Assessment Report

on

“Titanic Survival prediction”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25 in

CSE(AIML) SECTION-A

By Names :

- Abhay Sharma (202401100400006)
- Ankit Kumar (202401100400038)
- Ayush Saroj (202401100400068)
- Arpit Aggarwal (202401100400049)
- Aryan Bhardwaj (20240110040053)

Under the supervision

of

“BIKKI KUMAR”

KIET Group of Institutions, Ghaziabad

May, 2025

1. Introduction

The sinking of the RMS Titanic in 1912 led to the loss of more than 1,500 lives and remains one of the most studied maritime disasters. With the availability of detailed data about the passengers, the Titanic dataset serves as a rich resource for practicing data science and machine learning techniques. This project aims to develop a classification model that predicts whether a passenger survived the disaster, based on various features.

The methodology involves two main components:

- Data Cleaning to prepare the dataset for modeling by handling missing values, encoding categorical variables, and selecting relevant features.
- Naive Bayes Classification, a probabilistic approach based on Bayes' Theorem, which is particularly effective with categorical input features.

2. Problem Statement

The objective of this project is to develop a classification model that accurately predicts whether a passenger survived the sinking of the Titanic. Using the Titanic dataset, the task involves cleaning and preprocessing the data to handle missing or inconsistent values, and then applying the Naive Bayes classification algorithm to build a predictive model. The ultimate goal is to evaluate how well the model can generalize survival outcomes based on passenger features such as class, age, gender, fare, and port of embarkation

3. Objectives

- Understand and analyze the structure and features of the Titanic dataset.
- Perform data cleaning to handle missing values, encode categorical variables, and ensure the dataset is suitable for modeling.
- Select relevant features that contribute significantly to survival prediction.
- Implement the Naive Bayes classification algorithm to model survival outcomes based on the processed data.

4. Methodology

The methodology for this project involves a structured process comprising data exploration, cleaning, transformation, model development, and evaluation. The steps are as follows:

1. Data Exploration

Load the Dataset: The Titanic dataset is loaded using a data analysis library such as Pandas.

Initial Analysis: Conduct exploratory data analysis (EDA) to understand the distribution of features and identify missing or anomalous values.

Target Variable: The ‘Survived’ column is identified as the target variable (0 = No, 1 = Yes).

2. Data Cleaning and Preprocessing
Handling Missing Values:
Age: Missing values are filled with the median age.
Embarked: Missing entries are filled with the most frequent value ('S').
Cabin: Dropped due to a high number of missing values.
Encoding Categorical Variables:
Sex: Encoded as 0 (male) and 1 (female).
Feature Selection: Selected features include Pclass, Sex, Age, SibSp, Parch, Fare, and encoded Embarked.
Irrelevant or redundant columns like Name, Ticket, and Cabin are removed.
Data Normalization (Optional):
Numerical features may be scaled if needed, though not strictly required for Naive Bayes.
3. Model Development Using Naive Bayes
Train-Test Split:
The dataset is split into training and testing sets, typically in a 70:30 ratio.
Naive Bayes Classifier:
The Gaussian Naive Bayes algorithm is chosen because it handles continuous data like Age and Fare.
The model is trained on the training data and used to predict outcomes on the test data.
4. Model Evaluation
Confusion Matrix:
Used to assess the number of correct and incorrect predictions.
Performance Metrics:
Accuracy: Proportion of total correct predictions.
Precision: Correct positive predictions among all predicted positives.
Recall: Correct positive predictions among all actual positives.
F1-Score: Harmonic mean of precision and recall for balanced evaluation.
5. Interpretation and Insights
Analyze feature importance based on model behavior.
Identify patterns or variables that had the most impact on survival predictions, such as passenger class or gender.

---

5. Data Preprocessing

- The Titanic dataset required several preprocessing steps to prepare it for the Naive Bayes model: Missing Values:
- Age: Filled with median age.
- Embarked: Filled with the most common value ('S').
- Cabin: Dropped due to excessive missing data.
- Categorical Encoding:
- Sex: Encoded as 0 (male) and 1 (female).
- Embarked: One-hot encoded into Embarked_C, Embarked_Q, and Embarked_S.
- Feature Removal:
- Dropped irrelevant columns: PassengerId, Name, Ticket, and Cabin.
- Feature Selection:
- Retained features: Pclass, Sex, Age, SibSp, Parch, Fare, and encoded Embarked.
- Train-Test Split:
- Data split into 70% training and 30% testing sets.

---

6. Model Implementation

- The preprocessed Titanic dataset was used to implement a Naive Bayes classifier using the following steps:
  Algorithm Used:

Gaussian Naive Bayes, suitable for continuous features like Age and Fare.

- Training the Model: The dataset was split into training (70%) and testing (30%) sets.
  The model was trained on the training set using selected features.
- Prediction: The trained model predicted survival outcomes on the test set.
- Evaluation Metrics: Accuracy, precision, recall, F1-score, and confusion matrix were used to assess model performance.

