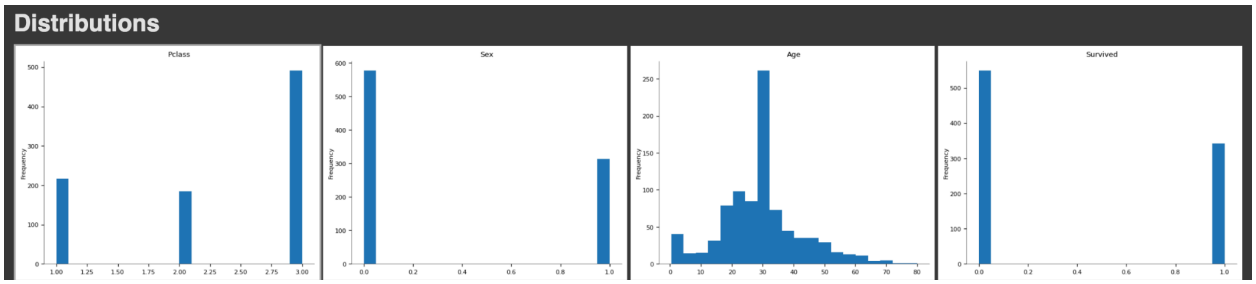
Report: Titanic Modelling and evaluation (Decision Tree and Random Forest)

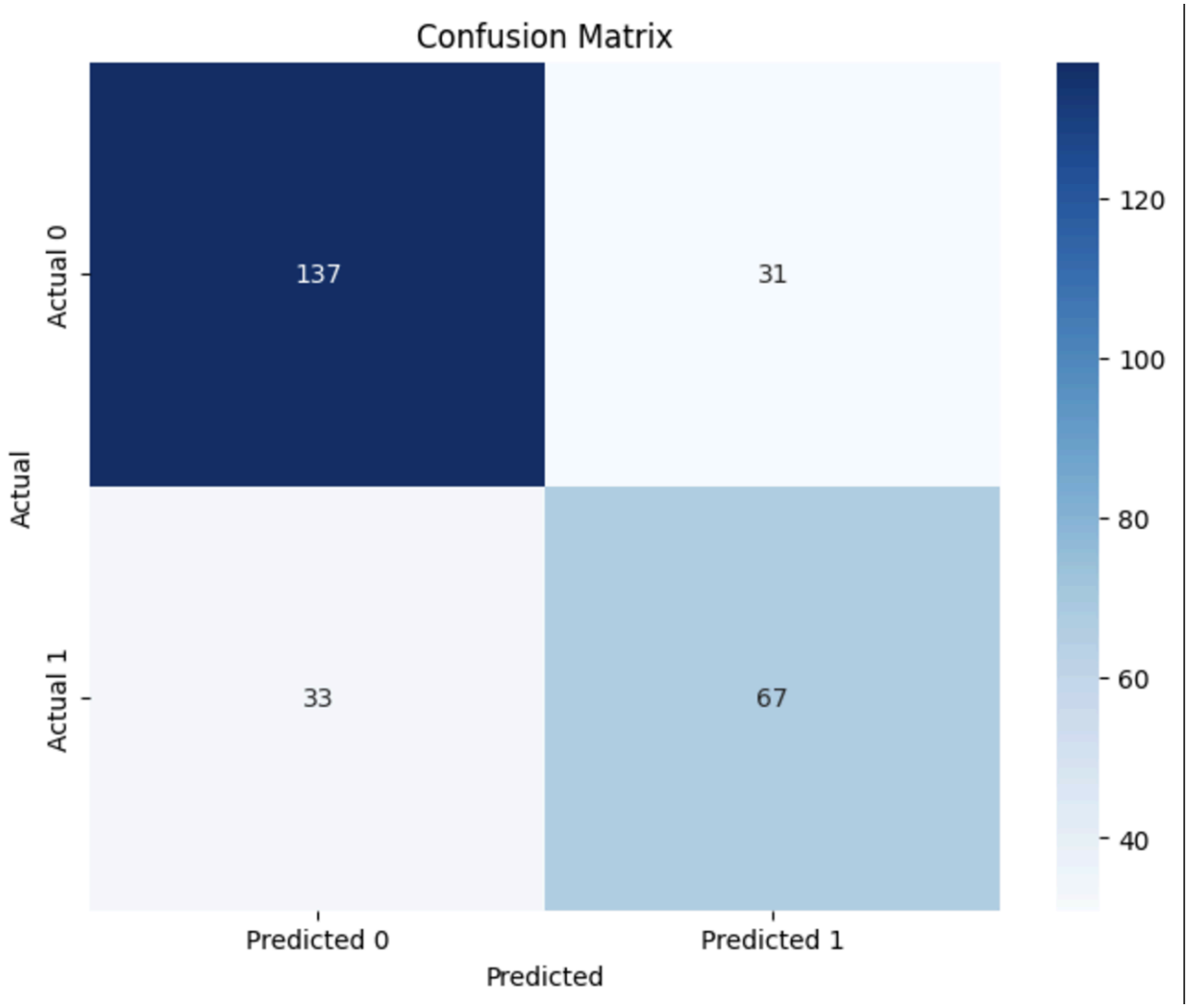
After analysing the findings of two models- decision tree and random forest, I observed that there are three different factors which contributed to analysing the passenger survival rate. These factors were 1) Gender of the passenger 2) Age of the passenger and 3) Pclass or passenger class (1st, 2nd and 3rd). The two models- decision tree and random forest and the individual analysis help us understand the accuracies and identify key factors of survival.



