Titanic Dataset Analysis: Factors Influencing Survival

Executive Summary

Our comprehensive data analysis on the Titanic dataset aimed to predict the survival of passengers and identify key factors influencing survival rates.

1. **Socioeconomic Status:** Passengers of higher socioeconomic status, as evidenced by their passenger class and embarkation port, had higher survival rates.
2. **Gender**: The survival rate was significantly higher for female passengers, reflecting the "women and children first" protocol implemented during the evacuation.
3. **Cabin Location**: Survival rates varied based on cabin sections, indicating that proximity to lifeboats or structural integrity during the sinking might have been influential factors.
4. **Age**: Younger passengers, particularly children, had higher survival rates, potentially due to evacuation protocol and physical ability.
5. **Family Size**: Smaller family units (2-4 members) had higher survival rates compared to larger families and solo travelers, suggesting a potential influence of family support during evacuation.

Based on these insights, we recommend that future safety protocols and emergency procedures should prioritize vulnerable groups, consider socioeconomic inequalities, account for cabin placement and access to lifeboats, provide family assistance, enhance training and preparedness, and improve data collection for analysis.

Introduction

The sinking of the RMS Titanic on April 15, 1912, remains one of the most infamous maritime disasters in history. This report aims to explore the factors that influenced the likelihood of survival for passengers on the Titanic, in order to better understand the characteristics of those who survived the tragedy. The objectives of this analysis are to:

1. The primary goal of this project is to predict the survival of passengers onboard the Titanic using machine learning techniques.

2. By analyzing and interpreting the available dataset, we aim to build a reliable model that can accurately predict whether a passenger survived or did not survive the disaster based on various features, such as age, gender, class, and more.
3. The insights gained from this project can help us understand the factors that contributed to a passenger's survival and provide a historical perspective on the tragedy.

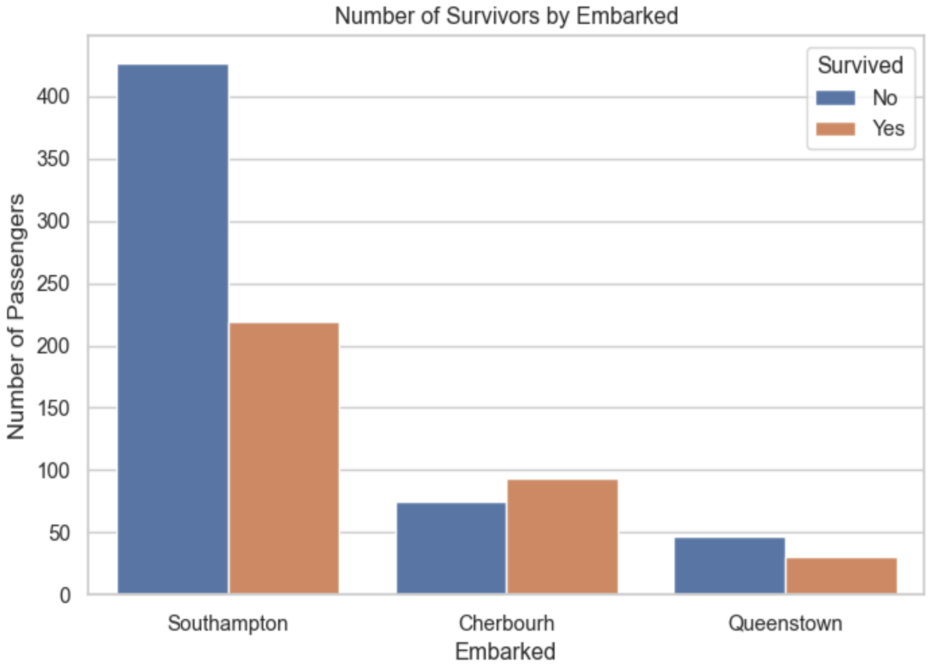
Variable	**Definition**	**Key**	**Type**
Survived	Survival	0 = No, 1 = Yes	Int
Pclass	Ticket class	1 = 1st, 2 = 2nd 3 = 3rd	Int
Sex	Sex		object
Age	Age in years	Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5	Float64
Sibsp	# of sibling/spouses aboard the Titanic	Sibling = brother, sister, stepbrother, stepsister; Spouse = Husband, wife (mistresses and fiances were ignored)	Int
Parch	# of parents/children aboard the Titanic	Parent = mother, father; Child = daughter, son, stepdaughter, stepson; Some children travelled only with a nanny, therefore parch= 0 for them.	int
Ticket	Ticket number		Object
Fare	Passenger fare		Float
Cabin	Cabin number		Object
embarked	Port of Embarkation	C = Cherbourg, Q= Queenstown, S = Southampton	Object

Result

The analysis of survival rate by different variables provides us with interesting insights into the factors that influenced survival rates onboard the Titanic. Here are the detailed observations:

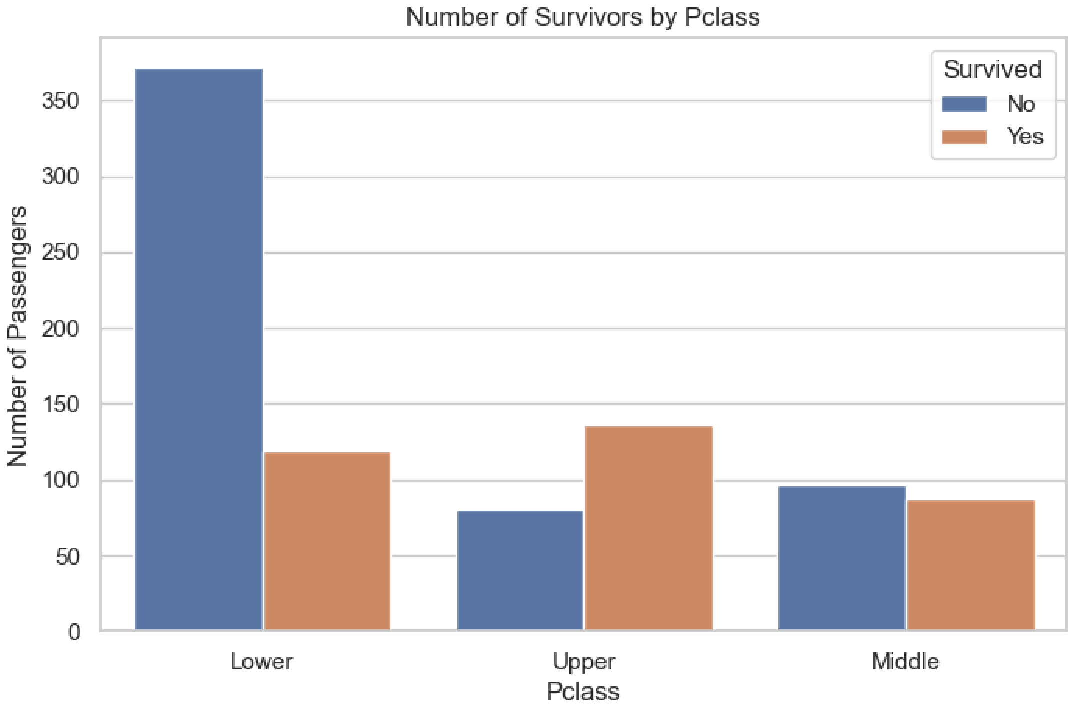
Embarked:

- Passengers who embarked from Cherbourg had the highest survival rate at approximately 55.36%, followed by Queenstown at 38.96%, and lastly, Southampton at 33.90%.
- **This could possibly be related to the socio-economic status of the passengers from these ports or the location of their cabins on the ship.**



