مبادرة رواد مصر الرقمية

DEPI

AI and Data Science

ALX2_AIS4_S5

Final Project Report

Healthcare Predictive Analytics Project

GitHub repository: https://github.com/Muhannad-Khaled/Heart-AI

Submitted by:

Mariam Ayman

Muhannad Khaled

Nourhan Ibrahim

Shrouq Ashraf

1. Executive Summary

Heart disease remains a leading cause of mortality worldwide. Early predictions of cardiovascular risk factors using data-driven methods can significantly enhance healthcare outcomes. This project focuses on developing a predictive analytics application to assist healthcare professionals in assessing the risk of heart disease in patients. Using machine learning models trained on real-world health data, we built a web application capable of predicting patient risk levels based on clinical inputs, visualizing health trends, and enabling model retraining.

The project aimed at:

- Analyze and clean real-world medical data.
- Handle outliers and missing values effectively.
- Build and compare multiple classification models.
- Evaluate performance using metrics such as accuracy, confusion matrix, and AUC-ROC.

2. Project Workflow Overview

Data Collection

Dataset: The project used a publicly available cardiovascular dataset containing 70,000 records of patients' data, 11 features + target, https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.

The primary dataset includes patient records with the following features:

- Demographics: Age, Gender, Height, Weight.
- Medical Metrics: Systolic & Diastolic Blood Pressure, Cholesterol, Glucose.
- Lifestyle Indicators: Smoking, Alcohol Intake, Physical Activity.
- Target Variable: Presence or absence of cardiovascular disease.

Feature	Description
Age	Age of the patient
Sex	Gender (1 = male, 0 = female)
Trestbps	Resting blood pressure
Chol	Serum cholesterol in mg/dl
Thalach	Maximum heart rate achieved
Oldpeak	ST depression induced by exercise
Restecg	Resting electrocardiographic results
Ca	Number of major vessels (0–3) colored by fluoroscopy
Thal	Type of thalassemia
Target	Heart disease presence (0 = No, 1 = Yes)

Data preprocessing involved:

- Handling missing values and outliers using interquartile range (IQR) and domain knowledge.
- Feature encoding categorical variables.
- Normalization and scaling of numerical features.
- Splitting data into training and testing sets.
- Feature selection and engineering to improve model accuracy.

Exploratory Data Analysis (EDA)

EDA was performed to understand the structure and distributions:

- Univariate Analysis: Histograms of age, chol, and trestbps showed the typical spread of values.
- Bivariate Analysis:
  thalach (maximum heart rate) was higher in non-disease patients.
  chol and trestbps were not clearly separable by target alone.

3. Model Development

Data Split: The dataset was split into training and testing sets (80/20).

Model Selection

Three models were evaluated: Logistic Regression, Random Forest, and Gradient Boosting (XGBoost).

Model	Accuracy	AUC-ROC	F1	Recall	Precision
Logistic Regression	0.80	0.85	0.84	0.93	0.77
Random Forest	0.80	Similar	0.84	0.93	0.77
Gradient Boosting	0.77	Lower	0.80	0.82	0.79

Logistic Regression and Random Forest both:

- ✓ Have **higher accuracy (0.80)**.
- ✓ Have **higher recall (0.93)** for class 1 — which is crucial if detecting positives is more important (e.g., medical diagnosis).
- ✓ Share the **same confusion matrix**, suggesting similar performance on this dataset.
- ✓ Logistic Regression has a **higher AUC-ROC (0.85)**, indicating slightly better class separability.

Gradient Boosting:

- ✓ Lower accuracy (0.77) and recall (0.82) for class 1.
- ✓ Slightly better balance in precision/recall, but not enough to outperform the others.

According to this, the best Model is ***Random Forest***; as it has:

- The Highest AUC-ROC (0.85): excellent at distinguishing classes.
- High recall (0.93) for class 1: captures most positives.
- Simpler and interpretable: a strong advantage in regulated fields like healthcare.

4. Application Deployment

Backend

- Flask web framework for routing, form handling, and API endpoints.
- Machine Learning model (trained and serialized using joblib) loaded at runtime.

