Heart Disease Health Indicators

Laura Valenzuela, Thomas Venner

May 13, 2023

Abstract

Heart disease (HD) is the leading cause of death in the United States[3]. The occurrence of heart disease in an individual is associated with a wide array of factors and their complex interplay. Thus, an analysis of the main associated factors to heart disease can offer insight into predicting the occurrence of heart disease in an individual with a given set of characteristics. Educating the public on the main associated factors of heart disease can help at risk individuals take action to reduce their chance of heart disease, and healthy individuals avoid the factors that are associated with heart disease. The predictive heart disease model can inform an individual based on their unique set of characteristics of their predicted status of heart disease, and thus motivate an individual to take action to change their set of characteristics if heart disease is predicted.

1 Introduction

In this paper, we perform a thorough Exploratory Data Analysis (EDA) of heart disease data, including looking at relationships between predictor variables on a holistic level through hierarchical clustering, and measuring predictive power of each predictor variable through Shannon entropy information. Odds ratios are also employed to give a cursory look at the relationships between predictors and the the response variable, Heart Disease. We use our findings in the EDA process to fit random forest classifier models on subsets of the most predictive variables. We chose random forest classifier models because of their ability to partition the response space while taking interactions into account without interaction hyper-parameter specification [6]. This ability is crucial given that interactions between predictors are usually common and pronounced in demographic and health data sets. We then use methods such as Cross Validation (CV) and model accuracy measurements to determine the predictive model with the highest accuracy. We then extract the most important predictors of the predictive model to uncover the factors most highly associated with heart disease.

2 Data Description

Our data set on heart disease comes from the Behavioral Risk Factor Surveillance System (BRFSS) from the CDC. BRFSS is a health-related telephone survey which collects health information from about 400,000 Americans each year. We will use the 2015 version of the telephone survey data, which has been cleaned and had its predictor pool narrowed down based on subject matter health knowledge of heart disease.

Initially, our cleaned data set (N = 253,680) was composed of a mixture of mostly binary categorical variables, with some continuous, ordinal, and discrete variables. For sake of analysis and interpret-ability, we changed all variables into binary categorical or ternary ordinal variables. We researched reasonable cutoffs to use to transform non-categorical variables into ternary ordinal variables, which we will list below. The superscripts denote reference numbers to research sources we used to chose the cutoffs. The naming of the variables is such to facilitate coding efficiency and readability. The prefixes 'Low', 'Med', and 'High' denote the magnitude of the ordinal variable name to which they are attached. Variables with no prefix are binary, taking value 1 if the variable name is fulfilled and 0 otherwise. If a variable has prefixes 'Yes' or 'No' it is a

labeled level of the binary variable (except for the variable 'NoDocBcCost' which includes the 'No' at the beginning by default)[1].

The first pattern to note about the data set is the strong class imbalance within the HD response variable as shown in Fig. 1.

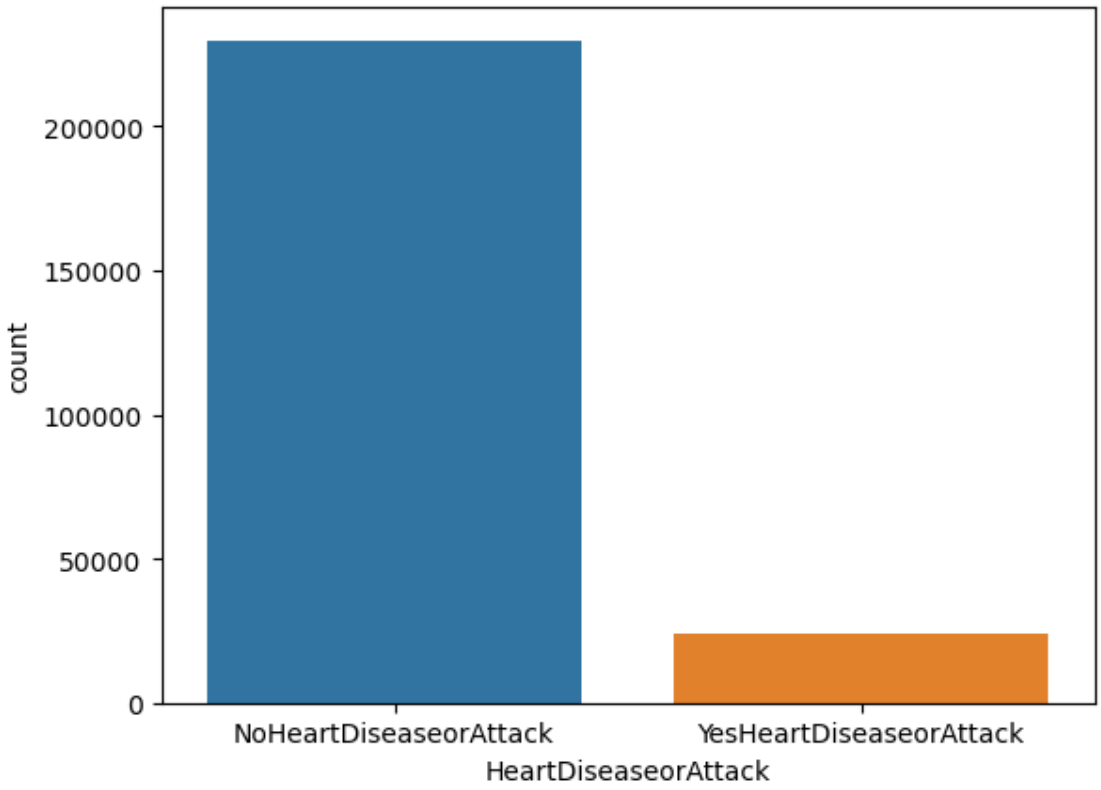

Figure 1: Histogram of class imbalances in unfused response

This it typical of health data sets, in general, most people are healthy and do not have the disease of interest. This class imbalance in the response variable is taken into account as we perform our EDA and model fitting, particularly in the EDA discussion of local Shannon conditional entropy on which predictor levels to value. In additional to the HD response variable, we are interested the fusion between the binary encoded variable 'Stroke' and HD as a response variable, as this variable encodes all binary combinations of having stroke and heart disease. We are interested in this fusion because these diseases are related. The histogram of this fused response variable is shown below in Fig.2.

