STA 160 Midterm Project Group 6 | Identity in Health

Rohan Arumugam
ID: 918891356

Jared Choy
ID: 920134917

Zain Khalid
ID: 919242548

Christopher Monzon
ID: 917207750

Abstract

Using data collected from the Behavioral Risk Factor Surveillance System (BRFSS), we evaluated and examined key risk factors that are associated with heart disease. Our goal was to see if there were any variables that heavily contributed to a person's chance of developing heart disease. Subsetting by health and identity, we examined the relationship between heart disease and other risk factors with conditional entropy and contingency tables. We were able to see how age, income, high blood pressure/cholesterol, gender, and stroke all played an influential role in determining a person's likelihood of developing heart disease. People who had other pre-existing health conditions were more likely to have heart disease compared to other healthy individuals. Identifying what factors play a strong role in predicting heart disease is paramount when it comes to creating methods to prevent heart disease in the general population, and help physicians with diagnosis.

1 Introduction

Data analysis as we know it, has become a field that is essential to find new discoveries, but also confirm what we already know. American industry is all too familiar with the application of both data science and data analysis, as many companies in the U.S. look to bolster this department [1] to aid in the decision making process. A field that can benefit substantially from analysis, is the practice of medicine. Medicine and diagnosis uses data given by the patient in order to come to a conclusion that can determine the quality of an individual's life. Due to the intersection between these two topics, let us apply this to the topic of heart disease in human life. Heart disease, especially in America, is a chronic illness that can strip a person of a normal life. The impact is not just limited to the personal scope, but also the global scope. In 2020, researchers from the European Study of Cardiology estimated that preventing all future cases of cardiovascular disease could save upwards of $15 billion in GDP, akin to $75,000 per each case [2].

When thinking of heart disease, it is crucial for a doctor to know the background and context of the patient. When we think of a patient's background, we think of variables such as age, gender, cholesterol, etc. However, we can subset these variables to elucidate patterns of the patient. This report aims to explicate the differences between a patient's identity and their health, and see how these can help a doctor in diagnosis.

The data was collected in a *2015 BRFSS (Behavioral Risk Factor Surveillance System)* study by the CDC and contains 253,680 survey responses [3]. Compiled and cleaned in Kaggle, we will use the data provided to us in order to explore and conclude our findings.

2 Exploratory Data Analysis

To begin, we first want to understand the dataset to get a picture as to what is being asked from each individual, and what does each variable entail. We will do so by looking at different topics within this branch in our various subsections.

2.1 EDA: Understanding the Dataset

Health Disease Indicators			
Feature	Values	Feature	Values
Heart Disease/Attack	0/1	HvyAlcoholConsump	0/1
HighBP	0/1	AnyHealthcare	0/1
HighChol	0/1	NoDocbcCost	0/1
CholCheck	0/1	GenHlth	[1, 5]
BMI	[12,98]	MentHlth	[0, 30]
Smoker	0/1	PhysHlth	[0, 30]
Stroke	0/1	DiffWalk	0/1
Diabetes	[0, 2]	**Sex**	0/1
PhysActivity	0/1	**Age**	[1, 13]
Fruits	0/1	Education	[1, 6]
Veggies	0/1	**Income**	[1, 8]

Table 1: Dataframe of Heart Disease and Health Variables

From our dataset, we wanted to emphasize mainly on the bolded variables. As Heart Disease or Attack as the response variable, we subsetted the rest of the variables by identity and health. In the identity category, we have Age, Sex, and Income. In the health category, we have Stroke, HighBP, and HighChol. As we can see from the table, Heart Disease and the health variables all are binary, categorical variables. This is extremely important to note, as it impacts the way that we can conduct testing and modelling on the data; this will be covered later on.

The variables follow a very easy pattern. For instance in the health variables, a zero for one of these variables would indicate an individual lacking this trait, while a one represents a positive occurrence. The identity variables are a little different. Sex is also represented as a binary variable, but a zero represents a male and a one represents a female. Age and income both exist in bounds [1, 13] and [1, 8] respectively. The numbers within these bounds represent different groups, and as the numbers get higher, so does the variable. For instance, an income group of 8 would indicate somebody is in the upper echelon of society, and has access to a bountiful amount of wealth. This also implies that an individual in the age group of 13 would be much older than the rest of the groups.

For the overall programming portion of our project, we used Python as the sole language. Python is a high performance language that is especially useful for pre-processing and machine learning. It allows us to utilize many statistical methods to explore our data.

2.2 EDA: Descriptive Statistics

In order to make sure we are making sound analysis, it is imperative we understand our data. To do so, let us find preliminary statistics. Focusing on our continuous variables first, we made density plots of each variable and its relation to heart disease.

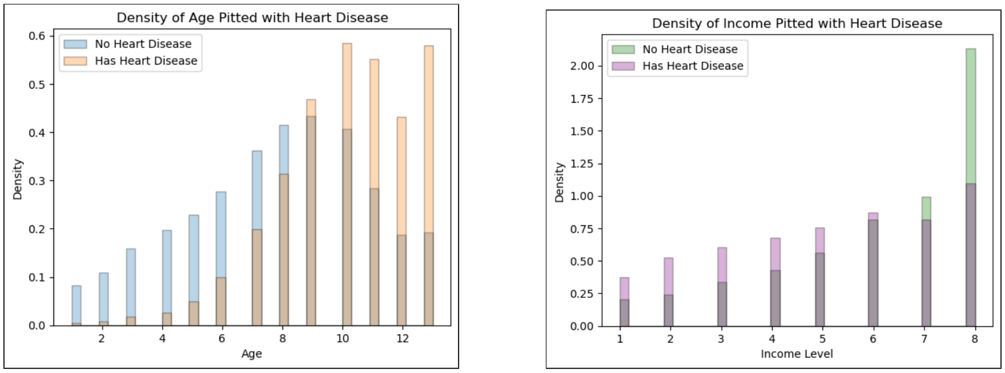


Figure 1: Density Plots - Age and Income

From these plots, we can notice a few things. Younger individuals generally do not possess heart disease. This is evident from the dominant blue bars in individuals in the six group and below, indicating that most of that population is less likely to have heart disease. When we get into the seventh group, it becomes a little more difficult to discern which heart disease value has the majority. Another interesting insight we can draw from this graph is that the survey results indicate a left-skewed population, which implies an older demographic. This makes sense, as we know that younger individuals are less likely to encounter and fall susceptible to disease[4]. Looking at the income plot, we see that our biggest group was the wealthiest class. For all of the other groups, we can observe that they are generally more probable for heart disease. However, the eighth group is seen to have a population more unlikely for heart disease. We found these findings intriguing from a societal perspective and gained a better understanding of them after conducting further research. From a *Lancet* research article, it was reported that "In high-income countries, low socioeconomic status is associated with an increased risk of cardiovascular disease and mortality"[5].

