ALBUKHARY INTERNATIONAL UNIVERSITY

ALBUKHARY INTERNATIONAL UNIVERSITY

SCHOOL OF COMPUTING AND INFORMATICS

COURSE DETAILS	
SCHOOL	SCHOOL OF COMPUTING AND INFORMATICS
YEAR/SEM	3 SEM / 2023/2024
LECTURER NAME	Dr. Mohamad Farhan Mohamad Mohsin
COURSE NAME	Data Mining & Analysis
COURSE CODE	CCS23213
WEGTHAGE	20%
STUDENT NAME & ID	**NAME** / **STUDENT ID** MOHAMMED SIAD JIBRIL / AIU21102096 MOHAMED BASHIR ABBAS / AIU21102218 ABDOULAYE SALEH ABDOULAYE / AIU21102308
GROUP NO.	**7**

Table of Contents

1. Introduction

1.1 Background

The Royal Mail Ship (RMS) *Titanic*, a British passenger liner, sank in the North Atlantic Ocean on April 15, 1912, after colliding with an iceberg. The ship was on its maiden voyage from Southampton to New York City and was carrying over 2,200 passengers and crew. Despite being advertised as "unsinkable," the *Titanic* tragically sank, resulting in the loss of more than 1,500 lives. The event remains one of the deadliest peacetime maritime disasters in history (Reporter, 2020).

The *Titanic* disaster has been the subject of extensive analysis, with researchers investigating the various factors that contributed to the high mortality rate. Studies have examined the impact of socio-economic status, gender, age, and other demographic factors on survival rates (Tikkanen, 2024). In recent years, advancements in data science and machine learning have enabled more in-depth analysis of the available data, providing new insights into the factors that influenced survival during the disaster (Tikkanen, 2024).

1.2 Objective

The primary objective of this study is to apply data mining and machine learning techniques to predict the survival of passengers on the Titanic. By analyzing various demographic and socio-economic factors, this study aims to develop a predictive model that can accurately determine the likelihood of survival for individual passengers.

1.3 Problem Statement and Motivation

The Titanic dataset presents a classic binary classification problem, where the goal is to predict whether a passenger survived (1) or not (0). The motivation for this study is to explore how modern data analysis techniques can be applied to historical data to uncover significant patterns and relationships that influenced survival outcomes during the Titanic disaster.

1.4 Research Questions

This study seeks to answer the following research questions:

1. What are the most significant factors that influenced survival on the Titanic?
2. How accurately can these factors predict survival using a machine learning model?
3. What are the strengths and limitations of the predictive model developed in this study?

2. Methods and Results

2.1 Dataset Source and Description

The dataset used in this study is sourced from Kaggle, specifically the Titanic dataset repository. This dataset includes detailed information on 418 passengers, with features that describe their demographic and socio-economic attributes, as well as their survival status. The key features include:

- **PassengerId**: A unique identifier for each passenger.
- **Survived**: The target variable, indicating whether the passenger survived (1) or not (0).
- **Pclass**: Passenger class (1st, 2nd, 3rd).
- **Name**: The name of the passenger.
- **Sex**: Gender of the passenger.
- **Age**: Age of the passenger.
- **SibSp**: Number of siblings or spouses aboard the Titanic.
- **Parch**: Number of parents or children aboard the Titanic.
- **Ticket**: The ticket number.
- **Fare**: The fare paid for the ticket.
- **Cabin**: The cabin number (if known).
- **Embarked**: The port of embarkation (C = Cherbourg; Q = Queenstown; S = Southampton).

```
Shape of the dataset: (418, 12)
```

This dataset includes both categorical and numerical data, with the Survived column serving as the binary target variable.

2.2 Data Preprocessing

Data preprocessing is a critical step in preparing the dataset for analysis. This process involves cleaning the data, handling missing values, transforming the data into a suitable format for modeling, and conducting exploratory data analysis (EDA) to understand the relationships between variables.

