LEPL1109 - Statistics and Data Sciences
HACKATHON 3 - Classification: Air Quality

Deadline: December 11, 2022

Lastname	Firstname	Noma
Pirot	Alexandre	53811900
Burignat	Bryce	35171700
Goffinet	Dylan	08471900
Defrère	Sacha	51621900
Hautier	Thomas	80162000
Ketels	Mathéo	23782000

Please, read carefully the following guidelines:

- Answer in English, with complete sentences and correct grammar. Feel free to use grammar checker tools such as LanguageTools free and open-source plugin;
- Do not modify questions, and input all answers inside `\begin {answer}...\end {answer}` environments;
- Each question should be followed by an answer;
- At the end of each question, there is the length of the expected answer. This is for your information but it is not too important if you do not respect these recommendations.
- Clearly cite every source of information (even for pictures!);
- For bonus material (additional figures, code, very long equations, etc.), use Appendices;
- Whenever possible, use the `.pdf` format when you export your images: this usually makes your report look prettier[1];
- Do not forget to also submit your code on Moodle.

Contents

[1] This is because `.pdf` is a vector format, meaning that it keeps a perfect description of your image, while `.png` and other standard formats use compression. In other words, this means you can zoom as much as you want on your figure without decreasing image resolution. For simple plots, vector formats can also save a lot of memory space. On the other hand, we recommend using `.png` when you are plotting many data points: large scatter plots, heatmap, etc.

Context and objectives

Description of the project

The air in cities is often polluted by human activities. Vehicular emissions, heating systems, industries... You probably remember those times when the air is so polluted that it is recommended to stay at home, and not do too much sport outside. It is often related to specific weather conditions (e.g. not enough wind to blow air pollution away).

Besides, one could wonder: is it healthy to go running in a city? Do the benefits of sport balance the fact that you are breathing polluted air? Studies show that the answer is yes, as long as you avoid busy roads with dense traffic.

Air quality is a crucial issue in our modern society, and its effect is poorly anchored in common knowledge. As engineers of tomorrow, you must be aware of the impact of thermic engines and industries on human health.

In this hackathon, we will learn to quantify air quality. Wikipedia says that smog (or smoke fog) is composed of nitrogen oxides, sulfur oxide, ozone, smoke and other particulates. How to quantify from those different features?

People use the Air Quality Index (AQI). This is a natural number running from 0 to 500+, the lower the better. AQI accounts for the concentration of important pollutants and particles. We usually distinguish 6 classes of air quality, from good to severe (see table below). In this hackathon, we will classify the AQI between only 2 classes: good or bad.

Good (0–50)	Minimal Impact	**Poor (201–300)**	Breathing discomfort to people on prolonged exposure
Satisfactory (51–100)	Minor breathing discomfort to sensitive people	**Very Poor (301–400)**	Respiratory illness to the people on prolonged exposure
Moderate (101–200)	Breathing discomfort to the people with lung, heart disease, children and older adults	**Severe (>401)**	Respiratory effects even on healthy people

Objective

The project aims to train a binary classifier to estimate the air quality index (AQI) based on the air concentration of certain pollutants. Note that the AQI is initially defined as a natural number between 0 and 500+. However, to better address the problem, we propose to classify the air quality as poor (labeled 0) or good (labeled 1).

Questions and Answers

1 Preprocessing

1.1 Import the data set

CELL N°1:

Import `gas_measurements.csv` in a **pandas data frame**. **Obtain** a brief description of the data (*size, variables type, missing values, etc.*).

Question 1.0 :

Describe, briefly, your data set (*size, variables type, missing values, etc.*).
Expected answer length : 2 or 3 lines.

Answer to 1.0 :

Our data set consists of a table of 44769 rows (*i.e.* measurements) and 11 columns which represents 7 concentrations of chemicals (type float64), 1 column for the index of the row of the measurement (type int64), one for the datetime of the measurement (type datetime object), one for the AQI (type AQI object), and one to indicate if the measurement has been done in a big urban city (type int64). There is also a total of 51768 missing values, all of which are from the chemicals concentrations columns.

CELL N°2:

Based on your observations, **justify and proceed** to the a priori deletion of **two** troublesome features.

Question 1.1 :

Which features did you delete? Justify.
Expected answer length : 1 line.

Answer to 1.1 :

To determine which two features we will delete, we will look at the percentage of NaN (*Not A Number*, or missing data) for each feature. We can see that for the features SO_2 and NH_3, the respective percentages of missing data are 51.75% and 51.37%, which leads us to delete those 2 columns.

1.2 Split the data set

CELL N°3:

Split the data set into a *training* and a *test*. The proportion of each subset is at **your own discretion**.

Question 1.2:

What are the drawbacks (if any) of choosing a small test set? On the contrary, what are the consequences (if any) of a relatively large testing set in this context?
Expected answer length : 6 lines.

Answer to 1.2:

It is important to find the good ratio between the sizes of the training set and the testing set: a too big training set means our model will 'learn by heart' the dataset on which it is training, and perform poorly on non-similar datasets as it considers too much non-relevant information; on the other hand, a too small training set means our model has not enough data to learn from to be robust, it is underfitting its data (did not have enough data to train correctly), and will perform poorly. We will choose a ration of 80/20 for Training/Testing set as it is the optimum values [1].

2 Exploratory data analysis

2.1 Target proportions

CELL N°4:

Analyze the distribution of the features AQI in the binary scenario. Namely, air quality is considered either

1. **bad** (AQI : Poor, Very Poor, Severe)
2. **good** (AQI : Good, Satisfactory, Moderate)

We conduct the analysis on the *training set*, avoiding therefore any modelling decision based on *unseen* data (*test set*). In most cases, we assume that the distribution of this latter set stays similar to the *training set*.

Question 2.1:

Are the binary classes balanced? What are the proportions of data in each class? Briefly, justify your answer and add a visualization.
Expected answer length : 1 line and 1 figure.

Answer to 2.1:

The binary classes are not balanced, as seen on the graph, the training dataset is favoring Good AQI.

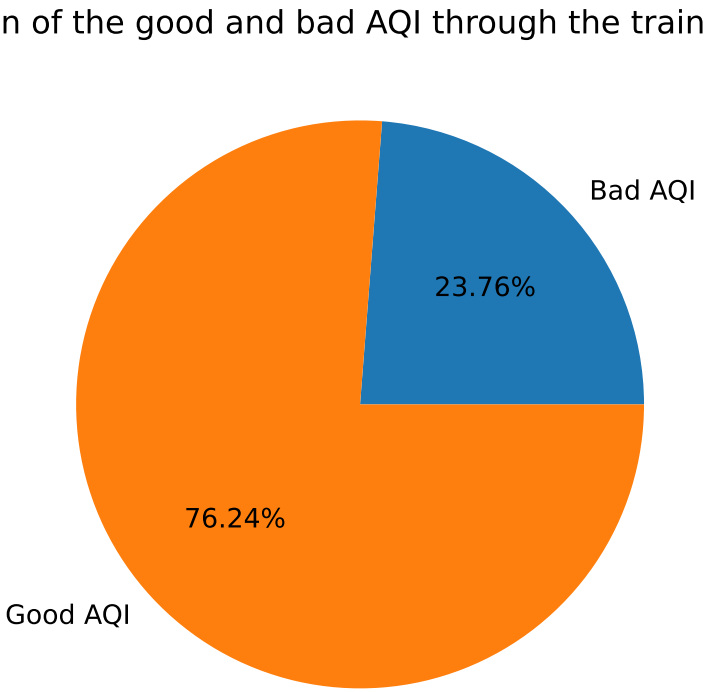

Figure 1: Distribution of the AQI

2.2 Cyclical features

CELL N°5:

Establish a *cyclical* feature transformation and store the new features in the corresponding variables. Many features are cyclical in nature. One example is time: months, days, weekdays, hours, minutes, seconds etc. are all cyclical.

Question 2.2:

Explain your transformation and the need to use cyclical features in some cases. What is the point to not simply encode features in a categorical way (e.g. morning=0, afternoon=1, evening=2, night=3)?
Hint: what happens between 23:00:00 and 00:00:00 ?
Expected answer length : 8 lines + 1 figure.

Answer to 2.2:

The transformation takes the column *Datetime* and extract the hour, before converting the hour into 2 other features: the sinus of the hour (by the formula Hour in sin form $= \sin\left(2\pi\frac{\text{hour}}{24}\right)$ (the same is done for cos) to highlight the periodicity of the results of the measurements. The need for cyclical features comes from the fact that lots of events have a recurring pattern. For example, we could say that the quality of the air should be the worse during the early-morning/morning as people go to work by taking their car. Less people drive at 2 in the morning, meaning that the AQI should be better. To summarise, we use cyclical features to see if there is any recurrence in a given time period,*i.e.* to see if the data follow a kind of sinusoidal/cosinusoidal pattern, which would explain the cyclic nature of them.

2.3 Correlation matrix

CELL N°6:

Compute and **plot** the correlation matrix.

Question 2.3:

With the help of your **plot** and **numerical results** of cell 6, **write** your observations down.
Expected answer length : 3 lines + 1 figure.

Answer to 2.3:

We see some pretty big correlation between $PM10$ and BigUrbanCity for example, between NOx and $PM2\_5$, BigUrbanCity, and $PM2\_5$, indicating that they give a pretty similar information.

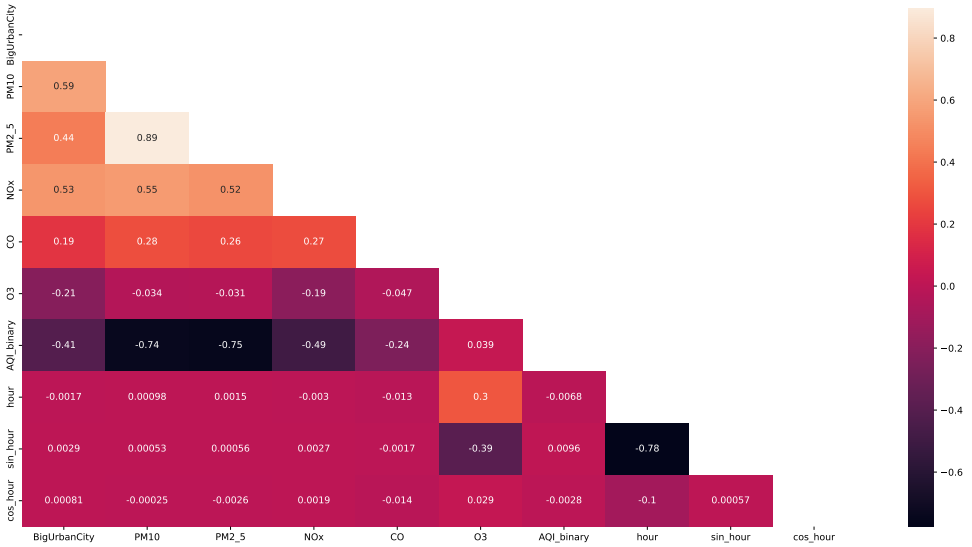

Figure 2: Correlation matrix between all the relevant features

2.4 Feature selection

CELL N°7:

Establish a feature selection strategy[a] and store the feature names in a list. Furthermore, investigate on redundant features which can be removed as well.

[a] Selection of the n first correlated features, and/or setting of a correlation threshold or any other rule.

Question 2.4:

Explain your strategy of cell 6. – *Expected answer length : 3 lines.*
Is it appropriate to use the correlation matrix to select (or not) cyclical features? – *Expected answer length : 1-2 lines.*

Answer to 2.4:

We will have as a reference the AQI_binary, and will choose the features with the highest pairwise correlation. We decide to take the 5 with the highest correlation, as the 5th has a correlation of 24%, and plummet to 4% for the 6th.

2.5 Data scaling and normalization

CELL N°8:

Split your *training* and *test* sets into their respective features set ($\mathbf{X}$) and a binary target variable ($\mathbf{y}$). **Standardize** the features sets.

Remark 1. The object scaler used to standardize the training set, should be the one used on the test set!

Question 2.5:

Why do we scale data? **Justify properly**, whether it is necessary or not for your *feature set* ($\mathbf{X}$), and which scaler you use.
Expected answer length : 5 lines.

Answer to 2.5:

We need to scale as all the variables used don't have the same scale. This means that they do not contribute equally to the model fitting. If we let the variables unchanged, it will end up biasing our model and giving more importance to the variables that have the highest scales. To get rid of the scale problem, we standardize our model, as by doing so, the mean of all our variables is 0 and their variation is 1. This means they are equally important.

3 Model selection

3.1 Precision, recall and F1 score

CELL N°9:

Implement the *precision, recall* and *F-measure* metrics based on the confusion matrix. Please follow the specifications in the provided template.

3.2 Model evaluation

CELL N°10:

Implement `evalParam`, which evaluates, using a **k-fold** cross-validation, a list of `scikit-learn` models. The score is the harmonic mean F1. The function must be **scalable**. Put differently, it must handle m methods, and a variable list of their possible hyperparameters configuration.

In addition to the list of *models* (**methods**) and their list of *hyperparameters* (**param**), the function takes as arguments the *features set* (**X**), *target variable* (**y**) and *the number of folds* (**cv**).

It returns an array *score* such that

$score[i][j]$ = average F1 over the folds, using method i with parameters configuration j.

Question 3.1:

Explain the idea of K-fold cross-validation and why it is useful. How the choice of K (in the cross-validation) impacts the bias and the variance of the scores obtained on the different folds? Choose and justify the number of folds you consider in this project.
Expected answer length : 18 lines.

Answer to 3.1:

The K-fold cross-validation allow us to get the best learning result. The idea behind it is to divide our data in K subsets of the same size. Then, we can do K different learning experiments, where each time one subset is use as the validation set and the $K-1$ remaining subsets are combined as one training set. One the K experiments are done, we can compute the average result of the test validation. Using this method, each fold is used once as a test and the other times as training data. This means that the entire data set is used to train our algorithm.
The choice of K impacts the bias and the variance of the score obtained as increasing the K as much as possible increases the bias and decreases the variance and take more time to compute (i.e. the higher the K is, the higher the bias is). However, if the K is too high, the test sample will not be broad enough to represent all the data, which is why the bias increases.

3.3 Model selection and parameters tuning

CELL N°11:

Run your function `evalParam` and **evaluate** the models: *linear regression*, *logistic regression* and *K-nearest neighbors*[a].

Investigate the effect of the following parameters: `n_neighbors`, `weights` and `p` in KNN; and `C` in logistic regression.

[a]The models are already implemented in `scikit-learn`.

Question 3.2:

Explain your methodology of model evaluation. More precisely, explain which hyperparameters you tune and the values you test for each of them. Next, provides the best hyperparameters configurations for **each models** as well as their CV F1 score.
Expected answer length : 12 lines.

Answer to 3.2:

Due to organisational issues on our part, we were not able to complete this question, nor here or in the notebook.

Question 3.3:

Based on your answers to previous questions, select a final model that you will keep as classifier. Justify.
Expected answer length : 3 lines.

Answer to 3.3:

As the logistic model is to be provided with more parameters, there is more possibility to fine-tune it and so the final model we would have selected if our code worked would have been the Logistic one.

3.4 Precision-Recall curve and thresholding

Question 3.4:

What happens to the precision and recall (of any method) when the threshold tends to 0? And when it tends to 1? Explain and, if possible, establish a link with Question 2.1.
Expected answer length : 8 lines.

Answer to 3.4:

Let's recall (no pun intended) that

$$\text{Precision} = \frac{TP}{TP + FP}, \text{ and Recall} = \frac{TP}{TP + FN}. \tag{1}$$

The threshold tending to 0 means that we consider all AQI to be good (as we decrease the number of False Negative and increase the number of False Positive). This leads to the Precision (the fraction of relevant information in all the information) will go to 0. The Recall on the other hand will go to 1 as as FN goes to 0, the fraction goes to 1. This situation means that we would have identified correctly all good AQI as good (Recall = 1), but would have no information on whether the AQI are *actually* good or bad (Precision = 0)
On the other hand, if the threshold tends to 1, we consider all AQI to be bad, we have the opposite scenario with the number of False Negative increasing and the number of False Positive Increasing. This leads, by a similar justification than above, to a Precision of 1 (all the good AQI retrieved are effectively good), and a Recall of 0 (we have missed an enormous amount of good AQI that have been classified as bad).

