BUT's 2nd year internship report

Determining the presence of heart disease in patients using random forest classification

From April 8th 2024 to June 14th 2024

Tutor: Professor Kazumasa Horie

Referent: Professor Cartier Sophie

Academic year: 2023-2024

By Martins Hugo

Abstract

Coronary Artery Disease is the most common heart condition, affecting over 18 million adults in the U.S. and having amounted to a total of 375,500 deaths in 2021, still, we are yet to be able to predict with 100% certainty if a certain person will have the heart disease or not. Certain conditions and lifestyle choices, known as risk factors, have naturally been identified in order to know what increases the likelihood of developping Coronary Artery Disease, however, the question only remains partially answered: some other conditions might not have been discovered yet, and the relevancy of the risk factors, the knowledge on how much they affect the development of the heart disease could also be lacking. For this internship, experiments using Machine Learning will be conducted to improve and reinforce the current knowledge of the heart disease, particularly in regards to the causes and symptoms of it. To achieve this, specific technology and methods will be used: random forest machine learning model using Python as the programming language through the Anaconda Distribution, and Visual Studio Code as the IDE. Various librairies related to machine learning and data visualization will also be used. This paper will be able to provide and reinforce knowledge on the causes of Coronary Artery Disease, as well as show the most important symptoms and sign of it; indeed, risks factors such as being an elderly male could be found in this research, which correlates with the existing knowledge.

Résumé

La coronaropathie est la maladie cardiaque la plus commune, touchant plus de 18 millions de personnes aux États-Unis et ayant résulté en 375 500 morts en 2021, cependant, nous ne sommes pas encore capable de prédire avec 100% de certitude si une personne va avoir cette maladie ou non. Certaines conditions et choix de vie, connus sous le nom de "risk factors", soit les facteurs de risque, ont naturellement étaient identifiés afin de savoir ce qui augmente la probabilité de développer la coronaropathie, cependant, la réponse reste partielle: certaine autres causes, conditions, n'ont peut-être pas encore étaient découvertes, et la pertinence des risk factors, la connaissance sur à quel point ils affectent le développement de cette maladie cardiaque pourrait la encore n'être que partielle. Pour ce stage, des expérimentations reposant sur le machine learning vont être conduites dans le but d'améliorer, de renforcer les connaissances actuelles sur la maladie, en particulier en ce qui concerne les causes et les symptômes. Afin de réaliser cela, des technologies spécifiques seront utilisées: random forest en tant que modèles de machine learning, en utilisant Python comme langage via la distribution Anaconda, avec Visual Studio Code comme IDE. Des librairies liées au machine learning et à la data visualization seront aussi utilisées. Ce rapport fournira et renforcera les connaissances sur les causes de la coronaropathie, et présentera les symptômes et signes de la maladie; en effet, les facteurs de risque comme être un homme âgé pourront être retrouvé dans cette recherche, ce qui corrèle avec les connaissances existantes

Acknowledgements

I would like to express my sincere gratitude to Professor Horie for guiding me throughout the project and for explaining and telling me about all the various aspects that I could work on for this subject. With his help, I was able to work on a Machine Learning subject while learning and discovering techniques related it. I would also like to thank Professor Aranha for all of his help preceding and during the internship, ensuring that each student would be able to conduct the internship at the University. Finally, I would like to thank both Professor Cartier and Professor Mery for giving to the student opportunities such as this one, and for the guidance and help, before and during the internship.

Table of contents

1 Introduction

In the medical field of expertise, Coronary Artery Disease is surely common knowledge and something that every cardiologist knows about. This heart disease is a blood vessel disease for which the major blood vessels that supply the heart struggle to provide to the heart muscle what it requires in order to function. It is the number one heart condition, and at the same time, it's also the number one killer of both men and women in the United States. With the prevalence of Coronary Artery Disease, it is expected to know what are the causes to it and how to be healthier, as to avoid this disease. While we know some causes, labeled as "risk factors", such as an old age, being a male, having high blood pressure, without forgetting a family history of heart disease, the fact remains that we only know some causes, or at the least, we cannot affirm that we know every single ones, and it is also hard to know what is the most important factors among the ones we know of. This is the main reason that lead to the beginning of this internship's subject: how can we improve and reinforce the already existing medical knowledge. To answer such question, we will be using Machine Learning techniques mainly for the following reasons:

- A Machine Learning algortihm will develop its own understanding of the data of the heart disease, this will offer another perspective which is important if we want to improve our knowledge
- Techniques emanating from Machine Learning will also be useful and directly related to the task at hand, such as displaying importance values graph representing a ranking of the features (such as the sex or age of a person) that affects the model the most.

To this end, we will be using Python through the Anaconda Distribution as our programming language, and work with Visual Studio Code as the IDE. Naturally, various Python librairies related to Machine Learning and Data Visualization will be used (such as Numpy, Pandas, Matplotlib, or scikit-learn among others. Throughout the internship, frequent meetings were carried out to ensure the quality and progress of the work, as well as to point towards future improvements. This report will be presenting in a first place the conditions of the internship, to thereafter chronologically present the work that was done and conclude on the project as a whole.

2 Internship's context

2.1 Integration at the University of Tsukuba

Located in Tsukuba Science City, and at sixty kilometers northeast of Tokyo, the University of Tsukuba was founded in 1973 with an aim to enable an open and new system of education and research in Japan, easily showed by the slogan of the university that we can find at many places on the campus: *Imagine the future*. The University of Tsukuba accepts students from over 100 countries and regions, and has around 16 500 students in total.

It has obtained different advantages and designations over the years, such as:

- the Designated National University Corporation: this designation signifies that the university shows a certain level of promise for world-class education and research activities. The way the university demonstrates those aspects is through the openness of the university, in which are mingled students and professor from various countries.
- the Top Type university status from the Top Global University Project of Japan. This means that they are conducting world-level education and research and have the potential to be ranked among the world's top 100 universities. This projects provide a funding for the university to lead internationalization of Japanese universities, which goes hand in hand with the university's philosophy.

On a more personal note, and as we can see on the picture above, bicycles are popular among students attending at the University of Tsukuba, and I also had the chance to rent one for the whole internship. This was a very practical and efficient way to cycle around the campus and even in the area around the campus, in which it is allowed to cycle on most sidewalks. Even though the meetings with Professor Horie were close to where I resided (at around 15min at a normal walking speed), it was still an advantage to use a bicycle.

I was thankfully able to get one after meeting with Professor Aranha on the 8th April. During this meeting, we were given reminders about the internship as well as other things related to our life at the University of Tsukuba, such as a personal card to use at the university's library to have some advantages. This is after this meeting that some of us had the opportunity to be accompanied by Japanese students to go rent a bicycle, and this is the reason that I was able to get mine.

2.2 Working environment

The realization of the project was done using my own personal laptop at my dormitory's apartment: as no particular location was designated for me to use, I decided to do the work from home. On the technological side, I used Anaconda as the package and environment manager for python (version 3.7), on which I installed various libraries useful for machine learning and deep learning, such as Tensorflow and Keras, without forgetting scikit-learn among others. The IDE I used to write, compile and execute code is Visual Studio Code, with Jupyter Notebook files, and all of this was executed on my laptop, through Windows 11.

