Syamsul Rizal Fany
saemfany@gmail.com

PREDICTING SURVIVAL

Titanic Dataset Analysis

Visit My LinkedIn
Visit My GitHub

About Me

My name is Syamsul Rizal Fany. I am a Mathematics graduate from the Faculty of Mathematics and Natural Sciences at Universitas Riau with a passion for technology, particularly in the field of data. Over the past year, I have immersed myself in the world of data, completing various online courses to enhance my skills in data analysis and visualization.

I am fascinated by how data can drive decision-making, uncover insights, and solve real-world problems. Currently, I am focusing on building my expertise in data science, learning tools like SQL, Python, and machine learning techniques.

My goal is to become a Data Scientist, leveraging my strong analytical background and my enthusiasm for continuous learning to contribute to impactful projects. I am always eager to collaborate and explore opportunities that challenge my skills and help me grow in the field of data.

Introduction

The Titanic disaster of 1912 remains one of the most tragic and widely studied maritime accidents in history. In this analysis, we delve into the Titanic dataset to understand the factors that influenced passenger survival rates and to build predictive models that can identify the likelihood of survival based on passenger characteristics.

Our goal is to analyze this historical data using machine learning techniques to uncover patterns and insights. By exploring features like passenger class, age, gender, family size, and fare, we aim to determine which factors played a significant role in survival chances.

Through this project, we compare two models—Decision Tree and Random Forest—to evaluate their effectiveness in predicting survival. This analysis not only sheds light on data-driven decision-making but also demonstrates the potential of machine learning in classification tasks.

Dataset Overview
Dataset Description

The Titanic dataset contains information about the passengers aboard the RMS Titanic, focusing on factors that may have influenced their chances of survival. This dataset includes 891 records, with the following key features:

- **PassengerId**: Unique identifier for each passenger
- **Survived**: Survival status (0 = No, 1 = Yes) - target variable for prediction
- **Pclass**: Passenger class (1 = 1st, 2 = 2nd, 3 = 3rd), a proxy for socio-economic status
- **Name**: Name of the passenger
- **Sex**: Gender of the passenger
- **Age**: Age of the passenger
- **SibSp**: Number of siblings/spouses aboard the Titanic
- **Parch**: Number of parents/children aboard the Titanic
- **Ticket**: Ticket number
- **Fare**: Fare paid for the ticket
- **Cabin**: Cabin number (often missing)
- **Embarked**: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

This dataset offers insights into demographic and socio-economic factors that may have influenced passenger survival during the tragedy, enabling predictive modeling and statistical analysis.

Dataset Overview
Key Features

1. **Pclass (Passenger Class)**: A proxy for socio-economic status, with three categories:
   - 1st class (Upper)
   - 2nd class (Middle)
   - 3rd class (Lower)
2. **Sex**: Gender of the passenger (Male/Female). Often a critical factor in survival analysis.
3. **Age**: Age of the passenger. Survival rates can vary significantly based on age groups.
4. **SibSp** (Siblings/Spouses Aboard): The number of siblings or spouses each passenger had on board.
5. **Parch** (Parents/Children Aboard): The number of parents or children each passenger had on board.
6. **Fare**: The fare paid for the ticket, providing insight into socio-economic status.
7. **Embarked**: Port of embarkation, which could affect survival due to cabin location and other factors:
   - C = Cherbourg
   - Q = Queenstown
   - S = Southampton

Dataset Overview

Objective

The objective of this analysis is to build a predictive model that estimates the likelihood of survival for passengers aboard the Titanic, based on demographic, socio-economic, and ticket-related features.

Key goals include:

- **Exploratory Data Analysis (EDA)**: Understanding patterns and relationships within the dataset, such as survival rates based on gender, age, and passenger class.
- **Feature Engineering**: Identifying and creating new features that may improve the model's predictive power.
- **Modeling and Evaluation**: Using machine learning models (such as Decision Tree and Random Forest) to predict survival, and evaluating the models based on accuracy, precision, recall, and F1-score.
- **Insights and Interpretation**: Drawing meaningful conclusions about the factors most strongly associated with survival, based on model performance and feature importance.

Data Preprocessing
Handle Missing & Duplicated Data

```
df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Age: Use median to fill missing values.

```
df.fillna({'Age': df['Age'].median()}, inplace=True)
```

Embarked: Use mode to fill missing values.

```
df.fillna({'Embarked': df['Embarked'].mode()[0]}, inplace=True)
```

Cabin: Since most values are missing, you can drop this column.

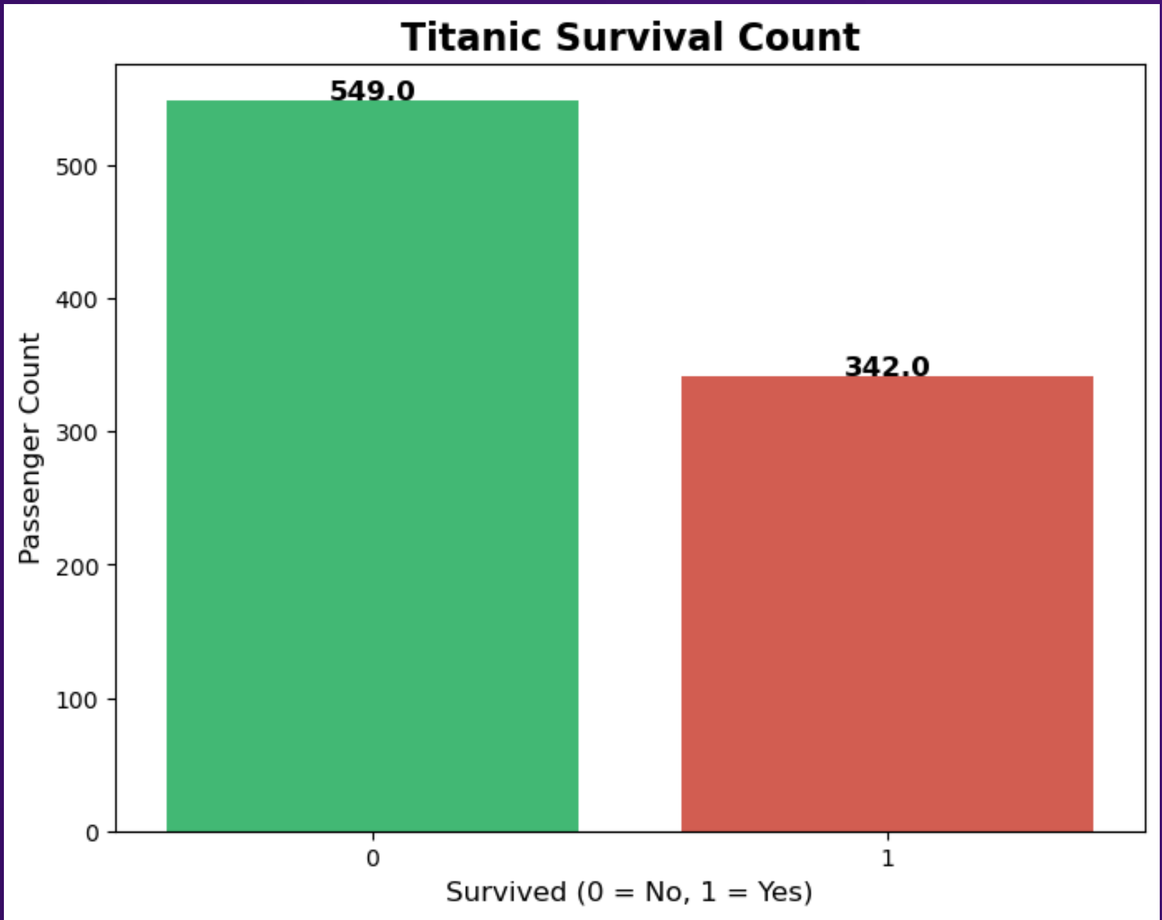
```
df.drop('Cabin', axis=1, inplace=True)
```

Handle Duplicated Data

```
df.duplicated().sum()
```

```
0
```

Exploratory Data Analysis
VIsualization



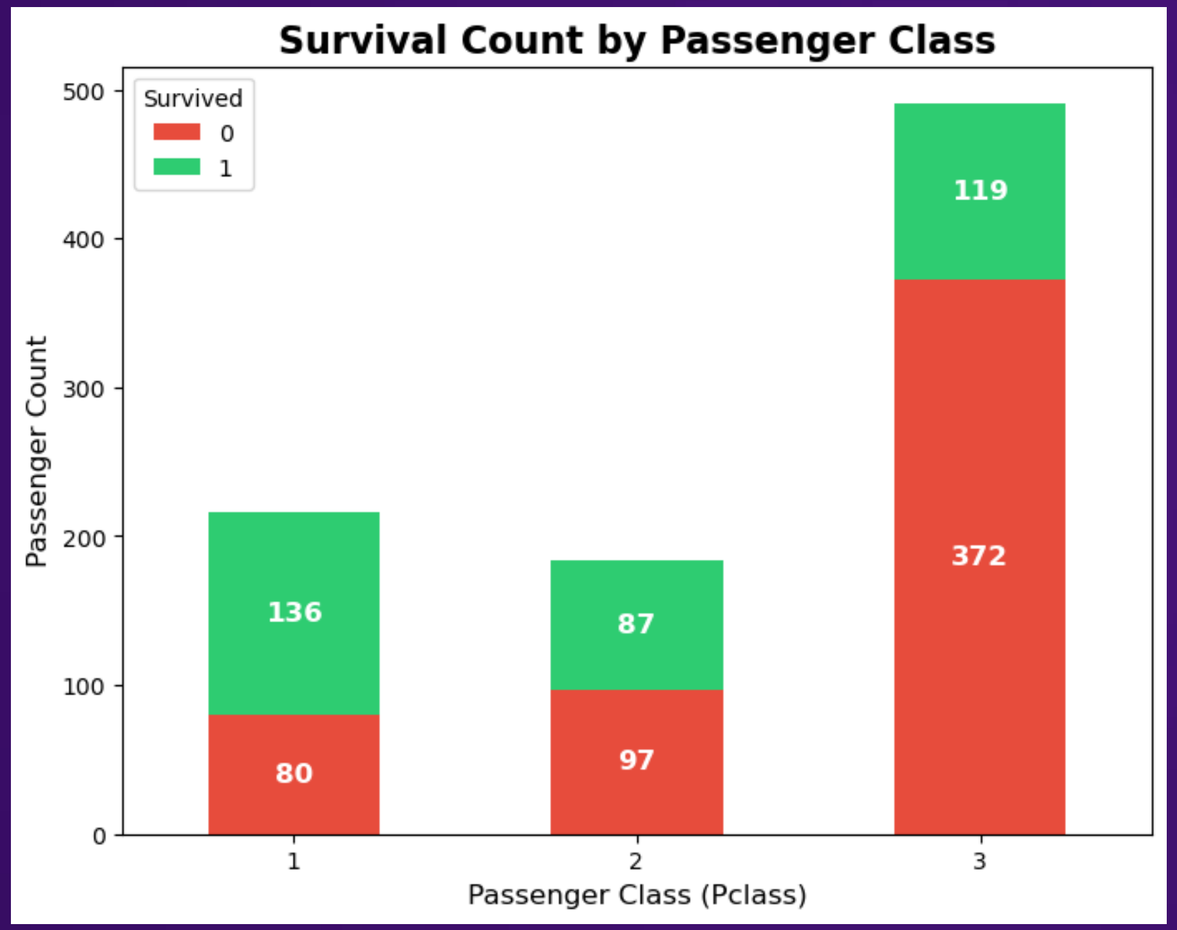
To determine if this is balanced or imbalanced, we can calculate the proportion of each class:

- Proportion of passengers who did not survive: 61.6%
- Proportion of passengers who survived: 38.4%

A balanced dataset typically has an approximately equal number of instances in each class (e.g., close to 50/50).

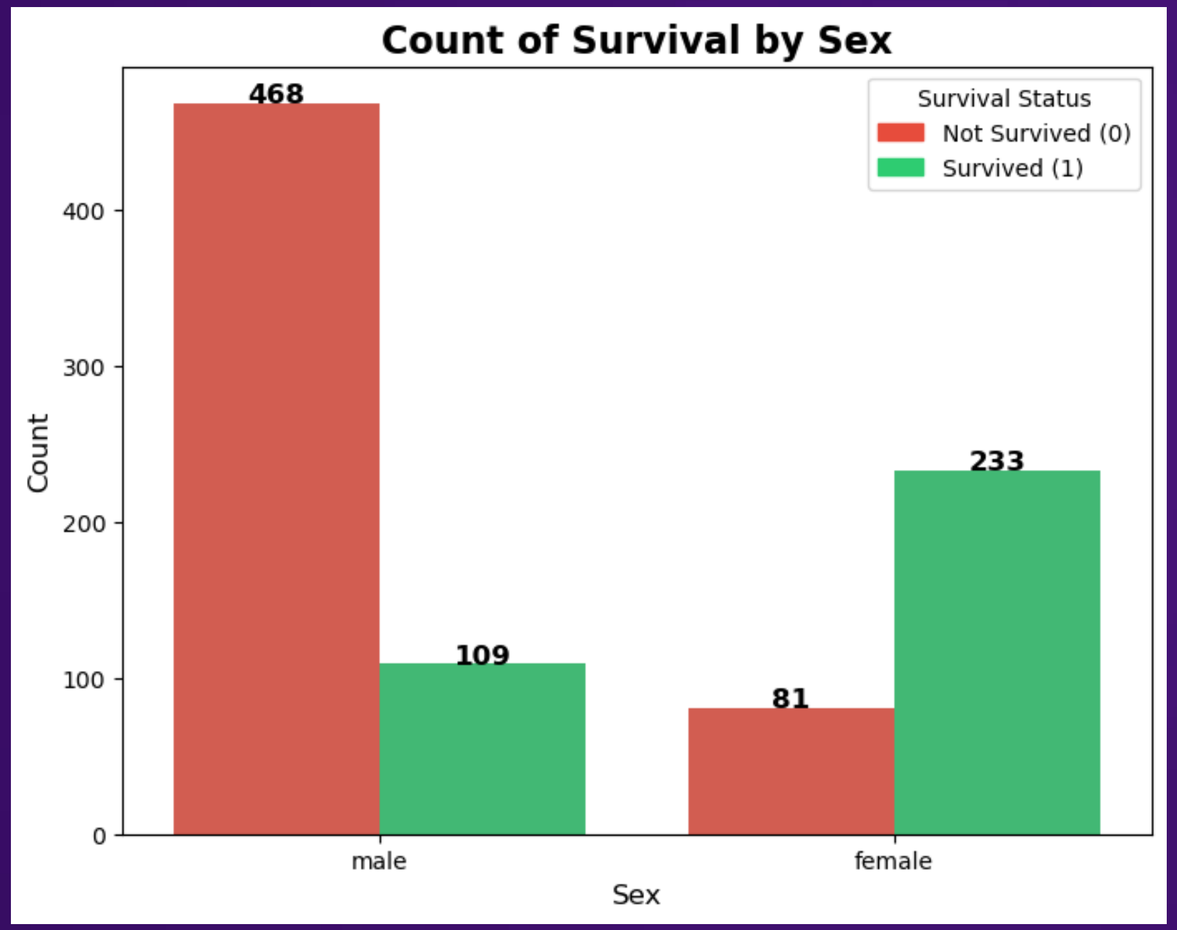
Since there are significantly more passengers who did not survive (61.6%) than those who did survive (38.4%), this dataset is slightly imbalanced. It's not a severe imbalance, but it still could affect the performance of some machine learning models, especially if you're using classifiers that are sensitive to class imbalance (e.g., decision trees or logistic regression). Techniques like resampling, class weighting, or using specialized algorithms can help if needed.

Exploratory Data Analysis
VIsualization



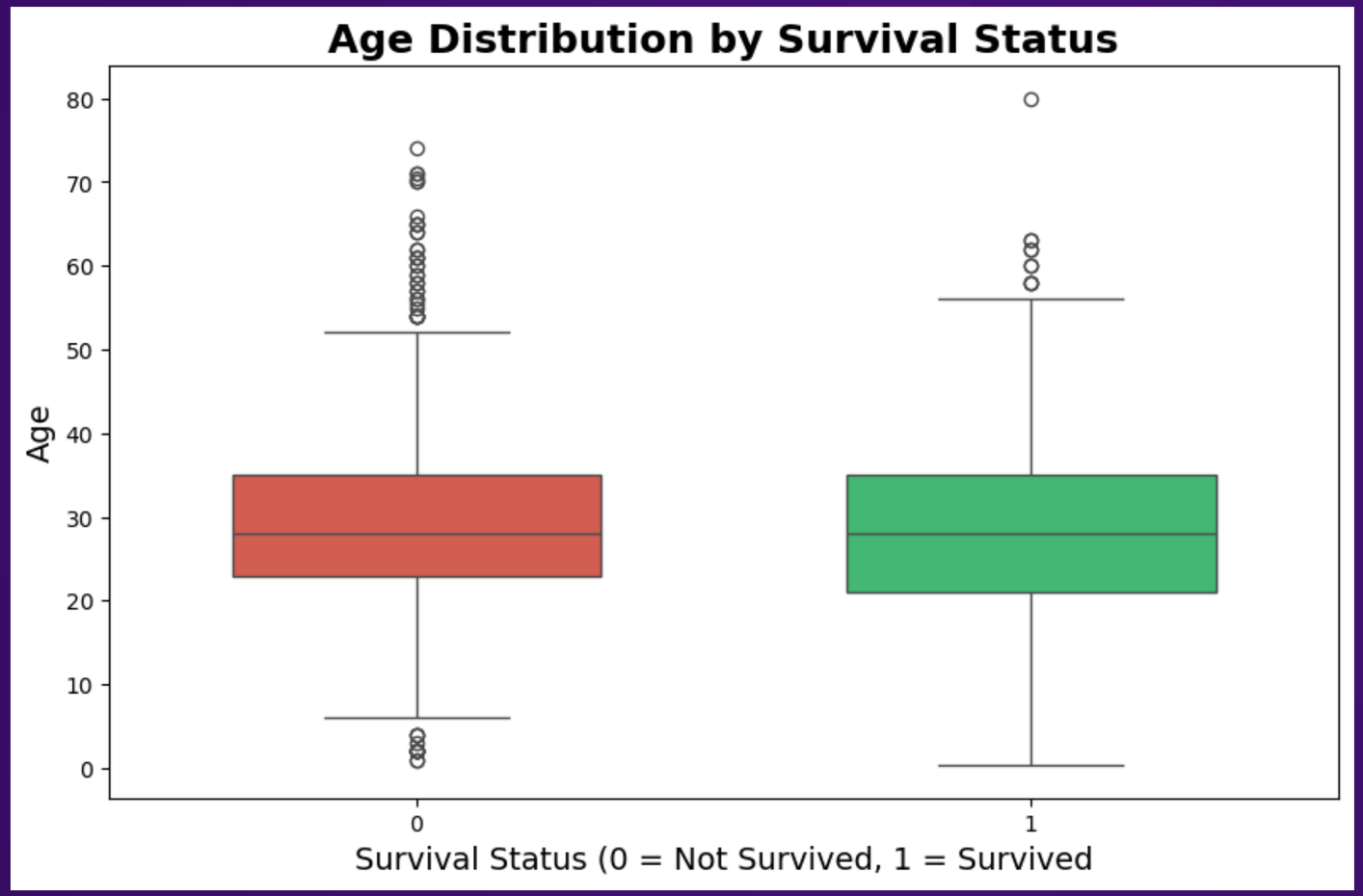
The analysis of Pclass and survival rates from the Titanic dataset clearly indicates that socio-economic status had a significant impact on survival. Passengers in higher classes had a much higher likelihood of surviving, reflecting broader societal inequalities that influenced the outcomes during this tragic event.

Exploratory Data Analysis
VIsualization



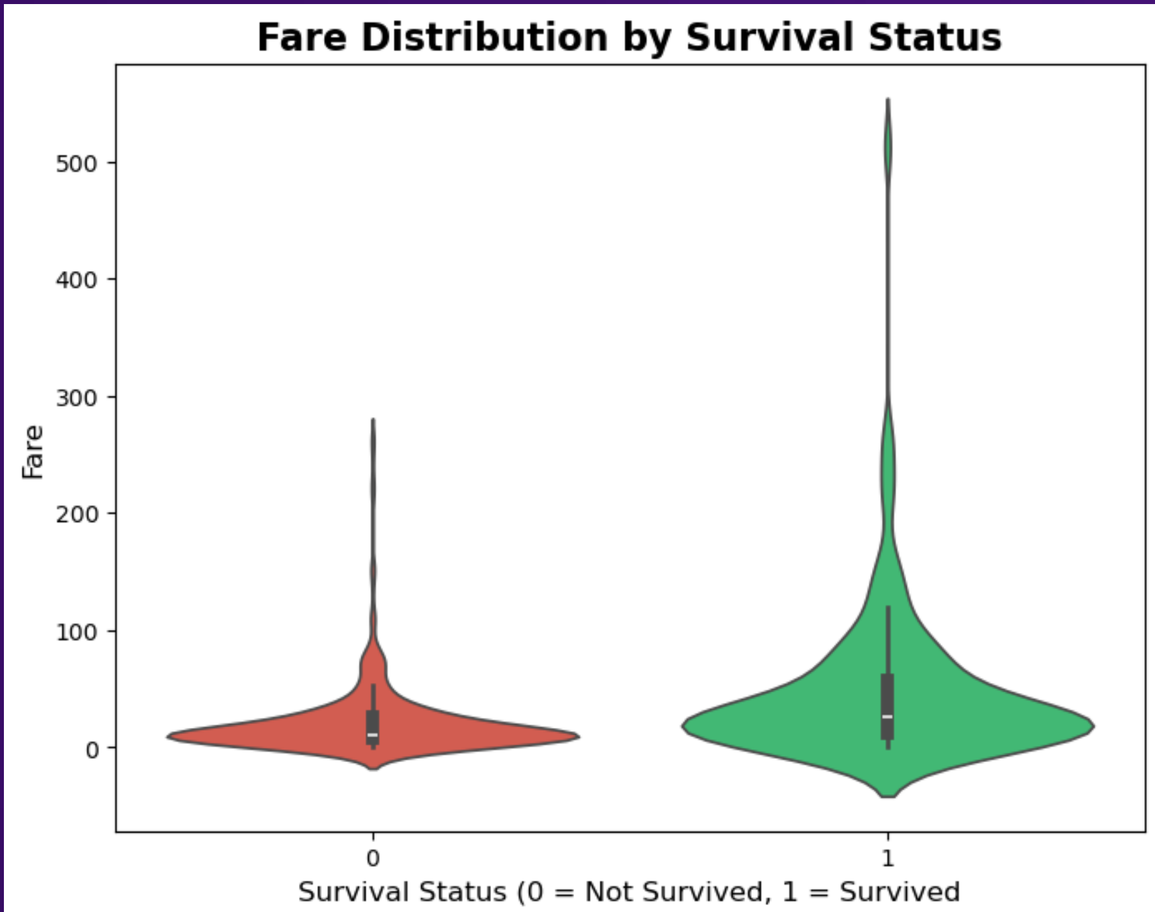
- Gender was a strong indicator of survival. Women were prioritized for lifeboats, which is a plausible explanation for the much higher survival rate among females compared to males.
