Examining Heart Disease: Insights from Associations and Covariate Interactions

Yucheng Zhao, Abhi Gupta, Alexander Yu, Brett Loy

Introduction

Heart disease continues to be a major global cause of death and illness, which presents a serious public health issue. It alone is the leading cause of death in the US, taking roughly 647,000 lives per year. Even though a great deal of research has revealed a number of risk factors linked to heart disease, the complex relationships between these factors still interest scientists and medical professionals. Comprehending the intricate association between heart disease and its predictor variables is vital in order to design effective prevention tactics, personalized treatments, and focused remedies.

In this project, we delve into the intricate web of associations and covariate interactions to enhance our understanding of heart disease prediction. Our analysis goes beyond simple associations, focusing on how specific categories and combinations of covariates influence the likelihood of heart disease. By examining both associations and covariate interactions, we aim to unravel the complexity of heart disease prediction, providing valuable insights for healthcare practitioners and policymakers. Through our comprehensive approach, we seek to contribute to the ongoing efforts to combat heart disease and improve public health outcomes.

Overall, we present our findings, detailing the methods used, results obtained, and implications of our analysis. By shedding light on the nuanced relationships between covariates and heart disease, we hope to advance the field of predictive modeling and pave the way for more effective prevention strategies.

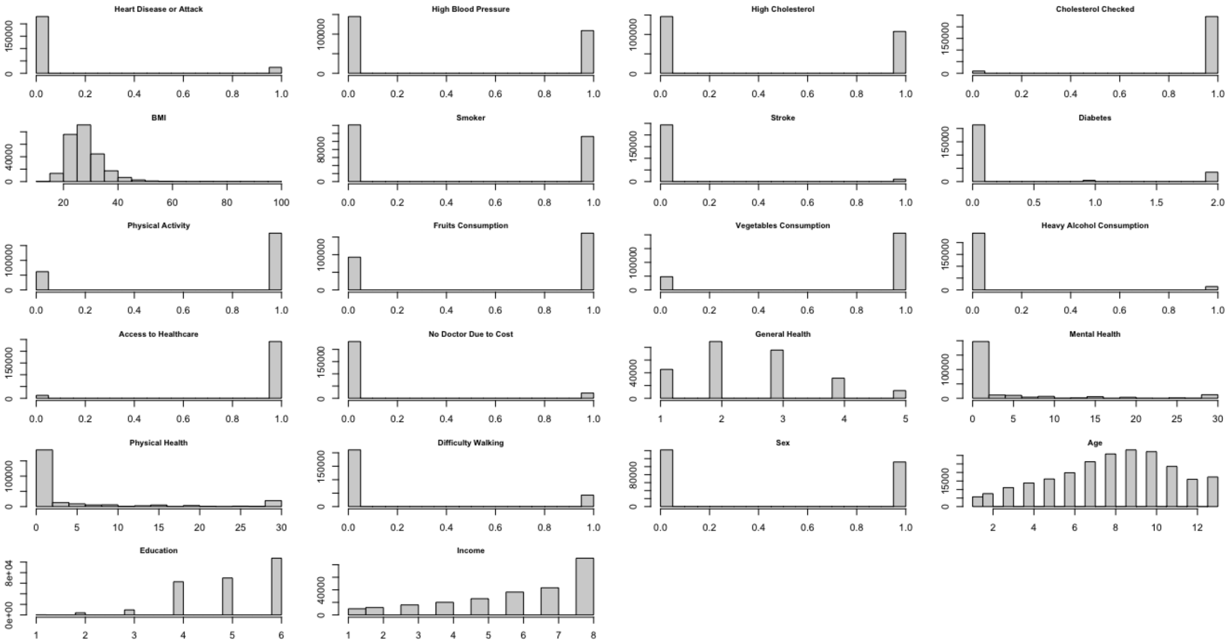
Data Cleaning and Summaries

The data used is the Behavioral Risk Factor Surveillance System (BRFSS) dataset for the year 2015, which is a health-related telephone survey conducted annually by the CDC. The BRFSS collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. This specific data set was cleaned before download and contains 253,680 survey responses with no missing values. It contains 22 variables which are the following:

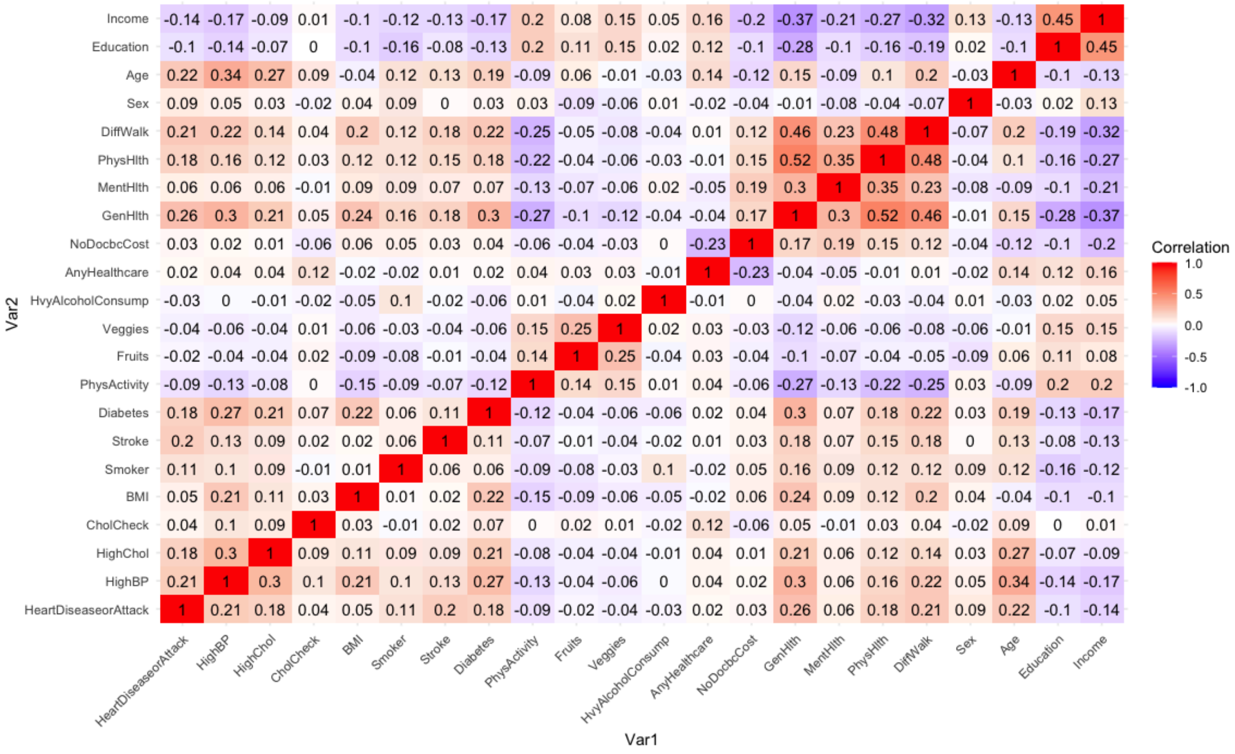
Variable	Description	Type
HeartDiseaseorAttack	Individuals who have reported having coronary heart disease or myocardial infarction	Binary Categorical
HighBP	Individuals diagnosed with high blood pressure by a health professional	Binary Categorical
HighChol	Individuals diagnosed with high cholesterol by a health professional	Binary Categorical
CholCheck	Individuals who have checked cholesterol level within the past five years	Binary Categorical

BMI	Body Mass Index	Continuous Numeric
Smoker	Individuals who have smoked at least 100 cigarettes in their entire life	Binary Categorical
Stroke	Individuals who have had a stroke	Binary Categorical
Diabetes	Individuals who have diabetes	Categorical (0-2)
PhysActivity	Individuals who have done physical activity or exercise in the past 30 days away from their job	Binary Categorical
Fruits	Individuals who consume at least 1 fruit per day	Binary Categorical
Veggies	Individuals who consume at least 1 vegetable per day	Binary Categorical
HvyAlcoholConsump	Individuals who consume heavy alcohol(more than 14 drinks a week for men and 7 drinks a week for women)	Binary Categorical
AnyHealthcare	Individuals who have any kind of health care coverage or government plans	Binary Categorical
NoDocbCost	Individuals who needed to see a doctor but could not because of cost in the past 12 months	Binary Categorical
GenHlth	Individuals' self-rated health	Categorical (1-5)
MentHlth	Number of days during the past 30 days where mental health was not good	Discrete Numeric
PhysHlth	Number of days during the past 30 days where physical health was not good	Discrete Numeric
DiffWalk	Individuals with serious difficulty walking or climbing stairs	Binary Categorical
Sex	Individual's sex	Binary Categorical
Age	Individual's age	Categorical (1-13)
Education	Individual's highest grade or year of school completed	Categorical (1-6)
Income	Individual's annual income level	Categorical (1-8)

To look at the trends and association of the data, we have created histograms of the interest and explanatory variables in the heart disease dataset which provide a detailed look at the frequency distribution of each variable's values, offering insights into the prevalence and distribution of heart disease risk factors and related health indicators. For example, the histogram for the HeartDiseaseorAttack variable reveals that approximately 10% of the observations correspond to individuals with heart disease or a heart attack (value of 1), while the remaining 90% correspond to individuals without these conditions (value of 0). This distribution suggests that heart disease or heart attacks are relatively less prevalent among the surveyed population. For explanatory variables we also can see their respective distributions as for binomial explanatory variables we can see which prevalence of their conditions while for continuous or other categorical variables, we can see their distribution across the population



The correlation heatmap below provides a visual representation of the relationships between various health factors based on the dataset. From each cell in the heatmap we can see the correlation coefficient between two health indicators, ranging from -1 to 1. We can see that a majority of the variables here have fairly low correlation, regardless of having a positive or negative relationship. Amongst correlations with our target variable heart disease, we can see that Age,General Health, High Blood Pressure and difficulty walking appear to be the most correlated while fruit consumption and having health care appear to be the least correlated.



Methodology

In this project, we utilized entropy to measure uncertainty and interaction effects between various health indicators (X) and heart disease (Y) in the heart disease dataset. Entropy is a measure of the unpredictability or disorder of a system, with higher entropy indicating greater uncertainty. For the discrete random variable X, representing each health indicator, we calculated the entropy CE(X) using the formula $-\Sigma p(x)\log p(x)$, where p(x) is the probability mass function. Similarly, for the outcome variable Y, representing heart disease, we calculated the entropy CE(Y) in the same manner. To understand the relationship between X and Y, we computed the conditional entropy CE[Y|X=x] for each value of X. This conditional entropy represents the remaining uncertainty in Y given a specific value of X. By averaging these conditional entropies over all values of X, weighted by the probabilities of each value of X, we obtained the conditional entropy CE[Y|X], which indicates the overall uncertainty in Y given X.

