Data Mining and Knowledge Discovery
Personal Key Indicators of Heart Disease

Felipe Cortes Jaramillo
email felipe.cortes.jaramillo@etu.univ-st-etienne.fr

April 2023

1 Introduction

According to the statistics from the American Heart Association, around 19.1 million deaths in 2020 were attributed to cardio vascular diseases. 244.1 million were living with Ischemic heart disease, which is a condition in the arteries that prevent the flow of oxygen-rich blood cells [2], which lead to high heart attack risk in the future. These are just some consequences of heart disease nowadays, but there are plenty misunderstandings in what exactly is this problem. Following the definition of Mayo Foundation for Medical Education and Research, refers to all kinds of conditions that affects the flow of the blood in the cardiovascular system, increasing the risk of strokes [6]. Therefore, it is necessary to recognize and predict the conditions or behavior that lead to this disease, in order to prevent it and reduce the amount of people that suffers from it daily. In this project we study a recompilation of data about people that have suffered heart disease, as well, as applying different techniques like PCA, data mining apriori algorithm, Random Forest, K-Nearest Neighbors, Logistic Regression and XGBoost to extract meaningful patterns and make important insights about the possible recognition and prevention of this disease. Likewise, perform metrics on imbalanced set of data, to create a model able to recognize a high risk patient that may suffer this condition.

2 Problem Understanding

The use of statistical methods and machine learning algorithms in the medical field has increased over the years, supporting the insights that the doctors made from a patient and give them the best treatment possible. Heart disease in particular, presents a higher problem because approximately half of all heart attacks and strokes occur to patients that have been categorized as low risk for heart disease [14]. Therefore, there is an urgent demand to create a way to detect accurately and fast if a person could suffer from heart disease or not, using different elements from their clinic history or previous medical conditions like: cancers, strokes, body mass index, physical activity, etc. So, the aim of this study is to identify those key factors that determine if a patient will have the conditions to acquire it. Highlighting which ones are the most important that lead to this pathology. Moreover, being able to use a predictor model in the medical field to avoid the complications and possible death of patients that have not been warned earlier.

3 Data Understanding

As the main goal is identify the different key indicators that may be related to the risk of heart disease, we used the data created by the center for disease control and prevention (CDC). This entity performs a series of annual surveys to recollect data on the health status of all 50 states in the U.S.[11]. It performs annually more than 400.000 adult interviews, asking important

health status questions. As well as, life habits that have the citizens creating a complete database overall. The original dataset had nearly 300 columns but it was filtered to 18 variables based on the questions that directly or indirectly influenced heart disease. It is currently available at Kaggle platform as: Personal Key Indicators for Heart Disease Database, which has in total 319795 rows and 18 columns and it is stored in a comma separated values file format with a total size of: 24,6 KB.

As mentioned above, we have in total 18 variables. Which means that among those we have our target variable (Heart Disease), that can have only values "yes" or "no". Also, giving us some insights that we need to perform a binary classification in this dataset. On the other hand, we have different kinds of variables that can be represented in two groups:

- **Numeric Variables:** We have BMI (Body mass index relationship), Physical Health (Any physical injury in last 30 days), Mental Health (Any episode of metal illness in the past 30 days), Sleep Time (Amount of hours of sleep).
- **Categorical Variables:** The other group containing variables that can only have yes or no as answer, like: Smoking, Alcohol Drinking, Stroke, Diff Walking (Difficulties when walking or climbing stairs), Diabetic, Physical Activity, Asthma, Sex (Not yes or no, but binary choice), Kidney Diseases and Skin Cancer. As well as, categorical groups with more than two options: Age category, Race (white, black, Alaskan native, etc.), Gen Health (General consideration of the health according to the patient, from: very good to poor).

3.1 Exploratory Data Analysis

Performing an initial data exploration (EDA), we could extract some meaningful information about the composition of the dataset. As well, as distribution of the variables. First, we encountered that we had an imbalanced case as seen in figure 1. This means that we have more rows that lead to a patient without heart disease than one that has it, with a difference among the two proportions of (**91% of non positive patients to 9% of positive patients with heart disease**). This scenario is commonly presented in the medical field.

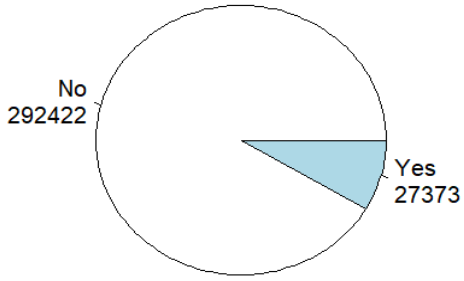

Figure 1: Binary distribution of target variable

As our main interest is identifying the key factors that are associated to having heart disease, we explore the distributions of variables among the patients that suffered from this disease (figure 2)[1]. Seeing some interesting elements, like that among the positive patients they are mostly older with two peaks above 70 years old, which explains why most of the medical results shows that above 65 years, the chances to have heart disease increases [1]. The majority of them are white males that smoked and do not consume daily alcohol. An interesting points comes from the distribution of the body mass index and sleep time, observing that a bast amount of patients who suffered heart disease had normal BMI but they usually slept few hours. Which can make some insights that sleeping a good amount of hours could reduce the risk of this pathology. All of this preliminary results are going to be explored in detail when analysing the correlation among the variables and modelling section.

4 Data Preparation

The previous section led us know that among the 17 independent variables, we have just 4 numerical ones. Therefore, it will be necessary to transform the

[1]All images of the report can be found at: Git Repository

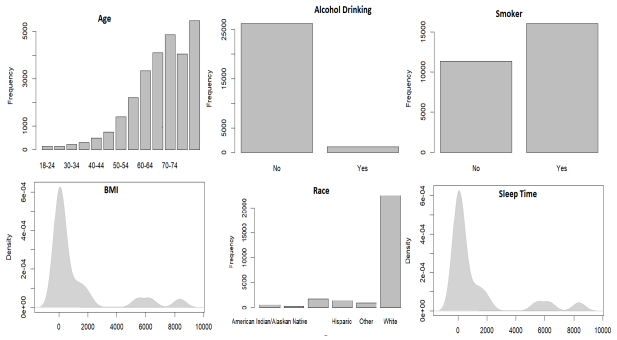

Figure 2: Variable distributions from diseased people

categorical variables into numbers for training the models and exploring the correlation among them. It was applied one-hot encoding [12] because most of the data was binary (yes and no the only possible options), so this kind of encoding is really useful to transform those values to number from 1 and 0 respectively. Moreover, when facing with variables that have more than two options, the encoding creates new variables in order to represent those special categories and uses a binary selection to indicate the value that each row has in this variable. Taking an example of age category, we will have 13 new variables with an interval of 4 years, like: 25 to 29 years, 30 to 34 years, etc. If that specific observation is an individual with 26 years, the value in the column from 25 to 29 years will be 1 and the others 0. This encoding has the trade-off of increasing the dimensionality of the data which most of the times can cause problems with the performance of some models. This technique took in total **1.81 seconds**.

Additionally, we use another technique to encode the categorical data that used more than two classes, using identifiers to represent the different kinds of groups. In the previous example, the age groups will be then labeled as 1 from the lowest to 13 the latest. In order to test another alternative that kept the dimensions of the original dataset but using the benefits of the one-hot encoding. Also, the time of this method took **2.87** seconds in total.

Finally, we ended the preparation applying a normalization of the data. In other words, we centered each value by the mean and scaled up to the standard deviations presented in the dataset. We applied this technique because this group our data, so when modelling it is much easier to generalized over a condensed set of data than in one with further extreme values. This process took in total **0.26 seconds**.

4.1 Correlations

One of the great questions that appears after encoding will be how is the correlation among the different variables. We observed that among the two encoding techniques the second approach was the best in order to visualize and understand the relationship among them. Furthermore, it kept the same dimensions as the original data. The analysis of the correlations as shown in figure 3, illustrates some important insights. First, we have 3 significant correlations in the dataset, between the variables Physical Health and Gen Health. This can be explained because when someone had an injury or suffered from an accident in the last 30 days, his overall feeling will be bad and will categorize his health as poorly or regular. Furthermore, when the individual has difficulties walking or using the stairs, it is probably because he/she suffered from an injury or accident involving his physical well-being (correlation Physical Health and Diff Walking). Finally, as the first relationship, if the person has trouble walking or climbing stairs, the general analysis of his health will be poorly due to the lacking performance in an activity that he needs to do every day (correlation Diff Walking and Gen Health)[2].

Even though the main correlations do not show anything surprising, we can see that the following relationships show us some interesting facts. The correlation among Physical Health and Mental Health shows that usually when someone is healthy physically, his mind or mental health will be good too, showing a relationship between being good and feeling good. Also, the correlation among Physical Activity and Gen Health, shows that when the persons exercises,

[2]It is worth mentioning that correlation is not causality, but in this case when we see one event the other event will be likely to exist.

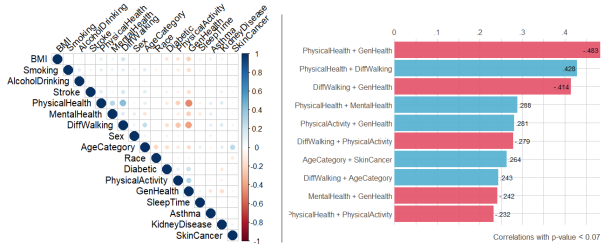

Figure 3: Correlation among encoded variables (left), Most correlated variables (right)

he will consider good his overall health, supporting that exercising is not only good for the body but mainly the mind [10].

4.2 Principal Component Analysis (PCA)

Now, taking into account that we have meaningful correlations among the dataset. It is important too see how to pick the best variables or reducing the amount of them but keeping the total variance. This is when PCA comes in. Before selecting the amount of components that will represent the 17 independent variables, it was important to do an analysis of how is the variance explained among the first components. The process to apply PCA in the data took **0.02 seconds**, giving the results in figure 4.

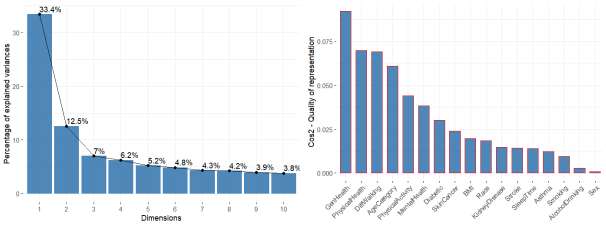

Figure 4: Explained variance among components (left), Variance per variable in new components (right)

We can see that in order to explain approximately 70% of the total variance present in the observations, it will be necessary to take at least 7 components. Which is a good amount of variance with the fact that we are cutting a total of 10 variables. Furthermore, we can see that among all the variables the ones that have their variance mostly represented in the new transformed components are Gen Health, Physical Health and Diff Walking, followed by Age. The main problem with this feature selection is that after extracting the new components we lose interpret-ability , because we do not have variables but condensed representations of them. Finally, based on this study we created a new dataset with only 7 columns (components), so we can see how it performs against the original data in the model refinement section.

5 Modelling

It was considered 5 methods during the learning steps to represent different algorithm and statistic approaches: two machine learning algorithms (Random Forest and XGBoost), a non parametric model (K-Nearest Neighbor), a linear model (Logistic Regression) and a data mining technique (Apriori algorithm). All of them were selected due to two main reasons: computing times and accuracy, which in the medical field will be crucial when detecting a patient with possible risks of heart disease as soon as possible. In each model we will see first some observations and insights brought after applying them. Then, in the evaluation section we will compare the results of the tuned models. Finally, for each of the following executions it was used 80% **(255731 rows)** and 20% for testing **(64064 rows)**.

5.1 Apriori Algorithm

This data mining technique was chosen to extract relevant association rules among the different variables in the dataset. Looking at all the rules present in the dataset we got initially **2923126 rules**. However, we reduced this number increasing the confidence to 90%,a minimum support of 25%, and looking for the rules that had in the right side our target variable. The results showed a total of **58496 rules** but all

of them were redundant, in other words, all of them had the same support and confidence compared to a general rule, so when filtering those rules by redundancy we ended with 0 relevant rules. Concluding that in this particular scenario the apriori algorithm was not good to extract meaningful patterns inside the data. The execution time of this algorithm was **1.11 seconds**.

5.2 Random Forest

This method is based on constructing multiple decision trees and takes the majority vote of all their predictions [3]. One of the benefits of applying Random Forest is that we can interpret easy the construction of the model based on the observations. Using a tuned version which consisted in 500 trees with the best hyper parameters, we got that with Random Forest we can see which one of the variables influence the most for the construction of the final model, as shown in figure 5.

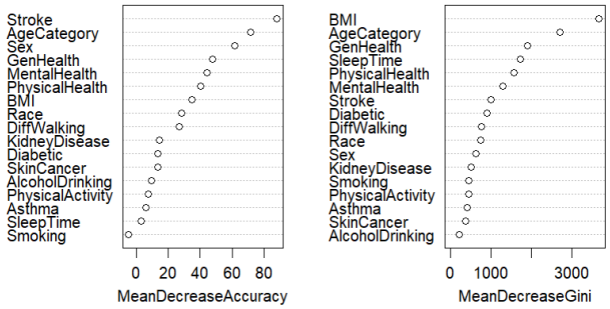

Figure 5: Influence of variables in Random Forest construction

From this we can see that among the variables that decreases the most the overall accuracy are: if the patient has ever had an stroke before, the age and the sex. Which shows that people who suffered a stroke can indicate a higher risk of heart disease, as well as, if they get older they will have more risk to acquire this kind of disease. On the other hand, the metric related to Gini refers how each variable contributes to the homogeneity of the final predictions [9], in other words, better helps to distinguish between the patients that have the disease or not at each split of the tree. Concluding that the age, the body mass index, stroke history and general health are four measures that can be useful when predicting if the individual will suffer a heart disease later.

5.3 K-Nearest Neighbor

Using non-parametric models like K-NN helps to get faster predictions, interpret easy the results with the clustering made by the algorithms and the high accuracy that it can achieve when the hyper parameters are well tuned. The main principle is locating the closest points to a new one and based on the majority vote assign the class that belongs to this new point. [8]. In this scenario, we could not find a good representation in the space that could show the clustering made by the algorithm between the variables.

5.4 Logistic Regression

An overall statistical model which is often used for classification. Their default attributes make them robust to any classification problem. Moreover, the accuracy that this model gives are often good, including this dataset. Through the model tuning it was possible to find the best value of the penalty applied to the model, in other words, the value which turns to zero the variables that are not usually used in the prediction [13].

5.5 XGBoost

XGBoost or extreme gradient descent, is a variation of gradient descent. But more efficient, because it is usually computes in parallel a linear method solver and a tree approach solver. Therefore, this method is much faster than the usual gradient descent, but keeping the predictive power from the original variation [5]. The selection of this model was given in order to have a more robust approach to the model line-up.

Metric	Random Forest	K-NN	Logistic Regression	XGBoost
Accuracy	0.917	0.9078	**0.9173**	0.9164
Kappa	0.0868	0.1027	**0.1377**	0.0553
Sensivity	0.0554	0.0855	**0.0929**	0.0345
Positive Pred Value	**0.5539**	0.317	0.5434	0.5344
Negative Pred Value	0.92	0.9214	**0.9227**	0.9185
Presicion	**0.5539**	0.317	0.5434	0.5344
Recall	0.0554	0.0855	**0.0929**	0.0345
F1 Score	0.1007	0.1347	**0.1588**	0.0649
Detection Rate	0.0046	0.0071	**0.0078**	0.0029
Balanced Accuracy	0.5256	0.53433	**0.5429**	0.5159
Time	703.09 sec	239.63 sec	14.26 sec	**0.14 sec**

Table 1: Performance metrics in models

6 Evaluation

All the results of the models are contained in table 1. As we are facing an imbalanced dataset the accuracy is not a proper measure to compare the performance among the models. Mainly, because the model can do a good performance predicting the class that is mostly present in the dataset, but doing poorly in the one that has fewer samples. In the prediction of heart disease is crucial the detection of true positives, because those patients should be warned and treated as soon as possible. So, we need to maximize the performance of this class that is less present. For that reason, we use metrics like Kappa and F1-Score to measure correctly the performance without the influence of this imbalance scenario. Cohen's Kappa metric relates the computation of the confusion metric and applying a relationship between the accuracy of the model and the the agreement between the model predictions and the actual class values as if happening by chance, which can acquire values between [-1,1] [4]. On the other hand, F1-Score is the combination of the precision of the model (accuracy of positive predictions) and recall (completeness of positive predictions), taking values between [0,1] [7]. Based on this two metrics, we can see that among the four models the best ones are: Random Forest and Logistic Regression. Both models were able to perform better than the others, where the Logistic Regression did it in **14.26 seconds**. Despite both models performed the best, we still achieved a lower score in the metrics that tell us if we are able to recognize the true positives. Therefore, we took the best model which in this case was logistic regression and we applied some improvements to the data in order to increase the performance.

6.1 Model Refinement

In order to improve the performance of the model, the dataset was balanced using both under-sampling and over-sampling. First, reducing randomly the amount of elements in the predominant class, then create randomly examples from the minority class until we got to a balance of both classes, from 233735 non positive and 21996 positive to an equal amount of **27525 rows per class**. The logistic model was then retrained and it achieved a kappa of **0.2388**, a F1-Score of **0.3385** (doubling the previous version), in a computation time of **0.39 sec**, which is still low but much better than the baseline models. Moreover, the second possible refinement was using the data with only 7 variables (after applying PCA), but the results were lower than the previous refinement with kappa **0.2318**, F1-Score **0.3301** and a computation time **0.24 sec**. Based on the previous analysis, we selected as our final choice Logistic Regression with an artificially balanced dataset.

7 Deployment

In the final iteration of the model, we would be able to predict in some scenarios the individuals that may suffer from heart disease in the future. This model can be deployed and used in the hospitals to warn the doctors about patients that have a high risk according to their history. Furthermore, send messages to those patients to be aware for checking often their health condition with a medical professional. Also, with the final Logistic Regression we were able to see some useful insights about key factors affecting the prediction of heart disease as shown in figure 6.

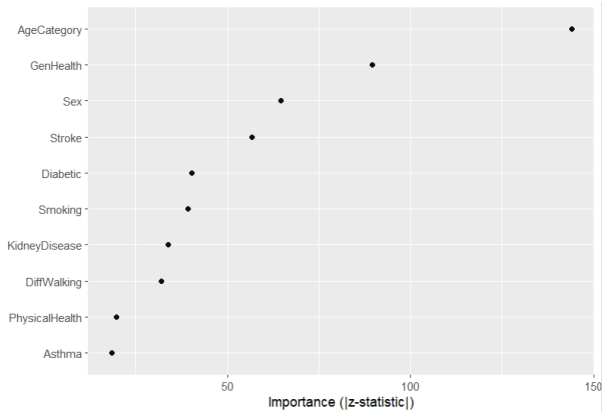

Figure 6: Influence of variables in final Logistic Regression

Which makes us conclude using figure 6 and the insights extracted from Random Forest (figure 5), that the main variables that can indicate if a person will suffer or not heart disease in the future is their age, stroke history and their overall feeling, this second one meaning how is their performance of the individual in a daily basis.

8 Further Work

Another field that will be important to take into consideration is adding the vaccination status of all the interviewed. In order to detect possible risk of heart diseases based on the last wave of vaccinations due to the COVID 19. As well as, increasing the accuracy of the model acquiring more data with true positives or trying other models that maximize the imbalance metrics as shown in this report, some alternatives: Neural Networks or Support vector machines.

9 Conclusion

We conclude that the main three factors that lead to heart disease are the age, stroke history and how the person feels his performance throughout the day. So, in order to prevent future risk of this kind of disease, the medical entities should prepare a series of diagnosis to high aged people in order to see their cardiovascular condition as they get older. As well as, flag the patient that had an stroke history to prevent a possible heart disease. A model was trained to predict if a person will suffer or not this pathology, taking into account different key indicators of their health. However, the performance was poor due to the imbalance dataset and the lack of generalization from the line-up of models selected. Then, we need to choose others that have less interpretability but more accuracy at prediction time.

References

[1] Aronow, W. S. Heart disease and aging. *Medical Clinics 90*, 5 (2006), 849–862.

[2] Association, A. H. 2022 heart disease stroke statistical update fact sheet global burden of disease. 2022 Statistics Diseases.

[3] Biau, G., and Scornet, E. A random forest guided tour. *Test 25* (2016), 197–227.

[4] Bland, J. Measurement in health and disease cohen's kappa. *Percentage agreement: a misleading approach York: University of York* (2008).

[5] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou,

T., ET AL. Xgboost: extreme gradient boosting. *R package version 0.4-2 1*, 4 (2015), 1–4.

[6] CLINIC, M. Heart disease - symptoms and causes. `https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118`.

[7] GOUTTE, C., AND GAUSSIER, E. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27* (2005), Springer, pp. 345–359.

[8] GUO, G., WANG, H., BELL, D., BI, Y., AND GREER, K. Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings* (2003), Springer, pp. 986–996.

[9] MARTINEZ-TABOADA, F., AND REDONDO, J. I. Variable importance plot (mean decrease accuracy and mean decrease Gini).

[10] MIKKELSEN, K., STOJANOVSKA, L., POLENAKOVIC, M., BOSEVSKI, M., AND APOSTOLOPOULOS, V. Exercise and mental health. *Maturitas 106* (2017), 48–56.

[11] OF DISEASE CONTROL, C., AND PREVENTION. Know your risk for heart disease. `https://www.cdc.gov/heartdisease/risk_factors.htm`.

[12] SEGER, C. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, 2018.

[13] SPERANDEI, S. Understanding logistic regression analysis. *Biochemia medica 24*, 1 (2014), 12–18.

[14] TOMOV, N., AND TOMOV, S. On deep neural networks for detecting heart disease. *CoRR abs/1808.07168* (2018).