- The significant gap between male and female survival rates suggests that the “women and children first” protocol was in effect during the Titanic disaster.

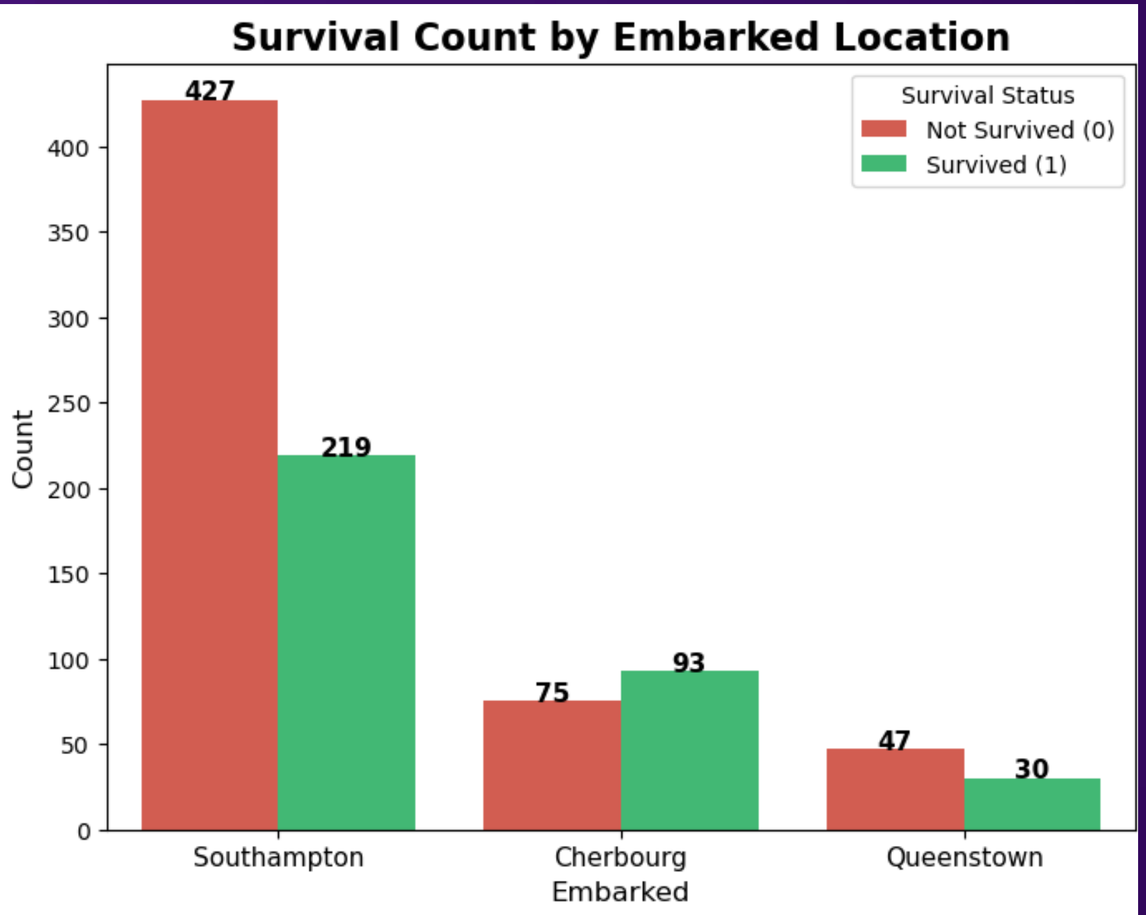
Exploratory Data Analysis
VIsualization



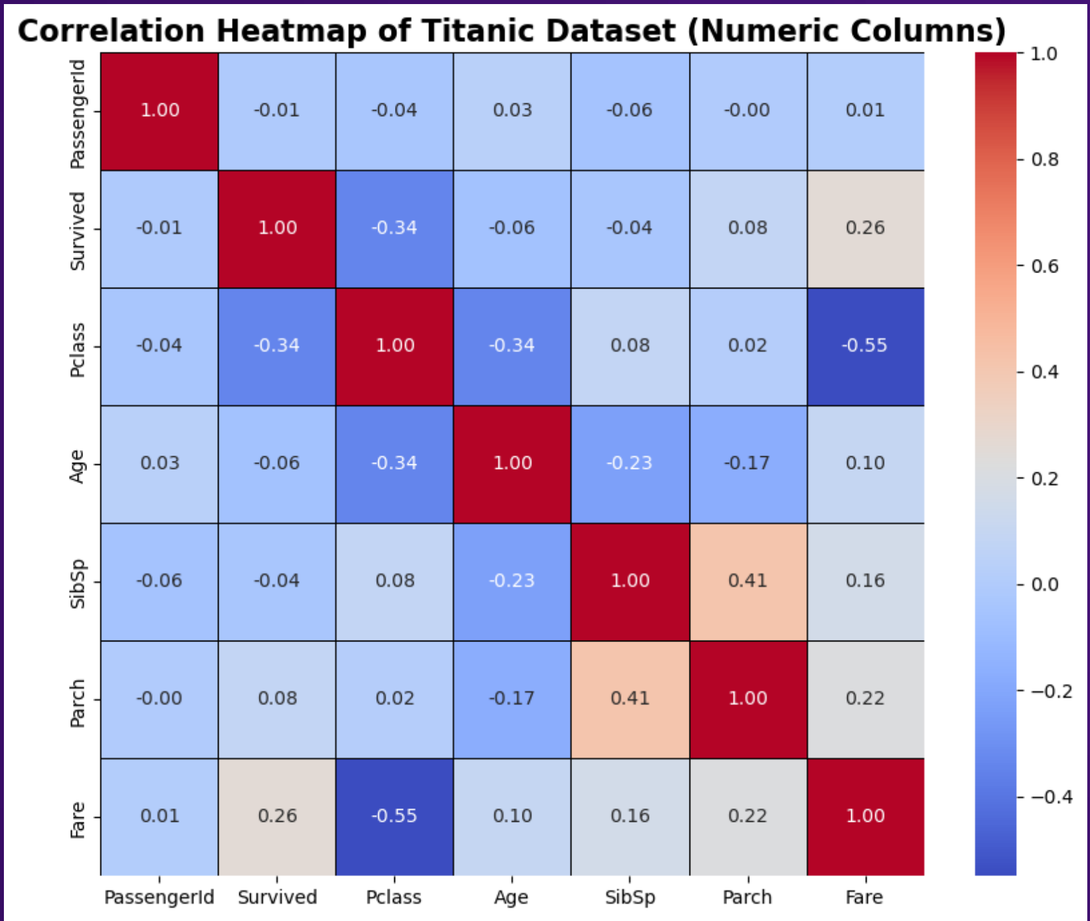
- While both groups show a similar spread of ages, there is a slight tendency for younger passengers to survive, as indicated by the lower median age of survivors.
- Age alone might not be a strong factor, but it could still play a role in survival, especially for outliers (such as very young or older passengers).

Exploratory Data Analysis
VIsualization





Exploratory Data Analysis
VIsualization



Let's say the correlation heatmap looks at Survived, Age, Fare, and Pclass:

- **Survived vs Pclass**: If you see a correlation like **–0.34**, this indicates a **negative relationship**. It suggests that higher-class passengers (lower Pclass value) had a better chance of survival.
- **Survived vs Fare**: A positive correlation, such as **0.26**, implies that passengers who paid a higher fare were more likely to survive (though not a strong correlation, it's still notable).
- **Age vs Survived**: A correlation of **–0.08** suggests that age doesn't have a strong linear correlation with survival, but there is a slight trend where younger passengers might have had better survival chances.

Feature Engineering

Create new features, such as FamilySize by combining SibSp and Parch.

```
df['FamilySize'] = df['SibSp'] + df['Parch']
df.drop(['SibSp', 'Parch'], axis=1, inplace=True)
```

Convert categorical columns like Sex and Embarked to numeric.

```
df = pd.get_dummies(df, columns=['Sex', 'Embarked'], drop_first=True, dtype=int)
df.head()
```

	PassengerId	Survived	Pclass	Name	Age	Ticket	Fare	FamilySize	Sex_male	Embarked_Q	Embarked_S
0	1	0	3	Braund, Mr. Owen Harris	22.0	A/5 21171	7.2500	1	1	0	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	38.0	PC 17599	71.2833	1	0	0	0
2	3	1	3	Heikkinen, Miss. Laina	26.0	STON/O2. 3101282	7.9250	0	0	0	1
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	35.0	113803	53.1000	1	0	0	1
4	5	0	3	Allen, Mr. William Henry	35.0	373450	8.0500	0	1	0	1

Model Selection

1. **Choosing Models for Binary Classification**:
   - Since the Titanic dataset's goal is to predict **survival** (yes or no), this is a binary classification problem.
   - Several algorithms are available for binary classification, including logistic regression, decision trees, random forests, and more.

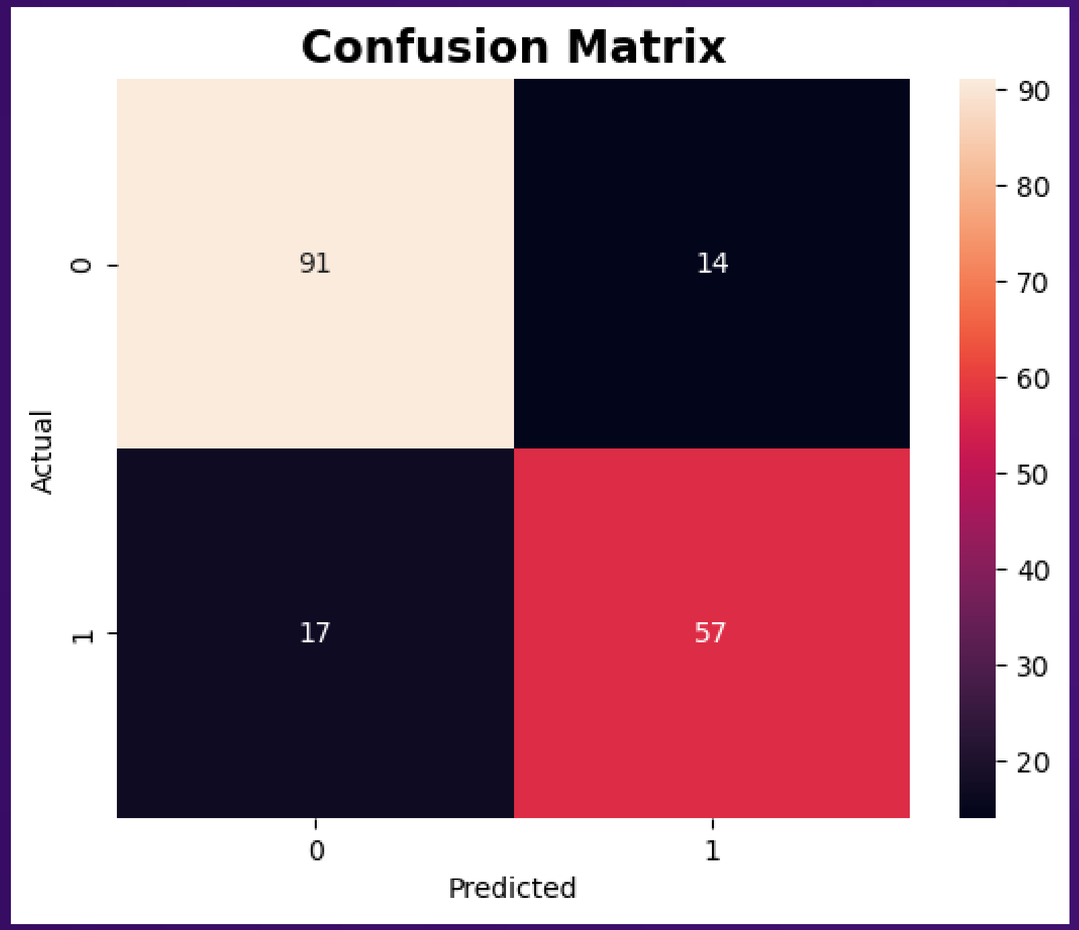
2. **Decision Tree Classifier**:
   - **Interpretability**: Decision Trees are highly interpretable, allowing us to visually trace decisions and understand feature splits, making it easier to explain why a passenger was classified as likely to survive or not.
   - **Feature Importance**: Decision Trees naturally rank features by importance, which is valuable for understanding which factors impact survival most.

Model Selection

3. **Random Forest Classifier**:
   - **Improved Accuracy**: Random Forest combines multiple decision trees (ensemble method) to reduce overfitting and improve predictive accuracy. It's particularly effective in handling noisy data.
   - **Robustness**: By averaging across many trees, Random Forest is less sensitive to data variability and more stable in its predictions.
