Machine Learning Assignment Report

1. Dataset Description

The Titanic dataset (`train.csv` from Kaggle) contains 891 passenger records with 12 initial features, used to predict survival (0 = Did not survive, 1 = Survived). Key features include:

Survived: Target variable (binary)

Pclass: Passenger class (1st, 2nd, 3rd)

Sex: Gender (male, female)

Age: Age in years

Fare: Ticket price

Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

SibSp: Number of siblings/spouses aboard

Parch: Number of parents/children aboard

2. Preprocessing Steps

Missing Values:

- Filled 'Age' with median ({:.2f}).

- Filled 'Embarked' with mode ('{}').

- Dropped 'Cabin' (77% missing), 'PassengerId', 'Name', and 'Ticket' (irrelevant for prediction).

- Encoding:

  - Label encoding: 'Sex' (male=1, female=0), 'Embarked' (C=0, Q=1, S=2).

  - One-hot encoding: 'Pclass' (Pclass_2, Pclass_3).

- Normalization: Standardized 'Age' and 'Fare' using StandardScaler.

- Data Split: 80% training, 20% testing (random_state=42).

3. Model Selection

Logistic Regression was chosen for its simplicity, interpretability, and suitability for binary classification. It predicts the probability of survival based on input features.

4. Evaluation Results

- Accuracy: {:.4f} (proportion of correct predictions)
- Precision: {:.4f} (proportion of positive predictions that were correct)
- Recall: {:.4f} (proportion of actual positives correctly identified)
- F1-Score: {:.4f} (harmonic mean of precision and recall)

Visualizations (saved as PNGs):

- EDA: Age distribution, survival count, survival by class, and sex (eda_visualizations.png).
- Confusion Matrix: Shows true positives, false positives, etc. (confusion_matrix.png).
- ROC Curve: AUC = {:.2f}, indicating model's ability to distinguish classes (roc_curve.png).

5. Insights and Improvements

- Insights:
  - The model performs well (AUC = {:.2f}), but struggles with recall, suggesting it misses some survivors.
  - Features like 'Sex' and 'Pclass' are strong predictors (seen in EDA).
  - 'SibSp' and 'Parch' may have limited impact without feature engineering.
- Improvements:
  - Feature Engineering: Combine 'SibSp' and 'Parch' into 'family_size', extract titles from 'Name' (e.g., Mr., Mrs.).
  - Model Selection: Try Random Forest or Gradient Boosting for potentially better performance.
  - Hyperparameter Tuning: Use GridSearchCV to optimize Logistic Regression parameters.
  - Cross-Validation: Implement k-fold cross-validation for robust evaluation.

- Outlier Handling: Address extreme 'Fare' values to improve model stability.

6. Conclusion

This assignment implemented a complete machine learning pipeline using the Kaggle Titanic dataset. The Logistic Regression model serves as an effective baseline, achieving reasonable performance. Future work could explore advanced models and feature engineering to enhance predictive accuracy.