Data Mining Report

(INF6028)

Student No: 200226686

Word Count: 2202

Abstract

Background: The most infamous stories of a shipwreck include the RMS Titanic, which is known worldwide and also has a movie based on it. But although the Titanic tragedy occurred well over a century ago, scientists and statisticians are still interested in learning why some passengers escaped and others did not.

Aim: The main purpose of this study will be to use the Titanic Dataset provided to predict the survival of the onboard passengers using different supervised machine learning algorithms.

Methods: Predicting the survival of onboard passengers of the titanic using supervised machine learning algorithms such as Random Forest, Gradient Boosted Tree, and Naïve Bayes with the help of different categorical and numerical variables present in the dataset. KNIME Data Analytics Software will be used to carry out the necessary operations for this study.

Results: Results from this study indicated that predictor variables like “AgeCategory”, “Sex”, “Fare”, “Embarked”, and “Ticket Class” contributed the most in predicting the survival of the passengers. And hence were the most important variables.

Conclusion: It was found that Gradient Boosted Tree performed the best among the three selected ML algorithms with an accuracy of 80.24% based on the confusion matrix. The same data also resulted in concluding that the model-based on Gradient Boosted algorithm was more accurate according to the ROC curve diagram.

Table of Contents

List of Figures

List of Tables

1. Introduction

Titanic's sinking was one of the deadliest maritime disasters in history. Even though luck played its part, it appears that certain people were more likely to succeed than others. To find valuable insights from this catastrophe, Supervised Machine Learning(ML) techniques were used to create a variety of predictive models to work out the variables that have a significant effect on survival, and we compared models to improve the accuracy of predicting the survival of passengers onboard.

Data used in this study is split into two files namely; “titanic_ticket_data .csv” and “titanic_personal_data.csv”. These two datasets are derived from the Kaggle website which hosts competitions. The variables present in the two datasets are as follows:

Table 1 titanic_personal_data.csv

Variable	Definition
PassengerId	Unique ID for passenger
Name	Passenger Name
Sex	Gender
Age	Age
SibSp	number of siblings/spouses where family relations are defined as follows: Sibling = brother, sister, stepbrother, stepsister ; Spouse = husband, wife
Parch	number of parent/children where family relations are defined as follows: Parent = mother, father; Child = daughter, son, stepdaughter, stepson. Some children traveled only with a nanny, therefore parch=0 for them
Salary	
Job	Job title

Table 2 titanic_ticket_data.csv

Variable	Definition	Key
PassengerId	Unique ID for passenger	
Survived	Survival	0= not survived 1=survived
Ticket	Ticket no	
Fare	Ticket price	
Cabin	Cabin Number	
Embarked	Port of Embarkation	C= Cherbourg, Q = Queenstown, S = Southampton

The main aim of this study is to predict the survival (0 or 1) of each passenger on-board the titanic based on the categorical and numerical variables present in the two datasets combined. The "titanic_personal_data.csv" contains 1269 entries and "titanic_ticket_data.csv" contains 1243 entries before applying any data pre-processing methods. The aim is to use exploratory data analytics to gain information from the dataset available and to determine the impact of each variable on passenger survival by using analytics between each column of the dataset and the "Survival" field.

Predictive analysis is a way of determining significant and valuable trends in vast amounts of data using analytical tools (Larose et. al., 2015). Survival is calculated using ML algorithms based on various combinations of functions. To build a predictive model, Supervised ML techniques have been implemented using the available data. ML algorithms like Random Forest, Gradient boosted tree and Naïve Bayes are used and compared to achieve the highest level of accuracy for the survival of passengers.

2. Literature Review

2.1 Predictive Analytics

The method of collecting information from massive data sets to make assumptions and projections about possible events is known as predictive analytics (Larose et. al., 2015). Many big firms have been collecting customer data since long ago which has resulted in an enormous amount of data being stored in their databases. To put this data to good use firms started using it to improve estimates, efficiency, and forecast decisions (Abbott, 2014). The best possible way to handle a large amount of data and gain useful insights from it was by using predictive analytics. With the help of Pattern recognition, Statistics, ML, AI, and data processing the basis for predictive analytics was developed (Abbott, 2014).

2.2 Supervised Machine Learning Algorithms

Predictive analytics algorithms are of two categories: Supervised or Unsupervised learning methods. But in this study only Supervised learning methods will be used.

Using the variables or columns in the dataset, supervised learning models attempt to predict a target variable, which is represented by a single class column in the dataset (Singh et. al., 2016). Predictive modelling is another name for supervised learning.

2.2.1 Random Forest

Random forest algorithm is similar to that of Decision tree but it has its own advantages over decision tree. With the data we have, decision trees would work well for Supervised ML because they are not susceptible to feature size and can accommodate both quantitative and qualitative attributes (Jha, 2017). But Random forests generate decision trees from randomly chosen data columns, receives predictions from each tree, and votes on the best solution with higher accuracy. To prevent overfitting problems in the decision tree, a random forest model was selected for classification in this study. The biggest drawback of random forest is that it might become too sluggish and useless for real-time prediction if there are too many trees.