To determine if there were interaction effects between the General Health and Age, which we found were the most significant predictors, we compared the difference between CE(Y) and CE[Y|X] with the sum of the differences between CE(Y) and CE[Y|X=0], and CE(Y) and CE[Y|X=1]. If the former difference was greater than the latter sum, we concluded that there were interaction effects betweenGeneral Health and Age.

Another method we used was odds ratio which is pivotal for quantifying the influence of various health indicators on the presence or absence of heart disease. For any health indicator, we can calculate the odds ratio (OR) to understand the strength of its association with heart disease.The odds of a health indicator (A) in the presence of heart disease (B) are calculated as P(A|B)/(1-P(A|B)), where P(A|B) is the probability of the health indicator given heart disease. Similarly, the odds of the health indicator in the absence of heart disease are P(A|¬B)/(1-P(A|¬B)), where P(A|¬B) is the probability of the health indicator given no heart disease. The odds ratio for the health indicator and heart disease is then the ratio of these two odds: OR = (P(A|B)/(1-P(A|B))) / (P(A|¬B)/(1-P(A|¬B))). This formula quantifies how much more likely a person with heart disease is to have the health indicator compared to a person without heart disease. After doing this we employed a two-way conditioned odds-ratio to explore the combined effects of specific indicators in heart disease risk

Overall, in both methods, we conducted each methodology, finding all the associations between heart disease and each specific indicator and then decided to take the significant indicators from each model to then conduct analysis on their interaction and association.

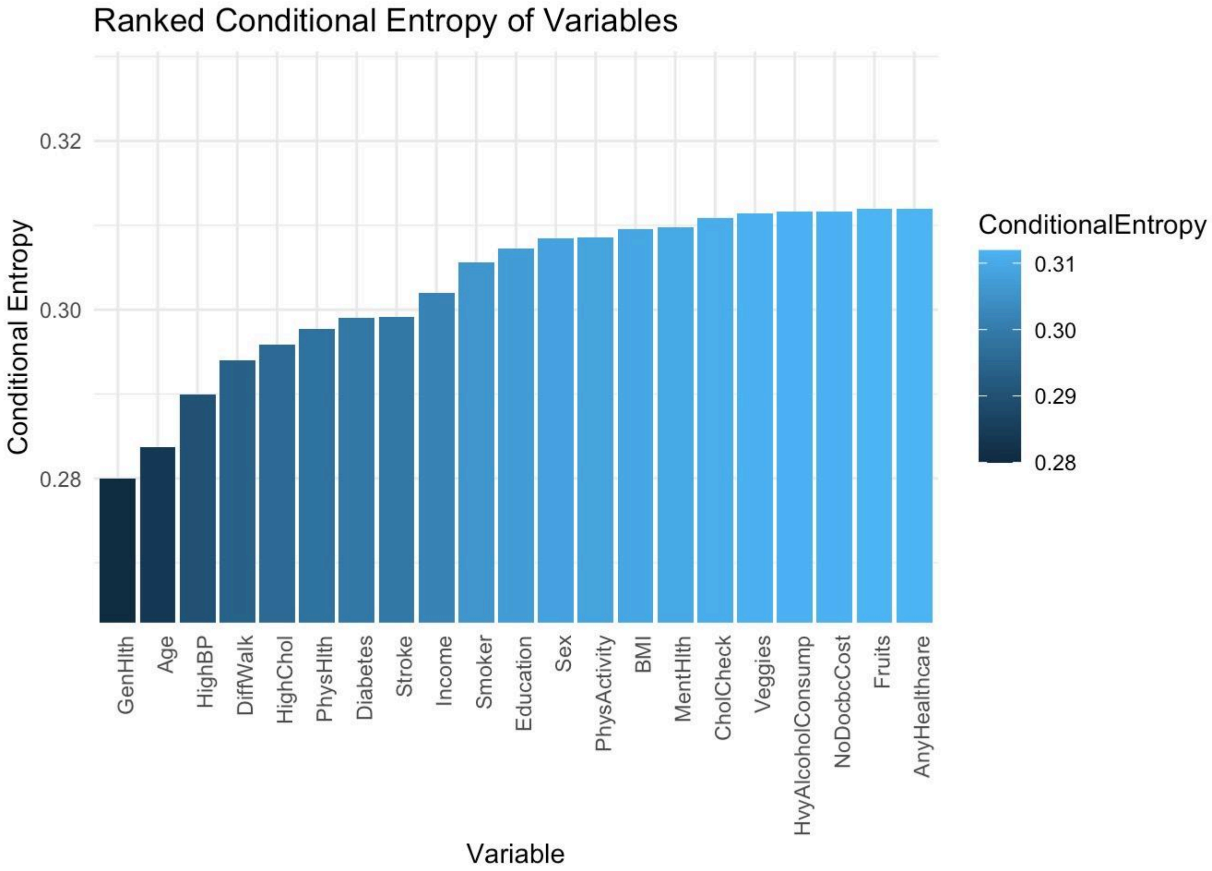
Results and Discussion

Conditional Entropy of each Covariate

This project explored the associations between heart disease and various covariates by utilizing conditional entropy to measure the uncertainty reduction provided by each covariate. By calculating the conditional entropy between the response variable “HeartDiseaseorAttack” and every other variable, we quantified the amount of uncertainty reduced in “HeartDiseaseorAttack” by each covariate and identified the key factors that are most predictive of heart disease.

