

# Behavioral Risk Factor Surveillance System

## Introduction

This report presents the findings and conclusions of the final project for the Data Mining course. The project focused on analyzing a dataset provided by the CDC, containing information from the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS conducts annual telephone surveys to gather data on the health status of U.S. residents, the dataset used in this project comprises information on 445,132 residents.

The initial presentation of this project covered the data cleansing process in detail. As such, this report will only mention the work done into the data cleansing procedures and will instead highlight the initial questions posed and the subsequent analyses conducted to address them.

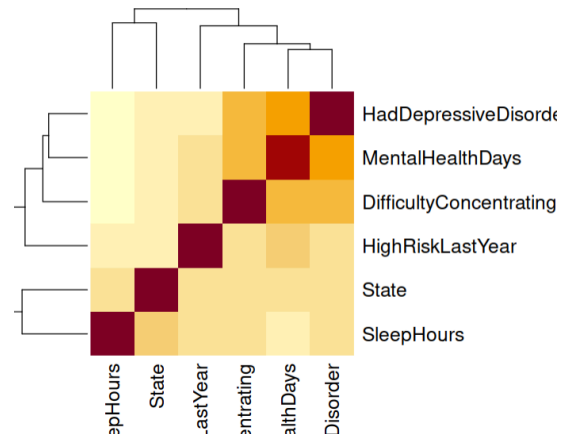
To process the data, we have binarized the categorical variables, removed the rows with five or more NA values and replaced NA values and outliers with the median of each variable. This cleansing can be more specified in the RMD files.

## I. Based on the variables related to mental illness, could we predict the state of residence of the individuals?

To try to answer this question we are going to take a few variables of the complete dataset that are related to mental health and build some classification models with the techniques studied along the course. These variables are:

- **State:** It will be our ground truth. [chr]
- **MentalHealthDays:** [num: 0, 1, ..., 30]
- **SleepHours:** [num: 0, 1, ..., 24]
- **HadDepressiveDisorder:** [chr: No, Yes]
- **DifficultyConcentrating:** [chr: No, Yes]
- **HighRiskLastYear:** This one has been chosen since we consider that the consequence of having suffered from this problem, can cause some kind of mental harm or trauma. [chr: No, Yes]

This is the correlation of variables for the new dataset:



To try to achieve some good results we are going to use as models: decision tree implementations, K-Nearest Neighbours and Naive Bayes algorithm.

To implement these models we have used the libraries: rpart (CART), e1071 (Naive Bayes), RWeka (J48-ID3) and class (K-NN).

To get a good accuracy we are going to split the new dataframe into training and a testing sets with a train/test ratio of 70% since we consider the dataset is large enough to get a good training with this percentage of the data.

First of all, we are going to see how the data performs only with the general data cleansing applied to the dataset.

The accuracy of the models are:

Model	Accuracy
Decision Tree (CART)	5.94 %
K-NN	Null: see note
Naive Bayes	5.46%

**Note:** For the K-NN model we can't get any results since the model can't finish due to having too many ties. We believe that this is caused because most of the variables are binary producing a lot of ties in distance. In the future we are going to try to avoid it.

We have seen that with the mental health dataframe as it comes from the whole dataset, we can't get any results. We believe that this is mainly because we have too many classes to predict and because of the nature of the data. The decisions that we have made to try to avoid this are:

## 1. Reducing labels

First of all, to reduce the number of labels to predict, we have done some research and based on the [USA census of 2020](#) we have divided the States into density of population, to be exact by number of people by km2. The classifications are:

- More than 100 people by km2 (23.34%)
- Between 100 and 43 people by km2 (24.88%)
- Between 43 and 20 people by km2 (25.81%)
- Less than 20 people by km2 (25.95%)

We have chosen this classification since we hope that 4 labels are few enough to get a good classifier. Another reason is that with this classification we get 4 balanced classes to predict, avoiding making a model better at predicting some class.

## 2. Normalize Data

The other technique to improve the nature of the data is to normalize it so all the values are between 0 and 1. Normalizing the data in a classification model is beneficial because it ensures that features with different scales contribute equally to the model, preventing dominance by variables with larger magnitudes.

After applying these changes and adding the ID3 decision tree algorithm this are the new accuracy:

Model	Accuracy
Decision Tree (CART)	26.51 %
Decision Tree (ID3)	25.81 %
K-NN	<b>Null</b> : Still too many ties
Naive Bayes	26.19 %

## Conclusions

Although we have improved the accuracy with the techniques applied, we consider that it still is pretty low so we can determine that the variables related to mental health are not enough to predict the state of living of a subject even if we group the different states.

Finally, we want to see if the class of density can be predicted with the whole dataset since we don't get good results with the mental health variables.

Model	Accuracy
Decision Tree (CART)	31.31 %
Naive Bayes	27.83 %

With the whole dataset, we can see how we improve

a bit the accuracy which it makes us think that are better variables for trying to predict the state of living than the ones related with mental health, that even though we chose this question because it was interesting for us trying to discover which could be the main variables that can classify a state, we have learned that not all the variables are good to predict certain things.

## II. Can we make a predictive model for the 'general health' label with all the available data?

In this second question, our objective is to construct a classification model to predict the general health status of individuals in our dataset. We intend to use all available columns in the dataset and explore multiple models to achieve the highest possible accuracy. It is important to state that the classification into five distinct classes is subjective, as these labels are based on responses obtained from a survey rather than professional medical assessments. It is acknowledged that the provided values may lack precision, given the potential for inaccuracies resulting from subjective participant responses.

To develop accurate classification models, we need to first train them using a portion dataset and then try to classify new data to obtain the real accuracy. Training involves enabling the models to learn patterns within the data. This division ensures a reliable evaluation of the models' ability to generalize beyond the training data and make accurate predictions for the five subjective health classes.

Our first challenge is to address the imbalance in class distribution within our dataset. Achieving a balanced distribution is crucial for optimal model performance, as an uneven class representation can introduce bias and compromise the effectiveness of our models.



To rectify this issue, we'll explore downsampling as an initial strategy. Downsampling involves reducing the number of instances in the majority class to align with the minority classes. This approach is relatively straightforward and serves as a first step.

The first model that we are going to use is **Naïve**

**Bayes**, a probabilistic classifier inspired by the Bayes theorem under the simple assumption that attributes are conditionally independent. This assumption reduces the computational cost greatly by only counting the class distribution. Even though the assumption is not valid in most cases since the attributes are dependent, such as our dataset, Naive Bayes is able to perform impressively well as we saw in the course and in the different assessments that we have done.

Constructing a model in R is a straightforward process, and within a few minutes, we can determine its accuracy on the test set, yielding an accuracy of approximately **42%**.

The following model in our analysis employs a decision tree with a recursive approach. Decision trees, as a supervised learning model, make decisions based on input feature values by recursively splitting the dataset into subsets. At each node, a decision is made regarding a specific feature, creating branches that lead to subsequent nodes. This recursive process continues until a stopping criterion is met, such as reaching a maximum depth or a minimum number of samples in a node.

To implement this approach in R, we leverage the `rpart` package with the `method = "class"` flag, indicating a classification task. Additionally, we set the control parameter to `1e-5` to regulate the complexity size of the tree, avoiding overfitting. Subsequently, we prune the tree using the best complexity parameter determined through cross-validation.

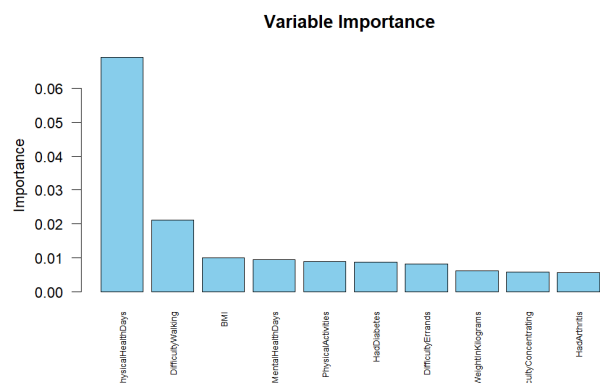
This **decision tree** model, configured with these parameters, yields a precision of **around 50%**, showcasing a substantial improvement compared to the Naïve Bayes model. This enhancement underscores the effectiveness of the decision tree's ability to capture intricate patterns in the data, leading to more accurate classifications.

The final model we will explore the application of a **Random Forest** model. Similar to the decision tree, a Random Forest is a supervised learning ensemble model that uses multiple decision trees to make predictions, and then take the best tree of them. To apply this method in R we will use `ranger`, a fast implementation particularly suited for high dimensional data where we are going to generate one-thousand random trees. With this approach we got a **48% accuracy**, getting a final verdict where the decision tree is the model that performs better for our objective.

## II.b Which are the illnesses that affect U.S. residents more?

Having constructed a predictive model for the General Health label, our subsequent goal is to determine the illnesses that predominantly affect residents of the U.S. In essence, we seek to identify the most prevalent health conditions that carry a significant weight in the algorithm's classification of the general health status label.

To achieve this, we will analyze the importance of features in our predictive model. By examining the importance of each feature, we can identify which illnesses contribute more or less to the prediction of general health status, in order to make a readable graphic we will only get the first ten features.



As we can observe the most influential feature is **PhysicalHealthDays**, nearly triple the importance of the next leading features, **DifficultyWalking** and **BMI**. This observation aligns with intuition, as more frequent absences due to physical health issues and difficulties with mobility often correlate with lower overall health quality.

To sum up, we employed three distinct models to predict the general status of individuals in our dataset, effectively addressing the class imbalance issue. The final model achieved a 50% accuracy rate in predicting subjective responses provided by interviewees, rather than healthcare professionals. While we find this outcome satisfactory, we believe that further refinement through enhanced and comprehensive data cleaning processes could yield even more promising results.

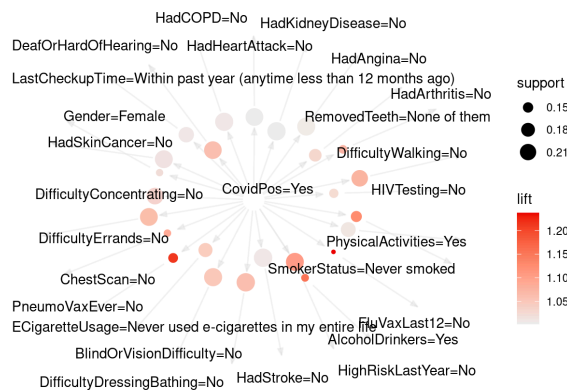
## III. Which are the characteristics of people that had covid? And the consequences?

In order to answer this question, we have considered that the best technique would be the use of association rules, which will allow us a strategic exploration to identify combinations of factors that are frequently associated with the presence of

COVID-19. In order to use this technique, we will discard all numeric variables and focus on the factor's ones.

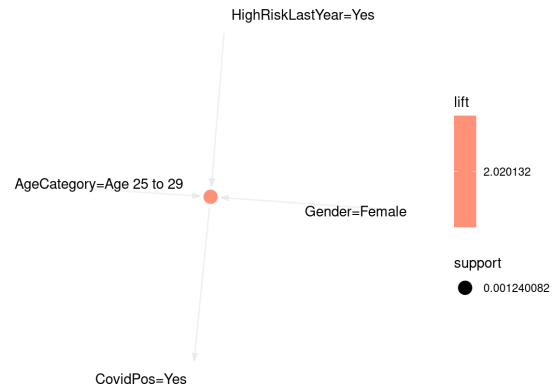
With the data clean and structured, the analysis progresses towards exploring associations through the Apriori algorithm, implemented using the **arules** library. The generated association rules provide an initial approach to the common characteristics of individuals who have had COVID-19, offering preliminary insights into potential risk factors or behavior patterns. The visualization of these rules, facilitated by **arulesViz**, allows for an intuitive interpretation and a quick identification of the most significant associations.

To identify the consequences, rules that contained the variable *CovPos=Yes* as the antecedent of the rule were generated.



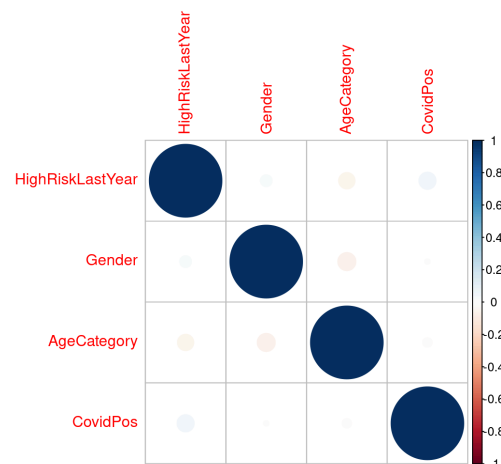
We observed that factors such as never having smoked, engaging in physical activities, and not having had certain diseases (such as angina, arthritis, or heart attacks) present a higher lift value in relation to having COVID. This suggests that people with a healthier lifestyle or without certain pre existing conditions might have had different experiences during the pandemic, possibly reflecting a lesser impact of COVID-19 on these groups.

On the other hand, the previous logic has been reversed to identify the characteristics of people affected by COVID, that is, the variable has been used as the consequent.



We can see that the combination of being female, aged between 25 and 29, and having been classified as high risk in the previous year, is significantly associated with having COVID-19, as evidenced by a lift value greater than 2. This implies that these characteristics together may increase the likelihood of having COVID-19 more than would be expected if they were independent.

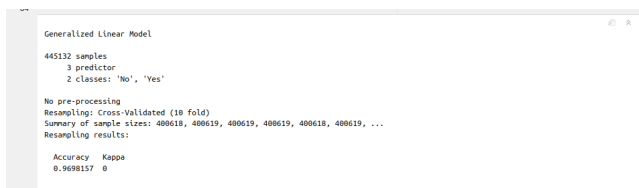
In order to verify the accuracy of the results, a logistic regression model will be used to test if these variables can indeed define individuals with COVID-19. But before that, a study of correlations will be carried out, taking the analysis a step further by quantifying the relationship between variables through a correlation matrix. This step is essential to understand not only the individual characteristics associated with COVID-19 but also how these characteristics interact with each other. By transforming categorical variables into dummy and numeric variables, the analysis ensures that the measured correlations are accurate and meaningful, thus providing a solid foundation for subsequent predictive modeling.



The variables *Gender*, *HighRiskLastYear*, and

*AgeCategory* show a notable correlation with *CovidPos*, indicating that these characteristics could be associated with the incidence of the disease. Specifically, there appears to be a stronger correlation between *HighRiskLastYear* and *CovidPos*, which could be interpreted to mean that individuals who were considered high risk in the previous year have a higher likelihood of having had COVID-19.

The use of a logistic model not only allows us to confirm the results obtained previously but, by modeling the probability of having had COVID-19 as a function of various characteristics, the model not only identifies the most significant variables but also quantifies their impact. The use of cross-validation in the model evaluation ensures that the results are not only statistically valid but also robust and reliable, thus providing valuable insights into the characteristics of COVID-19.



```

Generalized Linear Model
445132 samples
3 predictor
2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 400618, 400619, 400619, 400619, 400619, 400619, ...
Resampling results:

Accuracy  Kappa
0.9698157  0
  
```

In conclusion, although initially, very positive results might be observed, the discrepancy between a high accuracy and a kappa coefficient of 0 could suggest that, despite the high accuracy, the model does not have genuine predictive capability over both classes. In the context of the research question, this result could indicate that, although the model can correctly predict individuals who have had COVID-19, it might not be as effective in predicting those who have not had it, or vice versa. This suggests that there could be a severe class imbalance in the data.

#### IV. Based on our lifestyle, what's the probability of having depression?

The goal of this question is to predict if each of us is going to have depressive disorder and in what percentage. This way we'll know how likely we are. To do so, we are going to use linear regression technique where the dependent variable is *HadDepressiveDisorder* since that is the one we are predicting and the rest excluding the State are the dependent ones. We have deleted the State variable to do this prediction because we do not live in the United States so we did not want that variable in our model.

Before training the model we need to split the data in two sets, the training one and the test one. This way, when predicting and comparing to the real label we can estimate how good the predictions are from this model. We have splitted the data set into 70% for training and 30% for testing. We made this decision because we think the data set is big enough to be able to train with 70% of the observations and not a higher percentage. This way we can benefit from having more testing data.

The predictions of this model are going to be a number between 0 and 1. Since in our dataset having depression is a binary variable we need the results to be binary too so we can compare them. This means that we have overwritten the continuous predictions to binary by saying that every result higher than 0.5 is 1 and lower than 0.5 is 0.

We got that the precision from this model is about 83% which is a pretty high percentage. So, we can affirm that this model is a good fit.

Once we have our model, we need to answer the question and use the model with our own data. We have created a csv file with our data. The data we have inserted is already cleaned as in the given data set. Meaning that we have directly used 0 or 1 to indicate gender or if we had some disease, or used the age categories directly instead of putting our age. This means that this data looks exactly like the one we cleaned so we can use the same model we have trained to predict our personal result.

Finally, we have predicted using the model and we have got the following results:

**Sílvia:** 0.9999985

**Josep:** 1.0000000

**Martí:** 1.0000000

**Pau:** 1.0000000

**Macià:** 0.6622107

We see that Josep, Martí and Pau are the most likely to suffer depressive disorder and Macià is the least likely. However it is still more than 0.5 so binarizing it would give us a 1.

#### V. Is there any relation between the genetic race and the diseases of the individuals ?

We want to see if there is any relationship between different ethnicities and the probability of suffering

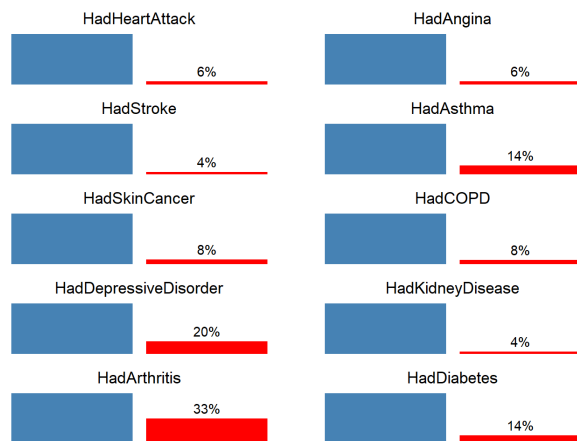


from various diseases collected in the dataset. To do this, we plan to use some clustering models ignoring the race label and then compare the clustering results with it.

As clustering algorithms are relatively expensive, and to avoid going completely blind, we decided to make a first pass using **logistic regression**. We left that variable in categorical mode so that the algorithm automatically creates dummy variables for each category.

Here came our first problem: to **avoid multicollinearity**, R decides to ignore the first level, as it can be explained linearly by the absence of the rest. We chose to keep the NA values in that column but label them to be their first level, thus being ignored by the “glm” algorithm.

We reviewed the **balance** of all diseases and the probability of suffering from them. In general, it is **around 4-8%**, and in the most extreme case (**HadArthritis**), it is **33%**. Since balancing by downsampling will get us a relatively small number of samples and by upsampling it will duplicate a lot of the data, we will continue as is, although we will create a balanced model for that disease and compare both results.



Some diseases may be directly related. For example, heart attack, angina, and stroke are circulatory problems, while asthma and COPD are respiratory/lung related. Therefore, we will include only one disease in the dependent variable, and the rest of the columns in the independent variable.

Once we have the results, focusing on the **p-value** and the **estimate**, we can draw several **conclusions**. We summarize them below, you can

find more details in the corresponding .Rmd and .html files.

```
[1] "GLM for HadArthritis unbalanced ( 12 )"

Call:
glm(formula = formula, family = binomial, data = df)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
GeneralHealth      -0.1903273  0.0046738  -40.722 < 2e-16 ***
HaddDepressiveDisorder 0.4516615  0.0107875  41.869 < 2e-16 ***
DifficultyWalking     0.7380827  0.0124843  59.121 < 2e-16 ***
Agecategory         -0.2453016  0.0016982 144.447 < 2e-16 ***
Gender              -0.3979480  0.0108109  -36.810 < 2e-16 ***
PhysicalHealthDays   -0.0178631  0.0005239   34.094 < 2e-16 ***
HadAsthma           -0.3456244  0.0110825  -31.187 < 2e-16 ***
Lastcheckuptime      -0.1418907  0.0047679  -29.760 < 2e-16 ***
SleepHours          -3.8579970  0.1337599  -28.843 < 2e-16 ***
PneumovaxEver       -0.0726057  0.0026032  -27.891 < 2e-16 ***
RemovedTeeth        -0.2307474  0.0089789  -26.400 < 2e-16 ***
Chestscan           -0.1945269  0.0082492  -23.581 < 2e-16 ***
Smokerstatus        -0.1917944  0.0083786  -22.891 < 2e-16 ***
HIVTesting          -0.1484133  0.0083060  -17.868 < 2e-16 ***
HaddKidneyDisease   -0.1384012  0.0092735  -17.081 < 2e-16 ***
HaddSkinCancer      -0.2028097  0.0130166  -15.581 < 2e-16 ***
DifficultyConcentrating 0.1884455  0.0139657  13.493 < 2e-16 ***
Deaforhardofhearing 0.1615307  0.0130313  12.396 < 2e-16 ***
PhysicalActivities   -0.1161866  0.0095789  -12.129 < 2e-16 ***
HaddKidneyDisease   -0.2113426  0.0178806  -11.820 < 2e-16 ***
BMI                 -0.0229594  0.0019932  -11.519 < 2e-16 ***
CovidPos            -0.1013863  0.0089643  -11.310 < 2e-16 ***
AlcoholDrinkers     -0.0917674  0.0082728  -11.093 < 2e-16 ***
WeightInkilograms   -0.0069588  0.0006459  -10.773 < 2e-16 ***
HaddCOPD            -0.1400427  0.0145011   9.657 < 2e-16 ***
FluVaxLast12        -0.0819770  0.0086118   9.519 < 2e-16 ***
RaceEthnicityCategoryWhite only, Non-Hispanic -0.2462317  0.0261286  -9.425 < 2e-16 ***
RaceEthnicityCategoryMultiracial, Non-Hispanic -0.1953790  0.0218216  -8.953 < 2e-16 ***
RaceEthnicityCategoryBlack only, Non-Hispanic -0.2675031  0.0346993  -7.709 1.27e-14 ***
DifficultyErrands    -0.1075132  0.0175578  -6.123 9.16e-10 ***
HadAngina           -0.0990223  0.0168418  -5.880 4.11e-09 ***
MentalHealthDays    -0.0030059  0.0005602   5.366 8.04e-08 ***
State               -0.0011400  0.0002452   4.650 3.32e-06 ***
HeightInMeters      -0.3236017  0.0775169  -4.175 2.99e-05 ***
DifficultyDressingBathing 0.0833145  0.0233335   3.571 0.000356 ***
TetanusLast10Tdap   -0.0404826  0.0113689  -3.561 0.000370 ***
HadStroke           -0.0617222  0.0179256  -3.443 0.000575 ***
HadHeartAttack      -0.0483571  0.0171411  -2.821 0.004786 **
RaceEthnicityCategoryBlack only, Non-Hispanic -0.0676925  0.0253319  -2.651 0.008019 **
HaddDiabetes        -0.0278666  0.0109342  -2.549 0.010817 *
HaddDiabetes        -0.0599146  0.0238690  -2.510 0.012068 *
RaceEthnicityCategoryOther race only, Non-Hispanic -0.0701777  0.0289765  -2.422 0.015440 *
BlindorvisionDifficulty 0.0332361  0.0173576  1.915 0.055519 .
ECigaretteUsage     -0.0107229  0.0098422  -1.089 0.275942

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 542171  on 426042  degrees of freedom
Residual deviance: 418308  on 425999  degrees of freedom
AIC: 418396

Number of Fisher Scoring iterations: 5
```

```
[1] "GLM for HadArthritis balanced ( 11 )"

Call:
glm(formula = HadArthritis ~ ., family = binomial, data = df_balanced)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
DifficultyWalking     0.7336674  0.0155397  47.213 < 2e-16 ***
Agecategory           0.2543665  0.0019608 129.725 < 2e-16 ***
GeneralHealth         -0.1988636  0.0055041  -36.130 < 2e-16 ***
HaddDepressiveDisorder 0.4438486  0.0128289  34.598 < 2e-16 ***
Gender               -0.3809506  0.0126232  -30.179 < 2e-16 ***
PhysicalHealthDays   -0.0191187  0.0006450   29.640 < 2e-16 ***
HadAsthma            -0.3458622  0.0132257  -26.151 < 2e-16 ***
Lastcheckuptime      -0.1400611  0.0053964  -25.955 < 2e-16 ***
SleepHours          -0.0714054  0.0030996  -23.037 < 2e-16 ***
PneumovaxEver       -0.2259960  0.0106361  -21.248 < 2e-16 ***
RemovedTeeth        -3.2365891  0.1569929  -20.616 < 2e-16 ***
Chestscan           -0.1923713  0.0096497  -19.935 < 2e-16 ***
Smokerstatus        -0.1952010  0.0098684  -19.780 < 2e-16 ***
HIVTesting          -0.1694579  0.0097951  -17.300 < 2e-16 ***
HaddSkinCancer      -0.1657986  0.0108261  -15.315 < 2e-16 ***
DifficultyConcentrating 0.2114828  0.0159254  13.280 < 2e-16 ***
Deaforhardofhearing 0.1991530  0.0168471  11.821 < 2e-16 ***
PhysicalActivities   -0.1824431  0.0160532  -11.365 < 2e-16 ***
HaddKidneyDisease   -0.1089865  0.0114002   9.560 < 2e-16 ***
AlcoholDrinkers     -0.0922440  0.0097259   9.484 < 2e-16 ***
CovidPos            -0.0984305  0.0104868   9.386 < 2e-16 ***
BMI                 -0.1649975  0.0180135   9.160 < 2e-16 ***
WeightInkilograms   -0.0217044  0.0023731   9.146 < 2e-16 ***
HaddCOPD            -0.0069667  0.0007661   9.094 < 2e-16 ***
HaddKidneyDisease   -0.1985421  0.0221779   8.952 < 2e-16 ***
RaceEthnicityCategoryWhite only, Non-Hispanic -0.2558798  0.0302612  -8.456 < 2e-16 ***
RaceEthnicityCategoryMultiracial, Non-Hispanic -0.1986944  0.0255158  -7.787 6.86e-15 ***
FluVaxLast12        -0.0782034  0.0100719  -7.764 8.20e-15 ***
RaceEthnicityCategoryOther race only, Non-Hispanic -0.2728735  0.0405743  -6.725 1.75e-11 ***
DifficultyErrands    -0.0037373  0.0006679   5.595 2.20e-08 ***
HadAngina           -0.1120495  0.0209820   5.340 9.28e-08 ***
State               -0.0989162  0.0216410  -4.571 4.86e-06 ***
HeightInMeters      -0.0103555  0.0002877  -3.599 0.000320 ***
DifficultyDressingBathing 0.03016369  0.0090817  -3.315 0.000915 ***
TetanusLast10Tdap   -0.0404774  0.0133594  -3.030 0.002446 **
HadStroke           -0.0607631  0.0221552  -2.743 0.006095 **
RaceEthnicityCategoryOther race only, Non-Hispanic -0.0881426  0.0334689  -2.634 0.008449 **
DifficultyDressingBathing 0.0740045  0.0294897  2.510 0.012090 *
HadHeartAttack      -0.0498154  0.0211353  -2.357 0.018425 .
BlindorvisionDifficulty 0.0395983  0.0212538  1.863 0.062446 .
HaddDiabetes        -0.0405738  0.0270361  -1.501 0.133427
RaceEthnicityCategoryBlack only, Non-Hispanic -0.0399967  0.0298394  -1.340 0.180114
HaddDiabetes        -0.0104482  0.0132150  -0.791 0.429160
ECigaretteUsage     -0.0075014  0.0115366  -0.650 0.515544

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 393361  on 283749  degrees of freedom
Residual deviance: 299512  on 283706  degrees of freedom
AIC: 299600

Number of Fisher Scoring iterations: 4
```

Different ethnicities are not in the top half of the list of variables that predict different diseases. General

```

call:
glm(formula = formula, family = binomial, data = df)

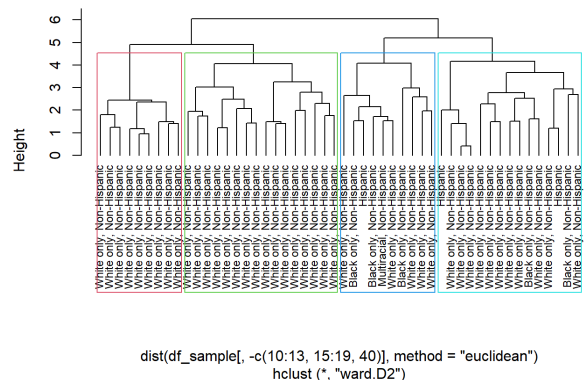
Coefficients:
Agecategory
(Intercept) -8.4424325 0.2279626 -37.034 < 2e-16 ***
RaceEthnicityCategoryBlack only, Non-Hispanic -2.1559455 0.0738637 -29.183 < 2e-16 ***
HadArthritis 0.2544783 0.0129034 19.024 < 2e-16 ***
PneumovaxEver 0.2627184 0.0141923 18.511 < 2e-16 ***
RaceEthnicityCategoryHispanic -0.9782193 0.0556312 -17.584 < 2e-16 ***
AlcoholDrinks 0.2232837 0.0128865 17.327 < 2e-16 ***
RaceEthnicityCategoryOther race only, Non-Hispanic -1.1170087 0.0669225 -16.691 < 2e-16 ***
PhysicianVisits 0.247231 0.0154265 15.789 < 2e-16 ***
FluVaxLast2 0.1866596 0.0141900 13.154 < 2e-16 ***
HadKidneyDisease 0.2911419 0.0226742 12.852 < 2e-16 ***
HeightInMeters 1.1515526 0.1302486 11.636 < 2e-16 ***
LastCheckupTime 0.0968944 0.0055511 11.323 < 2e-16 ***
chestXray 0.1389014 0.0130143 10.673 < 2e-16 ***
RaceEthnicityCategoryWhite only, Non-Hispanic 0.3673576 0.0383330 10.344 < 2e-16 ***
State -0.0036461 0.0003795 -9.607 < 2e-16 ***
RemovedTeeth -0.1200012 0.0129128 -9.293 < 2e-16 ***
CvdPos 0.1242243 0.0141489 8.780 < 2e-16 ***
PhysicalHealthDays 0.0056692 0.0007851 7.221 1.51e-13 ***
HearingInRt 0.1516344 0.0213099 7.116 1.11e-12 ***
HadAsthma 0.1084288 0.0175752 6.169 6.85e-10 ***
DeafForHardOfHearing 0.1023988 0.0172330 5.942 2.82e-09 ***
HIVTesting 0.0912601 0.0154930 5.890 3.85e-09 ***
HadDepressiveDisorder 0.1001597 0.0172949 5.791 6.99e-09 ***
DifficultyWalking -0.1021287 0.0168294 -5.542 3.34e-08 ***
EcigaretteUsage 0.0820093 0.0164332 4.995 5.87e-07 ***
RaceEthnicityCategoryMultiracial, Non-Hispanic -0.2993421 0.0676965 -4.422 9.79e-06 ***
HadDiabetes -0.0730963 0.0168033 -4.350 1.36e-05 ***
DifficultyErrands -0.1132567 0.0268281 -4.222 2.43e-05 ***
Gender -0.0716392 0.0173943 -4.119 3.81e-05 ***
WeightInKilograms -0.0046743 0.0011444 -4.085 4.41e-05 ***
SmokerStatus -0.0458282 0.0128156 -3.557 0.000375 ***
HadStroke 0.0797038 0.0243147 3.278 0.001045 ***
GeneralHealth -0.0233192 0.0074226 -3.142 0.01680 ***
HadHeartAttack -0.0618357 0.0231409 -2.672 0.07037 ***
DentalHealthDays 0.0022439 0.0009255 2.425 0.01532 ***
DifficultyDressingBathing 0.0768195 0.0337948 2.273 0.023019 ***
DifficultyConcentrating -0.0435165 0.0227850 -1.910 0.056149 .
BMI -0.0053036 0.0035504 -1.494 0.135223
TetanusLast10Tdap 0.0218751 0.0179329 1.220 0.225259
RightHskLastYear -0.0569897 0.0502703 -1.132 0.257566
HadCOPD 0.0199068 0.0203619 0.978 0.328248
SleepHours 0.0022989 0.0041367 0.556 0.578894
BlindOrVisionDifficulty -0.0026884 0.0263754 0.102 0.918315
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

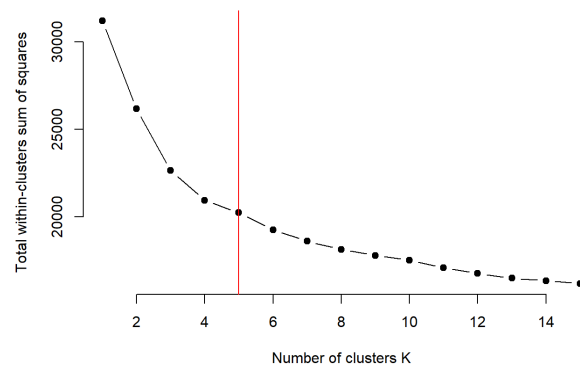
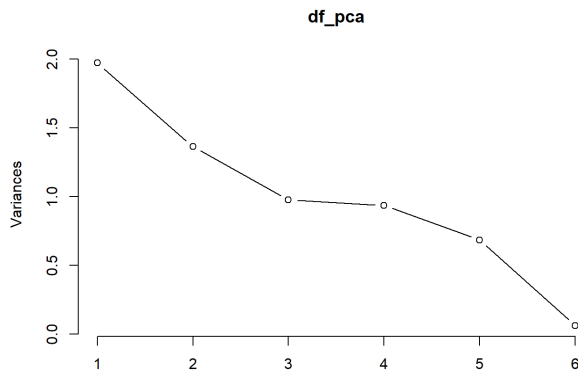
Null deviance: 239910 on 426042 degrees of freedom
Residual deviance: 200330 on 425999 degrees of freedom
AIC: 200418

Number of Fisher scoring iterations: 7

```



Finally, we also tried **K-means**. We started with a **PCA** analysis of the “real” numerical variables to try to reduce them, which are: “PhysicalHealthDays”, “MentalHealthDays”, “SleepHours”, “HeightInMeters”, “WeightInKilograms”, and “BMI”. Thanks to the **elbow plot**, we saw that we could reduce them to only 3 components. Another possible approach would have been to directly eliminate height and weight, as BMI is obtained using a formula with both of them.



Having already gathered enough evidence that there was no strong relationship between races and diseases, we decided to drastically reduce the dataset to only one disease, skin cancer, and reduce the number of rows to 4,000 records to avoid the relatively expensive K-means algorithm. Even so, our computer took almost an hour to calculate the Hopkins statistic.

In this reduced sample, we obtained an **H value of approximately 0.7**, indicating that the dataset was clusterable. We counted the number of unique races in the sample and used another elbow graph to determine the optimal number of clusters, obtaining a promising result: 5 is the number of ethnicities and also one of the points where the curve began to “flatten.”

Finally, we labeled each row according to the cluster in which it was grouped, and using a matrix, we checked if they corresponded to the race labels, obtaining a result consistent with what we saw with the previous methods: despite being able to be grouped into 5 clusters, they do not match the race values.

```
##
##           1   2   3   4   5
## Black only, Non-Hispanic    84  61  80  51  58
## Hispanic                    78  88 115  49  56
## Multiracial, Non-Hispanic   25  14  22  12  13
## Other race only, Non-Hispanic 67  68  45  18  39
## White only, Non-Hispanic   845 678 617 448 369
```

Since we have obtained compatible results using these approaches, we can conclude that although ethnicity can influence certain diseases, it is not a determining factor. Otherwise, we could have continued mining the data with other algorithms (PAM or DBSCAN) and even more methods (classification and association rules).