[1]For more information on the variables, please refer to the data accessibility section with a link to the data set at the end of this paper. For specific cutoffs of variable transformations, please refer to the code in a separate file.

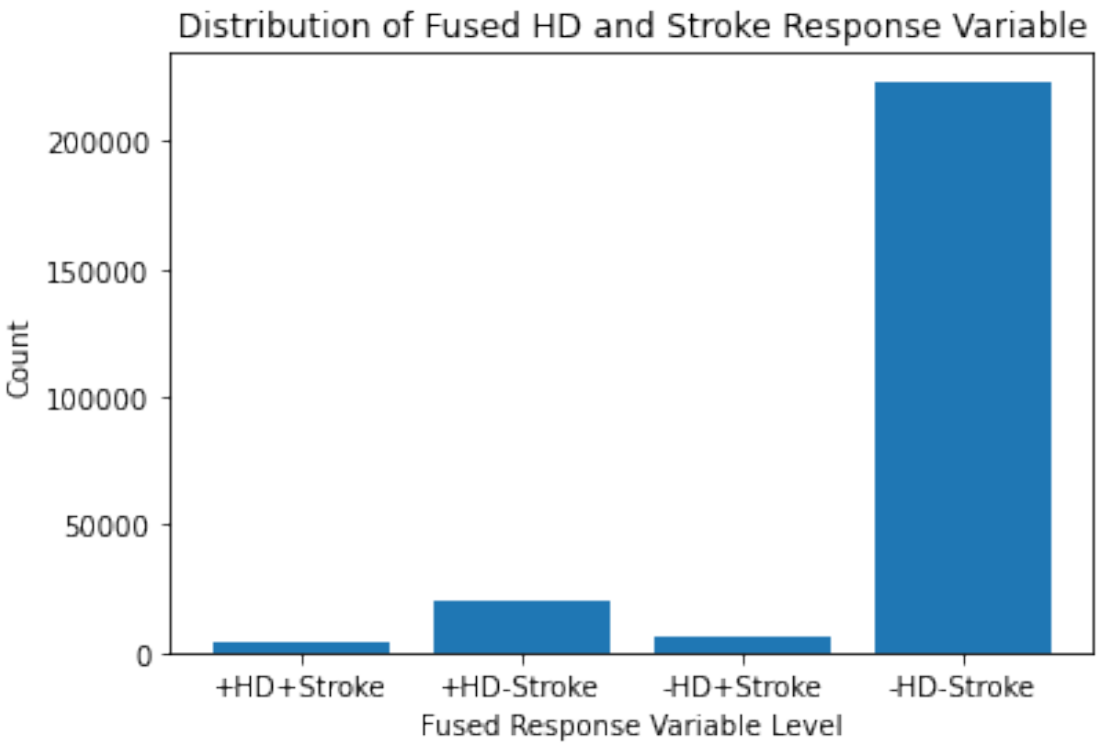

Figure 2: Bar Chart of Class Imbalance In The Fused Response Variable

3 Exploratory Data Analysis

3.1 Hierarchical Clustering (HC)

To explore the complex relationships between all variables in the data set, we use hierarchical clustering (HC). Hierarchical clustering is an agglomerative algorithm which relates variables by the distances between their column profiles. The output of (HC) for the response variable HD, shown below (Fig. 3), is a bifurcating tree structure which has most similar variables located in adjacent clusters and leaves. The colors visually aid in identifying the main clusters of the dendrogram. The nodes show how each cluster is related to other clusters.

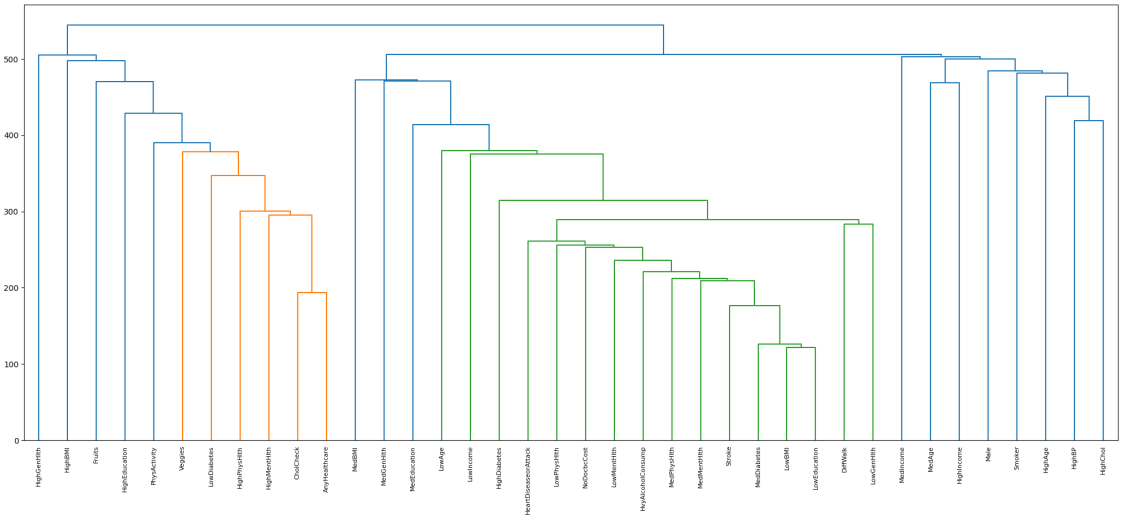

Figure 3: HD Response Dendrogram

Note that the HD response variable is most similar to Low Reported Physical Health, and is in the same cluster as 'NoDocbcCost', 'LowMentHlth', 'HvyAlchoholConsump', 'MedPHysHealth', 'MedMentHlth', 'Stroke', 'MedDiabetes', 'LowBMI', and 'LowEducation, which are listed in order of decreasing similarity. This initial look at the relationships between the response and other variables gives us a clue as to which variables could be the most related to the response.

