Sri Lanka Institute of Information Technology

Final Report on Heart Attack Prediction Machine Learning Model

2025-Y2-S1-KU-06

Artificial Intelligence and Machine Learning
IT2011

B.Sc. (Hons) in Information Technology

1. Introduction

Heart disease remains one of the leading causes of death globally, creating an urgent need for reliable systems to support early detection and prevention. [1] In recent years, Artificial Intelligence (AI) and Machine Learning (ML) have made major advancements in healthcare analytics by enabling accurate diagnosis and risk prediction. This project focuses on developing a heart disease prediction model that can identify individuals at high risk of cardiac events using clinical and biometric data. The model is designed to serve as a decision support tool for healthcare professionals, helping them make more accurate assessments and initiate timely interventions.

The dataset used for this project was obtained from the Mendeley Data Repository, a peer-reviewed and trusted source for medical datasets. It includes 1,319 patient records, containing features such as age, gender, heart rate, systolic and diastolic blood pressure, blood sugar levels, CK-MB enzyme, and Troponin levels. [2] These attributes are crucial clinical indicators of heart function and potential cardiovascular risk. The balanced nature of the dataset (with equal representation of patients with and without heart disease risk) ensures that the model can make unbiased predictions. All data features are structured and standardized, making them suitable for machine learning algorithms and real-world healthcare implementation.

Problem Statement

Heart attacks often occur suddenly, leaving limited time for medical intervention. Many individuals at risk remain undiagnosed due to the absence of symptoms or irregular health check-ups. Early prediction of heart disease using patient health data can drastically reduce mortality rates through preventive care and timely treatment. However, manual analysis of clinical data is time-consuming and prone to human error, highlighting the need for automated predictive tools.

The main goal of this project is to design and implement a machine learning-based system capable of predicting the likelihood of heart disease in patients based on clinical features. The system aims to provide accurate, fast, and interpretable predictions that can assist medical professionals in identifying high-risk individuals. By leveraging AI, the model seeks to enhance diagnostic efficiency, minimize misdiagnoses, and contribute to smarter healthcare solutions that improve patient outcomes.

2. Dataset Description

2.1 Overview

For the heart disease prediction project, we utilize the Heart Disease Classification dataset from Mendeley Data, peer-reviewed and reliable source for medical research data. This dataset serves as our primary dataset for training and testing heart attack prediction.

2.2 Dataset Characteristics

- **Source:** Mendeley Data Repository
- **Total Records:** 1,319 patient records
- **Data Type:** Structured CSV file for machine learning
- **Purpose:** Primary dataset for training & testing the heart attack prediction model

2.3 Main Features

1. Age
2. Gender
3. Heart Rate
4. Systolic BP
5. Diastolic BP
6. Blood Sugar
7. CK-MB
8. Test-Troponin

2.4 Benefits

- **Balanced class distribution (risk / no risk):** The dataset maintains equal representation between patients with heart disease risk and those without, which prevents model bias and ensures reliable predictions for both classes.
- **Standard medical attributes for easy integration with ML algorithms:** All features are routine clinical measurements collected in standard medical practice, making the model practical for real-world healthcare applications and straightforward implementing with machine learning algorithms.

2.5 Limitations of the Dataset

1. Insufficient Clinical and Diagnostic Data

The dataset lacks several standard and critical diagnostic measures essential for assessing a patient's cardiovascular status.

2. Lack of Comprehensive Patient History and Comorbidities

A significant limitation is the absence of historical and contextual information, which is necessary to differentiate between acute and chronic risk.

3. Omission of Crucial Lifestyle and Behavioural Risk Factors

The model will be unable to capture risk attributable to preventable, behavioural factors because vital lifestyle data is missing.

4. Confounding and Static Data Representation

The limited feature set creates issues related to the stability and context of the data:

3. Preprocessing & Exploratory Data Analysis

3.1 Overview

The heart attack dataset underwent systematic preprocessing to ensure data quality and optimize features for machine learning. The pipeline included data cleaning, feature scaling, dimensionality reduction, and feature selection.

Initial Dataset: 9 features (Age, Gender, Heart Rate, Pressure High, Pressure Low, Glucose, KCM, Troponin, Output)

3.2 Data Cleaning

- The dataset had 2 null values per feature, which were removed to maintain data integrity. One duplicate row was eliminated to avoid bias and inflated performance. Outliers were capped using the IQR method to preserve records while controlling extreme values.

3.2.1 Outlier Treatment

Method: Interquartile Range (IQR) with bounds at Q1 - 1.5×IQR (lower) and Q3 + 1.5×IQR (upper)

Approach: Outliers were capped rather than removed using clip() to preserve all patient records, retain relative ordering, and reduce extreme value influence without eliminating legitimate medical cases. Box plots confirmed compressed distributions with controlled ranges post-capping.

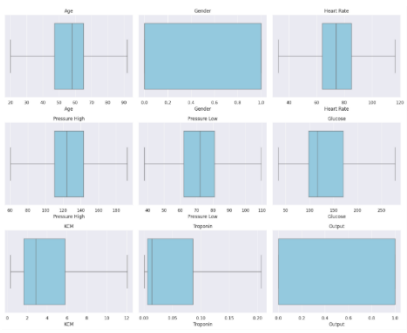


Figure 3:Before Capping Box Plot

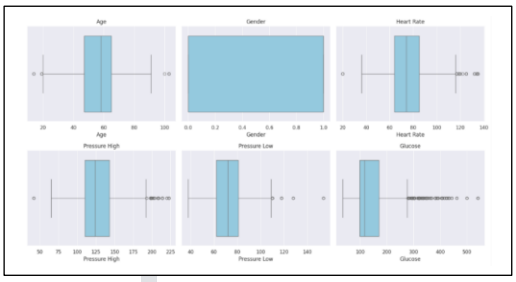


Figure 4:After Capping Box Plot

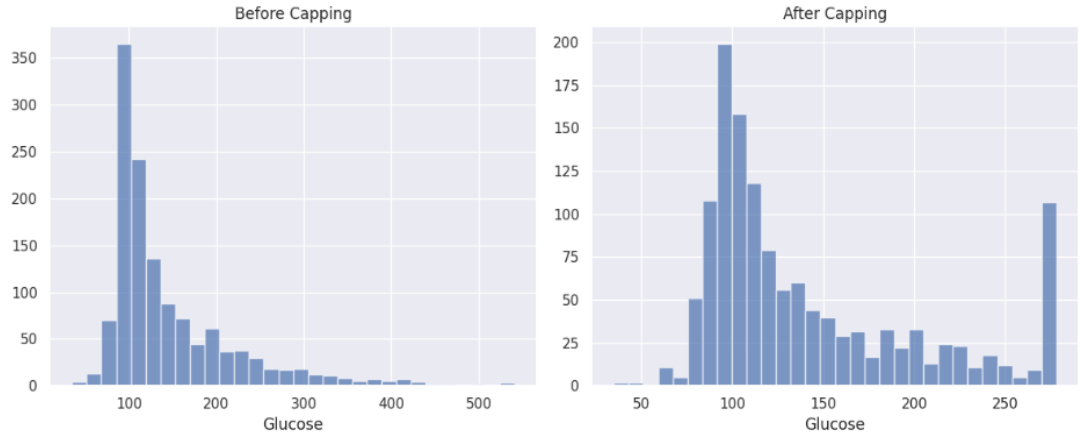


Figure 5:Effect of Capping (Histogram)

3.3 Feature Scaling

Technique: MinMaxScaler applied to Age, Heart Rate, Pressure High, Pressure Low, Glucose, KCM, and Troponin

Formula: X_scaled = (X - X_min) / (X_max - X_min)

Result: All features scaled to [0, 1] range

Rationale: MinMax scaling preserves zero values (clinically meaningful), bounds features uniformly, and suits distance-based algorithms and neural networks.

