Assignment: Logistic Regression on Study Hours vs Pass/Fail Dataset

Aim

To implement logistic regression for predicting pass/fail outcomes based on study hours, evaluate classifier performance, and visualize the prediction probabilities.

Lab Setup

- Software: Python 3.x environment (Google Colab, Jupyter Notebook, or local IDE)
- Required libraries:
    - pandas
    - numpy
    - scikit-learn
    - matplotlib
- Installation command (if needed):
- `bash`

```
pip install pandas numpy scikit-learn matplotlib
```

-

Input

- Dataset with:
    - Feature: Study Hours (numeric)
    - Label: Pass (binary: 0=Fail, 1=Pass)
- Example Entries:

Study Hours	Pass
1	0
2	0

3	0
4	1
5	1
6	1
7	1
8	1

Expected Output

- Accuracy of logistic regression on test data.
- Confusion matrix for classification results.
- Classification report showing precision, recall, F1-score.
- Plot of study hours against predicted pass probabilities.

Example output:

```
text
```

```
Accuracy: 1.0
Confusion Matrix:
 [[1 0]
  [0 1]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         1
           1       1.00      1.00      1.00         1
```

```
    accuracy                           1.00         2
   macro avg       1.00      1.00      1.00         2
weighted avg       1.00      1.00      1.00         2
```

Theory / Algorithm

Logistic Regression models the probability that input data belongs to a particular class by applying the logistic (sigmoid) function to a linear combination of features. The sigmoid function maps any real-valued number into a value between 0 and 1, interpreted as a probability.

Algorithm steps:

1. Initialize coefficients and intercept.
2. Compute linear combination
3. $z=\beta 0+\beta 1x$
4. $z=\beta$
5. 0
6. $+\beta$
7. 1
8. x.
9. Apply sigmoid:
10. $\sigma(z)=11+e-z$
11. $\sigma(z)=$
12. $1+e$
13. $-z$

14. 1

15. to get probabilities.
16. Use maximum likelihood estimation to fit
17. β
18. β parameters.
19. Predict class labels based on a threshold (usually 0.5).

Code

python

```
import pandas as pd
from sklearn.model_selection import train_test_split
```

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix,
classification_report
import matplotlib.pyplot as plt
import numpy as np

# Dataset: Study Hours vs Pass/Fail
data = {
    'Study Hours': [1, 2, 3, 4, 5, 6, 7, 8],
    'Pass': [0, 0, 0, 1, 1, 1, 1, 1]
}
df = pd.DataFrame(data)

# Features and label
X = df[['Study Hours']].values
y = df['Pass'].values

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.25, random_state=42)

# Train logistic regression model
model = LogisticRegression()
model.fit(X_train, y_train)

# Prediction on test set
y_pred = model.predict(X_test)

# Evaluation metrics
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
print("Classification Report:\n",
classification_report(y_test, y_pred))

# Plot logistic regression curve
x_values = np.linspace(df['Study Hours'].min(), df['Study
Hours'].max(), 100).reshape(-1, 1)
```

```
y_proba = model.predict_proba(x_values)[:, 1]  # Probability of
Pass

plt.scatter(df['Study Hours'], df['Pass'], color='blue',
s=100, label='Actual')
plt.plot(x_values, y_proba, color='red', linewidth=2,
label='Logistic Regression (Prob)')

plt.xlabel('Study Hours')
plt.ylabel('Pass Probability')
plt.title('Logistic Regression: Study Hours vs Pass')
plt.legend()
plt.ylim(-0.1, 1.1)
plt.grid(True, linestyle='--', alpha=0.5)
plt.show()
```

Conclusion / Discussion

The logistic regression model accurately predicted pass/fail outcomes on the test set, achieving perfect accuracy. Its probabilistic output allows flexible thresholding for classification decisions. The plotted curve shows the S-shaped sigmoid function mapping study hours to pass probabilities, clearly separating the classes.

Actual Output

`text`

```
Accuracy: 1.0
Confusion Matrix:
 [[1 0]
  [0 1]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         1
           1       1.00      1.00      1.00         1

    accuracy                           1.00         2
   macro avg       1.00      1.00      1.00         2
```

```
weighted avg       1.00      1.00      1.00         2
```

Graph / Analysis

The scatter plot and logistic regression curve visualize how study hours influence the probability of passing. The sigmoid curve fits smoothly between fail and pass groups, illustrating the decision boundary and model confidence in classification.