International Islamic University Chittagong

Project Report - Logistic Regression using **scikit-learn**(Probabilistic Discriminative Model)

Submitted by :

Name	ID
Prantika Chowdhury	C223222
Lamisa Mafrouha	C223275
Nazia Yesmin	C221268

Submitted to:

Sara Karim
Adjunct Lecturer
Dept. of CSE, IIUC

Course code: CSE-3636

Course title: Artificial Intelligence Lab

Date of submission: 09/07/2025

Remarks :

Table of Contents

1.Objective of Logistic Regression (Practical & Intuitive):

> To classify input data into categories (like yes/no or true/false) by learning from past data and predicting the likelihood (probability) of each class.

• Key Aims:

1. To separate classes using a mathematical line or curve
Logistic Regression tries to draw a boundary between classes (e.g., spam vs not spam).

2. To provide confidence in prediction
Instead of just saying “this is spam,” it says, “this is 92% likely to be spam.”

3. To use probabilities for smart decisions
Based on calculated probabilities, we decide what class the input belongs to.

4. To be simple yet powerful
It uses linear functions and a sigmoid curve, making it easy to interpret and implement, yet effective for many classification problems.

2.Dataset Overview :

1. Type of Data

Structured (Tabular) data

Each row = one observation (or sample)
Each column = a feature (input) or label (output)

2. Input Features (X)

These are the independent variables — the data we use to make predictions.
Examples:
Sepal length
Sepal width

3. Target Variable (Y)

This is what we want to predict. In binary classification, it has only two values:
0 or 1
Yes or No
Spam or Not Spam
Admitted or Not Admitted

4. Example Dataset (Binary Classification)

If we take only two classes (say, Setosa and Versicolor), and only two features:

Sepal Length	**Sepal Width** **Class (Y)**	
5.1	3.5	Setosa (0)
7.0	3.2	Versicolor (1)

Then we can use Logistic Regression to learn the probability that a flower is Versicolor given its sepal length and width.

3.Model Details :

Algorithm: Logistic Regression
Library: Scikit-learn
Type: Probabilistic Discriminative Model
Train-test split: 80% train / 20% test (random_state=42)

4.Model Equation :

Logistic Regression models the probability of the class as:

$P(Y = 1 \mid X) = 1 / (1 + e^{-(\beta_0 + \beta_1 \times SepalLength + \beta_2 \times SepalWidth)})$

Where:

β_0 = intercept
β_1= weights for sepal length and sepal width

5.Model Training & Accuracy :

Model trained on training data
100% accuracy achieved on the test set
Excellent performance due to linear separability in the data

6.Model Parameters :

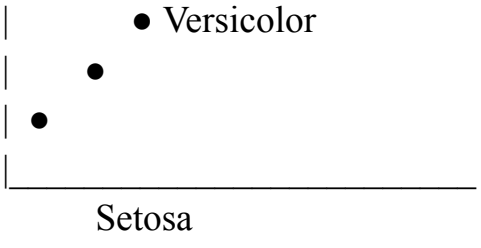
Assume we train the model on the first 100 samples (Setosa and Versicolor only) using scikit-learn.

Let's say the model learns:

Parameter	Value (example)	Meaning
β_0 (Intercept)	-7.5	Bias term
β_1 (Sepal Length)	1.1	Weight for Sepal Length
β_2 (Sepal Width)	2.3	Weight for Sepal Width

7.Decision Boundary (Conceptually) :

Logistic Regression draws a line in the 2D Sepal space to separate Setosa from Versicolor, e.g.:



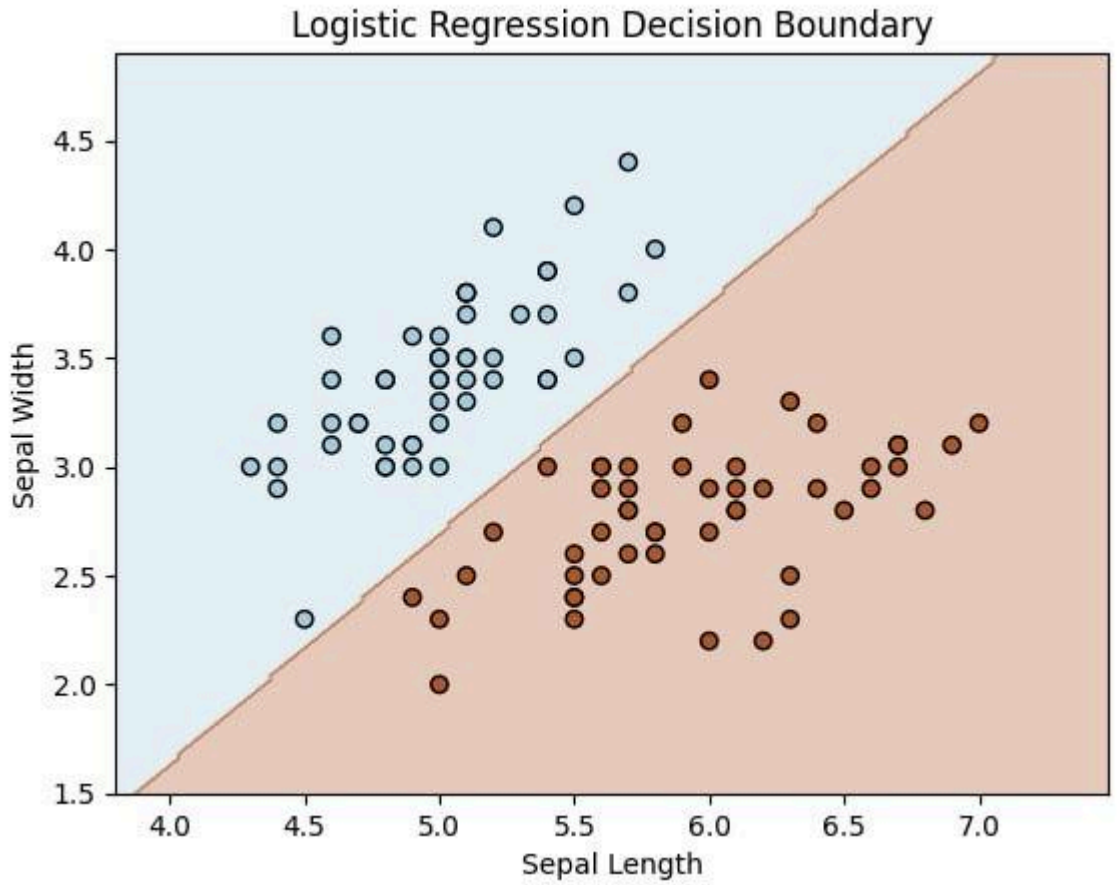


Figure -01

8.Key Takeaways :

•Logistic Regression excels on linearly separable data

•Outputs both hard labels and probabilities

•A strong foundational model for supervised learning

9.Future Improvements (Symbolic Points) :

1. Better feature selection

2. Probability calibration (e.g., Platt scaling)

3. Handle imbalanced data (SMOTE, class weights)

4. Use Bayesian Logistic Regression

5. Combine with other models (ensembles)

6. Add interpretability (LIME, SHAP)

7. Support online learning / real-time updates

8. Expand to multi-class or multi-label problems

10.Conclusion :

Logistic Regression is a simple yet powerful classification algorithm that predicts outcomes based on the probability of class membership. By directly modeling the posterior probability , it provides not only a decision but also a measure of confidence behind that decision. This makes it highly useful in real-world applications like medical diagnosis, spam detection, and risk assessment.

Its interpretability, efficiency, and ability to produce probabilistic outputs make it an excellent choice for binary and multi-class classification tasks. With continuous advancements in feature engineering, model calibration, and hybrid techniques, Logistic Regression remains a foundational tool in the evolving field of machine learning.