Would You Survive the Titanic?

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

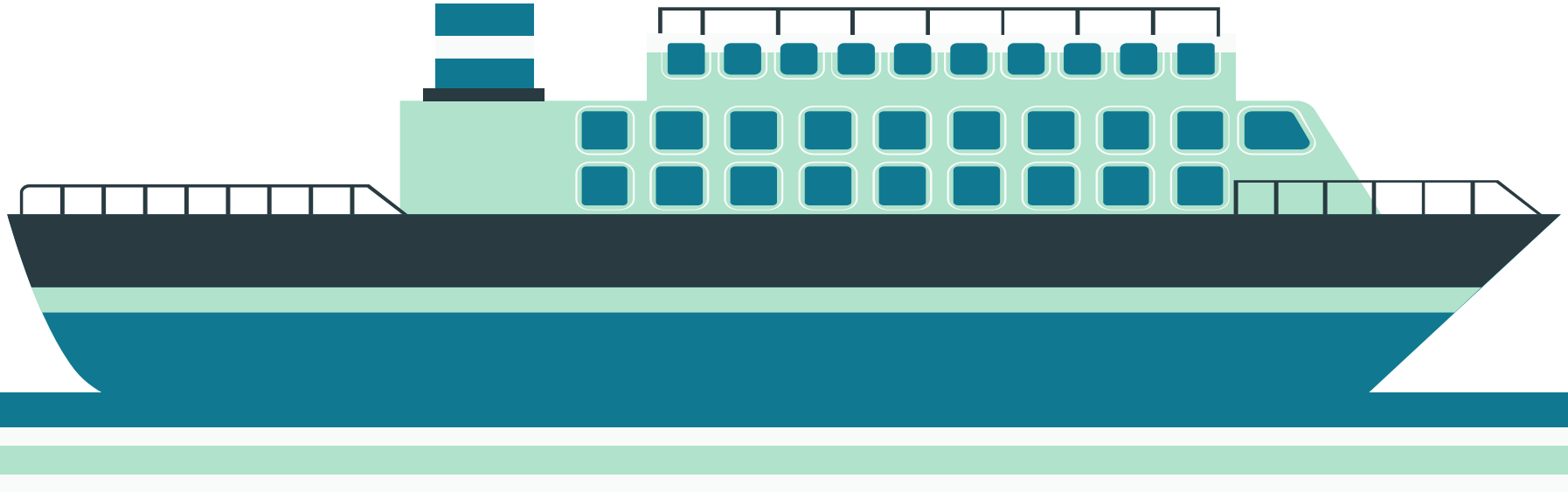
Team A4

Agenda

1. Selecting our Dataset
2. Methodology
3. Predictive Models
4. Interactive Test
5. Potential Improvement

01

Our Dataset



Introduction to our Dataset

Selecting our Dataset

Initial Interests: social media, dating applications, crime statistics
Titanic Dataset: we are all huge fans of the movie, wanted to explore this through an analytical lens

Problem Statement

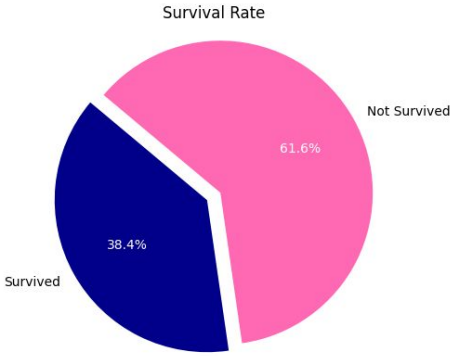
To understand why some people survived while others did not