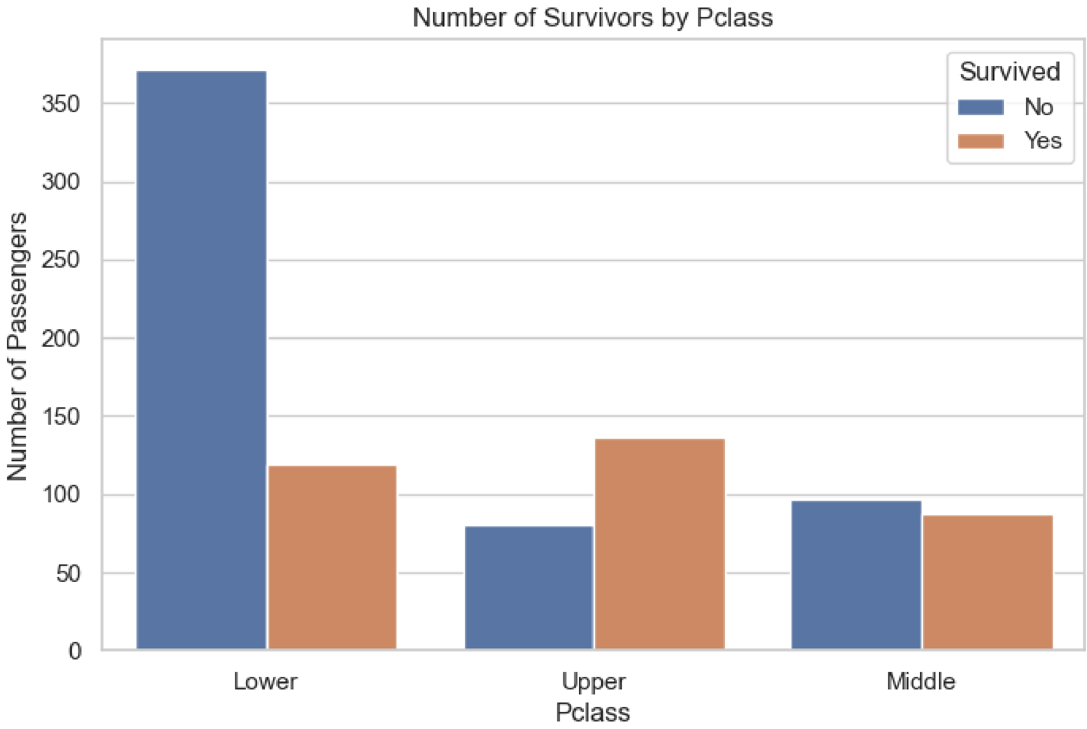
Pclass:

- The survival rate was highest for passengers in the Upper class (approximately 62.96%), followed by the Middle class (47.28%), and lowest for the Lower class (24.24%).
- **This suggests a strong correlation between socio-economic status and survival rate, with wealthier passengers more likely to survive.**



