**"Survival Prediction on the Titanic: A Comparative Analysis Using Machine Learning Models in R"**

**Name** : Kushal Dutia

**Roll No.** : A016

**SAP ID** : 86052200023

1. **Introduction**

The sinking of the RMS Titanic in April 1912 remains one of the most infamous maritime disasters in history. Out of the approximately 2,224 passengers and crew members aboard, more than 1,500 lost their lives when the ship struck an iceberg during its maiden voyage from Southampton to New York City (*Encyclopaedia Britannica.*, 2024).The tragedy not only marked a significant moment in maritime history but also revealed stark differences in survival rates among passengers based on their socio-economic status, age, gender, and class (Lord, 2005)

In the age of data science and machine learning, the Titanic dataset has become a widely used benchmark for learning and applying classification algorithms (*Titanic*, 2024).It offers a rich mix of categorical and numerical data, and allows for exploratory data analysis, visualization, and predictive modeling. By leveraging data-driven techniques, we can uncover patterns and relationships in the dataset that may have influenced the survival of passengers.

This project uses a subset of the Titanic passenger data to explore how different machine learning algorithms perform in predicting survival. We focus on three commonly used classification techniques—Logistic Regression, K-Nearest Neighbors (KNN), and Decision Tree—to develop models that estimate the probability of a passenger surviving the disaster. Each method offers unique strengths: logistic regression provides a probabilistic framework; KNN offers simplicity and intuitive logic; and decision trees bring interpretability and structure (James et al., 2013).

Beyond just model accuracy, this project aims to provide insights into the predictive power of various features and examine how well these algorithms generalize on unseen data. The findings can also demonstrate the potential and limitations of different algorithms when applied to real-world classification problems.

1. **Rationale**

The field of data science has revolutionized how we interpret, understand, and derive insights from complex datasets. One of the most popular datasets in the data science community is the Titanic dataset, which offers a historical yet structured setting to apply and compare various machine learning algorithms. Despite its simplicity, the dataset reflects real-world characteristics—missing values, categorical variables, and non-linear relationships—that make it ideal for learning and testing classification models.

This project aims to not only predict survival outcomes using machine learning models but also to critically examine the effectiveness of different algorithms in doing so. The rationale behind choosing this dataset lies in its educational value, interpretability, and the rich context it provides. It bridges the gap between theoretical concepts and practical implementation by offering a meaningful and historic scenario in which modern-day analytics can be applied.

Moreover, the Titanic dataset provides a balanced challenge for both beginners and intermediate practitioners. It includes variables such as age, gender, ticket class, and fare—which can be intuitively understood by humans—allowing for the visualization and justification of model decisions. This interpretability is essential in evaluating not just how models perform but *why* they make certain predictions.

Another key reason for choosing this problem is that it allows for experimentation with a range of techniques, from basic statistical models like **Logistic Regression** to more complex algorithms like **K-Nearest Neighbours (KNN)** and **Decision Trees**. Each method brings unique strengths to the table—logistic regression provides explainability, KNN emphasizes proximity-based learning, and decision trees excel at rule-based segmentation. This project allows for a comparison of these approaches in terms of accuracy, usability, and interpretability, which is valuable for anyone aiming to work with predictive analytics in real-world scenarios.

Ultimately, this exercise is not just about building predictive models, but about enhancing understanding of the data, improving problem-solving skills, and gaining a deeper appreciation of the role machine learning plays in data-driven decision-making.

1. **Aim and Objectives :**

**Aim:**

To build and compare classification models that predict the survival of passengers aboard the Titanic using different machine learning techniques in R.

**Objectives:**

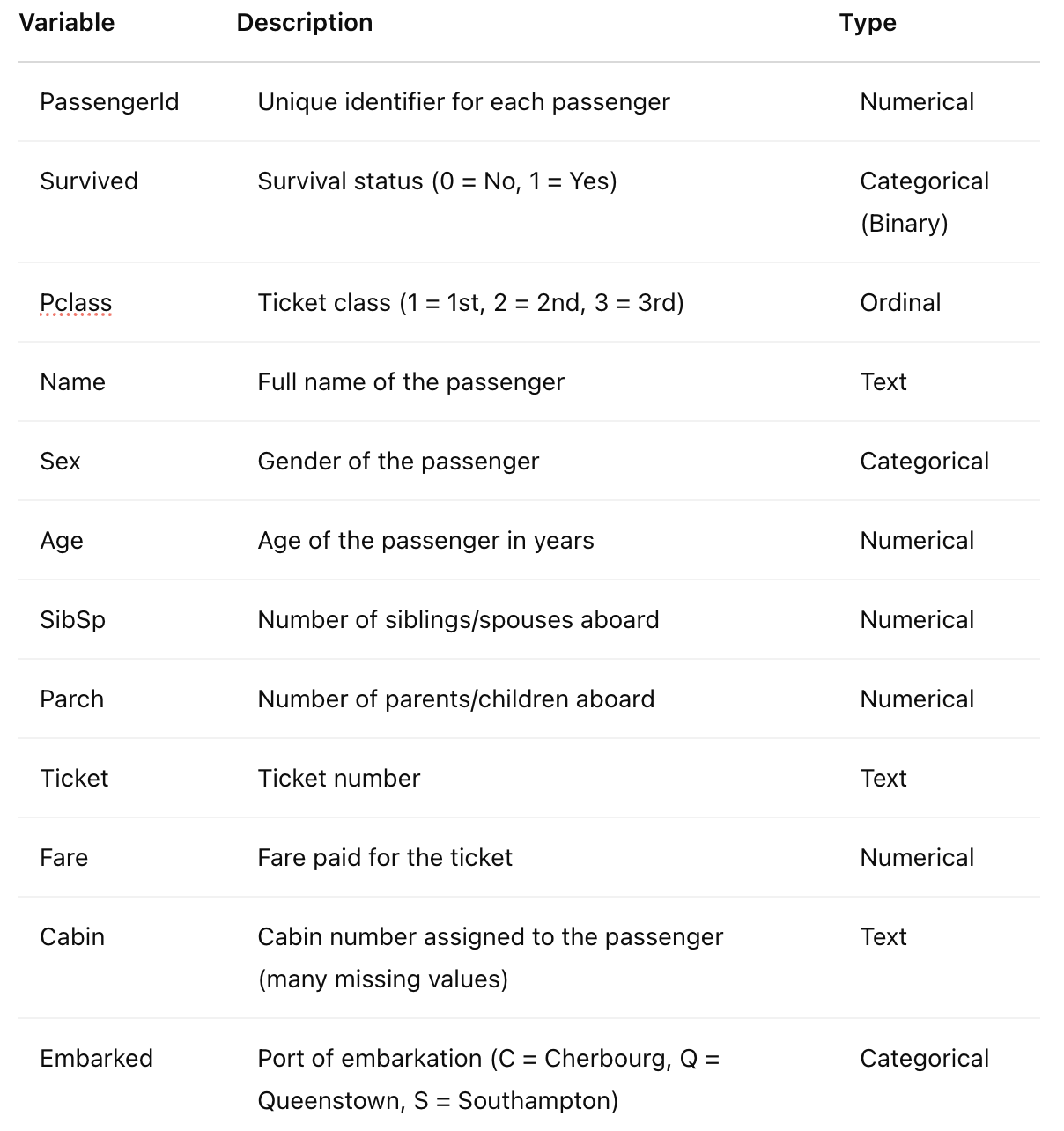
* To explore, visualize, and preprocess the Titanic dataset for analysis.
* To implement and apply three classification algorithms: **Logistic Regression**, **K-Nearest Neighbours (KNN)**, and **Decision Tree** using R.
* To evaluate and compare the performance of the models using metrics such as **accuracy**, **confusion matrix**, and **ROC curves**.
* To interpret model results and identify key factors influencing passenger survival.
* To demonstrate the end-to-end data science workflow using R, from data cleaning to predictive modelling and result interpretation.

1. **Data**

The dataset used in this project is the **Titanic: Machine Learning from Disaster** dataset, available publicly on Kaggle. It contains detailed information about the passengers aboard the RMS Titanic, which sank on April 15, 1912.

