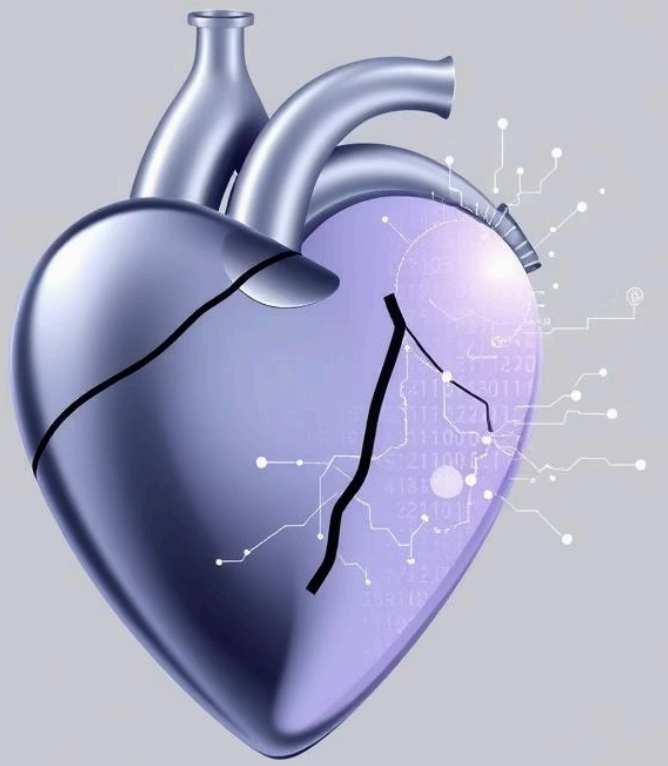
Cardiovascular Disease Prediction: A Data Science Approach



Our Team

MARIAM IBRAHIM MAHMOUD GOMAA

PERIHANE TAREK

MARIAM TAMER FAWZY

MOHAMED NEGM

MAHMOUD ABDELRAZEK

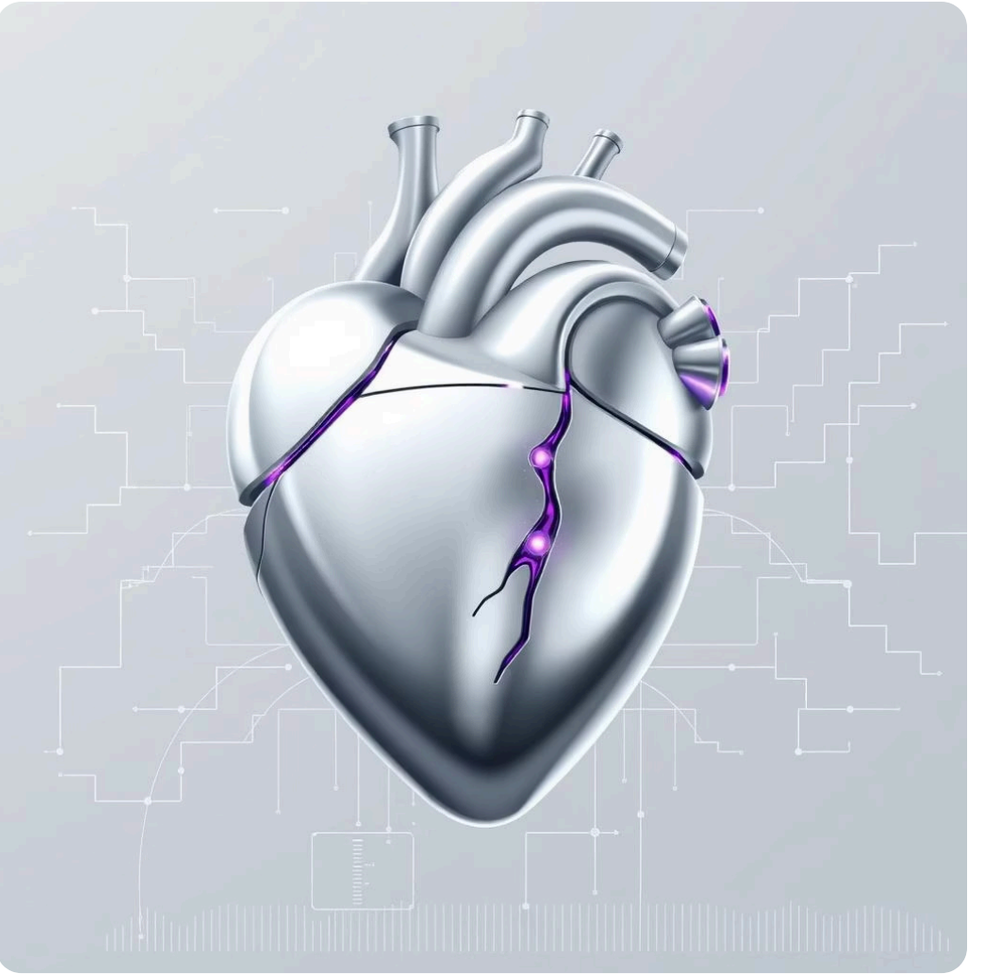
ADHAM ZAKARIA

Background & Motivation

Cardiovascular disease (CVD) covers conditions affecting the heart and blood vessels. Risk factors include modifiable elements (physical activity, smoking) and non-modifiable factors (age, genetics).

Traditional risk assessment often misses complex relationships between these factors.

Machine learning offers a powerful alternative, identifying patterns in large health datasets that human analysis might miss. This allows for earlier and more accurate identification of at-risk individuals.



Project Objectives

01

Data Preparation

Clean and preprocess cardiovascular health data for quality and reliability.

02

Exploratory Analysis

Visualize and understand relationships between health indicators and CVD prevalence.

03

Predictive Modeling

Develop and evaluate machine learning models to predict CVD presence.

04

Model Deployment

Create an accessible API for model predictions.

Dataset Description

Our dataset includes 70,000 patient records with 11 input features and a binary target variable for CVD presence.

Objective Features

- Age (converted to years)
- Height (cm)
- Weight (kg)
- Gender (categorical)

Examination Features

- Systolic blood pressure (ap_hi)
- Diastolic blood pressure (ap_lo)
- Cholesterol level (1-3)
- Glucose level (1-3)

Subjective Features

- Smoking status (binary)
- Alcohol intake (binary)
- Physical activity (binary)

Data Preprocessing & Cleaning

1 Data Loading & Exploration

Loaded CSV, confirmed 70,000 records, no missing values. Reformatted for compatibility.

2 Data Quality Assessment

Renamed columns (gluc→glucose, alco→alcohol, active→physically_active). Removed 'id' column.

3 Feature Transformation

Converted age from days to years for intuitive understanding.

4 Outlier Detection & Treatment

Applied IQR method and physiological filters for blood pressure, height, and weight.

5 Feature Engineering

Created BMI, Pulse Pressure, Health Index, Cholesterol-Glucose Interaction, and Hypertension Indicator.

Data Visualization & Exploratory Analysis

We developed a two-tier visualization system: a Jupyter notebook for reproducibility and an interactive Dash dashboard for dynamic exploration by clinical stakeholders.

Key metrics: 68,500 patients, 49.5% CVD prevalence, average age 53.3 years, 42.4% hypertension rate, 8.8% smoking rate.

Key Insights from Analysis

1

Blood Pressure Dominance

Systolic BP and pulse pressure are the strongest predictors.

2

Lifestyle Factor Limitations

Weak correlations for smoking and physical activity suggest data quality issues or non-linear effects.

3

Feature Engineering Success

Engineered features (BMI, pulse pressure, metabolic risk score, hypertension) show strong relationships.

4

Age-Disease Gradient

Strong monotonic relationship between age and CVD prevalence.

5

Class Balance Advantage

Nearly perfect 50-50 split eliminates sampling concerns.

6

Multicollinearity

High correlations between weight-BMI and BP measurements suggest need for regularization.

Machine Learning Model Evaluation

Logistic Regression

Moderate performance (F1=0.757), struggles with complex interactions.

Decision Tree

High training accuracy, but overfitting (F1=0.777) on test set.

Random Forest

Reduces overfitting, better than single tree (F1=0.801).

Gradient Boosting

High accuracy (F1=0.80), corrects mistakes step-by-step.

XGBoost

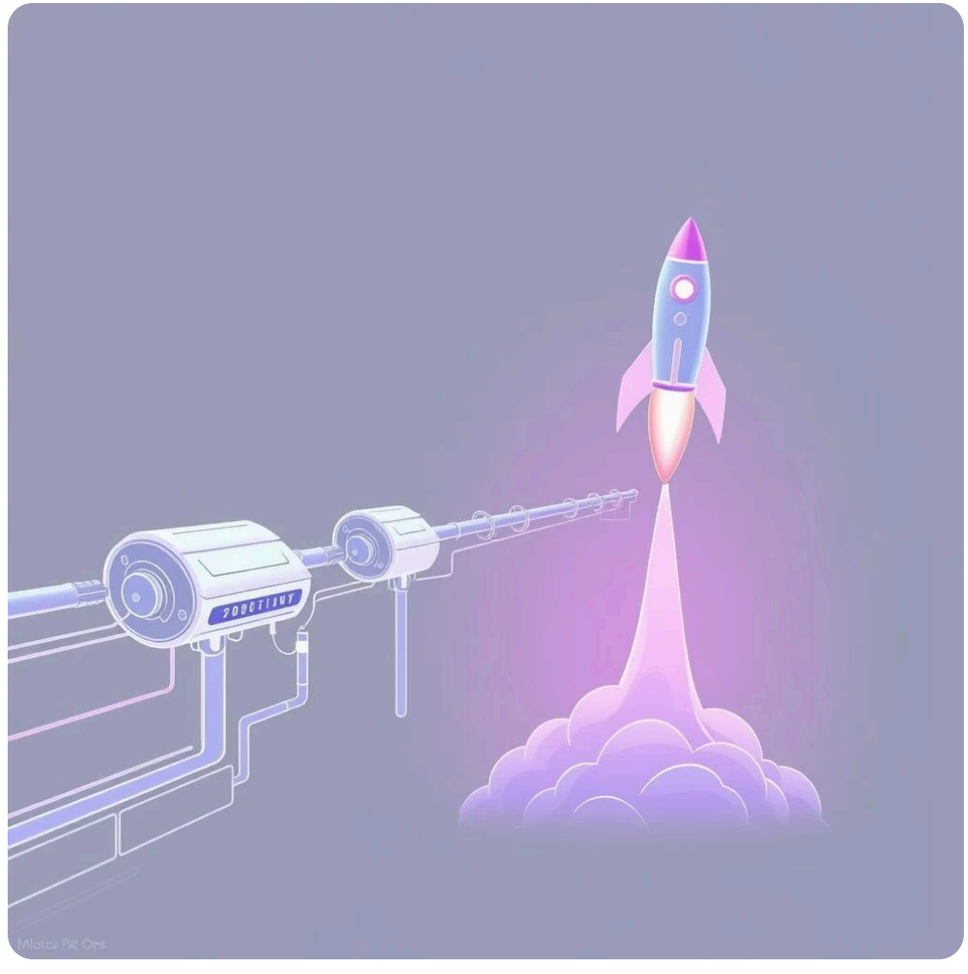
Best overall model (F1=0.814, ROC-AUC=0.909), balances precision and recall.

MLOps & Deployment

The final phase transforms trained models into an operational, scalable, and maintainable system for healthcare professionals.

MLOps, implemented with MLflow, tracks experiments, model versions, hyperparameters, and metrics. This ensures reliability and reproducibility.

The best-performing model (XGBoost) was automatically registered into the MLflow Model Registry, formalizing it for production.



Thank You for Your Attention

At depi we are dedicated to improving healthcare and enhancing lives. We look forward to serving you in 2030 and beyond.