BA 305 A4: Survival of the Titanic

Selina Manua, Julissa Mijares, Zoey Millstein,
Frances Josetta Sulistyo, and Sijia Zhan

Analysis Paper
QST BA305 at Boston University
Spring 2024

Introduction

Selecting a dataset

Starting this project, our team looked at various datasets and project ideas that we could explore. We initially looked at datasets that we believed most Boston residents would be familiar with, including social media applications, dating applications, and crime statistics. Then we found a Titanic dataset on Kaggle that interested all our team members. We are all big fans of the movie, with a general understanding of the history, and wanted to explore this topic through an analytical lens. Using the dataset from Kaggle, we found many variables to examine and predict passengers' survival, which would provide a compelling story.

Problem statement

The sinking of the Titanic, despite occurring over a century ago, remains a strong reminder of the unpredictability of disasters and the fundamental human instinct for survival. Our research focuses on analyzing the dataset, Survival of the Titanic, to understand why some people survived while others did not. We aim to answer the following questions:

1. How did variables such as age, gender, and fare impact survival probabilities?
2. What socio-economic and situational factors influenced the likelihood of survival among passengers?

Proposed analytic technique

Our team had various ideas on how to use the Titanic dataset for our analysis. We initially wanted to perform logistic regression to predict survival probabilities based on passengers' features. However, logistic regression may not capture complex patterns and interactions between features that other machine learning models like Random Forests and KNN can. We also considered implementing a neural network to handle the dataset's complexities; however, given the relatively small size of the dataset, there was a strong possibility of overfitting.

After considering these, we first decided to apply PCA to reduce dimensionality and focus on the most informative features, then create predictive models with KNN and Random Forests, and finally incorporate Cosine Similarity to predict survival rates of passengers based on 5 other passengers with the most similar feature vectors as them. The combination of these techniques enables us to create a strong survival prediction algorithm for our project.

Data Description

The Titanic Dataset

The Titanic dataset includes passenger information from the tragic maiden voyage of the RMS Titanic on April 15, 1912, capturing a combination of socio-economic, demographic, and familial details that are needed for predictive modeling and analysis (Tikkanen, 2024). This dataset contains 10 features, including both categorical (e.g. sex, embarked) and continuous variables (e.g. age, fare). The Kaggle dataset is split into training and testing data, where the testing data contains no information on passenger survival (Cukierski, 2012). Due to this, we decided to only use the training dataset for our study. The combined dataset has 1,309 passenger records, a snapshot view of the total of 2,240 passengers on the Titanic, and 891 of those records belong to the training dataset (Tikkanen, 2024). Hence, we will analyze the training dataset since it includes the column ‘Survival’. Our project aims to predict a passenger's likelihood of survival using this data, which provides a poignant reminder of the human aspects of the disaster but also serves as a basis for applying complex data science techniques.

Exploratory Analysis of the Dataset

First, let's explore the dataset and the percentage of the population that survived. Out of 891 passengers, only 38.4% survived.

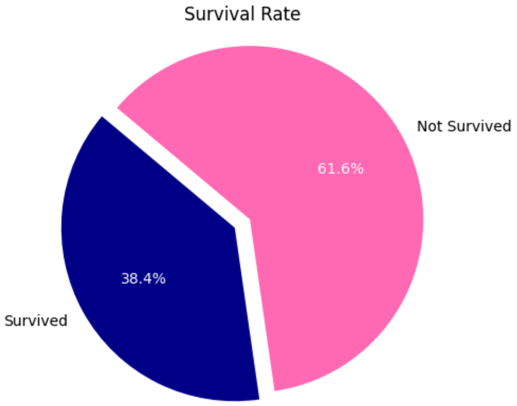


Figure 1: Pie chart showing the proportion of people who survived the Titanic

We hypothesized that sex and passenger class would be the most important features that affected survival rate because based on the movie, Jack, a working-class male, did not survive, while Rose, a female heiress, survived (Cameron, 1997). Hence, we plotted each of these graphs for our initial findings. For the variable sex, it was interesting to note that the total population of males is 577 and the female population is 314, but the survival rate for men is 18.89% and for women is 74.2%.

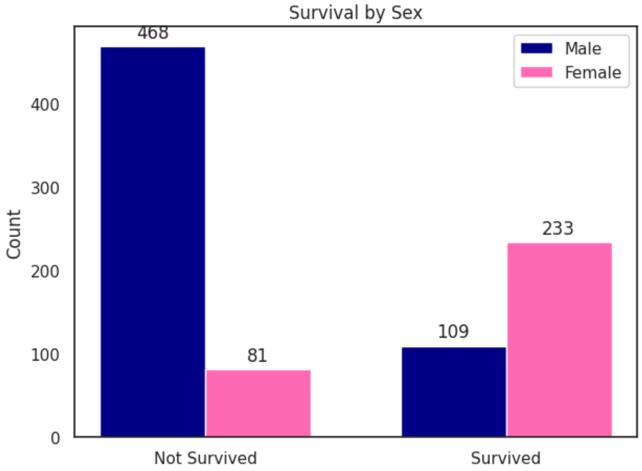
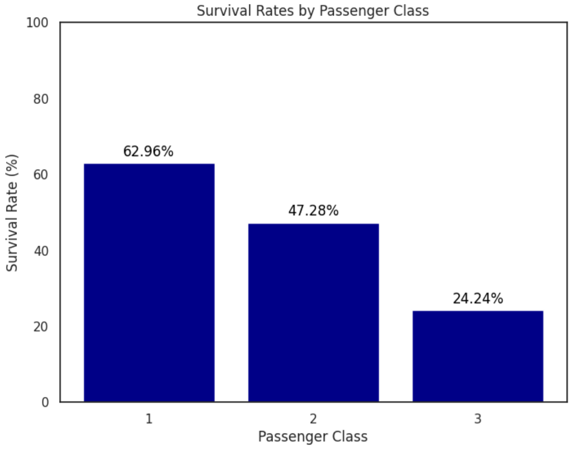


