Cardiovascular Disease Prediction

Biljana Novkovic

3/19/2023

INTRODUCTION

The term cardiovascular disease (CVD) refers to a range of diseases which affect the heart and blood vessels. They include heart attack (coronary heart disease), high blood pressure (hypertension), heart failure, and other heart diseases (Lopez et al. 2022).

CVD is the leading cause of death globally, and causes 1 of 4 deaths in the United States. CVD is also the most costly disease in the world with a calculated indirect cost of $237 billion dollars per year and a projected increase to $368 billion by 2035 (Lopez et al. 2022; Rana et al. 2021).

For decades, a lot of research and effort has been invested in trying to predict people's risk of heart disease. The most famous algorithm, called the Framingham Risk Score, has been developed based on the data obtained from the Framingham Heart Study, to estimate the 10-year risk of developing coronary heart disease (Wilson et al. 1998).

The Framingham Heart Study has been dubbed the most influential investigation in the history of modern medicine. It began in 1948 with around 5000 adults from Framingham, Massachusetts, and is now on its third generation of participants. A lot of what we know about heart disease is derived from this study, including many of the risk factors for heart disease (Hajar 2016).

The first Framingham Risk Score included age, sex, bad cholesterol (LDL cholesterol), good cholesterol (HDL cholesterol), blood pressure, blood pressure treatment, diabetes, and smoking. Other factors we know increase the risk of heart disease include unhealthy diets, physical inactivity, obesity, stress and family history/genetics (Lopez et al. 2022; Wilson et al. 1998).

Algorithms that help predict heart disease are important for heart disease prevention. They help both individuals and their doctors decide on lifestyle modification and preventive medical treatment. In this project we will analyze a large dataset to try and predict the development of cardiovascular disease.

ANALYSES

The CVD Dataset

The Cardiovascular disease dataset was obtained from kaggle (https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/code). It was chosen because of its size. Larger studies, in theory, provide stronger and more reliable results. They have smaller margins of error and lower standards of deviation. They also allow the researchers to decrease the risk of false-negative or false-positive findings, which is especially important in medicine.

The cardiovascular disease dataset contains 70,000 individuals, and 12 variables, one of which is the outcome of having/not having cardiovascular disease.

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	18393	2	168	62	110	80	1	1	0	0	1	0
1	20228	1	156	85	140	90	3	1	0	0	1	1
2	18857	1	165	64	130	70	3	1	0	0	0	1
3	17623	2	169	82	150	100	1	1	0	0	1	1
4	17474	1	156	56	100	60	1	1	0	0	0	0
8	21914	1	151	67	120	80	2	2	0	0	0	0

x
70000

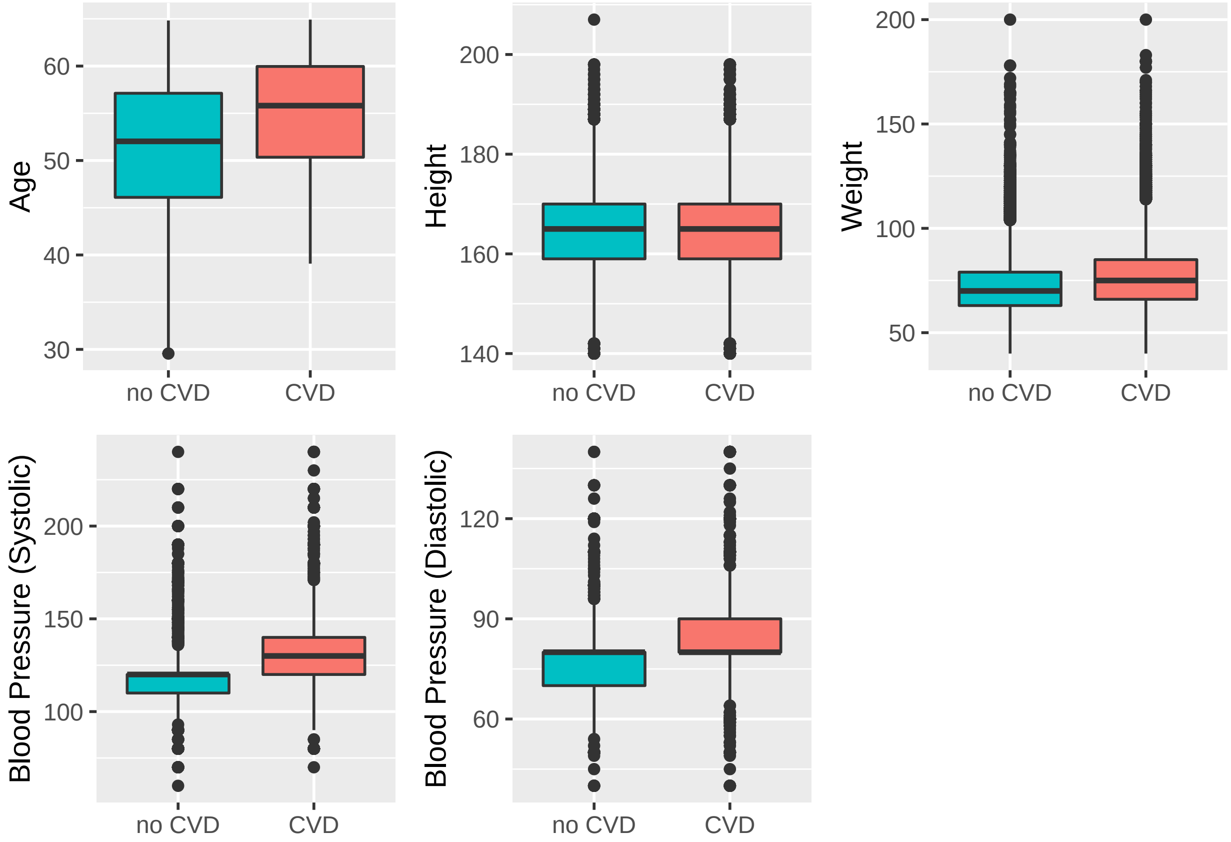
First, we will transform the variable “age” from days into years, for clarity. Second, upon inspection, our dataset includes some weird values that are implausible, such as negative blood pressure values. We will remove rows containing implausible values for variables height, weight, and systolic and diastolic blood pressure.

x
68555

Once the dataset has been cleaned, we are left with 68555 rows.

Exploring the Dataset

First, let's look at our numerical variables: age, height, weight, systolic and diastolic blood pressure.



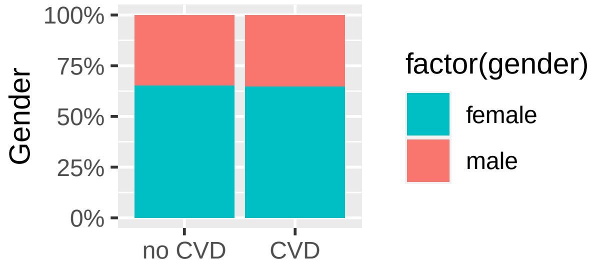
Boxplots suggest that people with CVD tend to be older, weigh more, and have higher systolic and diastolic blood pressure. A z-test confirms that all the variables are significantly different between non-CVD and CVD individuals.

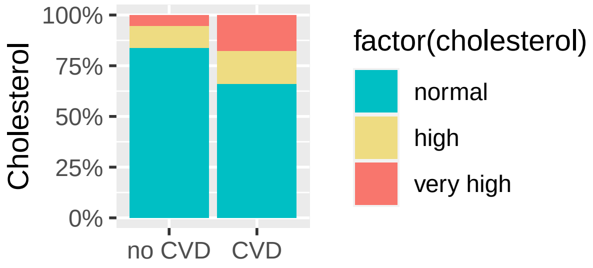
variable	p
Age	0.0000000
Height	0.0006057
Weight	0.0000000
Systolic Blood Pressure	0.0000000
Diastolic Blood Pressure	0.0000000

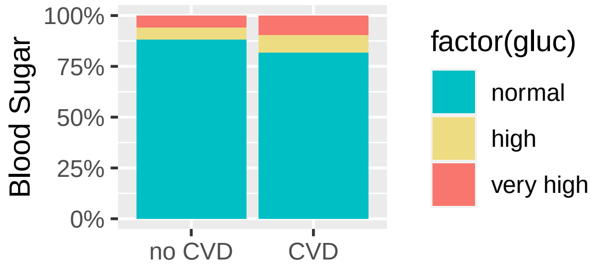
Next let's explore our categorical variables. For the purpose of exploration, let's turn the categorical variables into factors that we are familiar with: gender -> male, female; cholesterol -> normal, high, very high; glucose -> normal, high, very high; smoking -> non-smoker, smoker; alcohol -> non-drinker, drinker; and activity -> active, inactive.