Aside from our other variables, we also noticed some important variables that would impact our other ones. These variables include CholCheck and NoDocbcCost. The CholCheck variable accounts for a question asked by BRFSS regarding whether or not an individual has had the cholesterol checked. A value of zero indicates that they have not had their cholesterol checked, which automatically marks the HighChol column as being a zero. We did not want to include these values, as they could be potentially misleading to our data. Due to this, we eliminated all instances of a zero in our future study regarding health. The NoDocbcCost variable was a point of probing for us as we our concerned with the identity surrounding the surveyed. This variable is binary encoded, with a zero meaning that they do have a doctor, and a one meaning they don't because of the cost. When we counted how many values of ones were in each income level, we found some surprising results. The middle to higher income groups actually had the most values of ones, while the lowest income groups had the least amount of values. This however, could be the fault of the imbalance in our data. However, it could also be due to medical assistance offered to the elderly population based on income. This could also explain why it's mostly the middle class that occupy this variable.

From our exploratory data analysis, we discovered some very important aspects of our data. The density plots from before suggested that we were working with very imbalanced data. Due to this imbalance, we had to change the way we wanted to model the data.

3 Conditional Entropy

Conditional entropy is the measure of uncertainty of a random variable; our goal is to find most significant risk factors that affect heart disease.

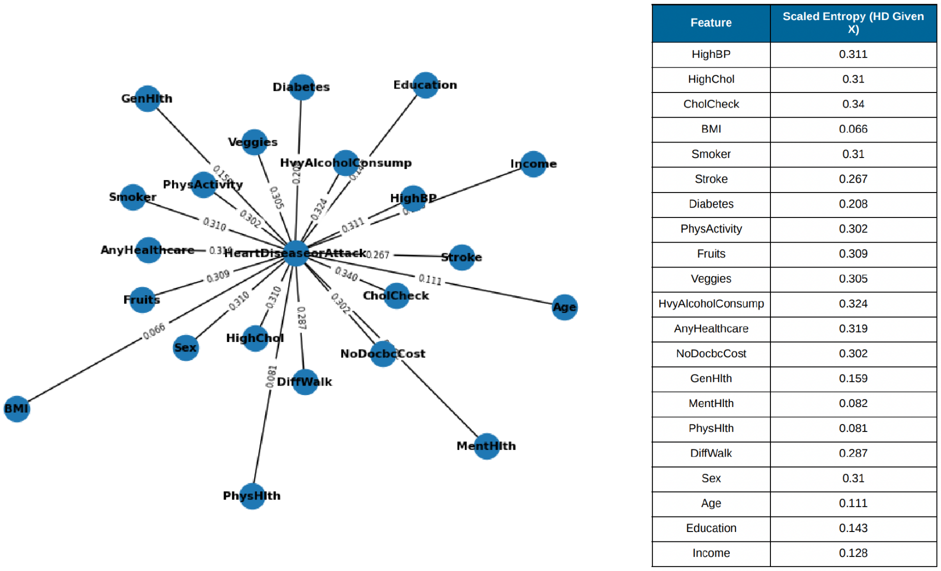
3.1 Conditonal Entropy Univariate

We calculated the conditional entropy of the 'HeartDiseaseorAttack' variable for every uni-variate combination. The purpose of the table below is to help us select the features with the highest amount of information to gain with respect to predicting a heart disease.

$$H(Y|X) = -\sum_{y\in Y}\sum_{x\in X} P(y,x)\log\frac{P(y|x)}{P(y)}$$

Here, $P(y,x)$ is the joint probability distribution of Y and X, $P(y|x)$ is the conditional probability of Y given X, and $P(y)$ is the marginal probability distribution of Y.

Luckily almost everything has a relatively low conditional entropy. This suggests that almost all of these features can comfortable predict heart disease by itself. It is also worth noting that BMI entropy is 0.066 only because BMI was a continuous variable so once we classified it the true BMI is 0.188



Feature	Scaled Entropy (HD Given X)
HighBP	0.311
HighChol	0.31
CholCheck	0.34
BMI	0.066
Smoker	0.31
Stroke	0.267
Diabetes	0.208
PhysActivity	0.302
Fruits	0.309
Veggies	0.305
HvyAlcoholConsump	0.324
AnyHealthcare	0.319
NoDocbcCost	0.302
GenHlth	0.159
MentHlth	0.082
PhysHlth	0.081
DiffWalk	0.287
Sex	0.31
Age	0.111
Education	0.143
Income	0.128

Figure 2: Table+Network Graph of Conditional Entropy of X given HeartDisease

3.2 Conditional Entropy Joint

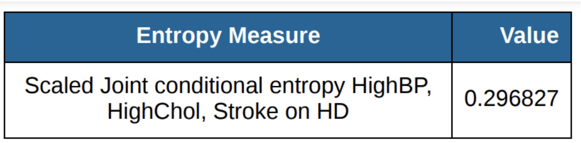
Using joint entropy, we can calculate the degree of dependency between two or more variables.

We decided to use the conditional entropy of HighBP, HighChol, Stroke on HeartDiseaseorAttack and Age, Sex, Income on HeartDiseaseorAttack.

$$H(Y|X_1, X_2, \ldots, X_n) = -\sum_{y\in Y} \sum_{x_1\in X_1} \cdots \sum_{x_n\in X_n} p(y, x_1, \ldots, x_n) \log p(y|x_1, \ldots, x_n) \tag{1}$$

where $H(Y|X_1, X_2, \ldots, X_n)$ is the joint conditional entropy of the output variable Y given input variables $X_1, X_2, \ldots, X_n$, $p(y, x_1, \ldots, x_n)$ is the joint probability distribution of $Y, X_1, X_2, \ldots, X_n$, and $p(y|x_1, \ldots, x_n)$ is the conditional probability distribution of Y given $X_1, X_2, \ldots, X_n$.

3.2.1 HighBP, HighChol, Stroke on HeartDiseaseorAttack



Entropy Measure	Value
Scaled Joint conditional entropy HighBP, HighChol, Stroke on HD	0.296827

Table 2: Joint Entropy Table 1

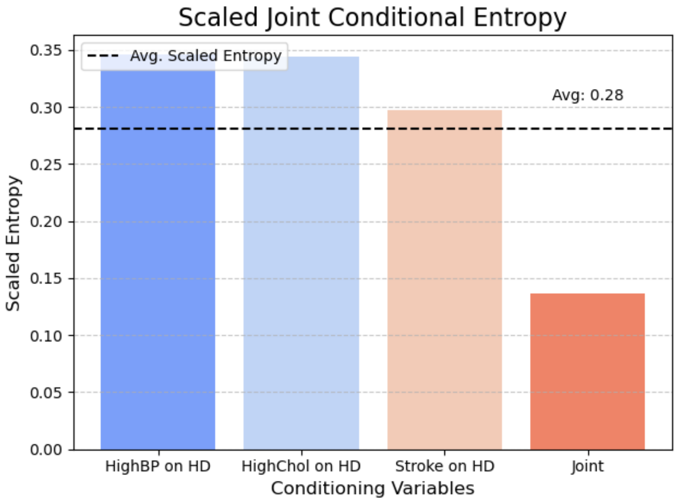
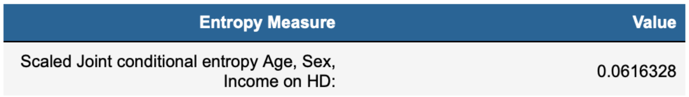


Figure 3: Joint Entropy Bar 1

From this we can see that the joint entropy has a significantly lower entropy compared to the uni-variate conditional entropy. Which means the joint conditional entropy has way higher predictability comparatively. This is expected because adding more data can only mean we add more information therefore we decrease the unpredictability of the data

3.2.2 Age, Sex, Income on HeartDiseaseorAttack



Entropy Measure	Value
Scaled Joint conditional entropy Age, Sex, Income on HD:	0.0616328

Table 3: Joint Entropy Table 2

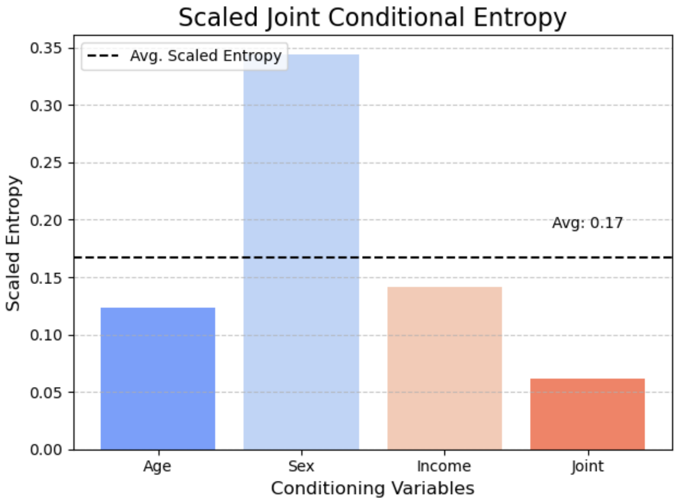


Figure 4: Joint Entropy Bar 2

It is noted that the joint entropy is much lower than the three uni-variate conditional entropy's. This suggests that the joint distribution of the two predictor variables provides more information about the target variable than either predictor variable alone.

Using the joint distribution, we can make predictions about the target variable with greater accuracy than by considering either predictor variable alone. This is because the joint distribution captures the relationship between the predictor variables and any dependence they have on each other. By considering this joint relationship, we can better understand the behavior of the target variable and make more informed predictions.

Overall, the use of joint and conditional entropy measures can provide valuable insights into the relationship between predictor variables and the target variable. By carefully analyzing these measures and their implications, we can make more accurate predictions and gain a better understanding of the underlying patterns in the data. Basically what this is saying is that we can model well, which will be discussed further later.

3.3 Sensitivity Anaylsis

Sensitivity analysis is a technique used to evaluate how sensitive a model's output is to changes in its input parameters. The main goal of sensitivity analysis is to identify which inputs have the most significant impact on the model's output, and how changes in these inputs affect the output.

In the context of sensitivity analysis, entropy can be used to identify the most important variables in a model, and to determine the range of values for these variables that have the most significant impact on the model's output. By performing sensitivity analysis using entropy, we can gain a better understanding of how our model works and how it can be improved.

$$P(Y|X_1, X_2, ..., X_n) = \exp(-\lambda H(Y|X_1, X_2, ..., X_n)) \tag{2}$$

In this formula, $P(Y|X_1, X_2, ..., X_n)$ represents the conditional probability of the output variable Y given input variables $X_1, X_2, ..., X_n$, $H(Y|X_1, X_2, ..., X_n)$ is the conditional entropy of Y given $X_1, X_2, ..., X_n$, and λ is a sensitivity parameter that controls the degree of sensitivity of the probability distribution to changes in the input variables.

