Data Analysis of Air Quality in Australia

Group 16 - GreenTech

Ran Jiang
ranj3

Zhanna Sakayeva
zhannas3

Mengjia Zeng
mengjia6

***Abstract*—Because of rapid urbanization, many developing countries have suffered from serious air pollution problems [6]. Governments and citizens continued to express increasing concerns on air quality, because it may cause long-term health problems and short-term health effects [3]. With these concerns in mind, air quality evaluation and prediction has become an important research area. Some previous studies have reported the effectiveness of some complex models, like different deep learning architecture. In this study, we used a less-used dataset collected in Australia, a developed country, to see whether the influence of pollutants on air quality would be different than that in the developed country. Furthermore, one limitation for previous studies was not including some trend analysis of chemicals and our project also included exploratory analysis of trend analysis of changes of chemicals, as well as PM2.5 to see whether air quality may be different among months and check further about its annual trend. Last, our project tried some easier models, including the regression models and classification models, instead of deep learning architectures in order to balance between the accuracy and processing cost among the process of predicting PM2.5 in the future.**

***Index Terms*—air quality, regression, exploratory analysis, classification**

I. Introduction

Today, air quality has become an overwhelming health-related problem in the world. Multiple types of research have been conducted to help prevent and predict air conditions worldwide. A clean atmosphere is one of the necessary conditions for human survival and plays an important role in the survival of human beings and living things. At the same time, exposure to bad-quality air is a major cause of mortality and other diseases like respiratory infections, allergic diseases, asthma, etc. Therefore, we are interested in doing research and analysis in favor of the improvement of air quality and the control of the emission of harmful gases.

In our project, our target is to build a model which ensures both interpretability and accuracy for predicting the AQI of PM2.5 using the chemicals in the present hour. This project will present a civil air quality analysis of Australia based on the hourly air quality data from a performance monitoring station. We will mainly focus on the variation of AQI (Air Quality Index) along with the time changes, and the prediction of impending air quality, also will display the findings in a clear and concise manner.

To achieve this target, this model should not be too complex and it should still have enough capability for predicting air conditions. We have conducted correlation analysis between all the air pollutants' Air Quality Indexes (AQI) including AQI_CO, AQI_NO2, AQI_O3_1hr, AQI_O3_4hr, AQI_PM10 and AQI_PM2.5 for better understanding the relationship between different chemicals. Additionally, visualization techniques like heatmap have been applied to the AQIs for us to better interpret their correlations. After careful data pre-processing and analysis, we start building two models which predict the AQI_PM2.5 in the next hour with OLS regression and evaluated these two models' performance with RMSE. In the regression model, we have achieved similar results from both of the OLS regression models. After that, we fit our data into classification models which are decision trees and random forests. By comparing the predicting accuracy, we concluded the random forest model predicts better among different levels of PM2.5.

II. Related Work

A. Published Work that relates to Our Project

As mentioned above, air quality evaluation and prediction has become an important research area and encouraged a lot of studies.

One study has reviewed various machine-learning techniques, like Artificial Neural Network, Random Forest Model and Decision Tree Model, used in the field of prediction of air quality. Those deep-learning architectures were constructed based on different kinds of datasets, which are mostly located in China, a developing country, with different accuracy varying from 0.55 to 0.89 [3]. A similar study also tried to develop a deep-learning technique, which was named as the stacked autoencoder model that incorporates autoencoders as building blocks to construct a deep network. By the experiments conducted at different stations in China, the air quality, as indicated by the concentration of PM2.5, were separated into different ranks and the overall prediction for ranks could be as round 0.82 [5]. Another similar study proposed a deep neural network (DNN)-based approach that consisted of a spatial transformation component and a deep distributed fusion network to predict air quality. Within this study, the accuracy could reach 0.81 in predicting the air quality in the next hour [6]. Moreover, Shwet (2022) proposed a model using Linear Regression based Recursive Feature Elimination with Random Forest Regression (RFERF) to predict the Air Quality Index (AQI) and Air Pollutant Concentration (NOx) levels. And for the model comparison, seven well-established machine learning models have been taken, which are aimed to compare with the proposed model to find the balance between accuracy and suitability. By using MAE, MSE, RMSE, and R2 scores to compare prediction models' performance, the proposed hybrid RFERF model is the best-suited and highly accurate. The Linear Regression and the Pearson correlation had been used by Wenrong for analyse the air quality in Shanghai, as well as considering the multicollinearity between the six factors, in order to reduce the influence of multicollinearity on the model [2]. Predicting air quality accurately and efficiently is an important problems of today's world. However, air quality prediction is very challenging because it is affected by many external factors, such as road networks and any other factors in time and space. This is also a major challenge. For instance, a polluted air may become clean in a few hours after a heavy rain, that shows the variability of air pollution [7]. Hence, PlumeCityNet, the forecasting engine that is based on a U-Net architecture was presented by Alléon et al. (2021). It was able to produce 24hour air quality forecasts with different spatial resolutions, from 50 meters to dozens of kilometers [1]. The engine covers the main atmospheric pollutants harming people's health nitrogen dioxide (NO2), ozone (O3) and particulate matter (PM2.5, PM10).

