Data analysis – PM 2.5 particles

Nikola Pupovac, IN31/2018, nikola.pupovac99@gmail.com

I. Introduction

Our goal with this data analysis is to find the strongest links between the increase or reduction in the amount of PM2.5 particles in the air in Chengdu, China, and weather conditions.

PM2.5 particles are air pollutants that can harm people if their concentration rises above a specific level (35.4 g/m3).

They are microscopic particles in the air that decrease visibility and make the air appear hazy in big quantities.

They've also been linked to an increase in lung cancer sufferers.

To put it another way, we will be monitoring and studying air pollution in Chengdu, China.

II. Database

The database of recorded meteorological conditions and PM2.5 particle levels in the air has 52584 samples; however, because many of these samples are invalid, this data set only includes 26861 valid samples for analysis.

PM2.5 particle concentration, condensation temperature, temperature, humidity, air pressure, wind direction, cumulative wind speed, precipitation amount, and cumulative precipitation amount were all measured every hour.

The wind direction column has category features, while the rest of the columns represent numerical features.

III. Data analysis

As most columns had missing samples that followed one another, samples that lacked values were removed from this set. By completing the data arbitrarily or using algorithms, we can give our machine learning algorithm a false picture of what the data should look like.

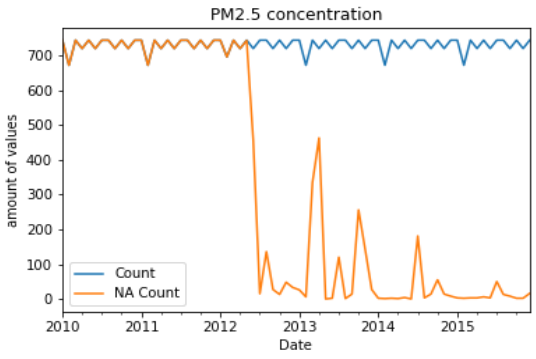

Figure 1: Comparison of the amount of samples with the amount of unspecified samples of the column "PM2.5 concentration"

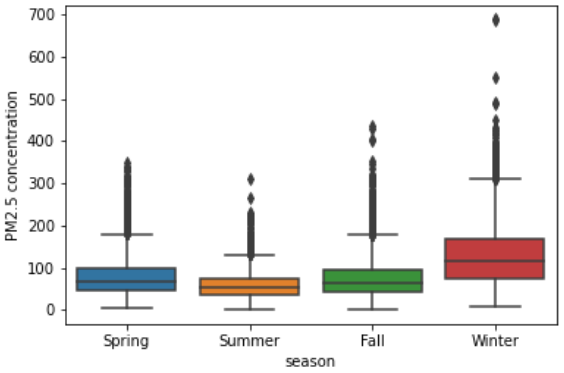

Figure 2: Amount of PM2.5 particles shown by seasons.

Figure 2 shows that most samples had PM2.5 particle values ranging from 0 to 300, depending on the season, and that deleting samples with PM2.5 particle levels greater than 380 removes only 0.01 percent of samples that we can consider outliers.

The number of PM2.5 particles is also lowest during the summer, equal throughout the spring and fall, and convincingly largest during the winter, as seen in the graph.

Significantly increasing fossil fuel consumption, which affects air pollution, could be the source of this.

PM2.5 particles analysis

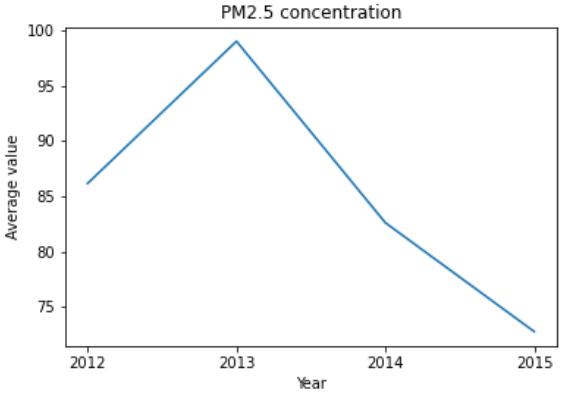

Figure 4: Graphical representation of the distribution of PM2.5 particles by year

In Figure 4 we can see the average number of PM2.5 particles over the years, where we observe that the air was the most polluted during 2013, as well as that it decreased significantly over the years. The average value of particles for all years was 84.31, while the standard deviation was 56.86. Of course, we must take into account that these are the values obtained after removing the outliers.

Analysis of the remaining attributes

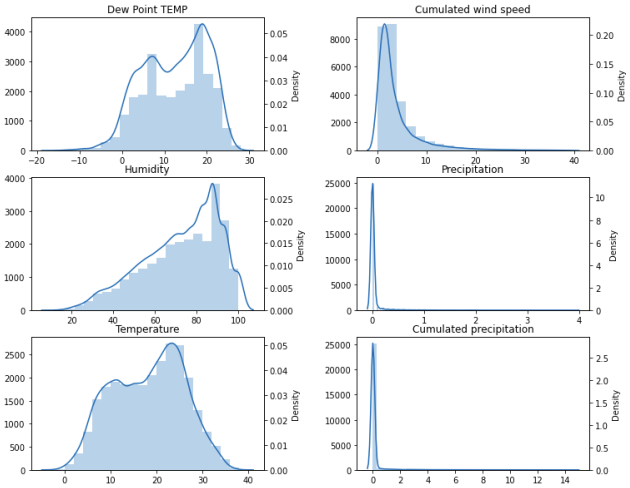

Figure 4: Graphical representation of attributes distribution

At the temperature we can notice that the normal distribution is valid, while at the condensing temperature as well as the air humidity it is tilted to the right. Cumulative wind speed, precipitation amount and cumulative precipitation amount have a distribution inclined to the extreme left.

Correlation analysis

Table 1: PM2.5 particles correlation

kolona	PM2.5 cumulation
Dew Point TEMP	-0.32
Humidity	0.15
Pressure	0.25
Temperature	-0.41
Cumulated wind speed	-0.19
Precipitation	-0.09
Cumulated precipitation	-0.11

The connections between the number of PM2.5 particles and the other columns are numerically presented in Table 1.

It is observable that the number of PM2.5 particles (that is, the inverse correlation) with temperature, as well as the condensation temperature, is the largest correlation, followed by the correlation with air pressure and the inverse correlation with cumulative wind speed.

We can see that the results make sense based on the foregoing, that is that the smallest amount of particles is found during the summer and the biggest during the winter.

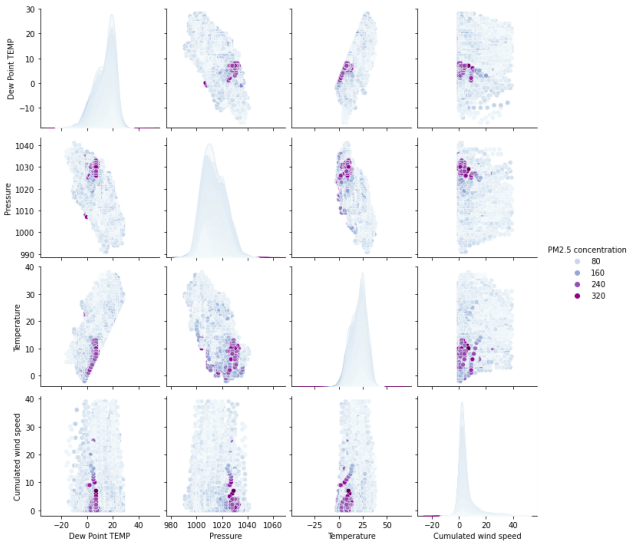

Figure 5: Graphically depicted correlation between the amount of PM2.5 particles and the most correlated attributes with it.

IV. PREDICTION

90% of the samples were used to train the model, and 10% of the samples were used to assess its performance, in order to forecast the amount of PM2.5 particles in the air from the data set.

Binary features have been created from categorical features.

Several algorithms were utilized during the training, with the Lasso model demonstrating the highest prediction accuracy.

Error analysis

Tabela 2: Error measurement

kolona	vrednost
MSE	2244.35
MAE	34.92
RMS	47.37
R2	0.25
R2 adjusted	0.25

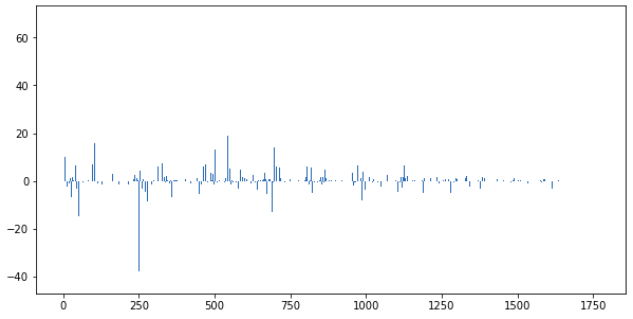

Figure 6: Graphical representation of Lasso coefficients of linear regression model

V. Conclusion

As can be clearly seen from the predictions based on the models used, with the given data we cannot perform a quality prediction of air pollution based on data on condensation temperature, temperature, humidity, air pressure, cumulative wind speed, precipitation and cumulative precipitation.