

**Course:**

**Machine Learning Lab**

Half-term Activities Report

**Methodology and Tools for Developing Machine Learning models**

# CONTENTS

[1 CONTENTS 2](#_Toc86100136)

[2 LIST OF FIGURES 4](#_Toc86100137)

[3 Methodology to develop machine learning models 5](#_Toc86100138)

[3.1 Methodological guidelines 5](#_Toc86100139)

[3.2 Methodology for developing Machine Learning models 5](#_Toc86100140)

[3.3 Tools to develop Machine Learning models 6](#_Toc86100141)

[4 Data preparation 6](#_Toc86100144)

[5 Exploratory Data analyses 7](#_Toc86100145)

[5.1 Regression 7](#_Toc86100146)

[5.2 Classification 7](#_Toc86100147)

[6 Developing machine learning models 8](#_Toc86100148)

[6.1 Splitting 8](#_Toc86100149)

[6.2 Evaluating metrics 8](#_Toc86100150)

[6.3 Training part 9](#_Toc86100153)

[6.4 Automation of the whole process 9](#_Toc86100154)

[7 CONCLUSIONS 10](#_Toc86100155)

[7.1 Methodology 10](#_Toc86100156)

[7.2 Framework and tools 10](#_Toc86100157)

[8 REFERENCES 11](#_Toc86100158)

# LIST OF FIGURES

Figure 1 - Developing Machine Learning models 5

Figure 2 – Read data with Panda 6

Figure 3 – Remove missing data 7

Figure 4 – Creating BMI index 7

Figure 5 – Scatter plot and correlation matrix 7

Figure 6 – Histogram and boxplot using Panda 7

Figure 7 – Splitting data 8

Figure 8 – Regression metrics 8

Figure 9 – Classification Metrics 8

Figure 10 – Training models with scikit-learn 9

Figure 11 – Automated process 9

# Methodology to develop machine learning models

## Methodological guidelines

Diagram

Description automatically generated

Figure 1 - Developing Machine Learning models

As it can be seen in Figure 1, I started with a first activity directed to clean and prepare the data for further use in developing Machine Learning models. To treat this problem, I first had to access the data I could use and visualised them in order to know what to do with. After this first step, I process it by retiring the missing data and did some features engineering.

Indeed, in this particular case, I created a BMI index with the aim to replace weight and height. This kind of features engineering is important because it simplify the problem for the algorithm, and it also give him data that are more relevant. For example, BMI is a better index to judge about AHI because the report between height and weight is better than having the information separately (note that in deep learning thig king of relationship is supposed to be found by the models itself).

Besides that, I addressed an Exploratory Data Analysis (EDA) using the Statistical tools I will describe in Section 5. Then I could have thinks about using some models based on a priory knowledge, train and test them using different and separated datasets to avoid overfitting. And I finally tried to change some hyperparameters of my models to get better results.

## Methodology for developing Machine Learning models

As said before, I train and test my models on separated datasets to avoid doing overfitting. Overfitting mean makes a model perfectly fitted with a dataset, but which won’t be able to have good result on others because there are not exactly the same. That’s why I choose, before creating and testing each model, to split my data into to set of 80%-20% and use one for training and another for testing my different models.

Even my results are not good enough affirmed that I correctly solve the problem, I can affirm that they are solid as they well represent how do my models understand the problem. I used different data for training and testing a repeat the whole process many times for each models getting almost the same results each time. So, if my models have to treats other data of those king in the future to predict AHI or to classify patient, they will do it with the same MSE or accuracy as those get during the test.

## Tools to develop Machine Learning models

Even if I know R and known how it is powerful to deal with the EDA part, I choose to do all my project in python because I am familiar with this language and libraries since I am using them for a long time now. I also choose it because I prefer using the Google-Colab environment where I am sure to get all the package I needs already installed. To finish, I took this decision because python is the most used language to deal with ML problem and I didn’t want to switch of language while using R to analyses and Python to create and evaluate model.

### Data preparation and EDA

Logo, company name

Description automatically generatedLogo

Description automatically generated

I used the panda to access my data because it is simple to visualise and treat the data right after reading it from a CSV file.

To explore and process my data, I had the option of using panda data frames or NumPy arrays. Since I already worked with NumPy in the past, I choose to do this part of my project using panda because I thought it was more proper and understandable, adding the fact that I my data was already in data frames form since I get them with penda.

### Model training and testing

Logo

Description automatically generated

As it is easy and intuitive to used, I decided to use the scikit-learn library to create, train and evaluate my models.

# Data preparation

As I explained before, I use the python library named panda to easily access the data which was content in csv files. After this I could have use my data inside the python code as a data Frames variables (types of data provide by Panda library).

A picture containing shape

Description automatically generated

Figure 2 – Read data with Panda

In order to prepare them, I then drop the rows which getting a row with no information to ensure that all my data are containing correct information.



Figure 3 – Remove missing data

And I finished my data pre-processing by creating the BMI column to use it instead of ‘Weight’ and ‘Height’.



Figure 4 – Creating BMI index

# Exploratory Data analyses

## Regression

While trying to create a ML model to predict the AHI, I chose to use panda to get the scatterplot and the correlation matrix of the data.

Text

Description automatically generated

Figure 5 – Scatter plot and correlation matrix

## Classification

In another hand, I used tools as histogram and boxplot for dealing with the classification approach.





Figure 6 – Histogram and boxplot using Panda

Those two statistical tools help me to find the more relevant feature in order to solve the problem. But since none of those showed a clear and evident relation with AHI, I decided to use all features I had.

# Developing machine learning models

## Splitting

To train my models avoiding overfitting, I used panda to split my data into to datasets:

Text

Description automatically generated

Figure 7 – Splitting data

## Evaluating metrics

To evaluate how the ML algorithm are doing with the problem, I used scikit -learn functions to easily compute the metrics I was needing.

### Regression

I used the MSE and the R2\_value for evaluate my regression algorithm

Text

Description automatically generated

Figure 8 – Regression metrics

### Classification

I used the accuracy metrics and the confusion matrix for classification’s one

Text

Description automatically generated

Figure 9 – Classification Metrics

## Training part

Here I also took benefits from the scikit-learn library while directly using provided functions to create and train the different models.

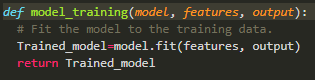
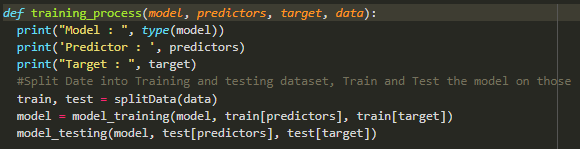


Figure 10 – Training models with scikit-learn

## Automation of the whole process

Note that I tried to automate the whole creating, splitting, training and testing process by creating function for each one of this part so I just need to call them in the order to try what type of model I want.



Text

Description automatically generated

Figure 11 – Automated process

# CONCLUSIONS

## Methodology

Looking back on the employed methodology for handling the problem, I believe that splitting the data was a really good choice to evaluate our models in a real-world condition and therefore have a relevant metrics on them. Perhaps I could have used more statistical tools and metrics to analyse and test my models, but I had to choose between many options and so I decided to take the most common ones so all peoples can understand. To choose the hyperparameters, I did this manually by testing combination one after another and I surely can improve this process by using method to determine those parameters in the future so I could explain why I take this value instead of another.

## Framework and tools

To deals with this problem, I convinced that python and the library I used help me to do it in a proper and clean way so my work can be understanding and be a good introduction to someone who want to introduce himself to this kind of ML problems. This point was really my priority while doing this project because I want it to be reusable, for me first for potentials future projects but not only.

In my opinion working with those frameworks what a good choice because I could have processed the problem in a really clear way and make my notebooks clean and understandable. The only bad thing I can say is about the fact that panda do not enable to do as much things as we want with the data and sometimes, I had to work around making a self-made function in order to go deeper in the EDA.

# REFERENCES

[1] – Overfitting: https://en.wikipedia.org/wiki/Overfitting

[2] – Google Colab: https://colab.research.google.com/notebooks/basic\_features\_overview.ipynb

[3] – Panda: https://pandas.pydata.org/

[4] – NumPy: https://numpy.org/

[5] – Scikit-learn: https://scikit-learn.org/stable/