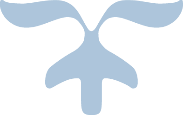


Titanic Dataset

CST8390\_23W Assignment 2



Algonquin College

Computer Engineering Technology

Business Intelligence and Data Analytics

Hoang, John – 040896360

Kingdom, Neil – 040967309

Prepared for:

Dr. Anu Thomas

Table of Contents

[Table of Figures 2](#_Toc128522330)

[1.0 Introduction 3](#_Toc128522331)

[2.0 Data Understanding 3](#_Toc128522332)

[2.2 Dataset Breakdown 4](#_Toc128522333)

[3.0 Data Preparation 6](#_Toc128522334)

[3.1 Data Selection 6](#_Toc128522335)

[3.2 Cleaning Data 7](#_Toc128522336)

[3.3 Constructing Data 7](#_Toc128522337)

[3.4 Integrating Data 10](#_Toc128522338)

[3.5 Formatting Data 10](#_Toc128522339)

[4.0 Modeling and Evaluation 10](#_Toc128522340)

[5.0 Discussion of Results 12](#_Toc128522341)

[6.0 Conclusion 14](#_Toc128522342)

[7.0 References 15](#_Toc128522343)

## Table of Figures

[Figure 1 Initial Titanic dataset Attributes 4](#_Toc128522285)

[Figure 2 Initial Titanic dataset Class 4](#_Toc128522286)

[Figure 3 Titanic Train set - Class and Potential Categorical Attributes 5](#_Toc128522287)

[Figure 4 Titanic Test set - Class and Potential Categorical Attributes 6](#_Toc128522288)

[Figure 5 Test Set Equal Frequency (Five bins) 9](#_Toc128522289)

[Figure 6 Test Set Equal Width (Five Bins) 9](#_Toc128522290)

[Figure 7 Updated Train Set Attributes 9](#_Toc128522291)

[Figure 8 Updated Test Set Attributes 10](#_Toc128522292)

[Figure 9 Classification Chart 11](#_Toc128522293)

[Figure 10 Confusion Matrix on 10-fold J48 default settings using Updated Train Set 11](#_Toc128522294)

[Figure 11 Decision Tree on 10-fold J48 default settings using Updated Train Set 11](#_Toc128522295)

[Figure 12 Predicted Class from Test Set 12](#_Toc128522296)

[Figure 13 Titanic Facts 12](#_Toc128522297)

[Figure 14 Inferred Facts 13](#_Toc128522298)

# 1.0 Introduction

In the early 1900’s, the Titanic was considered to be the largest vessel that had ever been afloat. Out of the estimated 2,224 passengers that boarded the ship, over 1,500 perished, making it the deadliest sea voyage up until that time 1. Kaggle, an online web platform hosted by Google for Data Scientists, released a competition in 2014 to analyse a dataset created from the known information of real-life passengers who boarded the Titanic. The objective of the competition is to use the given dataset to make a prediction regarding the number of passengers that would have survived or perished based upon the given “features” (descriptive classifications pertaining to a specific passenger) of each instance in the dataset. In order to accomplish this, a predictive model is built to determine the outcome of unknown survival classifications. In order to build a predictive model, the provided training set (which contains the known survival status of each passenger i.e., the “ground truth”) is analysed by a Machine Learning tool for correlations between the features of each passenger and their survival status. Once the predictive model is created, its accuracy is determined by attempting to predict the survival classification of each passenger in the given test set. The approach taken in this research paper was to use a supervised learning algorithm for classification using the J48 decision tree classifier in the Weka ML tool (created by the University of Waikato in New Zealand). A decision tree can be summarized as a learning model that is capable of determining the most probable classification for a given instance in the dataset by comparing it with pre-determined rules based upon the features of other instances in the test set. The ID3 (Iterative Dichotomiser) algorithm is used to divide the features into groups and select the determinant group with the best features by the highest Information Gain. On completing the goal, Weka’s J48 classifier is used to produce the underlying decision tree algorithm2. Once the predictive model is built, the next step is to supply the training set with the test set and run the classifier. The objective is to obtain the most accurate prediction possible based upon the results of the constructed predictive model3.

# 2.0 Data Understanding

The Titanic training dataset consist of a total of 889 instances and the test set contains 418 instances. The training set has 1 class labeled “Survived”. Both datasets share the same 11 distinct features. These 11 features are described as follows:

* Passenger Id (#1-889): The identification number of the passenger.
* Passenger Class (abbreviated as PClass): Passengers were divided into three separate classes base upon the price of their boarding ticket. The labels are abbreviated for simplicity and listed as follows: 1 represents “First Class”, 2 represents “Standard Class” and 3 represents “Third Class”.
* Name: The first and last name of each passenger. This feature may also include the title of the passenger. Titles include: Mr. ,Ms. , Miss., Mlle., Mme., Mrs., Master, Capt., Major, Col., Lady, and Rev.
* Gender: The passenger’s biological sex.
* Age: The passenger’s age in years.
* Siblings and Spouse (abbreviated as SibSp): The sum of a passenger’s siblings and/or spouse aboard the Titanic.
* Parent and Child (abbreviated as Parch): The sum of a passenger’s children and/or parent(s) aboard the Titanic.
* Ticket: The ticket number belonging to a passenger.
* Fare: The cost of a passenger’s ticket.
* Cabin: The room that the passenger vacated.
* Embarked: The port from where the passenger boarded the Titanic4. The labels are abbreviated for simplicity and listed as follows: C = Cherbourg, Q = Queenstown, S = Southampton.

|  |  |
| --- | --- |
| **Attribute/Feature** | **Type** |
| Passenger Id | Numeric Discrete |
| PClass | Categorical Nominal |
| Sex | Categorical Nominal |
| Name | String |
| Age | Numeric Continuous |
| Sibsp | Numeric Discrete |
| Parch | Numeric Discrete |
| Fare | Numeric Continuous |
| Embarked | Categorical Nominal |
| Cabin | String |
| Ticket | Numeric Discrete |

Figure 1 Initial Titanic dataset Attributes

|  |  |
| --- | --- |
| **Class** | **Type** |
| Survived | Class |

Figure 2 Initial Titanic dataset Class

## 2.2 Dataset Breakdown

A list of classes within the training dataset are described in detail below:

* 549 instances of non-survivors and 340 survivors.
* 214 First class passengers, 184 Standard class passengers and 491 Third class passengers.
* 312 female passengers and 577 male passengers.
* 167 passengers embarked at port C, 644 passengers embarked at port S and 78 passengers embarked at port Q.
* 606 passengers boarded without a spouse or any siblings, 237 boarded with 2 sibling(s) and/or a spouse, 16 boarded with 3 sibling(s) and/or a spouse, 18 boarded with 4 sibling(s) and/or a spouse, 5 boarded with 5 sibling(s) and/or a spouse and 7 boarded with 8 sibling(s) and/or a spouse.
* 676 passengers boarded without any parents or children, 118 boarded with parent(s) and/or children, 80 boarded with parent(s) and/or 2 children, 5 boarded with 3 parent(s) and/or children, 4 boarded with 4 parent(s) and/or children, 5 boarded with 5 parent(s) and/or children, and 1 boarded with 6 parent(s) and/or children.

In the training set there are 177 instances (20%) whose age is unknown and 687 (77.3%) whose Cabin number is unknown. In the test set, there are 86 instances (20.8%) whose age is unknown and 1 instance with a missing fare value.

|  |  |  |
| --- | --- | --- |
| **Class** | | **Count** |
| Survivor | Perished (0) | 549 |
| Survived (1) | 340 |
| **Categorical Attributes** | | **Count** |
| Sex | Male (1) | 577 |
| Female (2) | 312 |
| PClass | 1 | 214 |
| 2 | 184 |
| 3 | 491 |
| Embarked | C | 167 |
| S | 644 |
| Q | 78 |
| **Potential Numeric to Categorical Conversion** | | |
| Siblings and or Spouse | 0 | 606 |
| 2 | 237 |
| 3 | 16 |
| 4 | 18 |
| 5 | 5 |
| 8 | 7 |
| Parent and or Children | 0 | 676 |
| 1 | 118 |
| 2 | 80 |
| 3 | 5 |
| 4 | 4 |
| 5 | 5 |
| 6 | 1 |
| Age | 0.42-80 | 712 |
| Missing | 177 |
| Fare | 0-512.32 | 889 |

Figure 3 Titanic Train set - Class and Potential Categorical Attributes

|  |  |  |
| --- | --- | --- |
| **Categorical Attributes** | | **Count** |
| Sex | Male (1) | 266 |
| Female (2) | 152 |
| PClass | 1 | 107 |
| 2 | 93 |
| 3 | 218 |
| Embarked | C | 102 |
| S | 270 |
| Q | 46 |
| **Potential Numeric to Categorical Conversion** | | **Count** |
| Siblings and or Spouse | 0 | 283 |
| 1 | 110 |
| 2 | 14 |
| 3 | 4 |
| 4 | 4 |
| 5 | 1 |
| 8 | 2 |
| Parent and or Children | 0 | 324 |
| 1 | 52 |
| 2 | 33 |
| 3 | 3 |
| 4 | 2 |
| 5 | 1 |
| 6 | 1 |
| 9 | 2 |
| Age | 0-76 | 332 |
| Missing | 86 |
| Fare | 0-512.32 | 418 |

Figure 4 Titanic Test set - Class and Potential Categorical Attributes

# 3.0 Data Preparation

Following the guidelines of the CRISP-DM standard, both the test and training sets are analysed and modified. Data preparation is split into five stages: data selection, data cleaning, data construction, data integration, and data formatting.

## 3.1 Data Selection

The purpose of this report is to determine the most likely outcome in regard to survival of a given passenger by consideration of the particular features that the passenger possesses. With that said, not all features given in the dataset provided have a correlation with survival outcome. Out of the list of features, careful consideration is required when selecting the ones that pertain to survival and which do not.

The list of relevant features and the reasoning for their inclusion is as follows:

* Age: A passengers age would have had a heavy influence upon their physical health and modality. Weak or helpless passengers, such as the elderly and infants respectively, would have been prioritized for rescue.
* Passenger Class: It is likely that First Class passengers would have had quicker access to lifeboats, and due to their social status, there may have been a bias towards rescuing them prior to passengers in lower classes.
* Sex: In the same vain as age, women would have been deemed more helpless and less physically capable than men, meaning that their survival would have been prioritized above the survival of males.
* Relatives (sibling(s), spouse, parent(s), and/or children): The more relatives that a passenger had, the more likely they would have been to survive, as they would have been able to help one another. On the contrary, passengers without relatives would have had to fend for themselves, and may have been more likely to sacrifice themselves courageously.
* Fare: The fare price has a causal tie with passenger class, and thus, it would have affected a passenger’s chance of survival for the same reasons mentioned for the passenger class feature.
* Embarked: Observing the data in regard to the embarked feature reveals that the vast majority of passengers boarded the ship from Southampton, followed by Cherbourg and finally, Queenstown. Therefore, the likelihood of survival could be impacted by the port from which a passenger boarded.

## 3.2 Cleaning Data

As mentioned in the previous section, data that does not correlate with the survival of a given passenger ought to be removed. Passenger Id is included in the dataset for identification purposes, but no conclusions about the survival of a given passenger can be drawn from this feature, so it is removed. Likewise, the name of a passenger has no bearing upon the likelihood of survival, so it too is removed. The ticket number, if deciphered, may have some correlation with survival status, however, due to the fact that fare price is provided in the dataset, which provides a better metric for determining survival status based upon the ticket that was purchased by an individual passenger, warranting the inclusion of ticket number cannot be justified, hence its removal. Finally, due to the fact that the number of missing cabin numbers is so great, it cannot be included in the final dataset. To recap, passenger Id, name, ticket number, and cabin number are all discarded.

## 3.3 Constructing Data

In order to construct new data, a method known as Feature Engineering is used. Existing data may be classified into groups and placed into a new feature, or new data may be interpolated from the existing data using any number of methods.

Age is an important factor for determining the survival status of a given passenger, so removing it is undesirable. That being said, there are two problems that must be faced: the first is the substantial amount of missing data for this particular feature, and the second is the large range of possible age values. These issues can be resolved simultaneously by categorizing the data into groups. A new feature is created – AgeGroup. Age is categorized into six groups: Ages 2 and under are considered to be infants (the label used to represent this group is “Baby”), ages 3-11 are considered to be children, ages 12-24 are considered to be youth, ages 25-60 are considered to be adults, ages 60 and above are considered to be seniors, and finally, ages which are unknown are marked with the label “NK”, indicating that the age is “Not Known”. Once the age group feature is complete, the age feature is removed so that it does not obstruct or skew the final results.

The sibling/spouse and parent/child features can be combined and categorized similar to what was done for age. A new feature labeled “Relatives” is created. In order to determine the number of relatives of an individual passenger, the sum of sibling(s), spouse, parent(s), and/or children is taken. Once the quantity is known, it is categorized into the following groups: 0 relatives is marked as “None”, 1-2 relatives is marked as “Few”, and 3 or more relatives is marked as “Many”. For the same reasons aforementioned, both the SibSp and Parch features are removed afterwards.

Since the Fare is a continuous type it is possible to perform classification using a method known as binning. Before binning can be applied, however, it must be considered that one passenger in the test dataset (passenger with ID 1044) is missing a value for fare. Rather than remove the passenger from dataset completely, an approximate value can be estimated by taking the average fare rate for all passengers within the same class (Third Class in this case). In order to do this, passenger 1044 is temporarily removed from the dataset. The following Excel function can be used to retrieve the average fare rate of passengers in the Third Class: **=AVERAGEIF(B2:B419, "=3", I2:I419)**. This yields a result of 12.14 for the fare rate of the missing instance. The missing instance is re-included into the dataset and their fare feature is updated.

Now that all instances have a value for fare, binning can be applied. Two binning techniques are considered: equal width and equal frequency. On equal frequency binning, the value of Fare is arranged in ascending order and then split into approximately equally sized groups. A new feature is created called EqualFreq, which contains a number between 1 and 5 to represent the bin number that the entry belongs to. The number of bins is arbitrarily chosen depending upon the size of the data. If too many bins are created, then the data ends up being categorized too spersly, leaving more room for outliers. A bin number of 5 and 10 were both tested. A bin number of 5 happened to provide slightly more accurate results when predicting survival status on the test set. In Excel, the following formula can be used to calculate equal frequency: **=ROUNDDOWN(PERCENTRANK($F$2:$F$419, F2) \* N, 0) + 1**, where N = the number of bins. The PERCENTRANK() function takes an array (a range of cells) and computes the rank of a value as a percentage of the array. We must use the ROUNDDOWN() function to remove any remainder after the division and then optionally add 1 if it is preferable to begin indexing at 1 rather than 0. One caveat to this approach is that the final value in the EqualFreq column will always be categorized as N+1. To combat this, the Fare column is temporarily extended to include a new maximum value. The new maximum value will simply be the old maximum value plus some negligable amount that won’t affect the data. In this case, the old maximum value for Fare is 512.3292, so the new maximum will be 512.3293. We extend the range of the array from F419 to F420 and then remove the temporary new maximum value from Fare once finished. For equal width frequency, the approach is similar to equal frequency, however, rather than categorizing elements based upon position in the dataset, elements are categorized based upon whether or not their fare falls in between the maximum and and minimum fare value for a given bin. Taking this approach, each bin is more likely to vary in terms of the frequency of data. In order to perform equal frequency, Fare is once again sorted in ascending order. The width must be calculated in a separate cell using the following formula:

**=(MAX($F$2:$F$420) – MIN($F$2:$F$420)) / N**

Note that the same technique used in equal frequency (i.e., adding a temporary new maximum value) is used here, thus the extended range from F419 to F420. Once the width is determined, a new column is created called EqualWidth. Each cell in EqualWidth determines its bin number using the following formula:

**=ROUNDDOWN((G2 – MIN($F$2:$F$420)) / $H$2, 0) + 1**

Note here that $H$2 is a reference to the calculated width. Once again, the value must be rounded down since partial bin widths are not to be considered.

The results obtained from both equal frequency and equal width on both the training and test datasets with a bin number of 5 are displayed below:

|  |  |
| --- | --- |
| Bin Number | Count |
| 1 | 85 |
| 2 | 82 |
| 3 | 84 |
| 4 | 85 |
| 5 | 82 |

|  |  |
| --- | --- |
| Bin Number | Count |
| 1 | 387 |
| 2 | 13 |
| 3 | 0 |
| 4 | 17 |
| 5 | 1 |

|  |  |
| --- | --- |
| Bin Number | Count |
| 1 | 179 |
| 2 | 184 |
| 3 | 171 |
| 4 | 181 |
| 5 | 174 |

|  |  |
| --- | --- |
| Bin Number | Count |
| 1 | 836 |
| 2 | 33 |
| 3 | 17 |
| 4 | 0 |
| 5 | 3 |

|  |  |  |
| --- | --- | --- |
| Categorical Attributes | | Count |
| Sex | Male (1) | 577 |
| Female (2) | 312 |
| PClass | 1 | 214 |
| 2 | 184 |
| 3 | 491 |
| Embarked | C | 167 |
| S | 644 |
| Q | 78 |
| AgeGroup | Adult | 413 |
| NK | 177 |
| Youth | 210 |
| Senior | 21 |
| Baby | 14 |
| Child | 54 |
| Relatives | None | 535 |
| Few | 263 |
| Many | 91 |
| EqualFreq | 1 | 176 |
| 2 | 178 |
| 3 | 178 |
| 4 | 178 |
| 5 | 179 |

Figure 5 Updated Train Set Attributes

|  |  |  |
| --- | --- | --- |
| Categorical Attributes | | Count |
| Sex | Male (1) | 266 |
| Female (2) | 152 |
| PClass | 1 | 107 |
| 2 | 93 |
| 3 | 218 |
| Embarked | C | 102 |
| S | 270 |
| Q | 46 |
| AgeGroup | Adult | 190 |
| NK | 86 |
| Youth | 108 |
| Senior | 11 |
| Baby | 8 |
| Child | 15 |
| Relatives | None | 253 |
| Few | 131 |
| Many | 34 |
| EqualFreq | 1 | 82 |
| 2 | 84 |
| 3 | 83 |
| 4 | 84 |
| 5 | 85 |

Figure 6 Updated Test Set Attributes

## 3.4 Integrating Data

Data integration is used when multiple data sources apply to the same business problem. In this report, only one dataset was used, so this step is not applicable.

## 3.5 Formatting Data

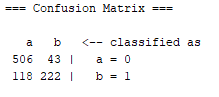
In Weka, a J48 decision tree with the ID3 algorithm is used to categorize the data by class depending upon the calculated information gain. Prior to doing this, the data must be formatted correctly. Weka incorrectly identifies the EqualFreq, PClass, and Survived attributes as having a numeric type. These are corrected to be nominal attributes. An additional column must be added to the test set for the survival class. All values are set to a value of ‘?’. This is a requirement due to the fact that a class must be selected for both the training and test sets in Weka.

# 4.0 Modeling and Evaluation

In order to create the predictive model in Weka, the pre-processed training set is first imported into Weka. J48 classifier is selected under Weka Classify tab and the default parameters are left untouched. Cross-validation is selected as the method of error detection with a value of 10 for k-folds on the Survive Class. All other settings for J48 Classifier are left as it’s default value mainly number of objects as 2 and un-pruning false. On applying the algorithm, a weighted average of 81.89% is obtained on correctly classified instances and 18.11% incorrectly classified resulting with the following confusion matrix and tree as shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **True Positive Rate** | **False Positive Rate** | **#Misclassification** |
| **0** | 0.922 | 0.347 | 43 |
| **1** | 0.653 | 0.078 | 118 |

Figure 7 Classification Chart



*Figure 8 Confusion Matrix on 10-fold J48 default settings using Updated Train Set*

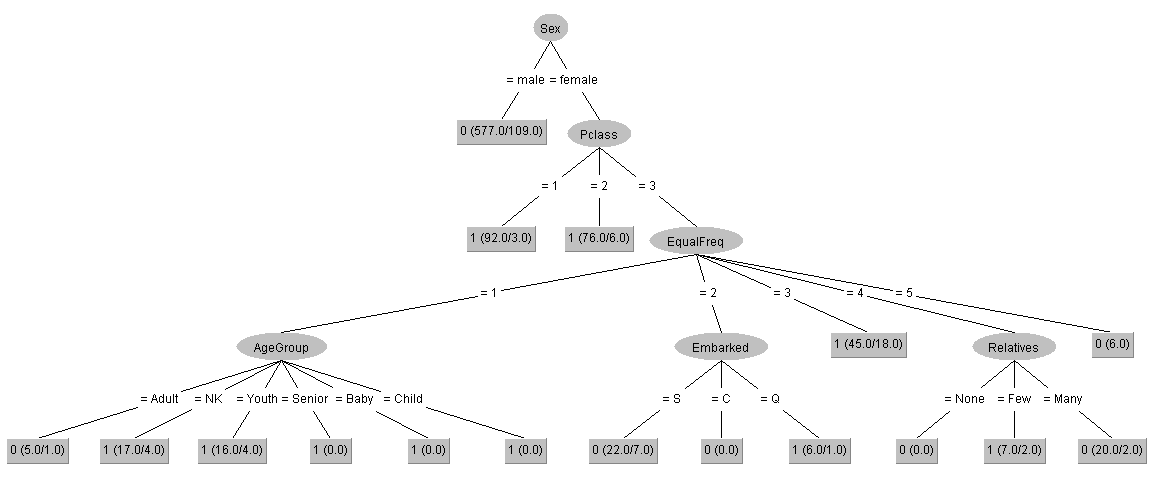


Figure 9 Decision Tree on 10-fold J48 default settings using Updated Train Set

Weka J48 uses entropy to determine the information gain of each given attribute. Since the Sex attribute has the highest information gain, it is used as the root of the tree. The J48 classifier in Weka uses a default object count of 2. Due to this fact, the male subtree ends with Survived as its only leaf node. It is determined that males are more likely to perish than to survive, and that other factors do not have enough influence to change this fact. On the female side of the tree, the PClass attribute has the highest information gain of the remaining categorical attributes. PClasses 1 and 2 are pruned and summarized as Survived (1). Third Class passengers are determined to have the highest information gain of the bunch, so another iteration of the algorithm takes place. As previously described, EqualFreq is a binning construction of the Fare attributes into 5 approximately equal size groups with bin 1 containing the lowest fare rates and 5 containing the highest fare rates. EqualFreq’s nodes 3 and 5 are pruned with group 3 having a predicted outcome of Survived (1) and group 5 predicted as Not Survived (0). The remaining 3 attributes are divided to 3 child nodes as group 1 for AgeGroup, group 2 for Embarked and group 4 for Relatives. Looking at the tree, it can be observed that the most likely candidates for survival are Third Class passengers who happen to be female . The following 3 criteria are exceptions to this rule:

* If the passenger is an adult, they are determined to perish
* If they embarked at ports S or C they are determined to perish
* If they had no relatives or many relatives on board the Titanic, they are determined to perish

Next, the same algorithm is ran using the pre-processed test file. A valid decision matrix and/or tree cannot be produced due to the fact that the Survived class values are unknown in the test set. The predictions for each instance are still recorded, however. These can be viewed by right clicking on the result record and clicking “Visualize classifier errors”. The results of the model’s predictions for survival status of the test set are shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Classification** | **Total Instances** | **Percentage** |
| **0 (Perished)** | 295 | 418 | 70.1% |
| **1 (Survived)** | 123 | 29% |

Figure 10 Predicted Class of the Test Set

# 5.0 Discussion of Results

The facts collected by Titanicfacts.net can be summarized as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Criteria** | **Female** | **Male** | **Passengers** |
| % Survived Passengers | N/A | N/A | 37% |
| # Survived Passengers | N/A | N/A | 492 |
| % 1st Class Survivors | N/A | N/A | 61% |
| % 2nd Class Survivors | N/A | N/A | 42% |
| % 3rd Class Survivors | N/A | N/A | 24% |
| Of Survivors % Male Survived | N/A | N/A | 20% |
| Of Survivors % Female Survived | N/A | N/A | 75% |
| 1st Class Youngest Female Survivor | 13yrs old | N/A | N/A |
| 1st Class Oldest Female Survivor | 64yrs old | N/A | N/A |
| 2nd Class Youngest Female Survivor | 10mths old | N/A | N/A |
| 2nd Class Oldest Female Survivor | 59yrs old | N/A | N/A |
| 3rd Class Youngest Female Survivor | 2mths old | N/A | N/A |
| 3rd Class Oldest Female Survivor | 63yrs old | N/A | N/A |
| 1st Class Youngest Male Survivor | N/A | 11mths old | N/A |
| 1st Class Oldest Male Survivor | N/A | 60yrs old | N/A |
| 2nd Class Youngest Male Survivor | N/A | 7mths old | N/A |
| 2nd Class Oldest Male Survivor | N/A | 62yrs old | N/A |
| 3rd Class Youngest Male Survivor | N/A | 5mths old | N/A |
| 3rd Class Oldest Male Survivor | N/A | 45yrs old | N/A |

Figure 11 Table of Facts Taken From Titanicfacts.net

Certain facts which can be inferred from the data above are summarized as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Inferred Facts** | **Female** | **Male** | **Passengers** |
| ∵ 37% Survived & 492 people Survived  ∴ 492/0.37 = Dataset Instances | N/A | N/A | 1330 |
| ∵ 20% Survivors were Male  ∴ 492\*20% = number Male Survived | N/A | 98 | N/A |
| ∵ 75% Survivors were Female  ∴ 492\*75% = number Female Survived | 369 | N/A | N/A |
| ∵ 75% Survivors were Female & 20% Survivors were Male  ∴ 100% - (75% + 20%) = Percentage/Number Unknown | N/A | N/A | 5% or 25 |
| ∵ 61% of survivors were 1st Class  ∴ 492\*61% = number 1st Class Survivors | N/A | N/A | 300 |
| ∵ 42% of survivors were 2nd Class  ∴ 492\*42% = number 2nd Class Survivors | N/A | N/A | 207 |
| ∵ 24% of survivors were 3rd Class  ∴ 492\*24% = number 3rd Class Survivors | N/A | N/A | 118 |
| ∵ Number of Survivors Exceed number of Survivors Based on % of Survivors from each Classes ∴ Absolute Error | N/A | N/A | 133 |
| ∵ Passenger Class Absolute Error 133  ∴ Passenger Class Relative Error | N/A | N/A | 27% |

Figure 12 Inferred Facts (Entries Highlighted in Red Contain Discrepancies)

According to Titanicfacts.net, there were 492 out of 1330 passengers (37%) passengers that survived the sinking of the Titanic. Within the test set, the learning model predicted that 123 of 418 instances would survive. Adding this to the training set data, (which contains 340 survivors out of 889 instances) gives an actual survivor count of 463 passengers out of 1307 instances (35%). This amount is off by 2%, which could be due, in part, to the difference between the inferred number of passengers that were on board the Titanic (1330) and the number of passengers within the dataset provided by Kaggle (1307). The calculated difference amounts to 1.72%.

Titanicfacts.net states that out of the passengers that survived 61% were First Class passengers, 42% were Standard Class passengers, and 24% were Third Class passengers. The passenger classes contained in the test set, as well as the corresponding predicted survival status for each class, are displayed below in Figures 15 and 16 respectively:

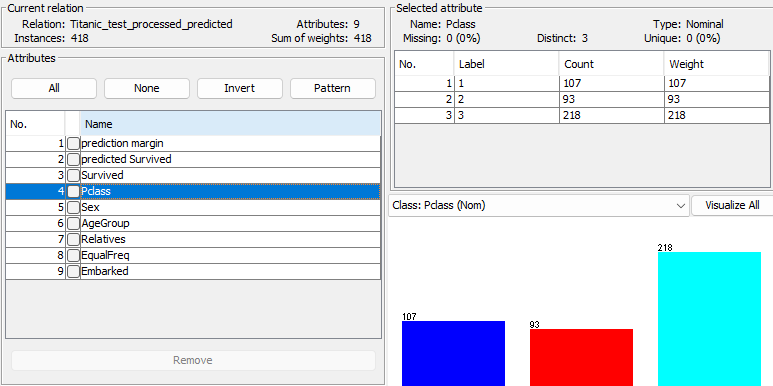


Figure 13 Weka Test Set Chart (Blue First Class, Red Standard Class, Teal Third Class)

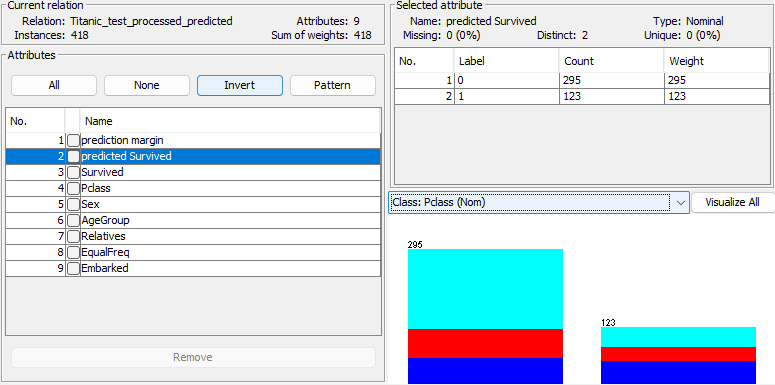


Figure 14 Weka Test Set Predicted Survivor to PClass (First bar for Perished and Second Bar for Survived)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Weka DB Feature on Test Set | | | | |
| Prediction | Count | Percentage to Test Set | Percentage Survived by Class | TitanicFacts.net Facts |
| 1st Class Survivors |  | 12.0% | 46.7% | 61% |
| 1st Class Passengers |  | 25.6% |
| 2nd Class Survivors |  | 7.2% | 32.3% | 42% |
| 2nd Class Passengers |  | 22.2% |
| 3rd Class Survivors |  | 10.3% | 19.7% | 24% |
| 3rd Class Passengers |  | 52.2% |

Figure 15 Test Set Titanic Survivors by Pclass

As shown in figure 17, the predictive model, in comparison to the known facts, differs by approximately 14.3 % on First class passengers, 9.7% on Standard Class passengers and 4.3% on Third Class passengers. However, there is a margin of error that can be applied, since, as shown in Figure 14, the percentage of survivors from each class contain some inconsistencies. Applying a 27% relative error to the number of survivors from each class produces 59.3%, 41% and 25.1% for First Class, Standard Class and Third Class, respectively. These percentages can be obtained by multiplying 1.27 with the number of Survivors from the passenger class, and then dividing by the number of passengers from said class. This analysis demonstrates that the predictive model has a 1-2% margin of error, which is consistent with the difference between the model’s predicted survivors (35%) in comparisson to the facts on Titanicfacts.net (37%).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Weka DB Feature Test Set and Train Set | | | | |
| Prediction | Count Test Set | Count Train Set | Percentage of Survivors Both Sets | TitanicFacts.net Percentage of Male/Female Survivor |
| Male Survivors |  |  | 12.9% | 20% |
| Male Passengers |  |  |
| Female Survivors |  |  | 76.3% | 75% |
| Female Passengers |  |  |

Figure 16 Combined Training Set and Test Set on Male and Female Survivors

According to the inferred facts, the number of passengers whos gender is unknown is 5% (approximately 25 passengers). Observing the facts from the datasets reveals that there is a significant discrepancy in male survivors. It may be theorized that the estimated 25 missing passenger’s gender on the Titanicfacts.net dataset were, in fact, male passengers which survived. By including 25 male passengers with a survival outcome of survived to the dataset’s prediction, the number of male survivors is increased to 134, and the total number of male passengers is increased to 868, making the percentage of male survivors 15.4%. With these assumptions included, there remains an approximate 4% margin of error, which is believed to come from the 2% difference in number of total dataset instances in addition to the observed misclassification of female survivors (1.3%).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Female** | **Female AgeGroup** | **Male** | **Male AgeGroup** |
| 1st Class Youngest Survivor | 13yrs old | Youth | 11mths old | Baby |
| 1st Class Oldest Survivor | 64yrs old | Senior | 60yrs old | Adult |
| 2nd Class Youngest Survivor | 10mths old | Baby | 7mths old | Baby |
| 2nd Class Oldest Survivor | 59yrs old | Adult | 62yrs old | Senior |
| 3rd Class Youngest Survivor | 2mths old | Baby | 5mths old | Baby |
| 3rd Class Oldest Survivor | 63yrs old | Senior | 45yrs old | Adult |

Figure 17 Youngest and Oldest Survivors by PClass

# 6.0 Conclusion

The Titanic dataset demonstrates the magnitude of impact that certain features can have upon outcomes as dire as one’s survival. This research paper attempted to create an model with accurate and reliable predictive power. In comparison to the ground truths of the actual event, it is concluded that the generated model succeeded in achieving this goal. In order to achieve this level of accuracy, the dataset had to be modified in accordance with CRISP-DM standards. Features such as Passenger Id, Name, Ticket Number and Cabin number were determined to have insignificant or no affect on the survival status of a passenger. Features such as Age, SibSp, ParCh, and Fare were modified to aid in classification with the J48 classifier in Weka. The data was formatted correctly and a predictive model was generated. A total of 80.54% of instances were correctly classified by the model. It was determined that the feature with the most influence upon survival was Sex. This was followed by PClass, and then Fare rate. The accuracy of the model could still be improved. Suggestions on how to do this include deciphering the ticket numbers, researching the cabin numbers of each passenger, and/or constructing a Title feature based upon the names of passengers. Exploring the Titanic dataset was a fascinating venture and a great opportunity for learning about techniques used in the field of Data Science.

# 7.0 References

|  |  |
| --- | --- |
| [1] | N. Donges, "Predicting the Survival of Titanic Passengers," Towards Data Science, 14 May 2018. [Online]. Available: https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8. [Accessed 13 February 2023]. |
| [2] | I. H. Witten, "The University of Waikato," 2013. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/slides/Class3-DataMiningWithWeka-2013.pdf. [Accessed 13 February 2023]. |
| [3] | T. Facts, D. Fowler and H. i. Numbers, "Titanic Survivor," Titanic Facts, 2023. [Online]. Available: https://titanicfacts.net/titanic-survivors/. [Accessed 13 February 2023]. |