Figure 2: Bar chart showing the proportion of people who survived or did not survive based on their genders

In Figure 3, First Class passengers have a higher survival rate, which could be attributed to the priority passengers in higher classes given to lifeboat assignments. This is interesting because the survival rate of the First Class is three times higher than the Third Class, even though the First Class is less than half the population count of Third Class, as seen in Figure 4.



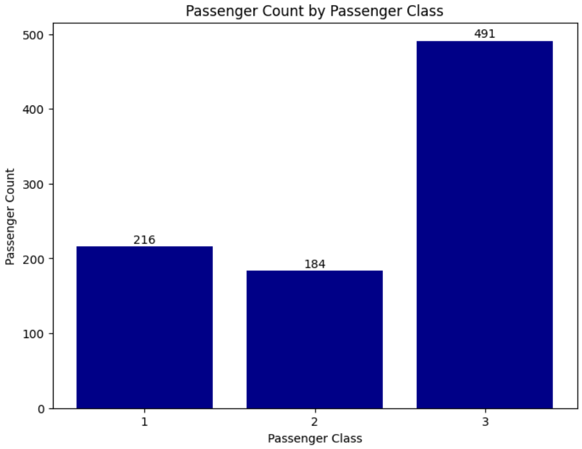


Figure 3 (left): Bar chart showing the survival rate of passengers by class
Figure 4 (right): Bar chart showing the count of passengers in each passenger class

Data Preprocessing

Identifying Missing Values:

After we were satisfied with our dataset selection, we began with data preprocessing. We identified the data type of each feature and the number of data available for each feature. *Age*, *Cabin*, and *Embarked* had 177, 687, and 2 missing values, respectively. *Age* was an important factor as it impacts mobility (older people may find it harder to run and escape) and maturity (younger children may not comprehend the depth of the situation), which would impact survival. Hence, we decided to fill the missing values in *Age* with the median so we could still utilize the feature. Since *Embarked* only has 2 missing values, we decided to replace the missing values with the mode of the feature in the entire dataset. We used the mode because it is the only central tendency measure that applies to categorical data. We dropped the *Cabin* column because 77% of passengers were missing this data, making it difficult to make justifiable replacements.

Creating a new Column called Title:

We recognized that *Name* shouldn't be an important column because every passenger had a unique name, so it should not impact their survival. However, in the *Name* column, there are also titles such as 'Mr', 'Ms', and 'Mrs' preceding the name of the passenger. We determined that this may potentially be a factor that impacts survival because titles may indicate marital status and social status. For example, titles like 'Mrs' signify a married woman, who might have been given priority to board on lifeboats with their children, or may have benefitted from the protective efforts of their husbands. Other titles like 'Dr', 'Capt', or 'Lady' may reflect a passenger's social status. Historically, people with higher social standing are given priority to resources, so they might have been given priority in lifeboat assignments or stayed in cabins closer to lifeboat decks. Therefore, we decided to extract the title from the ‘Name’ feature and create a new column called *Title*. Many French titles were used, so we replaced ‘Mlle’ (Mademoiselle) with ‘Miss’ and ‘Mme’ (Madame) with ‘Mrs’. We also replaced other less frequently occurring titles like ‘Lady’, ‘Countess’, ‘Capt’, ‘Col’, ‘Don’, and ‘Major’ with ‘Rare’ to indicate more special titles.

Dropping Unnecessary Columns:

We dropped the Name column since we already have the *Title* column from the information we extracted from our *Name* column. We also dropped the *Ticket* and *PassengerID* columns because both contained unique numbers for every passenger, so it would provide no meaningful information for predicting survival.

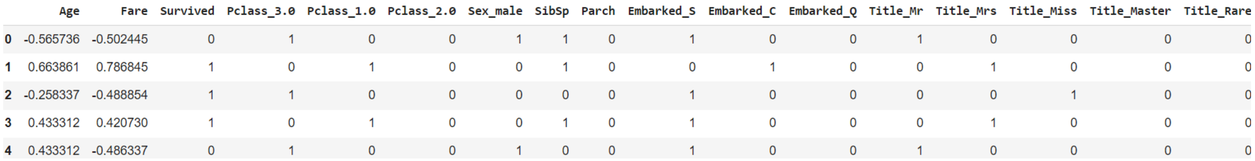
Creating Dummy Variables:

Afterward, we utilized OneHotEncoding to encode the remaining categorical data into dummy variables. The categorical variables are *Embarked*, *Sex*, *Pclass*, and *Title*. OneHotEncoding transforms a categorical value with N possible values into N binary variables, each representing one of the possible values. We chose to use OneHotEncoding instead of label encoding because it does not introduce a false ordinal relationship. It also tends to make the model’s predictions more interpretable because the impact of each category on the data can be individually assessed.

Scaling:

Finally, we used standardized scaling to scale our data. We chose standardized scaler over min-max scaling due to considerations of what would be best for the predictive models that we are going to create. K-Nearest Neighbors is a distance-based algorithm, so each feature needs equal weighting without being skewed by outliers. Min-max scaling only rescales features to a [0,1] range, which can be distorted by extreme values that could affect model performance. Standardized scaling is also preferred for Cosine Similarity as it ensures that each feature contributes equally, preventing the dominance of a single feature due to its scale. This normalization is important in Cosine Similarity, where the angle between the vectors is the focal point.

The final data frame that we processed looks like this:



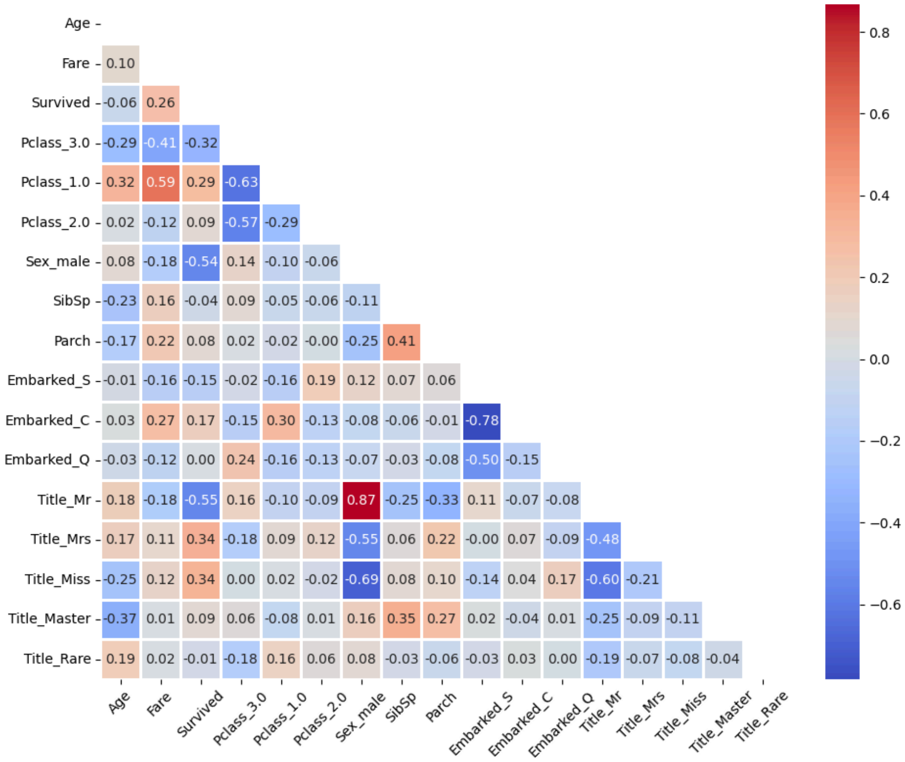
	Age	Fare	Survived	Pclass_3.0	Pclass_1.0	Pclass_2.0	Sex_male	SibSp	Parch	Embarked_S	Embarked_C	Embarked_Q	Title_Mr	Title_Mrs	Title_Miss	Title_Master	Title_Rare
0	-0.565736	-0.502445	0	1	0	0	1	1	0	1	0	0	1	0	0	0	0
1	0.663861	0.786845	1	0	1	0	0	1	0	0	1	0	0	1	0	0	0
2	-0.258337	-0.488854	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0
3	0.433312	0.420730	1	0	1	0	0	1	0	1	0	0	0	1	0	0	0
4	0.433312	-0.486337	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0

Figure 4: Dataframe after Data Preprocessing

Methodology

Following data preprocessing, we conducted a correlation analysis to explore the relationships among variables, focusing on identifying potential multicollinearity and understanding their impact on survival. Notably, we observed a significant correlation between *Fare*, representing ticket price, and the three passenger classes denoted by *Pclass_1.0, Pclass_2.0, and Pclass_3.0*. As anticipated, higher classes correlate with higher fares, suggesting redundancy between fare and class variables. Conversely, variables associated with embarkment points (*Embarked_S, Embarked_C,* and *Embarked_Q*), as well as the number of family members onboard (*SibSp* and *Parch*), exhibited minimal correlation. Therefore, those variables would be less useful in our prediction model.

