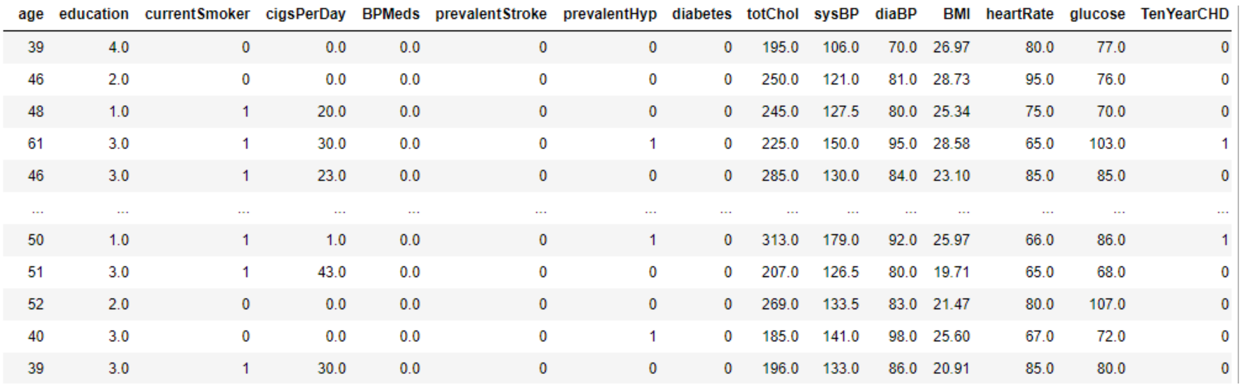
# The Framingham Heart Study dataset analysis

The following analysis refers to *The Framingham Heart Study* a public dataset *that contains* information collected by Boston University and the *National Heart, Lung and Blood Institute in the* U.S. in the late 1940s and its purpose is to assess which factors are at risk for developing some coronary artery disease (CHD). Download made on the site https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset. The machine learning - *Decision Tree method made* in *the* python programming language for dataset analysis was used.

The beginning of the study had just over 5000 patients, the people chosen were healthy, between 30 and 59, from the city *of Framingham,* Massachusetts. They would participate for 20 years of the study and, every two years, go to a medical center to perform tests and fill out questionnaires about their life habits, for example, whether they smoked or exercised.