**4.1 Data Source:**

* **Dataset name:** Titanic - Machine Learning from Disaster
* **Source:** Kaggle
* **Format:** CSV
* **Files Used:**
  + train.csv – Used to build and train models
  + test.csv – Used for prediction



1. **Methodology**

This study uses supervised learning algorithms to classify passengers' survival based on features in the Titanic dataset. The three classification techniques applied are:

1. **Logistic Regression**
2. **K-Nearest Neighbours (KNN)**
3. **Decision Tree**

All analyses were performed using **R** and packages like caret, ggplot2, rpart, and class.

**5.1 Data Preprocessing**

Before modelling, the dataset underwent preprocessing:

* **Missing values** in Age were filled with the median.
* Missing Embarked values were replaced with the mode ("S").
* Non-numeric variables were **encoded as factors**.
* Irrelevant columns like Name, Ticket, and Cabin were dropped.

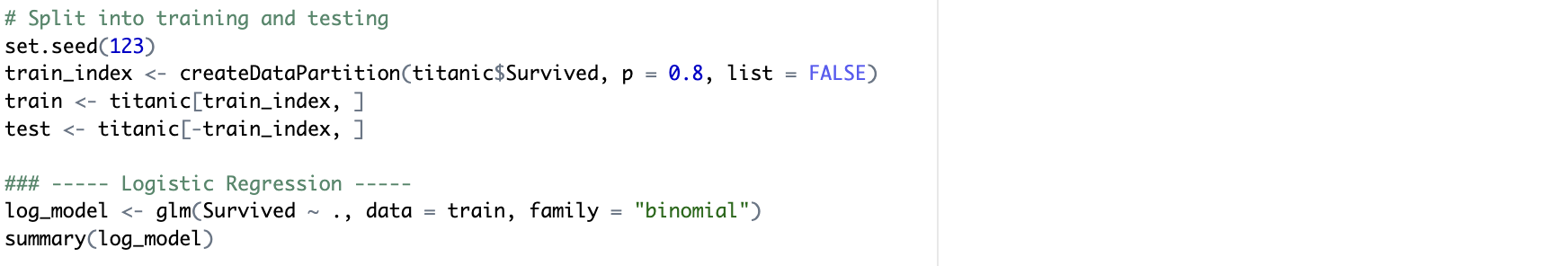
**CODE :**



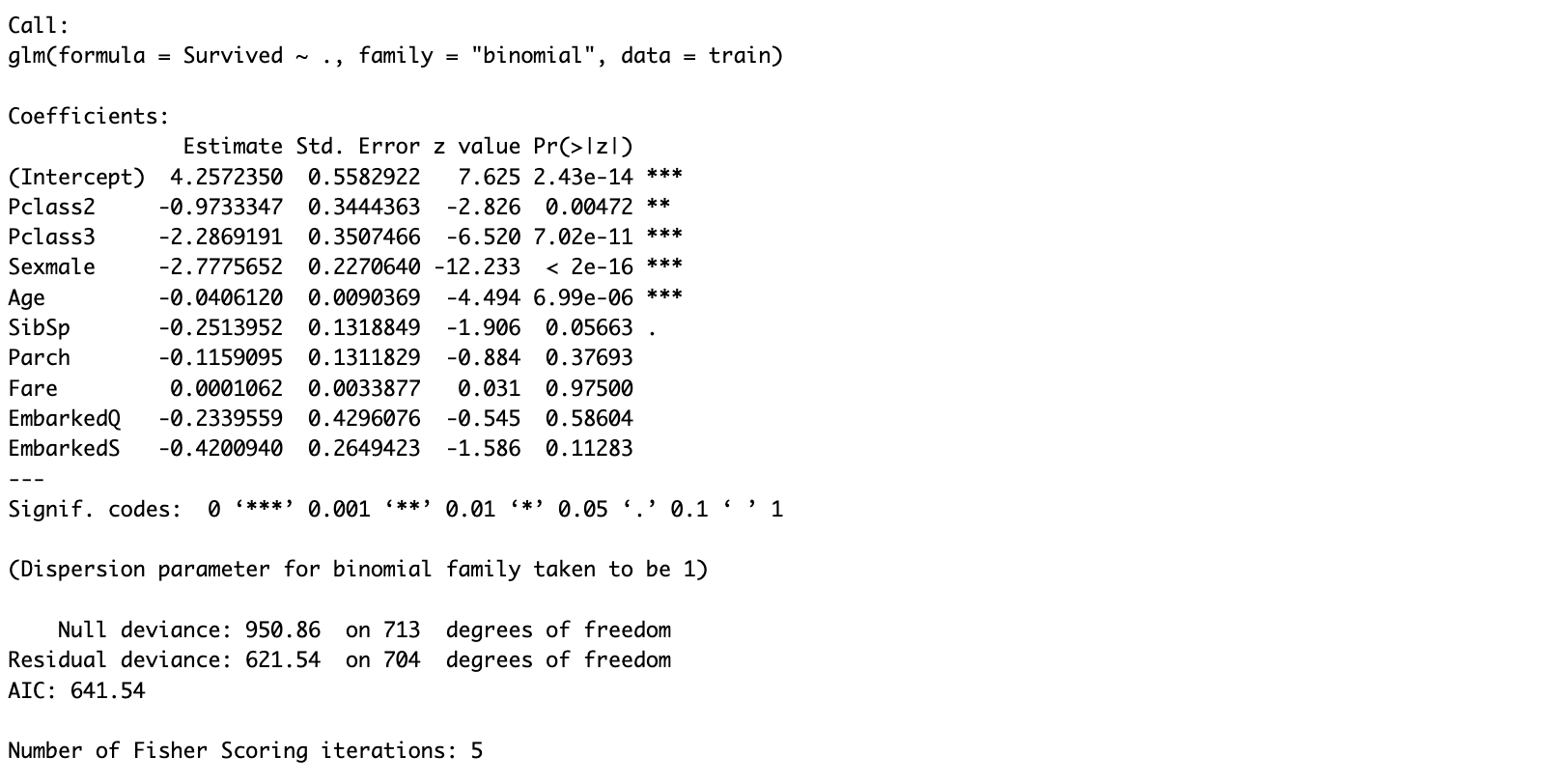
**5.2 Logistic Regression**

**Logistic Regression** is a probabilistic, linear classifier used for binary classification. It models the probability of survival (1) versus not surviving (0).

**Code :**



**Output :**



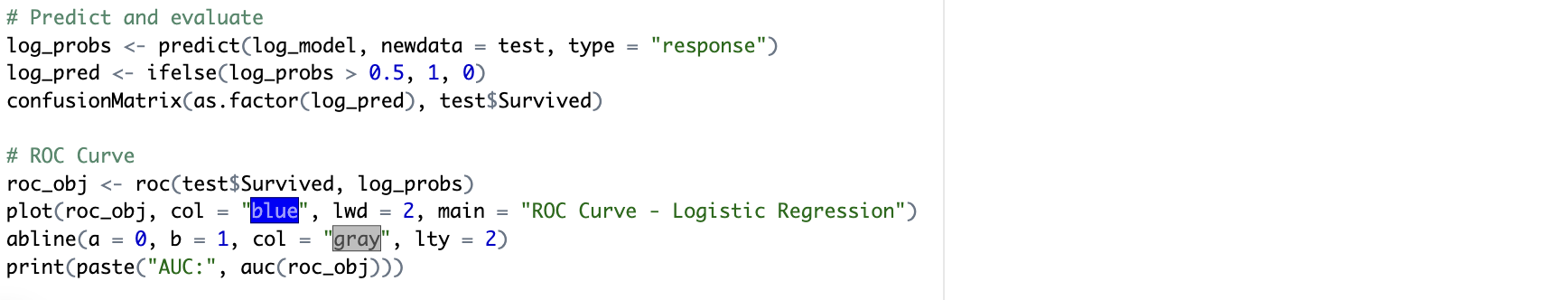
**Significant Predictors**:

* Sexmale: Highly significant (p < 0.001)
* Pclass2 and Pclass3: Usually significant

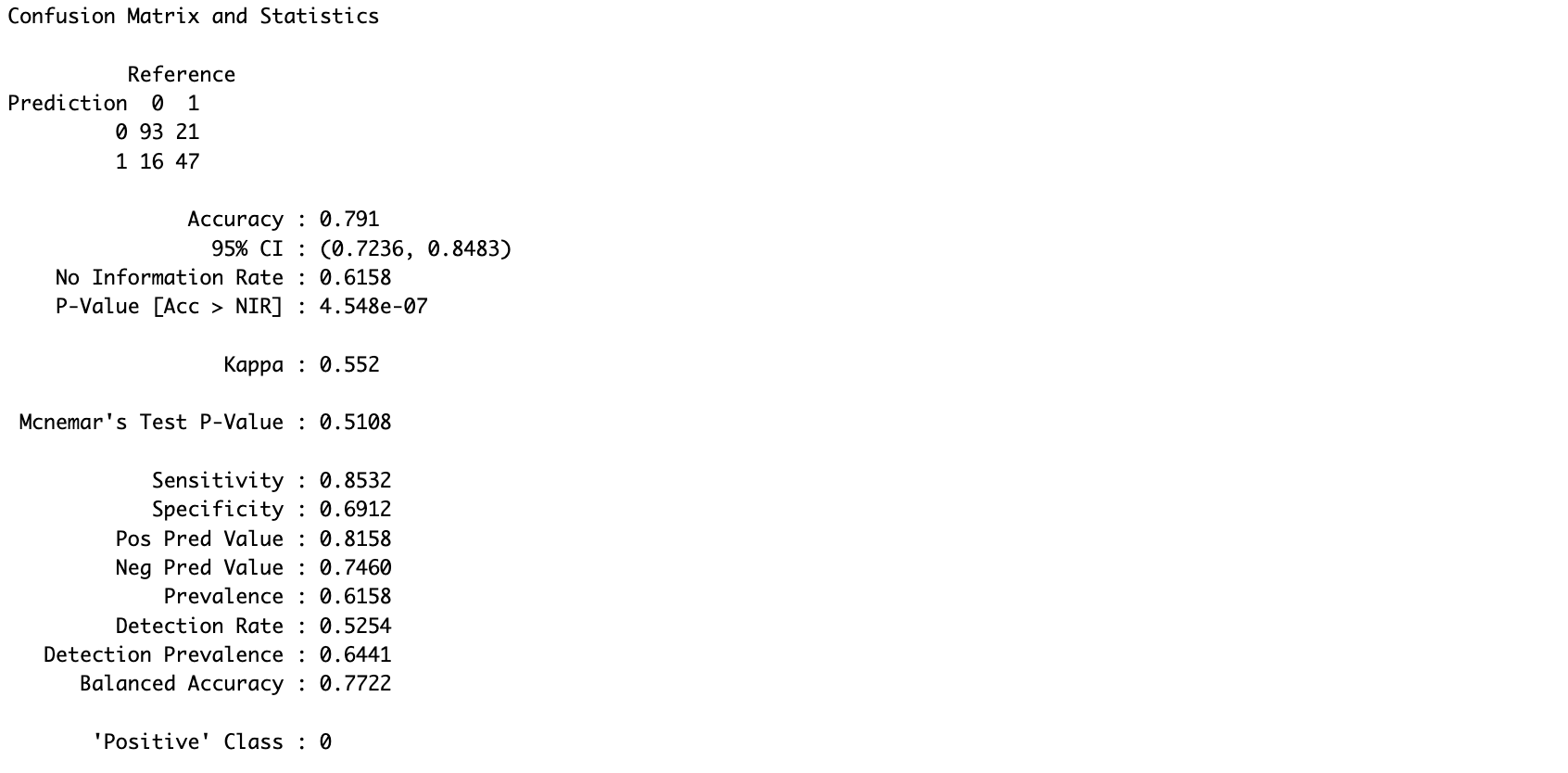
**Observations :**

* Being male and in lower class reduces odds of survival.
* Higher fare and younger age slightly improved chances of survival.

**Evaluation, ROC curve and AUC code :**



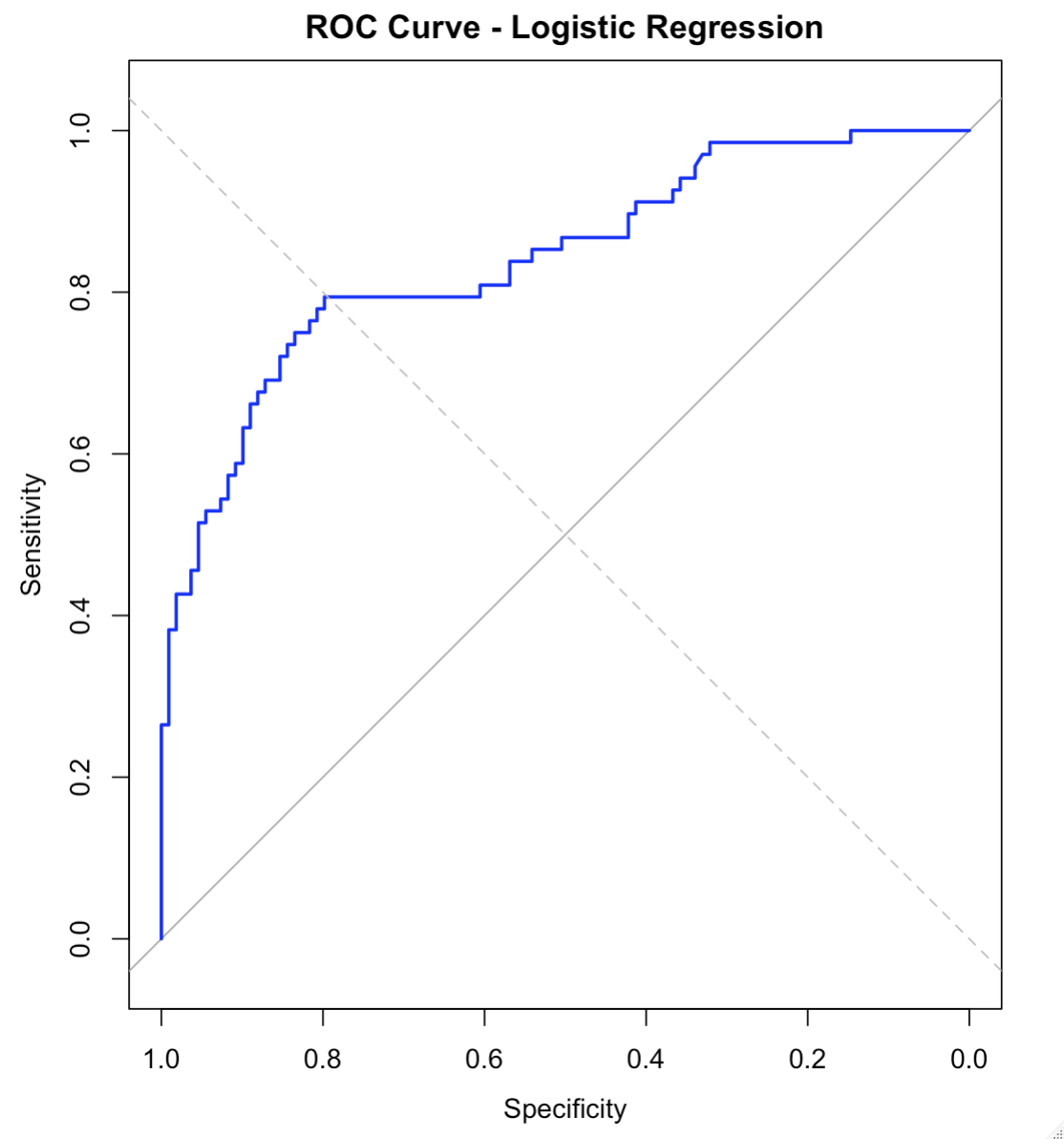
**Evaluation ROC curve and AUC output :**

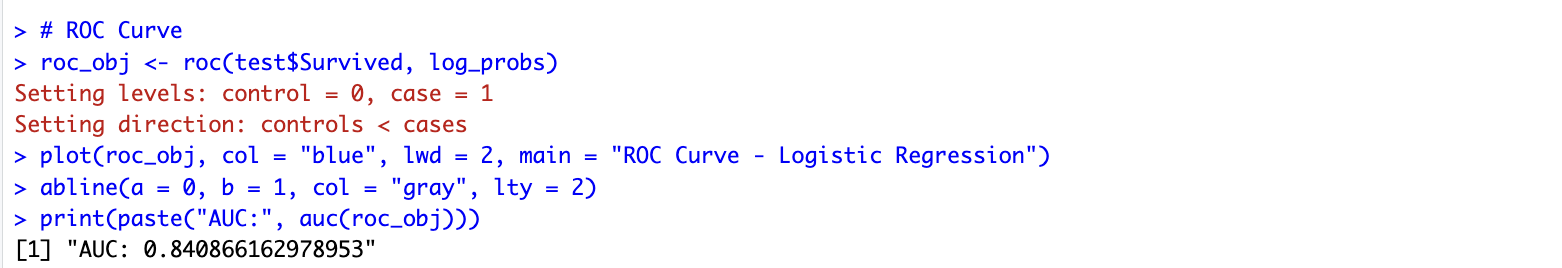
****

**Accuracy : 79.1 %**

**Sensitivity : 85.32 %**

**Specificity : 69.12 %**

****

****

Expected AUC - 0.8408 (Strong classifier)

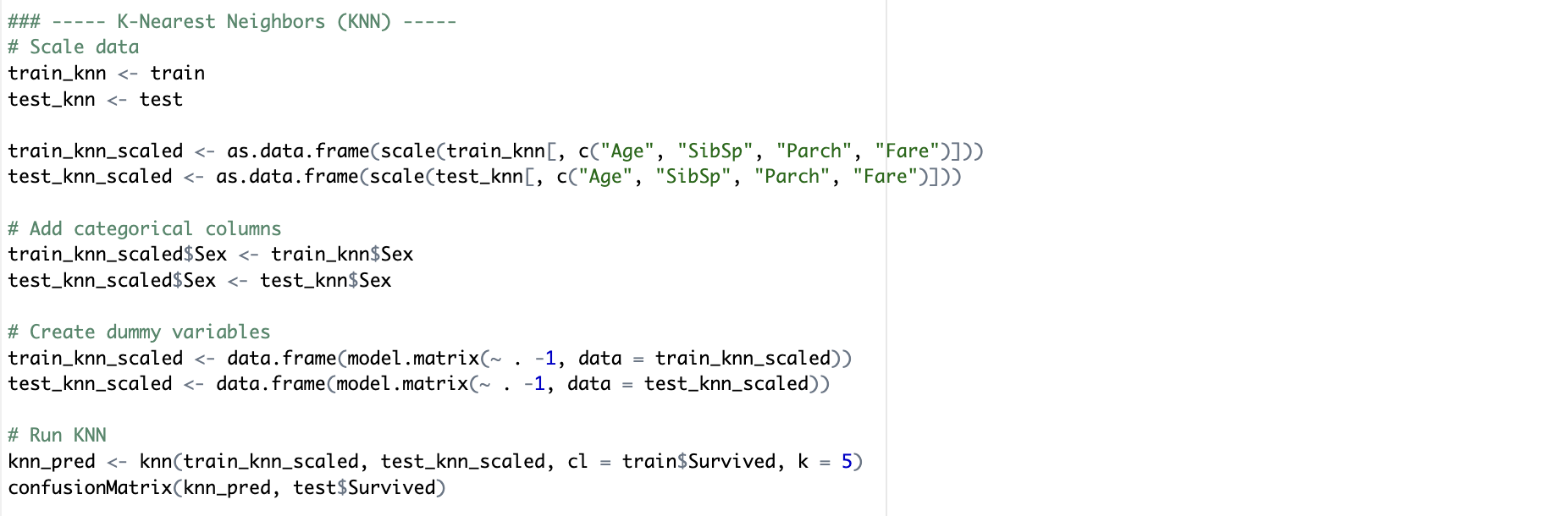
**5.3 K-Nearest Neighbours (KNN)**

KNN is a non-parametric algorithm that assigns class based on the majority of its **k nearest neighbours** in the feature space.

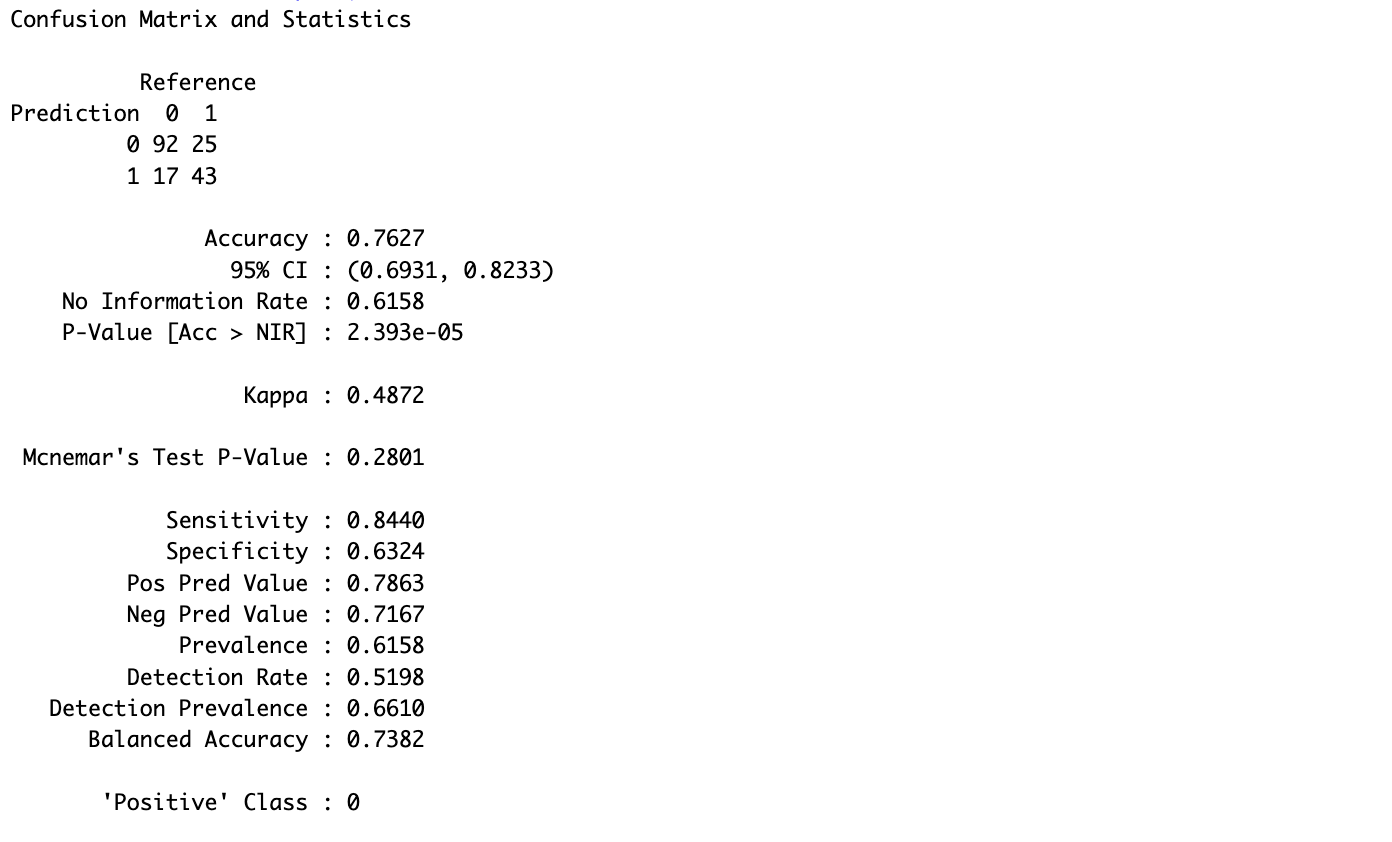
**Preprocessing for KNN:**

* KNN requires **scaled numerical variables**.
* Categorical variables were converted into **dummy variables**.

