Predicting Air Quality in Northeast Los Angeles with Machine Learning Methods

Joel Saarinen

December 2019

1 Introduction

Wildfires, industrial emissions, and a warm average temperature are among a few of the many factors that contribute to subpar air quality in Los Angeles, which consistently ranks as having some of the worst pollution in the nation. This is of course not ideal because has negative effects on human health.

In addition to breathing minimally polluted air, another dimension to a healthy life includes physical exercise, which is often fun to do outdoors. However, doing so in a city such as Los Angeles may carry additional risks, since one consumes more air (and the pollutants it contains) during exercise. The risks associated with outdoor activities such as running could be mitigated by choosing the day and time of day on which to exercise outside. Given the fact that exercise is important, and that many live busy lives, it may be helpful to have advance knowledge of when during a given week air conditions would be acceptable for exercise so that one can plan ahead. This can be done through a basic machine learning model.

2 Background Information & Aim of Model

There are many pollutants that might be present in the air, and I chose to work with PM2.5 (particulate matter 2.5), since it is a common pollutant for which data was widely available. This model aims to predict, for a given

hour on a given day, whether the PM2.5 concentration exceeds 20 μg/m3. The recommended quality of air that a person has breathed, on average, over the course of the year, is 12 μg/m3 (1). However, since in Los Angeles, the average PM2.5 concentration of air in a year is 18 μg/m3, living (and exercising) in a way that would result in breathing air with a PM2.5 concentration of 12 μg/m3 over the course of a year is difficult, assuming the average person does not have filtration devices in home environments (where people spend most time in the day) that improve the air quality. Since one tends to consume more air when doing outdoor exercise, and thus consume more PM2.5, living and running outdoors a few times a week in a given environment will result in a person having breathed, on average air with a concentration of PM2.5 that is slightly higher than the average concentration of PM2.5 in the air during that week. If the weekly average of air in Los Angeles is 18 μg/m3, for example, living in Los Angeles and exercising twice a week outside may mean the average air the person has breathed in the week is 20 μg/m3.

My target is, then, to have the average of the air that I have breathed throughout the week to have a PM2.5 concentration of 20 μg/m3. This is achievable since the average yearly PM2.5 concentration of Los Angeles air is 18 μg/m3, which could be broken down into 52 weeks of breathing air of 18 μg/m3 (although this is more unlikely, if there is some standard deviation). The task, then, is to figure out what combination, in terms of duration and amount of runs per week, of runs I could do such that the average air I breathe jumps from 18 μg/m3 (LA average) to 18 (goal).

We assume then that one breathes air with PM2.5 18 μg/m3 for an entire week. A person breathes about 7L of air per minute (2), which is 0.007 cubic meters. This means that I would be breathing approximately 0.007 x 18 = 0.126 μg of PM2.5 every minute, which would be 60 x 0.126 = 7.56 μg every hour, 181.11 μg every day, and 1270.08 μg each week.

Using the same method, for a person that is breathing air that is PM2.5 20 μg/m3 for the week, they would be breathing 0.007 x 20 = 0.14 μg/minute. This means 8.4 μg per hour, 201.6 μg per day, and 1411.2 μg per week.

After some guessing and checking, I arrived at the conclusion that I could do two runs a week in air with PM2.5 less than or greater to 18 μg/m3. This is because it is estimated that a person running at 70% VO2 max (easy running pace) for about three hours inhales the same volume of air as a sedentary person

would over the course of two days (3). A sedentary person in two days, in 18 μg/m3 concentrated air would then be consuming 2 x 181.11 μg/day = 362.22 μg, which is how much would be consumed during a three-hour run. I tend to go on 30-40 minute runs, so for each 40-minute run, I would be consuming 80.493 μg of PM2.5. For a 35-minute run, I would be consuming 70.432 μg.

In a day with a 40-minute run, I would be breathing: (23 hrs x 7.56 μg/hr = 173.88 μg) + 80.493 μg (40 min run) + (0.333hrs after run x 7.56 μg/hr = 2.52 μg) = 256.893 μg PM2.5

In a day with a 35-minute run, I would be breathing: (23 hrs x 7.56 μg/hr = 173.88 μg) + 70.432 μ (35 min run) + (0.4167 hrs after run x 7.56 μg/hr = 3.15 μg) = 247.462 μg PM2.5

So, the average air breathed over a week with one 40 and one 35-minute run would have: [256.893 + 247.462 + (181.11 x 5)] / 7 = 201.415 μg PM2.5 breathed per day.

