CPU5006-20: Artificial Intelligence:

Assessment 2: Machine Learning AI Scientific Research Paper

05th January 2025

GitHub Repository Link - https://github.com/RJay02/Artificial-Intelligence

Dataset Link - https://www.kaggle.com/competitions/titanic/overview

1. Introduction

There have been massive disasters throughout history, but the Titanic's disaster ranks as high as the depth at which it sank (Gupta and Saurabh, 2023). The RMS Titanic was a British passenger liner operated by the White Star Line, which, at the time of its maiden voyage on 10 April 1912, was promoted as the "unsinkable" pinnacle of 20th-century engineering and luxury. On April 15th, 1912, the Titanic sank when it struck an iceberg in the North Atlantic. Of the estimated 2,224 passengers and crew aboard the Titanic, over 1,500 died at the time the ship sank, making it one of the deadliest disasters in modern history. Yang Liu (2023). Even though there was a list of different reasons and even factors of luck influencing a person's survival during the tragedy, it is still presumed that according to socioeconomic status, gender, and class, some people were more likely to survive than others. (Singh et al., 2023)

For Example, first-class passengers had much higher survival rates than those in second and third classes, emphasising how great the divide between social classes was (Farag and Hassan, 2018). At the same time, ship evacuation priority for women and children reduced men's survival possibilities, especially within the classes of second and third priority (Li, Wang and Sung, 2008). These discrepancies can be further investigated using machine learning algorithms to determine survival probabilities from passenger attributes such as ticket class, fare paid, gender, and age. This paper similarly aims to conduct the same by developing two different predictive machine learning models and comparing the performance of these models to see which approach is more effective and, thus, better suited for this problem. We would utilise the Titanic dataset provided by Kaggle to build such models (Kaggle, 2025).

This dataset aligns well with our research because it allows for investigating real-world classification problems while comparing the effectiveness of different machine-learning approaches. This paper has used two supervised machine-learning techniques to predict survival outcomes. Logistic Regression and Random Forest were chosen as primary machine learning models for analysis because they have unique strengths. Logistic Regression finds the patterns in input features to predict continuous numerical values of the output variable (Chatterjee, 2017). This provides interpretability to show linear relationships between survival and predictor variables. At the same time, Random Forest, an ensemble model, is very efficient in capturing complex nonlinear interactions and highlighting the most influential features (Freeman, 2025). The Titanic Dataset is modelled using these algorithms, after which the efficiency and accuracy of these models are compared based on different unit metrics to see which of the two algorithms performs the best for this problem. Using these models, this research aims to show the main determinant factors for survival and demonstrate the efficiency of machine learning in historical data analysis.

2. Background & Literature Review

Extensive studies have been conducted on the Titanic catastrophe, especially in predictive analysis, where machine learning models have been used to classify survival chances depending on different passenger characteristics. Much research has been done over the years since the tragedy, and different classification algorithms have been evaluated to find the most suitable algorithm for the best survival predictions.

One of the prediction models is the Logistic regression algorithm. Chatterjee (2023) assessed numerous Linear Regression models on passenger survival; he reports that the logistics model outperforms these linear models by a maximum accuracy of 80.76%. This implies that Logistic Regression can be considered one of the foundational models in survival classification.

Random Forests and Decision Trees have also been studied vigorously and used in predictive modelling. When Dalta (2023) evaluated the use of the two algorithms, she established that Random Forest scored 81% while Decision Tress correctly labelled 84% of the predictive data. This suggests that although random forests, made using a combination of many decision trees, often perform better in capturing complicated connections in the dataset, decision trees tend to perform better when dataset complexity is low.

Other researcher's work involves using feature engineering to improve model accuracy. To enhance predictability, several other features have recently been introduced to the Titanic data set by the researchers for the machine-learning model: "child", "title", "family size", and "new fare" (Meyer et al., 2023). Surprisingly, their analysis showed that their feature-engineered attributes outperform conventional Titanic data set attributes, enhancing classification performance significantly.

Support Vector Machines, too, have been explored outside the tress-based models for the alternative classifier algorithm. Amongst the most reliable non-tree-based models is the SVM. For Ratsch et al., 2023, the assessment recorded an error rate of 22.4% when evaluated with the support vector machines (SVM). Similarly, Pea-Lei Tu and Jen Yao Chung, in 2023, introduced an improved decision tree classification algorithm, IDA, which considered global dependencies rather than local dependencies, producing a more efficient classification process.

On the other hand, Neural Networks have been used as an alternative deep learning model. Barhoom et al. (2023) used the application of ANNs on the Titanic dataset and obtained an impressive accuracy of 99.28%. While neural networks provide a much better predictive performance, they are hard to interpret. This is a problem in many real-world applications where feature importance is equally essential as achieving high accuracy.