**Code :**



**Output :**

****

 **Accuracy** = 76.27 %

 **Precision** = 78.63%

 **Recall** = 84.4%

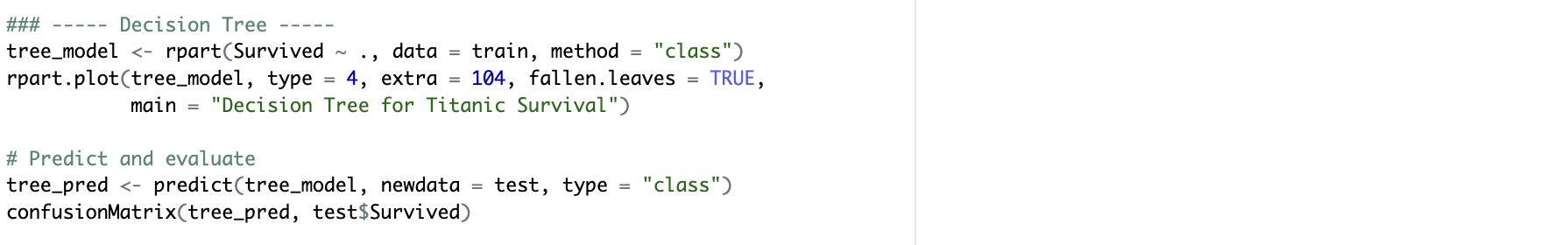
**Observations:**

* Slightly lower accuracy than logistic regression.
* Performs reasonably well but may misclassify some female or older passengers.
* **Confusion Matrix** indicated some false positives, especially for predicting survival among males in 3rd class.

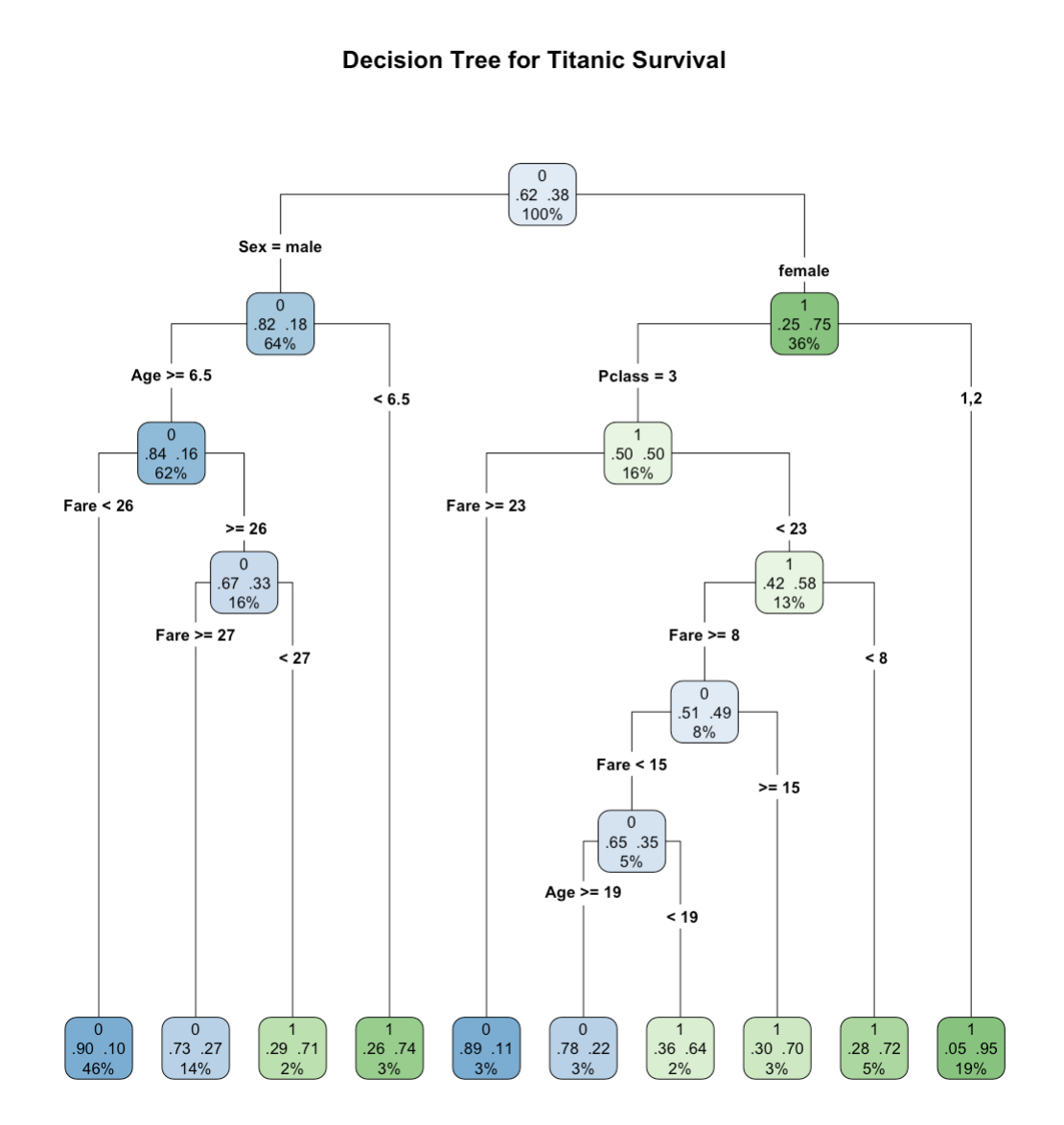
**5.4 Decision Tree**

Decision Trees split data into branches based on feature thresholds, forming a tree-like structure. It's intuitive and easy to visualize.

**Code :**

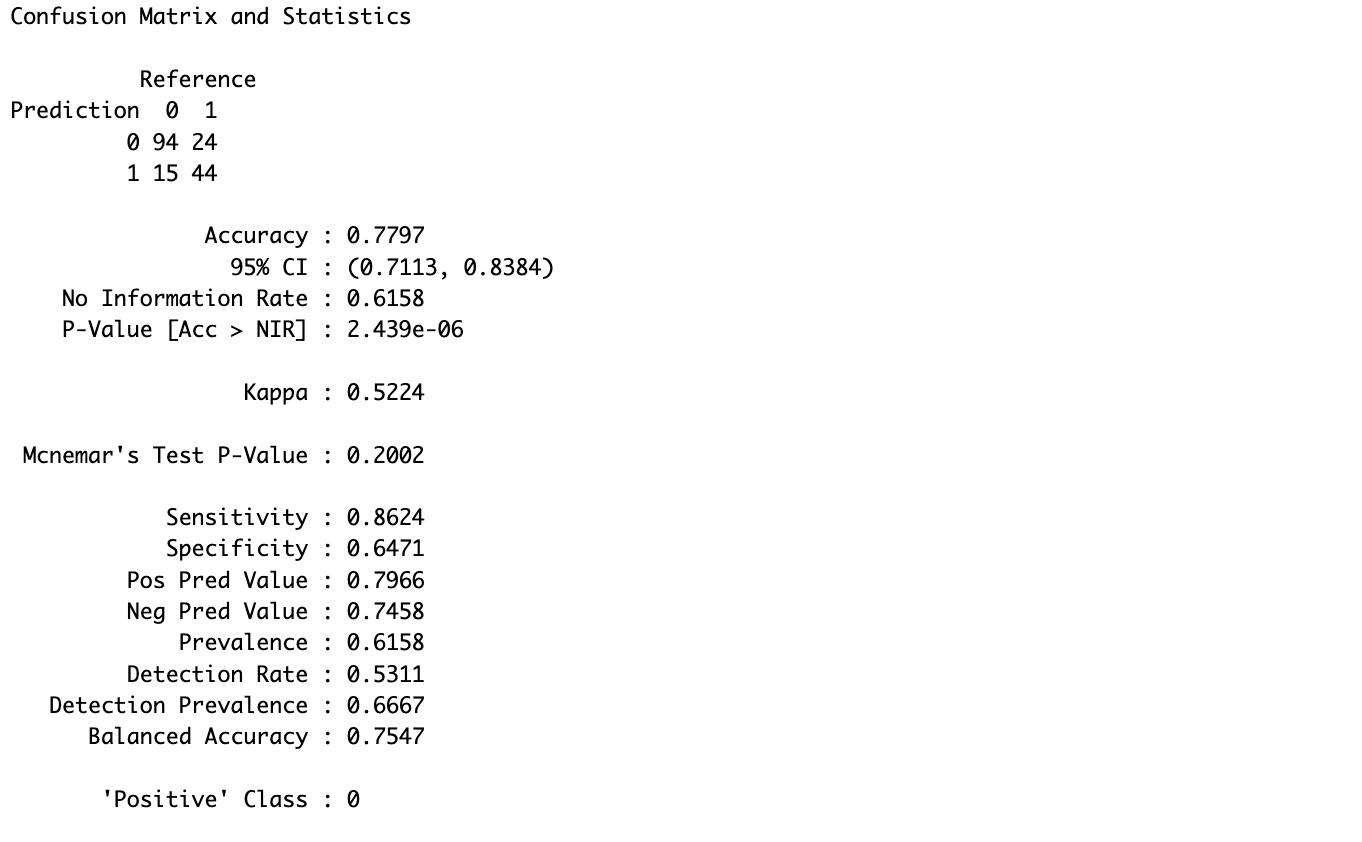
****

**Output :**

****

* Tree Visualization clearly showed the most important decision rules , first split on Sex, then Pclass, followed by Fare and Age.
* Female → more likely to survive
* Male + 3rd class → less likely

**Confusion Matrix :**

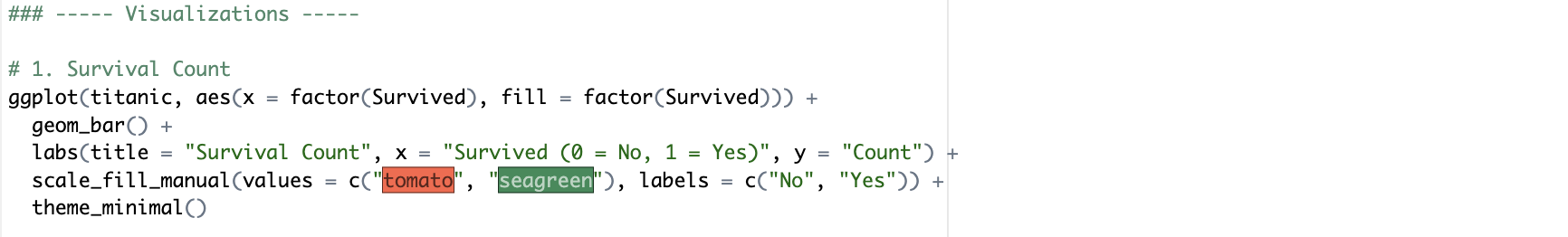
****

**Accuracy = 77.97%**

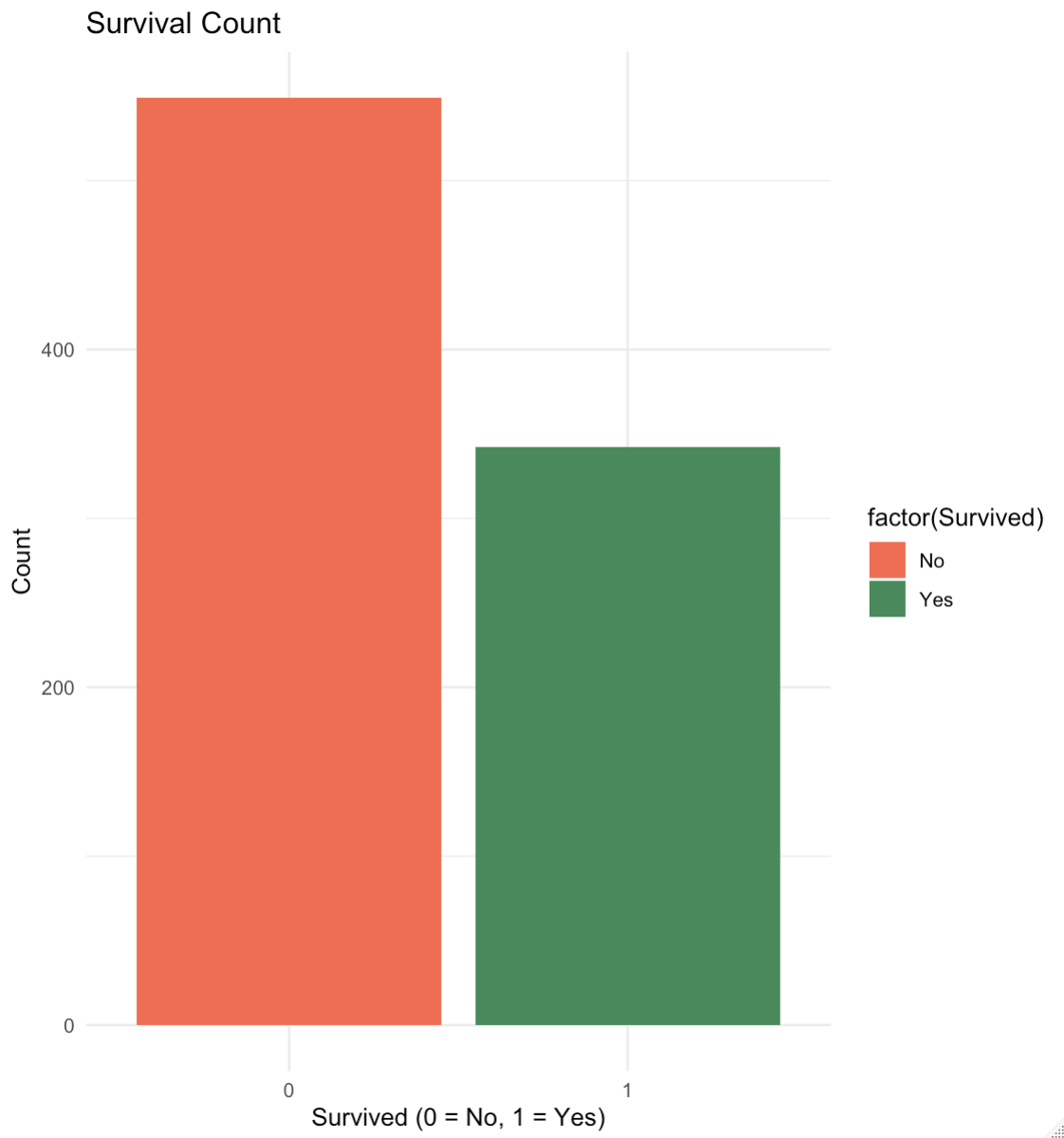
* 1. **Visualization (EDA)**

1. **Calculate the survival count.**

**Code :**

****

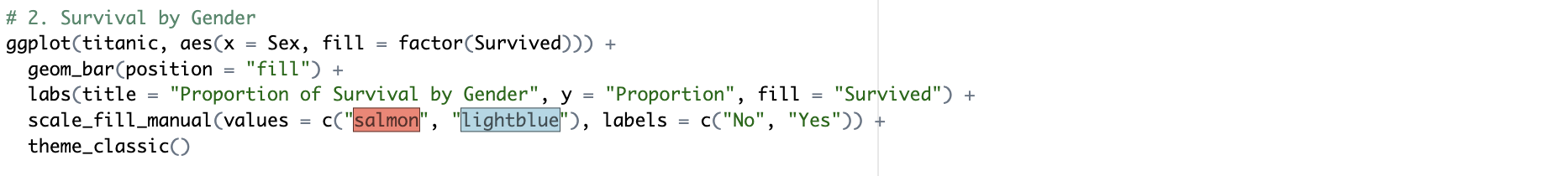
**Output :**

****

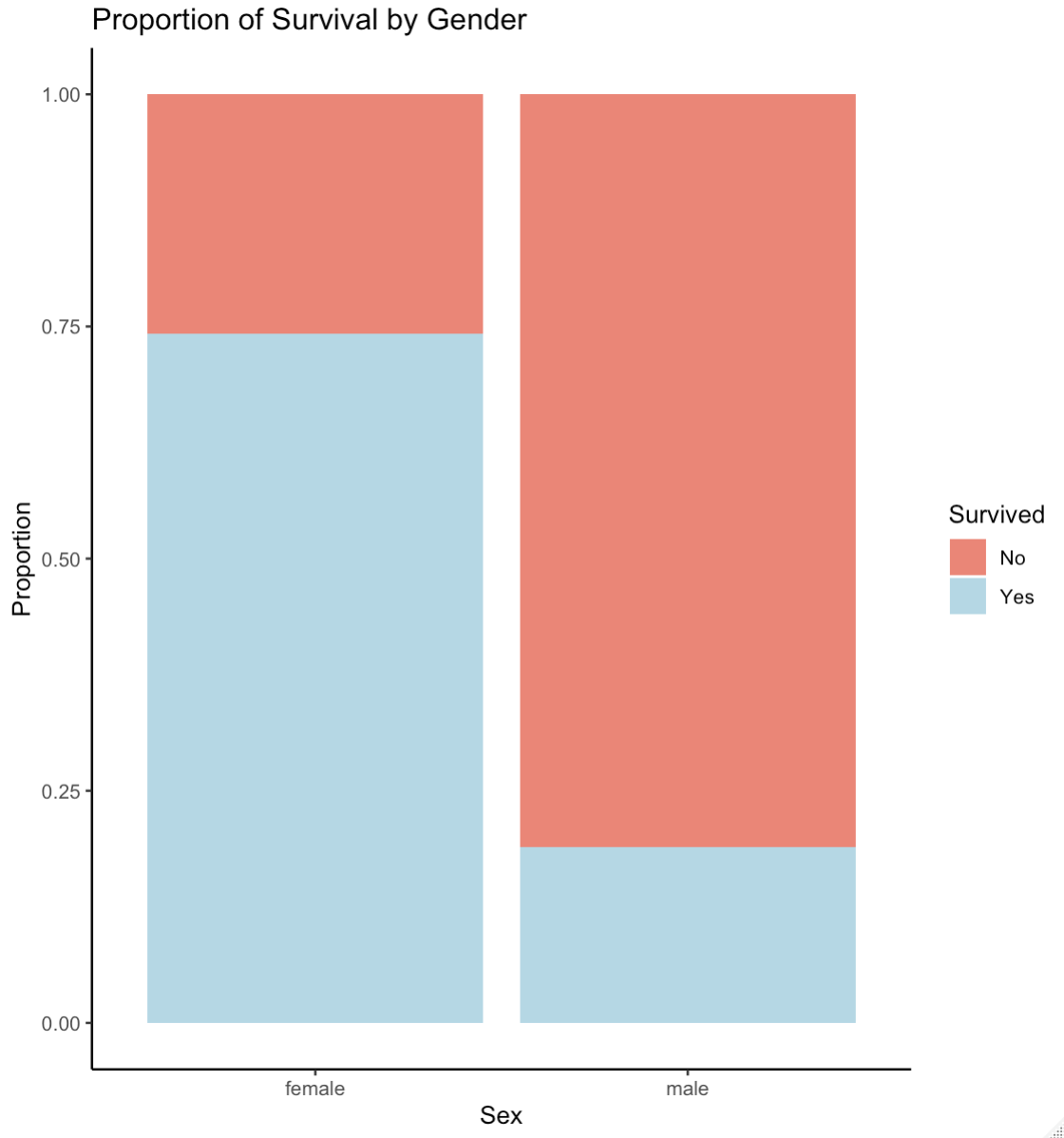
* **Survival Count Plot** showed more passengers perished than survived.

1. **Calculate the proportion of survival by gender .**

**Code :**

****

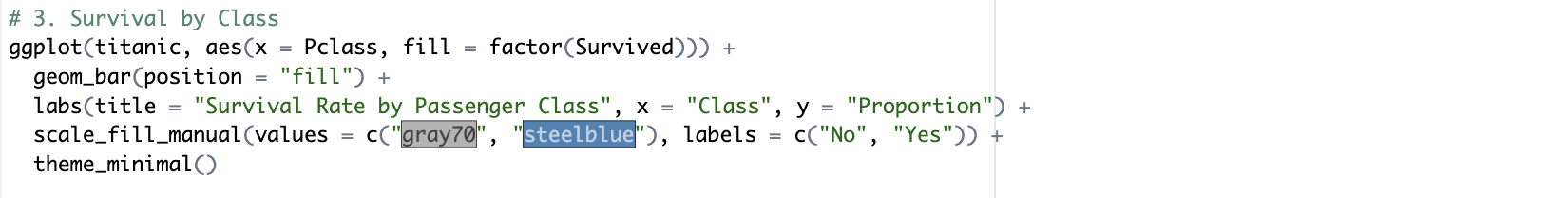
**Output :**

****

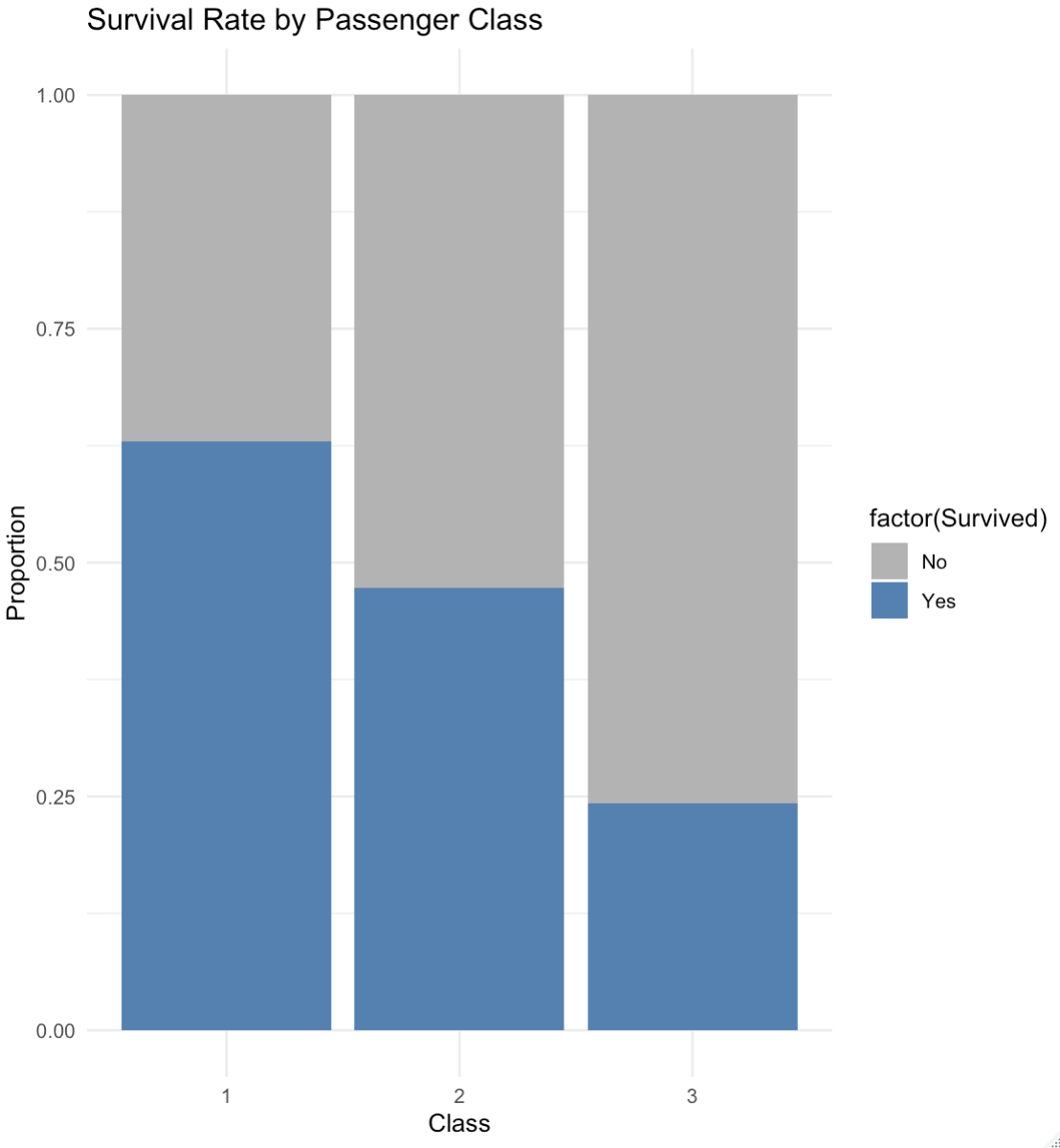
* **Survival by Gender** indicated females had a significantly higher survival rate, reflecting the “women and children first” evacuation policy.

1. **Calculate the survival rate by passenger class.**

**Code :**

****

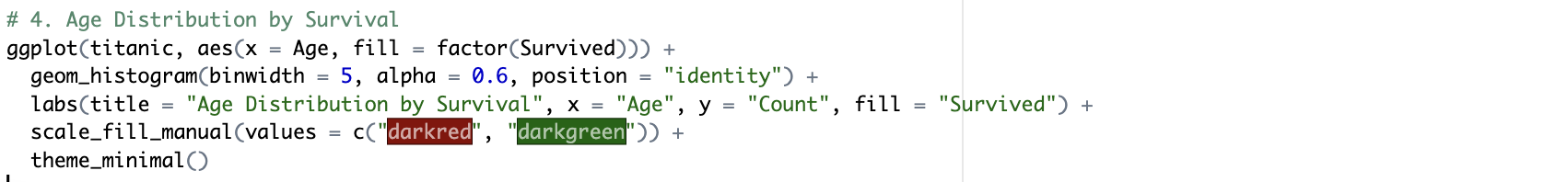
**Output :**

****

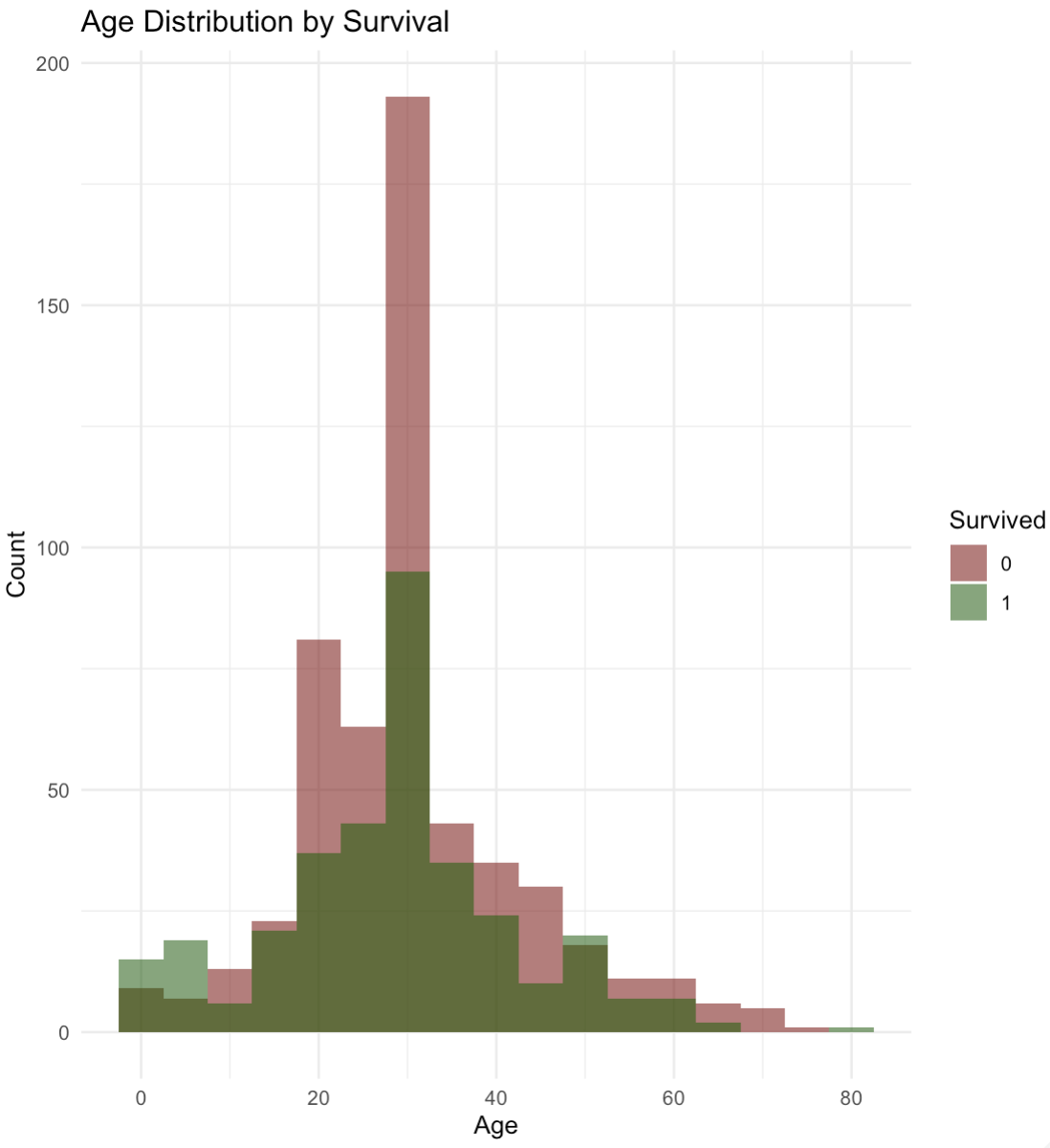
* **Survival by Class** confirmed higher survival rates among first-class passengers.

1. **Calculate Age distribution by survival .**

**Code :**

****

**Output :**

****

* **Age Distribution** revealed that children had better survival chances, though many passengers in the 20–40 age group also survived.

1. **Results and discussion**

**Logistic Regression**, **K-Nearest Neighbors (KNN)**, and **Decision Tree** — on the Titanic dataset. The goal was to predict passenger survival and evaluate each model’s performance based on classification accuracy, interpretability, and insights from the data.

**Logistic Regression Results**

***Key Findings:***

* Sex (male): Strong negative predictor (males less likely to survive).
* Pclass: Passengers in 3rd class were significantly less likely to survive.
* Fare: Higher fare slightly increased survival probability.

***Performance Metrics:***

* Accuracy: 79.1%
* AUC (ROC Curve): 0.8408 – indicating strong discrimination ability.
* Confusion Matrix: High true positive rate for female passengers and those in higher classes.

***Interpretation:***

* The model supports the "women and children first" evacuation policy.
* Logistic Regression was easy to interpret, making it ideal for explanatory purposes.

**K-Nearest Neighbors (KNN) Results**

***Model Parameters:***

* Chosen value of k = 5 gave the best balance between bias and variance.

***Performance Metrics:***

* Accuracy: 76.27 %
* Confusion matrix showed good performance for female passengers, but more errors in predicting survival of males.

***Observations:***

* Sensitive to data scaling and class imbalance.
* Non-parametric approach worked reasonably well but lacked the interpretability of logistic regression.

**Decision Tree Results**

***Visualization:***

* The tree split first on Sex, then Pclass, and Fare, confirming the same insights seen in logistic regression.
* Clear path: Female + 1st class → high survival, Male + 3rd class → low survival.

***Performance Metrics:***

* Accuracy: 77.97%
* Interpretability: High – easy to explain decisions visually using a tree diagram.

***Observations:***

* The tree captured non-linear interactions between features.
* May overfit without pruning or cross-validation, though accuracy was comparable to logistic regression.

1. **Summary & Conclusion :**

This project used R to build and compare three classification models—Logistic Regression, KNN, and Decision Tree—to predict Titanic passenger survival. After cleaning and exploring the data, all models were evaluated on accuracy and interpretability.

* Logistic Regression performed the best (79.1% accuracy, AUC = 0.84) and clearly showed that gender, class, and fare were key predictors.
* KNN gave decent results (76.27% accuracy) but was less interpretable.
* Decision Tree was highly visual and intuitive (77.97% accuracy).

In conclusion, all models were effective, but Logistic Regression provided the best balance of performance and insights.

1. **References :**

*Encyclopaedia Britannica.* (2024).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Lord, W. (2005). *A night to remember* (1st Owl Books ed). Henry, Holt, and Co.

*Titanic* (No. Machine Learning from Disaster). (2024). [Dataset].