MACHINE LEARNING-BASED SEASONAL AIR QUALITY SUSCEPTIBILITY MAPPING FOR MILAN CITY IN ITALY

Julius Maina

The author is a research student in data analytics at KCA University in Kenya

October 31st, 2023

Abstract *– The European Air Quality Index (AQI) serves as a vital tool for assessing air quality throughout Europe, using data from key pollutants like PM10, PM2.5, O3, NO2, and SO2. The AQI computation adheres to guidelines provided by the European Environment Agency. This paper explores the utilization of machine learning methods to predict AQI levels for Milan in 2022, with a primary focus on accounting for seasonal variations. The study aims to create AQI susceptibility maps for each season, taking into account various influencing factors, particularly meteorological conditions.*

Keywords – Air Quality Index, Seasonal Variations, Geospatial Data, Sustainable Development Goals, Interpolation Techniques.

1. INTRODUCTION

The city of Milan, nestled within the Lombardy region of Italy, presents a unique air quality challenge. Surrounded by the formidable Alps in the north and west, and the Apennines in the south, Milan encounters reduced wind circulation in the Po Valley. This geographical quirk leads to the accumulation of air pollutants, contributing to the region's reputation as a pollution hotspot.

Air pollution, defined as the presence of hazardous chemicals or compounds in the atmosphere at levels posing health risks, demands careful attention. Pollution can be evaluated individually for each pollutant or aggregated into an air quality index, offering a comprehensive view of the overall pollution landscape. In this paper the latter is pursued.

The primary aim of this paper is to showcase how we can harness the power of machine learning to produce city-level air pollution susceptibility maps with a remarkable 5-meter spatial resolution. These maps are instrumental for informed decision-making, enhancing the resilience of the city in the face of environmental challenges.

By addressing this challenge, we make significant contributions to Sustainable Development Goals (SDGs) 11 and 13, focusing on building sustainable cities and taking urgent actions to combat climate change and its repercussions. This endeavor aligns our efforts with broader global objectives and fosters a resilient and environmentally conscious future.

2. DATASET

The training dataset was recorded primarily from the metropolitan area and not the city itself because there were more training points in the metropolitan area. The locations for the train data are shown in figure 2 as blue dots.

The train data provided encompassed historical in-situ stations data, including meteorological and air pollution data, with a minimum hourly temporal resolution. Moreover, geospatial data detailing the topography of Milan, Italy, was provided. The data are explained below:

- ☐ Air Quality Data -a comprehensive time series dataset covering the years 2016 to 2021 was provided. It includes daily concentration values of key air pollutants: *Particulate Matter 10 (PM10), Particulate Matter 2.5 (PM2.5), Ozone (O3), Nitrogen Dioxide (NO2), Sulfur Dioxide (SO2), Carbon Monoxide (CO), Benzene and AQI (our target variable).* All these could be estimated, but only AQI was considered for this work.
- ☐ Meteorological Data-meteorological time series (2016-2022) was provided. This dataset comprises daily measurements of crucial meteorological parameters, including: *Temperature, Precipitation, Relative Humidity, Solar Radiation, Wind Speed, Wind Direction*
- ☐ Geospatial Data-Digital Terrain Model containing topographic information related to the terrain of the metropolitan area were provided. Key attributes include: *Elevation, Aspect, Hill Shade, Slope*. This data also included curvature and terrain indexes-information on *plan curvature,*

profile curvature, SPI (Standardized Precipitation Index), TRI (Terrain Ruggedness Index), and TWI (Topographic Wetness Index) at both station locations and 100-meter resolution.

- □ Land Cover Maps- detailed maps depicting land cover types within the region. Features such as *compass directions (E, W, N, S, NE, NW, SE, SW), locations* such as *lc_11.* These data were masked and no much details about them can be given.
- □ Geological Map - data showcasing the geological characteristics of the area e.g. *geologia, geo_0* etc. These data were also masked and no much details about them can be given.

In total the train data had a total of 62 columns including our target variable (AQI).

On the other hand, the test dataset had only 3 features: *latitude, longitude and season*. This was due to the limited number of monitoring stations within the city resulting in sparse data representation as shown in Figure 1.

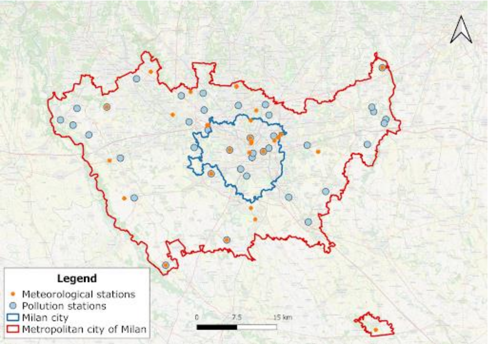

Figure 1-Location of the meteorological and air quality stations in the metropolitan city of Milan.

As such additional seasonal data were provided for the period 2022 (our target period). These data were a regular grid of interpolated data. The locations for these data are as shown in Figure 2.

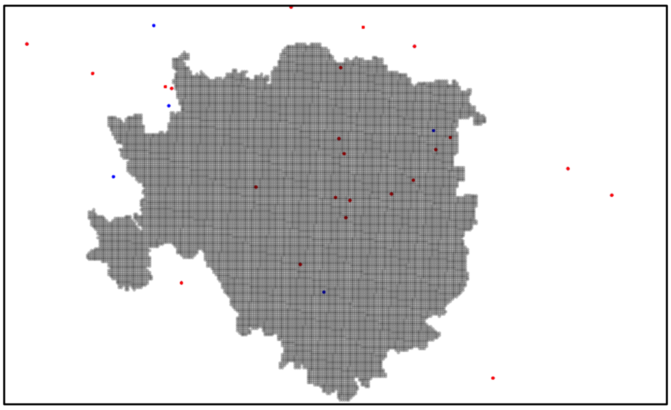

Figure 2 - Locations of seasonal interpolated data for year 2022 (shown in grey)

Additional topological data for Milan city were also provided. The locations for the test data are shown in Figure 2 as red dots.

2.1 Data Preprocessing

In this project, we focused on enhancing the test dataset by incorporating additional data, including meteorological and topological features. The seasonal interpolated data for the year 2022 provided a valuable resource, as it already contained 20 features, combining meteorological and land cover map data for four seasons. These 4 seasonal datasets of same shape were merged to create one comprehensive dataset.

We further enriched this seasonal dataset by including additional topographic data provided for Milan City, which contained 36 features. The compatibility of longitude and latitude values in both datasets enabled a seamless merger. As a result, we achieved a dataset with a total of 53 unique features.

To build up our final test dataset with these 53 features, we employed a custom Inverse Distance Weighting (IDW) interpolation technique. This method calculated missing values by considering the weighted average of neighboring known values, inspired by IDW interpolation widely used in geospatial analysis and environmental modeling [1]. Interpolation resulted into a test data with 51 features.

Following the interpolation process, our objective was to establish a set of common columns shared between the train and test datasets. Despite the initial distinctions, with the train dataset having 62 features and the test dataset containing 51 features, we successfully identified and retained 48 common columns for both datasets.

Handling categorical variables was another crucial step. Categorical features were converted into a machine-learning-friendly format using Label Encoding. This technique assigned a unique integer to each category within a categorical feature. Label Encoding is a common method for preparing categorical data for machine learning algorithms [2].

In addition to encoding, we performed feature scaling to normalize numerical features. Scaling ensured that numerical features were on a common scale, allowing for fair comparisons and improving machine learning model performance. Min-Max scaling, which rescales data to the range [0, 1], was selected as the scaling method.

2.2 Feature Engineering

We initiated the feature engineering process by generating the 'season' feature for the train dataset, as it was already present in the test dataset. This involved extracting this temporal information from the date column in the train data.

Additionally, we introduced a location-based feature, combining rounded latitude and longitude values. This novel feature not only enhanced data organization but also provided spatial context to the observations, enabling our models to consider proximity.

To explore alternative landscape perspectives, we applied coordinate rotation techniques to latitude and longitude, creating 'rot_45_x,' 'rot_45_y,' 'rot_30_x,' and 'rot_30_y.' These transformations aimed to diversify the data representation, potentially improving model performance.

Recognizing the value of geospatial insights, we leveraged a geocoder in the final phase of feature engineering. Geocoding, a well-established technique in geographic information systems [3], enabled us to enrich our dataset by inferring location-based details from latitude and longitude coordinates. This augmentation broadened our feature set to encompass place names, states, counties, and country codes. These additions equipped our models to account for geographic variations and patterns.

With the inclusion of these supplementary features, the dimension of both the train and test datasets expanded from 48 to 60 features. It is worth noting that not all these features were employed for the final model submission. After extensive experimentation, we carefully selected 40 features for modeling, excluding the 'geo_' and 'loc_' features (12 in total). A graphical representation of the feature importance rankings of our models can be found in Figure 4 within the discussion section.

Lastly, the dataset was partitioned into training and testing sets using the 'train-test-split' function, reserving 20% of the data for the test set. This division played a vital role in evaluating the performance of our machine learning models.