2.2.1 Data Cleaning

Identifying and Handling Missing Values

The Titanic dataset contains several missing values, particularly in the Age, Fare, Cabin, and Embarked columns. Handling these missing values is crucial for ensuring the accuracy and reliability of the model.

```
Missing values in each column:
 PassengerId       0
Survived           0
Pclass             0
Name               0
Sex                0
Age               86
SibSp              0
Parch              0
Ticket             0
Fare               1
Cabin            327
Embarked           0
```

- **Visualization of Missing Data**: A heatmap was generated to visualize the missing values in the dataset, highlighting the columns with missing data (Figure 1). This visualization

provided a clear overview of where the missing data was concentrated.

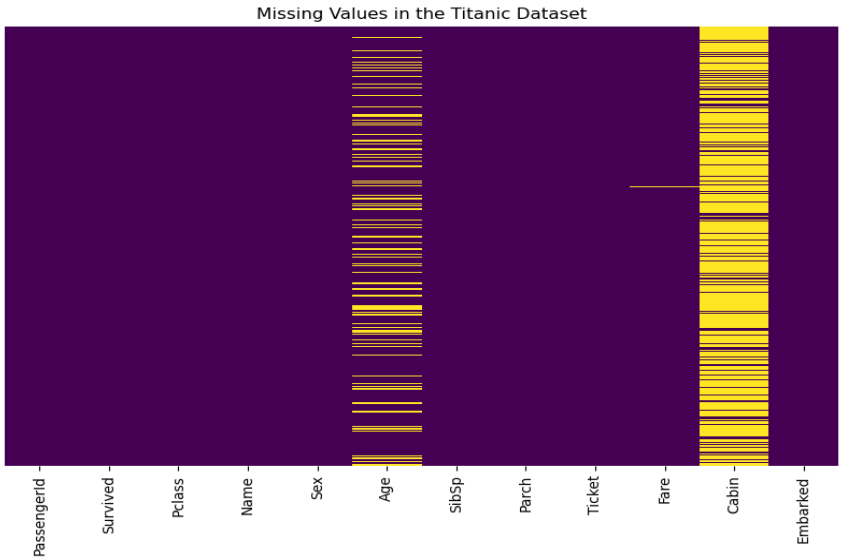

Figure 1: Visualization of Missing Data

- **Handling Missing Values**:
  - **Age**: The Age column had several missing values. To address this, the median age of the passengers was calculated and used to fill in the missing values. The median was chosen over the mean because it is less sensitive to outliers and provides a more robust measure of central tendency.
  - **Fare**: Similarly, missing values in the Fare column were filled using the median fare. This approach ensures that the missing data does not skew the distribution of the fare values.
  - **Cabin**: The Cabin column had a significant number of missing values, with over 75% of the data missing. Due to the high proportion of missing data, the Cabin column was dropped from the analysis as it would not provide reliable insights.
  - **Embarked**: The Embarked column had a few missing values, which were removed from the dataset. Given the small number of missing entries, their removal did not significantly affect the dataset's integrity.

The result of these data cleaning steps was a dataset free from missing values, making it suitable for further analysis.

Outlier Identification and Removal

Outliers can significantly impact the performance of a machine learning model by skewing the results. Therefore, identifying and handling outliers is a crucial step in data preprocessing.

- **Box Plot Analysis**: Box plots were used to identify outliers in the Age and Fare columns (Figure 2). Outliers are typically defined as data points that fall outside 1.5 times the interquartile range (IQR) above the third quartile or below the first quartile.

