A Comparative Study of LightGBM on Air Quality Data Across Multiple Locations

Martina Casari[1,*], Andrea Arigliano[1] and Laura Po[1]

[1]*Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia*

Abstract

In this paper, we present a novel approach utilizing LightGBM algorithms to estimate PM2.5 concentrations in two distinct geographical locations, Turin in Italy and Southampton in the UK. Our methodology integrates data from low-cost sensors co-located with reference stations in both locations, ensuring data reliability. Through a rigorous analysis encompassing diverse splitting techniques, learning pipeline components, and feature selection methods, our approach showcases remarkable performance across various scenarios, promising practical applicability. We initially train and test our model on the Turin dataset, followed by an assessment of its performance within the specific geographical context. Furthermore, we extend our investigation to the Southampton dataset without any adjustments, revealing disparities in performance. Additionally, we conduct comparative training on both datasets, offering insights into contextual factors influencing model efficacy within specific geographical areas. Our findings underscore the importance of contextual considerations for accurate air quality estimation and highlight the potential of our approach for real-world deployment. The datasets used in this study are publicly available, facilitating further research and validation.

Keywords

Particulate Matter, Low-cost sensors, Different Locations, LightGBM, Open dataset

1. Introduction

Airborne particulate matter (PM) refers to tiny particles in the air that can be composed of various materials such as dust, dirt, soot, smoke, and liquid droplets. These particles vary in size and can have different chemical compositions, originating from both natural and human-made sources [1]. Airborne PM consists of a heterogeneous mixture of solid and liquid particles suspended in air that varies continuously in size and chemical composition in space and time. PM is categorized based on the diameter of the particles, measured in micrometres (μm) [2]. The main classifications include PM1, PM2.5, PM4, and PM10, representing different size fractions, each of them causing different problems regarding both the environmental conditions, affecting ecosystems [3, 4], and human health [5] complications which mainly impact the respiratory and cardiovascular systems, also potentially affecting the bloodstream. Airborne PM can have severe environmental consequences. When it settles on the soil, it can have a detrimental impact on the nutrient cycling of plants and disrupt the ecosystem's balance. This can potentially lead to negative consequences on the entire food chain and have long-lasting effects on the environment. When it comes to health concerns, much attention has been given to the amount of PM that enters a person's body, which is referred to as the dose. Studies examining this dose have shown that exposure to high concentrations of PM can lead to damage at the cellular level, particularly in the lungs. Several possible reactions can occur in response to certain environmental and chemical exposures. One hypothesis is that the body may up-regulate its production of antioxidant enzymes to combat the negative effects of these exposures. In some cases, exposure can also result in cell death or an allergic immune response. Additionally, exposure can impair the body's ability to defend the lungs and cause DNA damage. It's important to note that these effects can also have a ripple effect throughout the body, impacting other systems such as the cardiovascular system. Given the detrimental impact that PM concentration in the atmosphere can have, accurately forecasting future PM levels based on current air conditions is a critical undertaking. This effort is essential in preventing the various issues associated with PM exposure and implementing effective measures like traffic and viability restrictions to address them.

The objective of this study is to demonstrate the effectiveness of the LightGBM algorithm in accurately forecasting PM2.5 levels using cost-effective sensors and various environmental parameters. Additionally, the study explores the applicability of the method across different locations, examining both homogeneous and heterogeneous approaches. The training process relies on PM2.5 measurements from reference stations, enabling the resultant model to predict and adjust measurement readings effectively.

The article is structured as follows: Section 2 introduces the dataset; Section 3 outlines the methodology, includ-

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy

*Corresponding author.

martina.casari@unimore.it (M. Casari); laura.po@unimore.it (L. Po)

0000-0003-0406-3036 (M. Casari); 0000-0002-3345-176X (L. Po)

ing the models used and the pipeline implemented; Section 4 presents the results and discussion; and Section 5 provides the conclusions.

2. Datasets

The datasets considered in this study are created by a collection of measurements captured in two different geographical areas, both by using SPS30 low-cost (LC) sensors as input and the co-located legal stations as reference:

- Turin (Italy): LC sensors capturing records with 15-minute frequency, reference station (RS) with hourly frequency based on Arpa weather stations [6];
- Southampton (UK): LC sensors capturing records with 2 minutes frequency, RS sensors with hourly frequency based on Fidas200s weather stations [7].

The data was obtained through individual sensor measurements, which were then used to construct the raw datasets for both Turin and Southampton. Subsequently, a thorough analysis of the LC and RS data was conducted to create a dataset linking each reference record with a low-cost measurement. To achieve this, the input datasets were resampled to match the hourly frequency of the reference datasets.

Initially, the resampling technique employed was averaging all the LC data over the RS hourly record. However, due to significant variations in the data within an hour, it was decided to assign the closest available LC record to each RS record instead. After this process, the raw datasets for both Turin and Southampton were created, and preprocessing techniques [8] were applied to uniformly adjust the data, preparing them for the training step. In the performance evaluation, just the preprocessed dataset was considered for comparison.

Incorporating contextual features based on time into the feature extraction process has allowed for a more thorough understanding of the data. This approach not only captures the original features but also encodes information about the time axis, enabling a fine and accurate representation of patterns that unfold over time. Ultimately, this results in more insightful and precise outcomes.

The final set of features included in the datasets comprises ”pm1,” ”pm2p5,” ”pm2p5 RF target,” ”pm4,” ”pm10,” ”wind speed,” ”pressure,” ”temperature,” ”relative humidity,” ”month,” ”day of the week,” and ”hour.” The correlation matrix is depicted in Figure 1.

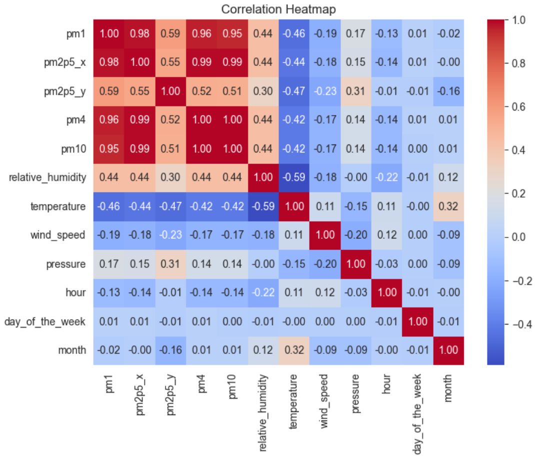


Figure 1: Correlation matrix of all the features.

3. Methodology

The research consisted of a methodical process with distinct stages. Firstly, a brute-force testing procedure was carried out to determine the most appropriate machine-learning model from a variety of options. Subsequently, the pipeline was created by examining the ideal dataset split, feature selection, and transformation techniques required for the specific task. Lastly, a thorough evaluation of performance metrics was conducted using the Turin dataset, including MAE, MSE, MdAE, and R2 metrics.

3.1. Model

The first step was to determine the appropriate model for the problem at hand. To accomplish this, a Bulk Regressor was implemented. This function tests a variety of regression models from popular Python libraries, such as *scikitlearn*, on the target dataset, ultimately producing a ranking of the most successful models based on average prediction accuracy metrics. Interestingly, the top-performing models were nonlinear, indicating that interpreting the features required an examination of nonlinear relationships between them. As a result, LightGBM was chosen as the model for this study.

LightGBM (Light Gradient Boosting Machine) is a powerful and efficient gradient-boosting framework developed by Microsoft researchers in 2017 [9]. It is designed to be efficient and scalable, making it particularly well-suited for large datasets and high-dimensional feature spaces. It utilizes the boosting framework, building an ensemble of weak learners (decision trees) sequentially to minimize the overall prediction error, thus ultimately combining multiple weak models to create a strong predictive model. Unlike depth-first tree growth in traditional gradient boosting frameworks like XGBoost [10], LightGBM adopts a leaf-wise tree growth strategy which chooses the leaf with the maximum delta loss to grow, which can lead to faster convergence and reduced computational cost. The trees are then used as usual, choosing the path that maximizes the information gain which is evaluated via the variance score of each node. Other characteristics are that it includes a feature selection process by itself and the loss used usually is the Mean Squared Error (MSE) Loss, Eq. 1.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

3.2. Split Techniques

Different split configurations were tested in order to obtain the optimal one for this case study, starting from a simple random split and going towards more complex splits based on the time period considered. The different splits considered are:

- Random Total Split (RTS): Random split among all the records in the domain of the whole dataset;
- Random Day Split (RDS): Random split obtained by grouping all the records by day, then randomly splitting in the subdomain of the single day;
- Random Month Split (RMS): Random split obtained by grouping all the records by month, then randomly splitting in the subdomain of the single month;
- Forecast Day Split (FDS): Forecast split obtained by grouping all the records by day, then assigning the first 75% to the train and the last 25% to the test in the subdomain of the single day;
- Forecast Month Split (FMS): Forecast split obtained by grouping all the records by month, then assigning the first 75% to the train and the last 25% to the test in the subdomain of the single month;

Every split considered kept a 75-25 ratio between the training and test set, simply varying the domain considered and whether the records were picked randomly or sequentially. Each of the aforementioned split techniques was tested over the preprocessed Turin dataset to choose the best-performing split for the next steps.

As it is possible to infer from results in Table 1, the RTS seems to achieve the best results all across the board, but since we are working with time series the best choice would be to not consider this split as it tends to overestimate the results due to the data nature. Therefore, the split technique considered in the next steps of this research is the RDS.

Table 1
Dataset split with performance metrics over the preprocessed Turin dataset

	MAE	RMSE	MdAE	R2
RTS	5.19	7.24	3.96	0.73
RDS	5.16	7.38	3.90	0.73
RMS	5.21	7.25	3.97	0.73
FDS	6.86	8.87	5.82	0.62
FMS	5.91	8.58	4.18	0.58

Table 2
Correlation of features with target variable

Feature	Absolute Correlation
pm1	0.588402
pm2p5	0.546849
pm4	0.521004
pm10	0.511054
temperature	-0.473645
pressure	0.306547
relative humidity	0.297370
wind speed	-0.227094
month	-0.159161
day of the week	-0.014929
hour	-0.006393

3.3. Pipeline

After selecting the model and dataset split method, the subsequent task involves determining the required data preparation techniques for this problem. The primary components of the data processing pipeline include feature selection and skewness transformation. It is unnecessary to scale the data given the characteristics of LightGBM.

3.3.1. Feature Selection

In this phase, the most representative features of the problem were extracted. Since there is a relatively low number of features, to begin with, the selection was done using a simple correlation method where the resulting features are the ones which correlate with the target variable higher than a chosen threshold.

As evident from Table 2, both ”day of the week” and ”month” exhibit weak correlations with the target variable.

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \tag{2}$$

The Pearson Correlation Coefficient (r) was utilized to assess these correlations, as indicated by Equation 2.

Consequently, even if a negative correlation with the target variable is obtained using this formula, it remains valuable as it signifies an inverse correlation, akin to inverse proportionality. Ultimately, the features selected by this method are those for which $|r| > 0.1$.

3.3.2. Skewness Transformation

Skewness is a statistical measure that describes the asymmetry of the probability distribution of a real-valued random variable. In simpler terms, it measures the degree and direction of skew (departure from horizontal symmetry) in a dataset. A skewness value of 0 indicates a perfectly symmetrical distribution, see Eq. 3. Positive skewness indicates a longer right tail, while negative skewness indicates a longer left tail.

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{3}$$

When dealing with regression problems, addressing highly skewed variables is crucial as they can impact the model's fit. This is primarily due to the assumption of linearity made by most regression algorithms, which presupposes linear relationships between features. By applying transformations such as power or logarithmic functions, this effect can be mitigated, especially considering that the chosen model inherently possesses nonlinear properties.

Additionally, highly skewed predictor variables can make the model overly sensitive to extremely high values, potentially resulting in a poor fit for the majority of the data.

To tackle this issue, a skewness transformation was incorporated into the pipeline. This transformation applies a predefined set of transformations to each feature in order to reduce its skewness. The set of transformations includes:

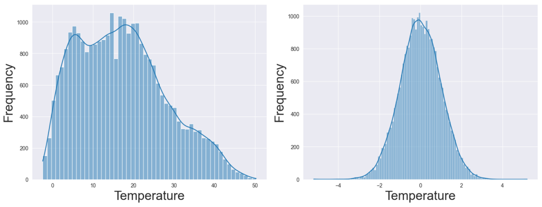
- Logarithm: $f_t = \log(f)$;
- Exponential: $f_t = e^f$;
- Square Root: $f_t = \sqrt{f}$;
- Quantile: $f_t = F^{-1}(f)$;

For each feature in the dataset, all transformations are tested, and the one selected is the transformation that minimizes the feature's skewness to 0.

An example of feature skewness transformation is depicted in Figure 2, illustrating the distribution of temperature data. The second figure demonstrates the attainment of a Gaussian distribution after applying the Quantile Transformer. Table 3 shows the best transformation found for each feature.

Table 3
Best transformation for each feature

Feature	Best Transformation
pm1	Log Transformation
pm2p5_x	Log Transformation
pm2p5_y	QuantileTransformer
pm4	Log Transformation
pm10	Log Transformation
relative_humidity	QuantileTransformer
temperature	QuantileTransformer
wind_speed	QuantileTransformer
pressure	QuantileTransformer
month	QuantileTransformer



(a) Before

(b) After

Figure 2: Distribution comparison with skewness transformer

4. Results and Discussion

By applying all the aforementioned techniques, the final pipeline is created and then trained on the preprocessed Turin dataset with the RDS split method.

Table 4
Performance metrics obtained from training the LightGBM model on the Turin preprocessed dataset.

Metric	Turin Train	Turin Test
MAE	0.3023	0.3369
RMSE	0.1467	0.1846
MDAE	0.2508	0.2775
R^2 Score	0.7435	0.6735

As evident from Table 4, the selected pipeline demonstrates strong performance on both the Turin training and test sets.

In Figure 3, the feature importance ranking for the constructed model is depicted. Observing the significance of meteorological features for the model's predictions is notable.

The results presented in Table 5 highlight the performance achieved when applying the model to a distinct dataset, the Southampton dataset. Here, it is evident that the model's prediction of outcomes is unsatisfactory. This suggests that while the model

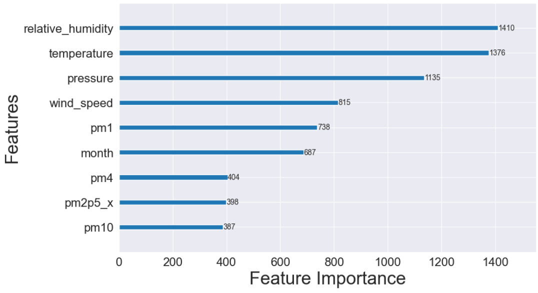


Figure 3: Feature importance ranking

Table 5
Southampton (UK) performance metrics obtained from training the LightGBM model on the Turin preprocessed dataset.

Metric	UK Dataset
MAE	6.4039
RMSE	82.9644
MDAE	4.5478
R^2 Score	-0.9130

reliably predicts where results should fall within their value range, it struggles to accurately forecast how they are distributed over time. Consequently, it can be inferred that the geographic location under study exerts a significant influence on PM forecasting.

To tailor forecasting models to specific geographic zones, it is essential to incorporate the studied area as a feature or consider creating independent models for each area under consideration. The challenge faced by the model in this scenario may stem from several factors, including the distinct nature of the datasets, their unique contextual considerations, and the temporal misalignment despite both datasets covering an entire year. Furthermore, the placement of the SPS30 sensors within different devices for Southampton and Turin introduces significant variability in the collected data due to positional and rotational differences.

To delve deeper into this issue, an additional test was performed by merging records from both the Southampton and Turin datasets. This merged dataset served as the comprehensive training and testing dataset with the RDS split and was subsequently processed through the aforementioned pipeline. The objective of this test was to develop a model capable of addressing both challenges simultaneously, by incorporating data from both geographical areas concurrently.

As we can see from the results in Table 6, this test provided surprisingly good results all across the board, with great values both in the distance metrics and in R2.

Table 6
Performance Metrics

Metric	MAE	RMSE	MdAE	R2
Merged Dataset	3.52	5.78	2.08	0.78

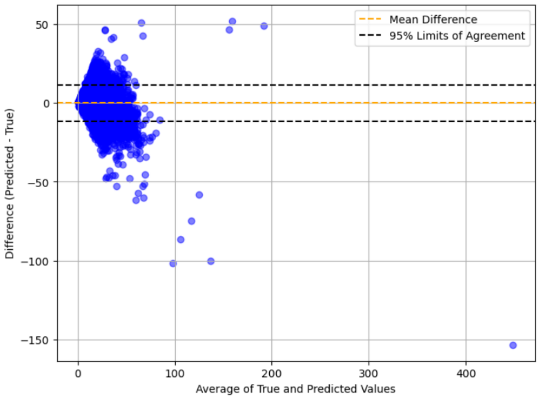


Figure 4: Bland–Altman plot for the merged dataset

However, upon analyzing the Bland-Altman plot in Figure 4, it becomes apparent that there exist relatively high absolute differences between the predicted and actual values, particularly within the first range of values where the majority of records are concentrated. This discrepancy implies that while the predictions generally fall within the desired range considering the wide scope of values (over 87k records), the model's precision in predicting exact values is suboptimal.

One possible explanation for this phenomenon is the variability of PM values across different geographical areas attributable to diverse environmental conditions. Without incorporating a feature that delineates between the two areas, the model treats the PM range as a unified domain for both datasets, endeavouring to predict within that domain without differentiation due to the absence of pertinent information. These findings underscore the original hypothesis, emphasizing the necessity to either incorporate features that encapsulate environmental conditions or devise distinct models for different areas, as the available features alone are insufficient to infer such information.

To conclude this discussion and affirm the thesis, a final test was conducted by creating a new independent model using only the Southampton data.

The latest results presented in Table 7 serve to reinforce the thesis that tailoring a model to a specific geographical area yields superior outcomes in accurately capturing and predicting PM levels using machine learning

Table 7
Performance metrics for Southampton model

Metric	MAE	RMSE	MdAE	R2
Southampton Model	1.73	3.04	1.01	0.88

techniques. The model trained exclusively on Southampton data demonstrates excellent performance across all metrics utilized, consolidating the argument for geographic specialization in PM forecasting models.

5. Conclusion

In conclusion, this paper presents a comprehensive study on the development of the LightGBM model for predicting PM levels, highlighting the crucial role of geographical considerations in the process. The study evaluates various dataset split techniques and identifies the RDS method as the most effective. The learning pipeline encompasses feature selection and skewness transformation. Remarkably, this pipeline achieves state-of-the-art results on both the Turin and Southampton datasets independently.
Furthermore, a comparative analysis is conducted on different combinations of data, as well as a merged dataset test incorporating data from both regions simultaneously. However, the findings suggest that creating independent models for distinct geographical areas yields the best performance for this case study, underscoring the significance of environmental conditions surrounding the utilized sensor.
This research endeavours to shed light on laying the groundwork for constructing models capable of generalizing, taking into account localized environmental factors in the predictive modelling of PM levels.

References

[1] K. R. Daellenbach, G. Uzu, J. Jiang, L.-E. Cassagnes, Z. Leni, A. Vlachou, G. Stefenelli, F. Canonaco, S. Weber, A. Segers, J. J. P. Kuenen, M. Schaap, O. Favez, A. Albinet, S. Aksoyoglu, J. Dommen, U. Baltensperger, M. Geiser, I. El Haddad, J.-L. Jaffrezo, A. S. H. Prévôt, Sources of particulate-matter air pollution and its oxidative potential in europe, Nature 587 (2020) 414 – 419. doi:10.1038/s41586-020-2902-8, all Open Access, Green Open Access.

[2] A. Mukherjee, M. Agrawal, World air particulate matter: sources, distribution and health effects, Environmental chemistry letters 15 (2017) 283–309. doi:10.1007/s10311-017-0611-9.

[3] X. Yue, Y. Hu, C. Tian, R. Xu, W. Yu, Y. Guo, Increasing impacts of fire air pollution on public and ecosystem health, The Innovation 5 (2024) 100609.

[4] D. Grantz, J. Garner, D. Johnson, Ecological effects of particulate matter, Environment International 29 (2003) 213–239. doi:https://doi.org/10.1016/S0160-4120(02)00181-2, future Directions in Air Quality Research : Ecological,Atmospheric,Regulatory/Policy/Economic, and Educational Issues.

[5] M. J. Mohammadi, B. F. Dehaghi, S. Mansourimoghadam, A. Sharhani, P. Amini, S. Ghanbari, Cardiovascular disease, mortality and exposure to particulate matter (pm): a systematic review and meta-analysis, Reviews on Environmental Health 39 (2024) 141–149. URL: https://doi.org/10.1515/reveh-2022-0090. doi:doi:10.1515/reveh-2022-0090.

[6] M. Casari, L. Po, L. Zini, Low-cost pm data, 2023. URL: https://doi.org/10.5281/zenodo.10037781. doi:10.5281/zenodo.10037781, https://-doi.org/10.5281/zenodo.10037781.

[7] F. M. J. Bulot, Characterisation and calibration of low-cost pm sensors at high temporal resolution to reference grade performances - dataset, 2022. URL: https://doi.org/10.5281/zenodo.7198378. doi:10.5281/zenodo.7198378, https://-doi.org/10.5281/zenodo.7198378.

[8] M. Casari, L. Po, Mith: A framework for mitigating hygroscopicity in low-cost pm sensors, Environmental Modelling & Software 173 (2024) 105955. doi:https://doi.org/10.1016/j.envsoft.2024.105955.

[9] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.

[10] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 785–794. doi:10.1145/2939672.2939785.

A. Online Resources

The Turin dataset used in this study is freely available through the Zenodo platform [6].