Lovedeep Bajaj

CUS 610

WHAT MODEL TO USE TO CLASSIFY CARDIOVASCULAR DISEASE

**Abstract:**

There are many machine learning models that exist for classification purposes, but which one is the best one to use for your data. The dataset I had was a cardiovascular disease dataset and it contained certain features of individuals that had cardiovascular disease and those that did not. The features of these individuals that were recorded for this dataset were age, gender, height, weight, systolic blood pressure (ap\_hi), diastolic blood pressure (ap\_lo), cholesterol levels, glucose levels, did they smoke, did they drink, and did they do any physical activity. These are the features that I must use to classify if someone has cardiovascular disease or not, but as I stated there are so many machine learning techniques for classification which one do you pick. Well for my data I have determined that using a decision tree yields you the best results.

**Related Work:**

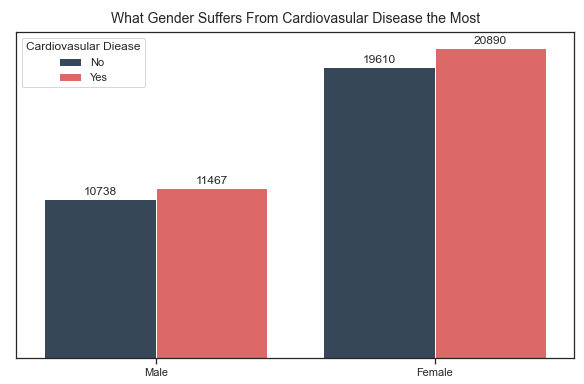
I obtained this dataset from the website Kaggle. If you go onto the Kaggle website and find this dataset you will see that many others have created classification models for cardiovascular disease (CVD). Most of the individuals use a decision tree, random forest, K nearest neighbor, or neural networks to classify CVD. Before creating the models most of the people have cleaned out the data first and removed all the duplicates. Afterward they use visualizations to get an understanding of the data and to see which one’s good predictors are to use for the model. Lastly, they create the models. My approach to this is somewhat alike for the most part, I plan on cleaning the data and removing duplicates and visualizing the data. But the place where I differ is that for the classification, I do not just want to look at the accuracy I want to look at the sensitivity of the model as well. That will tell you much more then accuracy would for medical data. Sensitivity formula is TP/(TP+FN) and when it comes to medical data you want to make sure the FN (False Negative) value is the lowest because you would rather tell someone who does not have cardiovascular disease that they do rather than tell someone who does have cardiovascular disease that they don’t. Thus, if the sensitivity value is too low that means that there was a lot of FN values and if sensitivity value good then it means that there was not that many FN values. On top of that I want to use ensemble methods to see if I can achieve a greater accuracy and sensitivity than I would be able to just by using one model on its own. I plan on using the VotingClassifier and trying to figure out what combination of models will give me the best results. To make sure that I have found the best model I also plan on doing a BaggingClassifier on the model that by itself gives me the best results. Then I will compare the results between the BaggingClassifier, VotingClassifier, and the models by themselves and evaluate which one is the best to use.

**Data:**

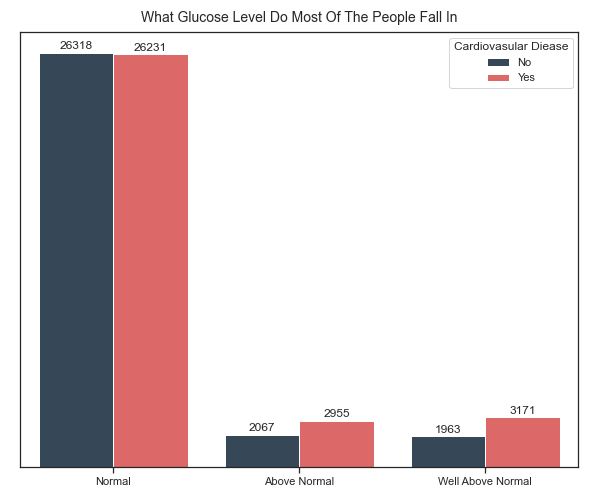
The first thing I had to do was convert some of the features values because they were in a form that I could not understand just by viewing it. For age it was originally in days, but the number of days someone has been alive is not the best to understand their age that is why I converted that into years. For height it was originally in centimeter, but I converted that into inches also I converted weight from kilograms into pounds. For gender I made 0 represent males instead of 2. After doing this conversion I also deleted all the duplicates in the data. This is what I had to do before I could start look for the outliers in the data so it would be easier for me to spot them.

For the pre-processing portion many of the features didn’t require any cleaning since they were binary and didn’t contain any missing values. The binary features in the data were gender, smoke, alco (alcohol intake), active (physical active), and cardio (cardiovascular disease). Also the features cholesterol and gluc (glucose levels) did not require any cleaning since they only contained 3 values to tell you the level of cholesterol and glucose that individual had. Age was a feature that the boxplot showed me 2 potential outliers which were individuals of the age 29 and 30. I decided not to remove them since none of the individuals that were in that 29 or 30 had cardiovascular disease which makes sense for someone of that age. Moving onto height, I removed all the females that were below the height of 4ft’8in (56inches) and that were above 6ft. The reason for this is because the average height for females in the US is 5ft’2in so I wanted to keep some of the females that were below that height and thus I decided 4ft’8in because not that many females were below that height. I decided that 6ft would be the cut off for females since less than 1% of them are over 6ft. For males I keep all the males that were above 5ft’2in and below 7ft. 14.5% of males are in the 6ft to 7ft in the US so I did not want to remove any of the 6ft individuals, and I decided to go with 5ft’2in because the average height for a male in the US is 5ft’4in. For the weight I decided to make the lowest weight for the females 60Ibs since 70Ibs is the ideal weight for 4ft’8in and I wanted to keep the females that were underweight. For the males I went with 90Ibs since the ideal weight for a male that is 5ft’2in is 102Ibs and I wanted to incorporate underweight individuals in my analysis. For ap\_hi (systolic blood pressure) I kept people whose ap\_hi value was 60 till 300. Since the below normal systolic blood pressure is 90, I decided 60 would be a good minimum value. I decided to make 300 the max because above 180 is horrible and I wanted to keep those individuals, so I went with 300 as my max. For ap\_lo (diastolic blood pressure) below normal is 60 and bad is higher than 120 I wanted to have those people in my analysis, so I made the min 40 and the max 190.

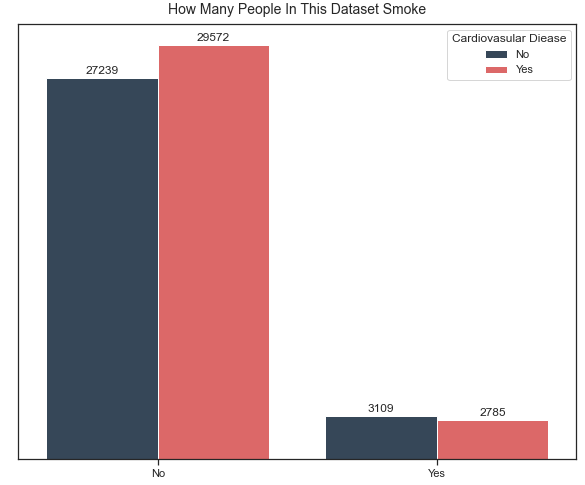
After the preprocessing was done, I created some visualizations that helped me decide which features I wanted to remove and see what it would do to the model’s accuracy and sensitivity. I removed height and weight and replaced it with BMI and other features that were removed were gender, glucose, smoke, and alco.

**Figure 1.**

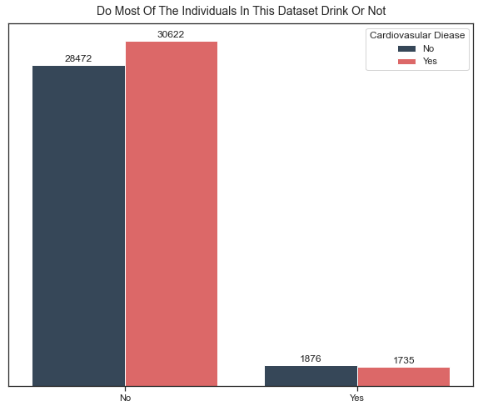
The image labeled Figure 1 shows you that most of the people in this dataset are females so gender might not play a role.

**Figure 2.**

In the image labeled figure 2 most of the people have normal glucose level which means it might not a good indicator for identifying cardiovascular disease.

**Figure 3.**

In the image labeled figure 3 most of the people do not smoke this might not a good indicator for identifying cardiovascular disease.

**Figure 4.**

In the image labeled figure 4 most of the people do not drink and that means it might not a good feature to be used for classification.

**Methodology:**

During the analysis phrase of my project, I used decision tree, random forest, naïve bayes, logistic regression, and support vector machine. I wanted to see how each model performed individually before combining them for the VotingClassifier as well as deciding which one to use for the BaggingClassifier. Before I could test these techniques and see which one gave me a better model, I needed to figure out how many trees to use for random forest, so I created a loop that tested 30 till 170 trees only going up by 10. I found out that 130 is the best number of trees to use. After that I needed to find out the max\_depth to be used for the random forest and decision tree. Once I had figured out the best max\_depth for the random forest (max\_depth = 9) and decision tree (max\_depth = 4) I created models with all the techniques. After that created models using the different techniques and I also looked for which features when removed gave me the best accuracy and sensitivity. Once I had done all of this, I created multiple VotingClassifiers to see which models combined gave me a better accuracy and sensitivity than I was able to achieve with one model. After I had obtained that data, I took the decision tree because it was the model that by itself gave me the highest sensitivity and used it in my BaggingClassifier. I plugged in a decision tree using Gini which is the default criterion and created another BaggingClassifier that used entropy to see which one yielded me better results.

**Results:**

The decision tree using the max\_depth of 4 and not including the gender, gluc (glucose level), height, and weight gave me the best results and is the one that I would use. Even though the base model for the BaggingClassifiers is the decision tree gave me similar results the sensitivity was just a bit lower than the decision tree. Also, all the other values such as accuracy, specificity, and precision were so close that they did not really make a difference in deciding between the model obtained from using the decision tree or either one of the models obtained from the BaggingClassifiers. I did not talk about any of the VotingClassifiers that I created because none of them were able to obtain higher sensitivity, but it did obtain a higher accuracy than the decision tree. But the accuracy was not that high that it would make me ignore the difference in the sensitivity values.

**References:**

1. [www.kaggle.com/sulianova/cardiovascular-disease-dataset](http://www.kaggle.com/sulianova/cardiovascular-disease-dataset)
2. [www.scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html](http://www.scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)
3. [www.scikit-learn.org/stable/modules/generated/sklearn.model\_selection.KFold.html](http://www.scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html)
4. [www.scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html](http://www.scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)
5. [www.scikit-learn.org/stable/modules/naive\_bayes.html](http://www.scikit-learn.org/stable/modules/naive_bayes.html)
6. [www.scikit-learn.org/stable/modules/tree.html](http://www.scikit-learn.org/stable/modules/tree.html)
7. [www.scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html](http://www.scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html)
8. [www.scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html](http://www.scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html)
9. [www.matplotlib.org](http://www.matplotlib.org)
10. <https://www.healthline.com/health/average-height-for-men#u.s.-height>
11. <https://www.healthline.com/health/womens-health/average-height-for-women#:~:text=How%20tall%20are%20American%20women,average%20weight%20is%20170.6%20pounds>.
12. <https://www.healthline.com/health/high-blood-pressure-hypertension/blood-pressure-reading-explained#normal>
13. <https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/blood-pressure/art-20050982>
14. <https://www.mayoclinic.org/diseases-conditions/low-blood-pressure/symptoms-causes/syc-20355465>
15. <https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings>