HEART DISEASE PREDICTION

Authors:

Erhard Rahm and Philip A. Bernstein

The VLDB Journal, Volume 10, Issue 4, 2001, Pages 334-350

Introduction

Heart disease remains one of the leading causes of death worldwide, necessitating advancements in early diagnosis and prediction tools. The Heart Disease Predictor Project aims to develop a predictive model using machine learning techniques to identify individuals at high risk of heart disease.

The project integrates data from multiple sources, including electronic health records (EHRs), patient surveys, and clinical trials, to build a comprehensive dataset. Inspired by the schema matching approaches discussed by Rahm and Bernstein (2001), this project employs advanced techniques to handle schema heterogeneity across diverse data sources.

Methodology

1. Data Collection and Preprocessing

- Data Sources:
  Aggregated health data from hospitals, public health databases, and fitness trackers.

- Schema Matching:
  Leveraged rule-based and machine-learning schema matching approaches to align and integrate diverse datasets. This ensured a consistent and unified schema for predictive modeling.

2. Feature Selection

Key features include age, cholesterol levels, blood pressure, glucose levels, smoking status, physical activity, and family medical history. Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) were used to identify the most influential predictors.

3. Predictive Modeling

- Algorithms:
  Logistic Regression, Random Forest, and Neural Networks were employed to develop and evaluate predictive models.
- Model Evaluation:
  Precision, recall, F1 score, and area under the ROC curve (AUC) were used to measure performance.

4. Implementation Framework:

The project was developed using Python libraries such as Pandas for data manipulation, Scikit-learn for machine learning, and Flask for deploying the web application.

Conclusion

The Heart Disease Predictor Project successfully demonstrates the potential of machine learning in early detection of heart disease. The integration of heterogeneous data sources through schema matching significantly improved the accuracy and reliability of predictions. Future work includes incorporating real-time data streams and enhancing the model with more diverse datasets to address population-specific variations.