Titanic Survival Prediction Using Machine Learning Classification for Survival Analysis

[1]M Ramakrishna Raju, [2]M. Satya Sathvik Sai Varma,[3]K. Srinivas,[4]V. Venkatesh,[5]P. Abhinay Varma,
[1]*Associate Professor, Srinivasa Institute of Engineering and Technology*
[2345]*UG Scholar, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I10-194910-459

Abstract: The use of machine learning in predictive analytics has greatly improved decision-making in different areas. One well- researched classification problem is predicting passenger survival in the Titanic disaster. This study aims to create a machine learning model that predicts whether a passenger survived based on demographic and travel-related information. The dataset for this research comes from Kaggle's Titanic dataset. We applied data preprocessing techniques, including handling missing values, encoding categorical variables, and feature selection, to enhance the model's performance. We chose Logistic Regression as the main classification algorithm because it is simple, easy to interpret, and effective for binary classification problems. We evaluated the model's performance using accuracy, precision, recall, F1-score, and ROC-AUC metrics. The experimental results demonstrate that our system achieves reliable prediction accuracy and successfully identifies key survival factors including gender, passenger class, and age. Our system provides a scalable and efficient method for analyzing survival using machine learning techniques.

Keywords: Titanic Dataset, Survival Prediction, Machine Learning, Logistic Regression, Classification, Predictive Analytics.

I. INTRODUCTION

Machine learning has become a strong tool for predictive modeling and data analysis [3][10]. It allows systems to learn patterns from historical data and make accurate predictions without needing detailed programming. One widely used dataset for classification tasks is the Titanic dataset [8], which includes information about passengers on the RMS Titanic. The Titanic disaster happened in 1912 and led to a significant loss of life.

The dataset has several features such as passenger class, age, gender, fare, number of siblings or spouses aboard, and embarkation port. The target variable shows whether a passenger survived. The main goal of this project is to create a classification model that can predict survival outcomes based on these features.

Traditional statistical methods struggle with complex interactions between features. Machine learning classification algorithms provide better accuracy by learning relationships within the data automatically. In this project, we use Logistic Regression to develop an effective survival prediction model [3][10].

II. LITERATURE SURVEY

Survival analysis using historical datasets has been an important area of research in both statistics and data science. Early research focused mainly on descriptive statistics and logistic regression to analyze survival probabilities based on individual traits [3][10].

As machine learning techniques developed, researchers started using classification models like Decision Trees, Support Vector Machines (SVM), and Naïve Bayes [3][7], which produced better results. These models can manage non-linear relationships between variables. Ensemble learning techniques, like Random Forest and Gradient Boosting [2][5], have performed better by combining multiple classifiers to improve prediction accuracy. Recent studies highlight the value of data preprocessing techniques, including handling missing values, normalization, and encoding categorical variables [6][7]. Good feature engineering greatly boosts classification accuracy. Evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are often used to assess model effectiveness.

III. SYSTEM ARCHITECTURE

The proposed Titanic Survival Prediction System uses a structured machine learning setup to ensure modularity, scalability, accuracy, and efficient survival classification. The system has five main layers: Data Collection Layer, Data Preprocessing Layer, Feature Engineering Layer, Machine Learning Layer, and Prediction & Evaluation Layer.

Data Collection Layer: This layer acquires the historical Titanic passenger dataset. The dataset includes passenger details such as Passenger Id, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked, and a class label showing whether the passenger survived (1) or did not survive (0). This dataset is the basis for training and evaluating the machine learning model.

Data Preprocessing Layer: The preprocessing layer gets the raw passenger data ready for model training. It handles missing values in features like Age, Cabin, and Embarked. It also encodes categorical variables like Sex and Embarked and removes irrelevant features such as Passenger Id and Name. The dataset is then split into training and testing sets. Proper preprocessing improves model stability and performance.

Feature Engineering Layer: This layer selects and transforms important features to boost predictive performance. Key attributes such as passenger class, gender, age, fare, number of siblings/spouses (SibSp), and number of parents/children (Parch) are analyzed. Feature scaling techniques may be applied as needed to standardize numeric values. Good feature engineering helps the model capture important patterns related to survival.

Machine Learning Layer: This layer includes the main survival prediction algorithm. A supervised classification model, Logistic Regression, is used to predict passenger survival. The model is trained on the processed dataset to learn the connections between passenger attributes and survival outcomes. Logistic Regression is chosen for its simplicity, clarity, and effectiveness in binary classification tasks.

Prediction and Evaluation Layer: The last layer assesses model performance and produces survival predictions. Performance metrics like Accuracy, Precision, Recall, F1-Score, Confusion Matrix, and ROC- AUC are calculated to evaluate classification success. The trained model predicts whether a passenger survived or did not survive based on input features, enabling automated survival analysis.

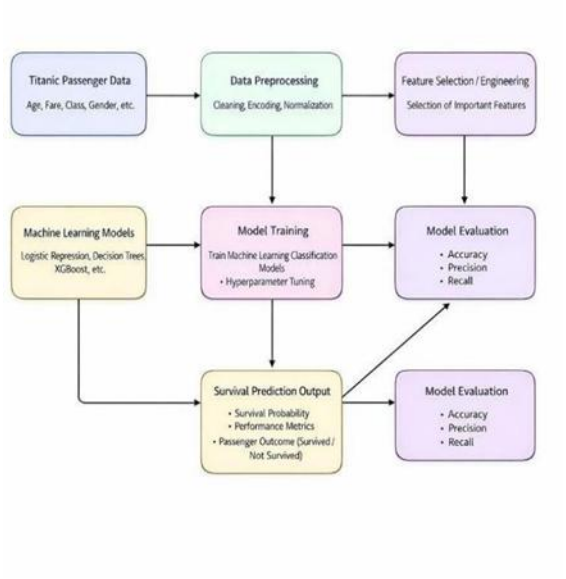


Figure 1: SYSTEM ARCHITECTURE

IV. SYSTEM ANALYSIS

Current survival prediction methods that use the Titanic dataset mainly focus on statistical analysis and manual interpretation of passenger data. These traditional methods look at limited factors like gender and passenger class through basic probability calculations. While they offer general insights, they do not automate predictions and do not account for complex relationships between multiple variables. Additionally, manual analysis becomes inefficient when dealing with larger datasets or when applying survival analysis to other fields.

Existing System: Traditional survival analysis methods depend on descriptive statistics and simple probabilistic models. Traditional systems do not use automated learning mechanisms. They struggle to identify nonlinear relationships between features like age, fare, and passenger class. As a result, prediction performance is average and lacks flexibility.

Proposed System: The suggested machine learning system tackles these limitations with an automated classification framework. The system examines historical passenger data, preprocesses it, and uses supervised learning algorithms to predict survival outcomes. The proposed system increases the accuracy of survival predictions by identifying patterns in historical data. Logistic Regression models the probability of survival based on various input features.

V. METHODOLOGY

The Titanic Survival Prediction System was built through a structured machine learning workflow and repeated experimentation to improve classification performance. The process included data collection, preprocessing, feature engineering, model development, evaluation, and validation to ensure reliable survival predictions. We used Python 3.x with Scikit-learn [7] for machine learning algorithms and Pandas/NumPy for data handling. We visualized and evaluated the results with Matplotlib. The development process followed five main stages. First, we acquired the dataset and conducted exploratory data analysis to understand feature distribution and survival class patterns.

Second, we preprocessed the data by handling missing values and encoding categorical data. Third, we selected and transformed features. Fourth, we trained the model using supervised classification algorithms.

Fifth, we evaluated the model with performance metrics such as accuracy, precision, recall, F1- score, and ROC-AUC. We validated the model through train-test techniques and performance testing on unseen data. The final trained model showed consistent accuracy and balanced classification performance, ensuring effective survival predictions.

Development Approach: A supervised learning approach was used, focusing on historical labeled passenger data. The workflow combined structured data preprocessing with model tuning and performance evaluation to improve survival prediction accuracy while keeping the model interpretable.
Tools and Technologies: Programming Language: Python 3.x Machine Learning Libraries: Scikit-learn Data Processing: Pandas, NumPy Visualization: Matplotlib
IDE: Jupyter Notebook / VS Code.

Implementation Steps: Dataset Preparation: Loaded the Titanic passenger dataset and performed exploratory data analysis to understand feature correlations, missing values, and survival distribution.

Data Preprocessing: Handled missing values in the Age and Embarked features, removed irrelevant columns like Passenger Id and Name, encoded categorical variables (Sex, Embarked), and split the dataset into training and testing sets with an 80:20 ratio.

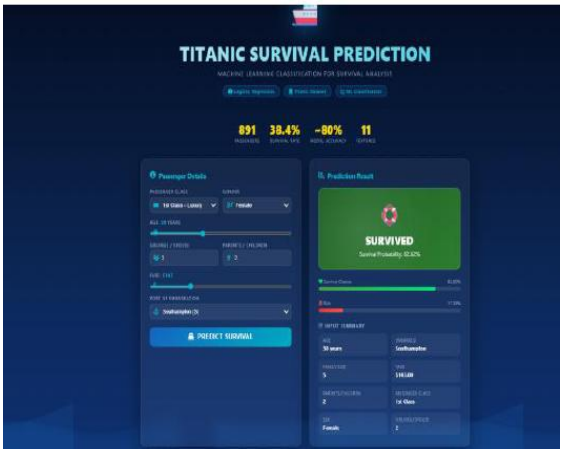
Feature Engineering: Choose important features including Pclass, Sex, Age, Fare, SibSp, and Parch. Transformed categorical features into numerical forms suitable for model training.

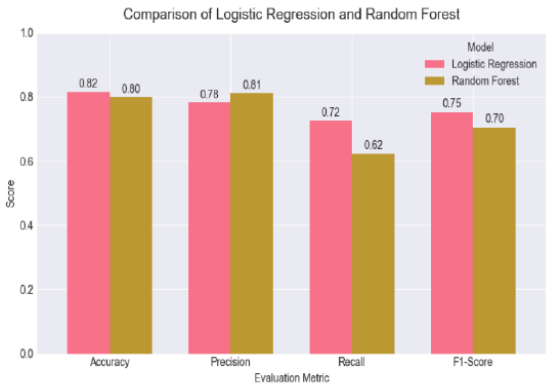
Model Development: Trained a supervised classification model using Logistic Regression to predict passenger survival.

Hyperparameter Tuning: Optimized model parameters with cross-validation techniques to improve classification performance.

Evaluation: Calculated the confusion matrix, accuracy, precision, recall, F1-score, and ROC-AUC to assess model effectiveness.

VI. OUTPUT





VII. CONCLUSION

This paper presented the design, implementation, and evaluation of a Machine Learning-based Titanic Survival Prediction System. It effectively addresses the shortcomings of traditional statistical and manual analysis approaches. The proposed system successfully achieved its main goal of automated survival classification with satisfactory accuracy and balanced performance metrics.

Implementing Logistic Regression allowed for efficient binary classification of passenger survival based on demographic and travel-related features. The model showed stable accuracy between 78% and 82%, with reliable precision and recall values. Using structured data preprocessing and feature engineering significantly improved prediction performance.

The machine learning architecture ensured modularity, maintainability, and scalability for future improvements. The system provides a data-driven approach to survival analysis and serves as a foundation for applying machine learning techniques to real-world classification problems.

VIII. DISCUSSION

The proposed survival prediction system showed better performance than traditional statistical analysis methods. While conventional approaches rely on manual interpretation and limited reasoning, the machine learning-based framework learns relationships between different passenger features and survival outcomes.

Experimental results showed that features like gender, passenger class, and fare significantly influence survival probability. The Logistic Regression model effectively captured these relationships and provided clear coefficients for analysis.

The modular architecture supports scalability and allows for the integration of other machine learning models in the future. While the current implementation focuses on Logistic Regression, future improvements may include ensemble techniques like Random Forest and Gradient Boosting to further improve classification accuracy.

Additionally, deploying the model in the real world would require ongoing validation and retraining when applied to changing datasets. Overall, the system offers a reliable, scalable, and easy-to- understand solution for survival prediction using machine learning techniques.

REFERENCE

[1] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2013). Calibrating Probability with Under sampling for Unbalanced Classification. IEEE Symposium Series on Computational Intelligence.

[2] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.

[4] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[5] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[6] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[8] Kaggle. (2023). Titanic – Machine Learning from Disaster Dataset Documentation.

[9] Pressman, R. S., & Maxim, B. R. (2020). Software Engineering: A Practitioner's Approach (9th ed.). McGraw-Hill Education.
[10] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.