Data Process second assignment

<Second Part>

CAMPILLO MORALES Salvador

HUC-LHUILLERY Alexia

GONG Seong-Min

TRIBOUT Antoine

Escuela Técnico Superior de
Ingenieros Informáticos

POLITÉCNICA
"Ingeniamos el futuro"

CAMPUS
DE EXCELENCIA
INTERNACIONAL

Universidad Politécnica de Madrid

Content

1. Introduction

Strategic planners should be able to achieve maximum effectiveness at minimum cost. On the other hand, it should be possible to manage and minimize the cost of such spending so that it does not leak through. The cost of healthcare for the people managed by the state is no exception. If the cost after a problem occurs is far higher than the cost spent for prevention, it is reasonable for us to predict the problem in advance and use it to reduce costs and promote investment or welfare elsewhere. One of the most effective ways to reduce costs is to identify current problems through data analysis, find appropriate methods, and apply them to problems to implement preventive activities.

We were given data on heart attacks in this project. From the data, we can identify the lifestyle patterns of people who have had heart attacks, and we can identify people who are at risk for heart attacks. In addition, given that the cost of treating a heart attack is EUR 50 000, while the lifestyle improvement program is EUR 1000 per person, which saves about 98% of the cost, the lifestyle improvement program can reduce heart attacks by about 75%.

In this project our goal is, given the classification models we have been able to develop, to determine the percentage of adherence to lifestyle intervention treatment needed to reduce the total amount spent on heart attacks by 20%. In our sample, we can use some kind of statistical methodology to estimate a range of heart attack incidence percentages with a 95% confidence interval. We can assume that the sample is more or less representative of the population, as it is the best estimator we have. With those percentages, calculate a hypothetical number of future heart attack cases with a sample of 100,000 patients and obtain total cost by multiplying by the cost of treatment. Having that, we can already estimate the actual cost that would suppose a 20% of net savings.

The entire process of data cleaning, preprocessing, and exploration for the task was centered on pandas, numpy, and matplotlib, which are Python packages that provide a diverse set of resources for that purpose. For modeling and deriving results, sklearn, which is a Python library composed of various machine learning modules, was mainly used.

2. Data processing

2.1. Data Pre-Processing

First, the BMI variables were classified into six ranges, and similarly, the ranges of PhysHlth and MentHlth were rearranged into six levels following 'GenHlth' to be categorized. they were categorized as follows :

<BMI>

- Underweight (1): BMI ≤ 18.5
- Normal weight (2): 18.5 < BMI ≤ 25
- Overweight (3): 18.5 < BMI ≤ 30
- Obesity I (4): 30 < BMI ≤ 35
- Obesity II (5): 35 < BMI ≤ 40
- Obesity III (6): 40 < BMI

<PhysHlth and MentHlth>

- Level 1: 1 to 5 points
- Level 2: 6 to 10 points
- Level 3: 11 to 15 points
- Level 4: 16 to 20 points
- Level 5: 21 to 25 points
- Level 6: 26 to 30 points

Looking at the age variable, it can be predicted that the value is a categorical variable, not an age, with a minimum value = 1 and a maximum value = 13. This suggests that each category includes a certain age range, but it is difficult to determine which age group is low and high. Thus, the top four categories and the bottom four categories in the category were divided and compared based on the incidence of heart attack. As a result, the incidence of heart attacks in the bottom four categories was 1.05%, and the incidence of heart attacks in the top four categories was 17.28%. In general, it was possible to estimate that the higher the range of the category type, the higher the age group, but the above results can support this hypothesis. After that, the age variable was encoded as a categorical variable.

Also, facilitate modeling, we convert all binary variables to categorical variables, such as "fruit", "vegetable", "HeartDisease or Attack", "HighBP", "HighChol", "CholCheck", "Smoker", "Stroke", "PhysActivity", "HvyAlcoholConsump", "Sex", and "Diffalk", as a preprocessing process. In addition, "Diabetes", "GenHlth", "Education","Income" those variables were encoded as categorical variables, converting types.

Finally, a copy of the data set is token and all variables are encoded into numeric variables. The Label Encoding approach is chosen instead of One Hot Encoding because numerical order is meant in the context of these explanatory variables.

2.2. Exploratory Data Analysis and Feature selection

For Exploratory Data Analysis and Feature selection, Visualizations such as histograms and heat maps were used to identify the variables which most related to the target variable.

