Cind 820: Big Data Analytics Initial Results
Adam Bruno
Student # 500904101
Dr. Tamer Abdou
Dr.Ceni Babaoglu
December 1, 2025

Literature review:

To better understand supervised learning and classifier models various researched articles were reviewed. These articles specifically covered utilization of modern algorithms to extrapolate and classify various ecological contaminants. The articles highlight various approaches to machine learning models as well as data processing. Review of past examples will assist in bolstering the overall integrity of the project. Similarly, the project attempts to explore the various distinct models in hopes of narrowing down an approach that best represents the dataset. These articles stress the importance of air quality index constraints that will utilise pollutant concentration variables such as PM2.5 to establish primary pollutant boundaries. Additionally, interpretation of the inherent relation between meteorological features and their impact on pollutant readings. This report will highlight past research experiences and how they compare with findings conducted in this study, illustrating how machine learning is becoming more prominent in the field of ecological studies.

Similar case studies can assist in approaching the question of whether machine learning could produce meaningful insights from air quality statistics. Liu et all (2024) highlights a distinct approach to analysing pollutant levels in the article *Air Quality Class Prediction Using Machine Learning Methods Based on Monitoring Data and Secondary Modeling.* The article delves into air pollutant variables present within Jinan, China from July 23 2020 through to July 13 2021. The research paper breaks down the fundamental steps to creating an effective predictive model, comparing various models to seek further proficiency. Firstly relevant data such as meteorological data and pollutant concentration data were obtained from the china air quality online repository during the previously specified time frame. The article illustrates the importance of proper data cleaning cautioning data validity loss due to external factors such as server outages. The article states “When building a prediction model, low-quality data may affect the prediction results. Therefore, it is necessary to clean the original data by eliminating noise and improve the data quality to improve the prediction accuracy.”(Liu et all 2024) After cleaning the data four models are tested including weighted model, light gradient boosted model, logistic regression model, and random forest model. The article further states the basic principles of each model specifically highlighting the LightGBM advantages including its high robusticity, accuracy, expandability and simple operation. It reports that the LightGBM demonstrated an accuracy rating of 97.5% as well as f1 scores of 93.3%. Comparatively, our model demonstrated better results using random forest algorithms than gradient boosted models.This illustrates the need to find data that best adapts to the datasets structure. Similarly, data preprocessing techniques heavily impact performance metrics as the journal articulates. Proper data cleaning techniques are vital to creating training data that assists the model in interpolating data. Ultimately the article highlights the aptitude of utilising gradient boosted techniques in python to effectively forecast air quality.

Expanding on the topic of utilizing supervised machine learning models to extract evaluation metrics from rich variable dense sources, Liu et al (2015) exemplified similar research on the application of air quality forecasting within urban landscapes. The article, *Forecasting Urban Air Quality via a Back-Propagation Neural Network and a Selection Sample Rul*e, seeks to provide precise and reliable predictions to local authorities in Guangzhou. Similarly to the case study previously discussed, observational data was constructed within a dataset that included both meteorological parameters as well as pollutant levels. The article provides thorough insight into the use of neural networks to craft in depth training algorithms. Liu et al specifically states the approach's ability to model highly nonlinear functions and their ability to be trained for accurate generalizations. However it is also stated that three main factors can potentially limit a neural network training algorithm. These include network topology, learning algorithms, and learning samples. As such the paper aimed to develop a learning sample method that could effectively predict containment concentration levels based on a similarity principle of meteorology and pollutants. The paper combines the use of sensitivity experiments for parameter selection and Back Propagation neural network for dataset function computation. The model provided insightful results in forecasting performance of distinct pollutant elements. Evaluations within the slated May 2011 to April 2022 time frame demonstrated a 4% in MAPE values of the element PM10. This article ultimately shows the emergence of new forecasting algorithms to accurately depict pollutant data generalizations.

Research questions:

1. Air pollutant quality can be measured through a wide range of contaminants.Growing technological integration can provide enhanced insights into various environmental conditions. As such, will the implementation of supervised learning models produce insightful discoveries into patterns and trends within modern living conditions?

2. How relevant and impactful are the identified discoveries from the model analysis? Is the data model sufficiently accurate can the model be used consistently or further

scalable to address future issues? Which model has the highest potential to provide the most meaningful context without large computational necessities?

