Predicting Cardiovascular Disease

Martin O'Sullivan

25/02/2020

Contents

1 Introduction and Summary

1.1 Description of the Project

This report describes a project carried out as part of the assessment for the HarvardX Data Capstone module of the Professional Certificate in Data Science. The specified task was to apply machine learning techniques to solve a problem (chosen by the candidate) using a publicly available dataset. The problem which the project addressed was how to predict cardiovascular disease (CVD) in people using health, lifestyle, socioeconomic and demographic information obtained from a publicly available dataset: the Behavioural Risk Factor Surveillance System (BRFSS) which is the USA's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services[1].

CVDs are the number one cause of death globally. An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. CVD is the name for the group of disorders of heart and blood vessels, and includes: coronary heart disease (heart attack), cerebrovascular disease (stroke) and other conditions such as heart failure and rheumatic heart disease. Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use

[1] https://www.cdc.gov/brfss/index.html

of alcohol using population-wide strategies[2]. However, people who are at high cardiovascular risk need early detection so that they can receive the necessary advice and preventive medicines if needed. Hence the need for methods for predicting which people are at risk. This project develops machine learning algorithms which seek to predict whether a person with certain attributes and risk behaviours will suffer from cardiovascular disease.

1.2 Description of the Dataset

The dataset used in this project is derived from the Behavioural Risk Factor Surveillance System (BRFSS) survey carried out in 2015. BRFSS, coordinated by Centers for Disease Control (CDC), is a collection of state health surveys conducted by the 50 U.S. states, the District of Columbia and three U.S territories. Over 400,000 non-institutionalised U.S. adults are interviewed each year about their lifestyle behaviors and preventive health practices including physical activity, diet, and health conditions. Survey respondents are contacted by landline telephone or cellular (mobile) phone.

Disproportionate stratified sampling (DSS) is used for the landline sample. DSS sampling of telephone numbers is more efficient than simple random sampling. Households are selected randomly and likewise the household member to be interviewed is also chosen at random. Cellular telephone respondents are randomly selected with each having equal probability of selection.

Social surveys in general, including the BRFSS study, are prone to various sources of bias, for example:

- non-response bias where participants and non-participants may differ in important ways;
- social desirability bias where respondents may exaggerate desirable behaviours or attributes while under-reporting less desirable ones.

Data weighting strategies deployed by CDC attempt to remove bias in the sample, using a dual weighting process: design weighting and iterative proportional fitting (“raking”). Moreover, BRFSS’s weighting protocols have the aim of ensuring that the survey data are representative of the population as a whole.[3] The survey is conducted every year and consists of core questions, which are asked by all states, and optional questions which states may or may not use, depending on their needs. Moreover, the core questions may be rotated between odd and even years. For this project, data from the 2015 survey have been selected as it contains some relevant questions which were not asked in subsequent years. The dataset was downloaded from the Kaggle website[4] and is also available on the CDC website[5].

The data required for the project were selected and saved as a .RDS file brfss.RDS which is available on the github repository for this project: https://github.com/mcos2405/brfssproj

1.3 Goal of the Project

The main goal of the project was to develop a machine learning classification algorithm which would take:

- the presence or absence of CVD as the outcome/response/target variable, and
- a set of predictor variables/features in the BRFSS dataset relating to health, demographic and socioeconomic risk factors.

An additional goal of the project was to compare the classification model with Cardiac Risk Scores currently used in medical practice with regard to the risk factors used and the predictive performance. CVD risk scores allow clinicians to integrate information from multiple CVD risk factors and quantitatively estimate a person’s absolute risk for, or likelihood of experiencing, a CVD event during a defined period of time[6].

[2] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[3] https://www.cdc.gov/brfss/annual_data/2015/pdf/overview_2015.pdf

[4] https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system#2015.csv

[5] https://www.cdc.gov/brfss/annual_data/annual_2015.html

[6] KarmaliKN et al, Risk scoring for the primary prevention of cardiovascular disease; Cochrane Database of Systematic Reviews 2017, Issue 3

1.4 Key Steps

In summary, the project consisted of the following key steps:

- the R packages required for the project were installed and relevant libraries were loaded;
- the dataset was downloaded from the Kaggle website as a .csv file, saved as a .RDS file and partitioned into a training (brfss1) set and a testing (validation) set;
- based on literature search an initial selection of potential risk factors was made;
- exploratory analysis and feature selection was carried out;
- after splitting the training (brfss1) set into training and testing sets, a number of possible models were investigated and compared; ultimately a Quadratic Discriminant Analysis model was derived and tested on the Validation set.

2 Methods and Analysis

2.1 Installing Packages

The R packages required for the project were installed and the relevant libraries were loaded.

```r
#Installing required packages and loading libraries

if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project.org")
if(!require(kableExtra)) install.packages("kableExtra", repos = "http://cran.us.r-project.org")
if(!require(tidyr)) install.packages("tidyr", repos = "http://cran.us.r-project.org")
if(!require(stringr)) install.packages("stringr",repos = "http://cran.us.r-project.org")
if(!require(forcats)) install.packages("forcats", repos = "http://cran.us.r-project.org")
if(!require(ggplot2)) install.packages("ggplot2", repos = "http://cran.us.r-project.org")
if(!require(plyr)) install.packages("plyr", repos = "http://cran.us.r-project.org")
library(knitr)
library(dplyr)
library(tidyverse)
library(kableExtra)
library(tidyr)
library(stringr)
library(forcats)
library(ggplot2)
library(data.table)
library(plyr)
```

2.2 Importing and Setting Up the Data

```r
brfss <- readRDS("./brfss.RDS")
```

2.3 Overview of the Dataset

The *brfss* dataset as contains data selected from the 2015 survey. There are over 441,000 rows and 19 columns. Each row represents a person who was interviewed, while each column represents the responses to a given question that was asked by the interviewer. Some columns (calculated variables) are derived from the answers in other columns.

```
dim(brfss)
```

```
## [1] 441456     19
```

All the data in the set are numeric and can be continuous or categorical. Details of the coding of the variables and the value labels are available from the 2015 Codebook[7] which also contains the codes for “don’t know”, “refused” and missing values. The Codebook also contains details of the response rates for each question. For most questions the non-response rate was 1% or less. (Notable exceptions were the questions on income where around 18% of respondents said they didn’t know, were not sure or refused to answer.)

2.4 Selection of Potential Predictors

The selection of the variables for inclusion in the machine learning model involves two stages:

- considering predictors based on domain knowledge and literature search, and
- a process known as *feature selection* where variables are further assessed on the basis of statistical criteria to determine if they are relevant and not redundant.

Risk factors for CVD have been categorised as modifiable (for example physical inactivity, smoking, high blood pressure, high cholesterol) and non-modifiable (including sociodemographic factors such as age, gender, ethnicity and socioeconomic status)[8]. Based on this categorization a preliminary set of 19 variables was selected. Variables relating to the income of respondents (which would normally be a useful indicator of socioeconomic status) were excluded due to the high non-response rate described above.

Details of these variables will be presented following some data pre-processing and cleaning tasks.

2.5 Data Preprocessing and Cleaning

Instances of the selected variables indicating “don’t know”, “refused” or missing values are removed.

```
#converting "don't know" and "refused" to NA
brfss[brfss == 7 | brfss == 9 | brfss == 77 | brfss == 99] <- NA
#removing all NA
brfss <- na.omit(brfss)
```

The variables indicating that a person has had coronary heart disease (CHD) or myocardial infarction (MI) or a stroke are merged to form *CVD* which will be the response/dependent variable in the classification model. The variables relating to heavy drinking and binge drinking have also been merged.

```
#merging cardiac disease and stroke variables
brfss<- brfss %>%
  mutate(CVD = ifelse(X_MICHD == 1 | CVDSTRK3 == 1, 1, 2))
#merging heavy-drinking and binge-drinking variables
brfss <- brfss %>%
  mutate(ALCPROB = ifelse(X_RFBING5 == 1 & X_RFDRHV5 == 1, 1, 2))
# removing variables which are no longer required
brfss <- subset(brfss,
                select = -c(X_MICHD, CVDSTRK3, X_RFBING5, X_RFDRHV5))
```

The transformed dataset resulting from the above operations *brfss* has been reduced to just under 300,000 rows and 17 columns.

```
dim(brfss)
```

```
## [1] 294618     17
```

[7] https://www.cdc.gov/brfss/annual_data/annual_2015.html

[8] https://www.world-heart-federation.org/resources/risk-factors/

The *brfss* dataset is now transformed further by replacing the numeric coding of the categorical variables with meaningful labels. For example in the SEX variable “1” and “2” are replaced with “Male” and “Female”.

```
#simplifying diabetes variable and allocating value labels
brfss <- brfss %>% mutate(DIABETE3 = ifelse(DIABETE3 == 1, "Diabetes", "No Diabetes"))
#allocating value labels to variables
brfss <- brfss %>% mutate(CVD = ifelse(CVD == 1, "CVD", "No CVD")) %>%
mutate(SEX = ifelse(SEX == 1, "Male", "Female")) %>%
mutate(X_AGE65YR = ifelse(X_AGE65YR == 1, "Age 18 to 64", "Age 65 or older")) %>%
mutate(X_RACEG21 = ifelse(X_RACEG21 == 1, "Non-Hispanic White", "Non-White or Hispanic")) %>%
mutate(X_EDUCAG = ifelse(X_EDUCAG == 1, "Not High School Graduate",
      ifelse(X_EDUCAG == 2, "Graduated High School",
          ifelse(X_EDUCAG == 3, "Attended College", "Graduated College")))) %>%
mutate(HLTHPLN1 = ifelse(HLTHPLN1 == 1, "Health Cover", "No Health Cover")) %>%
mutate(EXERANY2 = ifelse(EXERANY2 == 1, "Exercise", "No Exercise")) %>%
mutate(X_BMI5CAT = ifelse(X_BMI5CAT == 4, "Obese", "Not Obese")) %>%
mutate(X_SMOKER3 = ifelse(X_SMOKER3 == 1, "Smokes Every Day",
      ifelse(X_SMOKER3 == 2, "Smokes Some Days",
          ifelse(X_SMOKER3 == 3, "Former Smoker", "Never Smoked")))) %>%
mutate(ADDEPEV2 = ifelse(ADDEPEV2 == 1, "Depression", "No Depression")) %>%
mutate(X_VEGLT1 = ifelse(X_VEGLT1 == 1, "Vegetables", "No Vegetables")) %>%
mutate(X_FRTLT1 = ifelse(X_FRTLT1 == 1, "Fruit", "No Fruit")) %>%
mutate(X_RFHYPE5 = ifelse(X_RFHYPE5 == 1, "No Blood Pressure", "Blood Pressure")) %>%
mutate(TOLDHI2 = ifelse(TOLDHI2 == 1, "High Cholesterol", "Cholesterol OK")) %>%
mutate(ALCPROB = ifelse(ALCPROB == 1, "No Alcohol Problem", "Alcohol Problem")) %>%
mutate(CHCKIDNY = ifelse(CHCKIDNY == 1, "Kidney Disease", "No Kidney Disease"))
```

All the variables in the *brfss* dataset are converted to factors and the desired levels of variables with more than 2 categories is specified. (In general in R it is desirable that, for machine learning algorithms, the variables are factors.)

```
col_names <- names(brfss)
brfss[,col_names] <- lapply(brfss[,col_names] , factor)
#specifying the order of levels of variables with more than two levels
brfss <- brfss %>%
mutate(X_EDUCAG = ordered(X_EDUCAG,
  levels = c("Not High School Graduate", "Graduated High School",
             "Attended College", "Graduated College"))) %>%
mutate(X_SMOKER3 = ordered(X_SMOKER3,
    levels = c("Smokes Every Day", "Smokes Some Days",
               "Former Smoker", "Never Smoked")))
```

2.6 Creating Training and Validation Sets

The dataset is now split into a training set *brfss1* which will be used to develop the classification models and a testing set *validation* which will be retained and only used for testing the final model. (A training/test split of 80%/20% of the rows is used. For a large dataset like this, such a split is typical.)

```
# Validation set will be 20% of the brfss data
set.seed(1, sample.kind="Rounding")
test_index <- createDataPartition(y = brfss$CVD, times = 1, p = 0.2, list = FALSE)
brfss1 <- brfss[-test_index,]
validation <- brfss[test_index,]
```

The training set has almost 240,000 rows while the testing (validation) set has almost 60,000. Both sets have 17 columns.

```
dim(brfss1)
```

```
## [1] 235694     17
```

```
dim(validation)
```

```
## [1] 58924    17
```

2.7 Data Exploration and Visualization

The response variable and the 16 variables selected for further consideration as possible predictors are listed and described in the following table.

Variable Name	Description
SEX	Gender
X_AGE65YR	Age in two groups (under and over 65)
X_RACEG21	Ethnicity in two groups (White and Non-White)
X_EDUCAG	Education level in 4 groups
HLTHPLN1	Health Insurance or Other Health Cover?
EXERANY2	Exercise during past month?
X_BMI5CAT	Body Mass Index (BMI) in 2 groups
X_SMOKER3	Smoking status in 4 groups
DIABETE3	Ever told has Diabetes?
ADDEPEV2	Ever told has Depression?
X_VEGLT1	Consumes vegetables at least once a day?
X_FRTLT1	Consumes fruit at least once a day?
X_RFHYPE5	Ever told has high blood pressure?
TOLDHI2	Ever told has high cholesterol?
CVD	Ever told has cardiovascular disease?
ALCPROB	Heavy or Binge Drinker
CHCKIDNY	Ever told has Kidney Disease?

The *CVD* variable will be the response (outcome) while the remaining variables will be the potential predictors/features.

2.8 Exploring the Outcome Variable

Around 12% of survey respondents have been diagnosed with cardiovascular disease, while almost 88% have not. This indicates that the data are imbalanced, i.e the class proportions are skewed with "No CVD" the majority class and "CVD" the minority class. This imbalance will need to be taken into account when evaluating the results of the modeling phase.

```
brfss1$CVD %>%
   table() %>%
   prop.table() %>% {. * 100} %>%
   round(2)
```

```
## .
##    CVD No CVD
##  12.33  87.67
```

2.9 Feature Selection

Feature selection is a process for determining which predictors should be included in a model and is one of the most critical questions as data are becoming increasingly high-dimensional[9]. The process is primarily focused on removing irrelevant or redundant predictors from the model. Irrelevant features bring no useful information to the model while redundant features likewise are not useful either because they are correlated with other features or because they can be obtained by linear combination of other features. Retaining unnecessary features in the model can lead to reduced accuracy and increased training times. The aim is to include features which are strongly predictive of the target variable but are not related to other features. As the potential predictors here are all categorical pairwise Chi Square tests between all 136 combinations of the 17 variables are performed. As Chi Square tests can be unreliable if used with either large or small datasets a random sample of 1000 is taken from the training set and is used to carry out the tests.

```r
#taking a random sample of 1000 rows
set.seed(1, sample.kind="Rounding")
sample1 <- sample_n(brfss1, 1000)
```

```r
#pairwise Chi Square tests between all combinations of variables
combins <- combn(ncol(sample1),2)
adply(combins, 2, function(x) {
  test <- chisq.test(sample1[, x[1]], sample1[, x[2]])

  out <- data.frame("Row" = colnames(sample1)[x[1]]
                    , "Column" = colnames(sample1[x[2]])
                    , "Chi.Square" = round(test$statistic,3)
                    ,  "df"= test$parameter
                    ,  "p.value" = round(test$p.value, 3)
                    )
  return(out)
})
```

The results of the Chi Square tests between the target variable *CVD* and each feature are listed in the following table including the Chi Square statistics and the p-values.

FEATURE	TARGET	CHI SQUARE	df	p-value
SEX	CVD	6.738	1	0.009
X_AGE65YR	CVD	28.683	1	0.000
X_RACEG21	CVD	5.589	1	0.018
X_EDUCAG	CVD	14.018	3	0.003
HLTHPLN1	CVD	0.036	1	0.849
EXERANY2	CVD	20.261	1	0.000
X_BMI5CAT	CVD	0.691	1	0.406
X_SMOKER3	CVD	16.884	3	0.001
DIABETE3	CVD	11.149	1	0.001
ADDEPEV2	CVD	6.836	1	0.009
X_VEGLT1	CVD	0.007	1	0.933
X_FRTLT1	CVD	0.915	1	0.339
X_RFHYPE5	CVD	43.619	1	0.000
TOLDHI2	CVD	30.755	1	0.000
ALCPROB	CVD	4.989	1	0.026
CHCKIDNY	CVD	30.174	1	0.000

The higher values of the Chi Square statistic and low p-values indicate stronger relationships between the

[9] M. Kuhn and K. Johnson, Applied Predictive Modeling (p. 487), Springer New York 2013

variables. The criterion for the retention or discarding of features is set at a threshold Chi Square statistic value of 10. The variables relating to gender, depression, race, alcohol, BMI, health insurance, vegetable consumption and fruit consumption are removed from the analysis, leaving the target variable *CVD* and 8 features. Further inspection of the results of the pairwise tests indicate strong relationships between some features, for example between the Blood Pressure and Cholesterol variables. Education level and Physical Activity (Exercise) are also dependent on most of the other features. This is not desirable and suggests that some of these variables could be removed. However, in the context of this project, these are known to be important risk factors which are included in risk-scoring systems used by medical practitioners and these factors should be retained in the analysis for the present.

```
#removing predictor variables from both brfss1 and validation sets
brfss1 <- subset(brfss1,
          select = -c(X_RACEG21, HLTHPLN1, X_BMI5CAT, X_VEGLT1, X_FRTLT1, ALCPROB, ADDEPEV2, SEX))
validation <- subset(validation,
          select = -c(X_RACEG21, HLTHPLN1, X_BMI5CAT, X_VEGLT1, X_FRTLT1, ALCPROB, ADDEPEV2, SEX))
```

2.10 Modeling Approach

The modeling approach taken in this project is termed Supervised Classification. The aim of classification is to predict a target variable by building a classification model based on a training dataset, and then utilizing that model to predict the value of the target in unseen test data. This type of data processing is called supervised learning since the data processing phase is guided toward the target variable while building the model. After splitting the training set (*brfss1*) into a training set (*brfss_train*) and a testing set (*brfss_test*), three models are run on the training data using the caret package in R:

- a Logistic Regression model,
- a Linear Discriminant Analysis model, and
- a Quadratic Discriminant Analysis model.

The brfss1 (training) set is split, therefore, into training (80%) and test (20%) sets which will be used during the model development phase.

```
set.seed(1, sample.kind="Rounding")
train_index <- createDataPartition(y = brfss1$CVD, times = 1, p = 0.2,
                                   list = FALSE)
brfss_train<- brfss1[-train_index,]
brfss_test<- brfss1[train_index,]
```

These sets (brfss_train and brfss_test) have the following dimensions (rows and columns):

```
dim(brfss_train)
```

```
## [1] 188554      9
```

```
dim(brfss_test)
```

```
## [1] 47140     9
```

2.10.1 Logistic Regression Model

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. Mathematically, a binary logistic model has a dependent variable with two possible values (in this case "CVD" and "No CVD"). In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "CVD" is a linear combination of one or more independent variables ("predictors"); the independent variables can either be a categorical variable or a continuous variable (any real value)[10].

[10] https://en.wikipedia.org/wiki/Logistic_regression

```
train_glm <- train(CVD ~ ., method = "glm", data = brfss_train)
glm_preds <- predict(train_glm, brfss_test)
mean(glm_preds == brfss_test$CVD)
confusionMatrix(reference = brfss_test$CVD, data = glm_preds)
varImp(train_glm)
```

2.10.2 Linear Discriminant Analysis (LDA) Model

The Linear Discriminant Analysis method is credited to one of the best-known statisticians of the 20th century, R.A. Fisher, and dates from 1936[11]. This algorithm tries to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier. Linear Discriminant Analysis assumes the classes have equal covariance matrices but does not assume the data are normally distributed.

```
train_lda <- train(CVD ~ ., method = "lda", data = brfss_train)
lda_preds <- predict(train_lda, brfss_test)
mean(lda_preds == brfss_test$CVD)
confusionMatrix(reference = brfss_test$CVD, data = lda_preds)
varImp(train_lda)
```

2.10.3 Quadratic Discriminant Analysis (QDA) Model

Quadratic Discriminant Analysis (QDA) is a machine learning classification technique which is quite similar to Linear Discriminant Analysis (LDA). QDA, like LDA, uses Bayes Theorem to estimate the parameters of the underlying equation. In LDA, we assume that each class has a shared covariance matrix. However, in QDA, this assumption has changed and we now assume that each class has its own covariance matrix[12].

```
train_qda <- train(CVD ~ ., method = "qda", data = brfss_train)
qda_preds <- predict(train_qda, brfss_test)
mean(qda_preds == brfss_test$CVD)
confusionMatrix(reference = brfss_test$CVD, data = qda_preds)
varImp(train_qda)
```

2.11 Running the Final Model on the Validation Set

On the basis of Balanced Accuracy the Quadratic Discriminant Analysis has been chosen as the final model and is tested on the unseen data in the *validation* set.

```
val_qda <- train(CVD~., method = "qda", data = validation)
val_preds <- predict(val_qda, validation)
mean(val_preds == validation$CVD)
confusionMatrix(reference = validation$CVD, data = val_preds)
varImp(val_qda)
```

3 Results

The outcome of running the models including the testing on the validation dataset are presented in the following table:

	Accuracy	Sensitivity %	Specificity %	Balanced Accuracy
GLM	0.8769	6.8	99.1	0.5295

[11] Marsland, S: Machine Learning: An Algorithmic Perspective, 2nd Edition (Chapman & Hall Machine Learning & Pattern Recognition Series) 2014

[12] https://blog.quantinsti.com/quadratic-discriminant-analysis-optimize-intraday-momentum-strategy/

	Accuracy	Sensitivity %	Specificity %	Balanced Accuracy
LDA	0.8747	11.5	98.1	0.5482
QDA	0.8325	26.4	91.2	0.5584
QDA (Val)	0.8309	28.3	90.8	0.5957

The metrics used to assess the relative performance of the models were the Balanced Accuracy, Sensitivity and Specificity. As noted above the dataset was imbalanced. Therefore Balanced Accuracy is a more appropriate measure and can be defined as the average accuracy obtained on either class. It is calculated as follows:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right)$$

where TP = True Positive, TN = True Negative, P = Positive and N = Negative. Of the three models developed the QDA model had the highest Balanced Accuracy at 0.56 with Sensitivity and Specificity values of 26.4% and 91.2% respectively. Running the QDA model on the holdout validation data yielded a Balanced Accuracy of almost 0.6 and Sensitivity and Specificity of 28.3% and 90.8%. In a medical context, test sensitivity is the ability of a test (or prediction model) to correctly identify those with the disease (true positive rate), whereas test specificity is the ability of the test to correctly identify those without the disease (true negative rate). Sensitivity and Specificity frequently exist in a state of balance. Increased sensitivity – the ability to correctly identify people who have the disease — usually comes at the expense of reduced specificity (meaning more false-positives). Likewise, high specificity — when a test does a good job of ruling out people who don't have the disease – usually means that the test has lower sensitivity (more false-negatives)[13]. In the case of predicting cardiac disease there needs to be a balance between identifying people who are really at high risk of disease, while minimising the number of people wrongly identified as at risk and subjecting them to invasive tests and unnecessary medications.

The Variable Importance output of all models suggested that Age, Blood Pressure and Cholesterol are the most important features in this dataset, while Education level and concurrent Diabetes and Kidney Disease also have a high importance.

4 Conclusion

Data from the Behavioural Risk Factor Surveillance System (BRFSS) Survey carried out in 2015 were studied to develop a machine learning model to identify important risk factors and to predict Cardiovascular Disease (CVD) in the survey respondents. A subset of the data (80%) was used to investigate the data and to set up the possible models while an unseen subset (20%) was used to test the final model which used Quadratic Discriminant Analysis. A balanced accuracy of 0.6, sensitivity 28.3% and specificity 90.8% were recorded. It is noted that the risk factors highlighted as important in this study, and derived from a large telephone survey are the same as those used, for example, in the Framingham Risk Scores which were first developed in the USA and have been very influential both in the USA and around the world in identifying people at high risk of CVD so that preventive measures can be taken[14]. The results from this project are not directly comparable with risk scores used by doctors which involve clinical interaction with the patients including for example, blood pressure measurement and blood samples for cholesterol. The BRFSS surveys gather invaluable data every year which is used nationally and by individual states to measure the prevalence of chronic diseases and risk factors, to track changes in behavior and to measure progress towards achieving public health objectives.

[13] https://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/understanding-medical-tests-sensitivity-specificity-and-positive-predi

[14] D'Agostino, RB et al, General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Study; *Circulation* 2008; 117; 743-753