Modular structures include:

- ✓ modeling/train_model.py: Model training script.
- ✓ processing/preprocessing_utils.py: Preprocessing logic.
- ✓ processing/visualization_utils.py: Data visualization functions.
- ✓ monitoring/logging.py: Prediction logging.
- ✓ config/config.py: Centralized configuration including model path.

Frontend

- HTML templates rendered using Flask's Jinja2 engine.
- User-friendly input form with labeled dropdowns for categorical data
- Data file upload interfaces for visualization and training pages.

5. Functionalities

Prediction

- Input: Patient details (age, gender, blood pressure, glucose, smoking, alcohol, physical activity).
- Backend processes input via a trained model.
- Output: A message indicating whether the patient is at risk of heart disease.

Visualization

- Users upload a CSV file.
- Displays:
  - ✓ Data summary (head, missing values, descriptive stats).
  - ✓ Dataset metadata.
  - ✓ Target balance plot.
  - ✓ Distribution plots for numerical columns.
  - ✓ Correlation heatmap.
  - ✓ Pie charts for categorical features.
  - ✓ Boxplots for outlier analysis.

Model Retraining

- Users upload a new dataset.
- Preprocessing applied.
- Trains a new model and updates the features.pkl file to maintain consistent feature input.
- Model is reloaded after training.

REST API (/predict)

- Accepts JSON input.
- Returns prediction as JSON response.

Authentication

- Basic login/logout using Flask session management.
- Login required to access core functionalities.

6. Technologies Used

Category	Technologies
Web framework	Flask
Ml & data	Scikit-learn, pandas, NumPy
Visualization	Matplotlib, seaborn, plotly
Persistence	Joblib (for model storage)
Web templates	Jinja2
Ui components	Html, CSS, bootstrap
Authentication	Flask sessions
Deployment	Localhost (debug mode), ready for production
Environment	Python 3.11+, pip-managed dependencies

7. Key Challenges

- Data Quality: Required extensive cleaning and mapping for meaningful feature usage.
- Deployment Security: Integrated simple session-based login for access control.
- Handling user input errors: Reverse mapping with input validation and error messages in the UI.

8. Key Insights from the Predictive Model

- Systolic blood pressure, cholesterol, and age were the most influential predictors.
- Male patients and those with poor lifestyle indicators (smoking, alcohol use, physical inactivity) showed higher risk trends.
- Data visualizations revealed age related risk clusters and gender based differences in heart disease prevalence.
- The model achieved high accuracy and balanced performance, making it suitable for decision support.

9. Model Integration Recommendations

For Healthcare Professionals:

- Integrate the model into electronic health record (EHR) systems for automatic risk scoring.
- Use prediction results to prioritize high-risk patients for early intervention.
- Enable continuous model updates by regularly uploading recent patient data.
- Train medical staff on interpreting model predictions and incorporating them into decision-making.

