Data Mining Indicators of Heart Disease

Cyro de Lima
CSPB 4502
University of Colorado Boulder
cyfr7592@colorado.edu

Eric Hernandez
CSPB 4502
University of Colorado Boulder
Eric.Hernandez@colorado.edu

Lindsay Bordelon
CSPB 4502
University of Colorado Boulder
Lindsay.Bordelon@colorado.edu

ABSTRACT

Heart disease remains one of the leading causes of mortality worldwide, affecting individuals across various demographics. Despite advancements in healthcare, identifying and mitigating the risk factors associated with heart disease remains a critical challenge. This project aims to leverage machine learning techniques on a comprehensive dataset sourced from the Centers for Disease Control and Prevention (CDC) to predict the likelihood of heart disease based on key indicators.

Efficiently mining large-scale datasets like the CDC's Behavioral Risk Factor Surveillance System offers invaluable opportunities to uncover hidden patterns and insights relevant to heart disease prevention and management. By exploring factors such as high blood pressure, cholesterol levels, and behavioral attributes, we can extract meaningful correlations and predictive features. Through advanced machine learning techniques, we aspire to develop robust models capable of accurately identifying individuals at risk of heart disease. These models not only provide predictive capabilities but also offer actionable insights for personalized interventions and targeted healthcare strategies. By leveraging data mining methodologies, this project aims to contribute to the ongoing efforts in public health by empowering stakeholders with actionable intelligence to mitigate the impact of heart disease on individuals and communities.

KEYWORDS

Exploratory Data Analysis; Data Mining; Heart Disease; Data Visualization; Statistical Analysis

ACM Reference format:

Mohamed Elsayed, 2022. Heart Disease Prediction: Impact of Decision Trees and KNN Algorithms. https://www.kaggle.com/code/georgyzubkov/heart-disease-exploratory-data-analysis

1 Literature Survey

George Zubkov's work on exploratory data analysis sheds light on the primary factors contributing to cardiovascular diseases. Key issues identified include physical inactivity, mental health concerns, stress, and unhealthy habits such as alcohol consumption and excessive sugar intake.

In Mohamed Elsayed's research on heart disease prediction, Decision Trees and KNN algorithms are employed to forecast factors impacting heart conditions. The study emphasizes that individuals with difficult walking, a history of stroke, diabetes, and poor physical health are primary factors. The accuracy achieved using KNN was reported to be 0.71.

2 Proposed Work

2.1 Data Cleaning:
- Review and analyze data set
- Address missing values (if we use the NaN-included dataset)
- Update binary answers
- Remove duplicates & outliers
- Identify and correct data anomalies or inconsistencies
- Normalize numerical variables if necessary (e.g., scale or transform features to have a mean of 0 and a standard deviation of 1)

2.2 Data Preprocessing:
- Assign unique identifier
- Assign 'HEART ATTACK: Y/N' as 'HEART DISEASE: Y/N'
- Initial visualizations
- Split the dataset into training, validation, and testing sets
- Standardize or normalize numerical features to ensure they have similar scales

- Apply techniques to address multicollinearity among features
- Perform dimensionality reduction techniques (e.g., principal component analysis) if dealing with high-dimensional data

3.3 Data Integration:
- If enough time permits:
- Integrate health insurance rates per state to find correlations with heart disease
- Compare 2022 & 2020 analysis

3 Data Set

- Indicators of Heart Disease Data Set
400K+ Records (Some include NaN values)
Heath interview answers from 400K+ individuals
- "established in 1984 with 15 states, BRFSS now collects data in all 50 states, the district of columbia, and three u.s. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world." –cdc
- "a heart attack is often an indicator of heart disease."- google
- dataset will be downloaded on all members' machines and connected via github

4 Evaluation Methods

Cross-Validation: Cross-validation involves splitting the dataset into multiple subsets (folds), training the model on a subset of the data, and evaluating its performance on the remaining subset. This process is repeated multiple times, with different subsets used for training and testing, and the results are averaged to obtain a more robust estimate of the model's performance.

Holdout Validation: Holdout validation involves randomly splitting the dataset into a training set and a separate validation set. The model is trained on the training set and evaluated on the validation set. This method provides an estimate of the model's performance on unseen data.

Bootstrapping: Bootstrapping involves generating multiple random samples (with replacement) from the original dataset and training the model on each sample. The model's performance is then evaluated on the original dataset or a separate validation set. Bootstrapping allows for estimating the variability of the model's performance and obtaining confidence intervals for performance metrics.

Confusion Matrix: A confusion matrix provides a tabular representation of the model's predictions compared to the actual class labels. From the confusion matrix, various performance metrics such as accuracy, precision, recall, and F1 score can be calculated.

5 Tools

- KAGGLE
- GITHUB
- JUPYTERLAB
- EXCEL
- PYTHON
- PANDAS
- SKLEARN
- MATPLOTLIB
- SNS
- DISCORD

6 Milestones

Data Preparation:
- Clean the dataset by addressing missing values, outliers, and inconsistencies.
- Perform feature engineering and data preprocessing tasks.
- Split the dataset into training, validation, and testing sets.

Model Development:
- Choose appropriate data mining techniques and algorithms.
- Develop and train predictive models.
- Optimize model parameters and hyperparameters for improved performance.

Model Evaluation:
- Evaluate model performance using appropriate metrics.
- Validate models using cross-validation or holdout techniques.

Results Interpretation and Reporting:
- Interpret model results and analyze insights gained.
- Document the data mining process and prepare a final report summarizing findings.
- Present results to stakeholders, discussing implications for decision-making.

ACKNOWLEDGMENTS

We extend our gratitude to Professor Peterson and our team members, Cyro de Lima, Eric Hernandez, and Lindsay Bordelon, for their invaluable contributions to this project. Their dedication and collaboration have been essential in the completion of this work.

REFERENCES

[1] Georgy Zubkov, "Heart Disease Exploratory Data Analysis", Kaggle, Available:[https://www.kaggle.com/code/georgyzubkov/heart-disease-exploratory-data-analysis/comments]

[2] Mohamed Elsayed, "Heart Disease Prediction", Kaggle, Available: [https://www.kaggle.com/code/andls555/heart-disease-prediction/notebook#10|-Comparison]