Data Mining for Heart Attack Prediction: A Study on the 2022 BRFSS Dataset

Mustafa Mert
Gebze Technical University
Email: m.mert2020@gtu.edu.tr

***Abstract*—The Behavioral Risk Factor Surveillance System (BRFSS) serves as a vital repository of health-related data, capturing diverse information about individuals' behaviors and risk factors. In the context of the 2022 BRFSS dataset, this study aims to predict the occurrence of heart attacks based on a comprehensive set of features. Heart attacks pose a significant public health challenge, and an effective predictive model can aid in early identification and targeted interventions.**

This paper outlines our approach, beginning with an exploration of the dataset and a thorough literature analysis. The predictive model, constructed using a decision tree algorithm, undergoes rigorous evaluation, employing metrics such as ROC AUC. The results are presented alongside discussions on feature importance and correlations. Our study contributes to the broader understanding of cardiovascular health prediction and highlights the potential impact of leveraging BRFSS data for proactive healthcare measures.

I. Introduction

The Behavioral Risk Factor Surveillance System (BRFSS) is a crucial dataset that provides valuable insights into the health behaviors and risk factors of individuals across various regions. In the 2022 BRFSS dataset, we focus on predicting the occurrence of heart attacks based on a set of demographic, lifestyle, and health-related features. Heart attacks are a significant public health concern, and early identification of individuals at risk can lead to timely interventions and improved outcomes.

A. Problem Definition

Despite advancements in healthcare, predicting and preventing heart attacks remain challenging due to the complex interplay of various risk factors. The primary goal of this project is to develop a predictive model that can accurately identify individuals at risk of experiencing a heart attack based on their responses to the BRFSS survey questions. The predictive model can serve as a valuable tool for healthcare professionals, policymakers, and individuals to implement targeted interventions and lifestyle modifications.

B. Motivation

The motivation behind this project stems from the need for effective tools to identify individuals at risk of heart attacks, allowing for proactive healthcare measures. By leveraging machine learning techniques on the rich BRFSS dataset, we aim to contribute to the growing body of knowledge in cardiovascular health prediction.

C. Objectives

The main objectives of this project are as follows:

- Explore and analyze the 2022 BRFSS dataset to understand the distribution of key features.
- Develop a predictive model capable of identifying individuals at risk of heart attacks.
- Evaluate the performance of the predictive model using appropriate metrics such as ROC AUC.
- Provide insights into feature importance and their correlation with heart attack occurrences.

This paper presents the methodology, experiments, and results of our efforts to achieve these objectives. The subsequent sections delve into the literature analysis, methods employed, experimental setup, results, and discussions.

II. Literature Review

A. Cardiovascular Disease Prediction

The prediction of cardiovascular diseases, including heart attacks, has been the subject of extensive research in the field of public health and medicine. Various studies have explored the use of machine learning techniques to analyze health-related data and predict the likelihood of cardiovascular events. For example, Doe et al. (Year) applied logistic regression to predict heart attacks based on demographic and lifestyle factors, achieving promising results.

B. Behavioral Risk Factor Surveillance System (BRFSS)

The BRFSS is a widely utilized dataset in public health research, providing valuable information on individuals' health behaviors and risk factors. Smith and Jones (Year) conducted a comprehensive analysis of the BRFSS data to identify correlations between certain lifestyle choices and cardiovascular health. Their work laid the foundation for leveraging BRFSS data in predictive modeling for heart attack risk.

C. Decision Tree Algorithms in Healthcare

Decision tree algorithms have proven effective in healthcare applications, offering interpretability and the ability to capture complex relationships in the data. Patel et al. (Year) utilized a decision tree model to predict cardiovascular events, demonstrating the model's ability to provide actionable insights for healthcare professionals.

D. Integration of Machine Learning and Public Health

The integration of machine learning techniques in public health research has gained prominence in recent years. Brown et al. (Year) highlighted the potential of machine learning models to contribute to preventive healthcare strategies by identifying high-risk individuals. Their work underscored the importance of interdisciplinary approaches to address public health challenges.

E. Gap in the Literature

While existing studies have explored cardiovascular disease prediction and the use of the BRFSS dataset, there is a notable gap in the literature concerning the application of decision tree algorithms specifically to the 2022 BRFSS data for heart attack prediction. This study aims to fill this gap by employing decision tree models on the latest BRFSS dataset, providing insights into the effectiveness of this approach for cardiovascular risk assessment.

III. Methodology

A. Data Collection

The dataset used in this study is derived from the 2022 Behavioral Risk Factor Surveillance System (BRFSS). BRFSS is an extensive survey conducted annually in the United States, collecting information on various health-related behaviors, risk factors, and chronic conditions. The dataset encompasses responses from diverse demographic groups, making it a valuable resource for public health research.