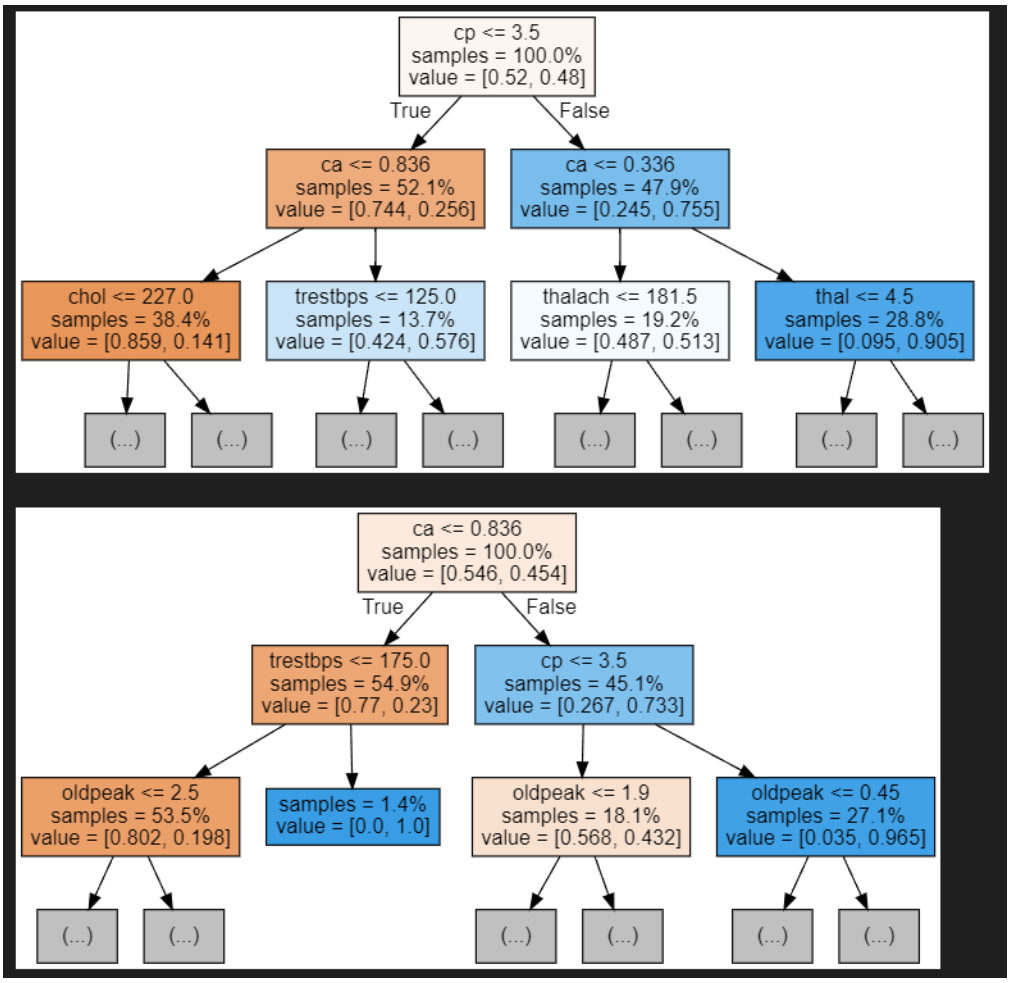
The way the work was done was the following: each Monday and Friday (albeit the days could slightly differ especially with holidays in mind), I would go meet with Professor Horie for one hour, during which we would talk about the progress that had been done and eventually the problems that had been encountered, to end the meeting on a new task to do before the next meeting. Basically, each meeting served as reviewing the work that had been done in the meantime and as giving the next step to go to in order to fully realize the project. Finally, the project realistically started after my first meeting with Professor Horie, which happened a Thursday afternoon, during which we discussed of what my assignment would be.

2.3 Assigned Project

To start off, I must mention that Professor Horie's researches deal with sensitive and private data (medical area) and as such, I unfortunately couldn't work on identical subjects. With that in mind, the mission I had was the following: choose an appropriate dataset from the UCI machine learning repository, to thereafter explore the different ways to create an appropriate model to accomplish the objective proper to the dataset.

For this internship, I decided to use the Heart Disease dataset: this dataset contains medical data about various patients, and we can notice a certain column, which is whether or not the patient has a heart disease or not. With this, I could understand that the goal would be to create a model that use every single pieces of data from the dataset apart from the column concerning the heart disease presence, to predict as accurately as possible if a particular patient has a heart disease.

To reformulate, the task was to create a model allowing to predict whether the patient has a heart disease or not as accurately as possible. To this end, the first model that was used was the random forest classifier, that I will talk about in details in the section corresponding to my work on the project. Nonetheless, to give a rapid overview and explanation of what it is, a random forest model in this context is, as the name indicates, a random forest composed of decision trees. In a task like classification, it uses each of those decision trees to predict the answer that we seek (in our case the presence of a heart disease in the patient) and the most recurrent prediction is chosen as the final prediction, as the answer. Below is an example of decision trees, directly taken from my work, that I will talk in further details in future parts of the report:


cp <= 3.5
samples = 100.0%
value = [0.52, 0.48]
True
False
ca <= 0.836
samples = 52.1%
value = [0.744, 0.256]
ca <= 0.336
samples = 47.9%
value = [0.245, 0.755]
chol <= 227.0
samples = 38.4%
value = [0.859, 0.141]
trestbps <= 125.0
samples = 13.7%
value = [0.424, 0.576]
thalach <= 181.5
samples = 19.2%
value = [0.487, 0.513]
thal <= 4.5
samples = 28.8%
value = [0.095, 0.905]
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
ca <= 0.836
samples = 100.0%
value = [0.546, 0.454]
True
False
trestbps <= 175.0
samples = 54.9%
value = [0.77, 0.23]
cp <= 3.5
samples = 45.1%
value = [0.267, 0.733]
oldpeak <= 2.5
samples = 53.5%
value = [0.802, 0.198]
samples = 1.4%
value = [0.0, 1.0]
oldpeak <= 1.9
samples = 18.1%
value = [0.568, 0.432]
oldpeak <= 0.45
samples = 27.1%
value = [0.035, 0.965]
(...)
(...)
(...)
(...)
(...)
(...)

3 Professional activity : 9 pages

3.1 General explanations on the tasks assigned

When first meeting with Professor Horie, our talk consisted in exploring the various type of tasks that could be done in an internship concerning machine learning and to try to choose what I would like to do for this internship, such as if there was a particular area I would like to explore and work on, or a specific task I would like to do. As I had no real precise idea and still had trouble to precisely distinguish certain aspects of machine learning, it was decided that my first mission would be to choose a dataset from the UCI ML Repository[1], and to create an appropriate random forest model. As I mentioned previously in the working environment, my future work conditions would be the following: realize the task, report the results and have a meeting on how to improve the current work and what others aspects, whether technical (such as the project itself from the programming side, or even the medical side of things) or theoretical (such as theoretical concepts, formulas) could be explored and talked about.

