COMP 379 -- Machine Learning: Final Project

Prediction of Air Quality in a Particular City

David Saffo and Mary Makarious

1. Introduction

The identification and regulation of air pollutant levels has been a major focus of the American government, and is becoming a task for other governments worldwide. While it is understood that meteorological data like weather and humidity play a role in the concentration of pollutants found in the atmosphere, the relationship is poorly understood. In order to better understand these connections, advanced techniques have been brought into air quality research.

This project's goal was to apply machine learning techniques to take in various meteorological and air pollutant data to classify any given day into one of five categories: the levels of concern outlined by the by the U.S. Environmental Protection Agency (EPA). The level of concern is determined by the overall Air Quality Index (AQI). This project aims to better understand the relationship of a particular city's AQI on a specific day depending on weather and air pollutant data. A chart can be found on the EPA website--

Air Quality Index (AQI) Values	Levels of Health Concern	Colors
When the AQI is in this range:	***..air quality conditions are:***	***...as symbolized by this color:***
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

2. Dataset Description

To effectively identify and determine the key parameters affecting air quality it is critical to systematically collect data and characterize the air quality. This is in an effort to propose appropriate preventative strategies and policies to be put in place and to more accurately determine what is healthy versus unhealthy.

The data includes three parts: a development set, a training set, and a test data set. Each row in the data set represents the weather conditions, the concentrations of the pollutants in the air, and the overall health concern per day.

The total data set covers five cities: (1) Chicago, IL (2) Los Angeles, CA (3) Denver, CO (4) Miami, FL (5) and New York City, NY covering 365 days in the year 2015. Currently the focus is on U.S. cities.

The data comes from the Environmental Protection Agency and from Weather Underground's data. There are several stations in every city and the data is recorded and reported daily to be put into these two datasets..

A variety of meteorological parameters affect the air pollution level, and the following features were then added to our dataset:

Previous Day Overall AQI Value and Level of Health Concern:

The air quality of the current day is influenced by the condition of the previous extent to a big extent. If the air pollution was high the previous day, the pollutants may stay and affect the air quality of the following day.

Overall AQI Value and Level of Health Concern:

The air quality and level of concern associated of the given day is what we are trying to teach the machine to learn. The overall AQI value puts the day into one of five categories: (1) Good, (2) Moderate, (3) Unhealthy for Sensitive Groups, (4) Unhealthy, and (5) Hazardous.

CO, SO2, NO2, Ozone, PM10, and PM2.5 AQI Scores with the Main Pollutant:

Each day, the concentrations of these six pollutants change, affecting the overall air quality for the day. By looking at these values, it is simpler to determine which pollutant is overpowering the overall change.

Average Temperature:

Temperature would affect the air quality because of temperature inversion. The warm air above the cooler air acts like a lid. This prevents mixing vertically and results in trapping the cooler air at the surface. As pollutants are emitted from various vehicles, industries, and appliances, the inversion traps these pollutants nearer to the ground.

Average Dew Point and Average Sea Level Pressure:

The dew point is the atmospheric temperature which varies according to pressure and humidity. It is the temperature below which water droplets can begin to condense, allowing dew to form.

Average Humidity:

Average humidity is calculated by averaging the morning's relative humidity with the evening's relative humidity. Humidity could affect the diffusion of contaminants, because the atmosphere would be heavy with water vapor.

Average Wind Speed and Wind Direction:

Wind speed plays a critical role in diluting pollutants. Broadly speaking, strong winds disperse pollutants, whereas lighter winds results in pollutants building up in the vicinity.

Average Visibility:

When pollutants accumulate and concentrations rise, it is even more likely that visibility is reduced. The air appears to be hazy when pollutant levels rise.

Precipitation and Any Associated Events:

Areas of low pressure generally develop clouds and precipitation. Climate change can affect the intensity and frequency of precipitation.

3. Baseline Approach Description

As this is a classification problem, our baseline approach was to take the frequency of the most common class for each city. For example, if the city in question had 60% “good” days, predicting the class to be “good” every time would yield 60% accuracy. By computing this, we can measure if our model is better than simply guessing the most common class for every classification. The most common class was “moderate” with the exception of Miami whose most common class was “good”. The baseline accuracy for each city is as follows.

