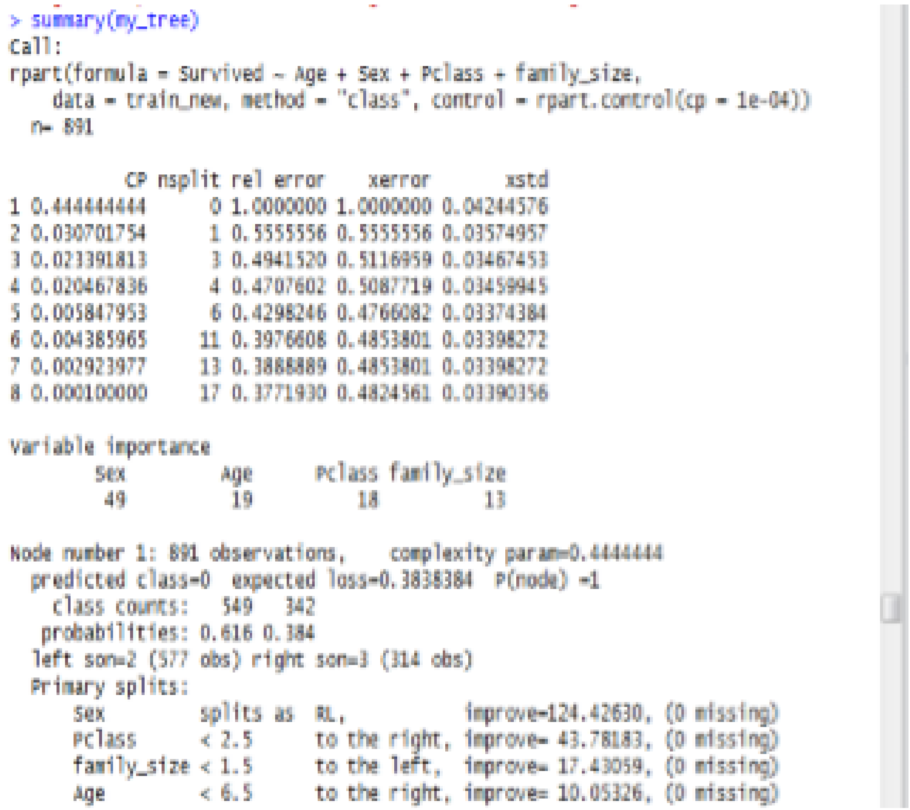
Classification technique will be applied for determining the class or category of a given observation to which it will belongs to. All nodes in the tree will indicate either the passenger survived or not survived using 0 and 1 respectively. As illustrated in the Fig. 4, at the first level the relation between the child nodes and the root node will indicate the gender of passenger. Further, the relation will be based on age, Passenger class and family size. The implementation will be done using the library called "rpart" in R programming. For every node, the survival status has to be displayed. The splits are categorized into primary and surrogate split. Each split is based on sex, Passenger class, family size, and age. Splits will appear only if a passenger has travelled with family. At every node, the number of observations will be displayed.