In terms of predicting survival outcomes, two noteworthy variables emerged. First, the variable *Sex_male* displayed a significant negative correlation (-0.54) with survival, indicating a lower likelihood of survival among males. However, it's important to note that *Title* is strongly correlated (-0.50) with *Sex_male*, potentially leading to multicollinearity issues. Second, *Fare* showed a moderate positive correlation (0.26) with survival, suggesting its relevance in predicting survival probabilities. In combination with *Fare*, looking more closely at Economic Class, we can see that *Pclass_3.0* (Lower Class) has a strong negative correlation (-0.41) and *Pclass_1.0* (Upper Class) has a strong positive correlation (0.59) indicating that class ranking played an important role in the survival rate.



	Age	Fare	Survived	Pclass_3.0	Pclass_1.0	Pclass_2.0	Sex_male	SibSp	Parch	Embarked_S	Embarked_C	Embarked_Q	Title_Mr	Title_Mrs	Title_Miss	Title_Master	Title_Rare
Age																	
Fare	0.10																
Survived	-0.06	0.26															
Pclass_3.0	-0.29	-0.41	-0.32														
Pclass_1.0	0.32	0.59	0.29	-0.63													
Pclass_2.0	0.02	-0.12	0.09	-0.57	-0.29												
Sex_male	0.08	-0.18	-0.54	0.14	-0.10	-0.06											
SibSp	-0.23	0.16	-0.04	0.09	-0.05	-0.06	-0.11										
Parch	-0.17	0.22	0.08	0.02	-0.02	-0.00	-0.25	0.41									
Embarked_S	-0.01	-0.16	-0.15	-0.02	-0.16	0.19	0.12	0.07	0.06								
Embarked_C	0.03	0.27	0.17	-0.15	0.30	-0.13	-0.08	-0.06	-0.01	-0.78							
Embarked_Q	-0.03	-0.12	0.00	0.24	-0.16	-0.13	-0.07	-0.03	-0.08	-0.50	-0.15						
Title_Mr	0.18	-0.18	-0.55	0.16	-0.10	-0.09	0.87	-0.25	-0.33	0.11	-0.07	-0.08					
Title_Mrs	0.17	0.11	0.34	-0.18	0.09	0.12	-0.55	0.06	0.22	-0.00	0.07	-0.09	-0.48				
Title_Miss	-0.25	0.12	0.34	0.00	0.02	-0.02	-0.69	0.08	0.10	-0.14	0.04	0.17	-0.60	-0.21			
Title_Master	-0.37	0.01	0.09	0.06	-0.08	0.01	0.16	0.35	0.27	0.02	-0.04	0.01	-0.25	-0.09	-0.11		
Title_Rare	0.19	0.02	-0.01	-0.18	0.16	0.06	0.08	-0.03	-0.06	-0.03	0.03	0.00	-0.19	-0.07	-0.08	-0.04	

Figure 5: Correlation Matrix for all variables. The more pigmented the color, the stronger the correlation. Blue represents a negative correlation, and red represents a positive correlation.

Dimension Reduction

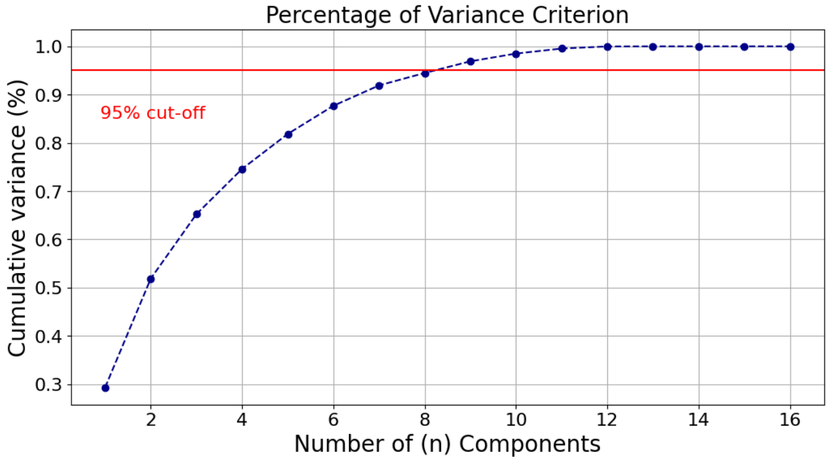


Figure 6: Graph showing the percentage of variance criterion in each PCA component

Since we determined some variables to be redundant in the correlation matrix above, we decided to perform Principal Component Analysis (PCA). Our team opted to conduct a Percentage of Variance Criterion test to determine the optimal number of n components for our models. This entire process was done using Python and importing the required sklearn packages. Initially, feature standardization was achieved through scalar transformation, followed by fitting using a PCA().fit() model. A visualization plot was created for us to determine where we should create the cut-off point. As illustrated in the chart above, our analysis revealed that employing 8 components effectively explained approximately 95% of the data's variance.