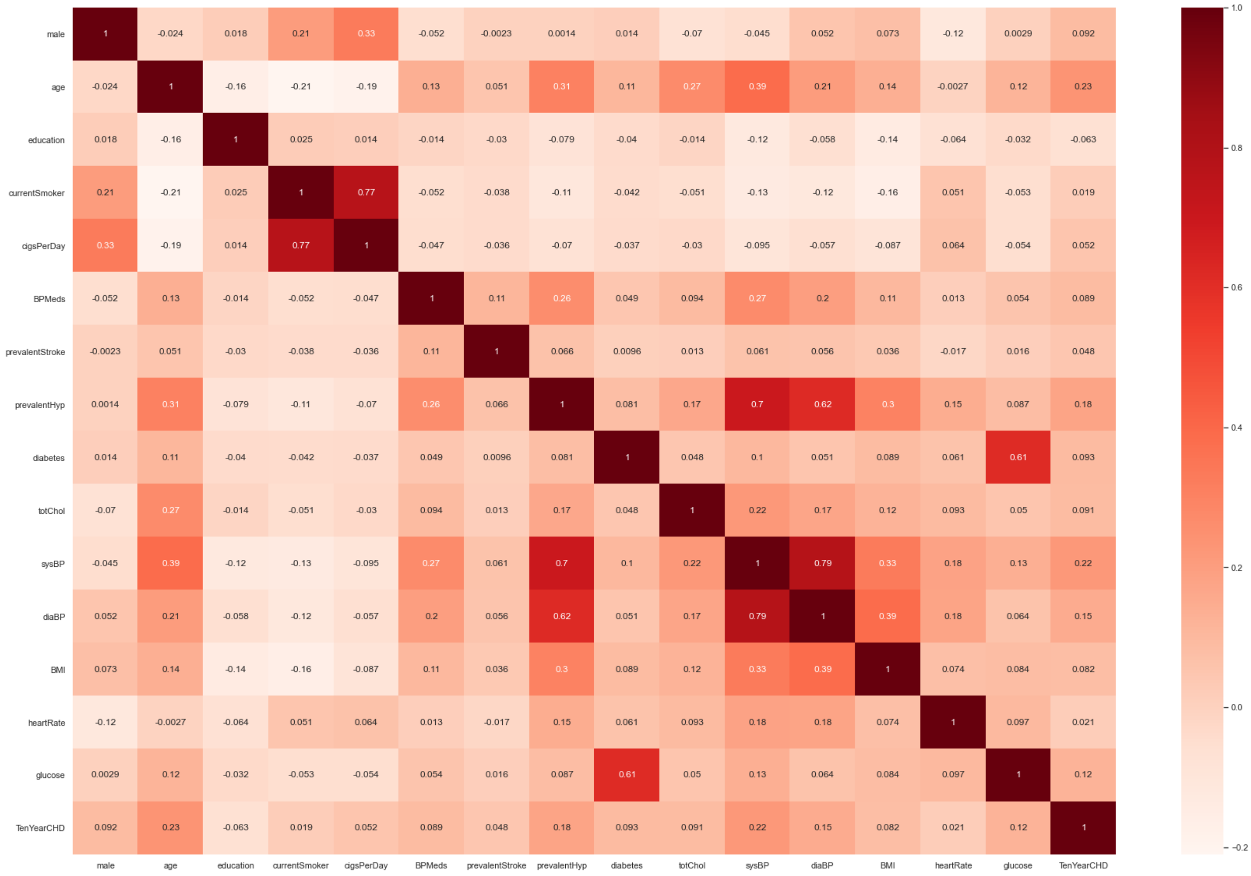
| age | education | currentSmoker | cigsPerDay | BPMeds | prevalentStroke | prevalentHyp | diabetes | totChol | sysBP | diaBP | BMI | heartRate | glucose | TenYearCHD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 39 | 4.0 | 0 | 0.0 | 0.0 | 0 | 0 | 0 | 195.0 | 106.0 | 70.0 | 26.97 | 80.0 | 77.0 | 0 |
| 46 | 2.0 | 0 | 0.0 | 0.0 | 0 | 0 | 0 | 250.0 | 121.0 | 81.0 | 28.73 | 95.0 | 76.0 | 0 |
| 48 | 1.0 | 1 | 20.0 | 0.0 | 0 | 0 | 0 | 245.0 | 127.5 | 80.0 | 25.34 | 75.0 | 70.0 | 0 |
| 61 | 3.0 | 1 | 30.0 | 0.0 | 0 | 1 | 0 | 225.0 | 150.0 | 95.0 | 28.58 | 65.0 | 103.0 | 1 |
| 46 | 3.0 | 1 | 23.0 | 0.0 | 0 | 0 | 0 | 285.0 | 130.0 | 84.0 | 23.10 | 85.0 | 85.0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 50 | 1.0 | 1 | 1.0 | 0.0 | 0 | 1 | 0 | 313.0 | 179.0 | 92.0 | 25.97 | 66.0 | 86.0 | 1 |
| 51 | 3.0 | 1 | 43.0 | 0.0 | 0 | 0 | 0 | 207.0 | 126.5 | 80.0 | 19.71 | 65.0 | 68.0 | 0 |
| 52 | 2.0 | 0 | 0.0 | 0.0 | 0 | 0 | 0 | 269.0 | 133.5 | 83.0 | 21.47 | 80.0 | 107.0 | 0 |
| 40 | 3.0 | 0 | 0.0 | 0.0 | 0 | 1 | 0 | 185.0 | 141.0 | 98.0 | 25.60 | 67.0 | 72.0 | 0 |
| 39 | 3.0 | 1 | 30.0 | 0.0 | 0 | 0 | 0 | 196.0 | 133.0 | 86.0 | 20.91 | 85.0 | 80.0 | 0 |

