Analysis of Heart Disease and High Blood Pressure

Kyle Dinh
University of California, Davis
Davis, CA 95616
kyvdinh@ucdavis.edu

Jacob Kim
University of California, Davis
Davis, CA 95616
jatkim@ucdavis.edu

Trevor Adam
University of California, Davis
Davis, CA 95616
tjadam@ucdavis.edu

Marco Oviedo
University of California, Davis
Davis, CA 95616
jmoviedo@ucdavis.edu

Abstract

In this paper, we will explore data obtained from the CDC Behavioral Risk Factor Surveillance System survey from 2015. We will first analyze the covariates of our data set to determine which contributes the most towards our selected response variable Heart Disease. This will be done by calculating and comparing the conditional entropy of the covariates with the response variable. In addition, we will also analyze the conditional entropy of the covariates with the fused response variable of Heart Disease and High Blood Pressure.

1 Introduction

According to the CDC, Heart Disease is the leading cause of death for men and women across most racial and ethnic groups in the United States, contributing to the death of 697,000 people annually [3]. The medical term Heart Disease is not a specific medical condition, but it is used as an umbrella term for several other heart conditions such as Coronary Artery Disease, Arrhythmia, Heart Failure, and Heart Attacks [1]. In this report, we will use data from the CDC's Behavioral Risk Factor Surveillance System survey from 2015 that was obtained from Kaggle to explore the relationship between Heart Disease and other covariates. This will be done by calculating and comparing the conditional entropies of the covariates with the response variables of Heart Disease and the fused variable of Heart Disease and High Blood Pressure. We will also find the conditional entropy of the response variables given the covariate + Age + Sex. Age was included in the calculations for the conditional entropies because it is a characteristic of the respondent that can not be changed. Additionally including Sex will allow us to get a deeper insight into potential Age and Sex-related findings in the diagnoses of heart disease. Our findings would help us determine which covariate has the strongest association with Heart Disease or Heart Disease fused with High Blood Pressure and in turn which covariates could be considered a possible risk factor or indicator for the response.

2 Data

The Behavioral Risk Factor Surveillance System (BRFSS) was developed by the CDC to gather health-related data by telephone survey across all 50 states each year. The data collected consists of chronic health conditions and risk behaviors that could affect physical and mental health [2]. The data set consists of 22 categorical variables and contains 253,680 observations. Table 1 below provides all the variables that will be looked at as well as their description.

Table 1: Data Set Variables and Description

Variable	Description
Heart Disease or Attack	Respondent has been diagnosed with Heart Disease
High Blood Pressure	Respondent has been diagnosed with High Blood Pressure
High Cholesterol	Respondent has been told that their blood cholesterol is high
Cholesterol Check	Respondent has had their blood cholesterol checked
BMI	Respondent's body mass index. Separated into four categories
Smoker	Respondent has smoked 100 cigarettes in their life
Stroke	Respondent has been diagnosed with a stroke
Diabetes	Respondent has been diagnosed with diabetes
Physical Activity	Respondent reported doing physical exercise in the last 30 days
Fruits	Respondent consumes fruit at least once a day
Vegetables	Respondent consumes vegetables at least once a day
Heavy Alcohol Consumption	Respondent has more than (Male: 14)(Female: 7) drinks per week
Health Care	Respondent has any kind of health care
NoDocBcCost	Respondent was not able to see a doctor because of cost.
General Health	Respondent's assessment of their general health
Mental Health	Days the respondent's mental health was not good in the last 30 days
Physical Health	Days the respondent's physical health was not good in the last 30 days
Difficulty Walking	Respondent has difficulty walking or going up stairs
Sex	Respondent's sex
Age	Respondent's age
Education	Respondent's highest level of education completed
Income	Respondent's annual household income

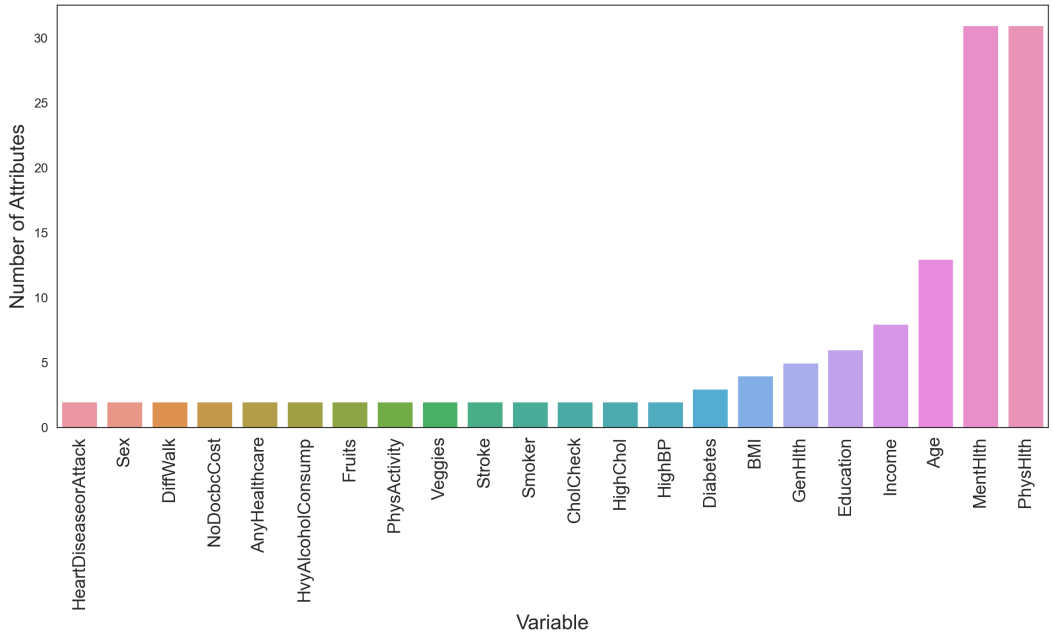

Figure 1: Categorical Variable Attributes

The majority of the variables in the data consist of binary values. All binary variables excluding sex have zero if the response was no and one if the response was yes. Education is made up of six attributes with one being "Never attended School" and increasing the level of education to six: "College Grad". General Health has five attributes ranging from one being "Excellent" and five being "Poor". Mental and Physical Health variables have thirty-one attributes for the number of days the respondent was not in good health. BMI has four attributes with one being "underweight" to four being "Obese". The Sex variable is binary with values of zero corresponding to Female and values of one corresponding to Male. The Age variable is categorical with one representing ages 18-24, two representing ages 25 to 29, three representing ages 30 to 34, and so on for five-year increments up to age 80 or older which is represented by 13. Another binary variable, HighBP (High Blood Pressure), has values of zero corresponding to no high blood pressure and values of one representing individuals with high blood pressure.

