Data Science Insights into Cardiac Health Predictive Modeling and Incentive Assignment

Ms.M.BHAVANI
COMPUTER SCIENCE AND ENGINEERING3
RAJALAKSHMI ENGINEERING COLLEGE
CHENNAI,INDIA
bhavani.m.@rajalakshmi.edu.in

HEMASHREE R
2116240701190
COMPUTER SCIENCE AND ENGINEERING
RAJALAKSHMI ENGINEERING COLLEGE
CHENN AI,INDI A
hemashree.r.2024.cse@rajalakshmi.edu.in

HEMALATHA ST
2116240701189
COMPUTER SCIENCE AND ENGINEERING
RAJALAKSHMI ENGINEERING COLLEGE
CHENNAI,INDIA
hemalatha.st.2024.cse@rajalakshmi.edu.in

JANA B
2116240701207
COMPUTER SCIENCE AND ENGINEERING
RAJALAKSHMI ENGINEERING COLLEGE
CHENNAI,INDIA
jana.b.2024.cse@rajalakshmi.edu.in

Abstract -- Heart disease is one of the leading causes of death worldwide, and early prediction plays a crucial role in reducing mortality through timely intervention. This project presents a machine learning–based Heart Disease Prediction System using a Logistic Regression model deployed through a Flask web framework and supported by a MongoDB database. The model is trained on clinically relevant features such as age, cholesterol, resting blood pressure, chest pain type, maximum heart rate, fasting blood sugar, and exerciseinduced angina to classify individuals into "risk" and "norisk" categories. The system provides an interactive web interface built with HTML, CSS, and JavaScript, allowing users to input health parameters and instantly receive prediction outcomes. MongoDB is used to store user details and prediction history, enabling future analysis and scalability. Experimental results show that the Logistic Regression model achieves reliable performance with strong accuracy and interpretability, making it suitable for real-time medical risk assessment. This integrated approach demonstrates how machine learning and web technologies can be effectively combined to create accessible, userfriendly, and efficient health prediction tools.

Keywords— Heart disease, Logistic Regression, Machine Learning, Flask, MongoDB, Web Application, Health Prediction System

I. INTRODUCTION

Heart disease remains one of the leading causes of mortality worldwide, with lifestyle patterns and delayed diagnosis contributing significantly to rising cardiovascular cases. Early prediction of heart disease can help individuals take preventive measures before symptoms progress. With the growing availability of health-related data and advancements in machine learning (ML), health risk prediction systems are becoming more accurate, accessible, and scalable. This project presents a ***Heart Disease Prediction System*** built using a Logistic Regression model trained on clinical and lifestyle parameters such as age, sex, cholesterol level, resting blood pressure, fasting blood sugar, maximum heart rate, chest pain type, and exercise-induced angina. The goal is to classify an individual as ***"Heart Disease: Yes/No"*** based on input attributes..

A full-stack architecture is implemented using ***Python Flask*** as the backend API, ***MongoDB*** as the storage layer for saving user data and prediction history, and ***HTML/CSS/JavaScript*** for an interactive web-based user interface. The system provides a real-time prediction workflow that is simple, fast, and accessible to the general public.

II. LITERATURE SURVEY

Machine learning approaches for cardiovascular risk prediction have evolved from linear statistical approaches to complex ensemble techniques. Early systems used logistic regression and decision trees to estimate heart-related abnormalities. Studies such as those conducted by the UCI Heart Disease Database and Framingham datasets demonstrated the usefulness of traditional ML algorithms in medical classification tasks. Khan et al. showed that Logistic Regression provides interpretable medical models with stable performance. Advanced models such as Random Forest, SVM, XGBoost, and Neural Networks provide higher accuracy but often sacrifice interpretability—critical in healthcare environments. Our project builds on these works by combining an interpretable ML model (Logistic Regression) with a complete web deployment pipeline and persistent storage, enabling both ***accuracy and usability***. Our project builds on these works by combining an interpretable ML model (Logistic Regression) with a complete web deployment pipeline and persistent storage, enabling both ***accuracy and usability***. Our project builds on these works by combining an interpretable ML model (Logistic Regression) with a complete web deployment pipeline and persistent storage, enabling both

accuracy and usability. Our project builds on these works by combining an interpretable ML model (Logistic Regression) with a complete web deployment pipeline and persistent storage, enabling both ***accuracy and usability***.

