Laboratory 4

Variant 4 – Survival prediction model in Titanic Crash

Lab group 105

Group 9 – Aditya Kandula, Martyna Wielgopolan

I. Introduction

The goal is to build models that can predict whether a given passenger survived the Titanic shipwreck, based on various features such as age, sex, ticket class, fare paid etc... This task uses real data collected from passengers aboard the RMS Titanic:

- **Demographics**: Age, sex, and passenger class (Pclass)
- **Travel Details**: Fare, number of siblings/spouses or parents/children aboard, embarkation port
- **Survival status**: The target variable (0 = did not survive, 1 = survived)

The objective is to explore and clean the dataset, engineer relevant features, and apply machine learning models to accurately predict survival.

Two models — **Logistic Regression** and **Random Forest** — are trained, validated using cross-validation, and compared based on performance metrics like accuracy, precision, recall, and F1-score.

By analysing model behaviour and the importance of individual features (e.g. gender), this task provides both practical insights into model interpretability and technical training in machine learning pipeline from data preprocessing to model evaluation.

II. Implementation

We built a classic machine learning pipeline that:

- Loads the Titanic dataset,
- Preprocesses it,
- Trains two models (Logistic Regression and Random Forest),
- Evaluates them using cross-validation and test accuracy,
- Visualizes results.

Data selection:

The dataset is loaded via Seaborn's `load_dataset("titanic")` utility. Several steps were taken to prepare the data such as:

- **Target Re-mapping**: The original "**survived**" column was renamed to "**target**" for clarity.
- **Column Pruning**: Columns like **deck**, **embark_town**, **alive**, and **who** were removed due to high redundancy or missingness.
- **Missing Values**: Median imputation was used for the age column, and the mode was used for the embarked column.
- **Categorical Encoding**: Categorical features like sex and embarked were converted into binary dummy variables using one-hot encoding (drop_first=True to avoid multicollinearity).
- **Type Consistency**: The target column was explicitly cast to integer to ensure compatibility with scikit-learn classifiers.

These steps were performed to standardize the dataset and prepare it for binary classification while minimizing information leakage or bias.

Model selection:

Two classification models were implemented and evaluated:

- **Logistic Regression**: Chosen as a simple, interpretable linear model suitable for establishing a performance baseline.
- **Random Forest Classifier**: Selected as a more complex ensemble method capable of capturing non-linear relationships and providing feature importance scores.

Both models were evaluated using 4-fold cross-validation on the training data to estimate generalization accuracy. Subsequently, the models were retrained on the full training set and tested on a held-out test partition. The following metrics were used for evaluation:

- **Accuracy**: For both cross-validation and test sets.
- **Classification Report**: Precision, recall, and F1-score for both classes.

- **Confusion Matrix**: Visualized using `seaborn.heatmap` to intuitively assess the distribution of predictions.
- **Feature Importance**: Extracted and visualized for the Random Forest model to understand which features contributed most to predictions.

All generated plots are saved automatically to a `plots/` directory for inclusion in the report or further analysis.

III. Discussion

Let us see the Logistic Regression and Random Forest Classifier and compare the results:

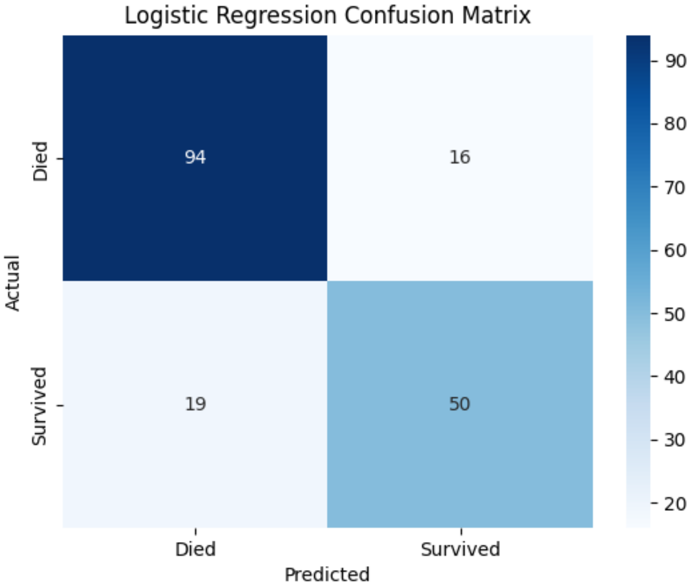
Model Performance:

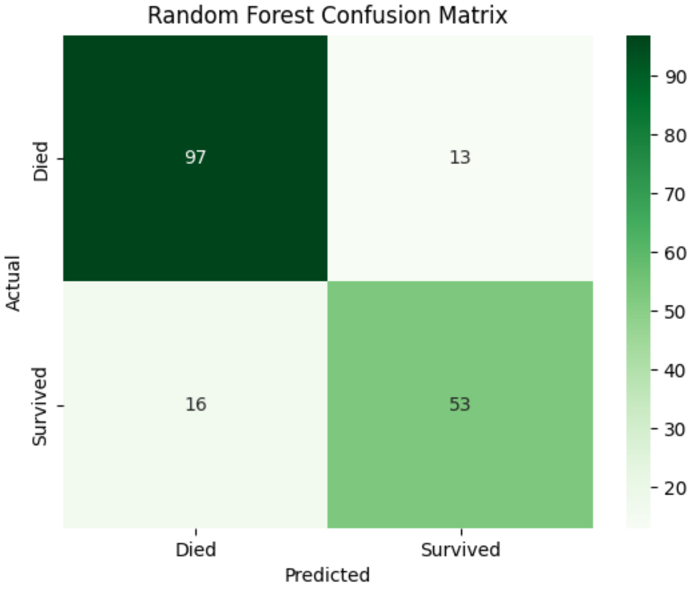
Both models achieved reasonable classification accuracy, with Random Forest slightly outperforming Logistic Regression. This result aligns with expectations, as ensemble methods like Random Forest are known to capture more complex interactions between variables.

- **Logistic Regression** demonstrated good generalization and interpretability, performing well given its linear nature. However, it was slightly less effective at correctly identifying survivors, which may be attributed to non-linear patterns in the dataset.
- **Random Forest Classifier** consistently produced higher cross-validation and test accuracies. It also showed better precision and recall scores, especially for the minority class (survived). This suggests that Random Forest was more robust to class imbalance and better at identifying subtler feature interactions.

Confusion Matrix Insights:

From the confusion matrices, both models showed high true negative rates (correctly predicting deaths), but the Random Forest model had fewer false negatives and false positives compared to Logistic Regression. This improvement translated into more balanced precision and recall for both classes, which is crucial in real-world survival prediction scenarios where both false alarms and missed detections carry consequences.





Feature Importance:

The feature importance plot for Random Forest revealed that `sex_male` was the most influential predictor, followed by `fare` and `age`. This corresponds well with historical records and domain knowledge, where gender and ticket price (a proxy for passenger class) were strong indicators of survival chances. Such consistency supports the validity of the model and the preprocessing approach.

Modeling Strategy Reflection:

The decision to compare a simple linear model with a more advanced ensemble model allowed us to explore the trade-off between interpretability and performance. Logistic Regression served as a transparent baseline, while Random Forest provided a more flexible and powerful alternative without extensive hyperparameter tuning.

The use of cross-validation ensured reliable performance estimation, and visualizations helped interpret model behavior and identify areas for further refinement.

IV. Conclusions

In this project, we implemented and compared two classification models—Logistic Regression and Random Forest—on the Titanic dataset. Both models demonstrated solid performance, with **Random Forest achieving slightly higher accuracy** and more balanced classification metrics.

Logistic Regression reached a cross-validation accuracy of **79.63%** and a test accuracy of **80.45%**, making it a reliable baseline with interpretable coefficients. In contrast, the Random Forest model yielded a cross-validation accuracy of **79.07%** and a higher test accuracy of **83.80%,** indicating its superior ability to generalize to unseen data.

The Random Forest model also produced better precision, recall, and F1-scores across both classes, particularly in distinguishing survivors (the minority class), where it reduced false positives and negatives compared to Logistic Regression. Feature importance analysis confirmed known survival predictors such as gender, fare, and age.

Overall, the results support the effectiveness of ensemble methods like Random Forest in handling moderately imbalanced classification tasks, while also validating the utility of simpler models like Logistic Regression for quick baselines and interpretability. The entire implementation adhered to reproducible and clean coding standards, and the outcomes provide a solid foundation for further exploration, including hyperparameter tuning or integration of additional features.