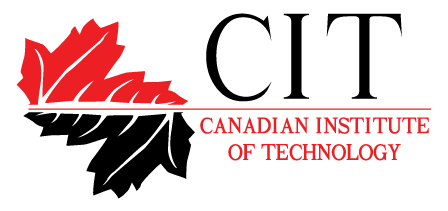
**June 2023 Data Mining & Data Analytics**



**Titanic**

**Machine Learning from**

**Disaster**

**Megi Menga, Paola Mallkas**

**Computer Engineering and Big Data, Faculty of Engineering,**

**Canadian Institute of Technology: CIT**

**Supervised by Dr. Adriola Faqolli**

### ABSTRACT

**Titanic**

### Machine Learning Study From Disaster

**Keywords**: Machine Learning, Titanic, Survival rate, Prediction accuracy

Machine learning plays an important role in the data science field nowadays. They can be used for classification problems. In this project, we are interested in understanding what kinds of people were more likely

to survive the sinking of the Titanic using different machine-learning methods.

Different predictors of passenger information were provided, and the survival chance of different passengers was predicted based on their covariates using different machine learning methods including Decision Tree, and Random Forest. Grid Search Cross-validation was used for calibrating the prediction accuracy of different methods. The Decision Tree model performs the best for our data with a prediction accuracy of about 89.75. We used R Studio for the whole analysis including cleaning the data, visualization, validation, and modeling.

### Acknowledgment

We thank Dr. Adriola Faqolli for their research assistance.

### Worked in this study

E-mail: [megi.menga@cit.edu.al](mailto:megi.menga@cit.edu.al)

paola.mallkas@cit.edu.al

Faculty of Engineering, Canadian Institute of Technology

# Introduction

The Titanic disaster occurred approximately 100 years ago, yet it continues to captivate scholars who want to comprehend and investigate how some passengers survived while others perished. The features of the passengers will be recognized in this work, and the relationship of survival chance from the tragedy will be discovered. Techniques for feature engineering will be used. The goal of this study is to obtain as trustworthy findings from raw and missing data as feasible using machine learning and feature engineering methods. Titanic, one of the most popular datasets in data science, is so exploited.

This dataset contains information about Titanic passengers, such as who survived and who did not. It was discovered that several missing and uncorrelated features reduced prediction performance. Kaggle, a popular data science website, compiled information about each Titanic passenger into a dataset and made it available in an effort to analyze the passengers. These algorithms' prediction and efficiency are heavily reliant on data analysis and the model. The paper describes an implementation that combines the advantages of feature selection and machine learning to accurately identify and distinguish travelers' age, class, cabin, and port. Based on prior knowledge and current evidence, the Decision Tree theorem and Random Forest can be utilized to generate predictions.

This project entails the use of data analytics and machine learning. The data will be analyzed using the applied algorithms, and the accuracy will be verified. Based on the performance of the indicated algorithms, Decision Tree and Random Forest, proved to be the best algorithm in this article, beating other employed algorithms for the Titanic classification problem since it acquired the highest accuracy.

# Initial Questions

When initially approaching the Titanic dataset, the following questions could be considered:

1. What factors influenced the survival rate of passengers aboard the Titanic?
2. How does the passenger class correlate with the chances of survival?
3. Did gender play a role in determining survival rates?
4. Were age and family size factors in the likelihood of surviving?
5. Did the embarkation port affect the survival rate?
6. Were there any notable differences in survival rates among different passenger categories, such as crew members, adults, and children?

As the project progresses and the analysis unfolds, these questions might evolve. For example:

1. Initially, we might look at the overall survival rate and then investigate which factors influenced it the most.
2. After examining the passenger class, we could delve into the economic and social implications of class differences and how they related to survival chances.
3. While considering gender, we could explore the "women and children first" principle and its impact on survival rates.
4. In terms of age and family size, we might analyze the effectiveness of the "survival in groups" strategy or the challenges faced by individuals.
5. The embarkation port might lead us to investigate potential differences in passenger demographics and survival training.
6. We might explore whether crew members had a higher or lower survival rate compared to passengers and whether adults or children had distinct chances of survival.

During the analysis, new questions might emerge based on the insights gained. Some potential new questions could include:

1. Did the availability of lifeboats impact the survival rate?
2. Were there any significant differences in survival rates among different deck locations?
3. How did the time of day or season of the year influence survival chances?
4. Did having a cabin or being in a certain area of the ship affect the likelihood of survival?
5. Were there any specific demographic factors that played a crucial role in survival, such as nationality or occupation?

By exploring these questions and potentially discovering new ones, a comprehensive analysis of the Titanic dataset can be conducted, providing insights into the factors that affected the survival of passengers onboard the ill-fated ship.

# Data Preparation

## Overview of the Titanic Dataset

The Titanic dataset is available for download from the Titanic machine-learning competition on Kaggle (https://www.kaggle.com/c/titanic/data). There are nine predictors, including age, sibsp (number of siblings and spouses), parch (number of parents and children), ticket (number), fare (passenger fare), and cabin (cabin number) as categorical variables and age, sibsp (number of siblings and spouses), parch (number of parents and children), ticket (number), fare (passenger fare), cabin (cabin number) as numeric variables. The outcome variable is "survival," a binary variable with values of 0 (did not survive) and 1 (did survive).

The original dataset was divided into two groups: training (891 observations) and testing (419 observations). The training set includes the outcome (or target variable) for a group of passengers as well as a variety of additional attributes such as age, gender, and so on. This is the dataset on which our prediction model must be trained. The test set, for which we must predict the now-unknown target variable using the other passenger attributes from both datasets.

The dataset for the test data is incomplete, which means that for numerous samples, one or more of the fields were unavailable and marked empty (particularly in the latter categories - age, fare, cabin, and port). However, all sample sites included at least gender and passenger class information.

## Data Discovery

First, we load the training data and look at its structure, summary, top and bottom rows.

*# Load training set*

setwd("C:/Users/paola/OneDrive/Documents/Data Analytics/Project")

train <- read.csv("train.csv")

test <- read.csv("test.csv")

str(train)

'data.frame': 891 obs. of 12 variables:

$ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...

$ Survived : int 0 1 1 1 0 0 0 0 1 1 ...

$ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...

$ Name : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...

$ Sex : chr "male" "female" "female" "female" ...

$ Age : num 22 38 26 35 35 NA 54 2 27 14 ...

$ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...

$ Parch : int 0 0 0 0 0 0 0 1 2 0 ...

$ Ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803

$ Fare : num 7.25 71.28 7.92 53.1 8.05 ...

$ Cabin : chr "" "C85" "" "C123" ...

$ Embarked : chr "S" "C" "S" "S" ...

There are 891 passengers and 11 attributes (PassengerId is just an index not an attribute of a passenger).

The attributes are:

* **PassengerId:** An identifier for each passenger.
* **Survived**: integer, binary indicator (Survived = 1) and the target outcome or dependent variable we are to predict.
* **Pclass**: The passenger class, integer, an ordinal variable, representing the socio-economic status

of the passenger

* **Name**, Factor w/ 891 levels (one level per passenger).
* **Sex**, Factor with two levels: “female”, “male”.
* **Age**, numerical, has 177 missing values coded as NA.
* **SibSp**, integer, an ordinal variable for the number of siblings or spouses.
* **Parch**, integer, an ordinal variable for the number of parents or children.
* **Ticket**, Factor w/ 681 levels.
* **Fare**, numerical, is in Pounds Sterling, a proxy for wealth or social status.
* **Cabin**, Factor w/ 147 levels, has 687 missing values.
* **Embarked**, Factor w/ 3 levels: “C”, “Q”, and “S” for the port of embarkation (Cherbourg, Queenstown, and Southhampton), has 2 missing values

*# Preliminary summary*

summary(train)

PassengerId Survived Pclass Name Sex Min.: 1.0 Min. :0.0000 Min. :1.000 Length:891 Length:891 1st Qu.:223.5 1st Qu.:0.0000 1st Qu.:2.000 Class: character Class :char Median :446.0 Median :0.0000 Median :3.000 Mode :character Mode :char Mean :446.0 Mean :0.3838 Mean :2.309 3rd Qu.:668.5 3rd Qu.:1.0000 3rd Qu.:3.000 Max. :891.0 Max. :1.0000 Max. :3.000

Age SibSp Parch Ticket Fare Min.:0.42 Min :0.000 Min.:0.0000 Length:891 Min.:0.00

1st Qu.:20.12 1st Qu.:0.000 1st Qu.:0.0000 Class :character 1st Qu.:7.91

Median :28.00 Median :0.000 Median :0.0000 Mode:character Median:14.45

Mean:29.70 Mean :0.523 Mean :0.3816 Mean:32.20 3rd Qu.:38.00 3rd Qu.:1.000 3rd Qu.:0.0000 3rd Qu.:31 Max. :80.00 Max. :8.000 Max. :6.0000 Max. :512.33

NA's :177

Cabin Embarked

Length:891 Length:891

Class :character Class :character

Mode :character Mode :character

NA's: 2

This first summary of our data is not very useful and helps us determine how to proceed with data pre-processing, converting appropriate variables into the categorical format, cleaning up variables, and imputing missing values as needed. A large number of missing values are in Age. The 2 missing values in Embarked, we are going to analyze later which is the most appropriate value to be filled.

head(train)

tail(train)

*# Look at number of people who survived*

table((train$Survived)

prop.table(table(train$Survived))

We see that in the training set, 342 passengers survived, while 549 died. While the proportion shows that 38% of passengers survived the disaster in the training set.

## Data Preprocessing

### 3.3.1 Missing Data Imputation

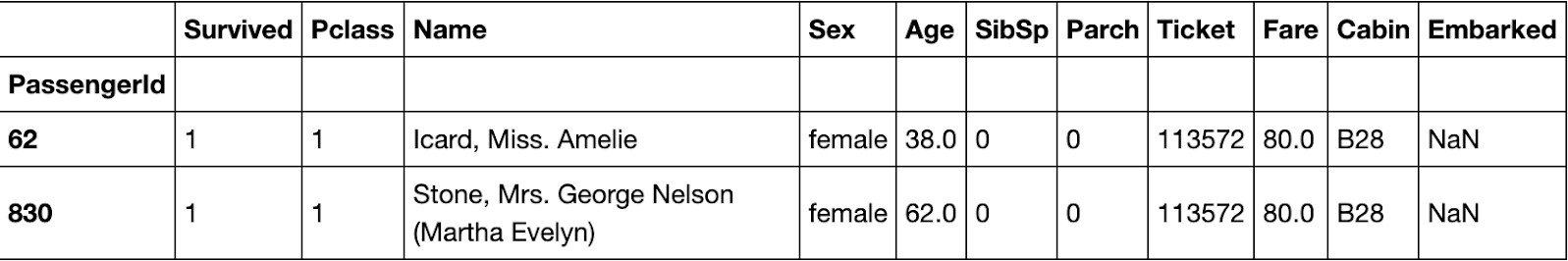
Before proceeding, it is essential to conduct a preliminary check on the prepared data obtained from a Kaggle competition to identify any missing information. Once the missing data has been identified, the next step involves determining the appropriate approach for handling it.