Some researchers Singh et al., 2023. suggest that Titanic survival rates may be predicted primarily by gender using the well-accepted "women and children first" evacuation strategy. Other factors, like ticket numbers, family size, and port of embarkation, have also been shown to have a minor influence on survival results. Passengers with ticket values starting with "PC 1755," for instance, had a greater survival probability (Eric Lam & Tang, 2023). These results stress the need to choose relevant features to maximise the performance of machine learning models.

Singh et al. (2023) evaluated Logistic Regression, Decision Trees, Random Forests, Naïve Bayes, and Support Vector Machines, concluding that hyperparameter-tuned Decision Trees achieved the highest accuracy at 93.6%. However, Kshirsagar et al. (2023) found that Logistic Regression had an accuracy of 95%, indicating that even simpler models may yield high-performance results on datasets with strong linear relationships.

This review lays down the basis for the current study, which aims to compare the Logistic Regression and Random Forest models in predicting the survival outcomes of the Titanic. The current research will use those findings to develop more effective feature selection, optimising model performance for valid and interpretable results.

3. Methodology

3.1 Dataset Description

The dataset used for this study is the Titanic-Machine Learning from Disaster provided by Kaggle. When downloaded, the dataset is already pre-divided into two files: a train.csv and a test.csv. The test.csv file, meant for competition submissions, does not include survival labels. Therefore, for this study, I only considered the train.csv file, which includes details about 891 passengers, of which only 350 passengers survived, as shown in *Figure 1*. Also, there are 12 independent variables, other than the target variable, Survived, as shown in *Table 1*. The dataset also comprises a good mix of categorical, numerical, and ordinal features, making it suitable for machine learning research.

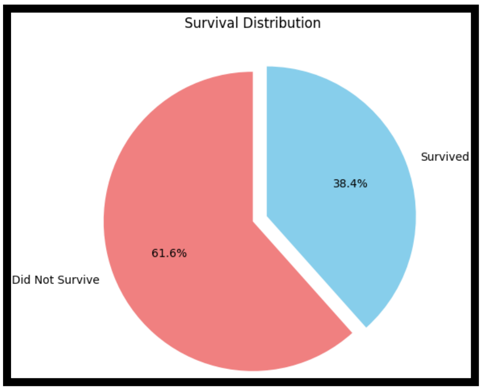

Figure 1: Survival Distribution of Titanic Passengers

Feature	Description	Value of Feature	Feature Characteristic
PassengerID	Identification no. of passengers.	1-891	Integer
Survived	Target variable (0 = perished, 1 = survived)	0,1	Integer
Pclass	Passenger class (1,2, or 3)	1-3	Integer
Name	Passenger Name	Name of Passengers	Object
Sex	Passenger Gender	Male, Female	Object
Age	Passenger age	0-80	Real
SibSp	Number of siblings or spouse on the ship	0-8	Integer
Parch	Number of parents or children on the ship	0-6	Integer
Ticket	Ticket Number	Ticket Number	Object
Fare	The price of the ticket	0-512	Real
Cabin	The cabin number of the passenger	Cabin Number	Object
Embarked	Port of embarkation (Southampton, Cherbourg, Queenstown)	S, C, Q	Object

Table 1: Overview of Titanic Dataset Features

3.2 Exploratory data analysis

After loading and reviewing our dataset in this research, the first step will be to perform exploratory data analysis. This process will allow us to inspect the dataset more closely and find key patterns and relationships. EDA helps find which feature impacts the survival rate most, find possible mistakes or inconsistencies in data, and check if any values are missing. This would help provide the necessary background for different preprocessing techniques and the selection of features to enhance performance for our machine learning models.

