Titanic Survival Prediction using Decision Tree and Random Forest

Github Link

1 Abstract

This report explores the application of Decision Trees and Random Forest classifiers to predict passenger survival on the Titanic dataset. We analyse feature importance, model accuracy, and data correlations to evaluate the effectiveness of these models. The results show that Random Forest performs better than a single Decision Tree due to its ensemble learning approach. A detailed exploration of dataset attributes, preprocessing techniques, and visual analysis of survival factors is also included.

2 Introduction

Machine learning techniques have been widely used to analyse structured datasets and make predictions. The Titanic dataset, a well-known dataset in machine learning, contains information about passengers, including demographics and survival status. Decision Tree and Random Forest classifiers are applied to this dataset to predict survival of passengers. This tutorial aims to compare their performances and understand which features contribute the most to survival predictions.

Supervised learning models such as Decision Trees provide interpretable decision-making processes, whereas ensemble methods like Random Forest improve accuracy by reducing variance. The goal of this tutorial is to determine whether ensemble learning outperforms a single decision tree when applied to real-world survival predictions.

3 Dataset Description

The dataset used in this study is sourced from the GitHub repository of Data Science Dojo [1]. It consists of the following key attributes:

- Survived: Target variable (1 = Survived, 0 = Did not survive)
- Pclass: Ticket class (1st, 2nd, 3rd)
- Sex: Gender of passenger
- Age: Age of passenger
- SibSp: Number of siblings/spouses aboard
- Parch: Number of parents/children aboard

- Fare: Fare paid for the ticket
- Embarked: Port of embarkation (C, Q, S)

4 Data Preprocessing

Data preprocessing is a critical step in any machine learning activity to ensure optimal model performance. The following preprocessing steps were applied to the dataset:

- Handling Missing Values: Missing values in the Age and Embarked columns were dropped to maintain data consistency.
- Encoding Categorical Variables: The Sex and Embarked variables were converted into numerical representations.
- Feature Scaling: Normalization was applied to numerical features such as Age and Fare to improve model performance.
- Splitting the Dataset: The dataset was split into 80% training data and 20% testing data to evaluate model performance.

5 Methodology

5.1 Model Selection

Decision Tree and Random Forest classifiers were chosen due to their interpretability and effectiveness in handling structured data.

5.2 Model Training:

- A Decision Tree Classifier was trained to learn decision rules based on passenger attributes.
- A Random Forest Classifier with 100 trees was used for improved accuracy through ensemble learning.

5.3 Model Evaluation:

Accuracy, classification reports, and feature importance were analysed. Various visualizations were created to understand data distribution and model insights.