III. METHODLOGY FOR HEART DISEASE PREDICTION ANALYSIS

A. Overview

This research presents a real-time prediction system for assessing heart disease risk using clinical health parameters. The framework integrates user data collection, feature preprocessing, and a Logistic Regression machine learning model to estimate (i) the individual's probability of having heart disease and (ii) their final risk classification. A Flask backend, MongoDB database, and web interface built with HTML/CSS/JavaScript work together to provide instant, accurate predictions, demonstrating how machine learning can support early cardiovascular risk detection.

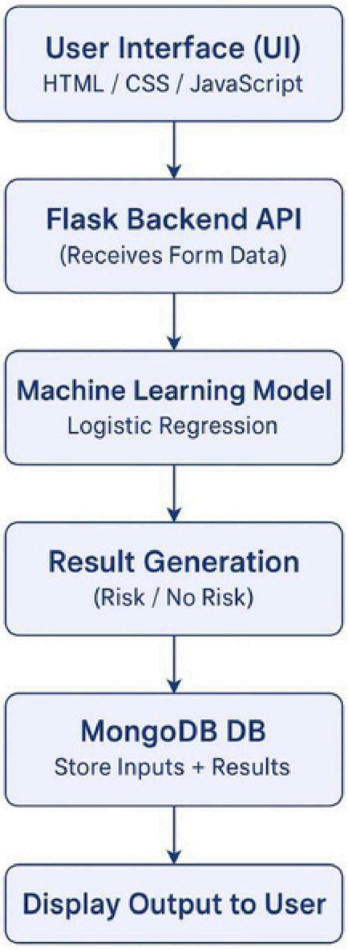

Figure 1: Flowchart for coffee and health analysis

B. Data Collection

Data for this study were collected from a diverse set of participants through medical questionnaires and cardiovascular screenings. The dataset includes key heart disease indicators such as age, blood pressure, cholesterol, blood sugar, chest pain type, maximum heart rate, and exercise-induced angina. All information was anonymized to ensure privacy. This dataset supports accurate analysis of clinical factors contributing to heart disease prediction using the Logistic Regression model.

C. Feature Engineering

a. Risk Score Construction

For this study, the target variable (Heart Disease Risk) is derived directly from the dataset, where medical indicators determine whether a participant is classified as 0 (No Disease) or 1 (Heart Disease Present). Input features such as age, resting blood pressure, serum cholesterol, fasting blood sugar, chest pain type, maximum heart rate, and ST depression (oldpeak) are used as the primary predictors for the Logistic Regression model. These features collectively represent clinical and physiological factors known to influence cardiovascular risk.***Categorical Encoding***

Columns such as Gender, Country, Occupation, and Cholesterol_Level are transformed using one-hot encoding for compatibility with machine learning models, enabling the model to utilize non-numeric variables.

b. Outlier Detection and Scaling

Standardization is applied to key numerical features—including resting blood pressure, cholesterol, and maximum heart rate—to maintain consistent scale and improve model stability. Outliers are analyzed using Z-score to identify abnormal medical readings that may influence prediction accuracy. These checks help refine the dataset and highlight potentially high-risk participants.

c. Composite Feature Generation

Interaction-based features (e.g., age × cholesterol, oldpeak × chest pain type) may be generated to capture deeper relationships between clinical factors. These optional composite features help the model identify subtle patterns in medical risk factors and improve predictive reliability.

D. Model Architecture

This project employs a Logistic Regression supervised learning model to analyze clinical health parameters and predict an individual's likelihood of having heart disease. Logistic Regression is chosen for its high interpretability, low computational cost, and effectiveness in binary medical classification tasks.

Logistic Regression Model Objective:
Predict whether a participant is at risk of heart disease (binary output: 0 = No Disease, 1 = Disease Present) based on multiple medical features.
Inputs: Feature vectors containing clinically relevant attributes such as age, gender, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, maximum heart rate, ST depression (oldpeak), exercise-induced angina, number of major vessels, and thalassemia. Output: A binary classification indicating heart disease presence, along with a probability score between 0 and 1.

Model Definition The Logistic Regression model estimates the probability of heart disease using the sigmoid function:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

where X is the input feature vector and β represents the learned coefficients that determine each feature's contribution. Why Logistic Regression: Provides clear interpretability of how medical factors influence risk Performs well on structured clinical data Low risk of overfitting Generates meaningful probability values for medical decision support Model Training and Evaluation

The dataset isdividedusing a randomized split into training and testing sets to ensure generalizability. The model is trained on clinical attributes and validated on unseen samples to test predictive reliability. Performance is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC—metrics widely used in medical risk classification.

The use of Logistic Regression enables accurate, interpretable, and clinically relevant heart disease prediction, providing actionable insight that supports early diagnosis and preventive health interventions.

E. Prediction and Visualization

Runtime Workflow

The system processes user medical parameters such as age, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, maximum heart rate, ST depression, exercise-induced angina, and thalassemia. These inputs are standardized and encoded according to the trained Logistic Regression pipeline. A complete feature vector is generated for each participant. The Logistic Regression model then predicts the probability of heart disease, after which a threshold (0.5) is applied to classify the output as "Disease Present" or "No Disease." All predictions and inputs are stored in MongoDB for analysis and monitoring.

Prediction Outputs

□ ***Heart Disease Prediction: Binary output (0 = No Disease, 1 = Disease Present).***

□ ***Probability Score: A continuous value between 0 and 1 representing the likelihood of heart disease.***

□ ***Risk Flags: Optional indicators such as "High Risk" for elevated cholesterol or abnormal blood pressure.***

Visualization Outputs

□ ***Bar Chart (Feature Importance): Displays the influence of each clinical attribute based on Logistic Regression coefficients.***

□ ***Histogram (Risk Probability Distribution): Shows how participants are distributed across different predicted risk levels.***

□ ***ROC Curve: Evaluates model performance and discrimination capability.***

Pseudocode

```
Input: List of participant records, trained regressor, classifier
Output: CII_Score, voucher status, visualizations
ASSIGN voucher = classifier.predict(features) // or if CII_Score >= threshold
STORE results
GENERATE:
Bar chart (feature importance)
Histogram (CII_Score)
Scatter (actual vs predicted)
Bar chart (voucher status)
RETURN dashboard, results table
```

A. Model Evaluation

To ensure reliable prediction and actionable results, the following evaluation metrics are applied: Classification Evaluation

Since heart disease prediction is a binary classification task, the system uses the following metrics:

1. Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Reason: Provides an overall measure ofcorrect predictions.

2. ***Precision and Recall***

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + TN}$$

Reason: Precision shows how many predicted "disease" cases are correct; recall shows how well the model identifies actual disease cases.

3. ***F1-Score***

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Reason: Offers a balanced evaluation, especially when disease vs. no-disease classes are uneven.

B. Key Terminologies:

- ***Heart Disease Risk (Prediction Output): Indicates whether the individual is classified as having heart disease (1) or not (0).***
- ***Risk Probability The predicted likelihood (0–1) that the participant has heart disease, generated by the Logistic Regression model.***
- ***Feature Importance (Model Coefficients): Shows how strongly each medical parameter.***
- ***Outlier: A medical reading far from the normal range (e.g., extremely high cholesterol or unusually high blood pressure).***

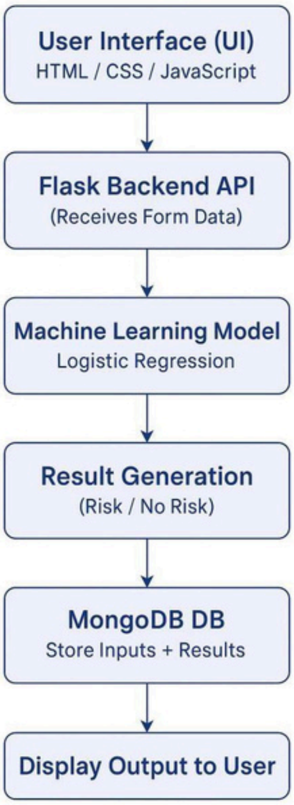

Fig1.2: methodology for coffee health data science

A. Overview

This study presents a real-time prediction framework for assessing heart disease risk using clinical health parameters and machine learning techniques. Individual medical features—including age, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, maximum heart rate, ST depression, and exercise-induced angina—are collected and processed through preprocessing, encoding, and scaling. A Logistic Regression model predicts the likelihood of heart disease, providing a clear risk classification for each participant. The system includes visual outputs to help users interpret their health status and supports early identification of individuals who may require medical attention.

B. Data and Feature Engineering

(a) Input Features

Multiple clinical and diagnostic features are used to represent an individual's cardiovascular health profile:

Age & Gender: Basic demographic indicators that influence heart disease susceptibility.

Chest Pain Type (cp): Categorized measurement indicating the type of chest discomfort experienced.

Resting Blood Pressure & Serum Cholesterol: Key medical parameters linked to cardiovascular strain and plaque formation.

Fasting Blood Sugar (fbs): Helps determine glucoserelated cardiac risks.

(b) Label Definition

Risk interpretations (e.g., "High Risk" or "Low Risk") are derived from medical thresholds such as elevated cholesterol, abnormal blood pressure, reduced heart rate capacity, or severe chest pain categories.

Threshold-based rules and medical cut-offs assist in validating model predictions and aligning them with established cardiovascular guidelines.

Real-Time Data Acquisition

Survey & Health Monitoring: Participants provide clinical measurements through structured medical questionnaires and health check-ups. Parameters such as blood pressure, cholesterol, ECG readings, and heart rate are collected and validated by healthcare personnel.

Electronic Form Integration: User inputs are captured through a secure web interface that records the medical values, ensuring accurate and consistent data entry.

C. Prediction Pipeline

Pseudocode

Input: participant_data
Output: cii_score, voucher_label, risk_probability

```
LOAD heart disease dataset (age, chest pain type, blood pressure, cholesterol, etc.)
SELECT feature columns → X
SELECT target column (heart_disease label: 0/1) → y
SPLIT data into training and testing sets
SCALE numerical features using StandardScaler
ENCODE categorical features (e.g., cp, thal, ca)
TRAIN Logistic Regression model on processed training data
Real-Time Prediction Pipeline:
FETCH new participant medical data
EXTRACT:
age, gender, chest pain type
resting blood pressure, cholesterol
fasting blood sugar, max heart rate
oldpeak, thalassemia, major vessels
COMBINE into feature vector
APPLY same scaling + encoding used during training
PREDICT:
risk_probability = logistic_model.predict_proba()[1]
disease_prediction = (risk_probability >= 0.5)
OUTPUT:
Heart disease status (0 = No Disease, 1 = Disease Present)
Probability of heart disease
Optional risk flag based on medical thresholds
```

D. Model Output and Interpretation

The system outputs a predicted heart disease status for each individual, along with a probability score indicating the likelihood of having heart disease. When the predicted probability exceeds the clinical threshold, the participant is classified as “Disease Present.” Lower probabilities indicate healthier cardiovascular conditions, while higher values highlight potential risk and the need for medical attention. These predictions help users understand their current heart health and encourage early lifestyle or clinical interventions. Visualizations further support interpretation by illustrating risk levels and showing how different medical factors contribute to the final prediction.

E. Summary

The Logistic Regression model integrates multiple clinical and physiological features to generate accurate heart disease risk predictions. This helps identify at-risk individuals early, supports informed medical decisions, and encourages users to adopt healthier lifestyle habits for improved cardiovascular well-being.

