1. **How did your EDA insights guide your preprocessing and modelling?**

From the EDA I was able to identify that the dataset contained some missing values in the following columns: age, embarked, and deck. This helped in guiding me in preprocessing, where I imputed age with the median, filled embarked with the modem, and dropped deck due to it having an excessive number of missing values. The correlation analysis also showed that sex and pclass had a strong relationships with survival, which influenced my feature selection. Additionally, since feature like fare and age were skewed, I decided to standardize them to improve the performance of distance-based models.

2. **Which model performed best overall and why?**

Among the models I used the Logistic Regression model performed the best overall. It achieved the highest balance between precision, recall, and F1-score, and it also had a strong AUC value in the ROC curve analysis. This shows that Logistic Regression was able to capture the underlying linear relationships between the features and the target variable effectively. The Decision Tree tended to overfit the training data, leading to weaker generalization, while kNN was more sensitive to scaling and did not perform as consistently. The strength of Logistic Regression lies in its ability to handle categorical predictors like sex and pclass well after encoding, which contributed to its superior performance.

**3. If your dataset had class imbalance, how did it affect your metrics?**

The dataset has a mild class imbalance, with approximately 62% of passengers not surviving and 38% surviving. This imbalance affected some of the evaluation metrics, particularly recall and precision, since certain models leaned toward predicting the majority class (non-survivors).

4. **Suggest one way to improve model performance (feature engineering, resampling, hyperparameter tuning).**

One way in which to improve the model it would be to apply Hyperparameter tuning, such as adjusting the maximum depth of the Decision Tree or the number of neighbors in kNN, could also yield better results.

**Conclusion**

The insights from the EDA directly influenced the preprocessing choices, such as handling missing values and scaling features. Logistic Regression emerged as the best-performing model due to its balance across multiple metrics. The mild class imbalance highlighted the importance of using precision, recall, F1-score, and ROC-AUC alongside accuracy. Finally, potential improvements such as hyperparameter tuning which identified ways to further enhance model performance**.**