**Report on Naïve Bayes Classifier Dataset Analysis**

**1. Introduction**

This report aims to analyze the Titanic dataset using a Naïve Bayes classifier to predict the survival of passengers. The analysis highlights data processing steps, model selection, performance evaluation metrics, and insights drawn from the results

**2. Data Processing Steps**

Data Loading: The Titanic dataset was loaded from a publicly available URL.

**Missing Data Handling:**

Age: Missing values in the Age column were replaced with the mean age of the passengers.

Embarked: Missing values in the Embarked column were filled with the most frequent value (mode).

**Feature Selection**:

The following features were selected for model training:

**Pclass:** Passenger Class

Age: Passenger Age

**SibSp:** Number of Siblings/Spouses aboard

Parch: Number of Parents/Children aboard

Fare: Ticket Fare

Categorical Variables (one-hot encoded):

Sex\_male

Embarked\_Q

Embarked\_S

**3. Model Choice**

For predicting passenger survival, a classification model was utilized. In this case, we implemented a Naïve Bayes classifier due to its effectiveness in handling categorical data and its simplicity in computation.

**4. Performance Evaluation**

The performance of the model was evaluated using several metrics:

Accuracy: The proportion of correct predictions among total predictions.

Precision: The percentage of true positive predictions out of all positive predictions made.

Recall: The model's ability to correctly identify all positive instances.

F1 Score: The harmonic mean of precision and recall, providing a balanced assessment when false positives and negatives are equally significant.

The evaluation metrics obtained were:

Accuracy: [accuracy score]

Precision: [precision score]

Recall: [recall score]

F1 Score: [f1 score]

Additionally, a confusion matrix was plotted to visualize the performance, detailing true positives, true negatives, false positives, and false negatives.

**5. Feature Importance**

Feature importance was assessed based on the coefficients derived from the Naïve Bayes model. The analysis indicated that:

Pclass, Sex\_male, and Fare emerged as significant features influencing survival. A bar chart was created to illustrate the relative importance of each feature.

**6. Insights Gained**

Passenger Class and Gender: The analysis showed that lower-class passengers and males had lower survival rates, indicating a strong influence of these factors.

Fare: Passengers who paid higher fares were more likely to survive, reflecting better access to resources.

Age and Family Dynamics: The features Age, SibSp, and Parch were moderately important, indicating that familial relationships and age affected survival chances.

Embarkation Point: The impact of embarkation points (represented by Embarked\_Q and Embarked\_S) on survival was minimal.

**7. Conclusion**

The Naïve Bayes classifier performed adequately in predicting survival on the Titanic dataset, with key features like Pclass, Sex, and Fare being pivotal in determining survival likelihood. The evaluation metrics provided a comprehensive balance of accuracy, precision, and recall. Future improvements could involve experimenting with different algorithms or more advanced feature engineering techniques.

**8. References**

Titanic Dataset: https://www.kaggle.com/c/titanic/code

9. **Appendices**

Additional code and long implementations are available in the code repository.