3.2 Datasets choices by browsing the content of the repository and comprehending it

The first task was quite straightforward, it simply consisted in choosing a dataset from the UCI ML Repository. I explored what the repository contained and chose a selection of 4 different dataset, in order to have a broader range of possible tasks to do. To be more specific, I choose datasets with missing values and without them, with few instances or labels and with many, with a goal of either classification or regression (the datasets consisted mostly in classification tasks). After this, I met with Professor Horie according to the schedule, and was told that it is possible to pick any of the ones I choose as each one could do for the internship. I decided to pick the Heart Disease Dataset for the following reasons:

1. The first reason is that if we were to use a scale of difficulty (how hard would it be to work on the dataset), the Heart Disease was probably the second most difficult out of the four, mostly because of the missing values, that the less difficult ones did not possess.

2. The second reason is that this dataset was more closely related to the researches of Professor Horie than the others, so I judged more fitting to do the internship's subject with this dataset and not the others, which were not related to the medical field.

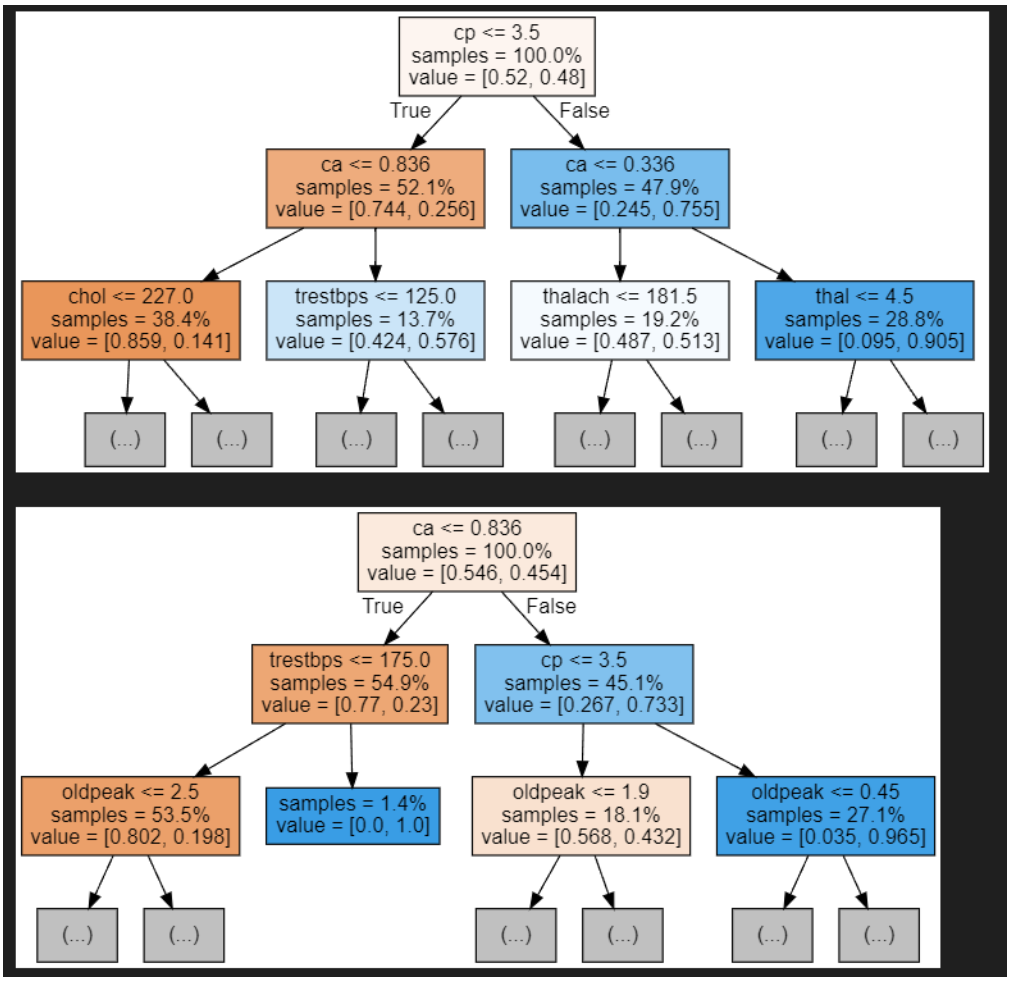
In the datasets, the instances were the patients, we can represent it like this: each row of the dataset, of the table, consists in a variety of information concerning a patient. In each row, these information, or to stay in the vocabulary of the domain, these features, represent medical information concerning the patient, with one particularly notable: whether the patient has a heart disease or not. The dataset was composed of 13 features (not including the presence of heart disease one), and 303 instances. This project was done on the processed Cleveland data file, and as the name indicates, the dataset was already processed so I did not need to turn the categorical values into numerical ones as this was already done. The missing values affected a total of 6 instances, spread across 2 columns.

3.3 Implementation of the random forest

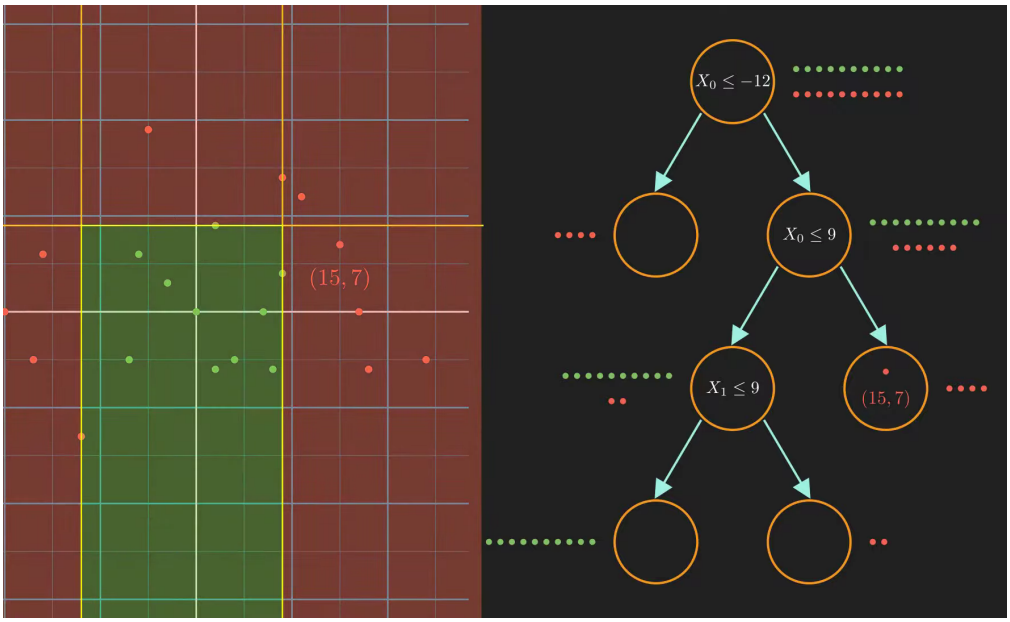
This was quite naturally the task following the selection of the dataset. To achieve it, I followed online tutorials on how to implement a random forest in Python. The core library to achieve this is *scikit-learn*, as this library proposes the *RandomForestClassifier* machine learning model; the library *imbalanced-learn* was also used along with the previous one to handle a problem that I will get back to later. The others libraries that were used in this work are the following: *pandas* and *numpy* for data processing, *matplotlib*, *seaborn* and finally *shap* for data visualization, or tasks related to understanding the model, notably for *shap*.

