Testing the Effectiveness of Machine Learning to Forecast Air Quality Index

AP Research

4845 Words

Author Note

This correlational study and research report were completed in partial fulfillment of requirements for AP Research.

Abstract

This research paper explores the effectiveness of using machine learning to forecast air quality. Air quality is measured through AQI(Air Quality Index), which ranges from 0 to 500. The specific index correlates to the concentration of five pollutants: ground-level ozone, particulate matter, carbon monoxide, sulfur dioxide, and nitrogen dioxide. Thus, a high AQI signals poor air quality. Conventionally, any AQI score above 100 points is considered unsafe. The air quality of a particular region can depend on hundreds of factors, therefore it is very difficult to forecast. Currently, almost all forecasts of air quality are done through intensive on-site research, simple statistical models or even just educated guesses. Machine learning models(complex computational mathematical algorithms), however, are widely regarded as one of the most effective methods to modeling relationships. In this scenario, a machine learning model was trained to identify the relationship between greenhouse gas emissions and population against the AQI of Los Angeles County, California. This was done by training the machine learning model to identify trends and patterns in historical data. Los Angeles County was chosen as the specific area of study because it is already known for having poor air quality. This is important because it allows the machine learning model to train on data that pushes its limits. After multiple rounds of testing, the machine learning model has proved to be effective with an average score of 91.62% accuracy. This discovery is integral to the field of air quality forecasting as it eliminates the uncertainty that comes with traditional air quality forecasts. This discovery also impacts the general population as people can feel more certain about their safety and are less exposed to health risks.

Introduction

Air quality is a critical environmental factor that significantly impacts public health and well being. It is formally defined as, “the degree to which the air is clean enough for humans and the environment”(AirNow 2021). The standard method to measure air quality is by plotting it on the Air Quality Index, otherwise known as the AQI. The specific index directly correlates to the concentration of harmful pollutants in the air. The five pollutants it is based on is ground-level ozone, particulate matter, carbon monoxide, sulfur dioxide, and nitrogen dioxide(AirNow 2021). The concentrations of each of these pollutants are conventionally measured through an air sample, and are plugged into a simple mathematical formula to compute the AQI of the sample. Because the AQI of a sample is directly linked to its pollutant concentrations, a high AQI signals a high concentration of harmful pollutants. The final AQI score ranges from 0 points to 500 points. Despite this, any score above 100 points is considered generally unsafe for select groups of people(American Lung Organization 2019). These select groups include children, senior citizens, and individuals with asthma or other respiratory diseases. Although it would generally take an AQI score of 150 or above to harm the regular person, this research paper will nonetheless consider any AQI above 100 to be a “harmful” score. This is done so that the results of this paper can truly be interpreted univerisally and not conditionally.

The overarching issue to be addressed in this research paper is the lack of concrete, standardized, and accessible methods for forecasting air quality. Accurate air quality forecasting is important because it provides vital data that can be used by public health protection agencies, environment conservation companies, and also urban planners. Moreover, it enables individuals to feel safe about their health and community for years to come. Nonetheless, forecasting air quality is a truly complex task because of how complex atmospheric, industrial processes, and human factors can be. For this reason specifically, the most that meteorologists can do is rely on basic computer models or even just estimates to forecast air quality.

In the field of data analysis, machine learning is an effective tool used to detect trends and patterns in datasets in order to make accurate future predictions. It essentially works by using a collection of recursive high-level computational mathematical algorithms. Additionally, machine learning is extremely scalable and can be used to compute very large amounts of data.

This research paper aims to test the effectiveness of using machine learning to forecast air quality. Prior to starting the research, I have come to the hypothesis that machine learning will, in fact, prove effective at accurately forecasting air quality.

Literature Review

Forecasting air quality index has been a rapidly growing point of interest in the field of meteorological research since the twenty first century. This is reflected by the vast amount of research conducted on the subject. For example, Stephen Vander Hoorn, a senior research fellow at the University of Auckland, worked extensively with chemical transport models (CTMs) in Western Australia to assess air quality and make future predictions. The CTMs he worked with can be described as complex computational chemistry simulation models that rely on extensive prior inputs and parameters that need to be collected by hand. Additionally, the data from the CTMs need to plugged into complex statistical models for final results. While it is definitely true that the combination of groundbreaking CTMs and complex statistical models has proved to be a valid method of air quality forecasting(Hoorn et al., 2022)., its lack of accessibility to those other than scientists can make it an obscure tool to use.

A study by Sarath K. Guttikunda, an atmospheric researcher at Desert Research Insitutue, attempted to forecast air quality with chemical mass balancing in Hyderabad, India. Guttikanda essentially used monitoring stations to capture air samples that would be used to calculate how much air pollution there was, as well as where it was coming from. This would be done through equations that modeled the flow of said pollutants. Guttikanda's method of forecasting air quality through chemical mass balancing proved effective and was even used as

the basis of legislation by policy makers in the Indian state Andhra Pradesh(Guttikanda et al., 2013). However, Guttikanda's approach was quite limited because he was only calculating for particulate matter(PM), only one of the five pollutants that AQI considers. Thus, Guttikonda's approach of chemical mass balancing is only effective for simple forecasts that only consider one pollutant at a time. In order to make an air quality forecasting model that can truly be used everywhere, it just needs to consider more variables.

Another common method of forecasting air quality is through the use of photochemical models. Heather Simmons, a US Environmental Protection Agency(EPA) officer, has proven its accuracy and effectiveness by repeatedly testing it from 2006 to 2012 in Arizona(Simmons et al, 2012). Her study modeled trends by using complex chemical simulations and chemical reaction equations based off a variety of inputs. These inputs included greenhouse gas emissions, weather patterns, chemical kinetics, etc. Out of the three air quality forecasting methods shown so far, Simmon's approach is the most scalable because it is also the most flexible. It essentially combined Hoorn's research(chemical simulation models) with Guttikonda's research(chemical mass balancing) to achieve the best of both worlds. Simmon's research shows the importance of incorporating a large variety of variables to truly model the complex processes involved in atmospheric chemistry and air pollution.

There have also been studies done to forecast air quality that focus on trends and predictive analysis rather than chemistry-based approaches. For instance, Yashon O. Ouma, a statistics proffessor at the University of Botswana, used such type of an approach to analyze air pollution in Africa. In particular, he focused on trends surrounding particulate matter(PM) concentration: similar to Guttikanda's research in Hyderabad, India. He used the *Mann-Kendall Test*, a simple mathematical formula to assess trends(EPA, 2011), to model the fluctations of particulate matter over time. Ouma indeed discovered that the *Mann-Kendall Test* perfectly modeled the historical upwards trend of particulate matter during the time(Ouma et al., 2021). Although Ouma's sole focus on particulate matter made his *Mann-Kendall* approach quite

limited, his research still demonstrates the simplicity, effectiveness, accessibility, and understandability of using trend-based approaches rather than pure chemistry-based approaches.

A study done by Sumita Gulati, a mathematics professor at Jain College, also conducted research on particulate matter that utilized a data-science based approach. Her research was focused on $PM_{2.5}$, or particulate matter with a diameter of 2.5 microns. Gulati successfully modeled the relationship between sulpher dioxide, nitrogen dioxide, and PM_{10}(particulate matter with a 10 micron diameter) using a basic neural network(Gulati et al., 2023). For context, a neural network is a computational framework composed of layers of interconnected nodes, inspired by the structure and function of human brains, capable of learning and performing tasks by processing input data through layers of interconnected neurons(Hardesty, 2017). More specifically, these layers of nodes have “weights” and “biases”, which are the parameters to complex mathematical formulas each node contains. When data is inputted into the neural network, the data travels node by node, constantly recalculating itself, until it reaches its final node. The neural network then computes the R^2(“R” is short for “error”) value, or the square of the difference between the actual output and the output that the neural network has created. Then, the neural network keeps iterating over itself until it computes the smallest possible R^2 value. This is done by using stochastic gradient descent(SGD), a concept in calculus where the weights and biases are changed with respect to the derivative of R^2. The process of constantly using SGD to modify weights and biases of nodes is called backpropogation. After the backpropogation is complete, the neural network is ready to be used to predict values based off the inputs provided. Gulati was able to confirm that her neural network was indeed effective with an R^2 value of only 9.65 units, which, in terms of her project, signals high accuracy and effectiveness. Gulati’s existing work with neural networks is important to this paper because neural networks are actually the foundation of machine learning. Her success with particulate matter, only one of the five components of AQI, demonstrates the likelihood of machine learning

being able to successfully forecast AQI as a whole. The specifics of how I implemented a neural network to forecast air quality will be defined in my methods section.

Wen-Tien Tsai, a researcher at National Pingtung University of Science and Technology, has conducted a trend analysis between AQI and greenhouse gases(GHG) in Pingtung County, Taiwan. By looking at historical AQI scores as well as historical GHG emissions in the past few years, he was quickly and easily able come to the conclusion that high GHG emissions have a correlation to poor AQI scores(Tsai et al., 2021). This hypothesis was confirmed when he specifically looked at data during the Covid-19 time period. Due to the global pandemic, most industrial activity had to be paused, which reduced the GHG emissions of the time. This change in GHG emissions paired with a decrease in an overall AQI score of the area ultimately confirmed and strengthened Tsai's argument. Thus, the correlation between GHG and AQI makes GHG a phenomenal indicator for predicting AQI. Finally, this information will be vital to designing the machine learning model later in the method section.

There has also been research conducted to identify the relationship between AQI and spatial distribution of humans and cities in relation to time. For instance, Renyi Yang, an atmospheric chemistry researcher at Texas A&M university, has conducted an analysis on the spatio-temporal evolution of AQI in China from 2014 to 2021. Yang employed the *Moran's I statistic*, a method used to calculate the degree as to which neighboring spatial units are spread apart within a geographic dataset(Environmental Systems Research Institute, 2023). It was used to create and study spatial distributions of both people and areas with high AQI(also known as a air pollution maps). First, he began by calcuating the *Moran's I statistic* for population distributions in China every year from 2014 to 2021. Next, he modeled the population distributions through population density maps. By definition, population density maps are visual representation of concentrations of human populations across a geographical area(GeographyRealm, 2017). He then calculated the *Moran's I statistic* for AQI data from over 2000 data sampling stations scattered across China. Again, by repeating this process each year

from 2014 to 2021, Yang was able to create comprehensive AQI maps that showed areas with high air pollution. These maps, put together, modeled the flow air pollution in China over time. Next, he compared the population distribution maps and airi pollution maps in China from 2014 to 2021. Yang discovered that the spatial distribution of humans and cities were perfectly correlated to the spatial distribution of high air pollution areas(Yang et al., 2022). Finally, he was able to confirm this argument by comparing the *Moran's I statistic* for both air pollution maps and population density maps over the years. He found that this comparison resulted in 97% overlap, thus proving his argument. In fact, Yang's discovery perfectly aligns with Tsai's discovery from earlier. To recap, Tsai discovered that greenhouse gas(GHG) emissions almost always results in an increased AQI of the area. This makes sense because, as Yang proved, areas with high population density also tend to have high AQI scores. This is because areas with an abundance of people are more likely to have industrial processes. These industrial processes creates GHG emissions which are deteriorating to the environment, as can be seen through increased AQI scores. Both Tsai's research and Yang's research are pivotal to this paper because it provides crucial inputs for the machine learning model. This will be discussed more extensively in the methods section.

Correctly forecasting AQI is indeed a very important task given with how it deals with the health and well-being of the general public. There have even been studies done on the impact of air pollution on an individual's health and happines. For instance, Lu Yiu, a researcher at the Wuhan Institute of Technology, found that air pollution directly impacts one's health as harmful chemicals and toxins can causeresipiory and cardiovascular issues(Yiu et al., 2021). Moreover, Yiu's study delves into the broader implications of health on overall happiness, revealing a strong correlation between the two factors. Such research emphasizes the importance of addressing air pollution not only for physical health but also for mental well-being. Given that public health protection agencies and environmental conservation companies are primarily responsible for mitigating air pollution, accurate AQI forecasting becomes essential. By working

with reliable and accurate forecasts, these entities can better inform policy decisions and interventions aimed at safeguarding public health and promoting environmental conservation. Thus, ensuring proper AQI forecasting is vital for protecting both the health and happiness of the general public.

In the past few years, open source software(OSS) has became a crucial tool to the public to accomplish various tasks and goals. It refers to software that is made by professional developers, yet still available to the public for anyone to use or modify on their own(IBM, 2023). Its emphasis on collaboration, accessibility, and transperancy fosters a dynamic environment of consant growth and advancement. In fact, there have even been studies done on this pheneomeon. For instance, Manuel Hoffmann, a research fellow at Harvard Business School, published a paper that looks at the value of open source software in terms of the economy and global productivity. He found that the demand value of all the open source software in the world is approximately $8.8 trillion and firms would be 300% busier developing software if OSS didn't exist(Hoffman et al., 2024). Although this finding doesn't directly deal with AQI forecasting, it still highlights the potential for advancement in the realm of machine learning and other softwares relating to it.

Gap

The "gap" that this research paper is trying to address is the lack of accurate, standardized, scalable, and accessible means of forecasting air quality. As seen in the literature review, almost all the existing methods of forecasting air quality have limitations that prevent it from checking all of these boxes. For example, some methods of forecasting air quality, such as photochemical modeling, were highly accurate, yet were not accessible to the general public. However, other methods, such as the *Mann-Kendall Test*, were accessible and easy to understand, but only worked on simple outputs like particulate matter(PM). This simplicity made such forecasting methods unpractical for widespread application.

In an attempt to fill the “gap”, I have decided to employee a machine learning model to forecast air quality. As seen in the literature review, machine learning models rely on the concept of “neural networks”: computational models designed to replicate the human brain to “learn” complex patterns within data. The concept of neural networks allow machine learning models to be scalable and accurate at identifying relationships. Additionally, the fact that machine learning relies on programs/code and not complex simulations or chemical equations makes it accessible due to the trends surrounding open source development. Although there has already been some research done on the application of using neural networks to forecast air quality, it only deals with particulate matter and not AQI in general. Although this application of using neural networks is a good start, it doesn’t even scratch the surface of what is possible with machine learning. In summary, the goal is quite simple: create a machine learning model that can forecast air quality to the best of its ability. The machine learning model will take GHG emissions and population of an area as inputs and will provide the forecasted AQI as an output. As explained in the literature review, GHG emissions and population magnitude have shown a positive correlation to poor air quality, making them perfect inputs to forecast AQI. This research, if done correctly, can lead to a whole new realm of possibilities in the field of machine learning in terms of open source development and public wellbeing.

Methodology

Here is the general procedure I am going to take to create a machine learning model to forecast air quality in terms of AQI:

1. Choose a specific location and time frame of data for the model to train on
2. Gather the necessary data: GHG Emmissions, population counts, AQI values
3. Organize the data using a spreadsheet and convert it into a data type that can be used in the program

4. Organize the data within the program into six categories: training input, training output, testing input, testing output, final testing input, final testing output
5. Declare, compile, and train the machine learning model
6. Use machine learning model to forecast values based off the final testing input
7. Compare the predicted output based off the final testing input to the final testing output. This comparison should result in a single percentage from 1-100%; the better the forecast, the closer the comparison, which means a higher overall percentage.
8. Repeat step 7 ten times and average all comparison percentages for one final comparison statistic.

Choosing Location and Time Frame

For my experiment, I chose Los Angeles County, California as the point of research. I did this because Los Angeles County is an area known to have many industrial processes. These industrial processes create greenhouse gas(GHG) emissions, which invariably lead to air pollution and increased AQI. On the other hand, increased awareness about air pollution and environmental conservation in the area has led to an overall decrease in AQI over the years(CleanLA, 2024). Overtime, this combination of human and industrial trends led to GHG emissions and AQI values that seem scattered. For this reason, I chose to look at data from the past 30 years. It is quite easy for a human being to look at such data and characterize it as random. However, there are many underlying trends and processes that happen every single day which contribute to the “randomness” of such type of data. This is ultimately the reason why I have decided to employ a machine learning model: it is literally designed to breakdown these patterns and model them.

Gathering the Necessary Data

Gathering the necessary data is arguably the most important step when it comes to working with machine learning models. This because the model is completely reliant on the data and uses zero outside knowledge. In order for a machine learning model to function, it needs, in simple terms, an input and an output. With this data, the model can then optimize its internal hyperparameters in order to identify the relationship between the input and output. From a very low-level perspective, all it is really doing is creating a mathematical function that takes the input(s) and returns the output(s). The caveat, however, is that these functions can be so complex that even a PH.D. in applied mathematics might not be enough for the average person to interpret it. Therefore, it is almost always done computationally.

The inputs that I will be using is the yearly greenhouse gas(GHG) emission values as well as the yearly population counts of Los Angeles County, CA. The reason I chose to use GHG emission values is quite self-explanatory: GHG emissions have harmful pollutants which contribute to air pollution. Moreover, Wen-Tien Tsai, a researcher who wrote a paper I studied in my literature review, confirmed this relationship by conducting a trend analysis between AQI and GHG emissions in Taiwan. The reason I used yearly population counts as one of my inputs is also quite self explanatory: where there is people, there is also pollution. Additionally, this phenomon was justified through a paper I studied in my literature review. This paper was written by Lu Yiu, a researcher at the Wuhan Institute of Technology, and involved comparing the flow of air pollution with that of people overtime. Keep in mind that I am not citing these authors because I have already cited them in the literature review itself.

The output that the machine learning model is going to ultimately forecast is the amount of days where the AQI is above 100 for that specific year. The reason I am choosing to do this output rather than just raw AQI is so that it can match up with my inputs in terms of time. My inputs are measured every year, so it fundamentally makes sense to also have outputs that are

also measured every year. The reason I specifically chose 100 on the AQI scale is because it is considered harmful to breathe in for select groups of people(AirNow 2021).

A challenge that I encountered while building my dataset was that I couldn't find the specific GHG emission values from Los Angeles County, CA. However, I came across a paper that helped me find an alternative. This research paper stated that the yearly GHG emission values of Los Angeles County were approximately 25% of the yearly GHG emission values of the State of California(Wenzel et al., 2018). Equipped with this information, I went ahead and simply divided the California GHG emission values by 4, as this data was much easier to find.

To build the dataset that my machine learning model was going to work with, I had to do extensive research and get my data for a variety of sources. First, I gathered the historical AQI values of Los Angeles County from Los Angeles Almanac(Laalmanac, 2023). Then, I combined the GHG emission values of California from 1990-1999, which I found in a research paper(Choate et al., 2000)., and 2000-2022, which I found on an environmental trends report(California Air Resources Board, 2023). Next, I found the historical population counts on Los Angeles Almanac from 1990-2022(Laalmanac, 2023). Lastly, I divided the GHG emissions from California by a factor of 4 to approximate the GHG emissions in Los Angeles County specifically.

Finally, I need to define final input and output values that are seperate to the machine learning model itself. This helps remove any sort of bias. In order to do this, I am going to only pass values from 2014 and before into the machine learning model while using values after 2015 for the final testing purposes.

Here is all the data I have collected and organized. I have put it into a format which I can directly input into the machine learning program.

Year	California Statewide GHG* Emissions(MMTCO2 Eq.*)	Los Angeles County GHG Emissions(MMTCO2 Eq.)	Population	>100 AQI Day Count
2022	370	93	9834503	94
2021	380	95	9942011	96
2020	370	93	10014009	137
2019	405	101	10163139	86
2018	410	103	10192593	108
2017	410	103	10181162	120
2016	412	103	10150386	104
2015	427	107	10124800	132
2014	428	107	10078942	108
2013	431	108	10025721	100
2012	435	109	9956888	130
2011	437	109	9881070	121
2010	440	110	9822121	114
2009	450	113	9801096	118
2008	480	120	9785474	122
2007	483	121	9780808	127
2006	477	119	9798609	129
2005	478	120	9816153	138
2004	485	121	9806944	158
2003	476	119	9756914	170
2002	476	119	9679212	179
2001	479	120	9590080	186
2000	460	115	9477651	175
1999	409	102	9330171	168
1998	398	100	9225813	118
1997	392	98	9147051	161
1996	391	98	9104719	170
1995	392	98	9103896	214
1994	397	99	9095157	219
1993	379	95	9072159	228
1992	383	96	9007999	240
1991	385	96	8908269	232
1990	399	100	8832500	230

Program

Now I am going to do a breakdown of the program that I have built.

```
#section 1
import pandas as pd
from tensorflow import keras
from keras import layers
import numpy as np

#section #2
df = pd.read_csv('AP_Research/data.csv')
df = df.sample(frac=1).reset_index(drop=True)


#section 3
input_factors = ['California Statewide GHG Emissions(MMTCO2 Eq.)',
                 'Los Angeles County GHG Emissions(MMTCO2 Eq.)',
                 'Population']
training_input = df.loc[8:24, input_factors]
training_output = df.loc[8:24, '>100 AQI Day Count']
testing_input = df.loc[25:32, input_factors]
testing_output = df.loc[25:32, '>100 AQI Day Count']
final_testing_input = df.loc[0:7, input_factors]
final_testing_output = df.loc[0:7, '>100 AQI Day Count']
```

In section 1, I am importing all the necessary libaries required for this project. A library is essentially just prewritten and open source code that anyone can use. The two main libraries are *Pandas* and *TensorFlow*. Pandas is a library meant for working with data tables while TensorFlow is the actual machine learning model I am going to implement.

In section 2, I am simply importing the data spreadsheet I have defined earlier. This was done through a specialized pandas function.

In section 3, I am defining what the specific inputs and outputs are. The output is the “>100 AQI Day Count” column and the inputs are all the columns of the spreadsheet except the “>100 AQI Day Count” column. The training and testing inputs and outputs are randomized to remove bias and are selected from the spreadsheet with a ratio of 7:3, respectively. Lastly, I

have defined the final testing inputs and output, which are data that is independent of the machine learning model. Again, this helps remove bias. The typical training/testing data for the model are all the values from 2014 and before while the final testing inputs and output is from 2015 and after.

```
#section 4
model = keras.Sequential([
    layers.BatchNormalization(input_shape=[3]),
    layers.Dense(6, activation='relu'),
    layers.BatchNormalization(),
    layers.Dense(6, activation='relu'),
    layers.Dropout(0.3),  # Dropout layer to prevent overfitting
    layers.Dense(1),])

#section 5
model.compile(optimizer='adam', loss='mae')

#section 6
history = model.fit(
    training_input, training_output,
    validation_data=(testing_input, testing_output),
    batch_size=3,
    epochs=25,
    verbose=0)
```

In section 4, I am declaring the actual machine learning model. As I have described before, this is done through the use of a neural network. However, they are extremely complex to both explain and understand. In fact, just to build a basic one from scratch, you would probably need a degree in applied mathematics. So for this reason, I am simply just using the prebuilt TensorFlow neural network. In this section, I am declaring the “shape” of the machine neural network, which plays a large effect on the actual machine learning process itself.

In section 5, I am allowing the machine learning model to “compile” itself. What this means is that the model is giving itself the initial hyperparameters which are randomized. It can sort of be viewed as the brain of a newborn baby.

In section 6, the actual “machine learning” is being done. Through complex mathematical algorithms and processes, it is understanding the underlying trends within the data. Again, this is extremely difficult to understand without an existing knowledge of high level mathematics, so I am not going to address this too much. But to put it simply, it is the computer equivalent of a child growing up and understanding the world around them.

```
41  #section 7
42  accuracy = 0
43
44  for i in range(8):
45      input_data = np.array([[final_testing_input.iloc[i,0] ,
46                              final_testing_input.iloc[i,1],
47                              final_testing_input.iloc[i,2]]])
48      p = model.predict(input_data)
49      prediction = p[0][0]
50      actual = final_testing_output[i]
51      difference = abs(actual-prediction)
52      accuracy += 1-difference/actual
53
54  print(str(round((100*(accuracy/8)),2))+'% Accuracy')
```

Lastly, in section 7, I am using the machine learning model to forecast the yearly AQI of 2015 and after so it can then compared with the actual yearly AQI of 2015 and after. In other words, it is meant to represent overlap. Here, I have built a simple algorithm to return that overlap in the form of a percentage.

Results

After running it 10 times, here are the final percentages the program has computed. As seen in the methods section, these percentages are meant to represent the overlap between the forecasted AQI and actual AQI of Los Angeles County, CA from 2015 and after. This overlap is used to measure accuracy.

Test #	Accuracy	Test #	Accuracy
Test #1	87.38%	Test #6	89.76%
Test #2	92.46%	Test #7	91.09%
Test #3	94.91%	Test #8	95.92%
Test #4	90.18%	Test #9	93.88%
Test #5	88.23%	Test #10	92.37%
Highest Accuracy			95.92%
Lowest Accuracy			87.38%
Range			8.54%
Average Accuracy			**91.62%**

As seen in the table, the average accuracy of my machine learning model was 91.62%. In terms of machine learning, this percentage signals high overlap and accuracy.

Discussion

Implications

The success of my machine learning program has many real world implications. From a fundamental standpoint, it has proved to the world that machine learning can, in fact, forecast air quality accurately. It is the first step to "filling the gap". The gap, as mentioned before, is the lack of scalable and widespread methods to forecasting air quality. As seen in the literature review, machine learning programs are scalable due to the nature of open source software

development. In other words, it is scalable because it is something that people can easily access, use, tweak, and improve. This is much better than the existing methods of forecasting air quality, which include labaratory grade chemical simulations or intensive chemcial mass balancing approaches.

Looking at the implications from a more specific standpoint, people can now use basic open source machine learning models to forecast air quality themselves. Governments and environmental protection agencies can also use it as the basis for taking environmental protection measures. Lastly, this discovery also helps advance the field of machine learning in general and gives a small look into the future of its development.

Limitations

There may have been some limitations that could have led to a flawed analysis. For example, industrial grade machine learning models work with hundreds if not thousands of lines of data. This is because they are funded by large companies which are able to buy this data off other companies or government officials. However, my machine learning model only utilized public data due to time and financial constraints. Because of these circumstances, I was only able to use about 30 lines of data.

Another limitation that could have flawed my analysis was the fact that I only forecasted the AQI of one place and not multiple. This could have created bias and ultimately a machine learning program that is less effective overall. However, by adding multiple places, I could have ran into an an opposing problem: a less focused study which is more prone to being incorrectly analyzed. For this reason, I chose to keep it simple and just forecast the AQI one place.

Conclusion and Future Research

In conclusion, the evidence supports my initial hypothesis that machine learning can, in fact, accurately forecast air quality. My machine learning program, as seen in the methodology, had an average accuracy of 91.62%, which signals that the experiment was ultimately successful. Also, as seen in the discussion section, this discovery has many powerful implications in the real world.

In terms of future research, I want to keep improving my machine learning model so it can be truly universal. For example, I could use data from more places, rather than just Los Angeles, CA. Moving forward, I could also take into account the specific sources of GHG emissions for a more in depth analysis. Also, I could also try testing various types of machine learning models rather than just the prebuilt one that TensorFlow has developed. All in all, through extensive research and development, machine learning has the potential to solve tommorow's problems.

References

Air Quality Days by Year for Los Angeles County, California. (n.d.). www.laalmanac.com. https://www.laalmanac.com/environment/ev01b.php

Air Quality Index | American Lung Association. (n.d.). Www.lung.org. https://www.lung.org/clean-air/outdoors/air-quality-index

AirNow. (2021). Air Quality Index (AQI) Basics. Www.airnow.gov. https://www.airnow.gov/aqi/aqi-basics/

Alford, J. (n.d.). Home. CleanLA. https://cleanla.lacounty.gov/

Altaweel, M. (2017, October 27). Density Mapping With GIS. Geography Realm. https://www.geographyrealm.com/density-mapping/

ArcGIS Pro. (n.d.). How Spatial Autocorrelation (Global Moran's I) works—ArcGIS Pro | Documentation. Pro.arcgis.com. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm

California Greenhouse Gas Emissions from 2000 to 2021: Trends of Emissions and Other Indicators. (2023). https://ww2.arb.ca.gov/sites/default/files/2023-12/2000_2021_ghg_inventory_trends.pdf

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis

evaluation. PeerJ Computer Science, 7(5), e623. ncbi. https://doi.org/10.7717/peerj-cs.623

Choate, A., Freed, R., Gibbs, M., Coe, D., Hyslop, N., & Franco, G. (n.d.). California's Greenhouse Gas Emissions and Trends over the Past Decade. https://www3.epa.gov/ttn/chief/conference/ei11/ghg/choate.pdf

Estimated Population for Los Angeles County, California. (n.d.). Www.laalmanac.com. https://www.laalmanac.com/population/po04.php

Finlayson-Pitts, B. J. (2010). Atmospheric Chemistry. Proceedings of the National Academy of Sciences, 107(15), 6566–6567. https://doi.org/10.1073/pnas.1003038107

Gulati, S., Bansal, A., Pal, A., Mittal, N., Sharma, A., & Gared, F. (2023). Estimating PM2.5 utilizing multiple linear regression and ANN techniques. Scientific Reports, 13(1), 22578. https://doi.org/10.1038/s41598-023-49717-7

Guttikunda, S. K., Kopakka, R. V., Dasari, P., & Gertler, A. W. (2012). Receptor model-based source apportionment of particulate pollution in Hyderabad, India. Environmental Monitoring and Assessment, 185(7), 5585–5593. https://doi.org/10.1007/s10661-012-2969-2

Hardesty, L. (2017, April 14). Explained: Neural networks. MIT News; MIT. https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

Hoffmann, M., Nagle, F., & Zhou, Y. (2024). The Value of Open Source Software. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4693148

Liu, Y., Zhu, K., Li, R.-L., Song, Y., & Zhang, Z.-J. (2021). Air Pollution Impairs Subjective Happiness by Damaging Their Health. International Journal of Environmental Research and Public Health, 18(19), 10319. https://doi.org/10.3390/ijerph181910319

Mann-Kendall Test for monotonic trend. Design Trend Mann-Kendall. (n.d.). https://vsp.pnnl.gov/help/Vsample/Design_Trend_Mann_Kendall.htm

Mishra, P., Pandey, C. M., Singh, U., Keshri, A., & Sabaretnam, M. (2019). Selection of Appropriate Statistical Methods for Data Analysis. Annals of Cardiac Anaesthesia, 22(3), 297–301. NCBI. https://doi.org/10.4103/aca.ACA_248_18

Ouma, Y. O., Keitsile, A., Lottering, L., Nkwae, B., & Odirile, P. (2024). Spatiotemporal empirical analysis of Particulate Matter PM2.5 pollution and air quality index (AQI) trends in Africa using merra-2 reanalysis datasets (1980–2021). Science of The Total Environment, 912, 169027. https://doi.org/10.1016/j.scitotenv.2023.169027

Simon, H., Baker, K. R., & Phillips, S. (2012). Compilation and interpretation of photochemical model performance statistics published between 2006 and 2012. Atmospheric Environment, 61, 124–139. https://doi.org/10.1016/j.atmosenv.2012.07.012

TensorFlow: A system for large-scale machine learning | Request PDF. (n.d.). ResearchGate. https://www.researchgate.net/publication/303657108_TensorFlow_A_system_for_large-scale_machine_learning

Tsai, W.-T., & Lin, Y.-Q. (2021). Trend Analysis of Air Quality Index (AQI) and Greenhouse Gas (GHG) Emissions in Taiwan and Their Regulatory Countermeasures. Environments, 8(4), 29. https://doi.org/10.3390/environments8040029

Vander Hoorn, S., Johnson, J. S., Murray, K., Smit, R., Heyworth, J., Lam, S., & Cope, M. (2022). Emulation of a Chemical Transport Model to Assess Air Quality under Future Emission Scenarios for the Southwest of Western Australia. Atmosphere, 13(12), 2009. https://doi.org/10.3390/atmos13122009

Wenzel, T., & Fischer, M. (2008, June 11). Spatial Disaggregation of CO2 Emissions for the State of California. Www.osti.gov. https://www.osti.gov/biblio/935753

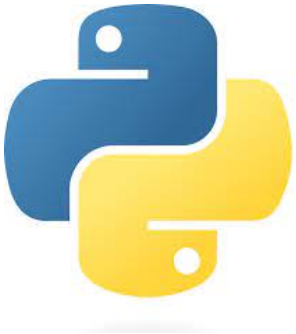
What Is Open Source Software? | IBM. (n.d.). Www.ibm.com. https://www.ibm.com/topics/open-source#:~:text=Open%20source%20software%20is%20software

Yang, R., & Zhong, C. (2022). Analysis on Spatio-Temporal Evolution and Influencing Factors of Air Quality Index (AQI) in China. Toxics, 10(12), 712. https://doi.org/10.3390/toxics10120712

Appendix A

Tools, platforms, etc.

Python Programming Language



Visual Studio Code Application

TensorFlow Library

Pandas Library

Appendix B

Initial Data

Year	California Statewide GHG* Emissions(MMTCO2 Eq.*)	Los Angeles County GHG Emissions(MMTCO2 Eq.)	Population	>100 AQI Day Count
2022	370	93	9834503	94
2021	380	95	9942011	96
2020	370	93	10014009	137
2019	405	101	10163139	86
2018	410	103	10192593	108
2017	410	103	10181162	120
2016	412	103	10150386	104
2015	427	107	10124800	132
2014	428	107	10078942	108
2013	431	108	10025721	100
2012	435	109	9956888	130
2011	437	109	9881070	121
2010	440	110	9822121	114
2009	450	113	9801096	118
2008	480	120	9785474	122
2007	483	121	9780808	127
2006	477	119	9798609	129
2005	478	120	9816153	138
2004	485	121	9806944	158
2003	476	119	9756914	170
2002	476	119	9679212	179
2001	479	120	9590080	186
2000	460	115	9477651	175
1999	409	102	9330171	168
1998	398	100	9225813	118
1997	392	98	9147051	161
1996	391	98	9104719	170
1995	392	98	9103896	214
1994	397	99	9095157	219
1993	379	95	9072159	228
1992	383	96	9007999	240
1991	385	96	8908269	232
1990	399	100	8832500	230

Final Results

Test #	Accuracy	Test #	Accuracy
Test #1	87.38%	Test #6	89.76%
Test #2	92.46%	Test #7	91.09%
Test #3	94.91%	Test #8	95.92%
Test #4	90.18%	Test #9	93.88%
Test #5	88.23%	Test #10	92.37%
Highest Accuracy			95.92%
Lowest Accuracy			87.38%
Range			8.54%
Average Accuracy			**91.62%**