Weather Forecasting and Air Quality Prediction: Leveraging Machine Learning for Environmental Insights

Aditya Arun, Vaibhav Bommisetty, Housheng Hai, Mark Sui, Zihao Yang

ARTICLE HISTORY:
Compiled March 16th, 2024

ABSTRACT

This study is dedicated to harnessing the capabilities of machine learning methodologies for the purpose of forecasting weather phenomena and estimating air quality indicators. Exploratory data analysis and predictive modeling methodologies are used to scrutinize datasets encompassing weather patterns and air pollution metrics. Specifically, polynomial regression and random forest regression models are deployed to anticipate precipitation levels and air quality index (AQI) values. Through a comprehensive exploration of the data, an analysis of feature importance, and rigorous model evaluation procedures, the study aims to interpret the intricate relationship between meteorological factors and air quality, thereby offering valuable insights for the advancement of environmental monitoring and management practices.

1 Introduction

Research Question: Which factors have significantly influenced precipitation and pollution levels in San Diego over recent decades?

Background: Precipitation and pollution are two important variables that have a strong impact on the environment. Both precipitation patterns and pollution levels significantly impact public health, with air pollution exacerbating respiratory issues and changes in precipitation affecting water quality and availability, potentially leading to waterborne diseases. Poor air quality can result in adults and children developing conditions such as asthma and bronchitis, and prolonged exposure can increase the risk of cardiovascular diseases. Changes in precipitation can affect the availability of water, overwhelming the sewer system and leading to the contamination of potable water. A large amount of precipitation also has the potential to be a danger to infrastructure as it could result in flooding and the destruction of property.

Precipitation and pollution are pivotal environmental parameters. Both of these parameters significantly impact public health and environmental sustainability. Air pollution, for instance, exacerbates respiratory ailments, while alterations in precipitation patterns directly influence water quality and availability, thereby causing potential waterborne diseases. The adverse health outcomes linked to poor air quality, including asthma and bronchitis in both adults and

children, alongside heightened risks of cardiovascular diseases, underscore the urgency of mitigating pollution levels. Moreover, variations in precipitation pose substantial challenges to water resource management, potentially overwhelming sewer systems and contaminating water sources. Additionally, excessive precipitation events, such as floods, are threats to infrastructure integrity and public safety, accentuating the need for robust environmental stewardship strategies.

Data:

Precipitation Dataset Hourly precipitation data for San Diego County from 2004 through 2014. This dataset can be used to analyze precipitation patterns in detail over ten years.[1] *Air Quality Datasets* US EPA's AirData contains annual concentration by monitor, annual AQI(Air Quality Index), and annual AQI by county for various years for the San Diego area.[2] The San Ysidro Air Study is San Diego's effort to collect community air pollution data using advanced, low-cost technology. Assess community air quality needs and concerns. Install 13 next-generation, low-cost sensors in the community to assess air quality. Collect air quality data on particulate matter (PM2.5), ozone, nitrous oxide, nitrogen dioxide, and carbon monoxide.[3]

Data Usage:

- Meteorological data: obtain historical precipitation data for san diego from reputable sources such as NOAA

- Access data on temperature, humanity, precipitation, and analyze the potential influence on precipitation level.

- Air quality data: gather air quality data from agencies like EPC Acquire information from such as PM2.5, nitrogen dioxide, ozone,

- Historical records and studies: review existing data research studies, environmental assessments, and reports on San Diego. Consult historical records, new archives, and academic literature.

2 Analysis

2.1 Method

Preparation: We import the necessary libraries; "pandas" and "numpy", and loads the precipitation dataset (https://data.sandiegodata.org/dataset/historic-precipitation-for-san-diego-county/) from a CSV file into a pandas DataFrame called "precipitation". Then, we load air pollution dataset (https://www.epa.gov/outdoor-air-quality-data/download-daily-data) from a CSV file into a DataFrame called "air_pollution".

Data Processing: Then we decide to filter the precipitation DataFrame to include only the data from January 1, 2019, onward, using a specific year as a training dataset to predict how the Precipitation trend be like over a year. It does this by checking if the 'Date' column values are greater than or equal to '2019-01-01'. Also, we replace all instances of -999.9 in the precipitation DataFrame with 0. This is typically done to handle missing or placeholder values in datasets. We filter "air_pollution" dataset to include only data from

the 'Chula Vista' site as a specific location we chose as a representation of whole San Diego weather performance. Lastly, we convert the 'Date' columns in both precipitation and air_pollution DataFrames from "string" type to pandas "datetime" type, enabling easier manipulation and comparison of dates.