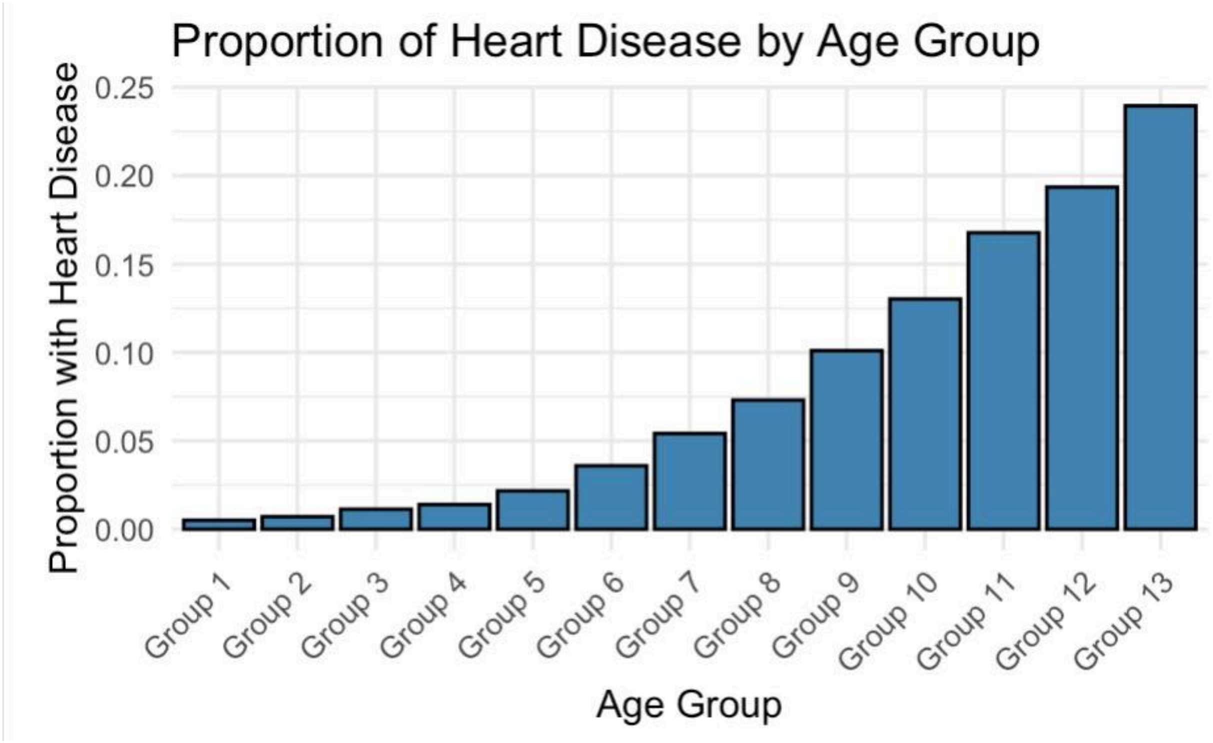
Our findings suggest that “GenHlth” and “Age” are the most significant predictors among all the covariates analyzed. The conditional entropy values for “GenHlth” and “Age” are 0.280 and 0.284, respectively. With lower entropy values indicating greater predictive power, the results suggest that knowing a participant’s general health and age significantly reduces the uncertainty about their occurrence of heart disease. At the same time, variables such as “NoDocbcCost,” “Fruits,” and “AnyHealthCare” have higher values, indicating they have less predictive power.

The bar chart below ranks each variable by its conditional entropy value:



Heart Disease by Age Group

To better understand the impact of age on heart disease, we plotted the proportion of heart disease occurrences across different age groups. The plot below shows that heart disease proportions increase with advancing age, indicating age is a critical factor.



Interaction Effect

To investigate the interaction effect between “GenHlth” and ”Age,” we fused the two variables and tested if the interaction effect presents according to the formula:

$$CE[Y] - CE[Y|X] > CE[Y] - CE[Y|X_1] + CE[Y] - CE[Y|X_2]$$

Calculations:
Conditional Entropy with Fused Variables (“GH_Age”) : CE[Y|X] = 0.257
Individual Conditional Entropy 1 (“GenHlth”): CE[Y|X1] = 0.280
Individual Conditional Entropy 2 (“Age”): CE[Y|X2] = 0.284
Total Entropy of Heart Disease (“HeartDiseaseorAttack”): CE[Y] = 0.312
Reduction in Entropy with Fused Variables: CE[Y] - CE[Y|X] = 0.055
Sum of Reductions with Individual Variables: CE[Y] - CE[Y|X1] + CE[Y] - CE[Y|X2] = 0.061

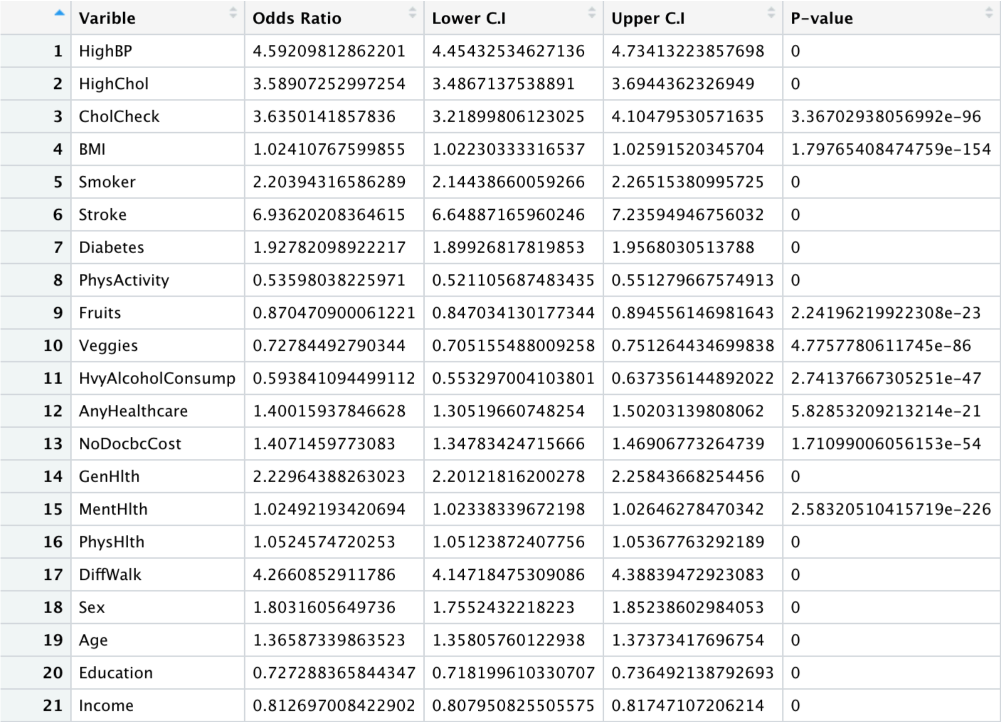
Since the reduction in entropy with the fused variables (0.055) is less than the sum of the individual reductions (0.061), no interaction effect is present between "GenHlth" and "Age." This indicates that their influences on heart disease are additive rather than synergistic.

In addition, we fused “Age” with another variable “PhysHlth” to calculate the combined factors that affect heart disease prediction. Our result suggests that while both factors are significant in predicting heart disease individually, their combined predictive power does not exceed their separate contributions. Since CE[Y] - CE[Y|X] = 0.042 < CE[Y] - CE[Y|X1] + CE[Y] - CE[Y|X2] = 0.043.

Odds Ratio

Another key statistical measure used is the odds ratio, seen from the table below. This revealed several key risk factors with varying odds ratios, indicating their impact on the likelihood of developing heart disease. The most significant were the individuals with a history of stroke as they had a high odds ratio of 6.94, suggesting that they are nearly seven times more likely for heart disease compared to those who have not had a stroke. High blood pressure was another significant risk factor, with an odds ratio of 4.59, indicating that individuals with high blood pressure are nearly five times more likely for heart disease compared to those with normal blood pressure. On the other hand, engaging in regular physical activity was associated with a lower odds ratio of 0.54, suggesting that physically active individuals have 46% lower odds for heart disease compared to sedentary individuals. Similarly, higher consumption of vegetables and fruits was associated with lower odds ratios of 0.73 and 0.87, respectively, indicating the benefits of a diet rich in these foods for heart health.

Additionally, we calculated 95% confidence intervals (CI) for each odds ratio. The CI provides a range of values within which we can be 95% confident that the true odds ratio lies. For example, the 95% CI for HighBP was [4.45, 4.73], indicating that if we were to repeat this study multiple times, we would expect the true odds ratio to fall within this range 95% of the time. Moreover, we assessed the statistical significance of the odds ratios using p-values. In our analysis, all the variables we examined had p-values less than 0.001, indicating a strong statistical significance.



	Varible	Odds Ratio	Lower C.I	Upper C.I	P-value
1	HighBP	4.59209812862201	4.45432534627136	4.73413223857698	0
2	HighChol	3.58907252997254	3.4867137538891	3.6944362326949	0
3	CholCheck	3.6350141857836	3.21899806123025	4.10479530571635	3.36702938056992e-96
4	BMI	1.02410767599855	1.02230333316537	1.02591520345704	1.79765408474759e-154
5	Smoker	2.20394316586289	2.14438660059266	2.26515380995725	0
6	Stroke	6.93620208364615	6.64887165960246	7.23594946756032	0
7	Diabetes	1.92782098922217	1.89926817819853	1.9568030513788	0
8	PhysActivity	0.53598038225971	0.521105687483435	0.551279667574913	0
9	Fruits	0.870470900061221	0.847034130177344	0.894556146981643	2.24196219922308e-23
10	Veggies	0.72784492790344	0.705155488009258	0.751264434699838	4.7757780611745e-86
11	HvyAlcoholConsump	0.593841094499112	0.553297004103801	0.637356144892022	2.74137667305251e-47
12	AnyHealthcare	1.40015937846628	1.30519660748254	1.50203139808062	5.82853209213214e-21
13	NoDocbcCost	1.4071459773083	1.34783424715666	1.46906773264739	1.71099006056153e-54
14	GenHlth	2.22964388263023	2.20121816200278	2.25843668254456	0
15	MentHlth	1.02492193420694	1.02338339672198	1.02646278470342	2.58320510415719e-226
16	PhysHlth	1.0524574720253	1.05123872407756	1.05367763292189	0
17	DiffWalk	4.2660852911786	4.14718475309086	4.38839472923083	0
18	Sex	1.8031605649736	1.7552432218223	1.85238602984053	0
19	Age	1.36587339863523	1.35805760122938	1.37373417696754	0
20	Education	0.727288365844347	0.718199610330707	0.736492138792693	0
21	Income	0.812697008422902	0.807950825505575	0.81747107206214	0

Two-Way Conditioned Odds Ratio

Since we found stroke and high blood pressure to have the highest odds ratios, we decided to examine the relationship between stroke, high blood pressure, and heart disease at a granular level by considering different combinations of these factors. The odds ratios for heart disease were calculated for each combination of stroke and HBP status. For instance, individuals with no HBP and no stroke had significantly higher odds of heart disease compared to those with stroke but no HBP (OR = 53.31). Similarly, the presence of HBP without stroke was also associated with increased odds of heart disease (OR = 13.27). However, the presence of stroke without HBP was surprisingly associated with lower odds of heart disease (OR = 0.0188), indicating a potentially protective effect. This approach allowed us to not only measure the association between variables but also assess how specific combinations of categories affect the likelihood of heart disease, providing a more detailed understanding of these complex relationships. Overall, the interaction between these factors is not straightforward, and the presence or absence of one factor does not always predict the risk of heart disease.

	HighBP=0	HighBP=1
Stroke=0	53.31233596	13.27265574
Stroke=1	0.01875738	0.07534287

Conclusion

In this project, we conducted a comprehensive analysis of associations and covariate interactions to enhance our understanding of heart disease prediction. By utilizing conditional entropy and odds ratios, we gained valuable insights into the factors influencing heart disease and their complex relationships. Our findings indicate that age and general health are significant predictors of heart disease, with lower entropy values suggesting greater predictive power. In contrast, variables such as healthcare access and certain lifestyle factors showed less predictive power. We also investigated the interaction effect between age and general health, finding that their influences on heart disease are additive rather than synergistic. Moreover, our analysis of odds ratios revealed several key risk factors for heart disease, such as a history of stroke and high blood pressure. Interestingly, the presence of stroke without high blood pressure was associated with lower odds of heart disease, indicating a potentially protective effect.

Overall, our study highlights the importance of considering both individual factors and their interactions in heart disease prediction. By providing a more nuanced understanding of these complex relationships, our findings contribute to the ongoing efforts to improve heart disease prevention and management strategies.

References

Teboul, A. (2022, March 10). *Heart disease health indicators dataset*. Kaggle. https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset?resource=download

Code

```
#EDA
library(ggplot2)
heart_disease_health_indicators_BRFSS2015 <-
read.csv("~/Downloads/heart_disease_health_indicators_BRFSS2015.csv")
```

```
data <- heart_disease_health_indicators_BRFSS2015
attach(data)
opar <- par(no.readonly=TRUE)
par(mfrow=c(6, 4), mar=c(3, 3, 1, 1))  # Set margins to leave space for
labels
hist(HeartDiseaseorAttack, main="Heart Disease or Attack", xlab="",
ylab="", cex.main=0.8)
hist(HighBP, main="High Blood Pressure", xlab="", ylab="", cex.main=0.8)
hist(HighChol, main="High Cholesterol", xlab="", ylab="", cex.main=0.8)
hist(CholCheck, main="Cholesterol Checked", xlab="", ylab="", cex.main=0.8)
hist(BMI, main="BMI", xlab="", ylab="", cex.main=0.8)
hist(Smoker, main="Smoker", xlab="", ylab="", cex.main=0.8)
```

```
hist(Stroke, main="Stroke", xlab="", ylab="", cex.main=0.8)
hist(Diabetes, main="Diabetes", xlab="", ylab="", cex.main=0.8)
hist(PhysActivity, main="Physical Activity", xlab="", ylab="",
cex.main=0.8)
hist(Fruits, main="Fruits Consumption", xlab="", ylab="", cex.main=0.8)
hist(Veggies, main="Vegetables Consumption", xlab="", ylab="",
cex.main=0.8)
hist(HvyAlcoholConsump, main="Heavy Alcohol Consumption", xlab="", ylab="",
cex.main=0.8)
hist(AnyHealthcare, main="Access to Healthcare", xlab="", ylab="",
cex.main=0.8)
hist(NoDocbcCost, main="No Doctor Due to Cost", xlab="", ylab="",
cex.main=0.8)
hist(GenHlth, main="General Health", xlab="", ylab="", cex.main=0.8)
hist(MentHlth, main="Mental Health", xlab="", ylab="", cex.main=0.8)
hist(PhysHlth, main="Physical Health", xlab="", ylab="", cex.main=0.8)
hist(DiffWalk, main="Difficulty Walking", xlab="", ylab="", cex.main=0.8)
hist(Sex, main="Sex", xlab="", ylab="", cex.main=0.8)
hist(Age, main="Age", xlab="", ylab="", cex.main=0.8)
hist(Education, main="Education", xlab="", ylab="", cex.main=0.8)
hist(Income, main="Income", xlab="", ylab="", cex.main=0.8)
par(opar)
library(ggplot2)
dev.off()
correlation_matrix <- cor(heart_disease_health_indicators_BRFSS2015[,
c("HeartDiseaseorAttack", "HighBP", "HighChol", "CholCheck", "BMI",

"Smoker", "Stroke", "Diabetes", "PhysActivity", "Fruits",
"Veggies", "HvyAlcoholConsump", "AnyHealthcare", "NoDocbcCost",

"GenHlth", "MentHlth", "PhysHlth", "DiffWalk", "Sex", "Age",

"Education", "Income")])
```

```
# Convert the correlation matrix to a data frame for plotting
correlation_df <- as.data.frame(as.table(correlation_matrix))
names(correlation_df) <- c("Var1", "Var2", "Correlation")
```

```
# Plot the heatmap
ggplot(correlation_df, aes(Var1, Var2, fill = Correlation)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint
= 0, limits = c(-1, 1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Correlation Heatmap of Health Indicators",
       x = "Health Indicators",
       y = "Health Indicators")
```

```
#oddsratio
variables <- setdiff(names(data), "HeartDiseaseorAttack")

odds_ratios <- list()
for (var in variables) {
  tab <- xtabs(~data[[var]] + HeartDiseaseorAttack, data = data)
  or <- oddsratio(tab, conf.level = 0.95)
  odds_ratios[[var]] <- or$measure[2, ]
}
odds_ratios_df <- do.call(rbind, odds_ratios)
print(odds_ratios_df)
```

```
library(epitools)
variables <- names(data)[!names(data) %in% c("HeartDiseaseorAttack")]
calculate_odds_ratio <- function(var_name) {
  tab <- table(data[[var_name]], data$HeartDiseaseorAttack)
  or <- oddsratio(tab, conf.level = 0.95)$measure[2, ]
  return(c(var_name, or[1], or[2], or[3]))
}

tab <- table(data$Stroke, data$HighBP)

odds_ratios <- matrix(NA, nrow = 2, ncol = 2)
for (i in 1:2) {
  for (j in 1:2) {
    odds_ratios[i, j] <- tab[i, j] / tab[3 - i, j]
  }
}
print("Odds ratios for heart disease conditioned on combinations of
categories:")
print(odds_ratios)
```

```
# Data

# Libraries
#install.packages("DescTools")

library(DescTools)
library(dplyr)

---

library(DescTools)
library(dplyr)

# Load data set
df <- read.csv("heart_disease_health_indicators_BRFSS2015.csv")

head(df, 10)
```

```
# Clean missing data
full_rows <- complete.cases(df)
df <- df[full_rows, ]
```

```
# Calculate the conditional entropy between heart disease and each other
variable
library(infotheo)
```

```r
df$HeartDiseaseorAttack <- as.factor(df$HeartDiseaseorAttack)

CE <- vector()
Y_variables <- setdiff(names(df), "HeartDiseaseorAttack")

# Calculate the conditional entropy for each variable
for (y in Y_variables) {
  variable_as_factor <- as.factor(df[[y]])

  # Calculate conditional entropy and store the result
  entropy_value <- condentropy(df$HeartDiseaseorAttack, variable_as_factor)
  CE <- c(CE, entropy_value)
}

names(CE) <- Y_variables
CE <- sort(CE, decreasing = FALSE)
print(CE)
```

```r
library(ggplot2)

CE_data <- data.frame(Variable = names(CE), ConditionalEntropy = CE)

ggplot(CE_data, aes(x = reorder(Variable, ConditionalEntropy), y =
ConditionalEntropy, fill = ConditionalEntropy)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Ranked Conditional Entropy of Variables",
       x = "Variable",
       y = "Conditional Entropy") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  coord_cartesian(ylim = c(min(CE_data$ConditionalEntropy) * 0.95,
max(CE_data$ConditionalEntropy) * 1.05))
```

```r
# Calculate CE[Y|X = X1, X2] where Y is HeartDiseaseorAttack, and X1 and X2
are GenHlth and Age
df$GH_Age <- paste(df$GenHlth, df$Age, sep=',')
tab_G_A <- xtabs(~HeartDiseaseorAttack + GH_Age, data=df)
tab_G_A = addmargins(tab_G_A)
```

```
probs = tab_G_A[3, 1:65]/sum(tab_G_A[3, 1:65])
H_yxx=apply(tab_G_A[1:2,1:65],2,function(o) Entropy(o,base=exp(1)))

CE_Y_X=sum(probs*H_yxx)
print(paste0('CE[Y|X]:',CE_Y_X))
```

```
# Test for interaction effect between GenHlth and Age
CE_Y_X1 <- 0.2799704
CE_Y_X2 <- 0.2836638

CE_Y <- Entropy(tab_G_A[1:2,66], base = exp(1))

diff_fused <- CE_Y - CE_Y_X
diff_ind <- (CE_Y - CE_Y_X1) + (CE_Y - CE_Y_X2)


if (diff_fused > diff_ind) {
  cat("Interacting effects are presentbetween Genhlth and Age: ",
diff_fused, ">", diff_ind, "\n")
} else {
  cat("No interacting effects present between Genhlth and Age: ",
diff_fused, "<=", diff_ind, "\n")
}
```

```
# Entropy (Manual)
## Y = HeartDiseaseorAttack, X = Age

tab1=xtabs(~HeartDiseaseorAttack+Age,data=df)
tab1=addmargins(tab1)
print(tab1)
```

```
# The number of people with heart disease or attacks increases with age

# Age groups 9 through 12 have particularly high numbers of individuals
reporting heart disease or attacks, suggesting a higher prevalence in these
older age groups.

# Load necessary libraries
library(ggplot2)
library(dplyr)
```

```r
df_plot <- as.data.frame.matrix(tab1[1:2, 1:13])

colnames(df_plot) <- paste("Group", 1:13)
rownames(df_plot) <- c("No", "Yes")

df_plot <- t(df_plot)
df_plot <- as.data.frame(df_plot)


df_plot$AgeGroup <- factor(paste("Group", 1:13), levels = paste("Group", 1:13))
```

```r
# Calculating the proportion of heart disease
df_plot$Proportion <- as.numeric(df_plot$Yes) / (as.numeric(df_plot$Yes) + as.numeric(df_plot$No))


df_plot <- data.frame(AgeGroup = df_plot$AgeGroup, Proportion = df_plot$Proportion)

# Plotting the proportion of heart disease by age group
ggplot(df_plot, aes(x = AgeGroup, y = Proportion)) +
  geom_bar(stat = "identity", fill = "steelblue", color = "black") +
  theme_minimal() +
  labs(title = "Proportion of Heart Disease by Age Group",
       x = "Age Group",
       y = "Proportion with Heart Disease") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  #
```

```r
library(DescTools)

# entropy CE(X)
CE_x <- Entropy(tab1[3,1:13],base=exp(1))
print(paste0('CE[X]:',CE_x))

# entropy CE(Y)
CE_y <- Entropy(tab1[1:2,14],base=exp(1))
```

```
print(paste0('CE[Y]:',CE_y))

# Calculate CE[Y|X=x] for each age group
ce_y_given_x <- numeric(13)

for (i in 1:13) {
  freq <- tab1[1:2, i]
```

```
  # Calculate entropy for each age group
  ce_y_given_x[i] <- Entropy(freq, base=exp(1))
}

print(ce_y_given_x)

# use apply function to obtain CE[Y|X=x]
H_yxx=apply(tab1[c(1,2),c(1:13)],2,function(o) Entropy(o,base=exp(1)))
print(H_yxx)


# obtain n_x/N
# Total counts for each age group
total_counts <- colSums(tab1[1:2, 1:13])

# Probability of each age group
p_x <- total_counts / sum(total_counts)

# CE[Y|X]
ce_y_given_X <- sum(p_x * ce_y_given_x)

print(ce_y_given_X)

cat("CE[Y|X]:", ce_y_given_X, "\n")
```

```
## Global view: Y = HeartDiseaseorAttack, X1 = Age, X2 = PhysHlth
# X2 = PhysHlth
tab2=xtabs(~HeartDiseaseorAttack+PhysHlth,data=df)
tab2=addmargins(tab2)
print(tab2)

# Calculate CE[Y|X2=x2] for each PhysHlth group
```

```
ce_y_given_x2 <- numeric(31)

for (i in 1:31) {
  freq <- tab2[1:2, i]

  # Calculate entropy for each age group
  ce_y_given_x2[i] <- Entropy(freq, base=exp(1))
}

print(ce_y_given_x2)

# CE[Y|X2]
# Total counts for each PhysHlth group
total_counts <- colSums(tab2[1:2, 1:31])

# Probability
p_x2 <- total_counts / sum(total_counts)

ce_y_given_X2 <- sum(p_x2 * ce_y_given_x2)


cat("CE[Y|X2]:", ce_y_given_X2, "\n")

CE_Y <- CE_y
CE_Y_X1 <- ce_y_given_X
CE_Y_X2 <- ce_y_given_X2

# fused variable Age_Phys
df$Age_Phys=paste(df$Age,df$PhysHlth,sep=',')
tab3=xtabs(formula = ~HeartDiseaseorAttack+Age_Phys,data=df)

counts_AP <- colSums(tab3)
p_x <- counts_AP / sum(counts_AP)

CE_Y_x3 <- apply(tab3, 2, function(freq) Entropy(freq, base=exp(1)))

# Overall conditional entropy CE[Y|X]
CE_Y_X3 <- sum(p_x * CE_Y_x3)

cat("CE[Y|X]:", CE_Y_X3, "\n")
```

```
# Interaction Effect Test
# Compare
diff_fused <- CE_Y - CE_Y_X3
diff_ind <- CE_Y - CE_Y_X1 + CE_Y - CE_Y_X2

if (diff_fused > diff_ind) {
  cat("Interacting effects are present between Age and PhysHlth: ",
diff_fused, ">", diff_ind, "\n")
} else {
  cat("No interacting effects present between Age and PhysHlth: ",
diff_fused, "<=", diff_ind, "\n")
}
```

```
# Entropy (Package)

library(infotheo)

# Conditional Entropy Y = HeartDiseaseorAttack, X1 = Age, X2 = PhysHlth

condentropy(X = df$HeartDiseaseorAttack, Y = df$Age_Phys)

condentropy(X = df$Stroke, Y = df$Chol_BP)


# Conditional Entropy Y = Stroke, X1 = Age, X2 = PhysHlth

condentropy(X = df$Stroke, Y = df$Age_Phys)


# Conditional Entropy Y = Diabeties, X1 = HighChol, X2 = HighBP

condentropy(X = df$Diabetes, Y = df$Chol_BP)


# Conditional Entropy Y = Diabeties, X1 = Age, X2 = PhysHlth
condentropy(X = df$Diabetes, Y = df$Age_Phys)
```

```
# Odds Ratio
```

```
library(epitools)


## Y = HeartDiseaseorAttack, X1 = HighChol, X2 = HighBP
tab1=xtabs(~Chol_BP+HeartDiseaseorAttack,data=df)
# calculate odds ratio
# oddsratio(x,method = c("midp", "fisher", "wald", "small"),conf.level =
0.95)
# x: input data can be r x 2 table
# method: method for calculating odds ratio and confidence interval
# conf.level: confidence level (default is 0.95)
oddsratio(tab1,conf.level = 0.95)

## Y = Stroke, X1 = HighChol, X2 = HighBP
tab1=xtabs(~Chol_BP+Stroke,data=df)
# calculate odds ratio
# oddsratio(x,method = c("midp", "fisher", "wald", "small"),conf.level =
0.95)
# x: input data can be r x 2 table
# method: method for calculating odds ratio and confidence interval
# conf.level: confidence level (default is 0.95)
oddsratio(tab1,conf.level = 0.95)
```