   - **Feature Importance Insights**: Random Forest also provides feature importance, which gives us insights into which factors are key in predicting survival across multiple trees, thus enhancing reliability.

4. **Why These Models?**
   - Both **Decision Tree** and **Random Forest** offer high interpretability and effectiveness in handling structured data with both categorical and continuous variables, like in the Titanic dataset.
   - **Complementary Strengths**: Decision Trees provide a clear, interpretable model, while Random Forest offers increased robustness and accuracy through ensemble learning.
   - **Comparison Goal**: By comparing these two models, we can better understand the trade-off between interpretability (Decision Tree) and predictive power (Random Forest), allowing us to choose the model that balances performance with interpretability for future decision-making.

Random Forest Classifer
Confusion Matrix



Confusion Matrix: The confusion matrix provides a detailed look at how well your model performs for each class:

- The rows represent the actual classes (0 and 1), and the columns represent the predicted classes (0 and 1).
- True Negatives (TN): 91 (correctly predicted class 0)
- False Positives (FP): 14 (incorrectly predicted class 1, when it was class 0)
- False Negatives (FN): 17 (incorrectly predicted class 0, when it was class 1)
- True Positives (TP): 57 (correctly predicted class 1)

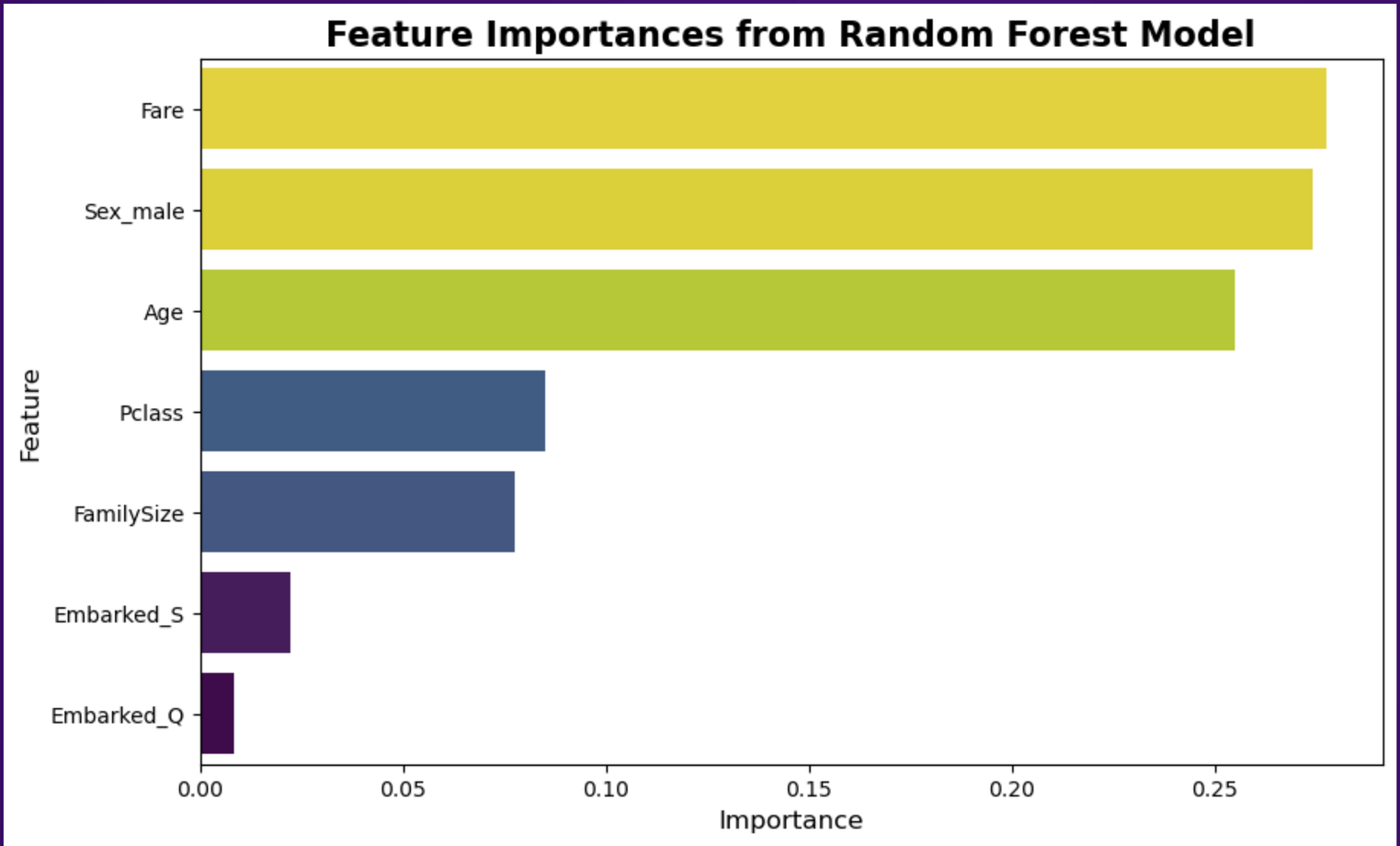
Random Forest Classifer
Classification Report

Classification Report:

	precision	recall	f1-score	support
0	0.84	0.87	0.85	105
1	0.80	0.77	0.79	74
accuracy			0.83	179
macro avg	0.82	0.82	0.82	179
weighted avg	0.83	0.83	0.83	179

1. Random Forest model performs better at predicting class 0 (non-survivors) than class 1 (survivors), with higher precision and recall for class 0.
2. The overall accuracy of 82.68% is a good sign, but the performance on class 1 (survivors) could be improved since the recall for class 1 is only 77%, meaning the model misses some survivors.