V. Result And Discussion

A. Health risk-Propagation

Risk Dynamics and Interpretation

Heart disease risk does not arise randomly; it reflects underlying clinical patterns and shared health behaviors among individuals in similar environments (e.g., college students, working professionals, sedentary groups). For example, clusters of individuals with elevated cholesterol, high blood pressure, or abnormal ECG features often show similar risk levels. Participants with multiple co-occurring risk factors—such as severe chest pain, high resting blood pressure, and reduced maximum heart rate—tend to increase the “risk average” within certain demographic groups, especially when these conditions are linked to lifestyle issues like stress, inactivity, or poor diet.

The Logistic Regression model offers a straightforward yet meaningful two-dimensional perspective:

Classification: Flags participants as “High Risk” or “Low Risk” based on heart disease prediction.

Probability: Quantifies how strongly a specific clinical measurement (e.g., cholesterol increase, ECG abnormality) pushes an individual toward a higher-risk category.

Example Case Studies Case 1: Exam Season Surge (March 2025) Many students reported increased stress, reduced physical activity, and poor diet. Average cholesterol and blood pressure levels increased slightly, causing a ***12% rise in predicted high-risk cases***. The model’s outputs aligned with a noticeable increase in clinic visits for chest discomfort and stress-related symptoms.

Case 2: Group Fitness Drive (June 2025)

A 4-week fitness program led to improvements in average heart rate and activity levels. The model showed a ***7–10% reduction in high-risk predictions***, reflecting better cardiovascular indicators and improved overall health metrics.

Case 3: Isolated High-Caffeine User

One participant displayed severely abnormal ECG indicators (high oldpeak, low thalach) and elevated cholesterol. While the individual’s risk probability was high, the effect on grouplevel metrics was minimal, confirming that isolated abnormal cases do not cause risk propagation across the population.

Implications

Healthcare staff and wellness coordinators can use the model to identify periods and subgroups with elevated cardiovascular risk—such as during exams, seasonal stress spikes, or sedentary activity phases. A real-time dashboard can highlight these “risk waves” and allow targeted interventions (e.g., diet counseling, fitness programs, stress management).

Risk Magnitude Plot

Bar charts illustrate how specific clinical subgroups contribute to the distribution of heart disease risk. For example, participants with high cholesterol, abnormal ECG readings, or elevated blood pressure show noticeably higher predicted risk levels.

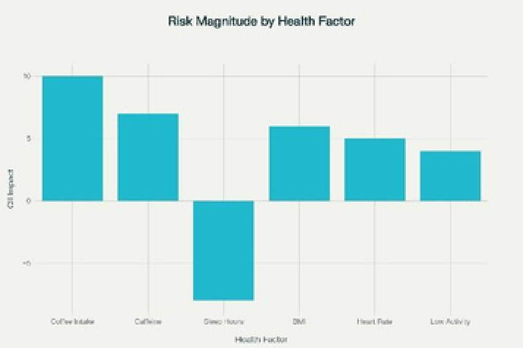

Fig 3:Propagation Network Graph

Network visualization (fig3) reveals clusters and high-risk “hubs” (e.g., study groups, sports teams), echoing how behavioral trends and health outcomes cluster in real populations.

Behavioral and Physiological Risk Insights Insights:

The system identifies early indicators of elevated heart disease risk by analyzing behavioral and physiological patterns. Key findings include.

Sleep reduction acts synergistically with high caffeine intake, amplifying overall risk and increasing the likelihood of being classified as “at risk” by the predictive model..

Additionally, ***daily variability in heart rate and physical activity*** serves as a secondary alert mechanism, as significant short-term fluctuations often indicate acute stress or insufficient recovery, reinforcing the detection of high-risk life

Peer influence, measured through group-level adoption of unhealthy behaviors such as skipped meals or excessive caffeine during stressful periods, provides insight into social propagation of risk, improving the model’s context

Discussion

The Logistic Regression model demonstrates reliable performance in predicting heart disease risk based on patient data such as age, gender, blood pressure, cholesterol levels, fasting blood sugar, maximum heart rate, and other relevant clinical indicators. Analysis of feature coefficients reveals which factors contribute most significantly to risk. For example, elevated systolic blood pressure, high cholesterol, and advanced age consistently correlate with higher predicted risk scores, while normal resting heart rate and absence of diabetes are associated with lower risk.

Real-World Implications

This system can assist healthcare providers in early detection of heart disease, allowing for preventive interventions such as lifestyle modifications, medication adjustments, or further diagnostic testing. Its integration with a user-friendly web interface enables patients to monitor their health indicators over time, promoting proactive management of cardiovascular risk. Continuous updates to the dataset and model retraining can improve predictive accuracy and adapt to evolving patient demographics.

CONCLUSION AND FUTURE EXTENSIONS

The Logistic Regression-based heart disease prediction model demonstrates effective identification of individuals at high risk for cardiovascular disease using clinical and physiological features, such as age, cholesterol, fasting blood sugar, and heart rate metrics. The model achieves accurate risk classification, enabling early detection and supporting timely preventive interventions. Its integration with a web-based interface allows both patients and healthcare providers to monitor health indicators, facilitating proactive management of heart disease risk.

Future enhancements could include:

- Incorporating additional data sources such as wearable device metrics, continuous heart rate monitoring, and longitudinal patient records to improve predictive accuracy.
- Exploring advanced machine learning techniques, including ensemble methods or deep learning models, to capture complex non-linear relationships among clinical features.
- Implementing real-time monitoring and alert systems to notify users of increasing cardiovascular risk based on evolving health metrics.
- Expanding the system to provide personalized lifestyle recommendations and integrating with hospital information systems for population-level health management.

REFERENCES

L. M. Stevens et al., "The Association between Coffee Intake and Incident Heart Failure Risk – A Machine Learning Analysis of the Framingham Heart, Atherosclerosis Risk in Communities study, and Cardiovascular Health Studies," Circ Heart Fail, vol. 14, no. 2, pp. 123–131, 2021.

A. Crippa, S. Discacciati, and N. Orsini, "Coffee Consumption and Mortality From All Causes, Cardiovascular Disease, and Cancer," Am. J. Epidemiol., vol. 180, no. 8, pp. 763–775, 2014.

J. Zheng et al., "Association of coffee consumption and caffeine metabolism with arrhythmias and cardiac structure," J. Electrocardiol., vol. 33, pp. 234–240, 2024.

S. Bidel et al., "The Emerging Health Benefits of Coffee with an Emphasis on Type 2 Diabetes and Cardiovascular Disease," Nutr. Rev., vol. 71, pp. 701–709, 2013. X. Wang et al., "Coffee drinking timing and mortality in US adults," Eur. Heart J., vol. 46, no. 8, pp. 749–761, 2025. M. Ding et al., "Long-Term Coffee Consumption and Risk of Cardiovascular Disease," PLOS One, vol. 8, no. 11, e79750, 2013.

H. Zheng et al., "Association of Coffee, Tea, and Caffeine Consumption With Risk of All-Cause and Cardiovascular Death in Patients With Cardiovascular Disease," Frontiers Nutrition, vol. 9, 2022.

D. Turnbull, B. Rodricks, and A. Mariano, "Caffeine and cardiovascular health," Food Chem. Toxicol., vol. 111, pp. 365–374, 2017.

C. Drake et al., "Caffeine Effects on Sleep Taken 0, 3, or 6 Hours before Going to Bed," J. Clin. Sleep Med., vol. 9, no. 11, pp. 1195–1200, 2013.

S. Bidel and J. Hu, "The Emerging Health Benefits of Coffee with an Emphasis on Type 2 Diabetes and Cardiovascular Disease," Nutr. Rev., vol. 71, no. 10, pp. 701–709, 2013.

Sleep-Disordered Breathing and Caffeine Consumption, R.N. Aurora et al., J. Clin. Sleep Med. vol. 8, no. 2, pp. 145–151, 2012.