We ran through some simple missing data codes in R and found that there is missing data.

Firstly, we began to work with the missing data in Age. There are about one-third of the missing data in the variable Age, and it’s impossible to drop all of them. From the Training Data, we are missing the age of 177 passengers of the 891 in the set. From the Test Data, we are missing the age of 86 passengers. For the missing ages, it has been a common practice to use the median age, to replace missing age values. For both datasets, the returned value was 28 for the train\_data and 27 for the test\_data, so we replaced the missing values with these numbers.

In the test data, there was one instance where the fare was missing. We created a subset of the train\_data and found that there was 3rd Class passenger, named Thomas Storey, who was a 60 year old male, who embarked from Scottland that had a missing fare value. The rounded mean fare for 3rd class passengers that embarked from Scottland was 7.90, and we replaced it with that value.

After that, we started addressing the two missing Embarked points from the training dataset.



**Table 1**. Two missing values of Embarked in the training dataset

We discover that the two missing data points in the training set share some characteristics: They were both women from Pclass 1 who had survived the accident. Furthermore, their tickets were identical, and they shared the same Cabin. This suggests that we should focus on individuals who have similar conditions and have the most likely Embarked values. The two missing data points may have the following characteristics: survived a disaster, female with the same fare resided in the same cabin at Pclass. We created a box plot of travelers who lived in Pclass 1 and paid $80 for their tickers. Both training and testing data indicate that these two missing data points are most likely related to Embarked C. Also, if we plot female travelers who paid $80 for their tickers, we get the same pattern, which confirms our hypothesis. As a result, we filled the missing values of the two passengers in Embarked to “C”.

Finally, in order to utilize the high efficiency of the given predictors to get the most accurate assumption of the missing values, we used feature engineering to create some variables based on existing variables to achieve our goal of missing data imputation.

### 3.3.2 Feature Engineering

Firstly, we created a variable named “FamilySize” by summing two variables SibSb and Parch that indicate the number of family members the passenger is traveling with. Basically, we added the number of siblings, spouses, parents, and children the passenger had with them, plus one for their own existence of course, and have a new variable indicating the size of the family they traveled with.

Also, we considered a large family having difficulty getting to lifeboats together, but perhaps some families had more difficulty than others. We could try to extract the passengers' surnames and group them to locate families, but a common surname like Johnson may have a few more unrelated persons aboard. In fact, there are three Johnsons in a family with size 3, and another three Johnsons who are most likely unrelated and are all traveling alone. However, combining the surname with the family size might solve this difficulty. On such a small ship, no two Johnson families should have the same FamilySize variable.

So we begin by extracting the passengers' surnames. To ensure uniqueness among families on the small ship, we combine the passengers' surnames with their family sizes. By extracting the last names from the names and appending the family size, we create a new identifier called FamilyID. To handle cases of small families (size two or less), we assign them the label "Small" as their FamilyID. This approach helps us identify and analyze different family groups on the ship, allowing us to examine the dynamics and potential challenges faced by families during the panic.

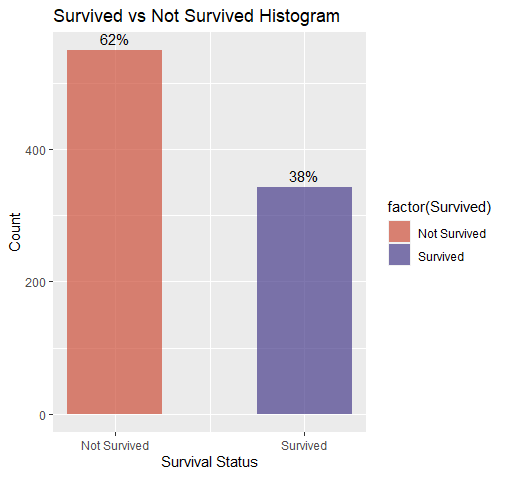
Lastly, we would like to extract some useful information from the name variable, which could not play an important role in building models itself. There are different kinds of titles in names, which could be divided into different groups. The majority of titles are Mr., Miss., Mrs., and Master. The rest of the titles are Dr. and nobility titles (Jonkheer, colonel, etc.) of small numbers, thus we combined them into one group and used it with the other four groups to create the “Title” variable with five levels.

# Exploratory Data Analysis

To begin with, our exploration of the data will focus on identifying patterns related to survival. In the provided dataset, there is a file called "gender\_submission.csv" which assumes that all females survived and all males perished. However, we will examine the training data to verify the accuracy of this assumption. In the test dataset, there were 314 women and 577 men. Among the women, 233 survived, accounting for a survival rate of 72.2%, while among the men, 109 survived, resulting in a survival rate of 18.89%. Hence, it can be concluded that the assumption of all women surviving and all men dying is incorrect. (Code in R)

## Outcome variable

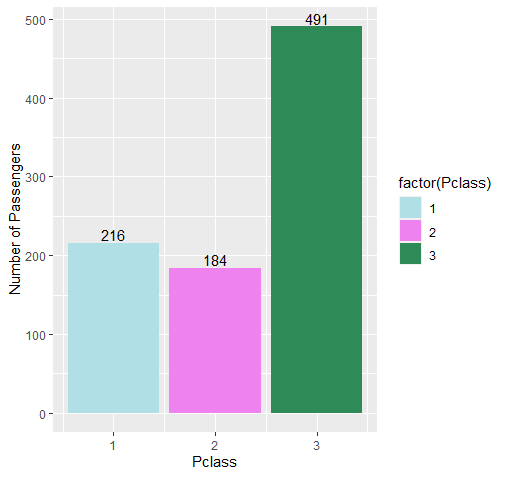
We drew the histograms of the outcome variable and all the predictors. The histogram (figure 1) of the survival or not comparison based on the training dataset (0 is not survived and 1 is survived) shows that **more than 60% of the passengers had died.**



**Figure 1.** Histogram of Survived or not Survived

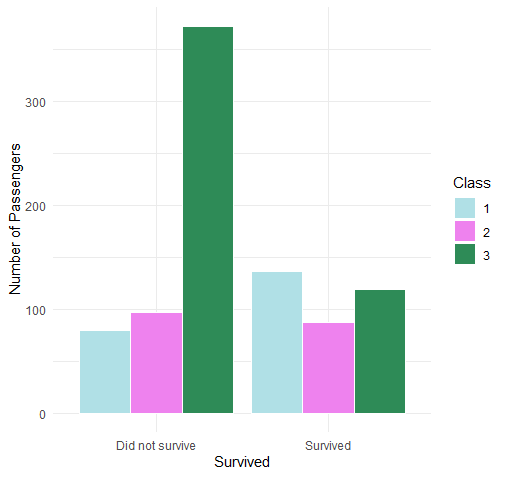
## Survival Rate by ticket classes

Based on Figure 2 (a), we can clearly see that most of the passengers came from the 3rd class (bought the cheapest ticket to board the Titanic), followed by the 1st and 2nd class. Surprisingly, the 1st class passengers are a bit more than the 2nd class.



**Figure 2.a.** Histogram of three classes

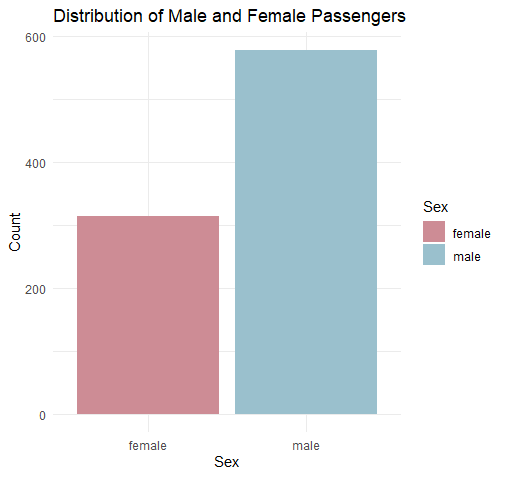
Figure 2.b. makes it very evident that passengers from the first class have a higher chance of surviving, while those from the third class have a lower chance of surviving.



**Figure 2.b.** Histogram of three classes between survived or not

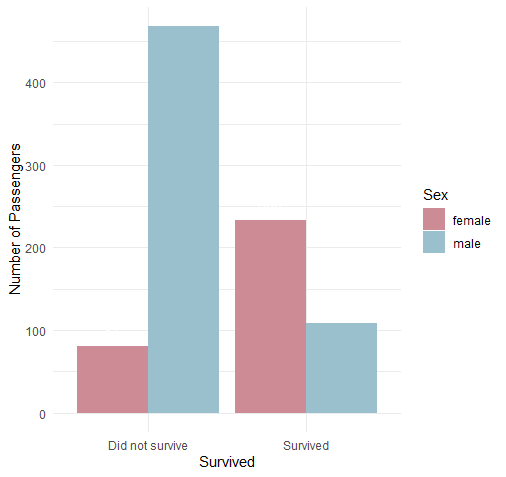
## Sex

Approximately 65% of the tourists were male while the remaining 35% were female.



**Figure 3(a).** Histogram of male and female

In Figure 3(a), the purple bar presents the female and it shows that females are more likely to survive.

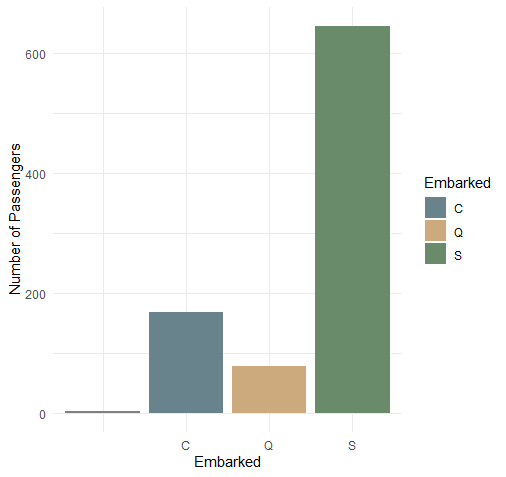


**Figure 3(b).** Histogram of male and female between survived or not

Nonetheless, the percentage of female survivors was higher than the number of male survivors.**More than 80% of male commuters died, as compared to around 70% of female commuters.**

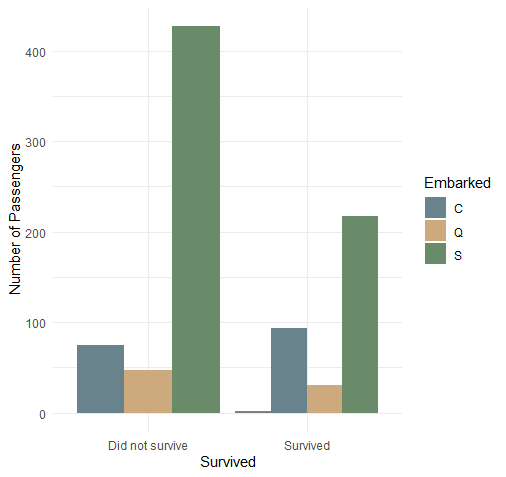
## Embarked

Embarked implies where the traveler mounted from. There are three possible values for Embark — Southampton, Cherbourg, and Queenstown. More than 70% of the people boarded from Southampton. Just under 20% boarded from Cherbourg and the rest boarded from Queenstown.  In figure 4(a), the blue bar represents Cherbourg, the brown bar represents Queenstown, the green bar represents Southampton.



**Figure 4 (a).** Histogram of Embarked

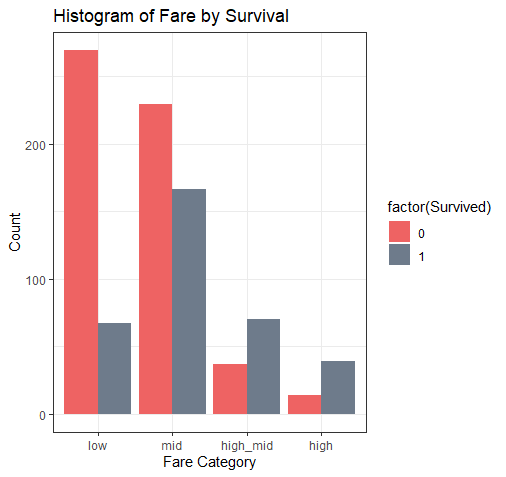
There are more people dead compared to people who survived generally, however, from Figure 4 (b), we can see that people who embarked from Cherbourg are more likely to survive.



**Figure 4 (b).** Histogram of Embarked

## Fare

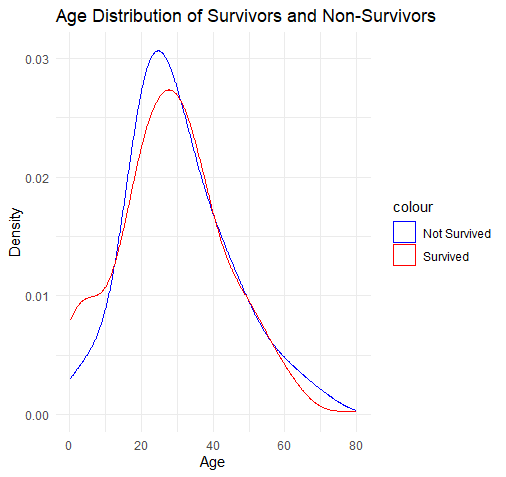
By splitting the fare amount into four categories, it was obvious that there was a strong association between the charge and survival. **The higher a tourist paid, the higher would be his chances to survive.**



**Figure 5.** Survival by Fare

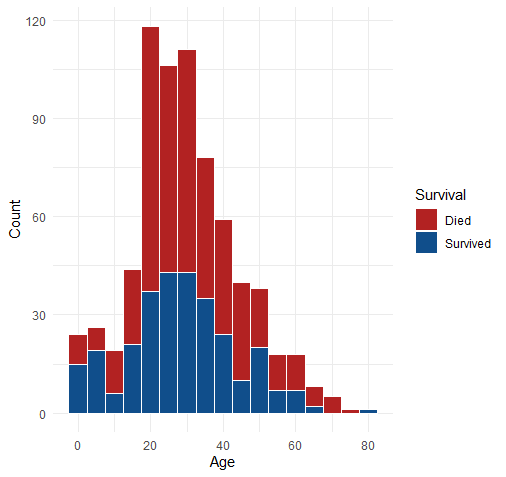
## Age

The age distribution of individuals who did not survive is represented by the blue line in Figure 5, while the age distribution of those who survived is represented by the red line. By observing the distribution, it becomes apparent that the center of the age distribution for those who did not survive is shifted more towards the left compared to the overall age group distribution. This indicates that a higher proportion of young adults died in the given population. Additionally, there is a small peak in the age range of 0-10 for the survivors, suggesting that a greater number of children or infants managed to survive.



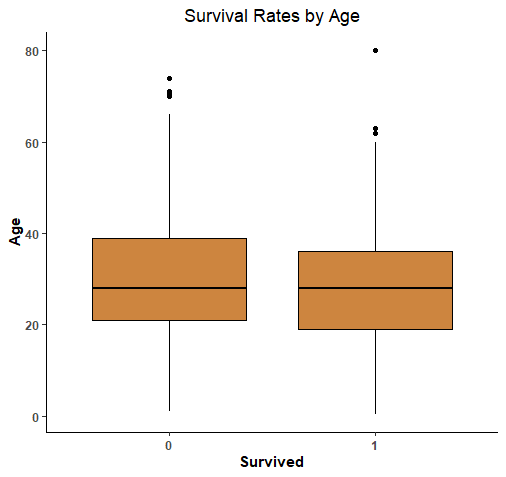
**Figure 5 (a).** Distributions of age between survived or not survived

The youngest traveler onboard was aged around two months and the oldest traveler was 80 years. The average age of tourists onboard was just under 30 years. Clearly, **a larger fraction of children under 10 survived than died** or every other age group, the number of casualties was higher than the number of survivors. More than 140 people within the age group 20 and 30 were dead as compared to just around 80 people of the same age range sustained.



**Figure 5 (b).** Distributions of age between survived or not survived

Children less than 5 years had higher survival chances. Passengers aged 20-40 were more likely to die. Passengers aged about 65 - 75 had an almost 0 survival chance. One passenger aged 80 years survived.

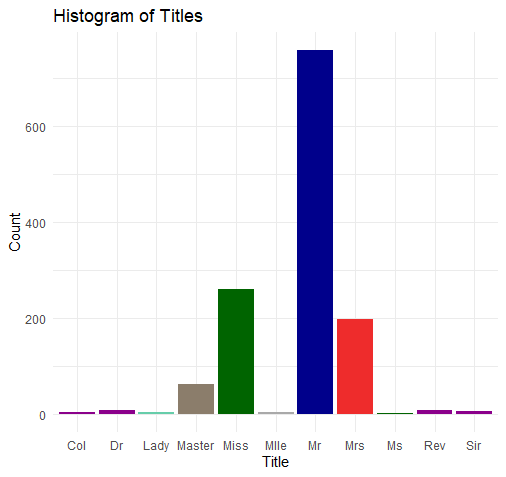


**Figure 5 (c).** Survival Rates by Age

Passengers who survived seems to have a lower median age.

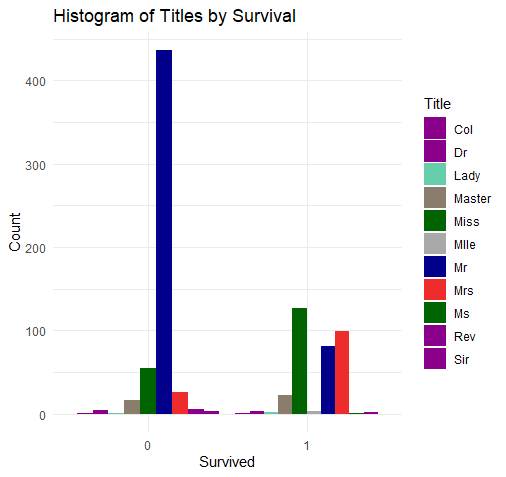
## Title

In Figure 7, the Blue bar is Mr., the Green bar is Miss, 2 the Red bar represents Mrs., the Grey bar represents Master, and the Purple bar is Dr & nobility titles.



**Figure 7 (a).** Histogram of Titles

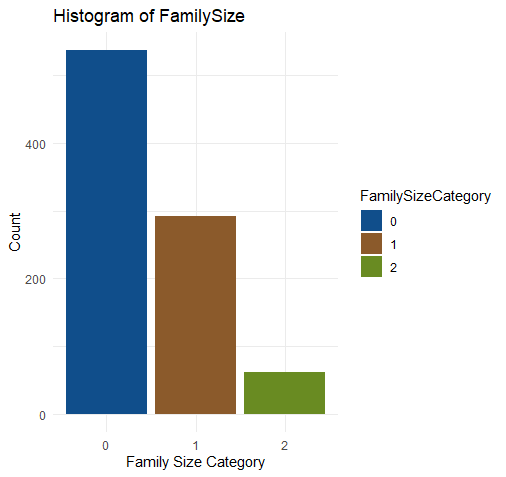
We can draw the conclusion that compared to Mr. and Dr. with nobility titles, Mrs., Miss, and Master (boys younger than 18 years old) are more likely to survive. The conclusion gave us that females and children are more likely to survive

****

**Figure 7 (b).** Histogram of Titles between survived or not

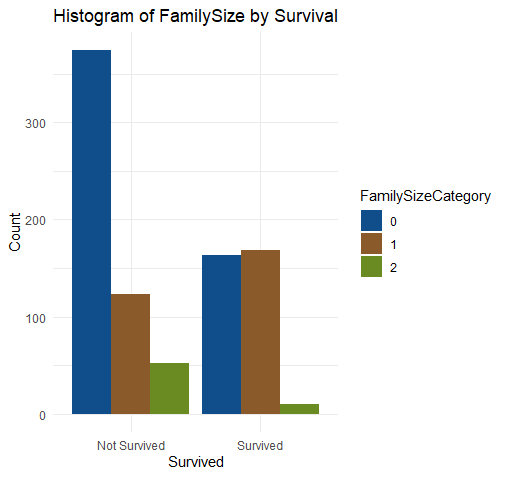
## Family Size

In Figure 8(a), 0 (blue bar) represents people with 0 or 1 family member, 1 (brown bar) represents people with 2-4 family members, and 2 (green bar) represents people with 4 members or more.