In a similar manner, we observe (in Fig.4) the dendrogram resultant from HC clustering on the data containing the fused response (and dropping the variables which combined to create the fused response).

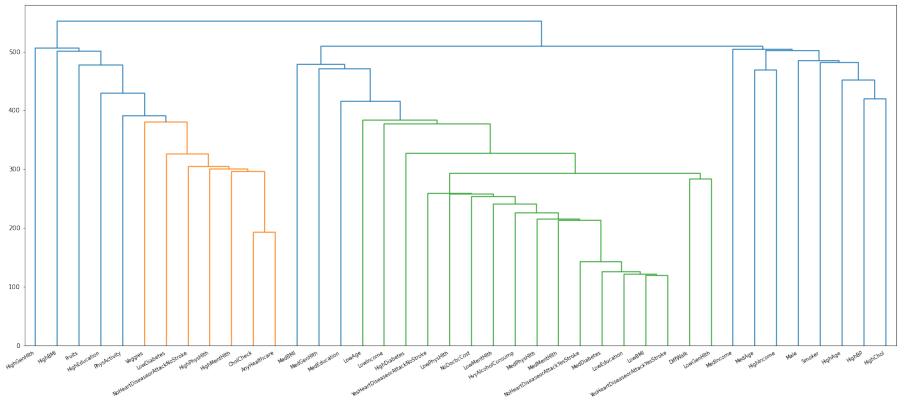

Figure 4: Fused Response Dendrogram

Note that the levels of the fused response variable in which at least one of the diseases is present are all grouped within the same cluster, and the other variables in this cluster which are most similar are 'LowPhysHlth', 'NoDocbcCost', 'LowMentHlth', 'MedPhysHlth', 'MedMentalHlth', 'MedDiabetes', 'LowEducation', and 'LowBMI'. Comparing these results to the dendrogram of the HD response variable, we see that the levels of the variable 'MentHlth', and the variables 'NoDocBcCost', 'LowEducation', and 'LowBMI' are included as being most similar to the HD response variable and the fused HD+Stroke response variable.

3.2 Odds Ratios

Stepping away from dendrograms, we now analyze the response variable HD from an odds-ratio standpoint. That is, we compare each predictor variable to the response variable using a contingency table. Then, we calculate the odds ratio of the odds of having HD based on one level of the predictor, over the odds of not having HD based on another level of the predictor. Odds ratios greater than 1 suggest that that particular level of the predictor variable increases the chance of having heart disease when compared with the chance of having heart disease at a different level of the same predictor variable. Odds ratios in descending order are shown in the histogram below. Note the meaning of the x-axis labeling. If the odds ratio of a binary variable is shown, then the odds of heart disease in the level of the binary variable corresponding to True or 1 is in the numerator, with the odds of heart disease in the level of the binary variable corresponding to False or 0 is in the denominator. If the odds ratio of a level of a ternary ordinal variable is shown, the odds ratio schemes are as follows:

HighVariable Numerator: HighVariable Odds of HD, Denominator: LowVariable Odds of HD

MedVariable Numerator: HighVariable Odds of HD, Denominator: MedVariable Odds of HD

LowVariable Numerator: MedVariable Odds of HD, Denominator: LowVariable Odds of HD

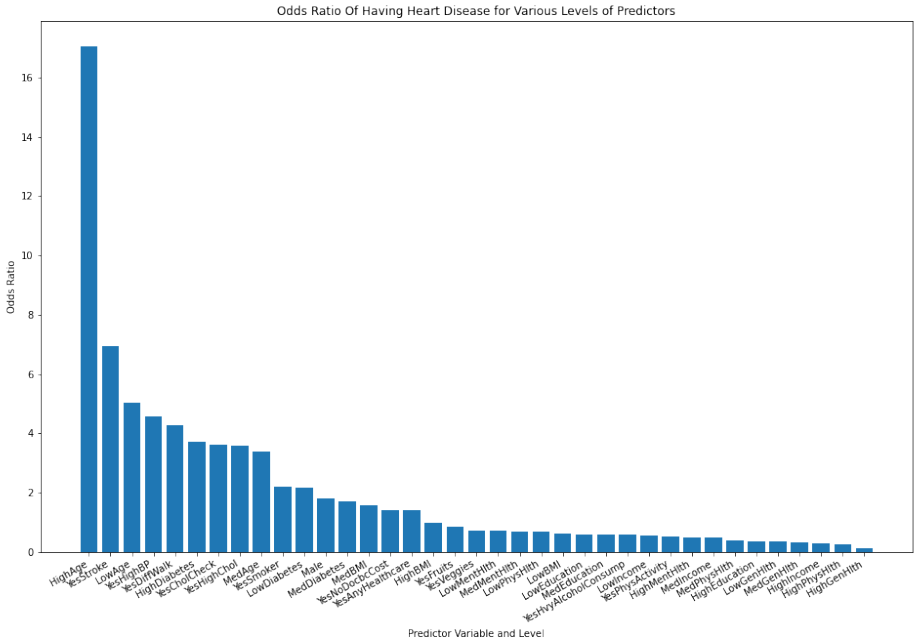

Figure 5: Odds Ratio for HD Response

Looking at the odds ratio bar chart in Fig. 5, we see that the odds of people over 60 ('HighAge') getting heart disease is 17 times the odds people less than 40 ('LowAge') getting heart disease. We note that the next factor level which increases odds of heart disease the most is 'Stroke'. We also note that the third factor level increasing odds of heart disease the most is middle age people compared to people of younger age. We conclude that the various levels of the predictor age seem to have the largest impact on whether a person has heart disease or not, and note the other predictor levels that appear high on the odds ratio bar chart.