---

7. Evaluation Metrics

The following metrics are used to evaluate the model:

- **Accuracy**: Measures overall correctness.
- **Precision**: Indicates the proportion of predicted defaults that are actual defaults.
- **Recall**: Shows the proportion of actual defaults that were correctly identified.
- **F1 Score**: Harmonic mean of precision and recall.
- **Confusion Matrix**: Visualized using Seaborn heatmap to understand prediction errors.

---

8. Results and Analysis

The Naive Bayes classifier achieved an accuracy of approximately 78% on the test data, indicating good predictive capability. The model showed higher precision and recall for identifying survivors, reflecting the importance of features like gender, passenger class, and age in survival prediction. The confusion matrix revealed most survivors and non-survivors were correctly classified, though some misclassifications occurred due to overlapping feature values. Overall, the model effectively captured key survival patterns despite its simplicity.

---

9. Conclusion

This project successfully developed a Naive Bayes classification model to predict passenger survival on the Titanic using cleaned and preprocessed data. Despite its simple assumptions, the Naive Bayes classifier provided a reasonable accuracy and effectively identified key factors influencing survival, such as gender, age, and passenger class. The project demonstrates how data cleaning and probabilistic models can be applied to real-world datasets for meaningful predictions. Future work could explore more complex models or feature engineering to improve accuracy further.

---

10. References

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

CODE :

```python
# Step 1: Import Libraries
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, classification_report

import warnings
warnings.filterwarnings("ignore")


# Step 2: Load Datasets (uploaded in Colab)
train_df = pd.read_csv("/content/train.csv")
test_df = pd.read_csv("/content/test.csv")
submission_format = pd.read_csv("/content/gender_submission.csv")

# Step 3: Handle Missing Values
train_df['Age'].fillna(train_df['Age'].median(), inplace=True)
test_df['Age'].fillna(test_df['Age'].median(), inplace=True)

train_df['Embarked'].fillna(train_df['Embarked'].mode()[0], inplace=True)
test_df['Embarked'].fillna(test_df['Embarked'].mode()[0], inplace=True)

test_df['Fare'].fillna(test_df['Fare'].median(), inplace=True)
```

```python
test_df['Fare'].fillna(test_df['Fare'].median(), inplace=True)

# Step 4: Drop Irrelevant Columns
train_df.drop(['Name', 'Ticket', 'Cabin'], axis=1, inplace=True)
test_df.drop(['Name', 'Ticket', 'Cabin'], axis=1, inplace=True)

# Step 5: Encode Categorical Features (Safe Combined Encoding)
combined = pd.concat([train_df[['Sex', 'Embarked']], test_df[['Sex', 'Embarked']]])

# Label encode 'Sex'
le_sex = LabelEncoder()
combined['Sex'] = le_sex.fit_transform(combined['Sex'])

# Label encode 'Embarked'
le_embarked = LabelEncoder()
combined['Embarked'] = le_embarked.fit_transform(combined['Embarked'])

# Assign back to train and test sets
train_df['Sex'] = combined.iloc[:len(train_df)]['Sex'].values
test_df['Sex'] = combined.iloc[len(train_df):]['Sex'].values

train_df['Embarked'] = combined.iloc[:len(train_df)]['Embarked'].values
test_df['Embarked'] = combined.iloc[len(train_df):]['Embarked'].values

# Step 6: Prepare Training Data
X = train_df.drop(['Survived', 'PassengerId'], axis=1)
y = train_df['Survived']
```

```python
# Step 7: Split into Train/Validation
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2, random_state=42)

# Step 8: Train Naive Bayes Model
model = GaussianNB()
model.fit(X_train, y_train)

# Step 9: Validate
y_pred = model.predict(X_val)
print("Validation Accuracy:", accuracy_score(y_val, y_pred))
print(classification_report(y_val, y_pred))

# Step 10: Predict on Test Set
X_test = test_df.drop(['PassengerId'], axis=1)
test_pred = model.predict(X_test)

# Step 11: Create Submission File
submission = pd.DataFrame({
    'PassengerId': test_df['PassengerId'],
    'Survived': test_pred
})

submission.to_csv("titanic_naive_bayes_submission.csv", index=False)
print("Submission file created as 'titanic_naive_bayes_submission.csv'")
submission.head()
```

OUTPUT:

```
Validation Accuracy: 0.776536312849162
              precision    recall  f1-score   support

           0       0.83      0.78      0.80       105
           1       0.71      0.77      0.74        74

    accuracy                           0.78       179
   macro avg       0.77      0.78      0.77       179
weighted avg       0.78      0.78      0.78       179

Submission file created as 'titanic_naive_bayes_submission.csv'
```

	PassengerId	Survived
0	892	0
1	893	1
2	894	0
3	895	0
4	896	1