201.415 ¡ 201.6 (the amount of PM2.5 breathed in during a week of living in an environment with average weekly PM2.5 20 μg/m3).

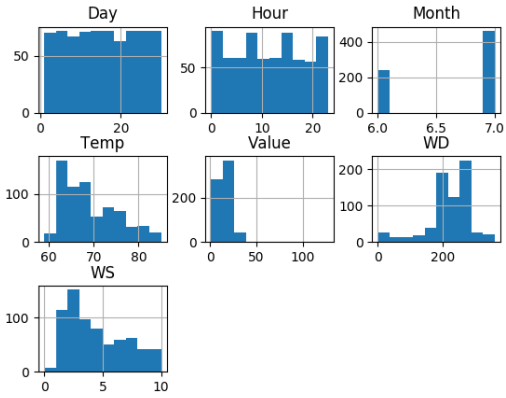
Thus, by running twice a week, for approximately 40 minutes at one time and 35 minutes the other, in an environment with constant PM2.5 concentration of 18 μg/m3, I would be hitting my goal of consuming as much PM2.5 as I would living in an environment with 20 μg/m3 PM2.5.

To know when I should go running, I would need to know whether the PM2.5 concentration during a given hour will be less than or equal to 18 μg/m3. The aim of my model is to predict this.

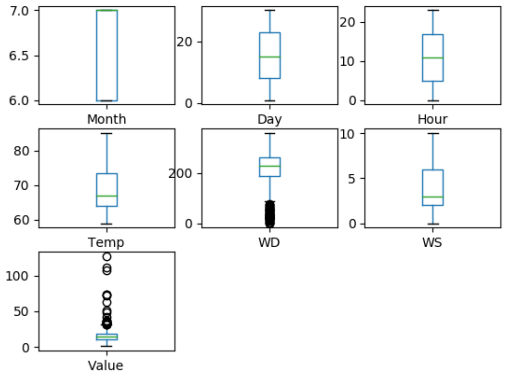
3 Data & Method

I modeled by experiment after Robert Ritz, who predicted PM2.5 pollution in Ulaanbaatar, Mongolia (4). Ritz identified a number of weather factors that affect PM2.5 pollution, including wind direction, temperature, and dew point. After extensive searching, I found appropriate Los Angeles data for temperature, wind direction, and wind speed, from South Coast AQMD (5). In addition to weather data, the date, month, and hour of day may be useful for making accurate predictions.

The data I gathered stretched from midnight on June 21, 2019 to 14:00 on July 20, 2019. (The limited data is due to technical issues I was having—see concluding remarks). Below is a summary of the different variables used to predict PM2.5, as well as their respective ranges:



(a) Bar graphs of feature ranges



(b) Box-and-whisker plot of feature ranges

Figure 1: Data on the ranges of the different features.

When these variables are placed in a scatterplot matrix, however, the correlation appears to be less obvious, since few variables form a straight line when mapped against the PM2.5 value during a given hour (labeled 'Value'):

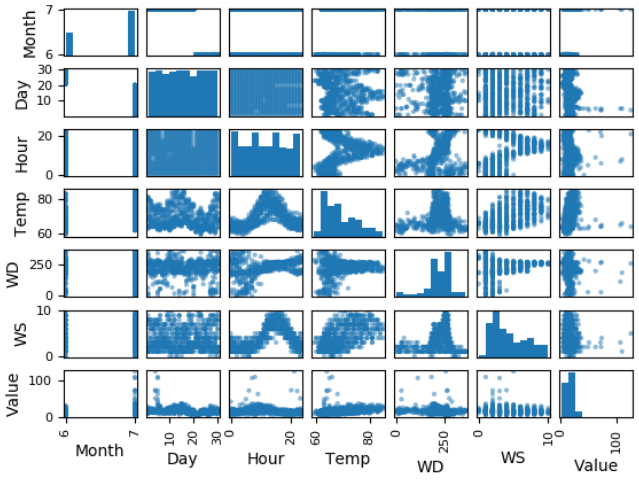


Figure 2: Correlation scatterplot of features.

