**Report on Titanic Dataset Analysis**

**Short Report on Titanic Dataset Analysis**

**1. Data Processing Steps:**

* **Data Loading**: We loaded the Titanic dataset from a publicly available URL.
* **Missing Data Handling**:
  + The Age column contained missing values, which were imputed by replacing the missing values with the mean of the column.
  + The Embarked column also had missing values, which were filled with the mode (most frequent) value.
* **Feature Selection**:
  + The features used for training the model were:
    - Pclass (Passenger Class)
    - Age (Passenger Age)
    - SibSp (Number of Siblings/Spouses aboard)
    - Parch (Number of Parents/Children aboard)
    - Fare (Ticket Fare)
    - One-hot encoded categorical variables:
      * Sex\_male
      * Embarked\_Q
      * Embarked\_S

**2. Model Choice:**

* A classification model was used to predict whether a passenger survived (Survived) or not.
* For this task, we used a machine learning model (logistic regression or another classifier).

**3. Performance Evaluation:**

* **Accuracy**: The model’s accuracy is calculated as the ratio of correct predictions to total predictions.
* **Precision**: Measures the percentage of true positive predictions out of all positive predictions made by the model.
* **Recall**: Measures the ability of the model to correctly identify all positive instances.
* **F1 Score**: A harmonic mean of precision and recall, providing a balanced metric when false positives and false negatives are equally important.

The model evaluation metrics are as follows:

plaintext

Accuracy: [accuracy score]

Precision: [precision score]

Recall: [recall score]

F1 Score: [f1 score]

* **Confusion Matrix**: A confusion matrix was plotted to visualize the performance in terms of true positives, true negatives, false positives, and false negatives.

**4. Feature Importance:**

The feature importance was calculated based on the model's coefficients or another form of feature importance extraction. The difference between coefficients for survival and non-survival classes provides insights into the importance of each feature.

* Features like Pclass, Sex\_male, and Fare had higher importance in predicting survival.

A bar chart was plotted to display the relative importance of each feature.

**5. Insights Gained:**

* **Passenger Class and Gender**: Pclass and Sex\_male were among the most influential features, with lower-class passengers and males being less likely to survive.
* **Fare**: Higher ticket fares were associated with a higher likelihood of survival, likely indicating that wealthier passengers had access to better resources.
* **Age and Family**: Features like Age, SibSp, and Parch showed moderate importance, suggesting that family dynamics and age also played a role in survival.
* **Embarkation Point**: The place of embarkation (captured by Embarked\_Q, Embarked\_S) had a smaller influence on survival.

**Conclusion:**

The model performed reasonably well in predicting survival on the Titanic dataset, with features like Pclass, Sex, and Fare playing key roles in determining survival likelihood. The evaluation metrics provide a good balance between accuracy and the trade-off between precision and recall. Further model improvements could be explored by experimenting with other algorithms or performing more advanced feature engineering.