3.3 Shannon Conditional Entropy

For our final EDA technique, we will analyze the Shannon conditional entropy of the predictors in relation to the response and the fused response on the global and local level, and taking into account interactions of the predictors with the highest mutual information with the response of interest. We chose to use Shannon condition entropy in our EDA because it can be readily used for predictor selection and data subsetting with which we can fit and compare our models.

3.3.1 Note about Displayed Shannon Entropy Values

All the Shannon entropies (conditional and unconditional) we display are the averages of a distribution of Shannon entropies for that particular predictor variable or predictor variable level. These entropy distributions are sampled from a multinomial distribution parameterized by observed values in the contingency table between the given predictor and response. Specifically, the vector of proportions parameterizing the multinomial distribution is the proportions in the two-way contingency table corresponding to the response variable given the predictor or level of the predictor. The N parameter of the multinomial distribution is the marginal corresponding to the vector of observed proportions in the contingency table. Each entropy distribution is the result of 1000 samplings from its specific parameterized multinomial distribution. Since the distributions had very low standard deviation ($\sigma \approx 0.001$), comparing the averages of these distributions when comparing entropy between predictors is sufficient. The word 'average' will be dropped for readibility, but anytime an entropy value is mentioned, it is the average value of the entropy distribution, unless otherwise stated. We calculated global Shannon conditional entropies to compare predictors which give the most information about the response, and the local Shannon conditional entropies to to compare predictor levels that gave the most information about the response. We also will explore Shannon entropy measures concerning interactions between the most important (highest information containing) predictors and predictor levels. The equations for calculating the global (2) and local (1) Shannon conditional entropies are shown below:

$$H(Y|X=x) = -\sum_{y \in Y} Pr(Y=y|X=x) log_2 Pr(Y=y|X=x) \tag{1}$$

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X=x) \tag{2}$$

3.3.2 Shannon Conditional Entropy Observations

First, we will observe the mutual information of each predictor at a global level for the unfused HD response variable. The mutual information $I(Y,X)$ for a predictor is given by the difference between the entropy of the response and the conditional entropy of the predictor as shown in (3), and is a measure of how predictive of the response a given predictor is.

$$I(Y,X) = H(Y) - H(Y|X) \tag{3}$$

Looking at the bar chart below (Fig. 6), we note the predictors containing the most mutual information.

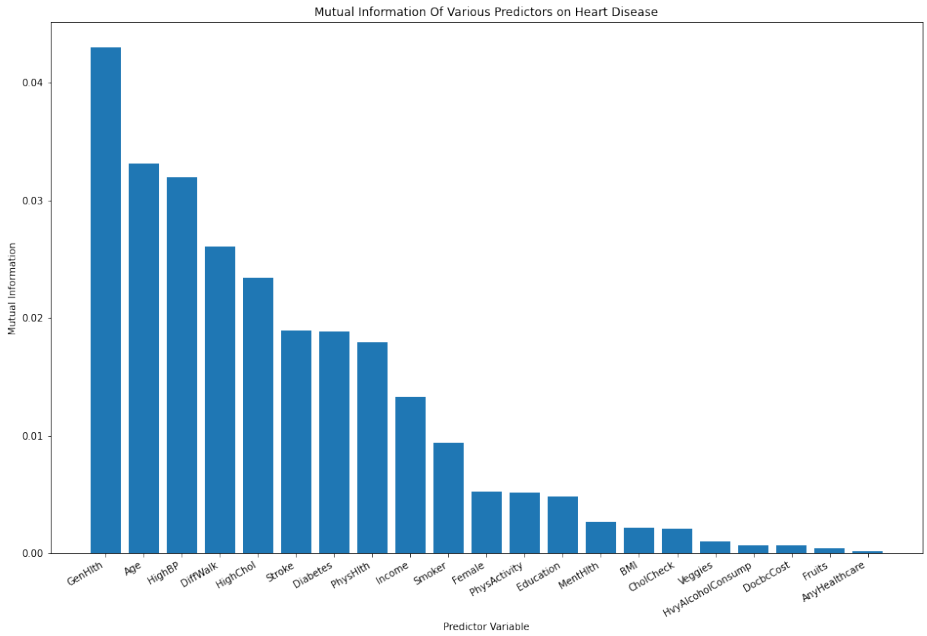

Figure 6: HD Response Mutual Information

We note that variables such as 'GenHlth', 'Age', and 'HighBP' make up the top three most informative variables on the response. Indeed, a person's age and high blood pressure are major risk factors for many diseases. Interestingly, 'GenHlth', which a more subjective measure in which participants reported on their general impression of their overall health, provides the most predictive information on the response. This suggests that survey participants' feelings about their general health is one of the most accurate predictors of heart disease.

Next, we'll look at the mutual information of predictors on the fused response of HD and Stroke, as shown below in Fig. 7:

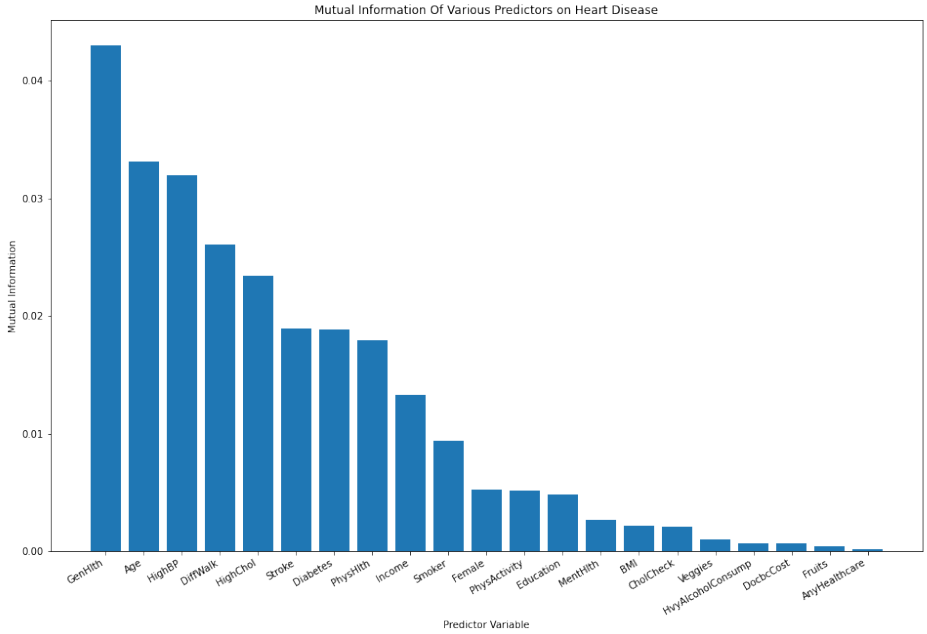

Figure 7: Fused Response Mutual Information

We note that the top 5 predictor variables in mutual information content on the fused response are identical to the non-fused HD response, and that the shape of the bar charts are similar.

Next, we will explore the predictors on a more fine grain level, which is at the level of each predictor. To do this, we look at the difference in entropy between the predictor level, and the baseline entropy, which is the entropy intrinsic to the response of interest. The difference in entropy equation is shown in (4). We chose to use difference in baseline entropy instead of mutual information, because at the local level, mutual information is not constrained to be greater than 0, and is thus less informative. Difference in baseline entropy is also a more straightforward interpretation of entropy values at levels of a predictor.

$$H_{diff} = H(Y|X = x) - H(Y) \tag{4}$$

Note that in contrast to mutual information, it is possible for the difference in local conditional entropy for a predictor to be positive or negative. Large positive values indicate that that particular level of the predictor has a much larger conditional entropy value than the entropy inherent in the response, and very negative values indicate the opposite. One might be tempted to pick only the predictor levels which have much reduced conditional entropy compared to the response, as these contain levels contain more information about the response. However, the predictor levels which have much greater conditional entropy compared to the response are of predictive value here as well. This is because the cause of their increased entropy is that they change proportions of the response variable the most and in the direction of balancing the proportions of the response. For example in the case of the HD response, predictors that greatly increase the odds of having heart disease increase the proportion of people with heart disease while decreasing the proportion of people without heart disease. Since entropy is greatest when the class proportions are equal, the conditional entropy of these predictors increasing the risk of people having heart disease is larger than baseline. Despite their increased conditional entropy due to the nature of how entropy is calculated and the strong class imbalance present in the response variable, we are still interested in these predictors as their change in entropy from baseline is indicative of their predictive power. Therefore, we look for variables whose absolute magnitude difference from baseline is largest.

Observe the bar chart of the entropy of predictor levels for the unfused case below in Fig.8:

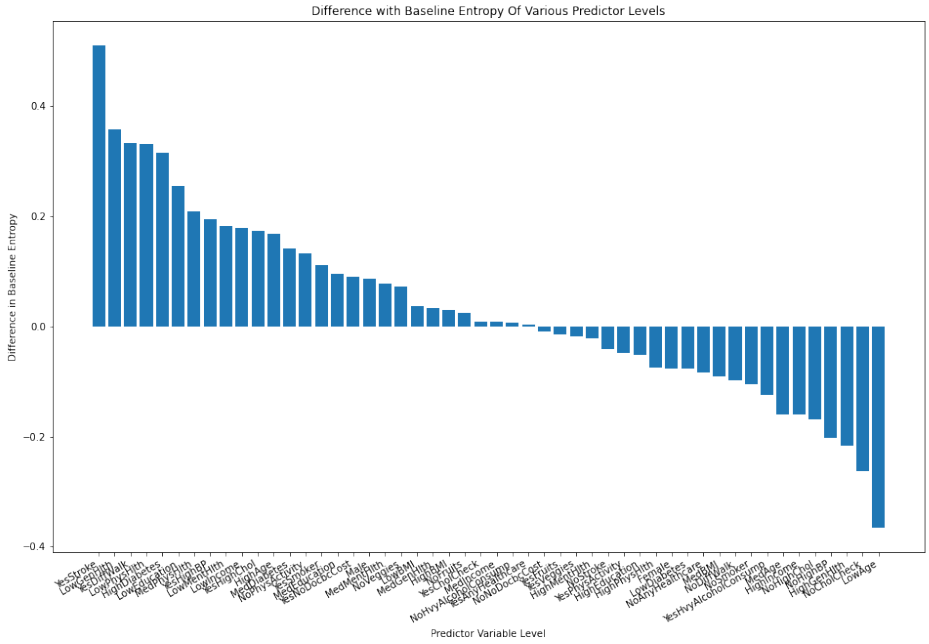

Figure 8: HD Response Difference in Baseline Entropy

As stated previously, the predictor levels with large positive values tend to appreciably increase the odds of heart disease, whereas the predictor levels with very negative values tend to appreciably decrease the odds of heart disease. Both of these types of predictors are of interest for their predictive power. We note that predictor levels with the largest positive difference in entropy are 'YesStroke' (the presence of stroke), 'LowGenHlth', 'YesDiffWalk', 'LowPhysHlth', and 'HighDiabetes'. Note that these levels of predictors come from predictors which we saw previously had the highest mutual information with the response at the global level. The predictor levels with the most negative difference in entropy are 'LowAge', 'NoCholCheck', 'HighGenHlth', 'NoHighBP', and 'NoHighChol'. Note 'HighGenHlth' is a complement level of 'LowGenHlth' meaning that 'HighGenHlth' decreases the odds of heart disease substantially whereas 'LowGenHlth' has the opposite effect of similar relative magnitude. Note that the negative difference entropy values are also levels of predictors who had the most mutual information at the global level.

