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 Kriz
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.

Interesting questions we sought to answer:
•Which states have the highest rate of heart disease?
• Does physical activity affect heart disease?
• Which features correlate with heart disease?
• If someone has heart disease, what is the largest & most common set of symptoms/features?
• Which demographics have heart disease?

What is a brief summary of your results?
During our data exploration phase, we utilized a matrix heat map to discern potential correlations among attributes. Initially, the group endeavored to apply a k-means model. However, this approach presented issues that demanded attention. Specifically, when I employed K-Means with the features HadAngina, HadStroke, and AgeCategory based on the correlation matrix findings, it became evident from the resulting visual representation that utilizing binary/categorical features for K-Means without dimension reduction led to a plot consisting of predominantly straight lines. Furthermore, we conducted a state-level analysis to investigate whether geographic location influences the prevalence of heart disease. Our findings reveal that Arkansas, Florida, Maine, and West Virginia have the highest percentage of their populations affected by heart disease, each comprising over 7% of the state's population.

Subsequently, we introduced a Bayesian network model to dissect attributes exhibiting chained correlations. This analysis clarified the primary interconnected attributes and their significance. By using different heat maps and plots to study the Conditional Probability Distributions (CPD), we obtained valuable insights. These insights helped us answer our questions and confirm that the most important factors influencing Heart Disease are 'HadAngina', 'DifficultyWalking', and 'AgeCategory', especially for individuals aged 50 and above.

INTRODUCTION

The primary aim of this data mining project is to explore several key questions pertaining to Heart Disease. These questions are of paramount importance due to the significant impact of Heart Disease on public health. Heart Disease remains one of the leading causes of morbidity and mortality globally, imposing substantial economic burdens on healthcare systems and society as a whole. Understanding the geographical distribution of Heart Disease prevalence, the influence of physical activity, correlated features, prevalent symptoms, and demographic predispositions can inform targeted interventions, resource allocation, and preventive measures. Therefore, investigating these questions is crucial for develop effective

strategies to mitigate the burden of Heart Disease and improve overall public health outcomes.

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

RELATED WORK

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.

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

DATA SET

- Indicators of Heart Disease Data Set
400K+ Records (Some include NaN values)
Heath interview answers from 400K+ individuals
- Mostly Yes/No survey responses
- "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.

MAIN TECHNIQUES APPLIED

○ Data clean/preprocess/etc.
○ Data Warehouse/cube/etc.
○ Classification/Clustering/etc.

- Correlation matrix was applied to initially discover which features were correlated with one another.
- K-Mean and PCA/MCA were applied to discover any feature clustering. However, during this process it was discovered that K-Means and PCA don’t work well for binary/categorical data.
- Bayesian Networks Analysis was employed to investigate correlations and dependencies among attributes associated with heart disease, offering insights into causal relationships and predictive modeling. This approach helped identify key attributes directly and indirectly linked to heart disease, facilitating the visualization of their correlations through graphical representation.

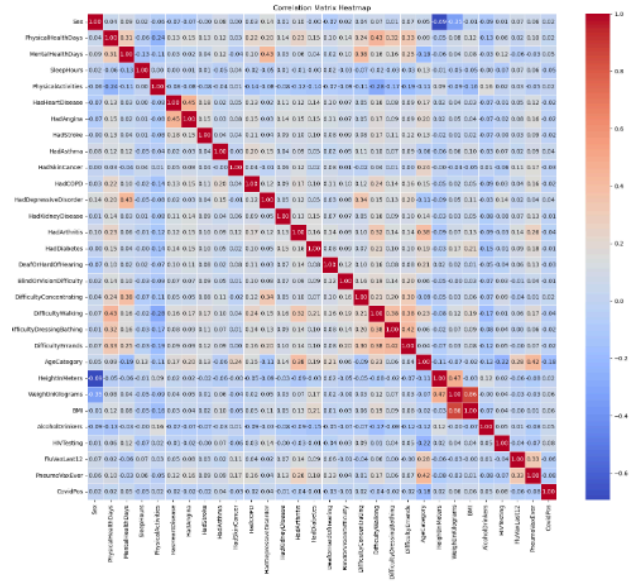
KEY RESULTS

- The rate of Heart Disease varies according to the geographical State.
- K-means is inconclusive for chained attributes.
- Heart Disease is strongly correlated with Angina and indirectly correlated with Difficulty Walking and Age Category.

Full plots can be viewed within the Jupyter notebook.

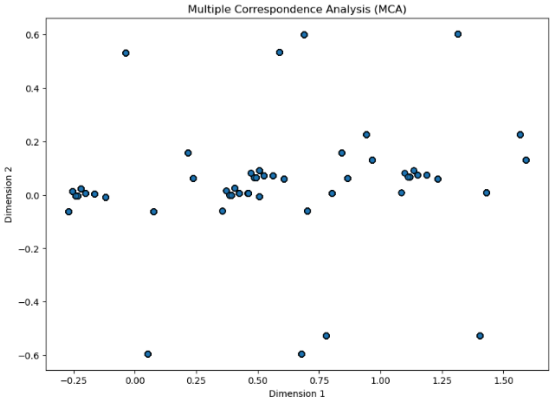
Correlation:

- Plotting the correlation matrix between the features showed which features are correlated and which features are not correlated. With a correlation cutoff of 0.15, HasHeartDisease has the greatest correlation to HadAngina, HadStroke, DifficultyWalking and AgeCategory. Angina is chest pain cause by reduced blood flow to the heart.



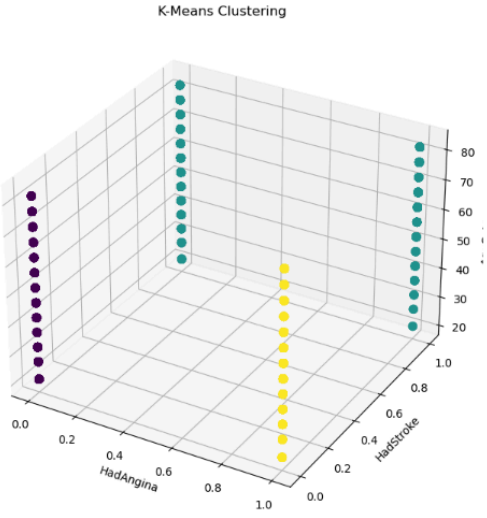
PCA/MCA:

- After attempting to apply PCA, we learned that PCA cannot be used on categorical/binary data. We found another method of dimension reduction for categorical/binary data – multiple correspondence analysis (MCA). MCA allows for categorical analysis across multiple features/categories.
- There isn't much guidance on how to implement MCA using Python. There are Python libraries for MCA – MCA and Prince, but neither have robust documentation and examples to aid in applying the methods. We applied PCA and created a scatter plot from the results, but the plot didn't show enough detail to allow for full analysis.
- The AgeCategory data had to be one-hot encoded to apply MCA.



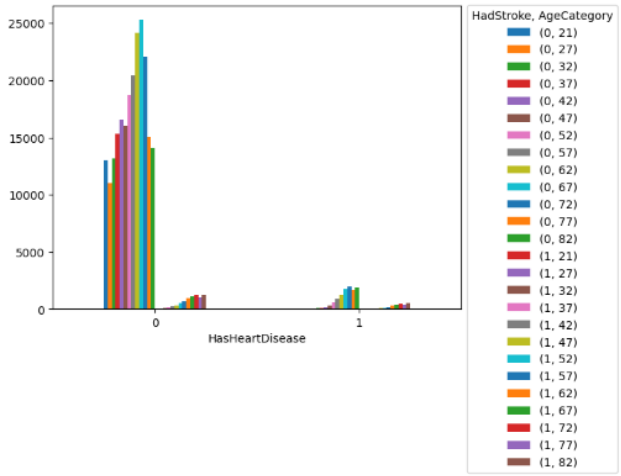
K-Means:

- The initial plan was to apply K-Means to the data to see which 'groups' of feature data were similar to each other. The groups would then show us the profiles of people based on the features reported that were most similar to each other. However, K-Means is not an appropriate model for categorical/binary data. Applying K-Means to the data set generated straight lines of 1/0 data and it was not helpful.



Contingency:

- Although PCA and K-Means were not helpful forms of analysis, visualizing the contingency table from MCA was helpful. Using the contingency table, we were able to visualize the relationship between HasHeartDisease, HadStroke and Age Category.
- Based on the plot visual, if someone hasn't had a stroke, they aren't likely to have heart disease. In addition, if they do have heart disease, they still aren't greatly likely to have had a stroke. However, AgeCategory does appear to have an effect on having heart disease and having a stroke.



Country Wide Results

Findings:
- Heart Disease rates are not necessarily dependent on geographic location in the country. The states with the highest and lowest rates of Heart Disease are spread out.

State	
Virgin Islands	3.36
District of Columbia	3.42
Colorado	4.07
New Jersey	4.08
Illinois	4.14
Arkansas	8.30
West Virginia	8.27
Florida	7.61
Maine	7.07
Nebraska	6.74

Something to take into consideration about this dataset is that each state does not have the same number of test subjects. This is purely dependent on the proportion of positive occurrences of heart disease and the amount of test subjects per state.

For example, Florida appears on the list for having some of the highest proportion of positive results at 557 out of 7315 test subjects. While the state of Washington has more positive detections at 701 but 15000 test subjects resulting in a low 4.67% detection of heart disease.

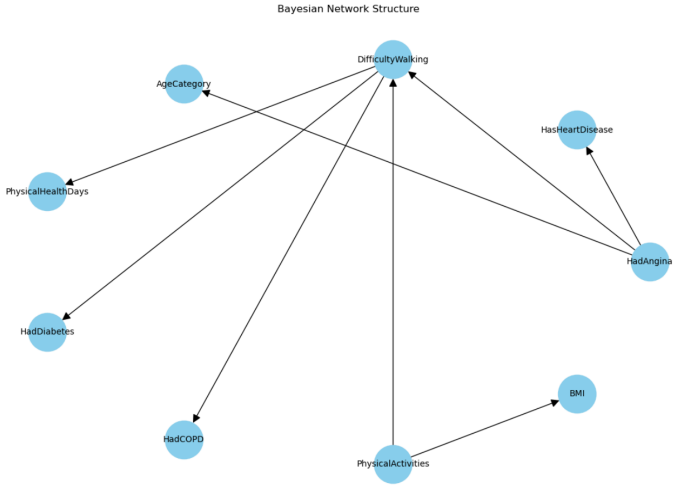
Florida	557	7315	7.61
Washington	701	15000	4.67

After calculating the country wide results, we could do more data cleaning, removing data items like the Virgin Islands and District of Columbia since they represent a miniscule portion of the total population. This would change the states with the lowest proportion of heart disease.

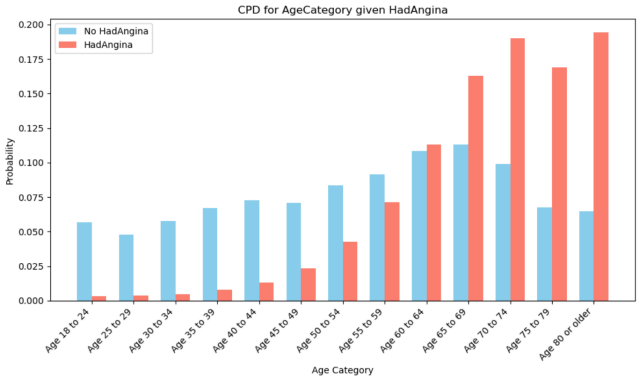
Bayesian Networks

- We trained the CPD values using Bayesian Networks Model to analyze the complex relationship of the data.

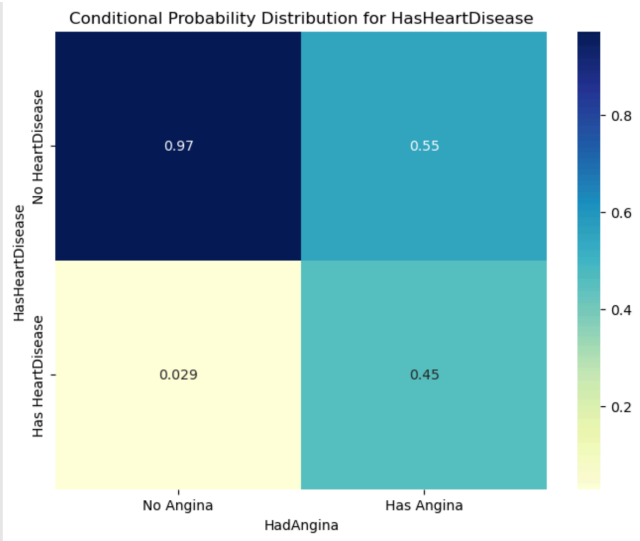
Directly Acyclic Graph showing the relationship of Heart Disease and Has Angina. The relationship of Has Angina with Difficulty Walking and Age Category.



After analyzing the data was possible to observe that Age Category 50+ years old had a higher probability to develop of Angina (Graph Bellow).

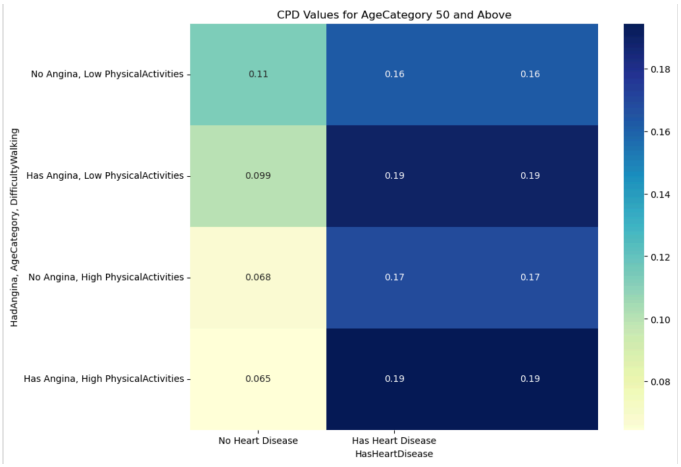


Now the model was trained using Age Category 50+ years old and using the Conditional Probability Values for HasHeartDisease attribute and HasAngina.

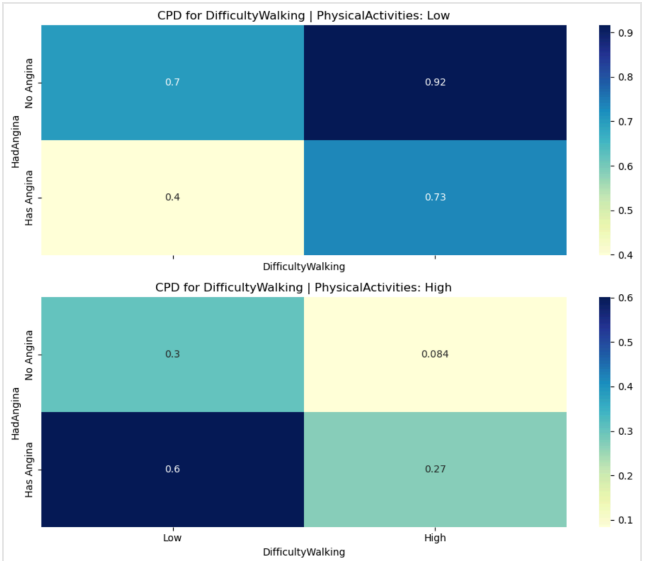


- HasAngina influences considerably HasHeartDisease by 0.45 of probability. Showing a strong correlation of the attributes.

The Heat Map bellow shows that Physical Activities are not directly correlated with HasHeartDisease. It is possible to observe, in dark blue, that HeartDisease rate remains the same for Low Physical Activity (0.19) and High Physical Activity (0.19).



However, the Heat Map bellow shows that Physical Activities attribute influences directly HasAngina, which is the main attribute for HasHeartDisease. Therefore, Physical Activities indirectly influences HasHeartDisease.



Above is possible to observe that there is a considerably difference for the rate of High DifficultyWalking and HasAngina given Low PhysicalActivities (bottom-right value in blue = 0.73) and High DifficultyWalking and HasAngina given High PhysicalActivities (bottom-right value in green = 0.27).

APPLICATIONS

- This new knowledge can easily be applied in our daily jobs:
- Allocating more time/resources to EDA before choosing a project direction
- Presenting EDA findings to leadership before moving forward

- This data mining project seeks to address critical questions regarding Heart Disease, given its profound impact on public health. By understanding the geographical distribution of Heart Disease prevalence, the influence of physical activity, correlated features, prevalent symptoms, and demographic predispositions, targeted interventions and preventive measures can be developed to mitigate its burden and enhance public health outcomes.

Tools

- KAGGLE
- GITHUB
- JUPYTERLAB
- EXCEL
- DISCORD
- PYTHON
- PANDAS
- SKLEARN
- MATPLOTLIB
- SNS
- KMEANS
- MCA
- PRINCE
- CONFUSION MATRIX
- PGMPY
- BAYESIANNETWORK
- MAXIMUMLIKELIHOODESTIMATOR

ACKNOWLEDGMENTS

We extend our gratitude to Professor Peterson and our team members, Cyro de Lima, Eric Hernandez, and Lindsay Kriz, 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]