3.3.1 HighBP, HighChol, Stroke on HD Sensitivity Analysis

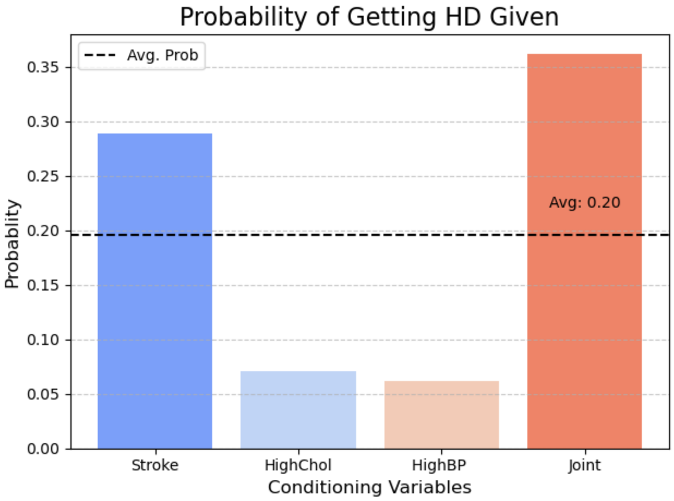
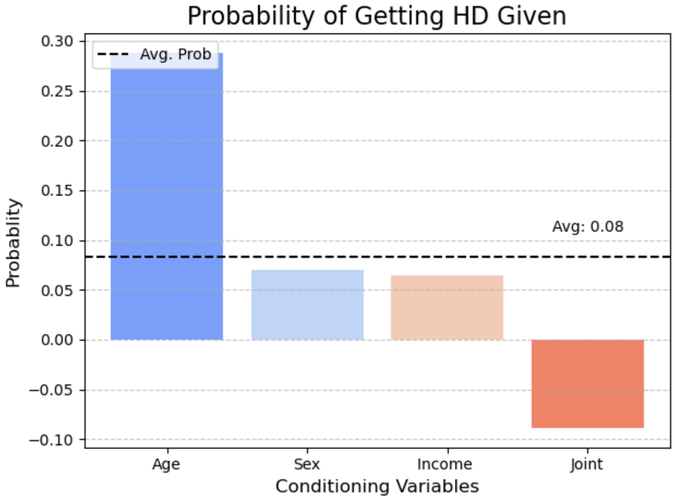


Figure 5: Prob. Of HD given Stroke,HighChol,HighBP Features

Using the method described before we were able to calculate the probability of getting a Heart Disease given if the user has a Stroke, HighChol, HighBP and all three. As we can see from the graph your likelihood to get a heart disease if you had all three is significantly higher than if you only had one of the features being 36 percent more likely if they have all. Which makes sense intuitively obviously a person is more likely to get a heart disease if they had a stroke, High Cholesterol and high BP compared to a person only having high cholesterol.

3.3.2 Age, Sex, Income Sensitivity Analysis



The relationship between age, income, and the probability of getting a heart disease is not straightforward. It has been observed that as age and income increase, the probability of getting a heart disease decreases. However, it is important to note that age and income are not binary variables, as they can take on a wide range of values. Therefore, relying solely on this probability may not provide the most accurate understanding of the relationship between age, income, and the probability of getting a heart disease.

To gain a more accurate understanding of the relationship between these variables, we can use contingency table to examine their joint distribution and calculate measures of association such as odds ratios. This approach can provide a more comprehensive picture of the relationship between age, income, and the probability of getting a heart disease. However it is also worth noting that a contigency table may be limited if it is not a 2x2 matrix because odds ratio will be super small due to the fact we are spreading our data to thin.

4 Contingency Tables

By using conditional entropy to identify the variables that yield the most information when predicting heart disease, we now have an idea of what variables could be good predictors for HeartDiseaseorAttack. We created contingency tables based on our selected variables to further explore the relationship between heart disease/heart attacks and the most important features identified from calculating the conditional entropy. We will first look at how HeartDiseaseorAttack interacts with the other variables in our dataset before evaluating any joint interactions that HeartDiseaseorAttack has with other risk factors measured during this study.

Before creating contingency tables, we reclassified the age and income variables into 2 separate categories. These categories represent whether a subject is young or old or if they are poor or rich. Age observations with a value less than or equal to 5 were classified with a "0", and any values greater than 5 were classified with a "1" to denote that the observation is old. Income observations with a value less than 5 were classified with a "0" to indicate a low income, whereas values greater than 5 were classified with a "1" to indicate a high income. Reclassifying these variables will make it easier to interpret how they affect a person's chance of developing heart disease/heart attack. Although we risk losing potentially insightful information when we reclassify these variables, that risk is mitigated by the increase in simplicity and the increase in our ability to interpret relationships between HeartDiseaseorAttack and Age and Income.

4.1 One Way Contingency Tables

The contingency tables below show the joint distribution between HeartDiseaseorAttack and the most important factors when it comes to predicting HeartDiseaseorAttack (Stroke, HighChol, HighBP, Age, Income, Sex). Below are contingency tables that explore the association between HeartDiseaseorAttack and other variables that we found to be important in identifying people at risk for heart disease/heart attacks. These contingency tables highlight how imbalanced the dataset is, as we can see many more observations of people without heart disease compared to people who do have heart disease/ have had a heart attack.

High Blood Pressure and HeartDiseaseorAttack Contingency Table			
HeartDiseaseorAttack	0 (No High Blood Pressure)	1 (Has High Blood Pressure)	Row Sum
0	138,886	90,901	229,787
1	5,965	17,928	23,893
Column Sum	144,851	108,829	253,680

Table 4: Odds Ratio: 4.592098602559366

High Cholesterol and HeartDiseaseorAttack Contingency Table			
HeartDiseaseorAttack	0 (No High Cholesterol)	1 (Has High Cholesterol)	Row Sum
0	138,949	90,838	229,787
1	7,140	16,753	23,893
Column Sum	146,089	107,591	253,680

Table 5: Odds Ratio: 3.589072560484596

Stroke and HeartDiseaseorAttack Contingency Table			
HeartDiseaseorAttack	0 (No Stroke)	1 (Had Stroke)	Row Sum
0	223,432	6,355	229,787
1	19,956	3,937	23,893
Column Sum	243,388	10,292	253,680

Table 6: Odds Ratio: 6.936202083608329

Sex and HeartDiseaseorAttack Contingency Table			
HeartDiseaseorAttack	0 (Female)	1 (Male)	Row Sum
0	131,769	98,018	229,787
1	10,205	13,688	23,893
Column Sum	141,974	111,706	253,680

Table 7: Odds Ratio: 1.8031605649849693

Age and HeartDiseaseorAttack Contingency Table			
HeartDiseaseorAttack	0 (Young)	1 (Old)	Row Sum
0	53,648	176,139	229,787
1	753	23,140	23,893
Column Sum	54,401	199,279	253,680

Table 8: Odds Ratio: 9.359796105132984

Income and HeartDiseaseorAttack Contingency Table			
HeartDiseaseorAttack	0 (Poor)	1 (Wealthy)	Row Sum
0	48,626	181,161	229,787
1	9,097	14,796	23,893
Column Sum	57,723	195,957	253,680

Table 9: Odds Ratio: 0.4365660550677109

These tables show a major imbalance in the data between people who have heart disease and people who don't have heart disease. This imbalance makes discovering associations between variables in the data difficult to do with traditional statistical tests, and using traditional methods to analyze this dataset will limit the reliability and power of the test. We will need to use other methods, such as evaluating the odds ratio, to analyze the relationship between HeartDiseaseorAttack and other variables in the dataset.