Now observe the bar chart of entropy predictor levels for the fused response variable (Fig. 9).

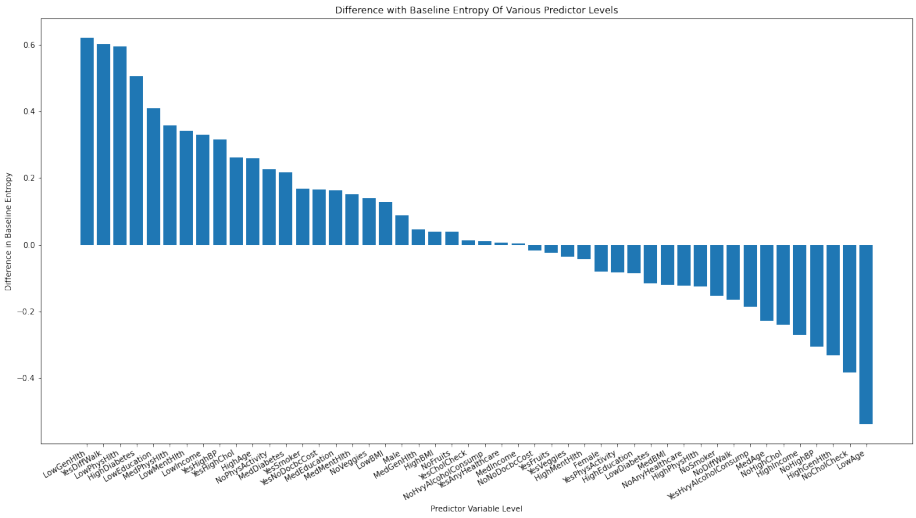

Figure 9: Fused Response Difference in Baseline Entropy

Note that the most extreme variables (including all variables mentioned explicity in the discussion of the local entropy of the non fused response variable) are in identical orderings and similar shape, except for the omission of 'stroke' since it was dropped to make the fused response variable, and the swapping of orderings of 'NoHighChol' and 'HighIncome' among the most negative predictor variables.

After reviewing the global and local entropy for the fused and unfused response, we two general observations. First, the most important predictor variables and levels globally and locally tended to be almost identically in the same order and have similar shape of mutual information when comparing the fused response to the unfused response. Second, the most important predictor levels tended to come from the predictor variables with the highest information. Hence, global information on conditional entropy is useful for selecting the best predictors to use for predictive modeling, and the local information is useful in selecting which levels of those predictors to use for predictive modeling. We will follow this logic when selecting subsets of variables for our random forest prediction models, and this illustrates how EDA can be used to generate better models.

3.3.3 Shannon Conditional Entropy Interactions

Since there are many possible pairwise interactions of predictors and their levels that are possible to pursue using entropy measures, and random forest classification does not need interaction hyper-parameter specification, we omit entropy interaction calculations from this paper for sake of brevity and concise reporting. We encourage you to look at our coding file which contains entropy tables exploring the interaction between the fused predictor variable 'GenHlth' and 'Age' two of the most predictive variables, and their levels.

3.4 Brief Summary of Observations from EDA

For the HD response, the clustering relationships from Hierarchical Clustering were much different than the predictors that the odds ratio and the Shannon conditional entropies at the local and global level deemed most predictive and similar to the response. This discrepancy can be explained by the fact that hierarchical clustering is an EDA tool to explore the the global relational similarity in shapes of distributions for each predictor variable[4], whereas odds ratios and Shannon conditional entropies are more concerned with predictive capabilities based on each predictor's individual relationship with the response. Additionally,

we did not use bootstrapping or re-sampling techniques to tease out the deterministic tree structures from the stochastic, simply due to the practical reason that we had a diminished group size and thus less teammates to carry out more analysis.

For the fused response, the Shannon conditional entropies at the local and global level aligned a bit more with the dendrogram in the predictors that were most similar to the response, but not by much.

4 Predictive Modeling: Random Forest Classification

As mentioned in the EDA section, we note the predictors with the highest mutual information, and predictor levels with the highest entropy difference from baseline when determining the subsets of predictors which we choose from the data for predictive modeling. Instead of comparing model accuracies or Cross Validation (CV) scores, we compared the average Out Of Bag (OOB) error between random forest model fits. The OOB error can be shown to approach the LOOCV generalization error for large N[6]. Any accuracy scores stated are OOB generalized error accuracy scores. Again we note that the random forest classifier is an effective classifier given the dataset, because it takes into account interactions without prior interaction hyper-parameter specification [7].

4.1 Overview of Subset Selection

For the modeling process below, we initially choose the top predictive variables and predictive levels based on entropy for the first model fit. For the second model fit, we dropped one of predictor levels that were associated with the same predictor, and were less important than the other predictor level as rated by the impurity reduction. We did this for all applicable predictors. This is because predictor levels from the same predictor are heavily dependent, and thus it is not necessary to have them both in the model (and having both in the model also reduces accuracy). For subsequent fits, we experimented with adding or subtract predictor variables based on entropy scores to determine the most parsimonious yet most accurate model. These competing model fits are displayed as additional EDA, but the best model fit is denoted and analyzed.

4.2 Subsetting the data

We ended up with two datasets. One was made up of the unfused variables with 'Heart Disease or Attack' being the target variable. And the other dataset was made up of the fused variables and we had four target variables this time. They where 'HD+Stroke', 'HD+NoStroke', 'NoHD+Stroke', and 'NoHD+NoStroke'. After getting the datasets, the next thing was to make subsets of the data using different variables. We chose the variables by entropy and made many subsets with different variations of the variables. We ran the model on them. Once we had our results, we put together the variables that were more important from the previous subsets and ran the model.

