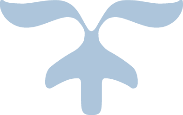


Titanic Dataset

CST8390\_23W Assignment 2



February 13, 2023

Algonquin College

Computer Engineering Technology

Business Intelligence and Data Analytics

Hoang, John – 040896360

Kingdom, Neil – 040967309

Prepared for:

Dr. Anu Thomas

Table of Contents

[Introduction 3](#_Toc127282987)

[Data Understanding 3](#_Toc127282988)

[Starting Dataset 3](#_Toc127282989)

[Dataset Breakdown 4](#_Toc127282990)

[Classes Train Set 4](#_Toc127282991)

[Attributes Train Set (Age Updated) 5](#_Toc127282992)

[Classes Test Set 5](#_Toc127282993)

[Attributes Test Set (Updated Age) 5](#_Toc127282994)

[Data Preparation 6](#_Toc127282995)

[Preprocessing – Classification 6](#_Toc127282996)

[Updated Training Dataset on Classification 6](#_Toc127282997)

[Updated Test Dataset on Classification 7](#_Toc127282998)

[Cleaning Up 7](#_Toc127282999)

[Removing Data 7](#_Toc127283000)

[Updating Missing Values 8](#_Toc127283001)

[Modeling and Evaluation 8](#_Toc127283002)

[Discussion of Results 9](#_Toc127283003)

[Conclusion 9](#_Toc127283004)

[References 10](#_Toc127283005)

[Appendix 1 11](#_Toc127283006)

[Training Set Male (Mr.) 11](#_Toc127283007)

[Training Set Female (Miss) 12](#_Toc127283008)

[Training Set Female (Mrs.) 13](#_Toc127283009)

[Training Set Unique 13](#_Toc127283010)

[Appendix 2 14](#_Toc127283011)

[Test Set All Missing Age Instances 14](#_Toc127283012)

[Test Set Male (Mr.) 15](#_Toc127283013)

[Test Set Female (Ms.) 16](#_Toc127283014)

[Test Set Female (Mrs.) 16](#_Toc127283015)

[Test Set Unique 16](#_Toc127283016)

[Test Set Fare Missing 16](#_Toc127283017)

# 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 using supervised learning 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 algorithm [2]. 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 model [3].

# 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 Titanic [1]. The labels are abbreviated for simplicity and listed as follows: C = Cherbourg, Q = Queenstown, S = Southampton.

## 2.1 Initial Dataset

|  |  |
| --- | --- |
| **Attribute/Feature** | **Datatype** |
| Passenger Id | Distinct Numeric |
| PClass | Nominal |
| Sex | Nominal Binary |
| Age | Continuous |
| Sibsp | Distinct Numeric |
| Parch | Distinct Numeric |
| Fare | Continuous |
| Embarked | Nominal |
| Cabin | Distinct Numeric |
| Ticket | Distinct Numeric |
|  |  |
| Survived | Nominal Binary |

## 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%) whos age is unknown and 687 (77.3%) whos Cabin number is unknown. In the test set, there are 86 instances (20.8%) whos age is unknown and 1 instance with a missing fare value, which can be seen in “Appendix 2”.

## 2.3 Training Set Classes

|  |  |  |
| --- | --- | --- |
| Category | | Count |
| Survivor | Survivor (1) | 549 |
| Non-Survivor (0) | 340 |
| Gender | Female (1) | 312 |
| Male (1) | 577 |
| Embarked | C | 167 |
| S | 644 |
| Q | 78 |
| 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 |

## 2.5 Test Set Classes

|  |  |  |
| --- | --- | --- |
| Category | | Count |
| Gender | Female (1) | 152 |
| Male (0) | 266 |
| Embarked | C | 102 |
| S | 270 |
| Q | 46 |
| 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 |

# 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 posseses. 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 sacrafice 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 Southhampton, followed by Cherbourg and finally, Queenstown. Therefore, the likelyhood 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 likelyhood 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:

## 3.3.1 Test Set Equal Frequency (Five bins)

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

## 3.3.2 Test Set Equal Width (Five Bins)

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

## 3.3.3 Training Set Equal Frequency (Five bins)

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

## 3.3.4 Training Set Equal Width (Five Bins)

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

## Once binning has been applied, the Fare feature is removed from the dataset. It was decided that using equal frequency binning was better suited for this dataset due to the fact that the spread of equal width had a heavy bias towards bin 1. This bias is the result of an outlier in Fare prices. The maximum value of Fare is 512.3292, which is well above the average fare price.

