Heart Attack Risk Prediction Using Data Mining Techniques Checkpoint

Michael Jung
michael.jung@emory.edu
Emory University
, Atlanta , Georgia , USA

Caleb Jennings
cbjenn2@emory.edu
Emory University
, Atlanta , Georgia , USA

Sungho Kim
sungho.kim@emory.edu
Emory University
, Atlanta , Georgia , USA

1 ABSTRACT

1.1 Topic background

Cardiovascular disease, particularly heart attacks, is a leading cause of death worldwide. Early detection and risk prediction are essential to reducing the burden on healthcare systems and improving patient outcomes. This project aims to apply data mining techniques to predict the risk of heart attacks in individuals, with a particular focus on identifying how risk factors vary across different age groups and health conditions.

1.2 Motivation

Our goal is to create a model predicting the risk of heart attacks for individuals, with a particular focus on identifying how risk factors vary across different age groups and health conditions. The dataset we use has over 50,000 instances with 20 related attributes which allows for wide-ranging model selection and unbiased results. With this large dataset we could use various data mining techniques to give insight and show helpful patterns related to cardiovascular diseases.

1.3 Results

The results collected from different prediction models will provide insights on heart disease risk factors

- **Top 5 risk factors**- Each model will output influential risk factors strongly associated with heart attack risk in general
- **Top 5 risk factors per age group** - The dataset is divided into 3 age groups and top 5 risk factors are calculated for each groups
- **Risk group classification** - Identification of age groups with higher or lower heart attack risk based on the input attributes, potentially finding trends across age groups

2 INTRODUCTION

2.1 Background

Heart attacks claim millions of lives globally each year. While considerable research has been conducted on heart disease prediction, gaps remain in how demographic and lifestyle factors contribute to risk across different age groups.

2.2 Challenges Faced

One of the primary challenges in this domain is managing noisy data while ensuring that predictive models generalize effectively across diverse patient populations. Moreover, biomedical datasets often contain features labeled with specialized clinical terminology—such as "ST-T wave abnormalities," "left ventricular hypertrophy," or "serum creatinine levels"—which can be difficult to interpret without domain-specific medical knowledge. This complexity poses additional barriers to data preprocessing, feature selection, and model explainability, particularly for interdisciplinary teams lacking clinical backgrounds.

3 RELATED WORK

3.1 Heart Disease UCI Diagnosis & Prediction

In a study utilizing the UCI Heart Disease dataset, which contains approximately 300 instances and 14 key attributes, researcher Hardick Deshmukh developed a predictive model aimed at assessing a patient's likelihood of experiencing a heart attack. By leveraging these attributes, the model was able to achieve an accuracy of 87%, determined by AUROC.

3.2 Early Prediction of Heart Disease Using PCA and HGA with k-Means

This research applies Principal Component Analysis (PCA) to reduce the dimensionality in the UCI Heart Disease dataset. It then uses k-means clustering to classify the data but improves upon it by incorporating a Hybrid Genetic Algorithm (HGA) to avoid local optima. By combining these techniques, the proposed model achieves an accuracy of 94.06

3.3 A Fast Algorithm for Heart Disease Prediction Using Bayesian Network Model

This research applies Bayesian Network (BN) modeling to analyze the relationships between 14 attributes in the UCI Heart Disease dataset. The study investigates how dependencies between attributes impact classification performance. The BN provides a clear graphical representation of these relationships and is capable of predicting new scenarios. The model achieves an accuracy of 85

4 METHODOLOGIES

4.1 Dataset

The original dataset we selected became unavailable for future use, leading us to switch to a new **Heart Attack Risk Dataset** from Kaggle. This dataset contains 8763 instances and 26 attributes, including features such as age, gender, heart rate, cholesterol levels, and lifestyle habits (e.g., smoking and alcohol consumption). Compared to the previous dataset, the current one offers more meaningful attributes, providing us with more data to improve the performance of our models.

