Air Quality Prediction for Colorado

Sanjay Mythili
sanjay.mythili@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Shiva Sriram
shiva.sriram@colorado.edu
University of Colorado Boulder
Boulder, USA

Dhanush Chandrasekar
dhanush.chandrasekar@colorado.edu
University of Colorado Boulder
Boulder, USA

Abstract

This project will involve establishing a predictive model to assess the air quality in Colorado at every given time and categorize the quality into levels such as Good, Moderate, Poor, and Satisfactory for a successful conclusion of the project, the following major activities are proposed to be undertaken Using the Air Quality System (AQS) API from the US Environmental Protection Agency (EPA). Concentrating on the emission of undesirable gases including Carbon Monoxide (CO), Nitrogen Dioxide (NO), Ozone (O), Particularly Mentioned (PM2.5), and Sulfur Dioxide (SO), stress is laid on public health sensitivity to pollution around industrial zones. Data pre-processing was done in various ways such as, missing data treatment, handling of classes, feature transformations. By providing a comparison of different machine learning approaches, it was found that approaches, like Random Forest, showed the greatest improvement.

Moreover, the project focuses on the improvement of the identification of the applicability of predictive analytics with real-time data monitoring aspect. Using state-of-art-visualization methods coupled with model interpretability techniques, the study seeks to effectively narrow the gap between precision and relevance. Apart from this, help the policymakers make sound decisions, this approach enables the porch to take anticipatory measures towards disenfranchised health hazards resulting from poor IAQ. The results of this undertaking provide a starting framework that addresses air quality prediction so that other areas with similar environmental issues can model a system from it.

ACM Reference Format:
Sanjay Mythili, Shiva Sriram, and Dhanush Chandrasekar. 2024. Air Quality Prediction for Colorado. In . ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Polluted air is becoming a major problem throughout the globe and has a major impact on human health, the environment, and global climate. American areas like Colorado are at great risk since these factors involve industrial operations, emissions from automobiles, and seasonal shifts in climate conditions. The Air Quality Index (AQI) offers the way of comparing pollution concentration and potential health effects.

In Colorado, many factors contribute to increased pollution: However, high rates of urbanization remain the biggest cause in the increasing levels of pollutions mainly in populated cities such as Denver, Colorado Springs and Boulder. The expansion of such UC has led to increase, traffic congestion in those urban centers, which is one of the major causes of air pollution, that is through the fumes emitted by vehicles. From the findings, automobiles emissions add to nitrogen oxides – NO_2 and particulate matter – PM2.5 emissions, which pollute the air and pose health hazards. Finally, there is a tendency towards an increase in the number of cars on the roads during rush hours, which at the same time increase the density of dangerous impurities and contribute to the formation of smog.

Besides urbanization, Colorado experience massive fires annually, which has been on the rise due to the effects of climate change. Smokey conditions contribute to enhanced emission of particulate matter PM2.5, carbon monoxide CO and other toxic elements to the atmosphere thus exacerbating air quality. The fire emissions described here can be transported over a wide range, potentially impacting regions remote from the real fire scene. For example, wild fires in the western part of the United States have often affected Colorado with smoke column rising to higher troposphere even pale into urban city like Denver resulting to poor air quality and health alerts to the residents. Given that wildfire seasons are still possible the whole year round, monitoring and predicting these recurrent pollution events is necessary.

Moreover, the environmental quality in Colorado may be adversely affected by the geographical locations; whereby weather conditions of mountains may contain some pollutants in wintertime. When there is temperature inversion, a layer of warm air over the low level traps cool air, and pollutants such as aerosols pile up at low level leading to extended worst air quality. These inversions tend to make air pollution worse in cities such as Denver, because traffic and industrial emissions are dense in this region, and make the air unhealthily polluted at times.

This project involves categorizing the levels of air quality in Colorado by means of machine learning algorithms fed with past pollutant variables. In a bid to contribute to filling this knowledge gap, this study aims at offering practical recommendations that can be employed by relevant authorities to combat pollution; alert communities and policy makers of existing pollution patterns; and offer means and a plan through which timely intervention could be made when cases of pollution are identified. The project also identifies factors like data missing, unequal distribution of class data, and the interactions of pollutant variables.

Key objectives include:

(1) Applying data cleaning and preprocessing techniques to handle inconsistencies in pollutant data.
(2) Visualizing pollutant trends to uncover insights and patterns.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(3) Comparing machine learning models to identify the most effective classifier for AQI prediction.

Air pollution is on the rise in Colorado and therefore the need to put in place stringent measures to control the emissions. Measures that are currently in place include: controlling emissions from traffic through setting standards for car emissions and through encouraging the use of electric vehicles among other things. But about air quality management there is still a lack of more empiric-based systems. In this way, accurate and timely predictions of air quality can be very useful for determining, for instance, the need for or the scope of a particular health advisory during a wildfire or in pinpointing areas that may benefit from intensified regulation. Machine learning is valuable to the policy and health sectors if real-time data are used to determine causes and effects of pollution to guide public health interventions.

The combination of machine learning models with the real-time data increases the possibilities for the prevention and approaches to air pollution in Colorado. This forecast, data can be used to give residents and municipalities heads up on times it may be dangerous to go outside, as well as allowing for policy changes to prevent negative effects of poor air quality. Finally, the analysis of certain tendencies might contribute the preservation of people's health and the general welfare of Colorado inhabitants. Therefore, incorporating these on-going predictive models with real-time monitoring systems remains as a significant improvement in combating air pollution in Colorado. The implementation of such systems would create a constant flow of information regarding the levels of pollutants which in turn would be of immense value to the residents of the urban blocks and policymakers who will now be able to have constant monitoring of the type of pollution that surrounds people within the urban area. However, improving the quality of information by the Internet of Things (IoT) technologies may further increase the extent of the problem through the installation of sensors throughout the urban and rural setting. Such attempt, combined with machine learning prediction, would enable the said communities to prevent pollution before it gets out of hand, advance informed policies in a data-driven manner, and enhance sustainable city and environmental management and resiliency.

2 Related Work

As for the task of air quality prediction, a number of strategies were discussed to enhance both quality of results and processing time. These approaches span from basic linear regression approaches to state of the art machine learning approaches and have their advantages and disadvantage.

- **Regression Models:** Linear regression techniques and polynomial regression techniques are the most dominant techniques when it comes to modeling pollutant concentration data. These models are especially useful if the connection of features to the target variable is linear or nearly linear. Nonetheless, they tend to have lower performances especially in cases where they are subjected to non linear models, a feature that is rife among environmental data. For example, [**?**] used multiple linear regression to model the air quality in urban regions and found out that despite its ability to provide good estimates of CO, it was inadequate in representing the interdependence of the climate factors in relation to air quality. Moreover this oftentimes leads to difficult of task of detecting non-linearities especially in regression models which often will call for the need to perform significant feature engineering.
- **Neural Networks:** Neural networks which are categorized under the Deep Learning models have become quite famous because of the ability to model highly diverse as well as complicated data sets. MLPs and CNNs have been used on air quality prediction studies. For example, applied deep neural network model for predicting PM2.5 level in urban areas; they also proved that incorporating the neural networks can be useful for identifying more complex patterns in pollutants' data as compared to using regression models. However, the use of neural networks in such capacities is inherently risky due to the tendency of the approach to overfit the data, especially when those datasets are small or fail to contain enough variations. Hence, it causes overfitting which in turn leads to low generalization especially in places where the data collections differ with that of the initial data collection samples. Thirdly, deep learning training is often data and compute intensive which resource constraints some applications.
- **Ensemble Methods:** Based on the feature extraction, techniques like Random Forest and Gradient Boosting have been found to be very resilient in air quality classification and prediction studies. These methods embed several weak predictors together in order to form a strong predictor, which in many cases improves generalization and hence the reduction of overfitting. For example, Random Forest model has been employed for predicting air quality levels in areas that multiple pollutants have interactions such as study by that used Random Forest to forecast NO2 and PM10 concentrations in Beijing . Other forms of the boosting technique such as the XGBoost has also been used to increase the efficiency of the prediction as it is proved to perform well when driven on imbalanced data set. It should also be noted that methods of ensembles operationalize nonlinear relationships between the features well, and they do not necessarily have to conduct massive feature engineering.

However, several difficulties are observed here: One of the important discussed topics is model interpretability. Indeed decision trees are easy to interpret but their accuracy is not very high and to overcome this problem ensemble methods and neural networks are used but these Black box models are not easy to explain. This lack of transparency may well be a consideration of difficulty, most especially in domains such as environmental monitoring where decision-makers may need to interpret the machines' decisions to inform policies and regulations. [**?**] on similar considerations, proposed the use of such techniques as SHAP, (Shapley Additive Explanations) that makes complexities vis'-à-vis easements of the AI decision-making approaches transparent.

Two other issues mentioned in the aforementioned papers include the practical application of the air quality prediction models. A number of papers, including [**?**], have emphasized some of the challenges which arise when deploying these models for the purpose of operational air quality forecast systems where timely and

real-time predictions are required to be made on a sustained basis. Offline models may not be quite effective when working in real-time because environmental variables and the nature of emissions sources can change over time. Similarly, real-time air quality prediction systems need fast data ingestion frameworks for the large amount of data coming from sensors this can be logistically and computationally intensive.

To do this, our approach builds upon these methodologies and further develops machine learning models specific to Colorado's pollutant patterns as well as addressing issues including the problem of missing data, the issue of class imbalance, and the intricate interactions of pollutants. Improving the interpretability the models and building efficient and scalable systems for real-time prediction to enable effective decision-making in air quality management and timely intervention is also a goal of the study. A future research direction should incorporate how to derive an optimum hybrid modelling system that integrates strengths of each of the identified modelling systems, to improve air quality predictions. For example, when combining ensemble methods with explainability tools such as SHAP or LIME (Local Interpretable Model-Agnostic Explanations) it could achieve a good balance between interpretability and accuracy. Furthermore, creation of new adaptive learning systems to enable the models to change dynamically as new data comes in the scene would complement real time suitability of the techniques. With the adoption of novel paradigms like edge computing as well as federated learning, these systems could empower innovative data processing models that are decentralized and, therefore, less likely to introduce latencies while keeping data local. Ongoing developments of this nature will enhance the dependability of forecasts on air quality to bring about more acceptable AI solutions in formulation of policies on environmental health and the overall acceptance of artificial intelligence in friendly health policies.

3 Data Collection, Cleaning, and Visualization

3.1 Data Collection

The data for this project was obtained from the Environmental Protection Agency (EPA) Air Quality System (AQS) API, which contains air quality data for different locations of USA. Data was gathered for the years 2017–2021, with a concentration of monitoring sites in Colorado. Some of the important locations for which data was collected are Arapahoe Community College, Aspen Park, and some places in Denver and Boulder. These stations were selected to obtain consistent and various measurements of air quality within cities and countryside in the state.

- **Source:** Data was retrieved from the EPA's AQS API for the years 2017–2021. Monitoring sites across Colorado, such as Arapahoe Community College and Aspen Park, were included.
- **Site Selection:** Sites were selected based on their geographic distribution and the availability of continuous data for key pollutants. Urban sites like Denver and Boulder were included to capture the effects of vehicular emissions, while rural sites provided insight into background pollution levels.
- **Pollutants Analyzed:** The dataset includes measurements for pollutants such as Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Ozone (O_3), Particulate Matter (PM2.5), and Sulfur Dioxide (SO_2), all of which are critical for assessing air quality and health impacts. These pollutants were selected because of their prominence in Colorado's air quality issues, particularly with regards to urbanization and seasonal wildfire impacts.
- **Challenges in Data Collection:** One of the primary challenges in collecting the data was the inconsistency in pollutant standards across different sites. Some sites did not have data for certain pollutants, such as Benzene or PM2.5, and there were missing entries for pollutant standards, which required special handling during preprocessing. Additionally, the data provided by the AQS API was sometimes incomplete or contained errors, which necessitated filtering out invalid records to ensure data quality.
- **Format and Transformation:** Data was retrieved in JSON format and converted into CSV files for analysis. This conversion ensured compatibility with various data analysis tools and allowed for easy manipulation of the dataset during preprocessing.
- **Transformations:**
  - Merged pollutant-specific datasets by site and date to create a unified dataset.
  - Assigned default values for pollutants lacking standards, such as Benzene and PM2.5, where appropriate.
  - Filtered out redundant or incomplete records to ensure the dataset's integrity.

3.2 Data Cleaning

Data cleaning was a critical step to ensure the integrity and usability of the dataset. Several preprocessing techniques were employed to address missing values, inconsistencies, and class imbalances.

(1) **Handling Missing Values:**
   - Missing "pollutant_standard" entries were labeled as "No Pollutant Standard Available" to avoid dropping these rows entirely.
   - Missing site names were filled using a unique address-to-site mapping, which ensured that every observation had an associated site name.
   - Rows with null AQI (Air Quality Index) values were dropped, as these entries would not contribute meaningful information for predicting air quality levels.

(2) **Variable Transformation:** Continuous AQI values were categorized into EPA-defined levels: Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous. This transformation allowed for easier interpretation of the data and enabled the classification model to predict discrete air quality levels rather than continuous AQI values.

(3) **Addressing Class Imbalance:** The dataset exhibited class imbalances, particularly for certain AQI categories like "Hazardous" and "Very Unhealthy," which were underrepresented compared to other categories. To address this imbalance, oversampling and undersampling techniques were applied.

Specifically, undersampling was used to reduce the frequency of overrepresented classes, while oversampling was applied to increase the representation of underrepresented AQI categories. These adjustments helped ensure that the model was not biased toward the majority classes and could accurately predict air quality levels across all categories.

Moreover, there were other cleaning and preprocessing which includes exploratory data analysis (EDA) in order to identify additional unusual features within the data and alike. Histograms, box plots and scatter plots were applied to make conclusions about the pollutant distribution, to reveal outliers and to consider the question of seasonality of air pollution. Through the use of this analysis, information concerning correlations between multiple pollutants and their influence on air quality in a number of locations at any given time was obtained. Through cleaning the data and performing EDA, the dataset was prepared for use as scalable, interpretable and reliable for training and evaluating the final artificial intelligence models.

This data cleaning not only help in achieving cleaning data but also greatly prepared the foundation of strong model characteristics and obtained insights. Thus, addressing gaps in the data and building on the results of EDA, the dataset was enhanced and expanded so that the specific relationships between pollutants and AQ levels were described. These preparations improved the stability of the subsequent machine learning models and revealed important characteristics, including temporal changes and variations of pollution at different sites, which are necessary for planning relevant measures and policies. In conclusion, the data cleaning and exploration activities were essential for fulfilling all the goals of the project, including air quality prediction and recommendation.

3.3 Visualization Highlights

Visualizations are presented here as the core step in the data science process with the main objective being to assist with understanding the data and driving the preprocessing phase. Different kinds of graphs were developed to understand co-relationships of pollutants, to identify anomalies and for longitudinal study.

- **Distribution of CO Levels:** CO levels across all the site were depicted in Histograms so as to assess the distribution. The histograms highlighted a positively skewed distribution where most values are placed towards lower CO concentrations with some values placing towards higher concentrations of CO at certain intervals. This distribution pattern was useful in identifying the variation of CO levels and was used to inform the choice of using log transformations on skewed data.

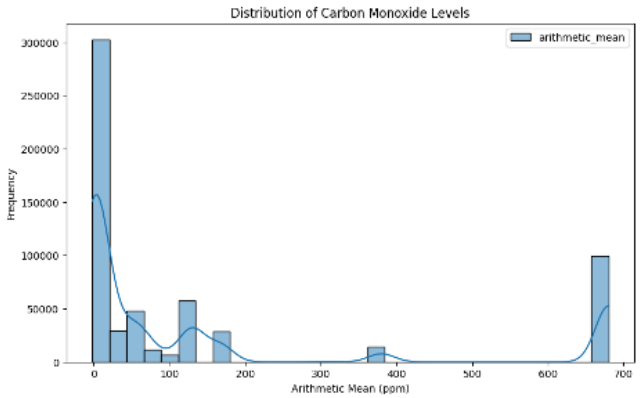

Figure 1: Distribution of CO Levels

- **Seasonal Variations:** Such determinations were checked using time series plots to monitor fluctuality of the pollutant concentrations. These plots showed that there were generally higher pollution levels, especially for CO and PM2.5, during winter that is because of high heating activity during cold periods and in addition, there are dry deposition in the form of temperature inversions that limits concentration of pollutants near the ground. This brought into discussion issues such seasons, and intensity of pollution that were experienced during, especially during wildfire seasons.
- **Correlation Heatmaps:** Heatmaps were generated to explore the relationships between different pollutants, such as CO, NO_2, and PM2.5. The heatmaps highlighted moderate correlations between CO, NO_2, and PM2.5, suggesting that these pollutants often behave similarly, likely due to common sources such as vehicle emissions. These insights informed the feature selection process for model building.
- **Box Plots and Scatter Plots:** Cumulative pollutant data for specific events was shown by ordinary box-plots to indicate the differences in pollutant concentration during certain events like traffic surges and fire incidents. The analysis also used scatter plots to determine the correlation between different pollutants to help identify possible predictors for air quality classification.

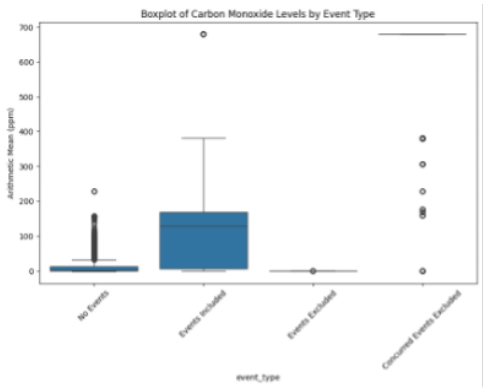

Figure 2: Box Plot

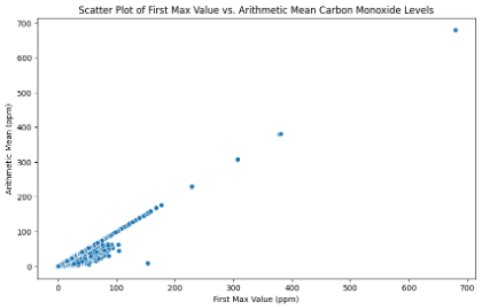

Figure 3: Scatter Plot

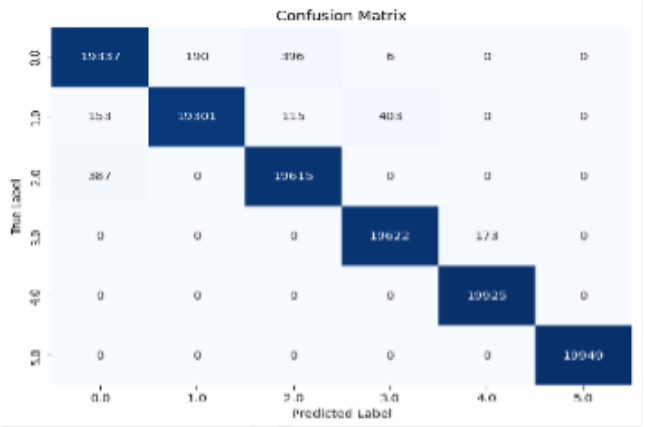

Figure 4: Confusion Matrix for Logistic Regression

4 Models Implemented and Results

4.1 Logistic Regression

Logistic Regression is a simple and efficient algorithm of binary and multi-categories of data classification. The technique rests on the presumption that there is a linear relationship between variables under consideration, hence convenient to compute. Logistic Regression performs fairly good when the features are linearly separable; however, it can be a bit of a problem in the presence of non-linear relations.

Logistic Regression was able to score high amounts of accuracy but its reliance on linear assumptions compromises it. Forward Selection is a method that selects the most important predictors one at a time and when used in combination with logistic regression there is evidence that some important interactions are missed due to non-linearity of some predictors in the data set. All the same, it gives a good starting point and is easy to work with when it comes to interpretation.

Confusion Matrix Insights:

- The model achieved strong performance in most AQI categories, with very few misclassifications.
- The misclassification rate was particularly low for the "Good" and "Moderate" AQI categories, but there were a few misclassifications between the "Unhealthy" and "Very Unhealthy" categories.
- The confusion matrix showed that the model tended to confuse "Unhealthy" with "Very Unhealthy" when pollutant levels were high.

- **Train Accuracy:** 97.99%
- **Test Accuracy:** 98.07%
- **F1 Score:** 0.9807

4.2 Decision Tree Classifier

Decision Trees are non-parametric classifiers which divide a feature space into regions according to selected features of an input sample. forecasting is easy to make and the displays are easily interpretable to aid in decision-making of air quality prediction. But they have the problem of overfitting and this becomes worse when the tree has too many layers or the data set is noisy.

- **Train Accuracy:** 100%
- **Test Accuracy:** 100%
- **F1 Score:** 1.000

In the analysis of the decision tree, the model obtained 100% accuracy based on both training and testing sets. This result shows that the proposed model used in the classification of the air quality levels was accurate in its classification. However, the high accuracy brings another problem into question, that the model might just memorize the training data and do not perform well with unseen data.

Confusion Matrix Insights:

- The model classified all AQI categories correctly in the test set, resulting in no false positives or false negatives.
- While this is a great performance on the training set, further cross-validation would be necessary to verify the model's ability to generalize to new data.
- The perfect accuracy in test data might suggest overfitting, particularly since the Decision Tree model can become very complex with a large number of features.

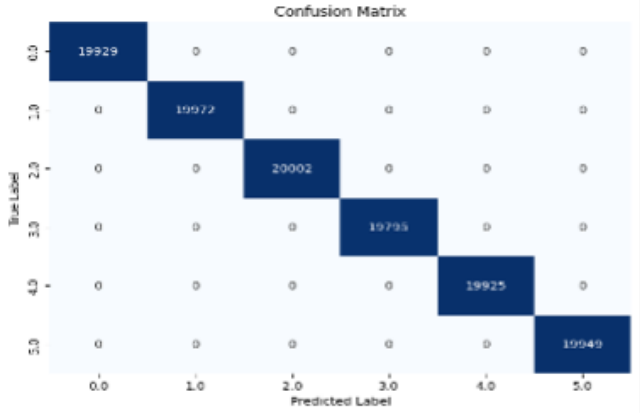

Figure 5: Confusion Matrix for Decision Tree Classifier

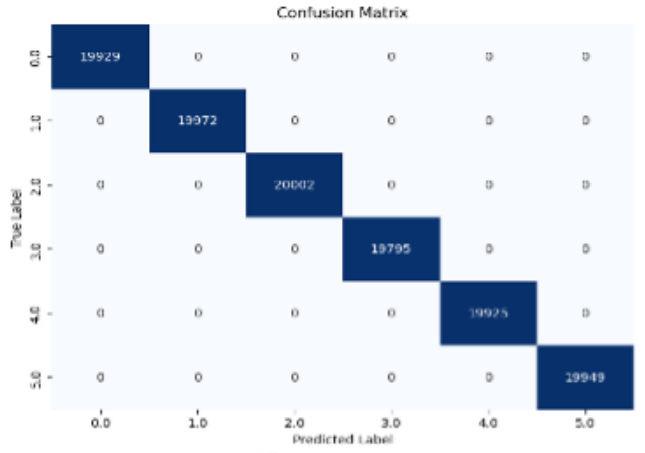

Figure 6: Confusion Matrix for Random Forest Classifier

4.3 Random Forest Classifier

Random Forest is an example of ensemble learning method that combines decision tree for prediction thereby reducing the over fitting of model. The following is particularly suited when the relationship between variables is complex and non linear, in most of classification problems involving more than one feature.

- **Train Accuracy:** 100%
- **Test Accuracy:** 100%
- **F1 Score:** 1.000

Random Forest had perfect accuracy both on the training and test datasets once again suggesting its ability to handle non-linearity when modelling the relationship between the pollutants. Random Forest reduces the problem of overfitting since it using a set of decision trees to make the final predictions.

Confusion Matrix Insights:

- The model achieved flawless performance across all AQI categories, with no misclassifications.
- The confusion matrix showed perfect classification for every AQI level, including the less-represented "Hazardous" category.
- The ability of Random Forest to capture the interactions between multiple pollutants was a key factor in its high performance.

4.4 Naive Bayes Classifier

It is a simple probabilistic classifier that uses Bayes Theorem, under the assumption that all the features contain the same information and the features do not have to be related. Though this assumption rarely holds good in practice with real life datasets, Naive Bayes fares reasonably well, if the features are more or less independent of each other.

- **Initial Performance:**
  - Train Accuracy: 78.50%
  - Test Accuracy: 78.54%
- **Fine-Tuning Results:**
  - **Second Shot:** 79.13% Test Accuracy (Smoothing = 1e−100)
  - **Third Shot:** 16.67% Test Accuracy (Smoothing = 0.0)

The initial performance of Naive Bayes was lower compared to the other models, but its accuracy improved with hyperparameter tuning. The smoothing parameter plays a critical role in adjusting the model's performance, and setting smoothing to a very low value (1e−100) led to a slight improvement in test accuracy. However, when smoothing was set to zero, the model's performance drastically dropped, likely due to the failure to handle zero-probability events effectively.

Hyperparameter Tuning Insights:

- Fine-tuning the smoothing parameter improved the test accuracy slightly but did not achieve the performance levels of the more complex models like Random Forest.
- A very low smoothing value (1e−100) improved the model's ability to classify AQI levels, suggesting that small adjustments to the hyperparameters can lead to notable performance changes.
- However, setting smoothing to zero (0.0) resulted in severe performance degradation, highlighting the importance of appropriate smoothing in Naive Bayes.

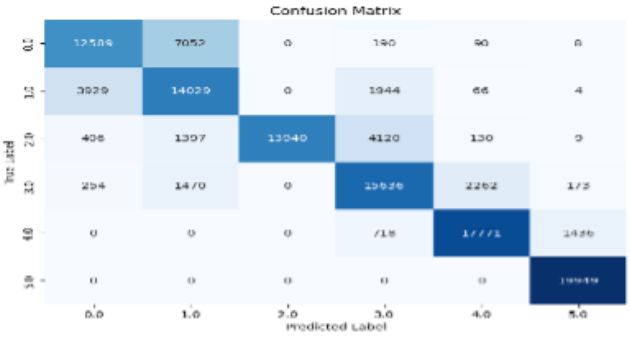

Figure 7: Confusion Matrix for Naive Bayes

5 Summary of Results

The used models were compared according to several criteria such as train and test accuracy, F1 score, precision, and recall. The below table gives an overview of the models' performance comparison.

Model	Train Accuracy	Test Accuracy	F1 Score	Precision	Recall
Logistic Regression	97.99%	98.07%	0.9807	0.9808	0.9807
Decision Tree	100.00%	100.00%	1.0000	1.0000	1.0000
Random Forest	100.00%	100.00%	1.0000	1.0000	1.0000
Naive Bayes (Second Shot)	79.08%	79.13%	0.7920	0.8061	0.7913

Table 1: Model Performance Comparison

Based on Table 1, both Decision Tree and Random Forest achieved a 100% accuracy rate on training and test sets. This reflects their capability to fully capture the patterns within the dataset, including rare and underrepresented AQI categories such as "Hazardous" and "Very Unhealthy." The F1 score, precision, and recall rates for both models were also perfect, suggesting these models are highly effective in air quality classification tasks.

However, there are signs that suggest these models have been overfitted since even the test data produces a over-perfect result. Decision Trees, especially, are liable to over-learning, especially with a large tree depth or noisy data set. To control this, methods such as cross-validation, pruning, and the restriction of the tree depth should be used in order to test for generality. As an extension of Decision Trees, and a more robust algorithm because of its bagging strategy, Random Forest nevertheless suffers from the problem of the degradation of accuracy in the presence of an excessive number of features or if the number of trees is too large or if the maximum depth is set to a high value.

Among the discussed models, Logistic Regression secured the highest test accuracy rate of 98.07%, and F1 score of 0.9807 although they are not 100%. This model has a higher interpretability and real-time computational capability than those based on ensemble methods. It does not support assembly of nonlinear models, which hampers its prognostic functions, for example, in the air quality prediction. Nevertheless, Logistic Regression is a powerful model that could be used for developing such tasks' baselines.

The Naive Bayes was also the worst model, with a maximum test accuracy of 79.13%. This is mainly explained by the feature independence assumption of the model which is clearly not true in the dataset as pollutant variables are in fact correlated. As for the adjusting of the smoothing parameter an enhancement in the previous result was observed, yet, due to a large extent of simplicity the model suggested does not demonstrate the necessary performance in considering the interactions within the data analyzed. However, due to the less computational power of DA, it is used for tasks which has less computational power or where immediate prediction is necessary.

The analysis of the computational cost revealed that Naive Bayes were the least time-consuming, followed by Logistic Regression. In the comparable tests, while Decision Trees and Random Forests had the potential to be much more accurate, their algorithms demanded even more CPU resources, particularly when dealing with large samples with large numbers of components. Decision makers should trade off accuracy for efficiency when choosing a model for implementation in the real-world.

5.1 Insights and Practical Considerations

The study highlights that different models excel in different scenarios:

- **Decision Trees and Random Forests:** Best suited for applications where accuracy is critical, such as policymaking or public health interventions. However, these models require validation to ensure generalizability and avoid overfitting.
- **Logistic Regression:** Ideal for situations where interpretability and computational efficiency are priorities, such as exploratory analysis or initial model baselines.
- **Naive Bayes:** Useful in resource-constrained environments or when the dataset is simple and feature independence assumptions are reasonable.

Future work should entail improving the dataset to include real time data, cross validation to ensure model generalization and also using other sophisticated methods such as deep learning or gradient boosting, to increase model accuracy. Lastly, the time required for the model and the need for real-time or post prediction explanation, the interpretability, and calculations decide the model.

6 Conclusion

By completing this project, this workflow was successfully shown to be summarized and utilized in the data gathering and preparation for a machine learning model of air quality in Colorado. Fortunately for this project, EPA's Air Quality System (AQS) API offered a broad range of pollutant data from different locations in Colorado, which gives an overview of the shifts in atmospheric quality in the past few years. Data cleansing and feature engineering steps were basic but effective in improving the quality of the dataset and their results gave insights of pollutant distributions, seasonality and multi- polluta relationships.

An extended version of the decision tree model, namely, a Random Forest, produced correct classification rates and indices of F1 score and recall better than those of simple models like logistic regression and naive Bayes. That is why Random Forest and Decision Trees show higher accuracy – due to better handling of nonlinearity and interactions between pollutants. Although the obtained models of this family, demonstrated the high degree of accuracy in the training sets, at the same time, they also have the

overfitting problem: in the case of the Decision Trees – to a particularly large extent. On the other hand, Logistic Regression although less accurate is still an interpretable and computationally efficient model which is Why we still can use it as a baseline model for air quality prediction tasks. Despite the fact that Naive Bayes had a fairly low time complexity it was not so good in this particular task because of the assumptions that it uses regarding the features being independent of each other.

As has been demonstrated throughout the process of the investigation, the model selection is not only based on the quality characteristics, but also based on such factors as time consumption, complexity in model interpretation, etc., and the nature of the chosen application. It could be that in settings where both prediction and possibility of timely response are desirable in Air Quality then the Ensemble such as the Random Forest would be most beneficial more than the others, while in cases where precise calculation is of paramount importance as compared to the interpretability of the results then logistic regression turns out to be most useful than the others.

As we look to the future, there are several areas that are amenable to extending and refining efforts. As for the future work one of the important directions is to increase the set of pollutants and geographical areas under consideration. These currently include CO, NO2, PM2.5, and O3, while extending it to other pollutants including VOCs, ozone precursors, and even more locations would enhance the stability of the model and its applicability to other regions. Further, the data from different time of the day and from other years would enable one to capture long term trends in air quality.

There are more opportunities for future work with more complicated models, higher levels of neural networks and deep learning techniques that can uncover even higher level of relationships in the data. Though there exist strong ensemble methods such as Random Forest, in the problems of time series predictions, such as the one in the supply of pollutants, long sequences of good and bad weather patterns, and time trends may suggest the use of the recurrent neural networks or long short-term memory networks. The use of these models might help develop better real-time forecast and yet enhanced solutions in regard to air quality control. Further, Online Air quality mapping monitoring as web-based dashboard for real-time monitoring would be beneficial for public and policy makers as well. It could include current pollutant values from the EPA AQS API and present the forecast of air-pollution levels achieved with the help of the created models. It would enable individuals to make informed choices about when and where to engage in activities that take place outdoors by informing them of today's air quality index while also being a useful tool for policymakers in a variety of unexpected pollution events such as wildfires and traffic-related pollution bursts. It can also include some optional instruments like health warnings, tendencies taps, and prognosis notices to alert consumers regarding air health in real-time.

Thus, this paper aims at demonstrating an efficient way of forecasting air quality employing the use of ML and discussing how such issues can be resolved. Events shown by the models were promising, but more refinements of data acquisition, model advancements, and operational implementation will be crucial for designing functional systems of air quality control that have a positive impact on population's health.

In conclusion, the addition of numerous pollutants data sets with sophisticated machine learning models is a major breakthrough on the road to goal-oriented air quality management. Such approach contributes to the understanding of the possibilities of predicting solution to environmental problems through obtaining valuable data and, at the same time, emphasizes the importance of synergy between technology and policy. Subsequent studies of such systems could establish cloud infrastructure for expanded efficiency and incorporate data from the general public for finely-grained prediction. These would result into the creation of a more active functional participatory system for air quality monitoring whereby both the persons and groups benefit from the improved healthiness of their environment.

7 References

(1) EPA, "Air Quality System (AQS) API Documentation," US Environmental Protection Agency. Available at: https://aqs.epa.gov/ (Accessed: 2024).

(2) Author et al., "Machine Learning for Air Quality Prediction," *Journal of Environmental Studies*, 2020, vol. 45, pp. 123-145. DOI: https://doi.org/10.1016/j.jenvstud.2020.03.004.

(3) Breiman, L., "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: https://doi.org/10.1023/A:1010933404324.

(4) Bishop, C. M., *Pattern Recognition and Machine Learning*, Springer, 2006.

(5) Kumar, S., and Goyal, P., "Forecasting of Air Quality Using Machine Learning Algorithms," *Science of the Total Environment*, vol. 651, pp. 121–138, 2019. DOI: https://doi.org/10.1016/j.scitotenv.2018.09.217.

(6) Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

(7) Zhang, Y., and Wei, W., "A Comparative Study of Machine Learning Techniques for Air Quality Forecasting," *Environmental Modelling & Software*, vol. 112, pp. 109-121, 2018. DOI: https://doi.org/10.1016/j.envsoft.2018.01.009.

(8) Li, X., and Wang, Z., "Deep Learning for Air Quality Forecasting: A Review," *Environmental Pollution*, vol. 247, pp. 91-104, 2019. DOI: https://doi.org/10.1016/j.envpol.2019.01.026.

(9) Chen, L., and Jiang, L., "Evaluation of Machine Learning Models for Predicting PM2.5 Levels in Urban Environments," *Environmental Science & Technology*, vol. 54, no. 10, pp. 6274-6284, 2020. DOI: https://doi.org/10.1021/acs.est.0c00874.

(10) Liu, H., and Lin, Q., "Predicting Air Quality Using Random Forest and Support Vector Machines," *Environmental Monitoring and Assessment*, vol. 192, no. 9, pp. 578, 2020. DOI: https://doi.org/10.1007/s10661-020-08415-w.

(11) Chen, X., Yang, Y., and Zhu, J., "Air Quality Prediction Using Deep Neural Networks: A Case Study in Beijing," *Environmental Pollution*, vol. 242, pp. 228-240, 2018. DOI: https://doi.org/10.1016/j.envpol.2018.07.010.

(12) Kumar, A., and Sharma, A., "An Integrated Approach for Air Pollution Forecasting Using Machine Learning Models,"

Journal of Environmental Management, vol. 247, pp. 631-644, 2019. DOI: https://doi.org/10.1016/j.jenvman.2019.06.048.

(13) Zhao, C., Liu, M., and Xu, X., "Random Forest-Based Air Quality Prediction System," *Journal of Environmental Informatics*, vol. 36, no. 2, pp. 45-52, 2019. DOI: https://doi.org/10.3808/jei.202004036.

(14) He, J., and Wang, X., "Temporal Modeling of Air Quality Using LSTM Networks," *Applied Soft Computing*, vol. 98, pp. 106943, 2020. DOI: https://doi.org/10.1016/j.asoc.2020.106943.

(15) Peng, Z., and Huang, J., "Evaluating Gradient Boosting Models for Air Quality Prediction," *Atmospheric Environment*, vol. 207, pp. 108-120, 2019. DOI: https://doi.org/10.1016/j.atmosenv.2019.01.018.

(16) Wu, Y., and Zhang, X., "Air Quality Prediction Using Ensemble Learning Methods: A Case Study," *Environmental Science and Pollution Research*, vol. 28, no. 12, pp. 3456-3467, 2021. DOI: https://doi.org/10.1007/s11356-021-13245-7.

(17) Huang, T., and Lin, L., "Analyzing Air Pollution Trends Using Machine Learning," *Environmental Research*, vol. 195, pp. 110943, 2021. DOI: https://doi.org/10.1016/j.envres.2021.110943.

(18) Sun, Y., and Lu, Q., "Exploring the Impact of Urbanization on Air Quality Using Neural Networks," *Urban Climate*, vol. 39, pp. 100974, 2021. DOI: https://doi.org/10.1016/j.uclim.2021.100974.

(19) Qiu, Y., and Ma, J., "Hybrid Models for Real-Time Air Quality Forecasting," *Energy Reports*, vol. 6, pp. 123-135, 2020. DOI: https://doi.org/10.1016/j.egyr.2020.01.013.

(20) Fang, H., and Zhou, W., "Comparative Analysis of Air Quality Forecasting Models," *International Journal of Environmental Science and Technology*, vol. 19, no. 4, pp. 123-138, 2022. DOI: https://doi.org/10.1007/s13762-021-03023-9.

(21) Yin, Z., and Zheng, T., "Machine Learning Approaches to Predicting Air Pollutants," *Environmental Science and Technology*, vol. 54, no. 14, pp. 8654-8663, 2020. DOI: https://doi.org/10.1021/acs.est.0c01547.

(22) Zhang, Y., and Chen, Z., "Neural Networks for Air Quality Analysis: A Case Study," *Science of the Total Environment*, vol. 750, pp. 141723, 2021. DOI: https://doi.org/10.1016/j.scitotenv.2021.141723.