## 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.

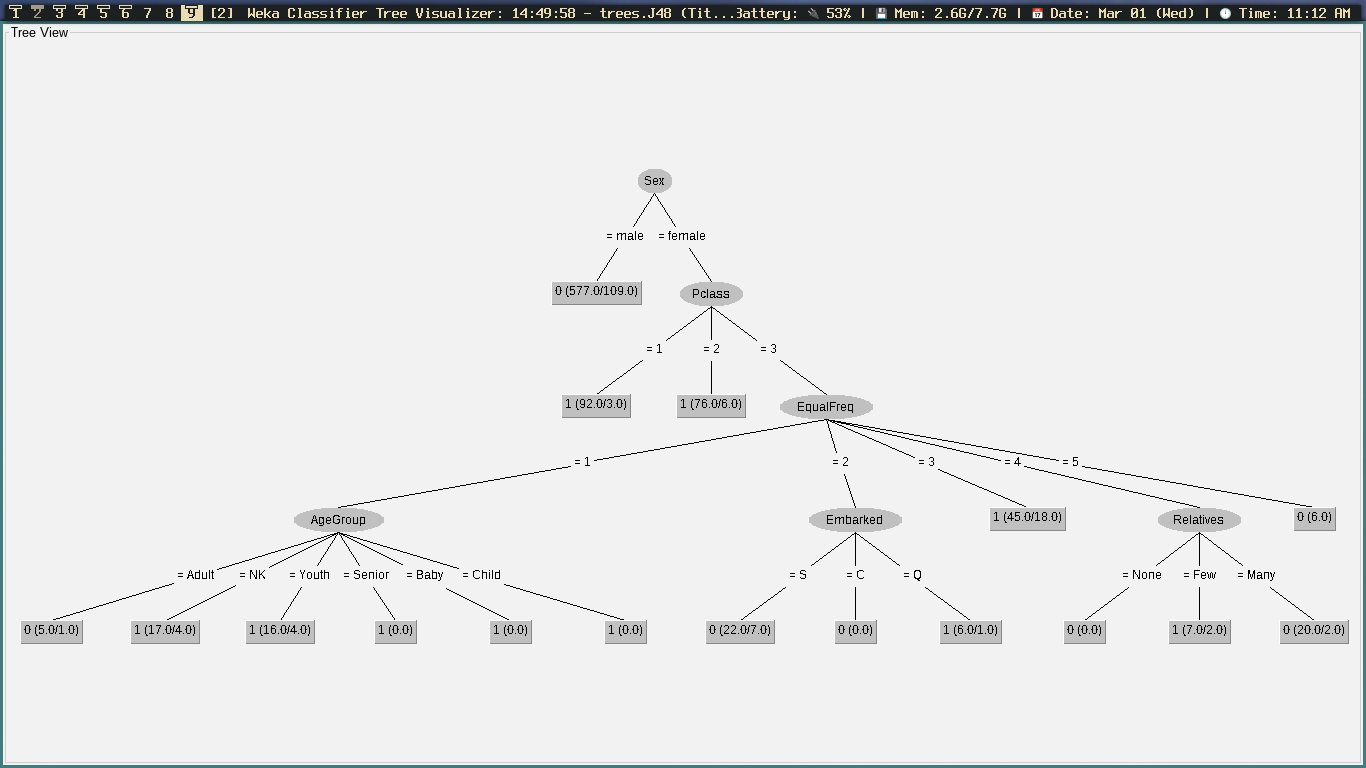
## 3.5 Formatting Data

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. Under the Classify tab, the J48 classifier is selected 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. Survived is selected as the class to be predicted. The algorithm is then ran. The resulting confusion matrix and tree shown below:

|  |  |
| --- | --- |
| Classified as Perished | Classified as Survived |
| 508 | 41 |
| 132 | 208 |

Figure 1: J48 Decision Tree for Survived Class

The ID3 algorithm calculates the entropy of the dataset in order to determine which attribute provides the highest information gain, and then uses that attribute as the root node for the current iteration. Figure 1. indicates that Sex is the predominant feature that affects the likelyhood of the passenger’s survival, followed by PClass, and so forth. The leaf nodes begin with a 0 or 1 indicating that the passenger will either perish or survive respectively.

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 uknown 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:

|  |  |
| --- | --- |
| Prediction Results | |
| Total number of instances | 418 |
| Number of persons predicted to survive | 121 |
| Number of persons predicted not to survive | 297 |
| Percentage of predicted survival | 28.94% |

# 5.0 Discussion of Results

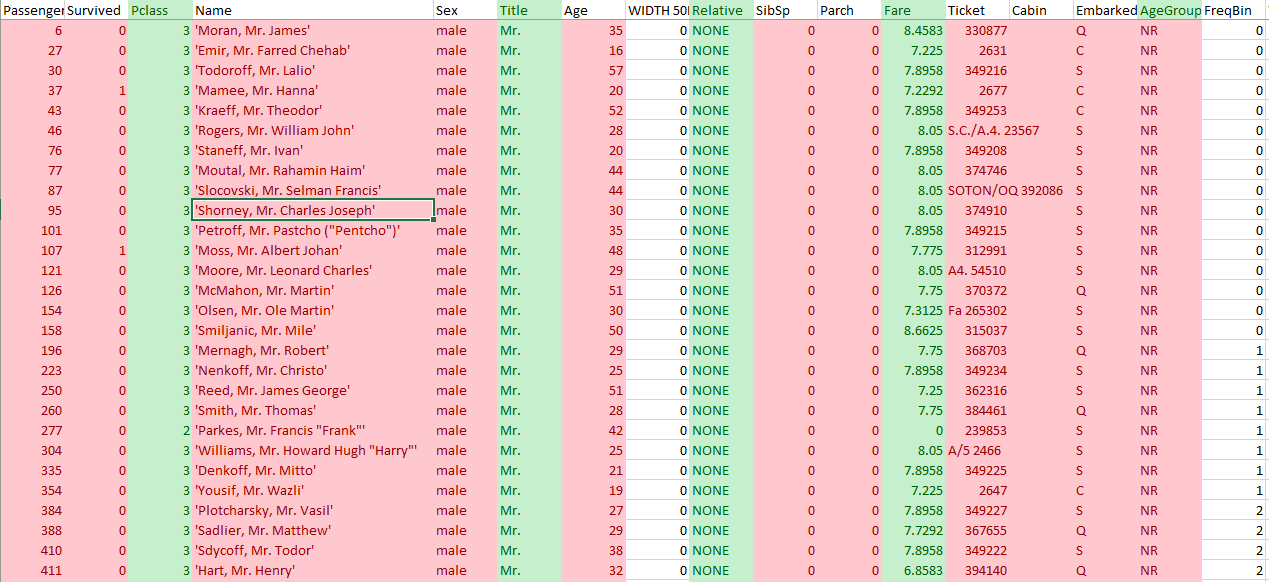
# 6.0 Conclusion

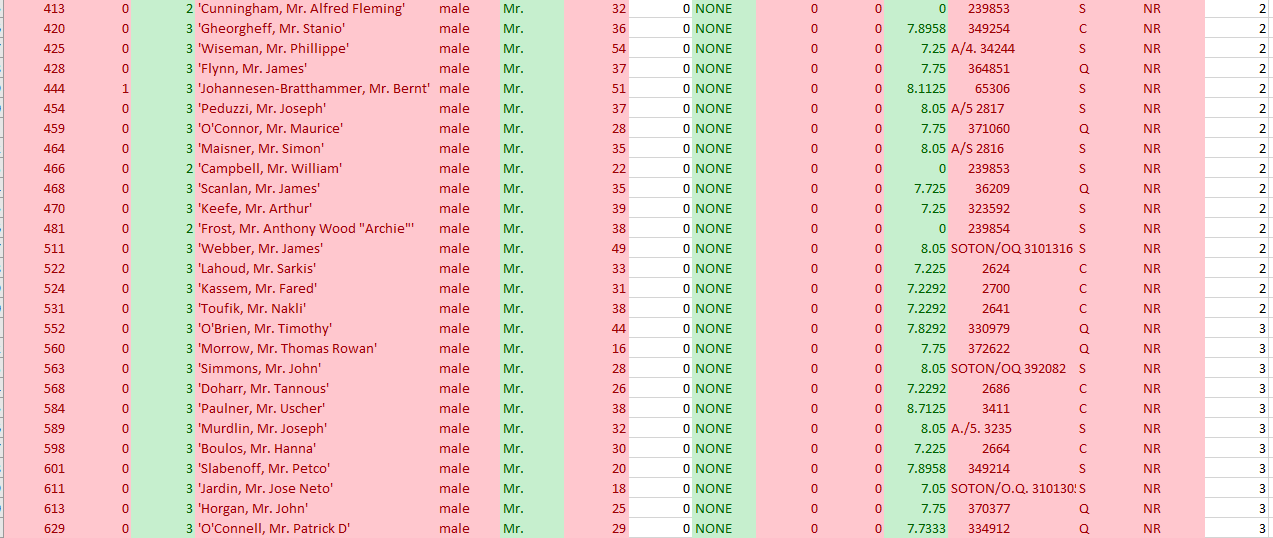
