Statistical analysis on key indicators of Heart Disease

Alfredo Funicello

July 20, 2022

Abstract

The following analysis takes place on survey lifestyle and health markers, gathered by the CDC on USA citizens, to explore the possibility of exploiting the data by employing supervised statistical learning techniques with the objective of pre-emptive diagnosis of Heart diseases in individuals. The analysis is composed by a deep exploration of reported incidence of heart diseases and the relationship with the other gathered markers; then the use of supervised classification techniques, such as Logistic Regression, Decision Trees and Random Forests, is examined to estimate the feasibility of a risk category prediction framework.

1 Problem & Data Presentation

Heart disease is the leading cause of death for men and women in the United States, one person dies every 36 seconds from a cardiovascular related disease and the number of caused yearly deaths is about 659000, which is 1 in 4 deaths. The estimated cost impact of the pathology on the USA economy between 2016 and 2017 is about $363 billion, constituted by health care services, medicines and lost productivity due to death[1].

The CDC definition of "heart disease" refers to several types of heart conditions. The most common type of heart disease in the United States is coronary artery disease (CAD), which affects the blood flow to the heart; decreased blood flow can in turn lead to heart attacks.

Symptoms of heart disease may be silent and the pathology might remain undiagnosed until the individual experiences major signs of a heart attack, heart failure or arrhythmia. Several medical conditions and lifestyle choices have been shown to be drivers of higher risk for heart disease; these information can be exploited by statistical methods for preemptive diagnosis and risk assessment studies.

The dataset analyzed in this study has been uploaded under the name "Personal Key Indicators of Heart Disease" on the Kaggle platform. The dataset consists of a split from the data obtained by the Behavioral Risk Factor Surveillance System (BRFSS) telephone surveys, employed to gather data on the health status of U.S. residents and conducted by the CDC annually. The data has been acquired through the entirety of 2020.

The dataset is composed by 319795 rows and features the following variables:

- HeartDisease, boolean dependent variable that represents respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
- BMI, Body Mass Index, discrete variable
- Smoking, boolean, the respondents answer to the question "Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]"
- AlcoholDrinking, boolean, heavy drinkers: adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
- Stroke, boolean answer to "Ever had a stroke?"

[1] Heart disease Facts, CDC https://www.cdc.gov/heartdisease/facts.html

- PhysicalHealth, discrete, "Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days)"
- MentalHealth, discrete, "Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days)"
- DiffWalking, boolean, "Do you have serious difficulty walking or climbing stairs?"
- Sex, boolean, "Are you male or female?"
- AgeCategory, categorical variable composed by 16 age ranges
- Race, categorical, the respondent race
- Diabetic, boolean, "Ever had diabetes?"
- PhysicalActivity, boolean, Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
- GenHealth, ordinal categorical variable, answer to the question "Would you say that in general your health is", values: "Excellent", "Very good", "Good", "Fair", "Poor"
- SleepTime, discrete, "On average, how many hours of sleep do you get in a 24-hour period?"
- Asthma, boolean, "Ever had asthma?"
- KidneyDisease, boolean, "Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?"
- SkinCancer, boolean, "Ever had skin cancer?"

2 Data Manipulation

The data has already been cleaned and doesn't present any missing or incoherent values.

The variable `Diabetic` is composed by 4 values in the initial dataset: `"No"`,`"Yes"`,`"No, borderline diabetes"`,`"Yes (during pr` The last two options appear less than 0.03% of time in the data; after some exploration on the impact these have on the dependent variable they are deemed not interesting to the work, reason for which they have been incorporated in the first two options. Instances of `"Yes (during pregnancy)"` have been assigned to `"No"` because diabetes during pregnancy is a transitory condition[2] unrelated to lifestyle choices or the incidence of underlying pathologies. Instances of `"No, borderline diabetes"` have been assigned to `"Yes"` because of the much higher risk of diabetes that characterizes individuals in the category compared to healthy individuals.

The variable `AgeCategory` which values describe age groups, starting from group "18-24" to "80 or older" in steps of 4 years, has been encoded as a progressive integer.

The variable `GenHalth` which values go from "Poor" to "Excellent" has been also encoded as a progressive integer.

[2]For most women with gestational diabetes, the diabetes goes away soon after delivery. https://www.cdc.gov/pregnancy/diabetes-gestational.html

3 Data Exploration

The data presents a big class imbalance on the dependent variable `HeartDisease` as we can see in Figure 1 91.4% of the respondents are Healthy individuals, while only 8.6% respondents have developed heart related diseases.

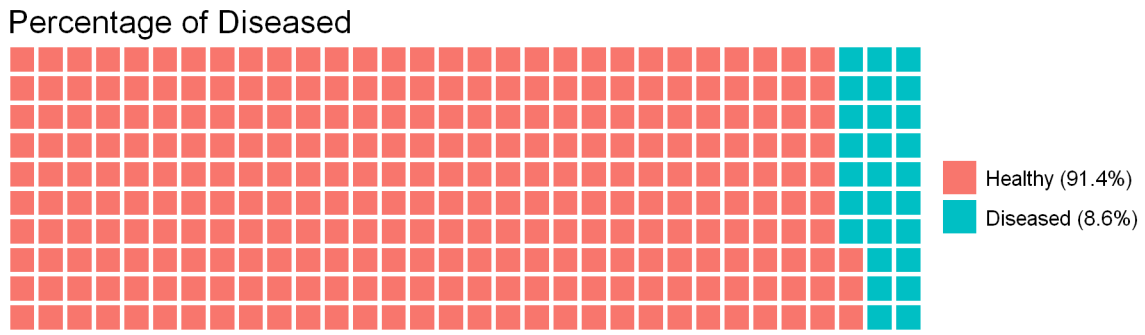

Figure 1: Dependent class imbalance visualized

The analysis starts from the variables `AgeCategory` and `GenHealth`. From what can be seen in Figure 2 there is a clear positive relation between the percentage of diseased respondents and their age. Around age group 9 (60-64) the percentage is over the 8.6% found in the survey population. From the literature we know that aging can cause changes in the heart and blood vessels that may increase a person's risk of developing cardiovascular disease[3].

The inverse seems to be true for `GenHealth`, the graphs in the second part of the Figure 2 showcase the inverse relation between the General Health group (ascendant from Poor to Excellent) and the percentage of diseased. Since the General Health category is self reported the analysis seems to point out the importance of both having good health and feeling healthy.

Both `AgeCategory` and `GenHealth` seem to be great predictors of the dependent variable.

[3]Heart health and aging, NIH https://www.nia.nih.gov/health/heart-health-and-aging

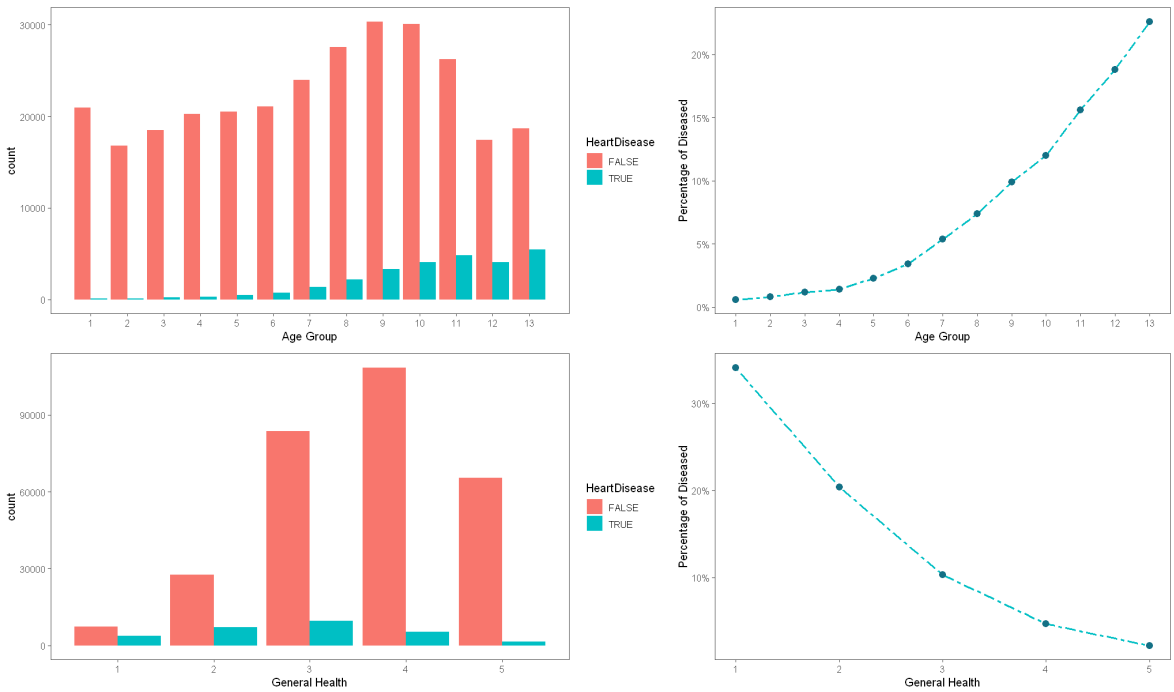

Figure 2: Analysis of the impact of Age and General Health on the percentage of diseased for each group.

The Figure 3 showcases the relations between the variables `MentalHealth`, `PhysicalHealth` and the percentage of diseased for each self-reported value. The X axis is the reported amount of days in the last month in which the respondent has felt that his health was not good.

We can notice a clear positive relation on the `PhysicalHealth` graph, people with more bad health days are more susceptible to heart diseases.

For `MentalHealth` the relation seems again to be positive, but much less linear in his behaviour. Stress and chronic stress have been shown to contribute to poor health behaviours leading to heart diseases and with high blood pressure, which can increase the risk for heart attack and strokes[4].

We notice a clear rise of the percentage at the 10 days marks; the reason is that people on average to have difficulty in keeping a clear record of their mental health and the issue is known to be severely underestimated[5] and most of the respondents are probably eye-balling a possible incorrect amount of days to answer.

It may be useful to discretize the variable in order to account for the variance in number of reports on different days which makes the relation less clear.

[4]Stress and Heart Health, American Heart Association https://www.heart.org/en/healthy-living/healthy-lifestyle/stress-management/stress-and-heart-health

[5]Global burden of mental illness underestimated, Harvard School of Public Health https://www.hsph.harvard.edu/news/hsph-in-the-news/global-burden-of-mental-illness-underestimated/

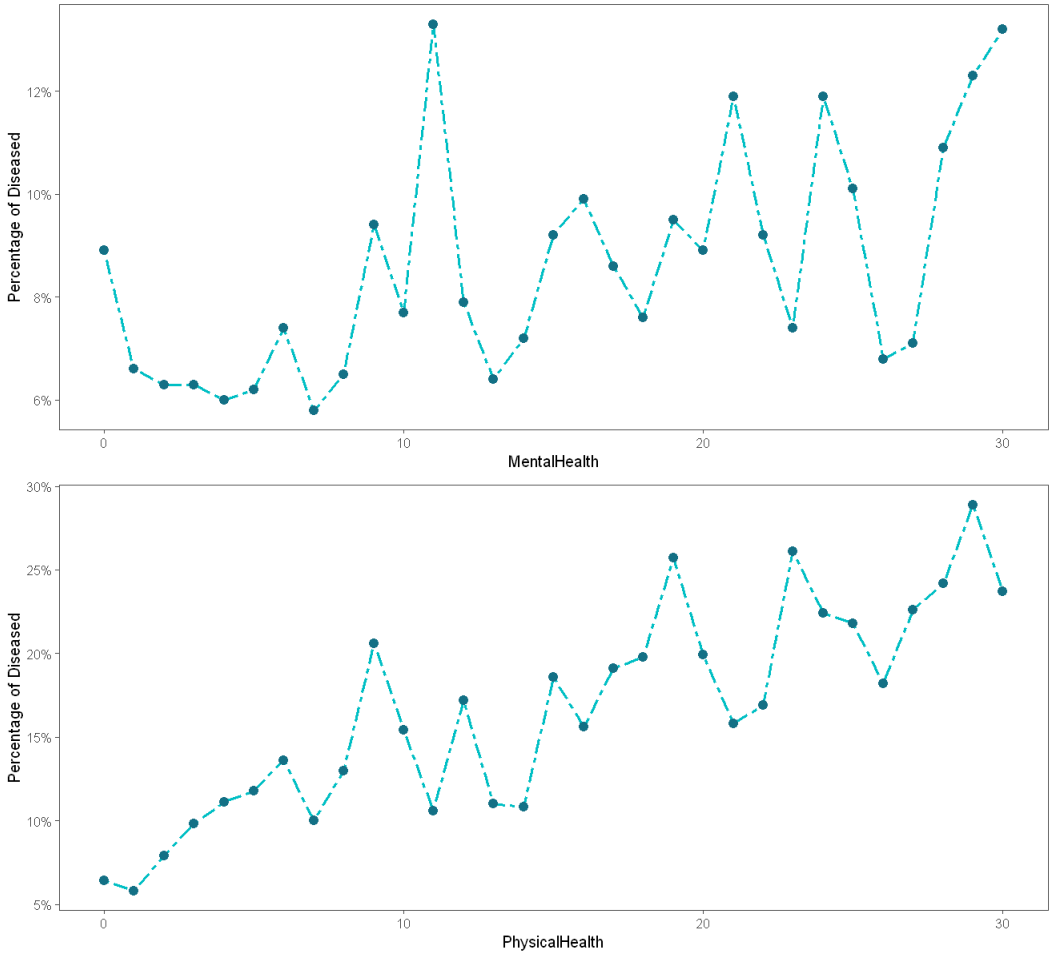

Figure 3: Analysis of the percentage of diseased for the variables MentalHealth and PhysicalHealth

We continue further by analysing the distributions of the variables `BMI` and `SleepTime`. Figure 4 showcases them.

The shown medians of the groups for `BMI` seems to suggest that people with an higher BMI in thee survey report higher incidence of heart diseases; hypothesis that is largely backed in the literature, it has been shown that lifetime risks for Cardiovascular diseases are higher in middle-aged overweight and obese adults [6].

Unfortunately the extreme outliers of the variable make the graph difficult to interpret, further analysis could be conducted by discretizing the variable in intervals.

The `SleepTime` is much more difficult to interpret because of the nature of its distribution; the sleep amount has been shown to be an important factor for general health problems, some of which raise the risk of heart diseases[7]. Further analysis work will be conducted in the next section.

[6] Association of Body Mass Index With Lifetime Risk of Cardiovascular Disease and Compression of Morbidity, https://pubmed.ncbi.nlm.nih.gov/29490333/

[7] How Does Sleep Affect Your Heart Health?, CDC https://www.cdc.gov/bloodpressure/sleep.htm

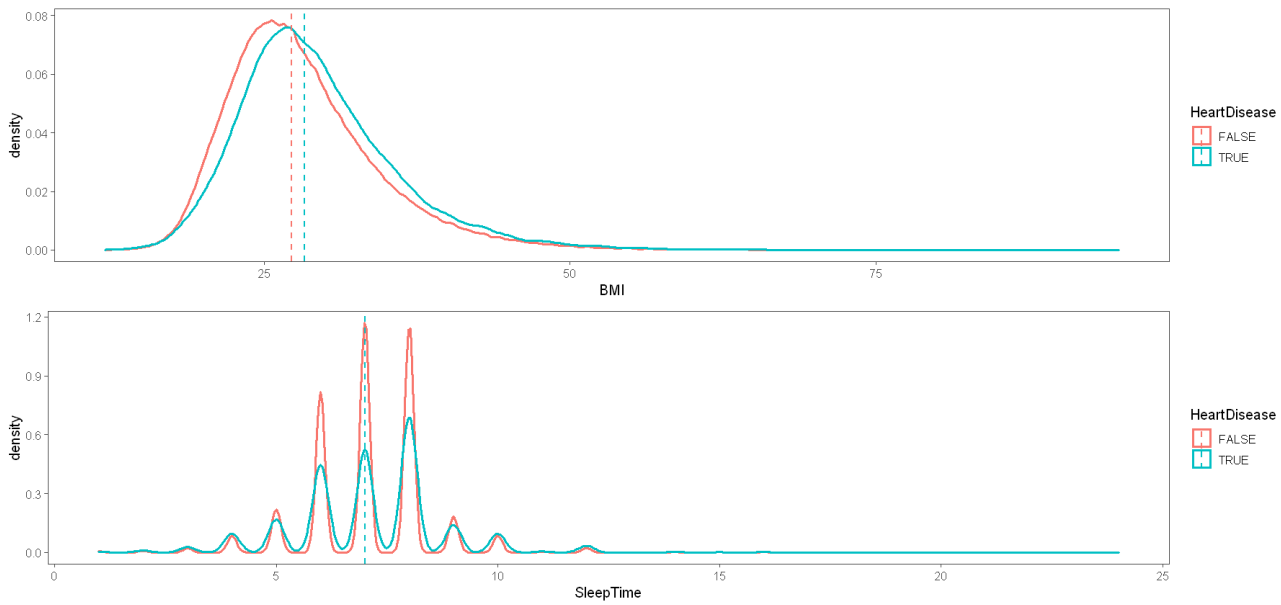

Figure 4: Analysis of the distribution density between diseased and non diseased of values for the BMI and Sleep Time variables.

Figure 11 illustrates the analysis on the percentage of diseased on the remaining variables. Pathology variables such as `Stroke`, `Diabetes`, `KidneyDisease` and `SkinCancer` show diversion from the 8.6% baseline percentage denoted by the white dotted line. The lifestyle related variables `Smoking` and `PhysicalActivity` seem to have less relative impact.

(smokers twice as likely)

The variable `DifficultyWalking` seems to have a big weight on the dependent variable, let's explore it. The baseline percentage of respondents with walking difficulties in the survey is 13%, shown in Figure 6 as the red dotted line. We can see the positive correlation between the variable and `AgeCategory`, around group 8 (55-59) the percentage soars. Another variable that is extremely tied with `DifficultyWalking` is `Diabetic`, 34% of the people having diabetes have difficulty walking, shown in Figure 5.

DiffWalking	Diabetic	%
TRUE	FALSE	65.9%
TRUE	TRUE	34.0%

Figure 5: Metrics for the basic weighted Classification tree

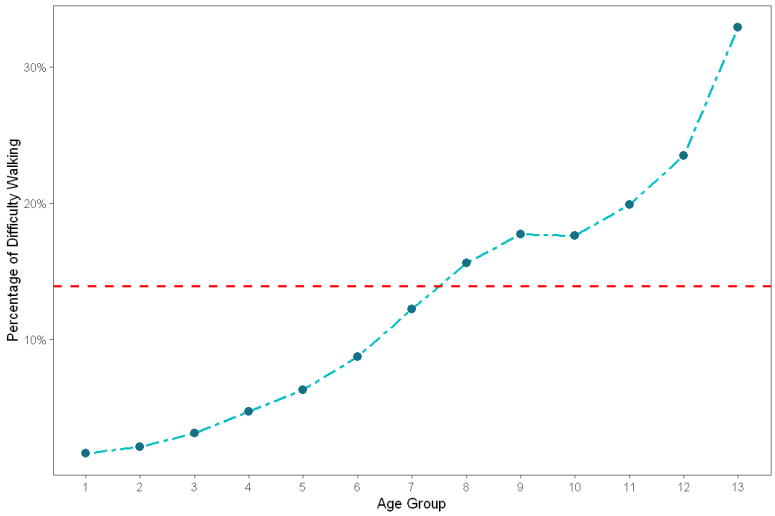

Figure 6: Analysis of percentage of respondents with difficulty walking

A very interesting outline that seems to be shown in the graphs in Figure 11 is what seems to be a positive effect of heavy Alcohol consumption on the percentage of people with heart diseases in the survey. Should we conclude that drinking alcohol prevents heart diseases? Further exploration in Figures 7 and 8 reveal an important characteristic of the heavy drinkers that explains their lower rate of heart diseases. As we can see in Figure 7 heavy drinking seems to be much more common in younger age groups, and becomes very rare going up in age; from what we know already, shown again in Figure 8, heart disease becomes a much more present pathology in older age groups. As we also know `Stroke` and pathology related variables are great predictors of Heart disease, Figure 9 shows that they appear more in older age groups; alcohol drinkers will be less affected by them.

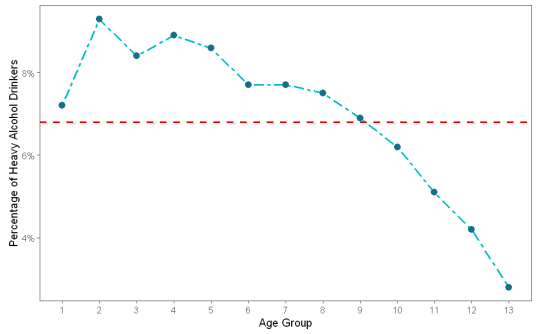

Figure 7: Percentage of drinkers by age group, red dotted line shows the survey population percentage

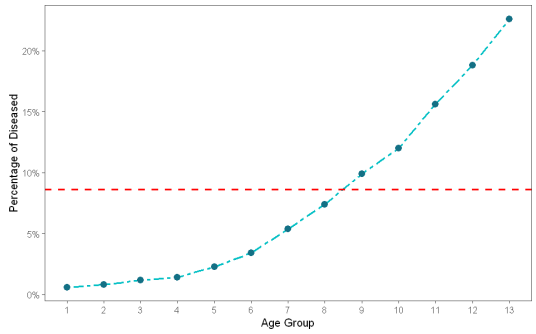

Figure 8: Percentage of diseased by age group, red dotted line shows the survey population percentage

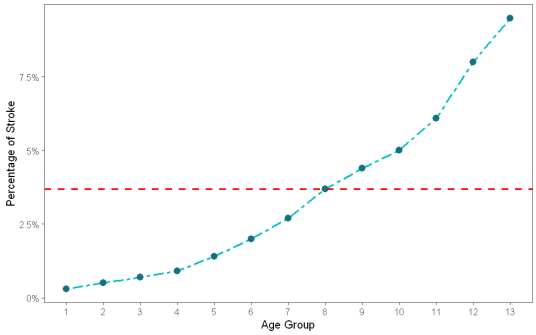

Figure 9: Analysis of percentage of percentage of people who suffered a stroke on their age group.

Figure 11 highlights the importance of good life habits, `Smoking` seems to be correlated with more incidence of heart disease; `PhysicalActivity` appears to be an essential healthy habit, people who don't practice it report higher incidence. Figure 10 displays the importance of being physically active in all age groups: on average active respondents report 3.9% less incidence, but it's true significance shows on older age groups, staying active from Age group 6 (45-49) on-wards nets an average 5.9% of less incidence.

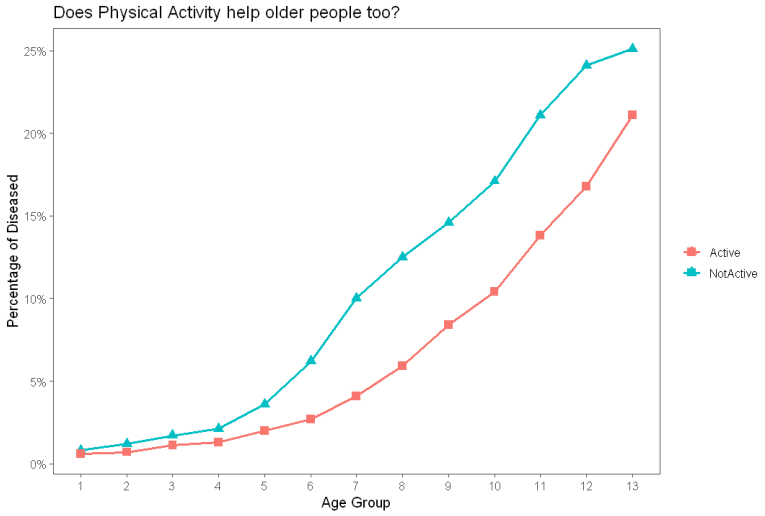

Figure 10: Analysis of percentage of diseased by age group between physically active and not physically active respondents.

Figure 11: Analysis of percentage of diseased for the different variables

Figure 12 visualizes the impact of `Sex` and `Race` on the percentage of diseased.

It seems that Asian people are less affected by heart diseases; according to the Office of Minority Health (OMH) Asian American adults have lower rates of being overweight or obese, lower rates of hypertension, and they are less likely to be current cigarette smokers[8].

Indian American and Alaskan natives have a recorded higher than population baseline incidence of heart disease; this findings are also a known and studied phenomenon in the literature [9].

An interesting finding is that females have a lower than baseline recorded incidence. Let's explore the distribution of the `Sex` variable to understand if there are any possible confounding variables that could explain the finding. Further exploration on the pathology variables shows that the distributions between the

[8]Heart Disease and Asian Americans, OMH https://minorityhealth.hhs.gov/omh/browse.aspx?lvl=4lvlid=49
[9]Heart Disease and American Indians/Alaska Natives, OMH https://minorityhealth.hhs.gov/omh/browse.aspx?lvl=4lvlid=34

sexes is not significantly different. As we can see in Figure 13 the distribution of age categories shows that we have more female respondents on the higher end of the Age spectrum, which from what the analysis has highlighted until now, should lead to higher incidence of heart diseases; but it's not the case here.

From what we have found healthiness of the individual seems to be the most important factor in predicting heart diseases, could it be that females are just healthier than males? Figure 14 disproves the hypothesis; there are more recorded females respondents than males at all `GenHealth` groups and the distribution matches the same exact trend as the male one. We could conclude that female respondents seem to be less affected by heart diseases independently of age and healthiness.

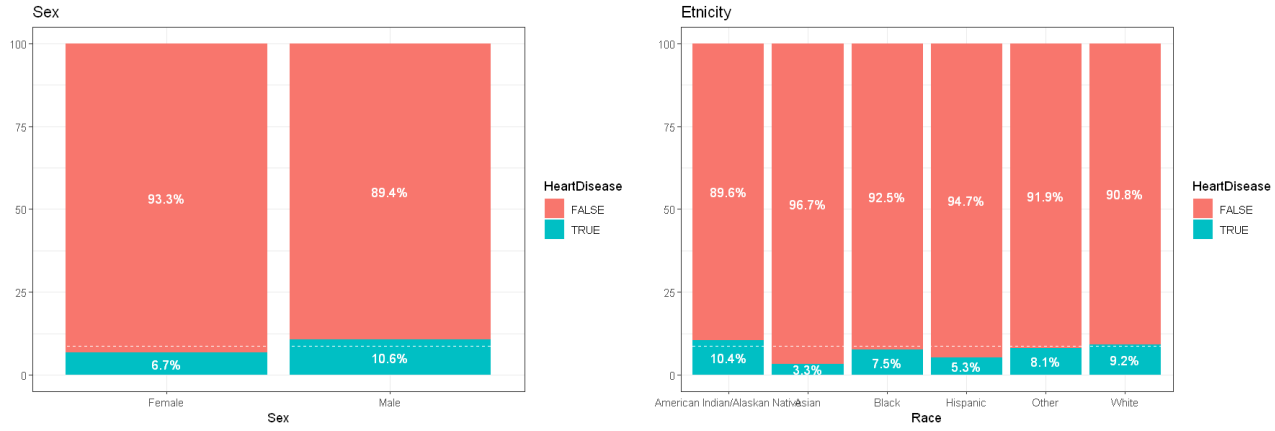

Figure 12: Analysis of percentage of diseased people based on Sex and Race

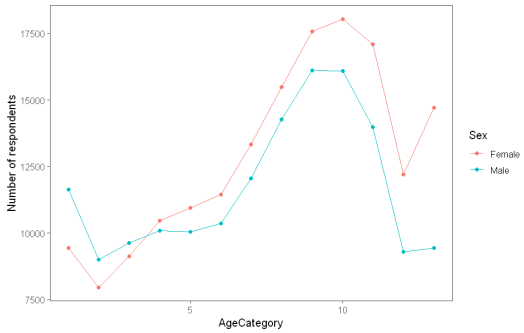

Figure 13: Distribution of respondents by Age Category grouped by Sex

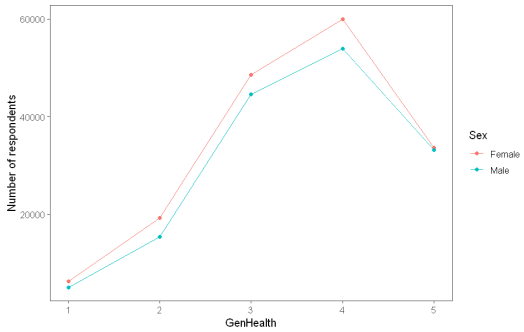

Figure 14: Distribution of respondents by General Health grouped by Sex

4 Further feature engineering

Up to this point the analysis has pointed out how some pathologies are strictly linked to frequent occurrence of heart diseases and how much the general healthiness of the individual plays a part. We move forward with the analysis by computing the variable `nComorbities` which is the count of how many pathologies between `Stroke`, `Asthma`, `KidneyDisease`, `SkinCancer` and `Diabetic` the respondent has reported. Figure 15 points out that multiple coexisting pathologies are often associated to heart disorders, 20% of respondents with 2 pathologies have reported suffering of heart diseases compared to the 8.6% population baseline; it gets even more critical as the number increases.

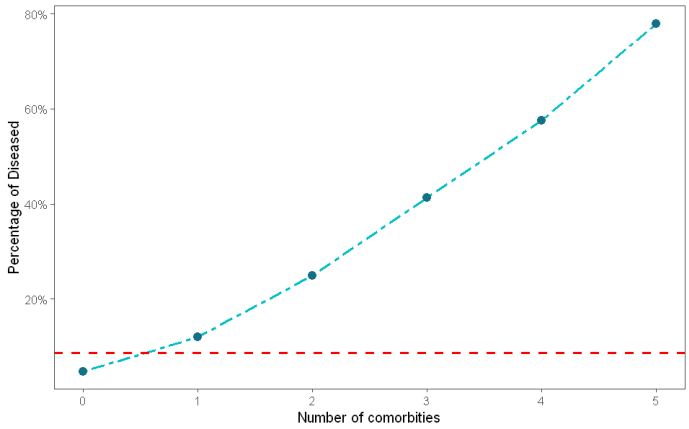

Figure 15: Analysis of percentage of diseased people on how many coexisting patholgies they reported. The red dotted line is the population baseline percentage.

Figure 16 showcases the percentages on the `BMI` variable split in categories following thee CDC BMI guidelines for classifying obesity[10] (1 - Underweight, 2 - Healthy, 3 - Overweight, 4 - Obese). The

The `SleepTime` variable is also been split in hour ranges (1 - ¡=4 hours, 2 - 5 to 7 hours, 3- 8 to 13 hours, 4 - over 13 hours). `PhysicalHealth` and `MentalHealth` have been split in intervals of 10 days.

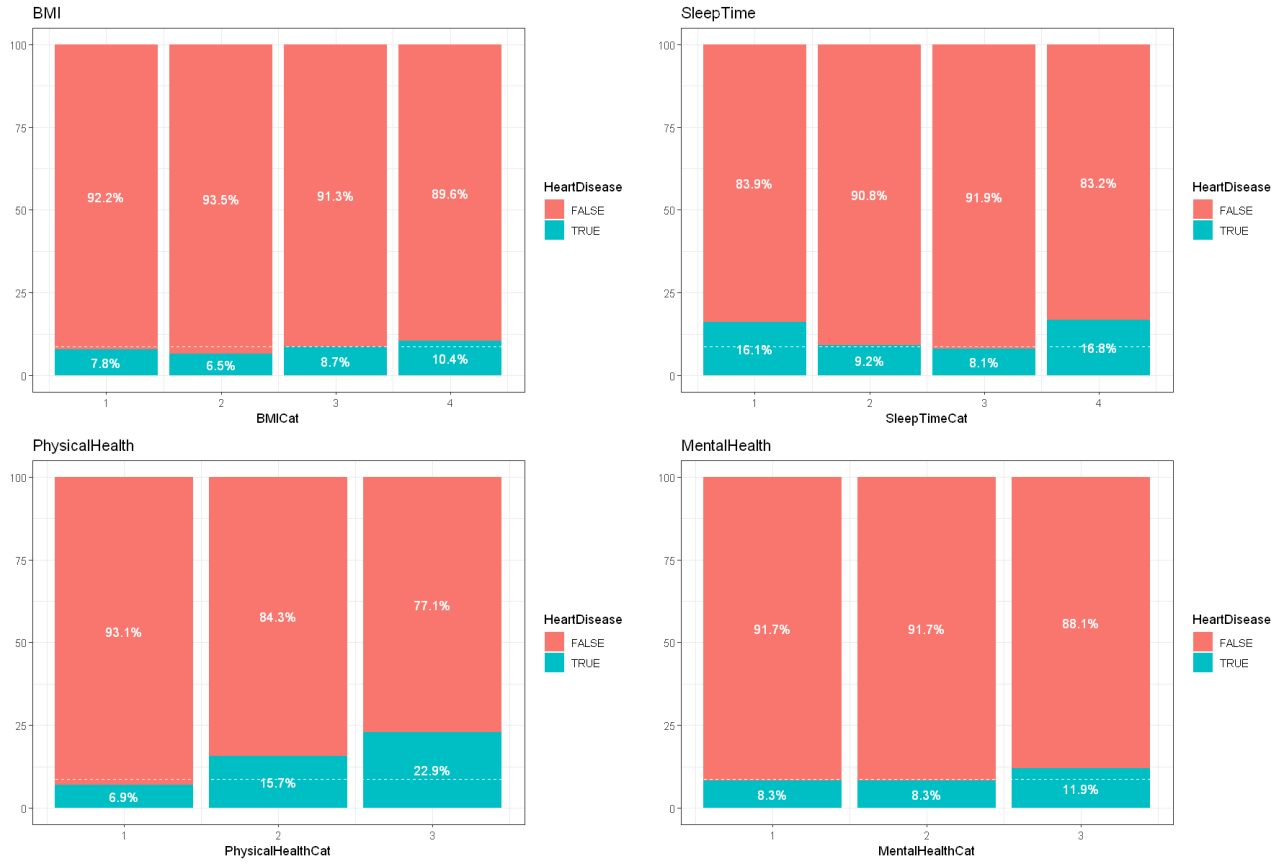

Figure 16: Analysis of percentage of diseased people based on discretized variables

Just as the `nComorbities` variable has been shown to offer significative insight we are going to compute a

[10]Defining Adult Overweight Obesity, CDC https://www.cdc.gov/obesity/basics/adult-defining.html

count of the number of bad lifestyle habits that a respondent has reported. `nBadHabits` will be the count of how many of bad lifestyle habits, defined as follows, the respondent has reported

- Is obese - `BMICat` is 4
- Has reported either less than 5 hours of sleep or more than 13 on average - `SleepTimeCat` is 1 or 4
- Has reported more than 10 days of stress in the last month - `MentalHealthCat` is 2 or 3
- Is a smoker
- Is not physically active

Figure 17 displays the percentage of diseased for each coexisting number of bad habits. As we can see, while from what we've seen singularly these lifestyle choices don't have a massive impact, when they pile up the diseased percentages rise very fast. We can deduct that a general healthy approach to lifestyle habits has a considerable weight.

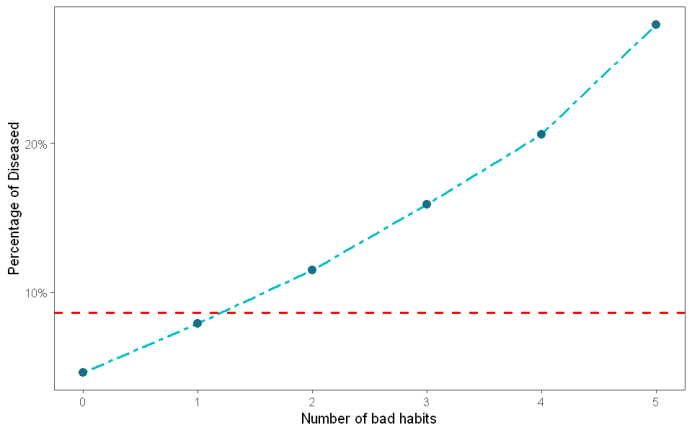

Figure 17: Analysis of percentage of diseased people based on the number of bad lifestyle habits they reported. The red dotted line represents the population baseline percentage

We conclude the exploration with the correlogram in Figure 18. The graph underlines some interesting observations. The variable `GenHealth` is negatively correlated with `nBadHabits` we can deduct that bad lifestyle choices take a toll on the physique; same logic is applicable for `BMI`; the variable is also negatively correlated with the Age. `AgeCategory` is positively correlated with the amount of bad physical health days reported, while positively correlated with the amount of bad mental health days reported; while the physique tends to become less reliable with age, older people may be subjected to less stress. `BMI` is very positively correlated with `nBadHabits`, we don't know the direction of the causality, bad lifestyle choices may lead to increased BMI and high BMI leads may lead to an increase in bad lifestyle habits. `MentalHealth` seems to be heavily tied to `PhysicalHealth`; in this instance we also don't know the causality direction but it would make sense that sickness would lead to stress and that stress can lead to physical pathologies.

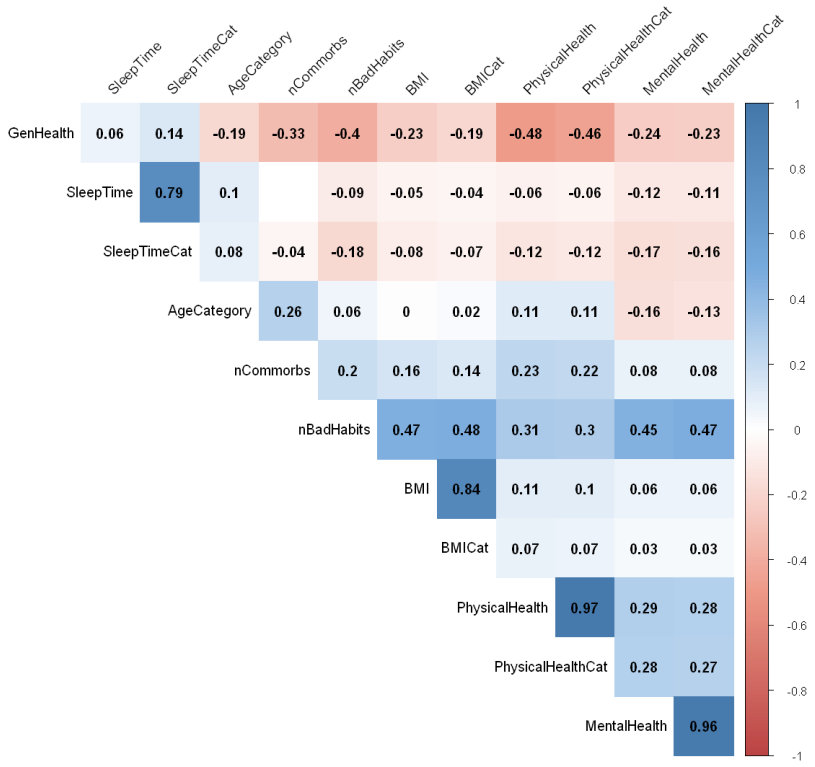

Figure 18: Analysis of the correlations between the variables.

5 Logistic Regression

We will apply the Logistic Regression algorithm to try to predict the dependent boolean variable `HeartDisease` by using the other variables in the survey as predictors.

As we can see from the listing below `Stroke` has the most relevance, followed by the pathology variables `Diabetic` and `KidneyDisease`. `SexMale` is another very important predictor, the previous analysis already highlighted it. All of the coefficient signs seem to be in the expected direction.

```
Call:
glm(formula = HeartDisease ~ . - BMI - PhysicalHealth - MentalHealth -
    SleepTime - nBadHabits - nCommorbs - AlcoholDrinking, family = binomial,
    data = train_aug)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1660  -0.4081  -0.2452  -0.1353   3.5809

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.046030   0.097689 -41.417  < 2e-16 ***
SmokingTRUE           0.348391   0.015986  21.793  < 2e-16 ***
StrokeTRUE            1.037363   0.025342  40.934  < 2e-16 ***
DiffWalkingTRUE       0.229963   0.020114  11.433  < 2e-16 ***
SexMale               0.711198   0.016276  43.696  < 2e-16 ***
AgeCategory           0.271033   0.003362  80.608  < 2e-16 ***
RaceAsian            -0.542023   0.094530  -5.734 9.82e-09 ***
RaceBlack            -0.335364   0.065116  -5.150 2.60e-07 ***
RaceHispanic         -0.240739   0.066178  -3.638 0.000275 ***
RaceOther            -0.051488   0.071828  -0.717 0.473481
RaceWhite            -0.072526   0.058240  -1.245 0.213022
DiabeticTRUE          0.467411   0.017870  26.156  < 2e-16 ***
PhysicalActivityTRUE  0.014400   0.017927   0.803 0.421833
GenHealth            -0.504777   0.009412 -53.634  < 2e-16 ***
AsthmaTRUE            0.275786   0.021519  12.816  < 2e-16 ***
```

```
KidneyDiseaseTRUE      0.576130   0.027406  21.022  < 2e-16 ***
SkinCancerTRUE         0.092585   0.021930   4.222 2.42e-05 ***
BMICat                 0.085734   0.010056   8.526  < 2e-16 ***
SleepTimeCat          -0.099303   0.015678  -6.334 2.39e-10 ***
PhysicalHealthCat      0.017122   0.012731   1.345 0.178636
MentalHealthCat        0.056586   0.013489   4.195 2.73e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 149328  on 255835  degrees of freedom
Residual deviance: 116029  on 255815  degrees of freedom
AIC: 116071

Number of Fisher Scoring iterations: 6
```

The output probabilities from the model have to be thresholded with a certain parameter. Since the dataset is highly unbalanced we have to choose that parameter based on an appropriate metric. Figure 19 shows the Specificity metric on different thresholds computed on the test set; when we set the threshold too high we lose the ability of classifying the heart diseases.

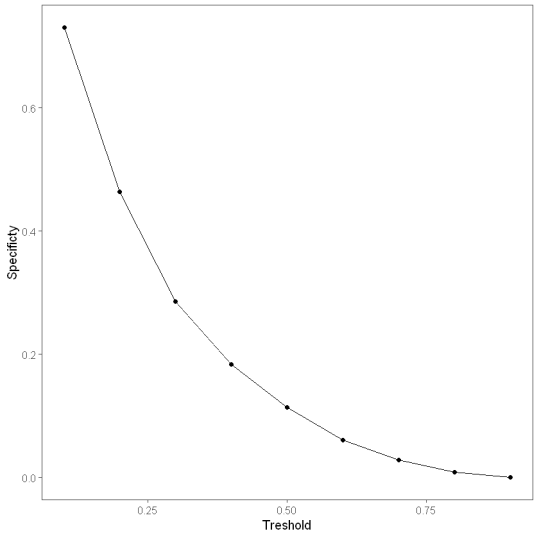

Figure 19: Precision metric on different thresholds

We run 10 fold cross validation on the whole dataset to find the best possible threshold on AUC.

Cv run	Threshold	AUC
1	0.06	0.7594933
2	0.08	0.7631360
3	0.09	0.7642782
4	0.08	0.7674162
5	0.09	0.7622858
6	0.08	0.7693481
7	0.07	0.7679158
8	0.08	0.7606776
9	0.09	0.7673325
10	0.09	0.7637488

We select the threshold by taking the mean of the best found thresholds in every cv run (0.082), with a mean AUC score of 0.765.

Figure 20 displays the metrics for the performance of the model on the test set (80/20 split).

	FALSE	TRUE
FALSE	42896	1182
TRUE	15547	4334

Precision	0.78
Recall	0.21
F1	0.203
AUC	0.759

Figure 20: Metrics for the Logistic Regression with the selected threshold on the test set

6 Trees

We continue our analysis by applying Classification Trees using the package `rpart`.

Given the unbalanced nature of the data creating a simple tree is going led to the tree trying to achieve the maximum accuracy by classifying all instances as Negatives.

	FALSE	TRUE
FALSE	290732	24085
TRUE	1690	3288

Precision	0.12
Recall	0.66
F1	0.203

Figure 21: Metrics for the basic non-weighted Classification tree

As we can see from the Confusion Matrix, from a run of the algorithm on the whole dataset, and the low Precision score shown in Figure 21 the algorithm is not able to recognize Positive instances in the database.

For this reason we are going to use class weights, set as 92 for the Positive class and 8 for the Negative class; the idea is using the distribution of the classes in the database to decide the weights (92% Negatives, 8% Positives).

We are going to prune the tree using the 1-SE technique, the errors for each of the cps are decided on a 10 fold cross validation on the entire data set.

```
           CP nsplit rel error  xerror       xstd
1  0.35231959      0   1.00000 1.00000 0.00047075
2  0.04545486      1   0.64768 0.64768 0.00043647
3  0.03625924      2   0.60223 0.60223 0.00042752
4  0.01381907      3   0.56597 0.56772 0.00041992
5  0.01267176      4   0.55215 0.55231 0.00041629
6  0.00890152      5   0.53948 0.53955 0.00041317
...
38 0.00016586     78   0.47753 0.48796 0.00039946
39 0.00015132     79   0.47737 0.48791 0.00039945
40 0.00014705     81   0.47707 0.48736 0.00039929
41 0.00014534     82   0.47692 0.48719 0.00039924
42 0.00014363     83   0.47677 0.48738 0.00039930
43 0.00014192     84   0.47663 0.48732 0.00039928
44 0.00013736     85   0.47649 0.48725 0.00039926
45 0.00013679     88   0.47608 0.48738 0.00039930
46 0.00013251     91   0.47567 0.48733 0.00039928
47 0.00012909     93   0.47540 0.48739 0.00039930
...
```

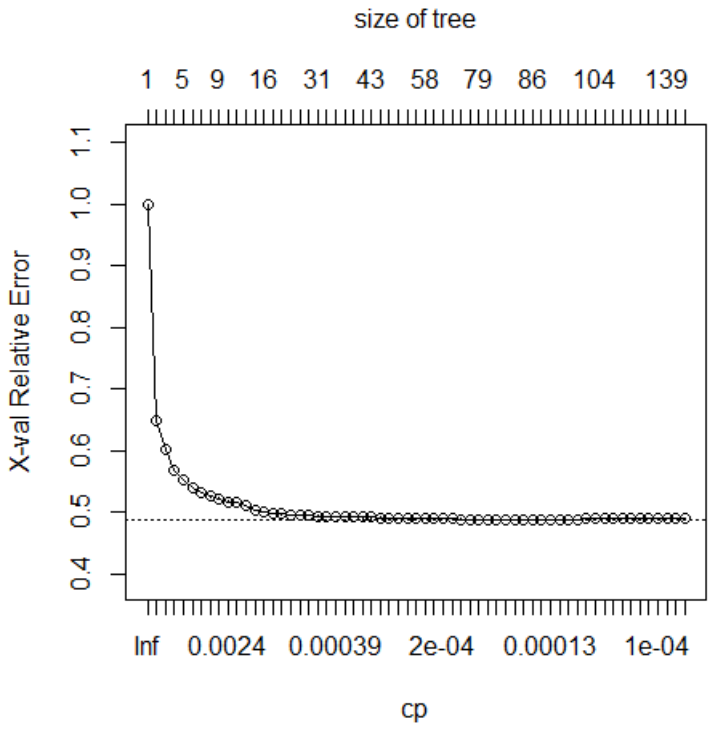

Figure 22: 10 fold CV error as the splits grow

From what we can see in Figure 22 the error gets fairly stable around 16 splits and forward. The lowest `xerror` appears with 82 splits, by using the 1-SE rule we select cp 0.00014705 which causes 81 splits for the pruning.

	FALSE	TRUE
FALSE	206933	4697
TRUE	85489	22676

Precision	0.82
Recall	0.21
F1	0.335

Figure 23: Metrics for the class weighted Classification tree

The performance as seen in the Figure 23 is not stellar but we are now able to discern the Positives, trading off the accuracy we had on the Negatives class.

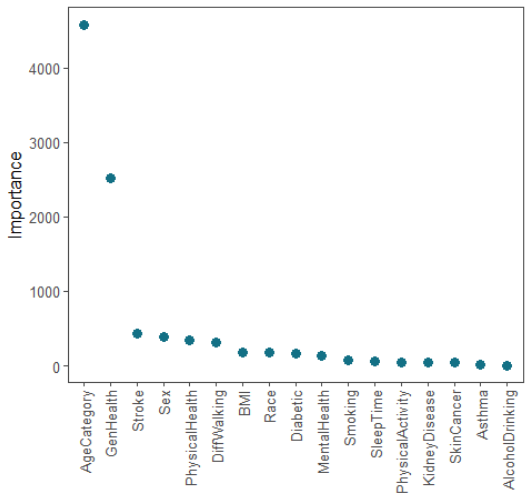

Figure 24: Variable importance overall

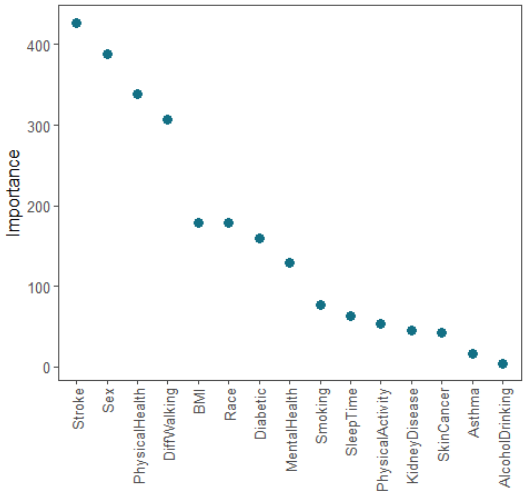

Figure 25: Variable importance omitting first two points

We can see in the Figure 24 that the `AgeCategory` is the most important variable, followed by `GenHealth`. Figure 25 is the same plot omitting the first two points, we can see that `Stroke`, `Sex`, `PhysicalHealth` and `DiffWalking` are the following most important variables.

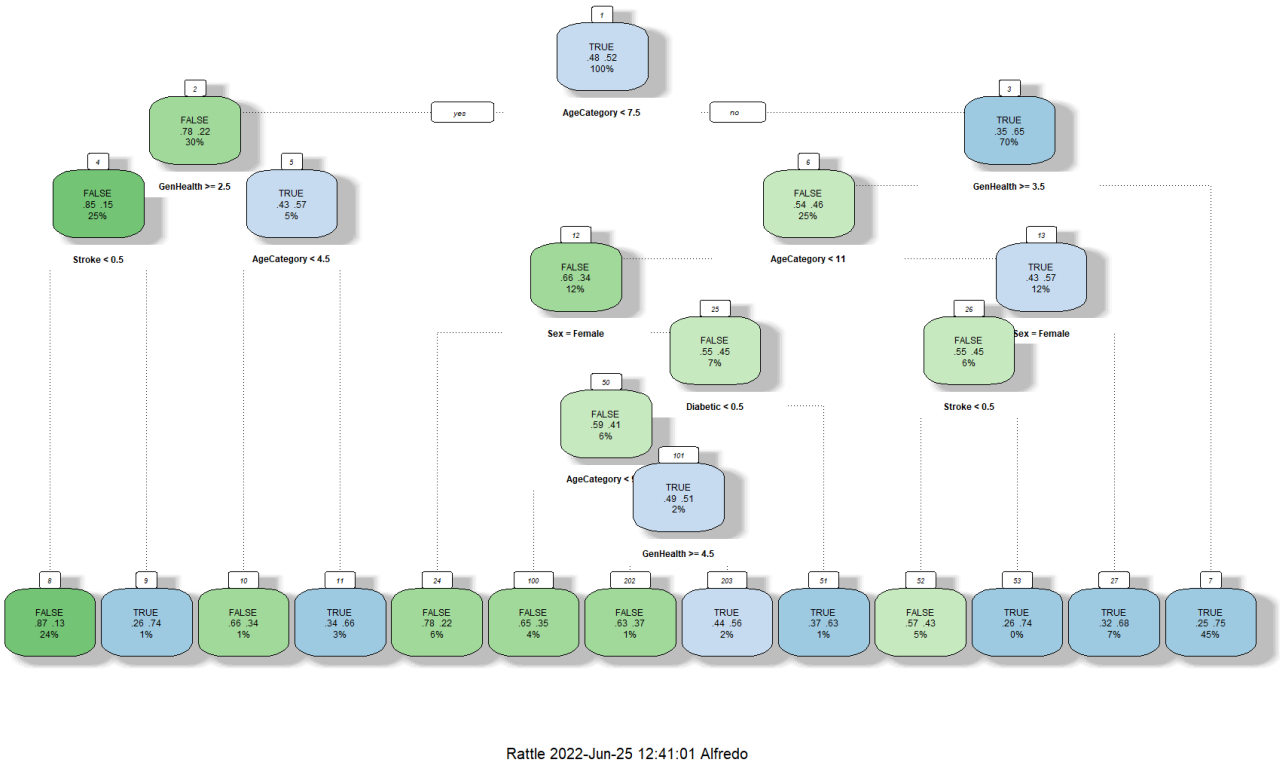

Figure 26: Graphical visualization of the class weighted Classification Treee

Figure 26 is a visualization of the tree (pruned at less splits for visualization purposes). The Classification

Tree is scoring 0.76 AUC on a 80:20 train test split, we can see its metrics in Figure 27.

	FALSE	TRUE
FALSE	41715	1061
TRUE	16728	4455

Precision	0.80
Recall	0.21
F1	0.33
AUC	0.76

Figure 27: Performance metrics from the 20% test split

7 RandomForest

The random forest approach didn't yield much better results; in Figure 28 it's possible to see the AUC performance as the number of trees rise. The maximum achieved AUC is 0.762.

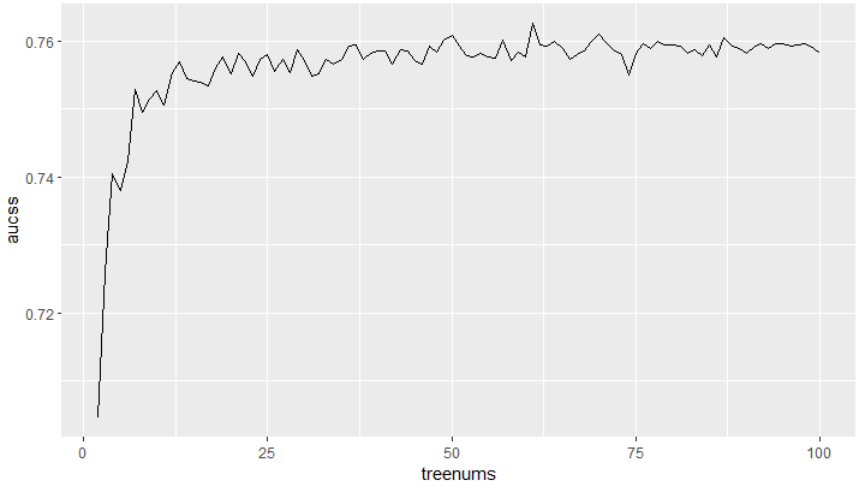

Figure 28: AUC as the number of trees in the ensamble go up