B. *How Our Approach will be Similar or Different from Others*

Based on the previous research, most of the study focused on developing a prediction model to predict PM2.5, and not paying super attention in the trend analysis part. There may be a specific trend of chemicals, as well as PM2.5 annually and air quality may even be different between weekdays and weekends, which was what this study would cover. Also, the generally used datasets are from stations based in China, a developing country, while our study will try a different dataset that is rarely used, as collected in Australia, a developed country, to check the trends and develop possible prediction models. Even though we will try to use AQI of PM2.5 as the response variable, as most of the studies did, we will try to fit linear regression models and also try to divide different concentration levels of PM2.5 into different ranks in order to develop some further classification models.

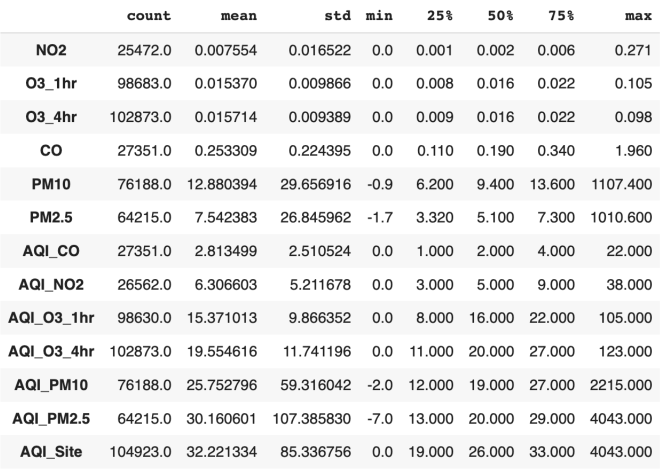
III. Data

In our project, we will be using the air quality data from a performance monitoring station. It was obtained from ACT Government Open Data Portal Australia Government. Our dataset Civic Air Quality Station mainly focuses on the air pollution data from Civic station which is based on Air Quality Monitoring Data Air Quality Monitoring Data — Open Data Portal. The website provides publicly available datasets that contain independent scientific measurements of various pollutants obtained from several active monitoring sites of Australia. The portal provides accessible and shareable data from the Australian Capital Territory and allows us to download datasets in csv or json forms Civic Air Quality Station — Open Data Portal. So, our dataset was available to download as a CSV file and was accessible through URL as well.The time period chosen was from January 1, 2011 and it is updated every hour. At this moment, the dataset contains 104925 records and 18 features. Examples could be seen in Fig.1.

Further we did some analysis on our dataset and performed data wrangling, which was followed by a data modelling stage. As we discussed, before constructing the model and drawing any insights from it we divided our dataset into train and test sets. We implemented our model and trained it with a training dataset and validated it on a testing dataset. The dataset was splitted in such a way that 0.75 was training data and the rest 0.25 was testing data. It was done so that the model could be first trained and then could be tested on the test set such that the error in prediction could be checked and proper results could be used. The variable to be predicted was AQI_PM2.5.

IV. Exploratory Data Analysis

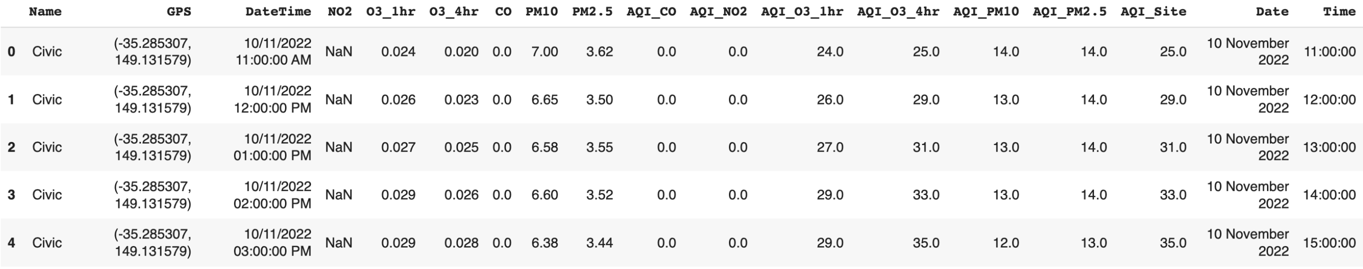
As a first step to get a good understanding of some data preprocessing steps, we provided the numerical summary of the data below.



	count	mean	std	min	25%	50%	75%	max
NO2	25472.0	0.007554	0.016522	0.0	0.001	0.002	0.006	0.271
O3_1hr	98683.0	0.015370	0.009866	0.0	0.008	0.016	0.022	0.105
O3_4hr	102873.0	0.015714	0.009389	0.0	0.009	0.016	0.022	0.098
CO	27351.0	0.253309	0.224395	0.0	0.110	0.190	0.340	1.960
PM10	76188.0	12.880394	29.656916	-0.9	6.200	9.400	13.600	1107.400
PM2.5	64215.0	7.542383	26.845962	-1.7	3.320	5.100	7.300	1010.600
AQI_CO	27351.0	2.813499	2.510524	0.0	1.000	2.000	4.000	22.000
AQI_NO2	26562.0	6.306603	5.211678	0.0	3.000	5.000	9.000	38.000
AQI_O3_1hr	98630.0	15.371013	9.866352	0.0	8.000	16.000	22.000	105.000
AQI_O3_4hr	102873.0	19.554616	11.741196	0.0	11.000	20.000	27.000	123.000
AQI_PM10	76188.0	25.752796	59.316042	-2.0	12.000	19.000	27.000	2215.000
AQI_PM2.5	64215.0	30.160601	107.385830	-7.0	13.000	20.000	29.000	4043.000
AQI_Site	104923.0	32.221334	85.336756	0.0	19.000	26.000	33.000	4043.000

Fig. 2: Numerical summary.

We may observe that the mean value for PM2.5 is 7.5442383 and it ranges from -1.7 to 1010.600. This may indicate the existence of outliers. In order to observe that and to get initial impressions about the data, the time series visualisation of the dataset was plotted. It can be seen that we have very high values(≈ 1000) of PM2.5 for the year 2020. This might indicate the greater the level of air pollution and health concern. Generally, the value that is below 50 might indicate good air quality and over 150 represents hazardous air quality.



	Name	GPS	DateTime	NO2	O3_1hr	O3_4hr	CO	PM10	PM2.5	AQI_CO	AQI_NO2	AQI_O3_1hr	AQI_O3_4hr	AQI_PM10	AQI_PM2.5	AQI_Site	Date	Time
0	Civic	(-35.285307, 149.131579)	10/11/2022 11:00:00 AM	NaN	0.024	0.020	0.0	7.00	3.62	0.0	0.0	24.0	25.0	14.0	14.0	25.0	10 November 2022	11:00:00
1	Civic	(-35.285307, 149.131579)	10/11/2022 12:00:00 PM	NaN	0.026	0.023	0.0	6.65	3.50	0.0	0.0	26.0	29.0	13.0	14.0	29.0	10 November 2022	12:00:00
2	Civic	(-35.285307, 149.131579)	10/11/2022 01:00:00 PM	NaN	0.027	0.025	0.0	6.58	3.55	0.0	0.0	27.0	31.0	13.0	14.0	31.0	10 November 2022	13:00:00
3	Civic	(-35.285307, 149.131579)	10/11/2022 02:00:00 PM	NaN	0.029	0.026	0.0	6.60	3.52	0.0	0.0	29.0	33.0	13.0	14.0	33.0	10 November 2022	14:00:00
4	Civic	(-35.285307, 149.131579)	10/11/2022 03:00:00 PM	NaN	0.029	0.028	0.0	6.38	3.44	0.0	0.0	29.0	35.0	12.0	13.0	35.0	10 November 2022	15:00:00

Fig. 1: First five observations from the whole dataset.

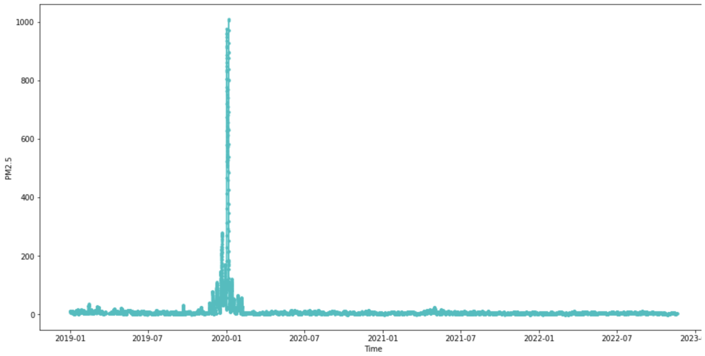


Fig. 3: Time series visualization from year 2020.

This could also be observed from the box plot in Figure 4, which represents PM2.5 by month in 2020.

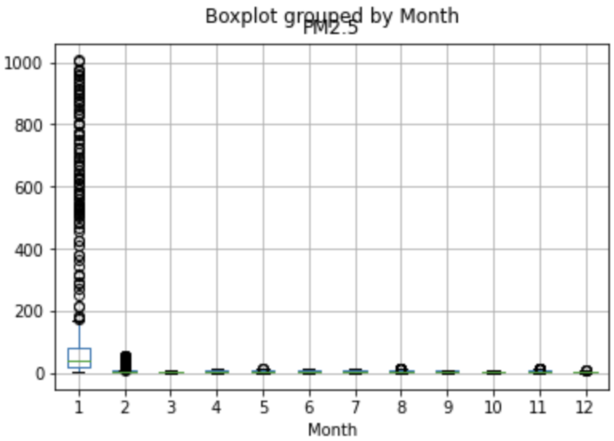


Fig. 4: PM2.5 by month in year of 2020.

Next, we could observe how PM2.5 has changed on average over the years from 2015 to 2022.

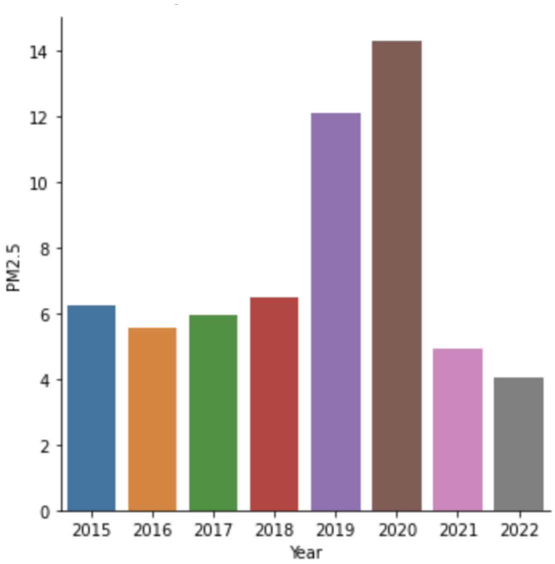


Fig. 5: PM2.5 from 2015 to 2020.

The monthly seasonal trend in Figure 6 demonstrates that there are high values for NO2 and CO and low values for PM2.5 and O3_4hr in Spring months. Hence, they are inversely related to each other.