Question 3.5:

Explain which precision/recall trade-off you prefer to have for the specific task asked in this hackathon: determining whether air quality is good based on some concentrations. How should you adjust the threshold of your model to bring it closer to the desired trade-off? Should it be above or below the default threshold value of 0.5?
Expected answer length : 4 lines.

Answer to 3.5:

We want to be sure that if the concentration is low, we have a good AQI (AQI_binary = 1). That means that out of all the link concentration-AQI that we have recuperated, we have the maximum that indicated an AQI_binary of 1 when the concentration is low. In other words, we want to maximise the precision, and so have a threshold superior to 1.

4 MODEL TESTING

CELL N°13:

Evaluate the performances $\{precision,\ recall,\ F-measure\}$ of your final model on your test set (with the default threshold 0.5 used in the cross-validation).

Question 4.1:

Use the test set to estimate the precision, recall and F1 score of your final model and validate its performance on unseen data. Observe if the scores are similar to the ones estimated with your cross-validation. Are you satisfied by the performance of your classifier, in view of the task for which it will be used?
Expected answer length : 3 lines.

Answer to 4.1:

Due to organisational issues on our part, we were not able to complete this question, nor here or in the notebook.

Appendix : LATEX tips

A Demo

Question Demo:

Can you show me what I can do?

Answer to Demo:

This is how I answer to Demo. I can cite [?] content that I use or refer to A. I can also reference images such as in Figure 3, or equations with (2) or Equation 2.

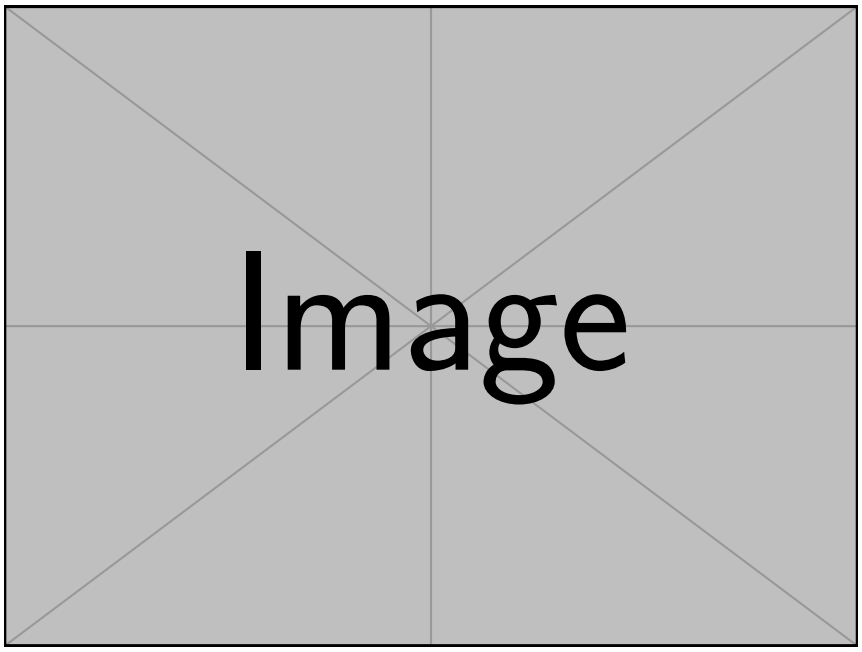

Figure 3: Demo caption.

$$E = mc^2 \tag{2}$$

If you wish to present code samples, you can either use the Listing 1 format or use inline code `import numpy as np; x = np.arange(10)` if this better suits your needs. However, we recommend putting your code in the Appendices.

```
1  import numpy as np
2
3  x = np.arange(10)
```

Listing 1: My super code.

Note: syntax highlighting for code is provided by the `minted` package. If you are not using Overleaf, you might need to install some requirements before it can work.

References

[1] Golamy et al. *Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation.* 2018.