I then split the data into a test and validation data set, where 80% of the total data belongs to the former and 20% to the latter. I used a 10-fold cross-validation to estimate accuracy, since using 10 for k is commonplace and often leads to accurate results.

I chose to test five different algorithms:

- Logistic Regression (LR)
- Linear Discriminant Analysis (LDA)
- K-Nearest Neighbors (KNN)
- Gaussian Naive Bayes (NB)
- Support Vector Machines (SVM)

The LR algorithm seemed to work the best for this problem, correctly predicting which hours the air would have an 18 µg/m3 PM2.5 or less concentration about 95% of the time. The results are displayed below:

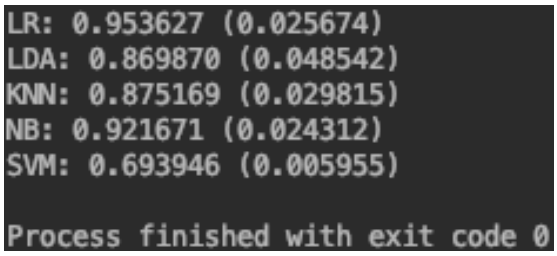


Figure 3: Algorithm results, where the decimal is the percentage of classifications predicted correctly.

4 Interpretation and Next Steps

There are a number of ways I plan to improve this model. First, the data I fed to the model only 704 instances of data. This is because I was getting strange errors that I could not resolve when I entered a number of instances greater than 704. It turned out this was a sloppy mistake in data entry on my part. Though the best-functioning algorithm still classified instances with up to 95% accuracy, performance could still be raised. The fact that not many instances or the optimal amount of features were used could be seen in a positive light, though, in that the model could more easily be used in other environments. It is easy to find data for a megacity such as Los Angeles, and people living in locations with fewer facilities/available data may still find that this model works to predict PM2.5 thresholds.

Furthermore, if there were more data instances implemented, a model to predict exact values of PM2.5 could be built. With this information, one could more accurately plan out their workout schedule. For example, you may know that for a given hour you want to run the concentration will be less than or equal to 18, but not know by how much. The air being 5 μg/m3 versus 17 μg/m3 could make a notable difference between how long would could run to meet their fitness and pollutant consumption goals.

The model to classify an hour as runnable or not runnable could also be improved by adding more features, most notably dew point, which Ritz found to be most correlated with PM2.5 concentration in his work in Ulaanbaatar. This is just a matter of doing more digging for data on a specific geographical location—in my case, Los Angeles.

It would also be useful to find a calculation to optimize how much one would have to exercise and when given one's personal fitness goals. As I mentioned, I typically go on 30-40 minute runs, and tried to find a goal (20 μg/m3) that would allow me to go on 2-3 of these runs per week in conditions that were slightly better than my goal (Los Angeles average of 18 μg/m3). In other words, I already had a good idea of what much of the information in the model would be. If, however, a person does not have a concrete number in mind for how much, or how often, they want to exercise in a week, it would be useful to have something that they could play with and generate different exercise scenarios for a week. A person who wants to run four times a week could, for instance, do four runs of 20 minutes, four runs of 40 minutes, at so and so times of day, and so on, which would make it easier for them to exercise in the first place.

Overall, this model is useful because it will help people exercise without fears consuming too many air pollutants. Though I picked a specific goal of having my yearly average air consumed be around 20 μg/m3 (largely dictated by the fact that I live in Los Angeles), this method could be used by anyone, anywhere, assuming there is appropriate weather data available for their location. One must simply pick a goal for how much pollutant they want to consume (ideally, at most, a yearly average of breathing air of PM2.5 concentration 12 μg/m3) and how much they want to exercise.

This software is available on my GitHub page.

References

[1] United States Environmental Protection Agency. NAAQS Table, 2012.

[2] Discovery Health. "How much oxygen does a person consume in a day?" `https://www.sharecare.com/health/air-quality/oxygen-person-consume-a-day`

[3] Inge Bos, Patrick De Boever, Luc Int Panis, Romain Meeusen. "Physical Activity, Air Pollution and the Brain". Sports Medicine, 14 August 2014.

[4] Robert Ritz. "Predicting Air Pollution in Ulaanbaatar, Mongolia". 2018.

[5] South Coast Air Quality Management District. Air Quality Historical Database. `https://xappprod.aqmd.gov/aqdetail/AirQuality/HistoricalData`