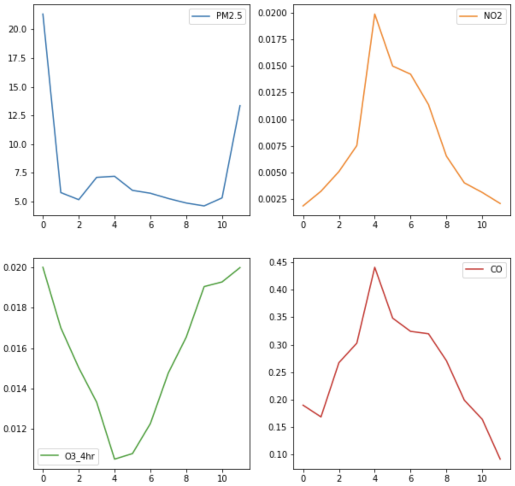


Fig. 6: PM2.5, NO2, CO and O3_4hr by monthly

Further, in order to construct models we will need to study the relationship between different features of our dataset. It can be done by constructing correlation matrix (heatmap). (Figure 7)

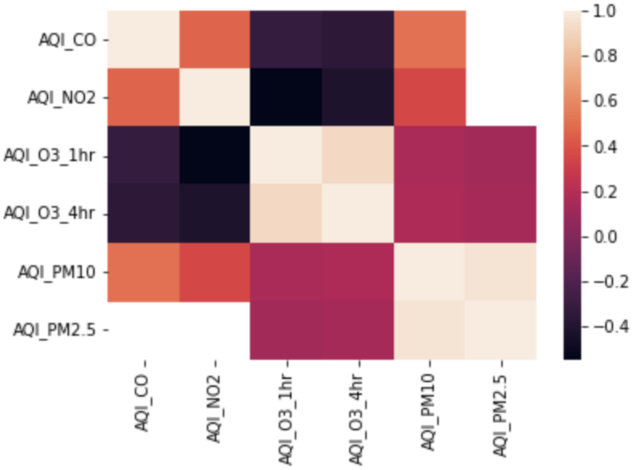


Fig. 7: Correlation matrix for the dataset.

It can be observed that there is some positive relationship between PM2.5 and AQI_O3_1hr, AQI_O3_4hr, AQI_PM10 variables.

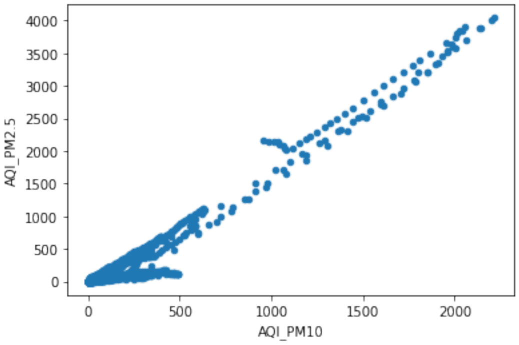


Fig. 8: The relationship between AQI_PM2.5 and AQI_PM10.

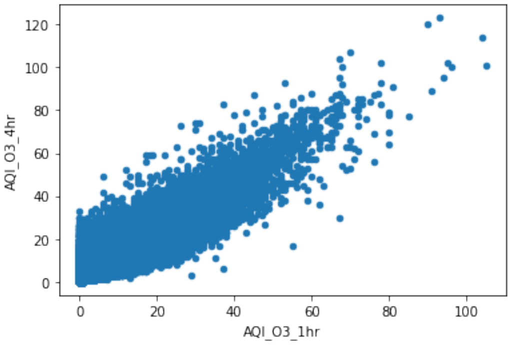


Fig. 9: The relationship between AQI_O3_1hr and AQI_O3_4hr.

By observing the above correlation coefficients, a scatter plot of partial variables is made to more intuitively observe the relationship between the variables.

From Figure 8, we can see the obvious linear relationship between AQI_PM2.5 and AQI_PM10. From Figure 9, we can see the linear relationship between AQI_O3_1hr and AQI_O3_4hr.

