Prediction of Nitrogen Dioxide Levels in London

Abstract - The main goal and contribution of this paper is to analyse Nitrogen Dioxide levels in relation to other air pollutant and gases, by using two machine models that are Decision tree and Linear Regression. These models were implemented with scikit-learn, a free tool used for predictive data analysis. [1]

The dataset used is from Kaggle [2], an open-source machine learning platform with over 50,000 public datasets. The dataset chosen is a .csv file named 'monthly-averages.csv' with one feature of type object and fourteen numerical features of type float64. The dataset contained over 100 data with some undefined values (Nan) that were substituted with the average values.
Data processing and normalisation, data analysis, and feature reduction, are described in this study.
Decision Tree Regression and Linear regression are suitable methods to predict Nitrogen Dioxide levels in London.

Keywords - Decision Tree Regression, Linear regression, Nitrogen Dioxide, Air quality.

INTRODUCTION

Is it possible to predict the levels of Nitrogen Dioxide in London using machine learning models such as Decision Tree Regression and Linear regression? This study aims to investigate and answer this question.

Air pollution is a significant global problem, especially in many urban cities such as Delhi, New York and London included. Air pollution often involves pollutants like Nitrogen Dioxide (NO_2), Particulate Matter (PM_{10} and $PM_{2.5}$.), Sulphur Dioxide (SO_2), Carbon Monoxide (CO), and Ozone(O_3) [3]. These pollutants are generated from sources such as industrial waste, vehicles, and natural events.
Why the need of investigating, monitoring and predicting the levels of these air pollutant? Because according to the World Health Organization – WHO, air pollution is responsible for 6.7 million premature deaths every year [3]; meaning that the effect that these pollutants have on human beings is fatal.

In this study, we will see how Nitrogen Dioxide may be assessed, predicted and evaluated using machine learning and data analytics methodologies. From problem structure to data analysis, data normalisation and transformation, and model selection and implementation, we can predict NO_2 concentrations without having to take recordings specifically targeted for this air pollutant, which can be expensive and ineffective due to

environmental factors and sensors being used.

OBJECTIVES

- Determine which features/coefficients are correlated with the concentration of NO_2 in the air.
- Use two regression models by implementing supervised machine learning algorithms (Decision Tree Regression and Linear regression) to predict NO_2 concentrations as the outcome.
- Implement exploratory data analysis to validate the correlation between certain features in the dataset.
- Evaluate and compare the accuracy of the regression models.

LITERATURE REVIEW

Air quality is an interesting and important topic that has been discussed and investigated in several research papers. In this section, the problem will be discussed and reviewed through four related studies.

The first source of our literature review is the report "Deep Learning Based Air Quality Prediction: A Case Study for London" [4] by Anil UTKU and Umit CAN from Munzur University, in Turkey.

Their study used measurement data from Eltham measurement station in London, to predict the short-term $PM_{2.5}$ values. The study compared a total of twelve successful machine learning and deep learning methods, resulting in Long short-term memory (LSTM), a deep learning model, being the most suitable model for their dataset (with a accuracy of 98.8%) thus outperforming the other models compared [4]. The paper, using both machine learning and deep learning models to address and predict air quality, highlights that air pollution values can be successfully obtained through artificial intelligence methods, thus taking measures in reducing pollution.

The second source is a research paper titled "Estimation of daily NO_2 with explainable machine learning model in China, 2007-2020" [5] by Yanchuan Shao and colleagues focuses on estimating daily ground-level nitrogen dioxide (NO_2) concentrations across China over a 14-year period using an explainable machine learning approach. This research offers valuable insights into NO_2 pollution patterns in China and showcases the potential of machine learning models in environmental monitoring and management. The model used demonstrated strong predictive capabilities, with R^2 (cross validation checks) of 0.75. They concluded that an explainable machine learning model can provide valuable guidance and insights into air pollution control.

The third source is "Forecasting the concentration of NO2 using statistical and machine learning methods: A case study in the UAE"

[6] by Aishah Al Yammahi, Zeyar Aung. The study used statistical and machine learning models ARIMA, SARIMA, NAR, and LSTM using the MAPE (Mean Absolute Percentage Error) evaluation criteria, resulting from excellent (MAPE of 8.64% at the Liwa station using the closed-loop architecture) to acceptable (MAPE of 42.45% at the Khadejah School station using the open-loop architecture).
They were able to demonstrate that the MAPE value is correlated with the relative standard deviation of Nitrogen Dioxide concentration values.

The final literature value is "Predicting air quality index based on meteorological data: a comparison of Regression analysis, artificial neural networks and decision tree" [7] by Akram Jamal and Ramin Nababizadeh Nodehi. In this article, various machine learning and artificial neural approaches were used to forecast next day air quality index and determine the correlation coefficient of 0.66 [6]. It was concluded that the application of the forecasting methods could be adopted for air quality management and protect public health.