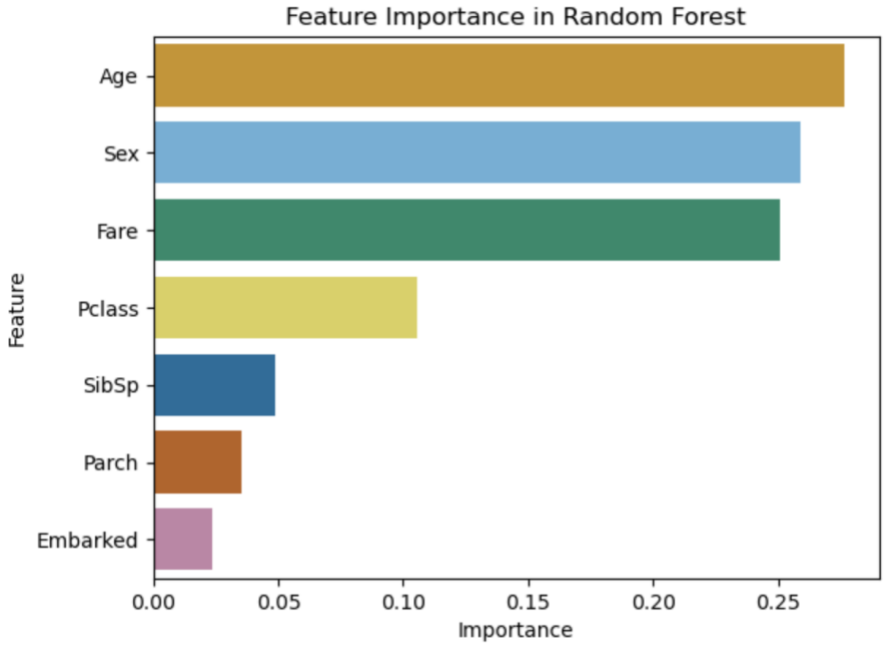
6 Results and Discussion

6.1 Model Performance

- Decision Tree Accuracy: 72%
- Random Forest Accuracy: 76.9%
- The Random Forest model outperformed the Decision Tree due to its ability to reduce overfitting by averaging multiple trees.

6.2 Feature Importance

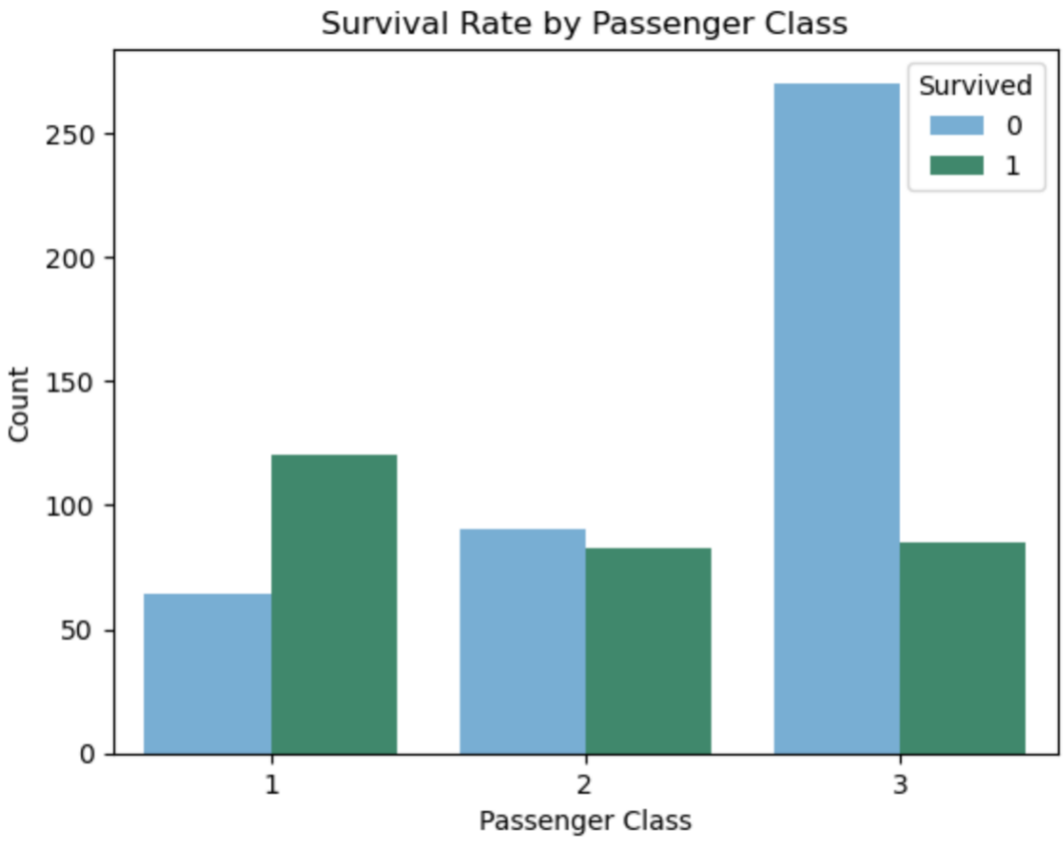
- The most effective features in predicting survival were Fare, Sex, and Pclass, as highlighted by the feature importance plot.



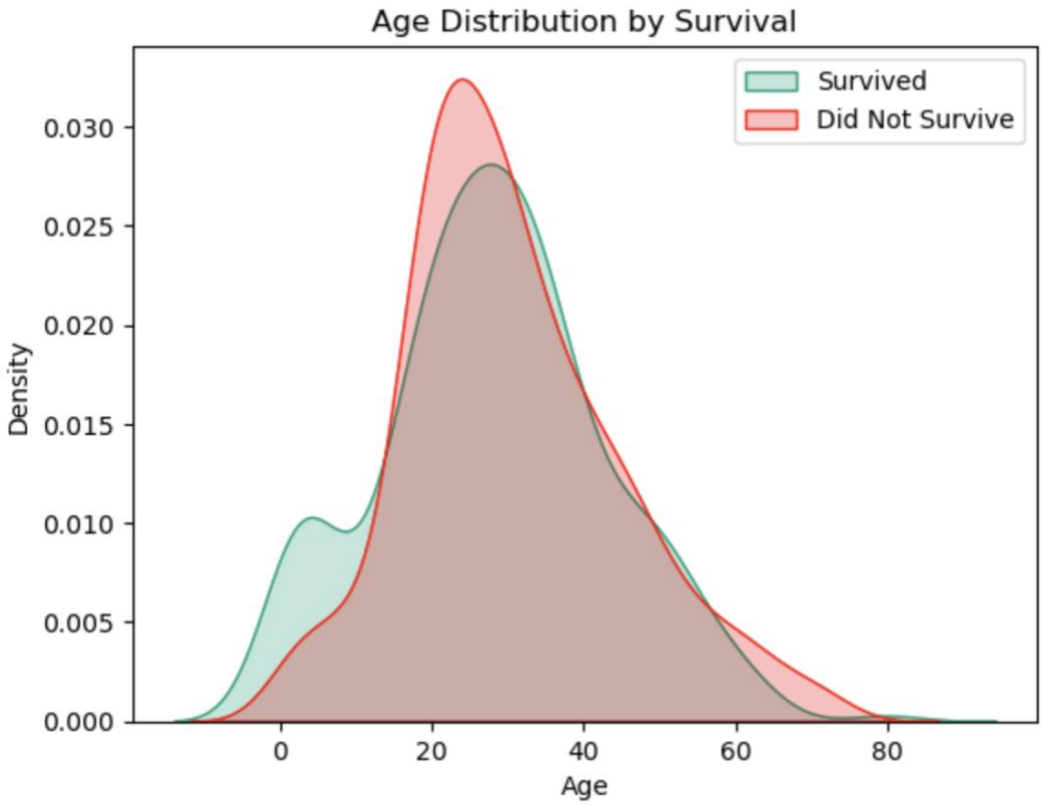
- Feature importance analysis suggests that social and economic factors played a significant role in survival likelihood.

6.3 Visual Analysis:

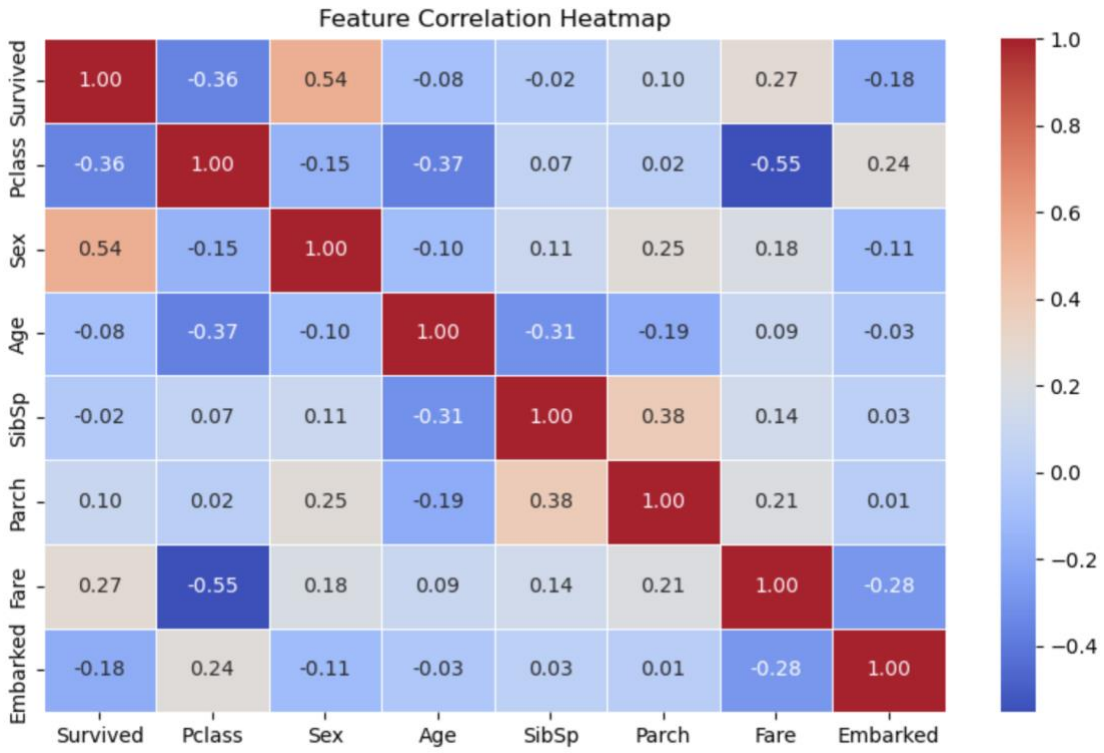
- Survival Rate by Passenger Class: Higher survival rates for 1st class passengers.



- Age Distribution by Survival: Younger passengers had a higher likelihood of survival.



- Correlation Heatmap: Fare showed a strong correlation with survival, which means wealthy passengers had more chances of survival.



6.4 Impact of Different Factors

- Socioeconomic Status: The correlation between class and survival suggests that wealthier passengers had better chances of survival. This shows the priority given to wealthier passengers during evacuation.
- Gender Influence: The well-documented "women and children first" protocol is evident in the survival rates. This aligns with historical accounts of the Titanic tragedy.
- Family Size and Survival: Larger families had a mixed impact on survival rates. While companionship may have helped some individuals, it may have hindered escape for others.

6.5 Comparison of Model Interpretability

- Decision Trees provide clear, interpretable decision paths but are more prone to overfitting.
- Random Forest enhances accuracy by averaging multiple decision trees, making it a more robust model.

7 Future Improvements

While the Random Forest model performed well, several enhancements can be made to improve prediction accuracy and model robustness:

- Hyperparameter Tuning: Further optimization of parameters such as tree depth, number of trees, and feature selection could improve performance.
- Feature Engineering: Creating new features based on passenger interactions, family groupings, or title-based information could provide additional insights.
- Advanced Models: Exploring gradient boosting techniques like XGBoost or neural networks might enhance accuracy further.

8 Accessibility Considerations

To ensure that this study is accessible to a wider audience, several steps were taken to enhance readability and usability:

- Colorblind-Friendly Plots: All visualizations use a color palette that is distinguishable for individuals with color blindness.
- Descriptive Labels: Clear axis labels and titles were added to each plot for better readability.
- Text Contrast: High contrast colors were used in all figures and textual descriptions to ensure visibility.
- Text Descriptions: Descriptions for plots are included so that screen readers can interpret the information.

9 Conclusion

This study demonstrates the effectiveness of Decision Tree and Random Forest classifiers for predicting Titanic survival. The Random Forest model performed better due to its ensemble approach, reducing overfitting and increasing prediction stability. Feature importance analysis provided insights into key survival factors, indicating that social class, gender, and fare amount significantly impacted survival rates.

Future work may explore hyperparameter tuning or additional feature engineering to further improve model accuracy. Additionally, incorporating advanced machine learning techniques, such as gradient boosting or neural networks, could enhance predictive performance.

References

[1] DataScienceDojo, "Github," [Online]. Available: https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv.

[2] T. T. R. &. F. J. Hastie, "The Elements of Statistical Learning. Springer.," [Online]. Available: https://link.springer.com/book/10.1007/978-0-387-84858-7.

[3] L. Breiman, "Random forests. Machine Learning," 2001. [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324.

[4] J. R. Quinlan, "Induction of Decision Trees. Machine Learning," 1986. [Online].

[5] F. e. a. Pedregosa, "Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research," 2011. [Online]. Available: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.