Predicting Air Quality
Using Machine Learning Techniques

By David S. and Mary M.

COMP 379 -- Final Project Presentation

Background Information

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.

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

The level of concern is determined by the overall Air Quality Index (AQI)

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).

Dataset Description

Discussing:

- Cities in the Dataset
- Variables in the Dataset

Cities (Currently) in the Database

Chicago, IL

Los Angeles, CA

Denver, CO

Miami, FL

New York City, NY

Currently, the focus is on cities in the U.S., since data was (relatively) more readily available, thanks to efforts by the Environmental Protection Agency (EPA)

The next city we'll be focusing on will be Beijing, China, because Beijing's air quality hits the "hazardous" classification, a classification not attained in any U.S. city.

Beijing, China

SOON!

Variables Included:

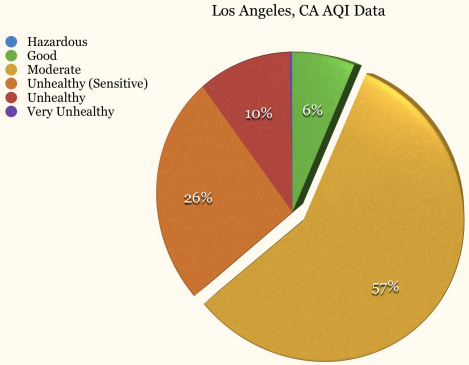
- ➔ Previous Day Overall AQI Value and Level of Health Concern
- ➔ Overall AQI Value and Level of Health Concern
- ➔ AQI Scores for CO, SO2, NO2, Ozone, PM10 and PM2.5 and the Main Pollutant
- ➔ Average Temperature
- ➔ Average Dew Point and Sea Level Pressure
- ➔ Average Humidity
- ➔ Average Wind Speed and Wind Direction
- ➔ Average Visibility
- ➔ Precipitation and Associated Events

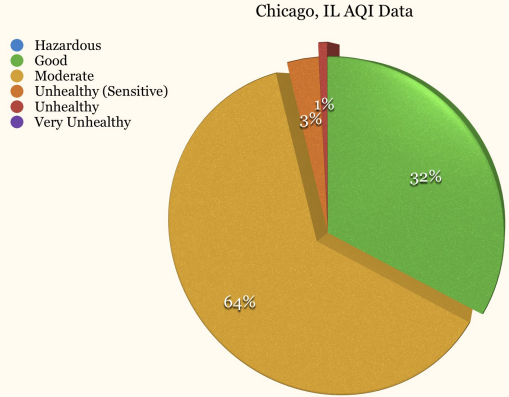
This dataset, while included all these variables, was not used in it's entirety.

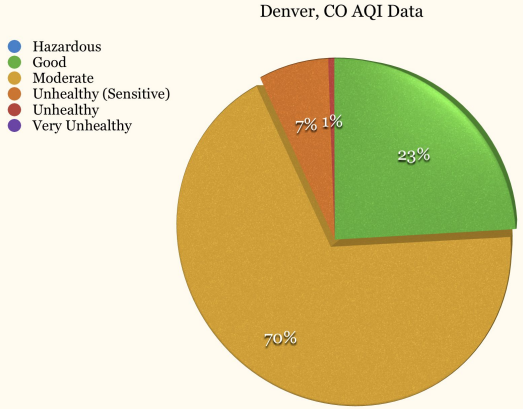
Depending on the city, certain variables made more of a difference than others.

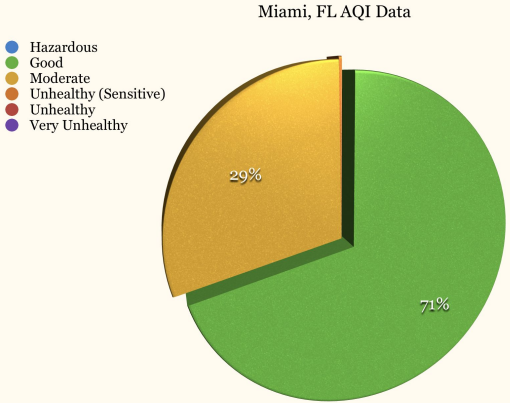
(Ex. Wind Speed in LA vs. Chicago)

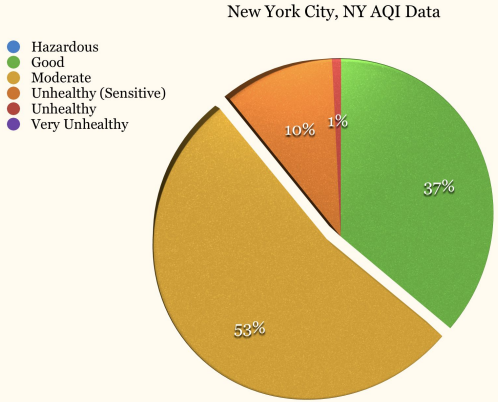
Baseline Approach











Baseline % for Each City

* Percentages are rounded // Baselines were computed to test if the machine was performing classification.

Method Description

Discussing:

- Feature Tuning
- Dataset Hold Out
- Model Fitting

Method

Feature Tuning:

- ➔ Avg Temp, Dew Point, Humidity, Wind Speed, Visibility, Sea Lvl Pressure, and Wind Direction
- ➔ Meta Estimator (Extra Trees Classifier)
  - Features were given scores between 0 and 1
  - Features with scores < 0.1 were removed for the given city

Hold Out:

- ➔ Data Split
  - 70% Training; 15% Developement; 15% Testing

Models:

- ➔ Logistic Regression
- ➔ Naïve Bayes
- ➔ Nearest Neighbors
  - ◆ Best k value was computed on the development set

Evaluation

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

City	Logistic Regression	Naïve Bayes	Nearest Neighbors	Baseline
Los Angeles	69% (+13%)	58% (+2%)	72% (+16%)	56%
New York	76% (+24%)	72% (+20%)	70% (+18%)	52%
Chicago	69% (+5%)	69% (+5%)	69% (+5%)	64%
Denver	65% (-4%)	65% (-4%)	69% (0%)	69%
Miami	76% (+6%)	81% (+11%)	78% (+8%)	70%

- ➔ 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%.

Discussion & Conclusion

Discussion

Notable Observations:

- ➔ For feature selection, the variables of visibility and average sea level pressure were often excluded.
- ➔ **Los Angeles** was the only city where wind speed was not an important factor in determining the AQI value.
  - ◆ Not all that windy in L.A.
- ➔ **Miami**'s classification was binary as the overall AQI value placed all the days as either in the "good" or "moderate" level of health concern category.
  - ◆ Resulted in higher accuracy and better F1 scores
- ➔ **Denver** was the only city where the models did not achieve a higher score than the baseline.
  - ◆ Something with the altitude/geography?
- ➔ While kNN was the most consistent, depending on which city was being evaluated other algorithms **sometimes presented better F1 scores.**

Conclusion

Did we Achieve Our Goal?

➔ Yes

What Can Improve?

➔ Add more data for previous years
➔ Add data for pollution contributors (industrial parameters like cars and factories)

What's Most Important?

➔ Quality and consistency of the data
➔ Precise selection of the important features

Any Questions?