4.2 Common Data Preprocessing - Feature Engineering

The data preprocessing phase is crucial for improving the performance and efficiency of the machine learning models. The following steps will be performed:

(1) **Dropping unnecessary columns:** We removed columns that were not relevant to the heart attack prediction task.
   - **Patient ID**
   - Locational data such as **Country**, **Continent**, and **Hemisphere**, which did not contribute valuable information for the model
(2) **Transforming Blood Pressure Data:** The Blood Pressure column was split into three distinct columns.
   - **SBP (Systolic Blood Pressure):** The top number in a blood pressure reading, which represents the pressure in the arteries when the heart beats and pumps blood.
   - **DBP (Diastolic Blood Pressure):** The bottom number, representing the pressure in the arteries when the heart is at rest between beats.
   - **MAP (Mean Arterial Pressure):** Calculated as the average pressure in the arteries, calculated with the formula:

$$MAP = \frac{SBP + 2 \times DBP}{2}$$

     MAP is important because it reflects the level of blood pressure required to ensure that the body's organs receive enough blood flow. Many experts view it as a more reliable measure of perfusion than SBP.
(3) **One-Hot Encoding Sex Column:** The Sex column was transformed into two binary columns using one-hot encoding. The encoding follows the mapping: Male: 1 / Female: 0
(4) **Mapping the Diet Column:** The Diet column, which initially had categorical values of Unhealthy, Average, and Healthy, was mapped to numerical values in the form of Unhealthy: 1 / Average: 2 / Healthy: 3
(5) **Converting Boolean Columns to Integers:** All Boolean columns (such as Diabetes, Obesity, etc.) were converted to integers for compatibility with the machine learning models like True $\rightarrow$ 1 / False $\rightarrow$ 0

4.3 Classification Models and further data preprocessing

Various machine learning algorithms will be employed to predict the risk of heart attacks. Each model used further preprocessing steps to increase its performance. Models we used include:

- **Logistic Regression**: A foundational classification technique for binary outcomes, useful for modeling the relationship between input features and the likelihood of heart disease.
- **Random Forest**: An ensemble learning method that combines multiple decision trees to improve the accuracy, robustness, and interpretability of predictions.
- **Neural Network**: A flexible and powerful model inspired by the human brain, capable of capturing complex, non-linear relationships between features and heart disease risk through multiple interconnected layers of computation.

4.4 Evaluation Metrics

Since the dataset's class is imbalanced, we focused on other metrics instead of accuracy. The model's performance will be assessed using the following metrics:

- **Accuracy**: To evaluate the overall performance of the model in making correct predictions.
- **Balanced Accuracy**: To evaluate the balanced performance of the model.
- **Precision and Recall**: To assess the model's ability to correctly identify true positives and true negatives, especially in the context of imbalanced classes.
- **F1-Score**: To provide a balanced measure of precision and recall, particularly useful when dealing with imbalanced datasets.
- **AUC (Area Under the ROC Curve)**: To assess the model's ability to distinguish between positive and negative classes, particularly useful in binary classification tasks.

5 EXPERIMENTS

5.1 General risk factor evaluation

To evaluate the most influential risk factors for heart disease, we conducted experiments using three different machine learning models: Logistic Regression, Random Forest, and Neural Network. Each model was trained on the same preprocessed dataset to ensure consistency across comparisons. In addition to the common preprocessing steps applied to all models, each classification model underwent additional model-specific preprocessing. For example, Logistic Regression required feature normalization using Z-score scaling, Neural Networks used Min-Max normalization to improve convergence, while Random Forest, being tree-based, performed best without further normalization. The goal of these experiments was twofold:

(1) **To assess model performance** using standard classification metrics such as Accuracy, Precision, Recall, F1-Score, and AUROC.
(2) **To identify the top 5 most influential features (risk factors)** for each model, based on the model's internal importance measures:
   - For Logistic Regression, coefficients with the highest absolute values were selected.
   - For Random Forest, feature importances were derived from the mean decrease in impurity.
   - For the Neural Network, permutation feature importance was used to interpret non-linear relationships.

Each model was evaluated using a stratified 80/20 train-test split to preserve class balance, and standard scaling was applied where needed. The top features were extracted from the trained models on the test data to reflect generalizable insights, not just training behavior. This evaluation provides both quantitative performance

insights and qualitative understanding of which features most contribute to predicting heart disease risk across different modeling techniques.

5.2 Risk factor evaluation for 3 different age groups

To evaluate the risk factors for heart disease across different age groups, the following steps were performed:

(1) **Data Division by Age Groups:** The dataset was divided into three age groups:
   - Young ($\leq$ 30 years)
   - Middle-aged (31-60 years)
   - Senior ($\geq$ 61 years)

(2) **Preprocessing and Evaluation:** The preprocessing and evaluation methods followed the same approach used for the general dataset.

