

**Course:**

**Predictive and Descriptive Learning**

Half-term Activities Report

**Developing Machine Learning models**

# CONTENTS

[1 CONTENTS 2](#_Toc86100212)

[2 LIST OF FIGURES 4](#_Toc86100213)

[3 Type of machine learning 5](#_Toc86100214)

[3.1 Self-Supervised learning 5](#_Toc86100215)

[1. DEVELOPING Machine learning models 6](#_Toc86100218)

[1.1. Problem description 6](#_Toc86100219)

[3.2 Machine Learning approach 7](#_Toc86100220)

[4 Data description 8](#_Toc86100224)

[5 EXPLORATORY DATA ANALYSES (EDA) 9](#_Toc86100225)

[6 Experimental setup 10](#_Toc86100226)

[7 machine learning models 11](#_Toc86100227)

[7.1 Regression: 11](#_Toc86100228)

[7.2 Classification: 11](#_Toc86100229)

[8 RESULTS 12](#_Toc86100230)

[9 CONCLUSIONS 12](#_Toc86100231)

[10 REFERENCES 13](#_Toc86100232)

# LIST OF FIGURES

Figure 1 – OSA Illustration 5

Figure 2 - CSV 1: Clinical Data 7

Figure 3 - CSV 2: Gender and Age 7

Figure 4 – Joined Data frame 7

Figure 5 – Adding of BMI 8

Figure 6 – Correlation Matrix 8

Figure 7 – Box plot Age Vs OSA 9

# Type of machine learning

## Self-Supervised learning

### What is Self-Supervised learning

Self-supervised learning is useful when having a lot of labels data is not possible. For example, it’s very used for NLP (natural language processing) It can be related to unsupervised learning because data are not labelled but the difference is that we don’t want the model to do cauterization. Indeed, it’s not possible to have all sentences that can exist labelled because there are too many combinations that we can make with words. So self-supervised learning is used to make the model understand the mechanism of language such as correlation between words. A model trained like this with a big amount of data can then be used or trained to complete a sentence with a missing word or a text with a complete sentence. It’s also can complete image with pixel or video with frame.

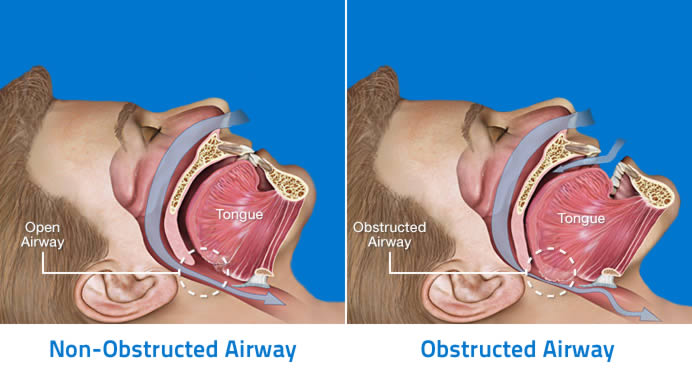
### How it is used

The huge improvement in NLP AI area is dues to this way of learning. A real application of this type of training is the algorithm of NLP called "Bert" and released in 2018 which has beat all other algorithms.

# DEVELOPING Machine learning models

## Problem description

**Obstructive Sleep Apnea (OSA)** [1] is a disorder in which breathing is repetitively interrupted during sleep due to collapse of the upper airway. it is the most common sleep-disordered breathing.



OSA illustration 1

Figure 1 – OSA Illustration

**AHI (Apnea-Hypopnea Index)** [2] measures the severity of the disease. it represents the number of times the patient stops breathing per hour.

The disease is prevalent as 100 million individuals worldwide have OSA but only 10%-20% are diagnosed because of the difficulty of its diagnosis. In fact, it implies to sleep one night at hospital with a lot of sensors and the waiting list for this can be very long depending on the country (1 year in Spain).

Because this disease is dangerous and can lead to death, it can be relevant to predict the AHI of a patient with data that are easy to get from an application (weight, height, age, gender, cervical perimeter, speech recordings, facial image or mouth images). This will be the aim of the project

## Machine Learning approach

We will try to solve the problem using both regression and classification approaches. The aim of the regression approach will be to estimate the AHI of a patient using the data that we collect from an imagined app (Actually the data we will use has been collected by a medical centre). The classification approach will be ordered in the same way as the regression on but the prediction we expected will not be the same. Instead of estimating the exact AHI, we want our ML model to put a patient into two classes which will be normal (AHI < 10) or severe (AHI > 30).

### EDA

We will first focus on the clinical data which are easier to use and can be treated by simple ML algorithms instead of a deep neural network which will be required to deals with data as image or speech recordings.

For both of those approach the first things to do in ML is to get the data, analyse it and transform it because it is important, even more in our medical case) to understand our data and find out which features can be relevant to predict the AHI. It avoids constructing a model as a black box that we can’t explain.

For this step I first want to have just one dataset instead of the two givens so the analyse of transformation will be easier. After this I also want to have only feature in my dataset because ML algorithm can only deal with number as word or letter doesn’t mean anything in term of mathematics. After having a clean dataset, I tried to find out which feature can be useful to predict the AHI of a person while using statistical tool and information I can get about this disease.

### Train, test method

Next step was to construct, train and test different model. Because we don’t want our model to do overfitting (try to fit perfectly to the dataset instead of getting a good understanding of it which can be used to other data), it is highly recommended to separate our dataset into 2 which will be used as a training and as a testing dataset.

To test my different models, I used different metrics depending on the approach. Indeed, since the output are not the same, we have to us different method to test our models. I used the MSE [4] and the R2\_value [5] to test my regression models where I used confusion matrix [6] and metrics such as accuracy, sensitivity and specificity to classification models.

### Machine Learning Models

For dealing with the regression approach, I first choose to use the simplest model which are the linear and logistic regression. With the result of those models, I then try to get better understand on the problem and then try to improve those result by using more complex models.

# Data description

The clinical data that we have access to are split into two CVS file:

Une image contenant table

Description générée automatiquement

Figure 2 - CSV 1: Clinical Data

Une image contenant table

Description générée automatiquement

Figure 3 - CSV 2: Gender and Age

Therefore, the first thing to do after import those into python is to fusion them into one panda data frame. To process this operation, we first must remove the incomplete row of our data so we can fusion them correctly.

We now have our only data frame having the columns “AHI, Age, Weight, Height, Cervical, Gender”.

Une image contenant table

Description générée automatiquement

Figure 4 – Joined Data frame

We are now able to analyse them in order to now which features will be relevant to predict the AHI or classify the patient.

# EXPLORATORY DATA ANALYSES (EDA)

The first thing to do when we have data like this is to think if we can create some new one, more relevant, from them. This procedure is called feature engineering and it have several benefits. The first is that we reduce the number of columns we want our model to use and therefore we are simplifying our ML problem and will be more able to understand what our model are doing. The second thing is that as human, we know the signification of our data and the impact they might have on the problem, but the model doesn’t. It’s good to help him by giving to him the features that are more susceptible to be helpful because we know they have a real impact on what we are wanting to predict. For example, according to our context and the info we have about the disease, we know that the BMI [3] is an index more relevant than having the weight and the height because what matter here is a metric that represent the obesity of a people.

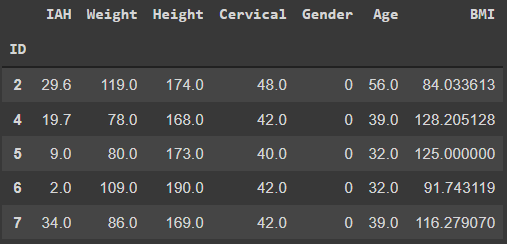


Figure 5 – Adding of BMI

After we did all, we can do to clean our data and try to improve them using our knowledge, it’s time to use statistical tools which are depending on the approach we use.

For built my regression model, I used scatter plot to see if I can find some obvious relation between data and correlation matrix too detect how features have an impact on the AHI. As it is a complex problem, I didn’t find any evidence, but I can at least know which data are more relevant to predict the output.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 6 – Correlation Matrix

In order to treat the classification problem, the tools I used was box plot, histogram and p value.

Chart, box and whisker chart

Description automatically generated

Figure 7 – Box plot Age Vs OSA

Using those statistical methods, I found in both case that the more correlated data are the BMI and the age, but I decided to use all of them at the beginning and then test if I can improve the accuracy by removing some useless data.

# Experimental setup

Once I get a good understood of my data, I was now able create my models and try them in order to get the best result. But the question is how do I evaluate my model? Like I said before, to test a model it is important to make him doing prediction from other data that those used to train it. It avoids having good result and think our model is good whereas it is just fitting our training data but can’t predict well with others. A model like this is useless because we want something that we can use in general, with any input. A thing to do is to split our dataset into to smaller one which will be use respectively for training and testing. In this case, I choose the proportion 80-20, because I want to keep enough data to train my model correctly.

After splitting our data, we can train our model and test it by doing prediction, but we are still not able to evaluate our model. Indeed, we still have any way to evaluate if predictions have been and how far have, they been good or not. The is why we need to use some metrics but as predictions are not from the same type if we are doing regression or classification, we must use different metrics depending on the approach.

To evaluate the predictions of my regression models, I used MSE and the R2 value which compare the MSE with the one we would have if we predicted the average of AHI for every input (simplest way to treat to problem).

To evaluate the predictions of my classification models, I used the accuracy and confusion matrix. Accuracy tells us in what proportion my models get right on the classification while the confusions matrix shows also the false positive and true negative. Confusion matrix is more complete than accuracy to evaluate predictions because we don’t want our model to predict a “severe” AHI on healthy people.

It is now possible to train, test and evaluate the different models we want to try to solve the problems of regression and classification.

# machine learning models

## Regression:

The first model I tried to predict AHI was Linear Regression [7] because it’s the simplest one and it can be used as reference to evaluate more complex model that I used after. The second I choose was SVM [8] model because this kind of model is knowing as a good one when it has to deal with multidimensional problem. As we have many features and therefore many dimensions, I thought this will get good result, but they were worse that those obtained previously.

Because of the bad result of the linear regression problem, I decided to try decisional tree [9] because it handles well non-linear problem and still understandable, but the result wasn’t really better. I therefore used a Random Forest [10] model because I concluded that the problem was very non-linear problem and there, I get good result. But the problem is that kind of model is hard to understand so we have to discuss if a gain of 50% on the R2\_value is good enough if we can’t understand what our model is doing.

## Classification:

Using the same logic for the classification approach, I first tried the simplest model which is the logistic regression [11] and then tried SVM, decisional tree and random forest which are very used and know models for classifications. The last models I tried is the KK-N [12] because find an approximative function is not required with this model, it simply using data. I get almost the same accuracy of 70%-73% with decisional tree, random forest and KNN models which is, taking account the complexity of our problem, not that bad.

# RESULTS

As expected, our models, even the most complex, didn’t get very good result. It’s explicable by the fact that the data we are using are not very correlated to the AHI. We might could have better result using more complex data such as speech recording or facial images and more complex as a deep neural network. In another hand, an accuracy of 70-73% for the classification approach is acceptable.

# CONCLUSIONS

During this project we saw how to implement a procedure to deals with a ML problem.

We first saw how to treat and analyse the data in order to use feature as most relevant as we can so that our models can train on data which have a real impact on what we are trying to predict or describe.

In a second order we saw how exploit our data to train and test our models in the best way. We also saw some examples of metrics that can be useful to know how good our model is doing.

During the third part I tried to explain how we choose an algorithm based on the type of problem we have (the basic model is often a good point of start before testing more complex one). We could have a good illustration of the black box problematics consisting in the fact that we something are given more importance in understand what our ML model is doing than having best result (especially in medical case for instance). Therefore, it’s not so easy to find the best compromise between the complexity and the efficiency of our model.

Finally, we saw that ML is not magic and even is we use the best model we won’t be able to get good result if we are not using the appropriate data.

# REFERENCES

[1] – OSA: https://en.wikipedia.org/wiki/Obstructive\_sleep\_apnea

[2] – AHI: https://en.wikipedia.org/wiki/Apnea%E2%80%93hypopnea\_index

[3] - BMI: https://en.wikipedia.org/wiki/Body\_mass\_index

[4] – MSE: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean\_squared\_error.html

[5] – R2\_value: https://en.wikipedia.org/wiki/Coefficient\_of\_determination

[6] – Confusion Matrix: https://en.wikipedia.org/wiki/Confusion\_matrix

[7] – Linear Regression: https://en.wikipedia.org/wiki/Linear\_regression

[8] - SVM: https://en.wikipedia.org/wiki/Support-vector\_machine

[9] - decisional tree: https://en.wikipedia.org/wiki/Decision\_tree

[10] – Random Forest: https://en.wikipedia.org/wiki/Random\_forest

[11] – Logistic Regression: https://en.wikipedia.org/wiki/Logistic\_regression

[12] - KNN: https://en.wikipedia.org/wiki/K-nearest\_neighbors\_algorithm