Most Common Class and Baseline Accuracy by City (Based on Overall AQI)

City	Most Common Class	Baseline Accuracy
Los Angeles	Moderate	56%
New York City	Moderate	52%
Chicago	Moderate	64%
Denver	Moderate	69%
Miami	Good	70%

4. Method Description

Using the scikit learn Python library, we implemented the following algorithms for classification and model selection; meta estimator (ExtraTreesClassifier), hold out (train_test_split), logistic regression (LogisticRegression), Naïve Bayes (GaussianNB), and nearest neighbors (KNeighborsClassifier). Based on knowledge of meteorological effects on pollution, we tentatively selected the following features to train; average temperature, dew point, humidity, wind speed, visibility, sea level pressure, and wind direction. To increase the accuracy of classification, we ran a meta estimator on the selected features for each city. The ExtraTreesClassifier class fits a series of randomized decision trees on random sets of the data to estimate the importance of each feature. Features that returned scores less than 0.1 were dropped from the feature list. Doing this removes

features that are not important for the city in question. Each city picked about the same features. After determining which features to remove for a specific city the data set was split 70/15/15 into a training, development, and testing set using the train_test_split class. Our main classification method was the nearest neighbors algorithm (kNN), however for the sake of evaluation we also implemented Naïve Bayes and multinomial logistic regression. The Naïve Bayes and logistic regression classifiers were left in scikit learn's default settings as this achieved the best and most consistent results. As for kNN, the k value changed drastically for each city. To find the best k value, kNN was run multiple times on the development set until the highest accuracy was found.

5. Evaluation

Due to the imbalance of labels in our data set, F1 score was used as our performance metric. F1 score is better suited for imbalanced data sets as it takes into account precision and recall instead of just the total accuracy of predictions. The results are as follows.

Total Accuracy of Predictions by City and Algorithm (F1 Score)

City	Logistic Regression	Naïve Bayes	Nearest Neighbors	Baseline
Los Angeles	69%	58%	72%	56%
New York	76%	72%	70%	52%
Chicago	69%	69%	69%	64%
Denver	65%	65%	69%	69%
Miami	76%	81%	78%	70%

All models (with the exception of Denver) performed better than the baseline predictions. In general, Naïve Bayes performed the worst in terms of score and consistency. Logistic regression and nearest neighbors performed similarly. However, logistic regression performed slightly worse averaging 71% across all cities while nearest neighbors averaged 71.6%. Overall, while Miami had the highest scores, it also had the lowest variance in AQI values. New York and Los Angeles had the highest variance in AQI values while Chicago and Denver both had moderate variance.

6. Discussion

After analyzing the results, a few notable observations were made in regards to the data and the models. For feature selection, the variables of visibility and average sea level pressure were often excluded. This came as a surprise as we anticipated visibility to be a good metric of the overall AQI value due to it being affected by smog levels, and that smog is a strong indicator for high pollutant levels. We also observed that Los Angeles was the only city where wind speed was not an important factor in determining the AQI value. We suspect that this is due to the variance in wind speed to be very low in the city. The classification for Miami ended up being binary as the overall AQI value placed all the days as either in the “good” or “moderate” level of health concern category. This resulted in Miami achieving the highest scores compared to the other cities. Denver was the only city where the models did not achieve a higher score than the baseline. We suspect that this has something to do with the cities altitude and unique geography. These factors were not well represented in our dataset and perhaps led to the poor performance of the models. Lastly, it is interesting to note that while kNN was the most consistent, depending on which city was being evaluated other algorithms sometimes presented better F1 scores.

7. Conclusion

The goal of this project was to predict the level of health concern classification of a given day from a particular city using meteorological and air pollutant data.
The dataset for this particular project is on the smaller end, and a way to improve the predictions would be to add more years for each city, keeping in mind air quality is a long-term problem that is affected by overall climate patterns. The dataset could also be improved by adding metrics for pollution for the specific day. For example, the amount of cars on the road or the total plant emissions for each day. Another thing to do, and what we were hoping to do, is add more locations, and expand globally. We set out to make a fairly accurate model using mostly meteorological data and we have achieved this goal. There is room for improvement and the next steps are clear. The EPA and the U.S. government’s combined efforts have led to a healthier atmosphere nationally. Models and datasets like ours help play a part in reducing the aqi and keeping the air we breath clean.

8. Appendix

Work was evenly and fairly distributed between the two members of the group.

9. References

[1] AQI Background Information: https://airnow.gov/index.cfm?action=aqibasics.aqi

[2] Weather History Archive: https://www.wunderground.com/history

[3] Daily Air Quality Data Archive:
https://www.epa.gov/outdoor-air-quality-data/download-daily-data

Background Information on

[4] Pollutants and AQI: https://airnow.gov/index.cfm?action=aqi_brochure.index

[5] Pressure and the Weather: http://www.weatherworksinc.com/high-low-pressure

[6] Humidity and the Weather: http://wonderopolis.org/wonder/what-is-humidity

[7] Visibility and the Weather: http://www.econet.org.uk/weather/visi.html