Count plots will be used to visualize the distribution of Heart Disease for the categorical variables. Due to the number of variables, it will be broken into two plots (Figure 2 and Figure 3).

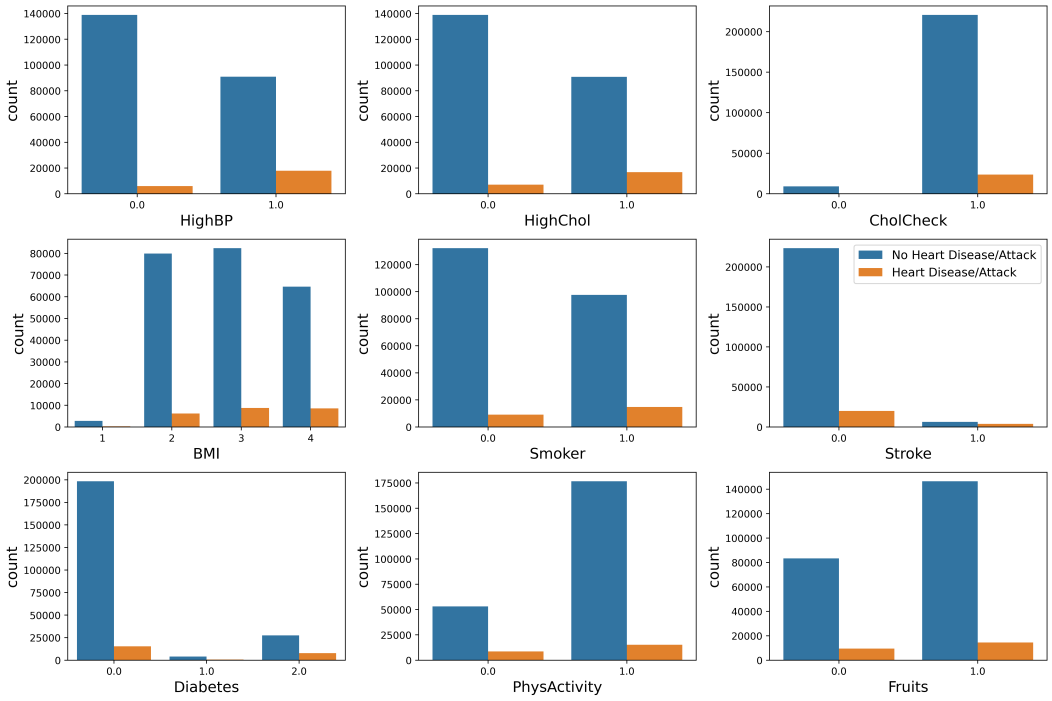

Figure 2: Heart Disease Distribution

For Figure 2, the number of respondents that responded positively for Heart Disease increased among those who also responded positively for the following variables: HighBP, HighChol, Cholcheck, Smoker, PhysActivity, and Fruits. For the variables of Stroke and Diabetes, the number of respondents decreased for being positive in both. We can also see that the number of respondents with heart disease increased with BMI.

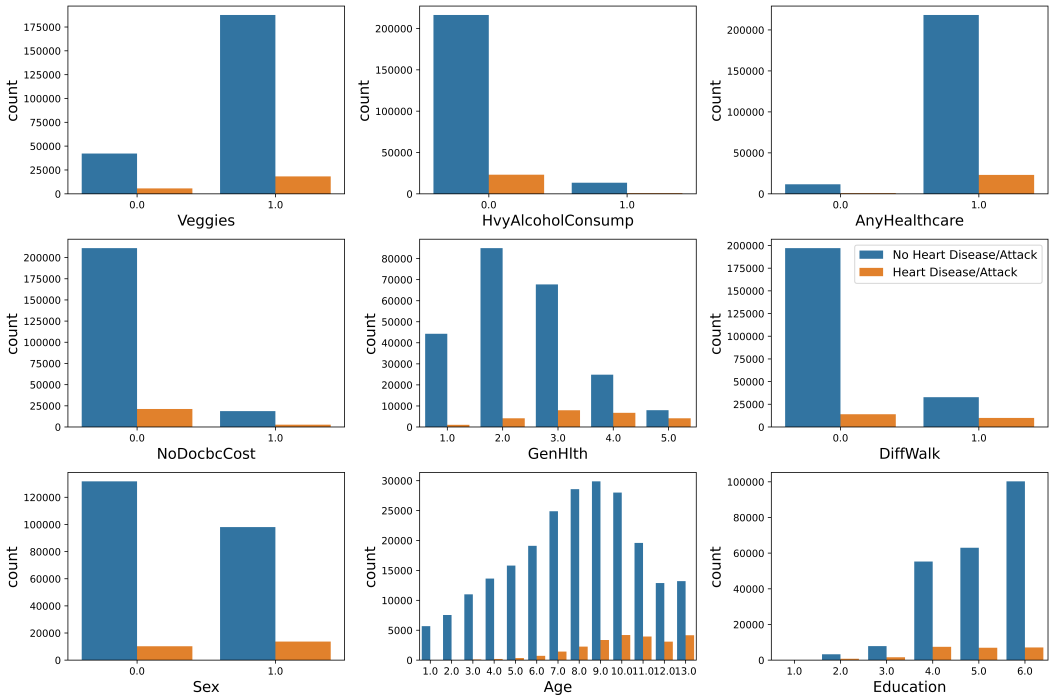

Figure 3: Heart Disease Distribution

For Figure 3, the number of respondents that responded positively for Heart Disease increased among those who also responded positively for the following variables: Veggies, and AnyHealthcare. In addition, the number of respondents with Heart Disease increased with Age and Education Level. Lastly, for the variables: HvyAlcoholConsump, NoDocbcCost, and DiffWalk the number of respondents with Heart Disease decreased when responding positively to the variable. Due to the large number of aspects with Mental and Physical Health, count plots will not be shown. However, they will still be presented and discussed in the report using alternative visuals in the Analysis section.

