Introduction:

One of the most famous disasters in history is the loss of the Titanic. In this study, we use machine learning techniques—more specifically, the Naive Bayes classifier—to predict the survival rate of passengers on the Titanic. A simple yet effective probabilistic classifier, the Naive Bayes classifier relies on the independence of characteristics and the Bayes theorem.

Dataset Description:

The dataset consists of 891 records of Titanic passengers, each with the following features:

- Passenger: A unique identifier for each passenger.
- Survived: A binary variable indicating whether the passenger survived (1) or not (0).
- Pclass: The passenger's ticket class (1 = 1st class, 2 = 2nd class, 3 = 3rd class).
- Name: The name of the passenger.
- Sex: The gender of the passenger.
- Age: The age of the passenger in years.
- SibSp: The number of siblings or spouses the passenger had aboard the Titanic.
- Parch: The number of parents or children the passenger had aboard the Titanic.
- Ticket: The ticket number of the passenger.
- Fare: The fare paid by the passenger.
- Cabin: The cabin number where the passenger stayed.
- Embarked: The port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

-The Titanic dataset is well-suited for the Naive Bayes classifier because it contains several categorical features, such as Sex, Pclass, and Embarked, which the classifier handles effectively. The binary nature of the target variable (Survived) aligns well with Naive Bayes' capabilities. Additionally, despite the independence assumption, Naive Bayes often performs well in practice, making it a robust choice for this classification task.

2. Data Exploration and Preprocessing:

Load Data:

We begin by loading the train and test datasets with pandas.

and exploring the features and checking for any missing values, The dataset contains information such as passenger class, age, sex, fare, and more.

Handle missing values:

i. We handle missing values by inputting median values for age and fare.
ii. filling missing embarkation points with the most common value.
iii. We drop the 'Cabin' and 'Ticket' columns as they contain too many missing values or do not provide meaningful information.

Dropping Irrelevant Features:

Drop the 'PassengerId', 'Name', and 'Ticket' columns.

Converting Categorical Features to Numerical:

Categorical features are converted into numerical ones using one-hot encoding.

- Identify categorical variables.
- Encode the 'Sex' column using mapping and one-hot encoding.
- Perform one-hot encoding on the 'Embarked' column.

Normalizing/Scaling Features:

I use features like 'Age' and 'Fare' using the StandardScaler function to standardize the data.

3. Feature Selection Rationale:

Choosing the right features is essential to creating a strong model. In this study, factors such as passenger class, age, sex, and embarkation site are selected based on their likelihood to significantly affect survival. Based on their availability in the dataset and significance to survival, these features were chosen.

4. Model Training and Parameter Tuning:

Using the preprocessed data

Splitting the Dataset:

The dataset is divided into features (X) and the target variable (y - 'Survived'). It is then split into training and testing sets.

we train a Gaussian Naive Bayes classifier. After the model is created, the training data is fitted to it, and it is then used to predict the test data. We use 5-fold split k-fold cross-validation to guarantee the model's robustness. This helps to reduce overfitting and assess the model's effectiveness across various data subsets.

5. Evaluation Results:

The performance of the model is evaluated using various metrics such as accuracy, precision, recall, and F1-score. Additionally, we examine the confusion matrix and classification report to gain insights into the model's strengths and weaknesses. Cross-validation results provide an estimate of the model's performance on unseen data, with average scores reported for each metric.

Accuracy:

The accuracy of the model was calculated for both the training and test sets:

- Train Accuracy: 0.7849
- Test accuracy: 0.7985

Confusion Matrix:

The confusion matrix provides a summary of the prediction results and the errors:

- True Positives (TP): 135
- True Negatives (TN): 79
- False Positives (FP): 29
- False Negatives (FN): 25
- A heatmap of the confusion matrix was also generated for better visualization.

6. Conclusion and Challenges Faced:

Challenges:

Handling Missing Values:

The dataset contained missing values in the 'Age' and 'Embarked' columns, as well as a significant number of missing values in the 'Cabin' column.

Feature Engineering:

Transforming categorical variables into numerical formats.

Model Evaluation:

Balancing between precision, recall, and F1-score for the 'Survived' class was essential.

Cross-Validation:

Ensuring that the cross-validation scores were consistent and reflective of the model's generalizability was essential.

In summary, the Naive Bayes classifier shows encouraging results when it comes to predicting survival on the Titanic dataset. However, the task ran into issues with feature engineering and model selection. While model selection included picking the best classifier for the job, feature engineering handled missing values and selected significant features. Despite these difficulties, the assignment underscores the importance of feature selection, data preprocessing, and model evaluation to creating successful machine learning models.