**Figure 8 (a).** Histogram of Family Size

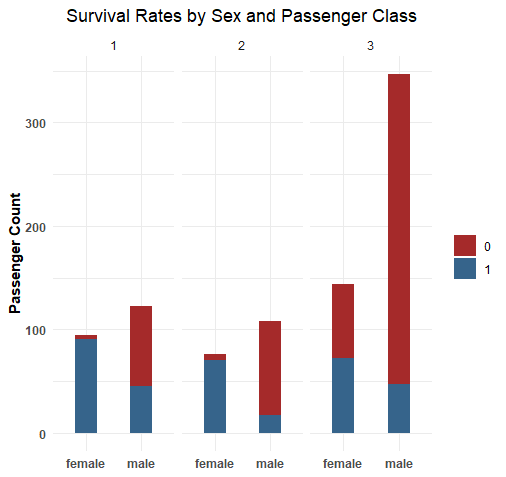
From Figure 8 (b), we can see that people with 2-4 family members are more likely to survive compared to other groups



**Figure 8 (b).** Histogram of Family Size between survived or not

## Survival Rate by Sex and Pclass

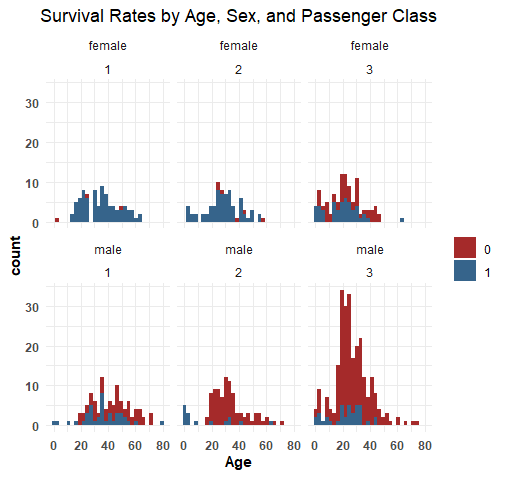
Finally, we will investigate how different variables interacted with each other and with the survival rate.



**Figure 9.** Survival rates by Sex and Pclass

Females in 1,2 and 3 class had 90%, 90%, and 50% survival chances respectively compared to 45%, 20%, and 20% respectively for their male counterparts.

## Survival Rates by Age, Sex, and Passenger Class



**Figure 10.** Survival rates by Age, Sex, and Pclass

## Conclusion

Graphs are powerful visualization tools. In the Titanic data, age, sex, and passenger class were important predictors of survival. Identifying such predictors is important for predictive modeling. From the exploratory data analysis, we concluded that in general there are more people dead than survived. The first class is more likely to survive and females are more likely to survive. People who embarked from Cherbourg are more likely to survive. Females and children are more likely to survive than males and nobility titles. People with 2-4 family members are more likely to survive compared to other family groups.

## Evolving Questions

Based on the questions and results from the Titanic dataset performed in R, we discovered that there are 891 observations (passengers) in the dataset, and 12 variables (columns). Also, the survival rate among the passengers is approximately 38.4%. The average age of the passengers is approximately 29.7 years. There are 314 female passengers and 577 male passengers. The highest fare paid by a passenger is $512.3292. Among the passengers, 168 embarked from port C, 77 from port Q, and 644 from port S.

Regarding the number of siblings/spouses aboard (SibSp), 608 passengers had none, 209 had one, 28 had two, 16 had three, 18 had four, 5 had five, and 7 had eight. In terms of parents/children aboard (Parch), 678 passengers had none, 118 had one, 80 had two, 5 had three, 4 had four, 5 had five, and 1 had six. These conclusions provide insights into various aspects of the Titanic dataset, such as survival rates, passenger demographics, ticket fares, and family relationships aboard the ship.

Based on the analysis of the Titanic dataset and the results of the scripts in R, the following conclusions can be drawn:

1. **Survival Rate by Passenger Class:** Passengers in higher classes (Pclass 1) had a higher survival rate (63%) compared to those in lower classes (Pclass 2: 47%, Pclass 3: 24%). Passenger class appears to be a significant factor in determining survival.
2. **Family Size and Survival:** Passengers with smaller family sizes (spouse, siblings, parents, or children) had a higher survival rate. Those with no family members aboard had a survival rate of 30%, while those with a family size of 3 had a survival rate of 72%.
3. **Embarked Port and Survival:** Passengers who embarked from the "C" port (Cherbourg) had a higher survival rate (55%) compared to those from the "Q" port (Queenstown: 39%) and the "S" port (Southampton: 34%).
4. **Age and Survival:** The density plot of age by survival shows that there is a higher concentration of survivors among children and young adults. However, the analysis also reveals missing age values, which could impact the accuracy of this observation.
5. **Title and Survival:** Passengers with certain titles, such as "Mrs" (79%) and "Miss" (70%), had a higher survival rate compared to others, including "Mr" (16%).
6. **Factors Affecting Survival:** Considering multiple factors together, the combination of passenger class, sex, and age appears to have a significant influence on survival rates. Females in higher classes had the highest survival rate (97.8%), while males in lower classes had the lowest survival rate (13.5%).
7. **Fare Category and Survival:** Passengers who paid higher fares had a higher survival rate compared to those who paid lower fares. The fare category analysis indicates a positive correlation between fare and survival.

# Data Mining Model

Decision trees can be a suitable choice for the Titanic dataset for several reasons, and they offer advantages over random forests and other data mining algorithms. Here are some reasons why decision trees might be a good fit:

1. **Interpretable and Explainable:** Decision trees provide a transparent and interpretable model. The resulting tree structure can be easily visualized and understood, allowing you to interpret the rules and decisions made by the algorithm. This interpretability can be crucial in domains where understanding the reasoning behind predictions is important.
2. **Feature Importance:** Decision trees inherently provide feature importance measures. By examining the tree, you can identify which features are the most influential in making decisions. This information can help in feature selection and understanding the underlying relationships between the predictors and the target variable. Based on the accuracy we achieved, we can say that the more important features of our dataset in determining whether a passenger survives are gender, class, and new variables such as title, family size, and family id.
3. **Handling Missing Values:** Decision trees can handle missing values in the dataset without requiring imputation. They split the data based on available values in the features, allowing the algorithm to make decisions even when some data points have missing values. This can be advantageous in cases where missing values are prevalent in the dataset. In the case of our dataset, we had some missing values for the age, fare, and cabin attributes as well as blank values for the embarked attribute.
4. **Fast Training and Prediction:** Decision trees are relatively fast to train and make predictions compared to more computationally intensive algorithms. They have a time complexity that is linear with the number of training instances, making them efficient for large datasets. This speed can be advantageous when working with datasets that have a substantial number of observations.
5. **Limited Preprocessing:** Decision trees do not require extensive data preprocessing steps, such as scaling or normalization of features. They can handle both categorical and numerical features directly without the need for encoding or transformation. This simplicity can save time and effort in the data preparation phase.

While decision trees have their advantages, it's important to consider that random forests, an ensemble method built upon decision trees, can offer improvements in certain scenarios. Random forests combine multiple decision trees to reduce overfitting, increase accuracy, and handle more complex datasets.

They can provide better generalization and robustness by averaging predictions from multiple trees. However, in our specific case, if the decision tree model already outperforms the random forest model, it suggests that the complexity introduced by the ensemble method may not be necessary for this dataset, and a simpler decision tree model is sufficient.

# Data Mining Evaluation

## Data Mining Technique Question

One question related to the decision tree technique that we have chosen to use is: *"How effectively can a decision tree model capture and represent the complex relationships between different attributes/features in the Titanic dataset to predict passenger survival?"* This question addresses the capability of the decision tree algorithm to handle the dataset's characteristics, such as categorical and numerical variables, missing values, and potentially nonlinear relationships.

It sets the foundation for exploring the strengths and limitations of decision trees as a data mining technique for this particular dataset.

The decision tree algorithm demonstrates the potential to effectively capture and represent the complex relationships between different attributes/features in the dataset. By recursively splitting the data based on features with the highest information gain, decision trees can identify relevant patterns and create branches that reflect distinct survival outcomes. This allows the models to make predictions based on attribute values and follow different paths through the tree structure.

The accuracy achieved by the decision tree models suggests that they have successfully captured significant patterns and relationships present in the Titanic dataset. Furthermore, the use of the random forest algorithm provides additional insights into the importance of each variable in the predictions. The random forest model, with an accuracy of 76.07%, identifies the newly created variables as influential factors in predicting passenger survival. This highlights the effectiveness of the decision tree-based approaches, including the New Decision Tree model, in capturing and utilizing additional information derived from the dataset to improve prediction accuracy.

Overall, based on the achieved accuracy and the ability to capture complex relationships, it can be concluded that the decision tree models, including the New Decision Tree model, exhibit a promising capability to predict passenger survival on the Titanic dataset.

Another question related to the decision tree technique is: "What are the most influential attributes/features in the decision tree model for predicting passenger survival in the Titanic dataset?" This question focuses on identifying the key factors that play a significant role in determining the survival outcome according to the decision tree algorithm.

By analyzing the importance or relevance of different attributes, we can gain insights into the factors that have the strongest influence on predicting survival and understand their relative importance in the model's decision-making process. As suggested by the random forest model as well as by the New Decision Tree model, we can conclude that the newly created variables such as Title and Family ID have a high impact on determining the passengers’ survival status. Nonetheless, other variables such as Gender, Class are found to be quite influential.

# Hypothesis

We have tried testing several hypotheses to gain insight if there exists a relationship/correlation between different variables of the titanic dataset:

1. **Determining if there is a relationship between PClass and Survived**

To do so we raise the Null hypothesis (H0) which states that there is no relationship between PClass and Survived as well as the alternative hypothesis (Ha) which states that there is some relationship between PClass and Survived. The result we obtained, 0.3398174, is the  correlation coefficient between the variables "Pclass" and "Survived" in the Titanic dataset.

A correlation coefficient of 0.3398174 suggests a moderate positive relationship between the variables "Pclass" and "Survived" in the Titanic dataset. The correlation coefficient ranges from -1 to 1, with 0 indicating no correlation, values close to 1 indicating a strong positive relationship, and values close to -1 indicating a strong negative relationship. In this case, a value of 0.3398174 suggests that there is a moderate positive relationship between the passenger class ("Pclass") and the survival outcome ("Survived") in the Titanic dataset.

