Feasibility of using machine learning algorithms in creating a shortened BRFSS questionnaire for heart disease and heart attack risk assessment

Friday 24th December, 2021

Daan Brugmans S1080742
Josse Nobel S1080646

Radboud University Nijmegen, Data Science Department

I. Abstract

Heart diseases and heart attacks are among the leading causes of death. Because of this, being able to assess the risks of developing cardiovascular diseases at an early stage is vital.

This research paper examines the feasibility of using machine learning models to create a short questionnaire based on the CDC's BRFSS questionnaire that can accurately determine the risks of developing cardiovascular diseases in participants by assessing the lifestyle choices of the participant. This study uses (gradient boosting) decision tree models and logistic regression models that are trained on a modified version of the CDC's 2015 BRFSS data set to predict the chance of a participant having experienced a cardiovascular disease. The research question for this study is as follows: "How can we create a shorter questionnaire from the existing CDC questionnaire such that the accuracy of predicting the risks for a heart disease or attack remains the same?".

The data set used contained 253,680 survey responses from different parts of the United States of America and was modified to only contain 22 categorical attributes. Machine learning models were trained on this data and used as guidelines to determine which questions are the predicting factors for heart diseases and attacks.

It was determined that the following nine factors included in the BRFSS questionnaire were the most vital factors for cardiovascular disease and heart attack risk assessment: (biological) sex, age, blood pressure, cholesterol, smoking, strokes, diabetic status, difficulties walking or climbing stairs, and a subjective general health score. From these results, a modified data set was created containing only the results of the questions pertaining to these nine factors. The models were trained again using the modified data set to determine the accuracy of only using these factors during the model training process.

Results showed that the accuracy of the models on both data sets showed a negligible difference in metrics, as metrics never differed more than 0.05% between models. The results of this paper indicate that these nine factors can be used by healthcare professionals to assess the risks of a patient developing cardiovascular diseases.

II. Introduction

Since 1984, the Center for Disease Control (CDC), a United States organisation, has been performing annual questionnaires under the collective name of the "BRFSS" to asses the physical and mental health of hundreds of thousands of American citizens across the US every year. One of the parts of this questionnaire concerns the risks of developing cardiovascular diseases. A lot of research has been performed in determining what lifestyle activities increase the risk of a heart attack and/or heart disease. The CDC uses many different factors to assess these risks. Risk assessment of heart disease and attacks at an early age or stage has the potential to be lifesaving and prevents any economic damage caused by a person getting a heart disease or attack. Cardiovascular diseases (CVDs) refer to several types of cardiovascular conditions, a general term for conditions affecting the heart or blood vessels. Heart diseases are the leading cause of death worldwide, taking an estimated 32% of all deaths every year. (World Health Organisation [8]). This fact highlights the importance of preventative measures and tests that can accurately predict if a person risks developing a heart disease.

Annual questionnaires from the Center for Disease Control (CDC) called the BRFSS contain many questions regarding the lifestyle of American citizens all across the nation, including whether of not citizens have had heart diseases or attacks in the past. The data collected from these questionnaires are published annually and may be accessed by anyone. A variety of lifestyle choices are recorded and all of them could have a say in a person's risk of getting heart disease or a heart attack. One could use such data to build a machine learning model that may predict a person's risk of heart disease or attack. However, since the amount of lifestyle factors taken into consideration for the data collected is fairly substantial, one cannot easily assign a small amount of lifestyle choices to the group of lifestyle choices that have a great affect on one's cardiovascular disease or attack risk. By analysing the BRFSS data, in addition to consulting previous literary works on the subject, one may be able to determine a small amount of lifestyle choices that have a strong connection with the increasing or decreasing of cardiovascular diseases and attacks. This paper talks about such analysis and describes one research done into reducing the amount of lifestyle factors taken into consideration when determining one's risk of heart disease or attack, which could lead to a reduction of the BRFSS's questionnaire size.

Much prior research has been done on determining the biggest risk factors for getting a heart disease or attack. For example, one study performed in 2000, based on a 1980's study, followed a large group of American women for 14 years and showed that there were specific correlations between different lifestyles and the frequency of heart diseases. This research also showed that different lifestyle factors, like smoking and drinking, could predict if a patient has a high risk of developing a heart disease (Primary Prevention of Coronary Heart Disease in Women through Diet and Lifestyle [7]).

Another study performed in 2006 on the same group of women as used by the previous study focused specifically on how BMI and obesity correlate to the risks of heart diseases and showed a significant correlation between a small weight gain in adulthood and heart diseases. By only measuring their hip-waist ratio and resulting BMI they could accurately predict if a woman had a high risk of getting a heart disease (Obesity as Compared With Physical Activity in Predicting Risk of Coronary Heart Disease in Women, [4]). These results and those of the previously discussed study coincides with preliminary analysis of our own data set, which shows correlations between the questions about BMI and smoking and the development of cardiovascular diseases.

A different study researching the possibility of using a two-item questionnaire asking about the amount of exercise and the amount of regularly ingested saturated fats to determine the risks of getting a heart disease in patients showed promising results. By asking just these two questions, the researchers were able to accurately predict the risks for getting a heart disease in patients (Feasibility of Using an Ultrashort Lifestyle Questionnaire to Predict Future Mortality Risk among Patients with Suspected Heart Disease, [6]).

A similar study to the two-item questionnaire study tried to assess the effects of five different lifestyles on developments regarding heart diseases and cancers. This study, performed on 291,778 different people from seven different European countries, showed a high correlation between different unhealthy lifestyles and the increase in heart diseases and cancers (Lifestyle factors and risk of multimorbidity of cancer and cardiometabolic diseases: a multinational cohort study, [5]).

Another similar study to this paper tried to verify the correlation between 27 different health factors such as smoking, blood pressure, genetics and more, with the risks of developing heart diseases. This study was focused on Irish men and women and showed strong correlations between certain lifestyles and the development of heart diseases. It also showed that there were a couple of factors which did not seem to accurately predict the risk of developing a heart disease. This suggests that only a certain amount of factors are needed to determine the risks of developing heart disease accurately (Comparison of the prediction by 27 different factors of cardiovascular diseases and death in men and women of the Scottish heart health study, [2]).

The results of the aforementioned studies suggest that the current BFRSS questionnaire used by the CDC can likely be shortened to a fraction of its current size, if one wished to use the BRFSS to query one's cardiovascular disease risk. Of the previously discussed studies, the most important factors in determining the risks of developing cardiovascular diseases were smoking, BMI, physical inactivity, and blood pressure. Preliminary analysis of the data set suggests that these factors are also the significantly correlated to the development of cardiovascular diseases. With the results of these studies in mind, the main question for this study is: "How can we create a shorter questionnaire from the existing CDC questionnaire such that the accuracy of predicting the risks for a heart disease or attack remains the same?"

In this research paper we will first examine the data set, which contains answers from a questionnaire given to 253,680 Americans in 2015, to determine what models we should create to predict the risks of getting a heart disease or attack. These models will be used as a baseline measure to determine which questions can be removed from the questionnaire while still giving the same accurate predictions. Using these models and our own analysis, we will determine which factors are the most important in predicting the risks of developing cardiovascular diseases. Using these results, we will create new models with these chosen factors to compare them against the baseline models. To conclude this research paper, we will present the best model and questionnaire that the CDC could use in the future to help American citizens that are at high risk of getting a heart disease or attack to get a preventative checkup. The results of this research paper should help the American Healthcare system to predict the risks of someone developing heart diseases or heart attacks. The results of this paper should also help with lowering costs and preventing unnecessary checkups without increasing the risk of developing a heart disease without the patient having got a checkup.

III. Methods

In this chapter, the chosen data set will be discussed, as well as the approach for answering the research question. A plan will be presented, which will be used to answer the main research question.

This study is a quantitative study focusing on a collection of American citizens, whose personal lifestyle data has been anonymously collected. This data will be analyzed for correlations between certain lifestyle choices and having experienced heart disease or attacks and such analysis is used in combination with previous literary work in order to produce a shortened BRFSS questionnaire.

The 2015 BFRSS data set consists of data from both landline call respondents and cell phone call respondents. This data set is an aggregate of data collected from all 50 US states, the District of Columbia, the territory of Guam, and the Commonwealth of Puerto Rico, in order to build a data set that represents the entirety of the American populace as correctly as possible. All data collected comes from adult respondents (at least 18 years of age). The total amount of participants recorded in the 2015 BRFSS data set is 253,680. Out of these participants, 23,893 reported ever suffering from a heart disease or attack, which is about 10% of the people surveyed, while 229,787 respondents reported never suffering from heart disease or attack.

The BRFSS questionnaire consists of a core component and optional modules. This questionnaire is conducted over telephone calls to a randomly selected set of telephone numbers from a known collection of telephone numbers. The target population for cellular samples consisted of people residing in private resident or college housing. States conducted questionnaire interviews during each calendar month, calling every day of the week, during daytime and evening hours.

As stated, the questionnaire consists of both a core component and optional modules. The core component is a standard set of questions used by all states during data collection. The contents of the core component are often taken from established national surveys like the National Health Interview Survey [3] and the National Health and Nutrition Examiniation Survey [1]. This is to ensure that questions included in the BRFSS have been tested thoroughly prior to usage.

During landline telephone number selection, the BRFSS divides telephone numbers into two groups. These groups are called strata. Strata are sampled separately. There are two strata for the 2015 BRFSS: high-density strata and medium-density strata. The strata a specific telephone number belongs to, depends on the number of listed household numbers in its hundred block, the set of 100 telephone numbers with the same area code, prefix, first two digits of the suffix, and all possible combinations of the last two digits. BRFSS puts numbers from these hundred blocks with at least one listed household numbers in either of the two strata. Both strata are then sampled in order to obtain a probability sample of all households with telephones. For cellular telephone numbers, the BRFSS's system is able to call random samples of numbers. The basis of the 2015 BRFSS cellular telephone number sampling is the Telcordia database of telephone exchanges.

For the 2015 BRFSS, all 53 states or territories used Computer-Assisted Telephone Interview (CATI) systems. Contractors and state health personnel conduct interview using CATI systems, following guidelines provided by the BRFSS.

For this study, a cleaned and modified version of the 2015 BRFSS questionnaire has been used. This data set was chosen, because it contains a large number of different questions, which each might have a different influence on the risks of developing a heart disease. This data set is a proper reflection of the real world, as the CDC has performed this survey since 1984 on around 400.000 different citizens every year. This data set thus gives a proper representation of the United States populace. By trying to improve this questionnaire, we hope to help the CDC, as well as other health professionals, in conducting surveys for assessing an individual's or group's risk of getting a heart disease or heart attack.

For this study, the used data set has been modified for it to be more suitable to both the reader and for machine learning purposes. In table III, all modifications for each variable are shown. All variables have been converted to categorical features in order to simplify the creation of the models. For all binary features, the categories ”0” and ”1” have been changed to ”no” and ”yes” respectively, to make it easier to understand the results of the created models.

Variable	**Modifications**
$HeartDiseaseorAttack$	’0’/’1’ → ’No’/’Yes’
Age	Converted categories expressed as numbers to categories expressed in words
$AnyHealthcare$	’0’/’1’ → ’No’/’Yes’
BMI	Converted numeric BMI values to 6 BMI categories based on the CDC's official BMI classification
$CholCheck$	’0’/’1’ → ’No’/’Yes’
$Diabetes$	’0’/’1’/’2’ → ’No diabetes’/’Pre-diabetes’/’Diabetes’
$DiffWalk$	’0’/’1’ → ’No’/’Yes’
$Education$	Converted categories expressed as numbers to categories expressed in words
$Fruits$	’0’/’1’ → ’No’/’Yes’
$GenHlth$	Converted categories expressed as numbers to categories expressed in words
$HighBP$	’0’/’1’ → ’No’/’Yes’
$HighChol$	’0’/’1’ → ’No’/’Yes’
$HvyAlcoholConsumption$	’0’/’1’ → ’No’/’Yes’
$Income$	Converted categories expressed as numbers to categories expressed in words
$MenthHlth$	Unaltered
$NoDocbcCost$	’0’/’1’ → ’No’/’Yes’
$PhysActivity$	’0’/’1’ → ’No’/’Yes’
$PhysHlth$	Unaltered
Sex	’0’/’1’ → ’Female’/’Male’
$Smoker$	’0’/’1’ → ’No’/’Yes’
$Stroke$	’0’/’1’ → ’No’/’Yes’
$Veggies$	’0’/’1’ → ’No’/’Yes’

TABLE I
MODIFICATIONS MADE TO DATASET

The initial data set analyzed contains a large class imbalance where 90.58% of participant has not developed cardiovascular diseases. To alleviate this imbalance, 80% of participants without cardiovascular diseases have been removed. The newly created data set that is used for the models contains 65.79% participants without cardiovascular diseases. While removing the data, the distributions across each answer have been preserved for people without heart diseases within a tenth of a percent.

A. Data Set

The data set is subdivided in three different subsets. The train set will be used to create the models and train them. The train set contains 60% of the total data used. The validation set will be used to assess the performance of a trained model. The validation set contains 30% of the total data used. The remaining 10% will be used as a test set to assess a model's performance after the model has been fine-tuned. The data has been divided in a stratified, random fashion among the three groups.

B. Research Question and Approach

The main research question for this study is: “How can we create a shorter questionnaire from the existing CDC questionnaire such that the accuracy of predicting the risks for a heart disease or attack remains the same?”

Our approach for this study to answer our research question is as follows: First, the previously discussed data set will be analysed to see if any clear correlations can be found between different variables and the target feature. Then, using this analysis, the data set will be split up into a train, validation, and test set, which will be used for training and testing the different models. We will train a logistic regression model in addition to a decision tree model. These models will be tuned with the validation set and when the models seem to predict the data by our set standards, we will test them a final time on the test set.

After these models have been created, we will combine the results from these models together with previous related studies as well as our analysis of the data set to determine the lifestyle factors that influence heart disease and attack risk assessment the most. With that knowledge, we will train different versions of these models, now only by using the feature selected. Lastly, we will compare the performance of these new models with the old models to determine if a model trained on a reduced number of features performs as well as a model performed on all features to determine if reducing the number of lifestyle choices considered for heart disease and attack risk assessment is feasible.

C. Tools

For this study, we will be using the R programming language with subsequent packages. In table III-C, all the used tools and software are displayed.

Software	**Version**	**Usage**
R programming language	4.1.2	data analysis, data preprocessing, model plotting, model creation
ggplot2 package	3.3.5	plotting distribution graphs, representing models graphically
scales package	1.1.1	plotting percentage based graphs
rattle package	5.4.0	model visualisation
dplyr package	1.0.7	data preprocessing, data manipulation
FactoMiner package	2.4	performing MCA, visualizing MCA
FactoExtra package	1.0.7	additions to FactoMiner package
caret package	6.0-90	machine learning model creation
MLeval package	0.3	drawing ROC curves

TABLE II
SOFTWARE PACKAGES USED FOR THIS STUDY

Two different types of algorithms will be used to create four different models. For both algorithms, one model has access to all variables of the questionnaire, while the other has access to only a select amount of features. The decision tree algorithm was chosen, because it will automatically try to choose the attributes that seem to have the most relevance in deciding the results of the target feature. The results of the decision tree model can then be used in creating a new questionnaire. The logistic regression algorithm was, chosen because the data set consists of a binary target value where a chance prediction might help understand the relation between the different attributes and the target values. The difference in the two different logistic regression models will tell us if there is a significant information loss and relationships among the different attributes in predicting the target value. Combining the analysis of the logistic regression models together with the results of the decision tree models will give the necessary results. With these results a final conclusion can be given on the quality of the new questionnaire.

D. Measurements

This study is divided into two parts. First, an analysis is performed on the current questionnaire to give us insight into the current correlations between the input features and the target feature. This analysis is included in appendix 1. To determine the correlation of these attributes, the data set is visualized in order to show the division per attribute value compared to the target value.

To measure the validity and performance of the different models in both the models created for the current dataset as well as the newly created dataset, some key indicators are taken into account. For every model, a confusion matrix will be shown, alongside a number of other metrics. III-D shows which measurements will be taken and how this will be presented in the results:

Reference → **Predicted ↓**	Positive	Negative
Positive		
Negative		

Accuracy	
Kappa	
P-value	
95% Confidence Interval	
No Information Rate	
Sensitivity	
Specificity	

TABLE III
CONFUSION MATRIX TEMPLATE

The accuracy, together with the sensitivity and specificity, will be used to determine the performance of the model. This will be compared to the no information rate to determine if the model actually performs better than random guessing. The p-value will be used together with the kappa value and the 95% confidence interval to determine if the results of the model can be considered significant compared to the accuracy. The kappa value was chosen, because we also want to see if the models have a high interrater value. Because this study is looking for the risks of developing a heart disease, a high kappa value is preferred. The sensitivity and specificity will be used for comparison to determine if the models are gravitating towards false positives or false negatives. Because the goal of this study is to create a shortened BRFSS questionnaire that can accurately determine the risks for developing coronary diseases, it is important that there is a balance between the sensitivity and specificity. Too many false positives, and more people than necessary will receive further medical assistance. Too many false negatives, and more people might develop coronary diseases without knowing.

E. *Decision Trees*

For this study, multiple models will be trained and compared to one another. In this section, the parameters of each of the build models are described and explained.

The first models that will be trained are the decision tree models. For this study, the decision was made to create two types of decision tree models. The first model will be a decision tree model trained using tenfold cross validation. The second model will be a decision tree that trained using a gradient boosting machine, or gbm.

The first decision tree model will be created using tenfold cross validation and will have five iterations. These parameters were chosen, because, across varying studies, they have been proven to give optimal results, while keeping the amount of computational power required low. This model was chosen as it gives us a clear indication of what attributes a base model will select based on correlation.

The second decision tree model will be trained using a gradient boosting machine (gbm). A gbm works by creating many smaller decision trees at the same time, using different variables, combining the results, and then iteratively improving upon the results. In our case, the model was trained with different amount of interaction depths, the amount of splits the tree may have, and also having a different amount of trees created alongside each other for each iteration. The gbm tries to create a model based on 0 through 60 separate trees with a step size of 5 on the interaction depths of 1, 2, 5 and 8. The number 60 was chosen based on the total amount of variable options our data set contains, namely the different categories per variable. The different interaction depths where chosen based on the results from the previous model, which showed that the model should have a tree that is sized between 2 and 8 layers.

IV. Results

As discussed in the previous chapter, four models were created. Two models are trained using all input features of the data set and two models are trained on a select group of input features. In this chapter, the results of these trainings are presented.

The first two models are a decision tree and logistic regression model with access to all input features. The results of these models, together with the results of the analysis of the data set, were used to reduce the amount of input features for the next models.

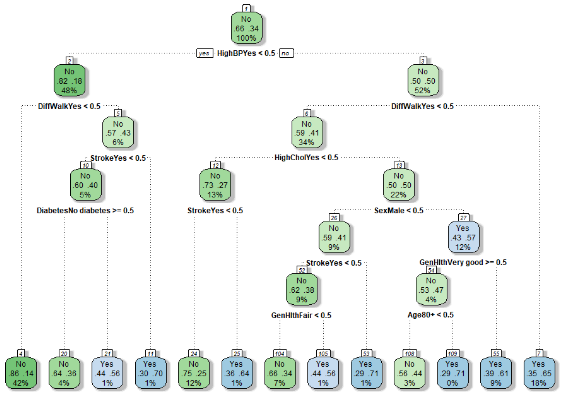
A. *Decision Tree model using all Input Features and Cross-Validation*

Reference → **Predicted ↓**	Positive	Negative
Positive	7574	2002
Negative	1628	2766

Accuracy	0.7402
Kappa	0.411
P-value	2.2e-16
95% Confidence Interval	(0.7328, 0.7474)
No Information Rate	0.6587
Sensitivity	0.8231
Specificity	0.5801

TABLE IV
Results of a decision tree with access to all features

Table IV-A shows the results from the decision tree model that was created using tenfold cross validation. The accuracy of 74.02% is a modest improvement above the no-information rate of 65.87%. However, the specificity, which determines the accuracy of guessing if a participant has a cardiovascular disease, is relatively low, with only 58.01%. Looking at figure 1 and 2, there are eight features that this model has used to assess the data: $HighBP$, $DiffWalk$, $Stroke$, $HighChol$, $Diabetes$, Sex, $GenHlth$, and Age. With an ROC of 0.75, it is better then just random guessing.



Rattle 2021-Dec-14 15:38:23 josse

Fig. 1. Decision Tree model with Cross-Validation

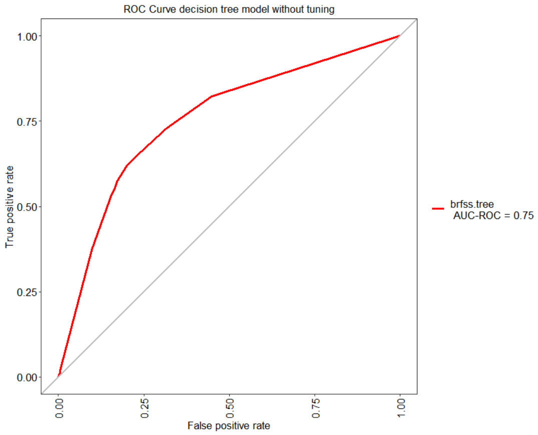


Fig. 2. ROC curve of Decision Tree model with Cross-Validation

B. GBM model using all Input Features

To get a more accurate model in the form of a decision tree, a Gradient Boosting Machine is used to create multiple decision trees at once and to iteratively improve the model. In table IV-B the results of the model are shown.

Reference → **Predicted ↓**	Positive	Negative
Positive	7830	1836
Negative	1331	2974

Accuracy	0.7733
Kappa	0.411
P-value	2.2e-16
95% Confidence Interval	(0.7663, 0.7802)
No Information Rate	0.6557
Sensitivity	0.8547
Specificity	0.6183

TABLE V
RESULTS OF A GRADIENT BOOSTED DECISION TREE USING ALL INPUT FEATURES

The accuracy of this model has marginally improved with 77.33%. Specifically, the specificity has improved from 58.01% in the previous model to 61.83%. Figure 3 also shows an improved ROC score of 0.85 compared to the previous models' 0.75.

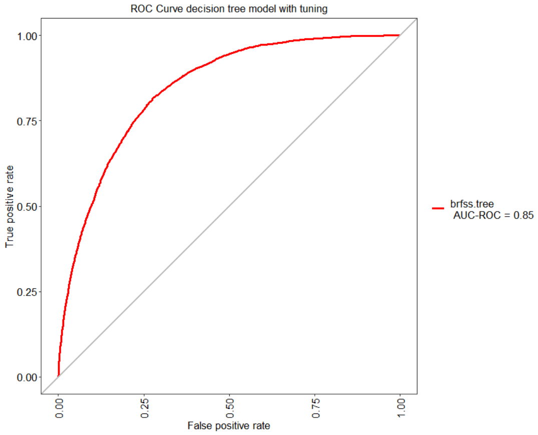


Fig. 3. ROC curve of decision tree model created with gradient boosting

Of this model, the 20 most important factors in determining if someone has a cardiovascular disease are shown in table IV-B.

Variable	**Importance score(0-100)**
HighBPYes	100.0000
HighCholYes	39.3165
DiffWalkYes	33.4618
StrokeYes	31.4177
Age80+	23.6511
GenHlthPoor	20.9974
GenHlthFair	20.6050
SexMale	20.1406
DiabetesDiabetes	16.5863
Age75 - 79	13.8969
SmokerYes	10.6232
GenHlthVery good	9.7804
Age70 - 74	8.5920
GenHlthGood	6.5387
Age65 - 69	6.0225
Age60 - 64	2.3644
Age35 - 39	2.3031
Age40 - 44	0.8068
Age45 - 49	0.6710
Age30 - 34	0.6304

TABLE VI
THE MOST IMPORTANT VARIABLES FOR A GRADIENT BOOSTED DECISION TREE ON THE WHOLE DATASET.

C. Logistic Regression Model on all features

In addition to the tree models, logistic regression models are also trained, which are used to predict an individual's relative chance of having suffered from a heart disease or heart attack. This logistic regression model had access to all input features of the data set and was trained using tenfold cross-validation.

Reference → **Predicted ↓**	Positive	Negative
Positive	11789	2879
Negative	1942	4345

Accuracy	0.7699
Kappa	0.4746
P-value	<2.2e-16
95% Confidence Interval	(0.7642, 0.7756)
No Information Rate	0.6553
Sensitivity	0.8586
Specificity	0.6015

TABLE VII
RESULTS OF A LOGISTIC REGRESSION MODEL USING ALL INPUT FEATURES

Table IV-D shows an overview of some metrics used to measure the performance of the logistic regression model. As may be seen, the model performs moderately well. With an accuracy of 76.99% and a sensitivity of 85.86%, the model seems to be fairly capable of producing reliable predictions. However, with a specificity of 60.15%, this model, just like the previous models, seems to have some more difficulties in reliably predicting the case of having experienced a heart disease or heart attack.

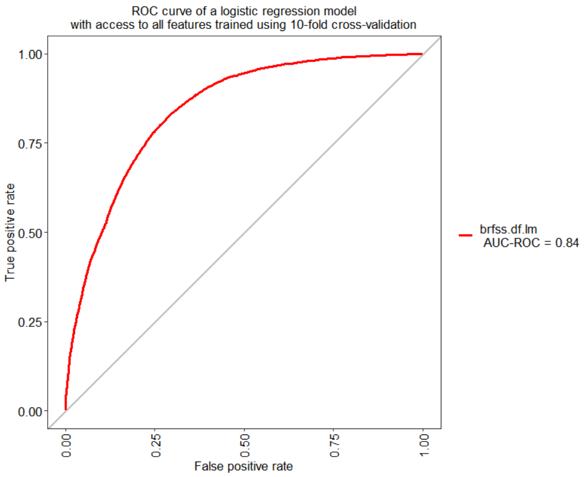


Fig. 4. ROC curve for the logistic regression model with access to all input features using tenfold cross-validation

In addition to this model, another, very similar, logistic regression model was trained. This model uses a boosting technique during training instead of a cross-validation technique. However, it seems that this did not benefit the model's performance, as its metrics were identical to the ones shown in IV-C. Its ROC curve is also identical to the one shown in 4. Its ROC curve is as so:

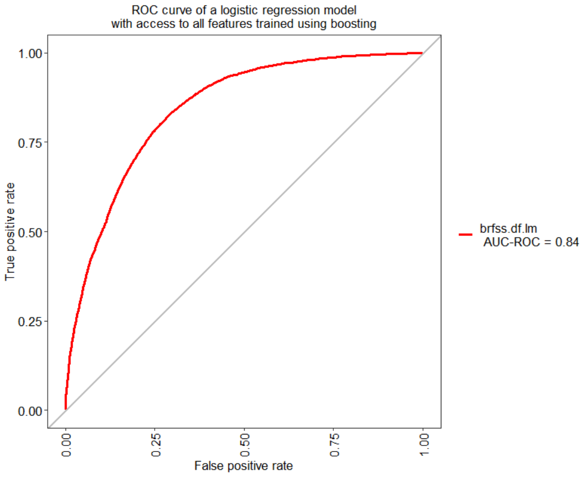


Fig. 5. ROC curve for the logistic regression model with access to all input features using boosting

D. Decision tree model using selected Input Features and Cross-Validation

Reference → **Predicted ↓**	Positive	Negative
Positive	7853	2275
Negative	1308	2535

Accuracy	0.7435
Kappa	0.4035
P-value	2.2e-16
95% Confidence Interval	(0.7362, 0.7508)
No Information Rate	0.6557
Sensitivity	0.8572
Specificity	0.5270

TABLE VIII
RESULTS OF A DECISION TREE USING SELECTED INPUT FEATURES

Table IV-D shows the results of the decision tree model build on the dataset with specifically chosen variables. The accuracy has increased marginally with 0.3% whereas the specificity has decreased by 4.66%. Figure 6 and 7 show the ROC curve and model of the decision tree. The amount of splits has increased as well as the leaf nodes. An increase from 13 to 17 leaf nodes has occurred.

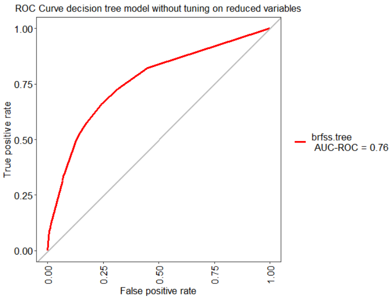


Fig. 6. ROC curve of decision tree model created with cross validation on the reduced variables

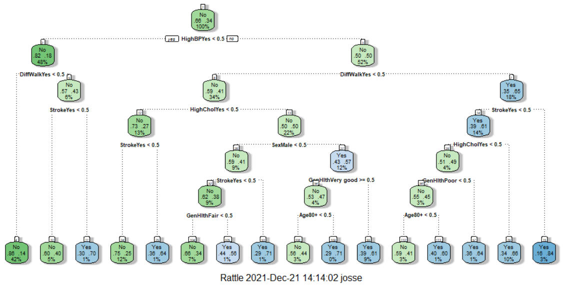


Fig. 7. Decision tree model with cross validation with reduced variables

E. GBM model using selected Input Features

Reference → **Predicted** ↓	Positive	Negative
Positive	7843	1854
Negative	1318	2956

Accuracy	0.7730
Kappa	0.4835
P-value	2.2e-16
95% Confidence Interval	(0.7659, 0.7799)
No Information Rate	0.6557
Sensitivity	0.8561
Specificity	0.6146

TABLE IX
RESULTS OF A GRADIENT BOOSTED DECISION TREE ON A REDUCED DATASET.

Table IV-E shows the result of the decision tree build with gradient boosting. with an accuracy of 0.7730 it is almost exactly as accurate as the model build including all of the variables. The specificity has decreased by 0.42% while the kappa value has increased by 0.0725. Figure 8 shows that the ROC score has been reduced a bit compared to the previous version. Table IV-E shows that the most important features still the same as with the model created on the whole data set.

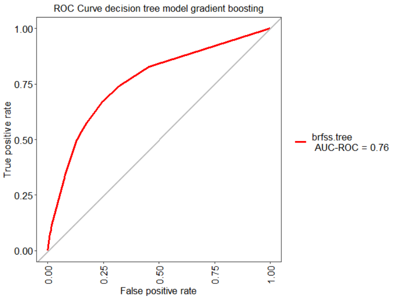


Fig. 8. ROC curve of decision tree model created with gradient boosting on the reduced variables

Variable	**Importance score(0-100)**
HighBPYes	100.0000
HighCholYes	39.3165
DiffWalkYes	33.4618
StrokeYes	31.4177
Age80+	23.6511
GenHlthPoor	20.9974
GenHlthFair	20.6050
SexMale	20.1406
DiabetesDiabetes	16.5863
Age75 - 79	13.8969
SmokerYes	10.6232
GenHlthVery good	9.7804
Age70 - 74	8.5920
GenHlthGood	6.5387
Age65 - 69	6.0225
Age60 - 64	2.3644
Age35 - 39	2.3031
Age40 - 44	0.8068
Age45 - 49	0.6710
Age30 - 34	0.6304

TABLE X
THE MOST IMPORTANT VARIABLES FOR A GRADIENT BOOSTED DECISION TREE ON THE WHOLE DATASET.

F. Logistic Regression Model on selected features

Just like the previously mentioned models, we trained a logistic regression model that had access to only a selected amount of features. These 9 features were perceived by us as the most important for heart disease and heart attack diagnosis. Using these 9 features, we reduced the amount of input features used by the model to less than half of the original. The goal of this dimensionality reduction is to produce a model that is (nearly) as capable of predicting one's heart disease and attack risk using only a small amount of factors. If such a model performed (nearly) as good as a model trained on all features, that could be proof that only a select group of BRFSS questionnaire questions are vital in heart disease and risk assessment. Our model, trained on the 9 features specified earlier, performed according to the following metrics:

Reference → **Predicted** ↓	Positive	Negative
Positive	11830	2907
Negative	1901	4317

Accuracy	0.7706
Kappa	0.4748
P-value	<2.2e-16
95% Confidence Interval	(0.7648, 0.7762)
No Information Rate	0.6553
Sensitivity	0.8616
Specificity	0.5976

TABLE XI
RESULTS OF A LOGISTIC REGRESSION MODEL USING SELECTED INPUT FEATURES

These metrics seem to be in favor of our goal. Compared to the logistic regression model mentioned in IV-C, this model's metrics seems to be nearly equal to the metrics of the previous logistic regression model. The accuracy and sensitivity have very slightly increased, from 76.99% to 77.07% and from 85.86% to 86.16% respectively. The specificity has slightly decreased, from 60.15% to 59.76%. These differences could be considered to be negligible, as any increase or decrease in these metrics between the two models are all less than .5 percent points. This means that one could conclude that the models perform practically equally, but such conclusions are more suitable for the Discussion section. For now, the ROC curve of this model is presented:

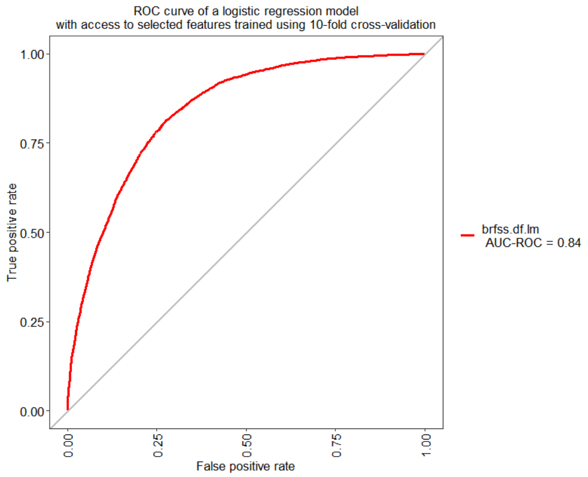


Fig. 9. ROC curve for the logistic regression model with access to selected input features using cross-validation

G. Summary

The results of the first three models gave nine variables in the data set that showed a strong relationship with determining if a participant has cardiovascular diseases. These are $HighBP$, $HighChol$, $Smoker$, $Diabetes$, $GenHlth$, $DiffWalk$, Sex, Age, and $Stroke$. Using these factors, new models were trained on the data set containing only the data from these variables. In table XII, XIII, and XIV the differences between the performance of each model before and after the reduction in the data set are shown.

	Before reduction	After reduction	difference
Accuracy	0.7402	0.7435	+0.0033
Kappa	0.411	0.4035	-0.0075
Sensitivity	0.8231	0.8572	+0.0341
Specificity	0.5801	0.5270	-0.0531

TABLE XII
TABLE SHOWING THE DIFFERENCE IN PERFORMANCE OF THE DECISION TREE MODEL TRAINED ON THE WHOLE DATA SET COMPARED TO THE REDUCED DATA SET

	Before reduction	After reduction	difference
Accuracy	0.7733	0.7730	-0.0003
Kappa	0.4111	0.4835	+0.0725
Sensitivity	0.8547	0.8561	+0.0014
Specificity	0.6183	0.6146	-0.0037

TABLE XIII
TABLE SHOWING THE DIFFERENCE IN PERFORMANCE OF THE GRADIENT BOOSTED DECISION TREE MODEL TRAINED ON THE WHOLE DATASET COMPARED TO THE REDUCED DATASET.

	Before reduction	After reduction	difference
Accuracy	0.7699	0.7706	+0.0007
Kappa	0.4746	0.4748	+0.0002
Sensitivity	0.8616	0.8561	-0.0055
Specificity	0.5976	0.6146	+0.017

TABLE XIV
TABLE SHOWING THE DIFFERENCE IN PERFORMANCE OF THE LOGISTIC REGRESSION MODEL TRAINED ON THE WHOLE DATASET COMPARED TO THE REDUCED DATASET.

V. DISCUSSION

During this study, we found that it is possible to train machine learning models for heart disease and heart attack risk risk assessment using the 2015 BRFSS questionnaire data and have these models perform well enough to be considered useful in cardiovascular disease risk assessment. We trained three different models, a decision tree, a gradient boosting machine, and a logistic regression model, both on all the BRFSS data set features and on a selected group of features that were deemed to have a close relationship with heart disease and attack risk, and compared the metrics of all models.

The results of the models created using all the input features showed accuracies ranging from 74% to 77.5% with specificities ranging from 52% to 62%. This is marginally better then the no information rate of 65.5%. The importance of the input features per model were compared and used to determine which variables contributed the most to determining if a participant has a cardiovascular disease. Nine attributes were determined to be strongly used by all models and were selected to create a new data set, and thus questionnaire, with only these nine factors. The new questionnaire is shown in table V-A.

A. Shortened Questionnaire

The results of the models trained on the new, reduced data set showed nearly identical accuracies and specificities for all models, with differences of less then 0.05 percent points when compared to the models trained on all the input features. This indicates that a questionnaire with only these questions/variables can still be used to predict if a patient has a high risk of developing cardiovascular disease as accurately as when the whole questionnaire is used.

Question	Possible answers
Do you have serious difficulty walking or climbing stairs?	Yes, No
Have you ever been told by a doctor, nurse or other health professional that your blood cholesterol is high?	Yes, No
Have you ever been told by a doctor, nurse or other health professional that you have high blood pressure?	Yes, No
Have you smoked at least 100 cigarettes in your entire life?	Yes, No
What is your diabetic status?	No diabetes, pre-diabetes, diabetes
What is your current age?	Age in years
What is your biological sex?	Female, Male
Have you ever had a stroke?	Yes, No
How would you rate your general health?	Poor, Fair, Good, Great, Excellent

TABLE XV
NEW QUESTIONNAIRE BASED ON THE RESULTS

Looking at the results compared to the literature study, the conclusion for our research question falls in line with previously performed research pertaining to this subject. The main question for this study was: ”How can we create a shorter questionnaire from the existing CDC questionnaire such that the accuracy of predicting the risks for a heart disease or attack remains the same?”. Using multiple different models to compare the performances on the different data sets gave us a resulting shortened questionnaire.

B. Possible Issues

For this study, we have reduced the class imbalance in the data set by removing 80% of the results of participants who did not have any cardiovascular diseases. With this, the distribution between participants with and without heart diseases is more balanced, as the ratios of no cardiovascular conditions/cardiovascular conditions went from 90.6%/9.4% to 65%/35% respectively. Because of the reduction and change in distribution, the results of this study may have been influenced. However, while changing the distribution, the ratios of the different attributes of the group without heart diseases were ensured to remain the same. Because of this, any patterns or relations between the different attributes and the development of cardiovascular diseases should have been preserved. A new study could be performed to ensure that the models will still accurately predict even with class imbalance.

An alternative solution to solving the class imbalance would have been to use imputation to add generated rows using the data already available to ”fake”, but realistic, participants that did develop heart diseases. However, because of the large size of this data set, we would have to generate tens of thousands of additional ”participants”. This would most likely influence the results heavily if not done properly. Because the original data set is already large with over 250.000 rows, removing a big portion of them still leaves us with over 65.000 rows.

Another important point of discussion is the lower specificity rates that all models produced. While all models performed similar, they all had relatively low specificities, meaning they were less accurate in predicting if a participant had a cardiovascular disease. This means that the models will produce relatively more false positives where a participant is still at risk of developing a heart disease. Because this study specifically looks into reducing the questionnaire while still having the same performance as with the whole data set, less research was spent on determining if this questionnaire is suitable to be used by a machine learning model. In a future study, this data set should be examined more closely to determine if a machine learning model can be used to more accurately predict the risks of a participant developing cardiovascular diseases.

The dataset used for this study was a modified version of the 2015 BRFSS’ questionnaire from the CDC. The specifics of this cleanup can be found using this link. The cleanup of this data set includes the removal of empty rows and dimensionality reduction. The data set used for this study consists of answers to questionnaires that were answered by anonymous participants. This means that the accuracy of the data could be put into question, especially for questions whose answers are on a loosely defined scale, such as questions regarding one’s general mental and general physical health. However, given that the population size of the number of participants interviewed is roughly a quarter of a million, it’s likely fair to say that represents the American populace.

VI. Appendix 1: Data Analysis

The data set used for this research is a subset of all data collected by the CDC’s 2015 BRFSS survey. Prior to using this data set for machine learning and prediction purposes, the data must first be preprocessed, analysed, and understood if one wishes to produce reliable prediction models. In this appendix, the data set is more thoroughly analysed than in the research paper itself, with the goal of allowing the reader to obtain a better understanding of the data used.

This appendix will start by giving a brief overview of the data set used, then it will dive more in depth into the features present in the data set. For every feature in the data set (except the target feature), two bar plots will be shown. For both plots, their bars are colored/separated according to the target feature, showing the share of target feature values per independent feature value. One plot shows the count of the distribution of all values for that feature and the other bar plot shows the relative share of the target feature values for every independent feature value present.

The data set used for this paper is a filtered data set of the CDC’s 2015 BRFSS survey. It has a binary target feature, 21 categorical features, and it contains no missing values. The target feature, $HeartDiseaseorAttack$, is a binary feature that tells us whether a person has experienced a heart disease or heart attack in the past. The other 21 features are attributes used to predict the target feature and consist of data regarding a person’s lifestyle choices and their current position in life. The values of a majority of features in the data set are answers to specific questions asked to BRFSS questionnaire participants, which means that a majority of the features have a question associated with them. For every feature where this is applicable, those questions will be noted as well.

At times, the plots features in this appendix may be difficult to read. If the reader wishes to see these plots in full size, they may follow this link that leads to the image directory of the corresponding GitHub repository.

$HeartDiseaseorAttack$ is the target feature of the data set. It is a binary (categorical) variable and records the amount of people that have suffered from a heart disease or heart attack in the past. Specifically, the question asked to participants was: ”Has a doctor, nurse, or other health professional ever told you that you had coronary heart disease or a myocardial infarction (heart attack)?”.

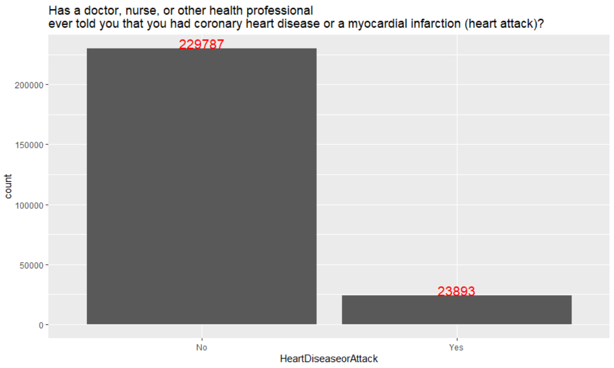


Fig. 10. Distribution of people that have suffered a heart disease or attack

Here, it becomes clear that only 23,893 people, about 10% of the total data set, have recorded ever having suffered from a heart disease or heart attack. This imbalance in target feature value spread should and has been accounted for during the machine learning process.

Age is a categorical variable that describes the age category of a participant. There are 13 different age categories, spanning all possible adult ages.

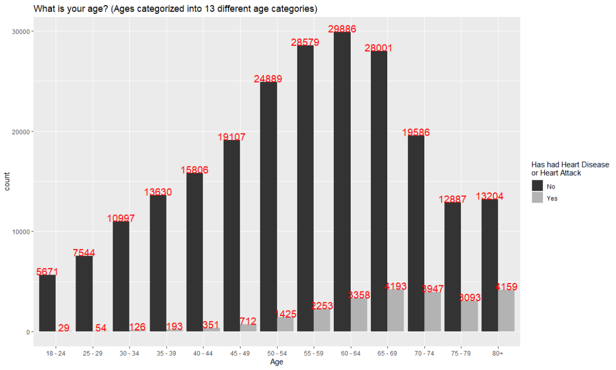


Fig. 11. Distribution of age categories among participants

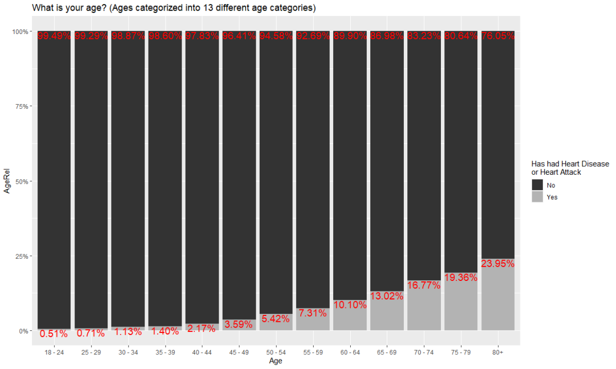


Fig. 12. Share of Yes/No to having suffered from a heart disease or attack among all age categories

Two things are swiftly noted: first, the age distribution seems Gaussian with a negative skew, the mode lying at 60 - 64. Secondly, the higher one's age, the more likely it is that one has experienced a heart attack or heart disease: less than 1% of all participants under 30 have experienced a heart attack or disease, while this share is well over 15% for participants aged 70 or older.

AnyHealthcare is a binary categorical variable that describes whether participants had any form of healthcare at the time of the questionnaire. Specifically, participants were asked: "Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service?"

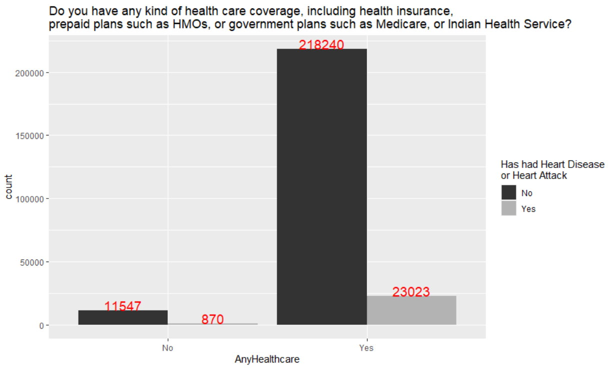


Fig. 13. Distribution of healthcare status among participants

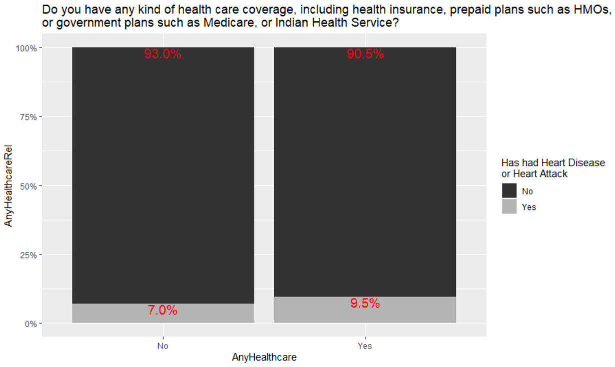


Fig. 14. Share of Yes/No to having suffered from a heart disease or attack among all AnyHealthcare categories

From the relative distribution, we may say that there is a minimal difference in one's cardiovascular disease risk between participants who did have a form of healthcare and those that didn't, as the difference is just 2.5 percentage points.

BMI is a categorical variable that shows the distribution of the BMI categories of the participants. These categories are based on the CDC's official body mass index categorisation and are as follows:

- Underweight ($BMI < 18.5$)
- Healthy weight ($18.5 \leq BMI < 25$)
- Overweight ($25 \leq BMI < 30$)
- Obese ($BMI \geq 30$)

The "Obese" category is further subdivided into three levels of obesity:

- Class 1 Obese ($30 \leq BMI < 35$)

- Class 2 Obese ($35 \leq BMI < 40$)
- Class 3 Obese ($BMI \geq 40$)

Participants were asked their actual BMI, not their BMI category, as the BMI values were categorized after the surveys had been performed.

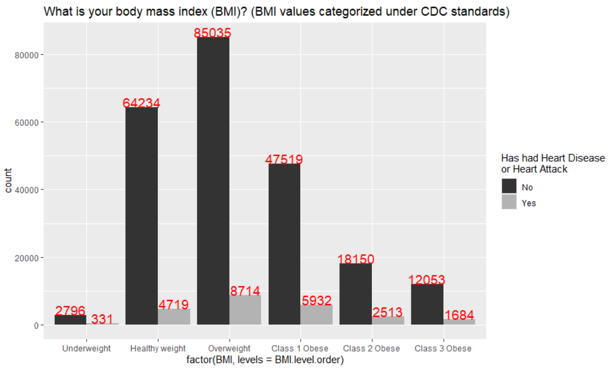


Fig. 15. Distribution of healthcare status among participants

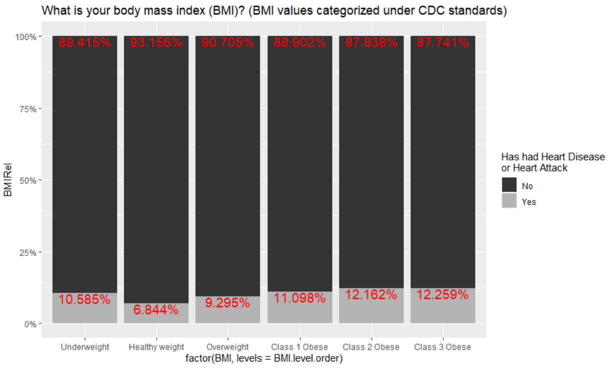


Fig. 16. Share of Yes/No to having suffered from a heart disease or attack among all BMI categories

For this data set, the mode of the BMI variable is overweight, with healthy weight being the second mode. It is also clear that the amount of underweight respondents is relatively small. Furthermore, some interesting findings can be found in the shares of $HeartDiseaseorAttack = Yes$: the share of people that has experienced a heart disease or attack is the lowest in the healthy weight category, and this share increases as one moves away from a healthy BMI. That fact holds true for both underweight and overweight participants. For overweight participants, the more severe their overweight BMI classification is, the more likely it is that they have experienced a heart attack or disease.

$CholCheck$ is a binary categorical variable that describes whether a person has had their cholesterol checked by a health professional. Participants were asked: ”Have you had a cholesterol check within the past five years?”.

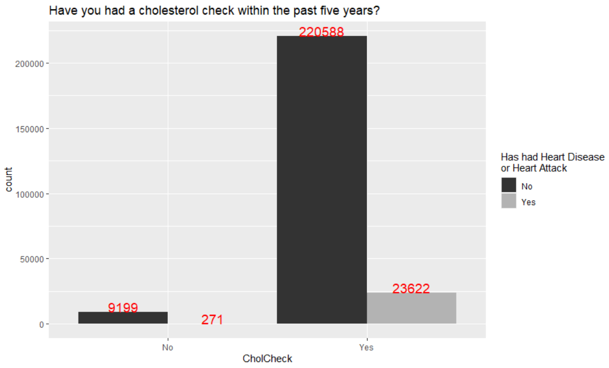


Fig. 17. Distribution of participants that had their cholesterol checked by a healthcare professional

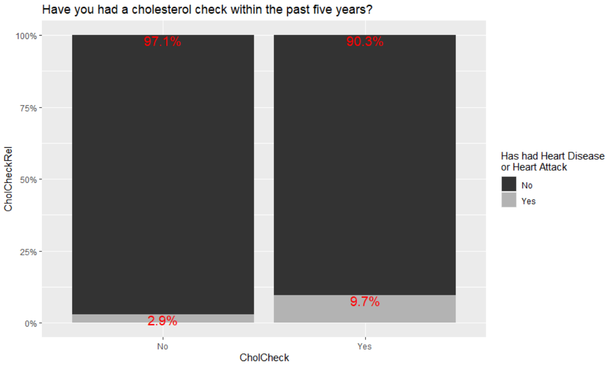


Fig. 18. Share of Yes/No to having suffered from a heart disease or attack among all CholCheck categories

From the second plot, it seems that the number of people that have suffered from a heart attack or heart disease seems lower for participants that haven't had their cholesterol checked. Two explanations for this phenomenon are as follows: the amount of participants that said ”no” to these questions is low, as may be seen in the first plot, meaning that this could be a statistical oddity, and that participants that hadn't had their cholesterol checked are more prone to not having other conditions checked, which could mean that this group of people have relatively less official diagnoses for heart disease or attack. Since participants were asked whether a health professional has told them that they had suffered from a heart attack or heart disease, participants that do not visit health professionals will have less diagnoses and are more likely to say ”no” to these types of questions.

$Diabetes$ is a categorical variable that describes whether a participants did not have diabetes, were in a pre-diabetic stage, or had diabetes at the time of interviewing. Officially, participants were asked: ”Were you ever told by a doctor, nurse, or other health professional that you had diabetes?”.

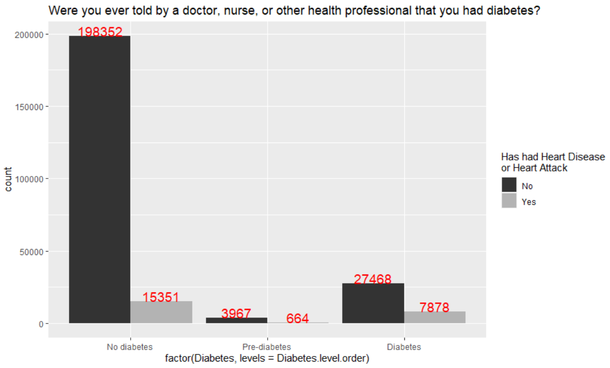


Fig. 19. Distribution of participants and their diabetic status

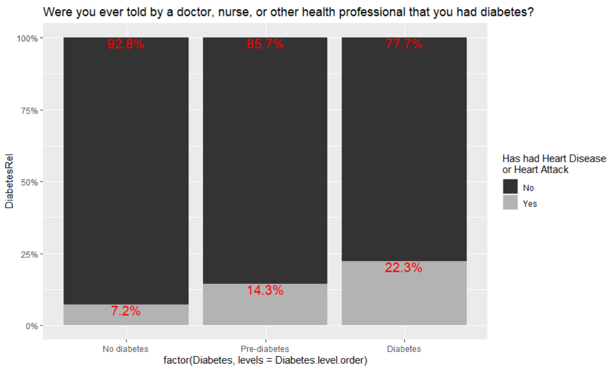


Fig. 20. Share of Yes/No to having suffered from a heart disease or attack among all Diabetes categories

From this data, there is a definite correlation between diabetes and heart disease and attacks: the share of participants with diabetes that had also suffered from a heart attack or disease in the past is thrice as high as the share of participants that didn't have diabetes. This number is twice as high for participants with pre-diabetes.

$DiffWalk$ is a binary categorical variable that describes whether a participant had difficulties walking and/or climbing stairs. Specifically, participants were asked: ”Do you have serious difficulty walking or climbing stairs?”.

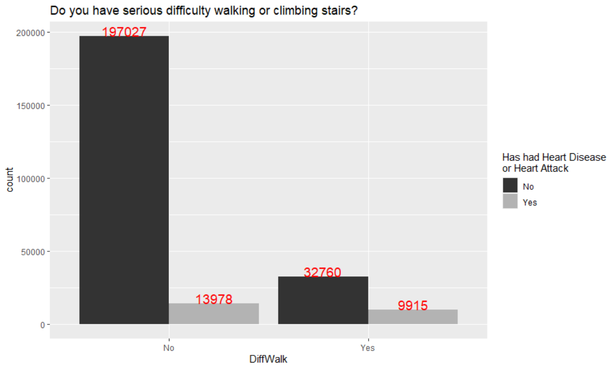


Fig. 21. Distribution of participants that had/hadn't any difficulties walking or climbing stairs

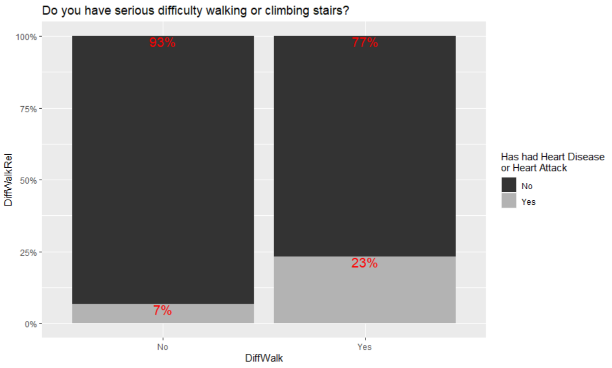


Fig. 22. Relative distribution of participants that had/hadn't suffered from a heart disease or heart attack and if they had difficulties walking or climbing stairs or not

For this data, it is clear that there is a connection between difficulty walking and the risk of heart disease and/or attacks, as participants that reported having difficulties with walking or climbing stairs were thrice as likely to have suffered from a heart disease or attack than participants who reported not having difficulties with walking or climbing stairs.

$Education$ is a categorical variable that describes the highest degree of education achieved by a participant. Participants' answers were categorized in 6 categories.

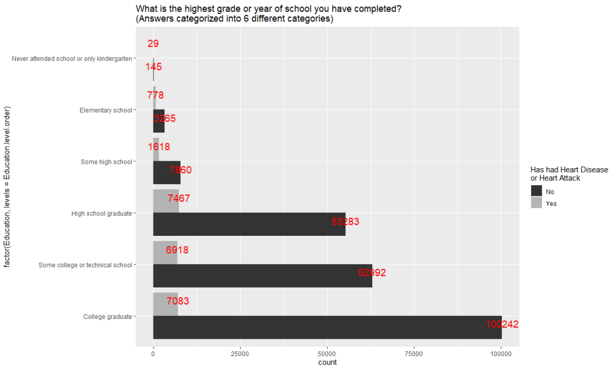


Fig. 23. Distribution of participants and their highest achieved academic level

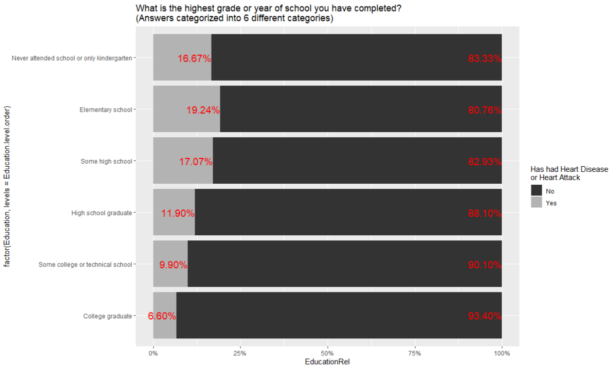


Fig. 24. Relative distribution of participants that had/hadn't suffered from a heart disease or heart attack and their highest achieved academical level

From these graphs, it may be concluded that, generally, the higher one's highest academical level achieved, the lower their risk of having suffered from a heart disease or attack. This is likely due to the fact that educated participants have had more opportunities to learn about lifestyle factors that influence one's health.

$Fruits$ is a binary categorical variable describing whether a participant consumes at least 1 portion of fruit daily. Participants were asked: "Do you consume at least 1 portion of fruit daily?".

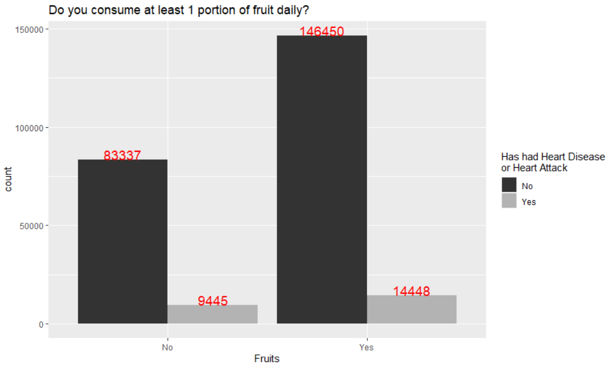


Fig. 25. Distribution of participants and whether or not they consume 1 portion of fruit daily

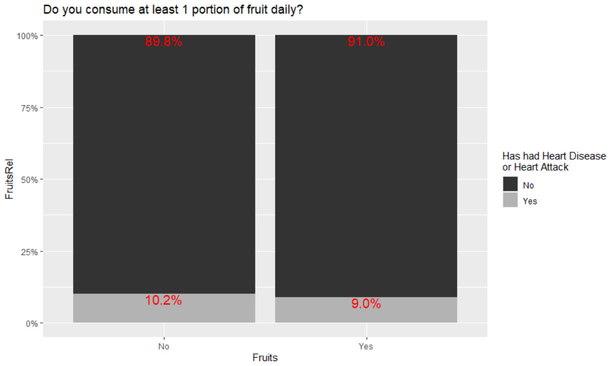


Fig. 26. Relative distribution of participants that had/hadn't suffered from a heart disease or heart attack and whether or not they consume 1 portion of fruit daily

From these graphs, there isn't a clearly definite correlation between $HeartDiseaseorAttack$ and $Fruits$, as the difference between shares of $HeartDiseaseorAttack$ is relatively small. Nonetheless, it seems that participants that ate fruit daily had a slightly smaller chance of having suffered from a heart disease or attack in the past.

$GenHlth$ is a categorical variable that describes the personal judgement of one's own health in general for a participant.

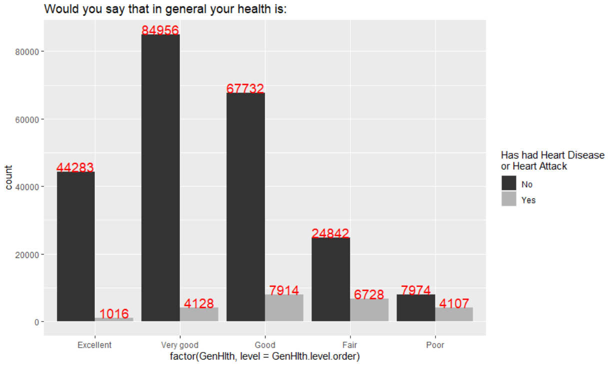


Fig. 27. Distribution of participants and their own perceived general health

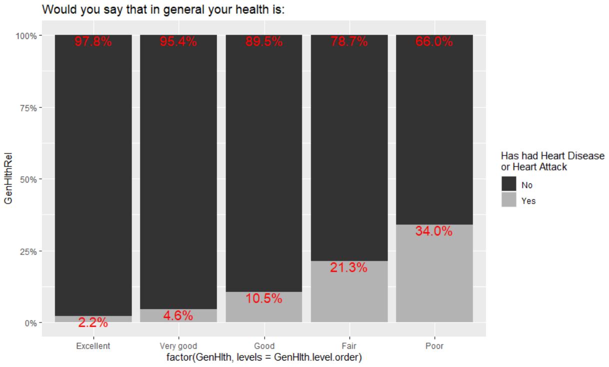


Fig. 28. Relative distribution of participants that had/hadn't suffered from a heart disease or heart attack and their own perceived general health

From these graphs, it is clear that participants that judged their own general health to be relatively worse, also had a higher chance of having suffered from a heart disease or heart attack in the past. Likely, participants have a solid grasp on their general health and this reflects in the number of participants that have experienced a heart attack or disease.

$HighBP$ is a binary categorical value and describes whether a medical professional has told a participant that they have a high blood pressure or not. Participants were asked: "Have you ever been told by a doctor, nurse, or or other health professional that you have high blood pressure?".

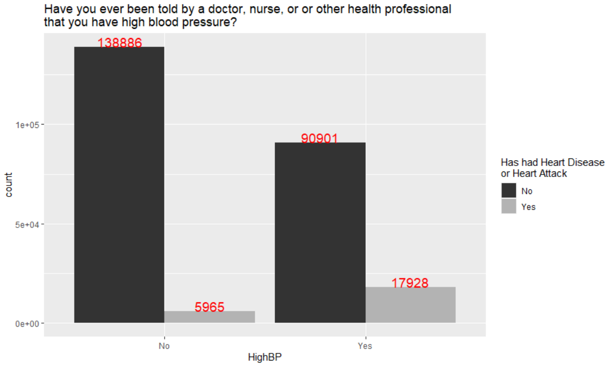


Fig. 29. Distribution of participants and if participants had a high blood pressure

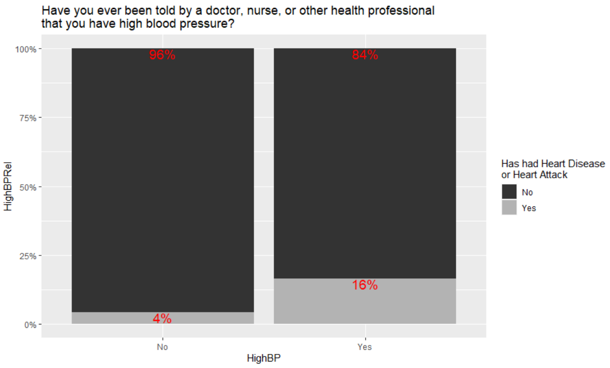


Fig. 30. Relative distribution of participants that had/hadn't suffered from a heart disease or heart attack and whether a health professional has told them they had a high blood pressure

This graph shows that there is a definite relationship between being professionally diagnosed with a high blood pressure and an increased chance of having experienced a heart disease or attack: the relative number of participants that had had such a cardiovascular condition in addition to having a high blood pressure is four times as high as participants who didn't report ever having a high blood pressure.

$HighChol$ is a binary categorical variable that describes whether a participant had a high cholesterol as assessed by a healthcare professional or not. Specifically, participants were asked: "Have you ever been told by a doctor, nurse, or other health professional that your blood cholesterol is high?".

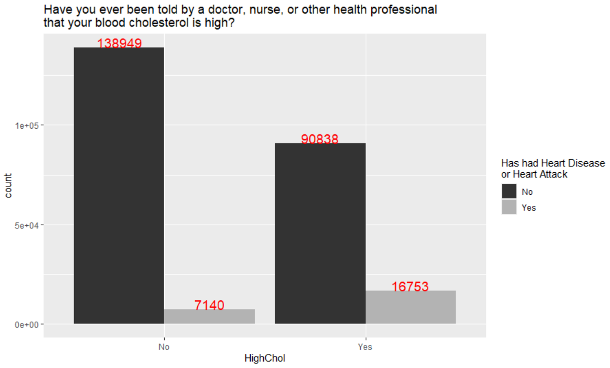


Fig. 31. Distribution of participants and if participants had a high cholesterol

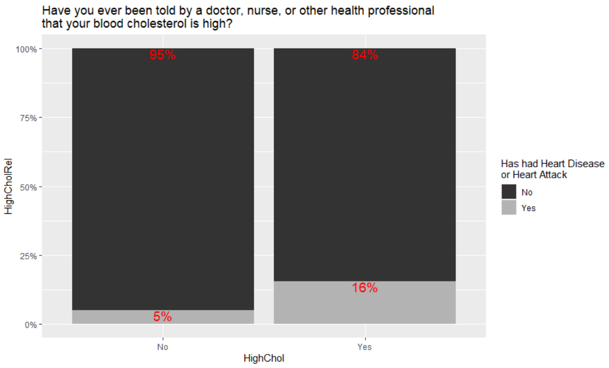


Fig. 32. Relative distribution of participants that had/hadn't suffered from a heart disease or heart attack and whether a health professional has told them they had a high cholesterol

This variable is very similar in both nature and values to the variable $HighBP$: there is a clear correlation between having been diagnosed with high cholesterol and having experienced heart attacks or diseases. This does not come as a revelation, however, as the relationship between cholesterol and the risk of cardiovascular diseases is a well-established fact.

$HvyAlcoholConsump$ is a binary categorical variable that describes if a participant consumed heavy amounts of alcohol. The set amount for being considered a "heavy drinker" is 14 drinks for male participants and 7 drinks for female participants. Participants were asked: "Is your weekly alcohol consumption considered heavy?".

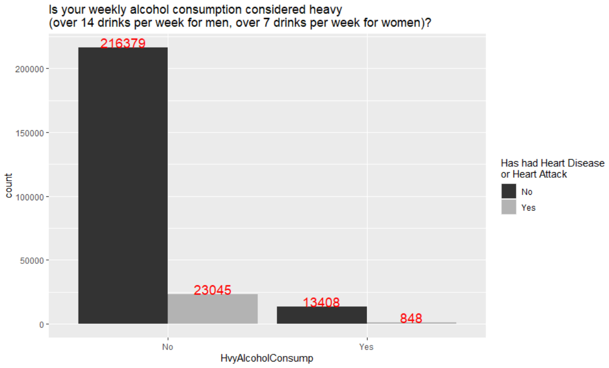


Fig. 33. Distribution of participants and whether they were considered to be heavy drinkers

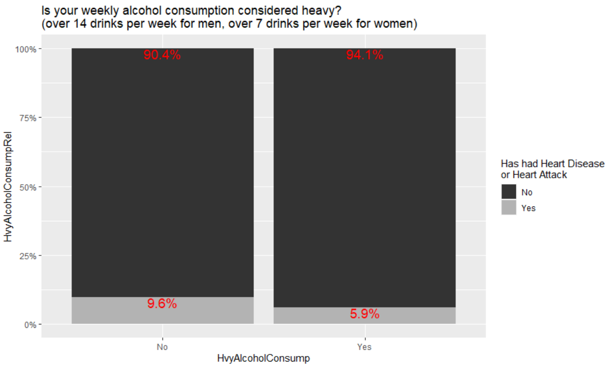


Fig. 34. Relative distribution of participants that had/hadn't suffered from a heart disease or heart attack and whether participants were considered heavy drinkers

These figures would imply that a heavy alcohol intake lowers the chance of having a cardiovascular condition. However, it is more likely that this is a statistical oddity and could be explained by the sampling size, possibly something else.

$Income$ is a categorical variable that describes the income category of a participant. There are 8 different income categories. The income of a participant is defined as the "annual household income from all sources".

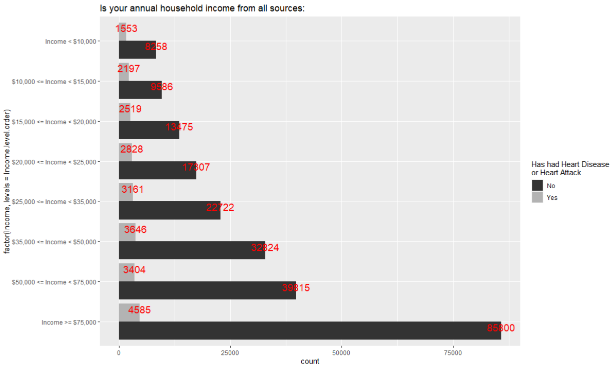


Fig. 35. Distribution of participants and their annual household income

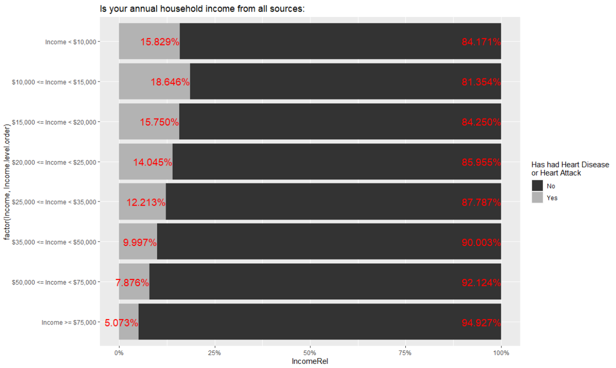


Fig. 36. Relative distribution of participants that had/hadn't suffered from a heart disease or heart attack and their income category

From these graphs, it may be said that, generally, participants with a lower household income have a higher chance to have suffered from a heart disease or heart attack in the past. This could be due to the lack of medical options low-income households have, as healthcare isn't a given for many American households, meaning that a relatively higher amount of low-income households do not have proper access to healthcare, decreasing the chances of preventing cardiovascular disease and increasing the risk of getting one.

$MentHlth$ is a categorical variable that describes the number of days a participant considered their mental health to be poor in the past 30 days as of the interview. Specifically, participants were asked: "Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?".

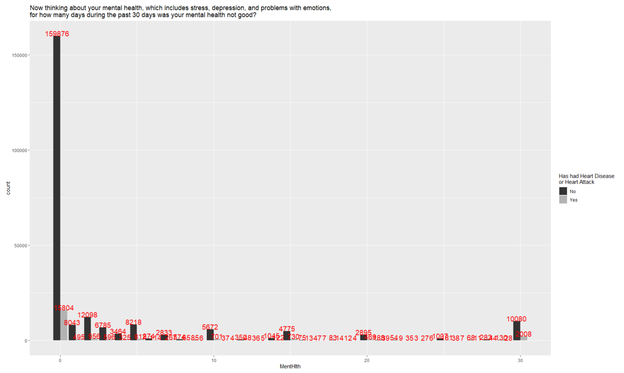


Fig. 37. Distribution of participants and the number of days a participant considered their mental health to be poor

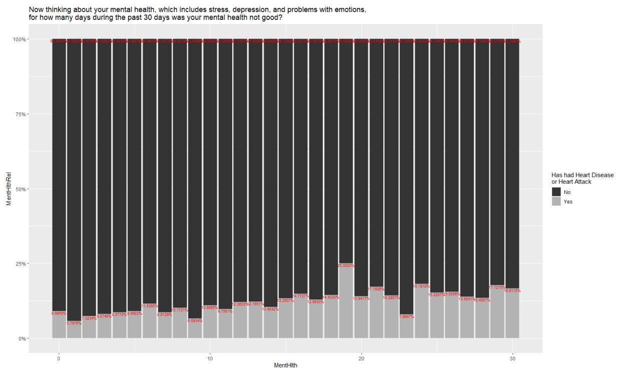


Fig. 38. Relative distribution of participants that had/hadn't suffered from a heart disease or heart attack and the number of days a participant considered their mental health to be poor

There isn't a lot one may discern from these graphs, since the distribution of participants is quite sparse for most days, with the notable exception of 0 days.

$NoDocbcCost$ is a binary categorical value that shows the share of participants who were not able to get healthcare in the past 12 months due to costs. Participants were asked: "Was there a time in the past 12 months when you needed to see a doctor, but could not because of cost?".

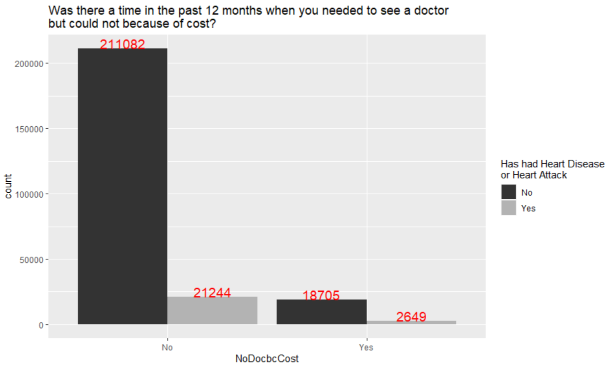


Fig. 39. Absolute values of share of heart disease among participants divided on the ability to pay for healthcare

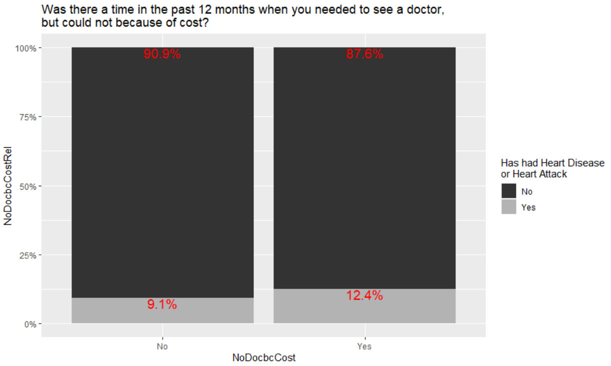


Fig. 40. Relative values of share of heart disease among participants divided on the ability to pay for healthcare

91.58% of participants did not have any economical problems getting healthcare. Of those, 9.1% have developed coronary diseases. Of the remaining 8.42% participants, 12.4% have developed coronary diseases. This is a slight increase compared to the other group, however, because the difference in group sizes is major, the correlation seems to be only slightly significant.

The $PhysActivity$ value is a binary categorical value where participants have to answer if they participated in any other physical activity or exercise other than their work. Participants were asked: ”During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?”.

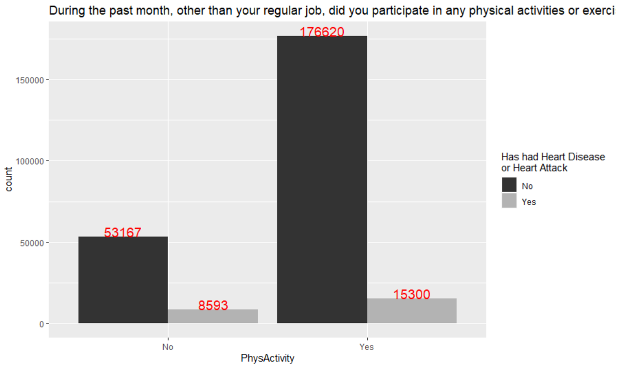


Fig. 41. Absolute values of share of heart disease among participants who participate in ant physical activities or exercise.

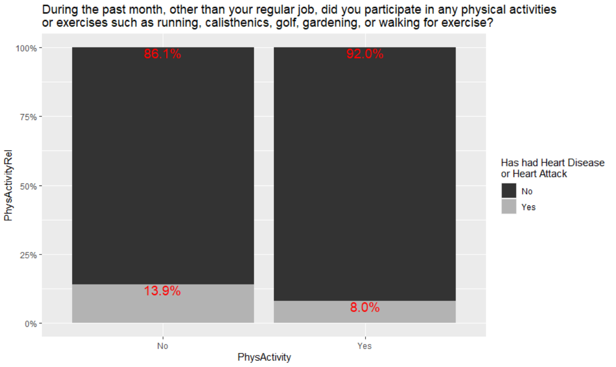


Fig. 42. Relative values of share of heart disease among participants who participate in ant physical activities or exercise.

75.56% of the participants participated at least once in a while in other physical activities or exercises. Of these, 8.0% have developed a coronary disease. Of the remaining 24.46%, 13.9% developed coronary diseases. This is a significant increase compared to participants who do partake in physical activities more regularly.

$PhysHlth$ concerns the amount of days given the previous 30 days on the day this question was asked when the participant was physically unwell. Participants were asked: ”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?”.

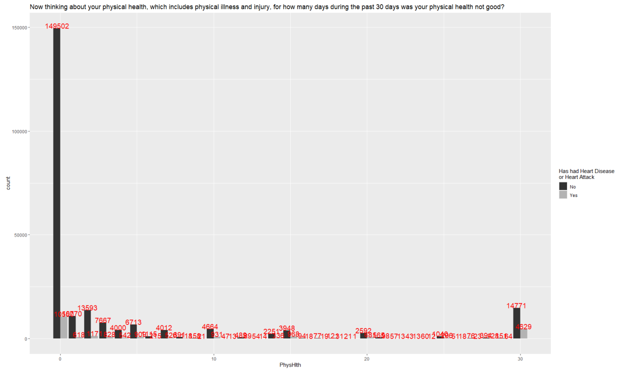


Fig. 43. Absolute values of share of heart disease among participants rating their own physical health

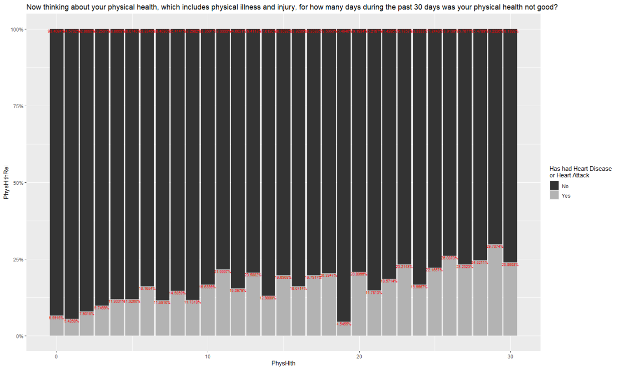


Fig. 44. Relative values of share of heart disease among participants rating their own physical health

63.09% answered that they felt physically unwell for zero of the previous 30 days. The rest of the participants answered with either a few days or a lot of days.

Sex is a binary categorical value determining a participant's biological sex.

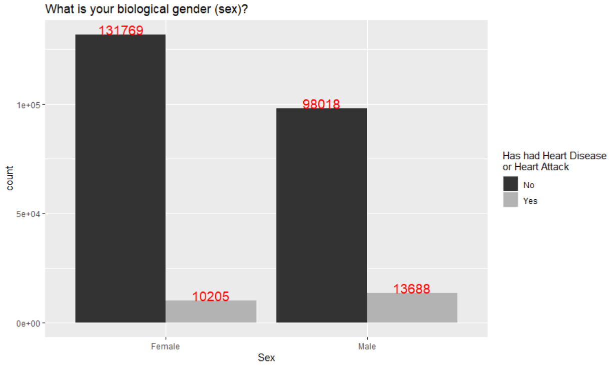


Fig. 45. Absolute values of share of heart disease among participants comparing women against men

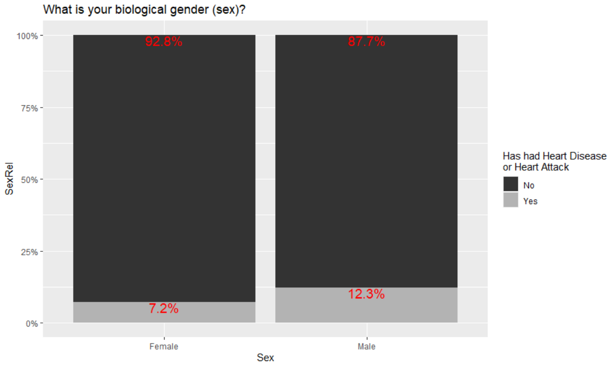


Fig. 46. Relative values of share of heart disease among participants comparing women against men

55.96% of participant are female, of which 7.2% have developed coronary diseases. The remaining 44.04% of participants were male, of which 12.3% have developed coronary diseases. One may notice that the share of those who have experienced a heart attack or disease is higher among male participants than female participants.

Smoker concerns participants who have consumed at least a 100 cigarettes (5 packs) throughout their life.

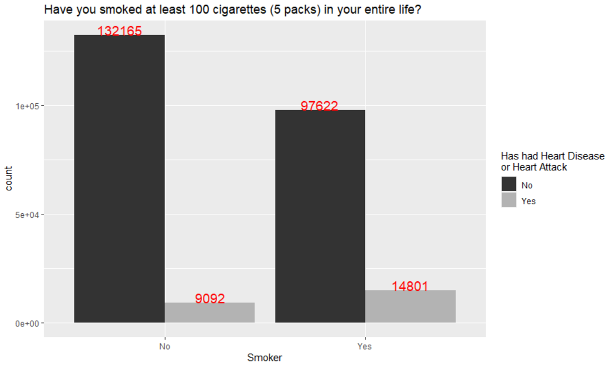


Fig. 47. Absolute values of share of heart disease among participants who are considered a regular smoker

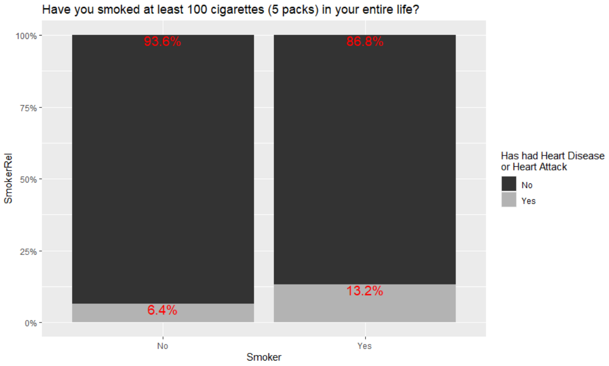


Fig. 48. Relative values of share of heart disease among participants who are considered a regular smoker

Of the total participants, 44.32% of participants had smoked at least a 100 cigarettes throughout their lifetime. Of these participants, 13.2% developed a coronary disease. Of the remaining 55.68%, 6.4% developed a coronary disease. This likely means that regular and/or intensive smoking may contribute to getting a heart disease or attack.

Stroke is a binary categorical value and shows the share of heart disease among participants who had a stroke.

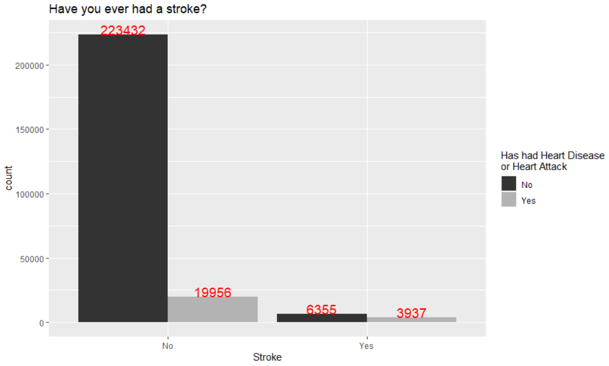


Fig. 49. Absolute values of share of heart disease among participants who had a stroke

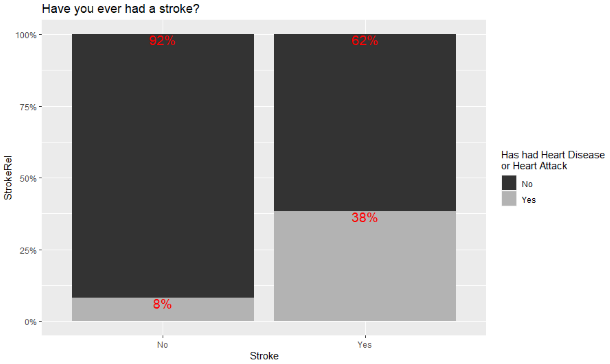


Fig. 50. Relative values of share of heart disease among participants who had a stroke

4.06% of participants reported ever experiencing a stroke, of which 38.0% developed coronary diseases. For the 95.94% of participants who never have had a stroke, only 8% developed a coronary disease. This means that their is a definite relationship between strokes and coronary diseases.

$Veggies$ is a binary categorical value that shows the amount of participants that (don't) consume a minimum of 100 grams worth of vegetables daily.

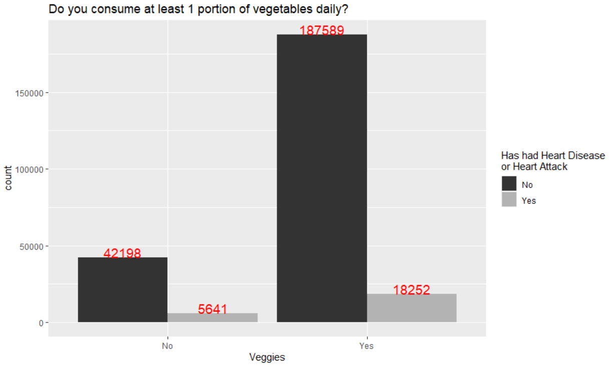


Fig. 51. Absolute values of share of heart disease among participants consuming vegetables on a daily basis

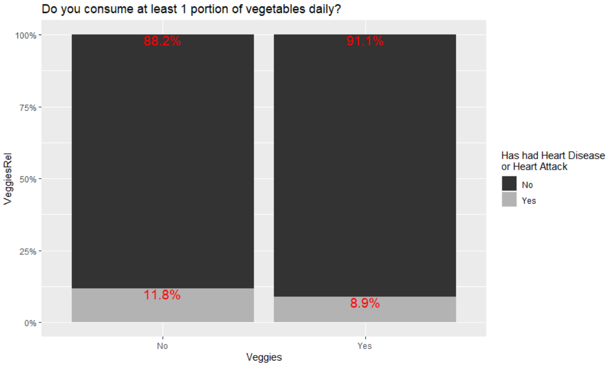


Fig. 52. Relative values of share of heart disease among participants consuming vegetables on a daily basis

In terms of distribution, 81.14% of people consume at least one portion of vegetables on a daily basis where 8.9% have developed a coronary disease. Of the remaining 18.86% of people who do not consume vegetables on a daily basis, 11.8% have developed a coronary disease. Both the relative and the absolute value graphs show a slightly higher percentage of people with coronary diseases for participants who do not consume vegetables on a daily basis.

A. *Multivariable Correspondence Analysis*

To get a general overview of the distribution and correlation of the different variables, we created three different Multivariable Correspondence plots showing how close the different variables and their categories are in comparison to the row values. Normally, a Principle Component Analysis would be performed, but because this dataset only consists of categorical values, an actual PCA is not possible. That is why a Multivariable Correspondence analysis is performed.

This part of the analysis is performed on the dataset after the amount of participants without cardiovascular diseases has been reduced. Because the ratios for this group has remained the same, the results of this analysis are also the same as with the complete dataset. This analysis has been performed after this reduction in size to make it easier to get an overview of the distribution of the data.

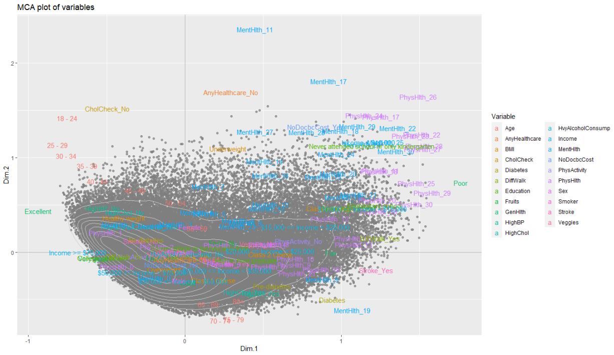


Fig. 53. Total MCA plot of all participants

Figure 53 shows the plot for the complete dataset, including both participants with and without cardiovascular diseases. The black dots indicate the individual participants and the distance between the different categories per column. Although this plot does not give a clear indication of which factors are the most common, it does show a relative dense structure of dots. This indicates that many of the individual participants are at least somewhat similar concerning the values for the different categories.

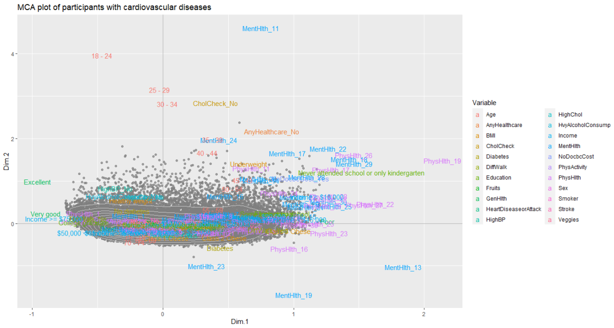


Fig. 54. MCA plot of all participants with cardiovascular diseases

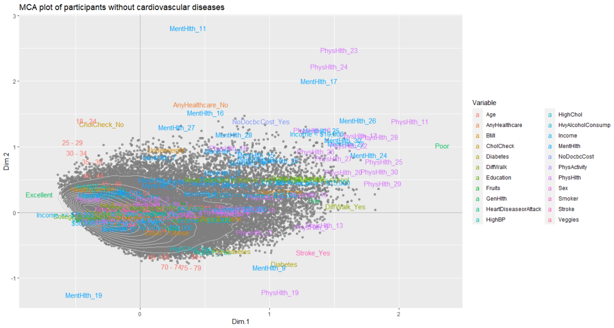


Fig. 55. MCA plot of all participants without cardiovascular diseases

Figure 54 and figure 55 show the difference in correlation between participants with and without cardiovascular diseases respectively. The plot showing participants with cardiovascular diseases seems to be slightly more concentrated with a few bigger outliers. The plot showing participants without cardiovascular diseases shows a more spread out graph. Also visible from these graphs is that figure 54 shows that more participants are in the older age categories whereas figure 54 shows that more participants are in the younger age category.

VII. Appendix 2: Code

For this study, R was used to create the models and graphs for the data analysis and results. In the following paragraphs, the steps taken for the preprocessing as well as the creation of the models are explained.

A. Data Preprocessing

To be able to use the data set for our research, a few modifications have been made. All variables have been converted to factor to make it easier for us to work with. Certain continuous variables like BMI and age have been transformed into groups. Below is an example of the conversion of BMI into groups.

```
BMI.to.categorical <- function(BMI.value){
    if (BMI.value >= 40) {
        BMI.value <- "Class␣3␣Obese"
    } else if (BMI.value >= 35) {
        BMI.value <- "Class␣2␣Obese"
    } else if (BMI.value >= 30) {
        BMI.value <- "Class␣1␣Obese"
    } else if (BMI.value >= 25) {
        BMI.value <- "Overweight"
    } else if (BMI.value >= 18.5) {
        BMI.value <- "Healthy␣weight"
    } else {
        BMI.value <- "Underweight"
    }
}

brfss.df$BMI <- sapply(brfss.df$BMI, BMI.to.categorical)
```

Fig. 56. Code for converting numeric BMI values to BMI categories

Next, most of the inserted values were re-coded to make it easier to read. For the most part, this means changin ”1” and ”0” to ”yes” and ”no” respectively. Next, to handle the class imbalance of the dataset, the data was split up into participants with a heart disease and without. The group without was split up again into a smaller portion of 20%. The ratios of attribute values of this group has remained the same during this process to ensure that the results are impacted as least as possible. The groups were then put together and randomized so they could be split up into train-, validation-, and testsets in the ratio 60, 20, 20 respectively.

```
amount.without.heartDiseaseorAttack.removed <- 0.8
brfss.df.with.heartDiseaseorAttack <- brfss.df %>%
        filter(HeartDiseaseorAttack == 'Yes')
brfss.df.without.heartDiseaseorAttack <- brfss.df %>%
        filter(HeartDiseaseorAttack == 'No')

sample <- sample.int(n = nrow(brfss.df.without.heartDiseaseorAttack),
    size = floor(amount.without.heartDiseaseorAttack.removed *
    nrow(brfss.df.without.heartDiseaseorAttack)),
    replace = F)

brfss.df.without.heartDiseaseorAttack.filtered <-
    brfss.df.without.heartDiseaseorAttack[-sample, ]

print(nrow(brfss.df.without.heartDiseaseorAttack.filtered))

# Validate that the ratios of all of the attributes have remained the same
for(i in 1:22){
    print(colnames(brfss.df.without.heartDiseaseorAttack)[i])

    print(cbind(freq = table(
        brfss.df.without.heartDiseaseorAttack[, i]),
        percentage = prop.table(table(
            brfss.df.without.heartDiseaseorAttack[, i])
        ) * 100))

    print(cbind(freq = table(
        brfss.df.without.heartDiseaseorAttack.filtered[, i]),
        percentage = prop.table(table(
            brfss.df.without.heartDiseaseorAttack.filtered[, i])
        ) * 100))
}
```

Fig. 57. Code for splitting the data set into a train, validation, and test data set

The distributions of the train-, validation- and test sets were verified to be the same to ensure that results would not be influenced by different distributions.

```
# Original data
cbind(freq = table(brfss.df$HeartDiseaseorAttack),
    percentage = prop.table(
        table(brfss.df$HeartDiseaseorAttack)) * 100)
#freq percentage
#No   45958    65.79433
#Yes 23893    34.2056

#Train
cbind(freq = table(train$HeartDiseaseorAttack),
  percentage = prop.table(
    table(train$HeartDiseaseorAttack)) * 100)
#freq percentage
#No   27595    65.84347
#Yes 14315    34.15653

#Validation
cbind(freq = table(validation$HeartDiseaseorAttack),
  percentage = prop.table(
    table(validation$HeartDiseaseorAttack)) * 100)
#freq percentage
#No   9202    65.86972
#Yes 4768    34.13028

#Test
cbind(freq = table(test$HeartDiseaseorAttack),
  percentage = prop.table(table(
    test$HeartDiseaseorAttack)) * 100)
#freq percentage
#No   9161    65.57154
#Yes 4810    34.42846
```

Fig. 58. Code for validating that the data was split in a stratified fashion

B. Models

Most of the settings for the different models have been discussed in the methods section of this paper. Below, two examples are given of the code used to create the models. The first code block shows the code for creating the decision tree using gradient boosting. The second code block shows the code for creating a logistic regression model.

```
tc <- trainControl(
    method = "boot",
    number = 20,
    classProbs = T,
    savePredictions = T
)

gbmGrid <-  expand.grid(
    interaction.depth = c(1, 2, 5, 8),
    n.trees = seq(0, 60, by = 5),
    shrinkage = 0.1,
    n.minobsinnode = 1
)

brfss.df.tree.tuned = train(HeartDiseaseorAttack ~ .,
                          data=train,
                          method="gbm",
                          trControl = tc,
                          tuneGrid = gbmGrid)

control <- trainControl(method="repeatedcv",
                          number=10,
                          repeats=5)

brfss.df.lm <- train(HeartDiseaseorAttack ~ .,
                          data=train,
                          method="glm",
                          trControl=control)
```

Fig. 59. (Part of the) code used for training the machine learning models

C. Results

We used every model to predict an individual's heart attack and heart disease risk, both as a "yes"/"no" value and a relative chance. The validation set was used to tune the model to improve results. The test set was used as the final set to test if the models were not over-fitted on the already used data sets. For each model, a confusion matrix was generated, as well as a list of features judged by importance and ranked as such. We also generated ROC curves for each model. In the code block below, a sample is given of code retrieving the results of a model's predictions.

```
  #Set type to "prob" for logistic regression probabilities
  brfss.df.lm.pred <- predict(brfss.df.lm,
                              newdata=validation,
                              type="raw")

  confusionMatrix(brfss.df.lm.pred,
                factor(validation$HeartDiseaseorAttack))
  glmImp <- varImp(brfss.df.lm, scale=T)

  brfss.df.lm.ROC <-
      data.frame(predict(brfss.df.lm, newdata=validation, type="prob"))
  brfss.df.lm.ROC$obs <-
      as.factor(validation$HeartDiseaseorAttack)
  brfss.df.lm.ROC$Group <- "brfss.df.lm"

  ROC <- rbind(brfss.df.lm.ROC)
  ROC.plot <- evalm(ROC,
      title="ROC␣curve␣of␣a␣logistic␣regression␣model\n
␣␣␣␣␣␣␣␣with␣access␣to␣all␣features␣trained␣using␣10-fold␣cross-validation")
```

Fig. 60. (Part of the) code for retrieving results from a trained model

VIII. Appendix 3: Note from the authors

This study was performed as a project for the course Data Mining on the Radboud University Nijmegen. For this project, parts of the abstract, introduction, and methods section have been reused from another research paper we have written for a different course: Academic Research and Writing. This research paper concerns the same subject of this paper, except that the research is focused on a literature study instead. The parts that were reused pertain to the information given about our data set and our methodology concerning the gathering of the data. Another part that was reused was the literature study, including all the sources, as the results and information of this study was of the same importance for both research papers.

List of Figures

LIST OF TABLES