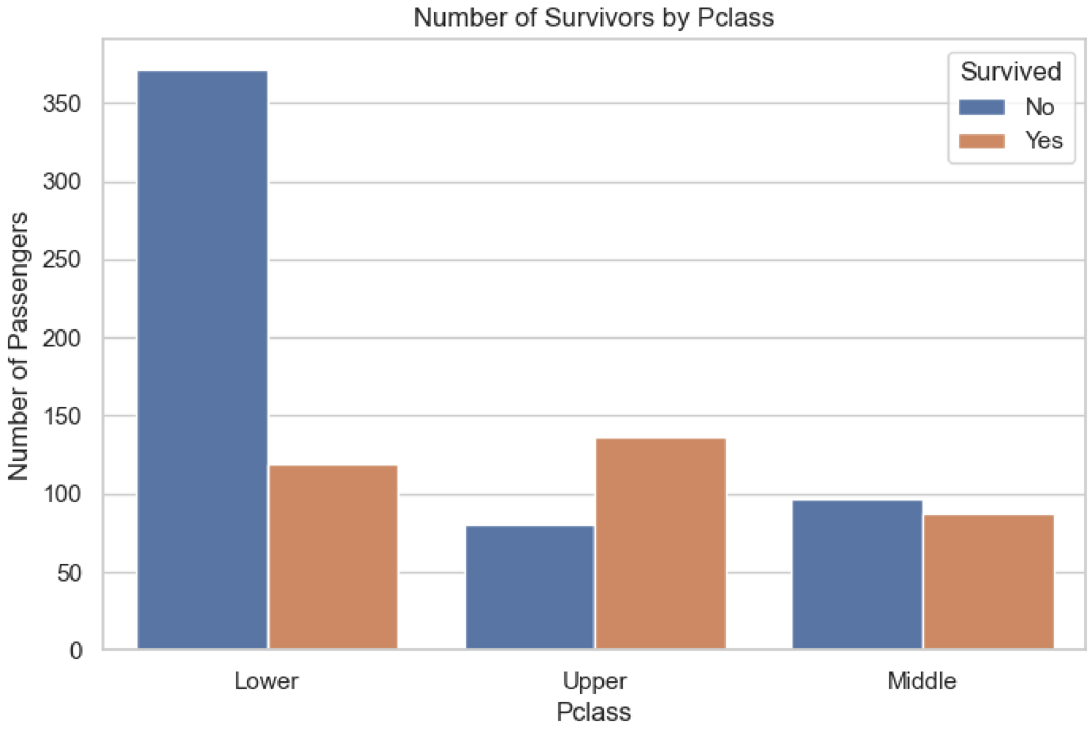
Sex:

- Female passengers had a significantly higher survival rate (74.20%) than male passengers (18.89%).
- **This likely reflects the "women and children first" protocol that was followed during the evacuation.**



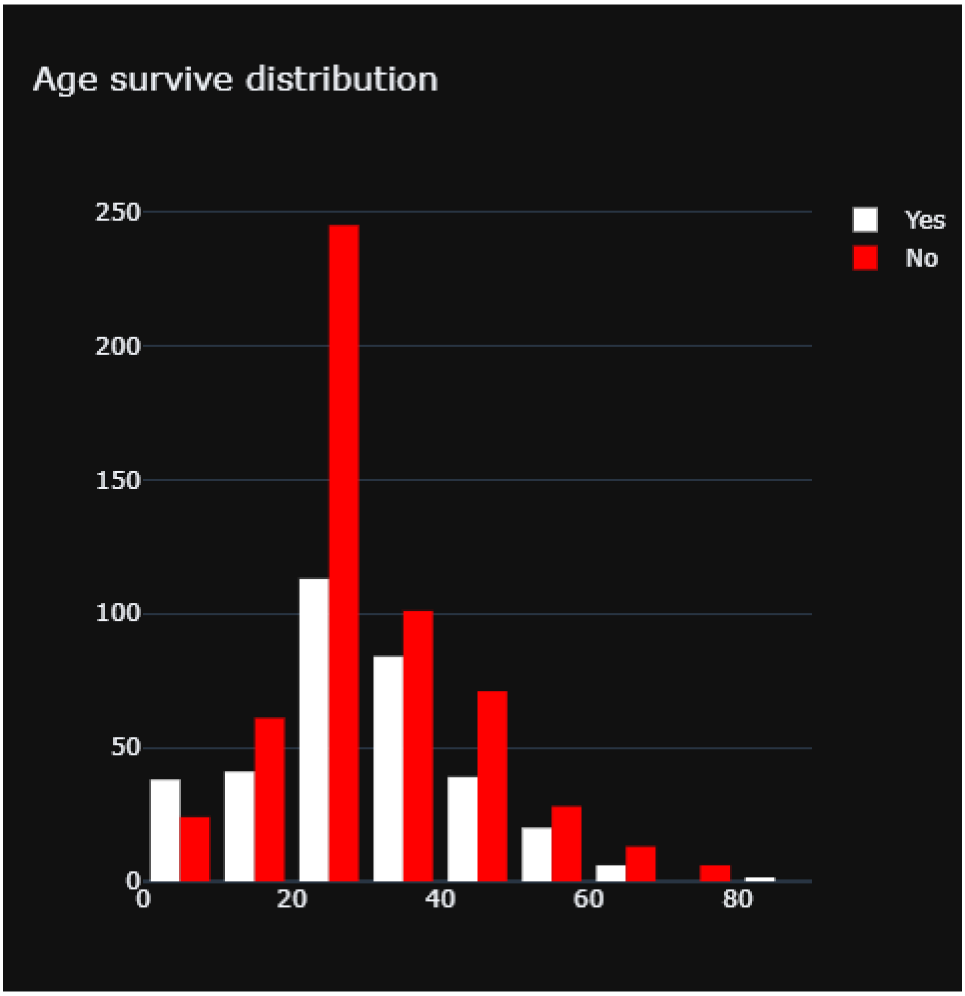
CabinSection:

