CARDIOPULSE AI: DATA-DRIVEN HEART DISEASE RISK ANALYSIS SYSTEM

Introduction to Data Science — Project Report

Submitted By:

Muhammad Wafa Abbas	2023-CS-454
Abdul Ahad	2023-CS-456
Abdul Rehman	2023-CS-463

Submitted To:

Dr. Tayybah Kiran

Department of Computer Science

Rachna College of Engineering & Technology, Gujranwala

(A Constituent College of UET, Lahore)

December 24, 2025

1. Introduction

Cardiovascular diseases remain a leading cause of global mortality, with early identification of high-risk individuals playing a crucial role in prevention and treatment. This project aims to analyze a heart disease dataset through statistical techniques, exploratory data analysis (EDA), predictive modeling, and performance evaluation to understand clinical factors contributing to heart disease and develop a predictive model.

The dataset includes **918 patient records** containing demographic, clinical, and diagnostic attributes such as age, sex, cholesterol levels, resting blood pressure, chest pain type, fasting blood sugar, ECG results, maximum heart rate, exercise-induced angina, ST depression, and ST segment slope. The target variable is **HeartDisease (0 = No Disease, 1 = Disease)**.

The main goals of this project are:

- Perform comprehensive data understanding and preprocessing.
- Conduct EDA using statistical summaries and visualizations.
- Evaluate relationships between categorical clinical indicators and heart disease using Chi-square tests.
- Build and compare machine learning models (Logistic Regression and Random Forest).
- Evaluate model performance using accuracy, confusion matrices, ROC–AUC curves, and feature importance.

2. Existing Tools and Techniques (with References)

2.1 Exploratory Data Analysis (EDA)

EDA techniques such as summary statistics, histograms, boxplots, correlation matrices, and scatter plots were used to identify patterns, outliers, and variable relationships.
Reference: Tukey, J. (1977). Exploratory Data Analysis.

2.2 Statistical Testing

The Chi-square test of independence was used to examine associations between categorical predictors and heart disease.
Reference: McHugh, M. L. (2013). The Chi-square test of independence.

2.3 Logistic Regression

A baseline machine learning classifier for predicting binary clinical outcomes.
Reference: Hosmer & Lemeshow (2000). Applied Logistic Regression.

2.4 Random Forest Classifier

A robust ensemble learning technique that constructs multiple decision trees and aggregates predictions. Effective for tabular medical datasets.
Reference: Breiman, L. (2001). Random Forests. Machine Learning.

2.5 Feature Scaling and One-Hot Encoding

Scaling (StandardScaler) and one-hot encoding were applied to ensure numerical stability and compatibility with predictive models.

Reference: Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python.

2.6 Correlation Heatmap

A correlation heatmap was used to visualize the strength and direction of relationships among numerical features, helping identify multicollinearity and important associations with heart disease.

Reference: Akoglu, H. (2018). User's guide to correlation coefficients.

3. Summary / Comparison Table

Technique / Tool	Purpose	Advantages	Limitations
EDA	Understand distributions & relationships	Visual insights	No formal statistical significance
Chi-Square Test	Check categorical dependencies	Simple, interpretable	Only works for categorical variables
Logistic Regression	Baseline classification	Interpretable, fast	Assumes linear separability
Random Forest	Ensemble prediction	High accuracy, handles non-linearity	Less interpretable
ROC–AUC	Evaluate classifier thresholds	Threshold-independent	Does not show calibration
Correlation Heatmap	Visualize correlations among numerical features	Identifies relationships and multicollinearity	Limited to linear correlations

Model Comparison Table: Comparison Between Logistic Regression and Random Forest

Aspect	Logistic Regression	Random Forest
Model Type	Linear classification model	Ensemble of decision trees
Interpretability	High (coefficients are easy to interpret)	Lower (black-box nature)
Ability to Handle Non-linearity	Limited	Strong
Feature Interaction Handling	Limited	Excellent
Sensitivity to Outliers	Moderate	Low
Computational Cost	Low	Higher
Accuracy (Project Results)	~89%	~92%
ROC–AUC Performance	Slightly higher AUC	Slightly lower AUC
Suitability for This Dataset	Good baseline model	Best overall performer

Based on accuracy and robustness, Random Forest was selected as the final model, while Logistic Regression served as an effective baseline for comparison.

4. Contribution / Features of System

This project presents the following key contributions:

- Complete preprocessing including missing value correction and outlier removal via IQR.
- Comprehensive EDA including histograms, boxplots, countplots, heatmaps, scatter plots, and pairplots.
- Statistical validation of categorical feature importance using Chi-square tests.
- Boilerplate-free one-hot encoding and feature scaling pipeline.
- Implementation and comparison of Logistic Regression and Random Forest models.
- Evaluation using accuracy, classification report, confusion matrix, cross-validation, and ROC–AUC.
- Feature importance analysis providing clinical interpretability.
- An interactive Power BI dashboard for visual analytics and result interpretation.
- A user-friendly web-based frontend (CardioPulse AI) enabling input-driven risk analysis and system interaction.
- End-to-end system integration combining machine learning, visual analytics, and frontend presentation.

5. Data Flow Diagram

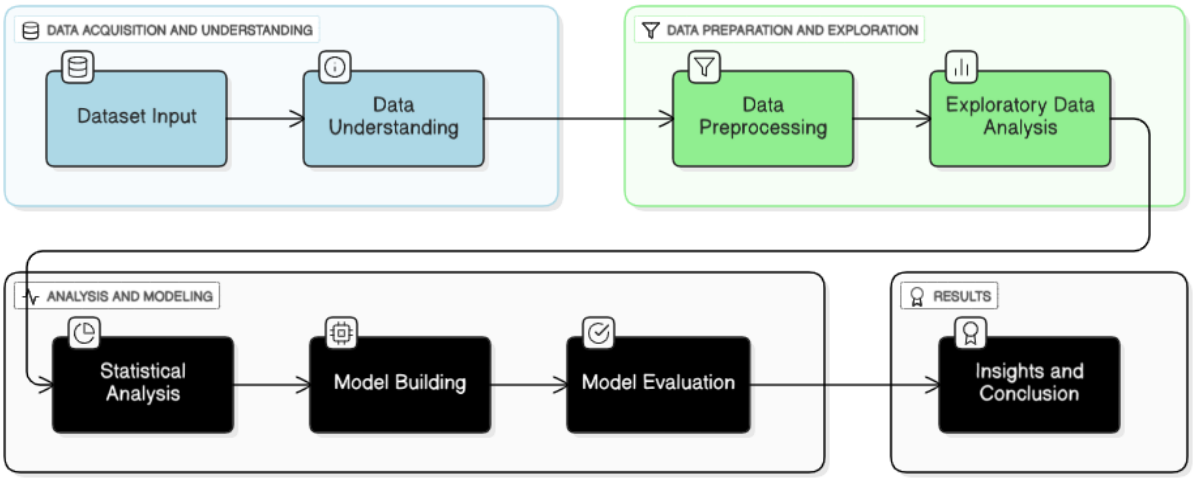

Figure 5.1: Data Flow Diagram of the Proposed Heart Disease Data Analysis System

5.1. Data Collection

- Import the dataset (918 samples, 12 attributes).

5.2. Data Understanding

- Check data structure, types, summary statistics.
- Identify missing values, invalid entries, and potential outliers.

5.3. Data Preprocessing

- Replace invalid 0 values in Cholesterol and RestingBP with mean values.
- Detect and remove outliers using the IQR method (82 removed → final 836 samples).
- Encode categorical variables using one-hot encoding.
- Scale numerical features using StandardScaler.

5.4. Exploratory Data Analysis (EDA)

- Visualize distributions (histograms, KDE plots).
- Create boxplots, scatter plots, pairplots.
- Analyze correlation heatmap.
- Compare heart-disease positive vs negative groups.

5.5. Statistical Analysis

- Perform **Chi-square tests** for categorical associations with heart disease.

5.6. Model Building

- Split dataset into training and testing sets.
- Train **Logistic Regression** and **Random Forest** classifiers.

- Apply **cross-validation** for performance validation.

5.7. Model Evaluation

- Evaluate using **accuracy**, **precision**, **recall**, **F1-score**.
- Generate **confusion matrix**.
- Plot and interpret **ROC–AUC curves**.
- Analyze **feature importance** (Random Forest).

5.8. Final Insights & Interpretation

- Summarize key predictors.
- Highlight clinically important patterns.
- Conclude which model is best suited.

6. Results / Execution

6.1 Data Overview

- Filled missing values in dataset.
- After imputation and IQR outlier removal → 836 rows retained.

Outlier removal summary:

Column	Removed
RestingBP	27
Cholesterol	40
MaxHR	2
Oldpeak	13

6.2 Exploratory Data Analysis Results

Distribution Plots

Histograms show:

- **Age** normally distributed around mean ~54.
- **Cholesterol** initially skewed due to zeros; corrected distribution smoother.
- **RestingBP** clusters around 120–140 mmHg.
- **MaxHR** shows normal-like pattern with mean ~137 bpm.

Boxplots

- Patients with heart disease tend to be **older** and have **lower MaxHR**.
- Outlier removal improved variable ranges and consistency.

Correlation Heatmap

Important findings:

- **MaxHR** negatively correlated with HeartDisease (**r = -0.40**)
- **Oldpeak** positively correlated (**r = +0.40**)
- **Age** moderately increases disease likelihood (**r = +0.28**)

Categorical Relationships

Visual countplots show:

- ASY chest pain type strongly associated with heart disease.
- Male patients show significantly higher heart disease count.
- ST_Slope category Flat highly related to disease.

6.3 Chi-Square Test Results

Feature	p-value	Result
Sex	1.02×10^{-19}	Significant
ChestPainType	3.20×10^{-54}	Highly Significant
FastingBS	6.32×10^{-16}	Significant
RestingECG	0.0037	Significant
ExerciseAngina	extremely small	Highly Significant
ST_Slope	1.10×10^{-69}	Extremely Significant

Conclusion: All categorical variables are strongly associated with heart disease.

6.4 Machine Learning Models

Model 1: Logistic Regression

- Accuracy: 0.89 (89%)
- Confusion Matrix:

Actual / Predicted	0	1
0	54	9
1	5	58

- ROC–AUC: 0.953
- Cross-Validation Mean: 0.84090

Model 2: Random Forest Classifier

- Accuracy: 0.92 (92%)
- Confusion Matrix:

Actual / Predicted	0	1
0	56	7

1	3	60

- ROC–AUC: 0.959
- Cross-Validation Mean: 0.855

ROC Curve Comparison

- Logistic Regression AUC: 0.953
- Random Forest AUC: 0.959

Random Forest provides higher accuracy, while Logistic Regression shows slightly better ranking performance (AUC).

6.5 Feature Importance (Random Forest)

Top predictors:

Feature	Importance
ST_Slope_Up	0.194979
Cholesterol	0.055863
MaxHR	0.067164
ST_Slope_Flat	0.123721
Age	0.053351
ChestPainType_ASY	0.100929
RestingBP	0.034464

Interpretation:

- **ST_Slope** is the strongest predictor.
- **Low MaxHR**, **high Cholesterol**, and **older Age** highly increase heart-disease likelihood.
- **ASY chest pain** strongly indicates presence of disease.

6.6 Interactive Power BI Dashboard

To enhance interpretability and interactive exploration of the results, an interactive dashboard was developed using Microsoft Power BI.
The dashboard consolidates key findings from exploratory data analysis, categorical statistical testing, and machine learning evaluation into four logical pages: Overview, Clinical EDA, Categorical Analysis, and Model Performance.
Interactive slicers allow filtering by sex, chest pain type, and disease status, enabling dynamic analysis. Visual elements such as KPI cards, distribution plots, correlation heatmaps, chi-square

summaries, ROC curves, and feature importance charts provide an intuitive understanding of clinical risk factors and model behavior.

7. References

Breiman, L. (2001). *Random forests*. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). John Wiley & Sons.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia Medica*, 23(2), 143–149. https://pubmed.ncbi.nlm.nih.gov/23894860/

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3), 91–93. https://pubmed.ncbi.nlm.nih.gov/30191186/

Scikit-learn developers. (2024). *Scikit-learn documentation*. https://scikit-learn.org

Pandas development team. (2024). *Pandas documentation*. https://pandas.pydata.org

Fedesoriano. (2021). *Heart failure prediction dataset*. Kaggle. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction