Cind 820: Big Data Analytics Project Design

Adam Bruno
Student # 500904101
Dr. Tamer Abdou
September 22, 2025

Topic Introduction:

This project serves to explore the drastic change in atmospheric contamination, highlighting classifications for a deeper understanding of emerging ecological risks. Modern industrialization and technological advancements have greatly shifted air quality conditions establishing an opportunity to implement data classification techniques. Industrialized consumption of fuel and burning of components have introduced environments to hazardous practical matters. A scientific hypothesis tested on the industrialised city of Hamilton reported a 36% higher concentration of PM2.5 than other non industrialised cities(Yassine et.al 2024). These various chemicals when left within the atmosphere can have adverse effects on individuals over time. An article published by Dr Bert Brunekreef explicitly states that " Exposure to pollutants such as airborne participate matter and ozone has been associated with increases in mortality and hospital admissions due to respiratory and cardiovascular disease"(Brunekreef 2002). Thus, the application of machine learning techniques will allow for assessment of growing air pollutant trends and retrieval of risk evaluations to serve local inhabitants. Similarly, implementation of modern data analytic methodology could benefit researchers by reducing time spent manipulating raw sensor data, ultimately offering greater support of formal research. Modern reliance on industrialized machinery has created the opportunity to utilize machine learning techniques to effectively identify, assess, and predict rapidly changing environmental trends.

Limitations and Problems Statement:

The goal of this project is exploring ways to express air pollution quality degradation through the lens of data classification analytics. As such various challenges and limitations will be encountered. These may range from poor data quality for training sets as a result of sensor inconsistency, large class imbalances and heavily skewed model interpretations, and clear definition of boundary parameters to create accurate thresholds. Data set imbalances pose a critical risk to the model assessment as it may interfere with training and test set results. This is demonstrated in an academic journal, *Atmospheric Research,* stating "If the problem of imbalanced data is overlooked, the training data may fail to represent real-world prediction situations, causing substantial estimation bias" (Tang 2024). As such it is vital to the project's integrity to understand these limitations so as not to dampen any prevalent discoveries.

Research Questions:

1. How accurately can the application of machine learning modelling, on a given environmental and meteorological data set, predict and interpret insightful trends and geographical patterns?
    a. Understanding how accurately models can perform the task of quantifying air quality insights is the core of this project and as such helps validate the practical significance of classification in environmental and public health planning.
2. Are there models that work more efficiently with the specified dataset? What impact does data imbalance and missing or anomalous value limitation pose on the performance of the classification model?
    a. Air pollution data can often be incomplete or skewed toward incorrect classifications. As such this may result in performance degradation of the class modelling. Therefore, investigating these problems is vital in determining how reliable the model findings are.
3. To what extent do factors such as temporal and spatial elements affect the overall accuracy of classification models for air quality levels.
    a. Air pollution is highly dependent on various factors such as location, season, time of day, etc. As such it is important to test whether the model generalizes classification with regards to different locations or times. This can be beneficial in identifying any geographical or weather related biases ultimately improving model performance

Dataset Definition:

This project will utilize the dataset found on the government of canada national air pollution surveillance program. Relevant documents will be downloaded as csv files from the following websites: National Air Pollution Surveillance (NAPS) Program - ECCC Data Catalogue. This resource is a rich dataset that demonstrates highly maintained local data to bolster the relevancy and reliability of the project. Additionally, the resource is data rich featuring multiple pollutant types, ultimately improving training machine learning models. Support of daily time delineation as well as geographical locations and weather indications will strengthen insights. The dataset is suitable for the concept of classification and can provide vital insights into potential pollutant trends. As such the project will aim to leverage the rich nature of the dataset through various tools,techniques, and methodologies.

Techniques, Methodology, and Tools:

Various techniques and tools will be implemented to handle the project's methodology. The project will utilize python and its different libraries to establish and implement the machine learning model. Data preparation and exploratory data analysis will involve libraries such as panda, numpy, and matplotlib to establish a general framework for the dataset. This serves to create a simple understanding of missing values as well as any data imbalances. Similarly, skewed pollutant data may heavily dominate the dataset as such resampling measures may need to be utilized. This rudimentary step is crucial to the project's success as improper data cleaning may damage the predictive models findings. Additionally, to establish the predictive model the library, scikit-learn, will be used for supervised learning algorithms. An article by Pedro Domingos at the University of Washington states that "As a rule, it pays to try the simplest learners first … More sophisticated learners are seductive, but they are usually harder to use, because they have more knobs you need to turn to get good results, and because their internals are more opaque"(Domingos 2012). As such various models will be utilized to gauge overall performance. For example, potentially creating naive bayers models as a baseline for more advanced classifiers. Furthermore, framing the data into categorical indexes (low, moderate, high) as well as applying cross validation to create a more accurate picture. Evaluation metrics such as accuracy, precision, recall, and confusion metrics will be implemented to interpret and authenticate any findings. This project will utilize various techniques and tools to create a comprehensive and detailed report on the utilisation of machine learning on air quality data sets.

References

Brunekreef, B., & Holgate, stephen T. (n.d.). *Air Pollution and health - sciencedirect*. Air pollution and health . https://www.sciencedirect.com/science/article/abs/pii/S0140673602112748

Yassine, M. M., Dabek-Zlotorzynska, E., Celo, V., Sofowote, U. M., Mooibroek, D., & Hopke, P. K. (2024a). Effect of industrialization on the differences in sources and composition of ambient PM2.5 in two Southern Ontario locations. *Environmental Pollution*, *341*, 123007. https://doi.org/10.1016/j.envpol.2023.123007

Tang, D., Zhan, Y., & Yang, F. (2024). A review of machine learning for modeling air quality: Overlooked but important issues. *Atmospheric Research*, *300*, 107261. https://doi.org/10.1016/j.atmosres.2024.107261

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

National Air Pollution Surveillance (NAPS) program. Open Government Portal. (2016, September 24). https://open.canada.ca/data/en/dataset/1b36a356-defd-4813-acea-47bc3abd859b