- The Cabin Section survival rate analysis reveals that passengers in cabins D, E, and B had the highest survival rates at 75.76%, 75.00%, and 74.47%, respectively.
- Conversely, passengers in the 'T' cabin section had a survival rate of 0%, and those with 'Unknown' cabin sections had a lower survival rate of 29.99%.



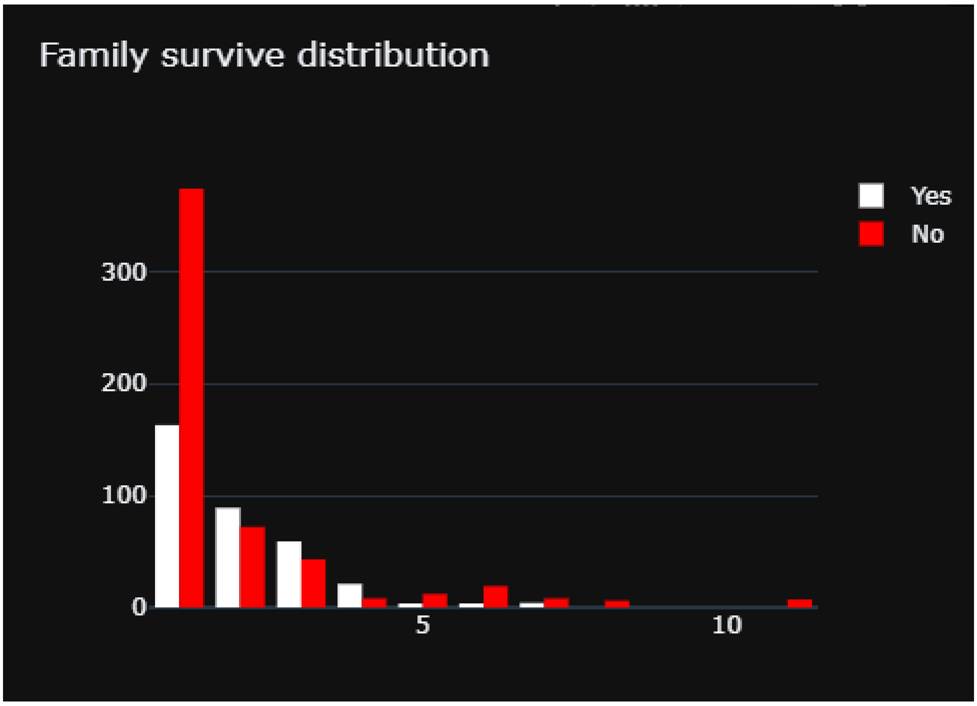
Age:

- The largest group of passengers fell in the age range of 19-29 years, with 358 people, followed by those in the 29-35 years range (185 people), and the 39-49 years range (110 people).
- The highest survival rate was for the youngest age group, 0-9 years, with a survival rate of 61.29%. **This supports the idea of "children first" during the evacuation.**
- Passengers aged 29-39 and 49-59 showed moderately high survival rates, at 45.41% and 41.67% respectively.
- The age group with the lowest survival rate was 69-79 years, which had a survival rate of 0%. **This suggests that older passengers were less likely to survive, potentially due to physical constraints.**
- The survival rate for the most populous age group, 19-29 years, was relatively low at 31.56%.