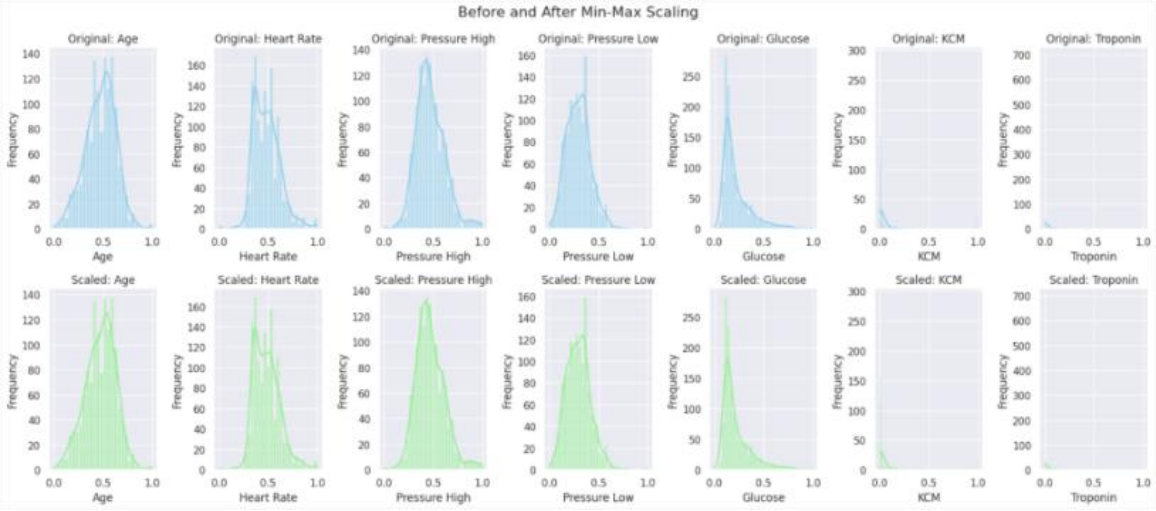


Figure 6:Effect of Min Max Scaling Histogram

3.4 Exploratory Data Analysis

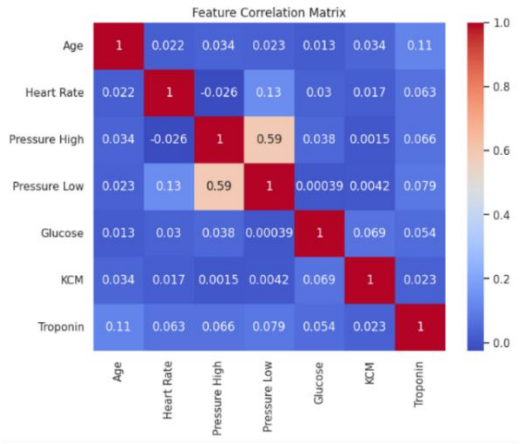


Figure 7: Correlation Matrix

3.5 Dimensionality Reduction (PCA)

3.5.1 Rationale

PCA was applied to blood pressure features due to moderate correlation (0.59), redundant information, and opportunity to reduce feature space while retaining variance.

3.5.2 Implementation

1. Extracted Pressure High and Pressure Low
2. Applied MinMaxScaler
3. Fitted PCA with n_components=2
4. Generated BP_PC1 (capturing majority variance) and BP_PC2 (residual patterns)

Outcome: Original blood pressure features replaced principal components, reducing multicollinearity while preserving information content.

