Relationship between Heart Disease and Health Indicators

Shraddha Jhingan **Tsubasa Lin**

Curtis Pan **Hazel Yu**

Abstract

This project examines a dataset from Kaggle which contains information regarding the occurrence of heart diseases and heart attacks of respondents to the Behavioral Risk Factor Surveillance System (BRFSS) survey administered in 2015. The dataset contains a binary classification of heart disease, with the value of 1 indicating the occurrence of heart disease, and 0 indicating the absence of heart disease. Along with the heart disease classification variable, there are other variables for each observation such as health statistics of each respondent, and also behavioral variables which are used in our exploratory data analysis. For example, we looked at variables such as BMI, age groups, alcohol consumption, cholesterol, stroke incidence, diabetes, and many others to find a relationship between the occurrence of heart disease and relevant factors. To begin our analysis, we looked into the mutual conditional entropy of the variables, particularly the conditional entropy with the chronic disease of Heart Disease or Attack as our response, and a fused situation with heart disease and stroke combined as one response. As heart disease is the leading cause of death in the United States, it will be beneficial to understand what factors have the greatest association with heart disease overall.

- The lowest mutual conditional entropy and thus most highly associated variables to heart disease is stroke, high blood pressure, difficulty walking, and high cholesterol, meaning the presence of these factors are more associated with heart disease and the information we can gain from studying heart disease given we know someone has at least one of these associated factors.
- When looking at heart disease and stroke as one response more, the general health, age, education, and income were the factors highly associated with the risk of heart disease.
- From the contingency tables, we found the odds ratios where it illustrated that individuals with high blood pressure are 4.59 more susceptible to heart disease than those without high blood pressure. The same could be said for those who have had a stroke, with 6.93 times more susceptible to heart disease than those who have not had a stroke. Those with difficulty walking also were 4.26 times more susceptible to heart disease than those without difficulty walking.

1 Background (Data)

Heart disease is a widespread health condition in America, and affects many Americans every year while proving to be very costly for the country. Heart disease is the leading cause of death in the United States, taking the lives of approximately 647,000 individuals every year. There are many factors to consider when determining the cause of heart disease. Larger coronary arteries, changes due to aging, chronic inflammation, high blood pressure, and diabetes can all contribute to the onset and risk of heart disease.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

The prevalence and impact of heart disease highlights the importance of preventative measures, research, and tests that can help predict the occurrence of heart disease and heart attacks. The Centers for Disease Control and Prevention identified three primary factors that cause heart disease: smoking, high blood pressure, and high cholesterol. In addition, the National Heart, Lung, and Blood Institute stated that other factors such as age, occupation, genetics, lifestyle habits, sex, and other variables contribute to heart disease. This particular research project takes a deep dive to find relationships between variables from the dataset and the occurrence of heart disease. We used exploratory data analysis in Python and R to achieve our goals.

The dataset for this project is the ‘Heart Disease Health Indicators Dataset’ and is from the 2015 BRFSS. It is an extremely large dataset, with 22 variables (columns) and 253,680 rows. Each row pertains to a survey response from the 2015 BRFSS. The variables include the binary classification variables ‘HeartDiseaseorAttack’, ‘HighBP’, ‘HighChol’, ‘CholCheck’, ‘Smoker’, ‘Stroke’, ‘Diabetes’, and more. ‘HeartDiseaseorAttack’ is our dependent variable for the project, with the value of 1 indicating the occurrence of heart disease, and 0 indicating the absence of heart disease. There are continuous variables as well, such as ‘MentHlth’ and ‘PhysHlth’. ‘Age’ is the only variable that is sorted in buckets of 5-year intervals. All variables are float variables, and there were no null or missing values.

2 Methodology

We used Python in Jupyter Notebook and R, while collaborating through a GitHub repository. We used a variety of packages such as infotheo in R, and NumPy, Pandas, Seaborn, MatPlotLib, Sci-kit Learn, and SciPy in Python. For the basic visualizations to explore the data, we used Python. To find the odds ratio, we used Python as well. For conditional entropy, we used R and Python. We also used the seaborn package in Python to create a heatmap. We also found the conditional entropies for fused variables using Python. To fuse the variables prior to calculating the conditional entropies, we created contingency tables for the fused variables, which we then used for the calculations.

Since the Heart Disease dataset contained many categorical variables and some continuous variables, we transformed certain variables, such as BMI, into a categorical variable with four levels in accordance with CDC guidelines. [1] This would also allow us to perform calculations of conditional entropy on all the now categorical variables of this dataset.

