Andrew Ku

08/18/2021

# Titanic – Machine Learning from Disaster

Table of Contents

[Titanic – Machine Learning from Disaster 1](#_Toc82797420)

[1. Summary 2](#_Toc82797421)

[2. Data Question and Problem Definition 2](#_Toc82797422)

[3. High Level Background 3](#_Toc82797423)

[3](#_Toc82797424)

[3.1 Missing Data 4](#_Toc82797425)

[4. Descriptive Statistics 6](#_Toc82797426)

[4.1 Numeric and Character Data 6](#_Toc82797427)

[4.1.1 Exploration of the continuous (numerical) data. 7](#_Toc82797428)

[5. Imputing NA values 10](#_Toc82797429)

[5.1 Embarked Attribute 10](#_Toc82797430)

[5.2 Fare Attribute 11](#_Toc82797431)

[5.3 Imputing Age Attribute 12](#_Toc82797432)

[12](#_Toc82797433)

[6. Feature Engineering 13](#_Toc82797434)

[6.1 Did each family sink together? 13](#_Toc82797435)

[6.2 Surname 14](#_Toc82797436)

[6.3 Father? Mother? Minor? 15](#_Toc82797437)

[6.4 which Gender Survived better? 16](#_Toc82797438)

[7. Machine Learning Models: Support Vector Machine and Random Forest 17](#_Toc82797439)

[7.1 Support Vector Machine 17](#_Toc82797440)

[7.2 Random Forest 18](#_Toc82797441)

[7.3 Variable importance 19](#_Toc82797442)

[8. Recommendation & Insights 20](#_Toc82797443)

[8.1 Recommendation 20](#_Toc82797444)

[8.2 Insights 20](#_Toc82797445)

[9. Appendix : Code 21](#_Toc82797446)

# 1. Summary

On April 15, 1912, the RMS Titanic, the largest and one of the most technologically advanced ship, sunk in the North Atlantic Ocean. This unfortunate event resulted in the death of 1502 out of 2224 people who were on the ship.

The dataset was acquired from the Kaggle website (<https://www.kaggle.com/c/titanic/overview>). This dataset one of the most popularized dataset, and widely used by beginner data scientists to start on their first project for the first time.

# 2. Data Question and Problem Definition

Data competitions sites such as Kaggle provide the problem to solve or questions to be answered with the given datasets (<https://www.kaggle.com/c/titanic>). Additionally, a few more questions were added as exploratory analysis was conducted. The questions available so far:

* What sorts of people were more likely to survive?
* Did each family sink together?
* Who were more likely to survive?
  + Father?
  + Mother?
  + Child?
* Do Surnames have any relation to survival?
* Which gender survived better?

From the website, both the training dataset (with survival information) and the test dataset (without survival information) were given. In this project, two of the classification Machine Learning algorithms was trained from the training dataset, and the model results were tested against the test dataset. The chosen algorithms were **Kernel SVM and Random Forest**. They were chosen with reasons of data with reasonable size, aim for accuracy, and predicting a categorical variable. Explanation of these algorithms will be explained in the later section.

# 3. High Level Background

# Text, letter Description automatically generated

***Image 1. Data Descriptive Statistics***

There are total of 1309 objects with 12 variables. Since a few of the variables are not easy to understand, data dictionary was provided from Kaggle:

Table

Description automatically generated

***Image 2. Data Dictionary.***

Variable notes were also provided:

Pclass: Socio-economic status (SES)

* 1st = Upper Class
* 2nd = Middle Class
* 3rd = Lower Class

Sibsp:

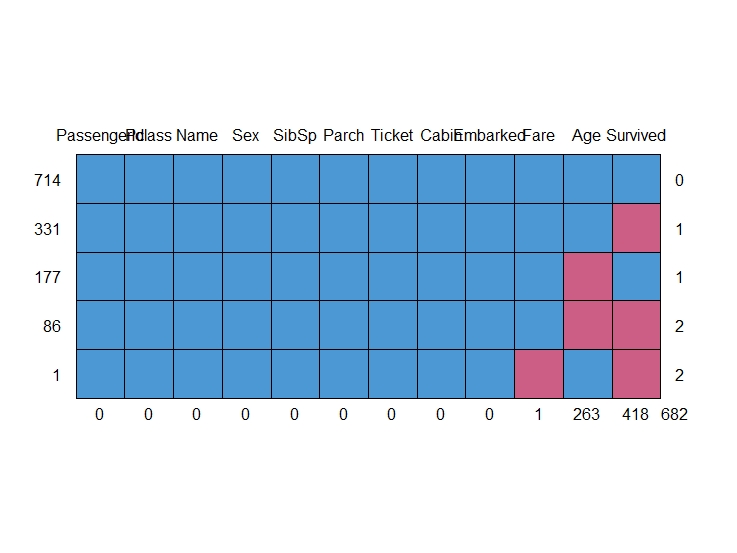
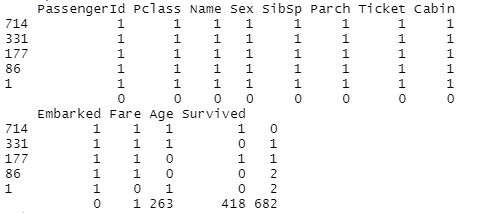
* Sibling = brother, sister, stepbrother, stepsister
* Spouse = husband, wife

Parch:

* Parent = mother, father
* Child = daughter, son, stepdaughter, stepson

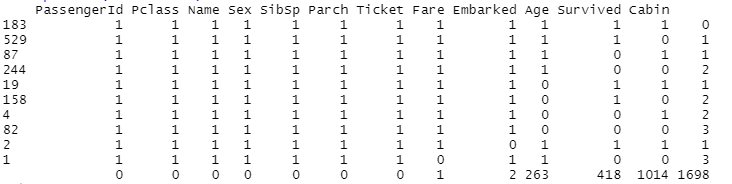
## 3.1 Missing Data

As seen in Image 1., there were a few blank values in the “Cabin” column. This is bad because computers will mis-interpret blank data that are meant to be NAs.



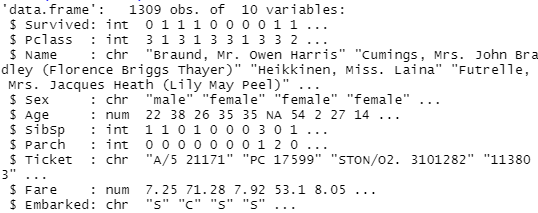
***Image 3. NA Values and Chart from Left to right, Respectively.***

Clearly, computer did not process blank data as NAs as seen in *Image 3*. This needed to be taken care of since some of the ML models would not like that.



***Image 4. Blank Values replaced with NAs.***

Blank values were replaced with NAs as seen in *Image 4.* This step was critical since without this process, 1014 NAs from the “Cabin” column would have been untreated. However, since there are too many NA values in the “Cabin” column, this attribute was decided to be excluded from the dataset. Additionally, the “passengerId” attribute was extracted as well since its only used as the primary key of the dataset (uniquely defines each row in the dataset).



***Image 5. Cabin and PassengerId Column Excluded from the Dataset.***

# 4. Descriptive Statistics

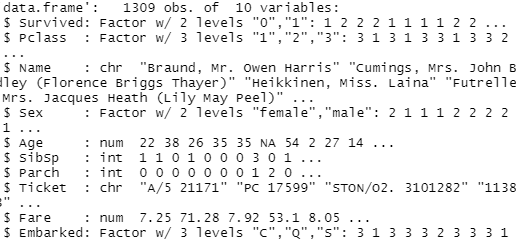
## 4.1 Numeric and Character Data

Threre are two Types of Data:

1. Categorical Data
   1. Nominal: values represent discrete units.
   2. Ordinal: values represent discrete ordered units.
2. Continuous Data
   1. Interval: ordered units with intermediate values.
   2. Ratio: ordered units with intermediate values (distance between the values is the same)

Identifying the type of data contained in a variable is a critical step. Taking this step thoroughly will ensure appropriate statistical procedures.

