Analysis of Risk Factors of Heart Disease or Attack, and Heart Disease or Attack and Stroke

Yuanyuan Gao, Yirong Xu, Rachel How, Kayla Pham
University of California, Davis
STA 160 - Practice in Data Science
Professor Fushing Hsieh
May 14, 2023

Abstract

The report consisted of two main sections that examined the relationship between heart disease and its major risk factors, and the relationship between heart disease and stroke and their major risk factors. The main objective is to develop predictive models that can accurately predict whether an individual will develop heart disease and stroke by using various risk factors. The research is based on the analysis of the 2015 Heart Disease Health Indicators dataset on Kaggle, which explores heart disease and its top influencing factors in the United States. The goal is to assist individuals in preventing heart disease more effectively. Our study revealed that BMI, age, income, general health, and high blood pressure are the five main risk factors associated with heart disease and stroke based on conditional entropy values. Nearly half of Americans experienced at least one of these factors [3]. In addition, we used supervised machine learning models to analyze the dataset of health indicators for heart disease and stroke. This research provides insights that could help develop effective preventative measures against heart disease and stroke.

1 Introduction

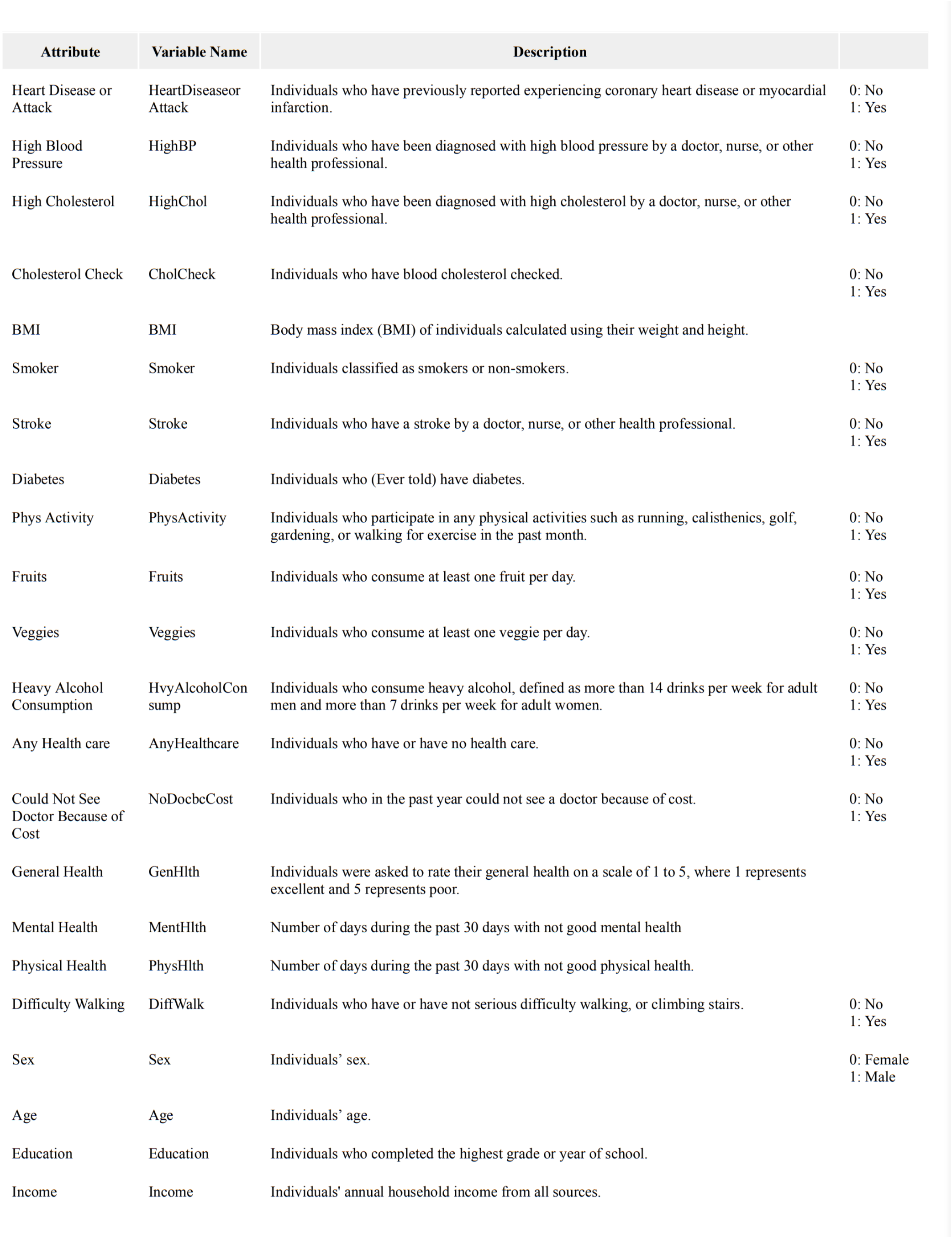
According to the CDC, heart disease and stroke are the first and fifth leading causes of death in the United States [2]. Understanding the risk factors associated with these conditions is crucial for prevention efforts. In our study, we will analyze the relationship between heart disease and its major risk factors, as well as heart disease and stroke, and their major risk factors in the United States using data from the 2015 U.S. Heart Disease Health Indicators dataset [4]. By examining the significant risk factors and using statistical models, we aim to identify preventable risk factors and their impact on the incidence of heart disease and stroke. The dataset includes responses from over 400,000 participants, covering various risk factors such as high blood pressure, high cholesterol, body mass index, and diabetes. By studying the links between these conditions and their major risk factors, we hope to let people be aware of the underlying forewarning factors of heart disease and stroke in the United States.