Exploratory Data Analysis

First we plotted histograms of the frequency of age and sex. Here are the results:

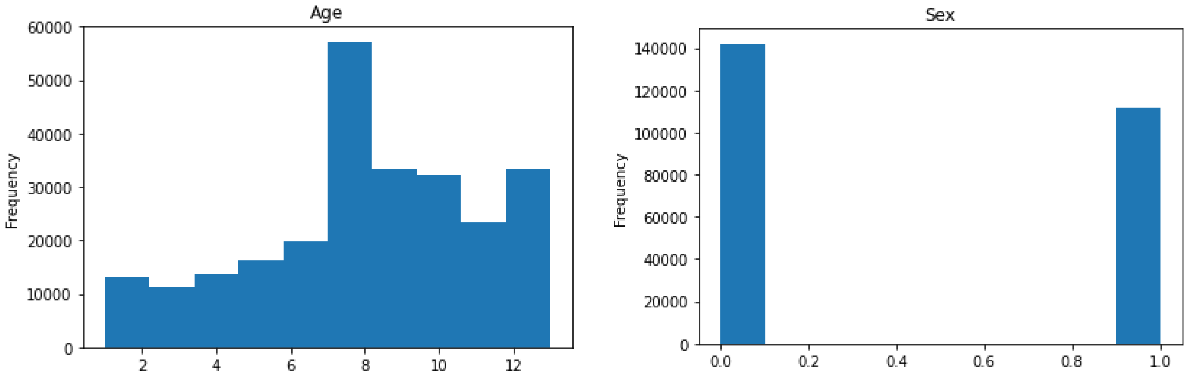


Figure 1. Histograms for age and sex

The histograms suggest that the sample looks quite alike the global population of a developed country, at least in what concerns demographic issues.

We also decided to print out a correlation heatmap so we can have a first idea on the different correlation patterns:

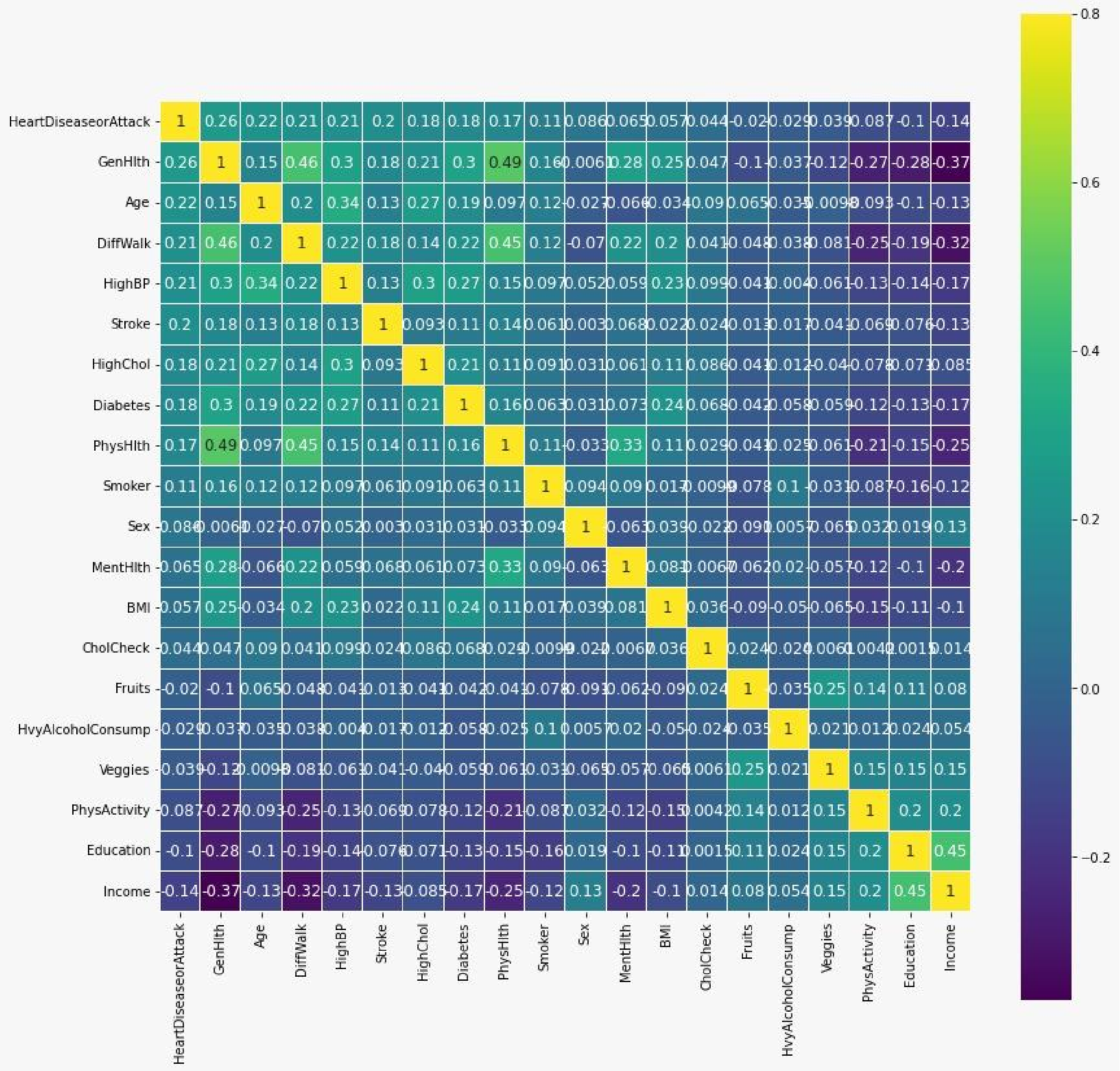


Figure 2. Correlation map for the original dataset

Some values clearly stand out such as: Education and Income (correlation of 0.45), PhysHlth and GenHlth(0.49), PhysHlth and MentHlth (0.33) which seems logical. Also we observe CholCheck isn't really correlated to any of the variables.

To understand better the relations with the response variable we represented relative values for the different levels for the categorical variables with higher |r| values, stratified by the response variable.

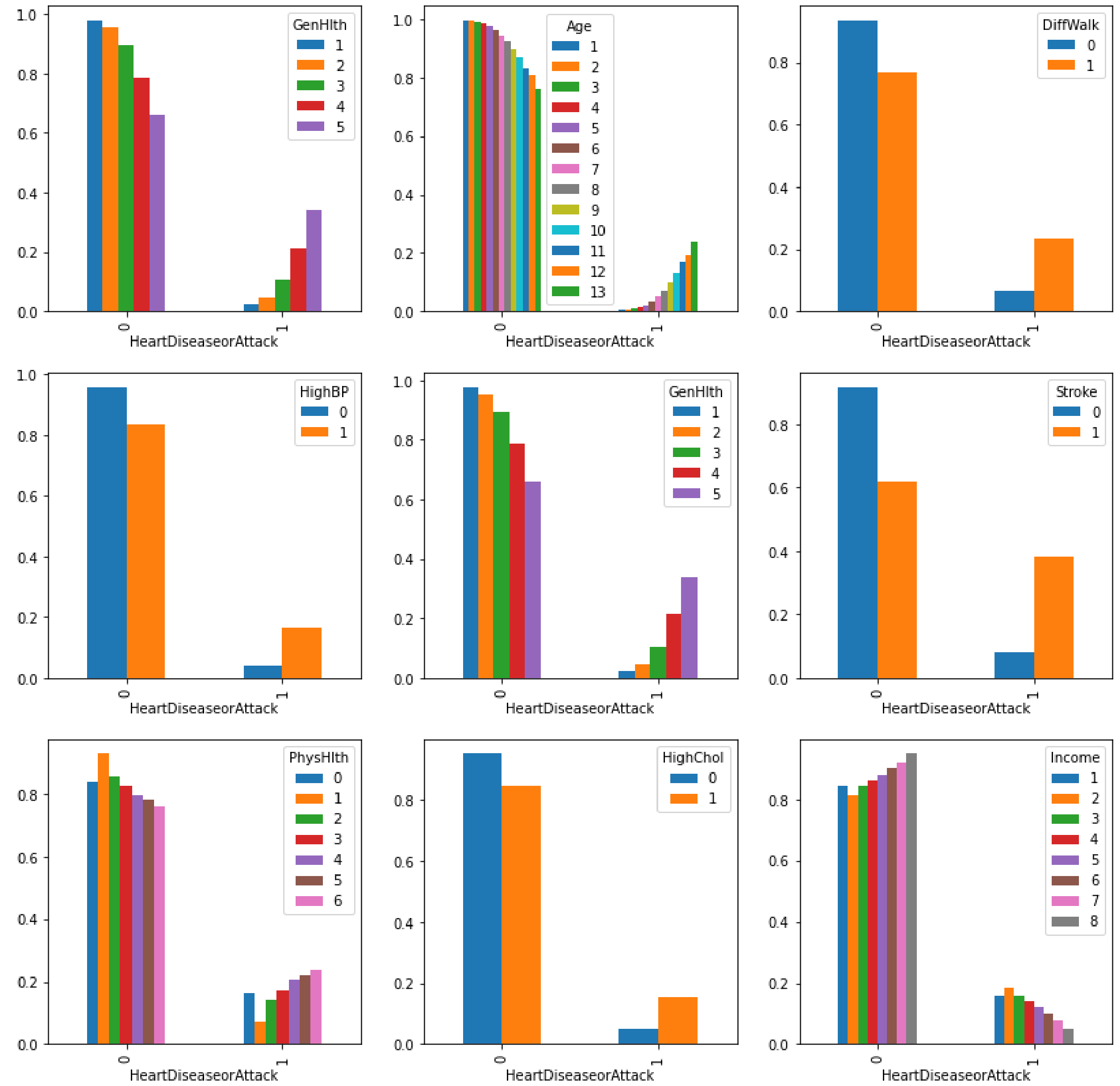


Figure 3. Relation between some explanatory and the response variables, segmented by explanatory variable possible values.

As we can see if we increase the age we increase the proportion of heart diseases or attacks. Same goes with HighChol, DiffWalk, HighBP, Stroke, GenHlth and PhysHlth. Also, the higher the income is the less a patient is inclined to get a heart disease/attack.

We then perform a chisquare test over each pair of variables in the dataset and represent the results in a heatmap:

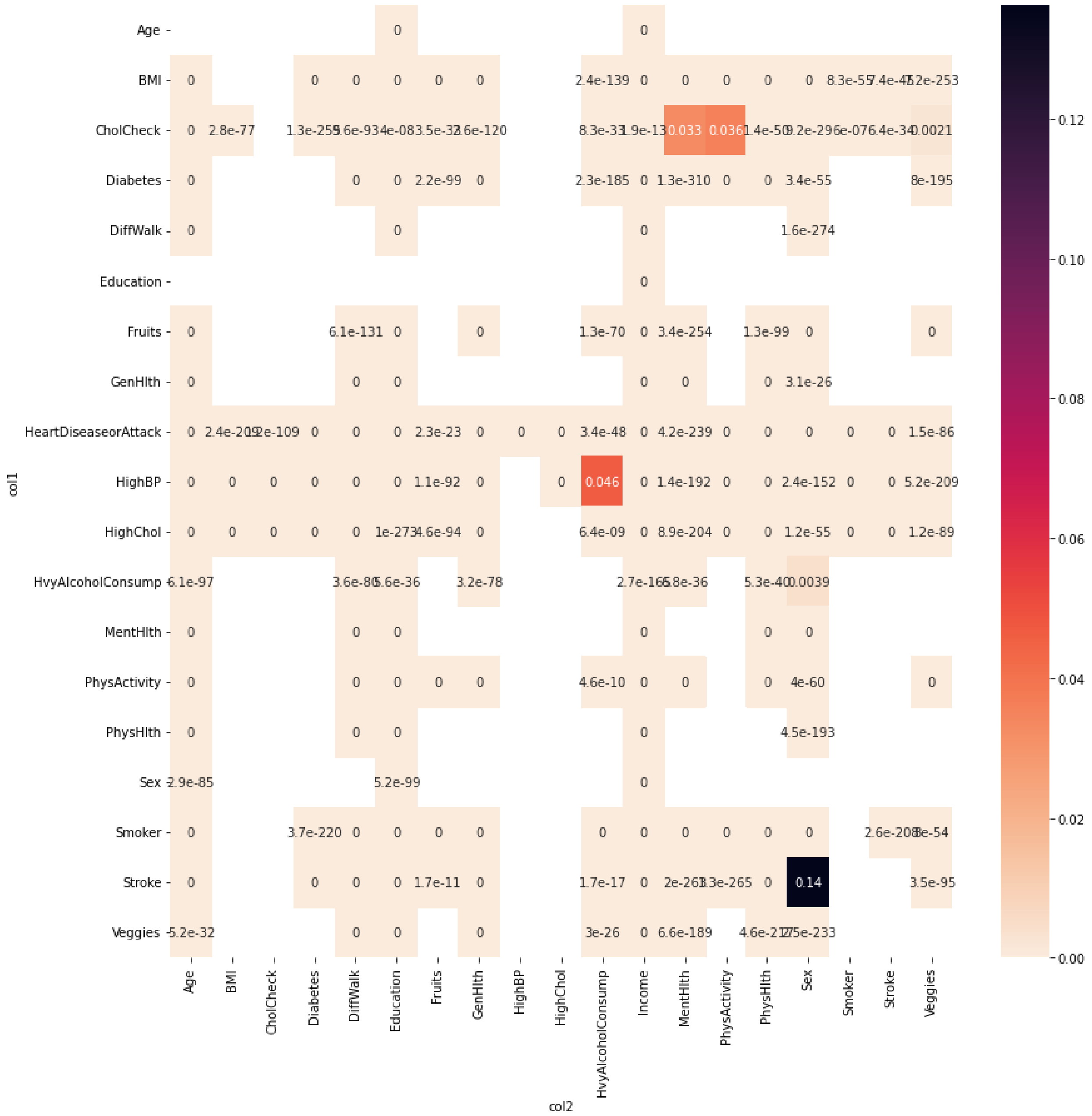


Figure 4. Chi square heatmap

As can be observed, all explanatory variables seem to have a statistically significant relationship with the response variable, as well as among them. That suggests we may face a certain level of redundant information. We need to apply feature selections.

Feature selection

As all p-values regarding combinations of an explanatory variable and the response variable are smaller than 0.05 (in fact by a wide margin), that threshold cannot be used to filter any dataframe column. Instead, what can be done is using the mutual information between explanatory variables and the target as criterion for selecting variables. Half of the columns (the ones with better scores) will be selected.

In addition, a wrapper subset selector will be implemented over the original dataset, in order to get an additional reduced version. A forward (start from null, populate then) method will be used, employing GaussianNB as the algorithm for finding the best possible subset. The reason for selecting a NB algorithm is that Naive Bayes is particularly sensitive to redundant information.

2.3. Model building

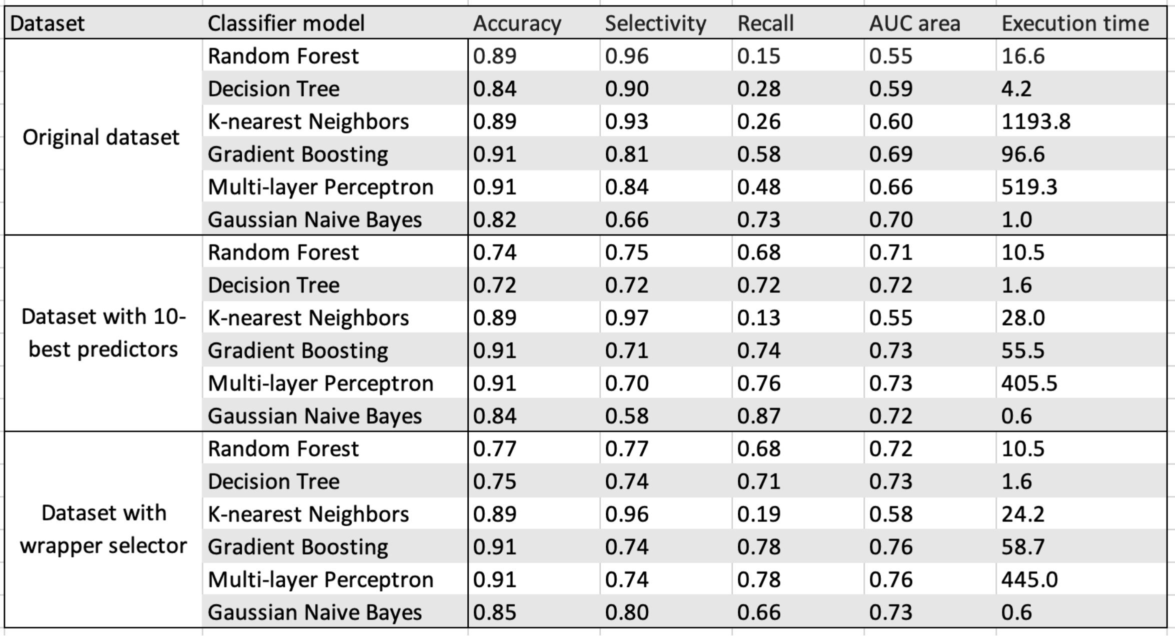
After the selection of our features, we have three datasets that we can use to compute a machine learning model to compute two classes 0 or 1 of HeartDiseaseorAttack, according to the predictors. These datasets are the complete one, one with the 10th best predictors according to the SelectPercentiles function, and a dataset reduced with a wrapper subset selector. For all of these dataset, the beginning of the processus is to split the data into a train and a test set, 70% of the data are used for training and 30% for testing at the end. Six models have been created for each dataset. All the models after are **classifiers**, because we want to find the right value between the two classes 0 and 1 that are present in the target variable.

