

**Course:**

**Predictive and Descriptive Learning**

Half-term Activities Report

**Developing Machine Learning models**

Name: **Gousset Marvin**

Date: 19/10/2021

CONTENTS

[1. Type of machine learning 5](#__RefHeading___Toc33139_644885603)

[1.1. Self-Supervised learning 5](#__RefHeading___Toc32370_644885603)

[1.1.1. What is Self-Supervised learning 5](#__RefHeading___Toc32372_644885603)

[1.1.2. How it is used 5](#__RefHeading___Toc32374_644885603)

[2. DEVELOPING Machine learning models 6](#__RefHeading___Toc33141_644885603)

[3. Data description 7](#__RefHeading___Toc33143_644885603)

[4. EXPLORATORY data analysis (EDA) 8](#__RefHeading___Toc33145_644885603)

[5. Experimental setup 9](#__RefHeading___Toc33147_644885603)

[6. machine learning models 10](#__RefHeading___Toc33149_644885603)

[7. RESULTS 11](#__RefHeading___Toc33151_644885603)

[8. CONCLUSIONS 12](#__RefHeading___Toc33153_644885603)

[9. REFERENCES 13](#__RefHeading___Toc33155_644885603)

LIST OF TABLES

[*Table 1. OSA dataset content* 5](#_Toc18514034)

LIST OF FIGURES

[Figure 1. Sleep Apnea (from [2]) 3](#_Toc18512579)

# 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 labeled but the difference is that we don’t want the model to do clusterisation. Indeed its not possible to have all sentences that can exist labeled because there is too much 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 others algorithms.

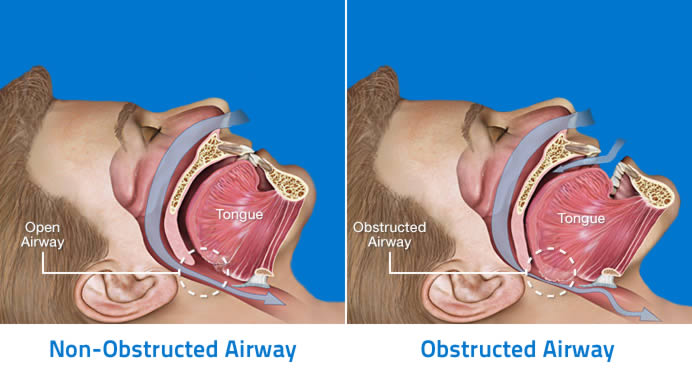
# DEVELOPING Machine learning models

## Problem description

**Obstructive Sleep Apnea (OSA)** 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.

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



The disease is really prevalent as 100 million individuals worldwide have OSA but only 10%-20% are diagnosed because of the difficulty of its diagnosis.

In fact, it imply to sleep one night at hospital with a lot of sensor and the waiting list for this can be very long depending of the country (1 year in Spain).

Because this disease is dangerous and can lead to death, it can be really 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.

## Regression

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

## Classification

The classification approach will be ordered in the same way as the regression on but the prediction we expected will not be exactly the same. Instead of estimate 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). Here we want our model to do a classification task.

For both of those approach the first things to do in ML is to get the data, analyze it and transform it because it is really important, even more in our medical case) to understand our data and find out out which features can be relevant to predict the AHI. It avoid to construct 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 given so the analyze 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.

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 use to other data), it is highly recommended to separate our dataset into 2 which will be used as a training and as a testing datasets.

To test my different models, I used different metrix 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 and the R2\_value to test my regression models where I used confusion matrixe and metrix such as accuracy, sensitivity and specificity to classification models.

For dealing with the regression approach, I first choose to use the simplest model which is the linear one.

# Data description

# EXPLORATORY data analysis (EDA)

# Experimental setup

# machine learning models

# RESULTS

# CONCLUSIONS

# REFERENCES