1. **Determining if there is a relationship between Sex and Survived**

To do so we raise the Null hypothesis (H0) which states that there is no relationship between Sex and Survived as well as the alternative hypothesis (Ha) which states that there is some relationship between Sex and Survived. The result we obtained, 0.3398174, is the correlation coefficient between the variables "Sex" and "Survived" in the Titanic dataset.

The correlation coefficient of 0.5409359 indicates a moderate positive relationship between the variables "Sex" and "Survived" in the Titanic dataset. This suggests that there is a tendency for female passengers to have a higher likelihood of survival compared to male passengers.

1. **Determining if there is a relationship between Cabin and Survived**

To do so we raise the Null hypothesis (H0) which states that there is no relationship between Cabin and Survived as well as the alternative hypothesis (Ha) which states that there is some relationship between Cabin and Survived. The result we obtained, 0.5200759, is the correlation coefficient between the variables "Cabin" and "Survived" in the Titanic dataset.

The correlation coefficient of 0.5200759 between the variables "Cabin" and "Survived" in the Titanic dataset suggests a moderate positive relationship. However, it is important to note that correlation coefficients are typically used to measure the strength and direction of linear relationships between continuous variables. Since "Cabin" and "Survived" are categorical variables, the interpretation of the correlation coefficient may not be straightforward.

In this case, the correlation coefficient might indicate that certain cabin categories or patterns are associated with a higher likelihood of survival. However, further analysis and consideration of other factors are necessary to fully understand the relationship between "Cabin" and "Survived" in the Titanic dataset.

In all these cases we have calculated chi-squared and correlation coefficient. The chi-squared statistic measures the discrepancy between the observed frequencies in the contingency table and the frequencies that would be expected if the variables were independent. In this case, a smaller chi-squared statistic suggests a weaker relationship or less association between the variables.

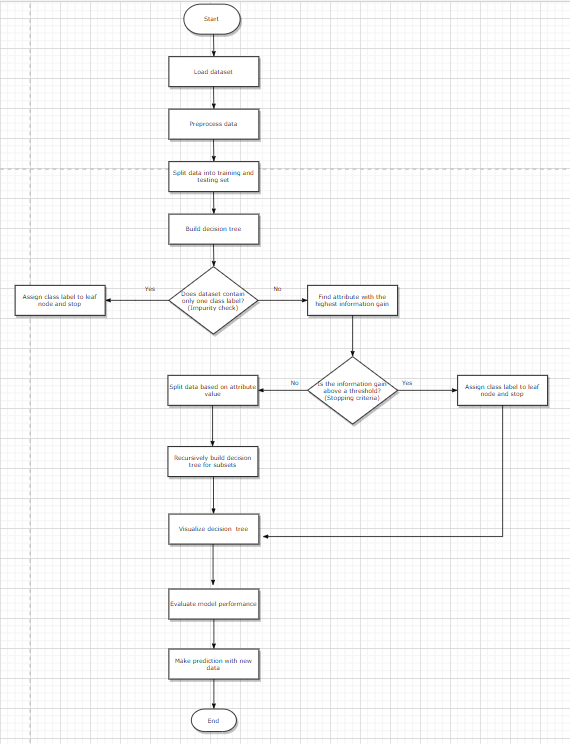
To determine whether this chi-squared statistic is statistically significant and indicates a relationship, we need to compare it to the critical value from the chi-squared distribution with appropriate degrees of freedom. The degrees of freedom for the chi-squared test in this case depend on the dimensions of the contingency table (in this case, 3-1 = 2).

If the obtained chi-squared statistic is greater than the critical value, we can reject the null hypothesis of independence and conclude that there is evidence of a relationship between the variables. However, if the obtained chi-squared statistic is smaller than the critical value, we would fail to reject the null hypothesis, indicating that there is no significant relationship between the variables.

In all the cases tested above, the obtained chi-squared statistic is greater than the critical value. Therefore, we can reject the null hypothesis of independence and conclude that there is evidence of a relationship between the variables.

# Flowchart

The provided flowchart illustrates the sequential process of building a decision tree model for the Titanic: Machine Learning from Disaster dataset. The flow starts with loading the dataset and preprocessing the data, followed by splitting it into training and testing sets. The decision tree is then constructed by recursively making decisions based on attributes with high information gain. At each node, an impurity check is performed to determine if the subset contains only one class label. If not, the process continues by selecting the attribute with the highest information gain and splitting the data accordingly. This recursive process continues until stopping criteria, such as reaching a subset with a single class label or information gain falling below a threshold, are met. The resulting decision tree is visualized, and the model's performance is evaluated using appropriate metrics. Later on, it can be used to make predictions with new data from the Titanic dataset.

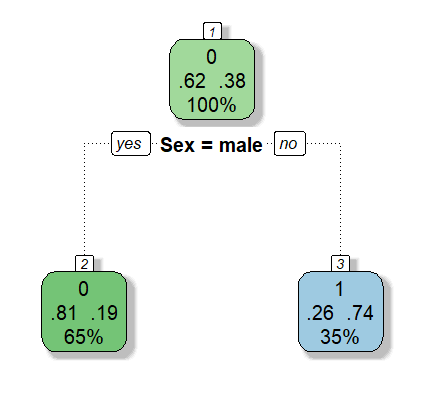


# Decision Tree

The decision tree algorithm is a popular and intuitive machine learning technique used for both classification and regression tasks. It is a supervised learning algorithm that learns simple decision rules from the training data and builds a tree-like model of decisions and their possible consequences.

In a decision tree, each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents an outcome or prediction. The algorithm partitions the data recursively based on the selected features, aiming to create homogeneous subsets with similar target variables. This process continues until the tree is fully grown, or a stopping criterion is met.

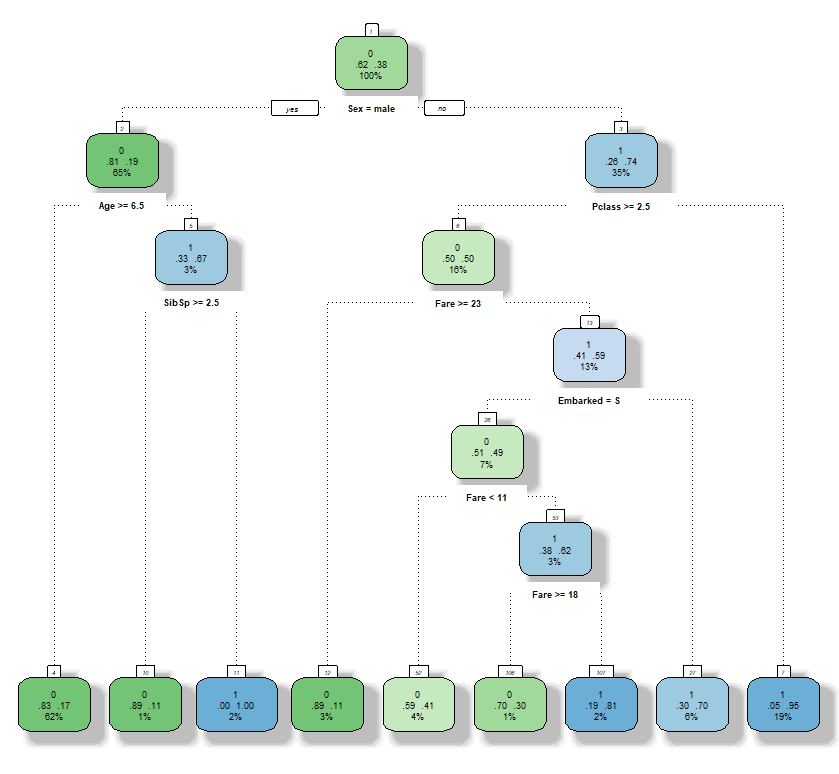
In our scenario, the decision tree algorithm has been utilized to predict the "Survived" variable in the Titanic dataset. First, we try to build a prediction model based only on one attribute, in this case Sex. The following decision tree displays our gender model.



The top node represents the overall survival rates: 62% of passengers died, while 38% survived. The number above the proportions shows the voting outcome at this node (we decided that everyone would die, coded as zero), and the number below indicates the population proportion in this node (100% at the top level).

Moving down the tree, we encounter subsequent nodes. If the passenger is male (based on the boolean choice below the node), we go left; if female, we go right. The survival proportions align with what we found in our previous analysis. For male passengers, only 19% survived, resulting in a vote for everyone in that bucket (representing 65% of passengers) to die. On the other hand, in the female bucket, where most passengers survive, the vote is in favor of survival. This decision tree precisely represents our gender model.

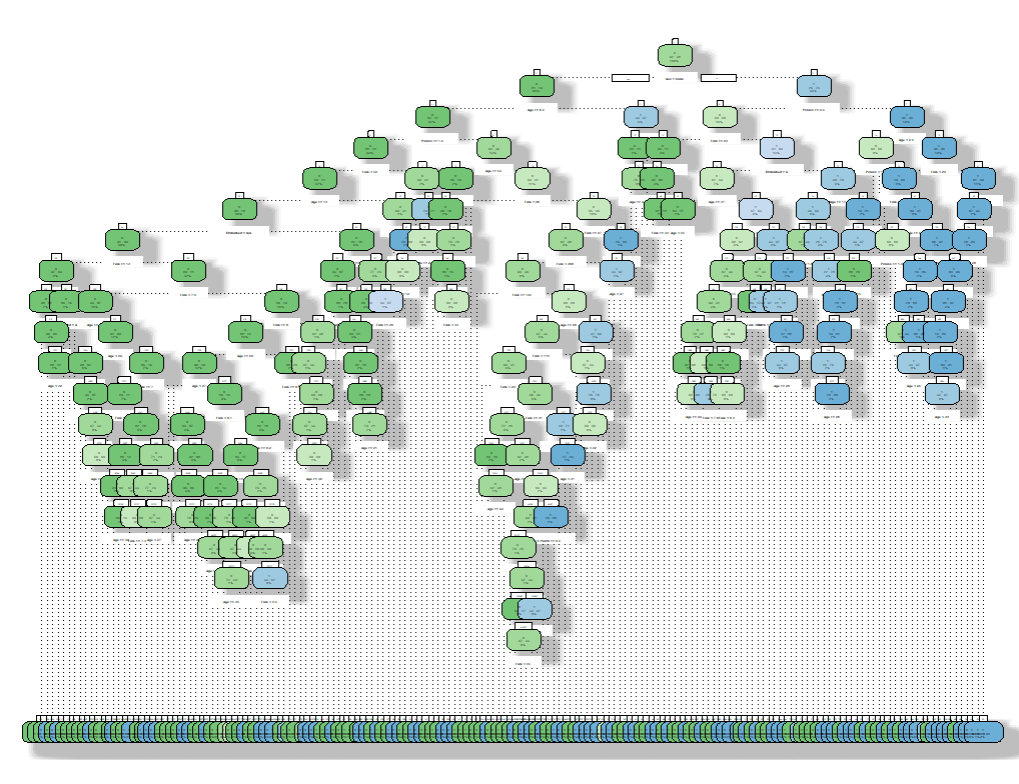
Secondly, we improve our decision tree by adding other attributes such as Pclass + Sex + Age + SibSp + Parch + Fare + Embarked to determine whether a passenger survives:



The discovered decisions in the current analysis are more extensive compared to our previous decision tree. Decisions have been identified for variables such as SibSp (number of siblings/spouses aboard) and even for the port of embarkation, which we didn't explore before. Additionally, on the male side, it has been observed that children below the age of six have a higher likelihood of survival, despite the relatively low number of such passengers.

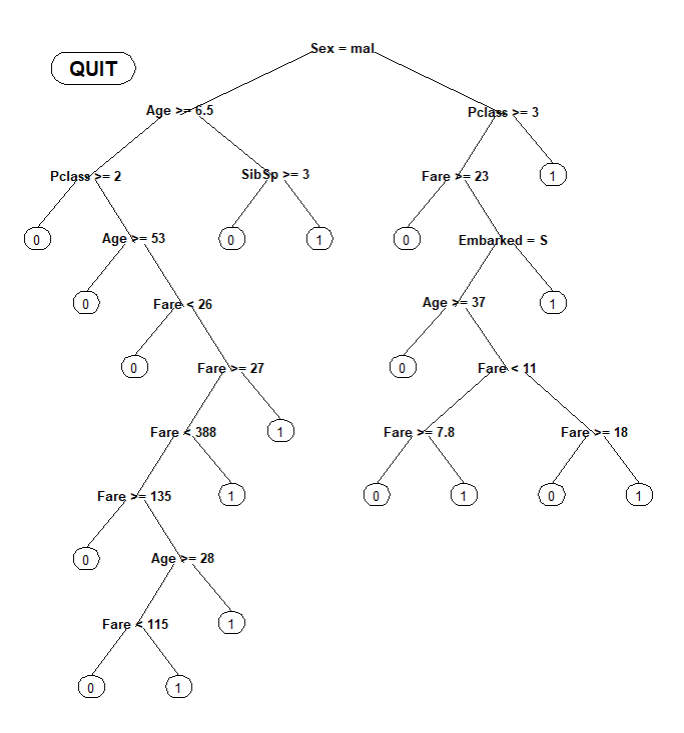
Thirdly, we proceed to display a fully grown decision tree by using:

* minsplit=2 which sets the minimum number of observations required in a node for further splitting. Nodes with fewer observations than this threshold will not be split further.
* cp=0 which sets the complexity parameter (cp) for pruning the tree. The complexity parameter controls the trade-off between the complexity (size) of the tree and its accuracy. A smaller value of cp leads to a more complex tree, while a larger value results in a simpler tree.

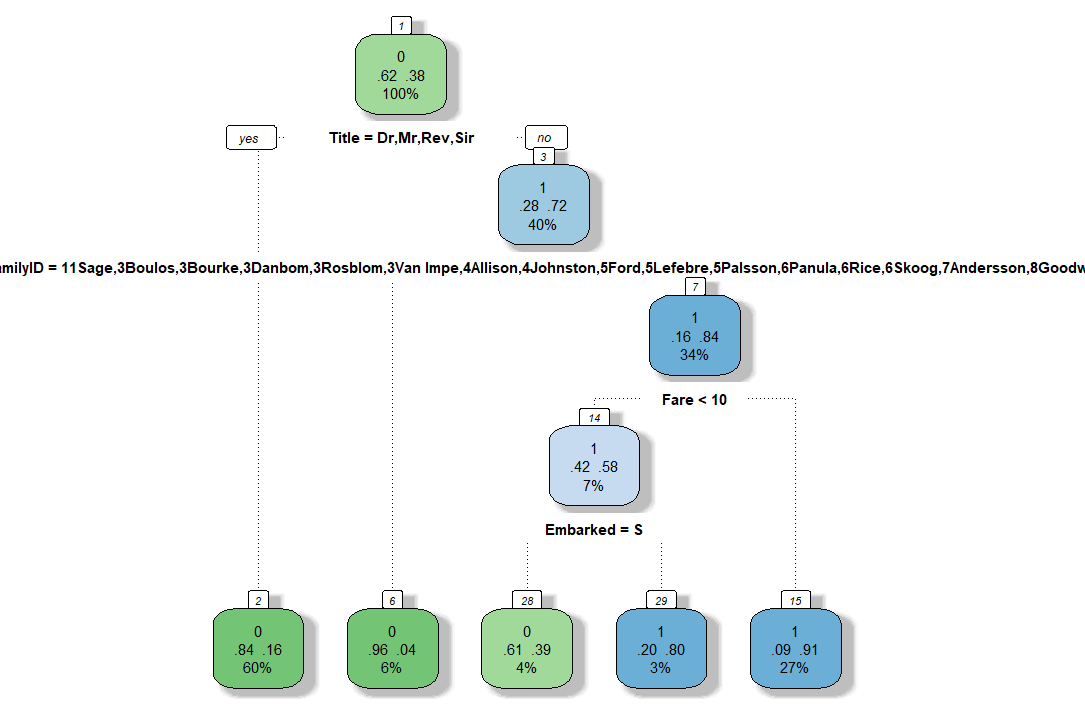


As we can see the complexity of the decision tree has increased. Furthermore, it is harder to read and not visually appealing.

In order to be able to determine which nodes we want to keep in the decision tree we can try building a manually editable decision tree. This way an interactive version of the decision tree will appear and we can manually trim it just by clicking on the nodes that we want to remove.



Next, after modifying the existing attributes of our dataset we were able to create 3 new variables such as Title, FamilySize, and FamilyID. Therefore, we will use these new variables to build a new decision tree. It will be based on Pclass + Sex + Age + SibSp + Parch + Fare + Embarked + Title + FamilySize + FamilyID attributes to determine whether a passenger survives:



**Information gain**

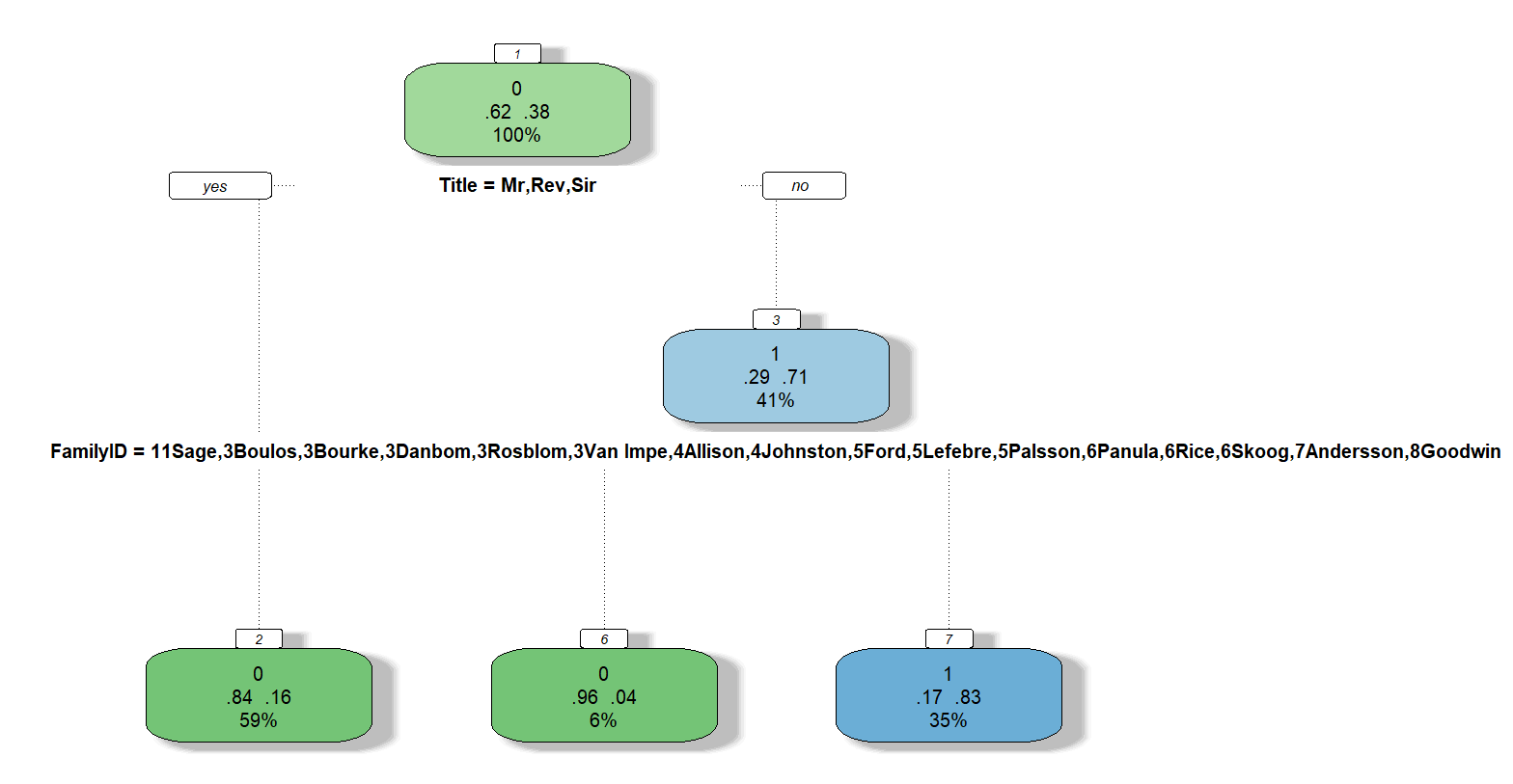
Information gain is a concept used in decision tree algorithms to measure the relevance or usefulness of a feature (attribute) in predicting the target variable. It helps in deciding the order in which features are evaluated and selected during the construction of a decision tree.

The goal of a decision tree is to split the data based on the features in a way that maximizes the predictive power or reduces the uncertainty in predicting the target variable. Information gain is a metric that quantifies the reduction in entropy (or impurity) achieved by splitting the data based on a particular feature.

Entropy is a measure of uncertainty or randomness in the target variable. A high entropy value indicates that the target variable has diverse and unpredictable values, while a low entropy value indicates a more predictable target variable. By selecting the attribute that maximizes information gain, we effectively choose the attribute that provides the most valuable information for reducing the uncertainty in the target variable.