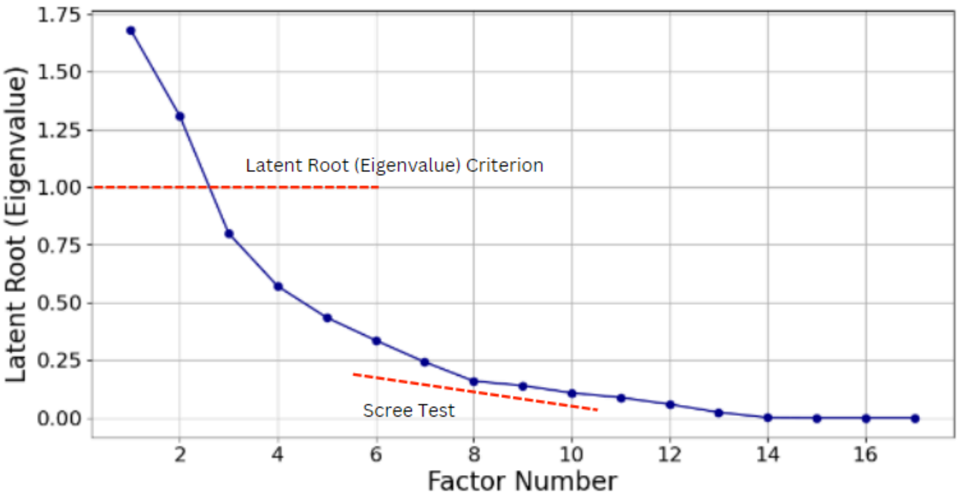
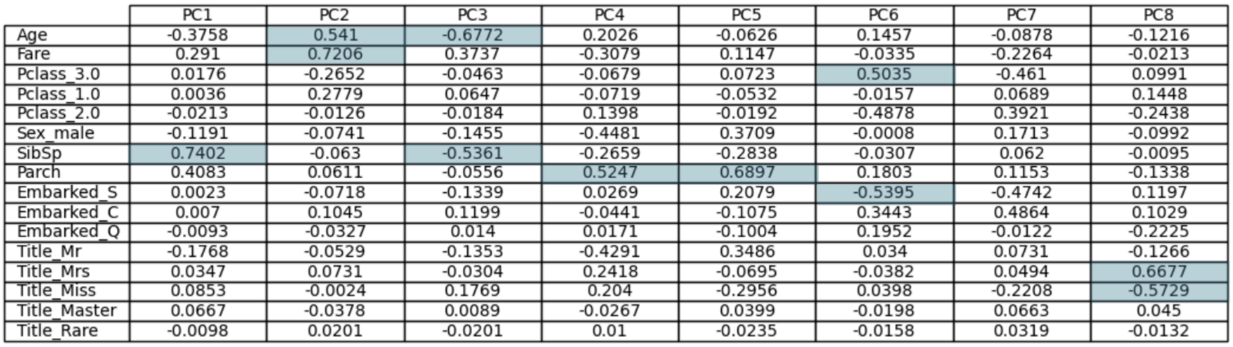


Figure 7: Latent Root Criterion Graph

The Scree Test Criterion aligns with the Percentage of Variance Criterion with an n value of 8. We did not choose the Latent Root Criterion because only 2 components had an eigenvalue greater than 1.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Age	-0.3758	0.541	-0.6772	0.2026	-0.0626	0.1457	-0.0878	-0.1216
Fare	0.291	0.7206	0.3737	-0.3079	0.1147	-0.0335	-0.2264	-0.0213
Pclass_3.0	0.0176	-0.2652	-0.0463	-0.0679	0.0723	0.5035	-0.461	0.0991
Pclass_1.0	0.0036	0.2779	0.0647	-0.0719	-0.0532	-0.0157	0.0689	0.1448
Pclass_2.0	-0.0213	-0.0126	-0.0184	0.1398	-0.0192	-0.4878	0.3921	-0.2438
Sex_male	-0.1191	-0.0741	-0.1455	-0.4481	0.3709	-0.0008	0.1713	-0.0992
SibSp	0.7402	-0.063	-0.5361	-0.2659	-0.2838	-0.0307	0.062	-0.0095
Parch	0.4083	0.0611	-0.0556	0.5247	0.6897	0.1803	0.1153	-0.1338
Embarked_S	0.0023	-0.0718	-0.1339	0.0269	0.2079	-0.5395	-0.4742	0.1197
Embarked_C	0.007	0.1045	0.1199	-0.0441	-0.1075	0.3443	0.4864	0.1029
Embarked_Q	-0.0093	-0.0327	0.014	0.0171	-0.1004	0.1952	-0.0122	-0.2225
Title_Mr	-0.1768	-0.0529	-0.1353	-0.4291	0.3486	0.034	0.0731	-0.1266
Title_Mrs	0.0347	0.0731	-0.0304	0.2418	-0.0695	-0.0382	0.0494	0.6677
Title_Miss	0.0853	-0.0024	0.1769	0.204	-0.2956	0.0398	-0.2208	-0.5729
Title_Master	0.0667	-0.0378	0.0089	-0.0267	0.0399	-0.0198	0.0663	0.045
Title_Rare	-0.0098	0.0201	-0.0201	0.01	-0.0235	-0.0158	0.0319	-0.0132

Figure 8: PCA Component Matrix

The component weights display the partial correlation between a particular variable and the component. Ideally, the absolute correlation should be greater than or equal to 0.50 in magnitude. Taking the squared weight would represent the amount of variance explained by the particular component. Looking at the figure above, we identified and highlighted the highest absolute weight across all components.