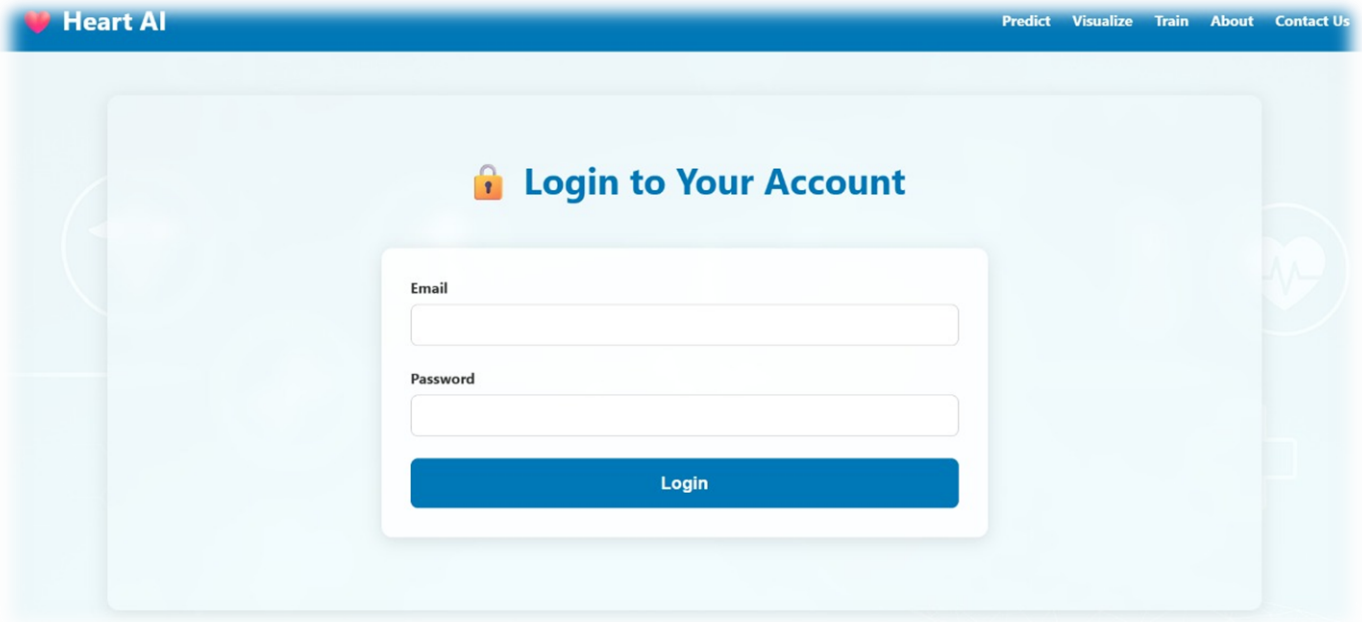
10. Conclusion and Future Improvements

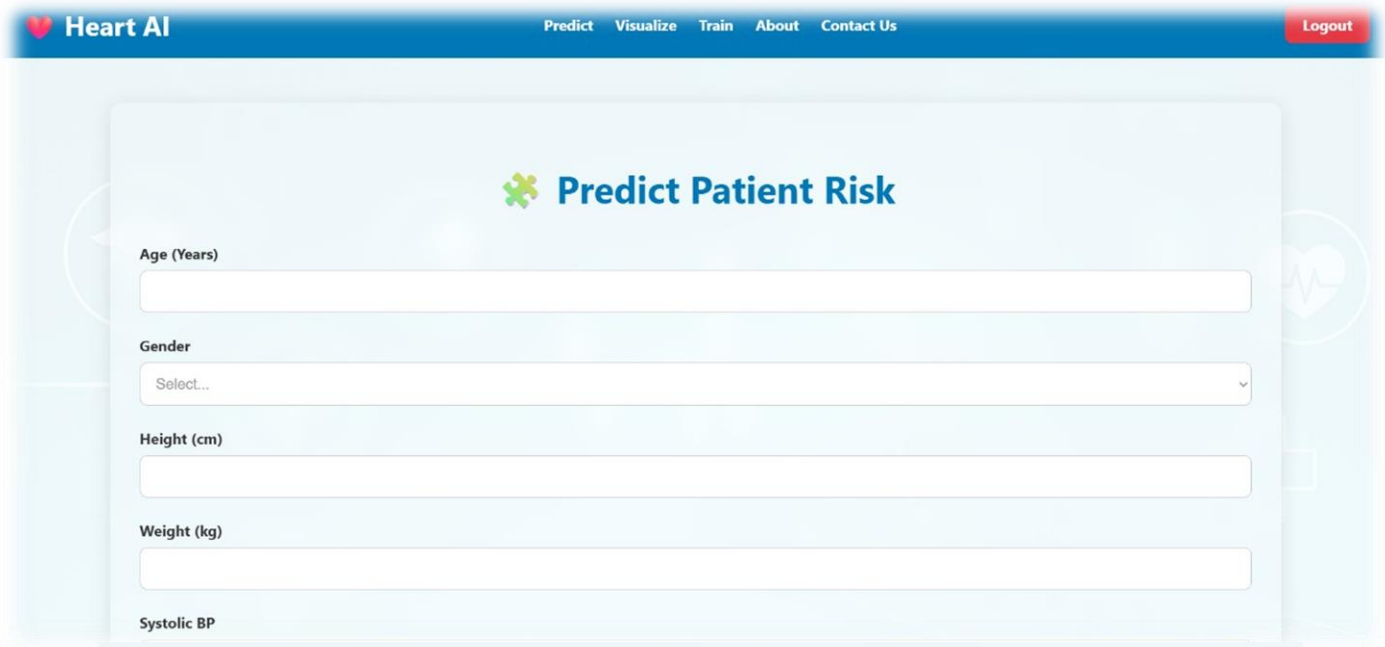
This project successfully developed and deployed a machine learning-powered web application for heart disease risk prediction. The model, combined with interactive visualizations and retraining capabilities, offers practical value for healthcare professionals aiming to improve patient outcomes through early risk detection.