Two classes have been used to classify the survival i.e- survived and not survived. After generating the tree using 'rpart' in R language, we can notice that totally 435 nodes will be generated [2]. Also level wise error is decreasing or accuracy is increasing. At first level, the accuracy is zero as error is 100 percent. The summary of the information from the decision tree has been shown in Fig. 5 where columns indicate the number of splits, relative error and expected error at each level. Conditional probability at each level will get reduced from root level to the leaf. Conditional level has been taken as 0.0001 while implementing. This indicates that, when the conditional probability of a node will reach this, then further node generation will get stopped. The expected output for the survival status for each passenger has been displayed in the Fig. 2 by 1 and 0 which are indicating survived and not survived passengers in the titanic disaster.



```
 $ family_size: num  1 2 1 1 3 1 1 3 1 3 ...
> my_tree <- rpart(Survived ~ Age + Sex + Pclass  + family_size, data = train_new,
 method = "class", control=rpart.control(cp=0.0001))
>
> my_tree
n= 891

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 891 342 0 (0.61616162 0.38383838)
    2) Sex=male 577 109 0 (0.81109185 0.18890815)
      4) Age>=6.5 553  93 0 (0.83182640 0.16817360)
        8) Pclass>=1.5 433  50 0 (0.88452656 0.11547344) *
        9) Pclass< 1.5 120  43 0 (0.64166667 0.35833333)
         18) Age>=36.5 81  22 0 (0.72839506 0.27160494)
           36) Age>=53 22   3 0 (0.86363636 0.13636364) *
           37) Age< 53 59  19 0 (0.67796610 0.32203390)
             74) Age< 47.5 46  11 0 (0.76086957 0.23913043) *
             75) Age>=47.5 13   5 1 (0.38461538 0.61538462) *
         19) Age< 36.5 39  18 1 (0.46153846 0.53846154)
           38) Age< 33.5 29  13 0 (0.55172414 0.44827586)
             76) Age>=28.5 8   2 0 (0.75000000 0.25000000) *
             77) Age< 28.5 21  10 1 (0.47619048 0.52380952)
              154) Age< 24.5 10   3 0 (0.70000000 0.30000000) *
              155) Age>=24.5 11   3 1 (0.27272727 0.72727273) *
           39) Age>=33.5 10   2 1 (0.20000000 0.80000000) *
      5) Age< 6.5 24   8 1 (0.33333333 0.66666667)
       10) family_size>=4.5 9   1 0 (0.88888889 0.11111111) *
       11) family_size< 4.5 15   0 1 (0.00000000 1.00000000) *
    3) Sex=female 314  81 1 (0.25796178 0.74203822)
      6) Pclass>=2.5 144  72 0 (0.50000000 0.50000000)
       12) family_size>=4.5 27   3 0 (0.88888889 0.11111111) *
       13) family_size< 4.5 117  48 1 (0.41025641 0.58974359)
         26) Age>=27.71591 25   9 0 (0.64000000 0.36000000) *
```

Fig. 4. Classification

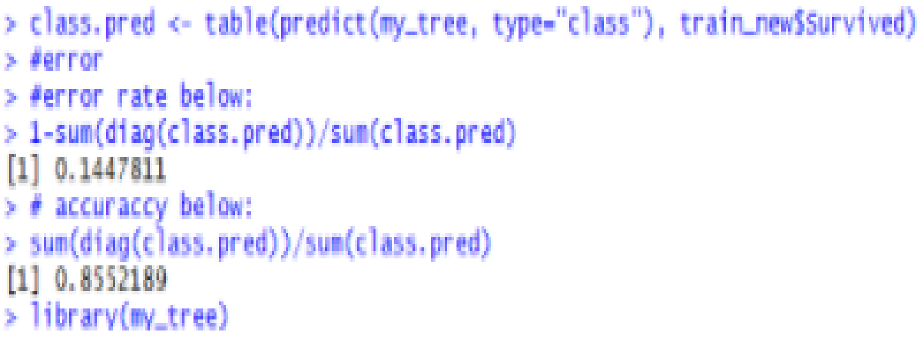


```
> summary(my_tree)
Call:
rpart(formula = Survived ~ Age + Sex + Pclass + family_size,
    data = train_new, method = "class", control = rpart.control(cp = 1e-04))
  n= 891

          CP nsplit rel error    xerror       xstd
1 0.444444444      0 1.0000000 1.0000000 0.04244576
2 0.030701754      1 0.5555556 0.5555556 0.03574957
3 0.023391813      3 0.4941520 0.5116959 0.03467453
4 0.020467836      4 0.4707602 0.5087719 0.03459945
5 0.005847953      6 0.4298246 0.4766082 0.03374384
6 0.004385965     11 0.3976608 0.4853801 0.03398272
7 0.002923977     13 0.3888889 0.4853801 0.03398272
8 0.000100000     17 0.3771930 0.4824561 0.03390356

Variable importance
        Sex         Age      Pclass family_size 
         49          19          18          13 

Node number 1: 891 observations,    complexity param=0.4444444
  predicted class=0  expected loss=0.3838384  P(node) =1
    class counts:   549   342
   probabilities: 0.616 0.384 
  left son=2 (577 obs) right son=3 (314 obs)
  Primary splits:
      Sex         splits as  RL,          improve=124.42630, (0 missing)
      Pclass      < 2.5      to the right, improve= 43.78183, (0 missing)
      family_size < 1.5      to the left,  improve= 17.43059, (0 missing)
      Age         < 6.5      to the right, improve= 10.05326, (0 missing)
```

Fig. 5. Summary of data

Decision tree can be generated using the prp function in R programming, where the type of decision has been selected as type 4. The decision tree helps to understand the survival status in a very user friendly way. At each level, the visualization is easier compared to reading decision tree summary. There is another way to generate more informative decision tree using "fancyRpartPlot" function in R programming. In this type of decision tree, the decision nodes, chance nodes and leaf nodes can be identified in different colors.

Overall accuracy of trained model for getting the survival status of passengers in the titanic disaster s 85.52 % as displayed in the Fig.6, which is more than that of the expected accuracy in our objective. Further in future work, we will be try to improve the accuracy of classification.



```
> class.pred <- table(predict(my_tree, type="class"), train_new$Survived)
> #error
> #error rate below:
> 1-sum(diag(class.pred))/sum(class.pred)
[1] 0.1447811
> # accuraccy below:
> sum(diag(class.pred))/sum(class.pred)
[1] 0.8552189
> library(my_tree)
```

Fig. 6. Classification accuracy

Clustering technique is used to group a set of observations in such a way that observations of similar types belong to a group [6]. There can be many groups defined while clustering. Clustering implementation has been performed using Python programming. Installation of numpy, pandas and scipy packages are required for its implementation in Python. KMeans algorithm will be implemented using KMeans library in sklearn. This will generate the survival predictions of all the passengers in the test data.

6. Conclusion

On training data set, modification has been performed to fill the missing values and other related errors for fulfilling the feature engineering technique. Then, the machine learning algorithms are applied for classification and clustering techniques. It can be concluded that, passengers travelled with small family size whose family size was from 2 to 4 survived more compared to death. The death expectation is more with the passengers who travelled alone and who travelled with large family size. Then, the classification with 2 classes has been performed using decision tree where survival got analyzed at every level. Also, the accuracy rate observed is 85%. Survival status was also analyzed using clustering techniques. Further as future work, Feature engineering has to be performed more accurately to gain more knowledge from the titanic data. Also, other classification and clustering algorithm can be applied to improve the accuracy value. The exploratory data analysis has been carried out on Titanic passenger data set in this work. The practical connotation will be extending or applying these algorithms on other real-world datasets like natural disaster data and covid-19 data set for further analysis and inference.

References

[1] Cicoria S, Sherlock J, Muniswamaiah M and Clarke L 2014 *Proc. of Student-Faculty Research Day (New York)* Classification of titanic passenger data and chances of surviving the disaster p 1-6

[2] Cortes C, Vapnik V 1995 *Support-vector Networks* (Machine Learning Vol 20) p 273–297

[3] Russell S J, Norvig P 2003 *Artificial Intelligence: A Modern Approach (*Prentice Hall) 2nd Ed

[4] Finlay S 2014 *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods* (New York: Palgrave Macmillan) 1st ed

[5] James G, Witten D, Hastie T and Tibshirani R 2013 *An Introduction to Statistical Learning with Applications in R* (New York: Springer) 1st ed

[6] Andrew Ng CS229 Notes 2012 (Stanford University)

[7] Jonathan W, Charlotta L and Renato U 2020 Automated machine learning: Review of the state-of-the-art and opportunities for healthcare (Artificial Intelligence in Medicine Vol 104 Elsevier)

[8] Guyon I, Elisseeff A 2006 *An Introduction to Feature Extraction* (Feature Extraction. Studies in Fuzziness and Soft Computing Vol 207 Springer, Berlin, Heidelberg)

[9] Heaton J *An empirical analysis of feature engineering for predictive modeling* 2016 Proc. SoutheastCon (Norfolk VA) pp. 1-6

[10] Khalid S, Khalil T and Nasreen S 2014 *A survey of feature selection and feature extraction techniques in machine learning* Proc. Int. Conf. on Science and Information (London) pp. 372 78

[11] Aparna U R., Paul S 2016 *Feature selection and extraction in data mining* Online Int. Conf. on Green Engineering and Technologies (Coimbatore) pp. 1-3

[12] Manal B, Liping Z 2019 *A review of machine learning algorithms for identification and classification of non-functional requirements* (Expert Systems with Applications: X Vol.1)

[13] Neeraj N, Shikha A and Sanjay S 2015 *Recent Advancement in Machine Learning Based Internet Traffic Classification (*Procedia Computer Science Vol. 60) pp. 784-791

[14] Youguo L, Haiyan W 2012 *A Clustering Method Based on K-Means Algorithm* (Physics Procedia, Vol. 25) pp. 1104-09

[15] Ahmed O, Fatima-Zahra B, Ayoub A L and Samir B 2018 *Big Data technologies: A survey* (Journal of King Saud University - Computer and Information Sciences Vol. 30 Iss. 4) pp 431-48