Findings and Implications

Naive Rule

Upon examining the dataset, 38.4% of passengers survived, and 61.6% did not survive. In the context of predicting survival outcomes, the Naive Rule serves as a baseline against the performance accuracy of predictive models. For our dataset, the naive rule would be to predict that all passengers did not survive, given that the majority class in the dataset comprises passengers who did not survive 61.6%. To ensure efficacy, our predictive models, KNN and Random Forest, used this percentage as a benchmark for assessing accuracy.

K-Nearest Neighbors

We chose to use K-Nearest-Neighbors because of its simplicity and interpretability. It makes predictions based on the 'nearest' examples in the training dataset which can be quite intuitive: passengers with similar characteristics (such as class, age, sex) are likely to have similar survival outcomes. It can also capture complex patterns from the data without needing explicit modeling of the decision boundaries.

In our analysis through KNN, K = 15 yielded the highest accuracy of 82.68%, ensuring a definitive outcome.

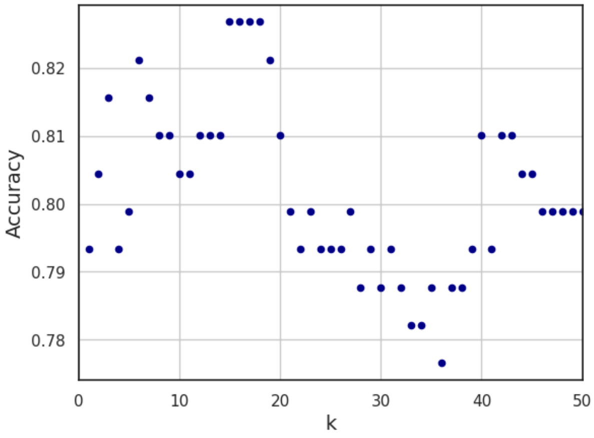


Figure 9: Graph identifying best K value

Additionally, our contextual understanding of the dataset revealed a strong correlation between fare and survival. By considering contextual factors K= 15, we leverage the assumption that passengers who paid similar fares likely resided in comparable areas of the steamship, possibly influencing their survival outcomes. This creates a smoother classification given that we are less sensitive to the noise of individual data points and there is more stability.

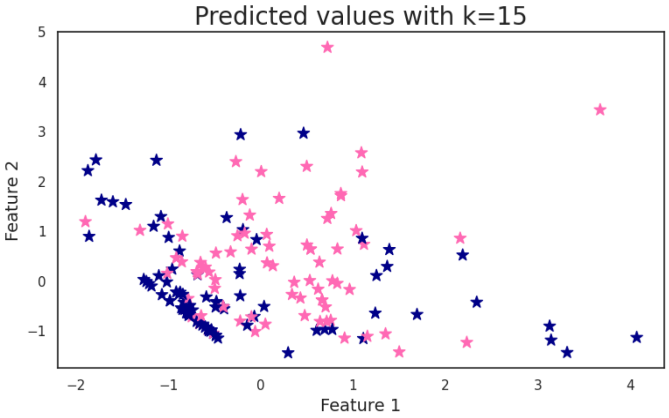


Figure 10: KNN Visualization

Given that survival on the Titanic might be influenced by a complex interaction of features (e.g., children in first class having a higher survival rate than adults in third class), KNN's flexibility in this regard can be beneficial.

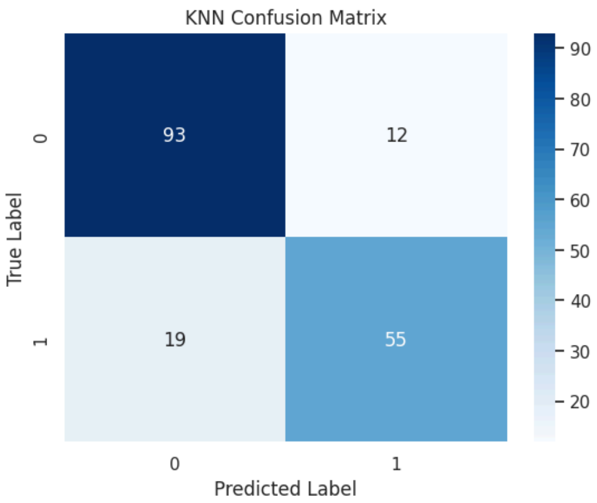


Figure 11: K-Nearest Neighbors Confusion Matrix

Our confusion matrix and accuracy score, reveal the overall correctness of the model's predictions, suggesting strong performance in predicting Titanic survival outcomes. We also made sure that our model does not overfit by comparing the performance of our model on both the training and testing datasets. The model's accuracy is 82.682% on the train data and 82.584% on the test data. This shows that our model does not overfit, as the difference in accuracy score is only 0.098%, indicating that the model is performing well considering unseen test data. This accuracy is also higher than our Naive Rule baseline indicating that our model is making more accurate predictions rather than just relying on the majority class.

While KNN provides value, it is important to acknowledge its limitations including sensitivity to outliers and the curse of dimensionality. In addition, our larger value of K can potentially pose bias in our predictions.

Random Forest

Besides K-Nearest Neighbors, we decided to develop a machine-learning model with a Random Forest classifier. We decided to use Random Forest instead of a Decision Tree because it is less likely to overfit than a single decision tree since it averages multiple trees that might overfit the data. We also chose to use Random Forest due to its ability to rank features in order of its importance in making predictions. This would enable us to determine which factors most significantly influenced the Titanic's survival, providing insights for model interpretation and understanding of historical context.

Without hyperparameter tuning, we achieved an accuracy of 78.21% for our Random Forest algorithm with the model's default settings. To increase the accuracy, we conducted hyperparameter tuning with GridSearchCV, iterating over a range of values for *n_estimators*, *criterion*, and *max_depth*. This led to an optimized model configuration of 200 trees, a 'gini' criterion for splitting, and a tree depth of 3, with an improved accuracy of 81%.

We also made sure that our model does not overfit by comparing the performance of our model on both the training and testing datasets. The model's accuracy is 84.971% on the train data and 81.006% on the test data. This shows that our model does not overfit, as the difference in accuracy score is only 3.965%. This accuracy is higher than our Naive Rule baseline, showing that Random Forest performs more accurately in making predictions than just relying on the majority class.

From there, we determined the importance of each feature in predicting survival in the Titanic based on the Random Forest algorithm.

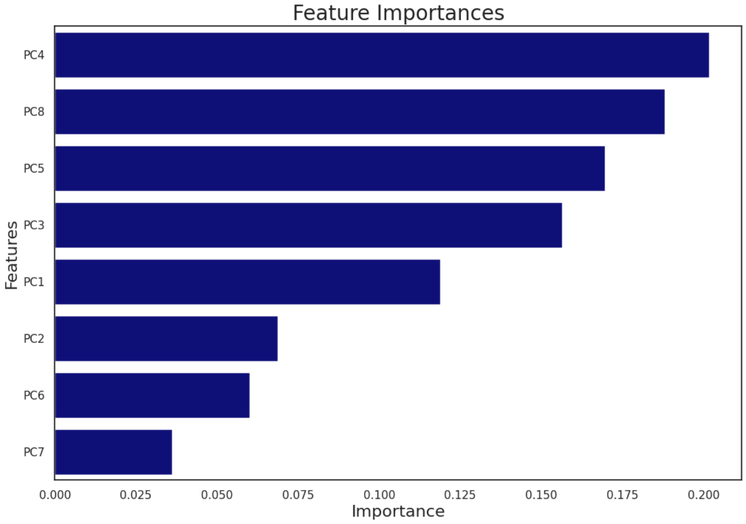


Figure 12: Graph showing feature importance for our Random Forest classifier

As seen in Figure 12, PC4 is the most important feature in determining a passenger's survival on the Titanic. Based on Figure 8, the three top contributing features of PC4 are *Parch*, *Sex_male*, and *Title Mr*. The positive coefficient on Parch suggests that having parents or children on board might have influenced survival, potentially because families might have been prioritized during the lifeboat loading process. However, the negative coefficients for *Sex_male* and *Title Mr.* imply that being an adult male was a survival disadvantage, consistent with the historical accounts of the "women and child first" policy.

However, there are some limitations of the Random Forest model. Building a Random Forest involves constructing multiple decision trees, which can be memory-intensive. This complexity also results in longer training times, especially as the number of trees or depth of each tree increases. Moreover, while Random Forest reduces the risk of overfitting than a single decision tree, it can still overfit if not properly tuned when handling noisy datasets. Despite providing feature importance scores, the ensemble nature of Random Forest makes it difficult to trace how specific decisions were made within the model.

Comparison of Models

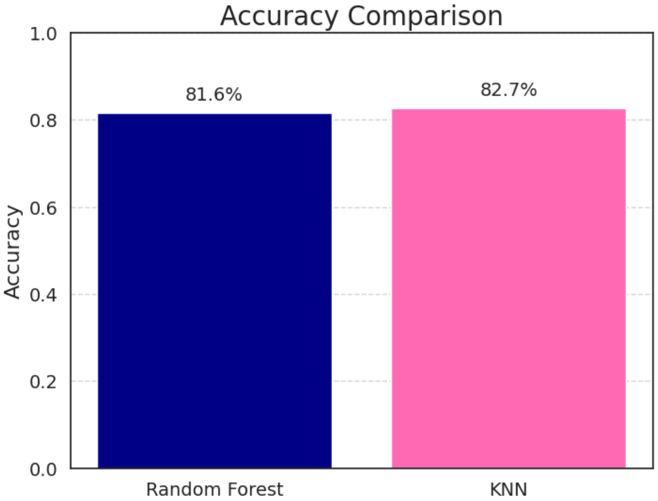


Figure 13: Graph comparing the accuracy of our KNN and Random Forest models

When we compare the accuracy of Random Forest and KNN, we can see KNN is performing slightly better with a score of 82.7% to Random Forest's 81.6%. This could be attributed to KNN's ability to capture local patterns within the data that are significant in predicting survivors. Conversely, Random Forest is also advantageous because of its robustness to outliers and ability to handle complex feature interactions through its ensemble of decision trees. Random Forest is also less sensitive than KNN to the curse of dimensionality and offers probabilistic assessments of survival, which can enhance interpretability for historical datasets. While KNN excels in simplicity and immediate neighborhood-based predictions, Random Forest offers broader analytical depth and better generalization. Both models significantly outperform the naive rule baseline of 61.6% accuracy, underscoring their respective strengths and utility for exploring the Titanic dataset, each bringing their respective advantages to this historical analysis.

Predictive Tool: Cosine Similarity

We developed an interactive survival prediction tool for the Titanic dataset that utilizes Cosine Similarity to estimate survival chances based on user-submitted profiles. This approach would let the general public explore hypothetical scenarios: how likely they would have survived the Titanic disaster given their attributes.

We chose Cosine Similarity because it focuses on the angle between attribute vectors, highlighting relational patterns in the data by comparing directions rather than distances. Manhattan Distance was also a choice we considered. However, we opted against it because Manhattan Distance is sensitive to small differences across numerous dimensions, which may obscure meaningful interactions between attributes.

The algorithm starts with users providing their details, such as age, title, sex, ticket class, and port of embarkation. These inputs are first processed to align with our model's requirements: categorical variables are one-hot-encoded and continuous variables are standardized with standard scaling. Upon processing, the algorithm computes the Cosine Similarity between the feature vector of the inputted user profile and those of the passengers in the dataset. The algorithm will identify the top 5 most similar passengers and calculate the average survival of these passengers, which will then be used to determine the user's survival. We decided on a cutoff value of 0.5, where if the survival rate is equal to or greater than 0.5, it means the user survived, and if not, then the user did not survive the Titanic. This provides a forecast that links the user's input to historically similar cases.

Potential Areas for Improvement

Survey in real-time

Our original project premise was contingent upon a questionnaire users would fill out that would put them in the shoes of passengers on the Titanic, and then receive an automatic result of survival or no survival. As our project was carried out we investigated many ways we could do this. Given more time, we would use APIs compatible with our Google Colab notebook that would run the survey results automatically through our model.

Investigating further into our predictive models

For K-Nearest Neighbors and Random Forest, we primarily focused on accuracy as an indicator of good model performance. To improve and strengthen our analysis, we could have explored further other evaluation metrics like precision, recall, and F1 score to gather deeper insights. While accuracy is a common metric for classification models, it may not provide the whole picture of performance.

Find more survival rates of ship datasets

Another potential way we could have explored further is to look for more ship datasets about survival to identify if there are similar features amongst ship survival rates. We know from history that one of the reasons for the low survival rates on the Titanic was insufficient lifeboats, so we are curious what variables are important amongst other datasets. With more data, we could compare the importance of features like sex, wealth, age to other survival rates.

Utilize a platform other than Google Colab

The main issue with Google Colab was the inability to actually collaborate with our teammates and code together. When one person was on the file and another person attempted to code, the file would have issues saving and updating in real-time. This resulted in slowed down work and inability to work on more than one computer at the same time. We managed to get around this by prioritizing in-person meetings so that we could collaborate together and vocally communicate who was using the colab and making new changes. We tried to keep it running primarily on one laptop to maintain a base saved colab.

Conclusion

Final thoughts on our project

We learned a lot about interpreting different models and understanding the bigger picture. We took what we learned in class and tried to combine it with data visualization to make it easier to understand. Reflecting on our project, we enjoyed diving deeper into the Titanic and growing our love for the movie from an analytical standpoint. We also were able to be creative and come up with different personas to test our recommendation algorithm. This was exciting for us, given our love for the movie, as we were able to revisit some of our favorite characters. We were pleased to see strong accuracy results from our predictive models and learn more about the historical context from that time. Our most surprising finding was our Random Forest feature importance, where *male* and the *Title Mr.* significantly decreased the likelihood of survival as we expected gender and wealth to be significant factors. Over the semester, our group grew in analytical understanding and interest in this topic and we are excited to share our findings.

Appendix A

Titanic Data Dictionary

Variable	**Definition**	**Key**
survival	Survival	0 = No, 1 = Yes
pclass	Ticket Class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	Age is fractional if less than 1
Age	Age in Years	
sibsp	# of sibling/ spouses aboard the Titanic	Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored)
parch	# of parents/ children aboard the Titanic	Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children traveled only with a nanny, therefore parch=0 for them.
ticket	Ticket Number	
fare	Passenger Fare	
cabin	Cabin Number	

embarked	Port of Embarkation	C = Cherbourg,
		Q = Queenstown,
		S = Southampton

AI Acknowledgement:

We would like to disclose the use of Chat GPT for assistance with our code on Google Colab.

Bibliography

Cameron, J. (1997). Titanic. Paramount Pictures.

Cukierski, W. (2012). *Titanic - Machine Learning from Disaster*. Retrieved from https://kaggle.com/competitions/titanic

Tikkanen, A. (2024, April 19). Titanic. Encyclopedia Britannica. https://www.britannica.com/topic/Titanic