Odds Ratio Formula

Assume that the contingency table below represents the probability of hypothetical events A and B occurring.

$$\begin{array}{c|cc} & A & A^c \\ \hline B & a & b \\ B^c & c & d \end{array}$$

$$\frac{a/b}{c/d} = \frac{ad}{bc} \tag{3}$$

The odds ratio is a measure of the association between two variables by calculating the ratio of the odds that events A and B occurring simultaneously compared to the odds that neither events A or B occurred divided by the odds of events B and A occurring individually. It tells us how much more likely it is for someone with one variable to have the other variable compared to someone without it. Odds ratios greater than 1 indicate that a person's likelihood of having heart disease is greater when the independent variable is present in the subject. For example, the odds ratio of High BP and Heart Disease is 4.592, which means that individuals with high blood pressure are 4.592 times more likely to have heart disease than those with normal blood pressure. This aligns with our intuition that high blood pressure is a risk factor for heart disease. Similarly, we can see that a person's odds of having heart disease/heart attacks increases if they have high cholesterol, have had a stroke, are male, and are considered old. Conversely, a person's likelihood of developing heart disease or a heart attack decreases if the person is considered young. We can further explore the dynamics behind common factors amongst people who have heart disease when we try and predict if a person has heart disease/heart attacks by modeling the data.

4.2 Two/Three-Way Contingency Tables

In order to further explore associations between variables in the dataset, we constructed some two and three way contingency tables in the hopes of uncovering insightful messages about the data. By fusing our original predictor variable with some of the response variables in the dataset, we may be able to see if there are any confounding relationships that may increase a person's chances of having heart disease or having a heart attack. Additionally, it may be easier to see if a person has heart disease if they also have another health condition, such as high blood pressure or high cholesterol. The tables below depict how HeartDiseaseorAttack interacts with other predictor variables, such as sex, age, and a fused variable between HeartDiseaseorAttack, high blood pressure, and high cholesterol.

Income Against HeartDiseaseorAttack+Sex			
HeartDiseaseorAttack	0 (Poor)	1 (Rich)	Row Sum
(0, 0)	32,689	99,080	131,769
(0, 1)	15,937	82,081	98,018
(1, 0)	5,248	4,957	10,205
(1, 1)	3,849	9,839	13,688
Column Sum	57,723	195,957	253,680

HighChol Against HeartDiseaseorAttack/Sex			
HeartDiseaseorAttack+Age	0 (None)	1 (High)	Row Sum
(0, 0)	43,481	10,167	53,648
(0, 1)	95,468	80,671	176,139
(1, 0)	405	348	753
(1, 1)	6,735	16,405	23,140
Column Sum	146,089	107,591	253,680

HighBP Against HeartDiseaseorAttack/Age			
HeartDiseaseorAttack+Age	0 (None)	1 (High)	Row Sum
(0, 0)	44,744	8,904	53,648
(0, 1)	94,142	81,997	176,139
(1, 0)	338	415	753
(1, 1)	5,627	17,513	23,140
Column Sum	144,851	108,829	253,680

Age Against HeartDiseaseorAttack/HighBP/HighChol			
HD + HighBP + HighChol	0 (Young)	1 (Elderly)	Row Sum
(0,0,0)	37,980	61,064	99,044
(0,0,1)	6,764	33,078	39,842
(0,1,0)	5,501	34,404	39,905
(0,0,1)	3,403	47,593	50,996
(1,0,0)	252	2,624	2,876
(1,0,1)	86	3,003	3,089
(1,1,0)	153	4,111	4,264
(1,1,1)	262	13,402	13,664
Column Sum	54,401	199,279	253,680

Sex and HeartDiseaseorAttack Contingency Table			
HD+HighBP+HighChol	0 (Female)	1 (Male)	Row Sum
(0,0,0)	16,963	82,081	99,044
(0,0,1)	6,991	32,851	39,842
(0,1,0)	10,056	29,849	39,905
(0,1,1)	27,968	23,028	50,996
(1,0,0)	984	1,892	2,876
(1,0,1)	946	2,143	3,089
(1,1,0)	1,743	2,521	4,264
(1,1,1)	5,424	8,240	13,664
Column Sum	57,723	195,957	253,680

Income and HeartDiseaseorAttack Contingency Table			
HD+HighBP+HighChol	0 (Low Income)	1 (High Income)	Row Sum
(0,0,0)	16,963	82,081	99,044
(0,1,0)	10,056	29,849	39,905
(0,0,1)	6,991	32,851	39,842
(0,1,1)	14,616	36,380	50,996
(1,0,0)	984	1,892	2,876
(1,1,0)	1,743	2,521	4,264
(1,0,1)	946	2,143	3,089
(1,1,1)	5,424	8,240	13,664
Column Sum	57,723	195,957	253,680

Due to the imbalanced data, it is hard to say definitively if there are any associations in the data after fusing the response variable HeartDiseaseorAttack with other variables in the dataset, such as Sex and Age. The ratios between those who do have heart disease to those who do not have heart disease is very small in this dataset, and predicting the likelihood of a person having heart disease with another illness becomes much more difficult without further sub-setting the data, which is something that we will do in the next section when we subset the data and create models to predict heart disease/attacks.

5 Prediction Accuracy

To highlight imbalance in the data, let us first look at the data as a whole. As we wanted to use machine learning to attain understanding, let us use the entire dataset and its variables to predict heart disease. When we did this, we saw that we could attain an accuracy rating nigh 95% using logistic regression, decision tree classification, and KNN clustering. However, this was due to the incredibly skewed ratio between those with heart disease and those without. The ratio of those without heart disease compared to those with heart disease was that of approximately 10:1. When looking at this ratio, it is perfectly understandable why we would have such a high accuracy. The model has a much higher chance of observing someone without heart disease, and as such it will continuously predict that an individual does not have heart disease. In order to extract meaningful results, we will use our aforementioned subsets of the variables to look for differences between groups, whilst harnessing the imbalance.

We selected the aforementioned machine learning techniques as they could be applied to categorical data. Logistic Regression was suitable for the prediction as we already have encoded data. Decision Trees were important to us as well, as they allow us to glean into the feature importance we discussed prior. It is important to note that we had to use the gower distance matrix with KNN, as euclidean distance would not be suitable for categorical data. With an 80/20 train test split for all techniques, we applied each method to a subset of the data. Starting with the identity variables, these were the following accuracy ratings.

Identity Subsets and Precision Scores				
Subset	Ratio	LogReg	Tree	KNN
Young | Female | High Income	138:20,140	0.99	0.99	0.99
Young | Male | High Income	186:18,579	0.98	0.98	0.98
Young | Female | Low Income	150:4,297	0.97	0.97	0.97
Young | Male | Low Income	105:2,362	0.96	0.96	0.96
Elderly | Female | High Income	1540:13,008	0.89	0.89	0.89
Elderly | Male | High Income	3828:11,873	0.75	0.75	0.75
Elderly | Female | Low Income	150:4,297	0.97	0.97	0.97
Elderly | Male | Low Income	829:1,896	0.68	0.60	0.68

Table 10: Ratio = Those with HD : Those without HD

As we can see from the table, the prediction accuracy scores are nearly identical for all groups, except for the last one. The biggest reason for this is because the ratio of those with heart HD and those without HD is exceedingly skewed. What happens is that the model has an easier time predicting heart disease, as the ratio is so skewed. In other words, heart disease response is so distinguishable between the classes that the classifiers end up predicting the majority class most of the time, resulting in similar accuracy scores. This is evident as the elderly + male + low income group saw the most balanced ratio and in turn had varied precision scores. From our findings, we believe that the combination of age and gender is what gives the greatest distinction between somebody who has heart disease and doesn't. However, it is interesting to note that the elderly + high income + male group had a lower prediction score compared to that of the elderly + low income + female group. This might be due in large part to the difference in ratio.

Health Subsets and Precision Scores				
Subset (x, y, z)	Ratio	LogReg	Tree	KNN
(0, 0, 0)	2,427:91,568	0.97	0.97	0.97
(0, 0, 1)	2,708:37,804	0.93	0.93	0.93
(0, 1, 0)	3,501:37,552	0.91	0.91	0.91
(0, 1, 1)	11,100:47,413	0.81	0.81	0.81
(1, 0, 0)	376:1,033	0.70	0.70	0.70
(1, 0, 1)	345:838	0.70	0.30	0.70
(1, 1, 0)	709:1,434	0.69	0.69	0.69
(1, 1, 1)	2456:2,946	0.55	0.44	0.55

Table 11: $(x, y, z) : x = Stroke, y = HighBP, z = HighChol$

When we look at the health variables, we can notice some key things. Firstly, we want to compare two groups: the group with no ailments and the group with all ailments. We first want to call attention to the sheer imbalance between these two groups. The zero ailment group is the biggest imbalance of HD and no HD within the health subset. Hence, it has the highest prediction scores as it is easier to classify. We see variance within the all ailments group, making it so that the prediction score is less accurate. From these results, it's possible to make some inferences about the data. Stroke is a more important feature compared to the other two when predicting heart disease. This can be seen in the fact that the groups where each trait was uniquely present saw stroke have the weakest accuracy score. Due to this, we can assume that the stroke variable creates the most divisive differences. Like before, we see that in tandem variables can influence whether or not an individual has heart disease greatly. However, it's interesting to note that on average the scores for the health subset was much lower than the identity ones. This suggests that when one of these is present, it's more likely to affect the chance of an individual contracting heart disease.

6 Clustering

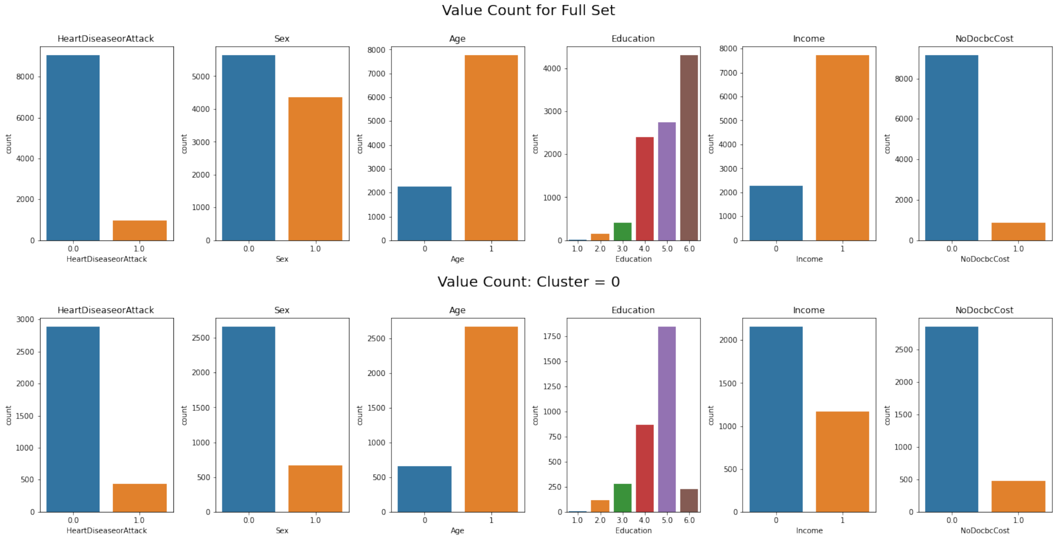
We want to be able to find a pattern in the data and compare variables among these patterns. In order to do that, we used clustering to separate the data by a pattern and compare the variables among the clusters.

First, we chose an appropriate clustering method given our data set. Due to the categorical nature of the data, we used k-modes which uses dissimilarity as a metric rather than distances, as distances are not suitable for categorical data. We then ran the k-modes algorithm on three different subsets of variables based on the types of variables. These subsets are based on demographic factors, health-related factors, and lifestyle factors. For each group of factors, we selected the number of clusters using the elbow method and examined the silhouette score, then separated the data by clusters. Having data clearly separated by clusters allowed us to examine which factors impact the presence of heart disease and how they have an impact.

Looking at the demographic info (Figure 6), we found that the clusters were most easily separated by sex and income. The first cluster captures most of lower income people in the data. While there is also a higher presence of females, the important factor in the cluster seems to be income. We find that this cluster has the highest presence of heart disease. The second cluster appears to be separated by all high-income males. This cluster has the second highest presence of heart disease. The third cluster is grouped by high-income females. This group appears to have the lowest rate of heart disease.

The clusters of health-related factors (Figure 7) appear to tell a different story. We find that one cluster has people who are generally considered to be healthier, where the average BMI is lower there is a much lower presence of high blood pressure, stroke, and type 2 diabetes. The graphs show that in the healthier cluster, there is a much lower presence of heart disease when compared to the less healthy cluster. This shows that the presence of other health problems often go together with heart disease. While these other problems may not all directly lead to heart disease, a confounding variable of worse overall health may explain the higher presence of heart disease in the cluster where people tend to have other health problems.

The subset of lifestyle factors (Figure 8) is clearly separable into clusters by people who tend to live a healthier lifestyle (better diet and exercise). When examining this, there does not appear to be clear difference in the presence of heart disease between the clusters



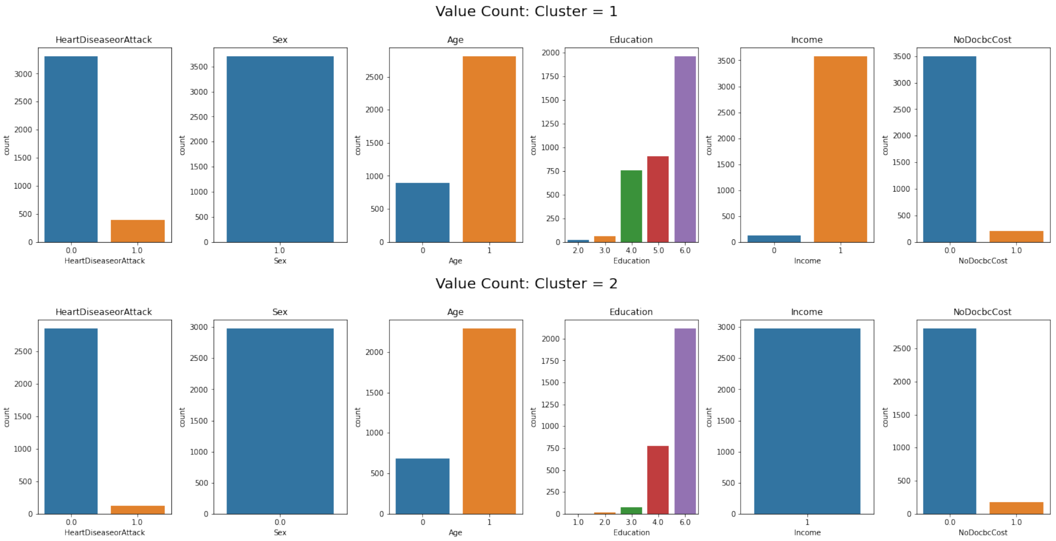


Figure 6: Value counts of demographic variables with clusters

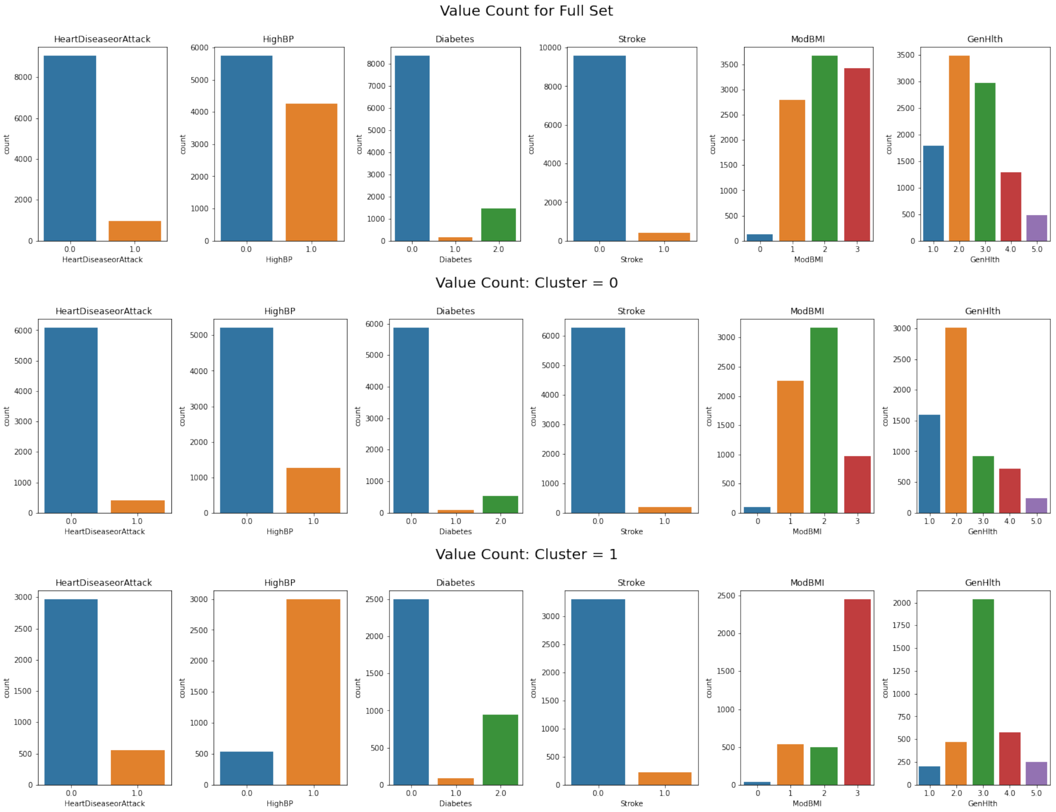


Figure 7: Value counts of health-related variables with clusters

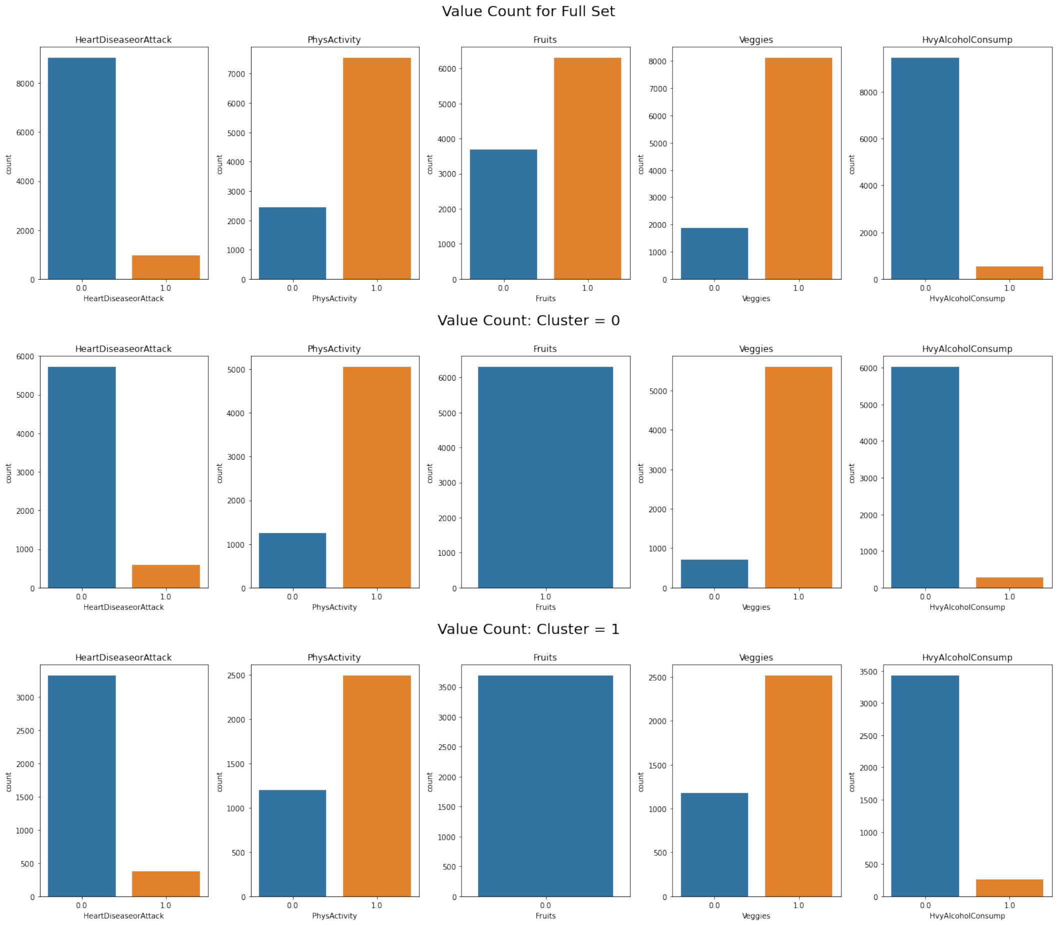


Figure 8: Value counts of lifestyle variables with clusters

These findings show that demographic and health-related variables appear to be the most important predictor for heart disease. It appears that one of the most important demographic factors for the presence of heart disease is income, as the low-income cluster has the highest presence of heart disease. As for the high-income clusters, the high-income male cluster appears to have second highest presence of heart disease whereas the high-income female cluster has the lowest presence. This shows that the next most important demographic factors is sex, where males tend to have a higher rate of heart disease. This corroborates the evidence found previously showing that low-income and males have a higher rate of heart disease. For health-related variables, the healthier cluster appears to have less heart disease. In particular, the presence of high blood pressure appears to be incredibly different, where the high blood pressure group has a greater presence of heart disease. This makes sense, as having high blood pressure forces the heart to work harder, likely leading to more disease.

7 Conclusion

Calculating conditional entropy we were able to see for certain that all variables can predict heart disease relatively well. This goes double for joint conditional entropy as expected adding more features given heart disease the conditional entropy drastically decreases. As for getting the probability of getting a heart disease depending on the features it is very straightforward if all the variables are binary like with HighBP, HighChol and Sex however like we saw with Income and Age we can not really confidently calculate the probability of getting a heart disease due to the complexity of the data.