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

Survival of the Titanic Dataset: — 891 Records of Passengers

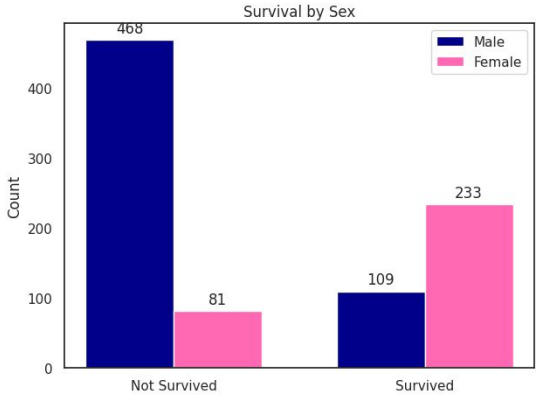
Exploratory Analysis

Overall Survival Rate



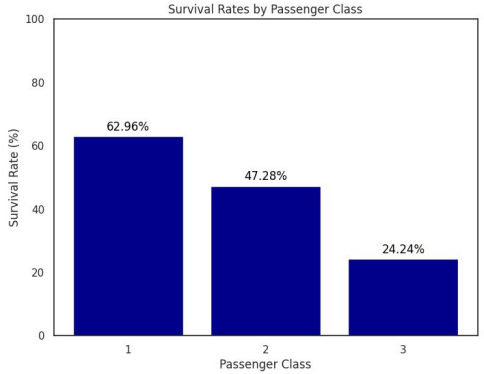
Out of 891 passengers, **only 38.4% survived** and therefore **61.6% acts as our naive rule** to compare our models' accuracies

Survival by Sex



Although there is a **higher population of males**, their survival count is **significantly lower than females**

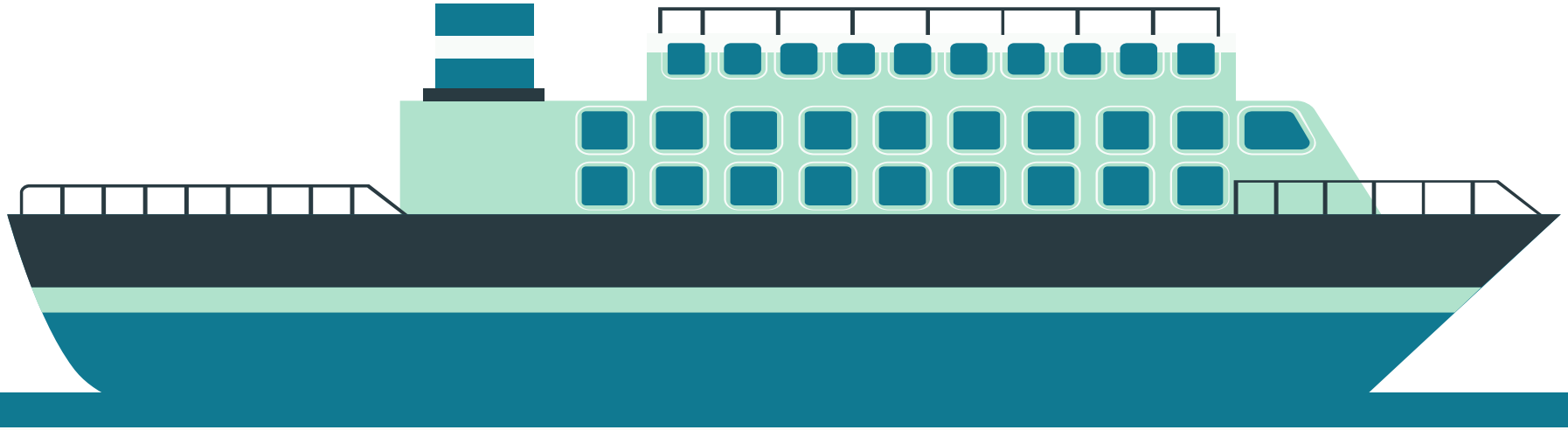
Survival by Class



Even though 3rd Class passengers are the **highest population**, their **survival rate is the lowest**

02

Methodology



Preprocessing Steps

Identify missing values

Fill missing values in *Age* and *Embarked*

Create new column

Split unique names from their *Titles*

Drop unnecessary columns

Drop *Name, Title, Cabin*

Create dummy variables

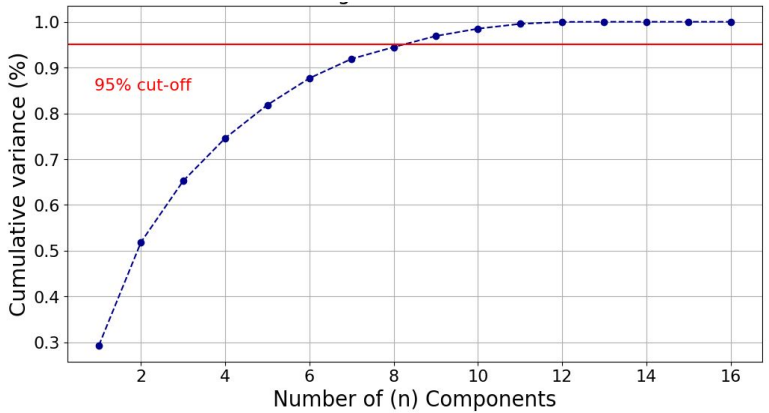
Convert categorical data into dummy variables

Normalization

Standardized Scaling

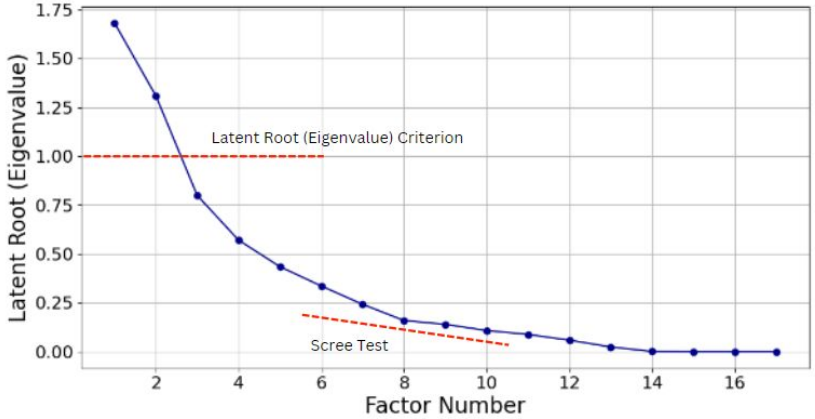
PCA Output

Percentage of Variance Criterion



Eight components explain 95% of the variance. While a high cut-off point, it did not overfit the data.

Latent Root/Scree Test Criterion



Only 2 components have an Eigenvalue greater than 1. For the Scree Test, we interpreted 8 to be the point of inflection.

