# Analysis of Heart Disease-Related Dataset

Xinhui Luo, Erick Arenas, Weilin Cheng

**Abstract - The first part project explores and explains the relationships between Heart Disease and multiple other factors that might cause heart disease. The second part is to find the relationships between Heart Disease + Stroke and other factors. By implementing contingency table-related methods, conditional entropy, and machine learning models, we aim to identify significant risk factors and interaction effects based on these factors.**

## I. INTRODUCTION

Heart disease is a pressing public health concern in the United States, devastatingly impacting the population. In 2020, a staggering 697,000 lives were claimed by this condition, representing an alarming 20% of all deaths, which places an overwhelming burden on the healthcare system.[1] Our project revolves around exploratory data analysis of the CDC survey data from BRFSS 2015. By carefully examining this rich dataset, we aim to uncover significant factors and patterns contributing to heart disease. Through this endeavor, we hope to generate valuable insights that can inform effective preventive measures and alleviate the burden of heart disease on society.

*Variable Explanation*

Table 1.
Variable explanation

| | |
|---|---|
| “HeartDiseaseorAttack” | 1 = if they have been diagnosed with Coronary Heart Disease or myocardial infarction, or 0 = if not diagnosed |
| “HighBP” | 1 = if they have High Blood pressure, or 0 = if they not |
| “HighChol” | 1 = if they have been diagnosed with High Cholesterol, or 0 = if |

| | |
|---|---|
| “HvyAlcoholConsump” | 1 = if they have 7-14 drinks per week, 0 = if they have less than 7-14 drinks per week |
| “AnyHealthCare” | 1= if they have healthcare, 0 = if they do not have healthcare |
| “NoDocbcCost” | 1= if they do not go to they doctor because is too expensive, 0 = if they go to the doctor |

| | |
|---|---|
| | they have not |
| “CholCheck” | 1 = if they have checked their cholesterol in the last 5 years, or 0 = if not. |
| “BMI” | Body Mass Index range from 12.0 - 98.0 |
| “Smoker” | 1 = if they have smoked more than 100 cigarettes, or 0 - if they have not smoked or smoke less than 100. |
| “Stroke” | 1 = if they have had a stroke, or 0 = if they have not. |
| “Diabetes” | 2 = if they have diabetes, 1= if they are pre-diabetic, 0 = if they have not diabetes |
| “PhysActivity” | 1= if physical activity in the last 30 days, or 0 = if the not physical activity at all |
| “Fruits” | 1 = if one piece of fruit per day, or 0 = if they do not consume fruit at all |
| “Veggies” | 1 = if they have consumed a piece of veggie a day, 0 = if they have not |

| | |
|---|---|
| | regardless of cost |
| “GenHlth” | General health range 1-5 where 1 is the best general health and 5 is the worst |
| “MentHlth” | Mental health range from 0 - 30 where 0 = if they have to feel good mentally for the past 30 days, and 30 = if they have not |
| “PhysHlth” | Physical Health range from 0 - 30, where 0 = if they have had good physical health in the last 30 days and 30 = if they have not. |
| “DiffWalk” | 1 = if they have had difficulty walking, 0 = if they have not |
| “Sex” | 1 = if gender male or 0 = if gender female |
| “Age” | Age range from 1 - 13. Where each category is an age range. |
| “Education” | Education level range 1 - 6. Where 1 = not school at all, and 6 = 4 years college or more |
| “Income” | Income level range 1 -8. Where 1 = 10,000 $ a year or less, or 8 = 75,000$ a year or more. |

## II. METHODS

*Conditional entropy*

Conditional entropy measures the uncertainty of a dependent variable, given the knowledge of other independent variables. It quantifies the average amount of information needed to describe one variable when the value of another

variable is known. A lower conditional entropy indicates a favorable scenario, reducing uncertainty between the dependent and chosen independent variables. By utilizing conditional entropy, we can gain insights into the relationship and dependence between variables, where lower entropy demonstrates higher predictability.[2]

$$H(Y|X) = \sum_i p(X = x_i) H(Y|X = x_i) = -\sum_i p(X = x_i, Y = y_i) log_2 \frac{p(X=x_i, Y=y_i)}{p(X=x_i)} \quad (1)$$

*Fused Variables*

We wanted to create new variables to analyze numerically with entropy and contingency table. We decided to fuse every variable except for Heart disease and stroke with each other, which gives us 210 combinations of columns.

*Contingency table-related methods*

*1. Odds plots*

Odds refer to the likelihood or probability of an event occurring compared to the likelihood of the event not occurring. It is a way to express the relationship between the chances of success and failure. Mathematically, it is expressed as: Odds = P(event) / P(not event). We can compare two groups within a given condition by knowing the odds.

2. *Contingency table heatmaps*

Since conditional entropy cannot present a relationship between all categories, odds usually only reveal the relationship between two groups. We use contingency tables to help us identify hidden associations and patterns between variables. Heatmaps provide a graphical representation of the tables and rank the distribution with distinctive colors to enhance the legibility of contingency tables.

*3. Hierarchical classification heatmaps*

We choose to use Hierarchical classification heatmaps due to the complexity of contingency tables based on two-way combinations. Hierarchical classification heatmaps based on contingency tables are visual representations that combine hierarchical clustering and classification using a contingency table. They display the relationships between categorical variables and the hierarchical structure of categories. The heatmap shows the frequency or probability of instances belonging to different categories, with colors indicating the strength of association. These heatmaps provide a concise and comprehensive overview of complex classification results in a hierarchical context.

*Decision tree modeling*

The decision tree is a model that creates all the possible outcomes of our dependent variable given a certain amount of random variables and creates a tree with nodes and leaves depending on the number of sets manually or the number of possible combinations. The model predicts based on the number of features available for each possible combination.

## III. ANALYSIS

*Part 1: Heart disease against one variable*

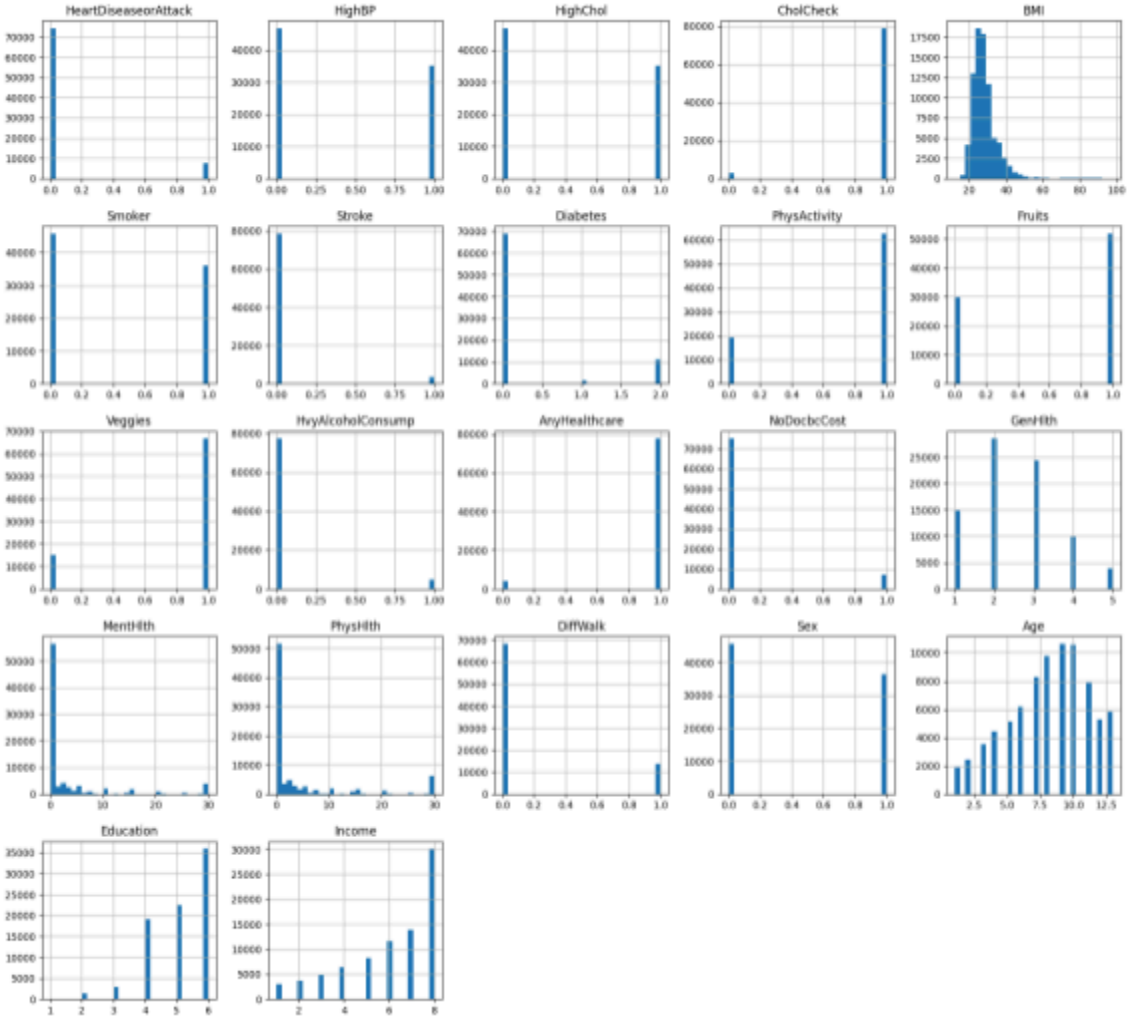



Figure 1. Histogram for each variable

Firstly, we plotted the distributions for all variables and observed that most were imbalanced. Specifically, in our response variable, “HeartDiseaseorAttack,” the proportion of 0 and 1 is approximately 10 to 1. Similar imbalances

are also present in variables such as Chol Check, Stroke, Diabetes, HvyAlcoholComcump, AnyHealthcare, and NoDocbcCost. Additionally, we have noticed that there are too many levels in variables like BMI, MentHlth, and PhysHlth. To address the issue of having numerous bunch levels of BMI, MentHlth, and PhysHlth, we have made the following decisions regarding categorization:

1. BMI: We have categorized BMI according to CDC standards [3] into four groups: "Underweight,"(labeled as 1) "Normal,"(labeled as 2) "Overweight," (labeled as 3), and "Obese." (labeled as 4)
2. Mental Health and Physical Health: We have chosen to categorize them naturally into five levels based on 3 weeks. Since we observed that there are many data points lying in the 0 categories, we have categorized them as follows: 0 (labeled as 0), 1-7 (labeled as 1), 8-14 (labeled as 2), 15-21 (labeled as 3), and 21+ (labeled as 4).

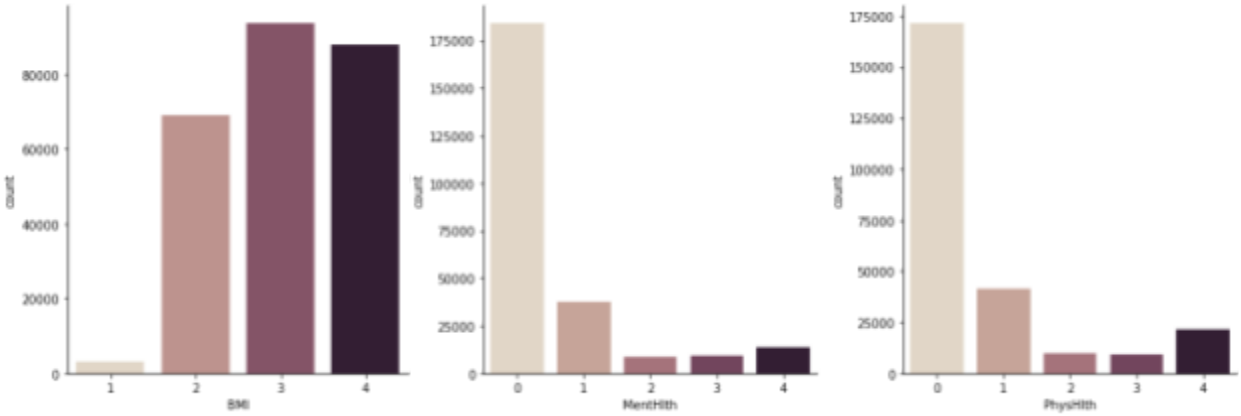


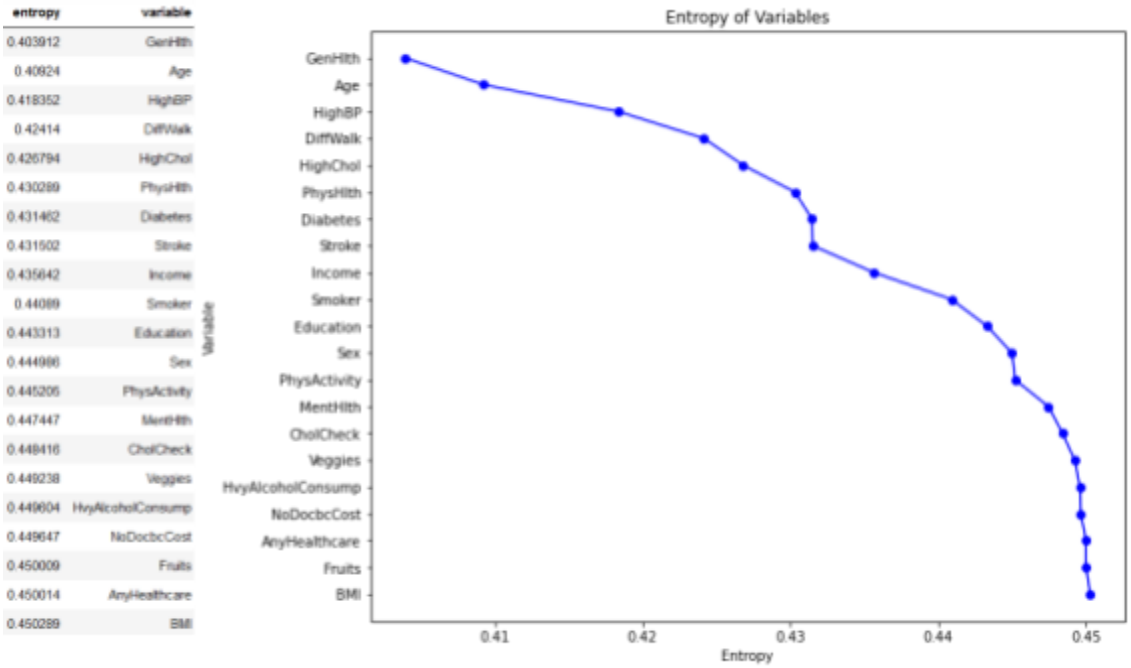

Figure 2. Histograms for BMI, MenHlth and PhysHlth

After categorizing BMI into four levels, named "Underweight," "Normal," "Overweight," and "Obese," we observed that the majority of the population in the survey data falls into the "Overweight" and "Obese" categories, specifically within the ranges (24.9, 29.9] and (29.9, inf] respectively, which indicates a higher prevalence of higher BMI values among the respondents. When examining the categorization of Mental Health (MentHlth) and Physical Health (PhysHlth), we noticed that a significant portion of the population is categorized as level 0, which implies that a large number of individuals reported not experiencing stress, depression, or problems with emotions, as well as physical illness or injury during the past 30 days.

Before conducting modeling and exploratory data analysis, our objective is to identify the most important variables associated with the response variable "HeartDiseaseorAttack." To achieve this, we will calculate the conditional entropy of "HeartDiseaseorAttack" given other variables.

By calculating the conditional entropy, we aim to measure the amount of information each variable provides in relation to "HeartDiseaseorAttack." This analysis will help us determine the variables that exhibit a strong association or predictive power with heart disease or heart attacks. Once we have identified the variables with significant conditional entropy values, we can focus on them during the modeling and exploratory data analysis phases. They are likely to play a crucial role in understanding and predicting the presence of heart disease or heart attacks in our dataset.

*Contingency Table Entropy Plot*



| entropy | variable |
|---|---|
| 0.403912 | GenHlth |
| 0.40924 | Age |
| 0.418352 | HighBP |
| 0.42414 | DiffWalk |
| 0.426794 | HighChol |
| 0.430289 | PhysHlth |
| 0.431462 | Diabetes |
| 0.431502 | Stroke |
| 0.435642 | Income |
| 0.44089 | Smoker |
| 0.443313 | Education |
| 0.444986 | Sex |
| 0.445206 | PhysActivity |
| 0.447447 | MentHlth |
| 0.448416 | CholCheck |
| 0.449238 | Veggies |
| 0.449604 | HvyAlcoholConsump |
| 0.449647 | NoDocbcCost |
| 0.450009 | Fruits |
| 0.450014 | AnyHealthcare |
| 0.450289 | BMI |



Figure 3. Entropy for heart disease against every variable sorted from lowest to highest and graphed

Upon examining the entropy results for the "HeartDiseaseorAttack" variable in relation to each column in our dataset, we have identified the top 6 variables that exhibit lower entropy values. These variables are GenHlth (General Health), Age, HighBP (High Blood Pressure), DiffWalk (Difficulty at Walking), HighChol (High Cholesterol), and PhysHlth (Physical Health). Lower entropy values suggest that these variables carry more

important information or are strongly associated with the occurrence of heart disease or heart attacks. Hence, they are likely influential factors in understanding the presence of these conditions. We have explored these top 6 variables with lower entropy to gain further insights. By delving into these variables, we aim to uncover potential reasons for their strong association with heart disease or heart attacks. This exploration will enable us to understand the relationships, patterns, and underlying factors contributing to these variables' lower entropy values.

We also noted that certain variables exhibit higher entropy values. Higher entropy values indicate a greater degree of uncertainty or randomness in their relationship with the occurrence of heart disease or heart attacks. While these variables may still hold importance, our initial focus will be exploring the top 6 variables with lower entropy to uncover potential insights related to heart disease or heart attack risks. We use contingency heatmaps to visualize variables with multiple levels and mosaic plots to visualize variables with only 0 or 1.

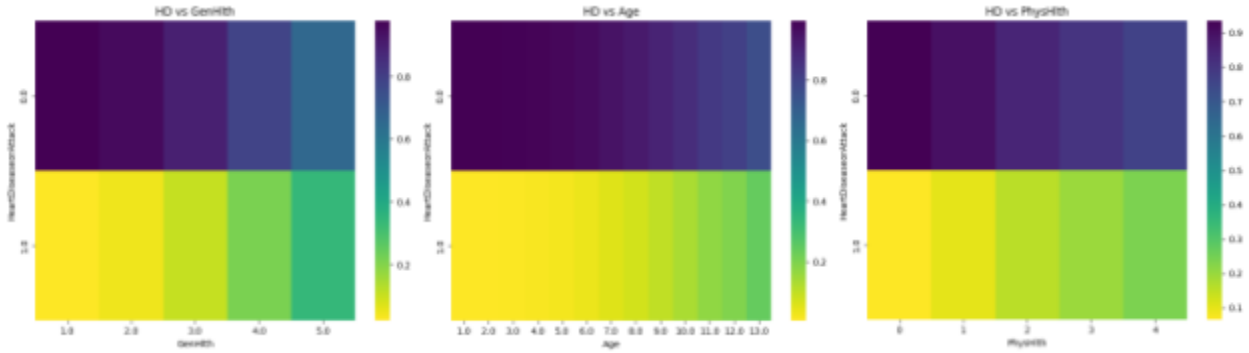



Figure 4. Heat maps for contingency tables of heart diseases against general health, age, and physical health

The most left plot is HeartDiseaseorAttack against GenHlth. Under five different GenHlth (General Health) scenarios, we can see a trend of color varies from level 1 to level 5. The bottom blocks represent the percentage of people with Heart Disease or Attack within five different General Health conditions. The color turns darker from left to right, meaning the people assigned to a terrible  General Health condition group, the proportion of people with Heart Disease or Attack higher.

The Heart Disease Against Age plot lies in the middle. There are thirteen different age levels, from level 1 (18-24) to level 13 (80+). Looking at age 1 (18-24), most people do not have heart disease or attack. While in the age group 13 (80+), there is a significantly increased amount of people with heart disease or attack compared to age group 1. Looking at the right side, we have Heart Disease against Physical health. Physical health has five levels, where 0 is a healthy person, and 4 is the least healthy. Looking at the heat map, we can see that healthy people are mainly in the 0 category of Heart disease. In contrast, people with the worst physical health of 4 have a more significant division between having Heart disease and not. There seems to be a clear indicator that people with good physical health are less likely to have heart disease.

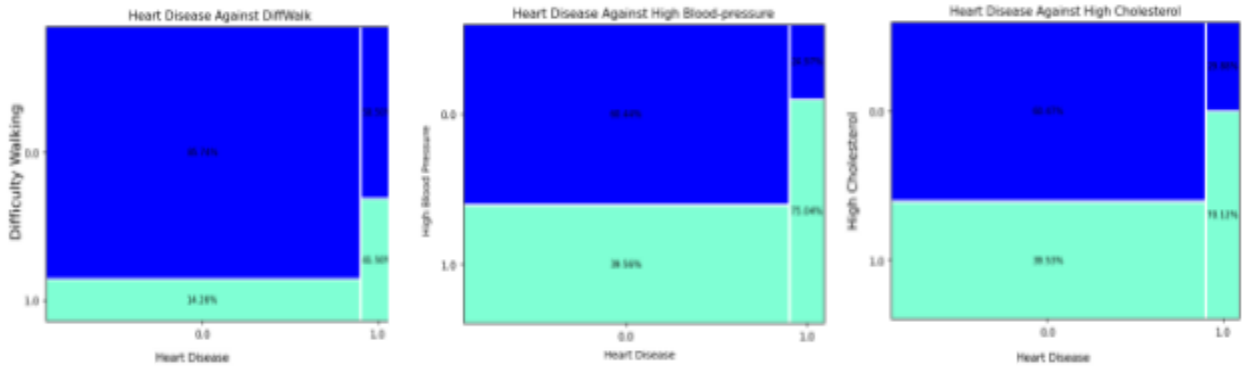



Figure 5. Mosaic plot for heart disease against difficulty walking, high blood pressure, and high cholesterol

In HeartDiseaseorAttack against DiffWalk (Difficulty at Walking) plot, we see that there is a higher population on zero than in one when the DiffWalk (Difficult at Walking) is zero. Meaning that people with no issue with their walking are mostly without heart disease, but we see that for people with difficulty with walking, there is a higher population with Heart Disease.

In the case of high blood pressure, we observe that the population of people with heart disease is more significant in the group with high blood pressure than in those without high blood pressure. Looking at the plot of Heart Disease against High Cholesterol, we find that people without High Cholesterol are likelier not to have heart disease. In contrast, people with High Cholesterol are more likely to have heart disease.

*Odds Plots for some interesting variables*

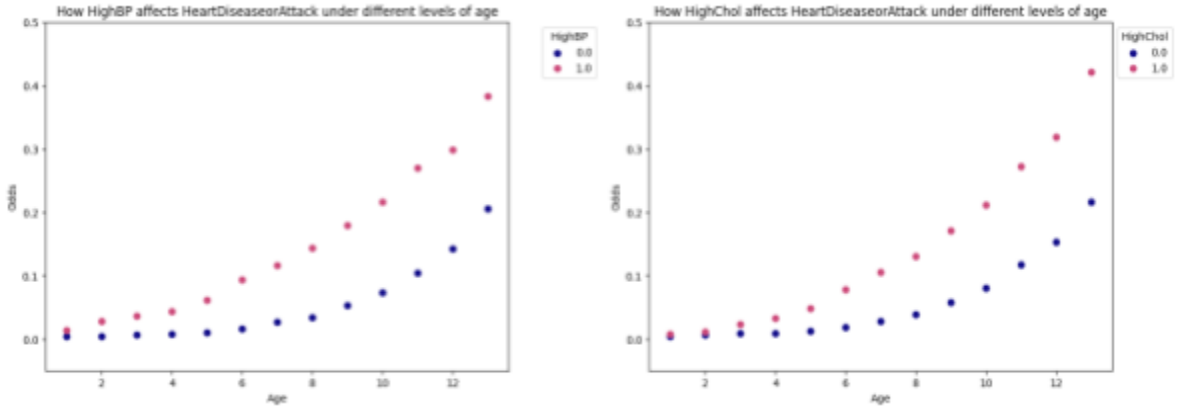



Figure 6. Odds plots for Heart Disease or Attack based on High Blood Pressure, High Cholesterol, and Age

Age emerges as the second variable with the lowest entropy in the entropy plot. Unveiling the odds associated with each age level is vital for several reasons. Age, a widely recognized risk factor for heart disease [4], demands meticulous exploration to comprehend its impact on the odds of developing this condition in the presence of HighBP and HighChol. By intricately examining the odds across distinct age levels, we can unravel age-specific nuances, thereby illuminating patterns and variations that shed light on the influence of HighBP and HighChol on heart disease within each unique cohort. Not surprisingly, we discovered that HighBP (High Blood Pressure) and HighChol (High Cholesterol) have similar effects on Heart disease or attack given different age groups. Also, one study published in the Journal of the American College of Cardiology by K. Shah et al. (2015) examined the association between cholesterol levels and blood pressure. [5] The study involved a large cohort of 4,385 participants and found a positive correlation between total cholesterol levels and blood pressure, which suggests that individuals with higher cholesterol levels were likely to have elevated blood pressure. Thus, we decide to analyze the difference between males and females further.

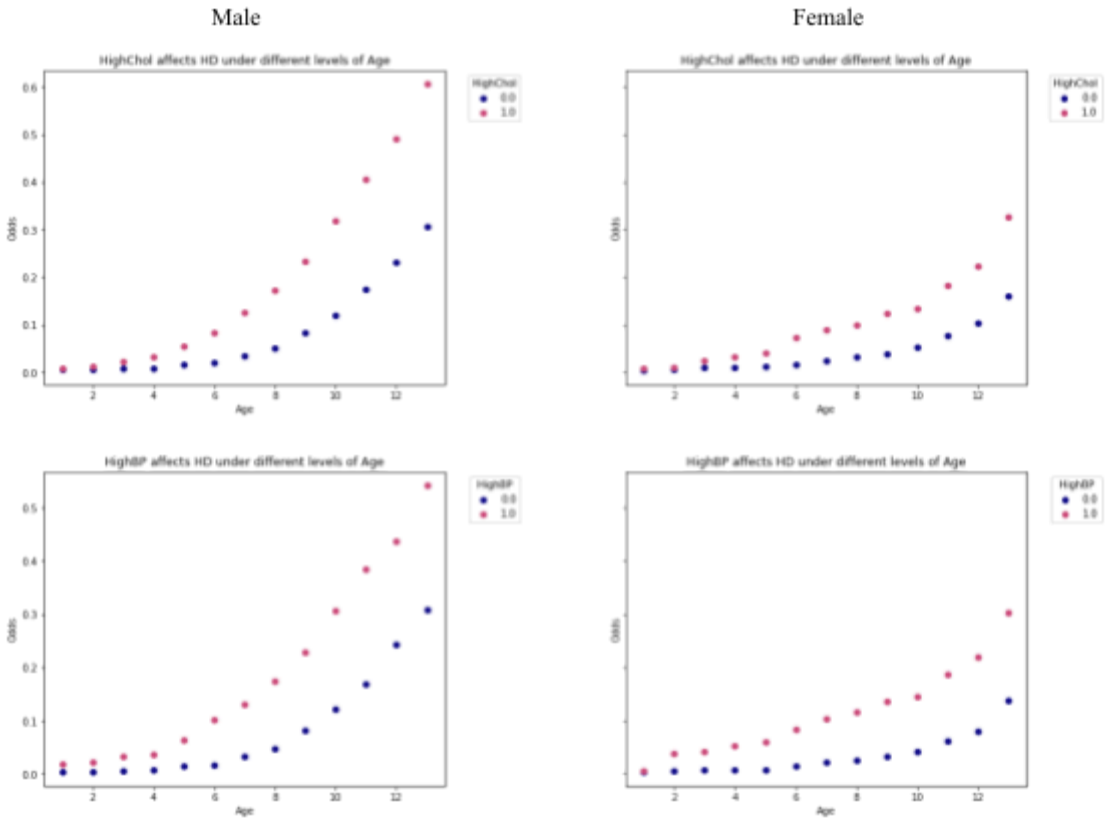



Figure 7. Odds plots for Heart Disease or Attack based on High Blood Pressure, High Cholesterol, and Age categorized by gender

We conducted an analysis to explore the relationship between gender, age, and the prevalence of heart disease. We aimed to identify any disparities in the impact of high blood pressure (HighBP) and high cholesterol (HighChol) on the occurrence of heart disease between males and females. We created four plots to visualize the data, with the left two plots focusing on males and the right two on females. The plots illustrate how HighBP and HighChol affect the likelihood of heart disease across different age groups. Upon comparing the plots horizontally, we observed a notable trend. For males, as age increases, individuals with high blood pressure or high cholesterol tend to have a higher proportion of heart disease or heart attacks. However, it is essential to note that the sample sizes for males and females are imbalanced, with more females included in the analysis. Based on this analysis, we hesitate to conclude that males have a higher portion of individuals with heart disease or heart attacks than females. The observed trend might be influenced by factors other than gender, such as the underlying health conditions or lifestyle habits within the studied population. To draw more accurate conclusions regarding the gender-specific impact on heart disease prevalence, further investigation with a more balanced sample size and consideration of additional

factors is necessary. Future research should aim to gather data from a more extensive and diverse population to provide a comprehensive understanding of the relationship between gender, age, and heart disease.

On the other hand, we want to explore more information in the BMI group. Despite BMI having the highest entropy, it is worth noting that the CDC acknowledges the heightened risk of various diseases and health conditions among individuals with obesity.[6] This risk encompasses high blood pressure, high cholesterol, and stroke. Further exploration is warranted to gain a comprehensive understanding of why BMI exhibits the highest entropy. By delving into this analysis, we can unravel the underlying factors contributing to the diverse distribution of BMI values and their associated health risks.

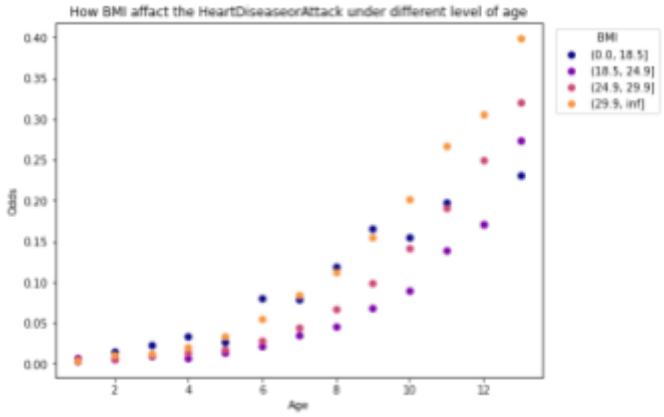



Figure 8. Odds plots for Heart Disease or Attack based on BMI and Age

BMI itself is not a significant predictor, as indicated by the conditional entropy. However, when we combine BMI with another important variable, "Age," we can observe some interesting patterns. In the Age groups 1-9, it appears that the underweight group has a slightly higher proportion of people with heart disease compared to those without heart disease. Conversely, in the age groups 10-13, the obese group exhibits a significantly higher ratio than the other three groups. It is important to exercise caution when interpreting the results for the underweight category. Many age groups within this category lack a representative population of individuals with heart disease or attacks. Particularly in age groups 1-5, there are only a few individuals with heart disease or attacks. Therefore, comparing the underweight group to the other BMI groups carries risks and should be done cautiously. Furthermore, when we compare only the normal and obese groups, we observe that the odds are almost double after age group 5 and continue to increase.

As a result, we have decided to consider BMI when comparing it with other variables under different levels of age. Upon analyzing the data, we try to find associations between different groups of BMI and HighBP (High Blood Pressure) and HighChol (High Cholesterol).

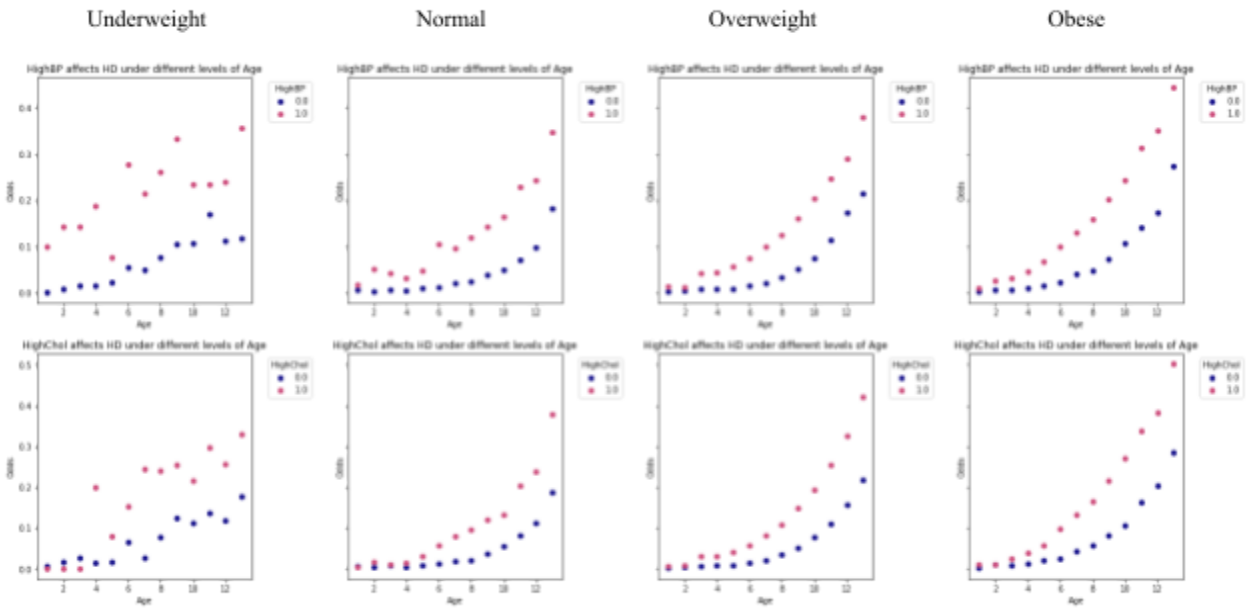



Figure 9. Odds plots for Heart Disease or Attack based on High Blood Pressure, High Cholesterol, and Age categorized by BMI

From the above plots, we divided our data into four different BMI levels and observed if there were any differences. However, this approach is only partially satisfactory due to the limited number of samples in the underweight category. Consequently, the odds plot for the underweight category exhibits a peculiar distribution. Therefore, we must carefully compare the underweight plot with the other three. On the other hand, the plots for the other three BMI levels consistently demonstrate a trend of higher odds with increasing age. When we conduct a horizontal comparison, we can observe a gradual increase in odds as Age increases. However, there is no significant difference in trend between the three different BMI levels, which is reflected in the high conditional entropy.

*Decision tree model 1*

We created a decision tree based on the top 6 lower entropy variables: GenHlth, Age, HighBP, DiffWalk, HighChol, and PhysHlth. The reason for choosing these variables is to determine if our exploratory analysis would work in our prediction model and how good these variables are at predicting the chance of having heart disease. We

balanced our data by using resampling and balancing out the data based on the length of heart disease equal to 1. The Resampling was done without replacement to evade bias.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.79 | 0.68 | 0.73 | 5974 |
| 1.0 | 0.72 | 0.82 | 0.77 | 5974 |
| accuracy | | | 0.75 | 11948 |
| macro avg | 0.76 | 0.75 | 0.75 | 11948 |
| weighted avg | 0.76 | 0.75 | 0.75 | 11948 |

Figure 10. Score table for the decision tree model with heart disease against top 5 low entropy

With a decision tree of depth 6, we find with 75% accuracy that our chosen predictor helps predict whether a person would have heart disease. That means that our top five lower entropy are associated with indicating if a person has heart disease or not. The Precision is high in both cases meaning that we correctly predict heart disease with 72% and no heart disease with 79%.

*Part 2. Heart Disease against fuse variables*

To get more insights between heart disease and fuse variables, we fused all 21 variables against each other, and by applying our conditional entropy function, we finally decided to choose the top 7 variables with the lowest conditional entropy values. As expected, the category with the lowest entropy is GenHlth_Age, since GenHlth (General Health) and Age were ranked as the top two in the individual variables.

| entropy | variable |
|---|---|
| 0.370143 | GenHlth_Age |
| 0.387881 | PhysHlth_Age |
| 0.390002 | HighBP_GenHlth |
| 0.390154 | HighChol_GenHlth |
| 0.391168 | DiffWalk_Age |
| 0.393248 | HighBP_Age |
| 0.394037 | GenHlth_Stroke |

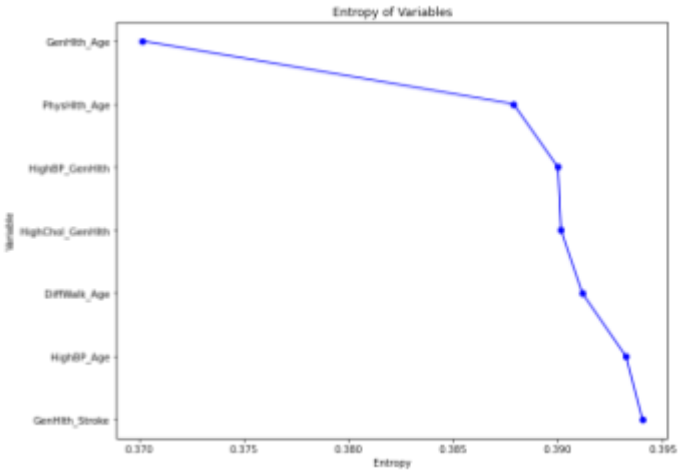



Figure 11. Conditional Entropy for Heart disease against fused variable sorted from lowest entropy to highest and graphed

After running our entropy model, we choose the top 6 variables with the lowest conditional entropy. In this case, we find fused variables 'GenHlth_Age', 'PhysHlth_Age', 'HighBP_GenHlth', 'HighChol_GenHlth', 'DiffWalk_Age', 'HighBP_Age' have the lowest entropy. These variables are similar to our original variables for our conditional entropy. Age seems to be a prevalent factor in conditional entropy which we would like to explore more in-depth and find out why this may be. However, 'GenHlth_Stroke' is the one that surprised us since stroke was ranked No.8. And we wonder how does the distribution might look like after we combine stroke with GenHlth.

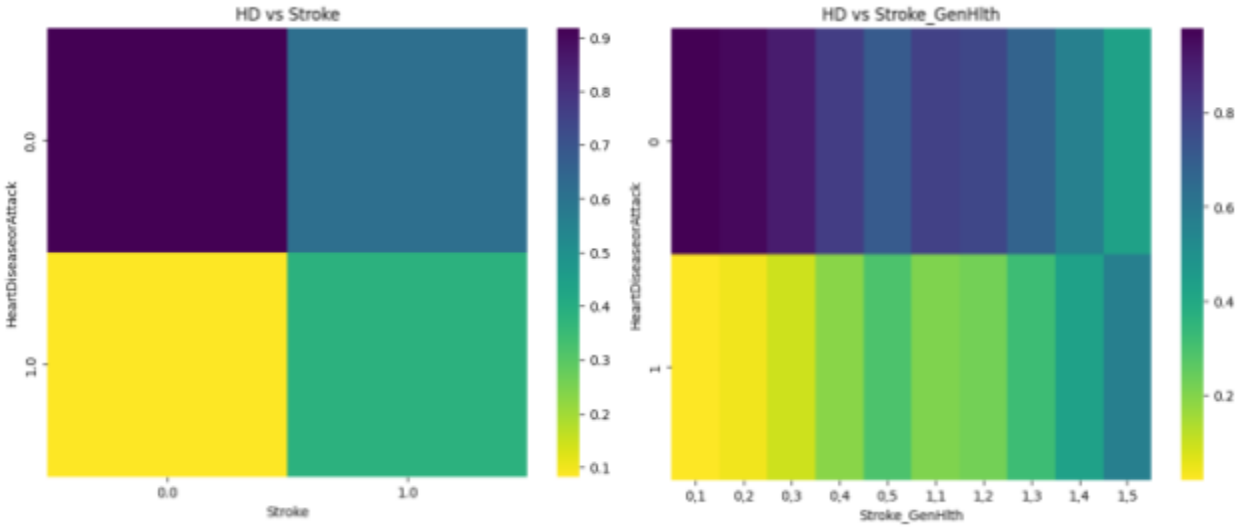



Figure 12. Heatmap for the contingency table of heart disease against stroke and stroke combined with general health

In the left plot for Figure 12, we can see a significant difference between the first and second columns, meaning that there are many more portion people who get Heart Disease if they had a Stroke. Furthermore, when we combine Stroke with GenHlth, we can see a trend either people who had a stroke or not. By looking at the lower trunk of the right side plot, we see that no matter whether people had a stroke or not, as their general health condition got worse, we can see the color of the heat map turn darker, meaning there is a more significant portion of people who have heart disease.

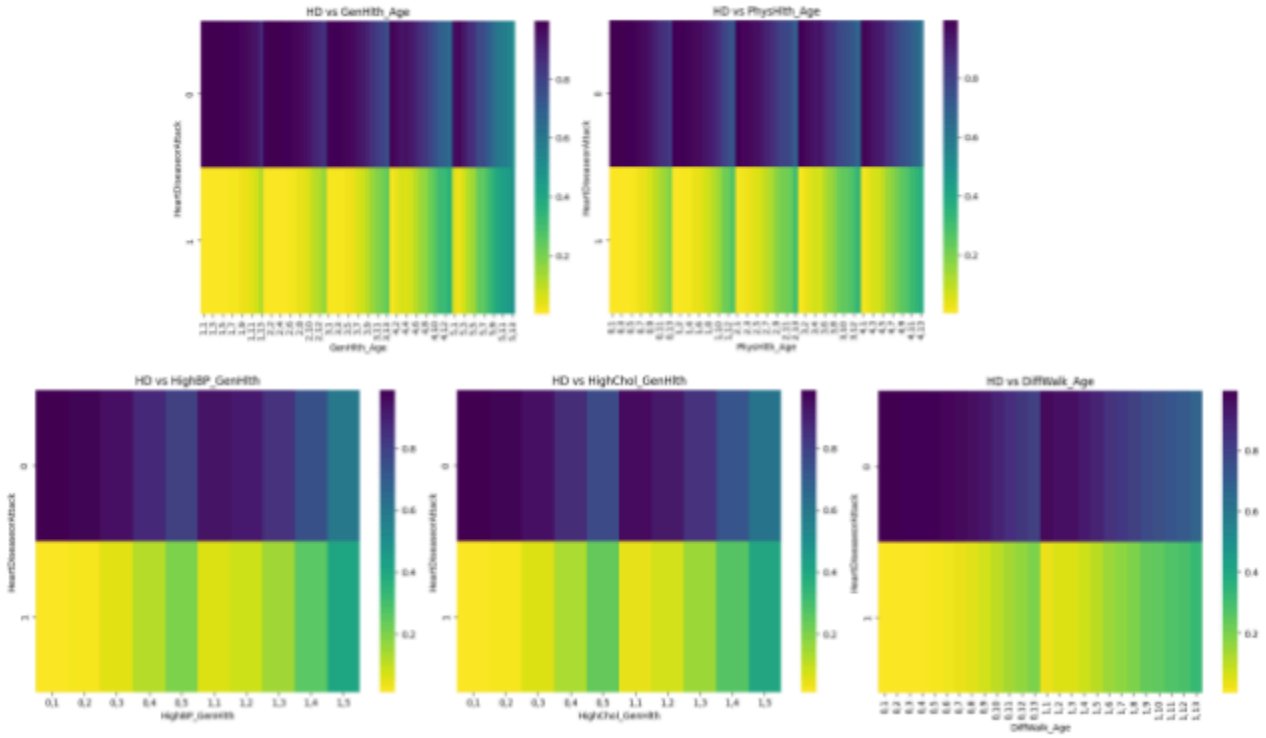



Figure 13. Heat map for the contingency table of heart disease against general health and age, physical health and age, high blood pressure and general health, high cholesterol and general health, and difficulty walking and age.

Looking at Fig. 13. the contingency heat map for heart disease against general health and age, we can see that for each group, the shift in a population happens toward having heart disease as age increases. People with good general health but old age still mainly focus on having heart disease, which is more noticeable when general health is the worst as age increases. The shift change is faster as age increases. Observing heart disease against physical health and age, we see a similar trend where there is a shift in the population having heart disease as age increases. When physical health is the worst, this phenomenon happens faster.

Watching heart disease against high blood pressure and general health. People with worst general health but not high blood pressure start concentrating on heart disease as a general decrease. The shift in population is faster when there is the presence of high blood pressure.

Regarding heart disease against high cholesterol and general health, we detect that as general health decreases when there is no high cholesterol, the population is relocated to having heart disease. This motion change in

population increases as well when there is the presence of high cholesterol and is almost even when the general health is the worst and there high cholesterol is present.

Finally, viewing difficulty walking and age against heart disease, we noticed that people without difficulty walking does not change the population concentration when age increases. On the other hand, when they have difficulty walking as age increases, the population change towards having heart disease. It is almost even when there is a presence of difficulty walking, and age is the highest.

*Decision Tree Model 2*

For the second decision tree model, we used heart disease against our top 5 five low entropy fused variables, this being; "GenHlth_Age", "PhysHlth_Age", "HighBP_GenHlth", "HighChol_GenHlth", "DiffWalk_Age", 'HighBP_Age'. This is because we have explored this variable enough to find the most essential understanding of what causes an increased chance of heart disease. For this model, we used sampling to balance the categories of heart diseases and resampling without replacement to evade bias.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.77 | 0.70 | 0.73 | 5974 |
| 1.0 | 0.73 | 0.80 | 0.76 | 5974 |
| accuracy | | | 0.75 | 11948 |
| macro avg | 0.75 | 0.75 | 0.75 | 11948 |
| weighted avg | 0.75 | 0.75 | 0.75 | 11948 |

Figure 14. Score for the Decision tree model for heart disease against the top 5 lowest entropy fused variables

Once we ran the model, we got an accuracy of 75%, which is similar to our simpler model, which is likely because the variables we used for it are the combinations of similar variables we used for our simple model. This Meaning that even though we have fused the variable, they still yell similar results. In this case we are predicting with more precision people with heart disease with 73% while not heartdisease with 77%. That meaning that our new predictors increase our chances of predicting accurately if a person has heart disease rather that not hear disease.

*Part 3. Heart disease and stroke against single variables*

Heart disease holds the position of being the primary cause of death in the United States, while stroke ranks as the fifth leading cause of death. [7] Therefore, we decided to explore the relationship between heart disease and stroke because there may be a relationship between a person having heart disease, stroke, or both. This means we want to find by combining if they influence each other. For this fused variable, we decided to combine heart disease with stroke and code it numerically so we could get the entropy by creating a new variable.

| Variable | Sign | Code |
|---|---|---|
| (No Heart disease, No stroke) | (-, -) | 0 |
| (Heart disease, No stroke) | (+,-) | 1 |
| (No Heart disease, stroke) | (-,+) | 2 |
| (Heart disease, stroke) | (+,+) | 3 |

Table 2. Variable explanation for fused of heart disease and stroke

*Conditional Entropy Plots for HD + Stroke vs other variables*

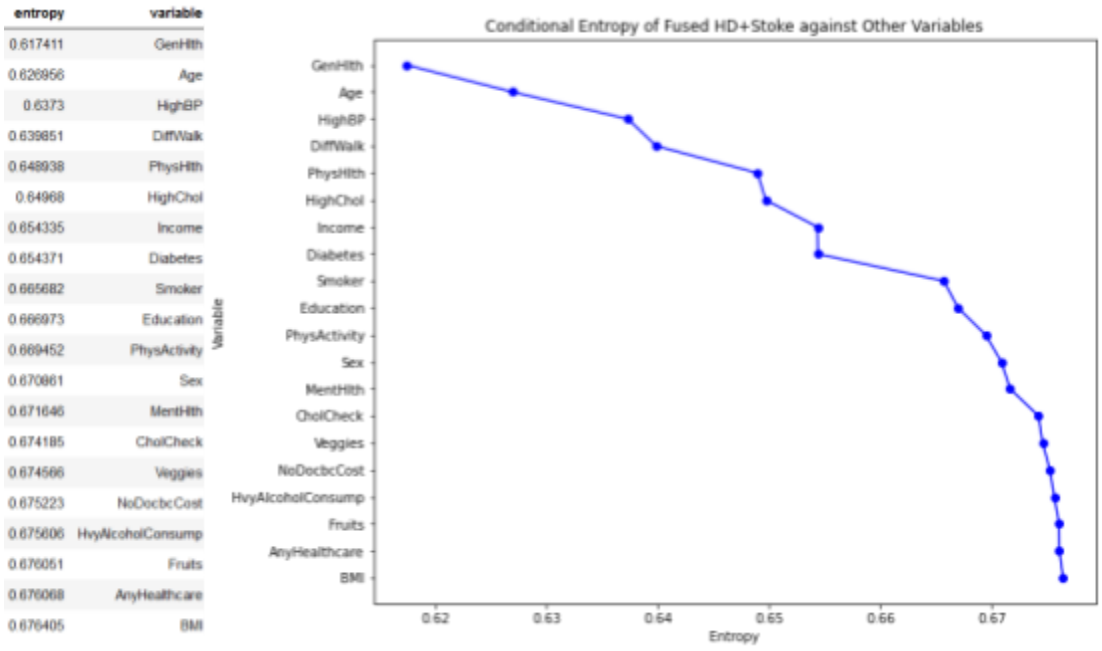



Figure 15. Conditional entropy for heart disease and stroke against one variable sorted from lowest to highest and graphed.

We find the entropy for the fuse variable to be similar to the entropy for the individual variable of heart disease. We find this interesting because it may show that there seems to be some type of relationship between a person who has or does not have heart disease and a stroke. We want to explore the contingency table of the top 5 lower entropy.

*Contingency table heatmaps*

These plots have a y-axis of HD_Stroke, a fused variable of Heart Disease or Attack and Stroke. It contains four levels: 0,0 (No Heart Disease or Attack nor Stroke), 0,1 (No Heart Disease or Attack but Stroke), 1,0 (Have Heart Disease or Attack but No Stroke), 1,1(Have both Heart Disease and Stroke)

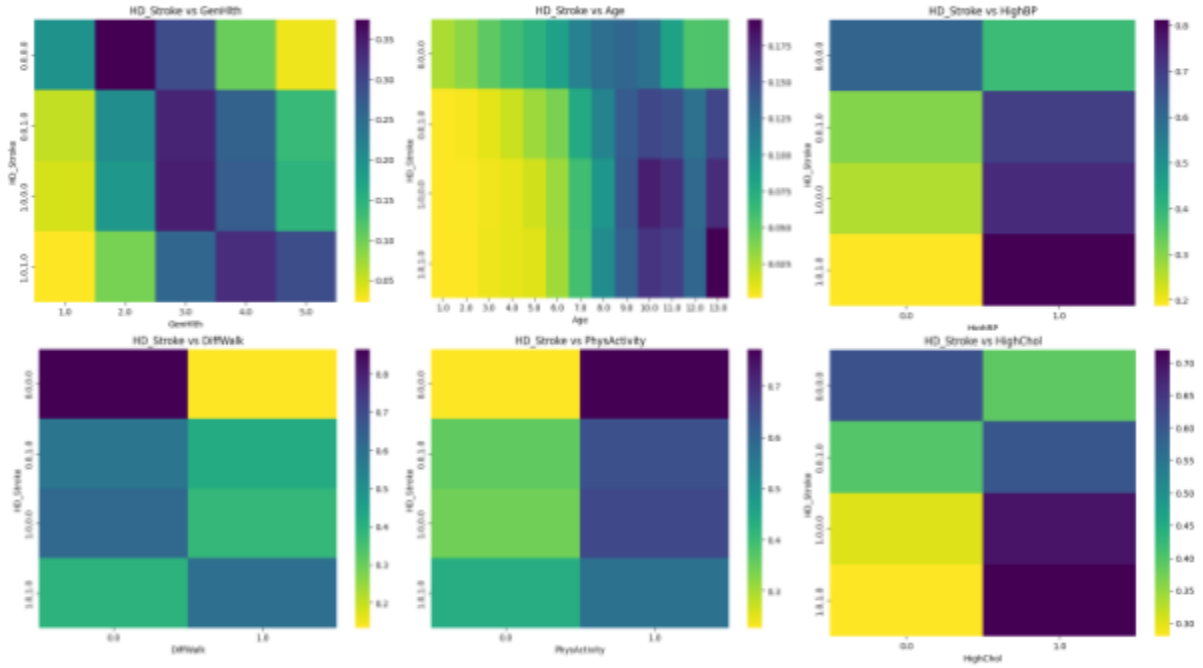



Figure 16. Heat map for the contingency table of Heart disease and stroke against general health, age, high blood pressure, difficulty walking, physical health, and high cholesterol.

We can see a trend in heart disease and stroke against general health. Whereas general becomes worse, the population starts shifting on concentration to having either heart disease, stroke, or both. In general health levels 4 and 5, most of the population is concentrated on heart disease and stroke. When we look at Heart disease and stroke against age, we see a trend that as age increases, the population concentration shifts from people not having heart disease or stroke to people having both when the Age group is labeled 13 (80 and beyond). Finally, observing heart disease and stroke against high blood pressure, we found that most people with no heart disease and no stroke are not diagnosed with high blood pressure. In contrast, people with either heart disease, stroke, or both are focused on high blood pressure.

Then viewing the plot of Heart disease and stroke against difficulty walking, we see that there does not seem to be a trend where the population is not centering in either. Most people have no difficulty walking, even with either disease. However, when both diseases are present, there is a shift in the population focus, and most people have difficulty walking when both diseases are present. When we view heart disease and stroke against high cholesterol, we see that when there is no presence of heart disease or stroke, the population is centered on no high cholesterol. At the same time, if either is present, the population has high cholesterol, meaning that either disease is associated with high cholesterol.

*Decision Tree Model 3*

We then decided to run our decision tree on the top 6 lower entropy variable against Heart disease fused with stroke. In this case, we did resampling with replacement due to the low count of (3 = a person has both heart disease and stroke).

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.81 | 0.80 | 976 |
| 1 | 0.73 | 0.65 | 0.69 | 987 |
| 2 | 0.65 | 0.66 | 0.66 | 1008 |
| 3 | 0.75 | 0.79 | 0.77 | 966 |
| accuracy | | | 0.73 | 3937 |
| macro avg | 0.73 | 0.73 | 0.73 | 3937 |
| weighted avg | 0.73 | 0.73 | 0.73 | 3937 |

Figure 17. Score for the Decision tree model for heart disease and stroke against the top 5 lowest entropy single variables.

We find in fig.17 an accuracy of 73% for all of our variables. However, we see that for the variable (2 = A person has heart disease but no stroke), we find that we are less likely to predict this category. This result is due to the more complex classification of our decision and the low count of all variables compared to 0. Thus we may find that even though the predictor we choose has the lowest entropy, the accuracy would be the lowest due to the uneven data distribution.

*Part 4. Heart disease and stroke against fused variables*

We then want to explore the relationship between heart disease and stroke against all the fused variables. In this case, we combined all other variables except for stroke and heart disease with each other. That leaves us with 190 combined variables to compare our fused variables. We want to do this because we want to see if there is a relationship between the fused information of heart disease and stroke and heart disease by itself. If we find similar variables in part 3, it would likely point out that there seems to be some type of relationship between them.

*Conditional Entropy of HD + Stroke vs other fused variables*

| entropy | variable |
|---|---|
| 0.575799 | GenHlth_Age |
| 0.597536 | HighBP_GenHlth |
| 0.598393 | PhysHlth_Age |
| 0.600297 | DiffWalk_Age |
| 0.602005 | HighChol_GenHlth |
| 0.60711 | HighBP_Age |
| 0.608843 | Age_Income |

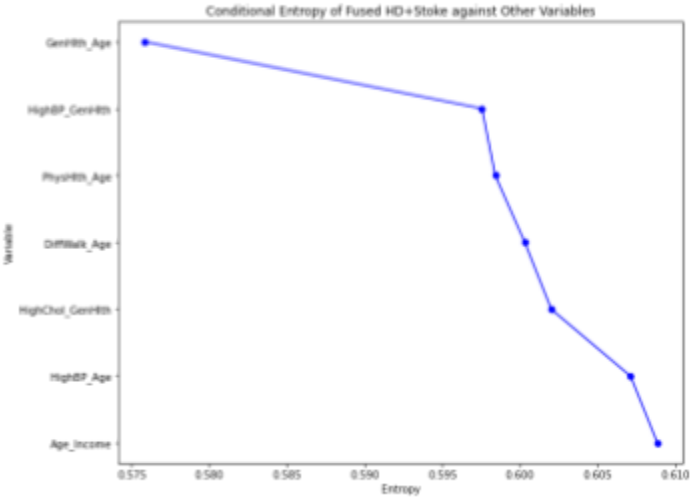



Figure 18. Conditional Entropy for Heart disease and Stroke against fused variable sorted from lowest entropy to highest and graphed

After finding the fused entropy for the fused model, we find again that we have similar values as heart disease against fused variables. Thus we have inferred that the relationship between heart disease and stroke is close, meaning that if a person has heart disease, stroke, or both, these factors are the closest to explaining why a person may have one disease or the other.

*Contingency Table Clustering Heatmap*

These plots have a y-axis of HD_Stroke, a fused variable of Heart Disease or Attack and Stroke. It contains four levels: 0,0 (No Heart Disease or Attack nor Stroke), 0,1 (No Heart Disease or Attack but Stroke), 1,0 (Have Heart

Disease or Attack but No Stroke), 1,1(Have both Heart Disease and Stroke). Among all the plots below, most of the groups' population is concentrated in the “No Heart Disease and No Stroke” Category due to the unbalanced data. According to the red lines in each of these plots, we can see that Hierarchical Clustering mainly divides the dataset into two main trunks: people who are more likely to have heart diseases or stroke or who do not have any of these diseases.

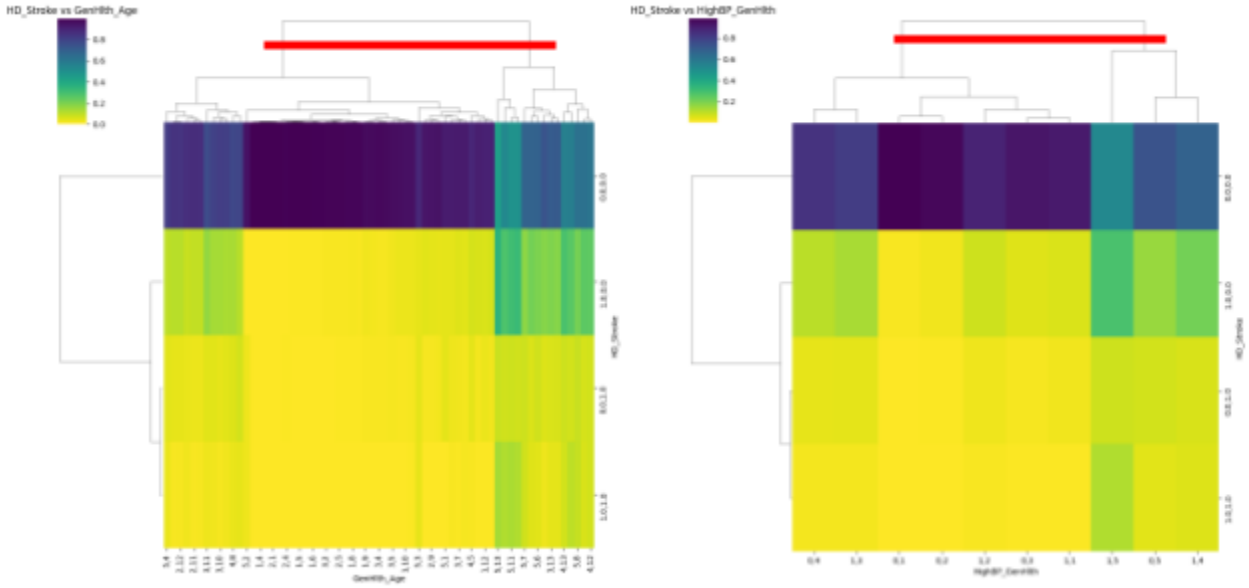


Figure 19. Clustering heatmaps based on contingency tables for HD+Stroke vs GenHlth+Age, HighBP+GenHlth.

In the plot of HD_Stroke vs GenHlth_Age, the clustering classifies which groups are more likely to have heart disease or attack based on the different levels of General Health and Age. The most distinctive group in this plot is (5,13), (5,12), (5,11), and (5,10) in GenHlth_Age. For these groups, it is less likely for them not to have any diseases. For example, when General Health Level is 5 (Poor General Health), and the Age group is 80+ (5,13) is most likely to have Heart Disease or Attack.

On the right side, we have the plot of HD_Stroke vs HighBP_GenHlth, the group (1,5) meaning having High Blood Pressure and poor General Health is more likely to have Heart Disease or Stroke.

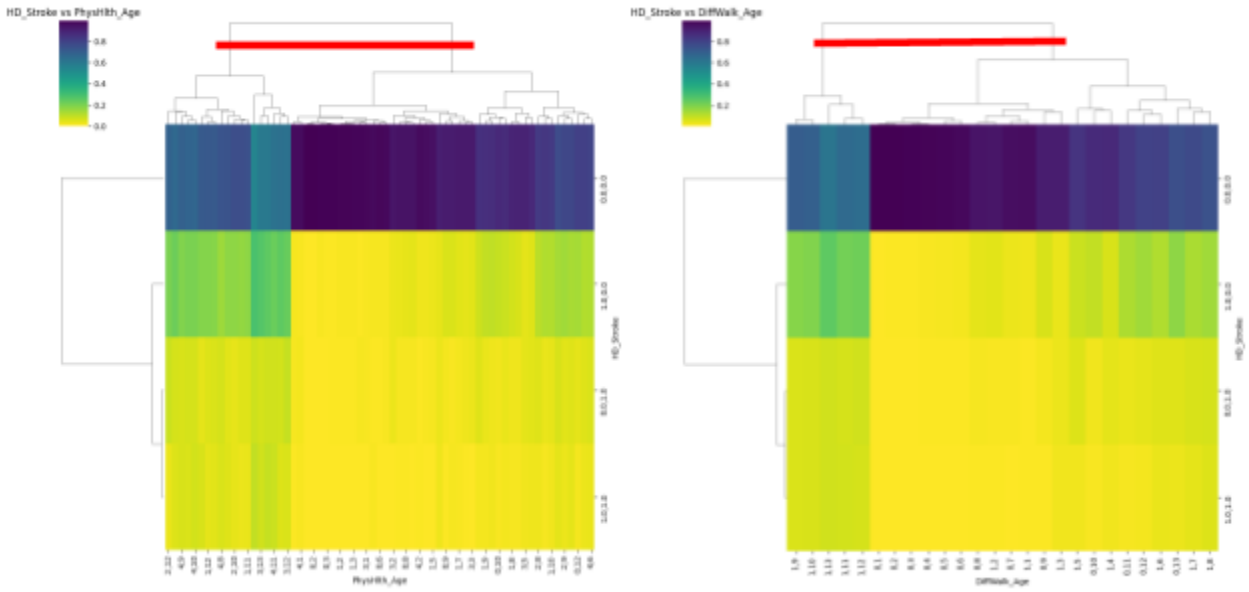


Figure 20. Clustering heatmaps based on contingency tables for HD+Stroke vs PhysHlth+Age, DiffWalk+Age.

In the plot of HD_Stroke vs PhysHlth_Age, the population is more likely to have diseases in the older age groups or have bad physical health. In the plot of HD_Stroke vs DiffWalk_Age, people who are 60 years old or older and have difficulty walking are likelier to have heart disease or stroke.

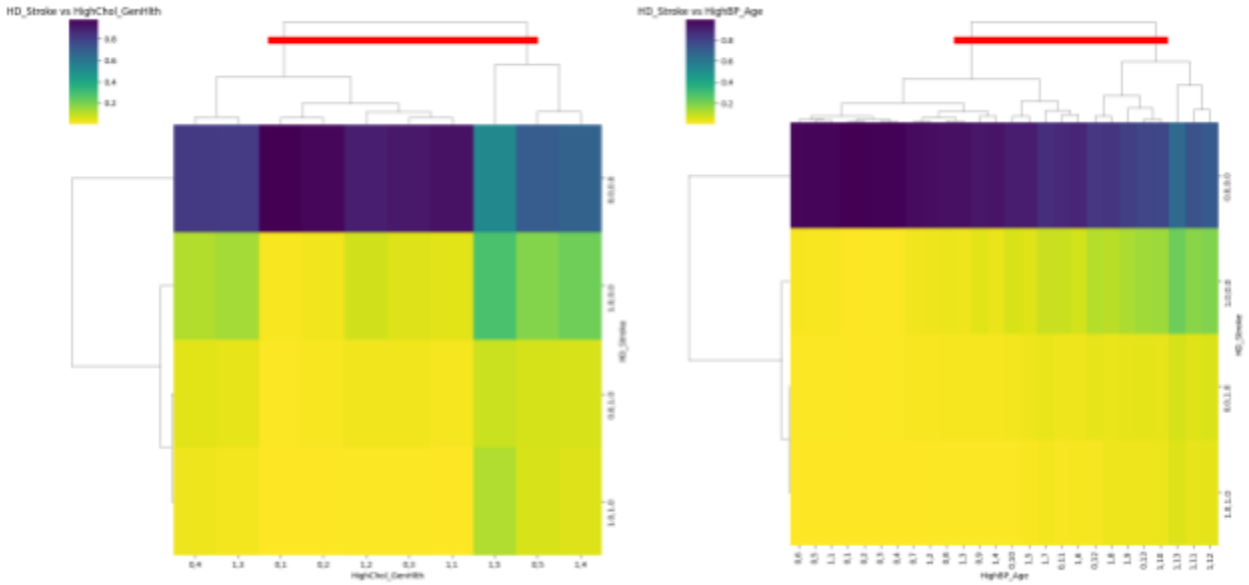


Figure 21. Clustering heatmaps based on contingency tables for HD+Stroke vs HighChol+GenHlth, HighBP+Age.

In the plot of HD_Stroke vs HighChol_GenHlth, three categories are more likely to have heart disease or stroke: (1,5)-people who have High Cholesterol and poor General Health, (0,5)-people who do not have High Cholesterol

but also poor General Health, (1,4)-people who have High Cholesterol and fair General Health. The plot of HD_Stroke vs HighBP_Age indicates that people who are older and have High Blood Pressure are more likely to have heart disease or stroke.

*Decision Tree Model 4*

We finally decided to run one more decision tree model based on our top 6 lowest entropy variables. In this case, we used the variables: "GenHlth_Age", "PhysHlth_Age", "HighBP_GenHlth", "HighChol_GenHlth", and "DiffWalk_Age" to create our decision tree model. We also resample with replacement based on the lowest variable (3 = a person has both heart disease and stroke).

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.76 | 0.77 | 1032 |
| 1 | 0.66 | 0.70 | 0.68 | 958 |
| 2 | 0.70 | 0.65 | 0.67 | 965 |
| 3 | 0.71 | 0.74 | 0.72 | 982 |
| accuracy | | | 0.71 | 3937 |
| macro avg | 0.71 | 0.71 | 0.71 | 3937 |
| weighted avg | 0.71 | 0.71 | 0.71 | 3937 |

Figure 22. Score for the Decision tree model for heart disease and stroke against the top 5 lowest entropy fused variables.

For this model, we got an accuracy of 71% which is 2% lower than our simple model, likely due to the complexity of the two fused variables brought against two fused variable responses. Especially when it comes to the variable (1 = Heart disease, not Stroke), we have the lowest precision, those we are less likely to predict a person with this condition. Even with the replacement and balancing of the variables, we still find it hard to predict.

## IV. CONCLUSION

After analyzing our decision tree model and contingency tables, there seems to be a clear trend for Heart disease where General health and Age seem to be the factors that give the most information and affect the likelihood of a person having heart disease. The simple trend seems to be that people with worse perceived general health are prone to being part of the Heart disease group. At the same time, people with better general health do not have heart disease. When we look at a person's age, there seems to be a higher concentration of people with heart disease when the age increases. Where is the opposite as age decreases, meaning that younger people would rarely be part of the heart disease group. If we look at High Cholesterol and High Blood Pressure, we find that being diagnosed allocated people more nearly the heart disease group.

Once we combined these four variables, we found that the more powerful group seemed to be general health and age, which helps us indicate a trend where people with both bad general health and old age are part of the heart disease group, whereas people in the other spectrum meaning that they are healthy and young are in the not heart disease group. There is some exception, such as people with bad general health who are young are in the heart disease category, but the counts tend to be low in those cases. Finally, we found that age has a low entropy combined with every other four essential variables, although age is not as important as general health. However, once combined with other variables, it becomes crucial. These trends appear when we combine heart disease and stroke, which would make us infer that the relationship for a person with heart disease or stroke seems to be highly based on their general health and age. Whereas high Blood pressure and High cholesterol, when presented, increase the chance of a person having heart disease, stroke, or both.

Furthermore, based on the entropy plot, although we have found that BMI is not as important as other variables, we still want to dive deeper due to CDC articles. We found that the underweight group had a different pattern than the other three groups when we observed each group individually based on the odds plots. However, we cannot conclude that the underweight group is more associated with heart disease because the sample size is way smaller than the other three groups.

All things considered, we would like to conclude that general health, age, high blood pressure, high cholesterol, and physical health are deciding factors that help us understand heart disease and stroke or both in ordinary life. In all cases, a person with good health and younger is less likely to suffer from heart disease or stroke, whereas the opposite is true.

## V. REFERENCE


[1] Centers for Disease Control and Prevention(CDC). 14 Oct. 2022. “Heart Disease Facts.” www.cdc.gov/heartdisease/facts.htm.

[2] Conditional Entropy function: https://datascience.stackexchange.com/questions/58565/conditional-entropy-calculation-in-python-hyx

[3] Centers for Disease Control and Prevention (CDC). (n.d.). Body Mass Index (BMI): Adult BMI. Retrieved from https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

[4] World Health Organization. (2022). Ageing and Health: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

[5] Shah K, et al. (2015). Relationship between Serum Lipids and Blood Pressure: The INTERGROWTH-21st Project. Journal of the American College of Cardiology, 66(4), S17.

[6] Centers for Disease Control and Prevention(CDC). 24 Sept. 2022. “Health Effects of Overweight and Obesity.” www.cdc.gov/healthyweight/effects/index.html.

[7] Murphy, S.L., Xu, J.Q., Kochanek, K.D., & Arias, E. (2018). Mortality in the United States, 2017. Retrieved from https://www.cdc.gov/nchs/data/databriefs/db328-h.pdf

[8] “Heart Disease Health Indicators Dataset.” www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset.