Volvo Practical Test

João Felipe Humenhuk

January 2022

1 Introduction

This document's goals are to summarize the procedures and analyses performed for the first exercise of the Volvo's Practical Test and the arrived conclusions. All graphs, models and calculations were performed using Python 3.9 and the data set used encompasses information about the Particulate Matter 2.5 (PM2.5). As described in the practical test document:

> Particulate matter (PM) - also known as Atmospheric aerosol particles - are microscopic solid or liquid matter suspended in the atmosphere of Earth. PM2.5 are fine particles with a diameter of 2.5 micrometers or less. They have impacts on climate and precipitation that adversely affect human health. In other words, it's used as a measure of pollution. PM2.5 readings are often included in air quality reports from environmental authorities and companies.

2 Data Set Description

The data set used in the analysis can be found in `https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data` and was made available by Song Xi Chen, csx '@' gsm.pku.edu.cn, Guanghua School of Management, Center for Statistical Science, Peking University. It was extracted using the BeautifulSoup4 library and contains hourly data of the PM2.5 concentration in Beijing, captured between Jan 1st, 2010 to Dec 31st, 2014. More information can be seen in Table 1.

It is possible to notice that there are 13 attributes, they are represented as time series, and the number of instances is 43824. The attributes encompassed in the data set are the row number (i.e., an identifier), year, month, day and hour of the data collection, the concentration of the PM2.5, the Dew Point (DEWP), the Temperature, the Pressure, the Combined Wind Direction (CBWD), the Cumulated Wind Speed (IWS), Hours of Snow (IS) and Hours of Rain (IR). There are a total of 2067 missing values, all encountered in the PM2.5 concentration attribute and they are denoted as "NA".

Data Set Characteristics	Multivariate, Time-Series
Attribute Characteristics	Integer, Real
Associated Tasks	Regression
Number of Instances	43824
Number of Attributes	13
Missing Values?	Yes

Table 1: The data set description found in the source website

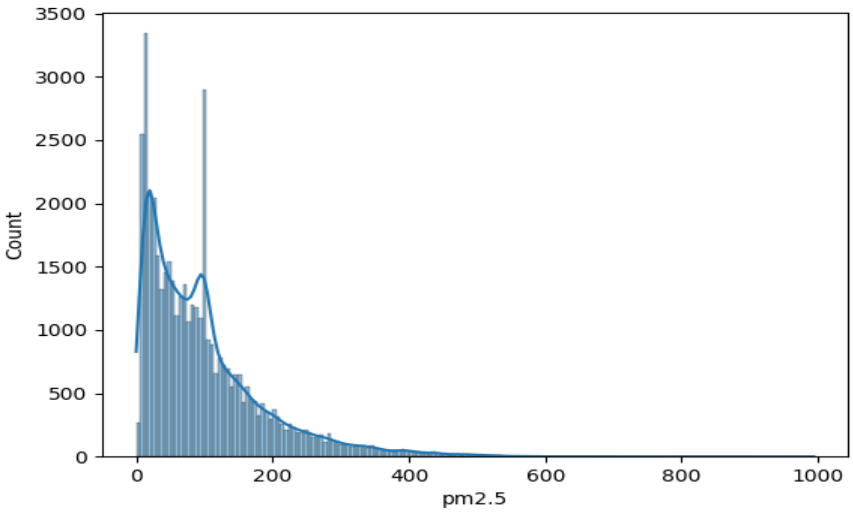

Figure 1: The PM2.5 presented in a histogram

3 Exploratory Data Analysis

The first step was to do an uni-variate analysis which can be seen in Figures 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 10, 12, 13. Starting with the attribute representing the PM2.5 concentration we can verify that the majority of the instances have a concentration lower than 200ug/m$\hat{3}$, it has a central tendency of approximate 30ug/m$\hat{3}$, low variance, and two modes more in evidence of 13 and 100ug/m$\hat{3}$. Looking at the Figures 1, 2 we can conclude that it is right skewed and has a high amount of outliers.

