Pace University

Department of Computer Science

PROJECT REPORT

On

"Data Mining Exploring and Application"

Submitted by :

NAME : Dhanush Sai Ram Bezawada.

UID : DB67875N

CLASS : Data Mining

Abstract

Heart disease is a leading cause of mortality worldwide. Predicting its occurence using clinical data can help improve early detection and treatment. This study applies data mining techniques on a heart disease dataset to explore relationships between clinical attributes and the presence of heart disease. We employed classification models like Decision Tree, Logistic Regression and Support Vector Machine as well as k-Means clustering with PCA to reduce dimensionality. The Decision Tree model achieved the highest accuracy (98.5%). The results highlight the effectiveness of machine learning in medical diagnosis Support systems.

1. Introduction

Heart disease encompasses a range of conditions affecting the heart. Early diagnosis can significantly reduce the risk of serious complications or death. The objective is to build predictive models and apply unsupervised clustering to gain insights into patters in the data.

2. Methodology

2.1 Dataset Overview

The dataset contains of 14 features and one target column indicating the presence (1) or absence (0) of heart disease.

The features include:

- Age, sex, chest pain type (cp), resting blood pressure (trestbps), cholestrol (chol)
- fasting, blood sugar (fbs), resting electrocardiographic results (restecg)
- Maximum heart rate achieved (thalach), exercise induced angina (exang), ST depression (oldpeak)
- Slope of peak exercise ST segment (slope), number of major vessels (ca), thalassemia (thal)

No missing values are found. All columns are of numerical type.

2.2 Exploratory Data Analysis

Initial EDA included visualizing the distribution of the target variable, showing a balanced dataset. key insights included:

- Negative correlation between thalach and target.
- Positive correlation between cp and presence of heart disease.
- exang and oldpeak also showed strong negative correlation with the target.

2.3- Preprocessing

All features were already numeric, categorical columns like cp, slope and thal were label encoded. feature scaling was performed using Standard Scaler for algorithms sensitive to magnitude.

2.4 Predicting Modeling

Decision Tree

- Accuracy : 98.5%
- Confusion matrix

	Predicted class 0	Predicted class 1
Actual class 0	102	0
Actual class 1	3	100

- Precision
  - class 0 : 0.97
  - class 1 : 1.00
- Recall
  - class 0 : 1.00
  - class 1 : 0.97
- F1 score
  - class 0 : 0.99
  - class 1 : 0.99

Logistic Regression

- Accuracy : 79.5%
- Confusion matrix

	Predicted class 0	Predicted class 1
Actual class 0	73	29
Actual class 1	13	90

- Precision
  - class 0 : 0.85
  - class 1 : 0.76
- Recall
  - class 0 : 0.72
  - class 1 : 0.87
- F1 score
  - class 0 : 0.78
  - class 1 : 0.81

Support Vector Machine

- Accuracy : 88.7%
- Confusion Matrix

	Predicted class 0	Predicted class 1
Actual class 0	85	17
Actual class 1	6	97

- Precision

  class 0 : 0.93 class 1 : 0.85

- Recall

  class 0 : 0.83 class 1 : 0.94

- F1 score

  class 0 : 0.88 class 1 : 0.89

2.4 Clustering results

- k-means clustering : Applied with 2 clusters to match the binary classification.
- Dimensionality reduced to 2D using PCA for visualization
- The 2D PCA projection visually indicated seperation between clusters.
- Adjusted Rand Index : 0.376.

3. Findings

- Decision Tree significantly outperformed the other classifier with 98.5% accuracy and high precision and recall
- SVM showed strong performance and balanced precision/recall
- Logistic Regression underperformed due to its linear nature

and inability to capture complex boundaries.

4. Discussion

The high accuracy of the Decision Tree model shows its suitability for heart disease prediction in structured data. Although logistic Regression is interpretable, it failed to match the accuracy of tree-based or kernel-based models. Limitations include potential overfitting in Decision Trees and limited generalizability of the ARI measure.

Challenges faced:

- Ensuring balanced data after preprocessing
- Avoiding overfitting in the Decision Tree model.
- Visualizing high-dimensional data through PCA.

5. Conclusion

This project successfully demonstrated the application of classification and clustering algorithms in heart disease prediction. The Decision Tree classifier provided the best results with 98.5.1. accuracy, making it suitable for medical diagnosis support. Clustering revealed meaningful but imperfect group seperation, as shown by the Adjusted Rand Index. The study confirms that machine learning can be a powerful tool in healthcare analytics when supported by quality data and thoughtful preprocessing. The frequent itemsets

and top rules revealed strong purchasing patterns, such as customers commonly buying themed products together. These insights offer practical applications for improving product placement, cross-selling strategies and inventory planning. The visualizations further illustrated the strength of relationships between key items. With accurate preprocessing and parameter tuning, association rule mining proves to be a powerful technique for discovering hidden patterns in transactional data. Overall, the results provide actionable insights that can enhance business decision-making.