**Naive Bayes Classifier Report**

**Dataset Overview**

For this project, we used the Titanic dataset. This dataset contains information about passengers on the Titanic, with details such as demographics, travel class, family connections, and embarkation details.

The dataset includes the following columns:

* PassengerId: A unique identifier for each passenger.
* Pclass: Passenger class (1st, 2nd, or 3rd).
* Name: Full name of the passenger.
* Sex: Gender of the passenger.
* Age: Age of the passenger.
* SibSp: Number of siblings or spouses aboard.
* Parch: Number of parents or children aboard.
* Ticket: Ticket number.
* Fare: Price of the ticket.
* Cabin: Cabin number (if available).
* Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

**Data Preprocessing**

1. **Handling Missing Values**

The dataset contained missing values in the Age, Fare, and Embarked columns. To handle these:

* + Missing values in Age and Fare were filled with the median.
  + Missing values in Embarked were filled with the mode.

1. **Encoding Categorical Variables**

The Sex and Embarked columns contained categorical data, which was converted to numerical values using label encoding to make them usable by the model.

1. **Feature Engineering**

Relevant features were selected to focus on key aspects of the dataset that could influence the target variable. The selected features are:

* + Features: Pclass, Sex, Age, SibSp, Parch, and Fare.
  + Target: Embarked, representing the embarkation port of the passenger.

1. **Splitting the Dataset**

We split the data into 70% for training and 30% for testing to evaluate the model's performance:

X = data[['Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare']]

y = data['Embarked']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

**Model Choice: Naive Bayes**

For this project, the Multinomial Naive Bayes classifier was chosen. Naive Bayes classifiers are probabilistic models based on Bayes' Theorem, assuming conditional independence between features, which simplifies computation.

The Multinomial Naive Bayes classifier is typically used for discrete features and works well with count data, making it suitable for categorical and integer-based features.

**Model Training**

The Naive Bayes classifier was trained on the training dataset as follows:

nb\_classifier = MultinomialNB()

nb\_classifier.fit(X\_train, y\_train)

**Model Performance**

After training, predictions were made on the test set, and performance was evaluated using a confusion matrix and classification report:

y\_pred = nb\_classifier.predict(X\_test)

**Evaluation Metrics**

The model's performance was evaluated using the following metrics:

* Accuracy: The percentage of correct predictions made by the model.
* Confusion Matrix: A table that provides a detailed view of true positives, false positives, true negatives, and false negatives.
* Classification Report: Includes precision, recall, and F1-score for each embarkation port (C, Q, S).

**Results**

* Accuracy: 39.7%
* Confusion Matrix:

[[ 8, 8, 11],

[ 0, 13, 1],

[12, 44, 29]]

* Classification Report:

Precision, recall, and F1-scores were generated for each class (embarkation port), highlighting areas where the model excelled or underperformed.

**Insights Gained**

* Feature Importance: Passenger class, age, and fare were key factors in predicting embarkation points, though additional feature engineering or alternative models may further enhance results.
* Model Limitations: The Naive Bayes classifier assumes independence among features, which may not hold true for all variables in this dataset. Continuous variables, such as Fare, may require feature scaling or other adjustments in alternative models.
* Data Imbalance: Some embarkation points have fewer passengers, impacting prediction performance for these classes. Techniques like class weighting or resampling could address this in future iterations.

**Conclusion**

This project demonstrated the application of the Naive Bayes classifier on the Titanic dataset to predict embarkation points. Although the model’s simplicity limits its accuracy, it performed reasonably well, highlighting useful patterns in the data. For further enhancement, models like logistic regression, decision trees, or ensemble methods could be explored.