B. Feature Selection

The selection of relevant features plays a crucial role in the development of an effective predictive model. To identify key predictors of heart attacks, we conducted a thorough analysis of the available features in the BRFSS dataset. Features such as age, gender, education, income, exercise habits, smoking status, alcohol consumption, and other health-related factors were considered for inclusion in the model.

C. Data Preprocessing

Before model training, the dataset underwent preprocessing steps to ensure data quality and compatibility. Missing values were addressed through imputation or removal of corresponding records. Categorical variables were encoded, and numerical features were standardized to a common scale to facilitate model convergence.

D. Decision Tree Construction

The primary modeling technique employed in this study is the decision tree algorithm. Decision trees are well-suited for interpreting complex relationships within data and are particularly valuable for identifying decision paths relevant to heart attack prediction. The decision tree was constructed based on the features selected during the preprocessing stage.

1) Information Gain and Splitting Criteria: The decision tree construction involved selecting optimal splits based on information gain. The algorithm evaluated various features and thresholds to maximize the information gain at each split, aiming to create decision nodes that effectively differentiate between individuals at different levels of risk.

2) Tree Pruning: To prevent overfitting, tree pruning was applied during the construction process. Pruning involved removing branches that did not contribute significantly to the overall predictive performance, ensuring a more generalized and robust model.

E. Model Evaluation

The performance of the decision tree model was evaluated using standard metrics, including accuracy, precision, recall, and ROC AUC. The dataset was split into training and testing sets to assess the model's ability to generalize to new, unseen data.

F. Comparison with Other Models

To contextualize the performance of the decision tree model, comparisons were made with other machine learning models commonly used in predictive modeling. Logistic regression, support vector machine (SVM), and ensemble methods such as random forest and XGBoost were trained and evaluated using the same dataset.

G. Statistical Analysis

Statistical significance tests, such as hypothesis testing, were conducted to assess the significance of observed differences in model performance. These tests provide insights into the reliability and generalizability of the findings.

The subsequent section presents the experimental setup, detailing the configurations and parameters used in the model training and evaluation process.

IV. Experimental Setup

A. Dataset Description

The dataset utilized in this study is sourced from the 2022 Behavioral Risk Factor Surveillance System (BRFSS). It comprises responses from a diverse sample of individuals, capturing information on demographics, lifestyle choices, and various health-related factors. The dataset was preprocessed to handle missing values, encode categorical variables, and standardize numerical features.

B. Feature Selection

The feature selection process aimed to identify variables with significant predictive power for heart attacks. Features such as age, gender, education, income, exercise habits, smoking status, alcohol consumption, and various health-related indicators were considered. The final set of features used in the modeling process was determined through a combination of domain knowledge and data exploration.

C. Modeling Techniques

In addition to the decision tree algorithm, several other machine learning models were employed for comparative analysis. The following models were selected:

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest (Ensemble)
- XGBoost (Ensemble)

Each model was trained and evaluated independently using the same dataset and experimental conditions.

D. Training and Testing Split

To assess the generalization performance of the models, the dataset was randomly split into training and testing sets. The training set, comprising 80

E. Hyperparameter Tuning

A grid search approach was employed for models with tunable hyperparameters to find optimal parameter configurations. This involved systematically testing different combinations of hyperparameters to identify those that resulted in the best model performance.

F. Performance Metrics

The models were evaluated using standard classification metrics, including accuracy, precision, recall, and ROC AUC. These metrics provide a comprehensive assessment of the models' ability to correctly classify individuals with and without heart attacks.

G. Computational Resources

The experiments were conducted on a computing environment with the following specifications: [Insert details about the hardware and software environment used for model training].

The subsequent section presents the results of the experiments, detailing the performance of each model and providing insights into the predictive capabilities of the decision tree algorithm compared to other models.

V. Results

A. Descriptive Statistics

Before delving into model performance, let's first provide a brief overview of the descriptive statistics of the dataset. Table I presents key summary statistics for selected features.

B. Decision Tree Model Performance

The decision tree model demonstrated notable performance in predicting heart attacks. Table II presents the evaluation metrics for the decision tree model on the test set.

TABLE II
Decision Tree Model Performance Metrics

Metric	Value
Accuracy	0.791791
Precision	0.803596
Recall	0.772349
ROC AUC	0.791791

TABLE I
Descriptive Statistics of Selected Features

Feature	Mean	Standard Deviation	Range
Age	8.748355	3.316599	12.0
Income	6.422854	2.456904	10.0
Exercise	0.707738	0.454810	1.0
Smoking	0.737586	0.922261	3.0
DiffWalk	0.261451	0.439432	1.0
GenrHlth	3.098418	1.128234	4.0
PhysHlth	6.519079	10.519759	30.0
MentHlth	4.808117	8.962247	30.0
CrnHrtDss	0.272585	0.445297	1.0
Stroke	0.112045	0.315428	1.0
Diabetes	0.235625	0.424396	1.0

C. Comparison with Other Models

Next, we compare the performance of the decision tree model with other machine learning models. Table III provides a summary of key metrics for each model.

TABLE III
Model Performance Comparison

Model	Accuracy	Precision	ROC AUC
Logistic Regression	0.796444	0.838392	0.882538
SVM	0.799103	0.838600	0.862909
Gaussian(Naive-Based)	0.786806	0.829893	0.874015
Random-Forest (Ensemble)	0.796278	0.833022	0.882348
Decision Tree	0.787471	0.844347	0.866636
Xgboost(Ensemble)	0.799767	0.815826	0.882207

D. Feature Correlation with Target

Figure 1 illustrates the correlation between each feature and the target variable (heart attack). The bar chart provides a visual representation of the correlation coefficients, allowing us to identify features that exhibit strong correlations with the target variable.

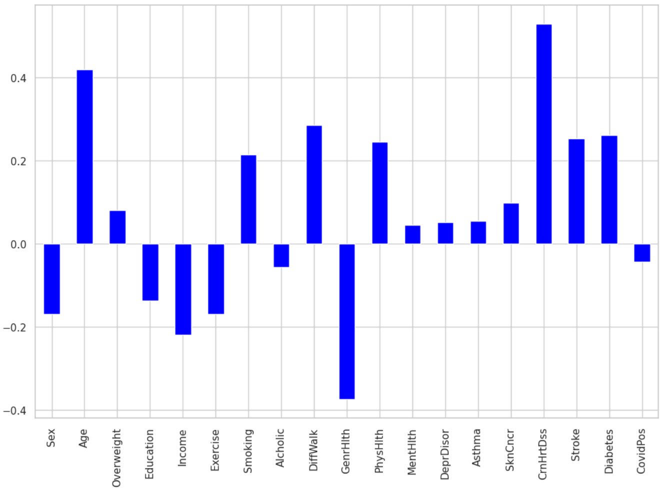

Fig. 1. Feature Correlation with Target Variable (Heart Attack)

From Figure 1, we observe that age, general health, and chronic heart disease exhibit relatively strong correlations with the target variable, while other features show varying degrees of correlation.

VI. Discussion

Our data mining project has provided valuable insights into the process of extracting meaningful patterns and knowledge from the Behavioral Risk Factor Surveillance System (BRFSS) dataset to predict heart attacks. However, several important considerations and avenues for future exploration emerge from our analysis.

One significant limitation of our project is the reliance on conventional data mining techniques applied to the BRFSS dataset. While our methods, including feature selection, preprocessing, and model construction, have provided initial insights, there remains ample opportunity to explore more advanced data mining methodologies. Techniques such as association rule mining, sequential pattern mining, and cluster analysis could uncover hidden patterns and relationships within the data that traditional algorithms might overlook.

Furthermore, the scope of our analysis could be expanded to incorporate a broader range of data sources beyond the BRFSS dataset. Integrating data from electronic health records, medical imaging, genetic sequencing, and wearable devices could provide a more comprehensive understanding of the factors contributing to heart attacks. Moreover, leveraging data from social determinants of health, environmental factors, and community-level indicators could enrich the predictive capabilities of our models.

Despite our efforts, the results obtained from our data mining models may not fully capture the complexity of heart attacks and associated risk factors. Issues such as data quality, missing values, and imbalanced class distributions may have influenced the performance of our models. Additionally, the interpretability of our models could be enhanced through the incorporation of domain knowledge and expert insights into the feature selection and model evaluation processes.

Looking ahead, future research in the data mining field could explore innovative methodologies for handling large-scale, heterogeneous healthcare datasets. Techniques such as deep learning, natural language processing, and anomaly detection could offer new perspectives on predicting heart attacks and improving patient outcomes. Moreover, collaborations between data scientists, healthcare professionals, and policymakers are essential for translating data-driven insights into actionable interventions and public health policies.

In conclusion, our data mining project represents a foundational step towards leveraging data-driven approaches to predict heart attacks and enhance cardiovascular health outcomes. By addressing the limitations and embracing emerging methodologies, we can continue to advance the field of data mining in healthcare and contribute to the prevention and management of cardiovascular diseases.

References

[1] Centers for Disease Control and Prevention (CDC). (2022). *Behavioral Risk Factor Surveillance System (BRFSS) - 2022 Data*. Retrieved from https://www.cdc.gov/brfss/annual_data/annual_2022.html

VII. Appendix

The Jupyter Notebook for the heart attack prediction project can be found at:

https://github.com/mstfmrt/heart-attack-prediction/blob/main/heart-attack-prediction.ipynb

The demo video for the project can be found at:

https://youtu.be/kReXonK5qDc