(3) **Feature Evaluation:** For each age group, the top 5 risk factors were identified based on feature importance. These risk factors were then compared across the age groups to understand how the risk factors differed with age and compared with the general risk factors to understand the risk factor shift over time.

6 PERFORMANCE

6.1 Model Performances for general risk factor evaluation

This section holds the model performances from using the entire dataset.

6.1.1 Logistic Regression.

- **Accuracy**: 0.4969
- **Recall**: 0.49
- **Precision**: 0.49
- **F1-Score**: 0.48
- **AUROC**: 0.4925

6.1.2 Random Forest.

- **Accuracy**: 0.6378
- **Recall**: 0.64
- **Precision**: 0.2667
- **F1-Score**: 1.24
- **AUROC**: 0.4896

6.1.3 Neural Network.

- **Accuracy**: 0.5499
- **Recall**: 0.3025
- **Precision**: 0.3514
- **F1-Score**: 0.3247
- **AUROC**: 0.4976

6.2 Top 5 general risk factors from each model

This section holds the top 5 risk factors that were predicted from each model

6.2.1 Logistic Regression.

- **Smoking**
- **Obesity**
- **Diabetes**
- **Sex_Male**
- **Cholesterol**

6.2.2 Random Forest.

- **Exercise Hours Per Week**
- **BMI**
- **Sedentary Hourse Per Day**
- **Income**
- **Triglycerides**

6.2.3 Neural Network.

- **Age**
- **Medication Use**
- **Family History**
- **Sleep Hours Per Day**
- **Physical Activity Days Per Week**

7 DISCUSSION

In our experiments on the general dataset, we tested three different models: Logistic Regression (LR), Random Forest (RF), and a Neural Network (NN). The results revealed clear differences in model performance and exposed some limitations in our current approach.

The Random Forest and Neural Network models achieved relatively high accuracy scores, suggesting that they were able to learn meaningful patterns from the data. However, both models exhibited surprisingly low AUROC (Area Under the Receiver Operating Characteristic) scores. This indicates that while they are good at predicting the majority class, they struggle to distinguish effectively between positive and negative cases — a potential sign of class imbalance or insufficient feature differentiation.

In contrast, Logistic Regression performed poorly in both accuracy and AUROC, with values below 50

These results raise important questions about model suitability. Although RF and NN appear better on the surface due to higher accuracy, the low AUROC scores indicate a need for further refinement. Possible next steps include exploring different feature engineering strategies, addressing class imbalance more directly, or testing additional models such as Gradient Boosting, Support Vector Machines, or ensemble approaches.

Ultimately, our goal is to develop a model that not only performs well on overall accuracy but also has strong discriminative power. Until we achieve both, we cannot confidently rely on these predictions for real-world applications.

8 FUTURE WORKS

Future work could expand the dataset with additional medical data to enhance prediction accuracy. We also propose investigating the impact of forced class imbalance on high-risk labels to improve the model's sensitivity toward underrepresented, high-risk populations, potentially improving early detection.

REFERENCES

[1] Md. Touhidul Islam, Sanjida Reza Rafa, and Md. Golam Kibria. Early prediction of heart disease using PCA and hybrid genetic algorithm with k-means. *CoRR*, abs/2101.00183, 2021.

[2] Arif Miah. heart attack risk dataset, 2025. https://www.kaggle.com/datasets/arifmia/heart-attack-risk-dataset/data.
[3] Mistura Muibideen and Rajesh Prasad. A fast algorithm for heart disease prediction using bayesian network model. *CoRR*, abs/2012.09429, 2020.
[4] Towards Data Science. Heart disease (uci) diagnosis prediction. *Medium*, 2025. Accessed: 2025-02-17.
[5] EMTPrep Staff. Map - understanding mean arterial pressure, 2022. https://emtprep.com/resources/article/map-understanding-mean-arterial-pressure.
[6] Chris Vincent. Systolic vs. diastolic blood pressure, 2023. https://www.verywellhealth.com/systolic-and-diastolic-blood-pressure-1746075.

[2] [5] [6] [3] [1] [4]