PCA Component Matrix

Correlation between a Particular Variable and the Component

	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

PC1

Family Size Aboard

PC2

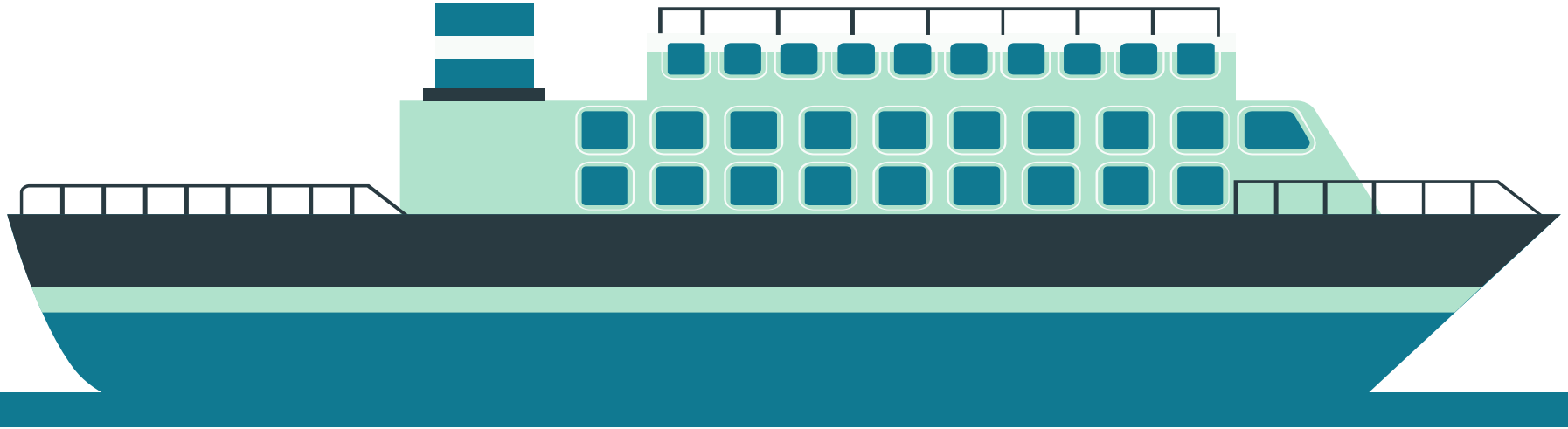
Fare type

PC3

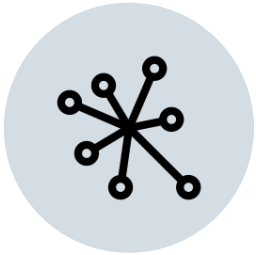
Age of Siblings

03

Predictive Models

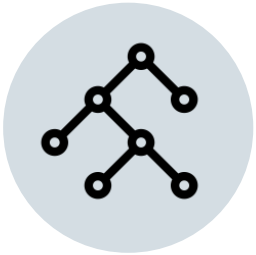


Predictive Models



K Nearest Neighbors

- Predictions based on nearest examples
- Simple and intuitive
- Capture local noise for medium dataset



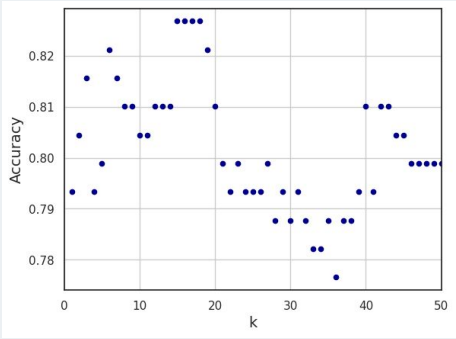
Random Forest

- Less likely to overfit than a decision tree
- Rank features by importance
- Less sensitive to outliers

Naive Rule Used the majority class (Not Survived) of 61% as our benchmark for comparison of accuracies

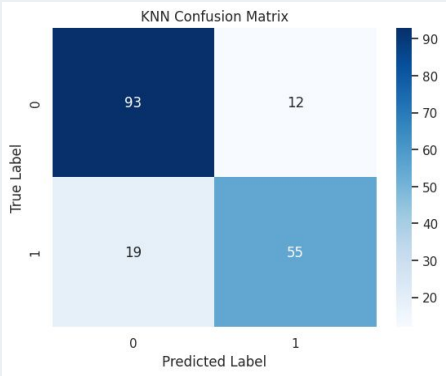
1: KNN

K = 15



K = 15 yielded the highest accuracy of 82.68%, ensuring a definitive outcome

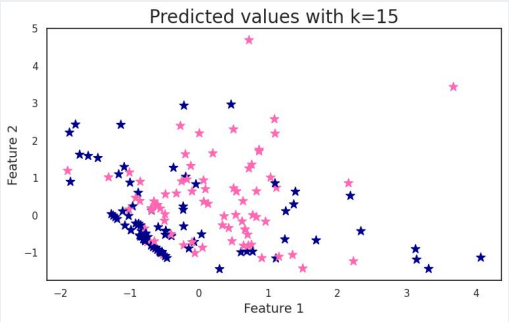
Overfitting?



Train accuracy: 82.682%

Test accuracy: 82.584%

Limitations



Sensitivity to outliers and curse of dimensionality

2: Random Forest

78.21%

Accuracy
without tuning

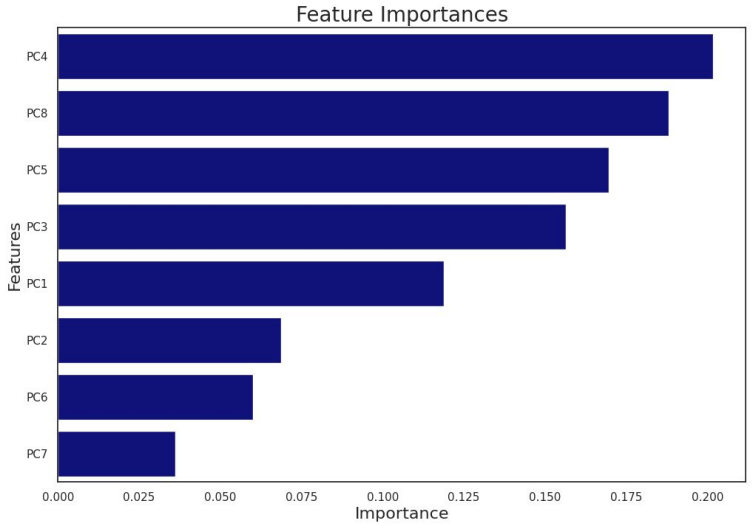
81.01%

Test Accuracy
Hyperparameter tuning & GridSearchCV

84.97%

Train Accuracy
for comparison

Feature Importance

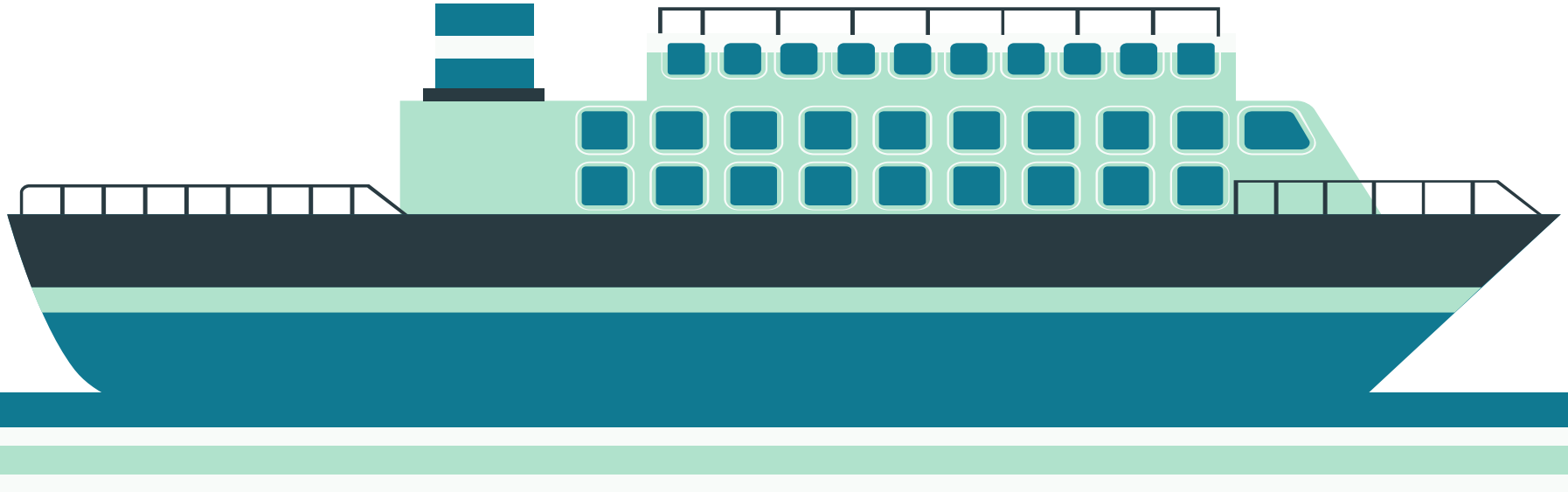


PC4: *Parch*, *Sex_male*, and *Title Mr* ⟶ Middle-class Father

PC8: *Title Mrs* and *Title Miss* ⟶ Women

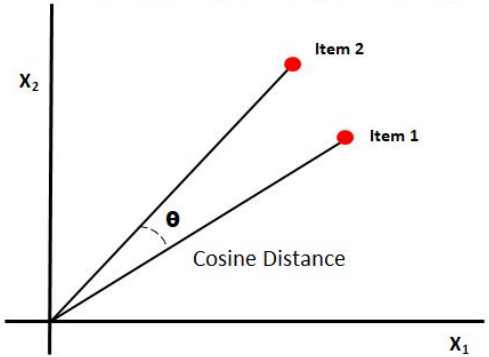
04

Interactive Test



Cosine Similarity

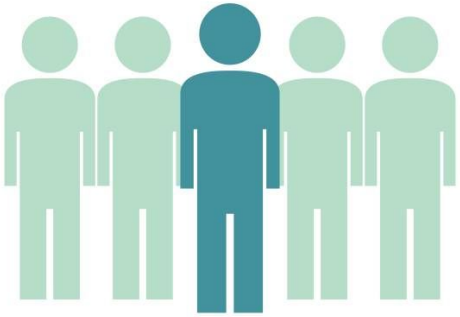
Reasons for Choice:



Focuses on **angle between vectors**

Opted against **Manhattan Distance** because of its **sensitivity to small differences across dimensions**

Recommendation Tool:



Chooses **top 5** passengers most similar to you

Calculates the **average** of their survival rate with cutoff of 0.5 to determine **YOUR** survival rate

Personas

Mr. Jack O'Malley
Age: 23
No Children
Passenger Class: 3
Fare: $10
Embarked: C

Mrs. Isadora Montgomery
Age: 45
Spouse Aboard: 1
Passenger Class: 1
Fare: $50
Embarked: S

Miss Rose Cordelia Fairchild
Age: 20
Parent Aboard: 1
Passenger Class: 2
Fare: $30
Embarked: S

Master Bartholomew Pembroke
Age: 35
Spouse Aboard: 1
Passenger Class: 2
Fare: $60
Embarked: C

Results

Mr. Jack O'Malley
Age: 23
No Children
Passenger Class: 3
Fare: $10
Embarked: C

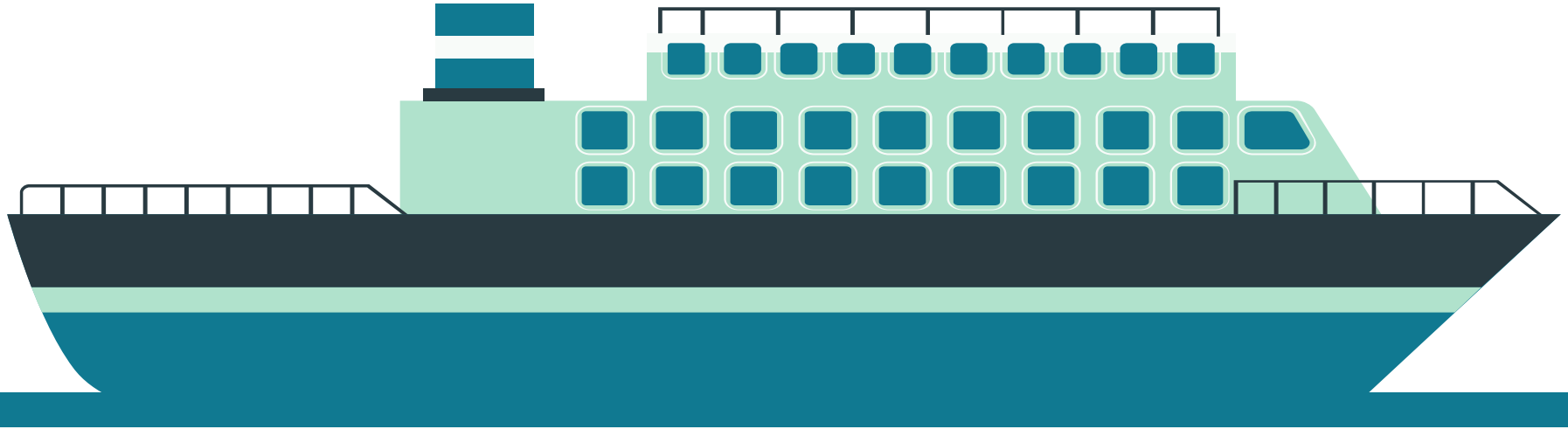
Mrs. Isadora Montgomery
Age: 45
Spouse Aboard: 1
Passenger Class: 1
Fare: $50
Embarked: S

Miss Rose Cordelia Fairchild
Age: 20
Parent Aboard: 1
Passenger Class: 2
Fare: $30
Embarked: S

Master Bartholomew Pembroke
Age: 35
Spouse Aboard: 1
Passenger Class: 2
Fare: $60
Embarked: C

05

Potential Improvement



Potential Areas of Improvement

Conduct survey in real time

Utilize APIs or Google Forms into our code

Investigate further into our models

Look at other evaluation metrics like precision, recall, F1 score

Find more ship survival datasets

Potentially identify similar features amongst ship survival rates

Difficulty with Google Colab

Inability to actually collaborate and code together

Thank You