Group Project: Titanic - Machine Learning from Disaster

Nick Hartnett (50475276), Sai Shanmukh Puppala (50537991), Utkarsh Mathur (50495131)

The State University of New York at Buffalo

EAS-503: Python for Data Scientists

December 21, 2023

Why is the problem important?

- The data we selected is from the famous sinking of the RMS Titanic.
- Our goal was to explore the data given the context of the 1910s to better understand the survival rates of the Titanic's passengers.
- This problem is important because we can analyze how disparities between socioeconomic class, age, and gender affect survival rates aboard shipwrecks.
- Analyzing these disparities can help provide historical context to other shipwrecks as well as serve as a point of contrast to modern shipwrecks to supply information to the discussion of progress as a societal whole.

How did you analyze the problem?

- We approached the problem firstly by parsing the raw data from CSV through Python and then loaded the parsed data into a normalized database table.
- To analyze disparities in survival rates by sex we created a survival analysis by gender (sex) using Pandas and created a bar chart with the counts of 'Dead Males', 'Survived Males', 'Dead Females', 'Survived Females'.
- To analyze disparities in survival rates by age we created a survival analysis by age groups (age) grouped them in bins using Pandas and created a stacked bar chart.
- To analyze disparities in survival rates by socioeconomic class we created a survival rate and predicted survival analysis by Passenger Class.
- To analyze the effects of community bonding we compared the survival rates and predicted survival rates based on Port of Embarkation.
- For our project freedom portion of the project, we decided to use machine learning.
- We created 5 models that we trained on the data categories Ticket class (Pclass), number of parents/children aboard the Titanic (Parch), number of siblings/spouses aboard the Titanic (SibSp), port of Embarkation (Embarked), Sex, and Age.
- We trained Logistic Regression, Support Vector Machine, Decision Tree, KNN, and Random Forest models and compared their predictions on test data with training set.

What are your findings, reference charts/graphs?

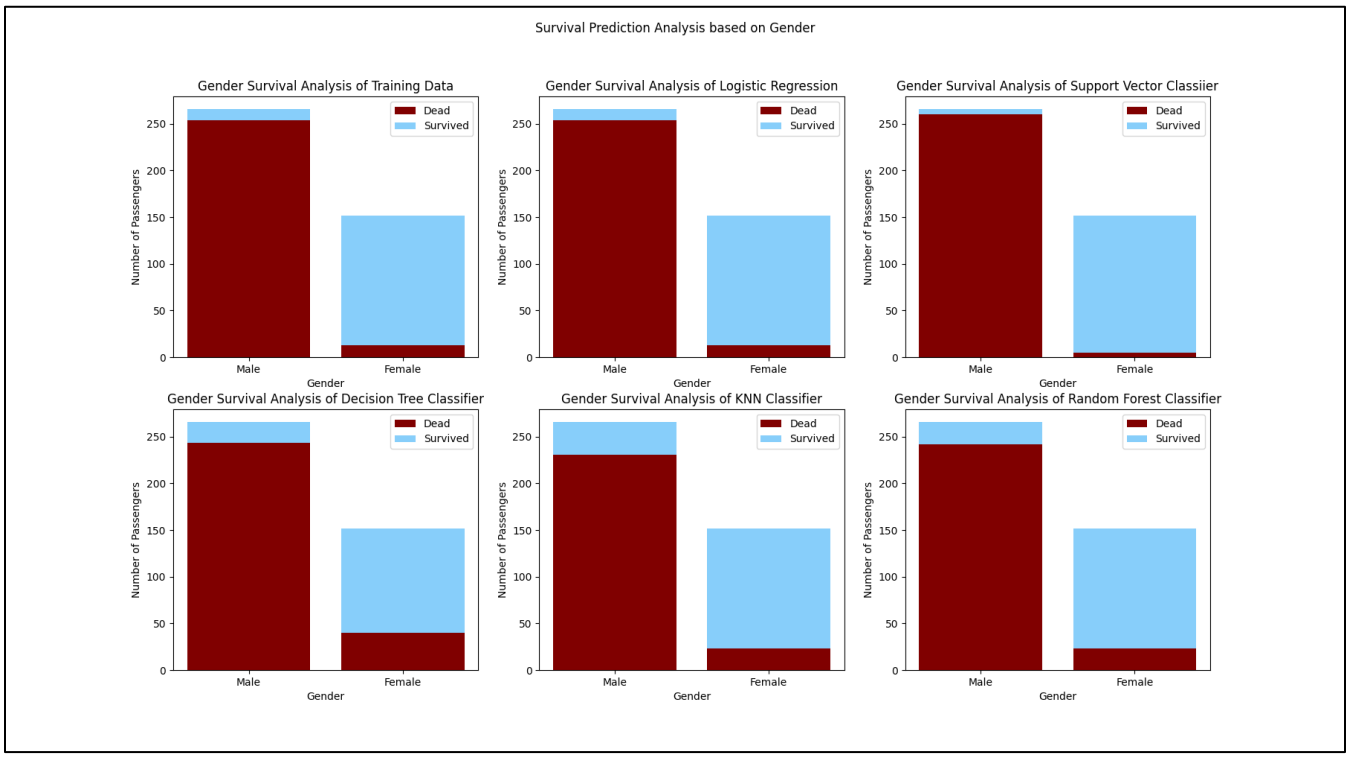
- In the analysis of disparities in survival rates by sex we found that women had a higher survival rate (Figure 1).
- In the analysis of disparities in survival rates by age we found that children had the highest survival rate, young to middle-aged adults died at similar rates, and the elderly had the lowest rates of survival (Figure 2).
- The survival rates of passengers in travelling class 1 were higher than that of class 2 and class 3 passengers. (Figure 3).
- The survival rates of passengers from Southampton port was very less as compared to that of Cherbourg and Queenstown however the total number of passengers from Southampton were also very high (Figure 4).
- The SVM model was performing better (based on validation accuracy, validation f-1 score, test accuracy) than all the other models (Table 1).

