COMPLEX COMPUTING PROBLEM 1.0

Titanic Dataset Survival Analysis & Prediction System

PROJECT REPORT

BS Artificial Intelligence, 3RD Year, 5TH Semester

Developed by:

Muhammad Waqas Ali
Roll# 23F-BSAI-59
Section: A – 1
Year: 3RD
Course: Machine Learning (2+1)
Semester: 5TH

Project Evaluator:

Engr. Humza Farooqui

Lecturer

Department of Artificial Intelligence

DAWOOD UNIVERSITY OF ENGINEERING & TECHNOLOGY

EXECUTIVE SUMMARY

This project presents a comprehensive machine learning–based approach for predicting passenger survival using the historical Titanic dataset. The primary objective is to analyze passenger attributes, apply effective data preprocessing strategies, and evaluate multiple supervised learning models to identify the most reliable predictive solution. The study emphasizes not only model accuracy but also robustness, generalization, and real-world applicability.

The dataset was systematically explored and preprocessed to ensure data quality and analytical consistency. Key preprocessing steps included handling categorical variables through one-hot encoding, scaling continuous features to improve model performance, and performing exploratory data analysis to uncover underlying patterns and relationships within the data. Visual and statistical analyses revealed strong correlations between survival outcomes and factors such as passenger class, gender, age, and fare.

Three supervised learning algorithms: Decision Tree, Naïve Bayes, and Support Vector Machine were implemented and evaluated using cross-validation to ensure reliable performance comparison. To further enhance predictive stability and reduce overfitting, an ensemble learning technique based on Bagging with Decision Trees was employed. Model performance was assessed using multiple evaluation metrics, including accuracy, precision, recall, F1-score, and confusion matrices.

The experimental results demonstrate that ensemble learning significantly improves predictive consistency compared to individual classifiers. The Bagging-based model achieved superior generalization performance, making it the most suitable model for this classification task. To support practical usability, the final model was saved and integrated into a Streamlit-based web application, enabling interactive and real-time survival predictions.

Overall, this project illustrates an end-to-end machine learning workflow, covering data exploration, preprocessing, model comparison, ensemble learning, and deployment readiness. The outcomes highlight the effectiveness of combining systematic data preparation with ensemble techniques to build reliable and scalable predictive systems suitable for real-world applications.

ABSTRACT

This project focuses on predicting survival outcomes of passengers aboard the Titanic using machine learning techniques. The data set consists of passenger demographic and travel information, including class, age, gender, and fare. Multiple models, including Decision Tree, Naïve Bayes, Support Vector Machine (SVM), and a Bagging Ensemble, were trained to handle noisy real-world data. Performance was evaluated using cross-validation and test set accuracy. The Bagging Ensemble of Decision Trees demonstrated superior generalization, achieving the highest predictive accuracy while mitigating overfitting. This study illustrates how ensemble learning enhances model robustness and reliability for classification tasks on structured datasets.

PROBLEM STATEMENT

Survival prediction represents a classic binary classification problem that closely resembles real-world decision-making scenarios. The primary problem addressed in this project is the prediction of passenger survival outcomes based on demographic and travel-related attributes. Given historical passenger data, the challenge is to accurately classify whether a passenger survived or did not survive the event using supervised machine learning techniques.

BACKGROUND

The sinking of the RMS Titanic in 1912 remains one of the most infamous maritime disasters in history, leading to significant loss of life. Over 1,500 passengers and crew perished that fateful night. Understanding the factors that contributed to survival can provide valuable insights into safety protocols and social dynamics during crises. In this project, we will leverage machine learning techniques to predict the survival chances of Titanic passengers based on various features, such as sex, age, and passenger class. Using the Random Forest classification algorithm, we aim to build a predictive model that will allow us to estimate the likelihood of survival for each individual aboard the Titanic.

OBJECTIVE

The primary objective of this project is to develop a machine learning model capable of predicting the survival status of Titanic passengers based on available data. By analyzing the features of the dataset provided, we seek to identify patterns that could influence survival rates and subsequently use these insights to make predictions on unseen data.

DATASET ANATOMY:

The Titanic dataset is publicly available from Kaggle and other open-source repositories, containing structured passenger information.

Dimensions:

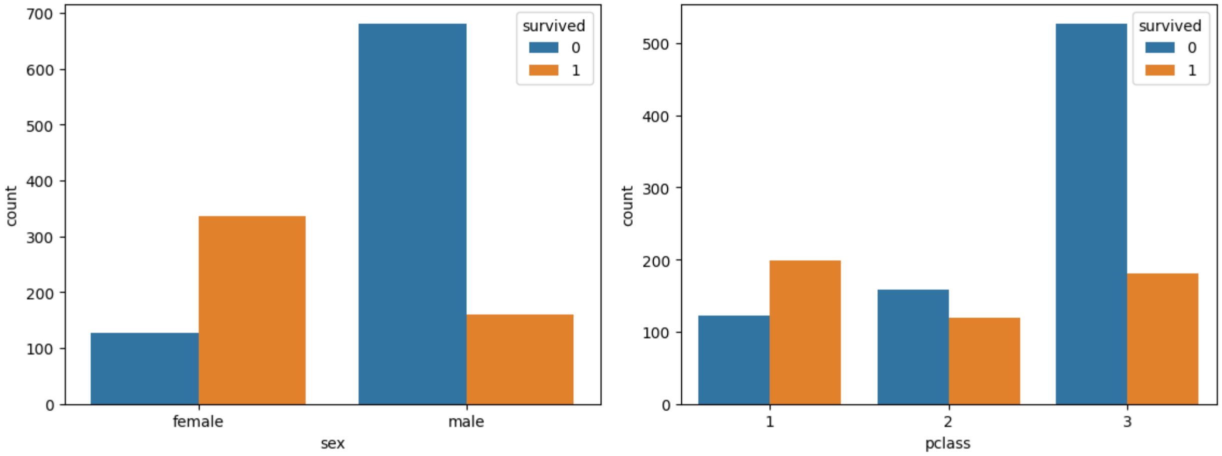
- **Number of records:** 891
- **Number of features:** 12
- **Target variable:** survived (0 = did not survive, 1 = survived)

Data Characteristics:

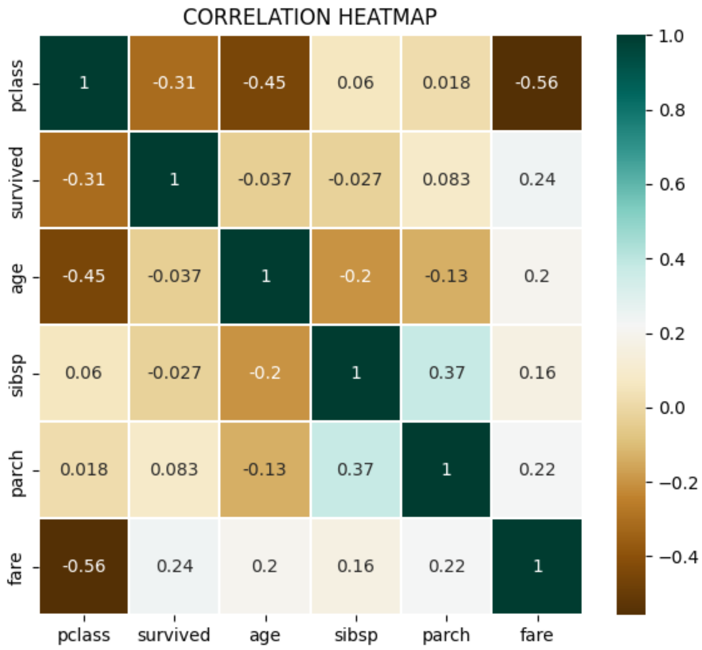
- **Categorical features:** sex, embarked, pclass
- **Numerical features:** age, fare, sibsp, parch
- Includes missing values in age and embarked.

