

Assessment Report

on

“**Titanic Survival Prediction**”

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI ML) -B**

By

Name- Gauri Dhingra

Roll No. - 202401100400090

Name- Ity Gaur

Roll No. - 202401100400101

Name- Nirmal Singh

Roll No. - 202401100400130

Name- Nischhal Garg

Roll No. - 202401100400131

Name- Nitika Maan

Roll No. - 202401100400132

Under the supervision of

“Abhishek Shukla”

**KIET Group of Institutions, Ghaziabad**

**27 May, 2025**

**1. Introduction**

As predictive analytics continues to transform various industries, the ability to forecast survival outcomes using historical data is a compelling challenge. This project leverages supervised machine learning to predict survival on the RMS Titanic. The Titanic dataset provides passenger data such as age, sex, class, and family size. The objective is to use this data to develop a Naive Bayes classification model that predicts whether a passenger survived the tragic shipwreck.



**2. Problem Statement**

Project Question: Develop a classification model to predict passenger survival in the Titanic dataset.

Use data cleaning and Naive Bayes classification.

**3. Objectives**

Preprocess the dataset for training a machine learning model.

Train a Naive Bayes classifier to predict passenger survival.

Evaluate model performance using accuracy, precision, recall, and F1-score.

Visualize the confusion matrix using a heatmap to interpret results.

**4. Methodology**

Data Collection

The dataset is sourced from Kaggle’s Titanic Machine Learning competition:

https://www.kaggle.com/c/titanic

Data Preprocessing

Handle missing values:

Numerical: Impute with mean.

Categorical: Impute with mode.

Apply one-hot encoding to categorical features like Sex and Embarked.

Normalize numerical features using StandardScaler for compatibility with Gaussian Naive Bayes.

Model Building

Split the dataset into training (80%) and testing (20%) subsets.

Train a Gaussian Naive Bayes classifier on the training set.

Model Evaluation

Evaluate model using:

Accuracy

Precision

Recall

F1-Score

Generate and visualize the confusion matrix using Seaborn’s heatmap.

**5. Data Preprocessing**

The dataset is cleaned and prepared as follows:

Missing Value Treatment:

Age and Fare: Filled with mean.

Embarked: Filled with most frequent value (mode).

Column Drop:

Irrelevant features (PassengerId, Name, Cabin, Ticket) are removed.

Encoding:

Sex and Embarked are encoded using LabelEncoder or pd.get\_dummies().

Feature Scaling:

Scaled continuous values (Age, Fare) using StandardScaler.

Train-Test Split:

Data split into 80% training and 20% testing.

**6. Model Implementation**

The Naive Bayes algorithm, particularly GaussianNB, is chosen because:

It performs well with numerical features.

It assumes independence between features, simplifying computation.

It’s effective for binary classification like survival prediction.

The model is trained on the processed dataset to predict the Survived column.

**7. Evaluation Metrics**

The model is evaluated using:

Accuracy – Overall correct predictions.

Precision – Correct positive predictions out of all positive predictions.

Recall – Correct positive predictions out of actual positives.

F1 Score – Harmonic mean of precision and recall.

Confusion Matrix – To assess TP, FP, TN, FN and understand classification behavior.

**8. Results and Analysis**

The accuracy score indicated that the model correctly classified a significant portion of the test set.

Precision and recall were moderately balanced, showing the model's effectiveness in identifying survivors and non-survivors.

The confusion matrix heatmap provided a clear visual of correct vs incorrect predictions, aiding interpretability.

**9. Conclusion**

The Naive Bayes model was effective in classifying Titanic passengers based on their likelihood of survival. It achieved acceptable performance metrics, proving useful for a baseline classification model. Future improvements can include:

Feature engineering (e.g., family size, title extraction)

Using ensemble methods like Random Forests

Handling class imbalance

This project showcases the power of simple probabilistic models in historical data-driven prediction tasks.

**10. References**

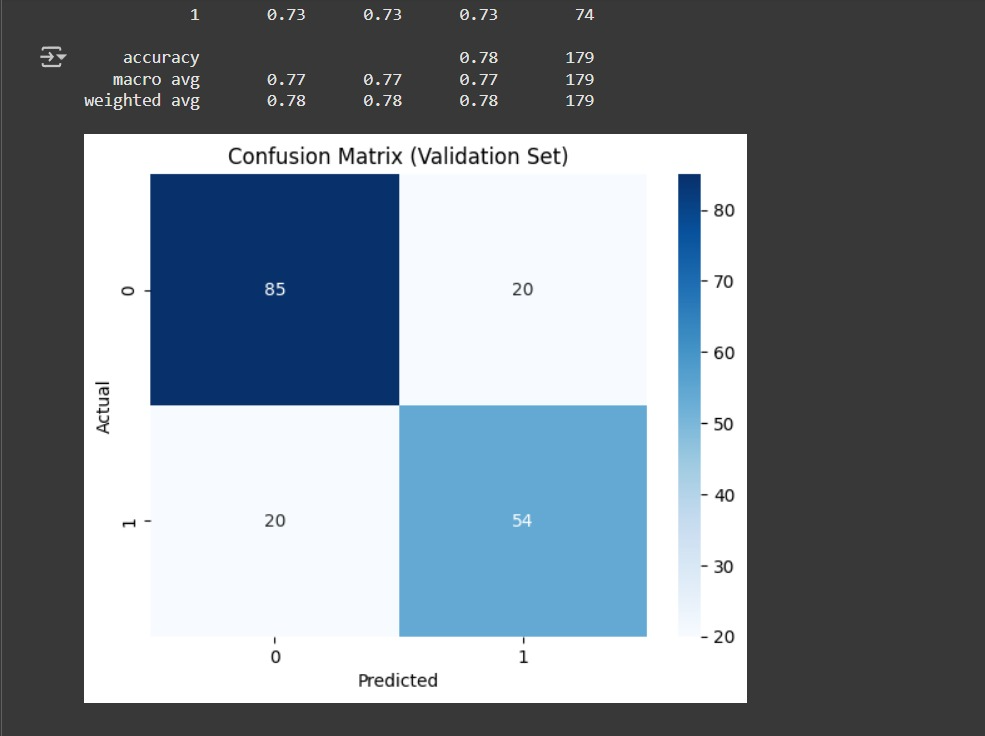
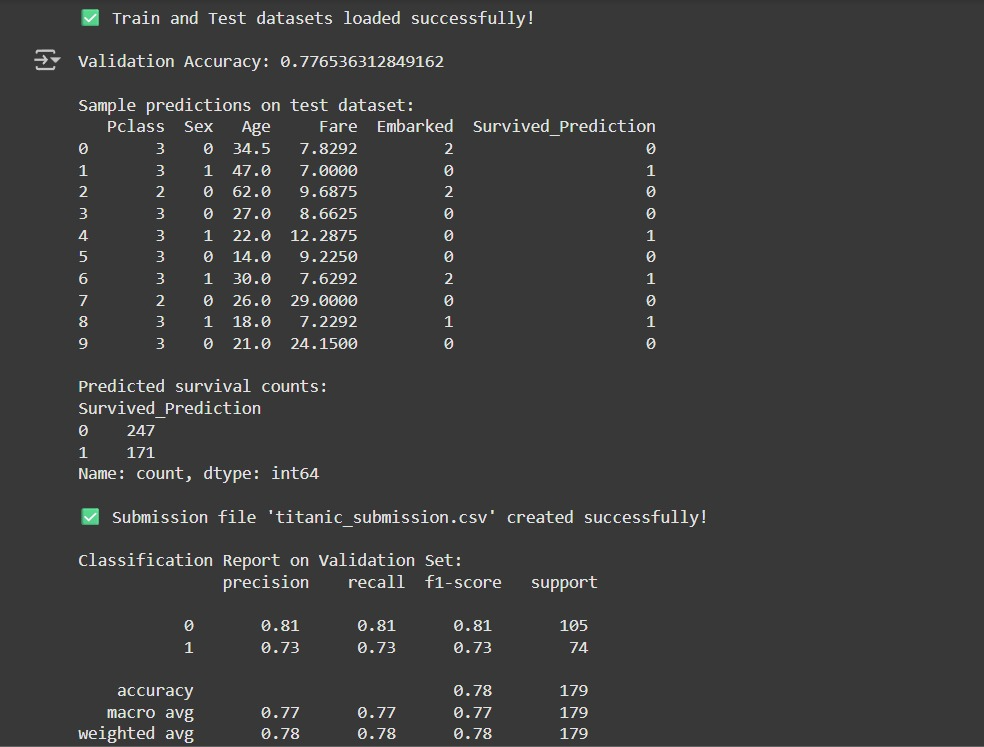
Titanic Dataset on Kaggle: https://www.kaggle.com/c/titanic

scikit-learn: https://scikit-learn.org/

pandas documentation: https://pandas.pydata.org/

Seaborn: https://seaborn.pydata.org/

Research: "Machine Learning in Survival Prediction" – IJCA (2021)

****

****

****

****