[1] University of California, Irvine, Machine Learning Repository

A random forest machine learning algorithm is composed of a multitude of trees. To be more specific, each tree is a *decision tree*, and the prediction of whether a patient has a heart disease or not can be illustrated as such: the random forest model asks each tree if there is a presence of a heart disease in the patient, and the results that the model return is simply the result that the decision trees returned the most. To be more precise, it is a majority vote using the values predicted by the decision trees, so if there is a majority of "1" predicted by the decision trees, then our model will predict "1". In my case, my main random forest model consists in 750 decision trees, with each tree being of depth 7 at most. The previously showed example of decision tree can make it easy to understand the structure:



With this, we can understand that a decision tree seeks to classify the instances in specific nodes by using conditions. For example, for the first node of the first tree in the figure above: the condition is whether the instance has a $cp <= 3.5$. By checking whether those conditions are respected or not, the decision tree can classify the instances. Here is an image that allowed me to understand more easily how the decision tree work:



It is essentially the same context, the part on the left could be considered as the instances of the model, the color of the dots as the zeroes and ones, and their position a combination of 2 of the instances (even though the model actually has 13 instances to work with). The part on the right represents the same thing as the previous figure.

3.4 Optimization of the random forest through a variety of approaches

During this meeting, I explained what I had done the previous time as well as showing the work itself. A variety of issues were discussed, those were the following:

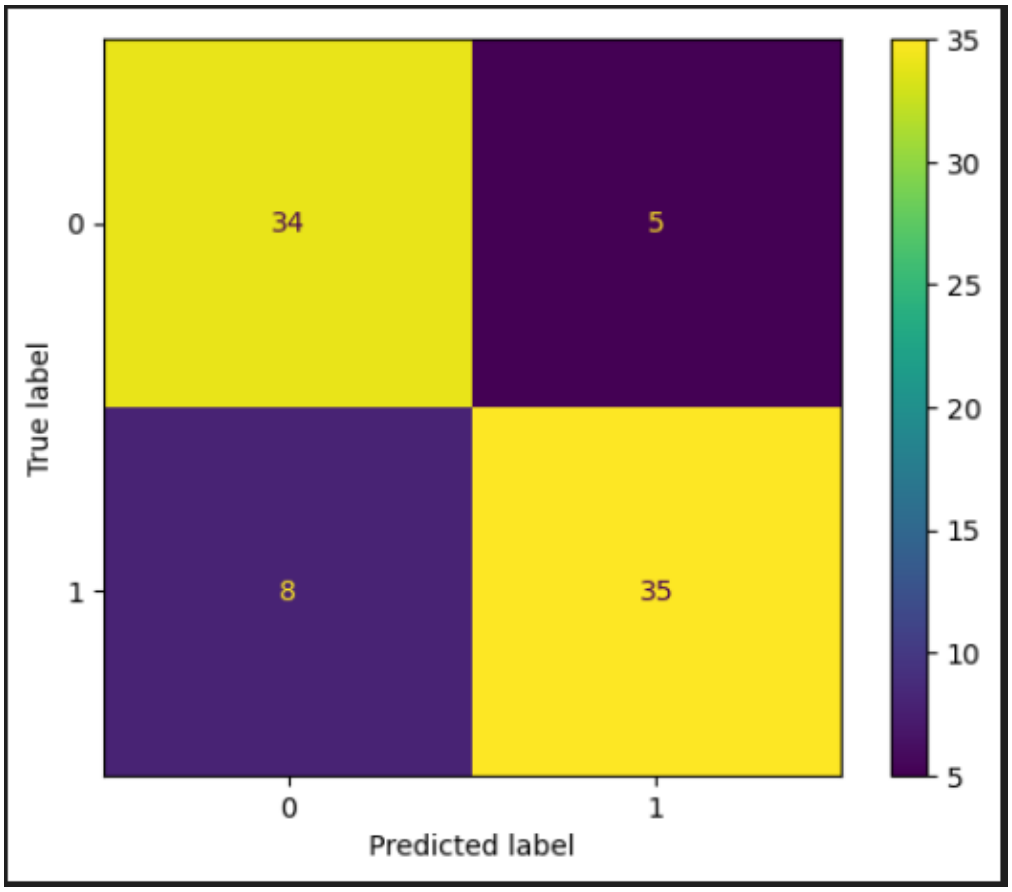
1. balancing the dataset
2. trying new ways to handle the missing values
3. looking up what cross-validation is

By balancing the dataset, what I mean is balancing the number of features to have an even number of features in which the patient has a heart disease and vice-versa. It is a very useful thing to do, as

> *With a greater imbalanced ratio, the decision function favors the class with the larger number of samples, usually referred as the majority class.*[2]

I could really understand this by looking at the confusion matrix of my model using an umbalanced dataset, as the zeros (signifying that the patient has no heart disease) were more present than the ones, and I could directly see that the model failed more when trying to predict the ones. Because there were more zeros overall, the model predicted more often zero and was more often wrong when trying to predict a one, which is a patient with a heart disease, so a very typical Confusion matrix was like the following:

[2]Directly taken from the official site of the *imbalanced-learn* library, `https://imbalanced-learn.org/stable/introduction.html`



On the figure above, we can see that the ones are indeed quite often predicted as 0; and in this case, even though there are more ones overall (43 ones compared to 39 zeroes, meaning that it is slightly umbalanced). We can see that the model doesn't even favor the majority class, so others factors must have an influence on the over-prediction of zeroes. To balance the dataset, I used *imbalanced-learn*'s functions to create an oversampler and an undersampler, their goal being to balance the dataset by either duplicating rows, or removing rows, to end up with a model that has a same amount of rows with healthy and unhealthy patient. When doing this, I had to take care not to have duplicates in both the training and the testing data, as otherwise, the model would never fail for this specific row, affecting the result in a really bad way.