Visualizations:

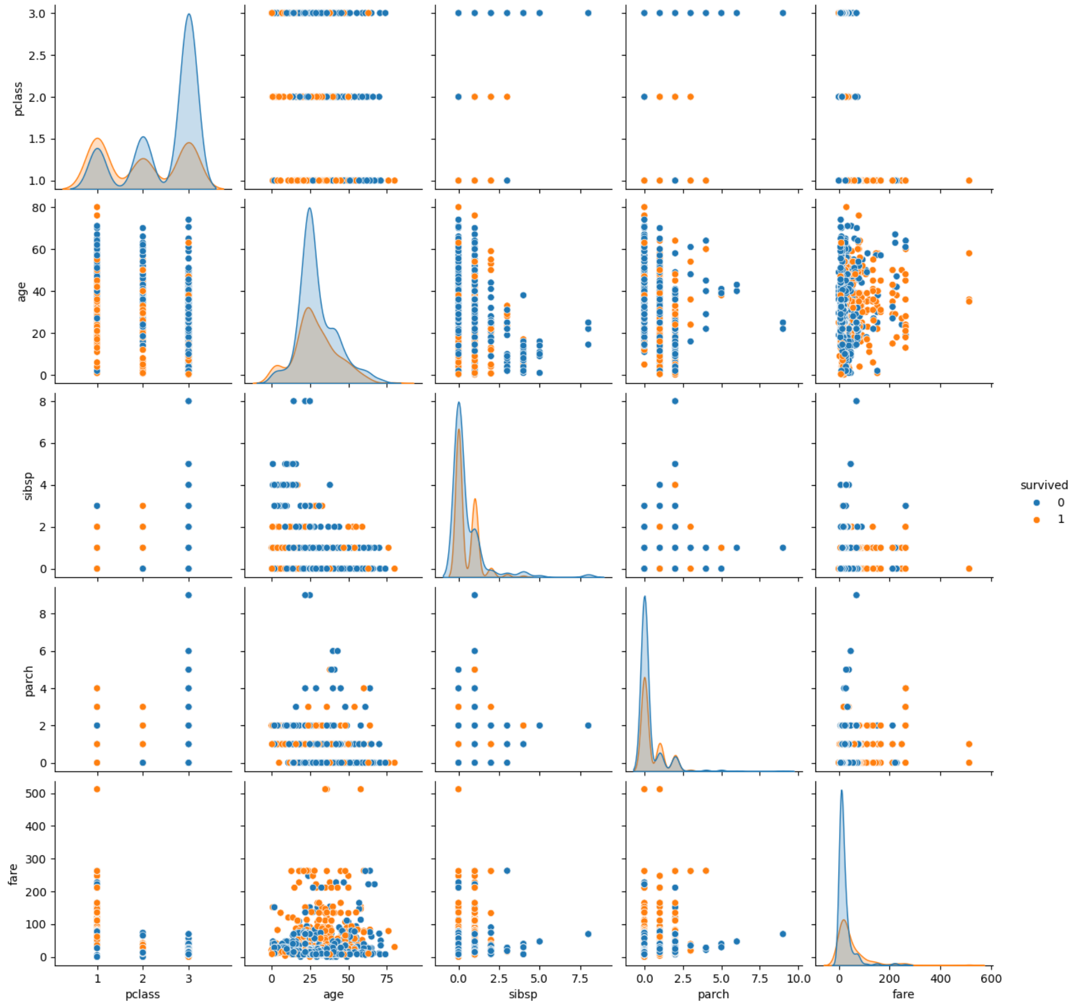
- Countplots for categorical features illustrate survival distribution.



- Heatmaps highlight correlations between numerical variables.



- Pairplots show feature interactions and class separation trends.



DATA PREPROCESSING & FEATURE ENGINEERING

Proper preprocessing is critical to ensure reliable model performance. The following steps were applied:

1. **Handling Missing Values:**
   - ➢ The ‘age’ and ‘sex’ feature had missing values, which were imputed using the median of each group to preserve distribution.
   - ➢ Rows with missing fare & embarked values were truncated because these columns contain 1% or 1.5% missing values.
   - ➢ Few features like 'body', 'boat' & 'home.dest' were dropped because these columns contain more than 60% of rows were empty.
2. **Removing Irrelevant Features:**
   - ➢ Columns such as name, ticket, and cabin were excluded, as they offered minimal predictivity as well as high cardinality.
3. **Encoding Categorical Variables:**
   - ➢ sex and embarked were transformed into numerical representations using One-Hot Encoding.
   - ➢ No categories were dropped (drop_first=False) to avoid losing critical information, especially for features with only 2–3 categories.
4. **Feature Scaling & Normalization:**
   - ➢ Numerical features age and fare were scaled using StandardScaler() to standardize ranges, improving model convergence and comparability.
5. **Dataset Split:**
   - ➢ The data was partitioned into training (80%) and testing (20%) sets, ensuring unbiased evaluation.

MODEL SELECTION & METHODOLOGY

Four models were selected based on their characteristics:

1. **Decision Tree:**

   Decision Tree was selected because it can handle both numerical and categorical features in parallel by providing interpretable decision paths but may overfit.

2. **Naïve Bayes:**

   Naïve Bayes is a probabilistic classifier and assumes independence: efficient & fast.

3. **Support Vector Machine (SVM):**

   SVM performs effectively with high-dimensional space as it uses RBF kernel to capture non-liner patterns.

4. **Bagging Ensemble (Decision Tree Base):**

   Since Decision Tree alone can overfit. However, Bagging Decision Trees combine multiple Decision Trees to reduce overfitting. It enhances robustness & generalization on unseen/test data.

Justification:
Models were chosen to compare simple interpretable approaches (Decision Tree, Naïve Bayes) against more complex classifiers (SVM, Bagging). Ensemble learning was included to demonstrate improvements in accuracy and stability.

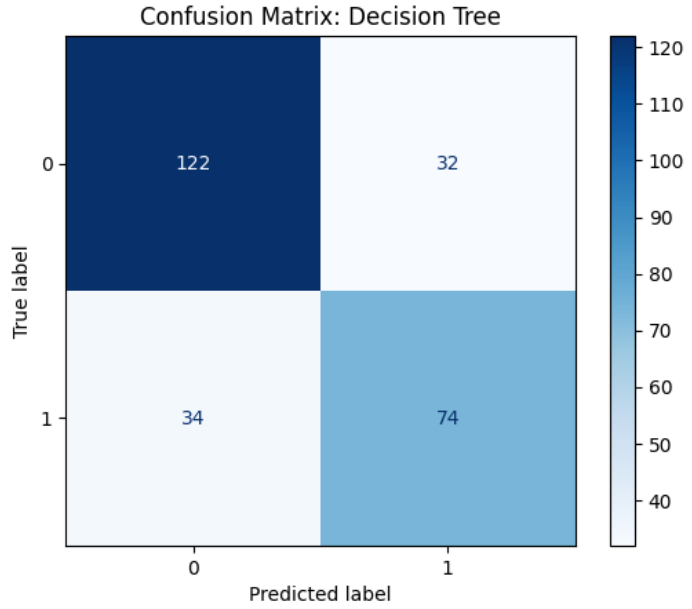
ACCURACY EVALUATION & MODEL OPTIMIZATION

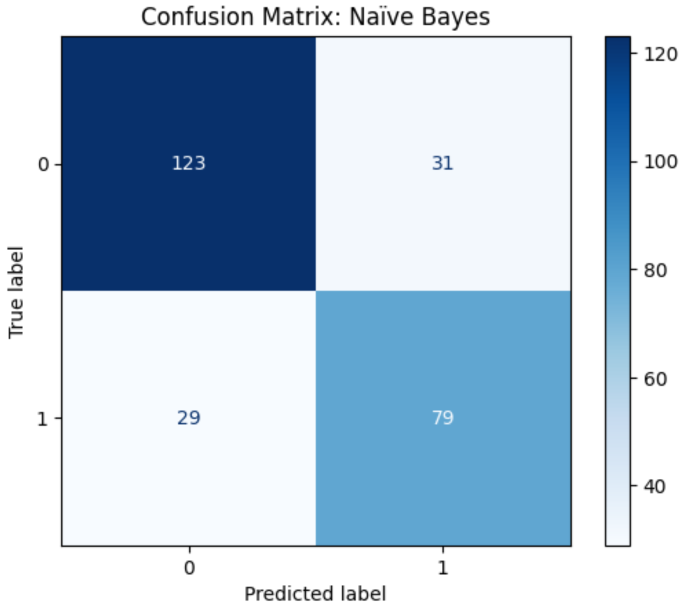
Evaluation Strategy:

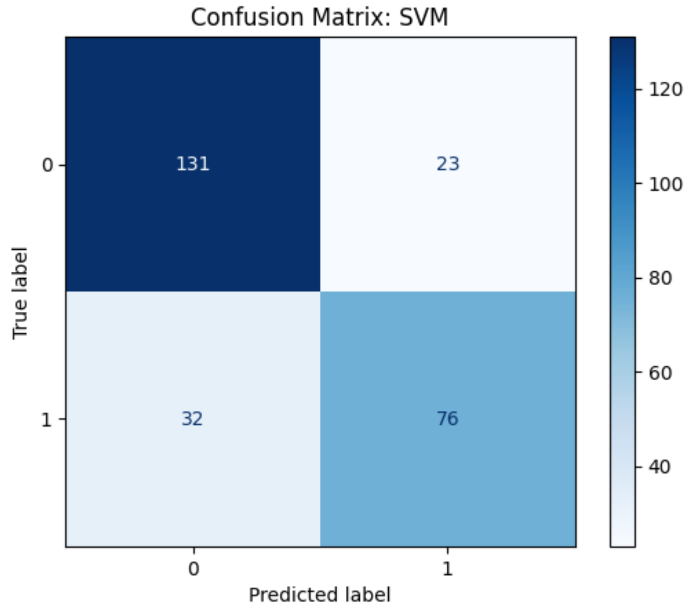
- 5-fold cross-validation on the training set provided robust accuracy estimates.
- Test set evaluation ensured generalization performance.

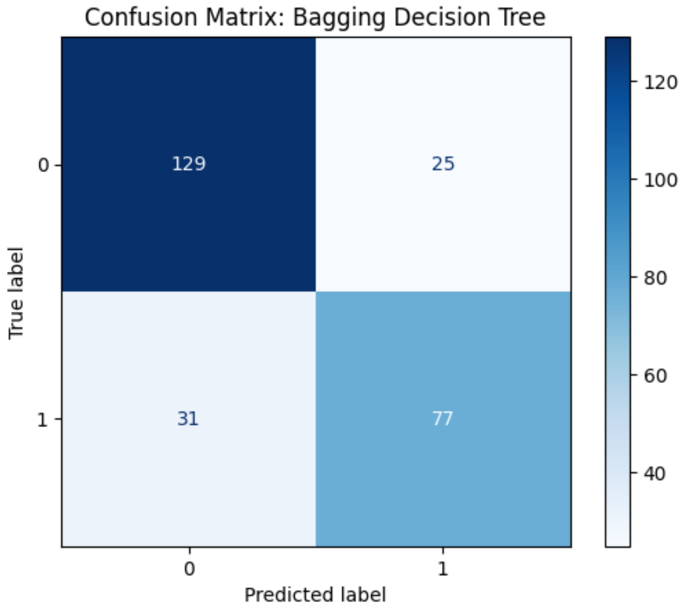
Performance Metrics:

- Accuracy, Precision, Recall, and F1-score were used to assess classifier performance.
- Confusion matrices provided detailed error analysis.









Results Summary:

- Bagging Ensemble achieved the highest accuracy (≈0.84) with balanced precision and recall.
- SVM performed well but required scaling and parameter tuning.
- Decision Tree exhibited slight overfitting; Naïve Bayes underperformed on mixed feature types.

MODEL COMPARISON AND JUSTIFICATION

Model	Training Accuracy %	Test Accuracy %	Overfitting Risk
Decision Tree	97.03	74.81	Medium+
Naïve Bayes	78.16	77.10	Low
SVM	82.09	79.01	Low
Bagging Decision Tree	78.63	77.87	Minimal

KEY INSIGHTS

- Passenger class, age, and sex were most predictive of survival.
- Single models like Decision Tree are prone to overfitting; ensembles mitigate this risk.
- Feature scaling and encoding significantly influence model performance.
- Cross-validation is essential to evaluate real-world model robustness.
- Bagging enhances model generalization and reduces variance on unseen data.

CONCLUSION

The Titanic Survival Prediction project demonstrates the effective application of supervised learning and ensemble methods on structured datasets.

Key takeaways:

- Rigorous preprocessing, including handling missing values and encoding categorical variables, is foundational.
- Ensemble learning, specifically Bagging with Decision Trees, provides superior accuracy and stability.
- Model evaluation using multiple metrics and cross-validation ensures reliable insights.

Final Model Selection:

Bagging Ensemble of Decision Trees is recommended for deployment due to its balanced performance and resistance to overfitting.

GitHub Repository URL: 🔗 https://github.com/MuhammadWaqasAli2211/Titanic.....

Live Demo: 🔗 https://titanic-dataset-survival-analysis-prediction-system-ml-project.....