3. What distinctions can be made between the various models and approaches? Are there any overt advantages or limitations of using one specific model over another? How does the EDA or data cleaning affect the models ability to perform? How much performance loss can be expected as a result of improper EDA implementation?

Findings and Interpretation Section:

Results presented through the study offer meaningful insights into the potential depth of machine learning in environmental studies. Performance metrics that will be investigated further in the model evaluation section demonstrate machine learning models ability to effectively categorize air quality assessments standards. Both models perform reasonably well with regards to overall project scope and limitations. Effective model implementation was discovered to offer critical insight into patterns and trends within modern environmental conditions. This was exemplified through the random forest algorithm demonstrating improved capability to assess pollutant readings for specified seasons. Although small, these findings still remain significant as they ultimately reveal a path for further integration within the field. However, data accuracy still remains a potential question as models rely heavily on sensor readings which can often be unreliable if not properly installed and maintained. As such, using models that offer greater robustness and scalability can help mitigate potential outlier biases. This was shown through the random forest models superior ability to handle outliers through evaluation metrics. Ultimately, the research demonstrated a necessity to continue integrating machine learning processes within environmental studies as they potentially supply new discoveries into particle matter pollutants. Further understanding can serve to impact the public through greater education of how categorized air quality can establish societal standards. This is exemplified through creating a groundwork for understanding how at risk individuals are when partaking in activities outside in conditions identified to be hazardous by the model. Major health implications exist as individuals with increased cardiovascular risks can rely on the model readings to avoid exacerbating risks. Development of accessible machine learning models within the environmental air quality field can improve further understanding of growing trends as well as offer bolstered awareness.

Model Evaluation section:

The capstone project aims to utilize classification models to further understand pollutant levels through machine learning. As such, this report focuses on implementation of machine learning techniques to distinguish a model that can be utilized to determine a suitable environmental approach. Specifically, two distinct approaches were utilised to assess the dataset and discover the validity of the models. This is demonstrated through the application of Random Forest and Gradient Boosted classifiers on the target feature to train and test machine learning. This is done to prove the integration of machine learning can be beneficial to understanding changing air quality conditions. The capstone project specifically highlights the random forests model's ability to evaluate pollutant detection across various seasons. The features for the distinct random forest include various meteorological characteristics such as humidity, temperature, etc. Combined wind speed was converted into a numerical format from a categorical format for model training purposes. Training and test split ratios remained the same as illustrated within the initial results, 80/20. This further allowed the model to be trained on a significant portion of the data ultimately improving the models ability to handle unseen data as well as portray more accurate performance and evaluation metrics. Specifically, the trained model achieved an overall accuracy score of 86% on the testing set. Similarly, the model performed generally well with precision and f1 metrics throughout the four seasons. However, the model struggled to determine high recall scores.This is shown in their 0.79 and 0.78 score in seasons spring and fall comparatively boasting high recall scores in summer and winter at 0.93. Overall the model performed well and was able to indicate varying performance across different seasons

The second model utilised gradient boost models with cross validation to create a more robust reading on model metrics. Previous research highlighted the usage of using GBM techniques to train the dataset. Research results indicated potential aptitude for classifying datasets using light gradient boosted models. However, this report's metrics found that the random forest model outperformed the gradient boosted model. This is evident as the accuracy score for the gbm was reported as .69. This is further validated as using 5 fold stratified cross validation of the training set achieved a score of .6777 with .002. On the held-out test set, the final gradient boosting model demonstrated an accuracy of 0.6882, a weighted precision of 0.6971, a weighted recall of 0.6882, and a weighted F1-score of 0.6746. Similarly, the model training time took under a minute with 50 seconds.

Limitations and Ethical Considerations Section:

As a result of limitations encountered throughout the project entirety, there are possibilities to still improve overall project functionality. Additionally, the machine learning models performance could continue to be improved with greater tuning, resources, and data quality. The first limitation this capstone project encountered was critical data quality issues. Poor sensor maintenance ultimately resulted in large portions of missing data within the dataset. As such, data handling through dropping missing values heavily impacted the models ability to effectively learn on the distinct sets. Proper meteorological and pollutant emissions reading are required to create accurate projections and analysis. This is highlighted through various columns representing particle matter stations (PM_DEUN) not presenting data. As such, the majority of the study utilized solely one station to conduct modelling techniques. The second limitation encountered was overall project computational constraints. This project was conducted with minimal computational resources. As such the overall scope of the project was minimized to better align with available resource hardware. The project was conducted using a mid range laptop with a dated intel i5. Biases within the data model can alter particle matter readings leading to real world implications. Improper training set data splits can lead to invalid results. A study conducted entitled, *The effects of data quality on machine learning performance on tabular data,* highlights the importance of proper training data stating "... incomplete, erroneous, or inappropriate training data can lead to unreliable models that produce ultimately poor decisions"(Mohammed et al 2025). As such prioritizing models with greater robustness is advised. Similarly, greater emphasis is placed on establishing improved model tuning to handle broader data structures. Additionally, distinct ethical considerations offer further project constraints. Poorly constructed models as previously stated can produce faulty analytical insights, impacting real world decisions on the condition of air quality. Ultimately, this can lead to misinformed decisions on how to most effectively combat rapidly deteriorating air quality. These limitations and ethical problems were highlighted within this capstone project. In time, the project scope will seek to develop solutions that further mitigate limitations.

Methodology and data Preprocessing:

Data preparation techniques are vital to creating a clean dataset for machine learning models to interpret. Proper data cleaning and scaling practices can have a significant impact on supervised learning model results. The project examines a dataset with readings dating from January 1st 2010 to December 31st, 2015. Within this time frame missing values or faulty reading are properly identified. Absent sensor readings are present and denoted as NaN signifying data cleaning requirements. The dataset contains 52,584 entries as well as 18 columns. From these columns significant missing values are present within various stations potentially demonstrating poor sensor maintenance. This is evident within the project's exploratory data analysis which found that 60% of readings with Dongsihuan station contained missing PM values within the dataset. Initial data cleaning results may deem that these stations be cut from the analysis. As such to remove, the missing values filling techniques were implemented. Specifically, utilising forward fill removed most missing values from the dataset. However, 23 missing values still remained from the PM_US Post station readings indicating the necessity for additional backfilling technique utilization. Similarly, data imputation techniques were introduced for the second model, gradient boost, to offer improved data cleaning quality and perform comparative analysis. Missing values within rows were effectively dropped and imputed using the median. Additionally, outliers were identified to be present within numerous columns.Furthermore, missing values were handled within the precipitation column by replacing the reading with 0 ultimately ignoring the reading. This capstone project will mainly focus on utilization of random forest as well as gradient boosted models to train the dataset . Additionally air quality outliers may be potential readings of severe weather conditions which could heavily alter finding. These machine learning models methodology offer various scalability and robust proponents that bolster the overall project outlook.

Future Work and Recommendations Section:

Moving past this report, future works should aim to provide focus on the integration of machine learning techniques with global environmental issues. Continued expansion of implementation needs to occur as more emphasis needs to be on creating proper machine learning infrastructure to handle bigger dataset readings. As for next steps, combining all aforementioned key topics to create more accessible modeling information could help improve awareness for growing air quality contaminants. Additionally, improved tuning and optimization standards need to be present to ensure proper resource management is occurring. Creating more eco-friendly options for technology integration also needs to be studied. Elevated processing demands often result in more power draw potentially creating more environmental harm. Further prioritisation on creating more sophisticated pre-processing models is also a key necessity. As raw data volumes increase the need to establish a clean dataset for model training is vital. Further development within the field is required as environmental issues continue to rise. Machine modeling integration is a fundamental necessity to understanding current and future meteorological conditions and their impact on society.

Github Repository Link:

albruno-droid/CIND820: Data Analytics Air Quality Project

References

Liu, Q., Cui, B., & Liu, Z. (2024). Air Quality class prediction using machine learning methods based on monitoring data and secondary modeling. *Atmosphere*, *15*(5), 553. https://doi.org/10.3390/atmos15050553

Et. al., C. K. (2021). Prediction of climate change using SVM and naïve Bayes Machine Learning Algorithms. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(2). https://doi.org/10.17762/turcomat.v12i2.1856

Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2025). The effects of data quality on machine learning performance on Tabular Data. *Information Systems*, *132*, 102549. https://doi.org/10.1016/j.is.2025.102549