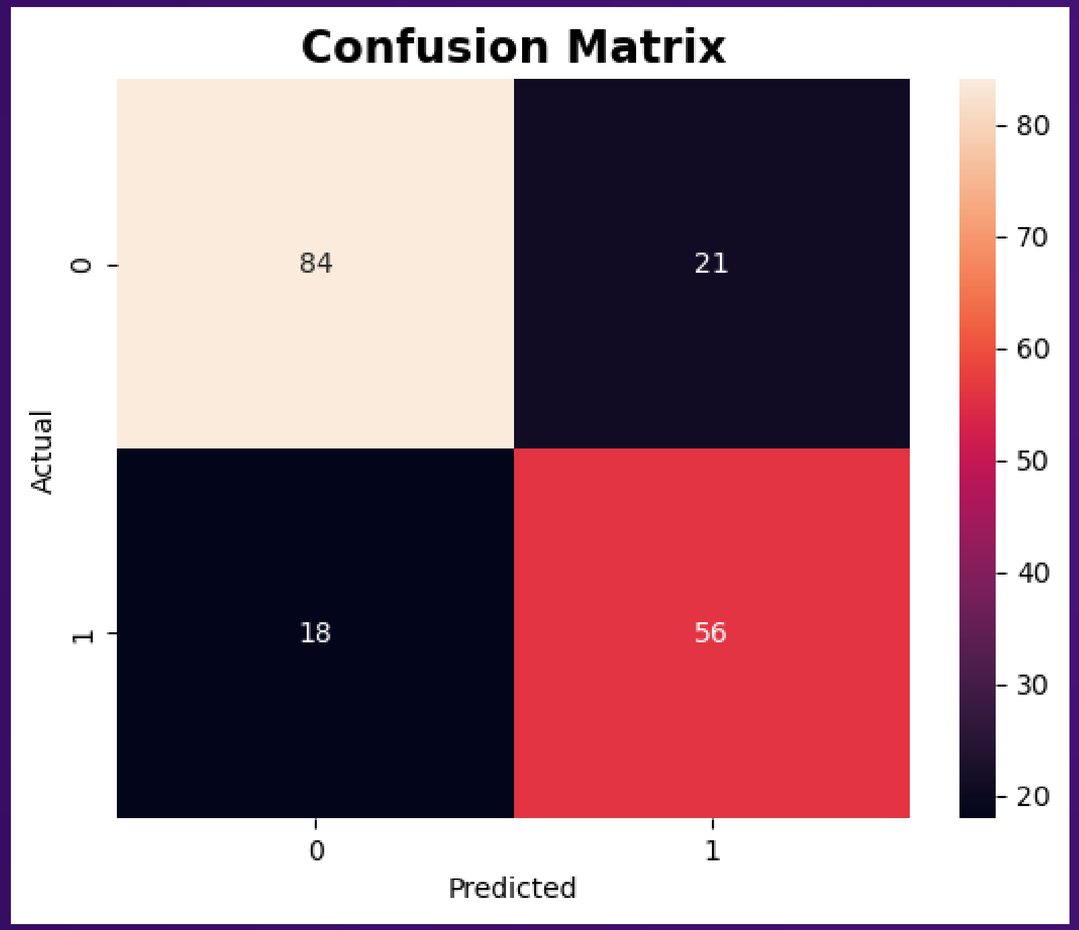
Random Forest Classifer
Feature Importances



1. Fare and Sex are the top contributors to survival prediction.
2. Age and Pclass also hold notable weight but are less influential than Fare and Sex.
3. FamilySize has some influence, while Embarked is the least significant feature.

These feature importances highlight the factors the model relies on most for prediction, helping you better understand the driving factors behind the survival rates in the Titanic dataset.

Decision Tree Classifer
Confusion Matrix



Confusion Matrix: The confusion matrix provides a detailed look at how well your model performs for each class:

- The rows represent the actual classes (0 and 1), and the columns represent the predicted classes (0 and 1).
- True Negatives (TN): 84 — The model correctly predicted 84 passengers who did not survive (Survived = 0).
- False Positives (FP): 21 — The model incorrectly predicted 21 passengers as surviving, but they did not (Survived = 0).
- False Negatives (FN): 18 — The model predicted 18 passengers as not surviving, but they actually survived (Survived = 1).
- True Positives (TP): 56 — The model correctly predicted 56 passengers who survived.

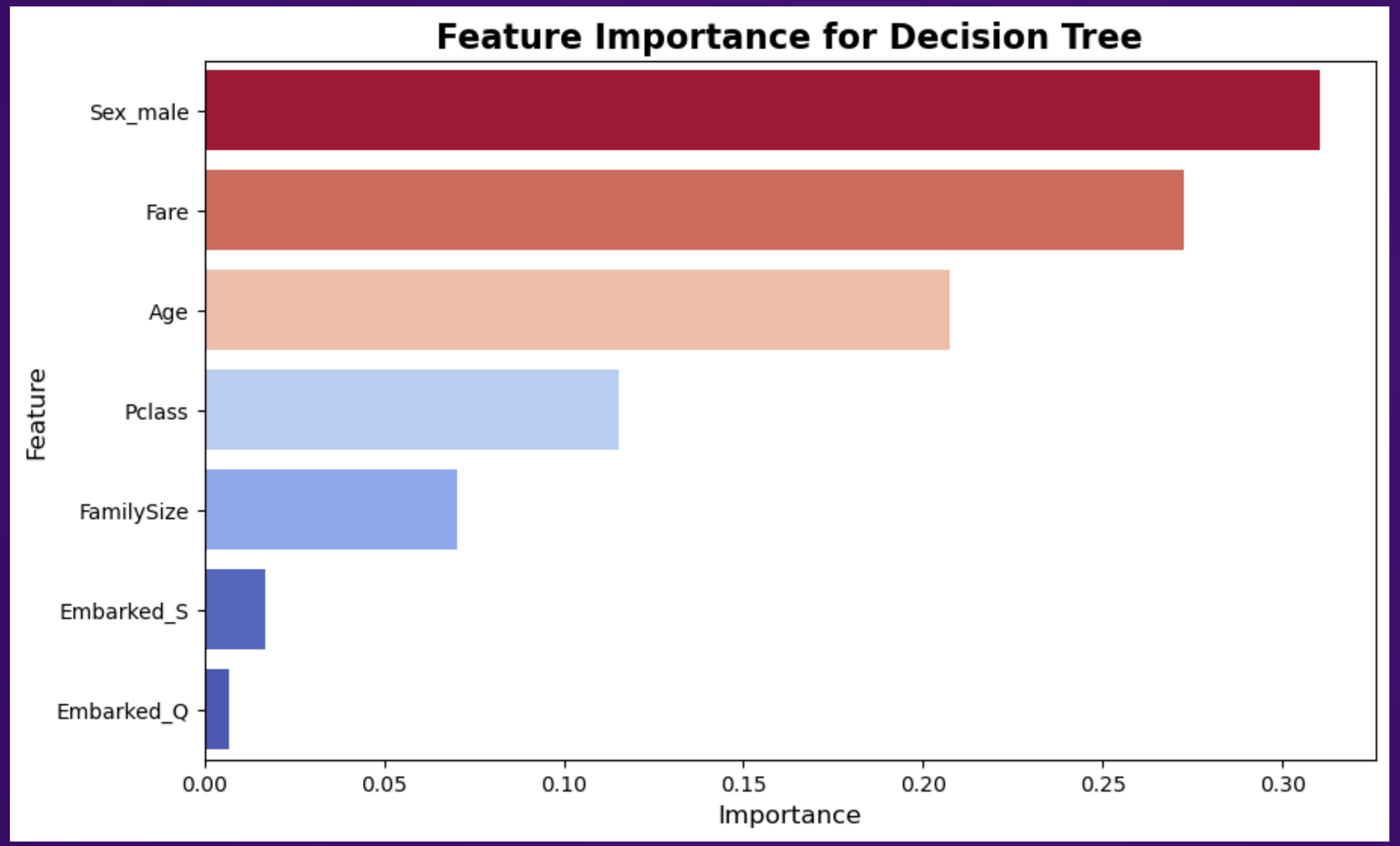
Random Forest Classifer
Classification Report

Classification Report:

	precision	recall	f1-score	support
0	0.82	0.80	0.81	105
1	0.73	0.76	0.74	74
accuracy			0.78	179
macro avg	0.78	0.78	0.78	179
weighted avg	0.78	0.78	0.78	179

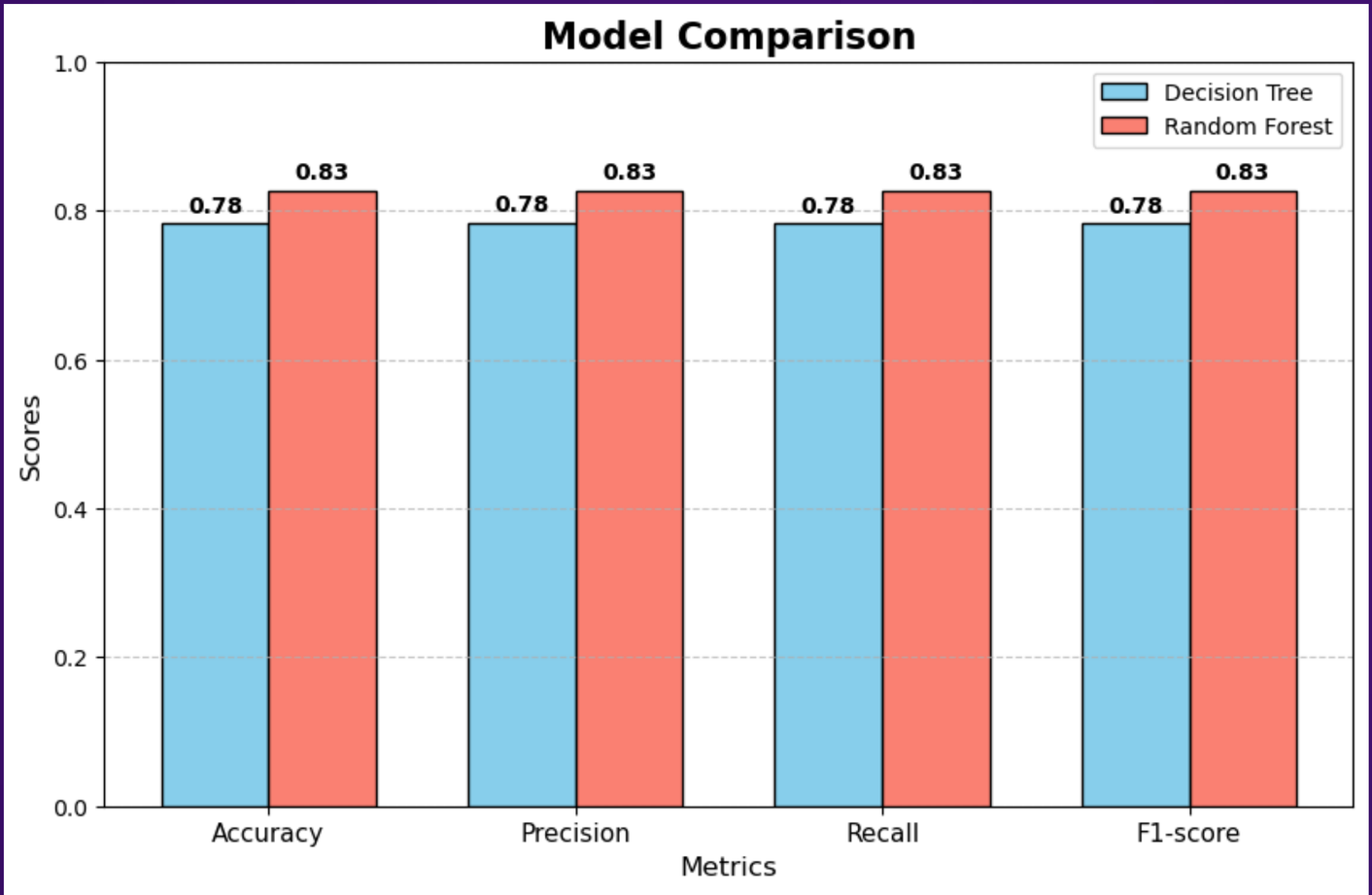
The model performs reasonably well, especially for predicting passengers who did not survive. However, it could be improved for better precision and recall for passengers who survived.

Random Forest Classifer
Feature Importances



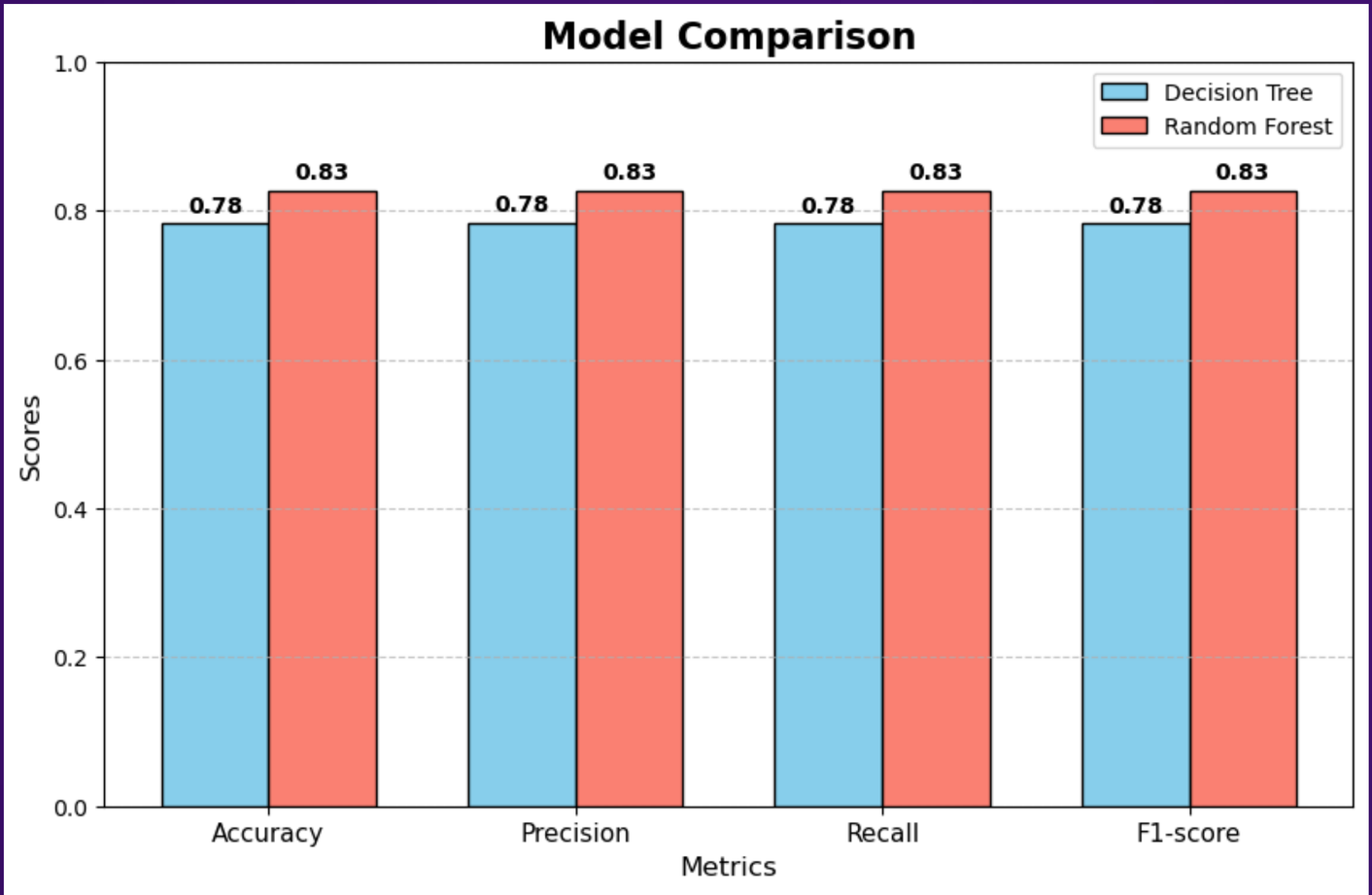
1. The model relies heavily on Sex, Fare, and Age for predictions, which aligns with the historical survival patterns on the Titanic.
