BAKU HIGHER OIL SCHOOL

INFORMATION TECHNOLOGY DEPARTMENT
INFORMATION SECURITY
DIVISION

Machine Learning

Home Assignment 2

Assignment name: Logistic Regression

Student name: Huseyn Abdullayev

Group number: CS 25

Instructor: Leyla Muradkhanli

Dataset Description and Preprocessing Steps

The dataset used for this assignment is the Titanic Survival Dataset. It contains 891 samples of passengers from the Titanic shipwreck, with information on whether they survived or not.

The dataset includes the following key columns:

• PassengerId: Unique identifier for each passenger.

• Survived: Target variable (0 = Did not survive, 1 = Survived).

• Pclass: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd).

• Name: Passenger name.

• Sex: Gender (male or female).

• Age: Age in years.

• SibSp: Number of siblings/spouses aboard.

• Parch: Number of parents/children aboard.

• Ticket: Ticket number.

• Fare: Passenger fare.

• Cabin: Cabin number.

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

The target variable is "Survived," which is binary (categorical).

Data Preprocessing Steps:

Loading Data:

The dataset was imported using pandas.read_csv('Titanic-Dataset.csv').

```
10  # Load the dataset
11  data = pd.read_csv('Titanic-Dataset.csv')
12
```

Data Cleaning:

Missing values were handled as follows:

Age: Filled with median value (177 missing).

Embarked: Filled with mode value (2 missing).

Fare: Filled with median value (0 missing, but checked).

Cabin: Ignored due to high missing values (687 missing) and irrelevance.

No other columns had missing values that required handling.

```
# Data cleaning: Handle missing values
data['Age'].fillna(data['Age'].median(), inplace=True)
data['Embarked'].fillna(data['Embarked'].mode()[0], inplace=True)
data['Fare'].fillna(data['Fare'].median(), inplace=True)
```

Encoding Categorical Labels:

Categorical features "Sex" (male/female) and "Embarked" (C/Q/S) were encoded into numeric labels using LabelEncoder from sklearn.preprocessing.

```
# Encode categorical features
le = LabelEncoder()
data['Sex'] = le.fit_transform(data['Sex'])
data['Embarked'] = le.fit_transform(data['Embarked'])
```

Feature Selection and Splitting:

Selected features: Pclass, Sex, Age, SibSp, Parch, Fare, Embarked (as X).

Target: Survived (as y).

The dataset was then split into 80% training data (712 samples) and 20% testing data (179 samples) using train_test_split() with random_state=42 for reproducibility.

```
# Select features and target
features = ['Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare', 'Embarked']
X = data[features]
y = data['Survived']

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Summary of Model Implementation

A Logistic Regression model from sklearn.linear_model was chosen for this binary classification task. The model was trained using the training data (712 samples) with max_iter=200 to ensure convergence.

```
# Train a Logistic Regression model
model = LogisticRegression(max_iter=200)
model.fit(X_train, y_train)

```

The model learned the following coefficients for each feature:

[[-0.930726595, -2.59670887, -0.0304300465, -0.293864908, -0.111820831, 0.00253307048, -0.218848023]]

Intercept: [4.55655971]

These values indicate how strongly each feature contributes to predicting survival. In particular, Sex (negative coefficient for male) and Pclass (negative for lower classes) had the largest magnitudes, meaning they play a key role in distinguishing survivors from non-survivors.

```
Coefficients: [[-9.30726595e-01 -2.59670887e+00 -3.04300465e-02 -2.93864908e-01
  -1.11820831e-01  2.53307048e-03 -2.18848023e-01]]
Intercept: [4.55655971]

Model Accuracy: 0.8100558659217877
```

Interpretation and Conclusion

The model was evaluated on the 179 test samples. Predictions were made on the test data, and performance metrics were calculated.

Overall Accuracy: 0.81 (81%).

The classification report showed:

```
Classification Report:
              precision    recall  f1-score   support

        Died       0.83      0.86      0.84       105
    Survived       0.79      0.74      0.76        74
        Died       0.83      0.86      0.84       105
    Survived       0.79      0.74      0.76        74

    accuracy                           0.81       179
   macro avg       0.81      0.80      0.80       179
weighted avg       0.81      0.81      0.81       179
```

The confusion matrix confirmed:

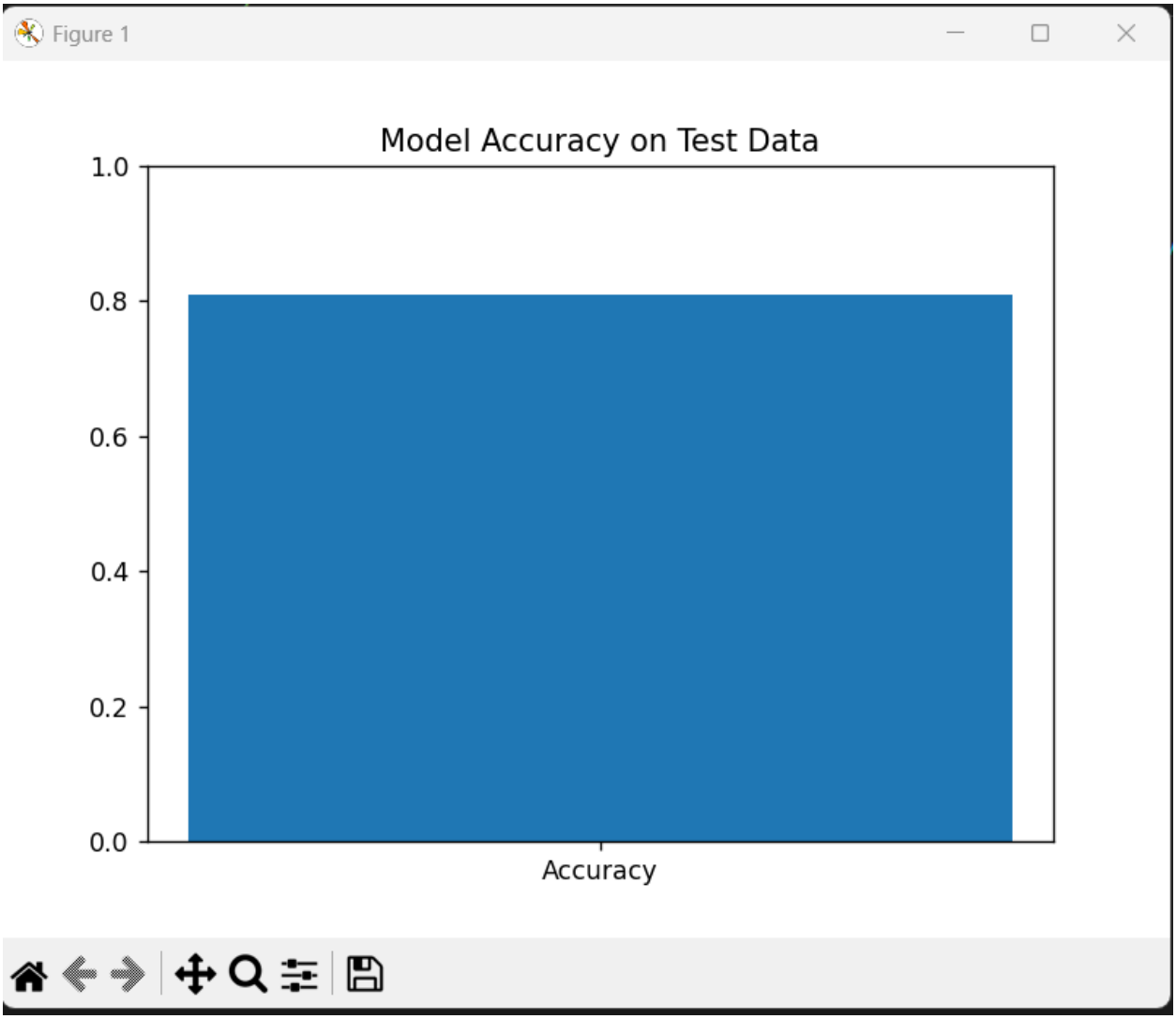
[[90 15]

[19 55]]

This means the model correctly predicted 90 non-survivors and 55 survivors, with 15 false positives and 19 false negatives.

```
Confusion Matrix:
[[90 15]
 [19 55]]
```

I also visualized the accuracy of the model on test data using pyplot from matplotlib:



Interpretation:

• Sex and Pclass are the most influential features for predicting survival, aligning with historical knowledge of the Titanic disaster (women and higher-class passengers had higher survival rates).

• The logistic regression model effectively captured the relationships between passenger attributes and survival outcomes, though not perfectly due to the dataset's complexity.

Conclusion:

The logistic regression classifier achieved 81% accuracy on the Titanic dataset. This demonstrates that logistic regression is suitable for binary classification tasks

with some linear separability, like predicting survival based on socioeconomic and demographic factors.

In real-world applications, while 81% accuracy is solid, further improvements could be made through advanced feature engineering (e.g., creating family size features) or using more complex models like random forests. This experiment highlights logistic regression's interpretability and effectiveness with structured data.