4.3 Random Forest Classification

Random forest is a way of averagng multiple deep decision trees, trained on different parts of the same training set, with the goal of overcoming the over-fitting problem of individual decision tree. It is an ensemble learning method for classification [1]. We applied the Random Forest Classification model to each subset data. And evaluated the importance of features on an artificial classification task. The blue bars are the feature importances.

4.4 Unfused Response Variables

We did five subsets of the unfused response variables. The first subset has nine variables. We fitted the model and plotted the feature importance on it. We got a 90.41% accucary. We discovered that the last

three variables were not very helpful for our model. So, we decided to drop them. After we modeled the second subset which included only the variables that were meaningful in our last subset. Which you can see below in Fig.10.

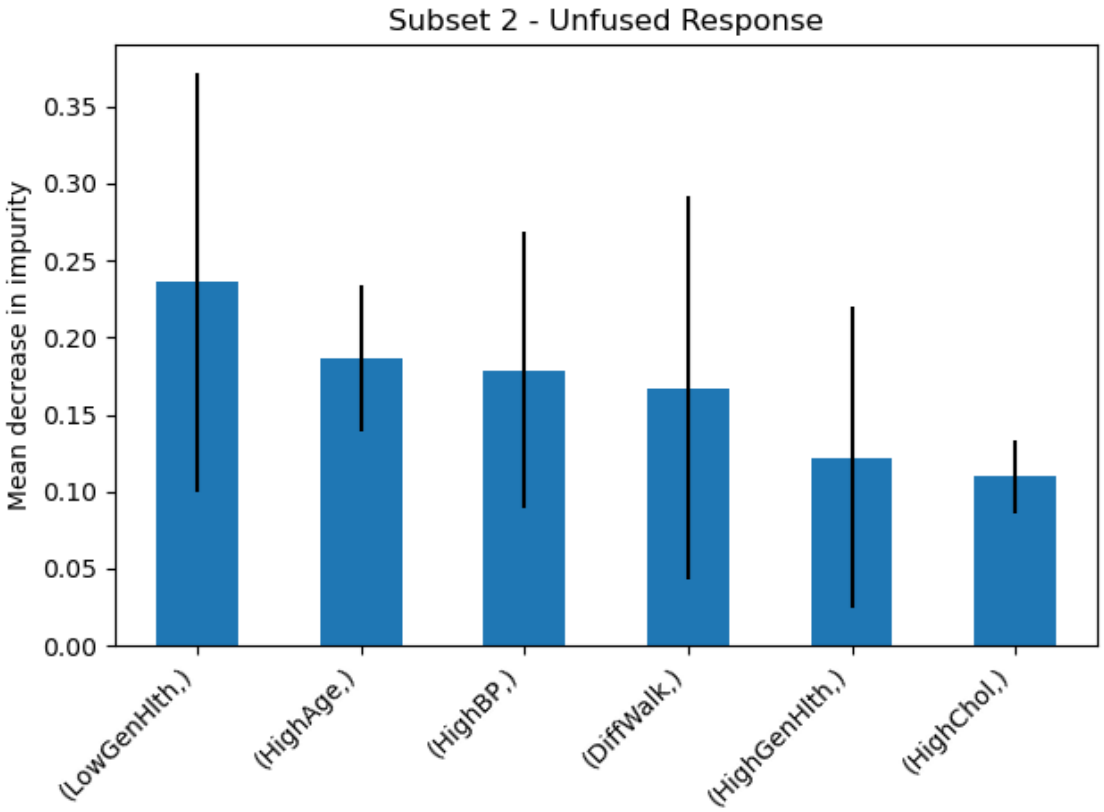

Figure 10: Subset 2 - Unfused Response

This modeled performed better with an accuracy of 90.57%. After seeing the performance we wanted to see if the subset would be better if we added the gender variable, and we got 90.55% accuracy. This means that gender is not as influential as we thought and our second subset is still the better model for us. For subset 4, we added other different variables and the accuracy went down to 90.49%. We noticed that adding more variables was not helping our model. For our last subset we removed 'HighChol' from subset 2 to see if fewer variables performed better and we got 90.54% accuracy. At the end, our chosen model was subset 2 - Unfused Response. After running the Random Forest Classifier in our subsets, we managed to get a 90.57% accuracy. This was because we decided to drop the variables that did not brought much importance to our model and kept the ones that performed better when using importance features. We noticed that the less variables there were, it was easier and simpler to interpret the model till a certain point where the accuracy was going down. But overall the accuracy of all of our subsets was consistent.

4.5 Fused Response Variables

We also did five subsets of the fused response variables. The first subset has eleven variables. We fitted the model and plotted the feature importance on it. We got a 87.50% accucary. We noticed almost half of the variables were not very helpful for our model. So, we decided to drop them. After we modeled the second subset which included only the variables that were meaningful in our last subset which where six. We see the

accuracy go a little bit high to 88.09%. Just like in the unfused variables, the accuracy grows higher with fewer variables. For our subset 3 we added the gender variable and the accuracy went down to 87.66%. And we concluded that the gender was not as important as some of our other variables like high bloop pressure and low general health.

