Logistic Regression

Group 5

WHAT IS LOGISTIC REGRESSSION?

Logistic regression is a supervised machine learning algorithm used for binary classification—i.e., classifying data into one of two categories such as spam/not spam, pass/fail, disease/no disease.

The output is either 0 or 1, but logistic regression actually predicts the probability that the outcome belongs to class 1 (i.e., the "positive class").

The sigmoid function is what is used to map values to a probability (0-1 range).

HOW DOES LOGISTIC REGRESSION WORK?

1 Calculates a weighted sum of inputs

First, logistic regression computes a weighted sum of the input features. Each input feature—like age, income, or GPA—is multiplied by a coefficient, and we add a bias term. The result ‘z’ is a number between negative and positive infinity

2 Applies the sigmoid function to convert z into a probability:

Instead of directly using ‘z’, we pass it into the sigmoid function, which squashes it into a value between 0 and 1. This is important because it lets us think in terms of probabilities—like, ‘there’s an 85% chance this student will pass’ or ‘a 20% chance this email is spam

3 Class prediction

We finally make a decision

- If $p \geq 0.5$, classify as 1
- If $p < 0.5$, classify as 0

UNDERSTANDING THE SIGMOID FUNCTION

What is Logistic Regression?

- Takes any real number and squashes it between 0 and 1 - Converts raw output into a probability

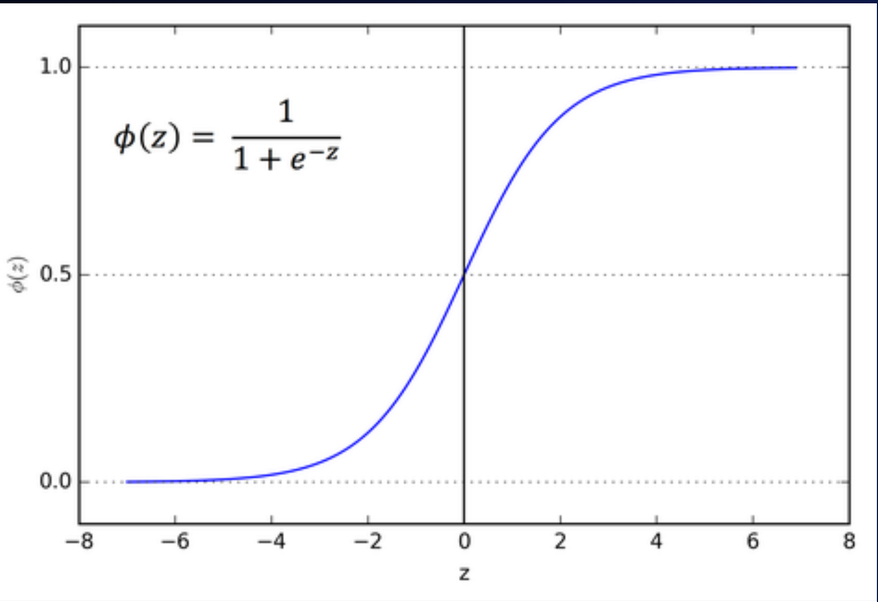
Sigmoid function

- σ(z) is the output between 0 and 1
- z is the linear combination of input features
- e is Euler's number (~2.718)

Why is it Important

$$P(y = 1 \mid x) = \sigma(w^T x + b)$$

- This gives us the probability that the output y is 1, given input x.
- If the result is greater than 0.5, we predict class 1
- If it's less than 0.5, we predict class 0



LOGISTIC REGRESSION VS LINEAR REGRESSION

Predict **continuous** values	Predict **categorical** outcomes (e.g. 0 or 1)
Real numbers (–∞ to +∞)	Probability (0 to 1)
$y = w^T x + b$	$P(y = 1 \mid x) = \frac{1}{1+e^{-(w^T x+b)}}$
Regression (e.g. house prices)	Classification (e.g. spam detection)
Mean Squared Error (MSE)	Log Loss (Cross-Entropy Loss)
Models a linear relationship	Uses a **sigmoid** on linear input

BUILDING A LOGISTIC REGRESSION MODEL

Collect and preprocess the data.

- Gather relevant data (e.g., GPA, family income)
- Clean it (remove missing or irrelevant values)
- Normalize numerical features and encode categorical ones (e.g., "rural"/"urban" as region of abode).

Split the data into training and test sets

- Divide data (e.g., 80% train, 20% test).
- Train set teaches the model; test set checks performance on unseen data.
- Optional: include a validation set for fine-tuning

Train the model using the training data

- Use LogisticRegressionCV to learn patterns
- Cross-validation helps find the best prediction
- Model predicts probabilities of scholarship eligibility

Evaluate model performance using metrics

- Accuracy: overall correctness
- Validation precision: how many predicted eligible were truly eligible
- Validation recall: how many truly eligible were correctly identified
- Confusion matrix shows detailed prediction breakdown

PERFORMANCE METRICS

Accuracy: Overall correctness.

- Measures how often the model is right.
- Formula: (Correct Predictions) ÷ (Total Predictions).

Precision and Recall: Performance in identifying true positives

- Precision: Out of all predicted eligible, how many were actually eligible?
- Recall: Out of all actually eligible, how many were correctly predicted?

Confucion matrix: Shows classification performance.

- Metrics help assess how well the model distinguishes between eligible and not eligible.
- Useful for improving model decisions and fairness.

HOW LOGISTIC REGRESSION CAN BE APPLIED

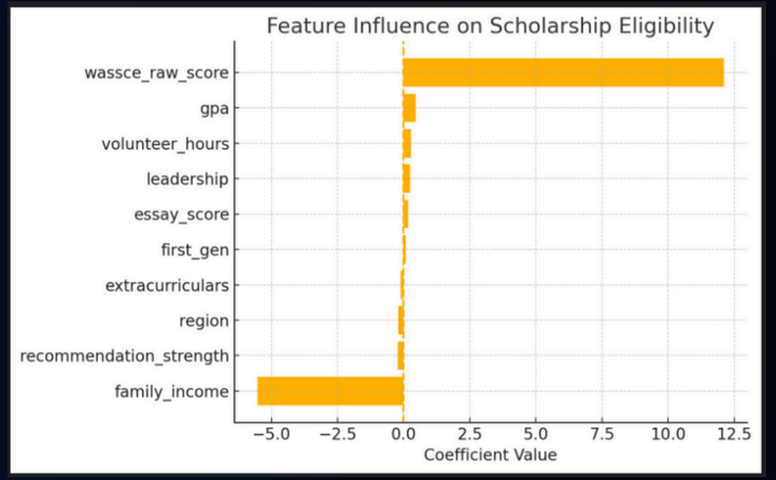
Dataset Prep & Features

- Dropped duplicates & nulls
- Removed sat_score, admission
- Encoded region numerically
- Min-Max scaled all other inputs

Features Used

gpa, family_income, extracurriculars, leadership, volunteer_hours, essay_score, recommendation_strength, region, first_gen, wassce_raw_score

Feature Impact



Model Formula

$$p(\text{eligible}) = \frac{1}{1 + e^{-(w_0 + \sum_i w_i x_i)}}$$

Example Prediction

Feature	Values
GPA	0.85
FAMILTY INCOME	0.10
VOLUNTEER HOURS	0.75
ESSAY SCORE	0.85
WASSCE SCORE	0.80
PREDICTED ELIGIBILITY	1.0

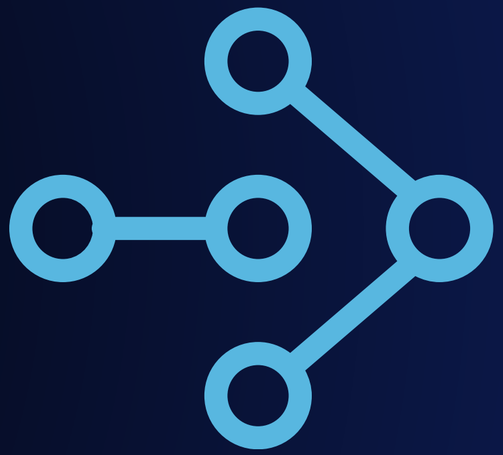
WHY USE LOGISTIC REGRESSION?

Pros

- Easy to use and understand
- Fast to train
- Good for binary outcomes
- Gives probability estimates

Cons

- Assumes linear relationships
- Doesn’t handle complex patterns well
- Needs feature scaling
- Sensitive to outliers



CONCLUSION