For my final project, my initial idea was to create a Streamlit application that would generate images based on user-input prompts. However, after running into technical issues with Streamlit and having a conversation with a friend about the very popular movie *Titanic*, I decided to shift directions. That discussion led to a long-running debate: could Jack have survived? Not on the door, but as a passenger on the *Titanic*, and what factors, such as class or gender, would have influenced his chances of survival. This project was my way of finally settling that debate.
To answer this question, I developed and evaluated predictive models to estimate passenger survival on the Titanic using historical data. The main task was a binary classification problem: predicting whether a passenger survived or did not survive the disaster. Using demographic and socioeconomic features commonly associated with survival outcomes, I trained and compared two machine learning models: Logistic Regression and Random Forest. Beyond standard model evaluation, I applied these trained models to a hypothetical "Jack-like" passenger inspired by the film, allowing for a clear and interpretable discussion of survival probabilities and the structural factors that shaped them.
Overall, the Random Forest model outperformed Logistic Regression in terms of predictive accuracy and overall performance. The results also confirmed well-established historical patterns, particularly the strong influence of gender, passenger class, and age on survival. The analysis of the "Jack" scenario further illustrates how individual survival outcomes were shaped less by personal choices and more by broader structural inequalities present on the ship.

The dataset used in this project is the Titanic passenger dataset, originally published on Kaggle and widely used for educational and benchmarking purposes. It contains records for 891 passengers and includes both survival outcomes and individual passenger characteristics.
The target variable, Survived, is binary, where 1 indicates survival and 0 indicates non-survival. The predictor variables selected for modelling were:

- Pclass: Passenger class (1st, 2nd, or 3rd)
- Sex: Passenger gender
- Age: Passenger age in years
- Fare: Ticket fare paid
- Embarked: Port of embarkation (C, Q, or S)

These features were chosen because they are interpretable, historically meaningful, and commonly used in Titanic survival analyses. The dataset does contain missing values, particularly in the Age and Embarked columns. Rather than removing rows with missing data, which could reduce the dataset size and introduce bias, missing values were handled during preprocessing.

Before training the models, the data were split into training (80%) and testing (20%) sets using a stratified split to ensure that the survival ratio remained consistent across both sets. Preprocessing was implemented using scikit-learn pipelines so that all transformations were applied consistently during training and evaluation.
Numeric features (Age and Fare) were processed using median imputation to handle missing values, followed by standardisation to ensure comparable feature scales. Categorical features (Sex, Embarked, and Pclass) were processed using most-frequent imputation and one-hot encoding, with the first category dropped to avoid multicollinearity.

Two classification models were trained for comparison. Logistic Regression was used as a baseline model because of its simplicity and interpretability. It estimates survival probability as a linear function of the input features. The second model, a Random Forest classifier, is an ensemble method that combines multiple decision trees to capture non-linear relationships and interactions

between variables. The Random Forest model was trained using 300 trees to ensure stable and reliable performance.
Model performance was evaluated using accuracy, precision, recall, F1-score, confusion matrices, and receiver operating characteristic (ROC) curves. These metrics provided both overall performance comparisons and more detailed insights into how well each model classified survivors versus non-survivors.
The Logistic Regression model achieved an accuracy of approximately 0.78 on the test set, while the Random Forest model achieved an accuracy of approximately 0.82. The Random Forest model also produced a higher ROC AUC score of around 0.84, indicating a stronger ability to distinguish between survivors and non-survivors across different classification thresholds. Confusion matrix analysis showed that the Random Forest model reduced both false positives and false negatives compared to Logistic Regression, particularly improving recall for the survivor class. This suggests that survival outcomes are influenced by non-linear relationships that Logistic Regression cannot fully capture.
Exploratory data analysis and model results consistently showed that Sex was the strongest predictor of survival, followed by Passenger Class and Age. Female passengers and first-class passengers had significantly higher survival rates, reflecting historical evacuation priorities and unequal access to lifeboats. Higher fares were also associated with increased survival, likely serving as a proxy for socioeconomic status.

To make the results more tangible, I created a hypothetical "Jack-like" passenger based on the movie: a 20-year-old male traveling in third class, paying a low fare, and embarking from Southampton. The predicted probability of survival for this passenger was approximately 0.10 using Logistic Regression and 0.03 using the Random Forest model. These low probabilities indicate that a passenger with Jack's characteristics belonged to one of the highest-risk groups aboard the Titanic.
Additional "what-if" scenarios helped illustrate this further. When the same passenger was modelled as female, the predicted survival probability increased dramatically, exceeding 0.80 in the Random Forest model. Assigning the passenger to first class also significantly improved survival odds, while increasing age reduced them further. These scenarios clearly demonstrate that survival was not simply the result of individual actions but was heavily shaped by structural factors such as gender norms and class-based access to safety.

In conclusion, this project shows that machine learning models can effectively capture and quantify historical survival patterns from the Titanic disaster. The Random Forest model outperformed Logistic Regression by modelling complex, non-linear relationships among passenger characteristics. The findings reaffirm the dominant role of gender, class, and age in determining survival outcomes.
The "Jack" case study provided a clear and engaging way to apply predictive modelling to a familiar narrative, turning abstract probabilities into a concrete conclusion. Based on the model's results, Jack's death aligns with statistically expected outcomes for passengers in his demographic group. In other words, Jack would have died regardless of Rose's selfishness.