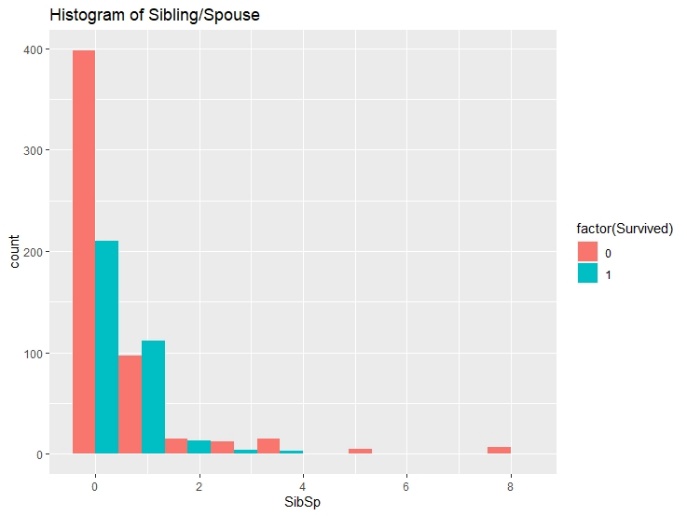
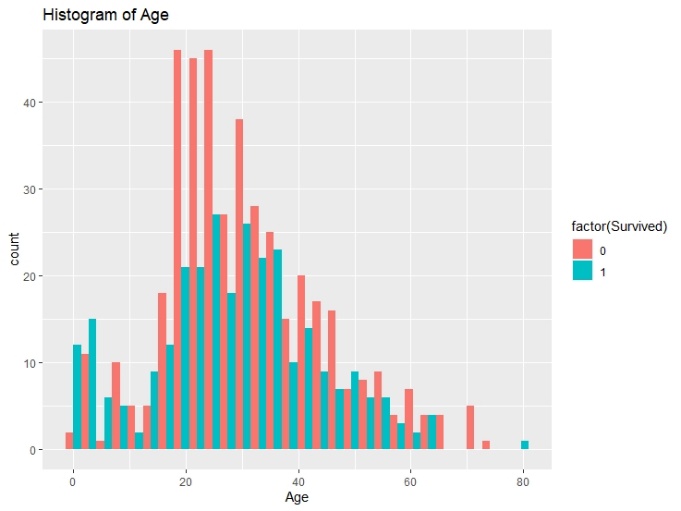
There were total of 4 continuous data, “Age”, “SibSP”, “Parch”, and “Fare”. On the other hand, there were six character data, “Survived”, “Pclass”, “Name”, “Sex”, “Ticket”, and “Embarked” as seen in *image 5*. From the character data, “Survived”, “Pclass”, “Sex”, and “Embarked attributes were converted to categorical data (Factors).

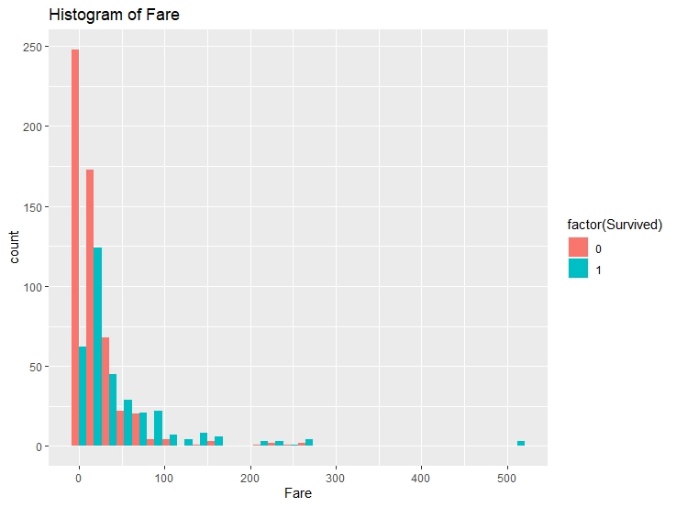
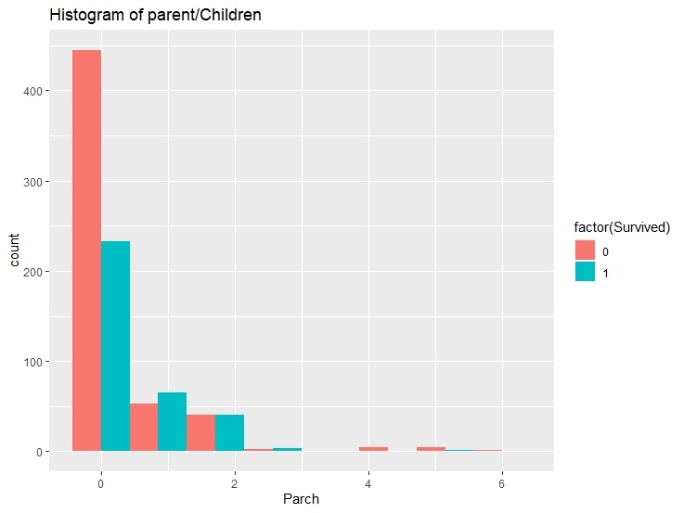


***Image 6. Converted Dataset.***

Next, each Numeric Data were lightly explored.

### 4.1.1 Exploration of the continuous (numerical) data.

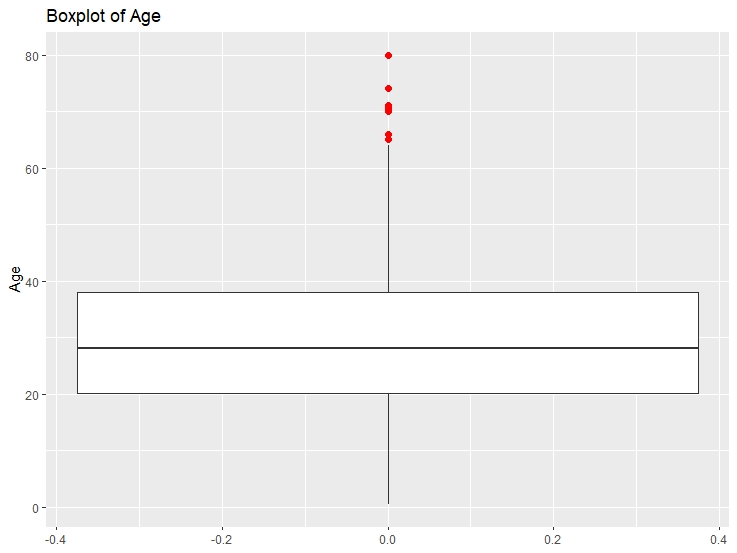




***Image 7. Histogram of Multivariate attributes (Age, SibSp, Parch, and Fare vs Survived, respectively).***

*Histogram* of Multivariate data between all numeric and Survived data are shown in *Image 7.* Explanation of each graph and deeper exploration was conducted into 3 sections: Exploration of Age Attribute, Exploration of SibSp and Parch Attribute, and Exploration of Fare Attribute.

##### Exploration of Age Attribute



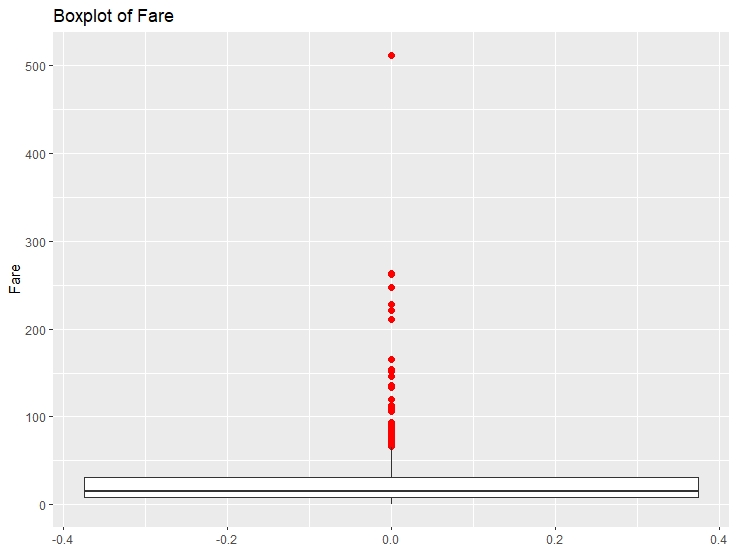
***Image 8. Boxplot of Age***

* Age between **20s to 30s** have shown the most frequency in not surviving as seen in *Image 7*.
* Age attribute had **many outliers**; hence, scaling may be needed later in the project as seen in *Image 8.*
* Age attribute had many **NA values**; hence, imputation is probably needed as seen in *Image 4.*

##### Exploration of SibSp and Parch Attribute

* People who did **not have any siblings or spouse** have shown **the most frequency in not surviving**, and only people with **1** sibling or spouse had **higher frequency of surviving than not surviving** as seen in *Image 7.*
* People who did **not have any parents or children** have shown **the most frequency in not surviving**, and only people with **1** parent or child had **higher frequency of surviving than not surviving** as seen in *Image 7.*
* From this exploration, it raised a question of how each of family member (father, mother, child) affected the survival outcome? Did families sink together? **Feature engineering** was conducted to answer this question.

##### Exploration of Fare Attribute



***Image 9. Boxplot of Fare***

* People with **lower fare amount (< $20)** have shown the **most frequency** in not surviving as seen in *Image 7.*
* People with **higher fare amount** had higher frequency of **higher frequency of surviving than not surviving** as seen in *Image 7.* This raises questions to if Fare amount is related to each passenger’s SES aka Pclass.
* Fare data also had many outliers; hence, scailing may be needed later in the project as seen in *Image 9.*

## 5. Imputing NA values

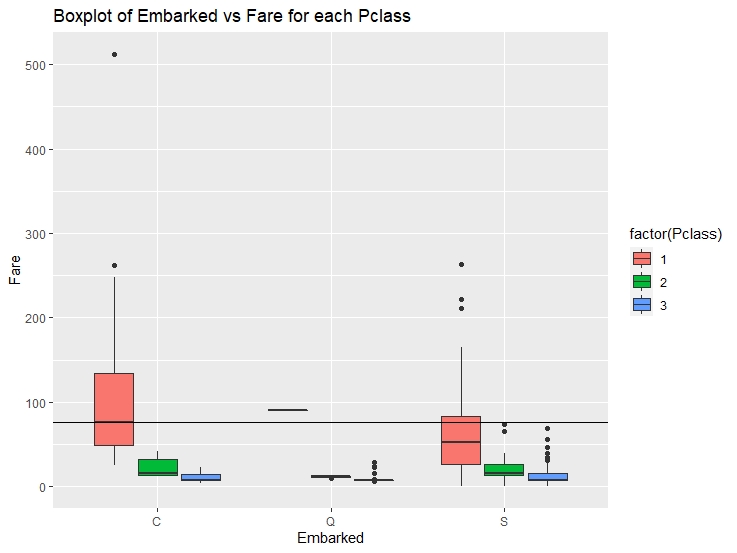
## 5.1 Embarked Attribute

There were total of 2 missing values in the Embarked attribute as seen in *image 4.*

**

***Image 10. Rows of Missing Embarked Values***

More specifically, rows of missing values in the Embarked attribute can be seen in *Image 10.*

**

***Image 11. Rows of Missing Embarked Values***

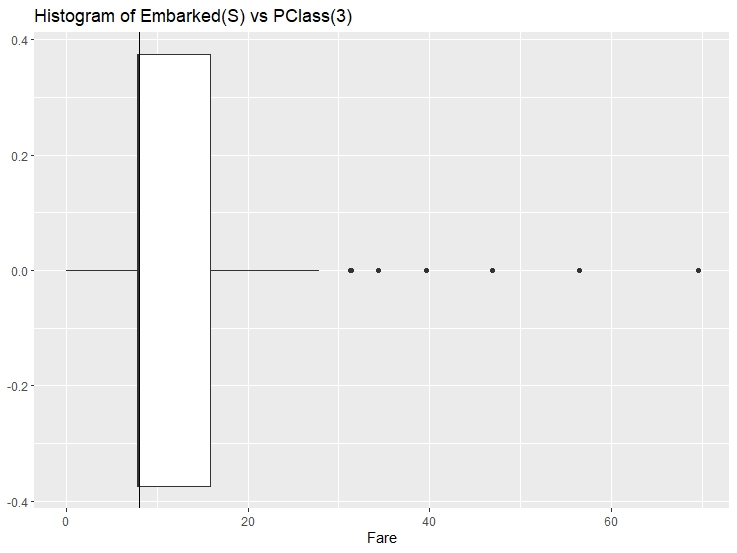
For both rows, Pclass = 1, Fare = $80, and Embarked = NA can be seen. From this information, boxplot was created as Fare vs Embarked filled with Pclass. From the box plot in *Image 12.,* Fare amount of $80 with a Pclass of 1 can be seen in the “C” embarkment location. Therefore, 2 NA values in in the “Embarked” attribute will be imputed with “C” values.

## 5.2 Fare Attribute

There was only 1 missing value in the Embarked attribute as seen in *image 4.* Though its only 1, carefully managing the NA value is critical. At the end, data needs to explain the story and not with gut feelings. More specifically, a row of missing value can be seen in *Image 12.*



***Image 12. A Row of Missing Fare Attribute***



***Image 13. Boxplot of Embarked(S) vs Pclass(3)***

From Image 13, it seems that median of fare with embarked values of S and Pclass values of 3 and median of all fare values (represented as a line in *Image 13.)* match well. Therefore, NA value in the Fare Attribute will be replaced with a median.

## 5.3 Imputing Age Attribute

There were total of 263 missing values in the Embarked attribute as seen in *image 4.* This was considered as an important factor so Imputing this attribute was a critical process.

## 

***Image 14. Imputation of the Age Attribute***

A picture containing text, receipt

Description automatically generated

***Image 14. Imputation Result***

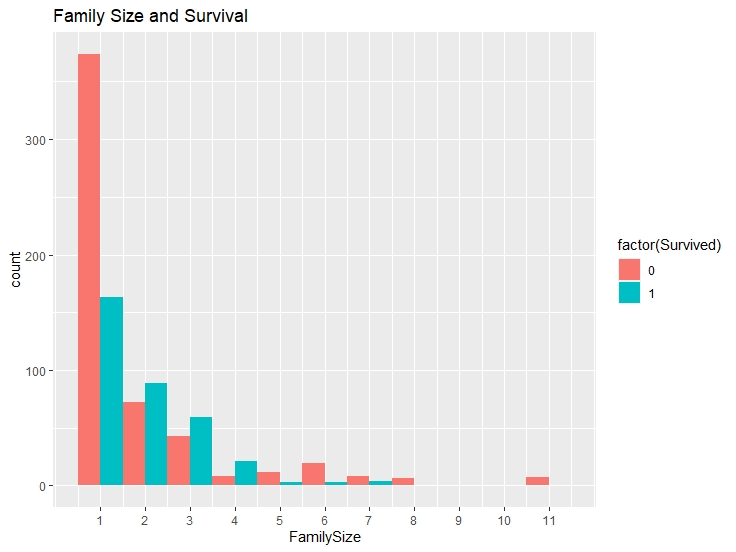
As seen in *Image 14.,* NAs were imputed and there are no NAs within the Age attribute.

## 6. Feature Engineering

After all NA values were taken care of, feature engineering was implemented to answer the following data questions.

* Did each family sink together?
* Who were more likely to survive?
  + Father?
  + Mother?
  + Child?
* Do Surnames have any relation to survival?

### 6.1 Did each family sink together?



***Image 15. Family Size vs Survival***

As seen in *Image 14,* people who did **not have any Family members** have shown **the most frequency in not surviving**, and people with family size of two, three, and four had **higher frequency of surviving than not surviving** as seen in *Image 15.*

### 6.2 Surname

Graphical user interface, text, application, email

Description automatically generated

***Image 16. Surnames***

Looking at the Name attribute in Image 16., the question was asked what information is there from this attribute. Then, the question of there might be a relation between the surname and the survival arose.

A picture containing text

Description automatically generated

***Image 17. Extracted Surnames***

As seen in *Image 17.,* surnames from the name attribute were extracted to its own column. rare surnames such as “capt”, “col”, “Donna”, “Lady, etc. were all combined into the other section.

### 

### 6.3 Father? Mother? Minor?

A picture containing graphical user interface

Description automatically generated

***Image 18. Columns of Father, Mother, and Minor***

Text

Description automatically generated with medium confidence

***Image 19. Code for Father, Mother, and Minor Feature Engineering***

As seen in *Image 18.,* columns of father, mother, and minor were created to see if there are any relation between the survival and father, mother, and minor.

### 6.4 which Gender Survived better?



***Image 19. Multivariant (sex & age) Histogram***

As seen in *Image 19.,* a multivariant histogram of sex and age is represented. The histogram infers that males age between 20 to 40 were more likely to not survive in this tragic event.

## 7. Machine Learning Models: Support Vector Machine and Random Forest

### 7.1 Support Vector Machine

A KSVM model takes the original low-dimensional data (2-D point data) and projects them into high-dimensional data (3-D point data). The input data (independent variables) from the train dataset are processed through a mapping algorithm called a kernel. Kernel algorithm will output the position of these variables into multidimensional space.

A picture containing text

Description automatically generated

***Image 20. KSVM Code***

Text

Description automatically generated

***Image 21. Result of KSVM***

KSVM model was used to predict Survived attribute as seen in *Image 20 & 21.* From the result, we can see that the training error obtained from the model was about **0.11 (11%)** and 0.21 **(21%)** cross validation error. Naturally, cross validation error will be higher because set of parameters never performs as well on subsequent data sets as it does with the original training set.

### 7.2 Random Forest

Random forest model runs many individual decision trees that combines at the end. Each decision tree outputs a class prediction, and the output with the most votes becomes the model’s prediction.

Text

Description automatically generated

***Image 22. Random Forest Model Code.***

Text

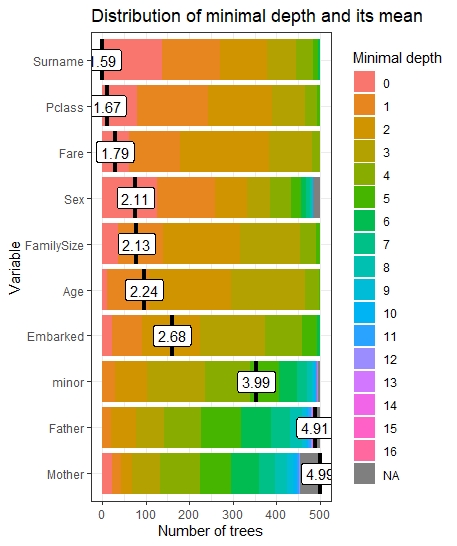
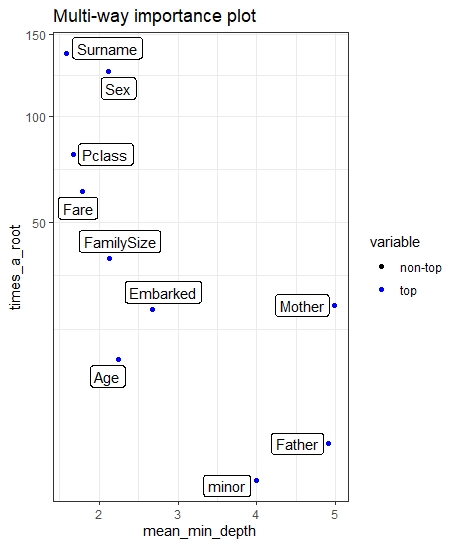
Description automatically generated

***Image 23. Random Forest Model Result.***

Random Forest model was used to predict Survived attribute as seen in *Image 21 & 22.* From the output, the model’s error rate was about 16.05%.

### 7.3 Variable importance

Relative variable importance by plotting the mean decrease across all trees.



***Image 24. Comparing input variables***

As seen in *Image 24.,* the highest relative importance out of all the variables was the surname variable. Surprisingly, the Sex variable was 5th amongst all variables. This variable was expected much higher than that.