DATA ANALYSIS OF HEART ATTACK RISK FACTORS: INSIGHTS FROM MACHINE LEARNING

Conference Paper · February 2024

DOI: 10.5281/zenodo.10651585

1 author:

Tolganay Muntinova
Campbellsville University
4 PUBLICATIONS **0** CITATIONS

SEE PROFILE

DATA ANALYSIS OF HEART ATTACK RISK FACTORS: INSIGHTS FROM MACHINE LEARNING

Tolganay Muntinova
Master's student at Campbellsville University
Data Science and Artificial Intelligence
ORCID: 0009-0007-8418-8205

Abstract

This study presents an exploratory data analysis (EDA) concentrated on recognizing risk factors connected with heart attacks by leveraging machine learning techniques. Given the complexity and multifactorial nature of heart disease, identifying and recognizing the most significant predictors are actually for early intervention and prevention strategies. Through feature importance ranking, correlation analysis, and predictive modeling, we highlight key factors contributing to increased heart attack risk. These findings underscore the importance of combining EDA with machine learning to enhance our understanding of heart disease dynamics, offering potential pathways for more personalized and effective preventive healthcare measures. The study targets not only to support the academic discourse on heart disease prediction but also to serve as a foundation for developing more sophisticated predictive models and health policy planning.

Keywords: *Exploratory Data Analysis, Machine Learning, Heart Attack Risk Factors, Predictive Modeling, Feature Importance, Health Data Analytics, R.*

Introduction

This study explores the intricate connection between various health indicators and their potential contribution to heart disease. It utilizes a dataset enhanced along with both categorical and numerical data. This dataset involves a wide scale of factors, including demographic details like age and sex, clinical parameters such as chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, and resting electrocardiographic results, among others. Each of these features provides a unique perspective on an individual's health profile page, providing essential insights into the multifaceted attributes of heart disease advancement. Incorporating an output variable showing the medical diagnosis of cardiovascular disease enhances the dataset's utility, paving the way for the development of predictive models targeted at very early detection and prevention.

Cardiovascular diseases, particularly cardiovascular disease, stand as the leading cause of mortality worldwide, providing a substantial public health challenge. The potential to accurately predict and alleviate the risk of heart attacks with the identification of crucial risk factors is of significant relevance in decreasing these statistics. Traditional strategies for recognizing cardiovascular disease have primarily depended on statistical analyses of risk factors such as age, sex, cholesterol levels, high blood pressure, diabetes mellitus, and also smoking practices.

This research specifies the dataset preprocessing steps to prepare the data for analysis and correlation analysis.

Exploratory Data Analysis (EDA)

This paper kicks off with an in-depth exploratory data analysis (EDA), setting the stage for understanding the dataset's structure and characteristics, which is vital for subsequent analytical phases. After meticulous data cleaning, the study examines a rich array of health indicators ranging from demographics like age and sex to specific medical measures such as chest pain type, blood pressure, cholesterol, glucose levels, and ECG results. The study further involves blood pressure, cholesterol, and glucose levels, recognized as conventional markers of heart health, and uses electrocardiographic outcomes and maximum heart rate during stress tests as indicators of potential cardiac issues. The research employs `dplyr` and `ggplot2` for data manipulation and visualization, presuming access to an existing dataframe or the necessity to import data for analysis.

Uniqueness: *The script loops through each column of the dataframe to count and print the number of distinct values using the n_distinct() function. This step is crucial to understand the cardinality of the data, which helps in identifying categorical variables and potential key identifiers.*

```
> cat("Distinct value counts for each column:")
> for (column in colnames(df)) {
  distinct_values <- n_distinct(df[[column]])
  cat(column, ": ", distinct_values, "distinct values")
 }
```

Handle Duplicates: *Duplicates are identified using the duplicated() function and then removed from the dataframe with df <- df[!duplicated(df),]. This ensures the data does not have redundant rows that could skew analysis results.*

```
> cat("Number of duplicate rows: ",sum(duplicated(df)),"\n")
> df <- df[!duplicated(df), ]
```

Completeness: *This part of the script calculates the percentage of missing values for each column using sapply() and reports the columns with missing data. Handling missing data is essential as it can impact the accuracy of statistical analyses and machine learning models.*

```
> missing_data <- sapply(df,function(x) sum(is.na(x))/length(x)*100)
> missing_data <- missing_data[missing_data > 0]
> cat("Missing Data Ratio:")
> print(missing_data)
```

Replace Values: *In the 'thall' column, zeroes are replaced with twos. This could be a data correction step where the value '0' is assumed to be a placeholder or an incorrect entry for 'normal' which is represented by '2' in the dataset.*

```
> df$thall <- ifelse(df$thall == 0, 2, df$thall)
```

Consistency: *The next block of code maps numeric codes to descriptive labels for several columns (cp, slp, thall, restecg, sex). This step improves data readability and may be required for certain types of analysis that expect categorical data in non-numeric form.*

```
> cp_mapping <- c('typical angina', 'atypical angina', 'non-anginal pain', 'asymptomatic')
> df$cp <- factor(df$cp, labels = cp_mapping)
> slp_mapping <- c('unsloping', 'flat', 'downsloping')
> df$slp <- factor(df$slp, labels = slp_mapping)
> thall_mapping <- c('fixed defect', 'normal', 'reversible defect')
> df$thall <- factor(df$thall, labels = thall_mapping)
> rest_ecg_mapping <- c('normal', 'ST-T wave abnormality',' left ventricular hypertrophy')
> df$restecg <- factor(df$restecg, labels = rest_ecg_mapping)
> sex_mapping <- c('female', 'male')
> df$sex <- factor(df$sex, labels = sex_mapping)
```

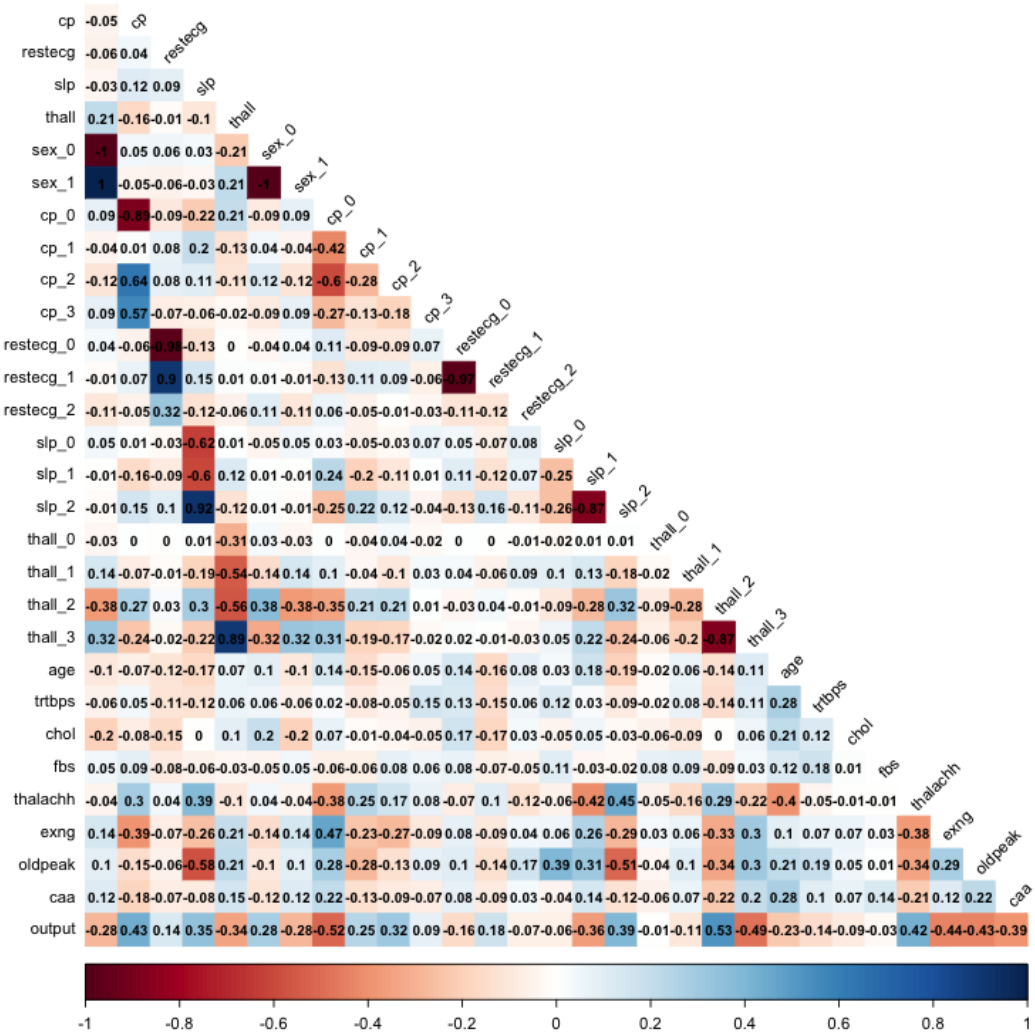
Correlation Analysis. *Correlation analysis was vital in this particular study, disclosing possible relationships amongst mathematical variables to deepen the understanding of cardiovascular disease risk factors. The dataset utilized is actually a detailed collection of medical data featuring various attributes connected to cardiovascular health. This analysis assisted in evaluating the strength as well as the direction of these relationships, where positive correlations show a direct relationship and negative correlations suggest an inverted one. These searchings are important for building predictive models but require more exploration to affirm causality. Machine learning takes advantage of these variables to forecast heart attack risk, striving to aid healthcare professionals in early identification as well as intervention. The comprehensive nature of the dataset, incorporated with complex analysis methods, might substantially improve public health results by informing better prevention and treatment of heart disease.*

```
> # Load necessary library
> library(corrplot)
>
> # Define the data and column types
> data <- df
```

```
> categorical_columns <- c('sex', 'cp', 'restecg', 'slp', 'thall')
> numerical_columns <- c('age', 'trtbps', 'chol', 'fbs', 'thalachh', 'exng', 'oldpeak', 'caa', 'output')
>
> # Define the function to plot a correlation heatmap
> plot_correlation_heatmap <- function(data, target_variable, title) {
  # Calculate the correlation matrix with complete observations
  cor_matrix <- cor(data, use = "pairwise.complete.obs")
  # Order the variables, putting the target variable at the end
  ordered_vars <- c(setdiff(names(data), target_variable), target_variable)
  cor_matrix <- cor_matrix[ordered_vars, ordered_vars]
  # Plot the correlation heatmap with specified parameters
  corrplot(cor_matrix, method = "color", title = title,
       type = "lower", order = "hclust",
       addCoef.col = "black", tl.col = "black", tl.srt = 45,
       diag = FALSE)
}
> # Call the function with your data frame
> plot_correlation_heatmap(df, 'output', 'Data Correlation Heatmap')
```

Key Findings and Implications. *The correlation heatmap gives a visual representation of the strength as well as the direction of relationships between numerous health indications as well as the possibility of heart disease. Notable searches coming from the analysis include the data emphasizing that while some individual factors are actually significantly correlated with cardiovascular disease, the ailment's multifactorial nature requires a much more holistic analysis. Our findings underscore the significance of certain health indicators in predicting heart disease. Variables like age, sex, chest aches, and serum cholesterol amounts emerged as crucial things in this context. The predictive models created via this study demonstrated promising precision, underscoring the possibility for very early discovery and treatment. This comprehensive analysis not only adds to the growing body of knowledge in cardiovascular disease research but also highlights the importance of leveraging data analytics in healthcare. The offered correlation heatmap is actually a highly effective tool for uncovering the correlation between numerous health and wellness signs and also the likelihood of heart problems. By taking a look at the correlation coefficients, it's easy to know just how each factor may help in the risk of heart problems, which is essential for developing helpful analysis as well as preventative solutions.*

According to Picture 1. age has a moderate positive correlation with **cholesterol (0.21)** *and* **resting high blood pressure (0.28).** *This is actually suggestive of a common physiological trend where developing age is connected with a boost in these cardio risk factors. As individuals age, their vascular system may go through improvements, including increased stiffness and reduced plaque collection, which is actually reflected in higher cholesterol and high blood pressure.*



Picture 1. R code results

Chest pain, *stood for listed here as 'cp', presents a strong positive correlation along with the* ***cardiovascular disease end result (0.43)****, suggesting its perspective as a significant indication of heart problems. Non-anginal chest pain is identified to have a substantial positive correlation with the presence of heart disease, indicated by a correlation coefficient of 0.316. Despite often being confused with less serious conditions such as indigestion or muscular pain, its significant association with cardiac issues suggests the need for careful evaluation to avoid delays in diagnosis and treatment. On the other hand, typical angina, reflected by a strong negative correlation coefficient of -0.516, implies that those experiencing classic heart diseases symptoms, like consistent chest pain or discomfort, may be more prompt in seeking medical care. Such timely intervention could be instrumental in curtailing the advancement of more severe heart disease.*

Resting blood pressure(trtbps) *and* ***cholesterol(chol)****, even being traditional risk factors for cardiovascular disease, do not show sturdy connections with the heart disease results within this dataset. This could be because of an assortment of main reasons such as the existence of well-managed cases along with medicine, or even it could possibly recommended that factors might play a more decisive role in the progression of heart disease within this detailed population.*

Not eating ***Fasting blood sugar****(fbs) additionally reveals a weak correlation with heart disease indications, which could suggest that while diabetic issues are actually a risk factor for cardiovascular disease, the fasting blood sugar level amount on its own may not be a strong standalone predictor of heart disease within this dataset.*

Resting electrocardiographic results*(restecg) and also* ***thallium stress tests****(thall) illustrate mild correlations along with some functions. While these tests are actually essential for diagnosing heart disease, their part as forecasters could be confined as well as potentially eclipsed by various other much more leading factors. For instance, a reversible defect detected during a thallium stress test, typically indicative of ischemia, correlates negatively with the incidence of heart disease, as indicated by a*

correlation coefficient of -0.486. This could imply that patients with reversible defects are effectively treated, thus diminishing their risk of advancing to serious heart conditions. Conversely, a normal outcome on a thallium stress test (thall_normal) presents a correlation coefficient of 0.526, associated with a higher probability of heart disease. Although thallium stress tests are designed to detect areas of diminished blood flow to the heart, potentially signaling coronary artery disease, this positive correlation might indicate that patients with "normal" results could still have a form of the disease that is in an incipient stage and may be progressing.

The maximum heart rate *achieved(thalachh) possesses a mild positive correlation with heart disease (0.42) as well as a moderate negative correlation with* ***age*** *(-0.39). This recommends that a lower optimum heart fee is related to much older age, which people who accomplish much higher heart fees in the course of workout could be at a higher risk of heart disease. This can easily show the heart's capacity to reply to stress, and also an impaired reaction may be a sign of heart problems.*

Exercise-induced angina(*exng****)*** *has a moderate negative correlation with* ***chest pain*** *(-0.44) and also a moderate positive correlation with* ***heart disease****(0.38). The damaging correlation with chest pain might be because of the truth that exercise-induced angina is actually a detailed form of chest pain set off by physical activity and also might be identified in a different way in medical assessments.*

The strong negative correlation between ***Oldpeak*** *(ST anxiety) and* ***The Slope of The Peak Exercise ST Segment*** *(slp) (-0.58) advises that these pair of ECG findings are actually carefully related. Both of these indicators are actually utilized to determine myocardial ischemia, and their correlation might also suggest that as one becomes a lot more obvious, the various others usually tend to observe, highlighting their integrated importance in evaluating heart disease risk. Furthermore, a positive correlation coefficient of 0.39 indicates a link between the presence of a downsloping peak exercise ST segment and an increased likelihood of heart disease. This pattern on an ECG may signal ischemia, pointing to the potential for compromised cardiac function.*

The number of significant ships colored by ***fluoroscopy*** *(caa, -0.39) is associated with a lower risk of heart disease. This could reflect good blood flow to the heart, reducing the likelihood of significant heart disease.*

Conclusion

The exploration and analysis of the heart disease dataset underscore the critical role of data-driven insights in advancing medical research. Through rigorous EDA, correlation analysis, cluster analysis, and the development of predictive models, this study offers valuable insights into the factors contributing to heart disease. These findings have the potential to inform clinical practices, guide further research, and facilitate the development of effective prevention and intervention strategies, contributing to the fight against heart disease. This study has demonstrated the power of correlation analysis and machine learning in understanding and predicting heart disease risk. The findings emphasize the multifactorial nature of heart disease and the importance of a holistic approach to its prediction and management. By continuing to leverage data analytics and machine learning, healthcare professionals can develop more effective early detection and intervention strategies, ultimately leading to better prevention and treatment of heart disease. This research not only contributes to the academic discourse on heart disease prediction but also highlights the critical role of data analytics in advancing healthcare outcomes.

References

1. Heart attack Analysis & Prediction Dataset, Kaggle, 2021. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/data.

SC
scientific conferences

VI international scientific conference
Toronto. Canada
06-07.02.2024

Scientific conferences

QUESTIONS. HYPOTHESES. ANSWERS: SCIENCE XXI CENTURY

Proceedings of the international Scientific and Practical Conference

06-07 February 2024

TORONTO. CANADA

2024

UDC 001.1

BBC 1

VI International Scientific and Practical Conference «Questions. hypotheses. answers: science XXI century», February 06-07, 2024, Toronto. Canada. 107 p.

ISBN 978-91-65423-54-1

DOI https://doi.org/10.5281/zenodo.10651585

Publisher: «SC. Scientific conferences»

Main organization: artmedia24

Editor: Hans Muller

Layout: Ellen Schwimmer

The sample of the citation for publication is Gugnin Aleksandr, Lisnievska Yuliia ANTI-ADVERTISING IN THE HOTEL BUSINESS // VI International Scientific and Practical Conference «Questions. hypotheses. answers: science XXI century », February 06-07, 2024, Toronto. Canada. Pp.9-11, URL: https://sconferences.com

Contact information

Website: https://sconferences.com

E-mail: info@sconferences.com

Philological sciences

Psychological sciences

Sociological sciences

Technical sciences