Random Forest algorithms were used for groundwater potential mapping as one of the predictive ML algorithm (Naghibi, et. al., 2017).

2.2.2 Gradient Boosted Tree

The “Gradient Boosting” classifier will produce a large number of weak, shallow predictor trees, which it will then merge, or boost into a powerful model. This model performs excellently on the data, but it has the disadvantage of being inefficient and difficult to refine, as the model is built sequentially and thus cannot be parallelized (Srivastava, 2015).

2.2.3 Naïve Bayes

“The Naive Bayes classifier is a simple probabilistic classifier that uses the Bayes theorem and strict (naive) independence assumptions. This method implies that input variables are independent of one another in terms of the output (class) vector” (Nettleton, 2014). This assumption is its advantage as well as disadvantage.

“Naive Bayes is a fast-learning algorithm. As a result, it may be used to make real-time forecasts. This algorithm is also known for its ability to forecast several classes” (Lowd, et. al., 2007). We can estimate the likelihood of several target variable groups here.

2.3 Evaluation of applied models

"The main aim of analysing classification models is to determine how closely the model's predicted classifications correspond to the case's actual classification. Depending on the observation process, there are many ways to assess the model's effectiveness" (Novaković, et. al., 2017).

For this study, as the main purpose of developing a predictive model here is to correctly identify whether a passenger onboard the titanic survives the sinking or not; the accuracy of the prediction matters the most. Hence, the most appropriate method to evaluate the performance of a model would be checking how the model performs based on the accuracy of the predicted outcomes.

$$\text{Accuracy} = \frac{\text{number of correctly classified examples}}{\text{total number of cases}}$$

3. Data Preparation

3.1 Overview of the Titanic Dataset

The Titanic dataset consists of 10 predictor variables which including sex, Job, and embarked as categorical variables, and Age, SibSp, Parch, Ticket, Fare, Cabin, and Salary as numeric variables. The outcome variable is "Survived", which is a binary variable with 0 (not survived) and 1 (survived) and other variables like PassengerId and Name.

3.2 Data Pre-processing

As the Titanic dataset is split into two different files with no direct link as specified in the problem statement, a suitable logic was required to be established between the two datasets to successfully merge them without any redundancy.

With the help of exploratory data analysis, it was discovered that passengers on-board the titanic with families were assigned the same ticket number which helped in linking the data of families from both the datasets. And for the remaining data was eventually related to all the passengers traveling as an individual.

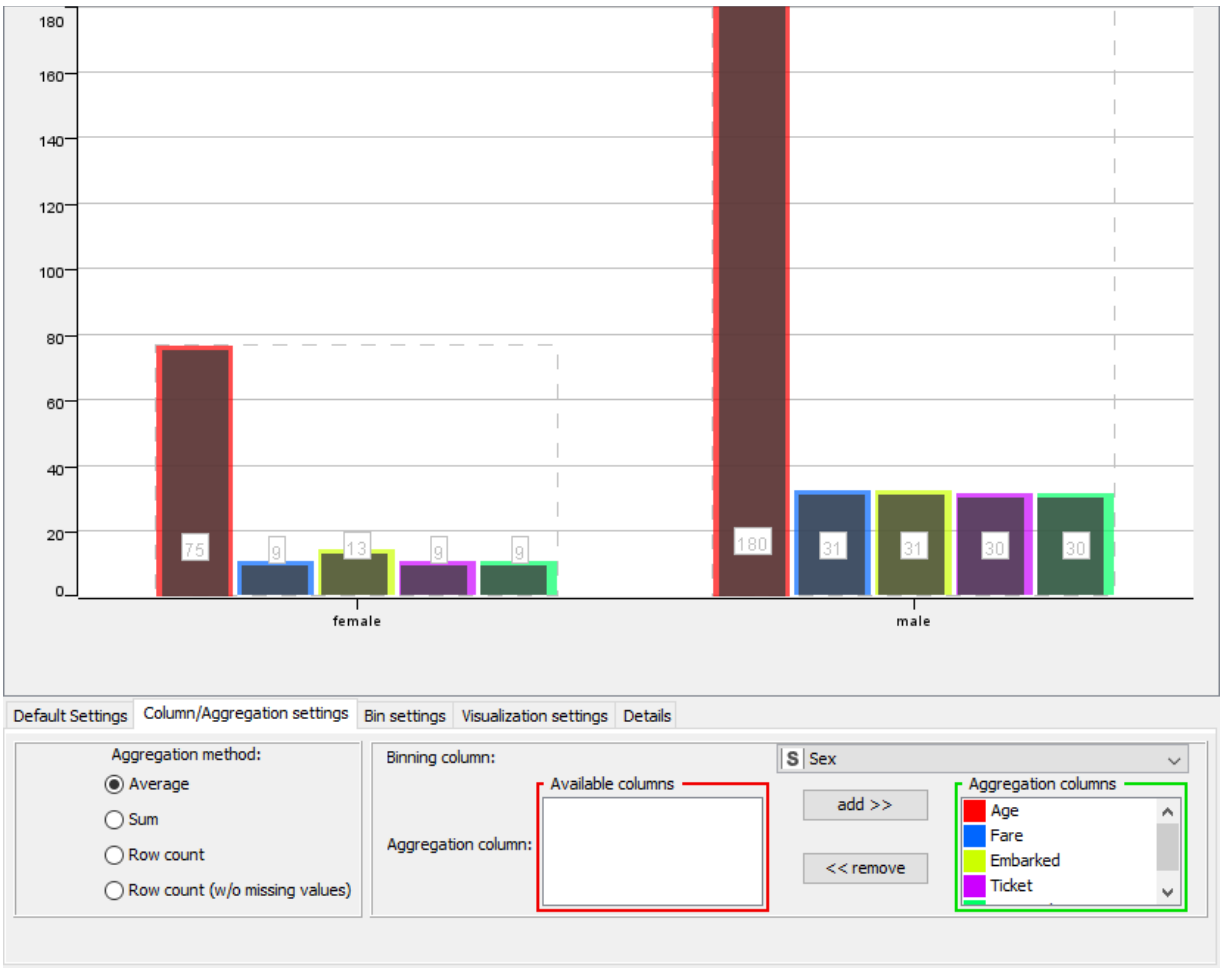
From the ticket data, to find the duplicate ticket numbers, the “Duplicate Row Filter” node was used without filtering out the duplicate rows and creating a new column assigning a value “Unique”, “Chosen” and “Duplicate” to each row. And then, rows with “chosen” and “duplicate” values were filtered using the “Row Filter” node. Then each ticket number was grouped and assigned a unique number (Group) using the "Column Expression" node and were ordered in descending manner.