We will use 3 different supervised machine learning algorithms including K-nearest neighbor (KNN), random forest (RF) and logistic regression (LR) to detect heart disease attacks and identify the characteristics that predicted heart disease and stroke most accurately. To achieve this, we will split our dataset into a 70% training set and a 30% testing set. We will train our machine learning models on a training set and evaluate their performance on a test set, and then utilize the machine learning algorithms to see whether the risk factors can accurately predict people getting heart disease and stroke. By employing multiple machine learning algorithms and selecting the most accurate one, we are able to identify the most important factors that contribute to heart disease and stroke and build predictive models. Our approach allows us to effectively detect and prevent heart attacks by identifying individuals with the highest risk.

2 Research Questions

1. How well can the variables predict people getting Heart Disease Attack in the United States?
2. What are the most significant variables that would affect Heart Disease Attack and Heart Disease Attack and Stroke based on the appropriate prediction model?

3 Dataset Description



Attribute	Variable Name	Description	
Heart Disease or Attack	HeartDiseaseor Attack	Individuals who have previously reported experiencing coronary heart disease or myocardial infarction.	0: No 1: Yes
High Blood Pressure	HighBP	Individuals who have been diagnosed with high blood pressure by a doctor, nurse, or other health professional.	0: No 1: Yes
High Cholesterol	HighChol	Individuals who have been diagnosed with high cholesterol by a doctor, nurse, or other health professional.	0: No 1: Yes
Cholesterol Check	CholCheck	Individuals who have blood cholesterol checked.	0: No 1: Yes
BMI	BMI	Body mass index (BMI) of individuals calculated using their weight and height.	
Smoker	Smoker	Individuals classified as smokers or non-smokers.	0: No 1: Yes
Stroke	Stroke	Individuals who have a stroke by a doctor, nurse, or other health professional.	0: No 1: Yes
Diabetes	Diabetes	Individuals who (Ever told) have diabetes.	
Phys Activity	PhysActivity	Individuals who participate in any physical activities such as running, calisthenics, golf, gardening, or walking for exercise in the past month.	0: No 1: Yes
Fruits	Fruits	Individuals who consume at least one fruit per day.	0: No 1: Yes
Veggies	Veggies	Individuals who consume at least one veggie per day.	0: No 1: Yes
Heavy Alcohol Consumption	HvyAlcoholCon sump	Individuals who consume heavy alcohol, defined as more than 14 drinks per week for adult men and more than 7 drinks per week for adult women.	0: No 1: Yes
Any Health care	AnyHealthcare	Individuals who have or have no health care.	0: No 1: Yes
Could Not See Doctor Because of Cost	NoDocbcCost	Individuals who in the past year could not see a doctor because of cost.	0: No 1: Yes
General Health	GenHlth	Individuals were asked to rate their general health on a scale of 1 to 5, where 1 represents excellent and 5 represents poor.	
Mental Health	MentHlth	Number of days during the past 30 days with not good mental health	
Physical Health	PhysHlth	Number of days during the past 30 days with not good physical health.	
Difficulty Walking	DiffWalk	Individuals who have or have not serious difficulty walking, or climbing stairs.	0: No 1: Yes
Sex	Sex	Individuals' sex.	0: Female 1: Male
Age	Age	Individuals' age.	
Education	Education	Individuals who completed the highest grade or year of school.	
Income	Income	Individuals' annual household income from all sources.	

Figure 1: Summary of All Variables in the Dataset

4 Data Visualization

The histogram of the distributions of each explanatory variable is to show the frequency distribution of the values of each explanatory variable in a dataset. This visualization allows us to observe the overall pattern and characteristics of each variable. For instance, in the case of the HeartDiseaseorAttack variable, the histogram reveals that it consists of only two distinct values, namely 0 and 1. From the histogram, we can observe that approximately 80% of the values correspond to 1, while the remaining 20% correspond to 0.

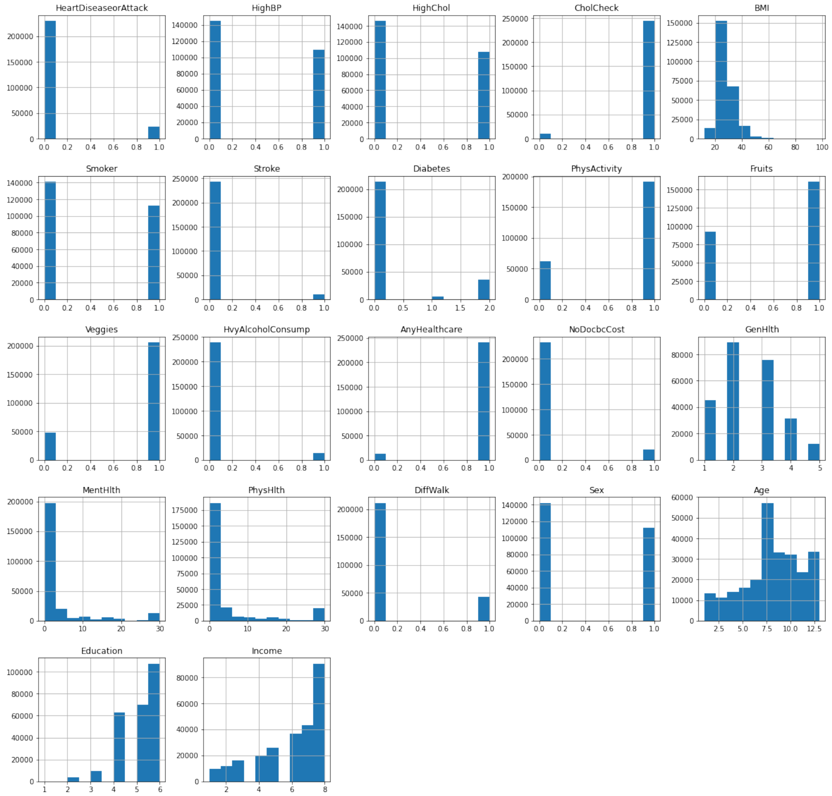


Figure 2: Histogram of the Distributions of Each Explanatory Variables

In Figure 3, we employed a heat map to visualize the correlations between two variables, with one plotted on each axis. By observing its color variations across each axis, we could also identify the correlations between the two variables. In our dataset, we observed that the all the independent variables exhibited moderate to low correlations, either positive or negative.

