Instituto Tecnológico y de Estudios Superiores de Monterrey

Intelligent Systems

Professor Benjamín Valdés Aguirre

Heart Failure Prediction using Decision Tree and Random Forest

Pablo Antonio Ortegón Ruiz A01702995

Due date: 31/05/21

Introduction

It is estimated that 17.9 million people die every year due to cardiovascular diseases (CVDs), making it the number 1 cause of death globally. A heart failure occurs when the heart cannot fill up with enough blood or on some occasions when it is too weak to properly pump the blood. A heart failure does not mean that the heart no longer works, but that medical care is required.

There are tools that allows us to quantify symptoms and test values, which can be used to perform analysis. In many situations the results of these analysis can be hard for doctors and medical professionals to correlate, therefore engineering methods such as artificial intelligence, have an important role in predicting the patient's survival using the data and can tell which features have a heavier weigh in the prediction's result. In this report the process of a heart failure prediction using the Decision Trees and the Random Forests methods, and the data analysis will be described.

The Supervised Learning Algorithms

Machine Learning systems can be classified according to the amount and type of supervision they get during training. During this project, the supervised learning method played an important role as the training data fed to the algorithms had expected results, which were then be compared to obtain the prediction accuracy. In this type of classification, the machine does not learn on its own, but relies on the programmer to *supervise* the results.

One A.I. method that is classified as supervised learning and capable of fitting complex datasets is the Decision Tree method. This type of algorithm predicts a result or a probability when the inputs follow a "branch" path according to certain statements or conditions being fulfilled by the data values, this means, every combination of values in the data follows a different "branch" as they fulfill or not different conditions. One of the many qualities of Decision Trees is that they require very little data preparation and can be very powerful regardless.

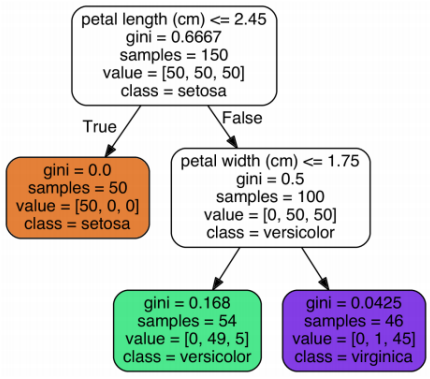


Figure 1.1 Example of Decision Tree (Geron, Aurelien)

When a branch of Decision Tree does not reach a full probability, it means that it has impurity as it is unsure of the complete result. A node is pure when its gini is equal to 0, meaning all training instances it applies to belong to the same class. It is calculated with the following equation:

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^{\ 2}$$

$\boldsymbol{p_{i,k}}$ is the ratio of class k instances among the training instances in the i^{th} node.

The second analysis was used implanting the Random Forest method. This method consists in a group or ensemble of Decision Trees, each trained on a different random subset of the training set, to get a more accurate prediction. This method works by obtaining the predictions of all individual trees, then calculate the average or get use the prediction with the most votes. It is thought to be one of the most powerful Machine Learning algorithms available today.

These algorithms, despite behaving in a similar manner due to their complementary backbones, help us classify features and classes with decent accuracy and do not required high complexity.

Dataset and Analysis

The dataset implemented during the analysis was provided by *Kaggle.com* and it is called *Heart Failure Prediction* uploaded by user *Larxel*. This dataser contains 13 features, which describe clinical, body and lifestyle information of 299 patients. Out of the 13 features, 5 are binary and the rest have numeric values of different magnitudes. In order to have a grasp of the weight of each feature in the death event, plots of data were applied.

To have a better understanding of the ratios in every plot, first, a correlation between patients who survived and did not was calculated and is presented in the following pie chart, letting us know that 2 thirds of the dataset population were alive after the data entry.

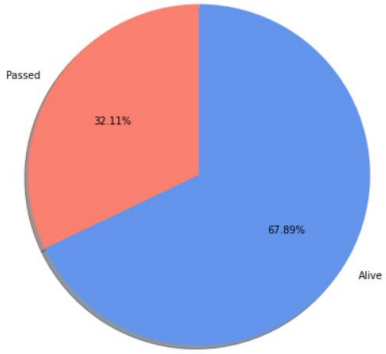
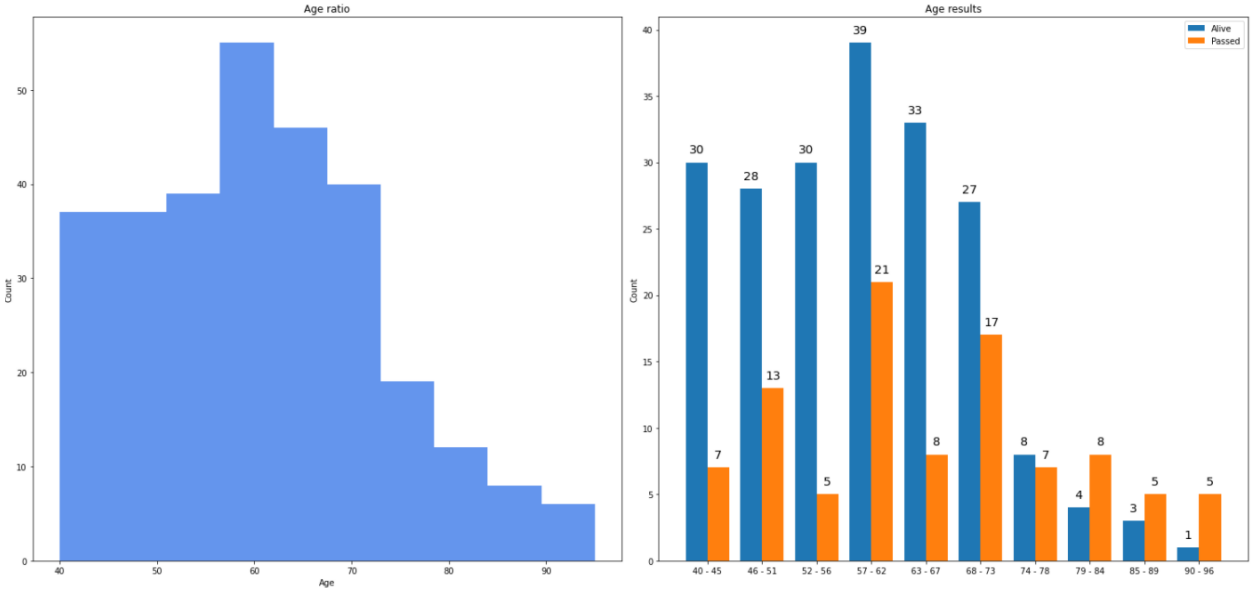


