Heart Attack Risk Prediction Using Data Mining Techniques Proposal

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 Contributions

This project utilizes data mining techniques to identify the patterns in heart attack risk across age groups using demographics and lifestyle attributes to uncover insights that could contribute to prevention strategies and personalized healthcare.

1.4 Results

The proposed models will provide accurate predictions and identify age-related risk factors.

- **Most influential risk factors**- The model will output influential risk factors strongly associated with heart attack risk in general and per age group
- **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
- **Predictive Risk score** - score indicating whether an individual is at high risk or low risk of a heart attack

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 Expected Challenges

The key challenges include handling noisy data and ensuring models generalize well across diverse patient demographics. Other challenges stem from the fact that identifying leading causes for heart attacks is already a well-researched area, and several labs with a large amount of resources have already unsuccessfully tackled this problem. So finding new insights and information will be very strenuous.

PCA will be used for dimensionality reduction and cross-validation will be used to avoid overfitting. The large datasets we have were gathered during extensive research and already come with many useful statistical insights that can help us to build off of others' previous work in order to yield new results.

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 Data Preprocessing

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

- **Feature Engineering**: Selecting, creating, or modifying features to improve the model's ability to make predictions, particularly focusing on identifying relevant risk factors.
- **Normalization**: Scaling the features to a consistent range, ensuring that no single feature dominates the learning process, particularly important for models like K-Nearest Neighbors.

4.2 Classification Models

Various machine learning algorithms will be employed to predict the risk of heart attacks. These 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 accuracy, robustness, and interpretability of predictions.
- **XGBoost**: A powerful boosting technique designed for performance and speed, particularly effective in handling complex patterns and imbalances in the data.
- **K-Nearest Neighbors (K-NN)**: A simple yet effective model for classification that assigns a label based on the majority class of neighboring data points, sensitive to feature scaling.

4.3 Clustering and Dimensionality Reduction

To explore hidden patterns and enhance model interpretability, the following techniques will be applied:

- **K-Means Clustering**: An unsupervised learning method used to identify groups of similar data points, which can reveal important patterns in the heart attack risk factors across different age groups.
- **Principal Component Analysis (PCA)**: A dimensionality reduction technique that will be used to reduce the number of features while retaining the most significant variance, improving computational efficiency and model performance.

5 EXPERIMENTS

5.1 Datasets

We will use the **Heart Attack Risk Dataset** from Kaggle, which includes features like age, gender, heart rate, cholesterol levels, and lifestyle habits (e.g., smoking and drinking) having a total of 50,000 instances and 20 attributes.

5.2 Metrics

The model's performance will be assessed using the following metrics:

- **Accuracy**: To evaluate the overall performance of the model in making correct predictions.
- **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.

6 PERFORMANCE

The performance of the model will be evaluated using several visualization techniques to gain insights into its effectiveness and interpretability:

- **Correlation Heatmaps**: To examine the relationships between input features and heart attack risk.
- **ROC Curves**: To visualize the tradeoff between true positive and false positive rates across different decision thresholds.
- **Confusion Matrix**: To assess classification performance, highlighting key metrics such as true positives, false positives, and overall accuracy.
- **Bar Charts or Histograms**: To display the distribution of key risk factors across different age groups and identify trends.
- **Clustering Visualizations**: To visualize clusters formed by K-Means, helping to identify groupings and patterns within the dataset.

7 DISCUSSION

This proposal aims to predict heart attack risk using data mining techniques on various attributes, focusing on age-specific risk factors for a more personalized healthcare approach. The results could aid early detection and prevention, helping reduce heart attack incidence by identifying risk thresholds across age groups.

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.

[3] [1] [2] [4]