Appendix

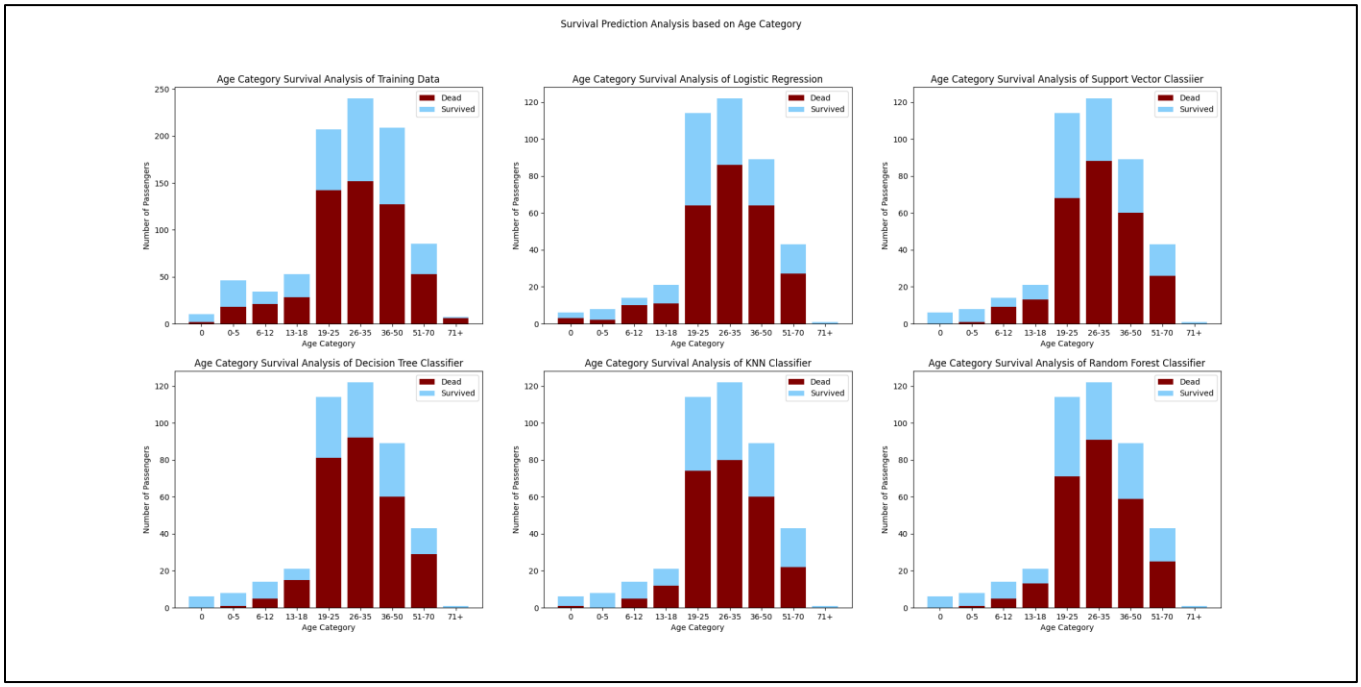
(**Table 1**: Model Performance Metrics)