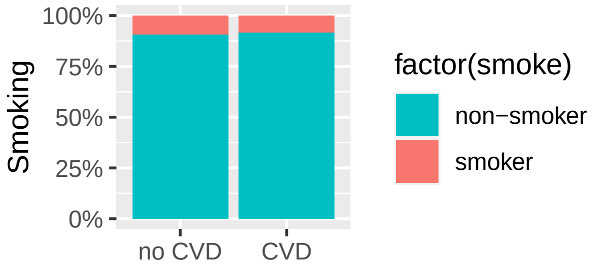
gender	cholesterol	gluc	smoke	alco	active	cardio
male	normal	normal	non-smoker	non-drinker	active	no CVD
female	very high	normal	non-smoker	non-drinker	active	CVD
female	very high	normal	non-smoker	non-drinker	inactive	CVD
male	normal	normal	non-smoker	non-drinker	active	CVD
female	normal	normal	non-smoker	non-drinker	inactive	no CVD
female	high	high	non-smoker	non-drinker	inactive	no CVD

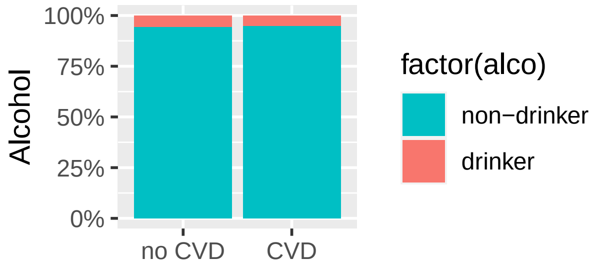
Now, let's generate some plots.

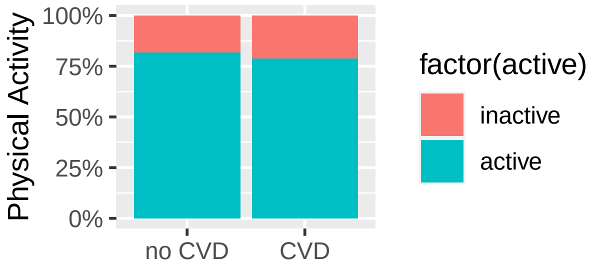










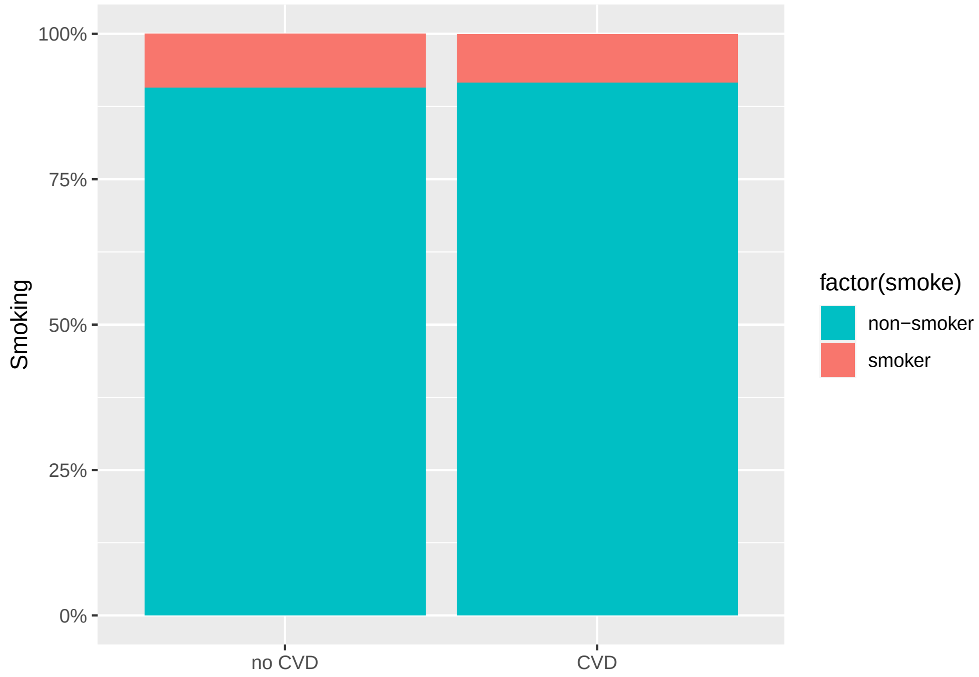


We can see that among people with CVD, there are slightly more males, there are more people with high and very high cholesterol, and high and very high blood sugar. There are also more inactive people among those with CVD. Surprisingly, there doesn't appear to be much difference between CVD cases and non-CVD, when it comes to drinking alcohol and smoking.

variable	p
Gender	0.0634007
Cholesterol	0.0000000
Blood Sugar	0.0000000
Smoking	0.0000204
Alcohol	0.0294288
Physical Activity	0.0000000

When we run a chi_squared test to see if there is a relationship between our variables and CVD state, we can see that age and alcohol are non-significant ($p > 0.01$), while other variables are significantly related to CVD.

Before we move on, let's briefly look at smoking.



Although we know that smoking is a risk factor for heart disease, in this dataset, we have more smokers among people without heart disease. This may be due to confounding. To prevent this variable from biasing our prediction, which would make our model perform poorly on real world data, we will remove this variable from the dataset before we proceed building models.

RESULTS

First, we will partition the dataset into a train (80%) and test set (20%). The size of our dataset allows us to do this.

Next, we will build a simple generalized linear model (GLM) for this dataset. GLMs allow for response variables that have error distribution models other than a normal distribution. This is useful to us because we can include our categorical variables into the prediction.

method	Accuracy
GLM	0.7294341

Let's try a different model to see if we can improve on the GLM. We will build a simple k-nearest neighbors model (k-NN) using our dataset.

method	Accuracy
GLM	0.7294341
kNN	0.6897608

Our k-NN model performs worse than our GLM model. It's known that the accuracy of the k-NN algorithm can be degraded when there are irrelevant features. Let's rerun the algorithm, but this time removing the gender and alcohol variables, which were not significantly related to the CVD outcome, and the height variable which was the least significant of the numerical variables.

method	Accuracy
GLM	0.7294341
kNN	0.6897608
kNN_2.0	0.6961785

The accuracy does improve, but is still not comparable to the GLM. Let's try and generate a random forest model. Random forest lends itself well to datasets with mixed numerical and categorical variables, such as ours.

method	Accuracy
GLM	0.7294341
kNN	0.6897608
kNN_2.0	0.6961785
RF	0.7310385

Random forest slightly outperforms GLM and it is our best model.

CONCLUSION

We have explored the cardiovascular disease (CVD) dataset and its variables and have built three different models to try and predict CVD using this dataset: GLM, k-NN and random forest. Random forest was the best performing model, with GLM as the close second. In the future, better accuracy can likely be achieved by (1) tuning the parameters of these models, which the author has attempted but couldn't finalize successfully due to the code timing out in R, likely because of the size of the dataset; (2) by building ensemble models.

Reference

Hajar R. (2016) Framingham Contribution to Cardiovascular Disease. Heart Views. Apr-Jun; 17(2):78-81. doi: 10.4103/1995-705X.185130

Lopez OE, Ballard BD, Jan A. Cardiovascular Disease. (2022) In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK535419/

Rana JS, Khan SS, Lloyd-Jones DM, et al. (2021) Changes in Mortality in Top 10 Causes of Death from 2011 to 2018. J Gen Intern Med. Aug; 36(8):2517-2518. doi: 10.1007/s11606-020-06070-z

Wilson PW, D'Agostino RB, Levy D, et al. (1998) Prediction of coronary heart disease using risk factor categories. Circulation. May 12;97(18):1837-47. doi: 10.1161/01.cir.97.18.1837