Figure 2.1 Death Event ratio

The first feature to be discussed is the age. Most patients were between 57 and 62 years old, the youngest been 40 and the oldest 96. According to the results, age played an important role, as the older patients had a greater death result than the younger ones.



The second feature is anemia, a disease in which the patient lacks enough red blood cells to transport a healthy level of oxygen to the body's tissues. 56.85% of patients presented this disease but according to the results it did not play an important role, as the ratio between the death event and anemia was not significantly different.

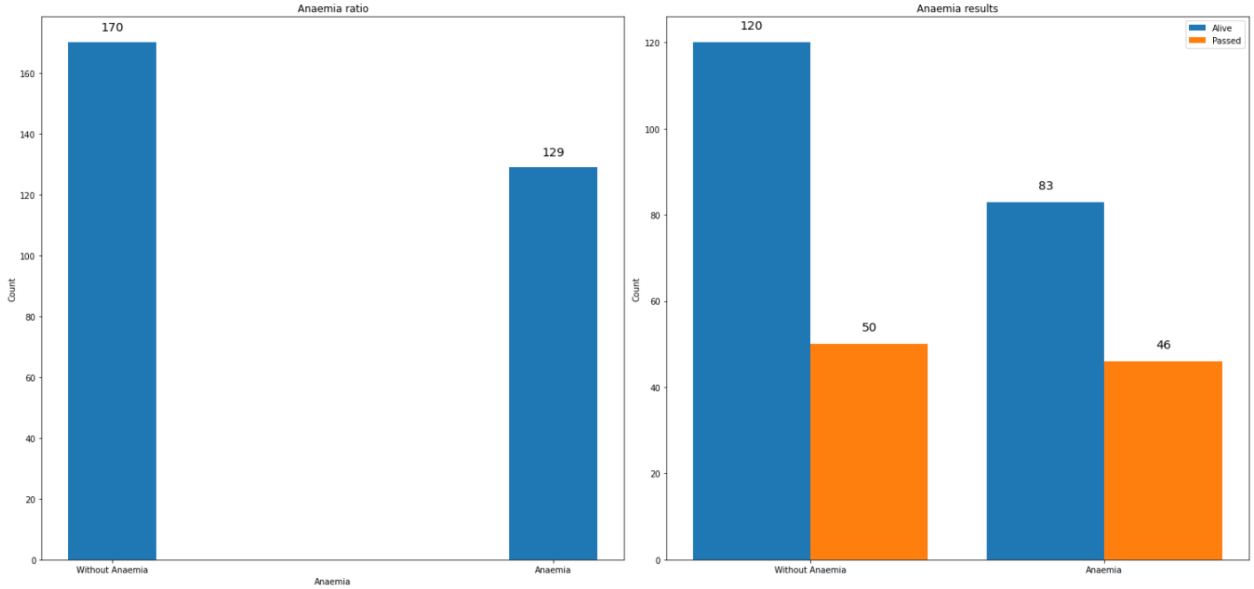


Figure 2.3 Anemia ratio

Next is the Creatinine Phosphokinase levels. This in the enzyme that flows into the blood when a muscle or tissue is damaged, therefore high amounts of this feature predict heart failure. According to the graph, most people were in the recommended amounts, (between 32 and 294 mcg/L) and as the amount increases so does the death event.

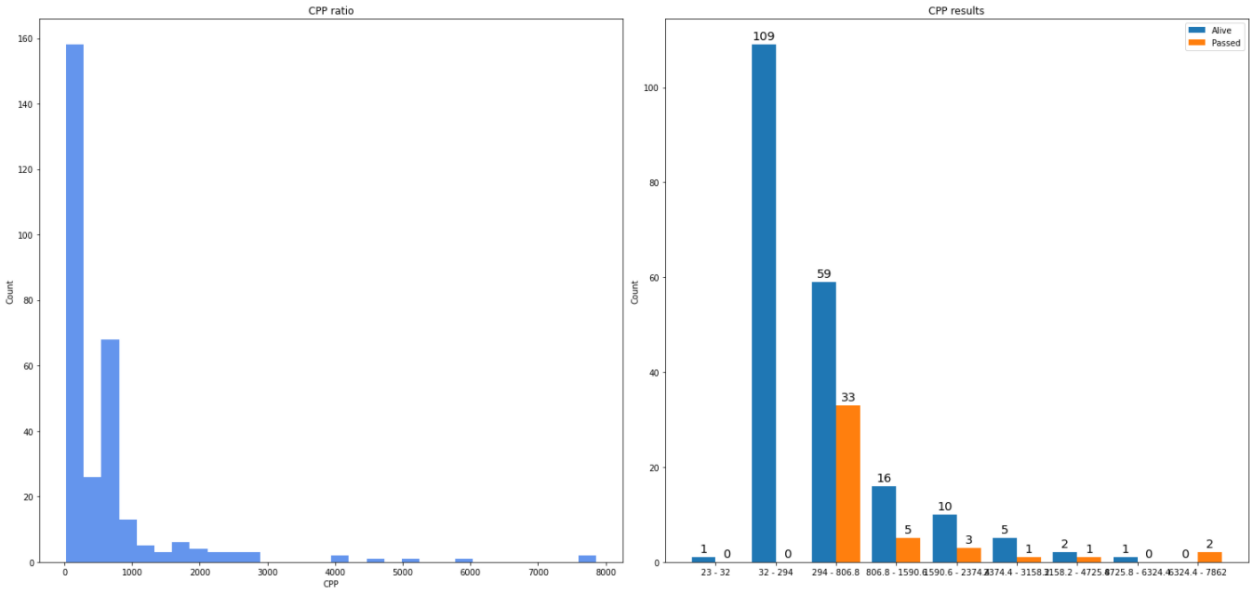


Figure 2.4 Creatinine Phosphokinase ratio

Diabetes is a disease in which the patient's glucose levels are high and can lead to not producing insulin. During this analysis, we can tell that diabetes does illustrate a big role, as the ratio between patients with diabetes and without diabetes was similar

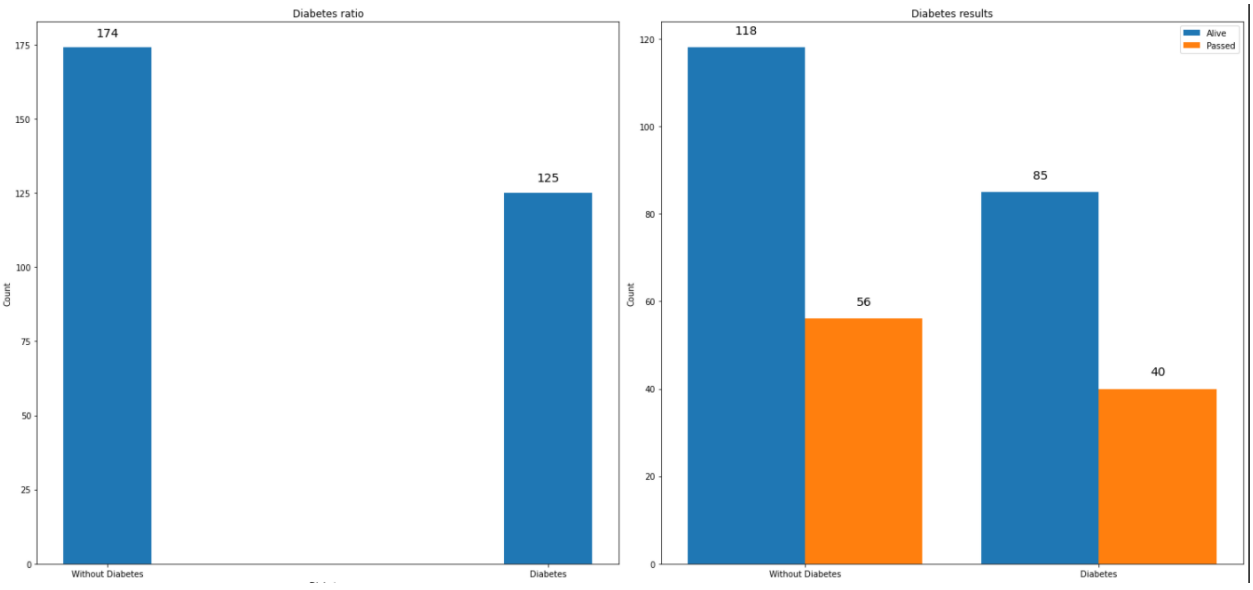


Figure 2.5 Diabetes ratio

The ejection fraction describes the percentage of blood leaving the heart after each contraction. The regular amounts are percentages between 53% and 73%, most of the patients presented values in this range but those who did not, had a very high death chance.

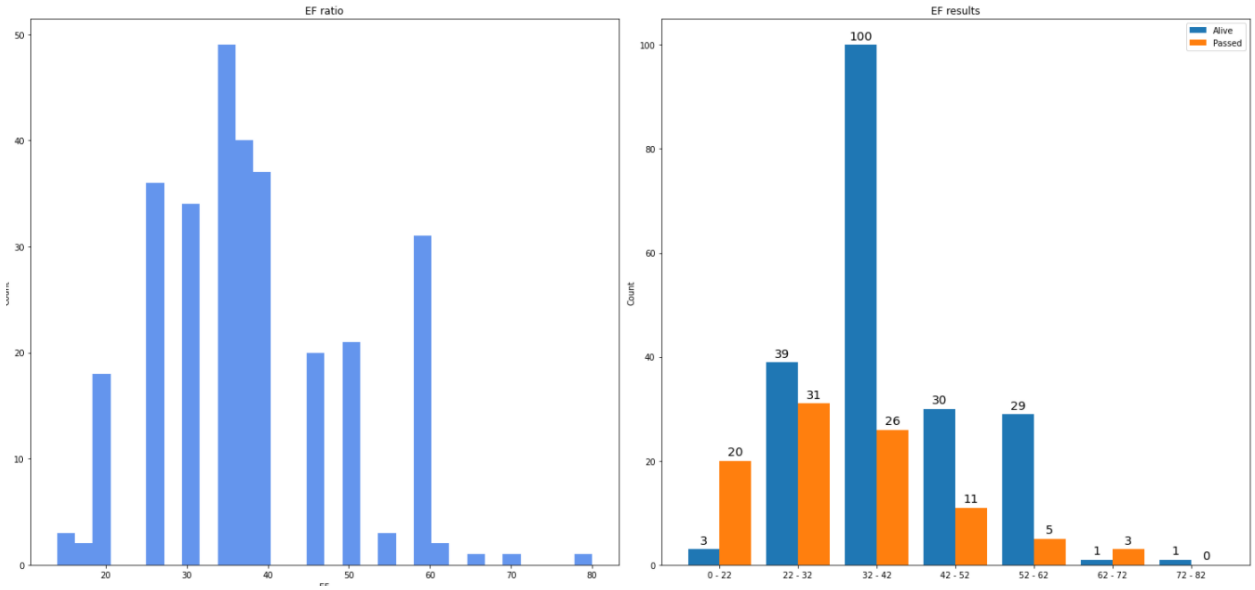


Figure 2.6 Ejection Fraction ratio

Talking about a High Blood Pressure, 35% of the patients presented this issue and the results that compare the chances of death are almost as twice when the value is existing than when the patient has a normal blood pressure. This tells us that HBP is a feature that persuades the death event when other factors apply.

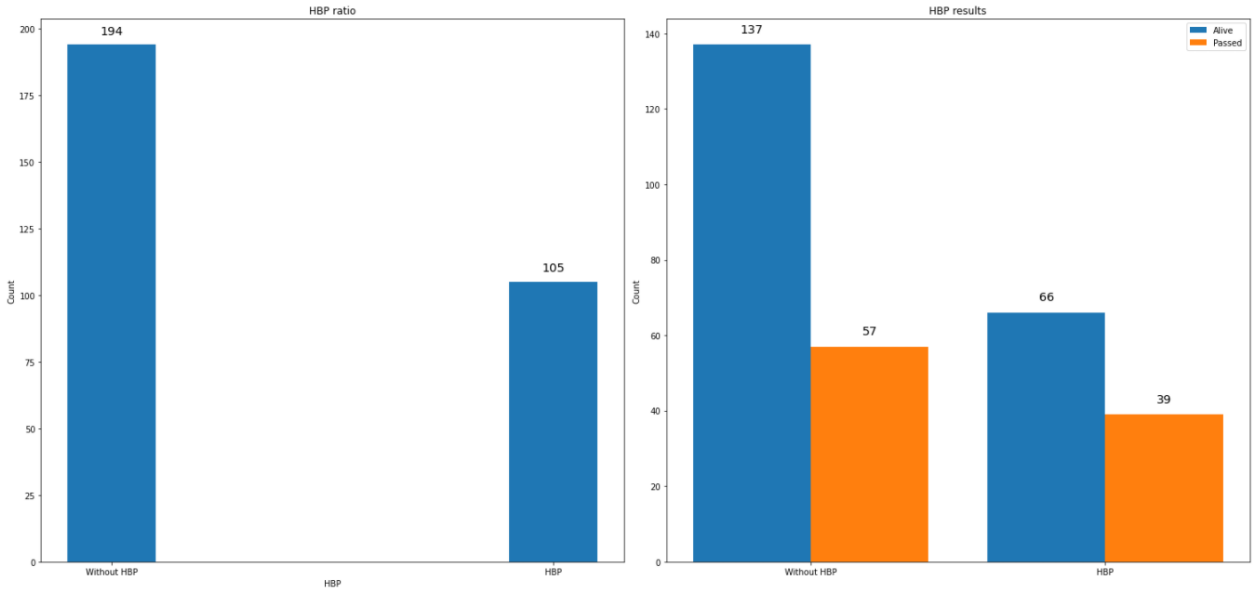


Figure 2.6 High Blood Pressure ratio

The seventh feature to be analyzed is the platelets amount. Platelets ar small cytoplasmatic fragments, whose work is to cycle in blood in order to help stop hemorrhagic processes. The healthy amount is anything between 150,000 and 400,000 mcL in blood, which most samples had and those who had a lesser amount had higher chances of death than those who did not.