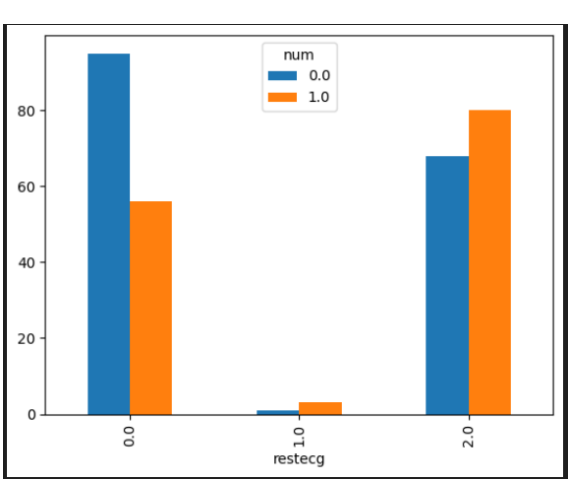
A second point was the handling of the missing values. Previously, I handled the missing values by putting the median value of the column. This way of replacing the missing values is not always the best option as in the end, if the value should not have been written as the median value, this would lead to mistakes done by the model. The way I changed this handling of missing values is by using models, random forests once again, to predict those missing values, similarly to what I had done for the main model that predicts if a patient is healthy or not. This way of managing the missing values, even if not always correct, would still be a better option than simply hoping for the median to be the missing value.

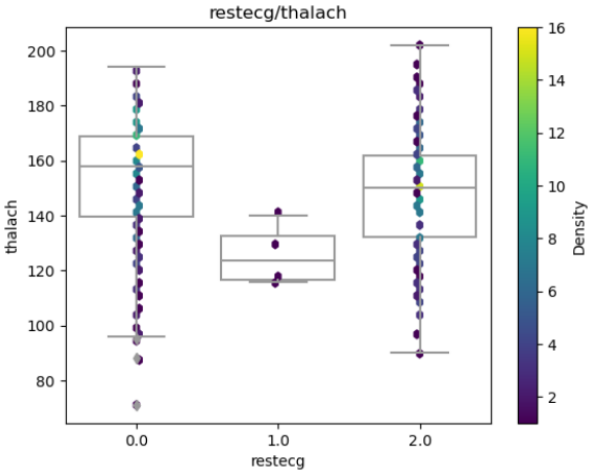
3.5 Observations concerning the dataset features and data visualization

Having now obtained better results and a theoretically (and also realistically) better model, the time came to understand why some features are more important than others and their direct relationship with the *num* feature, symbolizing the presence of heart disease. During the meeting concerning this new step, Professor Horie saw the features importance graph of my model[3], and a particularly interesting point could be found

[3]the *features importance graph* will be an important part of the project. The graph in question will be provided in the next section; it shows the importance of features for the ML model

within the bar diagram (among others that we will explore later). The "restecg" column did not affect much the results of the model according to the diagram, which was rather surprising as this feature consists in the *resting electrocardiographic results*, so it should have a direct link with the heart disease. As mentioned just before, I worked on visualizing features and their relationship with the *num* feature, but also did the same for the *restecg* feature and every others. To do this, I used libraries such as *seaborn* and *matplotlib* to visualize the data by creating simple bar diagrams when comparing a categorical feature to another of the same kind, and also fabricated slightly more complex graphs by combining boxplots and hexbin for comparing categorical features to numerical features; here are examples:

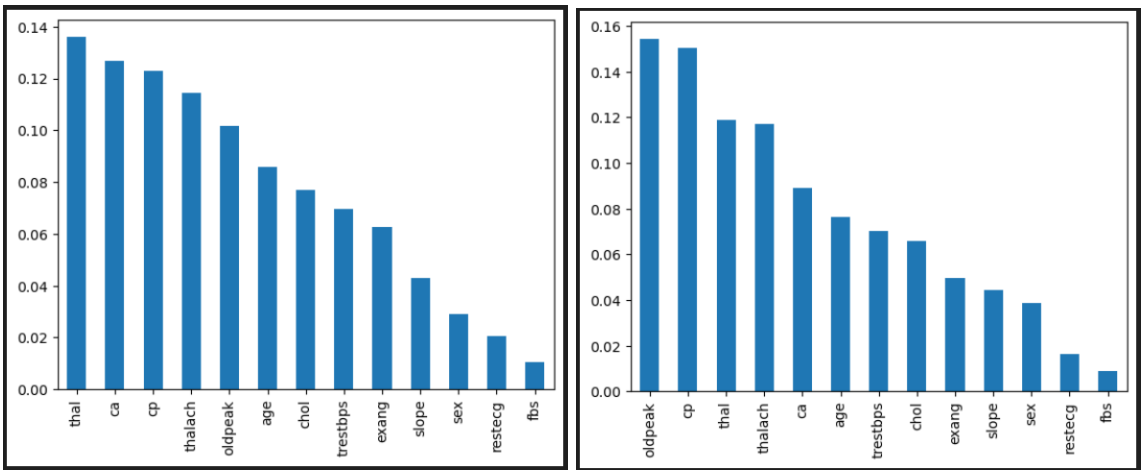




For the figures above, we can understand different things: firstly, the sample size for $restecg = 1$ is too low to be considered relevant, as there are approximately only 4 samples for this category. An other point is that besides this rather useless data, there is a low difference when comparing $restecg = 0$ and $restecg = 2$ with num, and there are basically no differences between $restecg = 0$ and $restecg = 2$ when comparing it to the *thalach* feature for the second figure.

3.6 External researches concerning the dataset and incremental/decremental approach to creating the model

The meeting that preceded this step of improvement of the model and of understanding of the features consisted in showing the work that I previously talked about. The observation that could be made is that indeed, *restecg* had a low direct relationship with the *num* feature, and that it also had only a few direct relationships with other features, as well as the direct relationship being not as important as some others. With those observations in mind, I was tasked to learn more about what the heart disease precisely is, because learning the actual name of the disease could allow us to actually search and understand from medical sources what the causes of it are, which would lead to understand why some features are more important than others. My other task was to use an incremental and decremental approach to create the random forest model, to see if manually adding/removing the features one by one could lead to a different order of features importance. What I must absolutely provide before showing and explaining the results is the features importance graph of the model:



What is essential to understand here is that there is no particular feature that helps showing in a noticeably better way than the others if a patient has a heart disease. Because of this, slight changes in the order had to be expected, as shown as the two different graphs; changes are caused by the fact that each time the code was executed, the random forest had the same parameters but the data passed in training and testing was different, hence the variation in graphs. For the figures above, we can see that the order for the first greatly differ than the one for the second: for the first figure, the feature *oldpeak* is at the fifth position while for the second it's at the first. What we should note however is that the most frequent graph that I encountered is the one from the first figure, most of the graphs that were generated throughout this project led to graphs similar to it, so we should consider graphs like the second figure to be the less frequently occurring and thus less representative ones. What we should really keep in my mind is that the graph was always slightly different.

The incremental and decremental approach consists in creating every possible models with respectively a single features, and every features but one. At each step, we add/remove once again a single feature to the best model found, until there are no features left. To ensure that I always picked the actual best model (and did not made any mistake, caused by the randomness of the splitting of the data), the best way I found to do so was to test the models multiple times using loops, and to try to pick the model that showed the best combination of high accuracy score (for both the average cross-validation score and the actual model average accuracy score) and low difference score (by difference score, I mean the difference between the split with the highest accuracy score and the one with the lowest, in cross-validation).