3 Analysis and Results

3.1 Visualizations

To begin the analysis, we first started with creating visualizations of the data to look for interesting patterns. We created histograms of the variables that were not binary, such as BMI and Age, to visualize the distribution.

Using the seaborn package in Python, we also decided to look at the distribution of every variable using boxplots. We then removed outliers by using a for loop and by calculating the interquartile range of each variable in the dataset. Then we plotted the boxplots for the dataset with removed outliers as well.

[1] Center for Disease Control and Prevention

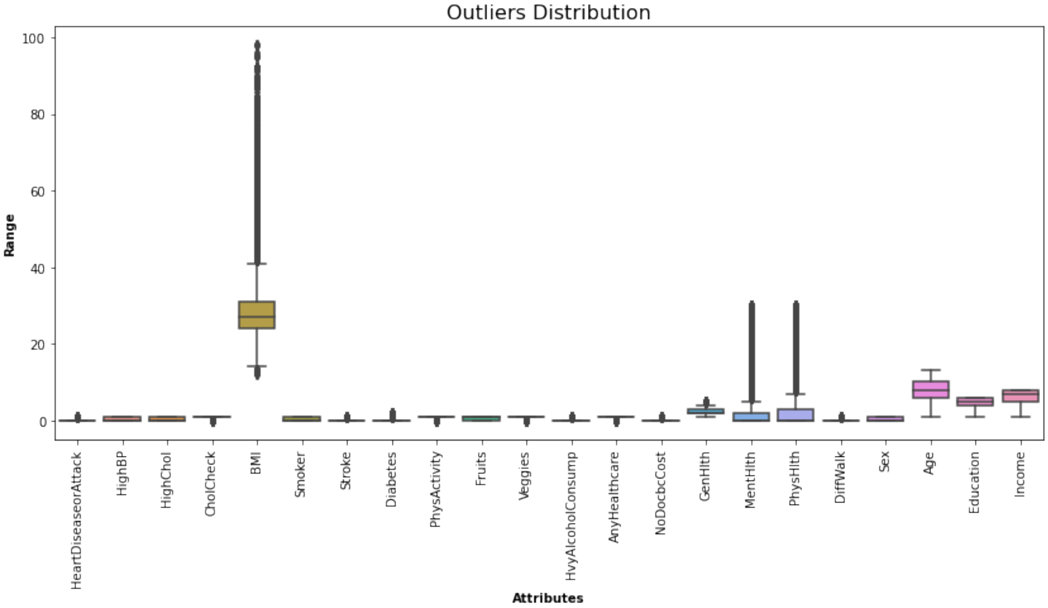


Figure 1: Boxplot with outliers

The first graph of boxplots shows that most of the variables in the dataset have similar distributions, presumably because many of the variables are binary variables or take numerically low values. The three variables worth noting with outliers are Body Mass Index ('BMI'), Mental Health Rating ('MentHlth'), and Physical Health Rating ('PhysHlth'), because they all have data points beyond their upper whiskers. The BMI variable has a higher overall distribution than all of the other variables, however this is because BMI values are nonzero and range from 15.0 and upwards.