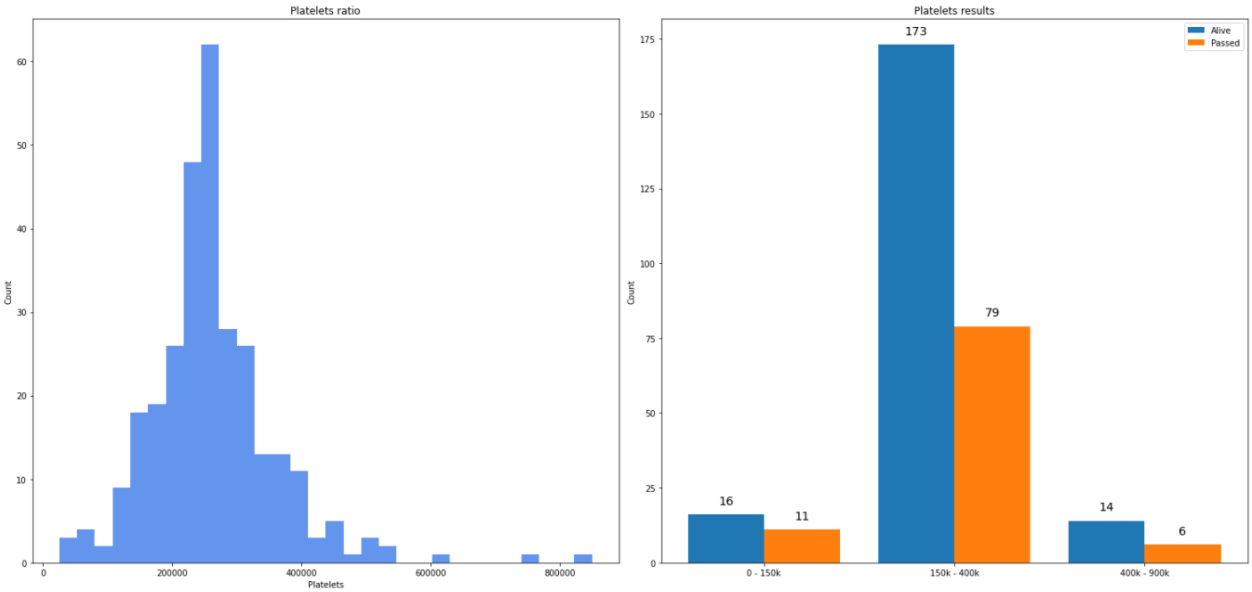


Figure 2.7 Platelets count ratio

Following platelets come the Serum Creatinine levels. Creatinine often generates waste product when a muscle breaks down, this is Serum Creatinine and doctors tend to use it to check kidney function, so if a patient has high levels of SC, it may indicate renal dysfunction. The regular amounts in the human blood for healthy kidneys is 0.9 to 1.3(mg/dL) for adult males and 0.6 to 1.1 (mg/dL) for adult females. Samples show that as the amount increases, and they leave the healthy amount, death event occurs more often, leading us to believe that this is a heavier feature when it comes to predicting the death probability.

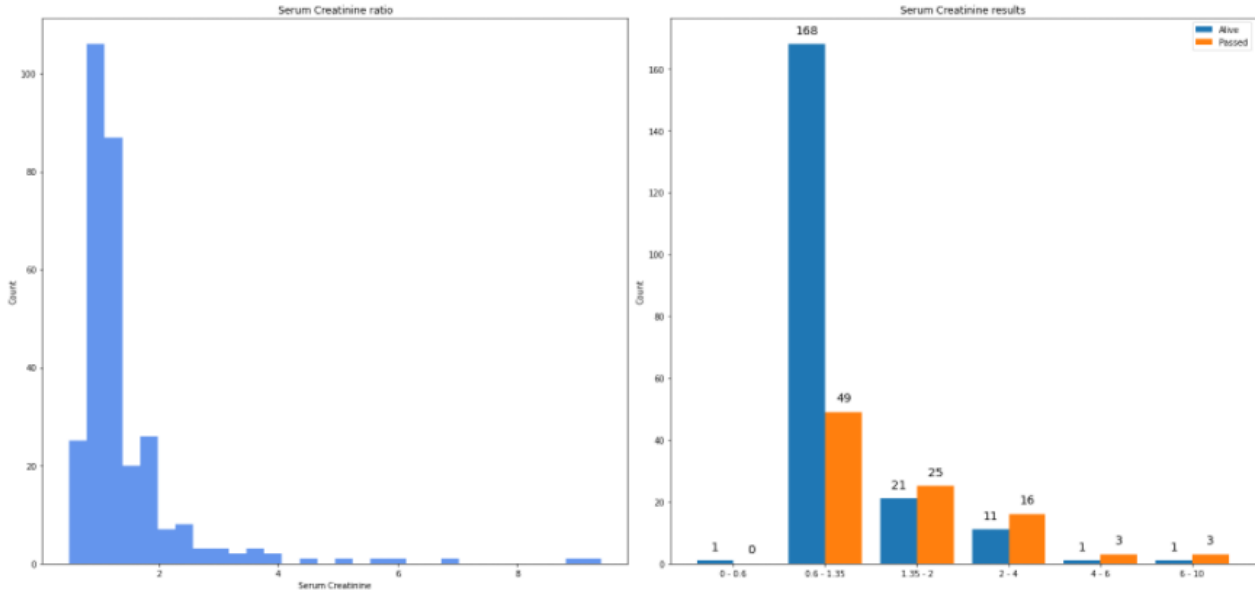


Figure 2.8 Serum Creatinine ratio

The Serum Sodium is another feature in this dataset, and it indicates the amount of sodium in blood, low levels might be caused by heart failure. Sodium helps maintain the right balance of fluids in the human body and strongly links to kidneys function. Normal levels are usually between 136 and 145 mmol/L, which most patients counted with, except for patients that with a higher death event chance, as the ratio was significantly different when comparing the regular levels with the lower levels.

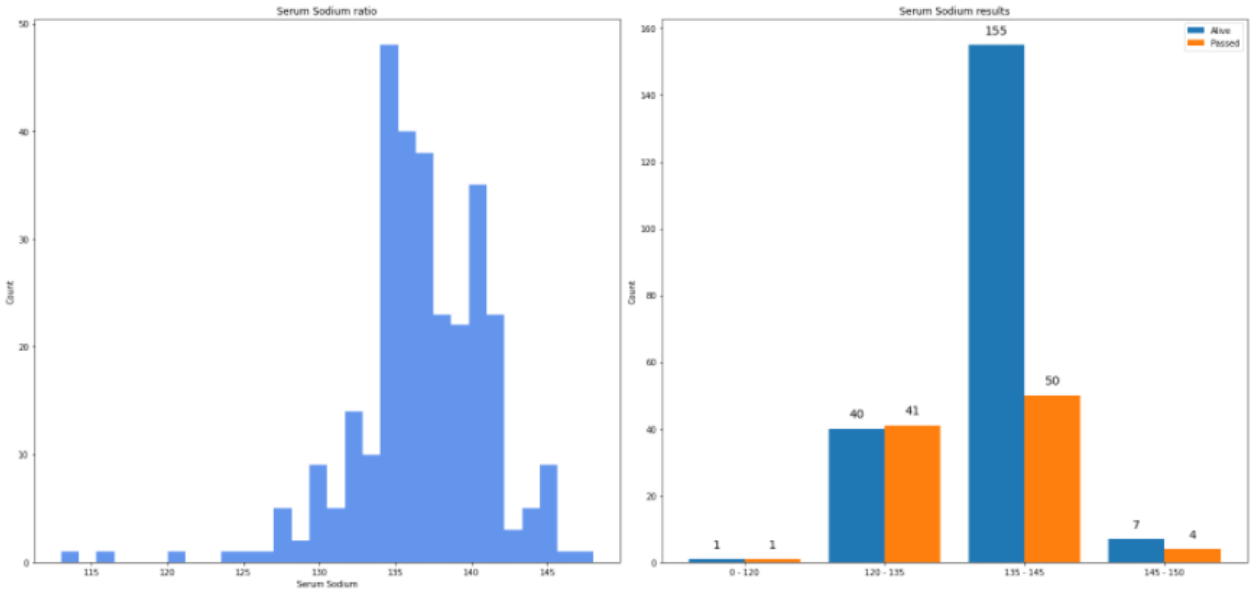


Figure 2.9 Serum Sodium ratio

Another feature asked to the patients was their sex, which in this case and after analyzing the results, did not have a significant impact as the ratio was around the 50% mark for both, female and male patients. The fact that 65% of the patients were male,

and 35% female, did not matter since the death events were equally distributed among them.

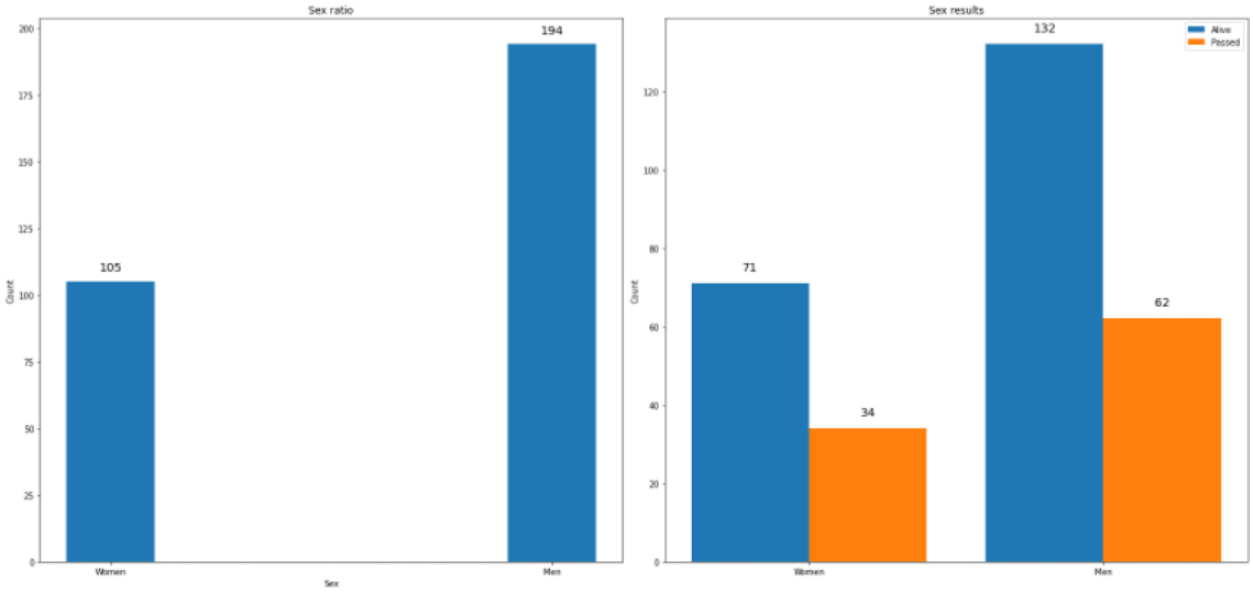


Figure 2.10 Sex ratio

The last binary feature is the smoke feature. This tells if the patient was a constant smoker or not, which would lead to lung damage and thus, oxygen distribution difficulties among other possible diseases.

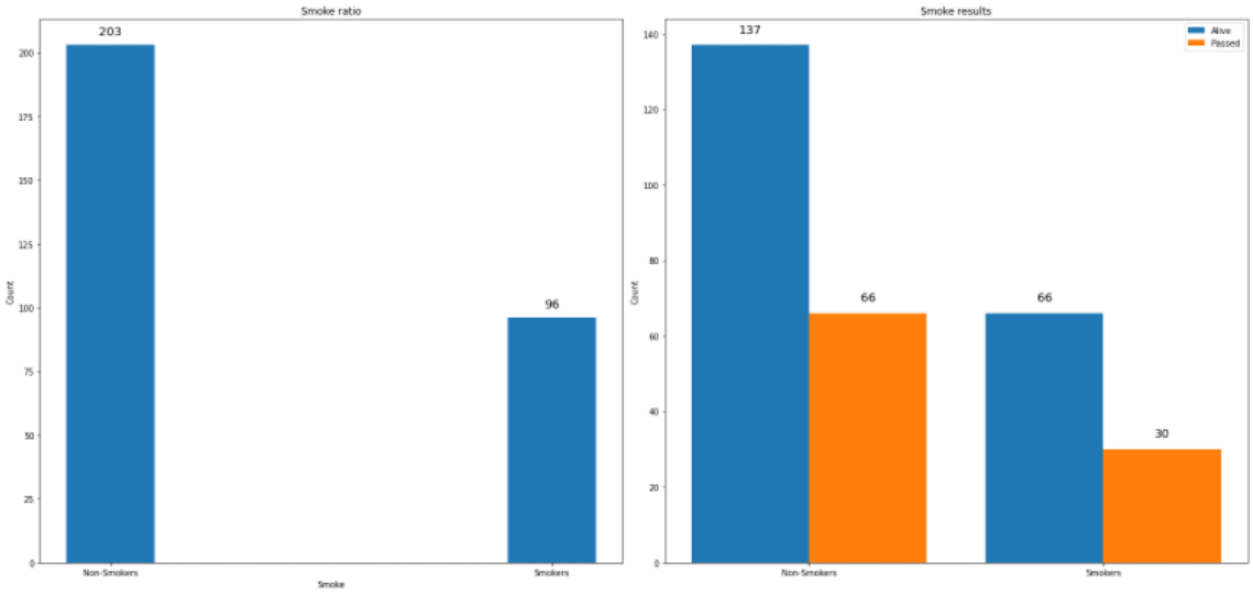


Figure 2.11 Smoke ratio

Finally, is the time feature, which shows the follow-up time after the heart failure occurred. This is the heaviest feature as the graph presents a decrease of deaths as time

passes, meaning that if a patient already survived 2 months after the heart failure, he or she is very likely to survive the rest of their days.

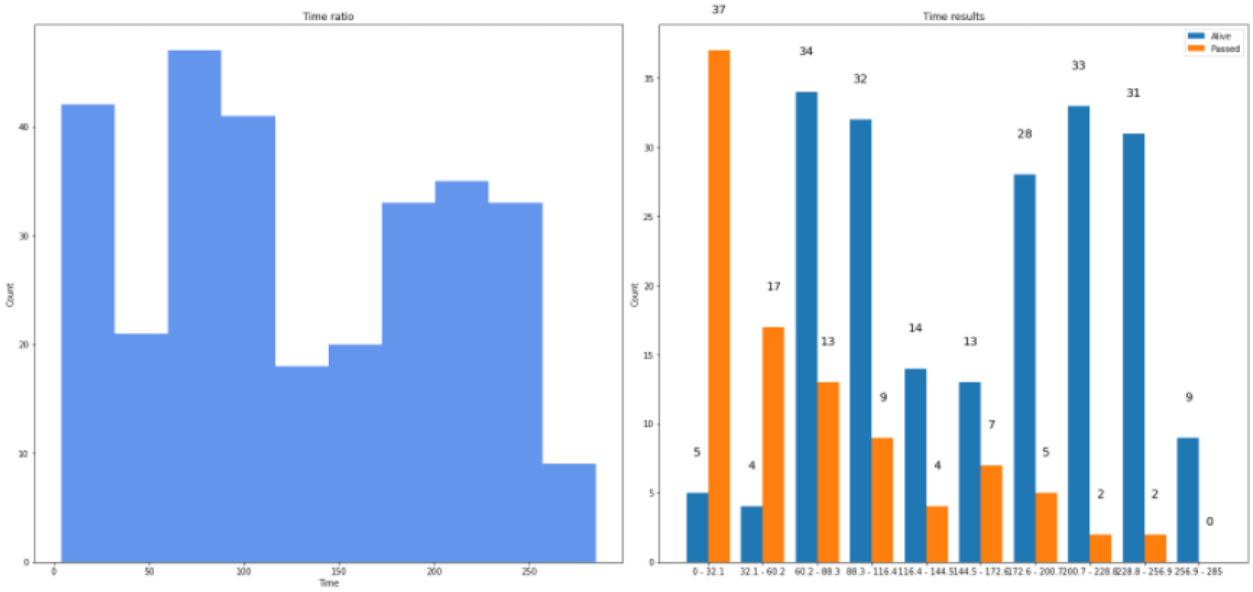


Figure 2.12 Follow-up Time ratio

Results

As previously stated, a Random Forest and a Decision Tree classifiers were implemented along this project and in this section the results will be discussed. It is important to know that for the data cleansing the public libraries *pandas* and *numpy* were needed, since they supply the required tools for the task. Now, starting with the Decision Tree Classifier's set-up, a loop was implemented to find the best values for the test size and the max depth, which would work together to fit the model in a very efficient way. Results showed that with a test size of 21% of the dataset and a max depth of 6 levels, the results would be 93.65% accurate according to the *sklearn* metrics. The result was the tree shown in the figure 3.1, displayed using *graphviz*.

Following with the Random Forest, the same test size values were implemented but as the Random Forest does not work with levels but with nodes, this value has to be grewater than 6 so it is not underfitted. The maximum leaf nodes were set to 13 and the forest consisted of 500 trees, leading to a result of a 95.23% accurate model. This is a very high percentage and is greater than the DTC result.

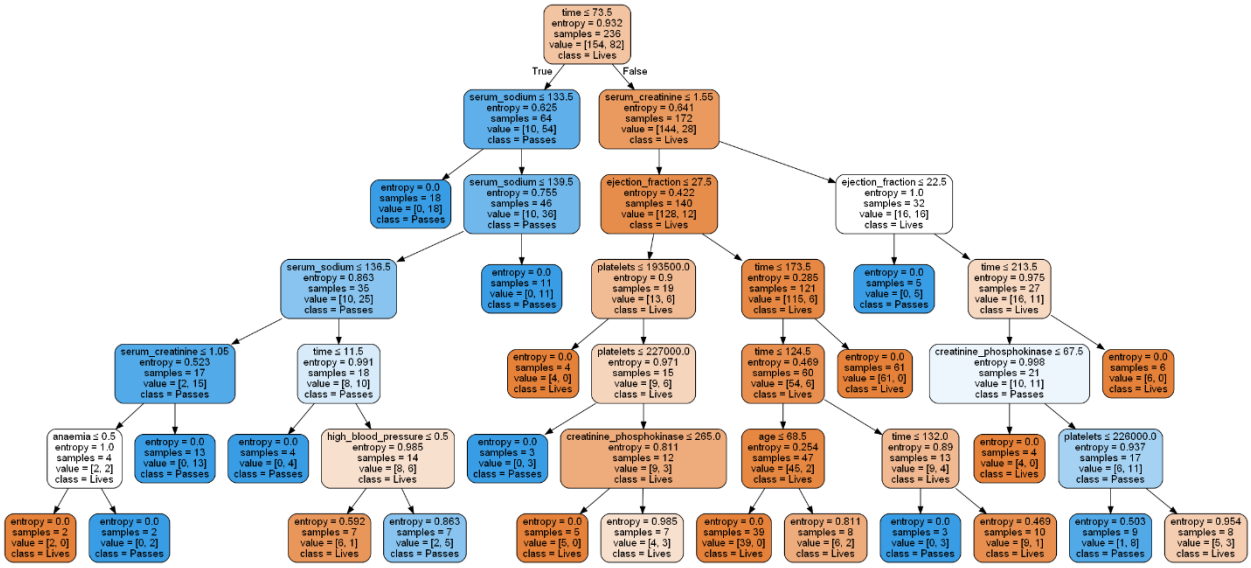


Figure 3.1 Decision Tree

Conclusion

Along this project, the Machine Learning Algorithms were implemented successfully and are able to predict the death event of a patient after being diagnosed with a Heart Failure. It is fascinating how even though the Decision Tree Classifier already had a big accuracy, the Random Forest was able to surpass it using the same dataset. It truly shows the difference between these 2 algorithms and it is very likely that comparing it to a linear regression algorithm the margin would be significantly larger, as these are the most powerful ways to solve it.

Even though we are on a early age of A.I. and this is a relatively small dataset to be accurate to every human being and be used as a official health service, we are truly in a right direction as human being to use computers to solve mathematical problems our brains would struggle with and I can only imagine what is ahead of us not only in a intangible way, but in a mechanical and physical way of implementing these new technologies.

References

Anemia - Síntomas y causas - Mayo Clinic. (2019, 14 diciembre). Mayo Clinic. https://www.mayoclinic.org/es-es/diseases-conditions/anemia/symptoms-causes/syc-20351360

Aujla, R. S. (2021, 20 abril). *Creatine Phosphokinase - StatPearls - NCBI Bookshelf*. NCBI. https://www.ncbi.nlm.nih.gov/books/NBK546624/

Chicco, D. (2020, 3 febrero). *Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone*. BMC Medical Informatics and Decision Making. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5

Diabetes. (2021). MedlinePlus. https://medlineplus.gov/spanish/diabetes.html#:%7E:text=La%20diabetes%20es%20una%20enfermedad,el%20cuerpo%20no%20produce%20insulina.

Sodium (Blood) - Health Encyclopedia - University of Rochester Medical Center. (2020). University of Rochester Medical Center. https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=sodium_blood#:%7E:text=Normal%20sodium%20levels%20are%20usually,are%20too%20high%20(hypernatremia).

Team, S. (2020, 18 diciembre). *Creatinine: Blood test, normal range, and how to lower levels*. The Checkup. https://www.singlecare.com/blog/creatinine-levels/#:%7E:text=In%20most%20cases%2C%20the%20normal,per%20deciliter%20for%20adult%20females.