From the personal data, to find the families "Column Aggregator" node was applied to find the total family members accompanying each passenger on-board by adding the "SibSp" and "Parch" columns and the passenger itself from the data. This was a part of the feature engineering process. Then, passengers with at least one family member were filtered, and using the "Cell Splitter" node the family name was extracted to group all family members, and each family name was assigned a unique value (Group) and ordered in descending manner. Using the numbers in the "Group" column both the tables were joined using the "Joiner" node.

A similar process was carried out for joining data of individuals on-board the titanic. And finally, the table for families and individual passengers was merged with the "Concatenate" node.

Handling the missing values in the final dataset was the first operation implemented to have robust data for applying the ML model later in the study. According to "Figure 1", the dataset consisted of a large number of missing values for the "Age" column. This was handled by taking the median of all values in the age column and assigning it in place of rows with a missing value. For the "Embarkment" column a mode of all values was used. Rows with missing “Ticket number” and “Fare” were removed from the data.

Figure 1 Missing Value Statistics



3.3 Feature Engineering

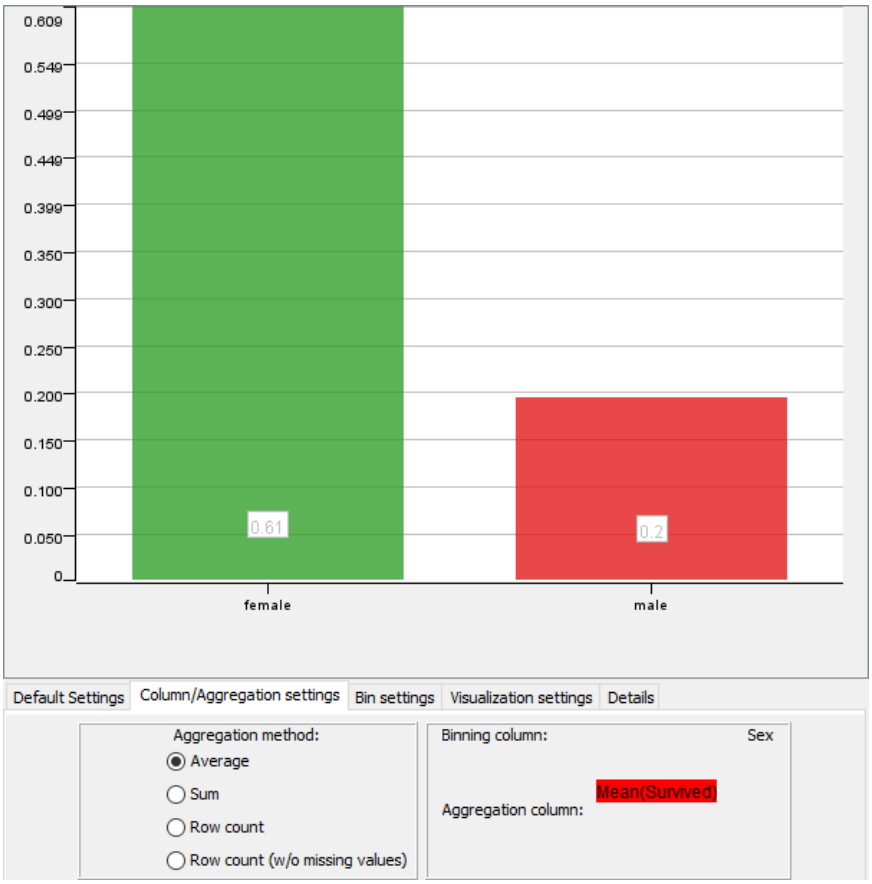
As mentioned in the previous section, that a new variable “no_of_family_members” was created by adding the family members assigned in the “SibSp” and “Parch” columns. Then we assigned age categories to each passenger based on their age groups with the help of the "Rule Engine" node. And finally, based on "Salary" column, the class for each passenger was assigned assuming they could afford the fare of the ticket.

3.4 Exploratory Data Analysis

3.4.1 Survival v/s Sex

From “Figure 2” it can be concluded that women onboard the titanic were more likely to survive the sinking compared to men. The average survival rate for women was just above 0.61 whereas for men it was around 0.2.

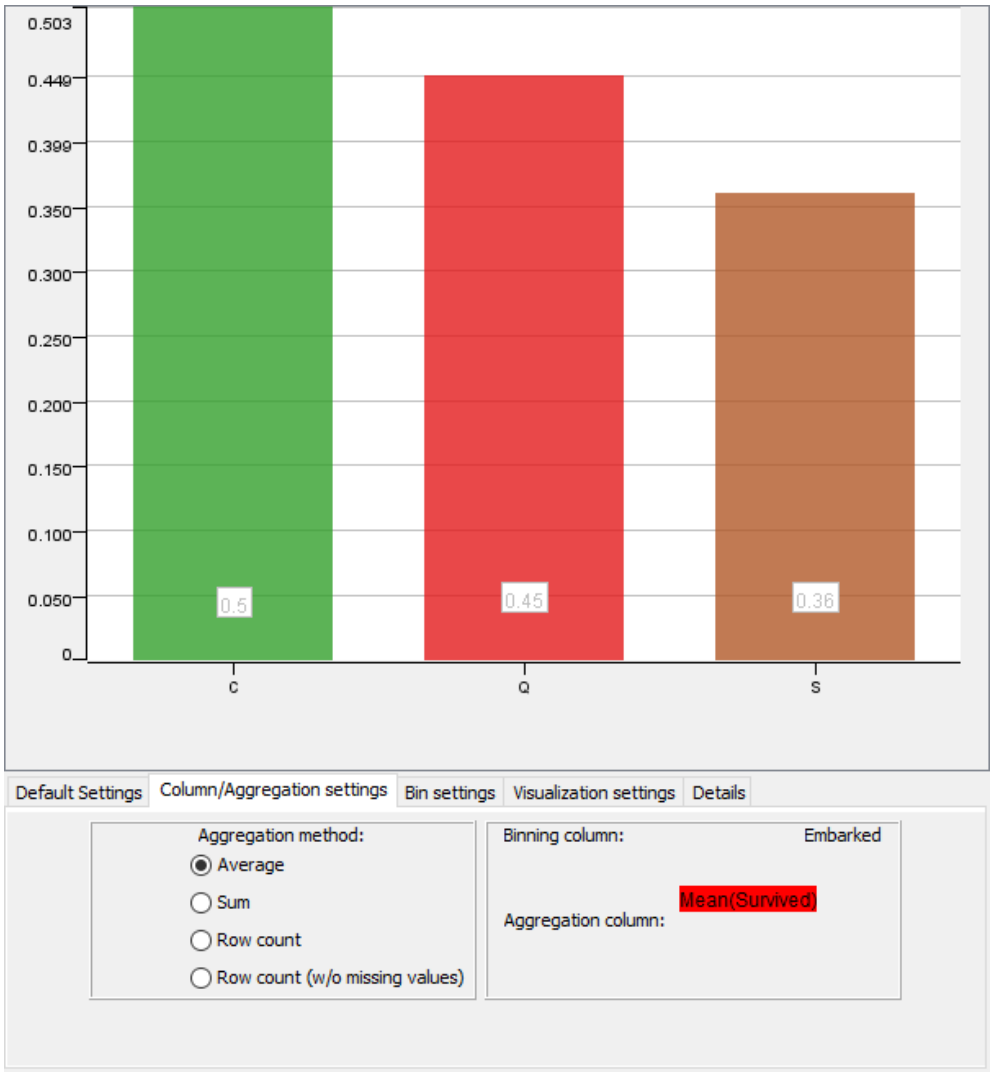
Figure 2 Survival v/v Sex



3.4.2 Survival v/s Embarkment

From “Figure 3”, it can be concluded that people boarding the Cherbourg (C) port were more likely to survive the sinking as compared to the other two ports, Queenstown (Q) and Southampton (S).

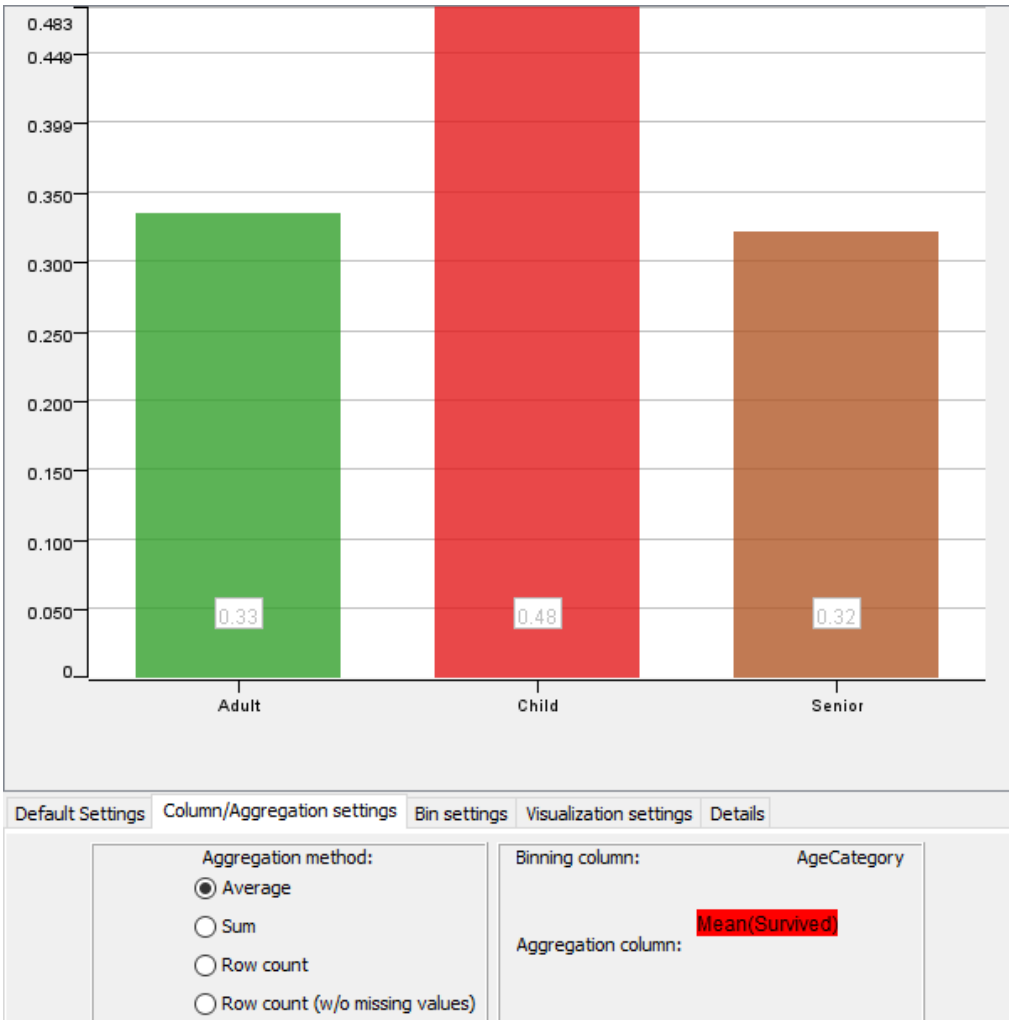
Figure 3 Survival v/s Embarkment Port



3.4.3 Survival v/s Age category

From “Figure 4”, it can be concluded that children onboard the titanic had a higher chance of survival than people in other age categories i.e. Adults or Seniors.

Figure 4 Survival v/s Age Category



3.4.4 Ticket class Distribution and Average Survival

"Figure 5" shows the distribution of passengers based on the class of the ticket they hold. And, "Figure 6" shows the probability of survival of the passengers from each class. From "Figure 6", it can be concluded that passengers from 1st class had the highest chances of survival following with the 2nd class and then the 3rd class passengers.

Figure 5 Class-wise Distribution (Ticket)

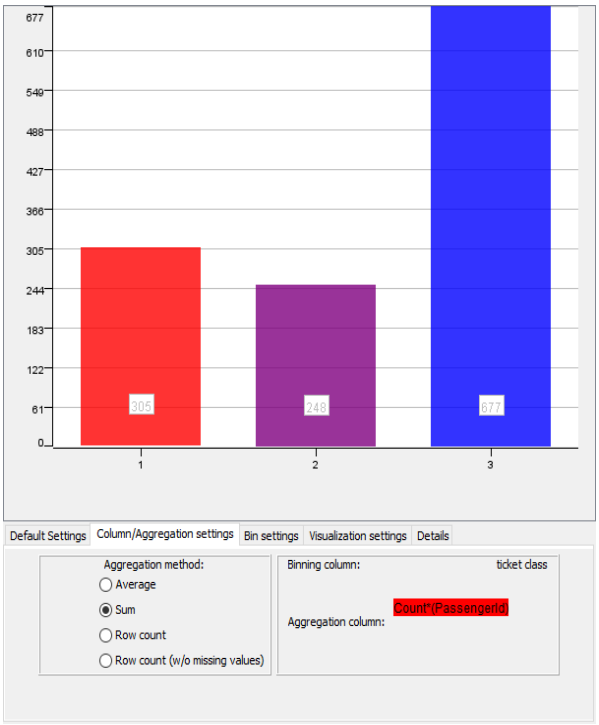
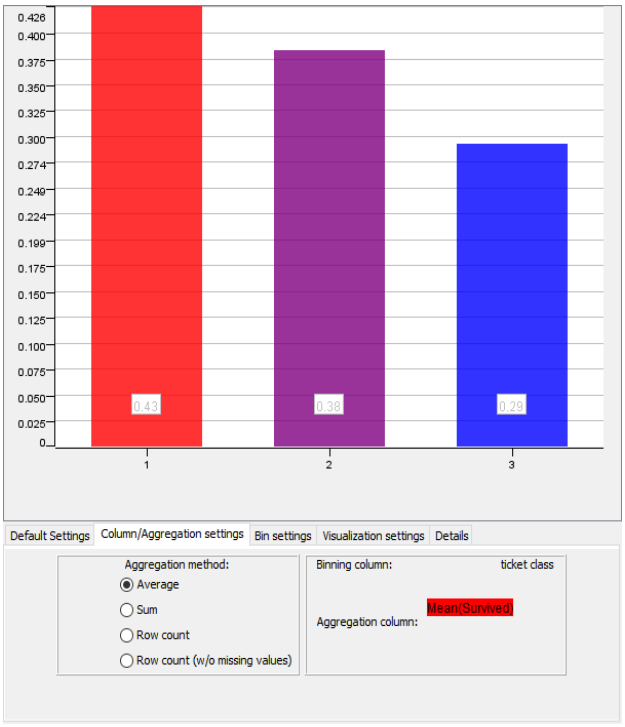


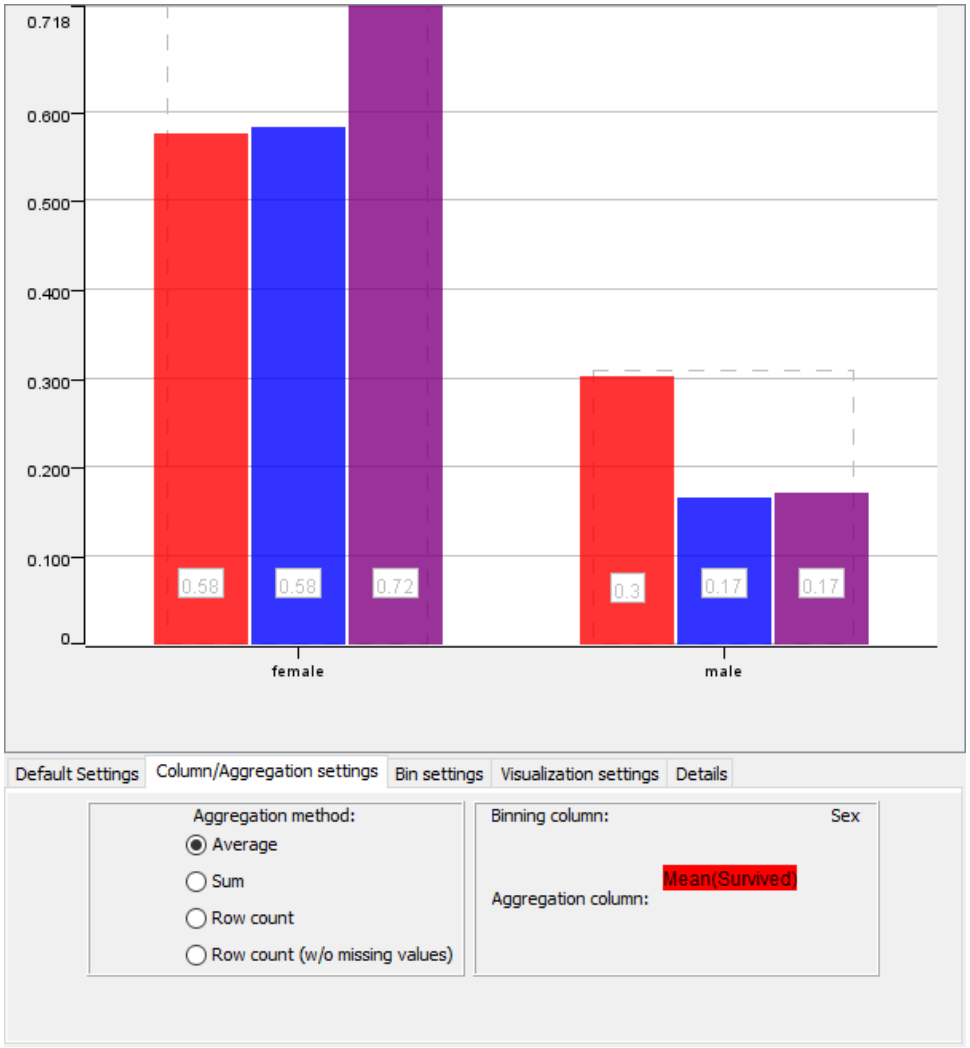
Figure 6 Survival vs Ticket Class



3.4.5 Survival v/s Ticket class and Gender

In “Figure 7”, the red, blue, and purple bar represents the “1st class”, “2nd class”, and “3rd class” ticket holders respectively. This graph shows the probability of survival based on gender in each class

Figure 7 Survival vs Ticket class and Gender



4. Experimental Setup

As our objective is to estimate whether the passengers on the Titanic will survive the sinking, we use the Supervised ML methods to solve this problem of classification. A predictive modelling task in which a class label is predicted for a given example of input data is referred to as classification (Kotsiantis, et.al., 2007). To develop a ML model, we use Knime Data Analytics Software. ML algorithms like Random Forest, Gradient Boosted Tree, and Naïve Bayes were used in this study.

In this study, to implement various ML algorithms, firstly, as there are many distinct values for predictor columns, we disable the domain restrictor using the "Domain Calculator" node, and then the dataset was split into two parts using the "X-partitioner" node i.e. by using the cross-tabulation technique into 5 parts stratified based on "Ticket class". "To overcome the problem of overfitting and avoiding selection bias we use cross-validation" (Berrar, 2019).

In Knime, for each ML algorithm, there is a learner and a predictor node. The learner node takes training data as an input and creates a predictor model based on it. And then, the test data is given as an input to the predictor node along with the model developed by the learner node to predict the value (Survival) for the test data. The variables like "Sex", "AgeCategory", "Fare", "Embarked", and "Ticket Class" were found to be good predictor variables. The learner nodes were configured to use those variables to develop a model. To keep uniformity in all the algorithms, the same variables were used in the learner node.

To merge the data from both training and test data "X-Aggregator" node is used with the target column set to "Survived" and the prediction column set to a newly generated column by the predictor node "Prediction (Survived)".

To find the best-performing model, the confusion matrix was evaluated. The confusion matrix was obtained using the "Scorer" node which compares the actual value of the prediction column against the predicted column to give the accuracy percent. As this is a classification problem, and the correct identification of prediction (survival) is the most important thing, the best-performing model was evaluated on the basis of the highest accuracy percentage.

To further compare the models, a "ROC curve" node is used to plot the probability of accuracy of predicting each passenger's survival. This is done by joining data of all positive class probabilities from all the predictor nodes into one table with the actual "Survived" class column.

5. Results and Discussion

5.1 Model Accuracy Comparison

Table 3 Model Comparison (Accuracy)

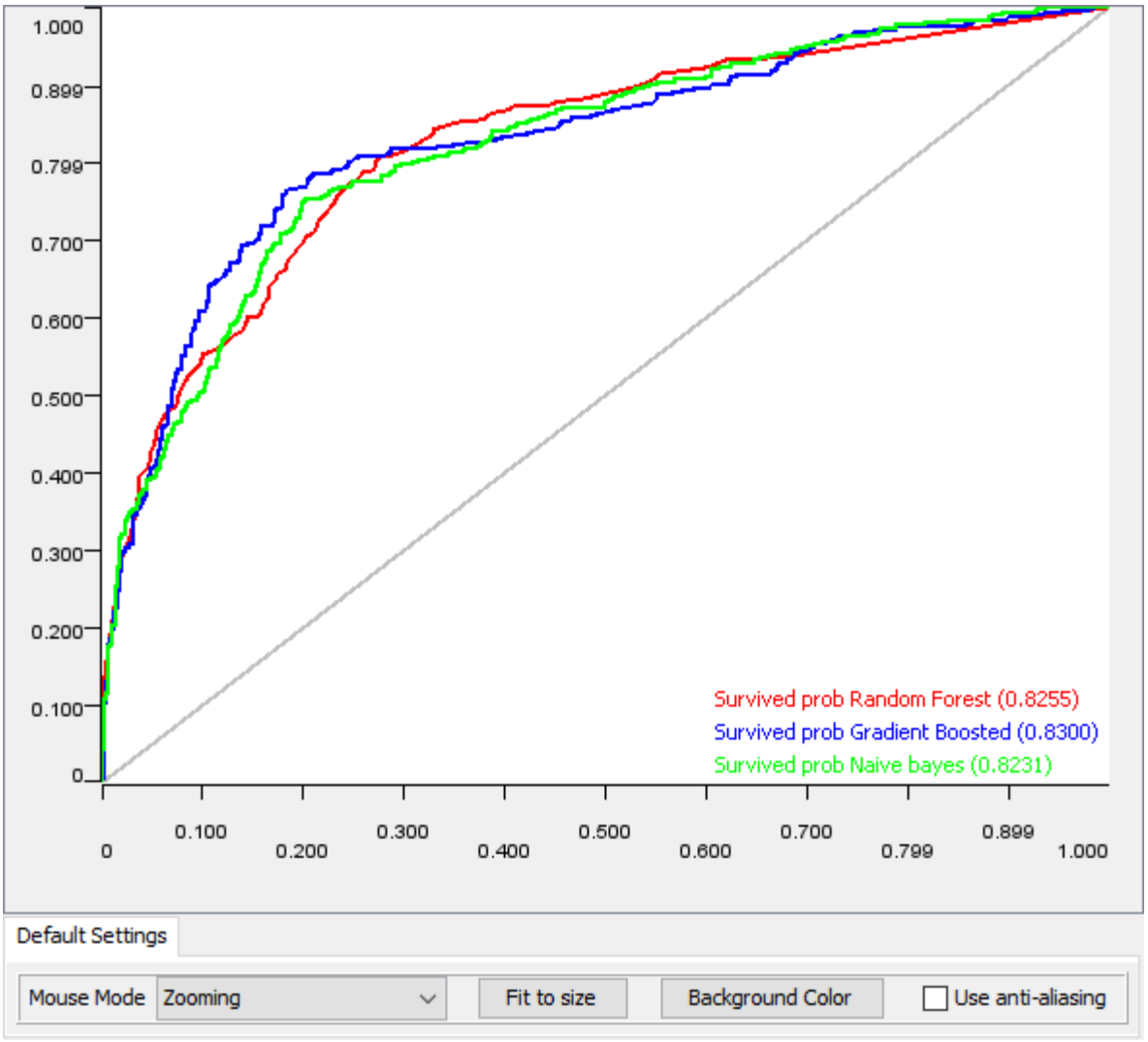
ML Algorithms model	Accuracy (%) for training data	Accuracy (%) for test data	Accuracy (%) for train+ test data	Cohen's Kappa (k) (final)
Random Forest	92.69%	76.42%	76.50%	0.499
Gradient Boosted tree	93.09%	81.71%	80.24%	0.544
Naïve Bayes	98.68%	79.27%	77.40%	0.489

"Table 3" presents the accuracy (%) for training, test, and both together of each ML algorithms used in this study and Cohen's Kappa Coefficient. It can be concluded that the Gradient Boosted Tree algorithm was comparatively more accurate than the other two models. This model is generally used with smaller datasets as it easy to implement and has a high computation speed and gives a good accuracy result (Srivastava, 2015).

5.2 ROC Curve Comparison

From “Figure 7”, we understand that the performance of Gradient Boosted is Slightly better than that of the other two models according to the outcomes of the true probabilities. As the value for Gradient Boosted is much closer to “1”, it predicts more accurately as per the ROC curve plot.

Figure 7 ROC Curve



6. Conclusion

The main purpose of this study was to predict the survival of the passenger on-board the Titanic. And by using various ML algorithms this has been achieved to a satisfactory level. The highest accuracy achieved was 80.24% by using the Gradient Boosted Tree.

From this study, it can be concluded that feature engineering plays an important role so as to get a more efficient prediction from the models. Future work that may be applied to this topic comprises understanding how to use multiple sophisticated algorithms in an attempt to create a superior solution and expand our knowledge of predictive analytics. Also, it will be advantageous to further work on refining and modifying variables to gain additional information from the available data.

7. References

Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. ProQuest Ebook Central https://ebookcentral.proquest.com

Berrar, D. (2019). Cross-validation. Encyclopedia of bioinformatics and computational biology, 1, 542-545.

https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/#:~:text=Naive%20Bayes%20uses%20a%20similar,with%20problems%20having%20multiple%20classes.

Jha, Vishakha (2017). Decision Tree Algorithm for a Predictive Model. retrieved on 21st May 2021 from https://www.techleer.com/articles/120-decision-tree-algorithm-for-a-predictive-model/#:~:text=The%20decision%20tree%20is%20an,outcomes%20based%20on%20certain%20conditions.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160(1), 3-24.

Larose, D.T., & Larose, C.D. (2015). *Data mining and predictive analytics*. ProQuest Ebook Central. https://ebookcentral.proquest.com

Lowd, D., & Domingos, P. (2005, August). Naive Bayes models for probability estimation. In Proceedings of the 22nd international conference on Machine learning (pp. 529-536).

Naghibi, Seyed Amir, Ahmadi, Kourosh, & Daneshi, Alireza. (2017). Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping. Water Resources Management, 31(9), 2761–2775. https://doi.org/10.1007/s11269-017-1660-3

Nettleton, D. (2014). Commercial data mining: Processing, analysis, and modeling for predictive analytics projects. ProQuest Ebook Central https://ebookcentral.proquest.com

Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., & Milica, T. (2017). Evaluation of classification models in machine learning. Theory and Applications of Mathematics & Computer Science, 7(1), 39-46.

Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1310-1315). Ieee.

Srivastava, T., (2015). Learn Gradient Boosting Algorithm for better predictions (with codes in R). retrieved on 05th June 2021 , from https://www.analyticsvidhya.com/blog/2015/09/complete-guide-boosting-methods/