3.2.1 Identifying Missing Data

The first in our EDA was identifying any missing values within the dataset. Missing data can skew model performance and require preprocessing and cleaning if not appropriately handled.

```
Missing Values Per Column:
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Table 2: Missing Values Per Column in the Titanic Dataset

3.2.2 Categorical Features Analysis

1. **Survival distribution by Sex:**

Figure 2 shows the vast difference between the survival rates of males and females. Out of almost 468 males, only 109 survived, showing a survival rate of approximately 23.3% for males, while 233 females out of 314 showed a much higher survival rate of approximately 74.2%. This shows that gender played an important role, and passengers at the time on the Titanic respected and abided by the "women and children first" protocol.

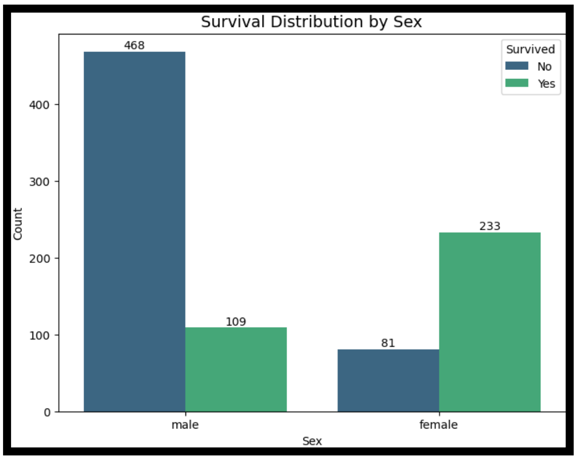

Figure 2: Survival Distribution by Sex

2. Survival Distribution by Passenger Class

The survival rate varied across the three-passenger class, with first-class passengers having the highest survival count, 136 out of 216, approximately 63%. Second-class passengers were seen to have near-equal survival distribution with 97 survivors, while third-class passengers had the lowest survival distribution with 372 non-survivors and only 119 survivors. This shows that money could buy anything in the world for some people.

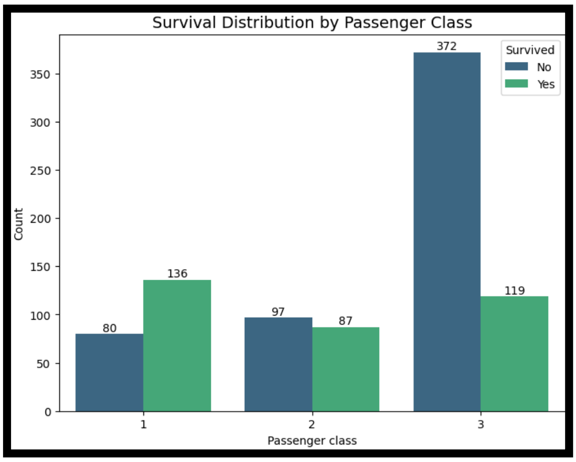

Figure 3: Survival Distribution by Passenger Class

3. Survival Distribution by Port of Embarkation

Passengers that were boarding from Southampton ("S") had the highest number of survivors (217) and non-survivors (427). For passengers embarking from Cherbourg ("C") and Queenstown ("Q"), the survivor-to-non-survivor distribution seems to be closely related as compared to Southampton, with passengers from Cherbourg having a survival rate of 55.4% while passengers from Queenstown having a survival rate of 38.9%.

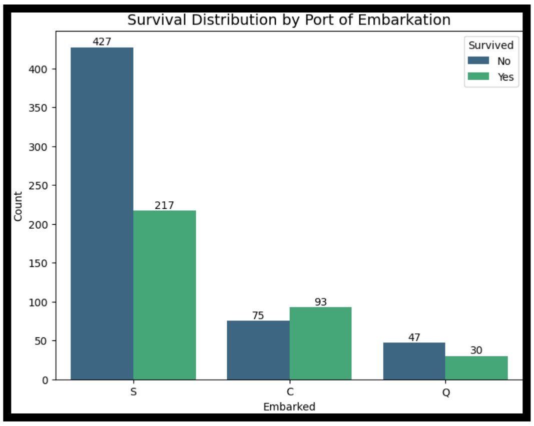

Figure 4: Survival Distribution by Port of Embarkation

3.2.3 Numerical Features Analysis

1. Age vs Survival

Data analysis shows that the median age for survivors and non-survivors is 28, indicating that age alone is not a strong predictor of survival. The age distribution is right-skewed, with most passengers aged between 20 and 30. Outliers are present, particularly among non-survivors. While age is not a decisive factor, it gains significance when combined with features like gender and class.

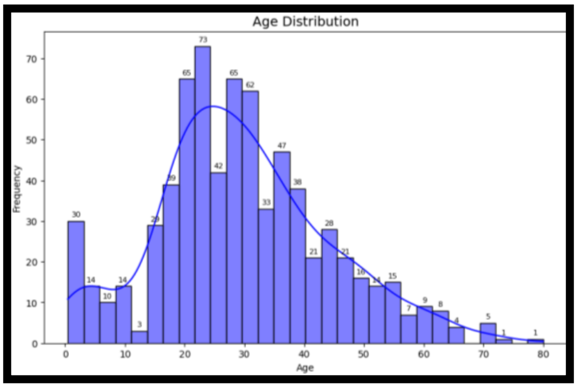

Figure 5: Age Distribution of Titanic Passengers

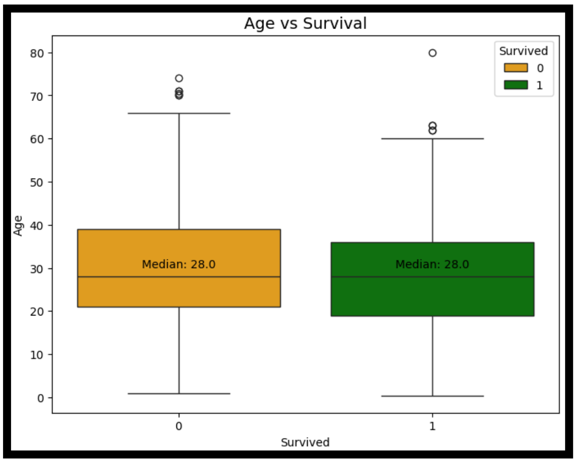

Figure 6: Age vs Survival Box Plot

2. Fare vs Survival

The fare distribution in the dataset is highly skewed, with most passengers paying lower fares. The histogram *below* indicates that over 500 passengers paid fares below 50, whilst only a tiny fraction paid over 100. This skewness reflects the socio-economic disparities among passengers.

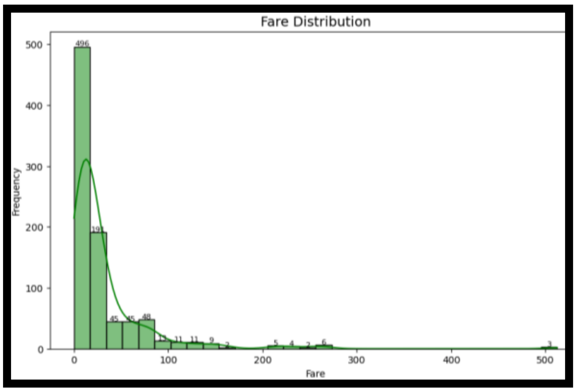

Figure 7: Fare distribution among passengers

The box plot *below* comparing fare with survival shows a noticeable difference between survivors and non-survivors. The median fare for survivors is 26, compared to 10.5 for non-survivors. Outliers in both groups indicate a wide range of fares, emphasising the socio-economic diversity on board.

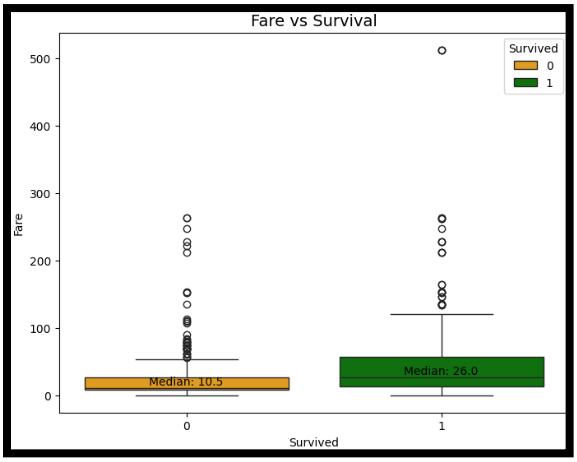

Figure 8: Fare vs Survival

3. Survival by Age and Passenger Class

The violin plot in *Figure 9* visualises the age distribution across passenger classes and survival status. First-class passengers show a broader age range among survivors, suggesting higher survival rates for people in this class regardless of age. This is contrasted by second and third-class passengers, who display more uniform distributions and survival skewed towards younger individuals.

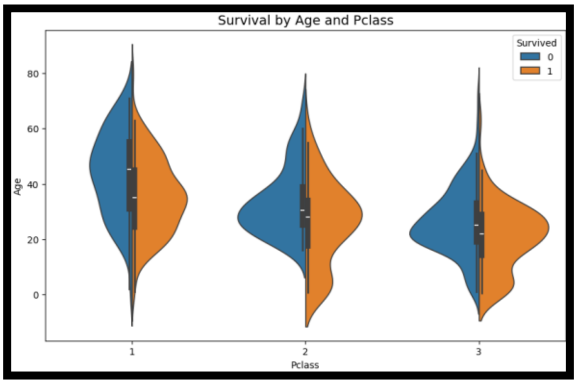

Figure 9: Violin plot showing survival distribution by age across passenger classes

3.3 Data Cleaning and Preprocessing

1. Data Correction: The Titanic dataset needed corrections to rectify errors and inconsistencies in the raw data. In the case of the Age column, the median age was calculated and inputted, as shown in *Table 3*.

Summary of corrections made:

	Title	Pclass	Fare	Age
0	Mr	3	7.2500	22.0
1	Mrs	1	71.2833	38.0
2	Miss	3	7.9250	26.0
3	Mrs	1	53.1000	35.0
4	Mr	3	8.0500	35.0

Table 3: Summary of Corrections Made

2. Data Completion: Missing values were addressed to achieve completeness and improve the dataset and its predictive power. The Fare column was carefully checked, and any missing entries were replaced with the median fare based on Pclass and Embarked, ensuring consistency. The data was reliable because it systematically addressed these gaps, enabling a robust foundation for developing the machine learning model.

3. Feature Engineering and Creation: Additional features were added to develop the predictive capability of the dataset by capturing the latent pattern and relationship. New variables like FamilySize were created to show whether a passenger travelled alone. An additional feature, Pclass_Fare_Interaction, represented the interrelationship between ticket class and fare on survival probabilities. Considering skew in the Fare column, LogFare transformation was done. The created features added depth to the data and made it complex for machine learning models to learn subtle interactions that would improve predictive accuracy.

4. Data Conversion: The dataset had to be prepared for machine learning algorithms by converting categories into numerical representations. Categorical variables include Sex, Embarked, and Title, all labelled and encoded in a format that can work with the models. Numerical variables include Age, Fare, and the created Pclass_Fare_Interaction, which needed scaling with StandardScaler to standardise their ranges. This made sure that all features contributed equally to model training without the dominance of variables with larger scales.

3.4 Model Training

Two supervised machine learning models, Logistic Regression and Random Forest, were selected for this study due to their distinct strengths and compatibility with the dataset. Logistic regression was chosen as a robust baseline model for its simplicity, interpretability, and effectiveness. At the same time, Random Forest was selected for its ensemble-based approach, which excels at capturing non-linear interactions and providing insights into feature importance. The train.csv dataset was split into training (80%) and testing (20%) subsets, and both models were trained on the pre-processed training data. Logistic Regression was implemented using default settings, while Random Forest was trained with 100 estimators and a fixed random seed to ensure reproducibility.