FamilySize:

- The majority of passengers traveled alone, with 537 individuals classified as a family size of 1. The next most common family sizes were 2 and 3, with 161 and 102 passengers respectively.
- Passengers traveling in a family size of 4 had the highest survival rate at approximately 72.41%, followed by family sizes of 3 and 2 with survival rates of 57.84% and 55.28% respectively.
- Family sizes of 7, 5, and 6 had lower survival rates of 33.33%, 20.00%, and 13.64% respectively.
- Surprisingly, passengers traveling alone (family size of 1) had a relatively low survival rate of 30.35%.
- Families of sizes 8 and 11 had the lowest survival rates, with no survivors.



Discussion

Based on the analyzed data, we have observed a number of interesting patterns:

1. **Socioeconomic Status and Survival:** A clear trend appears to show that passengers of higher socioeconomic status had a better chance of survival. This was evidenced by the higher survival rates for passengers who embarked from Cherbourg (a likely indication higher socioeconomic status) and those in the Upper passenger class.
2. **Gender and Survival:** The data significantly supports the historical account that the "women and children first" protocol was followed during the evacuation. Female passengers had a markedly higher survival rate compared to male passengers.
3. **Cabin Location and Survival:** The survival rate varied significantly based on the cabin section. Passengers in cabins D, E, and B had the highest survival rates, which might be due to these cabins' proximity to lifeboats or their structural integrity during the sinking.
4. **Age and Survival:** Age also played a crucial role in survival. Children and younger adults generally had higher survival rates than older adults, possibly due to the evacuation protocol or physical constraints among older passengers.

5. **Family Size and Survival:** Family size had an interesting relationship with survival rates. Passengers traveling with smaller families (2-4 members) had a higher chance of survival, while larger families and solo travelers were less likely to survive. This could be due to logistical issues during the evacuation or the support system available within smaller family units.

Recommendations

Based on analysis, we can draw several recommendations for future safety protocols and emergency procedures, not only for maritime disasters but for any situation where evacuation or rescue operations are required:

1. **Prioritize Vulnerable Groups**: Our analysis confirmed that women, children, and younger adults had higher survival rates, likely due to the "women and children first" protocol. This protocol should be enforced in all emergency situations, prioritizing the evacuation of vulnerable groups.
2. **Consider Socioeconomic Inequalities**: There were clear disparities in survival rates among different passenger classes, with wealthier passengers having higher survival rates. Emergency plans should take socioeconomic inequalities into account and ensure that all individuals, regardless of their socioeconomic status, have an equal chance of survival.
3. **Cabin Location and Access to Lifeboats**: The survival rate varied by cabin section, suggesting that proximity to lifeboats could have been a critical factor. Future ship designs should consider cabin placement in relation to lifeboat stations, ensuring that all passengers have quick and easy access to lifeboats in an emergency.
4. **Family Assistance**: Family size influenced survival rates, with smaller family units showing higher survival rates. Evacuation procedures should account for families and groups traveling together and provide assistance as necessary to keep families together during the evacuation process.
5. **Training and Preparedness**: The low survival rate for solo travelers suggests that many individuals may not have known what to do in the emergency situation. Regular drills, clear signage, and information on what to do in case of an emergency could help increase survival rates for all passengers.
6. **Data Collection and Analysis**: Our analysis was hindered by missing and incomplete data, particularly for the 'CabinSection' variable. Accurate data collection before and after emergencies can provide valuable insights for improving safety measures and emergency responses.

While these recommendations are based on a historical event, the insights gained can still be applied today, helping to improve safety procedures and reduce loss of life in future emergencies.

Variable Section and Modeling

To identify the most significant features for building a predictive model for passenger survival, we employed Recursive Feature Elimination (RFE) with logistic regression. This method works by recursively removing variables and building a model on the variables that remain, using model accuracy to identify which variables contribute the most to predicting the target variable.

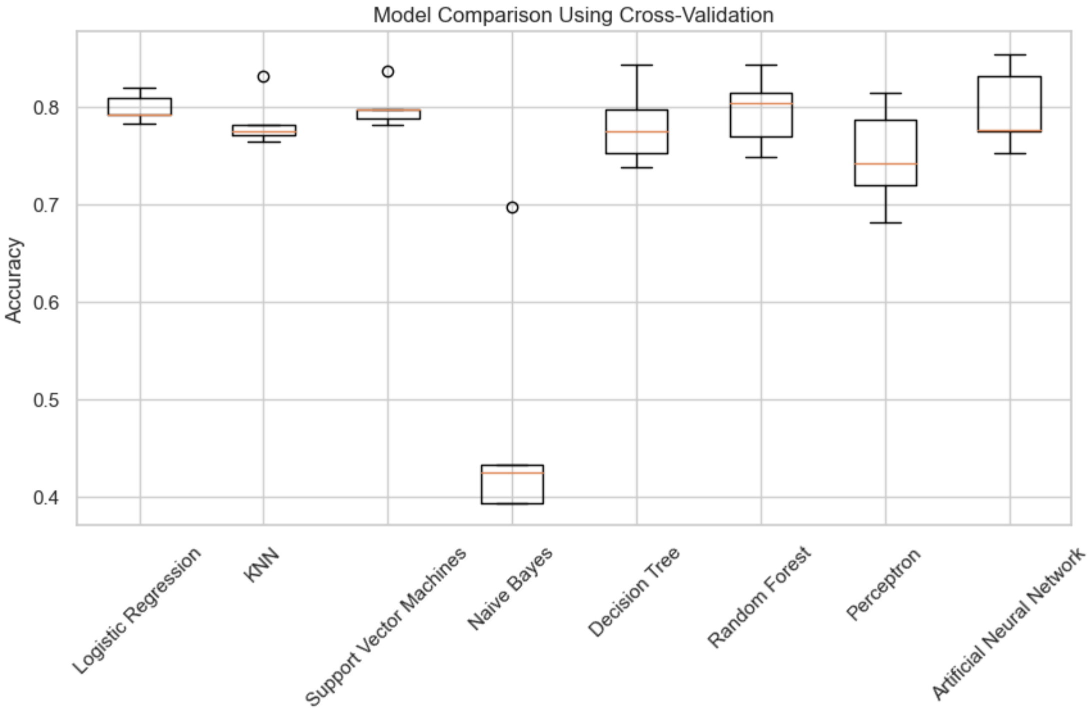
The RFE selected the following features as the most significant:

1. Pclass
2. Sex
3. Age
4. SibSp
5. Fare
6. FamilySize
7. IsAlone
8. Embarked_S
9. CabinSection_C
10. CabinSection_D
11. CabinSection_E
12. CabinSection_F
13. CabinSection_G
14. CabinSection_Unknown

These variables were used to build various predictive models, with the following results:

1. Logistic Regression: 0.799 (0.014)
2. KNN: 0.785 (0.024)
3. Support Vector Machines: 0.800 (0.019)
4. Naive Bayes: 0.468 (0.115)
5. Decision Tree: 0.781 (0.037)
6. Random Forest: 0.796 (0.033)
7. Perceptron: 0.749 (0.047)

8. Artificial Neural Network: 0.798 (0.038)



The ANN model has the highest mean accuracy (0.804) among the compared models. However, it also has a relatively higher standard deviation (0.049) compared to some of the other models.

The Support Vector Machines (SVM) model seem is a good choice also, due to its high accuracy and lower standard deviation.

The high accuracies of the Artificial Neural Network (ANN) and Support Vector Machines (SVM) models suggest that these models are able to effectively use the selected features to predict whether a given passenger would have survived the Titanic disaster. This demonstrates the potential of machine learning techniques for making accurate predictions even in complex and uncertain situations, which could be useful in a wide range of other applications. In other cases, if a vessel unfortunately has an accident, we can identify similar features to predict whether the survival of the passengers is guaranteed.