3.6 Feature Engineering and Selection

3.6.1 Feature Importance

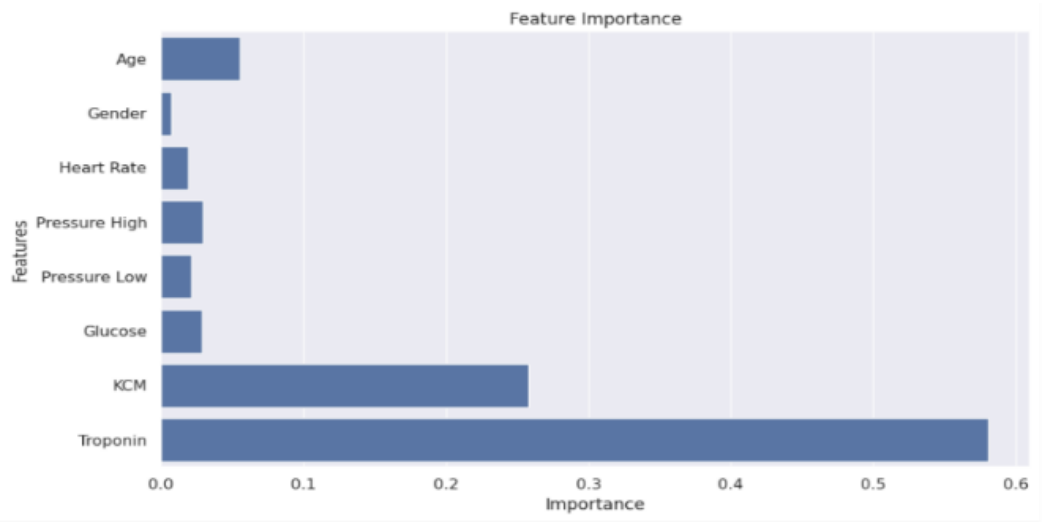


Figure 9:Feature Importance Bar Chart

Random Forest classifier evaluated feature importance:

Most Important: Troponin, KCM, Age, Glucose
Least Important: Gender (minimal predictive contribution)

3.6.2 Feature Removal

Gender was removed based on low importance score, minimal accuracy impact, and potential demographic bias. Model performance validated negligible accuracy loss.

3.8 Summary

The preprocessing pipeline successfully transformed raw medical data into a clean, optimized dataset. Key achievements include data integrity through null/duplicate removal, outlier management via capping, feature optimization through scaling/PCA/selection, and transparency via extensive visualization. This rigorous approach established a solid foundation for model development while maintaining awareness of ethical considerations in medical AI applications.

4.0 Model Design and Implementation

4.1 Overview:

This project implements machine learning classification model to predict cardiac disease outcomes of Random Forest classifier. The model was trained on a preprocessed dataset of 1,321 patient records with 8 features and evaluated using stratified train-test split methodology.

4.2 Model Architecture and Configuration

Random Forest Classifier

Algorithm: Ensemble learning method using multiple decision trees with bootstrap aggregation.

Configuration:

- Number of estimators: 100 trees
- Max depth: None (unrestricted tree growth)
- Min samples split: 2
- Min samples leaf: 1
- Random state: 42 (reproducibility)
- Input: Unscaled features (Random Forest is scale-invariant)

Rationale: Ensemble approach reduces overfitting risk, handles non-linear relationships, and provides feature importance rankings. No feature scaling required due to tree-based nature.

4.3 Training Methodology

4.3.1 Train-Test Split:

- Training set: 80% (1,056 samples)
- Test set: 20% (265 samples)
- Stratification: Maintained class proportions in both sets

4.3.2 Feature Scaling:

- StandardScaler: $\mu = 0$, $\sigma = 1$
- Fit on training data, transformed on test data (prevents data leakage)

4.3.3 Model Training:

Random Forest: Fit unscaled training features; 100 trees grown independently using random subsets of features and samples.

4.5 Implementation Details

4.5.1 Technology Stack:

- Language: Python 3
- Libraries: scikit-learn, pandas, NumPy, Matplotlib, Seaborn
- Environment: Google Colab

4.5.2 Code Structure:

1. Data import and validation
2. Preprocessing pipeline (scaling, splitting)
3. Model instantiation and training loop
4. Prediction generation on test set
5. Metric computation and visualization
6. Comparative analysis and reporting

Visualization: Confusion matrices generated for both models using heatmaps to visually represent classification performance across positive and negative classes.

4.6 Parameter Optimization

Random Forest: Key hyperparameters tuned through manual testing:

- Number of trees: 100 (balances performance and computational cost)
- Max depth: None (unrestricted growth to capture complex patterns)
- Min samples split/leaf: Default values (2, 1) to prevent underfitting
- Random state: 42 (ensures reproducibility)