3. MODEL EVALUATION AND SELECTION

In this section, we delve into our model selection process, ultimately opting for the LightGBM classifier as the most suitable choice for our task. Our primary evaluation metric was accuracy, which measures the ratio of correctly classified samples to the total samples, considering both cross-validation and leaderboard accuracy. The accuracy formula is depicted as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Additionally, we considered the model's training time and inference time as essential performance metrics to assess model efficiency. To ensure the effectiveness of our predictive model, we commenced with a comprehensive comparison of various machine learning algorithms, including XGBoost, HistGradientBoosting, LightGBM, and CatBoost. To maintain consistency in performance evaluation, we employed similar procedures and metrics across the four models, including:

- □ Default Parameters of each models used
- □ Train -80% of training data, 20% for validation
- □ 5-fold kfold splitting to help get cross validation scores (CV) to evaluate model
- □ Similar features across the models.

The results of this initial model comparison are summarized in Table 1, which includes accuracy scores for both cross-validation and the private leaderboard.

Table. 1 – Performance of the models with selected features

Model	**Accuracy CV**	**Accuracy Private LB**	**Train Time(s)**	**Inference Time (ms)**
XGBoost	**0.7743**	0.7857	593	3.0
LightGBM	0.7384	**0.8303**	**31**	3.0
HistGradientBoosting	0.7470	0.8035	64	16.1
CatBoost	0.7685	0.8125	385	2.5
Ensemble (4)	x	0.8035	1073	24.6

We further explored the impact of removing seasonal features from our model in Table 2, revealing the models' performance under these circumstances.

Table. 2 – Performance of the models with selected features *without* season feature

Model	Accuracy CV	Accuracy Private LB
XGBoost	**0.6996**	0.3571
LightGBM	0.6641	0.4642
HistGradientBoosting	0.6723	**0.4910**
CatBoost	0.375	**0.4910**
Ensemble (4)	x	0.4642

Furthermore, we conducted an investigation into the effectiveness of interpolation techniques by using only latitude, longitude, and season as features, as indicated in Table 3.

Table. 3 – Performance of the models without interpolation

Model	Accuracy CV	Accuracy Private LB
XGBoost	0.5836	0.9910
LightGBM		
HistGradientBoosting		
CatBoost		
Ensemble (4)		

The study ultimately demonstrated the superiority of LightGBM, leading us to explore the impact of various data on AQI estimation using this model, as detailed in Table 4.

Table. 4 –Performance of the LightGBM across various data

Data	Accuracy CV	Accuracy Public LB	Accuracy Private LB
Lat, Lon, Season (No Interpolation)	0.5836	**1**	**0.9910**
Meteorological Data	0.6371	0.4791	0.4464
Geospatial/topographic	0.4977	0.5	0.49
Land Cover Map	0.5414	0.5	0.49
Geological Map	0.4977	0.5	0.49
Engineered Features	0.5836	**1**	**0.9910**
All Features	0.7376	0.8541	0.8214
40 Selected Features	**0.7384**	**0.875**	**0.8303**

4. DISCUSSIONS

Discrepancies between cross-validation accuracy scores on the train dataset and leaderboard results on the test data were apparent. This variation could be attributed to the limited size of the test data, comprising only around 160 samples, rendering it less representative of the larger train dataset used for model training. The absence of real meteorological station data in the test dataset, replaced by interpolated data, may also have contributed to this disparity. The interpolated data exhibited a smoother pattern compared to actual station data, as visualized in Figure 3:

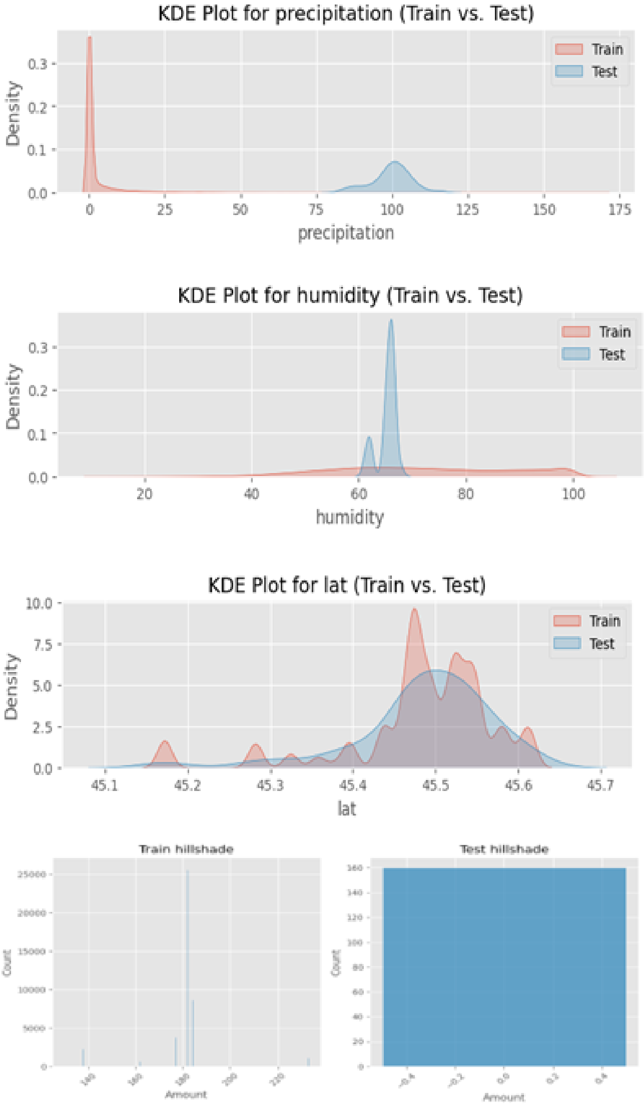

Figure 3 – Sample distributions of features between train and test data

Among the models evaluated, LightGBM outperformed with the highest accuracy of 83% and the shortest training time of 31 seconds (as displayed in Table 1). Ensembling the results of all four models using mode would only yield an accuracy of 80.35%, still falling short of LightGBM's performance. Hence, LightGBM emerged as the preferred model for this specific problem statement.

The absence of the season feature significantly impacted the models' performance, resulting in similar low accuracy scores of approximately 46% to 49% for all models, except XGBoost, which slightly underperformed at 35% (as shown in Table 2). This raises questions regarding the importance of the season feature in the models.

In Table 3, using the season feature alone led to nearly perfect 100% accuracy on the test data. However, this impressive test performance was accompanied by a sharp decline in performance on the train data. These inconsistencies highlight the potential challenges and nuances of utilizing the season feature, warranting further investigation in future work.

When individual data sources were examined separately, they yielded consistent results between train and test data, averaging around 50% accuracy, with the exception of meteorological features. The meteorological features excelled in the train dataset, achieving 65% accuracy, but underperformed in the test, resulting in 47% accuracy. This is shown in Table 4.The table also illustrates the impact of feature selection, showing that optimal feature choice can boost performance to 83%, compared to 82% when using all available features.

The final proposed solution underscored the importance of meteorological features, particularly when combined with other features, as depicted in Figure 4. Notably, LightGBM ranked the season feature as the fifth most important, while in other models, the season feature dominated as the most influential feature, overshadowing the rest.

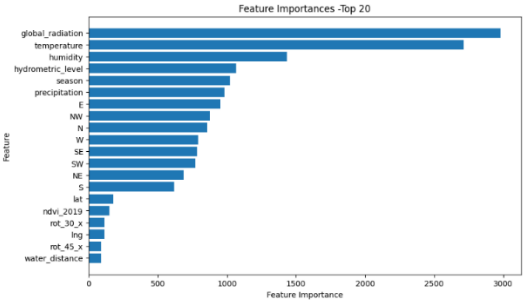

Figure 4 – Summary of Feature Importance in LightGBM for top features

5. FUTURE WORK

Future work could look at the following to improve results:

□ Incorporate real data instead of interpolated data and explore additional open-source resources, such as satellite and model-derived data like ERA5 or aerosol data, for a more robust analysis.

□ Consider predicting PM10, PM2.5, O3, NO2, and SO2 and utilizing them for AQI estimation as per the guidelines provided by the European Environment Agency.

□ With clear variable description provided, new additional more robust features could be engineered enriching the accuracy of the models.

□ Introduce additional hyperparameters for fine-tuning our models.

□ Investigate the application of Artificial Neural Networks (ANNs) and Deep Learning techniques.

6. CONCLUSION

In conclusion, this paper tackles the essential task of predicting 2022 Air Quality Index (AQI) levels in Milan while emphasizing the impact of seasonal variations. Leveraging machine learning techniques and a comprehensive dataset featuring air quality, meteorological information, and geospatial data, we've made significant strides in creating valuable AQI susceptibility maps. These maps provide season-specific air quality insights, accounting for meteorological influences. Notably, our predictive models, particularly the LightGBM classifier, achieve high AQI estimation accuracy despite the challenges of test data size and lack of real station data. The study's outcomes lay a strong foundation for improved air quality management in Milan and encourage comprehensive analyses, including satellite and model-derived data, for deeper insights.

REFERENCES

1. Li, X., Li, D., & Shi, Y. (2017). A Comparative Study of Inverse Distance Weighting and Radial Basis Function Networks for Spatial Interpolation of Precipitation. ISPRS International Journal of Geo-Information, 6(6), 188. DOI: 10.3390/ijgi6060188.
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.
3. "Reverse_geopy"library:https://geopy.readthedocs.io/en/stable/#nominatim.
4. This problem statement work can be found here: https://github.com/ITU-GeoAI-Challenge/GeoAI-Challenge-for-Air-Pollution-Susceptibility-Mapping