3 Methods

As shown above, most of the variables are categorical variables or have discrete values. Thus, our analysis will mainly consist of a categorical response variable and categorical covariates. An effective way to quantify the importance of the covariates with respect to the response variable is to use conditional entropy.

Entropy is a quantitative measure of a variable's uncertainty. We can calculate entropy as:

$$CE[Y] = -\sum_{i=1}^{n} P(Y_i) log(P(Y_i))$$

Similarly, conditional entropy is a quantitative measure of a variable's uncertainty when conditioned with another covariate. We can calculate conditional entropy as:

$$CE[Y|X] = -\sum_{i=1}^{n} P(Y|X = x_i) log(P(Y|X = x_i))$$

An important fact about conditional entropy is that the conditional entropy of a response given a covariate will always be less than or equal to the entropy of the response. In other words,

$$CE[Y|X] \leq CE[Y]$$

This implies that, given the covariate, the uncertainty of the response variable decreases.

For this report, we can determine which covariates are important for the response variable based on which conditional entropies are the smallest when considering the response variable of Heart Disease and the fused response variable of Heart Disease and High Blood Pressure. To find the probabilities, we will construct contingency tables of marginal probabilities.

4 Analysis

The following table illustrates the conditional entropy of heart disease given each covariate.

Table 2: Conditional Entropy given each Covariates

Heart Disease	0.312116
Variable	CE[HD | Variable]
GenHlth	0.279970
Age	0.283664
HighBP	0.289980
DiffWalk	0.293992
HighChol	0.295831
PhysHlth	0.297741
Diabetes	0.299066
Stroke	0.299094
Income	0.301964
Smoker	0.305601
Education	0.307281
Sex	0.308441
PhysActivity	0.308593
MentHlth	0.309818
BMI	0.310202
CholCheck	0.310818
Veggies	0.311388
HvyAlcoholConsump	0.311642
NoDocbcCost	0.311671
Fruits	0.311922
AnyHealthcare	0.311926

Looking at Table 2 the variable General Health (GenHlth) has the lowest conditional entropy. The General Health variable was a personal assessment of the respondent of how they viewed their own health. So it is possible that respondents with heart disease viewed themselves as having poorer health. This will be looked into with more detail later on in the paper.

Because it is useful to see variable interactions with sex and age, the conditional entropy of heart disease given each covariate, sex, and age is shown below.

Table 3: Conditional Entropy given Covariate + Sex + Age

Heart Disease	0.312116
Variable	CE[HD | Variable + Sex + Age]
GenHlth	0.250942
PhysHlth	0.263751
DiffWalk	0.264699
HighBP	0.268064
Stroke	0.268990
Income	0.269018
HighChol	0.270058
Diabetes	0.270701
MentHlth	0.271942
Education	0.274215
Smoker	0.275078
BMI	0.275728
PhysActivity	0.276055
NoDocbcCost	0.276181
Veggies	0.278191
Fruits	0.278262
CholCheck	0.278286
HvyAlcoholConsump	0.278469
AnyHealthcare	0.278640

Table 3 shows that including Sex and Age lowers the conditional entropy values for all variables. An interesting observation is that the three lowest conditional entropy values come from variables (GenHlth, PhysHlth, DiffWalk) that were personal assessments of the respondents rather than the variables that were medical diagnoses.

The conditional entropy tells us which variables provide information on heart disease but we still want to see how these variables affect the risk of heart disease. To do so, we can observe the conditional probabilities in our data. Without imposing any conditions, $\mathrm{P}(\mathrm{HeartDiseaseorAttack}) = 0.0942$. From the calculated conditional entropies, we know age provides a large amount of information on heart disease and there is a possible interaction of sex and age. We can find the conditional probabilities P(HD | Sex + Age) which are shown in the plot below.

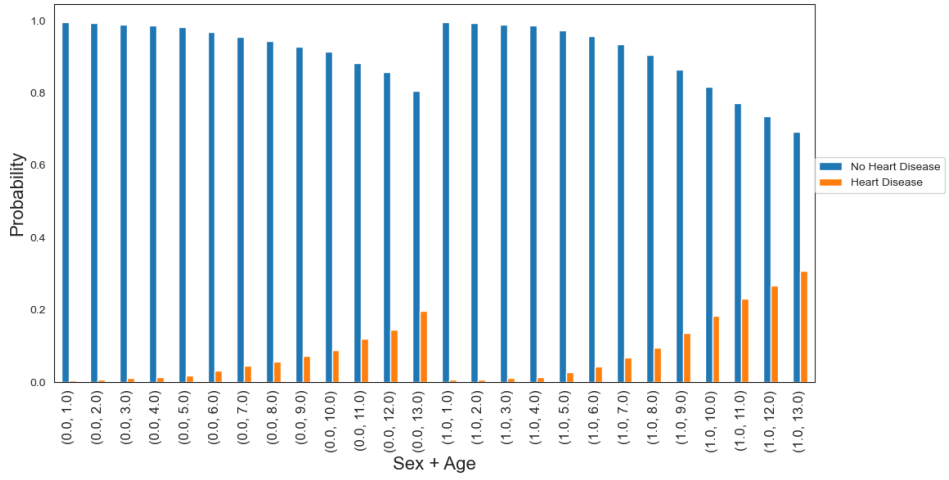

Figure 4: Conditional Probabilities of Heart Disease Given Sex + Age

It is clear that heart disease or heart attacks become more common as age increases. Through age groups 1 through 4, males and females experience close to an equal amount of heart disease. In the

rest of the age groups, men experience more heart disease than women. In our data, the probability of men having heart disease in age categories 9-13 is around double that of women.

Because general health is an important indicator, we will look at this variable more closely by looking at a heatmap.

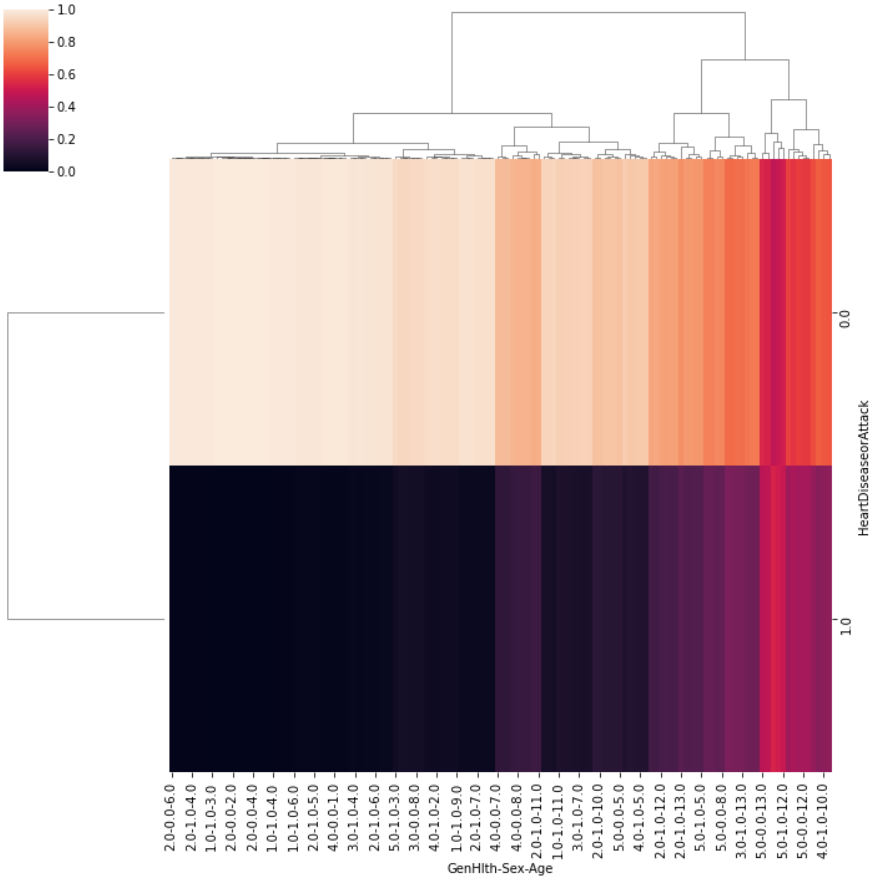

Figure 5: Heatmap of Heart Disease vs General Health + Sex + Age

The bottom right section of the plot in Figure 5 shows that people with a high general health score and high age group are more likely for heart disease where the highest probability of heart attack occurs in people with the worst general health (5), male (1), and of the highest age group (13). Interestingly there is a cluster in the middle bottom of the plot in Figure 5 which clusters females with bad general health (4) who are in the (7) and (8) age groups. In this cluster, there is also the group which consists of males with good general health (2) who are in the (11) age group. We can see that even in higher age groups, those with good general health have low levels of heart disease. In Figure 4, we saw that the probability of a male in the age group (13) having heart disease was around 30%. When general health is (1), the probability of heart disease is half of that and around two-thirds of that when general health is (2). Then, the upper categories of general health (4 and 5) see a 40-75% increase in heart disease compared to the average male in this age group (P(HD | Sex=Male + Age=13)). This trend of the probability of heart disease being lower than the average for general health (1) and (2) but much higher for groups (4) and (5) holds for all groups. Typically, general health group (3) is within a few percentage points of the probability when general health is not accounted for(P(HD | Sex + Age)).

Now we will look at the Physical Health covariate more closely.

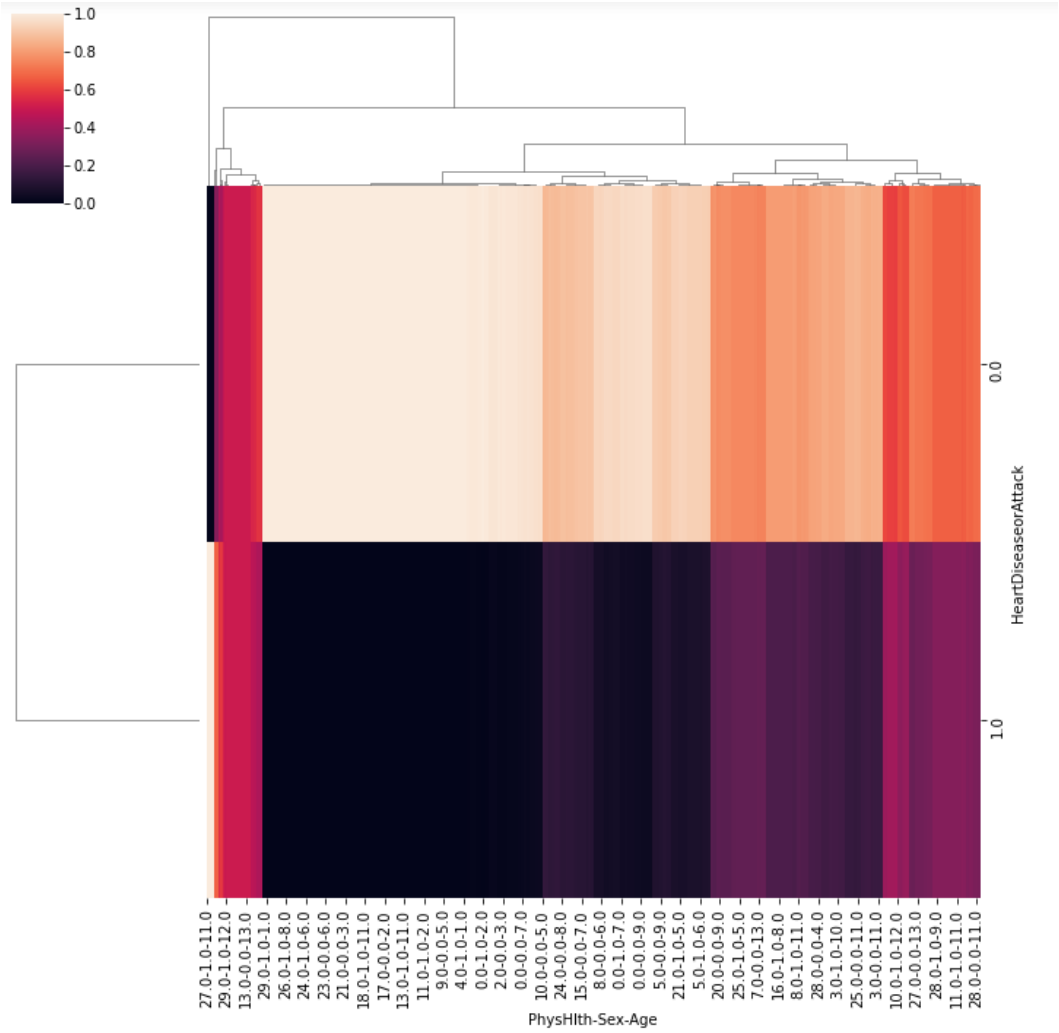

Figure 6: Heatmap of Heart Disease vs Physical Health + Sex + Age

One interesting thing to note is the cluster which consists of one group (27 days of bad physical health, male (1), and age group (11)) in the very bottom left of the heatmap of Figure 6. This group has a 100% probability of having heart disease. To be more specific, there were 0 instances of people in this group who did not have heart disease or attack and there were 2 instances of heart disease or attack.

The cluster just to the right of the cluster mentioned in the bottom left corner of the heatmap of Figure 6 shows that all of the people in this cluster are of high age groups (11), (12), and (13). Furthermore, most of the people in this cluster are males; however, there are few instances of females. The number of days of bad physical health were generally high in this cluster.

The bottom right part of the heatmap shows a large cluster which consists of a lot of groups. For example, it consists of people with very large amounts of days with bad physical health but of lower age groups and people with high age groups but do not have many days of bad physical health. This group also has many instances of male and female groups.

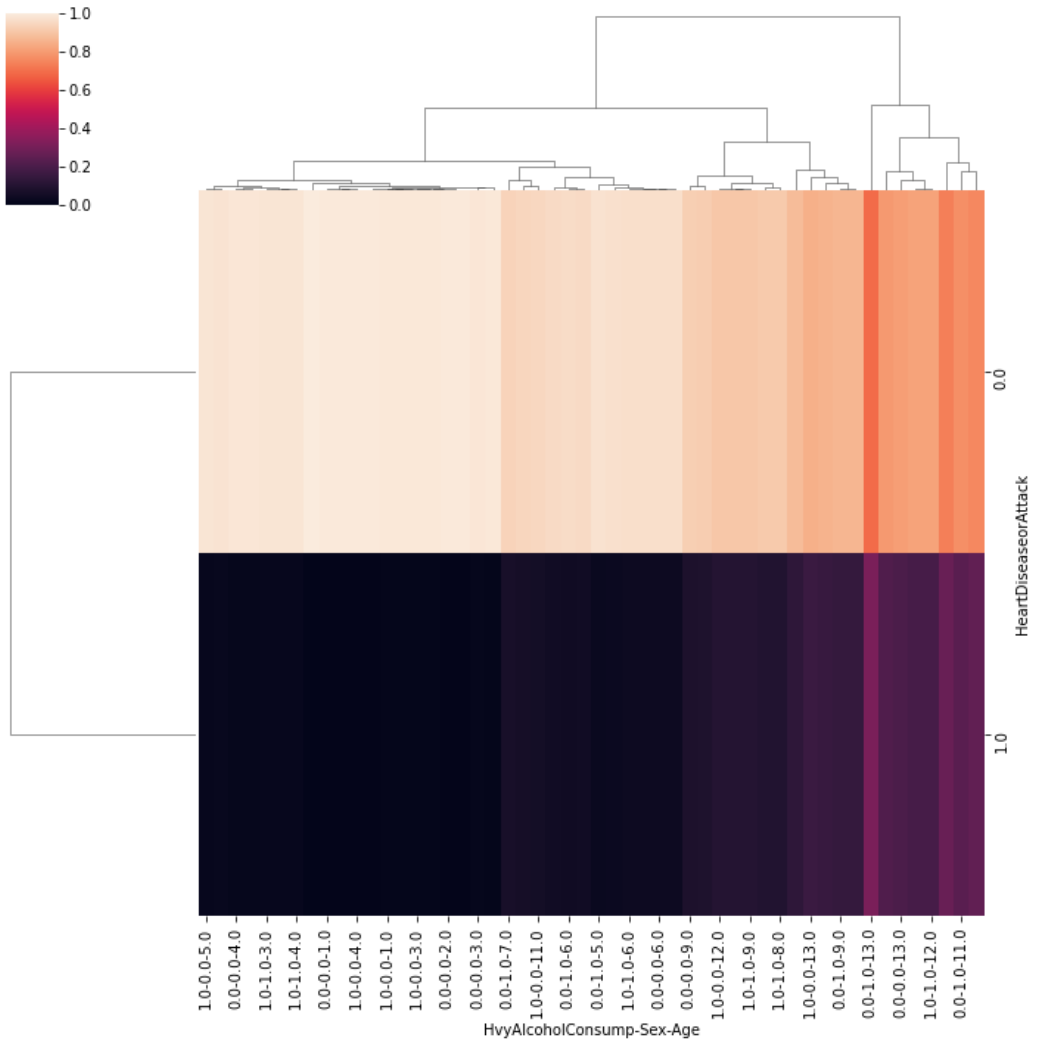

Figure 7: Heatmap of Heart Disease vs Heavy Alcohol Consumption + Sex + Age

Although heavy alcohol consumption is not a good predictor of heart disease, there is an interesting finding when looking at Figure 7. In particular, the highest probability of heart disease occurs in the group who does not heavily consumes alcohol, male, and the highest age group. This is contradictory to the known fact that heavy alcohol consumption and being older in age increase the risk of heart disease or attack [4].

Now we will consider the fused response variable of Heart Disease and High Blood Pressure. Alike the previous section, we can calculate the conditional entropy of this response variable given each of the covariates. The table below illustrates these conditional entropies.

Table 4: Conditional Entropy of given each Covariate

Heart Disease + High BP	0.973010
Variable	CE[HD + BP | Variable]
Age	0.890922
GenHlth	0.904106
HighChol	0.920298
Diabetes	0.929013
DiffWalk	0.936714
BMI	0.946115
PhysHlth	0.948685
Income	0.950711
Stroke	0.955020
Education	0.960125
PhysActivity	0.963329
Smoker	0.963511
CholCheck	0.967115
MentHlth	0.968426
Sex	0.968536
Veggies	0.970765
Fruits	0.972117
AnyHealthcare	0.972182
NoDocbcCost	0.972501
HvyAlcoholConsump	0.972534

Interestingly, Table 4 shows that age is the variable that is most useful in predicting the fused variable of heart disease and high blood pressure. This differs from Table 2 as the covariate that led to the lowest conditional entropy in the heart disease case was general health.