Table 1 - *dataframe*

In table 1 we have the data loaded. There are 16 variables, 15 *features* and 1 target. The 15 features are the risk factors that have been studied and the target *is* whether the person developed a CHD or not during the study. Risk factors are:

- *Male*: 1 indicates whether you are a man and 0 a woman
- *Age*: The age of the participant
- *Education Level*: 1-Incomplete high school, 2- Complete high school, 3-incomplete higher education, 4-Complete higher education
- *currentSmoker*: 1- Smoker, 0- Non-smoker
- *cigsPerDay*: Number of cigarettes per day
- *BPMeds*: How many pressure medications are you taking
- *prevalentStroke*: 0- Never had stroke, 1-Has history of stroke
- *prevalentHyp*: 0-Has no hypertension, 1-Has hypertension
- diabetes: 0-No diabetes, 1-Has diabetes
- *totChol*: Total cholesterol
- *sysBP*: Systolic pressure
- *diaBP*: Diastolic pressure
- *BMI*: Body mass index
- *Heart Rate*: Heart rate in beats per minute
- *Glucose*: Glucose in the blood (mg/dL)

And the target, *TenYearCHD*: 0 means that the person has not developed any disease and 1 has developed.



| | male | age | education | currentSmoker | cigsPerDay | BPMeds | prevalentStroke | prevalentHyp | diabetes | totChol | sysBP | diaBP | BMI | heartRate | glucose | TenYearCHD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| male | 1 | -0.024 | 0.018 | 0.21 | 0.33 | -0.052 | -0.0023 | 0.0014 | 0.014 | -0.07 | -0.045 | 0.052 | 0.073 | -0.12 | 0.0029 | 0.092 |
| age | -0.024 | 1 | -0.16 | -0.21 | -0.19 | 0.13 | 0.051 | 0.31 | 0.11 | 0.27 | 0.39 | 0.21 | 0.14 | -0.0027 | 0.12 | 0.23 |
| education | 0.018 | -0.16 | 1 | 0.025 | 0.014 | -0.014 | -0.03 | -0.079 | -0.04 | -0.014 | -0.12 | -0.058 | -0.14 | -0.064 | -0.032 | -0.063 |
| currentSmoker | 0.21 | -0.21 | 0.025 | 1 | 0.77 | -0.052 | -0.038 | -0.11 | -0.042 | -0.051 | -0.13 | -0.12 | -0.16 | 0.051 | -0.053 | 0.019 |
| cigsPerDay | 0.33 | -0.19 | 0.014 | 0.77 | 1 | -0.047 | -0.036 | -0.07 | -0.037 | -0.03 | -0.095 | -0.057 | -0.087 | 0.064 | -0.054 | 0.052 |
| BPMeds | -0.052 | 0.13 | -0.014 | -0.052 | -0.047 | 1 | 0.11 | 0.26 | 0.049 | 0.094 | 0.27 | 0.2 | 0.11 | 0.013 | 0.054 | 0.089 |
| prevalentStroke | -0.0023 | 0.051 | -0.03 | -0.038 | -0.036 | 0.11 | 1 | 0.066 | 0.0096 | 0.013 | 0.061 | 0.056 | 0.036 | -0.017 | 0.016 | 0.048 |
| prevalentHyp | 0.0014 | 0.31 | -0.079 | -0.11 | -0.07 | 0.26 | 0.066 | 1 | 0.081 | 0.17 | 0.7 | 0.62 | 0.3 | 0.15 | 0.087 | 0.18 |
| diabetes | 0.014 | 0.11 | -0.04 | -0.042 | -0.037 | 0.049 | 0.0096 | 0.081 | 1 | 0.048 | 0.1 | 0.051 | 0.089 | 0.061 | 0.61 | 0.093 |
| totChol | -0.07 | 0.27 | -0.014 | -0.051 | -0.03 | 0.094 | 0.013 | 0.17 | 0.048 | 1 | 0.22 | 0.17 | 0.12 | 0.093 | 0.05 | 0.091 |
| sysBP | -0.045 | 0.39 | -0.12 | -0.13 | -0.095 | 0.27 | 0.061 | 0.7 | 0.1 | 0.22 | 1 | 0.79 | 0.33 | 0.18 | 0.13 | 0.22 |
| diaBP | 0.052 | 0.21 | -0.058 | -0.12 | -0.057 | 0.2 | 0.056 | 0.62 | 0.051 | 0.17 | 0.79 | 1 | 0.39 | 0.18 | 0.064 | 0.15 |
| BMI | 0.073 | 0.14 | -0.14 | -0.16 | -0.087 | 0.11 | 0.036 | 0.3 | 0.089 | 0.12 | 0.33 | 0.39 | 1 | 0.074 | 0.084 | 0.082 |
| heartRate | -0.12 | -0.0027 | -0.064 | 0.051 | 0.064 | 0.013 | -0.017 | 0.15 | 0.061 | 0.093 | 0.18 | 0.18 | 0.074 | 1 | 0.097 | 0.021 |
| glucose | 0.0029 | 0.12 | -0.032 | -0.053 | -0.054 | 0.054 | 0.016 | 0.087 | 0.61 | 0.05 | 0.13 | 0.064 | 0.084 | 0.097 | 1 | 0.12 |
| TenYearCHD | 0.092 | 0.23 | -0.063 | 0.019 | 0.052 | 0.089 | 0.048 | 0.18 | 0.093 | 0.091 | 0.22 | 0.15 | 0.082 | 0.021 | 0.12 | 1 |

*Figure 1 - Heatmap of the dataset*

In Figure 1 we *have the heatmap* of *the dataset* to verify the relationship between the variables and find possible relationships to explore. 1 represents that the variables are very related and -1 that they are inversely related. We see, in Figure 1, that for *TenYearCHD* and the rest of the risk factors, there are no values close to 1 or -1, so it was concluded that it is not a single risk factor that contributes to the development of a CHD, but rather the set of them.

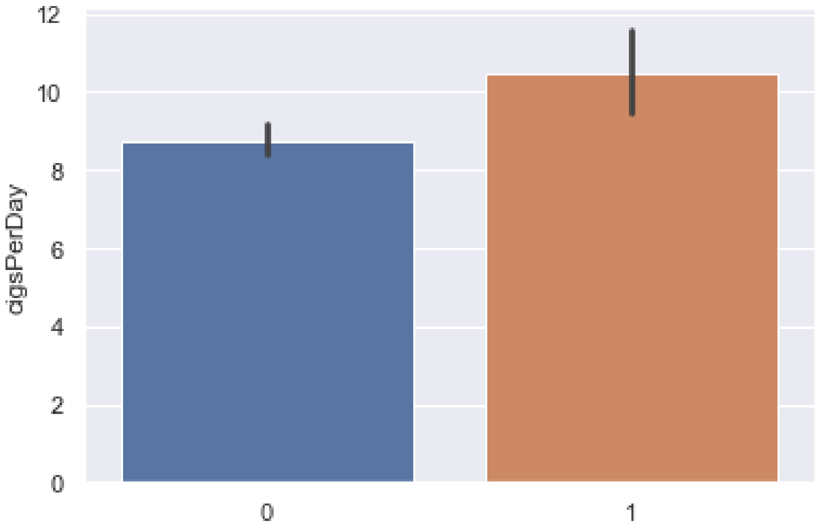


*Figure 2 - Bar Chart - Cigarettes per Day and TenYearCHD*

Figure 2 shows that people who smoke more cigarettes per day are more likely to develop a CHD, so smoking more is a risk factor.

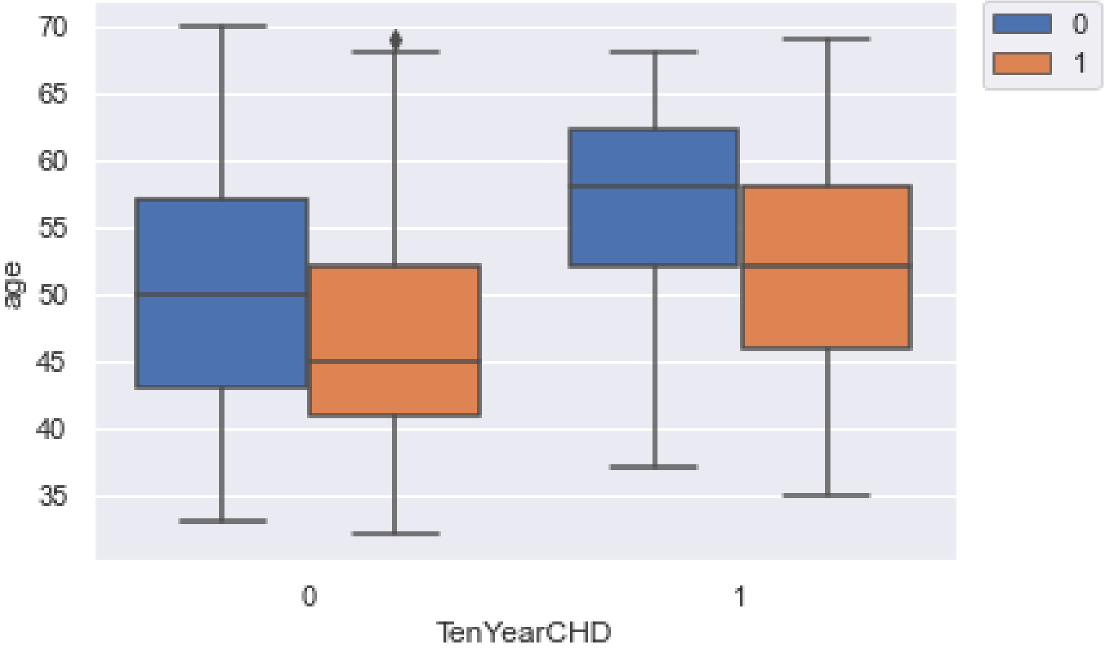


*Figure 3 – Boxplot of Age and TenYearCHD with hue=currentSmoker*

Older people are more likely to develop CHD, as seen in Figure 3. We can observe that 75% of people who smoke and develop CHD are aged up to 57-58 years, which is the median for 50% of the non-smoking group, therefore, smokers have a higher chance of developing CHD earlier, compared to the other non-smoking group.

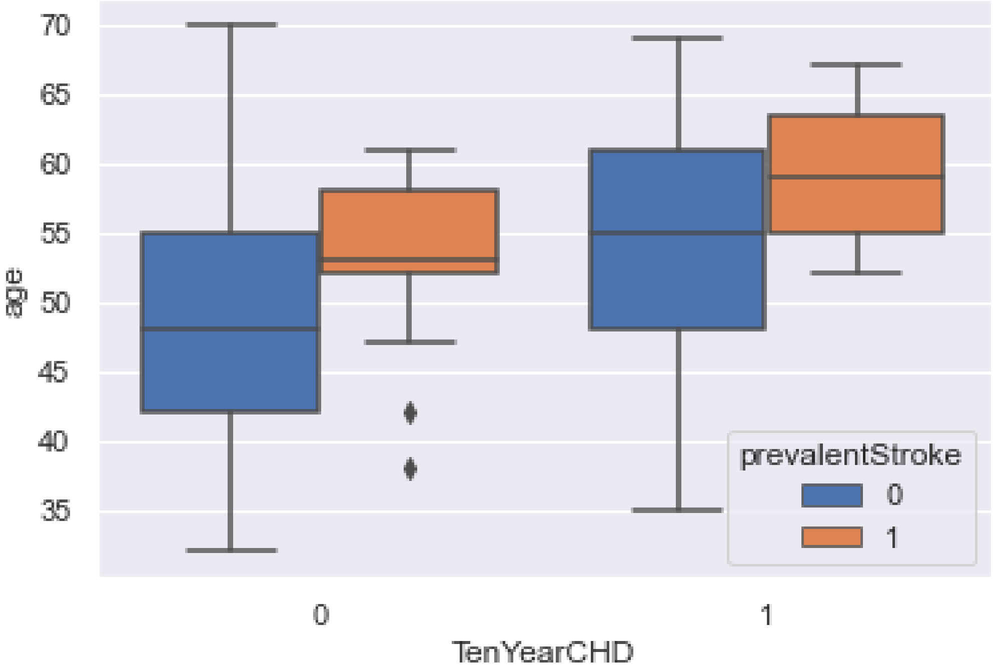


*Figure 4 - Boxplot Age and TenYearCHD with hue=prevalentStroke*

It is noted that strokes are more common at older ages, seen in Figure 4. We have that the minimum ages observed for stroke are 47 and 53 approximately, excluding the 2 outliers that have been detected.

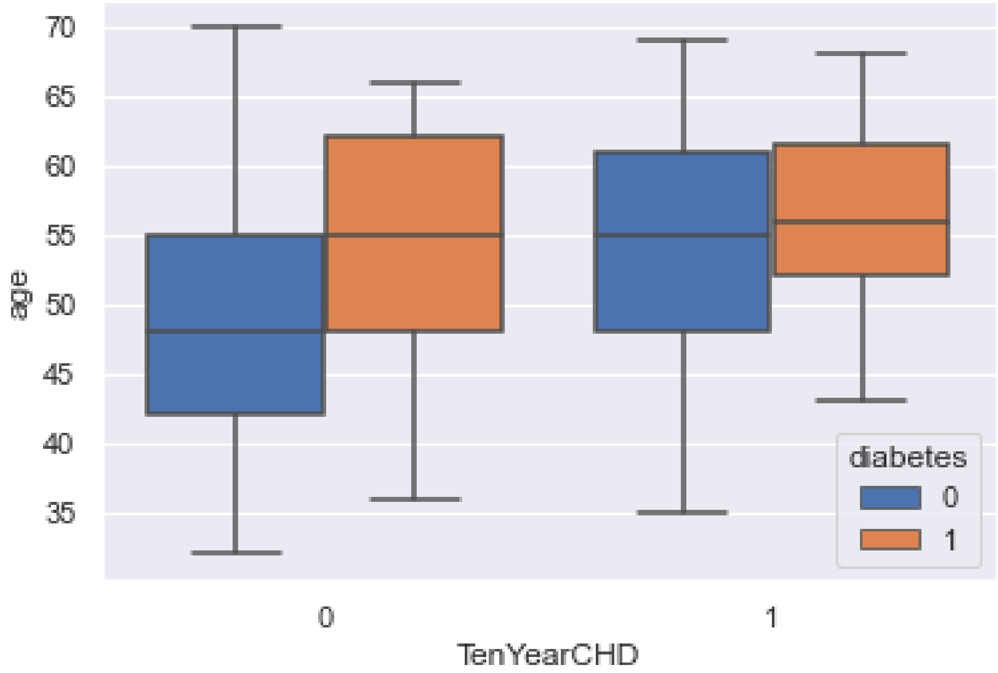


*Figure 5 - Boxplot Age and TenYearCHD with hue=diabetes*

In Figure 5, we observed the same behavior as the past figure, the presence of diabetes is more common in older people.

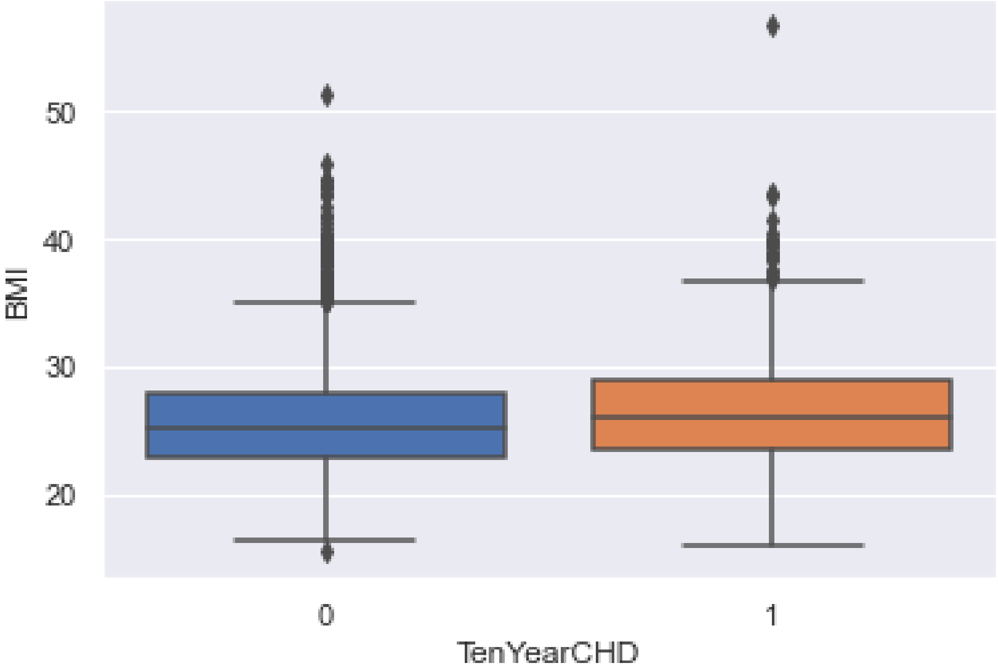


*Figura 6 - Boxplot BMI e TenYearCHD*

Figure 6 shows that BMI is not a risk factor for the development of BMI, as the observed difference is low.

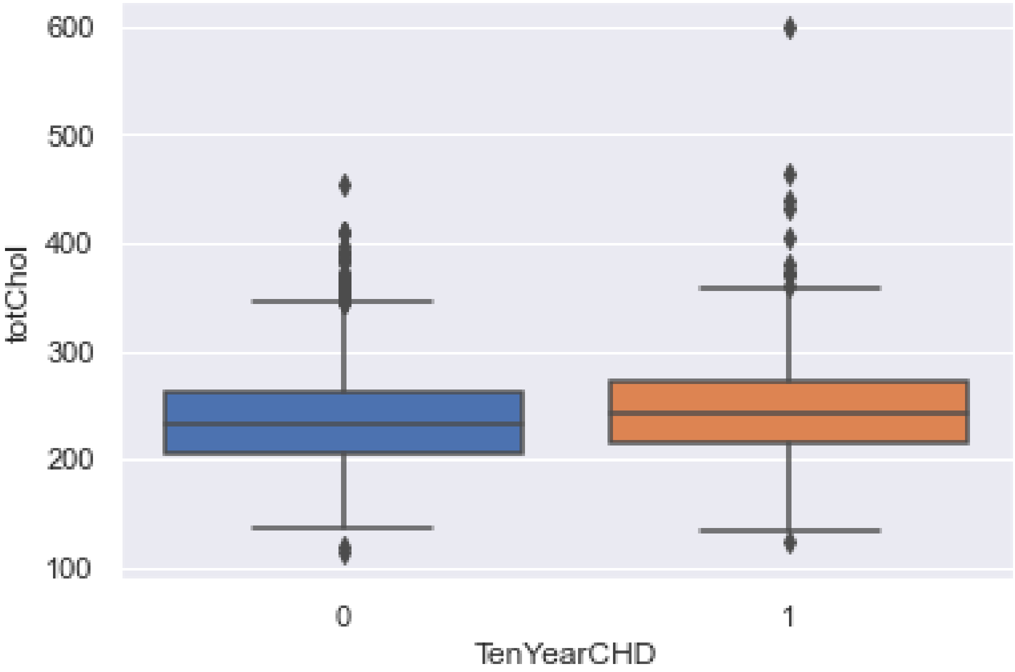


*Figure 7 - Boxplot totChol and TenYearCHD*

It is already known that cholesterol is a risk factor for heart disease, but it is LDL cholesterol that is the risk factor. HDL cholesterol is beneficial for health. In Figure 7, apparently cholesterol would not be a risk factor, but this is *due to that totChol* is the combination of the 2 types of cholesterol. If this variable were divinity in 2, LDL and HDL, we would probably have clearer results, indicating that LDL is a risk factor.

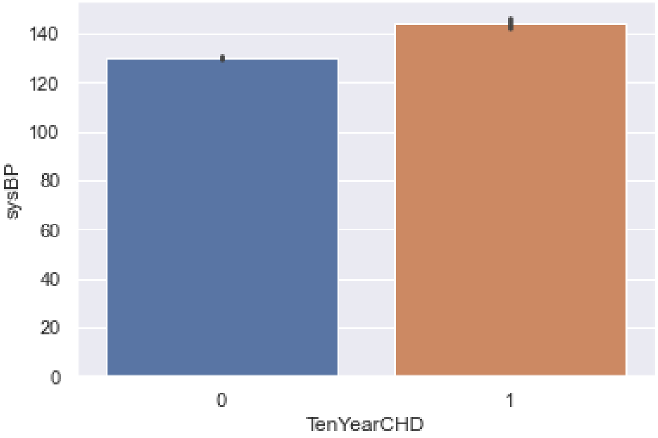


*Figure 8 - Bar Chart - sysBP and TenYearCHD*

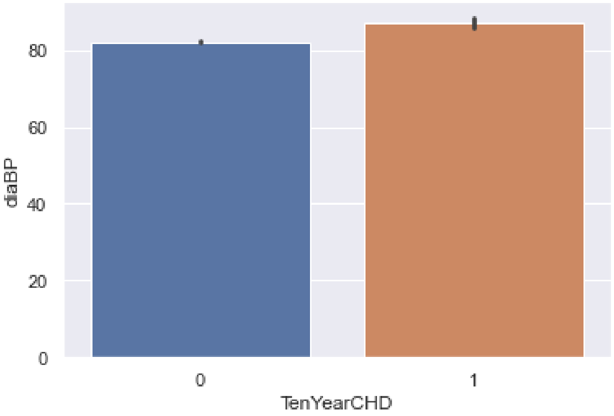


*Figure 9 - Bar Chart - diasBP and TenYearCHD*

In figures 8 and 9, it is observed that people with a higher blood pressure have a higher probability of developing CHD, that is, it is a risk factor.

In *the dataset,* there are more than 3600 observations and only 15.23 %, came to develop CHD, i.e., only 557 positive observations. This represents an unbalanced *dataset* that can generate problems for the model, for example, overfitting. This will be resolved by applying the SMOTE and Pipeline technique in conjunction with under sampling to get the best possible model.

| Accuracy | |
|---|---|
| SMOTE | Pipeline (SMOTE + RandomUnderSampling) |
| 69,67% | 80,38% |

Table 2 - Accuracy of the model

The model proposed for this study was *the Decision Tree* to predict whether or not a person will develop CHD, taking into account all the risk factors presented above. Factors were removed from the model: *education,* totChol*, BMI* and *heartRat,*in order to reduce the complexity of decision *tree* and having a gain in *accuracy, precision* and *recall.* By table 2, we have that with only the SMOTE technique *the accuracy* was lower than with the set of techniques coupled with the pipeline. Accuracy measures how many cases have been correctly identified, but does not identify false negatives and false positives. In this case of the Framingham study, the best measures *are precision* and revaluate identify false positives and false negatives, which are important for a study on disease development. We don't want to have a big rate of false negatives when it comes to diseases.

| | Dataset balancing technique | |
|---|---|---|
| | SMOTE | Pipeline (SMOTE + RandomUnderSampling) |
| Precision | 19,0% | 79,0% |
| Recall | 29,0% | 82,0% |

Table 3 - Precision and Recall for positive cases

Looking at the differences in the results presented in Table 3, it is important to note their differences. A model with 69.67% accuracy may be *acceptable,* but with *low results in precision and recall,* the large number of false negatives and positives very negatively impact the overall performance of the model, so it is not recommended to use it. The other model obtained better results in the 3 measures presented, accuracy, precision and *recall,* so this is the most indicated model for this study, since we are trying to predict the development of a CHD.