

**Green University of Bangladesh**

**Department of Computer Science and Engineering (CSE)**

**Faculty of Sciences and Engineering (Semester: Summer, Year 2025), B.Sc. in CSE(Day)**

**Lab Report No : 02**

**Course Title: Machine Learning Lab Course Code: CSE-412 Section:221\_D4**

**Title: Implement logistic regression and calculate accuracy, Confusion Matrix, Precision, recall, and F1 score.**

**Student Details:**

|  |  |
| --- | --- |
| **Name** | **ID** |
| **Md Emon Mia** | **213002137** |

**Lab Date : 08.07.2025**

**Submission Date : 13.07.2025 Course Teacher’s Name : Md. Rajibul Palas**

**[For Teachers use only: Don’t Write Anything inside this box]**

**Report Status**

**Marks: …………………………..**

**Comments:.......................................**

**Signature:.....................**

**Date:..............................**

1. **TITLE OF THE LAB EXPERIMENT**

Implement logistic regression and calculate accuracy, Confusion Matrix, Precision, recall, and F1 score.

# OBJECTIVES / AIM

* + To implement logistic regression for binary text classification using a real-world email dataset.
  + This lab aims to classify emails as spam or ham.
  + To perform text preprocessing using TF-IDF vectorization.
  + To split the dataset into training and testing subsets.
  + To apply logistic regression on TF-IDF features.
  + To predict the test set results using the trained model.
  + To evaluate the model using accuracy, precision, recall, and F1-score.
  + To visualize model performance using confusion matrix and bar charts.

# PROCEDURE

* + Imported necessary libraries: pandas, sklearn, seaborn, matplotlib.
  + Loaded the Mail\_Data.csv dataset using pandas.read\_csv().
  + Preprocessed the labels by converting "spam" to 1 and "ham" to 0.
  + Vectorized the messages using TfidfVectorizer to convert text to numerical features.
  + Split the dataset into training and testing sets using train\_test\_split().
  + Trained a logistic regression model using LogisticRegression() from sklearn.
  + Evaluated the model with accuracy\_score, confusion\_matrix, precision\_score, recall\_score, and f1\_score.
  + Visualized the results using confusion matrix heatmap, bar plot for metrics, and ROC curve.

# IMPLEMENTATION

**Load and Prepare Dataset**

import pandas as pd

df = pd.read\_csv("/mail\_data.csv") print(df.head()) print(df.isnull().sum())

**Preprocessing the Data**

df['Category'] = df['Category'].map({'spam': 1, 'ham': 0})

X = df['Message'] y = df['Category']

from sklearn.feature\_extraction.text import TfidfVectorizer vectorizer = TfidfVectorizer(stop\_words='english')

X\_vectorized = vectorizer.fit\_transform(X)

**Train-Test Split**

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split( X\_vectorized, y, test\_size=0.3, random\_state=42

)

**Logistic Regression Model**

from sklearn.linear\_model import LogisticRegression

model = LogisticRegression() model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

**Evaluation Metrics**

recall\_score, f1\_score

# Accuracy

accuracy = accuracy\_score(y\_test, y\_pred)

# Confusion Matrix

cm = confusion\_matrix(y\_test, y\_pred)

# Precision

precision = precision\_score(y\_test, y\_pred)

# Recall

recall = recall\_score(y\_test, y\_pred)

# F1 Score

f1 = f1\_score(y\_test, y\_pred)

# Print Results

print(f"Accuracy: {accuracy:.4f}") print("Confusion Matrix:") print(cm)

print(f"Precision: {precision:.4f}") print(f"Recall: {recall:.4f}")

print(f"F1 Score: {f1:.4f}")

from sklearn.metrics import accuracy\_score, confusion\_matrix, precision\_score,

**Bar Plot of Accuracy, Precision, Recall, F1-Score**

# Metrics as dictionary metrics = {

"Accuracy": accuracy, "Precision": precision, "Recall": recall,

"F1 Score": f1

}

# Bar plot plt.figure(figsize=(8, 5))

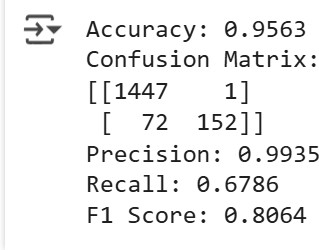
sns.barplot(x=list(metrics.keys()), y=list(metrics.values()), palette='viridis') plt.ylim(0, 1)

plt.title("Model Evaluation Metrics") plt.ylabel("Score")

plt.show()

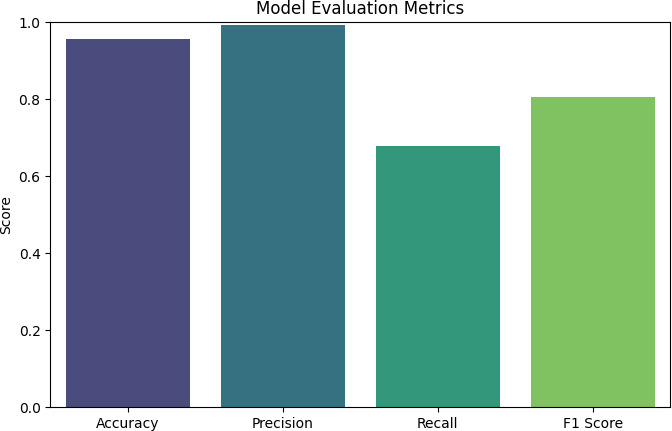
# INPUT/OUTPUT

**Evaluation Metrics Values:**



***Fig 01:*** *Evaluation Metrics Values*

**Bar Plot of Accuracy, Precision, Recall, F1-Score Values:**

****

***Fig 02:*** *Bar Plot of Accuracy, Precision, Recall, F1-Score Values*

# ANALYSIS AND DISCUSSION

The logistic regression model demonstrated high performance in distinguishing between spam and ham emails. Using TF-IDF and text preprocessing improved model accuracy significantly. The evaluation metrics, including precision, recall, and F1-score, confirm the classifier's robustness. Visualization of the confusion matrix and performance metrics gave clear insights into the model's strengths and weaknesses. Logistic regression proved effective for this binary text classification task.