Future work:

- Clinical Validation: Partner with hospitals to validate model predictions with real-world outcomes.
- Advanced Modeling: Explore deep learning models for longitudinal data analysis.
- Explainability: Integrate SHAP values to provide transparency for model decisions.
- Mobile Integration: Extend platform to mobile for field nurses and remote clinics.
- Multilingual UI: Support Arabic and other languages for broader usability.

Final project





Prediction Result:

The patient is at risk!

Entered Patient Data

Feature	Value
Age (Years)	58.0
Gender	Male
Height (cm)	185.0
Weight (kg)	100.0
Systolic BP	140.0
Diastolic BP	90.0
Cholesterol Level	Well Above Normal
Glucose Level	Well Above Normal
Smoking	Yes

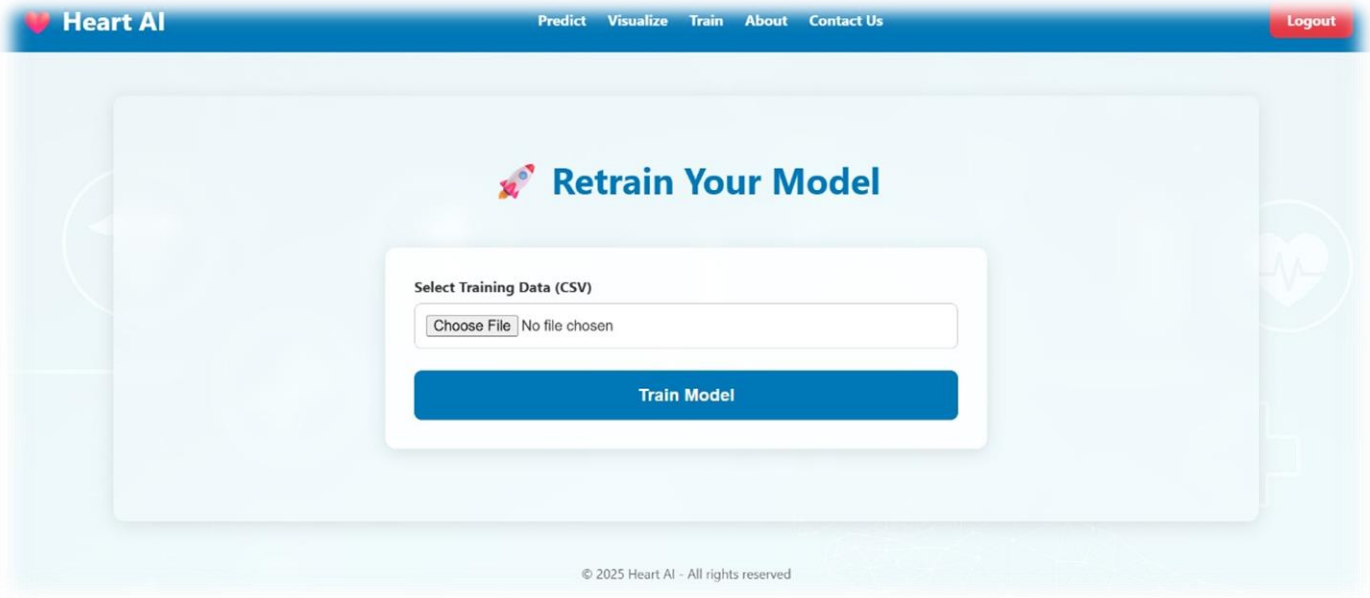
Prediction Result:

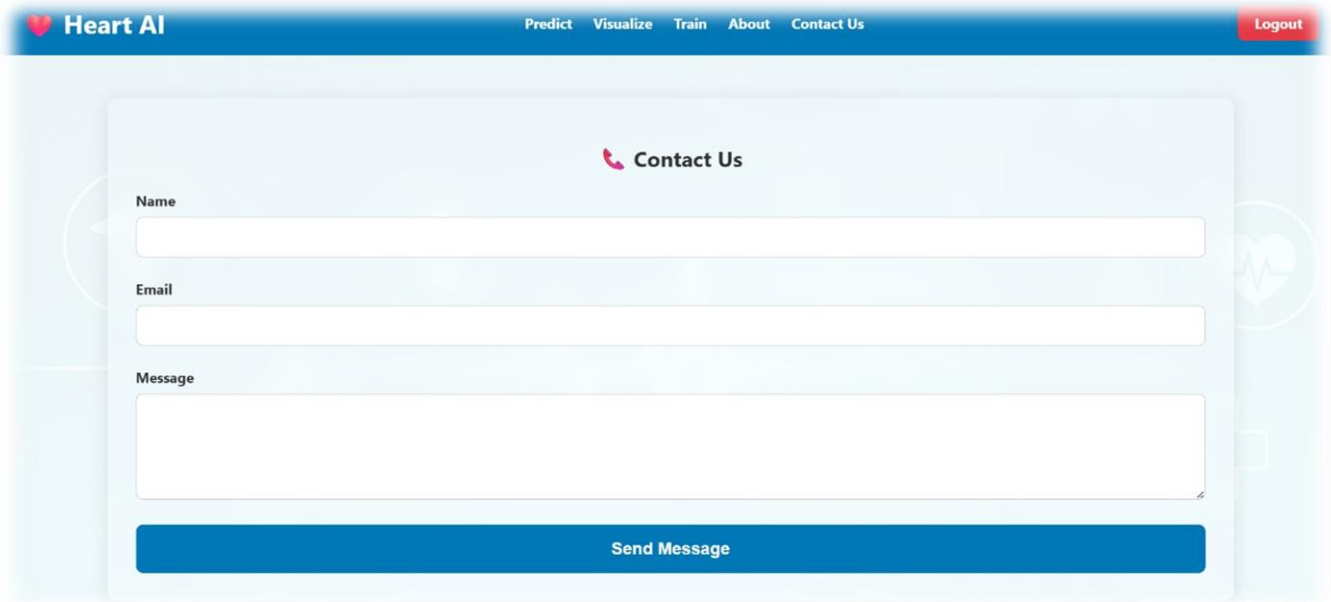
The patient is NOT at risk.

Entered Patient Data

Feature	Value
Age (Years)	50.3
Gender	Female
Height (cm)	168.0
Weight (kg)	62.0
Systolic BP	110.0
Diastolic BP	80.0
Cholesterol Level	Normal
Glucose Level	Normal
Smoking	No


Heart AI
Predict
Visualize
Train
About
Contact Us
Logout
Visualize Your Dataset
Choose File
No file chosen
Upload & Visualize
© 2025 Heart AI - All rights reserved


Heart AI
Predict
Visualize
Train
About
Contact Us
Logout
Retrain Your Model
Select Training Data (CSV)
Choose File
No file chosen
Train Model
© 2025 Heart AI - All rights reserved


Heart AI
Predict
Visualize
Train
About
Contact Us
Logout
Contact Us
Name
Email
Message
Send Message