Analysing the DEWP, temperature, and pressure we can see that they all have high variance, do not have a central tendency, and do not have outliers.

Since the CBWD is a polynomic attribute, a bar graph is more suitable for it. Looking at Figure 9 it is possible to notice that there is a greater amount of **SE**, followed by **NW**, **cv**, and **NE**, with the amount of **NE** being approximately half of **cv**, the second smaller amount.

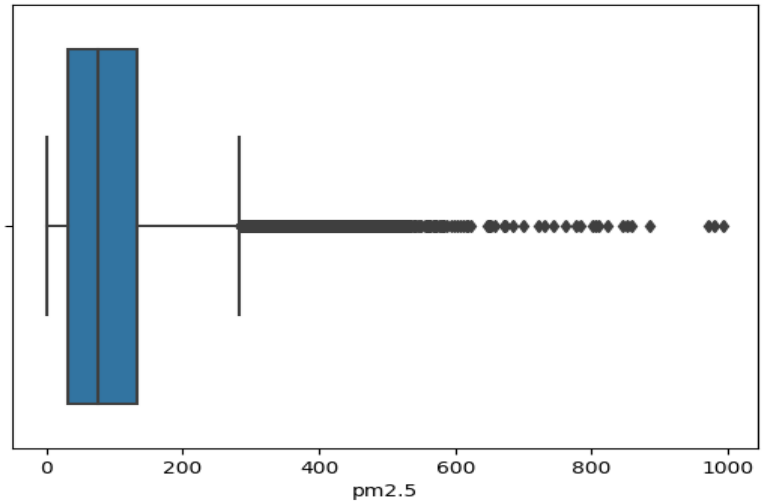

Figure 2: The PM2.5 presented in a boxplot

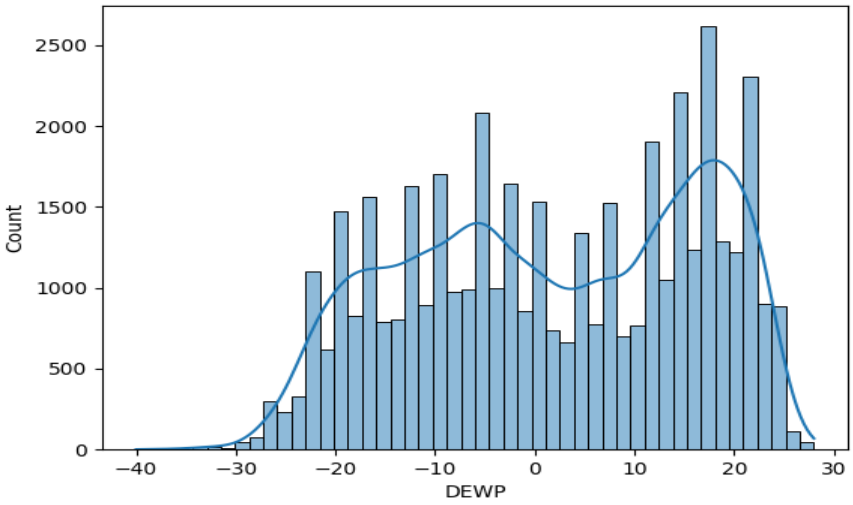

Figure 3: The DEWP presented in a histogram

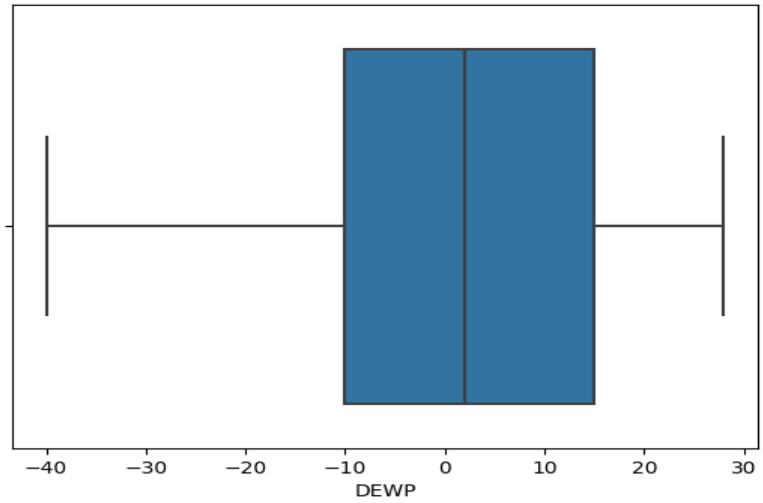

Figure 4: The DEWP presented in a boxplot

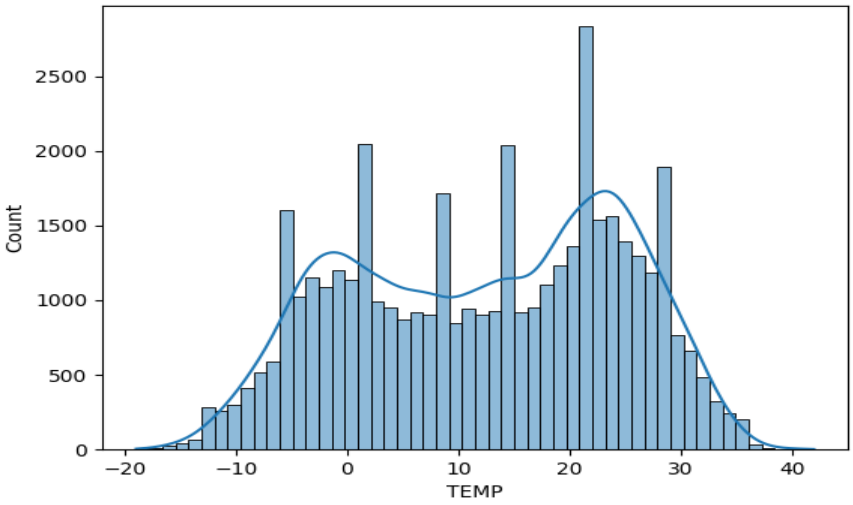

Figure 5: The temperature presented in a histogram

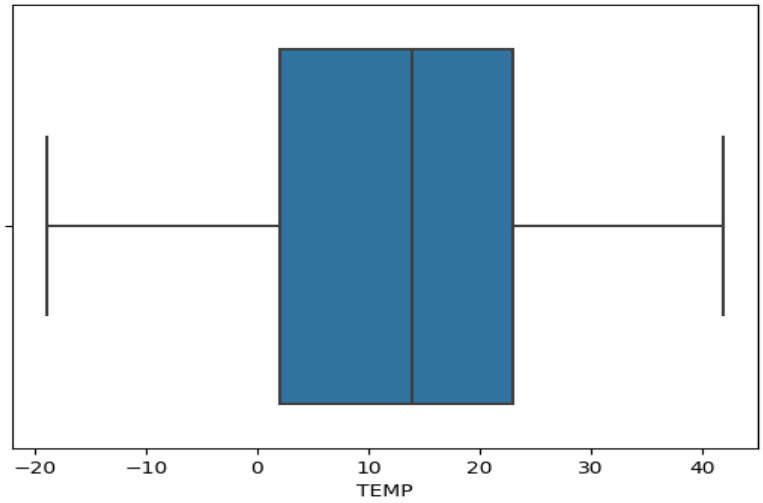

Figure 6: The temperature presented in a boxplot

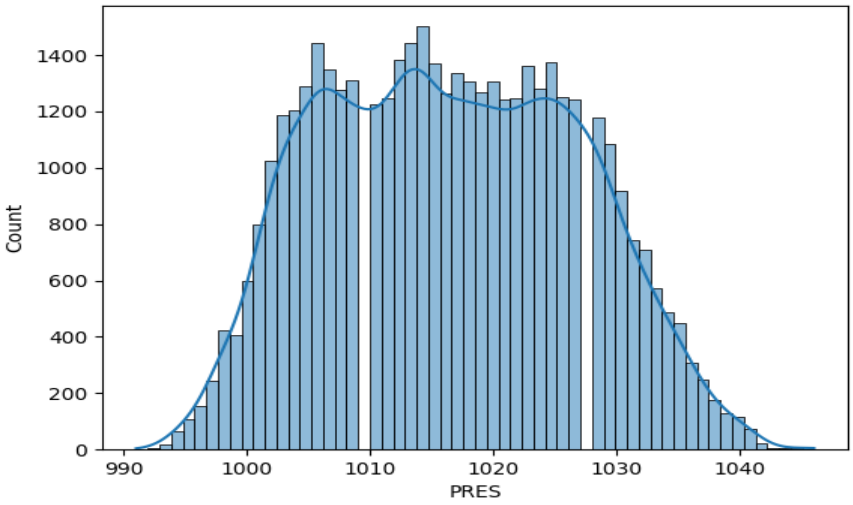

Figure 7: The pressure presented in a histogram

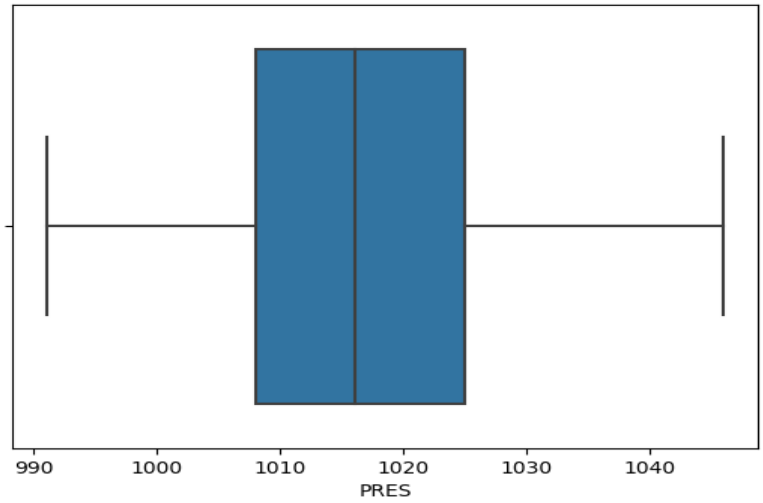

Figure 8: The pressure presented in a boxplot

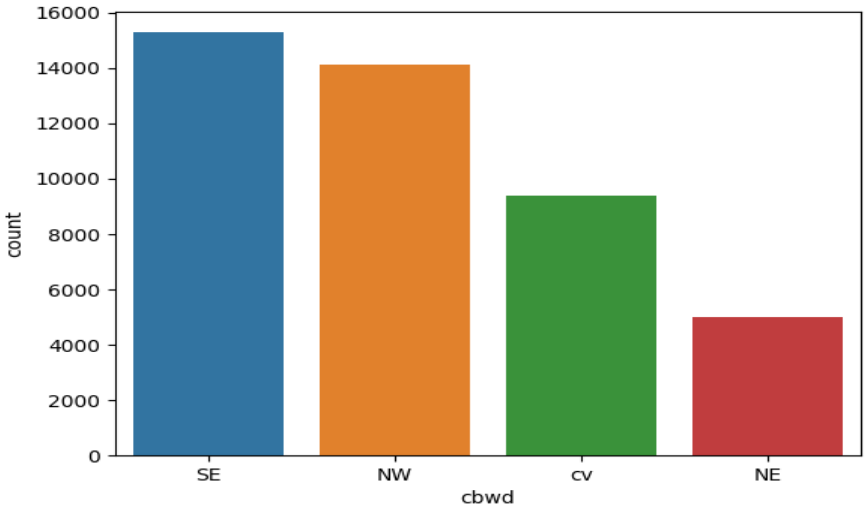

Figure 9: The CBWD presented in a bar graph

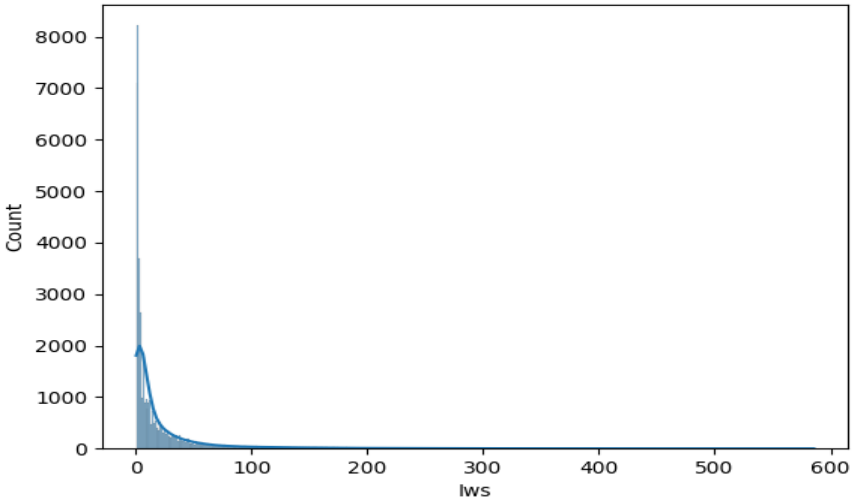

Figure 10: The IWS presented in a histogram

Analysing the IWS attribute, the only continuous variable, we can conclude that the majority of instances have a value lower than 13m/s, its central tendency is approximately 1m/s, it has low variance, and have a mode of 1m/s. Looking at the box plot it is possible to identify it with a right skew and a high amount of outliers.

The two left attributes IS and IR are very similar, both have lots of instances with zero hours, meaning that when they have some value aside from zero it is consider as an outlier.

Wishing to evaluate the attributes behaviors compared to the other variables, a correlation test was performed among them. The result is illustrated in Figure 14. By visualizing the heat map, it is possible to see that there is a high positive correlation between DEWP and the temperature, and two high negative correlations between the pressure and the DEWP, and the temperature with the pressure. These correlations are logical because these three attributes are correlated in Physics. Aside from these correlations there are not any other ones.

A pair plot was also deployed to analyse the correlation of the attributes, which can be seen in Figure 15. In this figure we can see the distribution of the attributes, already approached in the previous section, and the correlation between the attributes in a instance by instance view.

