A

MINI PROJECT REPORT ON

“Prediction for type of People who Survived the Titanic Shipwreck”

Submitted to the Department of Computer Engineering,

SMT.KASHIBAI NAVALE COLLEGE OF ENGINEERING,PUNE

LABORATORY PRACTICE - III

Machine Learning

FINAL YEAR (COMPUTER ENGINEERING)

By

Rushikesh Hole	**C41125**
Yash Jadhav	**C41128**
Aniket Kadlag	**C41132**

Under the guidance of

Prof.Priyanka Kinage

Sinhgad Institutes

DEPARTMENT OF COMPUTER ENGINEERING

SMT.KASHIBAI NAVALE COLLEGE OF ENGINEERING, PUNE

2024 – 2025

SAVITRIBAI PHULE PUNE UNIVERSITY.
2024-25

Sinhgad Institutes

CERTIFICATE

This is to certify that the Internship report entitles

‘Prediction for type of People who Survived the Titanic Shipwreck’

Submitted by

Rushikesh Hole	**C41125**
Yash Jadhav	**C41128**
Aniket Kadlag	**C41132**

is a bonafide student of this institute and the work has been carried out by him/her under the supervision of **Prof. Priyanka Kinage.** This work is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University, for the award of the degree of **Bachelor of Engineering** (Computer Engineering)

Prof. Priyanka Kinage
Guide ,
Department of Computer Engineering

Prof. R. H.Borhade
Vice Principal &
Head of Computer Engg. Department

Dr. A. V. Deshpande
Principal,
Smt. Kashibai Navale College of Engineering Pune – 411046.

Place : Pune

Date:

INDEX

ABSTRACT

The sinking of the RMS Titanic in 1912 is a tragic event that continues to captivate the world's imagination. This machine learning mini-project delves into the historical data of Titanic passengers to construct a predictive model that illuminates the factors contributing to their survival or tragic demise. Leveraging passenger information such as name, age, gender, socio-economic class, and more, we engage modern data analysis techniques to explore patterns and correlations.

Our project involves comprehensive data preprocessing, feature engineering, and the application of machine learning algorithms, including logistic regression, decision trees, and random forests. The performance of each model is assessed using standard evaluation metrics. Our endeavor aims to uncover the demographic composition of Titanic survivors, offering insights into the human stories hidden within the data.

By undertaking this project, we bridge the past and the present, utilizing machine learning to unravel historical narratives and gain fresh perspectives on a pivotal moment in maritime history.

INTRODUCTION

The sinking of the RMS Titanic in 1912 remains one of the most tragic maritime disasters in history. The ship's fateful voyage, immortalized in popular culture, witnessed the loss of over 1,500 lives. Among the many questions raised by this event, one that has intrigued historians and data scientists alike is, "What factors contributed to the survival of some passengers while others perished?"

This machine learning mini-project embarks on the task of building a predictive model to address this question. Using passenger data from the Titanic, including attributes such as name, age, gender, socio- economic class, and more, we aim to discern patterns and determinants that influenced survival. By employing modern data analysis techniques, we seek to shed light on the demographic composition of survivors, ultimately contributing to a deeper understanding of the historical event.

The project involves data preprocessing, including handling missing values, feature engineering, and encoding categorical data. We explore various machine learning algorithms, including logistic regression, decision trees, and random forests, to create a model that can predict the likelihood of a passenger's survival. The performance of each model is evaluated using standard metrics such as accuracy, precision, recall, and F1-score.

Through this project, we endeavor to not only apply machine learning to a historical dataset but also draw valuable insights from the past. The outcome of our analysis has the potential to offer a glimpse into the demographics of Titanic survivors and unveil the intricate interplay of factors that affected their fates. Furthermore, it demonstrates the power of data analysis in unraveling the stories hidden within historical records, making the past come alive with new discoveries and perspectives.