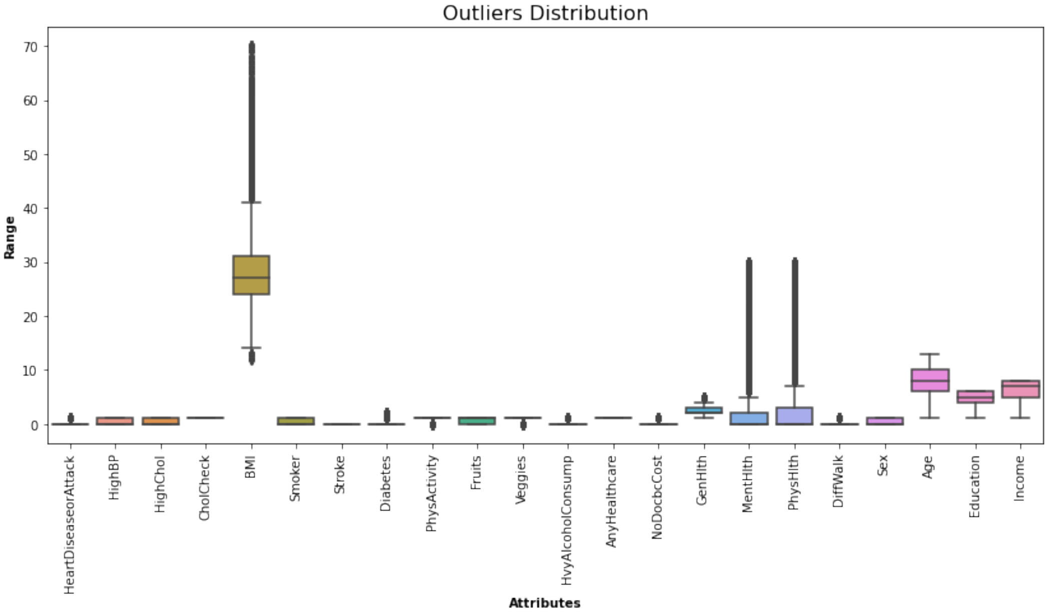


Figure 2: Boxplots without outliers

The second graph of boxplots shows boxplots that look very similar to the first graph, however this time outliers are removed. The only apparent visual difference is that the y-axis range changes from a scale of 0 to 100 to a scale of 0 to 70, and the BMI boxplot that had outliers beyond the value of 70 no longer has values past 70.

3.2 Conditional Entropy

We started off by taking a random subset with 60% of the data to balance and mitigate the inherent bias of the data set. We also converted BMI into a categorical variable following the CDC's categories:

Table 1: BMI Categories

BMI	Weight Status
below 18.5	Underweight
18.5 - 24.9	Healthy Weight
25.0 - 29.9	Overweight
30.0 and Above	Obesity

We created a new column 'BMI_cat' in our data which has values 0 for Underweight, 1 for Healthy Weight, 2 for Overweight, and 3 for Obesity.

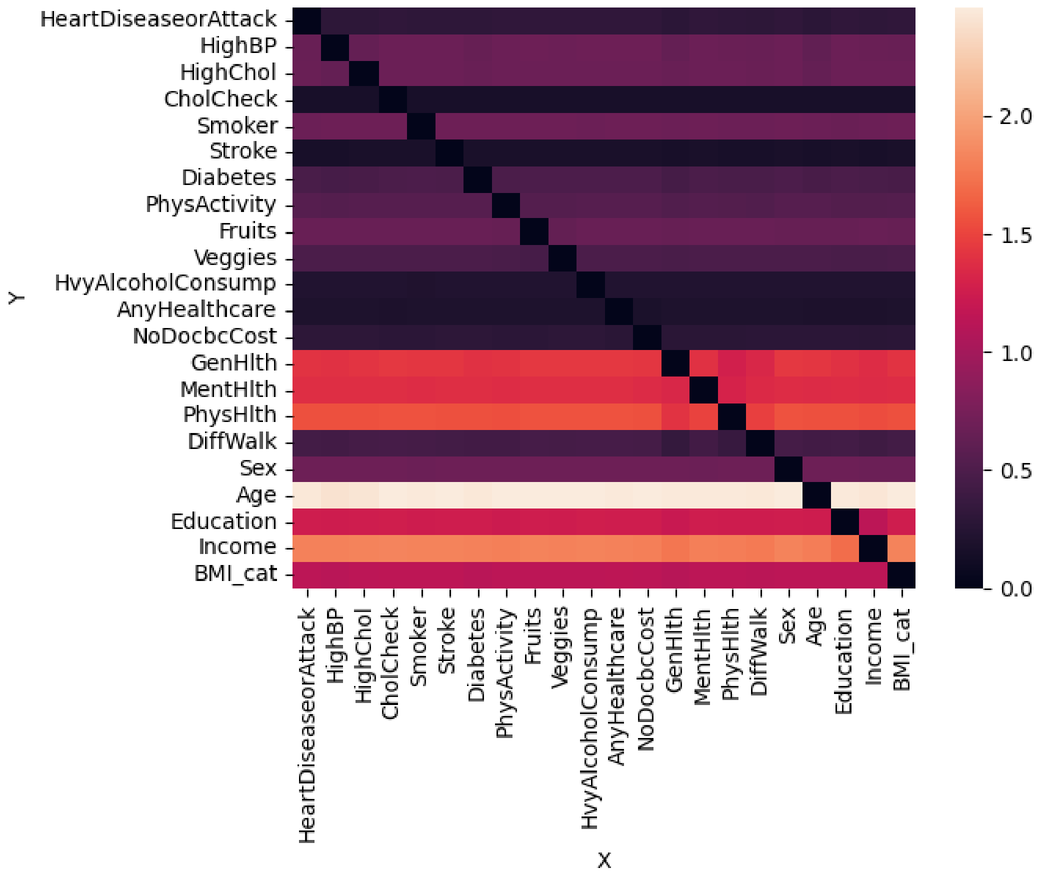


Figure 3: Conditional Entropy H(Y|X)

We found the conditional entropies of each variable given each of the other variables to create a heatmap for visualization. Conditional entropy of a variable Y given a variable X, H(Y|X), is the uncertainty about Y when X is known. Lower conditional entropy indicates more dependence between the two variables, while higher conditional entropy indicates more uncertainty and independence. Therefore, we want to look at variables that have a lower conditional entropy with the variable of interest. Looking more closely at heart disease, we found that the conditional entropy for heart disease given all independent variables were around the same. The five variables with the lowest conditional entropy were General Health ('GenHlth'), Age, High Blood Pressure ('HighBP'), Difficulty Walking ('DiffWalk'), and High Cholesterol ('HighChol').

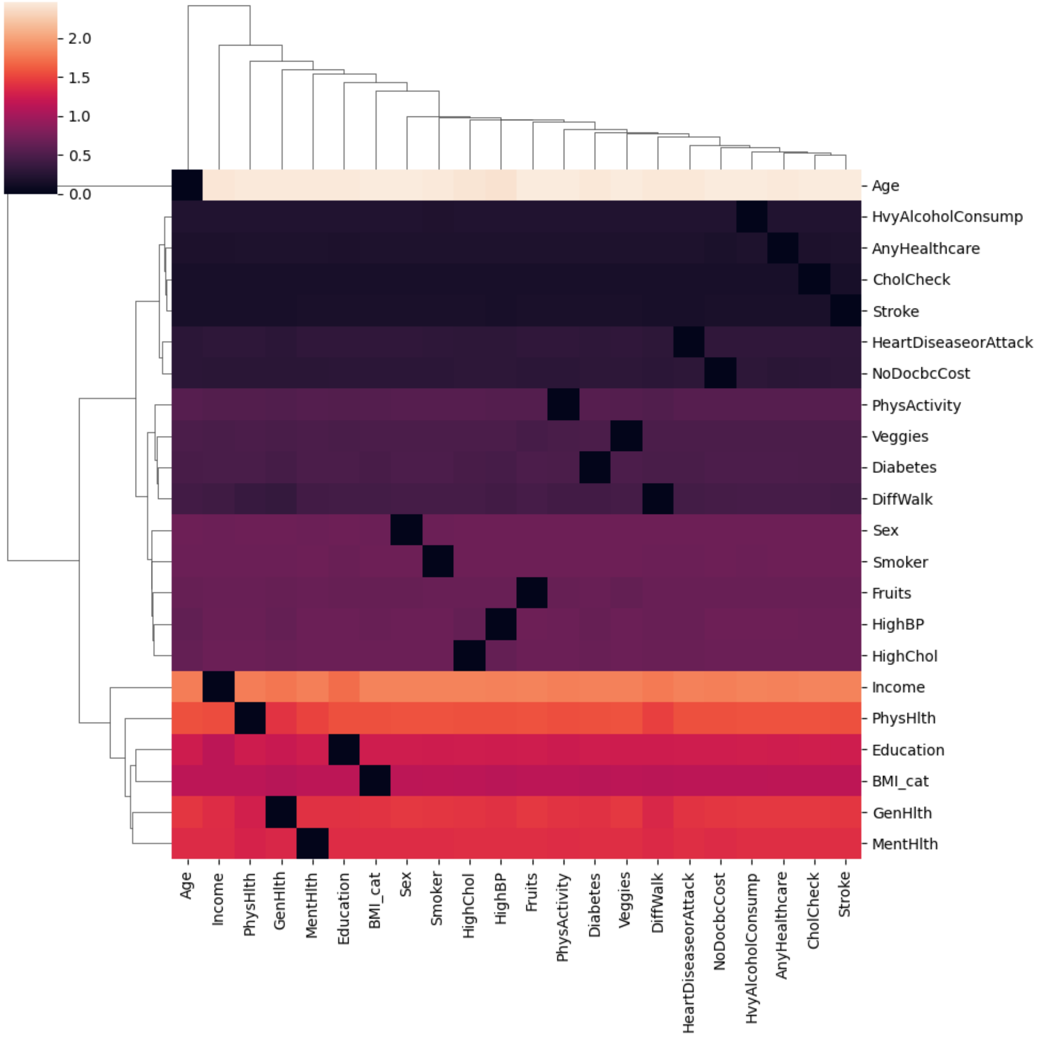


Figure 4: Conditional Entropy Dendrogram

From the Figure above we can see the synergistic features clustered together along with their conditional entropies on the response of heart disease. These synergistic feature groups such as economic and physical factors: income, education, general health, and mental health, all have a combined effort in affecting heart disease risk.

After finding the previous conditional entropies, we also found conditional entropies for the fused variables. We started off by fusing the variable 'heart.HeartDiseaseorAttack' with the 'Stroke' variable. We then found the conditional entropy between this fused variable and the variable 'GenHlth', which was found to be 1.77856531.

As mental health also plays an equally important role in a person's overall wellbeing, we also found the conditional entropy for the fused variable of 'heart.HeartDiseaseorAttack' and 'Stroke' with the variable 'MentHealth.' The value of the conditional entropy was 3.46417028.

In addition to physical and mental health, there are also societal factors that can affect the likelihood of someone being affected by heart disease, such as education and income. We found the conditional entropy for the fused variables of 'heart.HeartDiseaseorAttack' and 'Stroke' with the variable 'Education' to be 1.94399212. With the same fused variable and the 'Income' variable, the conditional entropy was 2.19513562.

From the conditional entropy heatmap earlier, we found that the variables that had the lowest entropy with 'HeartDiseaseorAttack' were 'GenHlth', 'Age', 'HighBP' and 'DiffWalk.' Thus, we decided to also find the conditional entropies with these additional variables for the fused variable 'HeartDiseaseorAttack' and 'Stroke'. The results for all the variables found thus far are summarized in the table below.

Table 2: Non-fused variable conditional entropy values

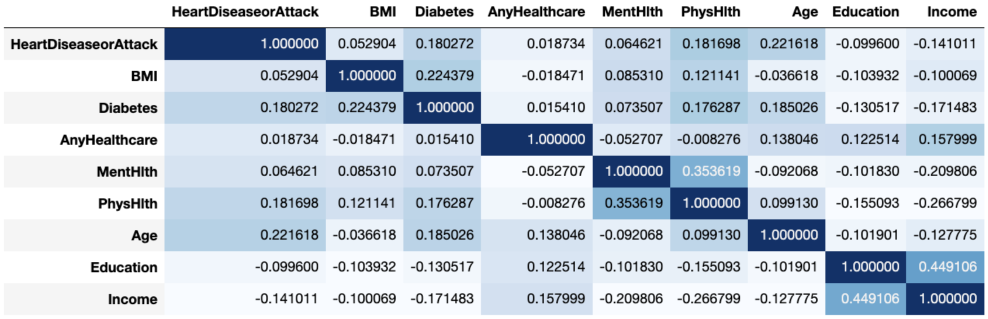
X	H(HeartDiseaseorAttack|X)
GenHlth	0.27997
Age	0.283664
HighBP	0.28998
DiffWalk	0.293992
HighChol	0.296831

Table 3: Fused variable conditional entropy values

X	H(HeartDiseaseorAttack+Stroke|X)
GenHlth	1.77856531
MentHlth	3.46417028
Education	1.94399212
Income	2.19513562
Age	2.63392152
HighBP	1.09609572
DiffWalk	1.09233492

Looking at the table, we can see that the variable that had the lowest conditional entropy with the fused variable 'heart.HeartDiseaseorAttack' and 'Stroke' was DiffWalk, followed by 'HighBP' and 'GenHlth.'

4 Correlation Analysis for Non Categorical Variables



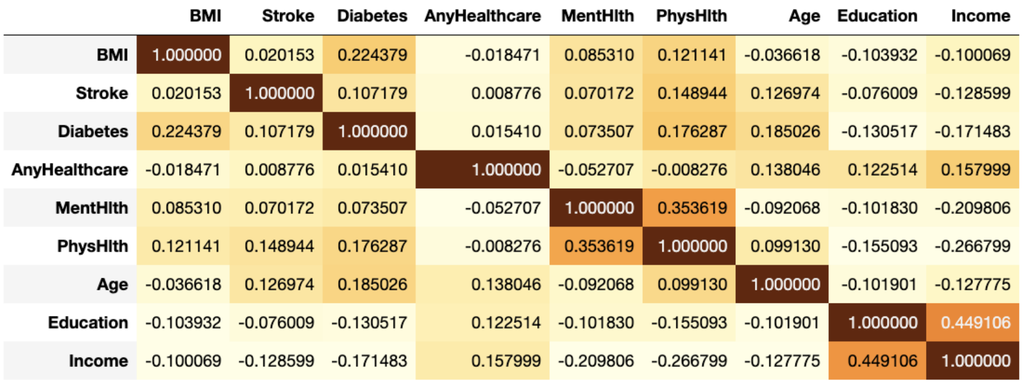
	HeartDiseaseorAttack	BMI	Diabetes	AnyHealthcare	MentHlth	PhysHlth	Age	Education	Income
HeartDiseaseorAttack	1.000000	0.052904	0.180272	0.018734	0.064621	0.181698	0.221618	-0.099600	-0.141011
BMI	0.052904	1.000000	0.224379	-0.018471	0.085310	0.121141	-0.036618	-0.103932	-0.100069
Diabetes	0.180272	0.224379	1.000000	0.015410	0.073507	0.176287	0.185026	-0.130517	-0.171483
AnyHealthcare	0.018734	-0.018471	0.015410	1.000000	-0.052707	-0.008276	0.138046	0.122514	0.157999
MentHlth	0.064621	0.085310	0.073507	-0.052707	1.000000	0.353619	-0.092068	-0.101830	-0.209806
PhysHlth	0.181698	0.121141	0.176287	-0.008276	0.353619	1.000000	0.099130	-0.155093	-0.266799
Age	0.221618	-0.036618	0.185026	0.138046	-0.092068	0.099130	1.000000	-0.101901	-0.127775
Education	-0.099600	-0.103932	-0.130517	0.122514	-0.101830	-0.155093	-0.101901	1.000000	0.449106
Income	-0.141011	-0.100069	-0.171483	0.157999	-0.209806	-0.266799	-0.127775	0.449106	1.000000