Furthermore, the second most useful variable for predicting heart disease (illustrated in Table 2) was PhysHlth; however, when we consider the fused variable of heart disease and high blood pressure, we find that PhysHlth is only the 7th most useful variable.

Similar to the previous section, we will consider the conditional entropy of heart disease and high blood pressure given the covariate, sex, and age. The table below shows the conditional entropies.

Table 5: Conditional Entropy of given Covariate + Sex + Age

Heart Disease + High BP	0.973010
Variable	CE[HD + BP | Variable + Sex + Age]
GenHlth	0.811572
PhysHlth	0.836685
BMI	0.847340
MentHlth	0.847544
Diabetes	0.849125
HighChol	0.851010
Income	0.852596
DiffWalk	0.854862
Education	0.867576
Stroke	0.870329
PhysActivity	0.872550
Smoker	0.875947
NoDocbcCost	0.877258
CholCheck	0.877861
Veggies	0.878724
Fruits	0.879116
HvyAlcoholConsump	0.879452
AnyHealthcare	0.882397

We see that GenHlth is the most useful predictor when combined with sex and age, which is the same when considering the single response variable of heart disease. When considering sex and age, PhysHlth and BMI become very good predictors; however, when considering PhysHlth and BMI individually, they aren't the best predictors, which may indicate some interaction effects.

From all of the tables constructed, it is interesting to note that AnyHealthcare is one of the worst variables in this dataset at predicting any of the response variables that we considered.

Because GenHlth is the most important predictor when combined with sex and age for predicting heart disease or attack and high blood pressure, we will look at this in more detail.

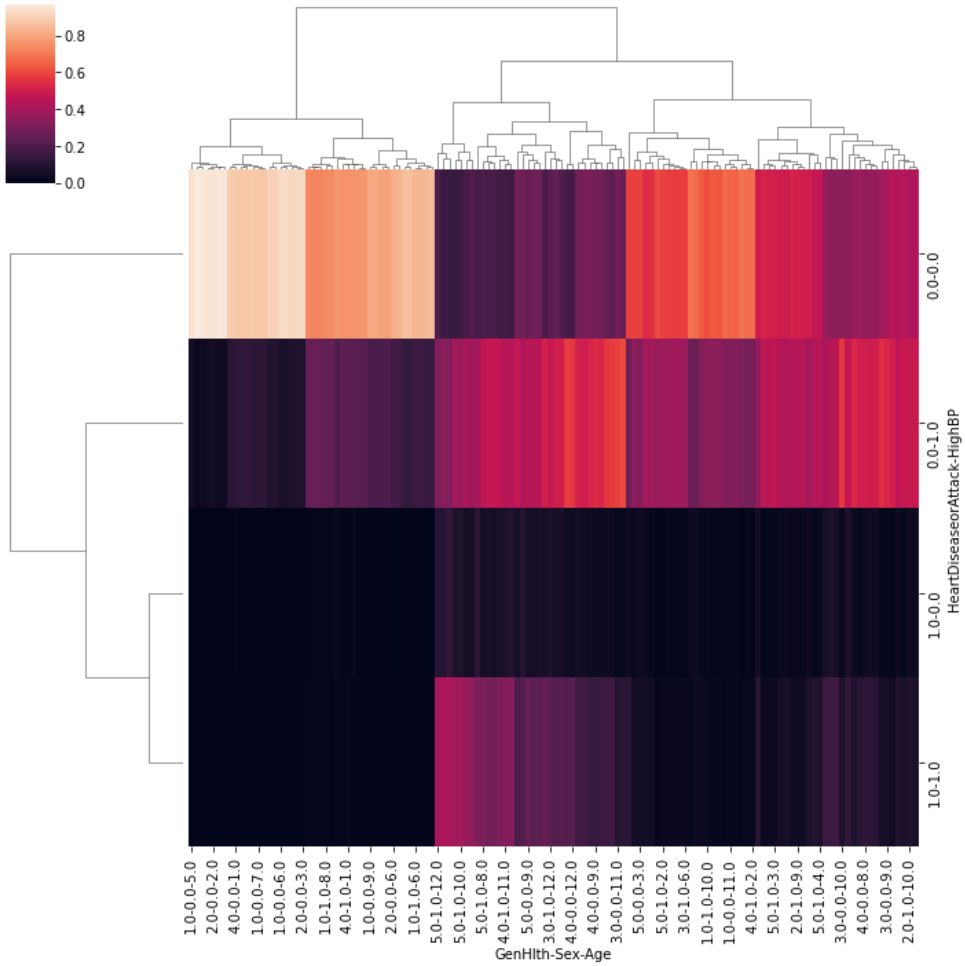

Figure 8: Heatmap of Heart Disease + High Blood Pressure vs General Health + Sex + Age

The fourth row in Figure 8 shows a cluster that is light purple that consists of high general health scores (4) and (5) and high age groups with the highest probability of heart disease or attack and high blood pressure belonging to the group who has the worst general health (5), male (1), and the highest age group (13). This is the same as the univariate response case of heart disease or attack when conditioned on the same covariates.