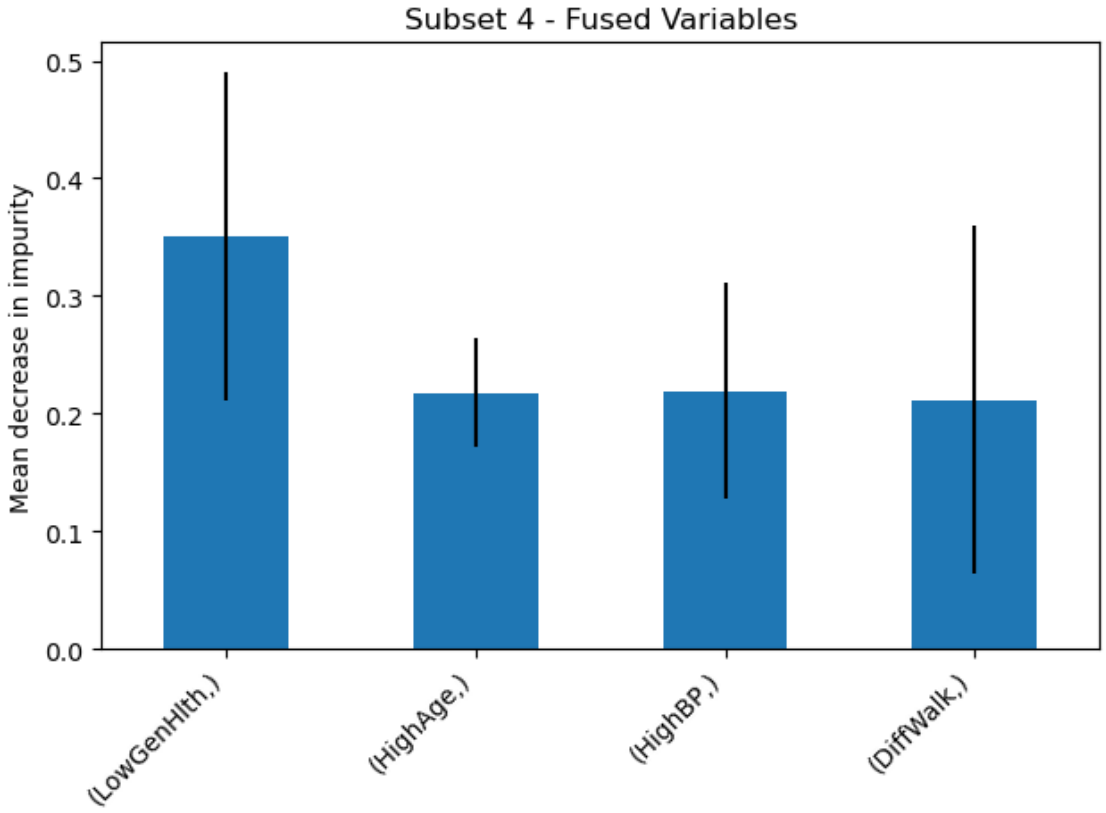

Figure 11: Subset 4 - Fused Response

In Fig. 11, we can see our subset 4. We reduced the model even more by having only four variables and the accuracy that we got was 88.20%. Which is the highest we have gotten so far. Making it a distinction from the unfused variables response where the model accuracy was going down if it had five or less variables. For our subset 5, we added the education and the income variables and our accuracy went down again to 87.75%. This concludes that our best model is Subset 4 - Fused Variables.

After running the Random Forest Classifier in all of our fused variable data we had an 85% accuracy. And after we ran it in our subsets, we managed to get a 88.20% accuracy with our fused variables. The most insightful variables were 'LowGenHlth', 'HighAge', 'HighBP', 'DiffWalk'. This model gave us the best accuracy. We dropped variables that were not very important in our model, and we managed to increased the performance by subsetting the data.

5 Conclusion

In conclusion, we first noted the strong class imbalance within the levels of the heart disease variable. We performed EDA via Heirarchical Clustering, Odds Ratios, and Shannon entropy (conditional and unconditional)

calculations on the HD response variable as well as the fused response variable. These EDA techniques allowed us to draw insights into the most predictive variable and variable levels in trying to build a predictive model for heart disease. Some of the most predictive variables include 'Age', 'Genhlth', 'HighBP', 'DiffWalk', 'Stroke', 'CholCheck', 'HighChol', and 'LowPhysHlth'. We chose to use a random forest classifier model because of its ability to pick up interactions without prior hyper-parameter specification. We used the variables and variable levels which had the lowest Shannon conditional entropy in choosing the subset of the predictors to use to fit the random forest models. We subtracted the least important predictors from each model fit based on impurity measures until we reached the most parsimonious yet highest accuracy based on OOB score classifier. In the case of the HD response variable, the variable and variable levels contained in the best model are, in order of decrease in impurity, 'LowGenHlth', 'HighBP', 'HighAge', 'DiffWalk', and 'HighGenHlth'. In the case of the fused response variable, the variables in the best model are the same (and in the same order) as the variables in the HD response variable.

Data Accessibility

The Heart Disease Health Indicators dataset is available at: https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

References

[1] Bhalla, Deepanshu. Listen Data. (2022). https://www.listendata.com/2014/11/random-forest-with-r.html. 14 05 2023.

[2] Developers, Scikit-learn. Scikit Learn. n.d. https://scikit-learn.org 14 05 2023.

[3] "FASTSTATS - leading causes of death," Centers for Disease Control and Prevention, https://www.cdc.gov/nchs/fastats/le causes-of-death.htm (accessed May 7, 2023).

[4] E. Chou, Y.-C. Hsieh, S. Enriquez, and F. Hsieh, "Evaluating reliability of tree-patterns in extreme-k categorical samples problems," Journal of Statistical Computation and Simulation, vol. 91, no. 18, pp. 3828–3849, 2021. doi:10.1080/00949655.2021.1951266

[5] Géron, Aurélien. Machine Learning with Scikit-Learn, Keras and TensorFlow. Sebastopol: O'reilly, 2018.

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, 2nd ed. Boston, Massachusetts: Springer, 2022.

[7] M. N. Wright, A. Ziegler, and I. R. König, "Do little interactions get lost in dark random forests?," BMC Bioinformatics, vol. 17, no. 1, 2016. doi:10.1186/s12859-016-0995-8