DATA PROCESSING

Dataset Overview

The data is a csv file called 'monthly-averages.csv', downloaded from Kaggle, in the dataset London Air Quality [2]. This contains air quality measurements for nitrogen dioxide (NO2) in London, recorded across various sites (roadside and background average readings) and times.
The dataset used 19.5+ KB memory and presented 132 entries (with some Non-a-Number-Nan data); from 01/01/2008 to 01/12/2018; and the total columns were originally 15 (14 for the air pollutants measured in ug/m3 and one for the Month).

Data pre-processing

Month	London Mean Roadside Nitric Oxide (ug/m3)	London Mean Roadside Nitrogen Dioxide (ug/m3)	London Mean Roadside Oxides of Nitrogen (ug/m3)	London Mean Roadside Ozone (ug/m3)	London Mean Roadside PM10 Particulate (ug/m3)	London Mean Roadside PM2.5 Particulate (ug/m3)	London Mean Roadside Sulphur Dioxide (ug/m3)	London Mean Background Nitric Oxide (ug/m3)	London Mean Background Nitrogen Dioxide (ug/m3)	London Mean Background Oxides of Nitrogen (ug/m3)	London Mean Background Ozone (ug/m3)	London Mean Background PM10 Particulate (ug/m3)	London Mean Background PM2.5 Particulate (ug/m3)	London Mean Background Sulphur Dioxide (ug/m3)
2008-01-01	NaN	55.502688	NaN	29.512097	24.969086	14.678763	4.217742	NaN	42.338710	NaN	36.942204	18.817204	NaN	3.572581
2008-02-01	NaN	75.922414	NaN	20.317529	39.477011	28.772989	7.553161	NaN	60.237069	NaN	26.425287	31.896552	NaN	6.734195
2008-03-01	NaN	55.610215	NaN	40.103495	21.569892	12.300135	3.868280	NaN	39.801075	NaN	50.227151	15.477151	NaN	2.286290
2008-04-01	NaN	61.756944	NaN	37.884722	28.740278	20.461111	4.475000	NaN	44.009722	NaN	50.133333	21.729167	NaN	3.236111
2008-05-01	NaN	62.903226	NaN	46.266129	34.611559	27.508065	4.634409	NaN	44.141129	NaN	60.512097	29.545699	16.576826	4.250000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2018-08-01	33.125000	38.950403	89.741935	22.498118	16.284946	7.897849	1.772446	5.893952	20.524462	27.563172	37.307527	11.926882	6.394624	1.947446
2018-09-01	39.063333	43.906389	103.810556	26.625556	18.896111	9.738333	0.343333	10.228472	25.935833	39.556111	40.955139	14.258750	7.894861	1.157083
2018-10-01	48.643414	46.765457	121.360887	19.501075	19.339247	11.509409	-0.794347	13.622715	29.816129	48.680780	30.284543	14.793817	8.969220	2.843952
2018-11-01	52.535833	47.584028	128.134306	17.159583	23.600000	15.207639	5.885035	18.285833	33.872639	57.820556	24.262361	19.425278	13.525694	2.393333

Figure 1Sample data from original dataset London Air Quality

- **Replacing Missing data**

The missing data found in the dataset could have been caused by environmental factors or by the monitoring sensors. It is common, when dealing with real-life values to have such issues.

There were a total of 100 missing values from the entire dataset. This count was achieved by the $isnull().sum()$ function in pandas, a Python library for data manipulation.

I replaced the missing value, for each column with missing data, with their

respective average column readings. This is done by utilising the function in panda $fillna$ and $mean$.

- **Duplicates removal**

My dataset had no duplicated rows. To check this I used the pandas' function $duplicated()$.

- **Feature selection**

To process my dataset, I had to reduce the number of features, without losing any valuable information.

Each of the 7 air pollutant readings were given with both background and roadside readings, hence having 14 reading types. To minimise this, I performed the average of the column base on their pollutant and created a new column for these new data, thus dropping the original columns with the function $drop$. I proceeded to delete the first column (the Month) as the date of the recordings weren't any valuable information to our dataset.

The obtained dataset was now a 132 by 7 dataset of type float64.

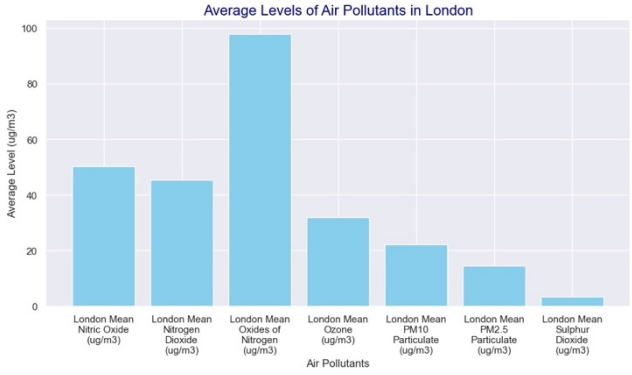


Figure 2 Average levels of air pollutant in London from the pre-processed dataset

For visualisation purposes, see its histogram in Figure 2.

- **Normality testing**

To check if my dataset and the features are normally distributed, I used a function in Seaborn called $displot$. This function provided the density distribution of my air pollutant. See below Figure 3 for the Density distribution of London Nitrogen Dioxide (ug/m3).

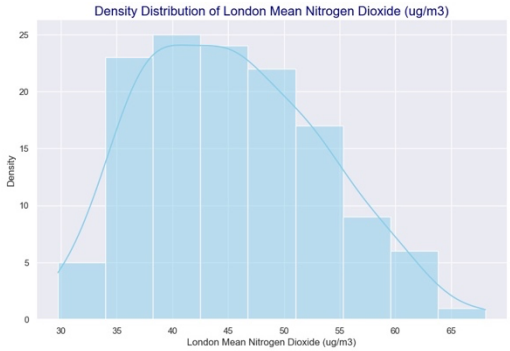


Figure 3 Density distribution of London Nitrogen Dioxide

The density distribution histogram with the bell-shaped curve, suggest that the data has a normal distribution with positive skewness [8].

The bell-shaped implies that the dataset has normal distribution; the positive skewness is suggesting high readings from adverse weather conditions. These are useful for processing our data, therefore they won't be further processed.

See below the Box-and-Whisker plot of NO_2 (Figure 4), achieved with the function $boxplot$. As you can see, there aren't any circle labels

from the boxplot meaning there are no outliers; this means there aren't values that differ from the mean over 3times more than the standard deviation for the data of London Mean Nitrogen Dioxide. The middle line (median) in the box is also quite cantered. This, plus the density distribution histogram with the bell-shaped curve, suggest that the data has a normal distribution.

To double check that our dependent variable which is going to be the NO_2 values are normally distributed, I plotted its Normal Q-Q plot. See the figure 4 below.

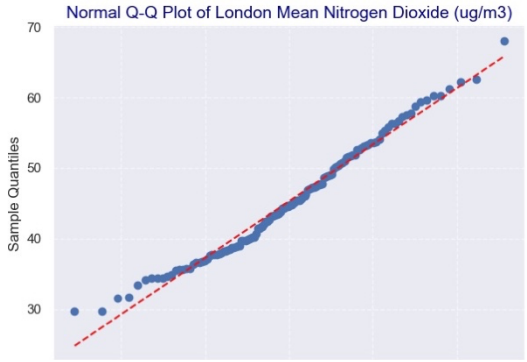


Figure 4 Normal Q-Q plot of London Nitrogen Dioxide

As seen, the points closely following the red dashed line indicate that the pollutant data is normally distributed; the S-shaped curve indicate the positive skewness present (as seen in the histogram); the few data points at the ends indicate the high and low outlier values caused by adverse weather conditions.

See below the density distribution for all air pollutant: they all have normal distribution with positive skewness.

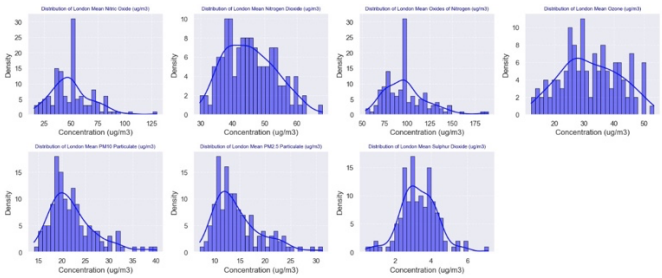


Figure 5 Density distribution for all air pollutants

METHODOLOGY

I used 80% of the dataset as training data and 20% as testing data.

- DECISION TREE REGRESSOR

I performed Decision tree regressor with sklearn. I used the in-built best parameter with best hyperparameters function to achieve a test accuracy of 80.5% and training accuracy of 89%.

See below comparison plot.

Figure 6 Decision Tree Regressor performance

We have obtained an array of accuracy of the model performance for each fold. See below:

```
Cross-validation R² scores: [ 0.394  0.722  0.625  0.348  0.646  0.011]
Mean R² score from cross-validation: 0.45777886552722274
Standard Deviation of R² scores: 0.2411388119740795
```

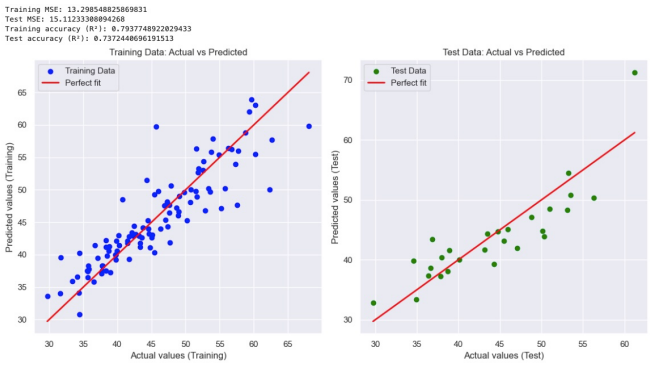
The standard deviation of 0.24113 suggests that the model's performance fluctuates between different subsets of data. The very low R^2 score from the sixth fold is poor because it suggests that the model is doing bad in that fold. But for this dataset, the model generally works well in the training dataset with accuracy of 89% and as we increae the size of the trainig dataset, it still performs well. The best folds are the second (score=0.722) and the fifth (score=0.646) as they have the highest score.

Decision tree might not the be the most suitable model for this type of dataset due to its complexity . Let's try Linear regression.

LINEAR REGRESSION

I used sklearn's Linear regression to build a model for my dataset. I used the default settings to find the best fit and I compared the scores for each feature.

See below images from the findings.



The training accuracy is 79% and the cross-validation accuracy is 24%. The model is good and I can now test and analyse my data.

Again, Feature 2 ("London Mean Oxides of Nitrogen (ug/m3)") and Feature 5 ("London Mean PM2.5 Particulate (ug/m3)") have the highest scores, 0.575 & 0.679 respectively; this suggest that they give higher prediction for nitrogen dioxide concentration.

Finally, with the computation of the Pearson correlation matrix we can confirm that the relationship between London Mean Nitrogen Dioxide (ug/m3) and the remaining air pollutant is as follows: "London Mean Oxides of Nitrogen (ug/m3)" and "London Mean PM2.5 Particulate (ug/m3)" have a higher correlation (0.76 and 0.75 respectively) as opposed to the remaining four pollutant.

Conclusion

After preprocessing the dataset , computing the decision tree methodology and linear regression, the air pollutant London Mean Nitrogen Dioxide (ug/m3) has some correlation of approximately 50-70% in relation to the London Mean

Oxides of Nitrogen (ug/m3) and London Mean PM2.5 Particulate (ug/m3). Due to this relationship , it is possible to predict Nitrogen Dioxide levels.

Works Cited

[1] [Online]. Available: https://scikit-learn.org/stable/.

[2] "kaggle - London Air Quality," [Online]. Available: https://www.kaggle.com/datasets/zsn6034/london-air-quality.

[3] "Type of pollutants - WHO," [Online]. Available: https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/health-impacts/types-of-pollutants.

[4] U. C. Anıl UTKU1, "Deep Learning Based Air Quality Prediction: A Case Study for London," [Online]. Available: https://dergipark.org.tr/en/download/article-file/2759608. [Accessed 17 03 2025].

[5] W. Z. a. ,. R. L. a. ,. J. Y. a. ,. M. L. a. ,. W. F. a. ,. L. H. a. ,. M. A. b. ,. J. B. a. c. ,. Z. M. Yanchuan Shao a b, "Estimation of daily NO2 with explainable machine learning model in China, 2007–2020," [Online]. Available: https://pdf.sciencedirectassets.com/271798/1-s2.0-S1352231023X00181/1-s2.0-S135223102300537X/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEOT%2F%2F%2F%2F%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJHMEUCIQCU77U0ZixRMYjyUgNppghUo0sJ5Bz4x3WC9rrOWMaAXwIgHQUcoWF6q75f. [Accessed 16 03 2025].

[6] Z. A. Aishah Al Yammahi, "Forecasting the concentration of NO2 using statistical and machine learning methods: A case study in the UAE," [Online]. Available: https://www.cell.com/action/showPdf?pii=S2405-8440%2822%2903872-5. [Accessed 16 03 2025].

[7] R. N. N. Akram Jamal, "PREDICTING AIR QUALITY INDEX BASED ON METEOROLOGICAL DATA: A COMPARISON OF REGRESSION ANALYSIS, ARTIFICIAL NEURAL NETWORKS AND DECISION TREE," [Online]. Available: https://japh.tums.ac.ir/index.php/japh/article/view/90. [Accessed 17 03 25].

[8] O. M. o. Dispersion. [Online]. Available: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/descriptive-statistics/other-measures-of-dispersion.html#:~:text=Negative%20skew%20would%20be%20indicated,hand%20side%20of%20the%20distribution.. [Accessed 17 03 2025].