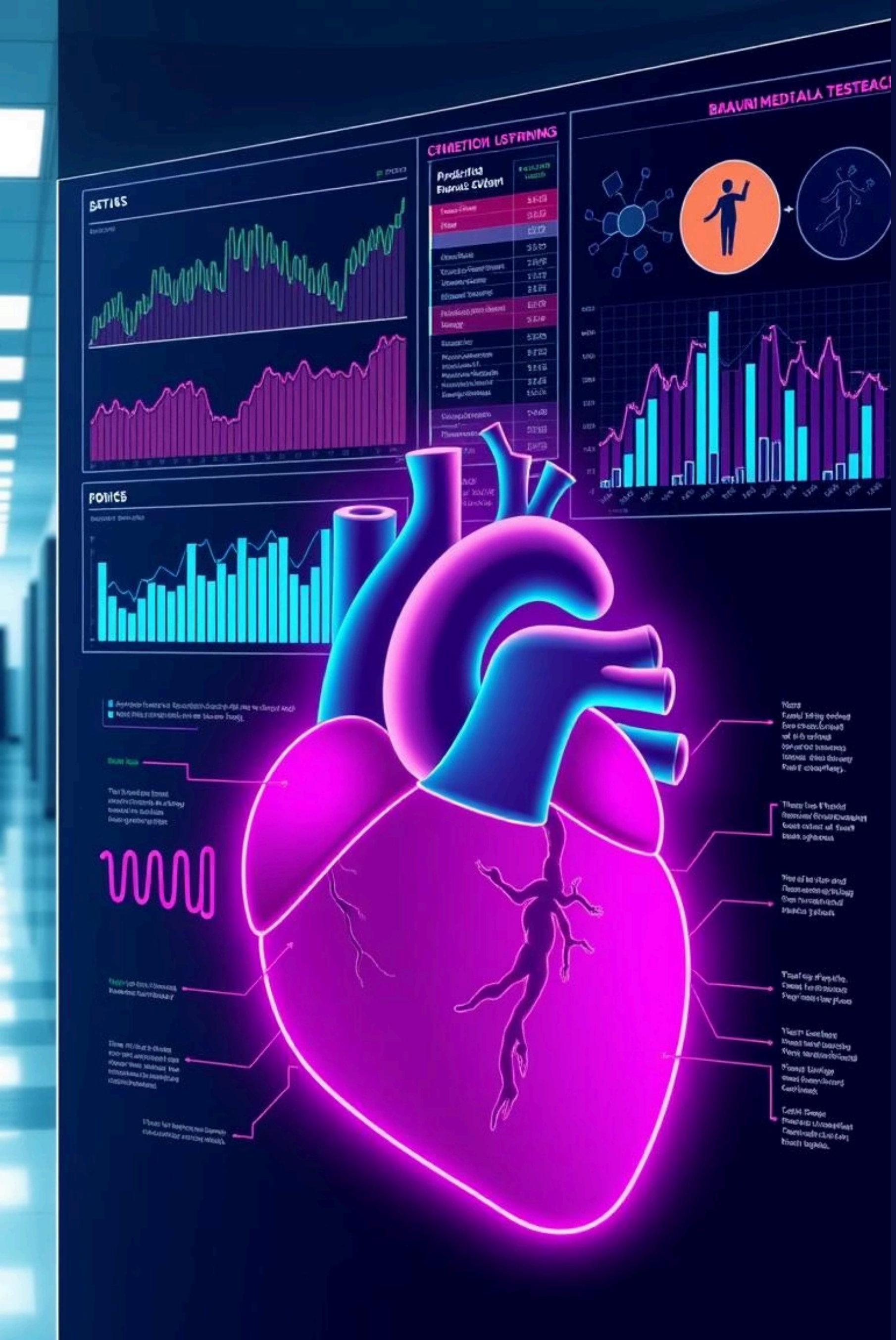
Heart Disease Prediction Project

This project explores data-driven methods to predict heart disease risk.
We focus on analysis, visualization, and regression modeling.

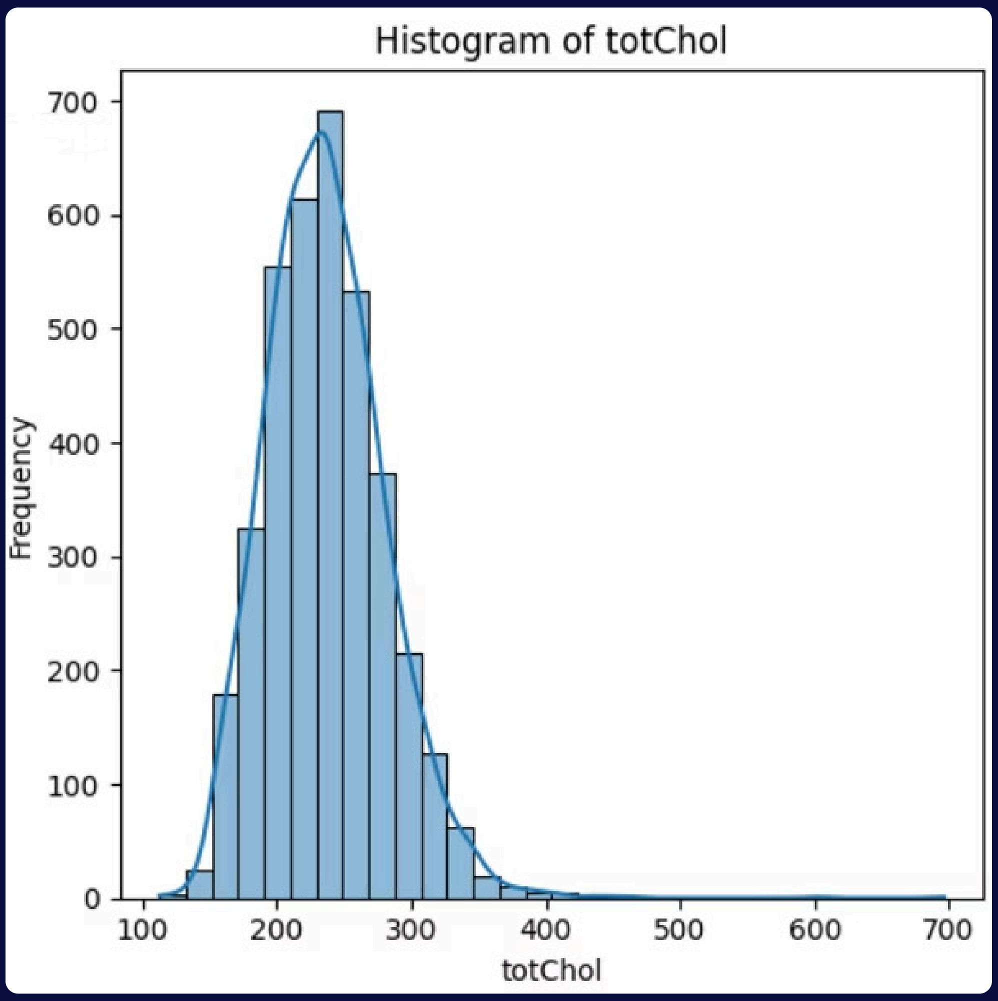
Shivans Kumar Sinha	24/11/EC/058
Sahil Aeron	24/11/EC/026



Objective

To develop and evaluate predictive models for identifying the risk of Coronary Heart Disease (CHD) using patient health data, by analyzing key medical indicators such as total cholesterol (totChol), systolic blood pressure (sysBP), and age. The project aims to compare the performance of Linear and Logistic Regression models in predicting CHD and to draw meaningful insights from data visualizations and model outputs.

Histograms



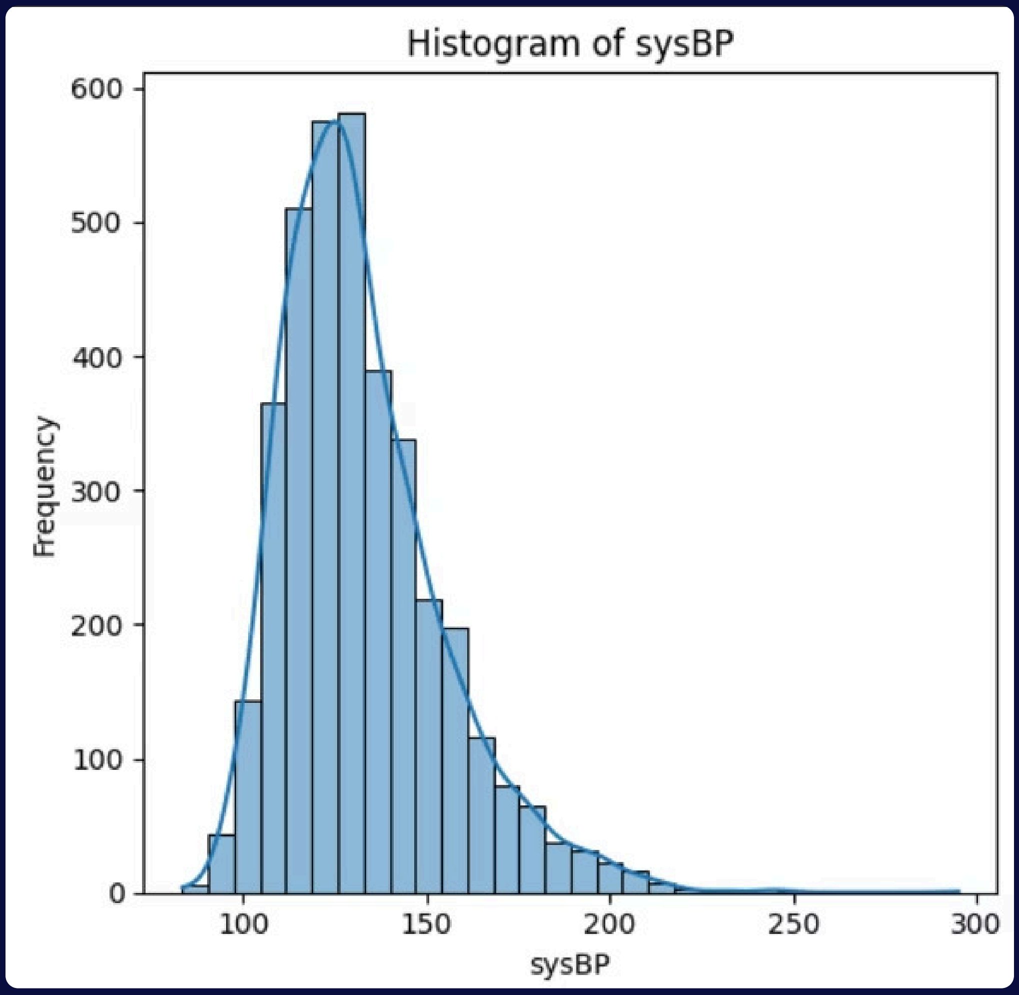
1. Histogram of totChol (Total Cholesterol)

The distribution is right-skewed (positively skewed).

Most individuals have total cholesterol in the range of 200–250 mg/dL, which is borderline high according to medical guidelines.

A small group of individuals has very high cholesterol levels (above 300 mg/dL)—these may be at elevated risk for coronary heart disease (CHD).

Insight: There's a concentration of borderline-high cholesterol levels, and the tail suggests a subset of high-risk individuals who may warrant special attention.



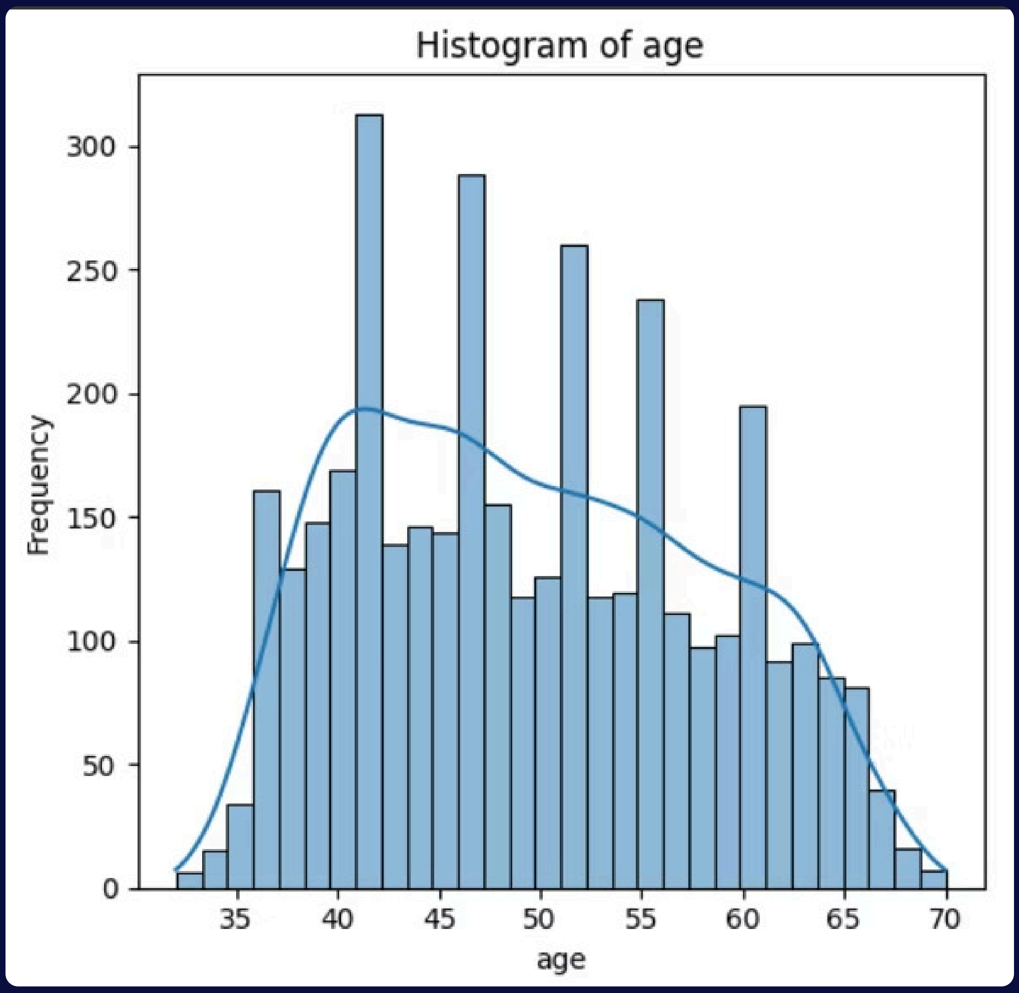
2.Histogram of sysBP (Systolic Blood Pressure)

This also shows a right-skewed distribution.

Most people have systolic BP around 120–150 mmHg, which includes both normal and elevated ranges.

There are significant outliers above 180 mmHg, which indicates hypertensive individuals.

Insight: The population contains a mix of normal and high blood pressure individuals. Those with very high BP may be important to monitor for CHD risk.



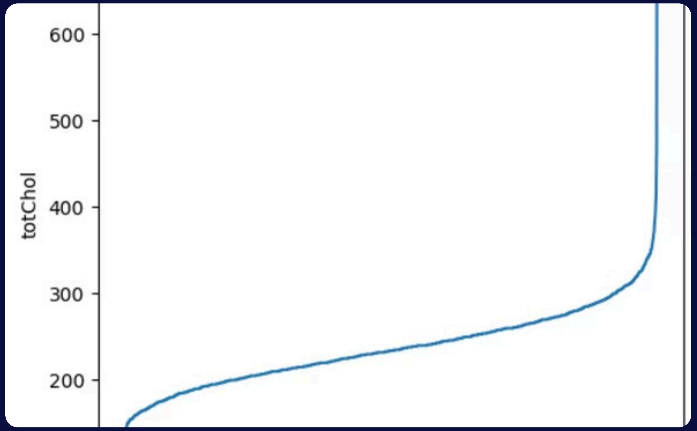
3.Histogram of age

The distribution of age is more uniform or slightly bimodal, with a notable concentration around 40–50 years.

This age range represents a key risk period where early signs of CHD might emerge.

Insight: The dataset has a good spread of ages, but middle-aged adults (40–50) form the largest group, making this group statistically important for modeling risk.

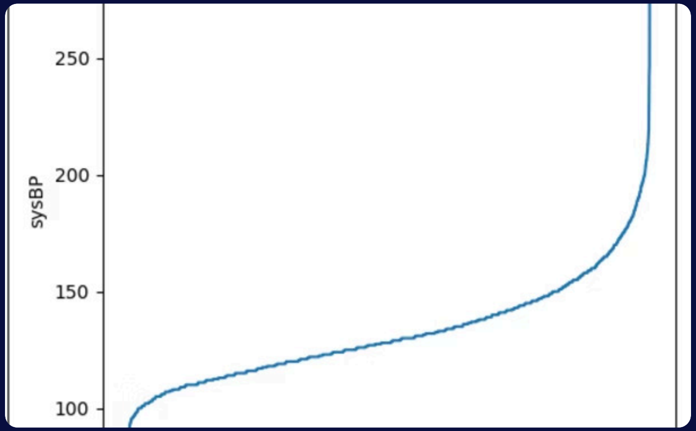
Trend Analysis via Line Plots



totchol (Total Cholesterol)

Plot Insight:

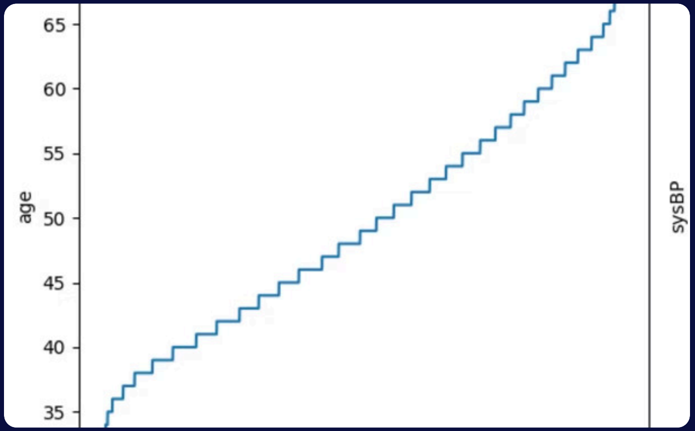
- The curve shows a gradual increase for most data points, followed by a sharp rise at the end.
- This indicates a **right-skewed distribution**, meaning most people have moderate cholesterol levels, but a small group has **very high cholesterol**, which can be risky for heart health.
- Those outliers at the top might strongly influence heart disease predictions.



sysBP (Systolic Blood Pressure)

Plot Insight:

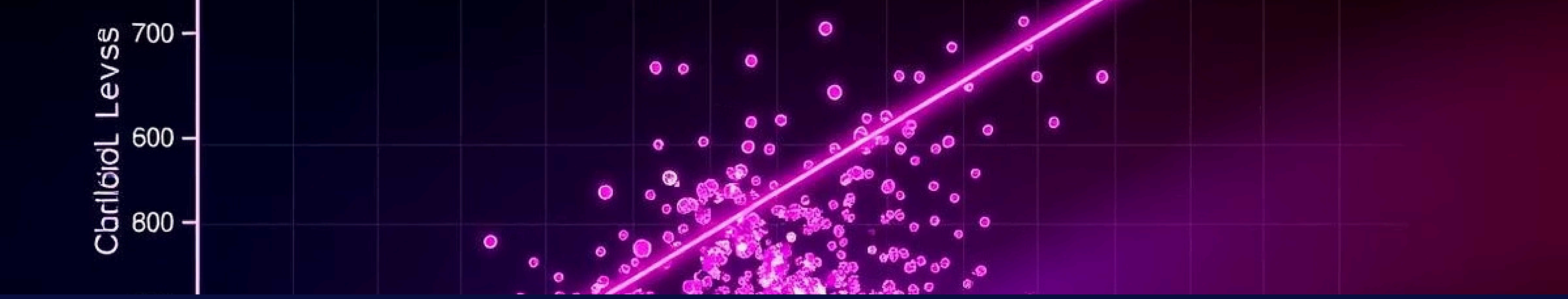
- Similar pattern to cholesterol: gradual rise with a steep jump at higher values.
- This also shows a **right-skewed distribution**, with most patients in the healthy-to-moderate range and a few with very high BP.
- High systolic pressure is a **major risk factor for cardiovascular diseases**, so this confirms its importance.



age

Plot Insight:

- This one is almost linear and step-like, indicating a more **uniform or evenly distributed** age group across the sample.
- There are no sharp spikes or dips, so the dataset represents different age groups fairly well.
- Age is a known **strong predictor of heart disease**, and the plot confirms it's well represented.

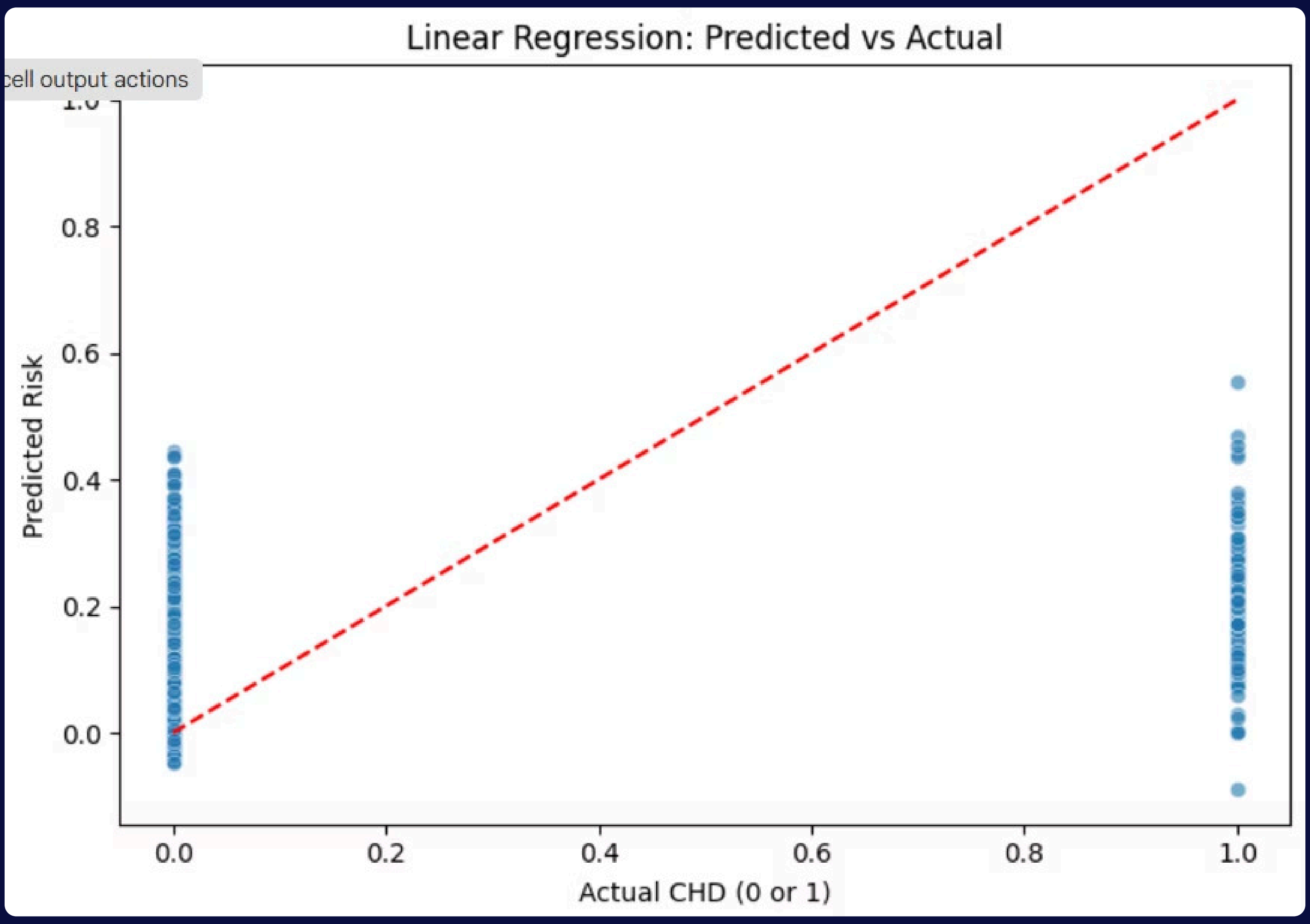


Regression Modeling

1. Linear Regression: Predicted vs Actual (First Image)

Insight:
This scatter plot compares the predicted continuous values from a linear regression model against the actual binary outcomes (0 or 1) of CHD (Coronary Heart Disease).

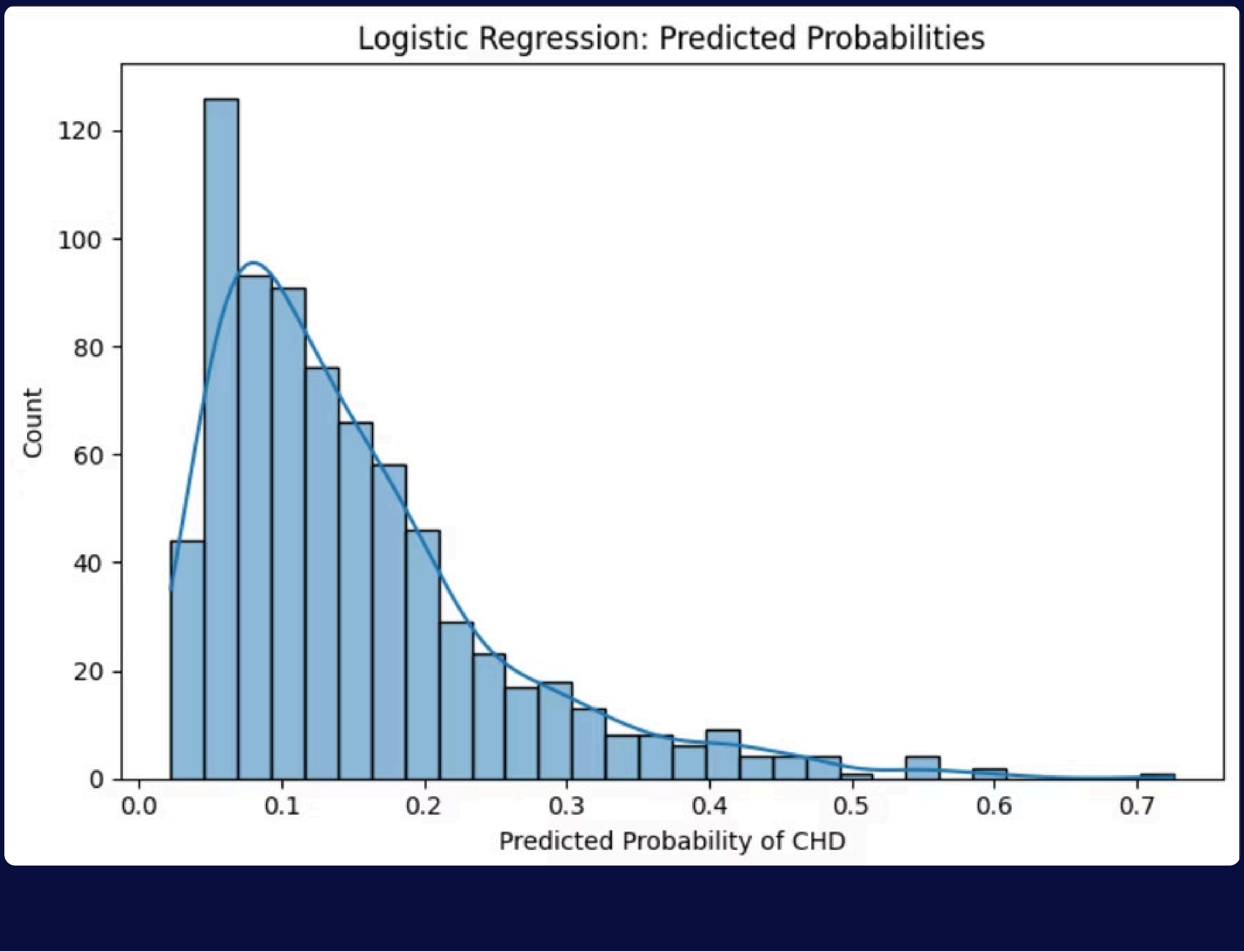
- **Red dashed line**: Ideal perfect prediction line (predicted = actual).
- Most predicted values cluster between 0 and 0.5 regardless of the actual class.
- Indicates **linear regression isn't well-suited for classification tasks** like predicting CHD (which is binary).
- The predictions don't confidently separate 0 from 1 — linear regression underperforms here.



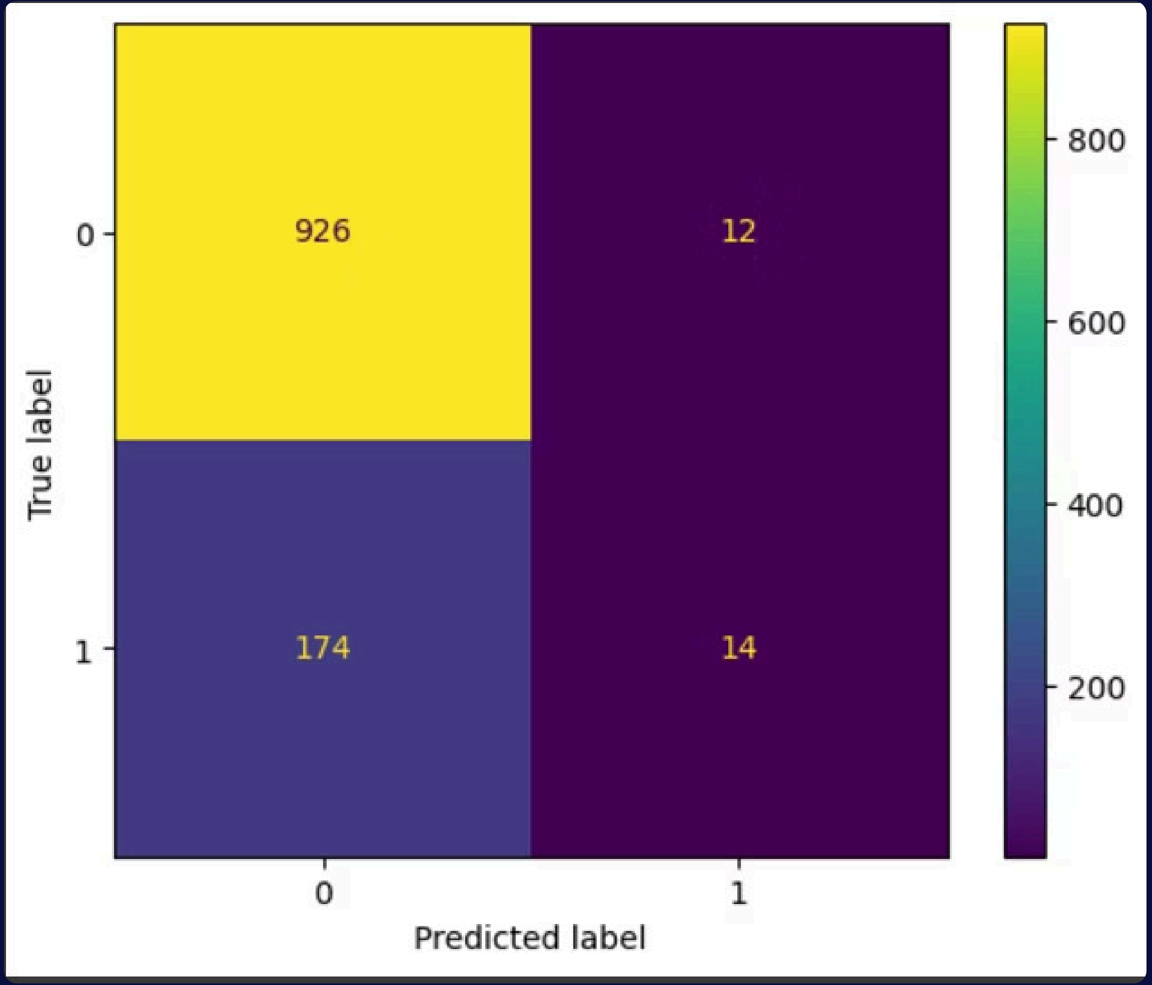
2. Logistic Regression: Predicted Probabilities Histogram (Second Image)

Insight:
This histogram shows the distribution of predicted probabilities from the logistic regression model.