No.	Models	Accuracy (Val)	Precision (Val)	Recall (Val)	F1-Score (Val)	Test Accuracy
1	Logistic Regression	77.78%	0.6585	0.8182	0.7297	75.59%
2	Support Vector Machine	84.44%	0.7436	0.8788	0.8056	71.05%
3	Decision Tree	80%	0.7419	0.6969	0.7186	74.4%
4	KNN	82.22%	0.7742	0.7273	0.75	77.99%
5	Random Forest	81.11%	0.7353	0.7576	0.7463	76.55%

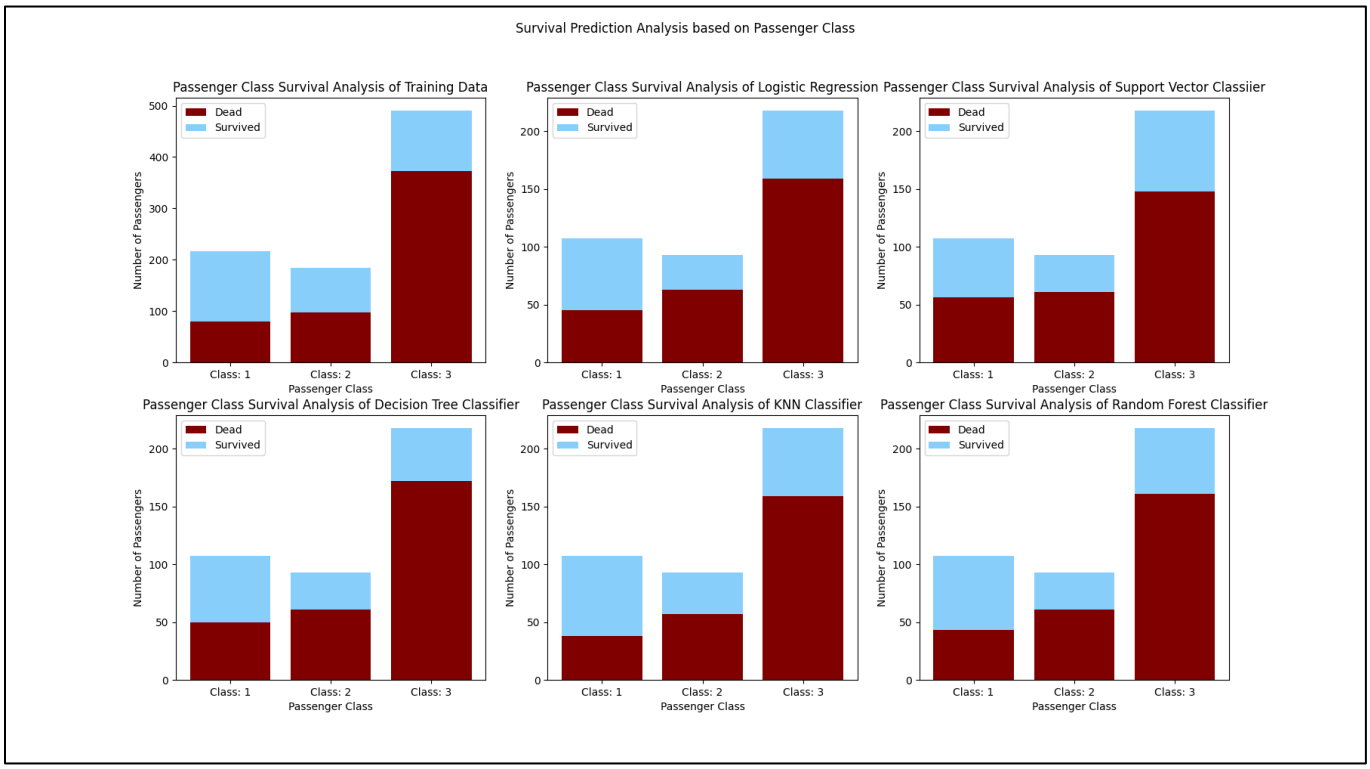
(Figure 1: Survival and Prediction Analysis by Gender**)**



(Figure 2: Survival and Prediction Analysis by Age**)**



(**Figure 3:** Survival and Prediction Analysis by Passenger Class)



(**Figure 4:** Survival and Prediction Analysis by Port of Embarkment)