Figure 5: Correlation Visualization for Heart Disease

In order to investigate which of the non categorical variables were correlated with heart disease, we subsetted the dataset to only include the non categorical variables. From this, we created a visualization to see which of the variables had the highest correlation with 'HeartDiseaseorAttack.'

From Figure 4, we can see that the variables that had the highest correlation with heart disease was Age, followed by PhysHlth and Diabetes. The values of these correlations were 0.221618, 0.181698 and 0.180272 respectively.

We also created a correlation visualization for the Stroke response variable, to see whether the variables that had the biggest correlation with heart disease also had a strong correlation with strokes.



	BMI	Stroke	Diabetes	AnyHealthcare	MentHlth	PhysHlth	Age	Education	Income
BMI	1.000000	0.020153	0.224379	-0.018471	0.085310	0.121141	-0.036618	-0.103932	-0.100069
Stroke	0.020153	1.000000	0.107179	0.008776	0.070172	0.148944	0.126974	-0.076009	-0.128599
Diabetes	0.224379	0.107179	1.000000	0.015410	0.073507	0.176287	0.185026	-0.130517	-0.171483
AnyHealthcare	-0.018471	0.008776	0.015410	1.000000	-0.052707	-0.008276	0.138046	0.122514	0.157999
MentHlth	0.085310	0.070172	0.073507	-0.052707	1.000000	0.353619	-0.092068	-0.101830	-0.209806
PhysHlth	0.121141	0.148944	0.176287	-0.008276	0.353619	1.000000	0.099130	-0.155093	-0.266799
Age	-0.036618	0.126974	0.185026	0.138046	-0.092068	0.099130	1.000000	-0.101901	-0.127775
Education	-0.103932	-0.076009	-0.130517	0.122514	-0.101830	-0.155093	-0.101901	1.000000	0.449106
Income	-0.100069	-0.128599	-0.171483	0.157999	-0.209806	-0.266799	-0.127775	0.449106	1.000000

Figure 6: Correlation Visualization for Stroke

From the correlation visualization above, we can see that the top three variables that had the highest correlation with the stroke response variable were PhysHlth, Age and Diabetes. The values of the correlations for these were 0.148944, 0.126974 and 0.107179 respectively. Thus, physical health, age and diabetes have a significant effect on the likelihood of an individual getting heart disease or strokes, for both cases.

4.1 Odds Ratio

The odds ratio measures the association between two events, A and B. It is the ratio of the odds of A when B is present and the odds of A when B is absent. If the odds ratio is equal to 1, it means that the events are independent. If the odds ratio is greater than 1, then A and B are associated and the presence of B increases the odds of A compared to the absence of B. If the odds ratio is less than 1, then A and B are negatively correlated. The presence of B decreases the odds of A.

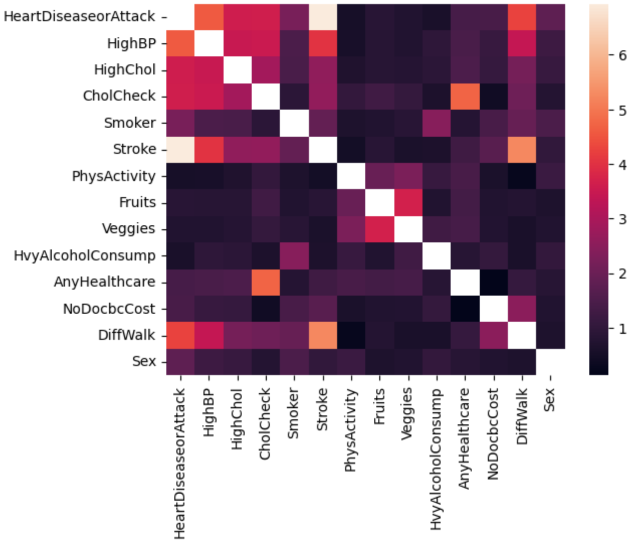
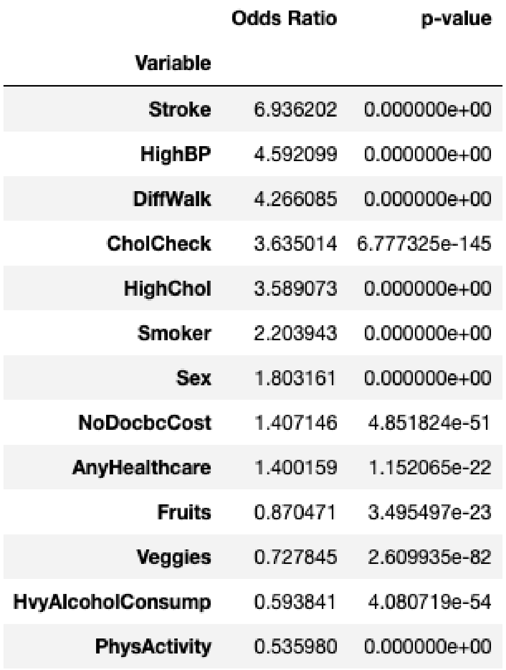


Figure 7: Odds ratio heatmap

We created a heatmap of the odds ratio between all of the binary variables. Looking closer at Heart Disease, we found that the top 5 variables closely associated with Heart Disease are Stroke, High Blood Pressure (‘HighBP’), Difficulty Walking (‘DiffWalk’), Cholesterol Check (‘CholCheck’), and High Cholesterol (‘HighChol’). The presence of these variables increases the odds of heart disease compared to the absence of these variables.



	Odds Ratio	p-value
Variable		
Stroke	6.936202	0.000000e+00
HighBP	4.592099	0.000000e+00
DiffWalk	4.266085	0.000000e+00
CholCheck	3.635014	6.777325e-145
HighChol	3.589073	0.000000e+00
Smoker	2.203943	0.000000e+00
Sex	1.803161	0.000000e+00
NoDocbcCost	1.407146	4.851824e-51
AnyHealthcare	1.400159	1.152065e-22
Fruits	0.870471	3.495497e-23
Veggies	0.727845	2.609935e-82
HvyAlcoholConsump	0.593841	4.080719e-54
PhysActivity	0.535980	0.000000e+00

Figure 8: Odds ratio table

The variables with odds ratio less than 1 are Fruits, Veggies, Heavy Alcohol Consumption (‘HvyAlcoholConsump’), and Physical Activity (‘PhysActivity’). These variables are negatively correlated with Heart Disease meaning the presence of these variables decreases the odds of heart disease compared to the absence of these variables.

4.2 Decision Tree

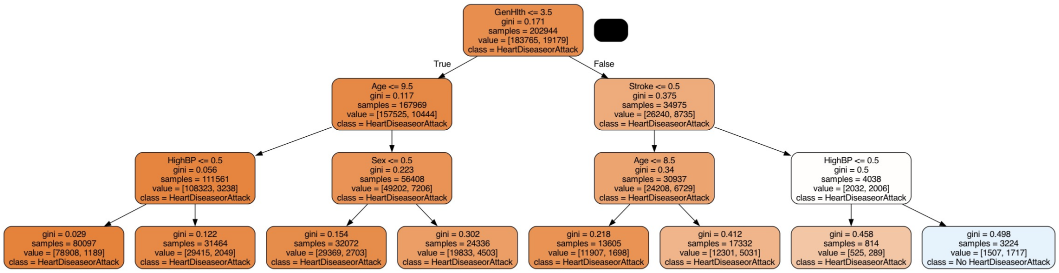


Figure 9: Decision Tree

Formula for information gain from conditional entropy of Y and conditional entropy of Y given X.

$$IG(Y|X) = CE(Y) - CE(Y|X)$$

Using the method to calculate the conditional entropy of the variables calculated in the previous sections, we constructed a classification decision tree that would try to minimize the amount of conditional entropy and thus maximize the information gain from these variables on heart disease.