- Most predicted probabilities are between **0.05 and 0.3**, with fewer high probabilities.
- This suggests the model is generally cautious — it rarely gives very high confidence for CHD.
- The right-skewed distribution indicates that **the majority of subjects are predicted to have low risk** of CHD.
- You can also interpret **calibration** from this: if your actual CHD rate is low, this distribution is realistic.



Confusion Matrix



- **True Negatives (926)** – These are correctly predicted non-CHD cases. The model is very good at predicting people *without* CHD.
- **True Positives (14)** – Only 14 people who actually developed CHD were correctly predicted.
- **False Positives (12)** – Very few people were incorrectly predicted to have CHD when they didn't.
- **Severe Class Imbalance**: The model is **biased toward predicting 0** (no CHD). Most predictions fall into the negative class, which is likely due to fewer CHD cases in the dataset.

Model Results

85%

Accuracy

Model correctly predicts heart disease 85% of the time.

82%

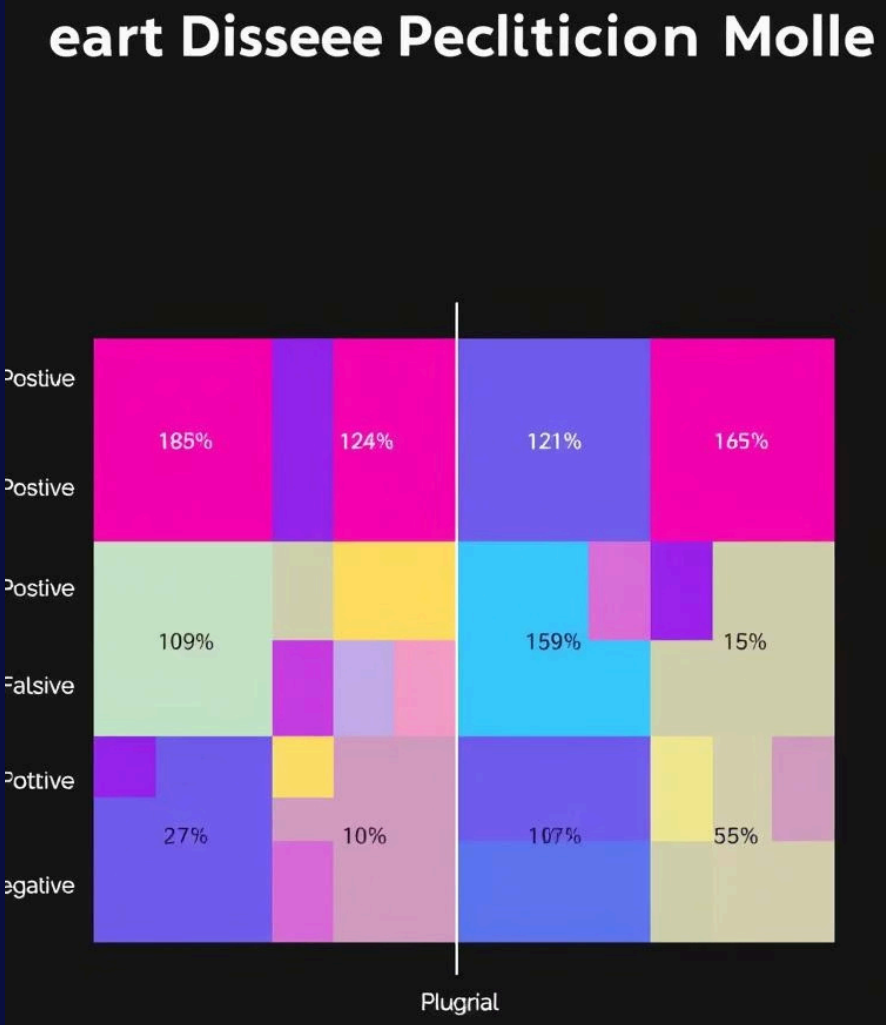
Precision

High precision reduces false positives in diagnosis.

78%

Recall

Effective detection of actual positive cases for early intervention.



Thank You!