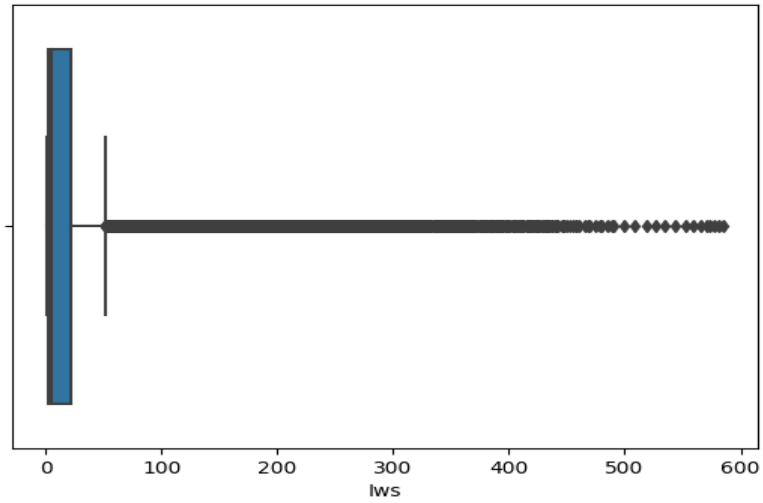

Figure 11: The IWS presented in a boxplot

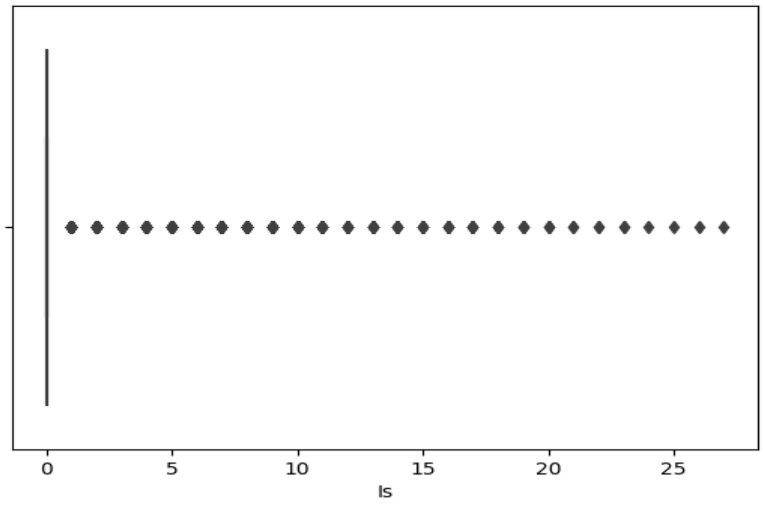

Figure 12: The IS presented in a boxplot

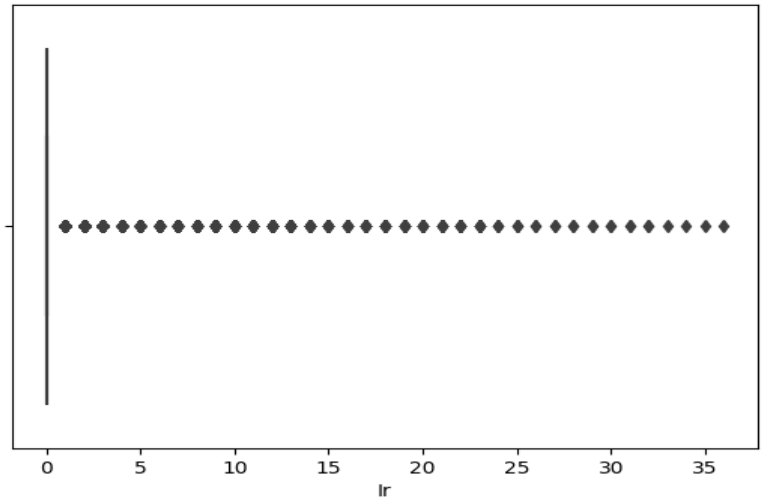

Figure 13: The IR presented in a boxplot

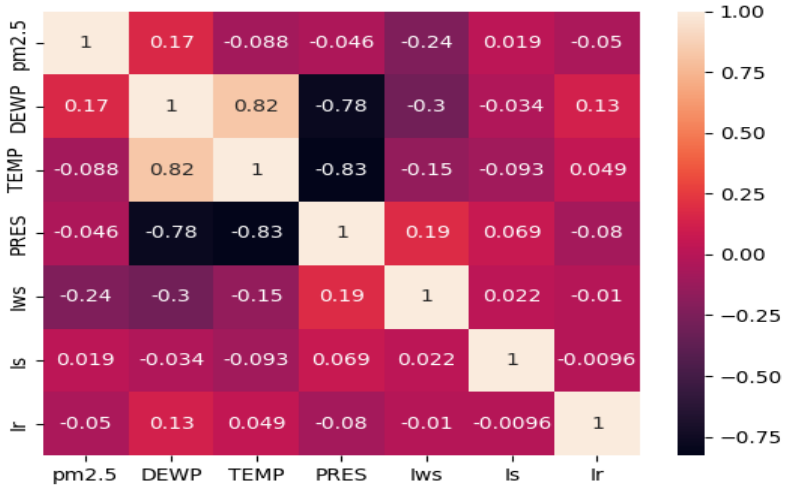

Figure 14: The correlation represented in a heat map plot

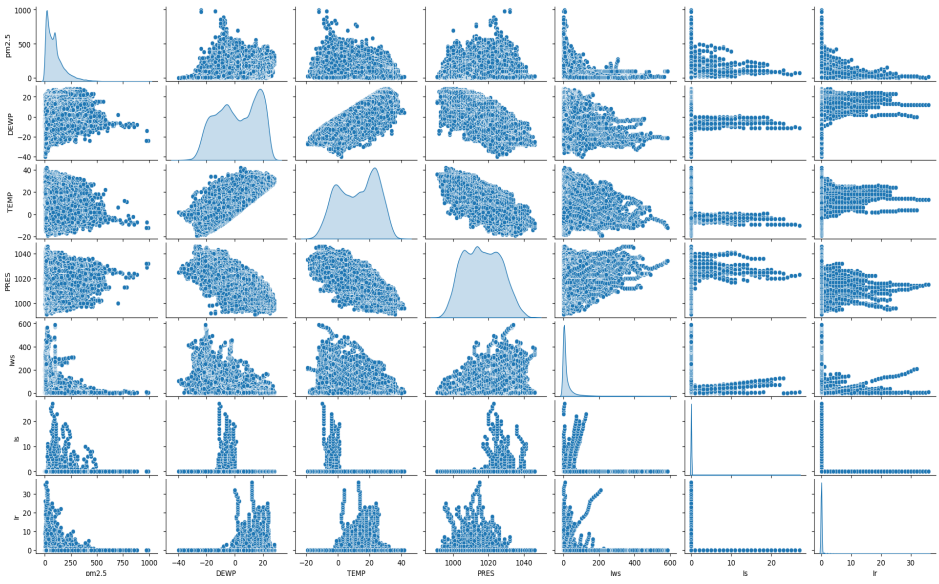

Figure 15: The correlation represented in a pair plot

4 Predictive Analysis

To predict the behavior of the PM2.5 attribute using the other attributes a regression model was created and trained. The first step consisted of encoding the polynomic attribute CBWD to numerical values, followed by the scaling of the attributes and the separation of the data into input, output, training and testing. Later a Long Short Term Memory (LSTM) model was created with a dropout percentage of 20% to diminish over-fitting. The loss function utilized was the Mean Absolute Error (MAE) function, and the optimizer was ADAM. The model trained with an early stopping policy, and its loss values for training and testing can be seen in Figure 16. It is possible to notice that they present a similar behavior which can indicate that the model could be general enough to predict unseen data. By deploying the test set using the trained model, the value of 33.05 was obtained by comparing the predictions with the true outputs and doing the square root of the Mean Squared Error (MSE). Since the values of the PM2.5 attribute vary between 0 and 300, removing outliers, the error is equivalent to approximately 10% of the maximum value which can be consider a good performance for a model trained in a single algorithm and without any fine-tuning.

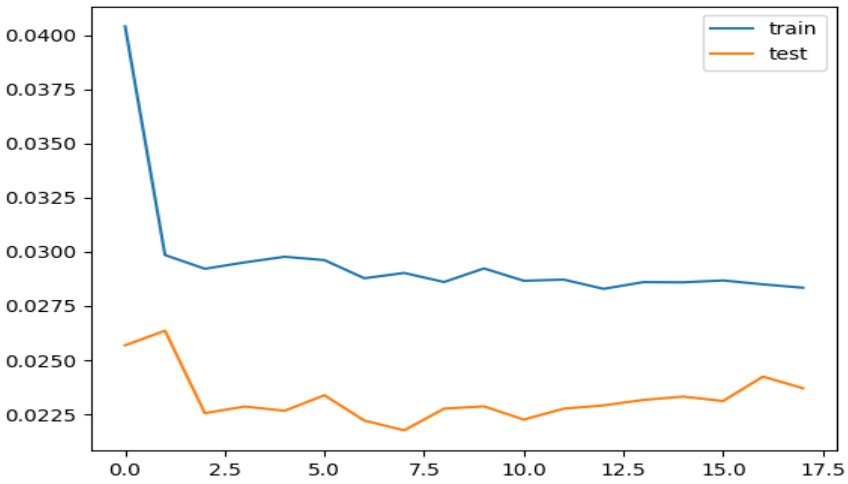

Figure 16: The training and testing loss values