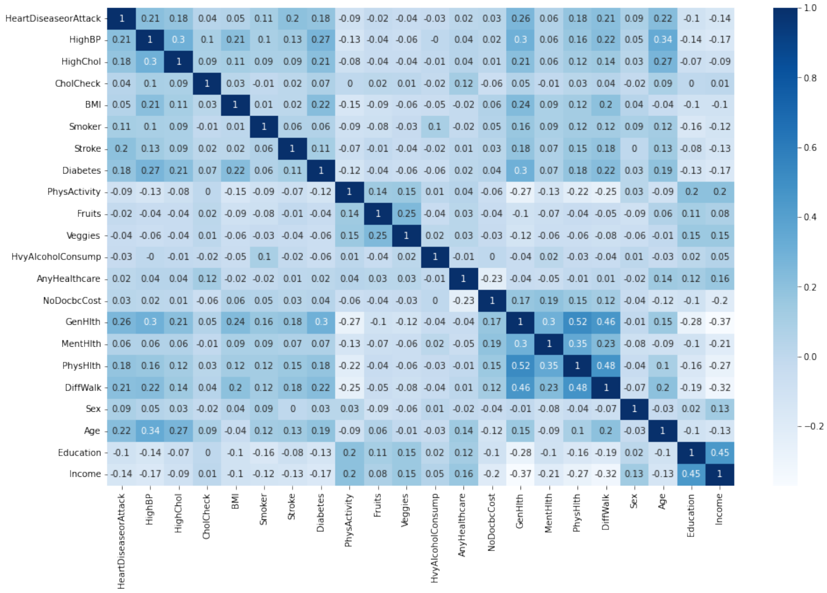


Figure 3: Exploring Data Relationships: Transforming a Correlation Matrix into a Heat Map

In Figure 4, we plotted the mean values of each independent variables for people with heart disease and people with no heart disease to find out the significant factors at a first glance. We can see that for people with heart disease, they tend to have a higher mean values of high blood pressure, high Cholesterol, smoker, stroke, diabetes, health problems, and age.

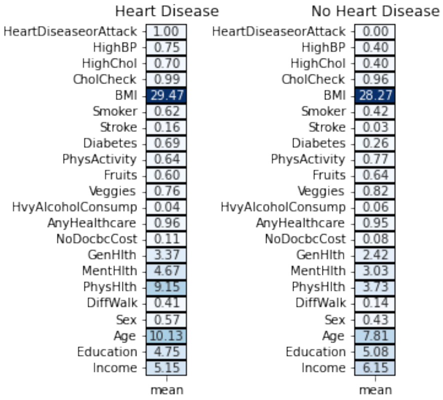


Figure 4: Mean Values of Each Explanatory Variable for Heart Disease vs. No Heart Disease

5 Methodology

Given that our dataset consists of binary, categorical, and numerical variables, we have chosen to apply several main methods for this project, including K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, and Conditional Entropy.

1. **K-NN Algorithm:** Non-parametric classifier which makes predictions regarding the classifications of data points.
2. **Logistic Regression:** Supervised machine learning model to estimate parameters of regression.
3. **Random forest:** Supervised machine learning model to determine the factors based on the accuracy scores and important features.
4. **Conditional Entropy:** The amount of uncertainty or randomness in a random variable Y, given the value of another random variable X. If H(Y|X) = 1, it means that the known value of X provides no information about the value of Y.

6 Data Analysis

6.1 K-Nearest Neighbor (K-NN)

6.1.1 K-NN of HeartDiseaseorAttack

We choose to perform a K-NN algorithm since it's a non-parametric method and it's suitable for classifications and predictions. We first choose HeartDiseasorAttack as the response variable, and separate the dataset into training set and testing set (70%:30%). We then iterated 15 k values and plotted the Accuracy vs. Number of neighbors plot, we can see that when k = 12, the best result is being captured, which is 0.9046699253652896. Thus, we used this score for the confusion matrix and classification report to visualize the accuracy of the prediction. For precision, it calculates the ratio of correctly predicted observations and the total predicted observations. We can see that it can classify 90% of the people that actually have no heart disease attack correctly, while it can only classify 48% of the people that actually have heart disease attack correctly. For recall, it calculates the ratio of correctly predicted observations to all observations in actual class. We can see that it works well with predicting the people that actually have no heart disease attack correctly, while it is not accurate for predicting people that actually have a heart disease attack. Lastly, F1 Score is the weighted average of Precision and Recall, and it is useful when having an uneven class distribution. Therefore, we can further see that predicting people with no heart diseases has a high accuracy, and that it is not useful for predicting people with heart diseases.

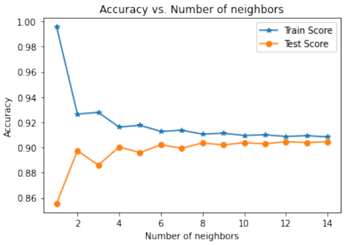


Figure 5: Accuracy vs. Number of neighbors

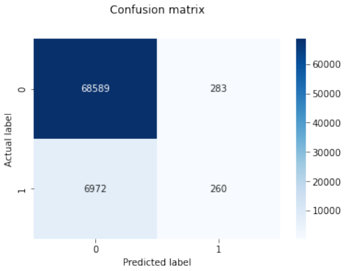
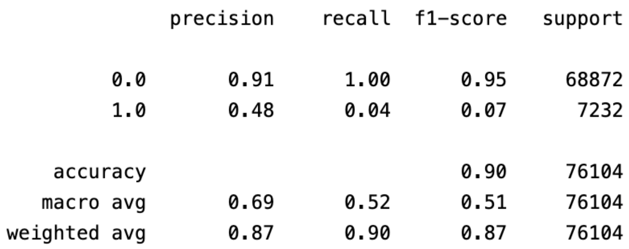


Figure 6: Confusion Matrix



```
              precision    recall  f1-score   support

         0.0       0.91      1.00      0.95     68872
         1.0       0.48      0.04      0.07      7232

    accuracy                           0.90     76104
   macro avg       0.69      0.52      0.51     76104
weighted avg       0.87      0.90      0.87     76104
```

Figure 7: Classification Report of K-NN for HeartDiseasorAttack

6.1.2 K-NN of HeartDiseaseorAttack and Stroke

We further performed k-NN by using both HeartDiseaseorAttack and Stroke as response variables, and we again separated the dataset into training set and testing set (70%:30%). We also iterated 15 k values and plotted the Accuracy vs. Number of neighbors plot, we can see that the best result is being captured when k = 14, which is 0.8789025543992431. Then, we used this score to perform the confusion matrix and classification report to visualize the accuracy of the prediction. For the confusion matrix, 0 represents (0,0), 1 represents (0,1), 2 represents (1,0), and 3 represents (1,1), and they are in the order of (Heart Disease, Stroke). For precision, it can classify 91% of the people that actually have no heart disease attack and no stroke correctly, while the classification accuracy for people having heart disease only or stroke only are 35% and 12% respectively, and the classification accuracy of people having both heart disease and stroke is only 11%. For recall, we can see that it works well with predicting the people that actually have no heart disease attack and no stroke correctly as it has a recall value of 1, while it is not accurate for predicting people that actually have a heart disease attack and stroke, or people that have either heart disease only or stroke only. Lastly, for F1 Score, we can further see that predicting people with no heart diseases and no stroke has a high accuracy of 94%, and that it is not useful for predicting people with heart disease attack and stroke, or people that have either heart disease only or stroke only.

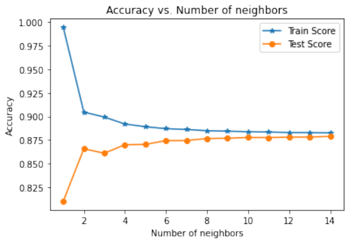
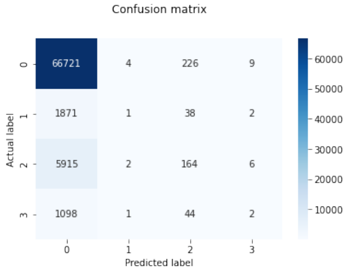
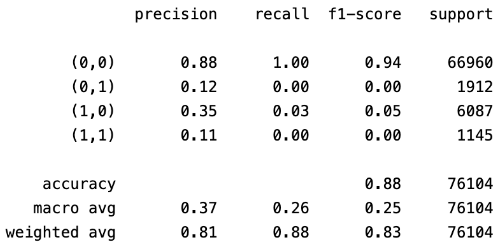


Figure 8: Accuracy vs. Number of neighbors



	0	1	2	3
0	66721	4	226	9
1	1871	1	38	2
2	5915	2	164	6
3	1098	1	44	2

Figure 9: Confusion matrix for HeartDiseasorAttack + Stroke



```
              precision    recall  f1-score   support

       (0,0)       0.88      1.00      0.94     66960
       (0,1)       0.12      0.00      0.00      1912
       (1,0)       0.35      0.03      0.05      6087
       (1,1)       0.11      0.00      0.00      1145

    accuracy                           0.88     76104
   macro avg       0.37      0.26      0.25     76104
weighted avg       0.81      0.88      0.83     76104
```

Figure 10: Classification Report of K-NN for HeartDiseasorAttack + Stroke

6.2 Logistic Regression

6.2.1 Logistic Regression of HeartDiseaseorAttack

To run a logistic regression model on the dataset, we first subsetted our data into our response variable and predictor variables. For our first process, we used all possible predictor variables given to us by the dataset in order to create a baseline for accuracy score. Then, we used cross validation to split the data into testing and training sets. For the first model, we used HeartDiseaseorAttack as the response variable. We used the training set in order to fit the data into a logistic regression model. Then, we predicted the response values of the test set based on the predictor values and compared that to the actual values. From there, we calculated the prediction score to compare accuracy and create a confusion matrix to visualize the accuracy of the model.

The prediction score of the first logistic regression model predicting HeartDiseaseorAttack is 0.9074555870913487.

