****

**Assessment Report**

on

**“Titanic Survival Prediction”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI)**

By

Name : Ayushman Bhowmik Roll Number : 202401100300089

Name : Harshit Singh Patel Roll Number : 202401100300118

Name : Harshita Kumari Roll Number : 202401100300119

Name : Harshita Sharma Roll Number : 202401100300120

Section: A

**Under the supervision of**

“Sir Shivansh Prasad”

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

**1. Introduction**

The Titanic disaster remains one of the most infamous maritime tragedies in history. Predicting the survival of passengers based on various features such as age, sex, ticket class, and fare using machine learning is a classical problem in data science. This project aims to apply preprocessing, feature engineering, and classification algorithms to predict survival outcomes.

**2. Problem Statement**

To predict whether a passenger survived the Titanic disaster using demographic and onboard information from the Titanic dataset.

**3. Objectives**

* Preprocess the dataset by handling missing values and encoding categorical features.
* Train a Random Forest classifier to predict the survival.
* Evaluate model performance using validation accuracy.
* Provide a function to make predictions based on new user input.

**4. Methodology**

* **Data Collection**:

Data is sourced from the train.csv and test.csv files provided by the Titanic dataset on Kaggle.

* **Data Preprocessing**:  
  + Missing values in 'Age', 'Fare', and 'Embarked' were imputed using median and mode.
  + Irrelevant columns such as 'Cabin', 'Name', and 'Ticket' were removed.
  + Categorical features like 'Sex' and 'Embarked' were encoded as numerical values.
  + Feature scaling using StandardScaler.
* **Model Building**:
* Features and target labels were split into training and validation sets.
* A Random Forest classifier was trained on the processed data.

**5. Data Preprocessing**

The dataset is cleaned and prepared as follows:

* Missing numerical values were filled with the median.
* Categorical variables were mapped to integers.
* Features were scaled using StandardScaler.
* Dataset was split into 80% training and 20% validation subsets.

**6. Model Implementation**

A Random Forest Classifier was selected due to its high performance and robustness in classification tasks. It was trained using 100 estimators with a fixed random seed for reproducibility.

**7. Evaluation Metrics**

The following metrics are used to evaluate the model:

* **Accuracy**: 82.1%
* **Validation Accuracy:** Achieved an accuracy score on unseen validation data.
* **Interactive Prediction:** Verified the model’s usability through a custom user input function.

**8. Results and Analysis**

* The model achieved satisfactory validation accuracy.
* Predictions were exported to submission.csv for review.
* The user-defined prediction function worked as expected for hypothetical passenger data.

**9. Conclusion**

The Random Forest model effectively predicted Titanic passenger survival based on available features. The project illustrates the practical application of preprocessing, feature scaling, model training, and prediction in machine learning pipelines. Future work could include hyperparameter tuning or using advanced models like XGBoost for potentially better performance.

**10. References**

* scikit-learn documentation
* pandas documentation
* Seaborn visualization library