In this assignment, I used the built-in Titanic dataset of Seaborn Library to visualize data and train the Logistic Regression model to answer my question: “Would I survive if I were in the Titanic Disaster?”

Author: The Hai Nguyen – 991745555

Class number: PROG25211

Assignment 1:

Would I survive if I were in the Titanic Disaster?

Contents

[I. Choose a dataset 1](#_Toc179122175)

[II. Ask a question 2](#_Toc179122176)

[III. Clean the dataset 2](#_Toc179122177)

[1. Remove irrelevant fields 2](#_Toc179122178)

[2. Remove all rows having missing values 2](#_Toc179122179)

[3. Convert categorical data into numerical values 2](#_Toc179122180)

[IV. Graph the dataset 3](#_Toc179122181)

[1. Bar plot 1: Survival rates by sex 3](#_Toc179122182)

[2. Bar plot 2: Survival rates by who 4](#_Toc179122183)

[3. Bar plot 3: Survival rates by sibsp 5](#_Toc179122184)

[4. Bar plot 4: Survival rates by parch 6](#_Toc179122185)

[5. Bar plot 5: Survival rates by male age groups 7](#_Toc179122186)

[V. Train and test the algorithm 8](#_Toc179122187)

[VI. Evaluate the model 8](#_Toc179122188)

[VII. Answer the question 8](#_Toc179122189)

[VIII. Conclusion 9](#_Toc179122190)

## Choose a dataset

In this assignment, I decided to use the dataset taught in the class: the [Titanic dataset](https://github.com/mwaskom/seaborn-data/blob/master/titanic.csv) (built-in dataset of Seabron library).

The Titanic dataset contains information about the passengers and crew of the RMS Titanic, a British passenger liner that sank in the North Atlantic Ocean in 1912 after striking an iceberg during her maiden voyage from Southampton, UK, to New York City, US.

## Ask a question

My question: Would I survive if I were in the Titanic Disaster?

Assumption: I was a 19-year-old adult male taking the second class and only went with my older brother.

## Clean the dataset

### Remove irrelevant fields

**- “embark\_town”** and **“embark”:** Those fields present the port of embarkation, which doesn’t have any impact on the survival rates.

**- “deck”** and **“fare”:** Those fields do not have any impact on the survival rates.

**- “class”** and **“alive”:** These columns are directly derived from other variables: "class" can be inferred from the "pclass" column, and "alive" can be inferred from the "survived" column. Those columns needed to be removed to decrease the redundancy.

### Remove all rows having missing values

Remove all missing values is to improve the performance of the Logistic Regression Model and avoid unexpected errors

### Convert categorical data into numerical values

- Replace values “male” and “female” in “sex” field by “0” and “1” respectively

- Replace values “man”, “woman”, and “child” values with “0”, “1”, and “2” respectively

**- Explanation:** Converting categorical data into numerical values is essential for training machine learning models because most algorithms are designed to work with numerical data.

## Graph the dataset

### Bar plot 1: Survival rates by sex

A graph of survival rates by sex

Description automatically generated

**Brief description:** The bar plot shows the survival rates of male and female passengers on the Titanic.According to the plot, there is a significantly higher proportion of females survived compared to males. Specifically, 75% of females survived, while only 21% of males survived. This suggests that women were given priority in lifeboat assignments during the evacuation.

### Bar plot 2: Survival rates by who

A graph with different colored bars

Description automatically generated

**Brief description:** The bar plot compares the survival rates of different passenger groups on the Titanic, categorized as "who." The results show that women had the highest survival rate (78%), followed by children (59%), while male adults had the lowest survival rate (17%). This confirms the "women and children first" policy implemented during the evacuation, as children and women were prioritized for lifeboat boarding, leading to their higher chances of survival.

### Bar plot 3: Survival rates by sibsp

A graph of survival rates

Description automatically generated

**Brief description:** The bar plot illustrates the relationship between the number of siblings/spouses (sibsp) and passenger survival rates on the Titanic. It shows that passengers with 1 sibling/spouse had the highest survival rate, while those with larger family groups faced lower chances of survival. This suggests that having more family members on board negatively impacted the ability to secure lifeboat spaces.

### Bar plot 4: Survival rates by parch

A graph of survival rates

Description automatically generated

**Brief description:** The bar plot illustrates the relationship between the number of parents/children (parch) on board the Titanic and passenger survival rates. It shows that passengers with 3 parents/children (parch = 3) had the highest survival rate, while those with larger family groups faced lower chances of survival. This suggests that having more family members on board negatively impacted the ability to secure lifeboat spaces.

### Bar plot 5: Survival rates by male age groups

A graph of a number of people

Description automatically generated

**Brief description:** The bar plot illustrates the survival rates of male passengers on the Titanic, categorized by age group. It reveals that younger males had significantly higher survival rates compared to older males. As age increased, the proportion of surviving male passengers declined, with the lowest survival rate observed for males aged 60+. This suggests that factors such as physical fitness, social standing, or prioritization in lifeboat assignments may have influenced the survival chances of male passengers based on their age.

## Train and test the algorithm

A screenshot of a computer

Description automatically generated

## Evaluate the model

* The trained Logistic Regression model achieved an accuracy of 0.805 (80.5%) on the testing data, indicating relatively good performance.
* Ideas to improve the accuracy of the model:
* **Collect additional data:** If possible, collect more data to improve the model's training. A larger dataset can help the model learn more complex patterns and make better predictions.
* **Random Forest:** Consider using ensemble methods like Random Forest, which combines multiple decision trees to improve prediction accuracy and reduce overfitting.

## Answer the question

A black rectangular object with white text

Description automatically generated

According to the model’s prediction, I would not survive during the Titanic Disaster. The result is as good as I expected because based on all the graphs above, my survival rate is relatively low.

## Conclusion

In this assignment, I decided to use the Titanic dataset to visualize data and train the Logistic Regression model. The goal of this assignment is to answer my question: “Would I survive If I were in the Titanic Disaster?”. Unfortunately, the trained Logistic Regression model with the accuracy of 80.5% predicts that I would not survive after the Titanic Disaster.