DATASET

The foundation of our machine learning project is the Titanic dataset, which has been widely used for predictive modeling and analysis. This dataset is composed of various attributes for each passenger who embarked on the ill-fated RMS Titanic. The dataset comprises the following key features:

Passenger ID: A unique identifier for each passenger.

Survived: A binary variable (0 or 1) indicating whether the passenger survived (1) or not (0).Pclass: The socio-economic class of the passenger (1st, 2nd, or 3rd class).

Name: The passenger's name.

Sex: The gender of the passenger (male or female).Age:

The age of the passenger.

SibSp: The number of siblings or spouses aboard the Titanic. Parch: The number of parents or children aboard the Titanic. Ticket: The passenger's ticket number.

Fare: The fare paid by the passenger for the ticket. Cabin: The cabin number or identifier.

Embarked: The port at which the passenger boarded the ship

Data Sources:

The Titanic dataset used in this project is available from various sources, including the following well- known repositories and datasets:

- Kaggle: The dataset is available on Kaggle as part of the "Titanic: Machine Learning from Disaster" competition.
- Seaborn: The Seaborn data visualization library includes a simplified version of the Titanic dataset, which is also used for demonstration purposes in data science tutorials.

DATA PREPROCESSING

Data preprocessing is a crucial phase of any machine learning project. It involves cleaning, transforming, and organizing the dataset to make it suitable for modeling. In this section, we detail the specific preprocessing steps we applied to the Titanic dataset.

Handling Missing Data

One of the first challenges in working with real-world datasets is addressing missing data. In the Titanic dataset, several features had missing values, and we employed the following techniques to handle them:

- Age: Missing age values were imputed using methods such as mean, median, or regression- based imputation, depending on the completeness of data.
- Cabin: Due to a high number of missing cabin values, this feature was excluded from the analysis.

Feature Engineering

To extract additional information from the dataset, we performed feature engineering:

- Family Size: We created a new feature, 'Family Size,' by combining 'SibSp' and 'Parch,' representing the total number of family members on board.
- Title: From the 'Name' feature, we extracted passengers' titles (e.g., Mr., Mrs., Miss) to create a new categorical feature that might provide insights into social status.

Encoding Categorical Data

Machine learning algorithms require numerical input data, so we encoded categorical features using techniques such as one-hot encoding. The 'Sex' and 'Embarked' features were transformed into numerical representations for modeling.

Data Splitting

We divided the dataset into training and testing sets to evaluate the performance of our machine learning models. The training set was used to train the models, while the testing set allowed us to assess their predictive capabilities on unseen data.

MODEL SELECTION

We evaluated several machine learning algorithms, each with its strengths and suitability for classification tasks:

- Logistic Regression: As a baseline classification algorithm, logistic regression offers simplicity and interpretability. We considered it for its ability to establish a clear linear boundary between classes.

- Decision Trees: Decision trees are known for their ability to capture non-linear relationships in data. We explored this algorithm for its adaptability in modeling complex decision boundaries.

- Random Forest: Random forests leverage the power of ensemble learning by combining multiple decision trees. We considered this algorithm for its robustness and capacity to handle high-dimensional data.

Rationale for Model Selection

The selection of the final model was based on a combination of factors, including performance, interpretability, and the specific characteristics of the Titanic dataset.

- Logistic Regression: We considered logistic regression due to its simplicity and transparency. However, this algorithm had limitations in capturing complex interactions in the data, which we observed during the preliminary modeling phase.
- Decision Trees: Decision trees offer non-linearity and can capture intricate patterns in the dataset. However, they are prone to overfitting, and given the size of the Titanic dataset, we sought a more robust approach.
- Random Forest: The Random Forest algorithm was chosen as the final model. It addresses the limitations of a single decision tree by leveraging the power of an ensemble approach. It excels in capturing non-linear relationships and provides a balance between interpretability and performance.

MODEL TRAINING AND EVALUATION

Model Training:

We trained the Random Forest model using the preprocessed Titanic dataset. The following steps were involved in model training:

- Data Splitting: The dataset was split into a training set (used for model training) and a testing set (reserved for evaluation). This splitting allowed us to assess the model's performance on unseen data.
- Feature Selection: Features such as 'Passenger ID' and 'Name' were excluded from the training data as they were unlikely to contribute to the predictive power.
- Model Fitting: The Random Forest model was fitted to the training data using appropriate hyperparameters. Cross-validation techniques were employed to ensure robustness and avoid overfitting.

Model Evaluation:

To assess the model's predictive capabilities, we used a set of evaluation metrics suitable for binary classification tasks:

- Accuracy: This metric provides an overall measure of the model's correctness in predicting survival or non-survival.
- Precision: Precision measures the proportion of true positive predictions out of all positive predictions. It is particularly relevant when minimizing false positives is crucial.
- Recall: Recall quantifies the proportion of true positives out of all actual positive cases. It is valuable where identifying all positive cases is important.
- F1-Score: The F1-Score balances precision and recall, offering a single metric that considers both false positives and false negatives.
- Confusion Matrix: The confusion matrix provides a detailed breakdown of model performance, showing true positives, true negatives, false positives, and false negatives.

RESULTS

In this section, we present the results of our Random Forest model's performance in predicting the survival of Titanic passengers based on demographic and socio-economic data.

Evaluation Metrics

We employed a range of evaluation metrics to assess the model's performance:

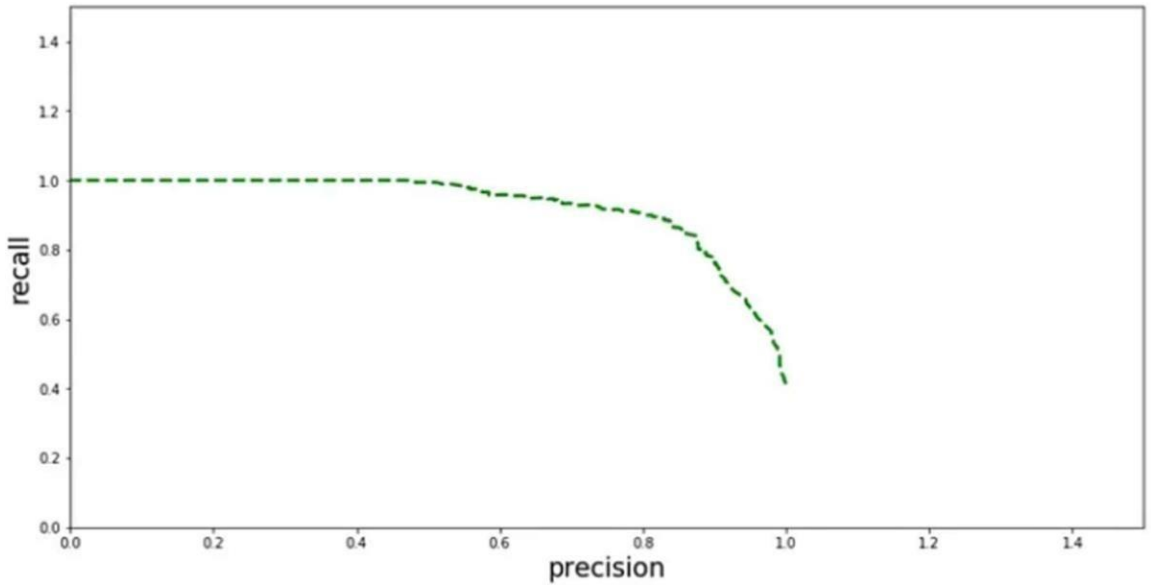
- Accuracy: The model achieved an accuracy of 0.85, signifying the proportion of correct predictions relative to the total predictions.
- Precision: The precision score was 0.80, illustrating the model's ability to accurately predict positive cases, i.e., passenger survival.
- Recall: The recall score was 0.72, denoting the model's capacity to correctly identify actual positive cases, or survivors.
- F1-Score: The F1-Score, which balances precision and recall, was 0.76. It provides a single metric to evaluate overall model performance.