To explain in the context of the work, for the incremental approach, I created 13 models for each features to test. For each model, I tested the cross-validation scores and differences a certain number of times, then tested the accuracy of the actual model, once again a certain number of times, by scrambling the data to train and test the model on. This was done to see the actual variation and mean accuracy of the model; by variation, I mean the difference like the one for the cross-validation. Once I found the best model, I would keep it as the base for the new model, and create 12 new models already containing this model's feature + one of the 12 features left.

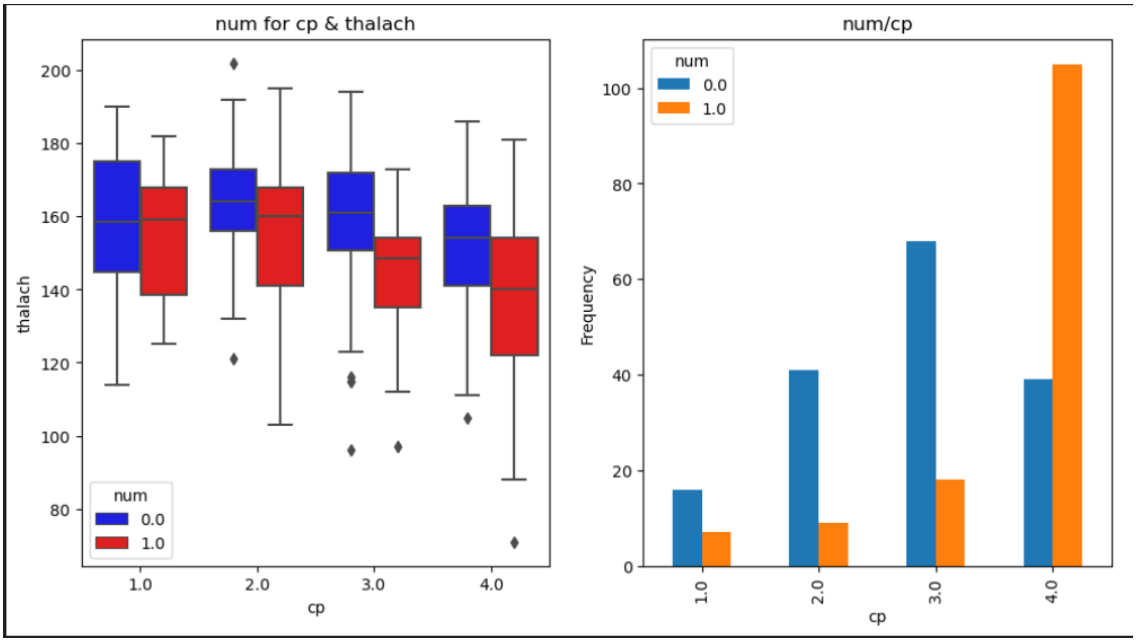
Through this scrambling (which is simply calling a function that randomly splits the dataset into data to train and test the model on), I could ensure that the accuracy results of the models would be for each iterations different, which helps finding a good model. Indeed, it means that an actually good seeming model would have more chance to be realistically good: if I did not do that, sometimes, a very good model could show up with high accuracy, but if I scrambled the data once, it could show a terrible accuracy score, which would mean that the accuracy of the model obtained before scrambling the data was obtained through sheer luck of a beneficial data split (into training and testing data). This is why testing the model and scrambling the data multiple times is useful: to avoid a model that showed good/bad accuracy by pure coincidence.

3.7 Visualization of specific features and understanding features relationship

From the previous task, a relatively new and more solid order of features importance could be obtained, particularly for the 6 most important values, which were in the almost same order for the incremental and decremental approach; for the decremental approach, the order was $thal > ca > cp > slope > sex > age > exang$. One thing to note is that the position of *exang* was at the third place instead of sixth for the incremental approach, and that the features after the 6 most important were not in any particularly reliable order.

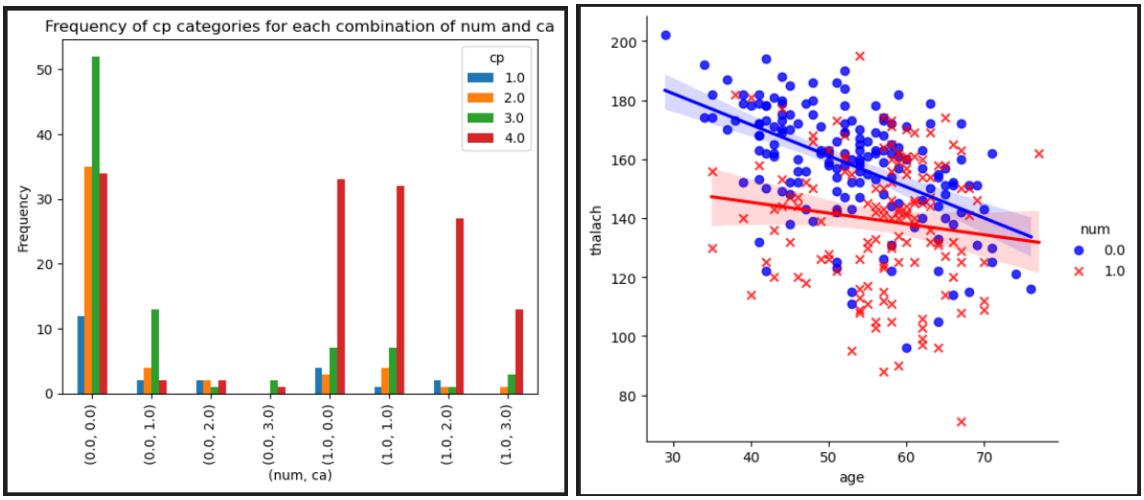
With this new order, interesting differences could be found from the previous graph: indeed, features like sex and exang grew to be more important, while thalach and oldepeak grew to become less important. The new goal from there on would be to visualize the 6 most important values and their relationship with other features in accordance with *num*, as well as look at why thalach is not considered important.

For both part, we can observe the following graph to understand that our dataset shows clear conditions on whether a patient has a heart disease or not:



The graph on the left shows the relationship between cp, thalach and num. By comparing cp to thalach, the aim was to understand why cp would be preferred to thalach in the importance order. The graph on the right is relevant to understand that for example, the sample size for $cp = 1 and num = 1$ is a quite low, compared to $cp = 4 and num = 1$. This indicates that there are more patients with a heart disease if $cp = 4$, but more importantly that the corresponding box for the other figure might not be a proper representation of thalach for $cp = 1$, as the sample size is very low.

Here are some other graphs to consider, for the first and second part respectively:



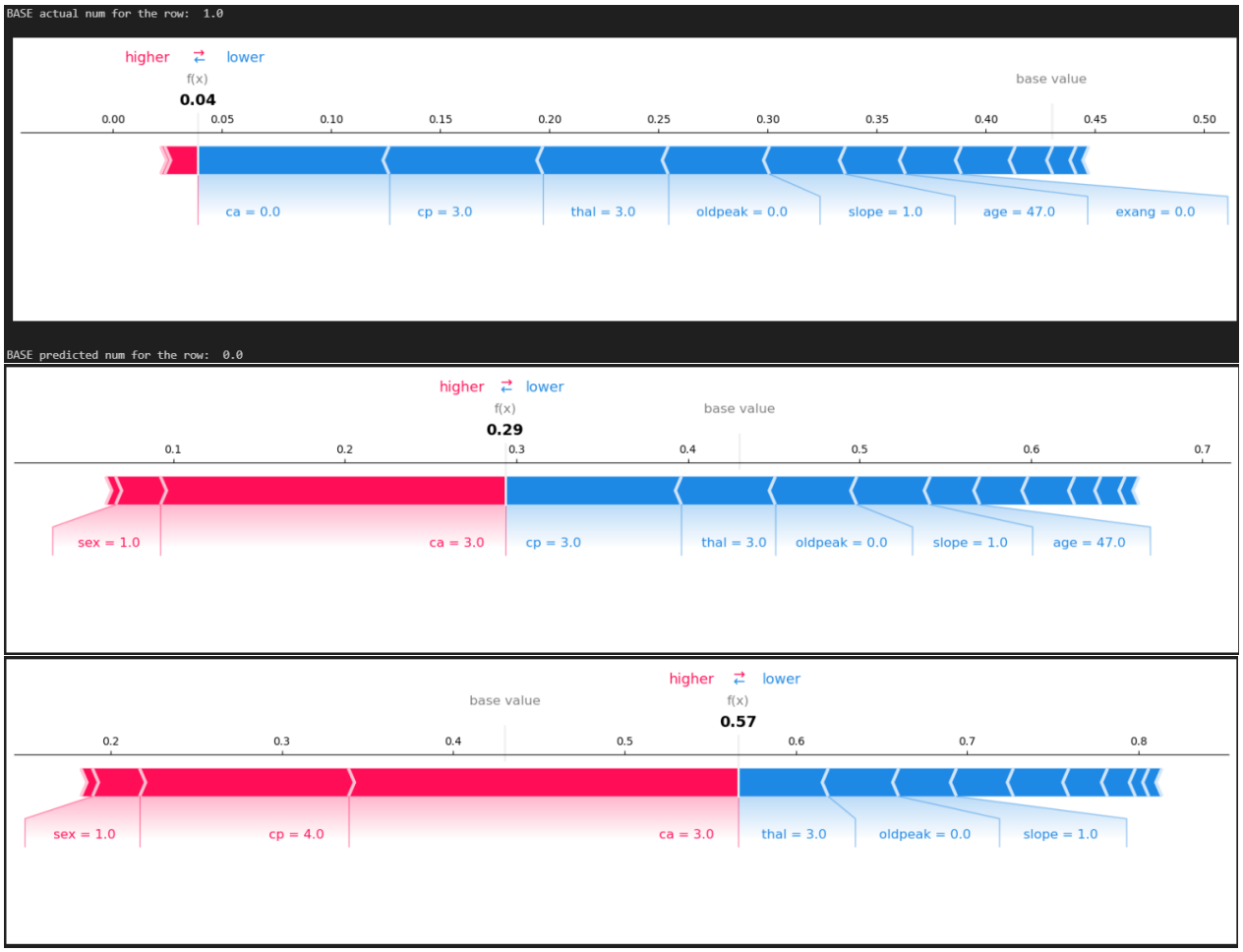
For the first graph, we understand that by avoiding $cp = 4$ and having $ca = 0$, there are very high chances for the patient to be healthy. The extremely clear conditions for a patient to be healthy or not, as opposed to the previously displayed graph comparing the almost non-existent relationship between the fbs and num feautres, are most definitely the reason as to why cp and ca will be highly ranked in the feature importance graph provided by SHAP, which we will see in the next section. For the second graph, we can observe that there is a high correlation between the age of a person a the thalach score. The more a person ages, the less their thalach score will tend to be. This could be a reason of why adding thalach (for the incremental approach) would not particularly improve the model if it already comported the age feature.

3.8 Observing the impacts of changing the value of specific features, and usage of SHAP values

Now that we could get some kind of better understanding of the features, time came to see if we actually properly understood it by directly confronting it to other kinds of test and by using new tools. For the first part, I was tasked to see how cp and ca would affect the prediction made by the model. If I could indeed understand the graph comparing cp categories to ca and num categories, then I should be able to use this knowledge to affect the result of the model. What I did is the following:

1. Use the random forest model to make a prediction for a specific row
2. Manually change the cp value, then the ca value to see how it affects the prediction of the model

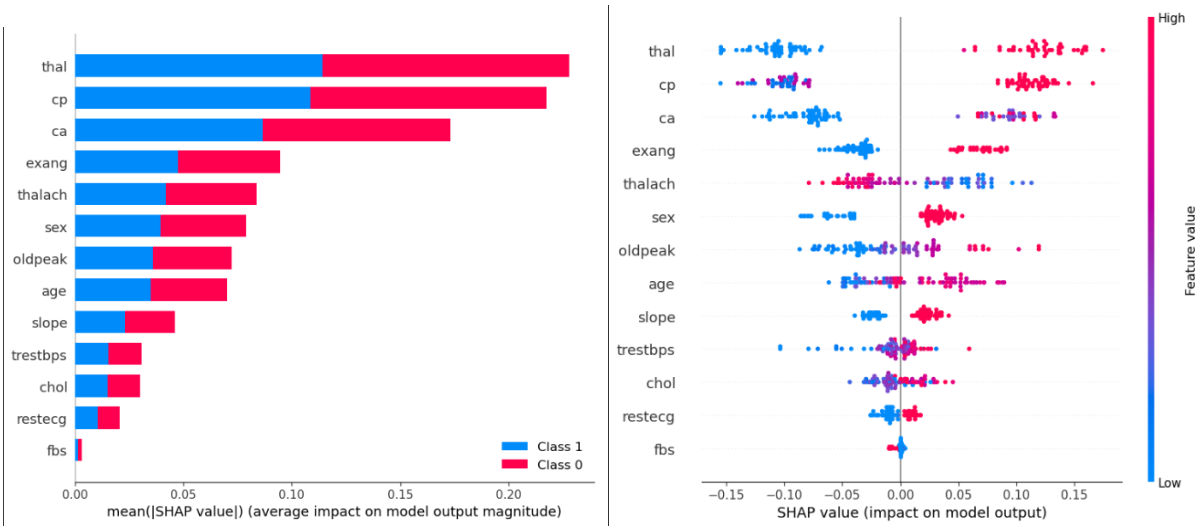
This step actually served two different purposes: ensure that the knowledge obtained from the graph was relevant, as well as see how changing the values of the features would lead to the model predicting the patient as healthy or not, which is still in alignment with the goal of understanding the causes leading to the Coronary Artery Disease. Here are graphs obtained through the shap library, displaying the result of the aforementioned experience:



Let's now break down the meaning of those 3 graphs; the first one shows that our model predicts $num = 0$, with a force of 0.96 (shap values go from 0 to 1). The closer the shap value is to 0, the more the model will predict the instance as to have $num = 0$. What is actually displayed is that the force for $num = 1$ is of 0.04, so it won't predict $num = 1$ as the force is below 0.5.