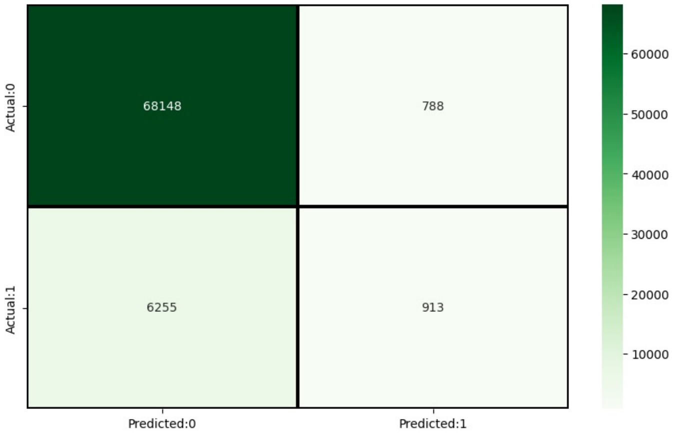


Figure 11: Logistic Regression Confusion Matrix of HeartDiseaseorAttack

6.2.2 Logistic Regression of HeartDiseaseorAttack and Stroke

Next, we wanted to see if there is a covariate dynamic between HeartDiseaseorAttack and Stroke through using those both as response variables for a multinomial logistic regression. We followed the same steps as the first model. Our results showed an accuracy of 0.8797960685377904. Below is the confusion matrix of the multinomial logistic regression model's accuracy for HeartDiseaseorAttack.

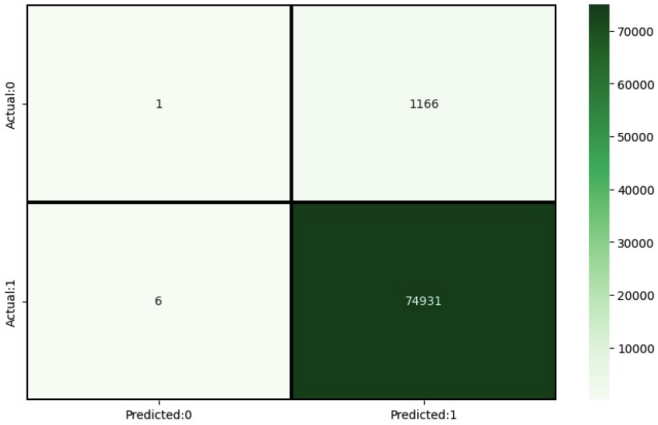


Figure 12: Multinomial Logistic Regression Confusion Matrix of HeartDiseaseorAttack

From this confusion matrix **(Figure 12)**, we can see that the model was very good at predicting when a subject has HeartDiseaseorAttack. However, there are many false positives that the algorithm predicted because it only predicted one person to not have HeartDiseaseorAttack that actually did not have it and predicted 1166 people to have HeartDiseaseorAttack when they actually did not. Although the logistic regression model has many false positives, having false positives in this instance is better than having false negatives. However, I would not recommend this method in order to predict HeartDiseaseorAttack.

Below **(Figure 13)** is the confusion matrix of the multinomial logistic regression model's accuracy for Stroke.

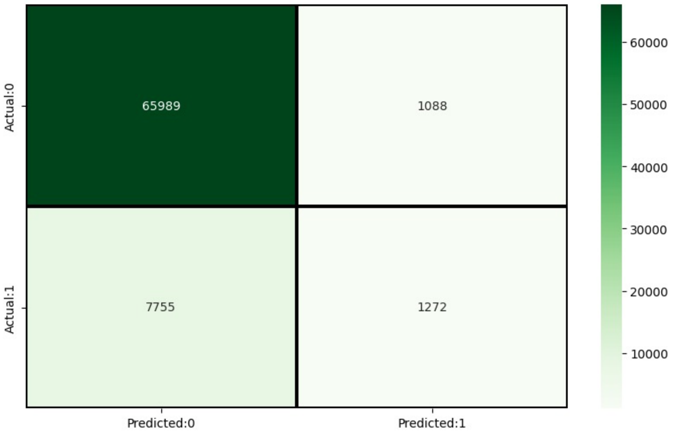


Figure 13: Multinomial Logistic Regression Confusion Matrix of Stroke

Contrary to how the multinomial logistic regression model predicted HeartDiseaseorAttack, it predicted many false negatives. This is dangerous because subjects could be under the idea that they will not get strokes based on the algorithm but end up getting them. From both confusion matrices, we concluded that the multinomial logistic regression model was not a good fit in predicting HeartDiseaseorAttack and Stroke.

Furthermore, we also performed a sensitivity analysis to assess the robustness of our logistic regression model. We varied the threshold probability for classification and observed how it affected the accuracy of the model. The results showed that our model can still provide accurate predictions even if the classification threshold is set at different values.

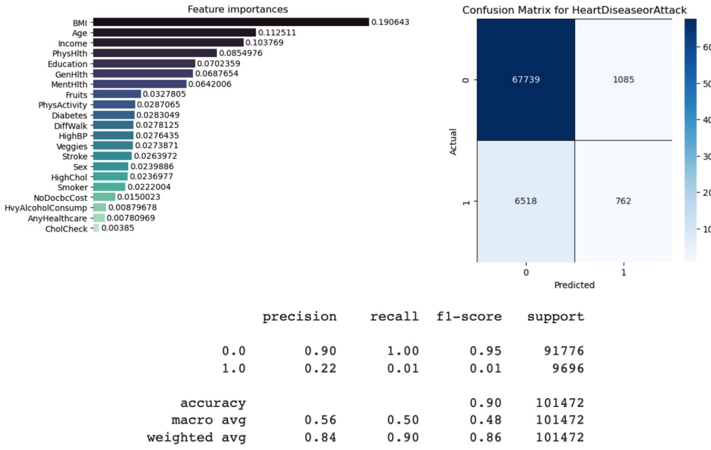
In summary, our study used logistic regression modeling to explore the relationship between heart disease, stroke, and their major risk factors. We first created a baseline model using all possible predictor variables, and then used sensitivity analysis techniques to improve the accuracy of the model. While a multinomial logistic regression model was not a good fit for predicting heart disease or stroke, our logistic regression model produced promising results and can provide insights into effective prevention strategies to reduce the incidence of heart disease and stroke in the United States.

