
# Report for the Second Assignment (2nd Part)

## Data Processes

Academic year: 2022/23

Víctor Morcuende Castell

Guillermo Nájera Lavid

Javier Rocamora García

Antonio Ruiz García

# TABLE OF CONTENTS

# 1. Introduction

The purpose of this project is to **analyze the cost** of a national healthcare system regarding the expenses **associated to heart attacks**/diseases. To carry out this task, several guidelines will be followed: first, the team will **create a classification model** to properly distinguish between subjects with heart attack risk and those without. Then, we will **analyze the specific cases** (healthy/sick subjects, subjects that accept the proposed plan and adhere to it or not…) that make up the whole cost analysis. Finally, the team will **determine the percentage of adhere** necessary to reduce the national healthcare system cost to a specific amount, which is the goal of this problem.

# 2. Model Selection and Evaluation

For the **classification model**, the team thought it would be beneficial to **start with something simple** and expand the architecture the more the needs of the classification problem grew. That is why we started by building a simple **Neural Network classifier with 3 layers**, two 'Dense' layers with the 'relu' activation function as parameter and one third with the 'sigmoid' activation function, which process the output neuron into a class probability, instead of outputting a binary result, which we thought could be of use later in the assignment.

After training our first-generation Neural Network, we evaluated it with our test dataset and were given promising results at first, with an **accuracy value** in the test dataset **of around 90%**. However, in order to evaluate how the model behaved in terms of bias, we built and plotted the confusion matrix of the test set predictions, comparing them to the true labels, and we encountered what we had been suspecting. **The model was extremely biased**, meaning that since the dataset we were working on had way too more healthy subjects than sick ones, the classifier tended to bias its classifications to the healthy class, predicting as healthy not only the healthy subjects, but also most of the sick ones, giving as a result high overall accuracy, but also utterly high bias.

For the **second model**, we tested the idea of using a **decision tree classifier**, since we investigated that it would bring beneficial results for the bias of the model. The decision tree classifier did indeed offer **better results in terms of bias**, but not very promising, and at the expense of **lowering the overall accuracy** of the model to an **85%**. Knowing this, we were able to understand how the F1 score was so low, at around 25%, given that this metric considers the overall accuracy of the model, but also the recall, which in this case was extremely low, due to again, the **high bias towards healthy subjects** of the model.

At this stage, we knew we were facing a huge problem, we really needed to step up and think of something that could help us significantly reduce the bias, given that either the use of Random Forest, K-FOLDS or doubling the weight of the sick class could not provide a satisfying result, throwing very similar results as the Decision Trees.

After **thorough investigation and deep understanding** of the problem, we encountered the **SMOTE approach**, which consists in artificially inflating the minor class, in order to force the classifier to learn and train for predicting both classes in equal importance, eliminating completely the bias of the model, while also retaining a very respectful **91% of** overall **accuracy with the Decision Tree**, which then threw also an **F1 score of 91% percent**.

With these results, we determined the classifier to be fully developed and started to work on the business goal more precisely, developing an algorithm that could determine, with the results given by the model, the adherence percentage needed to reach our goal of 20% cost decrease.

# 3. Cost analysis

Once the Decision Tree with SMOTE approach was created and evaluated, the next step was to create the analysis of the national healthcare system cost.

First, we decided to take the **actual cost** of the national healthcare system as the **number of sick subjects** that appears in the provided dataset **multiplied by the worst case**, that is, the heart attack treatment (**50.000€**). In this scenario, there are 23.893 subjects, which would give a cost of **1.104.650.000€** (1.194,65M€). We decided to use this number instead of utilizing the number of sick subjects provided by the SMOTE approach because this function creates synthetic samples, therefore the cost would become fictitious instead of based on real data.

After that, thanks to the evaluation metrics of the classification model we could start working on the different cases:

Regarding the **healthy subjects**, we used the **specificity metric** provided by the model, which in our case is approximately **0,91** (results may vary depending on the runtime). The decision behind using this metric is because it gives you, out of all the truly healthy subjects, the percentage of healthy ones which were correctly predicted. For these subjects, the cost analysis is simple (**0€**), as they do not suppose any expense since they are healthy and labelled as so.

On the other hand, for the **1-specificity metric** (**0,09**), which tells the percentage of subjects labelled as sick when they are actually healthy, we have to take into consideration two scenarios: as they are treated as sick, they are also offered the plan, so **the ones that do not accept the plan do not suppose any expense**, since although they are marked as sick, they are actually healthy. However, **the subjects that accept the plan would suppose a cost** for the national health care system (1.000€ each plan). In our case, this cost adds a quantity of **17.888.794,95€** (again, this may vary depending on the runtime).

In relation to the **sick subjects**, we used the **recall metric**, which describes out of all the truly sick subjects, the percentage of sick ones which were correctly predicted. In our case this value is **0,92**. Before commenting on the recall subjects, we will describe the analysis of the **1-recall metric** (**0,08**), as it is simpler. For this metric (percentage of subjects labelled as healthy when they are actually sick), because they are treated as healthy **no plan is offered**, which **eventually concludes in a heart attack, causing the national healthcare system to spend** 50.000€ in each of these subjects (in our case, the cost for this scenario is **91.368.279,34€**).

Going back to the subjects belonging to the recall "branch", there are multiple scenarios/branches: first, as they are sick and labelled as so, they are offered the plan. For those **subjects who decline the plan**, they **eventually end up with a heart attack, causing the national healthcare system to spend** 50.000€ in each of these subjects (in our case, the cost for this scenario is **165.492.258,10€**). On the contrary, for those subjects that accepted the plan (85% of the subjects labelled as sick and truly sick), some of them adhered to it and some of them did not.
At this point, we decided to introduce an iterative loop in order to find the best percentage of adherence that satisfies the goal of the problem: reducing the national healthcare system cost by 20%. Once we did this, we ended up with a **39,18% of adherence** (again, this may vary depending on the runtime). Therefore, continuing with our analysis, the percentage of **subjects which accepted the plan but did not adhere to it** (60,82%) would ultimately suffer a heart attack, causing the national healthcare system to spend 51.000€ for each (as they are given the

plan and the treatment), being the **total cost** for this scenario **581.763.816,67€**. Nevertheless, for the **subjects which stuck with the plan** (39,18%), 75% of them properly **succeeded to reduce the risk of heart attack**, meaning that only 1.000€ were spent (the cost for these subjects adds a total amount of **5.511.491,69€**). But, for the **ones that failed** (25%), the amount of 51.000€ was spent (as they are given the plan and ultimately the treatment), causing the national healthcare system to spend **93.695.358,79€**).

To conclude, considering the Decision Tree with SMOTE approach and the 39,18% of adherence that the team has used, if we sum up all the total cost of the national healthcare system we obtain the amount of **955.719.999,53€** spent, which reduces by 20% the actual cost, properly achieving the goal of the problem.