The distributions show key patterns in the titanic dataset which reveals that most of the passengers belonged to the 3rd class, followed by 1st and 2nd class. This reflects a large portion from the lower socio-economic class. The gender distribution shows a significant amount of female passengers compared to males had a higher priority during evacuation. The age distribution highlights a concentration of passengers between 20 and 40 years old and with fewer children and older individuals, indicating age as a likely factor in survival. The survival distribution shows that the majority of the passengers did not survive, reinstating the tragic outcome of the disaster.

Model-1: Decision Tree

To identify the performance of the model, I analyzed the confusion matrix, which provides a detailed breakdown of predictions. The model correctly predicted 137 non-survivors (true negatives) and 67 survivors (true positives), while it misclassified 31 non-survivors as survivors (true negatives) and 33 survivors as non survivors (false negatives). This indicates that the model performs better at identifying passengers who did not survive, as reflected by a higher recall of 82% of the “Not Survived” class compared to 67% for the “survived” class. These results suggest the model’s predictive strength lies in identifying the majority class (non-survivors), but it struggles with the minority class (survivors), which may be due to class imbalance as we discussed above.



Hyperparameters:

`max_depth = 3`

`random_state = 1`

Evaluation Scores:

Accuracy: 76%

Precision:

- Class 0 (Not Survived): 0.81
- Class 1 (Survived): 0.68

Recall

- Class 0 (Not Survived): 0.82

- Class 1 (Survived): 0.67

F1-Score:

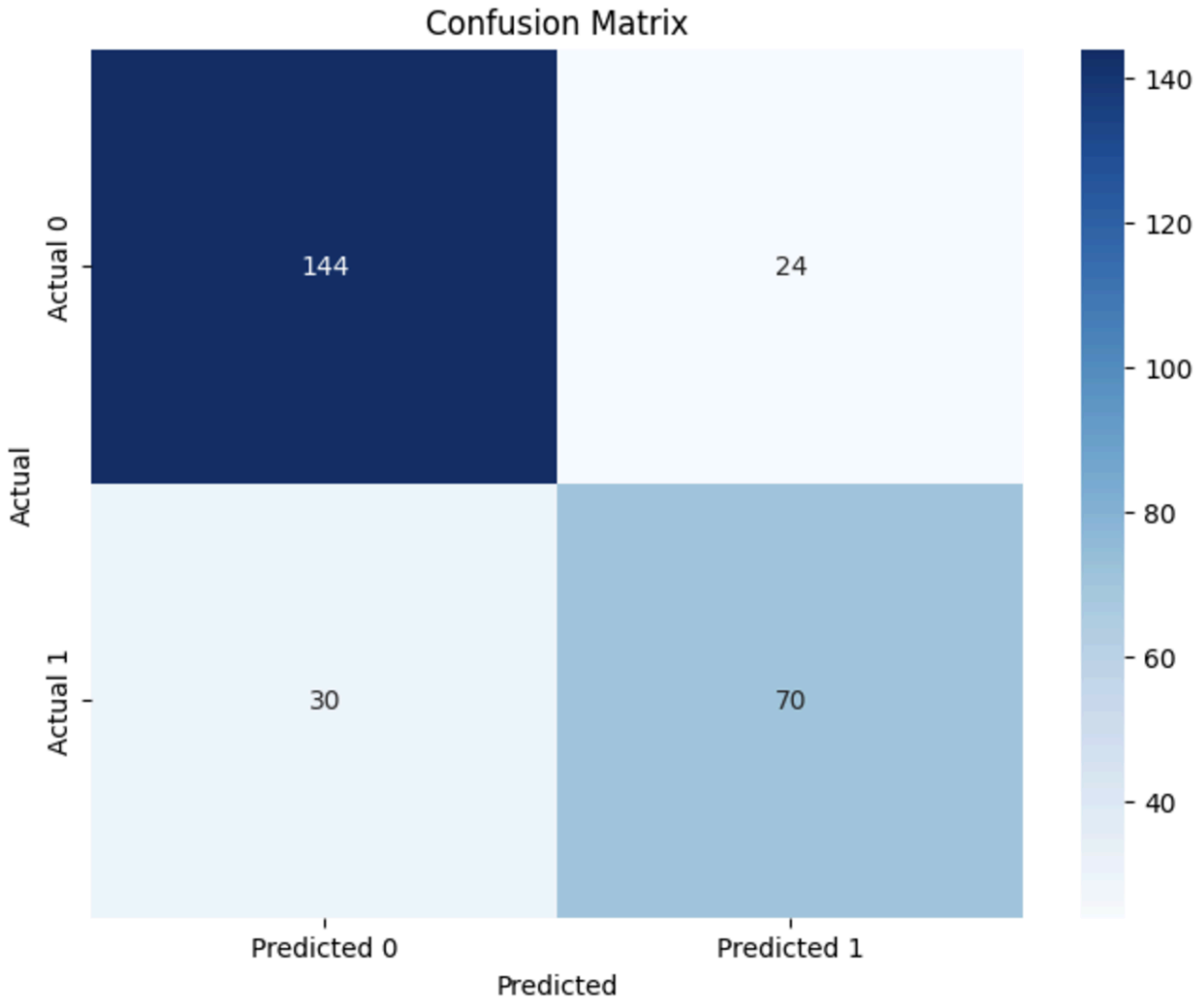
- Class 0 (Not Survived): 0.81
- Class 1 (Survived): 0.68

Model-2: Random Forest

To evaluate the Random Forest Classifier, I analyzed the feature importance bar chart and the confusion matrix.

The feature importance bar chart highlights the most influential predictors in the Random Forest model. The chart reveals that Sex is the most significant feature, followed by Age and Pclass. This indicates that the likelihood of survival is heavily influenced by gender, with females having a higher probability of survival, while age and passenger class also play important roles in predicting outcomes.

To assess the model's performance, I examined the confusion matrix, which shows the relationship between actual and predicted outcomes. The matrix indicates that the model correctly predicted 144 non-survivors (True Negatives) and 70 survivors (True Positives). However, it misclassified 24 non-survivors as survivors (False Positives) and 30 survivors as non-survivors (False Negatives). These results demonstrate an improvement over the Decision Tree model, with the Random Forest achieving a higher accuracy and better handling of the minority class (survivors). This highlights the Random Forest model's ability to capture complex relationships in the data, making it more reliable for predicting survival outcomes.



Hyperparameters:

```
n_estimators = 100

random_state = 1
```

Evaluation Scores:

Accuracy: 80%

Precision:

- Class 0 (Not Survived): 0.83
- Class 1 (Survived): 0.74

Recall:

- Class 0 (Not Survived): 0.86

- Class 1 (Survived): 0.70

F1-Score:

- Class 0 (Not Survived): 0.84
- Class 1 (Survived): 0.72

Key Findings

The analysis revealed some significant patterns in the survival outcomes of Titanic passengers:

1. Gender (Sex): Women had a much higher chance of survival than men, reflecting the "women and children first" principle followed during the evacuation.
2. Age: Younger passengers, especially children aged 14 and under, were more likely to survive, as they were prioritized during rescue efforts.
3. Pclass: Passengers traveling in 1st class had the greatest likelihood of survival, highlighting how socio-economic status played a crucial role in access to lifeboats and safety.

Conclusion

The Random Forest model proved to be the better option for predicting survival on the Titanic. It outperformed the Decision Tree model by delivering higher accuracy and effectively capturing the complex relationships within the data. Both models agreed that Sex, Age, and Pclass were the most important factors influencing survival.

This analysis not only reaffirms well-documented historical patterns of the Titanic tragedy but also demonstrates how machine learning can uncover insights from real-world events. By combining data with predictive models, we can better understand the factors that shaped one of the most well-known maritime disasters in history.