6.3 Random Forest

6.3.1 Random Forest of HeartDiseaseorAttack

When performing Random Forest, we are hoping to predict factors that could affect Heart Disease or Attack. We decided to use accuracy scores and features importance to determine the factors. First, when accessing the data, we found that the dataset is tidy, clean and has zero missing values. We used a model evaluation procedure and decided to split our dataset into two pieces using a training set and testing set. The training set is used for the Random Forest model and the training set is used to evaluate the fit to our Random Forest model. We split the testing set and training set to a ratio of 30:70. To maximize our model, we will be looking at the classification metric to evaluate the algorithm, which consists of precision, recall, f-1 score, support, accuracy score and confusion matrix.

For the first model using HeartDiseaseorAttack as the response variable, we found that the accuracy score is 0.9012.



```
              precision    recall  f1-score   support

         0.0       0.90      1.00      0.95     91776
         1.0       0.22      0.01      0.01      9696

    accuracy                           0.90    101472
   macro avg       0.56      0.50      0.48    101472
weighted avg       0.84      0.90      0.86    101472
```

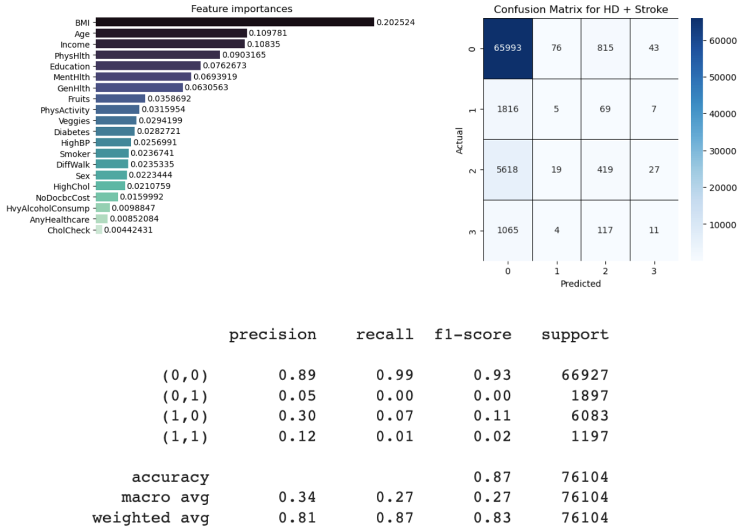
Figure 14: Importance Feature graph, Confusion Matrix and Classification Report of Random Forest for HeartDiseaseorAttack

- **Precision** is the fraction of retrieved values that are relevant to the data. The precision is the ratio of tp / (tp + fp)
- **F1-score** is the harmonic mean of precision and recall. Where an fscore reaches its best value at 1 and worst score at 0
- **Recall** is the fraction of successfully retrieved values that are relevant to the data. The recall is the ratio of tp / (tp + n)
- **Support** is the number of occurrences of each class in y test.

Looking at the importance feature for our response variable, BMI is the highest. In the classification report, both precision and recall are high for class 0 but low for class 1. From the classification report, the recall has a value of 1.00 and it tells us that of all individuals who actually had heart disease, the model correctly identified 100% of them which means the model has a high ability to detect positive cases. The F1-score is also relatively high with 95%, which is ideal for this model. By looking at the confusion matrix, the true positive shows that the model was able to predict that 67739 people who reported having heart disease actually do have indicators of heart disease.

6.3.2 Random Forest of HeartDiseaseorAttack and Stroke

Next, we will be using heart disease or attack and Stroke as our response variables since HeartDiseaseorAttack and Stroke are highly correlated towards one another. To perform the model, we fused response variable HeartDiseaseorAttack and Stroke together and found that the accuracy score is 0.872858.



```
              precision    recall  f1-score   support

       (0,0)       0.89      0.99      0.93     66927
       (0,1)       0.05      0.00      0.00      1897
       (1,0)       0.30      0.07      0.11      6083
       (1,1)       0.12      0.01      0.02      1197

    accuracy                           0.87     76104
   macro avg       0.34      0.27      0.27     76104
weighted avg       0.81      0.87      0.83     76104
```

Figure 15: Importance Feature graph, Confusion Matrix and Classification Report of Random Forest for HeartDiseaseorAttack + Stroke

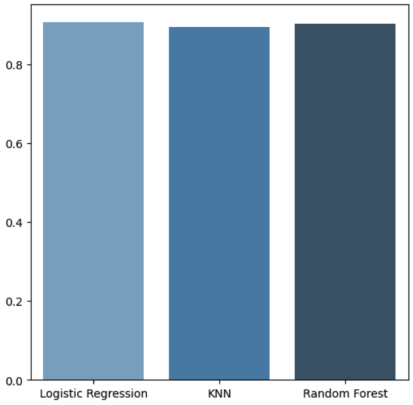
Similar to the first model, the importance feature for this model tells us that BMI is the highest which makes sense because having a high or low BMI could be a factor to Heart Attacks and Strokes. Looking at the confusion matrix and classification report, the score represents 0 represents (0,0), 1 represents (0,1), 2 represents (1,0), and 3 represents (1,1), and they are in the order of (Heart Disease, Stroke). For precision, it seems that the model can classify 89% of the people actually have neither heart disease and stroke, 5% of them have stroke, 30% of them have heart disease and only 12% of them having both heart disease and stroke. For recall, it seems that the model works well in predicting people with no heart disease and stroke. However, for people with heart disease and stroke are really low with an accuracy of 1%. Finally, F1 score has a similar result as recall, the model accuracy is high in predicting people with no heart disease and stroke but it is low for predicting people with both heart disease and stroke.

Overall, both models seem to have a good ability to detect people with no Heart Disease and Stroke since it has a high recall value and F1 score. However, the false positive is also relatively high, in this case we may need to adjust the model to prioritize recall or precision. That way we could predict people who has both Heart Disease and Stroke.

6.4 Classification

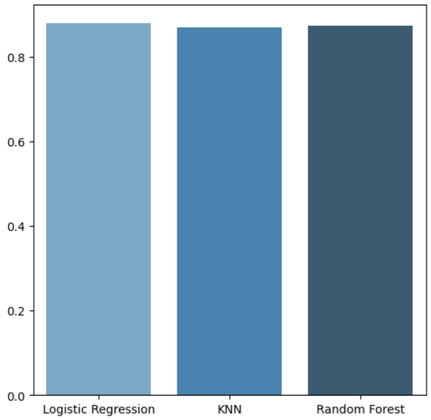
Machine Learning classification allows us to effectively make predictions from our dataset. From our dataset, we have found that there is no missing value. We decided to split our data into a training and testing set. Since we have sufficient data, we will split the data with a ratio of 70:30 for training and testing respectively.

We decided to look at three different predictive models to find the best models to use based on the accuracy each model gives us. After building all of our models, we can now see which model performs the best. From above we can see that Logistic Regression will be the best in predicting Heart Disease and both Heart Disease and Stroke.



```
{'Logistic Regression': 0.9068314411857458,
 'KNN': 0.8953405865657521,
 'Random Forest': 0.9024558498896247}
```

Figure 16: Predictive Model Accuracy score for Response Variable HD



```
{'Logistic Regression': 0.8789419741406497,
 'KNN': 0.8694286765478818,
 'Random Forest': 0.872779354567434}
```

Figure 17: Predictive Model Accuracy score for Response Variable HD + Stroke

7 Conditional Entropy and Contingency Table

Part I: Conditional entropy values for Heart Disease or Attack and Heart Disease or Attack + Stroke, sorted from lowest to highest.

We used conditional entropy to quantify the level of uncertainty in our dataset, calculating the conditional entropy values for all predictors with respect to both the response variable HeartDisease-orAttack and the fused response variable HeartDiseaseorAttack and Stroke. We then ranked these values from smallest to largest, with a focus on identifying the predictors with the lowest conditional entropy values. As the lower the value, the more predictable that independent variable is.

Our analysis revealed that when the predictors were fitted onto the response variable HeartDiseaseorAttack, all variables demonstrated relatively low conditional entropy values in the range of 0.4 to 0.45. This level of conditional entropy suggests moderate uncertainty in the predictive power of the predictors. However, when fitted onto the fused response variable, the conditional entropy values increased to a range of 0.617 to 0.676, indicating a relatively higher level of uncertainty.

Overall, our findings highlight the importance of choosing response variable when assessing the predictive power of predictors in our dataset. By selecting appropriate response variables, we can identify the predictors with a high predictive power and develop more accurate predictive models.

Indep. Variables	Conditional Entropy
GenHlth	0.40391188201780215
Age	0.4092402986338385
HighBP	0.4183520169099112
DiffWalk	0.42414012708185544
HighChol	0.42679414153717476
PhysHlth	0.42954996606384704
Diabetes	0.4314615128659573
Stroke	0.43150163116895246
Income	0.43564204845089943
Smoker	0.4408895526524299
Education	0.4433130644606779
Sex	0.44498596682676717
PhysActivity	0.445205307902124
BMI	0.4466642856440235
MentHlth	0.4469728133498111
CholCheck	0.4484162092066758
Veggies	0.44923826878202616
HvyAlcoholConsump	0.4496043657515937
NoDocbcCost	0.4496468455296574
Fruits	0.45000873012353737
AnyHealthcare	0.45001419887106564

Table 1: Conditional Entropy for Response Variable HD

Indep. Variables	Conditional Entropy
GenHlth	0.617410743597923
Age	0.6269558812515654
HighBP	0.637299946017686
DiffWalk	0.6398510858234705
PhysHlth	0.647888012991303
HighChol	0.649679834087291
Income	0.6543346940213803
Diabetes	0.6543711417199365
Smoker	0.66568153414044
Education	0.666973049185613
PhysActivity	0.6694522681616484
Sex	0.6708608046291098
MentHlth	0.6709428672703581
BMI	0.6719681599498072
CholCheck	0.6741852529776382
Veggies	0.6745662640747261
NoDocbcCost	0.6752233317765843
HvyAlcoholConsump	0.6756064628968568
Fruits	0.6760505214822498
AnyHealthcare	0.6760677134353594

Table 2: Conditional Entropy for Fused Response Variable HD + Stroke

Part II: Contingency Tables for some Significant Response Variables vs HeartDiseaseorAttack

We chose some predictors that have a relatively low conditional entropy value for HD response variables and chose the predictor with the lowest conditional entropy value to plot the contingency tables.

From Table 3, 4, and 5, we can obtain the odds of each categories of the predictor, which shows the ratio of the observations in the two categories of the response variable HeartDiseaseoAttack, which are non-diseased and diseased. We can further obtain the odds ratio of each pair of observation, which reveals the strength of association between the two variables. Since they all have an odds ratio larger than 1, we can conclude that they have a strong association with the response variable HeartDiseaseoAttack.

From Table 6, we can more directly to learn the patterns between one independent variable and the fused response variables HeartDiseaseorAttack and Stroke.

HD/HighChol	**-**	**+**	**Row-Sum**
non-diseased	138949	90838	229787
diseased	7140	16753	23893
Col-sum	146089	107591	253680
odds	0.051386	0.184427	3.589073 (odds-ratio)

Table 3: HD vs HighChol Contingency Table

HD/HighBP	**-**	**+**	**Row-Sum**
non-diseased	138886	90901	229787
diseased	5965	17928	23893
Col-sum	144851	108829	253680
odds	0.042949	0.197226	4.592099 (odds-ratio)

Table 4: HD vs HighBP Contingency Table

HD/Diabetes	**-**	**+**	**Row-Sum**
non-diseased	197027	32760	229787
diseased	13978	9915	23893
Col-sum	211005	42675	253680
odds	0.070945	0.302656	4.266085 (odds-ratio)

Table 5: HD vs Diabetes Contingency Table

HD+Stroke/GenHlth	**1**	**2**	**3**	**4**	**5**	**Row-Sum**
(-, -)	43929	83666	65579	23160	7098	223432
(-, +)	354	1290	2153	1682	876	6355
(+, -)	925	3754	6894	5431	2952	19956
(+, +)	91	374	1020	1297	1155	3937
Col-sum	45299	89084	75646	31570	12081	253680

Table 6: HD + Stroke vs GenHlth Contingency Table

8 Discussion of Data Analysis Result

The k-NN algorithm showed a high prediction score of 90.47% when predicting Heart Disease or Attack as a single response variable, and its accuracy when predicting with both Heart Disease or Attack and Stroke as response variables is 87.89% , a little lower than the prediction score of Heart Disease or Attack only.

Logistic regression showed a high prediction score of 90.75% when predicting Heart Disease or Attack. This model is a better fit for the data than linear regression due to the amount of binary predictor variables in the data set. However, the accuracy reduced when Stroke was added as a response variable.

Random Forest showed a high prediction score of 90% when predicting Heart Disease or Attack. However, the prediction score decreased to 87.29% when we predict both Heart Disease or Attack and Stroke as response variables.

9 Conclusion of Data Analysis

How well can the variables predict people getting Heart Disease or Attack in the United States?

The three methods that we used (k-NN, logistic regression, and random forest) all have prediction scores of around 0.90. Therefore, the data set includes strong predictors that may help predict Heart Disease or Attack in the United States; however, the inaccuracy is still very high for a serious issue. False negatives are dangerous in this prediction process because that means that the algorithms predicted someone to not have Heart Diseases or Attacks but they end up getting it. Therefore, it is dangerous because people cannot prepare or be warned of these health issues. As a result, we conclude that although the prediction scores are fairly high for all three statistical methods, it is not high enough to predict Heart Disease or Attack which is a serious issue that heavily affects lives.

What are the most significant variables that would affect Heart Disease or Attack and Heart Disease or Attack and Stroke based on the appropriate prediction model?

Based on the information from our random forest clustering method, the most significant variables that affect Heart Disease or Attack are BMI, age, and income. From these results, we can draw a conclusion that higher body mass and older age can lead to possible Heart Disease or Attack and/or Stroke. Surprisingly, income is also a strong predictor in Heart Disease or Attack and/or Stroke. It is possible that people who have lower incomes do not have as much access to proper healthcare, thus leading to higher chances of Heart Disease or Attack and/or Stroke.

Based on the conditional entropy part, the most significant variables that affect Heart Disease or Attack and Stroke are general health, age, and high blood pressure. The different results we got from different methods are probably due to the variance in training and testing sets and the possible outliers of the dataset.

In conclusion, the findings that BMI, age, income, general health, and high blood pressure are some of the significant predictors of Heart Disease or Attack and Heart Disease or Attack and Stroke in the United States have important implications for understanding and addressing this health issue. The high prevalence of obesity and high blood pressure, an aging population in the US, coupled with potential barriers to accessing healthcare for those with lower incomes, suggest that Heart Disease or Attack is likely to remain a significant health challenge in the country. The use of statistical methods such as k-NN, logistic regression, and random forest can provide insights into potential risk factors. However, it is clear that further research is needed to develop more accurate prediction models and improve our understanding of the complex factors that contribute to Heart Disease or Attack in the US.

Acknowledgments

We would like to thank our TA, Shuting Liao, for her guidance and support throughout this project. Additionally, we extend our appreciation to our groupmates for their invaluable contributions towards the success of this project.

References

[1] (n.d.). Number of deaths for leading causes of death – United States, 2015-2020. Retrieved May 7, 2023, from https://www.cdc.gov/nchs/data/health$_{p}olicy/Leading-Causes-of-Death-for-2015-2020.pdf.$

[2] Heart Disease and Stroke | CDC. (n.d.). Centers for Disease Control and Prevention. Retrieved May 14, 2023, from https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm.

[3] Heart Disease Facts | cdc.gov. (2022, October 14). Centers for Disease Control and Prevention. Retrieved May 7, 2023, from https://www.cdc.gov/heartdisease/facts.htm.

[4] Heart Disease Health Indicators Dataset. (n.d.). Kaggle. Retrieved May 14, 2023, from https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset.

[5] 3.1. Cross-validation: evaluating estimator performance — scikit-learn 1.2.2 documentation. (n.d.). Scikit-learn. Retrieved May 7, 2023, from https://scikit-learn.org/stable/modules/cross$_{v}alidation.html.$