The second and third graph shows how the prediction of the model is impacted, going increasingly up to a force of 0.57 for $num = 1$ after changing respectively ca to 3 and cp to 4 (the worst values possible, as understood from the graph that we previously talked about). With this experiment, we could indeed directly observe the impact of those two features on the model's prediction, being capable to completely turn around the prediction if both were modified.

Now comes the time to display the importance values graph obtained once again with the shap library, here showed with two variants of the *summamry plot*:



What we can observe is pretty clear for the first graph: there are 3 really important features in the model, those being thal, cp and ca, while there is one that shows a really low importance, this one being fbs. Class 0 and Class 1 represent patient without a heart disease and with it respectively.

As for the second graph, it shows with the same importance order, how each instance (represented by the dots) impacts the shap value based on the value of the feature. For example, and similarly to the three graphs above, if we take for example a low value for ca such as $ca = 0$ (the values of ca ranges from 0 to 3), we can see that it leads to a lower shap value (around -0.12), so a bigger force towards the prediction of $num = 0$, meaning a healthy patient. The bigger the gap in the middle, the more the feature affects the shap value, so the model's prediction. We can see that thal has the biggest gap in the middle as well as the biggest impact on the shap values, which means that there is a clear distinction between the thal value of a patient with and without the heart disease, while also meaning that the value, whatever it is, will be highly important in determining the presence of the disease. Overall, it can be seen as the complete opposite of a feature like fbs, with no gap and no impact.

4 Sustainable Development and Social Responsibility

As we read on the website, the University of Tsukuba *sets environmental goals for chemical safety management, energy saving, resource saving, recycling, and green purchasing, and strives to achieve them.* In other words, the University is implicated in sustainable development. It can easily be observed by the following facts:

- In 1977, a specialized graduate school focusing on environmental education was founded at the University, which permitted to undertake the master's program in Environmental Sciences or it's continuation for the doctoral program. In a nutshell, the program aims is to nurture talents who are keen on finding solutions to current and future environmental problems. The course ranges from a *fundamental understanding of the natural environment to a comprehensive evaluation of human impacts on the environment with corresponding solutions*, which allows us to see the extent of the program, with the scales of subjects going from cellular level on life activities up to global scale.
- Another course about similar subjects is the master's and doctoral program in Biology. It comports 8 research fields, such as Ecology, Plant development & physiology, or Animal development & physiology. Especially for the ecology field, we can easily see how it would lead to an environment respecting Sustainable Development and Social Responsibility norms and practices

5 Conclusion

What are the signs that someone has Coronary Artery Disease, and which ones are the most relevant ? How to ensure that we are able to stay healthy, or if that's not the case, how to recover from a heart disease and treat it ? This paper tried to answer the following questions by using Machine Learning methodology and techniques. Indeed, from the implementation of the Random Forest, to the usage of SHapley Additive exPlanations, without forgetting the most useful Data Visualization graphs that were generated, we could observe the understanding of a machine learning model towards the Coronary Artery Disease heart condition, and from this emerged the possibilities to aquire either new-found knowledge or reinforce the already existing one.

To start off, various observations and comparisons can be done with the results that have been obtained throughout the project, so I would like to first list off the correlations, based on the book *Be Heart Smart: Understand, Treat and Prevent Coronary Heart Disease (CHD) by Waqar Khan*:

- Males are more likely to develop the heart disease; this piece of knowledge can easily be observed with the graphical figures displayed in the SHAP values section
- An older age also goes hand in hand with increasing the chance for the heart disease to form, and it is also related to the maxium heart rate achievable (represented by the thalach feature in the dataset)

Here are peculiarities and intriguing aspects that could be observed, as well as obsevations on the importance of certain features:

- While ca (the number of major vessels, from 0 to 3) and thal (Thalassemia blood disorder) are not as frequently mentionned as other risk factors, they were two of the three most important features in the model (the last one, Chest Pain, is more of a symptom and is frequently mentionned)
- We could also observe the total inefficiency of the fbs feature, having an extremely low importance despite being one of the fourteen features preselected in the dataset (that originally comported up to 76 features, depending on the clinic from which the data was obtained)

With those results in mind, we could answer to a certain extent some of the questions that we had at the beginning; however, some improvements could still be made to improve the results obtained and try to better answer the questions. Indeed, problems such as the low sample size of certain categories of certain features, as well as the low overall instances amount and the fact that all of this data came from a single clinic and not multiple ones, are all reasons to why the answers could be relatively innacurate or at least, not as accurate as it could be. As such, working on improving those aspects could be a direct continuation of the work that has been done in order to improve it. Another aspect that could be improved upon is the usage of different Machine Learning techniques, or a most extensive use of them. For example, doing similar experiences but using a different Machine Learning Model could be a way to challenge the results obtained to ensure that they are reliable and observe the differences between the models.

Appendix

hugo.martins5733@gmail.com
GRADIGNAN (33)
07 61 72 36 15

Martins Hugo

OBJECTIF

DEVELOPPEUR MACHINE LEARNING

PROFIL

Actuellement étudiant en 2ème année de BUT informatique à l'IUT de Bordeaux.
Durant ces 2 années, mon sérieux, mon engagement et mon adaptabilité ont pu être mis en valeur dans les divers projets auxquels j'ai participé, comme la création d'un site web avec requêtes SQL via Doctrine en utilisant le framework Symfony, ou encore la création d'une application C# avec requêtes sur une base de données via des scripts Python.

COMPÉTENCES

- Langages Informatiques (Python, Java, C#, SQL, HTML CSS, PHP, Javascript)
- gestion de projet informatique (méthode agile, cycle scrum)
- Utilisation de frameworks et librairies (NumPy, TensorFlow et autres pour Python, Symfony et React pour le développement web)

HOBBIES

- Langues et cultures étrangères (Japon, pays slaves)
- Art digital (tablette graphique, logiciel de dessin *Krita*)

PARCOURS

En cours 2022-2025

BUT Informatique

I.U.T INFORMATIQUE DE BORDEAUX

- Parcours A : réalisation d'applications - conception, développement, validation
- Parcours international (anglais)

2019-2022

Baccalauréat général - mention bien

LYCÉE GÉNÉRAL ET TECHNOLOGIQUE JEAN-MOULIN

- Section européenne
- Spécialités maths et informatique

PROJETS

Janvier 2024

DÉVELOPPEMENT D'UN SITE WEB

- Réalisation d'une application full-stack de suivi, notation et critiques de séries avec consultation des avis, notes et profils des utilisateurs
- Méthode agile de gestion de projet (Scrum), et utilisation de Git

STAGE INFORMATIQUE

avril 2024 à juin 2024

JAPON - UNIVERSITÉ DE TSUKUBA

- Machine learning - Random Forest model
- Étude des features du dataset - data visualization et SHAP values

EXPERIENCE

juin 2023 à juillet 2023

CHATEAU DU HAYOT

- Travail en autonomie et en équipe