Predicting Heart Disease: A KDD Approach Using the Heart Disease UCI Dataset

Aishly Manglani
San Jose State University
San Jose, CA

September 30, 2024

Abstract

Heart disease remains a leading cause of mortality worldwide. Early detection and intervention are crucial for improving patient outcomes. This paper presents a Knowledge Discovery in Databases (KDD) process applied to the Heart Disease UCI dataset, aiming to predict the presence of heart disease based on various medical attributes.

1 Introduction

Heart disease continues to be one of the leading causes of death globally. Early detection and effective management are crucial for improving patient outcomes. The Heart Disease UCI dataset provides a rich source of data for predictive modeling in this domain. This paper outlines the KDD process applied to this dataset, including data selection, preprocessing, transformation, mining, interpretation, and knowledge representation.

2 Dataset Overview

The **Heart Disease UCI** dataset consists of 303 samples with 14 features related to heart health. These features include:

- **Age**: Age of the patient
- **Sex**: Gender of the patient
- **CP**: Chest pain type (0-3)
- **Trestbps**: Resting blood pressure
- **Chol**: Serum cholesterol in mg/dl
- **Fbs**: Fasting blood sugar ¿ 120 mg/dl (1 = true; 0 = false)

- **Restecg**: Resting electrocardiographic results (0-2)
- **Thalach**: Maximum heart rate achieved
- **Exang**: Exercise induced angina (1 = yes; 0 = no)
- **Oldpeak**: ST depression induced by exercise
- **Slope**: Slope of the peak exercise ST segment (0-2)
- **Ca**: Number of major vessels colored by fluoroscopy (0-3)
- **Thal**: Thalassemia (1 = normal; 2 = fixed defect; 3 = reversible defect)
- **Num**: Diagnosis of heart disease (1 = presence; 0 = absence)

3 KDD Process

The KDD process involves several key phases: data selection, preprocessing, transformation, mining, interpretation, and knowledge representation. Below is an overview of each phase.

3.1 Data Selection

The first step in the KDD process involves selecting the appropriate dataset for analysis. The Heart Disease UCI dataset was sourced from a publicly accessible URL. Initial exploration of the dataset helps understand its structure and the types of features available for analysis.

3.2 Data Preprocessing

Data preprocessing is crucial to ensure the quality of the dataset before analysis. This step involves checking for missing values and assessing the data types of each feature. In this dataset, some features had missing values that needed to be addressed. Additionally, categorical features may need to be converted to a suitable format for analysis.

3.3 Data Transformation

Once the dataset is preprocessed, the next step is data transformation. This involves normalizing numerical features to ensure they contribute equally to the analysis. Transforming categorical features into numerical representations, such as one-hot encoding, is also necessary for compatibility with machine learning algorithms.

3.4 Data Mining

In the data mining phase, various machine learning algorithms are applied to discover patterns and relationships within the data. The Random Forest algorithm was utilized to predict the presence of heart disease based on the available features. The dataset was split into training and testing sets to evaluate the model's performance effectively.

3.5 Interpretation

The interpretation phase involves analyzing the results obtained from the data mining step. The model achieved an accuracy score of approximately 85%. A classification report provides detailed metrics such as precision, recall, and F1-score for both classes (presence and absence of heart disease). This information is essential for understanding the model's strengths and weaknesses.

3.6 Knowledge Representation

Finally, knowledge representation involves visualizing the results and presenting the findings in a comprehensible manner. Feature importance analysis helps identify which features contributed most to the model's predictions. Visualizations, such as bar charts, can effectively communicate these insights, aiding stakeholders in understanding the key factors influencing heart health.

4 Conclusion

The KDD process applied to the Heart Disease UCI dataset demonstrates the potential of machine learning to assist in predicting heart disease. The Random Forest model achieved satisfactory accuracy, and feature importance analysis revealed critical factors affecting heart health. Future research could explore additional models and refine feature selection to enhance predictive performance.

5 References

References

[1] UCI Machine Learning Repository. *Heart Disease Dataset*. [Online]. Available: `https://archive.ics.uci.edu/ml/datasets/heart+Disease`.

[2] Kaggle. *Heart Disease UCI Dataset*. [Online]. Available: `https://www.kaggle.com/ronitf/heart-disease-uci`.