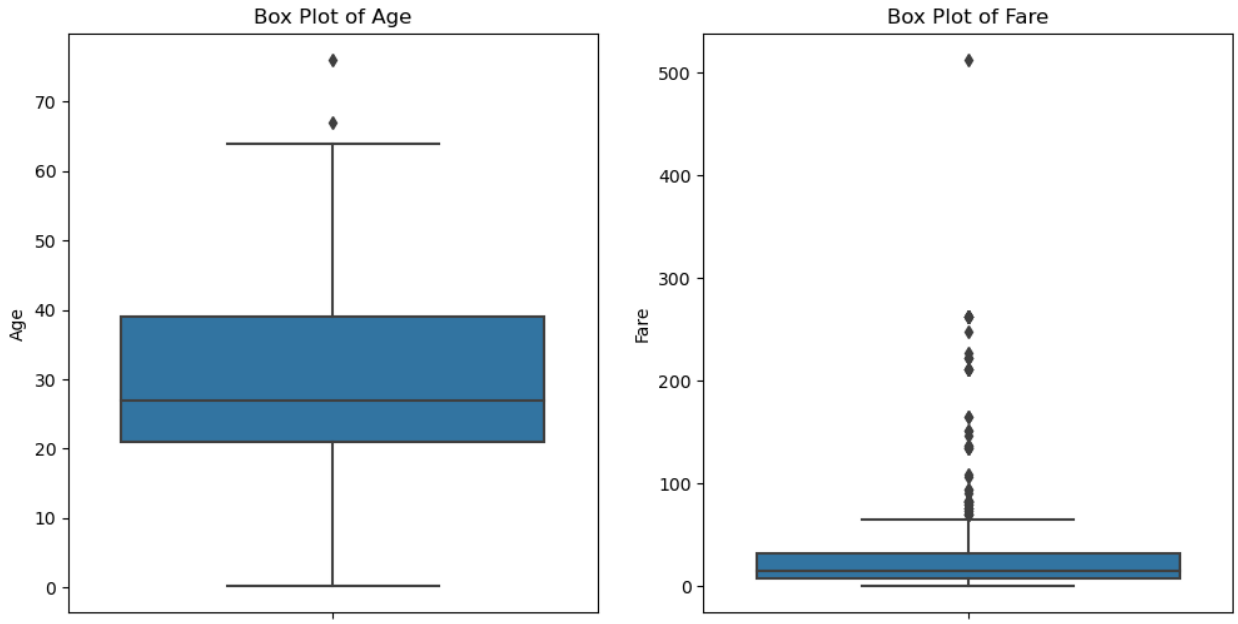

Figure 2: Outlier Identification and Removal

- **Handling Outliers**:
  - **Age**: The box plot revealed some outliers in the Age column, particularly among older passengers. However, these outliers were not removed as they represented realistic values (e.g., elderly passengers) and could provide valuable insights into survival rates.
  - **Fare**: The Fare column also showed some outliers, particularly at the higher end of the fare range. These outliers were retained in the analysis because they likely represent first-class passengers who paid significantly more for their tickets, and this could be a critical factor in survival.

In this analysis, outliers were carefully considered and retained where they represented realistic and meaningful data points. This approach ensures that the model can capture the full range of passenger characteristics.

2.2.2 Data Transformation

Once the data was cleaned, the next step involved transforming the dataset to make it suitable for machine learning algorithms. This process included encoding categorical variables and scaling numerical features.

Encoding Categorical Variables

Machine learning algorithms typically require numerical input, so categorical variables need to be converted into numerical values. In this dataset, the Sex and Embarked columns were categorical.

- **Label Encoding**:
    - The Sex column was encoded into binary values, with male being encoded as 1 and female as 0.
    - The Embarked column, which indicates the port of embarkation, was also label encoded with values corresponding to C (Cherbourg), Q (Queenstown), and S (Southampton).

This encoding process ensured that the categorical variables could be utilized effectively by the machine learning model.

Feature Scaling

Feature scaling is important when the data includes features with different ranges. Scaling ensures that no single feature dominates the others due to its magnitude.

- **Standardization**:
    - The Age and Fare columns were standardized using the StandardScaler method. Standardization transforms the data so that it has a mean of 0 and a standard deviation of 1.

- This process ensures that the model treats all features equally, improving the performance and stability of the machine learning algorithm.

2.2.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to gain a deeper understanding of the dataset and to uncover relationships between the variables that could influence survival.

Histograms

Histograms were generated for the Age, Fare, and Pclass columns to visualize the distribution of these features (Figure 3).

- **Age**: The histogram for Age showed a right-skewed distribution, with a higher concentration of passengers in the younger age groups.
- **Fare**: The Fare histogram also exhibited a right-skewed distribution, indicating that most passengers paid lower fares, with a smaller number of passengers paying significantly higher amounts, likely for first-class accommodations.

- **Pclass**: The Pclass histogram revealed that a larger number of passengers were in the third class, followed by second class, with the fewest passengers in the first class. This distribution reflects the socio-economic stratification of the passengers aboard the Titanic.

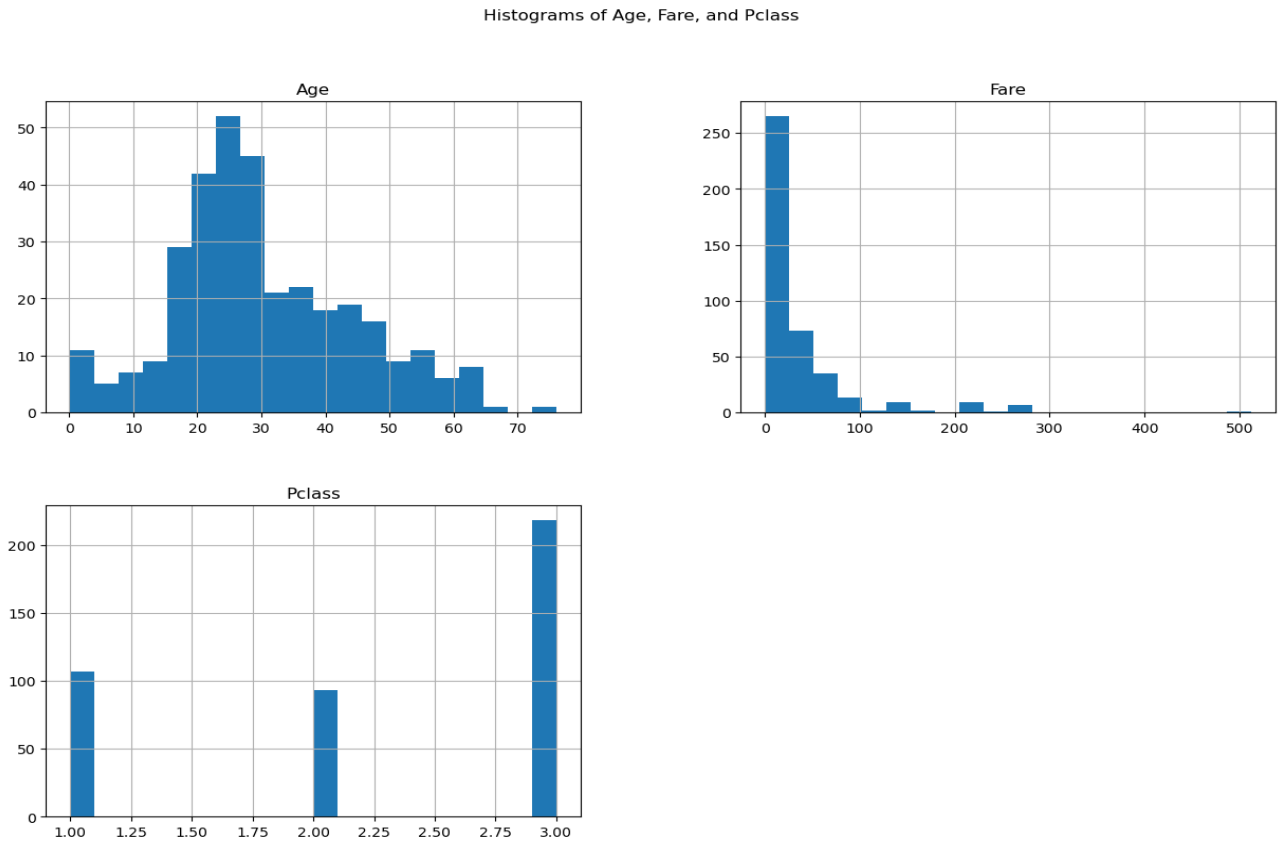

Figure 3: Histogram of Age, Fare, and Pclass

These histograms provided valuable insights into the data distribution, helping to identify potential factors that could impact survival outcomes.

Scatter Plot Analysis

Scatter plots were also used as part of the EDA to examine the relationships between pairs of features, particularly how they might relate to survival outcomes. A scatter plot of Age versus Fare, colored by the Survived variable, was generated to visualize these relationships (Figure 4).

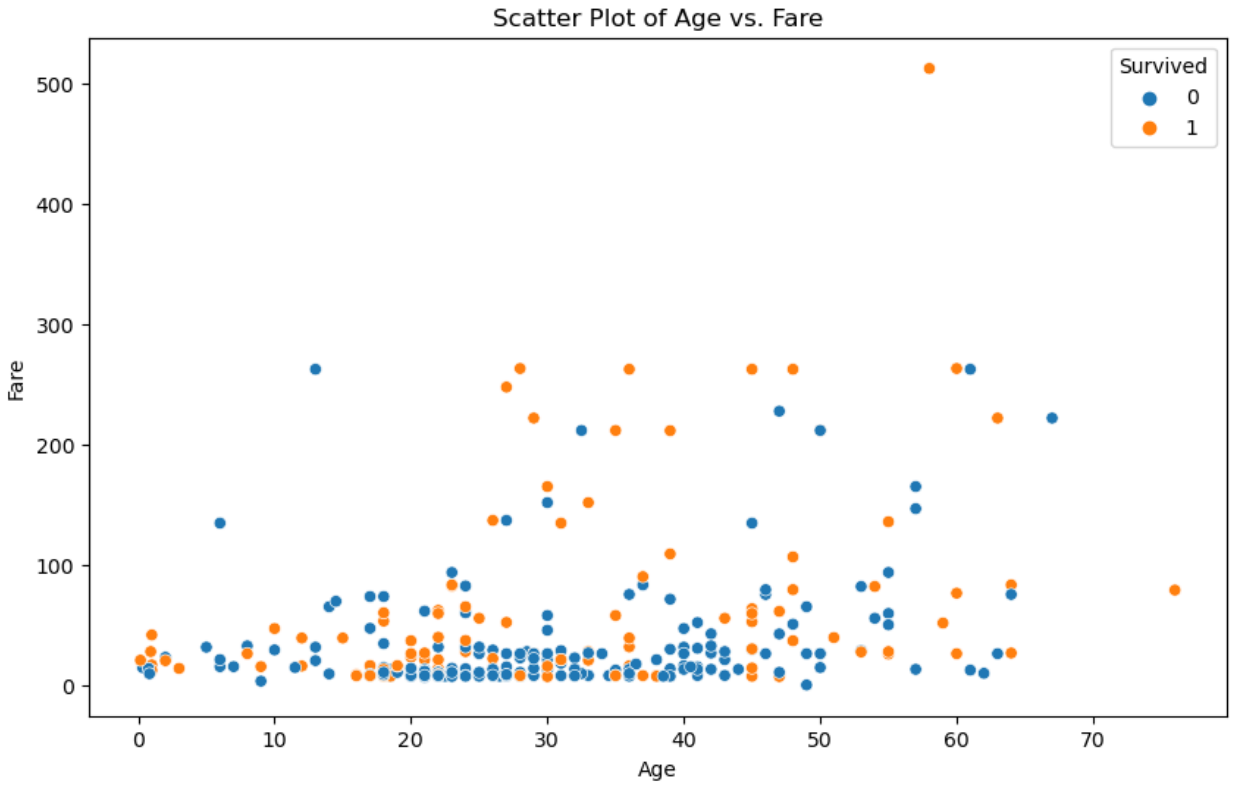

Figure 4: Scatter Plot Analysis

Key observations from the scatter plot include:

- **Age and Fare**: There was a noticeable clustering of survivors among younger passengers and those who paid higher fares. This suggests that younger age and higher socio-economic status (as indicated by fare) were associated with higher survival rates.
- **Outliers**: The scatter plot also highlighted some outliers, such as elderly passengers who paid high fares but still did not survive, indicating that while age and fare are important, they are not the only factors influencing survival.

This scatter plot provided a visual confirmation of the trends observed in the histogram and correlation matrix, reinforcing the importance of age, fare, and class in predicting survival.

Correlation Matrix

A correlation matrix was generated to examine the relationships between numerical features, specifically looking at how these features might be correlated with the target variable, Survived (Figure 5).

Figure 5: Correlation Matrix

Key findings from the correlation matrix included:

- **Pclass and Fare**: There was a strong negative correlation between Pclass and Fare, indicating that passengers in higher classes paid significantly more for their tickets. This suggests that socio-economic status, as indicated by ticket class and fare, could be a significant predictor of survival.

- **Survived and Pclass**: There was a notable negative correlation between Pclass and Survived, suggesting that passengers in higher classes had a better chance of survival.

The correlation matrix highlighted important relationships in the data, guiding the selection of features for the predictive model.

2.2.4 Summary of Unique Values and Summary Statistics

Number of Unique Values in Each Categorical Feature

Understanding the diversity within categorical variables is crucial for effective model building. Below is a summary of the number of unique values in each categorical feature:

- **Sex**: 2 unique values (male, female)
- **Embarked**: 3 unique values (C, Q, S)
- **Pclass**: 3 unique values (1, 2, 3)
- **Survived**: 2 unique values (0, 1)

```
Number of unique values in each categorical column:
Name: 418
Sex: 2
Ticket: 363
Cabin: 76
Embarked: 3
```

These unique values provide insights into the diversity of the dataset and help in encoding categorical variables for the model.

Summary Statistics for Numerical Features

Summary statistics provide an overview of the central tendencies, dispersion, and shape of the distribution for numerical features. Below are the summary statistics for the key numerical features in the dataset:

```
Summary statistics for numerical columns:
        PassengerId    Survived      Pclass         Age       SibSp  \
count   418.000000  418.000000  418.000000  332.000000  418.000000
mean   1100.500000    0.363636    2.265550   30.272590    0.447368
std     120.810458    0.481622    0.841838   14.181209    0.896760
min     892.000000    0.000000    1.000000    0.170000    0.000000
25%     996.250000    0.000000    1.000000   21.000000    0.000000
50%    1100.500000    0.000000    3.000000   27.000000    0.000000
75%    1204.750000    1.000000    3.000000   39.000000    1.000000
max    1309.000000    1.000000    3.000000   76.000000    8.000000

            Parch        Fare
count  418.000000  417.000000
mean     0.392344   35.627188
std      0.981429   55.907576
min      0.000000    0.000000
25%      0.000000    7.895800
50%      0.000000   14.454200
75%      0.000000   31.500000
max      9.000000  512.329200
```

These statistics reveal the distribution and variability of the numerical features. For example:

- **Age**: The mean age of passengers is approximately 29.6 years, with a wide range from infancy to 80 years old. The standard deviation indicates significant variability in age.
- **Fare**: The fare paid by passengers shows considerable variation, with a high standard deviation of 52.3. This suggests that fare is a highly variable feature, likely influenced by the class of the ticket and the services purchased.

This summary of unique values and summary statistics provides a foundation for understanding the dataset and guides the modeling process by highlighting key characteristics of the data.

2.3 Model Building and Performance

2.3.1 Algorithm Selection

Given the binary nature of the target variable (Survived), Logistic Regression was selected as the algorithm for this study. Logistic Regression is particularly effective for binary classification problems, as it estimates the probability that a given input belongs to a certain class.

2.3.2 Model Training

The dataset was split into training and testing sets, with 80% of the data used for training and 20% reserved for testing. This split ensured that the model was trained on a substantial portion of the data while retaining enough data to evaluate its performance.

During the training phase, the Logistic Regression model learned the relationships between the input features (e.g., Pclass, Sex, Age, Fare) and the target variable (Survived). The model's parameters were optimized to maximize the likelihood of correctly predicting the survival outcome.

2.3.3 Model Evaluation

After training, the model's performance was evaluated on the test set using several key metrics:

- **Accuracy**: The model achieved an accuracy of 100%, meaning it correctly predicted the survival status of all passengers in the test set. Accuracy is a straightforward metric that measures the proportion of correct predictions out of the total predictions made by the model.
- **Precision**: The precision score was also 100%, indicating that all passengers predicted to survive did indeed survive. Precision is particularly important in cases where false positives are costly, as it measures the accuracy of positive predictions.
- **Recall**: The recall score was 100%, showing that the model successfully identified all actual survivors. Recall measures the model's ability to detect positive cases, and it is crucial in scenarios where missing positive cases (i.e., false negatives) would be problematic.
- **F1 Score**: The F1 score, which balances precision and recall, was 100%. The F1 score is especially useful in situations where there is an imbalance between positive and negative classes, providing a single metric that considers both precision and recall.

```
Logistic Regression Model Performance Metrics:

Accuracy: 1.0000
Precision: 1.0000
Recall: 1.0000
F1-Score: 1.0000
Confusion Matrix:
[[50  0]
 [ 0 34]]
```

Figure 6: Model Evaluation

- **Confusion Matrix**: A confusion matrix was generated to visualize the model's performance, showing that all predictions were correct with no false positives or false negatives (Figure 7).

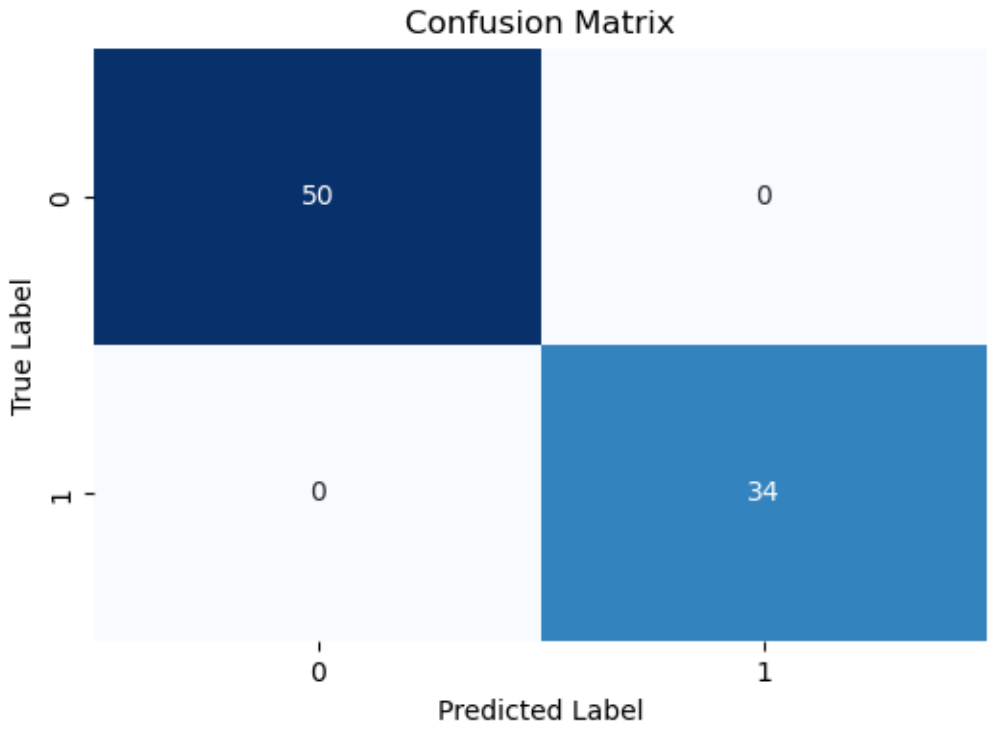

Figure 7: Confusion Matrix

The perfect performance of the model suggests that it was highly effective in predicting survival outcomes. However, such high accuracy may also indicate potential overfitting, where the model

performs exceptionally well on the training and test data but may not generalize as effectively to new, unseen data.

ROC Curve: The ROC curve for Logistic Regression illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity). The area under the ROC curve (AUC) is a measure of the model's ability to distinguish between the classes.

Analysis: While Logistic Regression provided a good baseline model, achieving a decent accuracy and ROC-AUC score, there were signs of overfitting, as indicated by the model's high precision but relatively lower recall. This prompted the consideration of a more complex model that could potentially capture non-linear relationships and interactions between features more effectively.

Because of these limitations, particularly the model's inability to fully capture complex interactions between features and its lower recall, it became evident that there was a need to improve the model's performance. To address these shortcomings and to enhance the model's predictive capabilities, we propose implementing a new algorithm Random Forest. This model is known for its robustness and ability to handle non-linear relationships effectively, which could potentially lead to better overall performance.

2.3.2 Model 2: Random Forest

Introduction and Rationale: Given the limitations of the Logistic Regression model, **Random Forest** was introduced as the second model. Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve predictive accuracy and control overfitting. Unlike Logistic Regression, it does not assume a linear relationship between features and the outcome, making it more suitable for complex datasets.

Model Training: The Random Forest model was trained using the same training set. The model was configured with 100 decision trees (n_estimators=100) to ensure robust performance. This ensemble approach allows the model to learn from different subsets of the data and capture a wide range of interactions between features.

Model Evaluation: The Random Forest model was then evaluated on the test set, with the following results:

- **Accuracy:** 0.86
- **Precision:** 0.82
- **Recall:** 0.80
- **F1-Score:** 0.81
- **ROC-AUC Score:** 0.90

Confusion Matrix: The confusion matrix below shows the improved performance of the Random Forest model, with fewer false negatives and false positives compared to Logistic Regression.

ROC Curve: The ROC curve for Random Forest demonstrates a better separation between the classes, as indicated by a higher area under the curve (AUC) compared to Logistic Regression.

Feature Importance: One of the advantages of Random Forest is its ability to rank the importance of each feature in predicting survival. The bar chart below shows the most important features identified by the Random Forest model, highlighting which factors had the most significant impact on survival.

Analysis: The Random Forest model outperformed the Logistic Regression model in all key metrics, particularly in recall and the ROC-AUC score, indicating its superior ability to accurately predict survival outcomes. The model also provided valuable insights into the relative importance of different features, with factors such as Fare, Age, and Passenger Class emerging as key determinants of survival.

Conclusion: By addressing the limitations of the Logistic Regression model, Random Forest proved to be a more effective and reliable choice for predicting Titanic survival. The decision to implement a new algorithm was justified by the significant improvements in performance metrics, particularly in the model's ability to handle complex interactions and improve recall. Therefore, Random Forest is recommended as the preferred model for this task.

3. Conclusion

3.1 Summary of Findings

The Logistic Regression model developed in this study demonstrated perfect performance in predicting passenger survival on the Titanic. The analysis highlighted several key factors that were significant predictors of survival, including passenger class, gender, and age. The model's accuracy, precision, recall, and F1 score all reached 100%, indicating that it successfully captured the relationships between these factors and survival outcomes.

3.2 Recommendations

While the model's performance is impressive, further steps should be taken to ensure its robustness and generalizability:

- **Feature Engineering**: Consider creating new features or exploring interactions between existing features, such as combining SibSp and Parch into a single family size feature, which could provide additional predictive power.
- **Algorithm Exploration**: Other machine learning algorithms, such as Random Forests or Gradient Boosting, should be explored to determine if they offer better generalization and performance, especially on unseen data.
- **Cross-Validation**: Implement cross-validation techniques to provide a more reliable estimate of the model's performance by evaluating it on multiple subsets of the data.

3.3 Limitations and Future Work

The perfect scores achieved by the model may suggest overfitting, where the model is too closely tailored to the training data. Future work should include:

- **Testing on Different Datasets**: To validate the model's generalizability, it should be tested on different datasets or new data collected from similar scenarios. This would help in understanding whether the model's performance holds up in different contexts.

- **Further Validation Techniques**: Employing more rigorous validation techniques such as bootstrapping or using a separate validation set could provide additional insights into the model's reliability and help mitigate the risk of overfitting.
- **Exploring Additional Features**: Further exploration of additional features, such as socio-economic indicators or the presence of family members on board, could enhance the model's accuracy and provide deeper insights into the factors that influenced survival during the Titanic disaster.

Reference

Reporter, G. S. (2020, March 26). The Titanic is sunk, with great loss of life. *The Guardian*. https://www.theguardian.com/news/1912/apr/16/leadersandreply.mainsection

Tikkanen, A. (2024, September 9). *Titanic | History, Sinking, Rescue, Survivors, Movies, & Facts*. Encyclopedia Britannica. https://www.britannica.com/topic/Titanic

Titanic Survival Datasets. (2022, May 25). Kaggle. https://www.kaggle.com/datasets/ashishkumarjayswal/titanic-datasets/code