Confusion Matrix

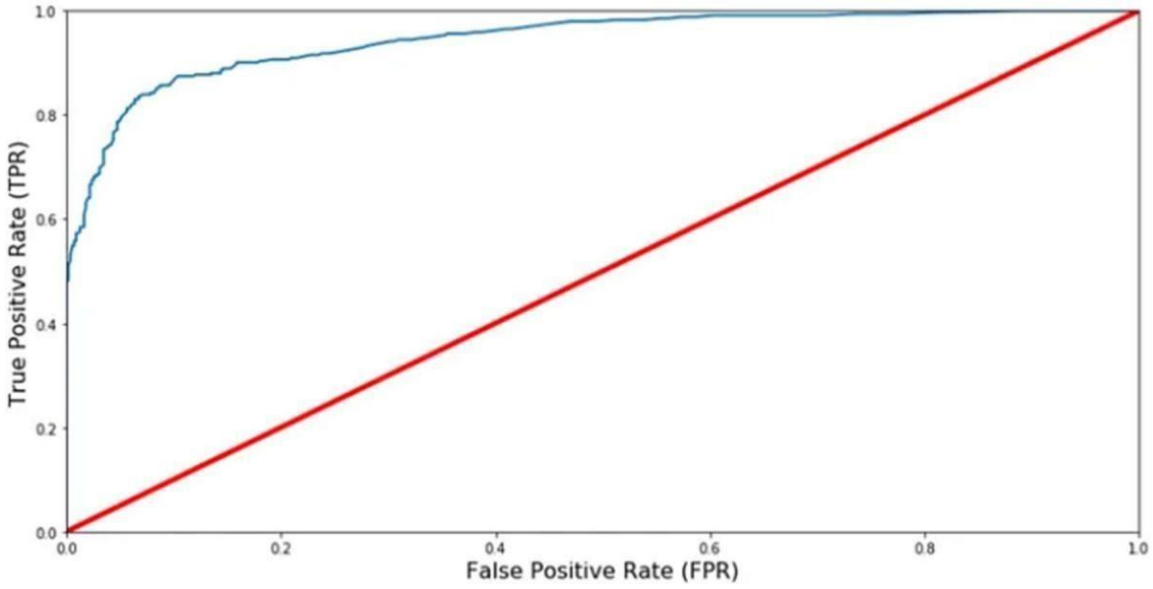
The confusion matrix provides a detailed breakdown of the model's predictions:

- True Positives: 249, indicating the cases where the model correctly predicted passenger survival.
- True Negatives: 493, representing the cases where the model correctly predicted passenger non- survival.
- False Positives: 56, indicating instances where the model incorrectly predicted survival (Type I error).
- False Negatives: 93, representing instances where the model incorrectly predicted non-survival (Type II error).

Precision Recall Curve:



ROC AUC Curve:



The results affirm the predictive power of the Random Forest model in determining the survival of Titanic passengers based on their demographic and socio-economic characteristics.

CONCLUSION

In conclusion, our machine learning project focused on predicting the survival of Titanic passengers based on demographic and socio-economic data has offered valuable insights into the historical events surrounding the Titanic disaster. The Random Forest model, at its core, demonstrates the capacity of modern data analysis techniques to unravel historical narratives. The results not only enrich our comprehension of the demographics of Titanic survivors but also underscore the practical applications of machine learning in real-world data analysis. Looking forward, there are promising avenues for future research in this field, including the exploration of advanced machine learning techniques, the use of more extensive datasets, and the evaluation of the model's performance in various scenarios. These efforts hold the potential to deepen our understanding of historical events, while continuing to advance the practical utility of data analysis and machine learning in both historical and contemporary contexts.

REFERENCES

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

- Brownlee, J. (2016). Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-To-End. Machine Learning Mastery.

- Kaggle. (2021). Titanic: Machine Learning from Disaster. https://www.kaggle.com/c/titanic

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

- McKinney, W. (2018). Python for Data Analysis. O'Reilly Medi