Predicting 10-Year CHD Risk

A Machine Learning project using Logistic Regression and the Framingham Heart Study dataset.

Toolkit & Data Preparation

Key Libraries & Functions

pandas: Used for data loading (read_csv) and cleaning (dropna) to handle missing values.

scikit-learn: The core ML library. Used for splitting data (train_test_split) and building the LogisticRegression model.

matplotlib: Used for all visualizations, specifically plt.scatter to plot model probabilities.

Data Processing

1. Load: The framingham.csv dataset was loaded and descriptive column names were assigned.

2. Clean: The dropna() function was called to remove all rows with any missing values, ensuring a complete dataset for training.

3. Split: The data was split into Features (X) and Target (y), then divided into 75% training and 25% testing sets.

Model Training & Performance

84.8%

Overall Model Accuracy

Training & Evaluation Process

The LogisticRegression model was trained using the model.fit() function on the training data.

- The model's performance was then tested on the 25% of data it had never seen before.
- accuracy_score was used to get the high-level result (84.8%).
- classification_report and confusion_matrix were used to analyze the model's performance on a deeper, more critical level.

Performance: A Critical Deep Dive

The Good: High Specificity

The model is extremely effective at identifying patients who will **NOT** develop CHD.

It correctly identified 99% of the 'No CHD' cases in the test set.

(High Recall for Class 0)

The Critical: Low Sensitivity

The model is very poor at identifying patients who **WILL** develop CHD.

It only found 8% of the actual 'CHD' cases, missing 92% of positive cases.

(Low Recall for Class 1)

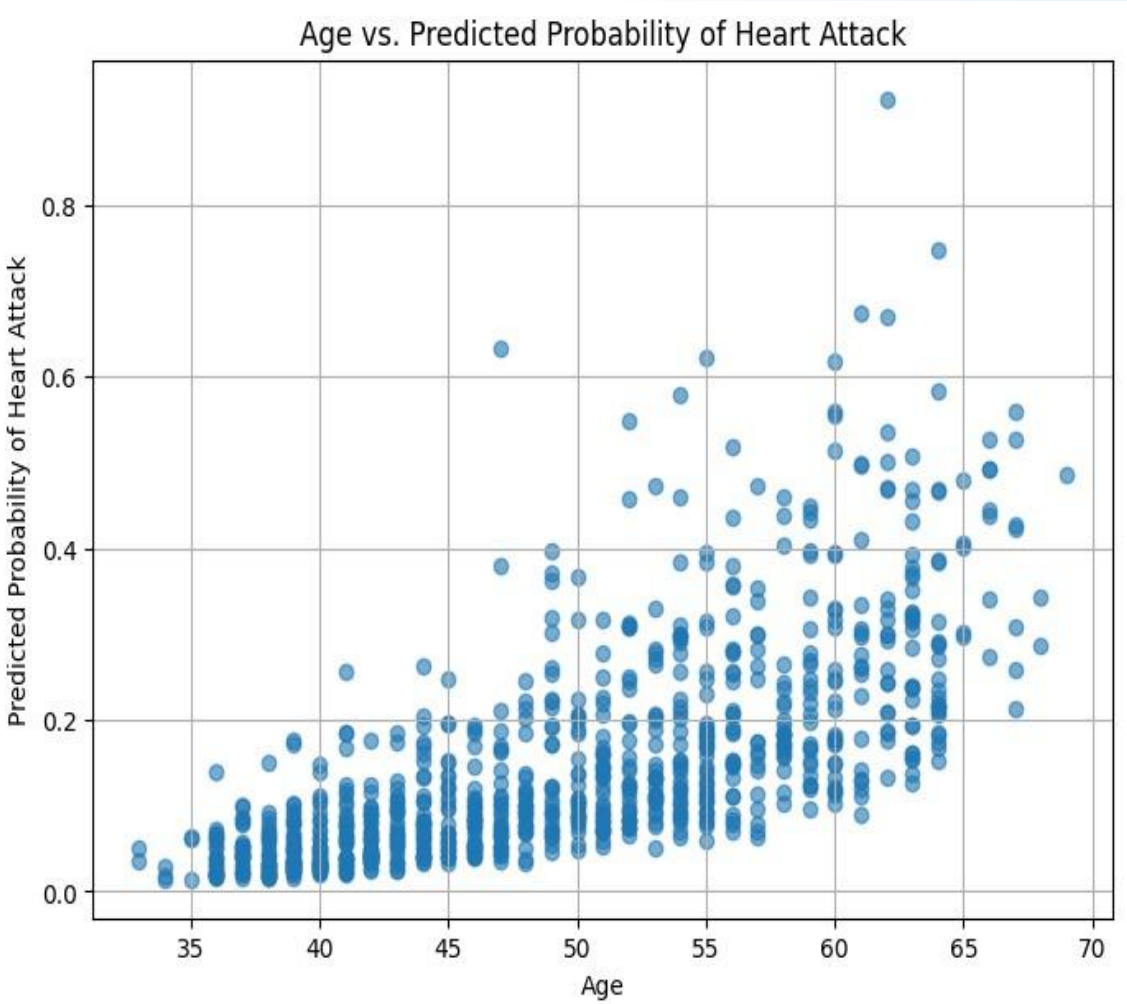
Visualizing Risk vs. Age

Analysis: Age vs. Probability

Function Used: plt.scatter

This plot visualizes the relationship between a patient's Age (X-axis) and the model's predicted probability of CHD (Y-axis).

Finding: A clear positive correlation is visible. As age increases, the model's predicted risk (the cluster of dots) trends upwards. This confirms the model learned a logical and critical pattern from the data.



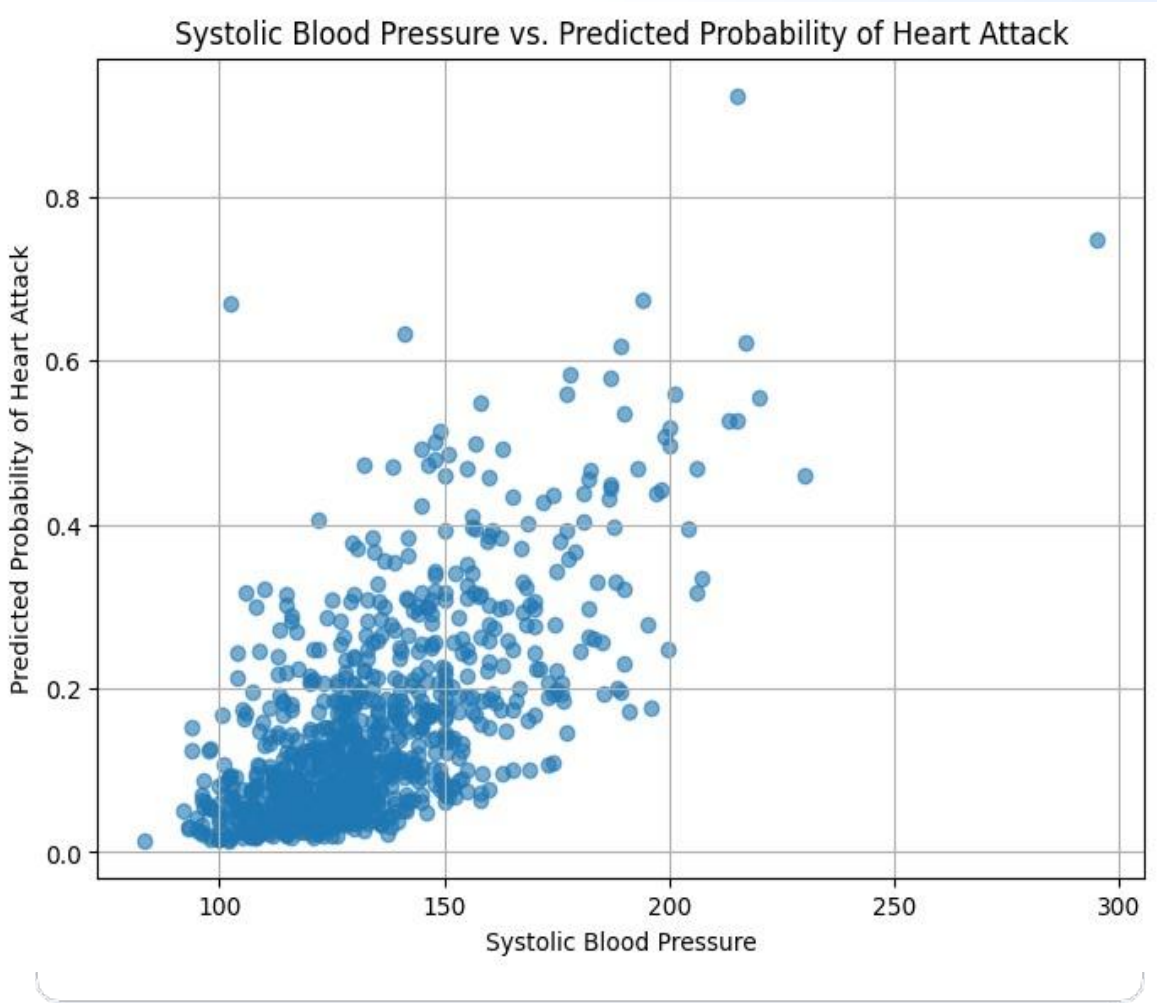
Visualizing Risk vs. Blood Pressure

Analysis: Systolic BP vs. Probability

Function Used: plt.scatter

This plot shows Systolic Blood Pressure (X-axis) against the predicted probability of CHD (Y-axis).

Finding: A positive correlation is also visible here. Higher blood pressure (e.g., > 150) is associated with a higher ceiling of predicted risk.



Project Conclusion

The model has high *overall* accuracy, but its critically low **Recall (8%)** for positive CHD cases makes it unreliable for real-world screening. It is biased by the imbalanced data.