The first ones are a **random forest** model and a **decision tree** model. These two models are working on computing a decision graph to find the best way among predictors to explain the target value. These models contain a parameter class_weight which when it is in balanced mode allows to avoid an imbalance problem of the dataset. Actually, in all the datasets there is an issue of the number of 0 there is compared to the 1 of the column HeartDiseaseorAttack. Then the parameter random_state controls the randomness of the bootstrapping of the samples when building trees and the sampling of features to consider when looking for the best split for a node in a random forest model, and it controls the randomness of splitting nodes for a decision tree model. The parameter ramdom_state will be the same as for splitting the values to generate similar results. The random forest model is also using n_jobs parameter, which allows it to compute on all processors available in parallel, because the dataset is large so it allows to gain time computation.

For the others models, the parameter class_wieght doesn't exist so the function SMOTEN is used to replace that. This function allows us to create a new dataset with more values close to the existing ones of the less represented cluster of our data. So the new dataset contains approximately the same number of observations for 0 and 1 in the column HeartDiseaseorAttack, to avoid keeping an unbalanced dataset in the models. The new values are computing with the neighbors of the existing values, it is like the function duplicates some information to equilibrate the two clusters.

The next model is the **k-nearest neighbors** classifier, which computes the model in comparing the individual and people with close values in their variables. Here only the 3 nearest neighbors are keeped for the rapidity of the computation. A **gradient boosting** classifier has been created, it works with a gradient descending method to find the best estimators' coefficient to classify the data. The next model is the **multi-layer perceptron**, it is a model based on neural networks. It optimizes the loss function of the model. The learning rate is adaptive to have better performances in time between all the computation steps, with an initial value at 0.01, and the random state stays the same, like for the gradient boosting model. The last model is the **gaussian naive Bayes** model, based on Bayes' theorem to compute it.

For each of the models the main **scores** are printed to compare their performances. The first ones are the mean of the accuracy and the standard deviation of the cross-validation score computing on the all dataset. This cross-validation splits the data into 4 subsets, and fourth time the model is computed with 3 out of 4 datasets and the score is computed on the last one, then the mean of the accuracy is printed with the standard deviation of the scores obtained. Then the selectivity of the model is calculated thanks to the confusion matrix, which corresponds to the number of true positives divided by the sum of the numbers of true positives and false positives. Then the recall score is computed, which corresponds to the number of true positives divided by the sum of the numbers of true positives and false negatives. The ROC AUC area is shared from prediction scores, it indicates the area under the curve score for the model. Then the confusion matrix is printed to compare the number of true or false positives or negatives. And to finish, the time the model needs to run the fitting, the prediction on the test dataset and the cross-validation score is printed to have an idea of the time needed to compute for each model.



Dataset	Classifier model	Accuracy	Selectivity	Recall	AUC area	Execution time
Original dataset	Random Forest	0.89	0.96	0.15	0.55	16.6
	Decision Tree	0.84	0.90	0.28	0.59	4.2
	K-nearest Neighbors	0.89	0.93	0.26	0.60	1193.8
	Gradient Boosting	0.91	0.81	0.58	0.69	96.6
	Multi-layer Perceptron	0.91	0.84	0.48	0.66	519.3
	Gaussian Naive Bayes	0.82	0.66	0.73	0.70	1.0
Dataset with 10-best predictors	Random Forest	0.74	0.75	0.68	0.71	10.5
	Decision Tree	0.72	0.72	0.72	0.72	1.6
	K-nearest Neighbors	0.89	0.97	0.13	0.55	28.0
	Gradient Boosting	0.91	0.71	0.74	0.73	55.5
	Multi-layer Perceptron	0.91	0.70	0.76	0.73	405.5
	Gaussian Naive Bayes	0.84	0.58	0.87	0.72	0.6
Dataset with wrapper selector	Random Forest	0.77	0.77	0.68	0.72	10.5
	Decision Tree	0.75	0.74	0.71	0.73	1.6
	K-nearest Neighbors	0.89	0.96	0.19	0.58	24.2
	Gradient Boosting	0.91	0.74	0.78	0.76	58.7
	Multi-layer Perceptron	0.91	0.74	0.78	0.76	445.0
	Gaussian Naive Bayes	0.85	0.80	0.66	0.73	0.6

Performances of the models

3. Computing the minimum adherence required

3.1. Premises and followed methodology

As said at the beginning, the main goal of this practical application is to determine, given the developed machine learning models, the minimum percentage of adherence to the proposed lifestyle treatment to reduce the aggregated cost of heart attack treatments by 20%.

A set of premises will be used for that purpose.

- The sample used as reference for building the models has been properly gathered, following the appropriate statistical methods. So that, it is a good estimator of the whole population. However, as it is not possible to create samples fitting perfectly with its associated population behavior, confidence intervals of 95% will be computed over the statistic "proportion of positive cases of HeartDiseaseorAttack". Adherence percentage will be computed in both cases. Histograms over demographic data (sex and age) suggest that this hypothesis is reasonable.
- 85% of patients that are going to be prescribed with the preventive treatment will accept it at the beginning, the remaining 15% will reject it.
- The cost associated with a heart attack is 50k€. The cost associated with a lifestyle treatment is 1k€.

- A conservative approach will be used regarding treatment effectiveness. If a patient doesn't follow the lifestyle treatment until the end, the treatment will be considered completely ineffective. Moreover, if a patient adheres a treatment at the beginning and then quits it before its end, the cost of the treatment is 1000€, as occurs with patients that follow the treatment till the end
- Partitions will be considered homogeneous ones. For example, if 85% of the patients classified as prone to suffer a heart attack follow the treatment, it will be supposed that the percentage of initial adherence to the treatment is the same both for patients correctly and incorrectly classified.
- The lifestyle treatment reduces the possibility of suffering a heart attack by 75%.
- For each of the chosen models, sensitivity and recall will be projected over two hypothetical samples of 100k patients (one with the number of patients prone to suffer a heart attack equivalent to the upper bound of the confidence interval, the other with this number of patients equivalent to the lower bound), thus generating references for the computing process.

The adherence computing will be performed as follows

(1) Original cost: oc=patients_prone_to_heart_attack*50000

(2) Maximum cost to achieve the goal: gc = oc * 0,8

(3) oc = n*(1-r)*50000+n*r*0,15*50000+1000*n*((s-1)+r)+n*r*0,85*0,25*a+ n*r*0,85*b

Where n≡100k (the total sample size), r≡recall, s≡selectivity, a≡proportion of people following the treatment until the end, b≡proportion of people quitting the treatment before the end of it. The terms of the equation represent, in this order:

- The cost associated with cases of patients highly prone to heart attacks that have been misclassified.
- The cost associated with cases of patients highly prone to heart attacks that have been correctly classified, but that have left the treatment before reaching satisfactory results.
- The global cost associated with lifestyle interventions.
- The cost associated with cases of patients highly prone to heart attacks that have followed the treatment till the end, but that have suffered a heart attack even so.

The minimum required percentages of adherence and non-adherence (a and b) can be computed using a two equation system: on the one hand the equation (3) and, in the other hand, a+b=1

3.2. Selected models

Among the different built models, two of them have been selected for this task, due to their performance.

3.2.1. Gaussian Naive Bayes + filter feature subset selection with mutual information

Although the wrapper forwarding feature subset selection algorithm was implemented using GaussianNB algorithm (so then, the global accuracy is higher for GaussianNB + the wrapper), the recall is particularly high in this case. A high recall can be considered a particularly capital feature in this context, because of the importance of avoiding as possible false negatives (both for public health and economic reasons).

- Accuracy: 0.84
- Recall: 0.87
- Selectivity: 0.58
- AUC area: 0.72

3.2.2. Multilayer Perceptron +wrapper feature subset selection based on Gaussian NB

The recall for this second model is lower than the one for the first but the selectivity is quite better . So this model is expected to incur in less false positive misclassifications than the first with a little worse performance regarding false negatives. The AUC area for this model is the highest among all built models so a good performance can be expected.

- Accuracy: 0.91
- Recall: 0.87
- Selectivity: 0.58
- AUC area: 0.72

3.3. Results obtained

Next, the results achieved are shown. The code used for reaching the shown results, as well as the results themselves, can be also found at the codebook.

3.3.1. Confidence intervals

The proportion of positive cases in the original sample as well as the sample size have been used as the main parameters for computing the confidence intervals. For computing that, a studentized t-intervals algorithm has been used (mainly because of its capacity of getting accurate results independently of the skewness within the sample).

Proportion of heart attack cases: (0,091 0,0973) (95% CI)

3.3.2. Projected confusion matrices

Case 1: Gaussian Naive Bayes + filter feature subset selection with mutual information

Lower bound

Predicted values / Actual values	0	1
0	52722	38178
1	1183	7917

Upper bound

Predicted values / Actual values	0	1
0	52357	37913
1	1265	8465

Case 2: Multilayer Perceptron +wrapper feature subset selection based on Gaussian NB

Lower bound

Predicted values / Actual values	0	1
0	64539	26361
1	1729	7371

Upper bound

Predicted values / Actual values	0	1
0	64092	26178
1	1849	7881

3.3.3. Results obtained

Once configured and solved the two unknown equation system previously explained for each of the 4 cases, the results obtained are the following:

Minimum % of adherence required for achieving the proposed goal (Model 1, lower bound): 54.31%
Minimum % of adherence required for achieving the proposed goal (Model 1, upper bound) is: 53.30%
Minimum % of adherence required for achieving the proposed goal (Model 2, lower bound) is: 53.10%
Minimum % of adherence required for achieving the proposed goal (Model 2, upper bound) is: 52.29%

The second model requires a slightly lower adherence to reach the proposed goal. Anycase, an adherence of about ~53% seems to be required.

4. Results interpretation and conclusions

Taking a deeper look at the results, several conclusions can be inferred.

4.1. The value of Feature Subset Selection

In order to refine machine learning models and achieve the best possible results for the kinds of problems that m-l algorithms can help to take care of, discarding noisy, redundant and invariant features plays a capital role, especially for those algorithms that are particularly sensitive to some of these undesired circumstances. The purpose of FSS is this. Depending on the particular context of the dataset analyzed, one or another subset selector will be selected, based on the performance expectancy.

Applied to this particular case, the use of subset selectors have resulted in models much more suitable for solving the raised business problem than the ones corresponding with the original dataset.

4.2. Data science and ethics: the tradeoff between public health and economy

Comparing the results for the two chosen models regarding adherence percentages, the second model seems slightly more efficient regarding money saving. A little lower percentage is needed for achieving the desired goal in the second model than in the first. At the same time, the second model presents slightly higher false negative rates, which could lead to a little bit more heart attacks than the first model.

In that particular case, differences are narrow. We could simply choose the first model, assuming that the money gap between the two models is sufficiently low to privilege the possibility of avoiding a small additional number of heart attacks.

But as the gap between these two concepts becomes higher, things become more difficult. This kind of situation is quite common when working in “sensitive” areas, especially clinical ones, and remarks the role that ethics play in data science and its importance.

4.3. More complex algorithms do not always mean achieving better results

If we take a look at the two selected algorithms (considered the most appropriate ones) one of them (the Multilayer Perceptron) lays on a much more complex development than the other (based on the Bayes theorem) and requires more computation. However, the results of both are almost interchangeable. We have even determined that in this specific context GaussianNB could be somewhat more suitable than MultilayerPerceptron. Other discarded algorithms (Decision Tree Classifier, Gradient Boosting), also based on simpler premises, have also achieved a quite decent performance

4.4. Accuracy is not all in m-l classification problems

This is a paradigmatic case where other metrics (especially the recall) are much more useful than the raw accuracy. That applies especially in the cases in which there is one kind of error (or a subset of them) that has more severe consequences than the other. This situation can be expected to be found very usually in the clinical domain, but also in a wide spectrum of fields where the decisions taken have an economic impact.