Predicting Heart Disease Using Data Mining Techniques

Advanced Big Data and Data Mining (MSCS-634-B01)

Pawan Pandey

August 22, 2025

Project Overview

- Dataset: Heart Disease UCI, 1026 records, 14 features
- Objective: Predict presence/absence of heart disease
- Features: Age, Sex, Chest Pain Type, Cholesterol, Max Heart Rate
- Goal: Provide actionable healthcare insights

Data Preprocessing & Feature Engineering

- Checked for missing values and removed duplicates
- Encoded categorical features, scaled numeric variables
- Created derived features: chol_high, bp_high
- Ensured dataset ready for modeling

Regression Analysis

- Models: Linear, Ridge, Lasso Regression
- Evaluated using RMSE and R^2 metrics
- Ridge Regression performed best

Classification Analysis

- Models: Logistic Regression, KNN, Decision Tree, Random Forest, SVM, Naive Bayes
- Evaluated with Accuracy, F1-Score, ROC-AUC
- Random Forest & Decision Tree were top performers
- Models help identify high-risk patients

Clustering Analysis

- Algorithm: K-Means (n_clusters=2)
- Evaluated with Silhouette Score
- Clusters partially aligned with heart disease presence
- Helps segment patients for targeted interventions

Association Rule Mining

- **Algorithm:** Apriori for discovering frequent patterns
- **Features used**: chol_high, bp_high, target
- High cholesterol & high BP → higher heart disease risk
- Provides actionable insights for preventive care

Practical Recommendations

- **Use Random Forest/Decision Tree for clinical prediction**
- **Monitor key features: cp, thalach, oldpeak, chol**
- **Apply clustering for patient segmentation**
- **Leverage association rules to guide preventive strategies**

Challenges & Future Improvements

- Mixed data types and feature interpretation
- Class imbalance and overfitting in trees
- Need larger datasets and more granular features
- Improve feature engineering for better predictions

Conclusion

- Multi-technique approach provides comprehensive insights
- Early detection and preventive strategies enabled
- Regression estimates risk, classification predicts presence
- Future work: retrain with new data for improved accuracy

THANK YOU