# 7.0 References

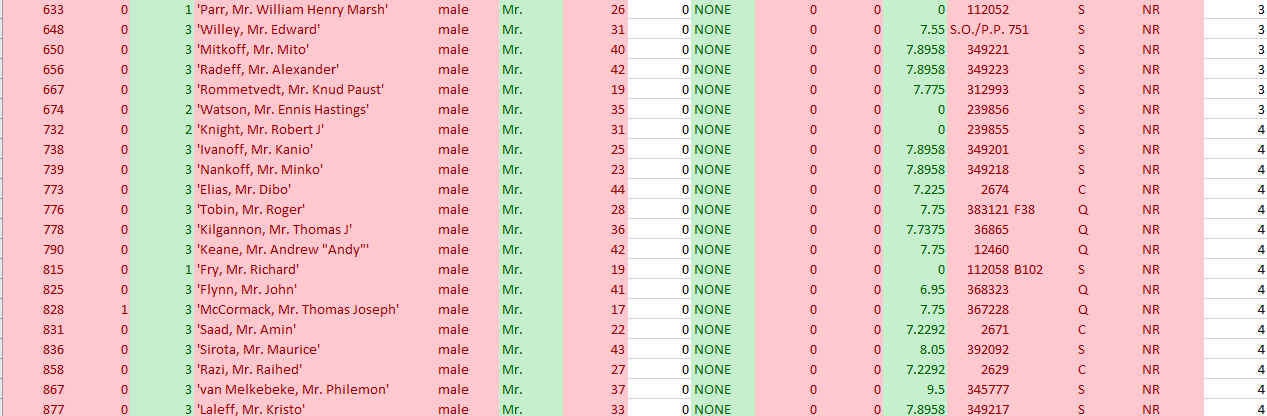
|  |  |
| --- | --- |
| [1] | N. Donges, "Predicting the Survival of Titanic Passengers," Towards Data Science, 14 May 2018. [Online]. Available: https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8. [Accessed 13 February 2023]. |
| [2] | I. H. Witten, "The University 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]. |

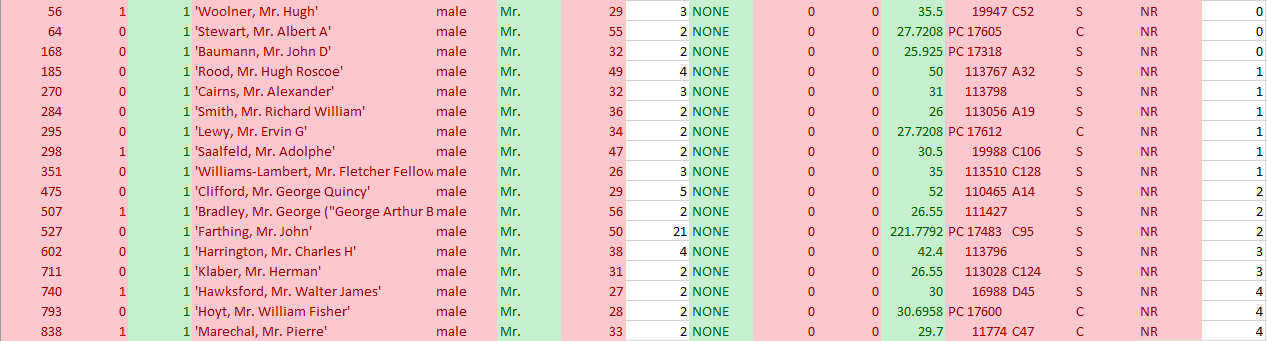
# Appendix 1

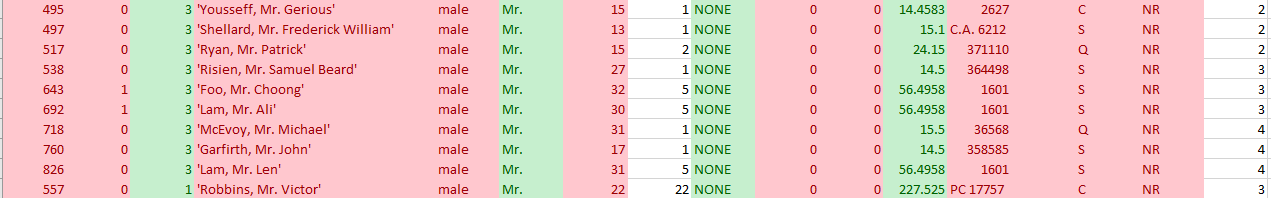
### Training Set Male (Mr.)

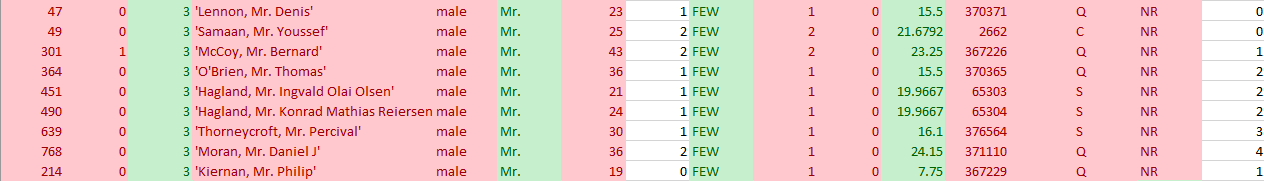


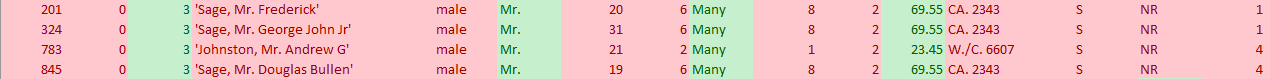




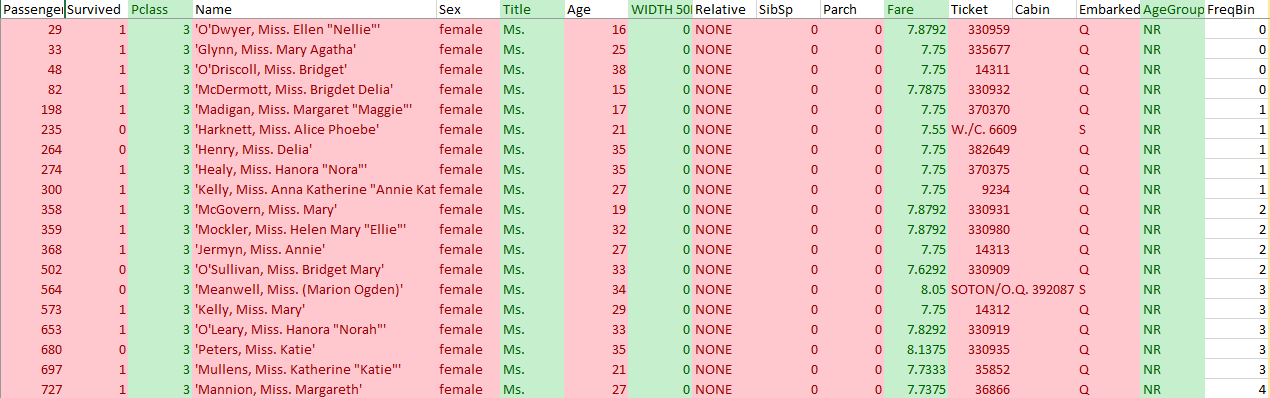








### Training Set Female (Miss)

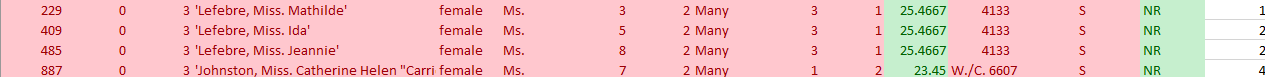












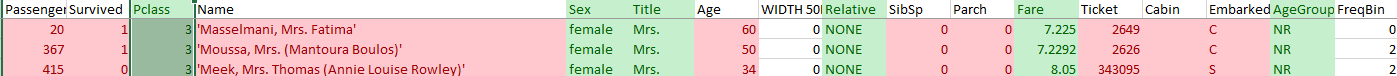


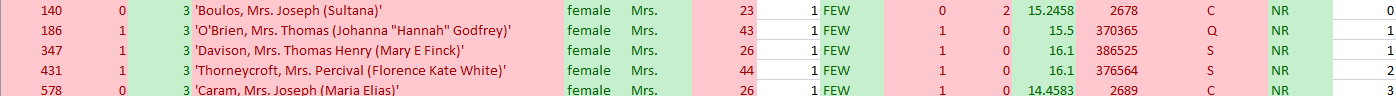






### Training Set Female (Mrs.)













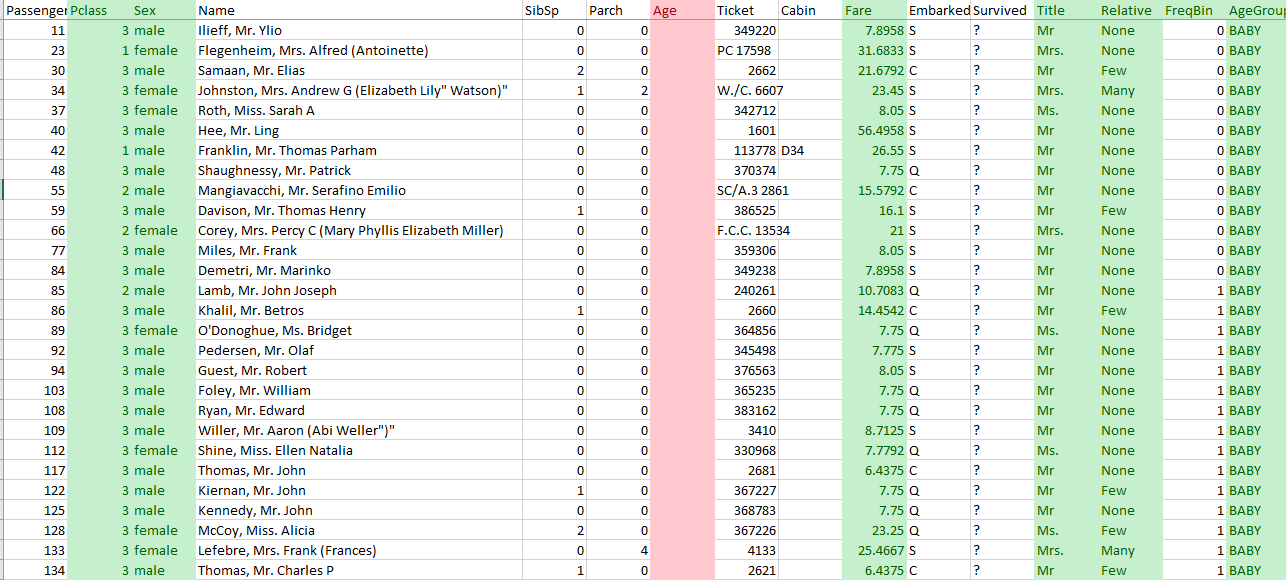
### Training Set Unique

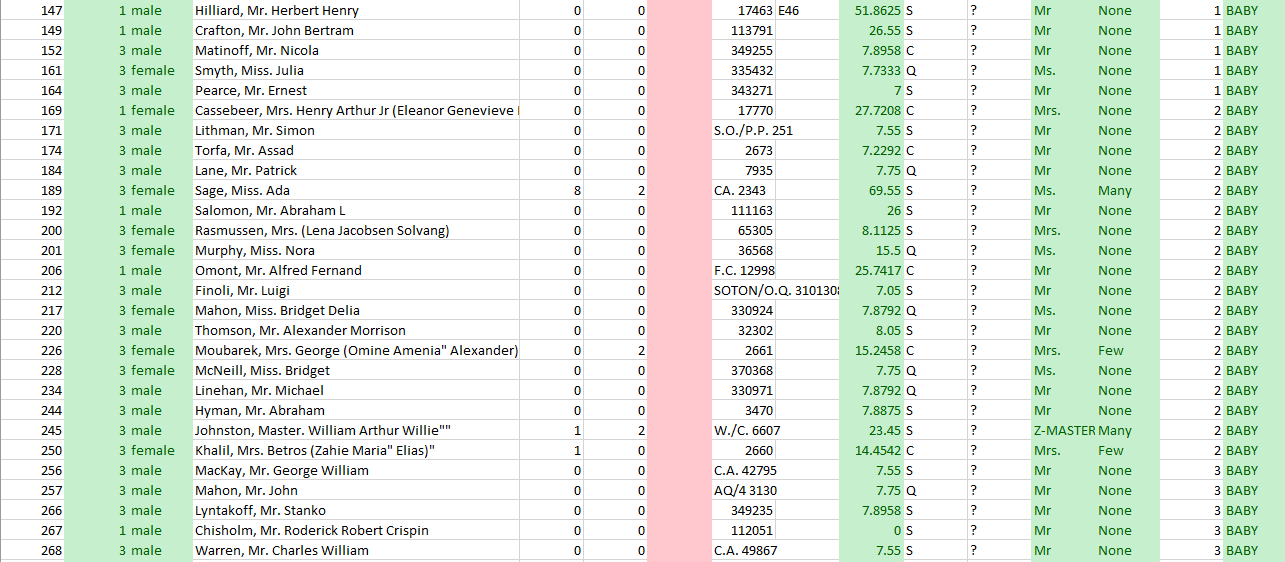


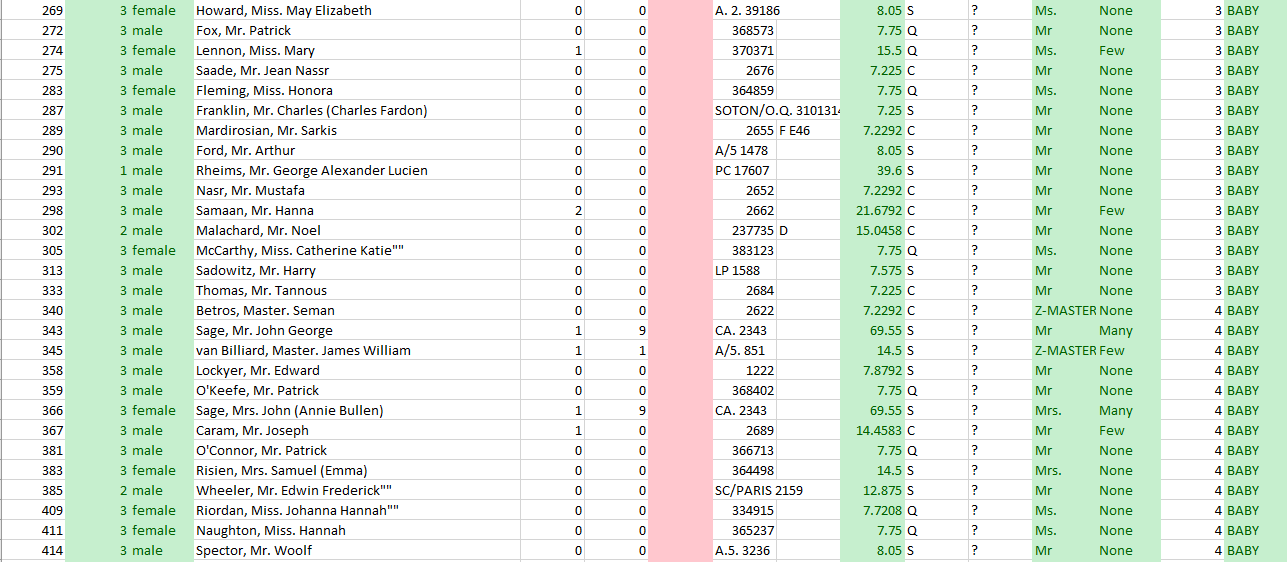


# Appendix 2

### Test Set All Missing Age Instances

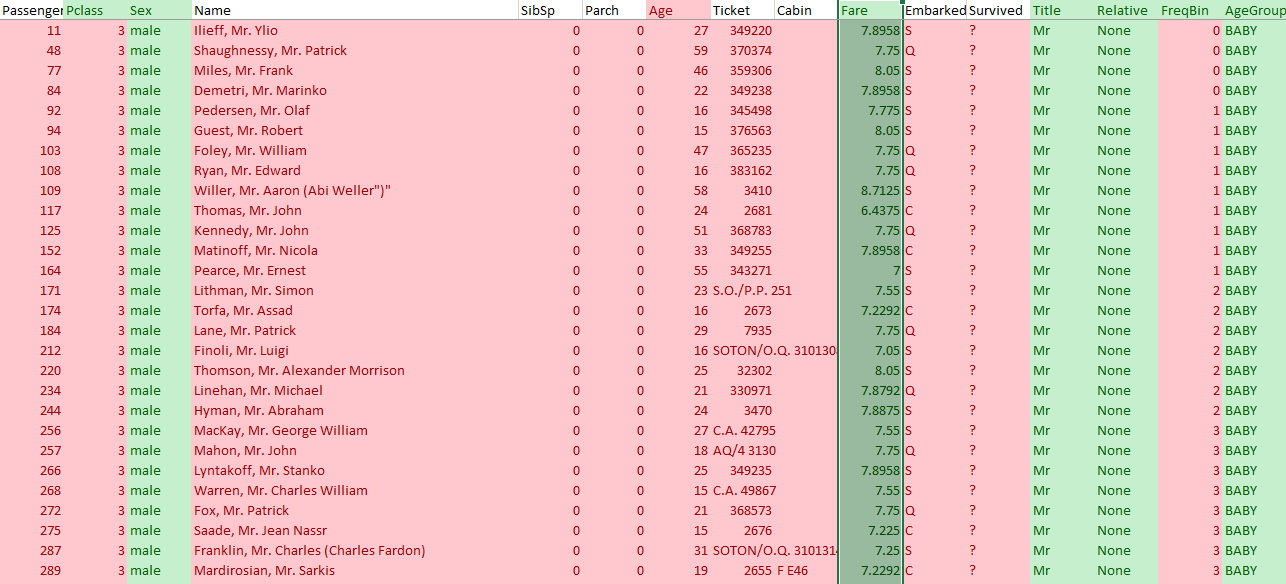


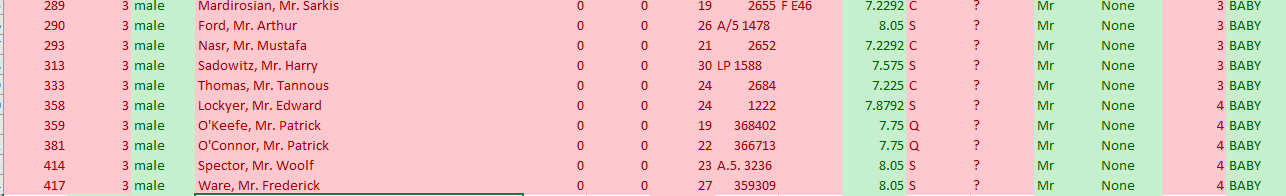






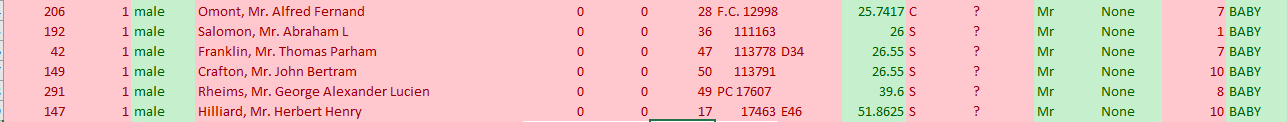
### Test Set Male (Mr.)









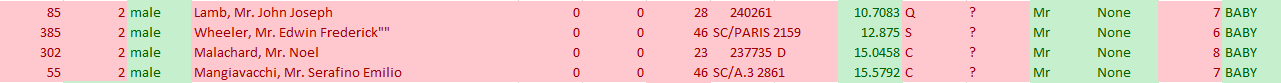




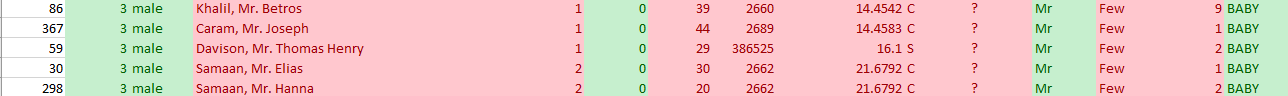






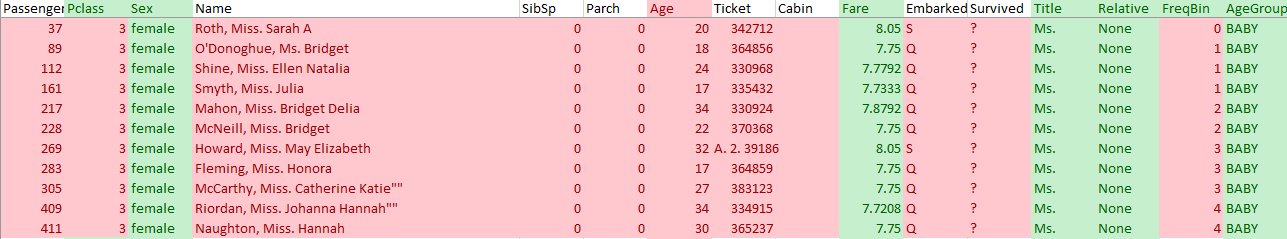








### Test Set Female (Ms.)







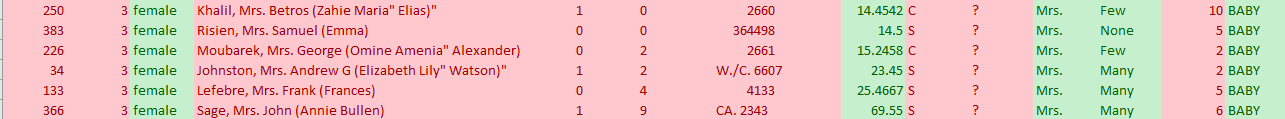




### Test Set Female (Mrs.)















### Test Set Unique





### Test Set Fare Missing