As seen in Figure 9, the classification tree splits at subsets of the data where the homogeneity will increase and the entropy (gini index) lowers in order to correctly predict the response. The model had a 90% accuracy rate based on the confusion matrix tabulated.

However, this model may not be entirely accurate as we could further increase the accuracy through boosting methods - Random Forest, LASSO - or even simply subsetting the data in a smaller sample size, since the dataset itself was inherently biased that this would pose of large risk in the efficacy of the model to accurately classify heart disease or attack.

5 Conclusion

After finding the conditional entropies between heart disease occurrence and every variable in the dataset, we found five independent variables worth noting that had the lowest conditional entropies out of every pair observed. Namely, the variables with the lowest conditional entropy were General Health (‘GenHlth’) at 0.279970, Age at 0.283664, High Blood Pressure (‘HighBP’) at 0.289980, Difficulty Walking (‘DiffWalk’) at 0.293992, and High Cholesterol (‘HighChol’) at 0.295831. The low conditional entropy values implies that there is dependence between each of the five variables and the dependent variable of heart disease occurrence.

We then proceeded to analyze the conditional entropy values of fused variables, and found that the lowest conditional entropy values with the the dependent variable and the fused variable ‘Stroke’ was ‘DiffWalk’ at 1.09233492, followed by ‘HighBP’ at 1.09609572 and ‘GenHlth’ at 1.77856531.

From the analysis of the correlation visualizations, we found that the three variables that had the highest correlation with heart disease were age, physical health and diabetes. The three variables that had the highest correlation with the response variable stroke were in order: physical health, age and diabetes. From this we can conclude that these three variables have the biggest effect on the likelihood of a person getting heart disease or stroke.

Subsequently, we continued our analysis using odds ratios by creating a heatmap of the odds ratio between all of the binary variables. The top 5 variables that have a strong correlation with Heart Disease are Stroke, High Blood Pressure (‘HighBP’), Difficulty Walking (‘DiffWalk’), Cholesterol Check (‘CholCheck’), and High Cholesterol (‘HighChol’). The odds ratio values for these were, in order: 6.936202, 4.592099, 4.266085, 3.635014 and 3.589073. From these results we can infer that these five variables increase the odds of heart disease, as opposed to when they are not considered as predictors of heart disease. Similarly, we were also able to find the variables that have an odds ratio less than one, indicating that they have a negative correlation with heart disease and thus their presence decreases the risk of heart disease. These variables were Fruits, Veggies, Heavy Alcohol Consumption (‘HvyAlcoholConsump’), and Physical Activity (‘PhysActivity’). Surprisingly, based on the odds ratio analysis, heavy alcohol consumption was indicated to decrease the risk of heart disease. The odds-ratio values for these four variables are in order: 0.870471, 0.727845, 0.593841 and 0.535980.

Given these points, we can conclude that the most important factors in understanding heart disease are prevalence of stroke, difficulty walking, health factors such as high cholesterol, and others that have been discussed throughout this paper. Further studies into these specific factors, can help determine exact effects on the outcomes of heart disease or stroke and help prevent people from being vulnerable to heart disease.

Acknowledgments

Thank you to Dr. Fushing Hsieh and Shuting Liao.

References

[1] Acerace.py, et al. “A Better Way to Calculate Odd Ratio in Pandas.” Stack Overflow, stackoverflow.com/questions/43261747/a-better-way-to-calculate-odd-ratio-in-pandas. Accessed 14 May 2023.

[2] “Assessing Your Weight.” Centers for Disease Control and Prevention, 3 June 2022, www.cdc.gov/healthyweight/assessing/index.html.

[3] “Conditional Entropy Computation.” R, search.r-project.org/CRAN/refmans/infotheo/html/condentropy.html. Accessed 14 May 2023.

[4] Holtz, Yan. “Dendrogram with Heat Map.” The Python Graph Gallery, www.python-graph-gallery.com/404-dendrogram-with-heat-map. Accessed 14 May 2023.

[5] Lutes, Jeremiah. “Entropy and Information Gain in Decision Trees.” Medium, 3 Feb. 2021, towardsdatascience.com/entropy-and-information-gain-in-decision-trees-c7db67a3a293.

[6] Orac, Roman. “The Fastest Way to Visualize Correlation in Python.” Medium, 24 Aug. 2021, towardsdatascience.com/the-fastest-way-to-visualize-correlation-in-python-ce10ed533346.

[7] Thorn, James. “Decision Trees Explained.” Medium, 26 Sept. 2021, towardsdatascience.com/decision-trees-explained-3ec41632ceb6.