Incorporating information gain into the decision tree construction process allows the algorithm to make informed decisions about which attribute to split on at each node. By selecting attributes with higher information gain, the decision tree can learn more discriminative patterns in the data, leading to better predictions. Therefore, we tried incorporating information gain in the tree displayed above, which is the model that performs better according to the accuracy calculations.

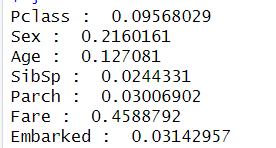
After incorporating information gain, we displayed this new decision tree for the newly created variables:



**To determine the accuracy, we used two methods. The first method involved utilizing Kaggle, while the second method involved writing the code in R.**

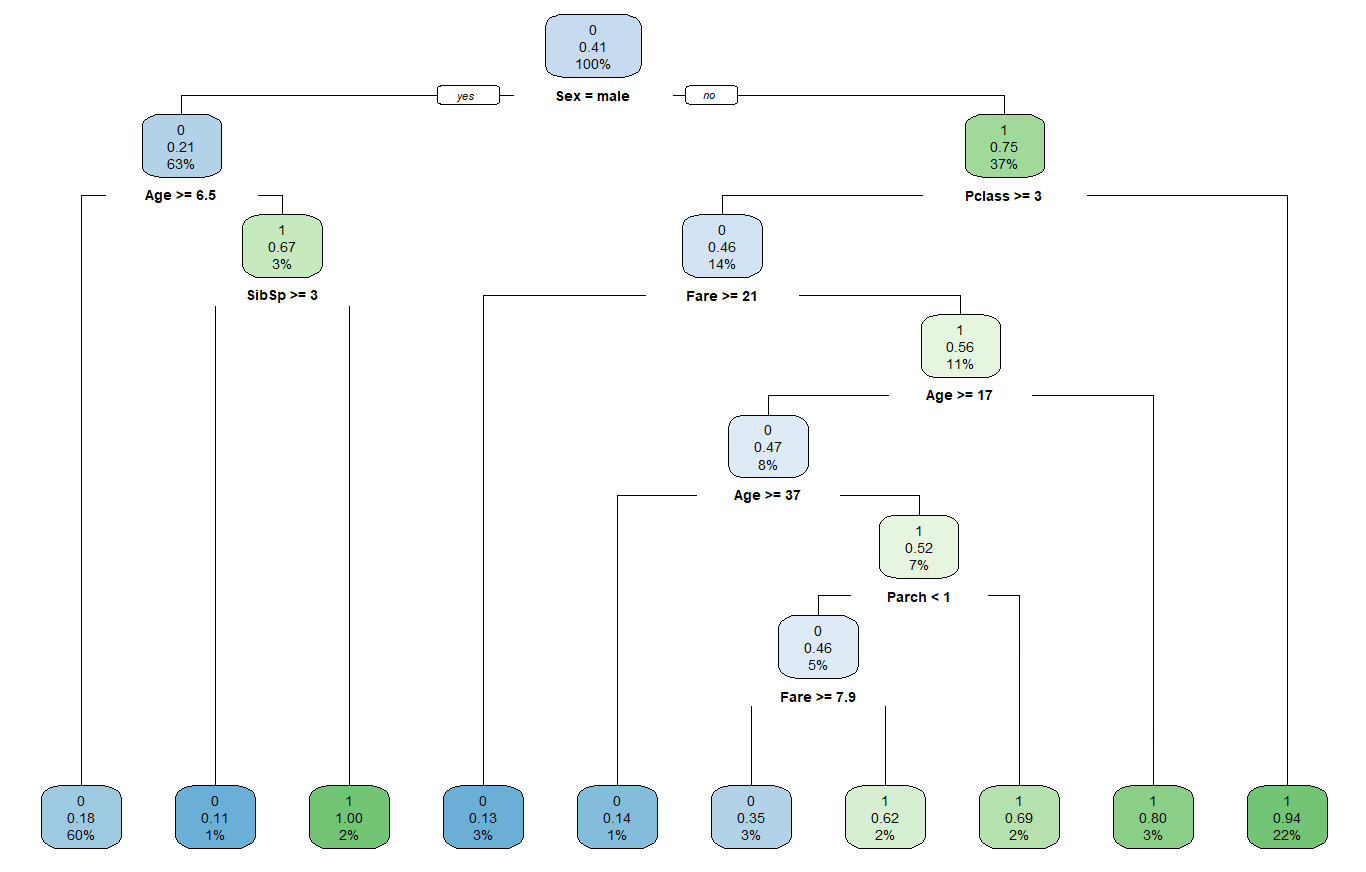
We noticed an increase in accuracy after incorporating information gain to this model. The new accuracy of the model after uploading it to **Kaggle is 78.70%,** while the accuracy output by **R code is 82.52%.**

Next, we proceeded with calculating the information gain for each attribute of the train dataset in order to better understand their influence in determining whether a passenger survives or not:



As we can see, Fare is the attribute with the highest information gain while SibSp with the lowest. Since we are not gaining enough information from SibSp and Parch attributes we can justify our previous decision to combine them in order to create new attributes such as FamilySize and FamilyID, which can be m ore beneficial in terms of information gain and accuracy when building the model.

Since Fare is the attribute with the highest information gain, we would think that it would have been placed as the root of the tree, but when building the decision tree in RStudio we got the following model:

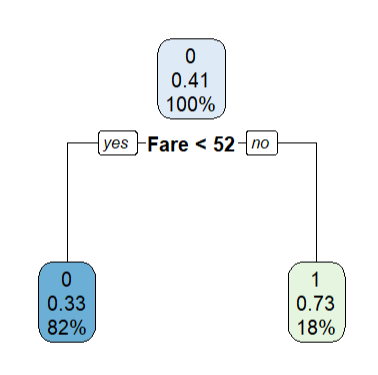


The rpart function in R does not necessarily guarantee that the attribute with the highest information gain will be chosen as the root attribute. The decision tree construction algorithm in rpart follows a different approach called the "classification and regression tree" algorithm.

In the algorithm used by rpart, the decision tree construction is based on a combination of factors, including information gain, Gini index, and other heuristics. These factors help determine the optimal split points and attributes for creating the decision tree.

Therefore, even if the attribute with the highest information gain is identified, it does not necessarily mean that it will be chosen as the root attribute in the resulting decision tree. However, we wrote another R Script to build the decision tree with the attribute that has the highest information gain as the root attribute, by modify the rpart function call accordingly.

In order to do so, we created a subset of the training data subset\_train by including only the attribute with the highest information gain (root\_attribute) and the target variable (Survived). The decision tree is then built using the rpart function with Survived ~ ., indicating that the attribute with the highest information gain should be the root of the tree, and the remaining variables should be considered as potential split variables. This is the displayed tree:



By modifying the subset\_attributes vector, we can include additional attributes to have multiple branches in the resulting decision tree.

# Random Forest

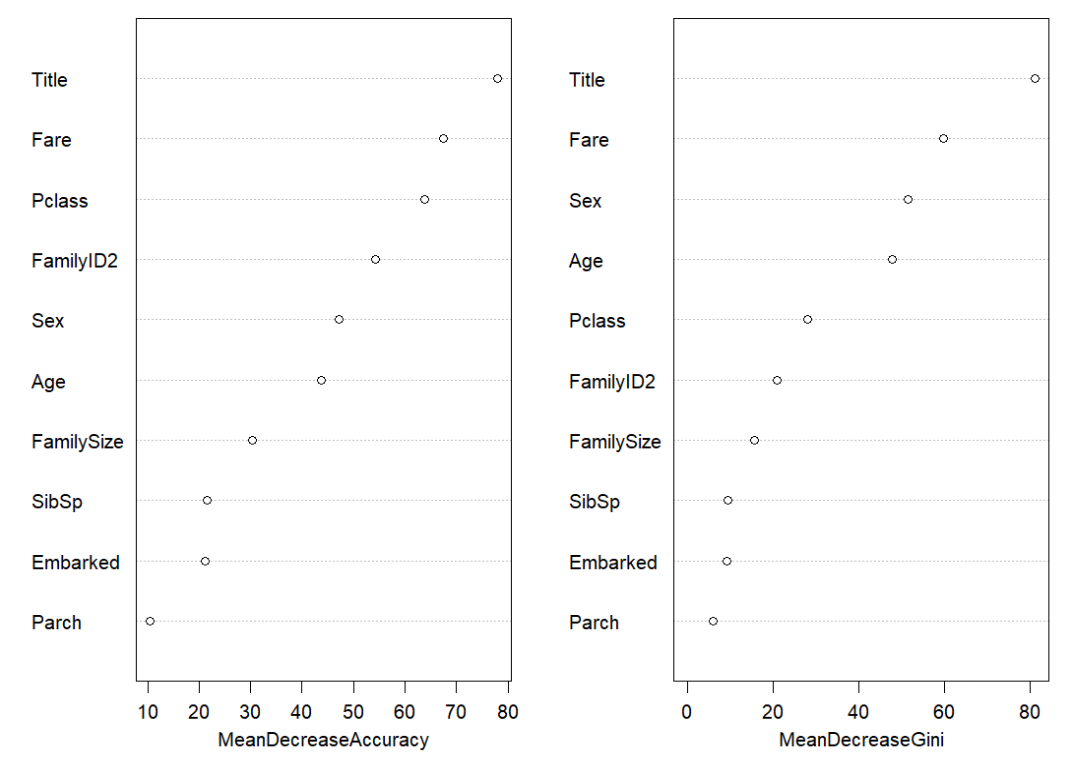
Random Forest is a powerful machine learning algorithm that combines the predictions of multiple decision trees to make accurate predictions or classifications. It belongs to the ensemble learning methods, which aim to improve the performance of individual models by combining them.

In our code, you have employed the randomForest package in R to build a random forest model. The purpose of this model is to predict the "Survived" variable in the Titanic dataset, based on various input variables such as "Pclass," "Sex," "Age," "SibSp," "Parch," "Fare," "Embarked," "Title," "FamilySize," and "FamilyID2."

After training the random forest model, we visualized the variable importance using the varImpPlot() function. This plot provides insights into the relative importance of each variable in predicting the survival outcome. The importance measure considers the MeanDecreaseAccuracy and MeanDecreaseGini metrics, which quantify the impact of each variable on the accuracy of predictions and the reduction of Gini impurity, respectively.

The resulting table of variable importance shows the contribution of each variable to the prediction accuracy and Gini impurity reduction for each class (0 and 1). Higher values indicate greater importance, suggesting that the corresponding variable plays a significant role in predicting the survival outcome.

By leveraging the random forest algorithm and analyzing variable importance, we have gained valuable insights into which factors are most influential in determining the survival of passengers in the Titanic dataset.



Variable Importance Measure: The x-axis represents the measure of variable importance. By default, random forests use the mean decrease in accuracy to calculate the importance measure. This measure indicates how much the model's accuracy decreases when a particular variable is randomly permuted, providing an indication of the variable's importance in making accurate predictions. A greater decrease in accuracy suggests that the variable carries more information or predictive power. Variables with higher mean decrease accuracy are considered more important in the model's decision-making process.

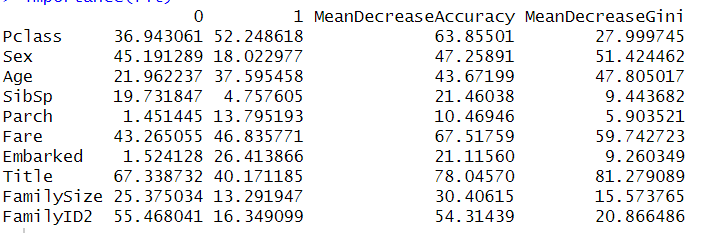
Another alternative method that can be used is mean decrease gini. Mean decrease Gini assesses the importance of a variable by measuring how much the Gini index of the nodes in the decision trees changes when the variable is used for splitting.

Mean decrease Gini measures the impact of a variable on the purity of nodes in decision trees within the random forest. A larger decrease in the Gini index indicates that the variable is more effective at separating the classes and is thus considered more important in the model's decision-making process.

Variable Importance Ranking: The y-axis represents the variables included in the random forest model. The variables are ranked in descending order based on their importance measure. The variable at the top has the highest importance, while the variable at the bottom has the lowest importance.

Length: The length between the dots represents the magnitude of the importance measure for each variable. Large length indicates variables that have a greater impact on the model's predictions.

Additionally, we can extract the numerical values of variable importance using the importance() function on the fit object. This will provide a ranked list of variables and their corresponding importance measures, allowing for further analysis if needed.



MeanDecreaseAccuracy: This metric measures the decrease in model accuracy when a specific variable is randomly permuted. Higher values indicate that the variable has a stronger impact on the model's predictive accuracy. Variables with higher MeanDecreaseAccuracy values are more important in making accurate predictions.

MeanDecreaseGini: This metric represents the total reduction in the Gini impurity index attributed to a specific variable. The Gini impurity is a measure of how well a split separates the classes in a classification problem. Higher MeanDecreaseGini values indicate that the variable has a greater ability to improve the purity of the splits. Variables with higher MeanDecreaseGini values are more important in creating meaningful splits in the decision trees.

Now, let's interpret the results for each variable:

* Pclass: This variable has a MeanDecreaseAccuracy of 36.943061 and MeanDecreaseGini of 52.248618. It is an important predictor and contributes significantly to the accuracy and purity of the model.
* Sex: This variable has a MeanDecreaseAccuracy of 45.191289 and MeanDecreaseGini of 18.022977. It is a highly important predictor and has a strong impact on the model's accuracy and the purity of the splits.
* Age: This variable has a MeanDecreaseAccuracy of 21.962237 and MeanDecreaseGini of 37.595458. It is relatively important in predicting the outcome.
* SibSp: This variable has a MeanDecreaseAccuracy of 19.731847 and MeanDecreaseGini of 4.757605. It has a moderate impact on the model's accuracy and purity.
* Parch: This variable has a MeanDecreaseAccuracy of 1.451445 and MeanDecreaseGini of 13.795193. It is less important compared to other variables.
* Fare: This variable has a MeanDecreaseAccuracy of 43.265055 and MeanDecreaseGini of 46.835771. It is an important predictor and contributes significantly to the model's accuracy and purity.
* Embarked: This variable has a MeanDecreaseAccuracy of 1.524128 and MeanDecreaseGini of 26.413866. It is relatively important in predicting the outcome.
* Title: This variable has a MeanDecreaseAccuracy of 67.338732 and MeanDecreaseGini of 40.171185. It is highly important and has a strong impact on the accuracy and purity of the model.
* FamilySize: This variable has a MeanDecreaseAccuracy of 25.375034 and MeanDecreaseGini of 13.291947. It is moderately important in predicting the outcome.
* FamilyID2: This variable has a MeanDecreaseAccuracy of 55.468041 and MeanDecreaseGini of 16.349099. It is relatively important compared to other variables.

The numbers displayed in each row of the first two columns indicate the variable importance measures for each class. Specifically:

* The number in the "0" column represents the variable importance measure for the class where passengers did not survive.
* The number in the "1" column represents the variable importance measure for the class where passengers did survive.

These measures indicate the contribution of each variable to the prediction accuracy or Gini impurity reduction specifically for each class. For example, taking the first row as an example:

* In the "0" column, the number 36.943061 represents the variable importance measure for the class "0" (did not survive). This means that the variable "Pclass" contributes significantly to predicting passengers who did not survive.
* In the "1" column, the number 52.248618 represents the variable importance measure for the class "1" (did survive). This means that the variable "Pclass" also contributes significantly to predicting passengers who did survive.

Similarly, the numbers in the subsequent rows indicate the variable importance measures for the corresponding classes in the "Survived" variable.

These variable importance measures provide insights into the relative importance of each variable in predicting each class and help identify which variables are more influential in distinguishing between the classes.

Overall, this analysis provides insights into the relative importance of each variable in the Random Forest model. Variables with higher MeanDecreaseAccuracy and MeanDecreaseGini values are considered more influential in predicting the "Survived" outcome.

# Evaluation

We built several prediction models for Titanic: Machine Learning from Disaster dataset. They were built using 2 data mining techniques: decision tree algorithm as well as random forest. Furthermore, different attributes were taken into consideration when building the model:

1. **Gender Model**

This prediction model uses the survived variable as the dependent variable or target variable, representing the survival outcome of the passengers and the sex variable as the independent variable or predictor variable to predict whether a passenger in the Titanic is more likely to survive or die. After submitting this model to **Kaggle,** we calculated its accuracy which was **76.55%,** while after running the code in **RStudio** the accuracy was **80.45%.**

1. **Gender Class Model**

This prediction model uses the survived variable as the dependent variable or target variable, representing the survival outcome of the passengers. It also uses the sex and class variable as the independent variables or predictor variables to predict whether a passenger in Titanic will survive. The accuracy of this model is **77.27% by Kaggle** and **82.50% by R code**.

1. **First Decision Tree**

This prediction model uses the decision tree algorithm as a data mining technique to split the data into smaller subsets based on the features that have the highest information gain. The accuracy of this model is **77.51% by Kaggle**, while by **R code the accuracy is 84.95%**.

1. **Full Decision Tree**

This prediction model allows the decision tree to grow to its maximum. Its accuracy is **74.64% by Kaggle** and **78.53% by R code**

1. **New Decision Tree**

This prediction model uses new variables as independent variables or predictor variables.

These newly created variables are developed by further processing the existing variables of the dataset in order to gain more powerful insights. The accuracy of this model is 77.75%. We also noticed an increase in accuracy after incorporation information gain to this model. The new accuracy of the model is **78.70%** **by Kaggle** and **89.75% by R code**

1. **Random Forest**

This prediction model uses the random forest algorithm as a data mining technique to provide insights into the importance of each variable in the model's predictions. By examining the plot, we can identify the most influential variables in the random forest model, which seem to be the newly created variables. The accuracy of this model is **76.07%** **by Kaggle and 82.70 % by R code.**

Based on the accuracy scores alone, the "New Decision Tree" model has the highest accuracy at 89.75%. This suggests that it performs slightly better than the other models in terms of predicting survival outcomes for Titanic passengers. However, to make a definitive conclusion, it's essential to consider additional factors such as interpretability, model complexity, and the generalization capability of the model.

The "New Decision Tree" model stands out as it incorporates newly created variables that provide more powerful insights. These variables are developed through further processing of existing dataset attributes, potentially capturing more nuanced patterns and improving the predictive performance.

It's important to note that the decision tree algorithm, in general, offers advantages such as interpretability, handling of both numerical and categorical data, and capturing non-linear relationships. On the other hand, random forests provide an ensemble approach that combines multiple decision trees to improve prediction accuracy and robustness.

Considering the higher accuracy and the inclusion of more powerful insights through the creation of new variables, the "New Decision Tree" model appears to be the best choice among the options provided.

# Conclusions

Apart from accuracy, there are several other reasons that suggest the "New Decision Tree" model may perform better than the other models:

1. Interpretability: Decision trees are highly interpretable, allowing us to understand the decision-making process and the rules that lead to predictions. The "New Decision Tree" model, with its newly created variables, may provide more transparent and meaningful insights into the factors influencing survival on the Titanic.
2. Feature Importance: Decision trees provide information about the importance of features in the prediction process. By examining the structure of the "New Decision Tree" model, we can identify which variables are deemed most influential in determining survival outcomes. This can offer valuable insights for further analysis and decision-making.
3. Flexibility: Decision trees allow for the inclusion of a variety of attribute types (categorical and numerical), making them flexible in handling different data types. The "New Decision Tree" model may leverage this flexibility to capture complex relationships between the newly created variables and survival outcomes.

Regarding the comparison between decision trees and random forests, some advantages of decision trees over random forests include:

* 1. Simplicity: Decision trees are relatively straightforward to understand and implement. They require minimal parameter tuning and can be easily visualized, making them a preferred choice when interpretability is important.
  2. Faster Training: Decision trees typically have faster training times compared to random forests. This is because random forests involve constructing multiple decision trees and combining their predictions, resulting in a more computationally intensive process.

* 1. Handling of Outliers: Decision trees can handle outliers effectively by creating dedicated branches or nodes to capture these unique instances. Random forests, on the other hand, rely on the consensus of multiple trees, which may dilute the impact of individual outliers.
  2. Relationship Exploration: Decision trees provide clear visualization of the decision-making process, allowing analysts to explore and interpret relationships between variables easily. This makes decision trees useful for uncovering insights and generating hypotheses.

It's important to note that the performance of decision trees and random forests can vary depending on the specific dataset and problem at hand. While decision trees have certain advantages, random forests excel in reducing overfitting, handling high-dimensional data, and providing robust predictions through the ensemble of multiple trees. In our case, we found decision trees as the most appropriate algorithm to be used in solving this prediction problem.