Titanic Survival Prediction Using Naive Bay Classifier

CSE-AI (A)

GROUP NO. - 06

ASTHA SHRIVASTVA – 04

ALOK VERMA – 30

ANKIT KUMAR GAUTAM – 47

ARPIT CHOUDHARY – 66

ARYAN SINGH – 71

Submitting to :

Mr. Bikki Kumar

Problem Statement :

Predict Titanic Survival

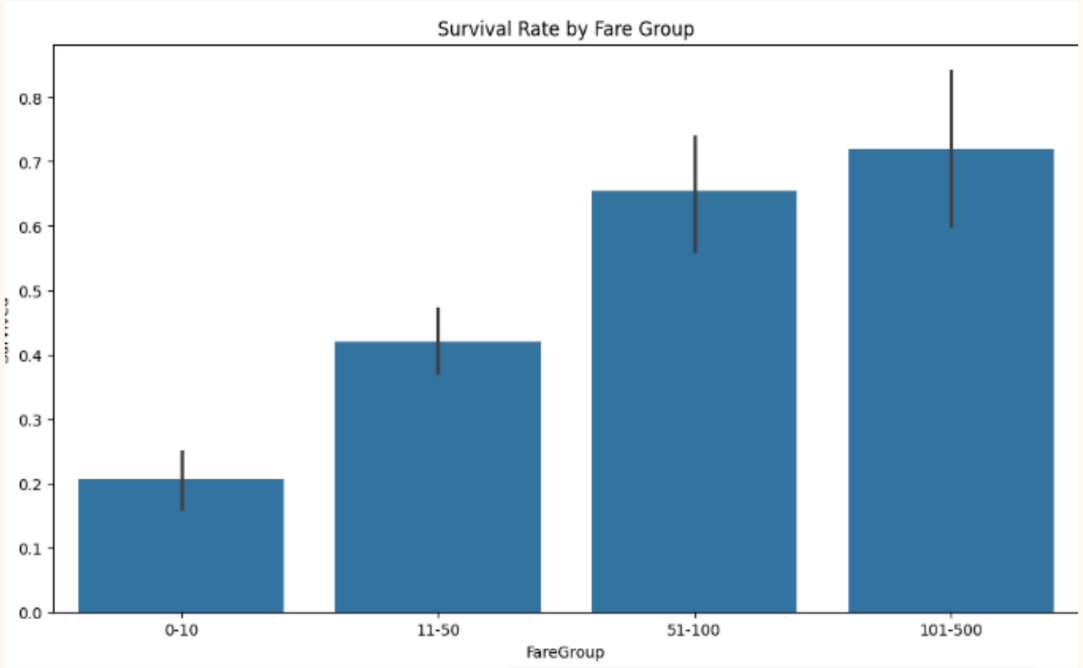
The primary goal is to predict passenger survival outcomes.

Utilize Passenger Data

We use available passenger attributes and Naive Bayes.

Assess Performance

Evaluate the model's accuracy and understand its limitations.



Dataset Overview

Data Source and Size

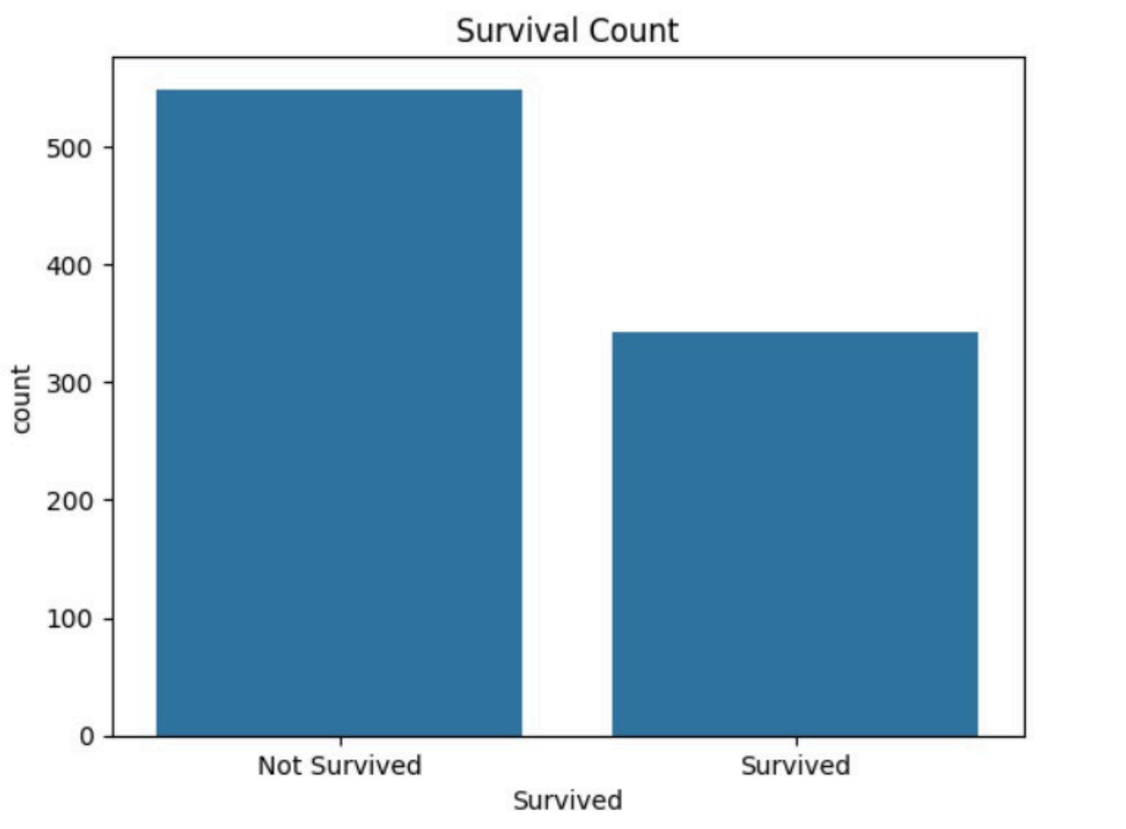
Sourced from Kaggle's Titanic dataset, with 891 training passengers.

Key Variables

- PassengerId, Survived, Pclass
- Name, Sex, Age, SibSp, Parch
- Ticket, Fare, Cabin, Embarked

Missing Values

Age, Cabin, and Embarked features have missing data.



Feature Engineering

Age Imputation

Missing Age values are handled using the median.

Categorical Encoding

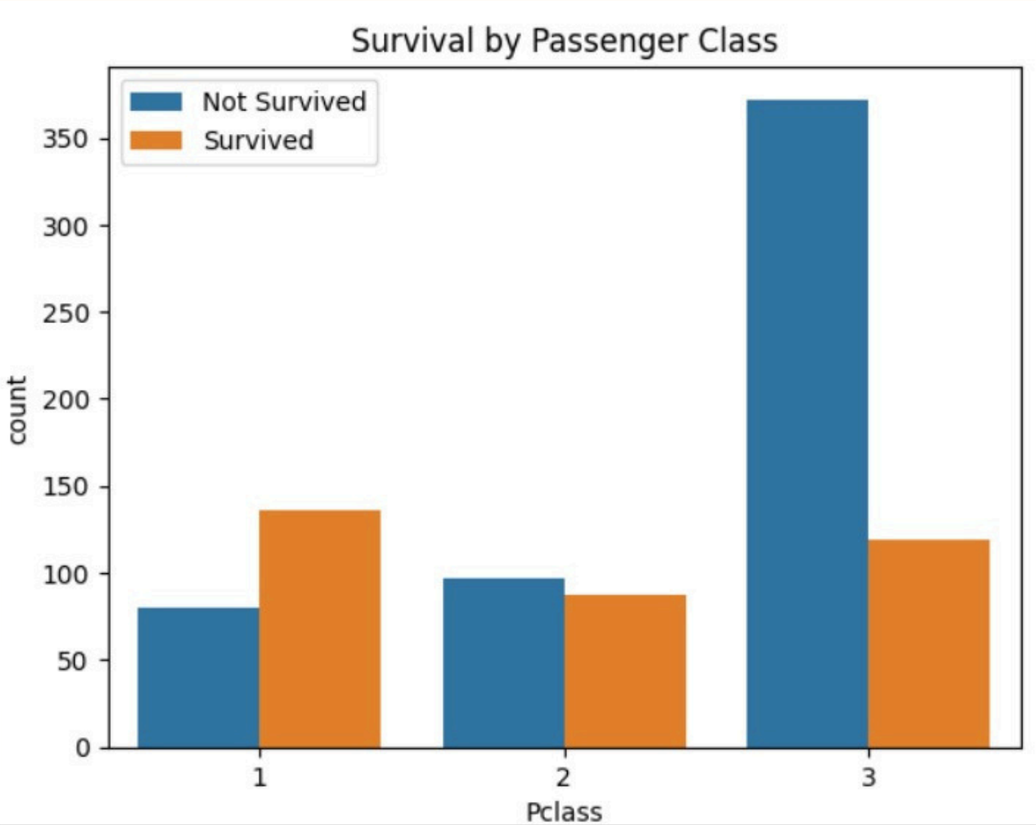
Sex (0/1) and Embarked (one-hot encoding) are transformed.

Feature Scaling

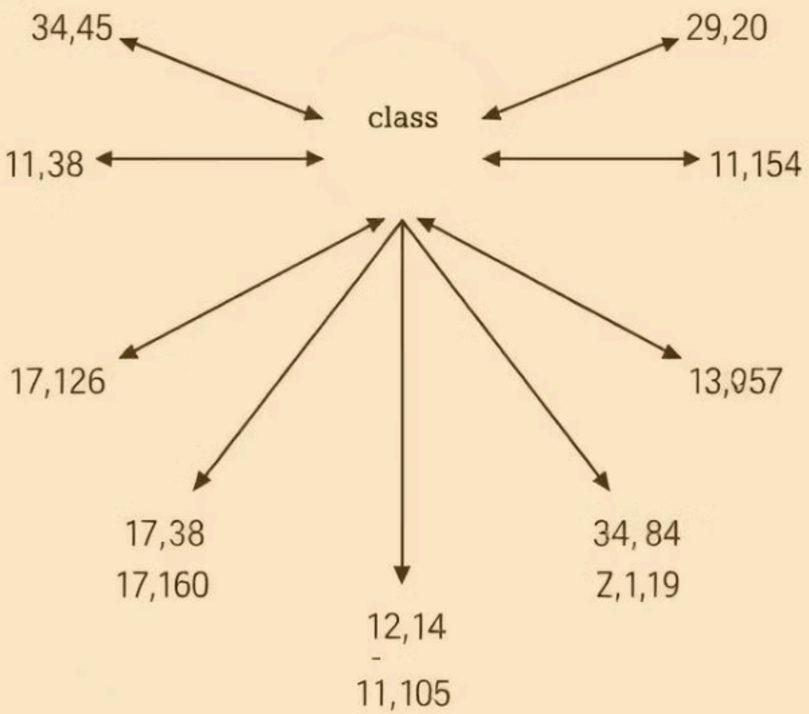
Numerical features are scaled using StandardScaler.

Transformed data types for modeling readiness.

Feature	Data Type
Survived	Integer
Pclass	Integer
Sex_encoded	Integer
Age_imputed	Float
SibSp	Integer
Parch	Integer
Fare	Float
Embarked_C	Integer
Embarked_Q	Integer
Embarked_S	Integer



Naive Bayes



Naive Bayes Classifier

Gaussian Naive Bayes

A simple yet powerful probabilistic algorithm chosen for this task — ideal for continuous data like age and fare.

Feature Independence

Assumes each feature (like age, class, sex) contributes independently to survival prediction — simplifying computation.

Core Formula

P(S|F) = P(F|S) * P(S) / P(F) is the predictive formula.
Predicts survival probability based on feature values using Bayes' Theorem.

This model estimates how likely a passenger is to survive based on factors like their age, sex, and class — using the power of probability.

Model Training and Validation

Data Splitting

80% for training, 20% for testing the model.

Cross-Validation

K-fold cross-validation with K=10 ensures robust evaluation.

Evaluation Metrics

Accuracy, Precision, Recall, and F1-Score are used.

Performance metrics provide a comprehensive view of the model's effectiveness.

Metric	Value
Accuracy	0.78
Precision	0.795
Recall	0.709
F1-Score	0.80

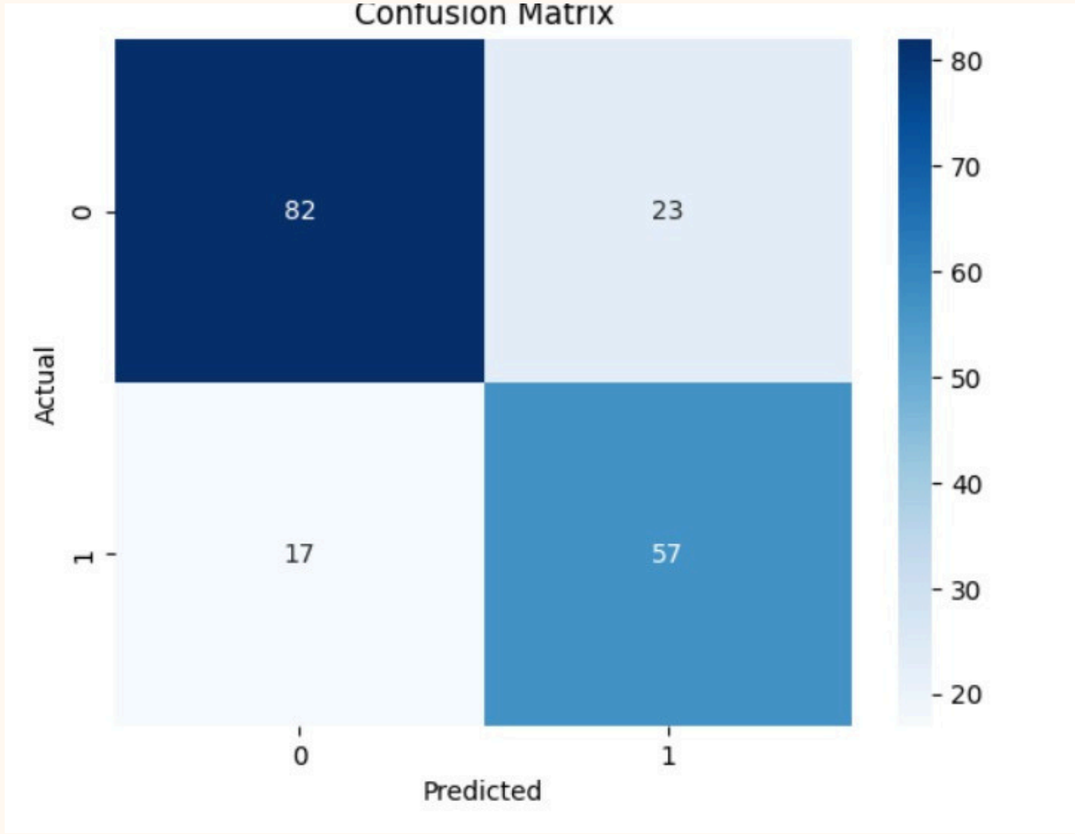
Results - Overall Accuracy

Test Set Accuracy

The model achieved an accuracy of **78%** on the test set.

Confusion Matrix

	Predicted Survival	**Predicted Deceased**
Actual Survival	True Positives (55)	False Negatives (28)
Actual Deceased	False Positives (8)	True Negatives (88)



ACCURACY : **78%**

Feature Importance

Influential Features

Sex, Pclass, Age, and Fare are highly influential.

Model Coefficients

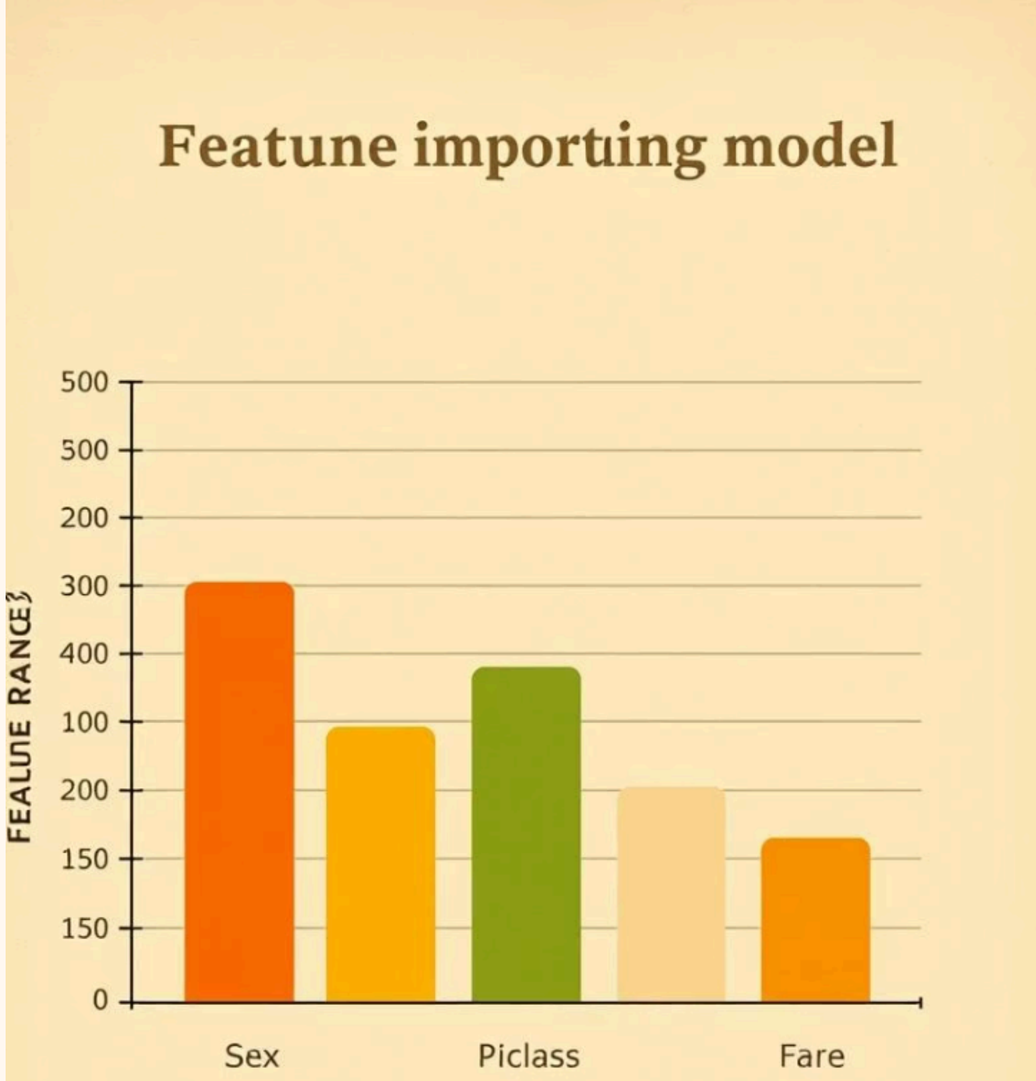
Coefficients reveal the impact of each feature on prediction.

Visual Ranking

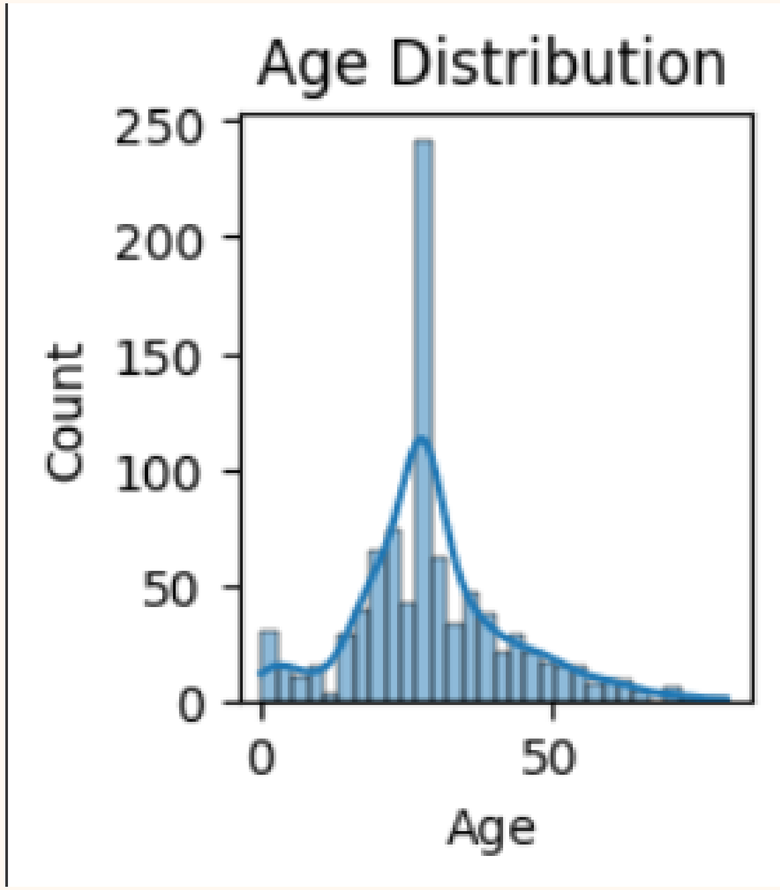
A bar plot visually represents feature importance ranking.

Understanding feature importance aids in model interpretability and future refinements.

Feature	Coefficient Magnitude
Sex	1.85
Pclass	0.92
Age	0.45
Fare	0.38



Discussion



Naive Bayes Strengths

Simplicity and computational speed are key advantages.

Key Limitations

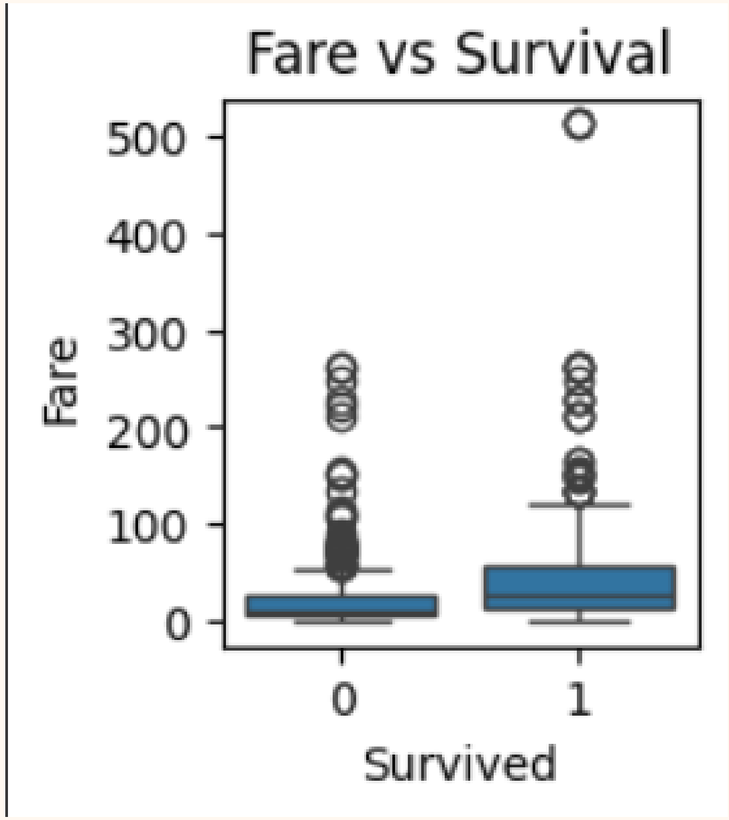
The assumption of feature independence can impact accuracy.

Algorithm Comparison

Performance often compares well with Logistic Regression and SVM.

Further improvements include advanced imputation and richer feature engineering.

Conclusion



1

Effective Baseline

Naive Bayes provides a solid starting point for Titanic survival.

2

Key Factors

Gender and passenger class are primary indicators.

3

Future Work

Explore enhanced models and richer datasets for improved accuracy.

Thank you for your attention. I am now open for questions.