```
Training set shape: (712, 12)
Testing set shape: (179, 12)

Target distribution in training set:
Survived
0    0.616573
1    0.383427
Name: proportion, dtype: float64

Target distribution in testing set:
Survived
0    0.614525
1    0.385475
Name: proportion, dtype: float64
```

Figure 10: Dataset split

4. Results and Discussion

4.1 Overview of Model Performance

The research aimed to assess the prediction of passenger survival on the Titanic dataset using Logistic Regression and Random Forest. These two models were considered in multiple metrics: accuracy, precision, recall, F1-score, ROC-AUC, and error rates (MAE, RMSE). These criteria gave a full view of the strengths

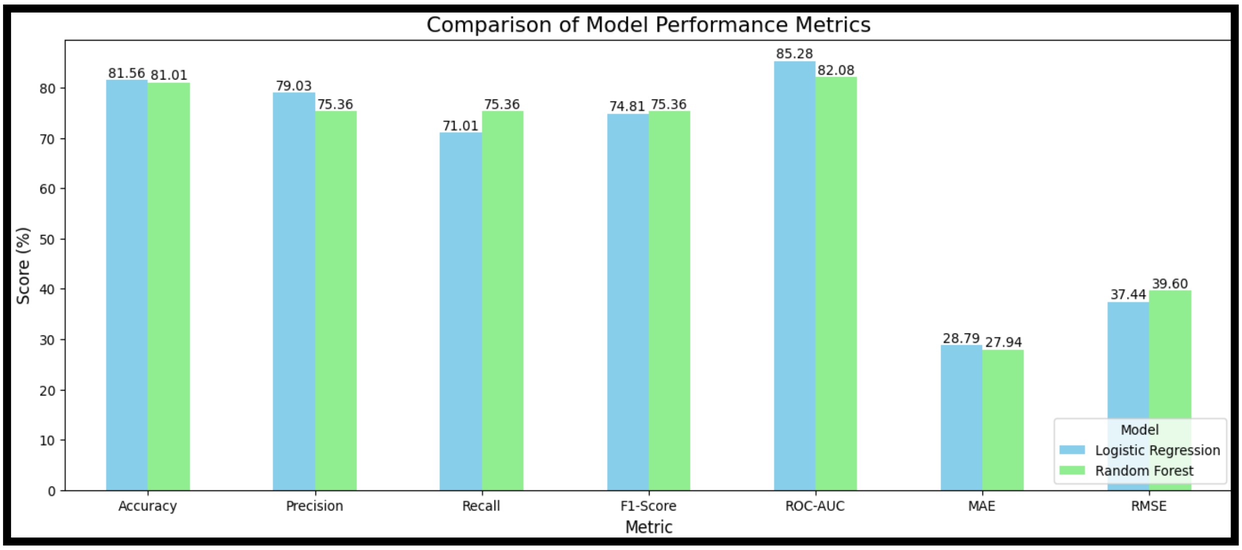

Figure 11: Bar chart comparing model performance metrics

and weaknesses of each model in classifying survivors.

According to *Table 4*, Logistic Regression performed slightly better, with an accuracy of 81.56%, followed by Random Forest at 81.01%. Both models gave relatively good results, though their comparative analysis showed significant differences in interpretability, error rates, and feature importance.

Metric	Logistic Regression (%)	Random Forest (%)
Accuracy	81.56	81.01
Precision	79.03	75.36
Recall	71.01	75.36
F1-Score	74.81	75.36
ROC-AUC	85.28	82.08
MAE	28.79	27.94
RMSE	37.44	39.60

Table 4: Performance metrics comparison

The higher ROC-AUC score of Logistic Regression (85.28%) suggests it is better at distinguishing survivors from non-survivors than Random Forest. However, Random Forest exhibited a slightly lower Mean Absolute Error (MAE), indicating a marginally improved prediction stability.

4.2 Key Findings and Feature Importance Analysis

Logistic Regression performed better in metrics like ROC-AUC, precision and accuracy than Random Forest, with interpretable coefficients showcasing linear relationships.

While Random Forest ran robustly on both recall and F1-score, it handled nonlinear relationships and feature interactions very well. Feature importance based on the Random Forest identified Sex, Age, Pclass_Fare_Interaction and FamilySize as strong predictors of survival. Among them, gender was the most influential factor. The Pclass and Fare further support the evidence of socio-economic differences; first-class passengers who paid higher fares had a better survival rate.

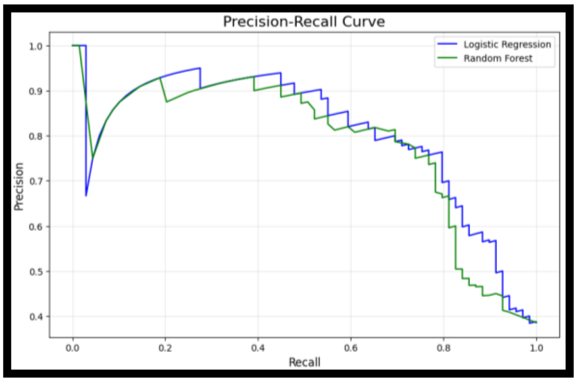

Figure 12: Precision-Recall curve

4.3 Model Comparison and Discussion

Logistic Regression and Random Forest models have different strengths, coming with their trade-offs. Logistic regression was excellent in interpretability and had precision at 79.03%; therefore, finding applications in domains where transparency of results is a key factor. It generally had a lower recall than

Random Forest, 71.01%, missing some actual survivors. On the other hand, Random Forest robustly handled a non-linearly interacting feature, showed higher recalls, 75.36%, and correctly targeted more survivors while sacrificing precision by the same, 75.36%.

The ROC curve analysis showed Logistic Regression with an AUC score of 85.28%, while Random Forest had a very competitive 82.08%. Logistic Regression is better at balanced classification problems, while the Random Forest approach is good at recall.

Error analysis highlighted misclassified cases, such as third-class passengers with high fares and lower-class children. Logistic Regression struggled with non-linear relationships, while Random Forest was less interpretable and prone to overfitting. Addressing these limitations through advanced models like Gradient Boosting or Neural Networks could enhance future predictions.

4.4 Comparison with existing research

These results are very similar to the previously conducted studies on the Titanic survival prediction problem. Singh et al. (2023) obtained a very impressive accuracy of 93.6% with Decision Trees, which outperformed Random Forest and Logistic Regression models. Similarly, Kshirsagar et al. (2023) identified the high reliability of Logistic Regression, obtaining as high as 95% accuracy. Our combined insight from Logistic Regression and Random Forest agrees with historical accounts of the Titanic disaster, putting together comprehensive insight into the factors that drive survival.

4.5 Limitations and Implications

The dataset's relatively small size and inherent biases, such as incomplete records and unrecorded factors (e.g., physical fitness, decision-making), may impact generalizability. Additionally, Logistic Regression assumes linear relationships, which can constrain its performance on non-linear data, while some latent factors influencing survival, such as behaviour during the disaster, remain unquantifiable despite feature engineering efforts.

Nonetheless, this research underscores the value of the Titanic dataset as a benchmark for testing predictive models and demonstrates the effectiveness of machine learning in analysing historical events. This contributes to the broader discourse on predictive modelling and its applications.

5. Conclusion

The RMS Titanic is a fascinating dataset that allows researchers to use modern techniques in machine learning with historical events. This study used logistic regression and random forest models to predict passenger survival based on socio-economic, demographic, and travel-related factors. Both models performed well and had strong correlating results aimed at serving different analytical needs.

The study has shown that feature engineering, scaling, and handling missing values are key to optimising performance in data preprocessing. It also points toward balanced model selection, where trade-offs between interpretability and robustness must be made considering the research objectives.

Although these findings provide beneficial insights into the determinants of survival on the Titanic, this study recognises several limitations regarding dataset size and unrecorded factors that may influence survival. Further work could be done on these findings by analysing additional datasets, using more advanced ensemble methods, or incorporating other contextual factors than those with the Titanic dataset.

Ultimately, this research shows how well machine learning can uncover hidden historical insights and provides a backbone for further studies where data science meets history. The methodologies and findings contribute toward a greater understanding of predictive modelling for structured data analysis and provide a benchmark for similar applications within diverse contexts.

References

1. **Kakde, Y., 2018.** *Predicting Survival on Titanic by Applying Exploratory Data Analytics and Machine Learning Techniques*. ResearchGate. Available at: https://www.researchgate.net/profile/Yogesh-Kakde/publication/325228831_Predicting_Survival_on_Titanic_by_Applying_Exploratory_Data_Analytics_and_Machine_Learning_Techniques/links/5c068f63a6fdcc315f9c0bb9/Predicting-Survival-on-Titanic-by-Applying-Exploratory-Data-Analytics-and-Machine-Learning-Techniques.pdf [Accessed 5 January 2025].
2. **Farag, N. & Hassan, G., 2018.** *Predicting the Survivors of the Titanic*. The British University in Egypt & Ain Shams University. Available at: https://d1wqtxts1xzle7.cloudfront.net/58810000/farag2018-libre.pdf [Accessed 5 January 2025].
3. **Li, X., Wang, L. and Sung, E., 2008.** AdaBoost with SVM-based component classifiers. *Engineering Applications of Artificial Intelligence*, 21, pp.785-795.
4. **Acun, N., 2018.** A Comparative Study on Machine Learning Techniques Using Titanic Dataset. *ResearchGate*. Available at: https://www.researchgate.net/profile/Neytullah-Acun/publication/324909545_A_Comparative_Study_on_Machine_Learning_Techniques_Using_Titanic_Dataset/links/607533bc299bf1f56d51db20/A-Comparative-Study-on-Machine-Learning-Techniques-Using-Titanic-Dataset.pdf [Accessed 5 January 2025].
5. **Shetty, J., Pallavi, S., & Ramyashree, B., 2023.** Predicting the Survival Rate of Titanic Disaster Using Machine Learning. *DCS DataPro*. Available at: https://dcs.datapro.in/wp-content/uploads/2023/11/3-Predicting-the-Survival-Rate-of-Titanic-Disaster-Using-machine-learning.pdf [Accessed 5 January 2025].
6. **Chatterjee, T., 2017.** Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms. *International Journal of Emerging Research in Management & Technology*, 6, pp.1-5.
7. **Köse, E. & Gökkaya, T., 2020.** *Survival Prediction on the Titanic Dataset Using Machine Learning Algorithms*. DergiPark. Available at: https://dergipark.org.tr/en/download/article-file/1256672 [Accessed 5 January 2025].
8. **Datla, M.V., 2015.** Bench Marking of Classification Algorithms: Decision Trees and Random Forests – A Case Study using R. In: *Proceedings of I-TACT15*, pp.1-7.
9. **Gupta, S. & Saurabh, S., 2023.** *Ensemble Approach for Titanic Survival Predictor*. ResearchGate. Available at: https://www.researchgate.net/profile/Swati-Gupta-71/publication/376530685_Ensemble_Approach_for_Titanic_Survival_Predictor [Accessed 5 January 2025].
10. **Meyer, D., Leisch, F. and Hornik, K., 2003.** The support vector machine under test. *Neurocomputing*, 55, pp.169-186.
11. **Ali, M., 2023.** *Survival Analysis Using Logistic Regression and Random Forests*. PhilPapers. Available at: https://philpapers.org/archive/ALHCOS.pdf [Accessed 5 January 2025].
12. **Yang Liu (2023).** *Casualty on the Titanic based on Machine Learning Methods*. ResearchGate. Available at: https://www.researchgate.net/publication/369869556_Casualty_on_the_Titanic_based_on_Machine_Learning_Methods [Accessed 5 January 2025].
13. **Bauer, E. and Kohavi, R., 1999.** An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning*, 36, pp.105-139.
14. **Liang, G., Zhu, X. and Zhang, C., 2011.** An empirical study of bagging predictors for different learning algorithms. In: *Proceedings of AAAI'11*, pp.1802-1803.
15. **Singh, A., Saraswat, S. & Faujdar, N., 2017.** *Analyzing Titanic disaster using machine learning algorithms*. In: *Computing, Communication and Automation (ICCCA), 2017 International Conference*. IEEE, pp. 406-411.

16. **Simpson, H., 2004.** *Dumb Robots*. 3rd ed. Springfield: UOS Press, pp. 6-9.
17. **Russell, S. & Norvig, P., 2015.** *Artificial Intelligence: A Modern Approach*. Pearson Education.
18. **Kaggle, 2025.** *Titanic - Machine Learning from Disaster*. [online] Available at: https://www.kaggle.com/competitions/titanic/data [Accessed 5 January 2025].
19. **Freeman, L.D., 2025.** *A Data Science Framework to Achieve 99% Accuracy*. [online] Kaggle. Available at: https://www.kaggle.com/code/ldfreeman3/a-data-science-framework-to-achieve-99-accuracy [Accessed 5 January 2025].
20. **Kshirsagar, V. and Phalke, N., 2019.** *Titanic Survival Analysis using Logistic Regression*. [online] *International Research Journal of Engineering and Technology (IRJET)*. Available at: https://d1wqtxts1xzle7.cloudfront.net/60580261/IRJET-V6I81520190913-115024-jrj0l4-libre.pdf [Accessed 5 January 2025].

References

1. **Kakde, Y., 2018.** *Predicting Survival on Titanic by Applying Exploratory Data Analytics and Machine Learning Techniques*. ResearchGate. Available at: https://www.researchgate.net/profile/Yogesh-Kakde/publication/325228831_Predicting_Survival_on_Titanic_by_Applying_Exploratory_Data_Analytics_and_Machine_Learning_Techniques/links/5c068f63a6fdcc315f9c0bb9/Predicting-Survival-on-Titanic-by-Applying-Exploratory-Data-Analytics-and-Machine-Learning-Techniques.pdf [Accessed 5 January 2025].
2. **Farag, N. & Hassan, G., 2018.** *Predicting the Survivors of the Titanic*. The British University in Egypt & Ain Shams University. Available at: https://d1wqtxts1xzle7.cloudfront.net/58810000/farag2018-libre.pdf [Accessed 5 January 2025].
3. **Li, X., Wang, L. and Sung, E., 2008.** AdaBoost with SVM-based component classifiers. *Engineering Applications of Artificial Intelligence*, 21, pp.785-795.
4. **Acun, N., 2018.** A Comparative Study on Machine Learning Techniques Using Titanic Dataset. *ResearchGate*. Available at: https://www.researchgate.net/profile/Neytullah-Acun/publication/324909545_A_Comparative_Study_on_Machine_Learning_Techniques_Using_Titanic_Dataset/links/607533bc299bf1f56d51db20/A-Comparative-Study-on-Machine-Learning-Techniques-Using-Titanic-Dataset.pdf [Accessed 5 January 2025].
5. **Shetty, J., Pallavi, S., & Ramyashree, B., 2023.** Predicting the Survival Rate of Titanic Disaster Using Machine Learning. *DCS DataPro*. Available at: https://dcs.datapro.in/wp-content/uploads/2023/11/3-Predicting-the-Survival-Rate-of-Titanic-Disaster-Using-machine-learning.pdf [Accessed 5 January 2025].
6. **Chatterjee, T., 2017.** Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms. *International Journal of Emerging Research in Management & Technology*, 6, pp.1-5.
7. **Köse, E. & Gökkaya, T., 2020.** *Survival Prediction on the Titanic Dataset Using Machine Learning Algorithms*. DergiPark. Available at: https://dergipark.org.tr/en/download/article-file/1256672 [Accessed 5 January 2025].
8. **Datla, M.V., 2015.** Bench Marking of Classification Algorithms: Decision Trees and Random Forests – A Case Study using R. In: *Proceedings of I-TACT15*, pp.1-7.
9. **Gupta, S. & Saurabh, S., 2023.** *Ensemble Approach for Titanic Survival Predictor*. ResearchGate. Available at: https://www.researchgate.net/profile/Swati-Gupta-71/publication/376530685_Ensemble_Approach_for_Titanic_Survival_Predictor [Accessed 5 January 2025].
10. **Meyer, D., Leisch, F. and Hornik, K., 2003.** The support vector machine under test. *Neurocomputing*, 55, pp.169-186.
11. **Ali, M., 2023.** *Survival Analysis Using Logistic Regression and Random Forests*. PhilPapers. Available at: https://philpapers.org/archive/ALHCOS.pdf [Accessed 5 January 2025].
12. **Yang Liu (2023).** *Casualty on the Titanic based on Machine Learning Methods*. ResearchGate. Available at: https://www.researchgate.net/publication/369869556_Casualty_on_the_Titanic_based_on_Machine_Learning_Methods [Accessed 5 January 2025].
13. **Bauer, E. and Kohavi, R., 1999.** An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning*, 36, pp.105-139.
14. **Liang, G., Zhu, X. and Zhang, C., 2011.** An empirical study of bagging predictors for different learning algorithms. In: *Proceedings of AAAI'11*, pp.1802-1803.
15. **Singh, A., Saraswat, S. & Faujdar, N., 2017.** *Analyzing Titanic disaster using machine learning algorithms*. In: *Computing, Communication and Automation (ICCCA), 2017 International Conference*. IEEE, pp. 406-411.
16. **Simpson, H., 2004.** *Dumb Robots*. 3rd ed. Springfield: UOS Press, pp. 6-9.
17. **Russell, S. & Norvig, P., 2015.** *Artificial Intelligence: A Modern Approach*. Pearson Education.
18. **Kaggle, 2025.** *Titanic - Machine Learning from Disaster*. [online] Available at: https://www.kaggle.com/competitions/titanic/data [Accessed 5 January 2025].
19. **Freeman, L.D., 2025.** *A Data Science Framework to Achieve 99% Accuracy*. [online] Kaggle. Available at: https://www.kaggle.com/code/ldfreeman3/a-data-science-framework-to-achieve-99-accuracy [Accessed 5 January 2025].
20. **Kshirsagar, V. and Phalke, N., 2019.** *Titanic Survival Analysis using Logistic Regression*. [online] *International Research Journal of Engineering and Technology (IRJET)*. Available at: https://d1wqtxts1xzle7.cloudfront.net/60580261/IRJET-V6I81520190913-115024-jrj0l4-libre.pdf [Accessed 5 January