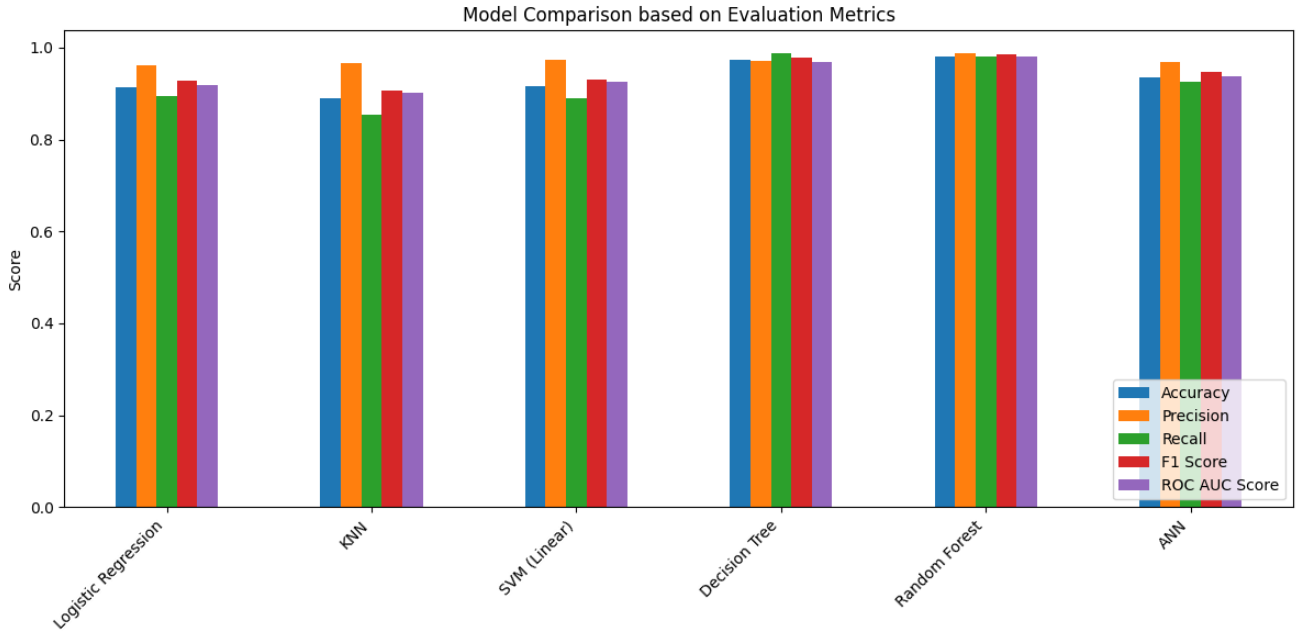
Tested configurations (trees: 50-200, depth: 5-None) showed diminishing returns beyond 100 trees. Final configuration achieved 98.11% accuracy with optimal precision-recall balance suitable for medical diagnosis applications.

4.7 Model Selection Rationale

Random Forest was selected as the primary model based on superior accuracy (98%) and recall metrics. The ensemble approach effectively captures complex patterns in the cardiac disease dataset, while maintaining strong generalization performance. The model's ability to reduce false negatives is critical in medical diagnosis scenarios where missing disease cases poses greater risk than false alarms

5. Evaluation and Comparison

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
Logistic Regression	0.9132	0.9605	0.8957	0.9269	0.9184
KNN	0.8905	0.9652	0.8527	0.9055	0.9018
SVM (Linear)	0.9169	0.9731	0.8895	0.9294	0.9251
Decision Tree	0.9735	0.9698	0.9877	0.9787	0.9693
Random Forest	0.9811	0.9876	0.9815	0.9846	0.9809
ANN	0.9358	0.9679	0.9263	0.9467	0.9386



Based on the evaluation metrics given above, it can be clearly seen that the Random Forest model produces the best overall results. Therefore, it is taken as the primary model for the implementation.

6. Ethical Considerations & Bias Mitigation

This section provides a comprehensive ethical analysis and bias mitigation plan for the Heart Attack Prediction Model, ensuring its development and deployment adhere to the core ethical principles of fairness, transparency, accountability, and privacy.

1. Data Privacy and Consent

The model utilizes sensitive health attributes such as age, cholesterol levels, resting blood pressure, and fasting blood sugar. To uphold the Privacy & Data Protection principle, all personally identifiable information is removed before analysis. Data handling strictly adheres to data protection regulations like GDPR or equivalent local standards. Crucially, in real-world clinical deployment, explicit patient consent must be obtained for any data collection or predictive modelling activities, reinforcing the commitment to secure handling of personal health information.

2. Fairness and Bias Mitigation

Bias in medical prediction systems is one of the most pressing ethical challenges, as it can lead to systematic unfairness and discriminatory outcomes for certain groups.

- **Bias Detection:** To ensure **Fairness**, the model is rigorously evaluated across demographic subgroups, specifically gender and age, to detect potential **Data Bias** (where historical clinical data may underrepresent certain populations) or **Sampling Bias** (where the training set does not fully represent the diversity of the target population).
- **Mitigation Strategy:** If imbalance or systematic prejudice is detected, techniques such as oversampling underrepresented groups or using stratified sampling are applied.
- **Evaluation Metrics:** We will monitor fairness-aware metrics (like demographic parity and equal opportunity difference) to assess **equity** in model outcomes, ensuring the system treats individuals and groups equitably and avoids indirect discrimination.

3. Transparency and Explainability

Transparency and Explainability are critical to prevent "black-box" models and foster trust in healthcare AI. The model's decisions must be interpretable to medical professionals.

- **Explainable AI (XAI) Techniques:** Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) are used to explain feature contributions.
- **Usability:** Visualizations of feature importance allow clinicians to understand which

factors most influence predictions, improving trust, usability, and enabling **Human Oversight** over critical diagnostic support.

4. Accountability and Responsibility

The model serves strictly as a **decision-support tool** rather than a replacement for medical expertise. This upholds the principle of **Human Oversight** in critical systems.

- **Verification:** All predictions must be verified by qualified healthcare professionals before diagnostic or treatment decisions are made.
- **Governance:** Model development, evaluation, and deployment are documented via **Model Cards or Datasheets** for full traceability and to ensure that **Responsibility** is clear if the system causes harm.
- **Auditing:** Periodic audits and retraining ensure continuous improvement, ethical compliance, and robustness against errors.

5. Social and Ethical Impact

The deployment of this system must promote **Beneficence** (well-being) and **Non-Maleficence** (avoiding harm).

- **Intended Use:** The results of the model should be used solely to enhance patient care, health awareness, and early intervention.
- **Preventing Misuse:** We actively prevent potential misuse, such as discrimination in insurance, employment, or allocation of healthcare resources, ensuring the system does not perpetuate or amplify social inequalities.

6. Continuous Monitoring and Responsible AI

Ethical AI requires ongoing monitoring after deployment to maintain integrity.

- **Drift Detection:** The model's accuracy, **Fairness**, and interpretability are continuously assessed to detect **bias drift** (a shift in bias over time) or ethical violations post-deployment.
- **Refinement:** Feedback from healthcare practitioners and patients is formally integrated to refine the system, completing a continuous loop of evaluation, accountability, and transparency to ensure the model remains safe, fair, and trustworthy.

6. Reflections and Lessons Learned

6.1 Process

Initially, the issue of the need to have a tool to predict heart attacks was identified. Secondly, suitable datasets with potential to be trained to address this issue were collected. After careful analysis, a primary dataset was selected based on the quality of the data. Afterwards, the dataset was cleaned by utilizing various preprocessing techniques. Next, the preprocessed dataset was studied and the features with highest importance were extracted. Then various models were trained using the same dataset to obtain evaluation metrics. After an extensive comparison between the performances of each model, Random Forest was selected as the best model. Next, the model was optimized through hyperparameter tuning and steps were taken to prevent model overfitting and underfitting.

6.2 Challenges

- Overfitting during model training
- Collaboration and task assignments in group
- Selecting the right hyperparameters to be tuned
- Comparison between Random Forest and Decision Tree before tuning hyper parameters

6.3 Future Improvements

- Use advanced models (e.g: XGBoost)
- Deploy the model as a web application
- Automate hyperparameter tuning with GridSearchCV

6.4 Team Insight

Each team member had different strengths and weaknesses. Therefore, taking that into consideration the tasks were divided in a fair manner. Regular online meetings allowed each member to participate actively and share their opinions in a space that treated each opinion equally. The collaboration significantly improved the technical knowledge of each member and provided the members with the ability to plan and execute a machine learning project effectively

References

[1]
"Heart Disease Facts | Heart Disease | CDC." Accessed: Oct. 1, 2025. [Online]. Available: https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

[2]
S. S. Maghdid and T. A. Rashid, "An Extensive Dataset for the Heart Disease Classification System," vol. 1, 2022, doi: 10.17632/65GXGY2NMG.1.

[3]
"Cardiovascular Diseases - Our World in Data." Accessed: Oct. 1, 2025. [Online]. Available: https://ourworldindata.org/cardiovascular-diseases