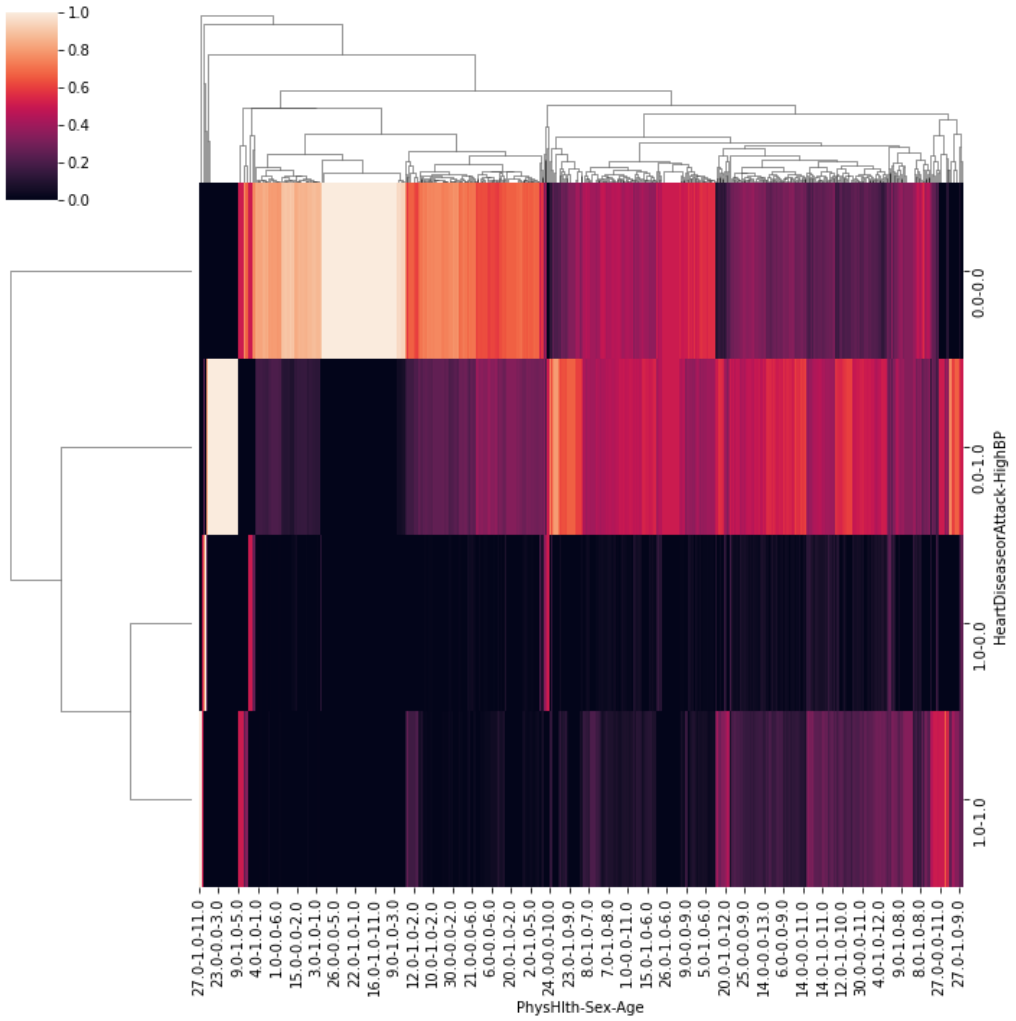

Figure 9: Heatmap of Heart Disease + High Blood Pressure vs Physical Health + Sex + Age

When predicting heart disease or attack and high blood pressure, we find a small cluster in the bottom right corner of the plot of Figure 9 which consists of a large number of days with bad physical health and high age groups. This is similar in the univariate case of heart disease as many days of bad physical health and high age groups led to an increased chance of heart disease or attack.

Comparing this Figure 9 to Figure 8, we see a relatively large probability of predicting heart disease but not high blood pressure in Figure 9 (indicated by the red lines in the third row of Figure 9). This is interesting because Figure 8 shows that it is hard to predict just heart disease and not high blood pressure as the row is very dark in color.

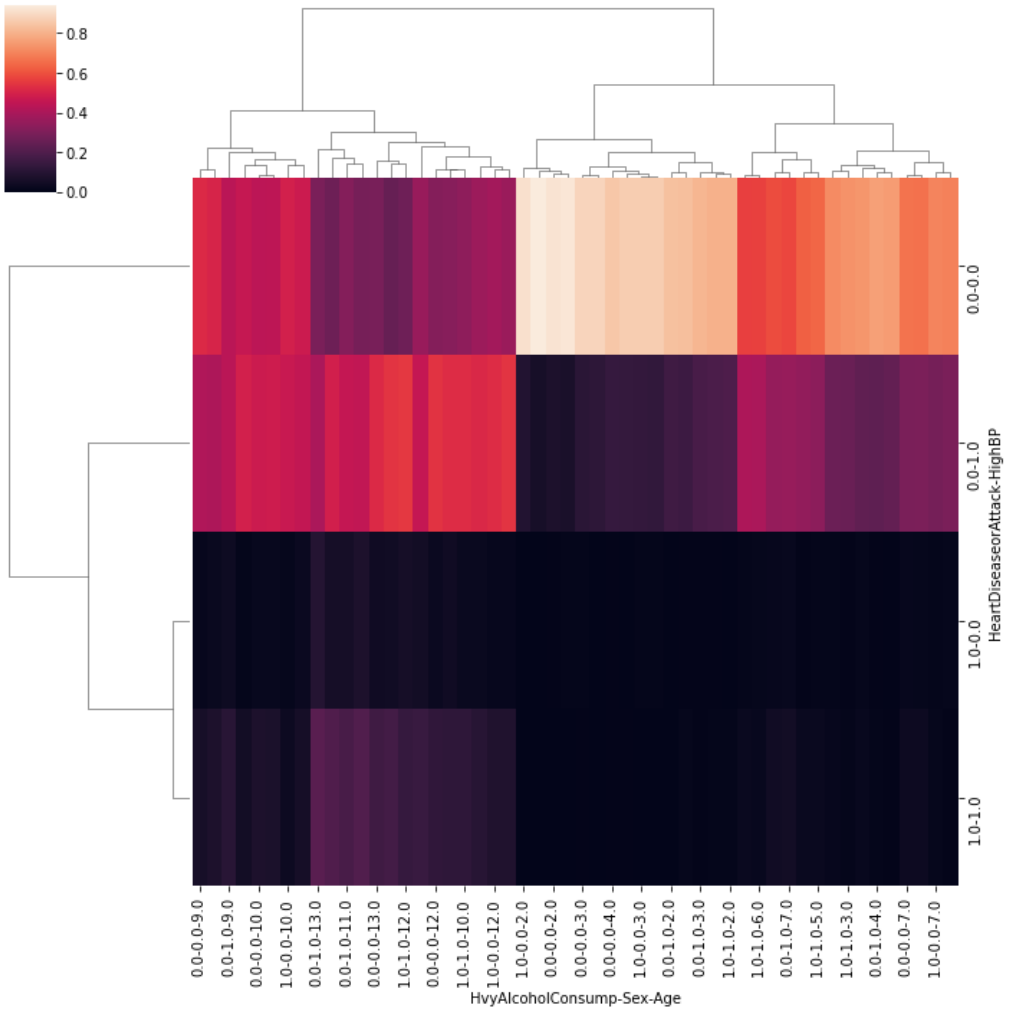

Figure 10: Heatmap of Heart Disease + High Blood Pressure vs Heavy Alcohol Consumption + Sex + Age

Alike in the heart disease case, we find that the largest probability of heart disease and high blood pressure occurs in the group with no heavy alcohol consumption, male, and age group (13).

When we predict no heart disease or attack but high blood pressure, we find that the cluster in Figure 10 with the highest probability occurs generally in people with high age groups (the orange part of the second row). The groups in this cluster contains many groups of males and females as well as heavy alcohol consumers and non heavy alcohol consumers.

From the plots above, it seems like predicting only heart disease or attack and no high blood pressure seems very unlikely.

5 Discussion

As stated before, conditional entropy can loosely translate to predictive capability when regarding exploratory and response variables. Generally speaking, if the conditional entropy of the response differs significantly from the conditional entropy of the response given a particular covariate, then we can say that the covariate would be useful in predicting our response variable in a model setting. All of this is keeping in mind that the conditional entropy of the response will always be greater than the conditional entropy of the response given a covariate. In the context of our findings, the probability of entropy of our response, Heart Disease, is greater than the probability of entropy of our response given a covariate GenHlth (General Health). This may be an indicator of GenHlth is useful in predicting Heart Disease. Also, we found that the probability of entropy of Heart Disease is almost the same as the probability of entropy of Heart Disease given Fruits. This is an example of a covariate like Fruits not being so useful in predicting Heart Disease. From our initial findings, we found that it was beneficial to fuse covariates Age and Sex along with other covariates to determine the conditional entropies. From the conditional probabilities of Heart Disease given Sex and Age, we found that in general Heart Disease is more likely to occur in Men than Women. Additionally, the probability of Heart Disease increases as an individual ages. This aligns with the CDC's findings of Heart Disease between males and females [3] and aligns with our general understanding that Heart Disease would be more common in individuals with older ages due to their bodies not being able to maintain the same level of metabolism. Referring back to fusing the covariates with Sex and Age, the probability of conditional entropy with PhysHlth (Physical Health) sees a decrease. When using PhysHlth alone in conditional probability, its entropy is not very good. However, combined with Age and Sex, the entropy decreases and shows that PhysHlth helps with predicting Heart Disease. Furthermore, this could mean that there is some kind of interaction effect between PhysHlth and Age/Sex. It would be worthwhile to investigate the covariates that yield better entropy by building a statistical model that would predict Heart Disease/Attack if the goal of this paper was prediction. It is also noteworthy to notice that HvyAlcoholConsump (Heavy Alcohol Consumption) has conditional entropy that is almost the same as the entropy of Heart Disease alone. This implies, for this set of data, that heavy alcohol consumption is not informative of someone having Heart Disease or not. This finding could be of interest considering that cardiovascular disease is attributed to bad diet and exercise habits.

Our findings from our response being Heart Disease/Attack can be informative in telling us which covariates are deterministic of a person contracting heart disease or not. In the case of this data set, there are some covariates that could help us further understand heart disease if we were to be combined as the response variable. In our study, we decided to move forward with combining HighBP (High Blood Pressure) with Heart Disease for our response and comparing the entropies of the remaining covariates. What we discovered is that the fused response variable sees a decrease in entropy for HighChol (High Cholesterol). However, this entropy seems to increase when fusing Sex and Age with HighChol. This could be interpreted as saying that high levels of cholesterol can play a role in the chances of high blood pressure or heart disease, but not so much when considering age and sex. HvyAlchoholConsump remains to be a poor predictor of heart disease even when adding the response variable of HighBP. This further shows that heavy alcohol consumption does not attribute to Heart Disease. Something that is worth noting is that in all cases, GenHlth (or General Health) remains to be one of the best covariates with one of the lowest entropies. This would lead us to further explore this covariate as it seems that when using Heart Disease as a response, regardless of fusing with other variables, has high predictive capabilities.

To further evaluate our findings, we created some heatmaps of our response variable, the fusion of Heart Disease and High Blood Pressure, along with some fusion of covariates. From the GenHlth heatmap, we were able to conclude that older Men who rated themselves poorly through the General Health variable were the most susceptible to Heart Disease. This makes sense with our current knowledge that Men and old age are big factors in Heart Disease. With the heatmap, we are able to further deduce that people's assessment of their health can also be indicative of Heart Disease. From our other heatmaps, we are able to say that poor physical health can lead to Heart Disease or Attacks, and heavy drinking does not have an impact.

6 Conclusion

From our findings and analysis, we have explored the covariates of this dataset and how it pertains to Heart Disease. Through the use of conditional entropy and the data provided to us through Kaggle via the CDC, we were able to better understand Heart Disease/Attacks. We can infer that Heart Disease risk is increased depending on one's sex or age. For example, if someone is older in age, then they are at higher risk. Additionally, Men are more at risk than Women are in general. If someone is not physically active, they are again at higher risk of HD. Also, if one responds to a health survey by admitting their quality of general health is poor, their blood pressure may be higher, and more at risk of heart disease. On the other side, if someone was to be a heavy drinker, there is no higher risk of heart disease. Our findings about Sex and Age are coherent with the findings of the CDC concerning heart disease and attacks. Knowing this, our analysis with the Kaggle data set allows us further insight into how Heart Disease and Blood Pressure are affected by various factors. If we were to build a model, we could use the information about the entropy of the variables to help choose our predictors. It should be noted that from our findings that it would make sense to subset the data as creating a model to encapsulate the whole data frame would be difficult and cumbersome.

References

[1] Centers for Disease Control and Prevention. *About Heart Disease*. https://www.cdc.gov/heartdisease/about.htm. Accessed: May 7, 2023. 2021.

[2] Centers for Disease Control and Prevention. *About the Behavioral Risk Factor Surveillance System*. https://www.cdc.gov/brfss/about/index.htm. Accessed: May 11, 2023. 2021.

[3] Centers for Disease Control and Prevention. *Heart Disease Facts*. https://www.cdc.gov/heartdisease/facts.htm. Accessed: May 7, 2023. 2021.

[4] Centers for Disease Control and Prevention. *Know Your Risk for Heart Disease*. https://www.cdc.gov/heartdisease/risk_factors.htm. Accessed: May 11, 2023. 2021.