2. Lower importance for FamilySize and Embarked suggests these features have a smaller impact on survival in this model.

Comparison of the Random Forest and Decision Tree Models



The bar chart effectively compares the performance metrics (Accuracy, Precision, Recall, and F1-score) of the Decision Tree and Random Forest models. The Random Forest model consistently outperforms the Decision Tree across all metrics, achieving higher scores (0.83) compared to the Decision Tree's scores (0.78). This visualization highlights the overall advantage of the Random Forest model in terms of classification performance on this dataset.

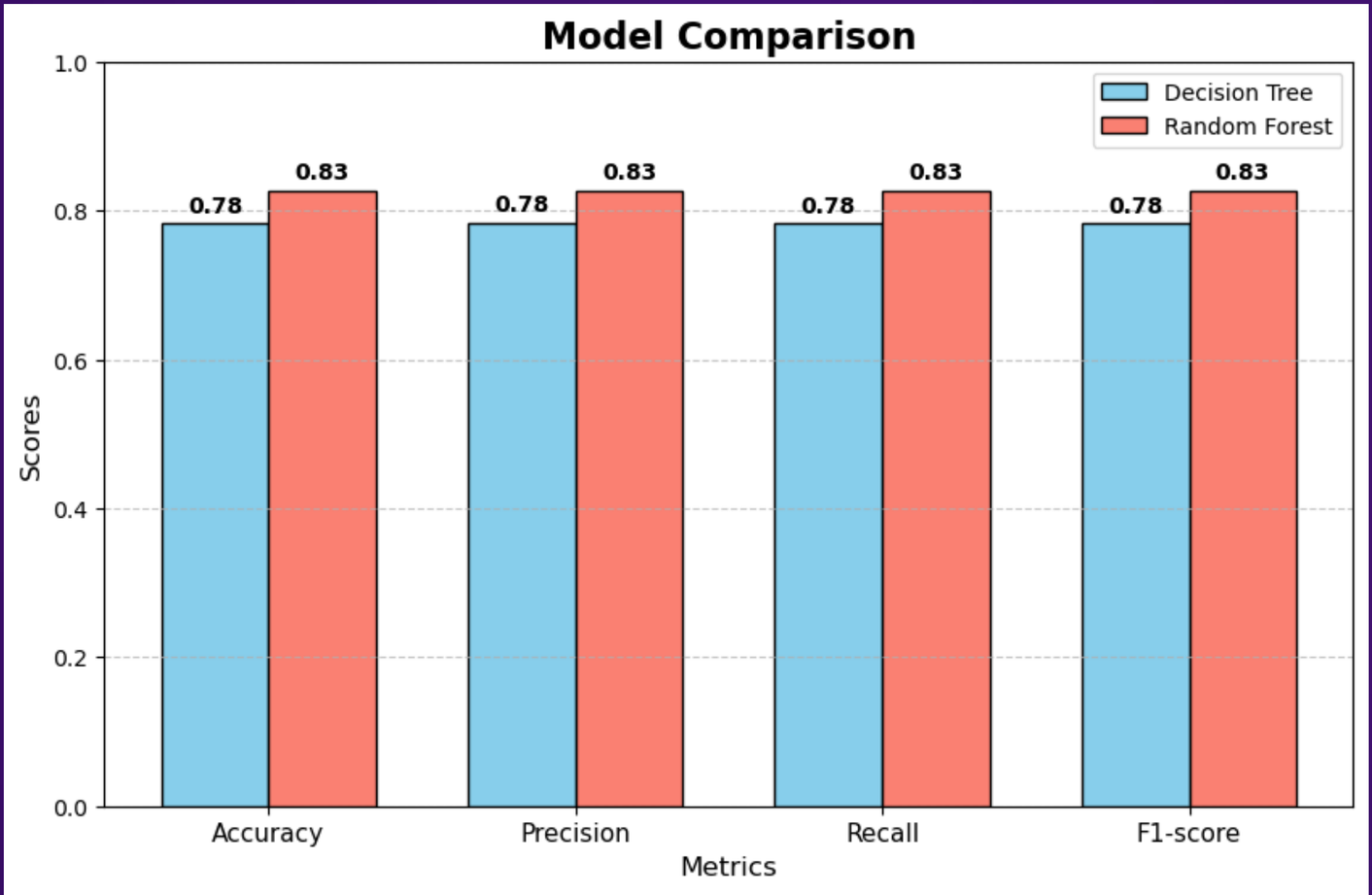
Comparison of the Random Forest and Decision Tree Models



Interpretation of Metrics

1. **Accuracy**
   - **Definition**: Accuracy is the proportion of correct predictions (both true positives and true negatives) out of all predictions.
   - **Interpretation**: Random Forest achieved an accuracy of **83%**, whereas the Decision Tree had **78%**. This suggests that the Random Forest model correctly classified a higher proportion of observations. Higher accuracy in Random Forest can be attributed to the ensemble effect, where multiple trees reduce variance and improve generalization.

Comparison of the Random Forest and Decision Tree Models

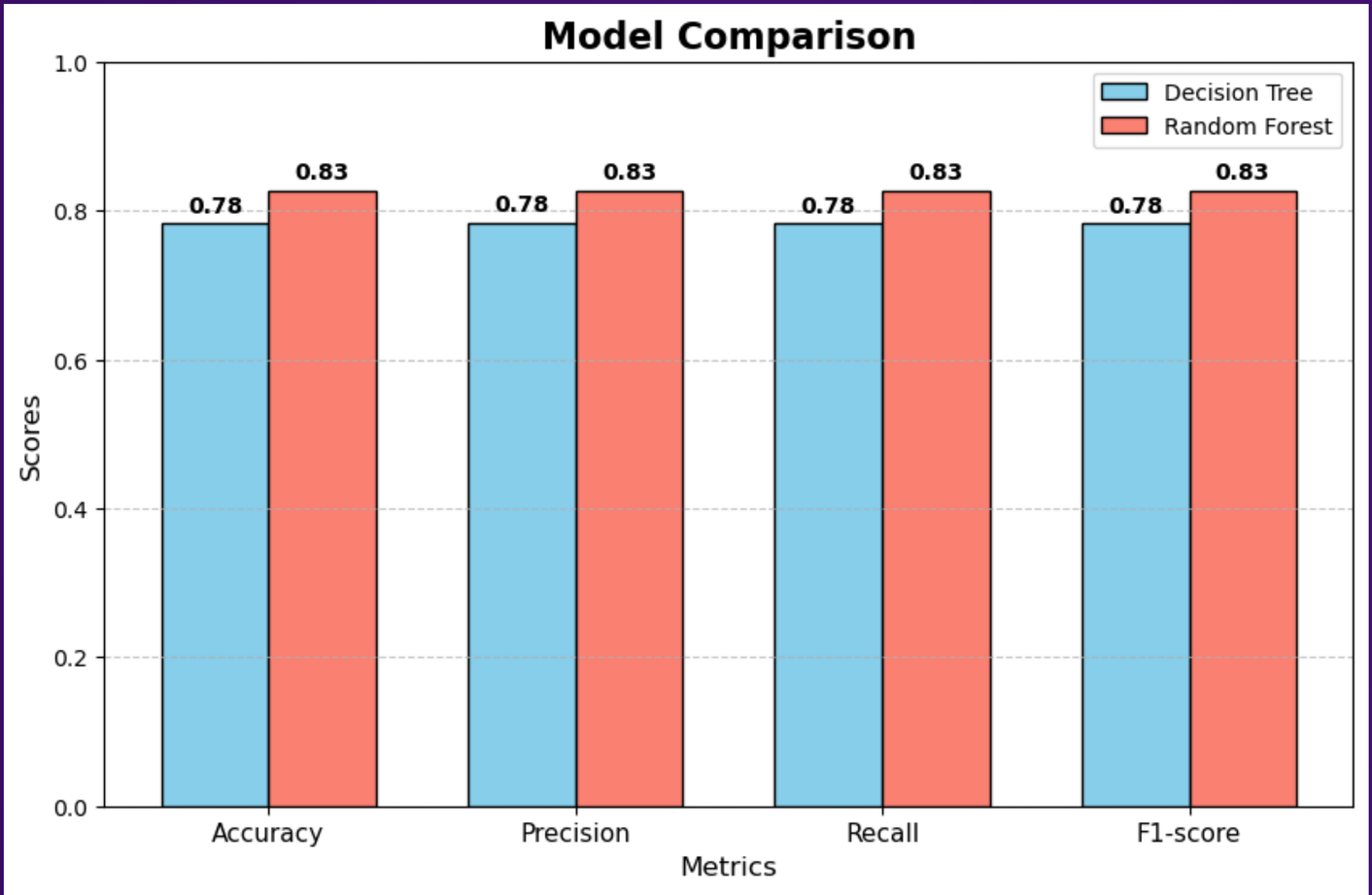


Interpretation of Metrics

2. Precision

- **Definition:** Precision (for a specific class) is the proportion of true positive predictions out of all positive predictions made by the model.
- **Interpretation:** Random Forest achieved a higher precision (0.83 vs. 0.78 for the Decision Tree). Precision is particularly important when you want to minimize false positives. In this context, Random Forest might be a better choice if you want more confidence that when it predicts a specific class (like survival), it's likely correct.

Comparison of the Random Forest and Decision Tree Models

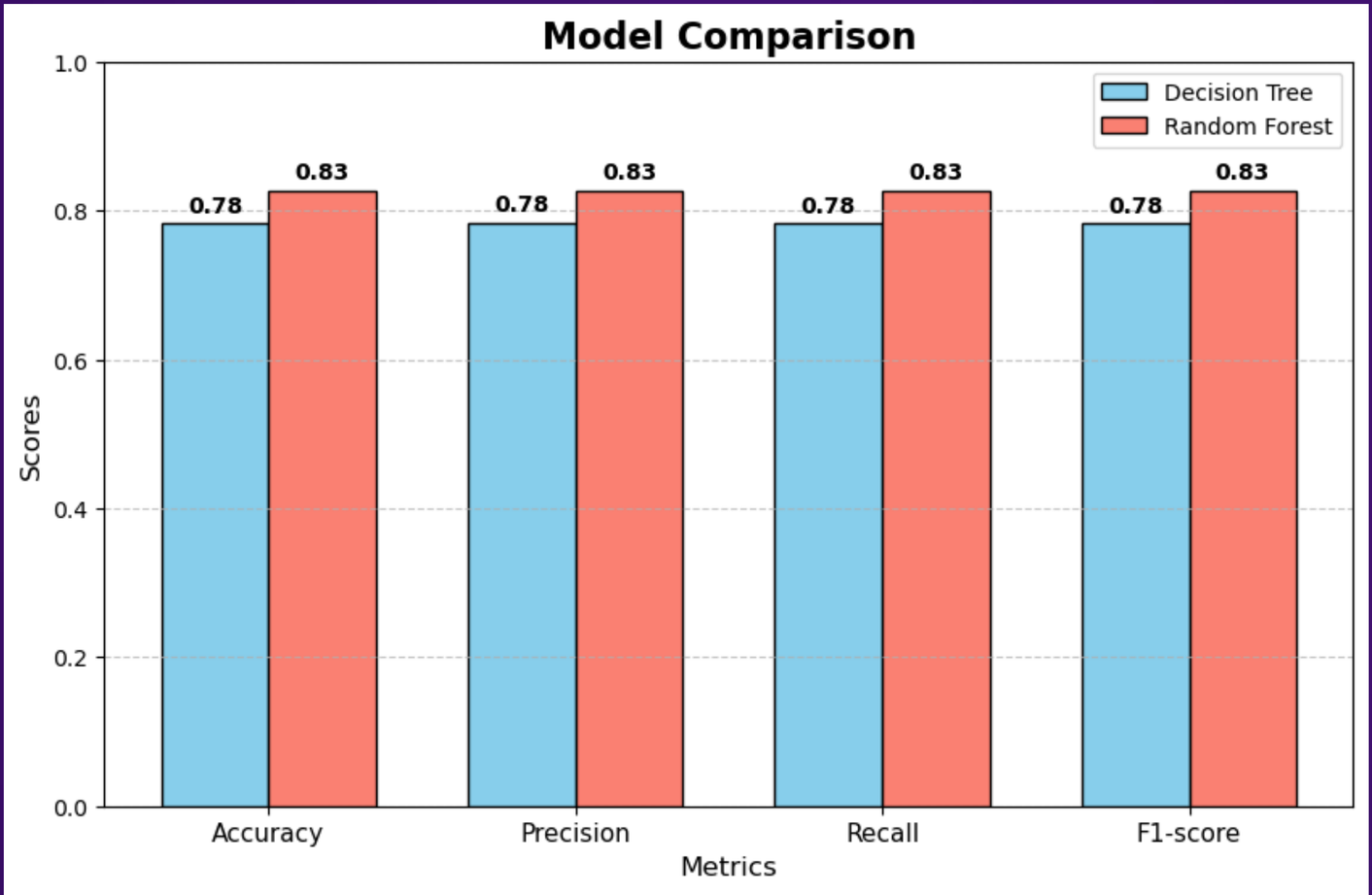


Interpretation of Metrics

3. Recall

- **Definition:** Recall is the proportion of true positive predictions out of all actual positive instances.
- **Interpretation:** Both Random Forest and Decision Tree performed the same in terms of recall (0.83 for Random Forest vs. 0.78 for Decision Tree). This indicates that Random Forest is better at capturing more actual positives, which can be valuable when missing a positive case (e.g., a survivor) is costly.

Comparison of the Random Forest and Decision Tree Models

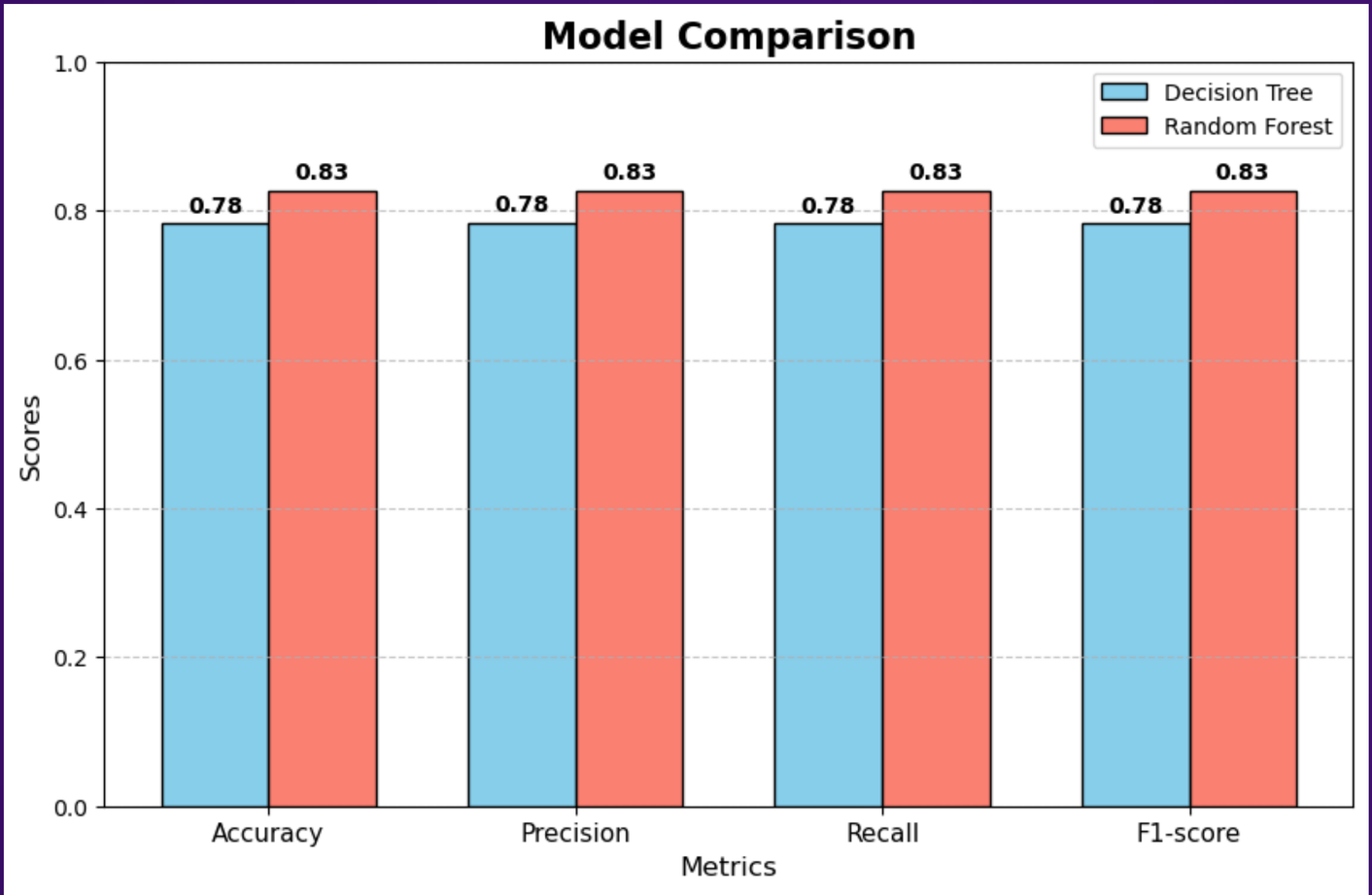


Interpretation of Metrics

4. **F1-score**
   - **Definition:** F1-score is the harmonic mean of precision and recall, balancing the two metrics.
   - **Interpretation:** Random Forest had a higher F1-score (0.83 vs. 0.78 for Decision Tree), meaning it provides a better balance between precision and recall. This is particularly useful if both false positives and false negatives have significant consequences.

Overall, **Random Forest is outperforming the Decision Tree in all four metrics.** This suggests that Random Forest is a more robust model for this dataset, likely because it reduces overfitting by averaging multiple trees and thus provides better generalization on unseen data.

Comparison of the Random Forest and Decision Tree Models



Summary

Model Performance: Random Forest outperformed the Decision Tree across all metrics, suggesting it's a better choice for balanced and reliable predictions.

The **Random Forest's ensemble approach** enables it to capture a wider set of patterns, which explains its stronger performance and better utilization of multiple features to make balanced predictions. If interpretability is important, the Decision Tree's simple structure is easier to understand, but for accuracy and reliability, Random Forest is the better model.

Syamsul Rizal Fany
saemfany@gmail.com

Thank You

For Your Attention

Visit My LinkedIn
Visit My GitHub