Heart Disease Logistic Regression Synthetic Data Failure

Objective

This project aimed to implement logistic regression to predict heart attack risk using patient health data. While the pipeline was technically correct and all modeling steps were properly followed, the **dataset itself proved unreliable**, resulting in a model that was **flawless in execution but useless in outcome**.

What Worked Flawlessly

All machine learning components were correctly applied:

- ✅ Cleaned the dataset and verified zero missing values

- ✅ Encoded categorical variables properly

- ✅ Transformed blood pressure into numeric features

- ✅ Standardized numerical data

- ✅ Used stratified train-test split

- ✅ Trained a regularized logistic regression model

- ✅ Evaluated using precision, recall, F1-score, accuracy, and AUC

- ✅ Visualized results clearly with boxplots, heatmaps, and coefficient charts

Why the Model Failed

Despite perfect modeling steps, the results revealed a deeper issue:

Symptom	Explanation
Correlation values near zero	Features showed no real relationship with the target variable
Logistic model predicted only class `0`	Class imbalance or lack of pattern
Evaluation metrics were all zero	Model couldn't find a boundary to separate risk from non-risk
Coefficients were small and medically implausible	No true signal to learn from — only noise

Likely Causes

- **Synthetic dataset**: The values look plausible but don't reflect any real clinical relationships.

- **Random label assignment**: The target variable `Heart Attack Risk` appears to be loosely or randomly generated.

- **Overprocessed or AI-generated source**: Structure without substance — classic "flawless garbage."

Conclusion

This project is a perfect case study in why clean modeling isn't enough. A predictive model is only as strong as the signal in the data. This notebook was useful as a demonstration of correct technique and critical evaluation, but it fails as a predictive tool.

More proof if needed:
https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset/discussion/508204