Data Merging: The precipitation and air pollution datasets are merged on the 'Date' column, resulting in a combined dataset that includes both types of data on matching dates. Therefore, we cleaned and merged two dataset together.

2.2 Exploratory Data Analysis

Exploratory Data Analysis allows for insights into the distributions of features, relationships among features, and notably, aids in the selection of features for model development. Subsequently, our model is leveraging machine learning to predict precipitation and pollution levels in San Diego. Plotting the features facilitates which features could be the most impactful for our model.

Data Visualizations

The histograms below allow for distributions of key features to be viewed with ease. These plots tell us that apart from the distribution of precipitation, PM2.5, and AQI daily values, the distributions are fairly normal.

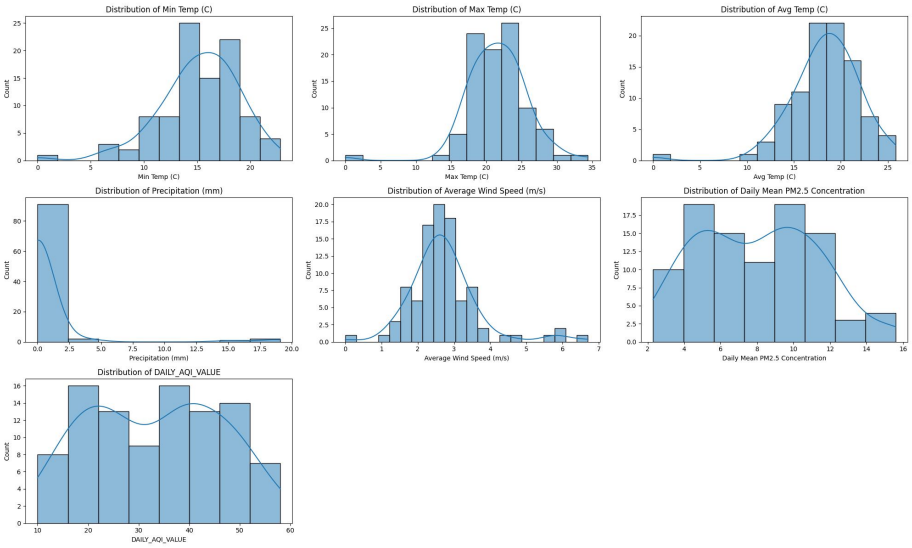

Figure 1: Histograms of All Features

The pair plot below presents scatter plots between all the features. It shows what the relationship between certain features can look like. Since we are focusing on precipitation levels and AQI levels, we should look at the fourth and seventh row. For the fourth row, none of these scatter plots seem to indicate a strong correlation; however, in the seventh row, we can see a strong linear correlation between AQI and PM2.5. This is the case because AQI levels are measured directly by PM2.5 levels.

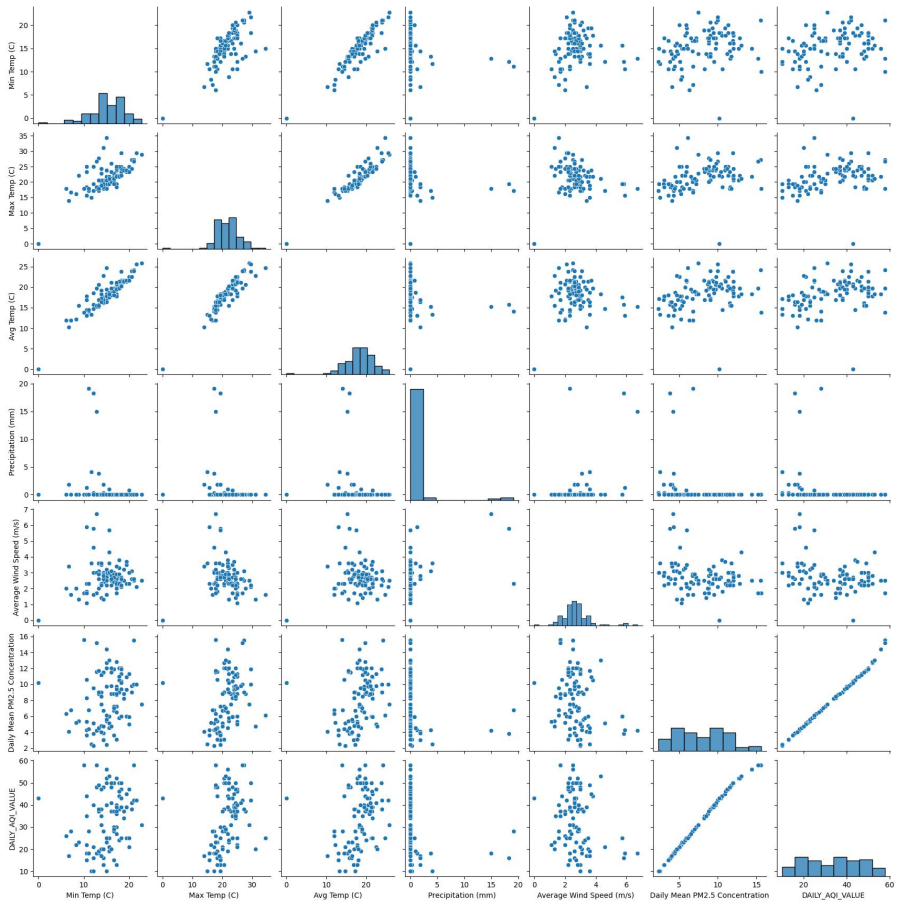

Figure 2: Pair Plot of All Features

As the previous graph did not provide clear insights into the relationships between features, a correlation heatmap was generated for insights into the relations among all features:

The heatmap below confirms existing knowledge, chiefly revealing weak associations between the features. As anticipated, the average temperature exhibits

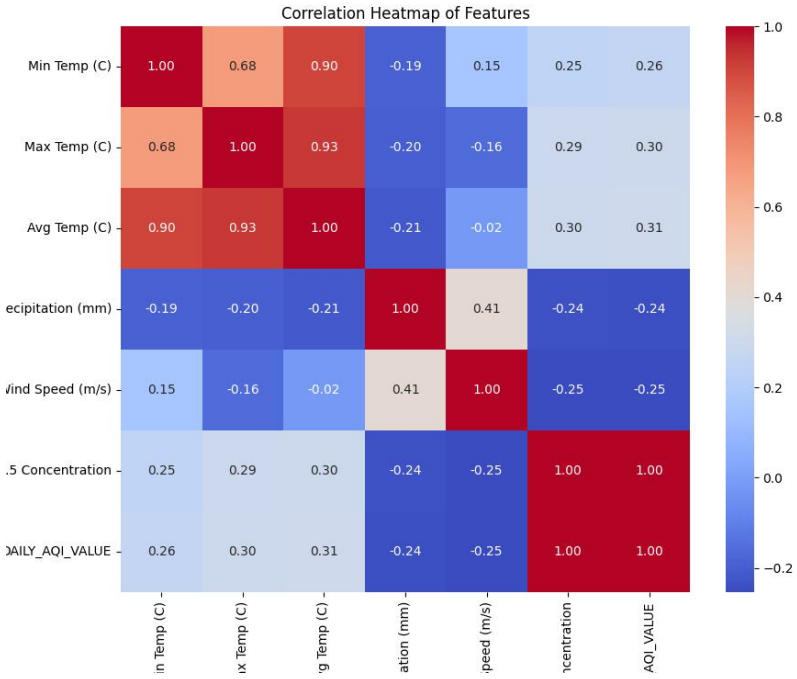

Figure 3: Correlation Heatmap of All Features

a strong correlation with both high and low temperatures throughout the day. However, a moderate correlation between precipitation and wind speed is observed. This insight holds potential significance for the construction of our model.

2.3 Predictive Modeling

For our research question, the core of our study is to develop and evaluate forecasting models using polynomial regression and random forest regression. Because our goal is to predict precipitation and to predict the value of the Air Quality Index (AQI) based on various environmental factors.

Polynomial Regression Models Our exploration into polynomial regression revealed distinct outcomes for precipitation and pollution levels. For precipitation, negative R^2 value of -0.5066785312099893. The Mean Squared Error (MSE) stood at 26.042335740552183. The pollution model with an R^2 value of 0.9924422042250075. The MSE for pollution was significantly lower, at 1.4097367514199173.

Random Forest Regression Models

3 Feature Importance Analysis

3.1 Precipitation Prediction Model

Feature importances from the Random Forest regression model trained to predict precipitation levels show the relative importance of each input feature:

- **Min Temp (C)**: 2.75%
- **Max Temp (C)**: 8.39%
- **Avg Temp (C)**: 4.43%
- **Average Wind Speed (m/s)**: 84.43%

Insight: For precipitation prediction, **Average Wind Speed (m/s)** is the most significant predictor, contributing to approximately 82.82% of the model's decisions. This suggests a strong association between wind speed and precipitation levels. Temperature variables (min, max, and avg) play minor roles.

3.2 Pollution Prediction Model

Feature importances from the Random Forest regression model trained to predict pollution levels (Daily AQI values):

- **Min Temp (C)**: 14.04%
- **Max Temp (C)**: 27.53%
- **Avg Temp (C)**: 26.92%
- **Average Wind Speed (m/s)**: 31.52%

Insight: For pollution prediction, the **Average Wind Speed (m/s)** is the most significant predictor, accounting for approximately 31.52% of the importance. Other features, including temperature variables and wind speed, contribute separately. Thus, it means this model doesn't rely on one specific parameter and all attributes work with the same importance.

3.3 Summary

- The **precipitation model** indicates that **wind speed** is the primary factor influencing precipitation levels, suggesting significant effects of wind speed variations on precipitation.
- The **pollution model** highlights the **wind speed** as the primary factor influencing AQI predictions, but also illustrating all other attributes has high importances concurrently, showing the importance of every single parameter to the air quality.

4 Conclusion

Interpretation of the results:

Looking at the analysis above, the 2 key things to note are: the relationship between various environmental factors (minimum temperature, maximum temperature, average temperature, average wind speed, and daily mean PM2.5 concentration) and the key outcomes which is the precipitation and pollution levels in San Diego over the period 2004-2014. From the polynomial regression analysis, the model's performance differs significantly between precipitation and pollution outcomes. The precipitation model shows a negative R^2 value (-0.5066785312099893), which indicates that the model does not fit the data well. This could be due to several factors such as other influencing factors not included in the dataset. The Mean Squared Error for precipitation is relatively high (26.042335740552183), further supporting the conclusion that the model's predictive capability is poor. On the other hand, the pollution model shows an R^2 value of 0.9924422042250075, which indicates an excellent fit between the model predictions and the actual data. The MSE for pollution is quite low (1.4097367514199173), suggesting that the model's predictions are very close to the observed values. The Random Forest analysis provides additional insights, particularly regarding the relative importance of each predictor variable. For precipitation, the most influential factor is average wind speed (0.82820667), suggesting a strong relationship between wind patterns and precipitation levels. For pollution, the daily mean PM2.5 concentration is overwhelmingly the most significant predictor (0.99045862), highlighting the direct impact of particulate matter on air quality. The scatterplot from the 'time series of hourly precipitation in San Diego (2004-2014)' shows that high precipitation events are relatively rare and do not follow a clear seasonal or yearly pattern. The histogram of 'Distribution of hourly precipitation in san Diego (2004-2014)' indicates that most of the precipitation measurements are close to 0, which aligns with San Diego's generally dry climate but also highlights the occasional occurrence of significant precipitation events.

Conclusion and discussion for future work:

Our project has provided a mixed picture of the factors influencing precipitation and pollution levels in San Diego. The pollution model has proven to be highly effective, demonstrating a clear link between PM2.5 levels and pollution outcomes. This finding underscores the importance of controlling particulate matter emissions to maintain good air quality. Future work in this area could explore additional pollutants and their sources, further refining our understanding of air quality dynamics. The precipitation model performed poorly on the other hand, suggesting that the factors considered in this analysis are not sufficient to explain precipitation patterns in the region. This result indicates that other variables, possibly including oceanic conditions, atmospheric pressure, or more localized climate factors, might play significant roles in affecting San Diego's precipitation (since San Diego is a coastal Pacific city). Future research can consider incorporating these factors into the model to better understand and predict rainfall patterns. Exploring the interactions between pollution

and precipitation could also provide valuable insights, as pollutants can affect cloud formation and, consequently, precipitation patterns. Understanding these complex relationships could significantly improve predictive models and inform environmental policy and planning. Further analysis should look at when and how intense rainfall and pollution events happen, as knowing when these extreme events occur is very important for public health and city planning. This is especially true as the climate changes and cities grow.

References:

[**1**] Busboom, Eric. "Historic Precipitation For San Diego County." National Oceanic and Atmospheric Administration https://data.sandiegodata.org/dataset/historic-precipitation-for-san-diego-county. San Diego, May 5, 2019. Precipitation Dataset

[**2**] Environmental Protection Agency. "Outdoor Air Quality Data." Environmental Protection Agency, https://www.epa.gov/outdoor-air-quality-data/download-daily-data. San Diego, 2019. Air Quality Dataset

[**3**] "San Ysidro Community Air Study." Office of Environmental Health Hazard Assessment, 9 Jan. 2018, oehha.ca.gov/calenviroscreen/general-info/san-ysidro-community-air-study.