Using contingency tables to further explore the relationship between HeartDiseaseorAttack and the other metrics collected during this survey showed how imbalanced the data was, as there was significantly fewer people who reported having some sort of heart disease or a heart attack within the dataset compared to people who reported not having heart disease or a heart attack. This issue with the imbalanced dataset was only compounded when we fused HeartDiseaseorAttack with other variables in the dataset, making it difficult to draw more accurate predictions when it comes to predicting heart diseases or heart attacks. However, despite this limitation, we can still draw some meaningful inferences into what causes heart disease/heart attacks by looking at how different factors interact with one another to increase/decrease a person's chance of heart disease. We saw a slight association between gender and pre-existing health conditions and heart disease, but we were unable to use traditional tests to examine this association due to the the imbalance in the data. The limitations behind how imbalanced the data is can be further explained by the prediction accuracy.

From the prediction accuracy, we can get a glimpse of how imbalanced some of the data is. It's important to understand that this can be harnessed to understand heart disease and who is affected by it. When we first started this project, we wanted to see the difference between identity and health. However, as we progressed further along it became more about the intersection between the two and heart disease. Overall, our findings highlight the importance of finding methods to deal with an imbalanced data set and it also increases our understanding of heart disease and how it affects the people in our sample group. Hopefully, the findings that we have uncovered by exploring this data set will be very useful in any future discussion about preventative measures for heart disease.

Demographic factors such as income and sex, as well as health-related variables such as high blood pressure, appear to be important predictors for the presence of heart disease. Low-income and male individuals are at higher risk, while healthier individuals tend to have lower rates of heart disease.

References

[1]“What’s Driving the Demand for Data Scientists?” Knowledge at Wharton, knowledge.wharton.upenn.edu/article/whats-driving-demand-data-scientist/. Accessed 14 May 2023.

[2]“Preventing Heart Disease Could Keep More People Employed and Save Billions for the Economy.” European Society of Cardiology, www.escardio.org/The-ESC/Press-Office/Press-releases/Preventing-heart-disease-could-keep-more-people-employed-and-save-billions-for-the-economy. Accessed 14 May 2023.

[3]Teboul, Alex. “Heart Disease Health Indicators Dataset.” Kaggle, 10 Mar. 2022, www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset.

[4]“Heart Disease Risk Factors for Children and Teenagers.” The Texas Heart Institute, 3 Dec. 2021, www.texasheart.org/heart-health/heart-information-center/topics/heart-disease-risk-factors-for-children-and-teenagers/.

[5]Socioeconomic Status and Risk of Cardiovascular Disease in 20 Low ..., www.thelancet.com/journals/langlo/article/PIIS2214-109X(19)30045-2/fulltext. Accessed 14 May 2023.