Body signals of smoking

Andrea Cardinali and Adonis Kingsley Granita

Table of Contents

[Introduction 1](#_Toc159063224)

[Dataset 1](#_Toc159063225)

[Exploration phase 2](#_Toc159063226)

[Preprocessing 2](#_Toc159063227)

[Classification models 2](#_Toc159063228)

[Evaluation metrics 3](#_Toc159063229)

[Conclusions 4](#_Toc159063230)

[Bibliography 5](#_Toc159063231)

## **Introduction**

Smoking has multiple serious effects that impact the health of smokers and those around them. By gathering and analysing health data of smokers and non-smokers and using the KNIME platform, we can create a classification model that can predict an individual’s smoking status, depending on other variables.

## **Dataset**

The available training dataset, which will be used to train the classification model, is rather big, containing more than 14000 rows and 26 columns. The test set that will be used to validate the accuracy of the model, has one more column related to the binary variable of smokers.

Here are the 26 variables we’re dealing with:

1. ID: that’s the ID of the individual.
2. Gender: 1 for males and 0 for females.
3. Age
4. Height: measured in cm.
5. Weight: measured in cm.
6. Waist: measured in cm.
7. Eyesight left: from 0 to 1.
8. Eyesight right: from 0 to 1.
9. Hearing left: 1 for yes, 2 for no.
10. Hearing right: 1 for yes, 2 for no.
11. Systolic: the pressure in the arteries when the heart beats [1].
12. Relaxation: also known as diastolic pressure. That’s the pressure in the arteries between beats [1].
13. Fasting blood sugar: 0-99 normal level of sugar; 100-125 prediabetes; 126+ diabetes [2].
14. Cholesterol: from 0 to 200 it’s a normal level, from 201 to 239 the subject is at risk, and a 240+ level it’s dangerous [3].
15. Triglyceride: type of fat in the body, used to gain energy [4].
16. HDL: the “good” cholesterol [4].
17. LDL: the “bad” cholesterol [4].
18. Haemoglobin: it transports oxygen to the tissues [5].
19. Urine protein: measures if traces of protein are found in urine [6].
20. Serum creatinine: measure of how good the kidneys are working [7].
21. AST: A high level means the liver is damaged [8].
22. ALT: A high level means the liver is damaged [9].
23. Gtp: indicates liver disease [10].
24. Oral: oral examination status
25. Dental caries: 0 for no, 1 for yes.
26. Tartar: either yes or no.

## **Exploration phase**

First, we checked whether there are missing values and, if so, how to replace them.

Luckily for us, the dataset does not contain any missing value. We can then proceed to see how many variables are correlated, so to avoid redundancy.

We opted to remove the carriable “oral” from our analysis encompassed solely affirmative responses. This rendered the variable redundant for our predictive modeling purposes, as it failed to contribute meaningful variability to the model.

We decided to remove it from the training set together with the ID variable, to make the process a little bit less demanding.

There’re couples of variables that show some degree of correlation, like systolic and relaxation or weight and waist. However, there’re two variables that are highly correlated: cholesterol and LDL (or bad cholesterol). We decided to remove the LDL variable as well to keep only the cholesterol.

## **Preprocessing**

Before beginning with the training phase, we encoded the variables "gender" and "tartar" into binary representations. Specifically, "gender" was transformed into a binary variable where 0 signifies females and 1 denotes males. Similarly, for "tartar", we assigned a value of 1 to indicate "yes" and 0 to signify "no".

Furthermore, we conducted data normalization exclusively for the multilayer perceptron, omitting this step for logistic regression.

## **Classification models**

We implemented four classification models: logistic regression, the multi-layer perceptron, a decision tree classification, and a random forest algorithm.

Logistic regression is a generalized linear model that takes in input the explanatory variables to predict the probability of the class attribute smoking.

The logistic regression model was chosen for its widely recognized effectiveness in binary classification tasks.

To compute the accuracy, we opted to classify probabilities greater than 0.5 as 1.

The multi-layer perceptron is an artificial neural network consisting of linked neurons. Those neurons can be:

* Input neurons: associated with the explanatory variables.
* Hidden neurons: can be a set or multiple set.
* Output neurons: associated with the class attribute.

The signal goes from the input neurons to the hidden neurons. Once all the neurons receive the signal coming from all the input neurons, they start sending the signal to the next hidden neurons (if there are any) or directly to the output neuron, where the class attribute is registered.

The number of hidden neurons and hidden layers is not fixed as it influences the accuracy measure. In this case we choose one hidden layer with twelve hidden neurons, as from multiples tries it gives us the best accuracy.

We have decided also to use an "AttributeSelectedClassifier" node to selects the optimal subset (feature selection) using a wrapper approach and then executes a classification model called j48 that performs a decision tree.

A decision tree classification model is a hierarchical model that uses a supervised approach to classify data. It consists of

* A rood node, which has no incoming edges.
* Internal nodes, one incoming edge and two, or more, outgoing edges.
* Leaf, which have one incoming edge but no outgoing edges.

Decision trees can be used not only for interval and ratio attributes, but also for nominal and ordinal ones, making it suitable for this dataset.

A random forest algorithm is a heuristic model, meaning that it searches for an approximate solution by constructing different decision trees and the solution is the class selected by most trees [11]. This algorithm has the theoretical advantage of not overfitting the data, as it usually happens in a single decision tree.

A wrapper takes the dataset, selects different subsets of attributes which are then used as input to a classification model. The performance of the classification model is evaluated based on these subsets of attributes, and the list of optimal attributes is determined based on the model's performance.

Is an iterative process and different subsets of attributes are evaluated to find the subset that has the best performance according to a specific criterion.

On our case we used the AUC as criterion.

## **Evaluation metrics**

To assess the quality of our models we used different measures that state how good they fit to the data.

The first measure we used is the AUC, or Area Under the Curve. It measures the area under the ROC Curve (that’s the Receiver Operating Characteristics Curve).

The AUC measures how a classifier model makes better predictions than a random one. Its values range from 0 to 1: the higher the value, the better the evaluation.

Immagine che contiene testo, schermata, Diagramma, linea

Descrizione generata automaticamente

Figure 1: ROC Curves

After reviewing the ROC curves of the four models in Figure 1, we concluded to discard the decision tree model and keep the other three to compare their accuracy.

The second measure we used is the accuracy. It measures the number of correctly classified records divided by the total number of classified records. The accuracy formula is:

To get a more reliable and robust estimate of how well a model will generalize to unseen data compared to a simple train/test split, we performed a 10-fold cross-validation by splitting the data into 10 subsets and training the model on different combinations of these subsets, ensuring that each partition of the dataset maintained consistent proportions of the class attribute. This method allowed for a comprehensive evaluation of model performance across multiple folds, mitigating the risk of bias introduced by uneven class distributions. Subsequently, we compared the cross-validation accuracy with the accuracy obtained from our prior train/test split, shown in Figure 2.

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

Figure 2: Accuracy Box Plot

In the iterated holdout, the original training set is divided into two distinct datasets, a training set, and a test set. The process is then iterated multiple times to find a better accuracy value.

After evaluating both models based on accuracy and Area Under the Curve (AUC) shown in Figure 1 and Figure 2, we determined that the Random Forest outperformed the other models, making it our top choice.

Once chosen the RF, it’s necessary to apply the model to the test set to finally predict the smoking variable.

## **Conclusions**

We generated a cumulative gain chart that illustrates [in Figure 3] the percentage of positives identified by a model within specific subsets.

And finally, we conducted an Iterated Holdout to enhance the robustness of our AUC assessment [in figure 4].

Immagine che contiene testo, schermata, Diagramma, linea

Descrizione generata automaticamente

Figure 3: Cumulative Gain Chart

Immagine che contiene testo, diagramma, schermata, Rettangolo

Descrizione generata automaticamente

*Figure 4: AUC box plot*

From Figure 4 we see that the Accuracy calculated just on the single training set is slightly higher than the interval computed with the iterated holdout. A reason could be that the iterated holdout uses a smaller training set, thus leading to less accurate results.

Despite this, the result of the iterated holdout is a consistent interval, as the variance is small.

# Bibliography

|  |  |
| --- | --- |
| [1] | NHS, "What is blood pressure?," NHS, 12 12 2022. [Online]. Available: https://www.nhs.uk/common-health-questions/lifestyle/what-is-blood-pressure/#:~:text=As%20a%20general%20guide%3A,to%20be%20below%2090%2F60mmHg. |
| [2] | CDC, "Diabetes tests," 28 02 2023. [Online]. Available: https://www.cdc.gov/diabetes/basics/getting-tested.html#:~:text=Fasting%20Blood%20Sugar%20Test,higher%20indicates%20you%20have%20diabetes.. |
| [3] | Cleaveland Clinic, "Cholesterol Numbers and What They Mean," Cleaveland Clinic, 22 07 2022. [Online]. Available: https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean. |
| [4] | CDC, "LDL and HDL Cholesterol and Triglycerides," 16 05 2023. [Online]. Available: https://www.cdc.gov/cholesterol/ldl\_hdl.htm#:~:text=HDL%20(high%2Ddensity%20lipoprotein),it%20back%20to%20the%20liver.. |
| [5] | Britannica, "hemoglobin," Britannica, 19 12 2023. [Online]. Available: https://www.britannica.com/science/hemoglobin. |
| [6] | Santagostino, "Test per le proteine nelle urine. Cos’è e perché farlo," Santagostino, [Online]. Available: https://www.santagostino.it/it/santagostinopedia/proteine-nelle-urine. |
| [7] | Mayo Clinic, "Creatinine test," Mayo Clinic, 28 02 2023. [Online]. Available: https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646. |
| [8] | Cleaveland Clinic, "Aspartate Transferase (AST)," Cleaveland Clinic, 24 11 2021. [Online]. Available: https://my.clevelandclinic.org/health/diagnostics/22147-aspartate-transferase-ast. |
| [9] | Cleaveland Clinic, "Alanine Transaminase (ALT)," Cleaveland Clinic, 04 11 2021. [Online]. Available: https://my.clevelandclinic.org/health/diagnostics/22028-alanine-transaminase-alt. |
| [10] | MedlinePlus, "Gamma-glutamyl Transferase (GGT) Test," MedlinePlus, 16 11 2021. [Online]. Available: https://my.clevelandclinic.org/health/diagnostics/22055-gamma-glutamyl-transferase-ggt-test. |
| [11] | Wikipedia, "Random forest," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Random\_forest. |