V. Method

A. Fit the Regression Model

Because the numeric concentration PM2.5 was the response variable in this project and the predictor variables selected were also the numeric ones, this study first used the multiple linear regression model in order to predict the numeric response variable.

Multiple linear regression, to be simple, refers to a statistical technique that uses two or more predictor variables to predict the outcomes.

Before coming into fitting the regression model, this project first did some data wrangling and data cleaning. First, this study used AQI of chemicals, AQI_CO, AQI_NO2, AQI_O3_1hr, AQI_O3_4hr and AQI_PM10 as the predictor variables and AQI of PM2.5 as the response variable. After checking the summary statistics of the non-null PM2.5, this study used all the observations with non-null AQI_PM2.5. Second, this study wanted to develop a prediction model using the chemicals in the present hour to predict the AQI level of PM2.5 in the next hour. Therefore, the AQI_PM2.5 in the next hour is considered as the response variable, while chemicals at the present hour are considered as the predictors. One more point to mention was that the number of non-null CO and non-null NO2 were too small, which may not be useful when adding into the model.

Therefore, after choosing observations based on the criterion above, this study then divided all the observations into train and test dataset with a division as 0.75 to 0.25. And this study first started with fitting multiple regression with three predictor variables, AQI_O3_1hr, AQI_O3_4hr and AQI_PM10. And the p-values from this multiple regression showed that all the three variables are considered as significant, which is denoted as Model1. After checking the autocorrelation between AQI_O3_1hr and AQI_O3_4hr, it seems that they are correlated with each other and shared with quite high VIFs. Also, since AQI_O3_1hr and AQI_O3_4hr are both related to the concentration level of O3, therefore, this study further tried a smaller model with only AQI_O3_4hr and AQI_PM10 as the predictor variables, which is denoted as Model2. This study kept AQI_O3_4hr as the indicator of the concentration level of O3, since the p-value of AQI_O3_4hr is much smaller, indicating its significance.

This study would use RMSE as the performance metric in order to compare the performance of two models.

After fitting the Model2, this study also conducted some analysis for assumptions check. The mean for the residuals was approximately near 0, while the VIFs were all around 1, which may indicate few multicollinearity. The only one assumption that may be offended was the homogeneity of variance, indicating that in the future, this study could conduct some other models to check the results, like non-parametric linear regression or some linear mixed effects models.

B. Fit the Classification Model

The study further divided the concentration levels of PM2.5 into different categories, good, moderate, unhealthy for sensitive people and unhealthy as in the following table.

Categories	PM2.5
Good	Less Than 50
Moderate	50 to 100
Unhealthy for Sensitive People	100 to 150
Unhealthy	Higher than 150

TABLE I: Changing numeric PM2.5 into Categorical Level

After changing the numeric concentration of PM2.5 into different categories, this study tried to fit two different classification models with those available data, the decision tree model and random forest model.

Decision tree was considered as an non-parametric supervised learning algorithm that could be utilized in the classification task and random forest was considered as the combination of decision trees.

VI. Results/Discussion

Take the observation on 02/21/2016 09:00:00 as an example, for the regression model, the predicted concentration for PM2.5 is 17.5481, while the actual concentration is 18, which are quite close to each other.

As for the classification models, the two models, decision tree and random forest, will predict the actual category for PM2.5 to be good, which are the same as the actual category.

Therefore, following this one observation, those three different models would perform quite well.

As shown by the table below, the two regression models have similar RMSE. Since Model2 realized more with our simplicity concern, this study used Model2 to analyze further. As indicated by the coefficients of the multiple regression, with a positive coefficient of AQI_PM10 to be 1.6438, PM10 exerts a positive influence on the concentration level of PM2.5. However, as indicated by the negative coefficient of AQI_O3_4hr, O3 reversely exerts a negative influence on the concentration level of PM2.5.

Model	Model1	Model2
RMSE	29.5087	29.5127

TABLE II: RMSE with Two Multiple Linear Regression

As for the classification models, after training these two models, the study have compared the accuracy and found that random forest model would do better in predicting the categories of PM2.5 in the next one hour.

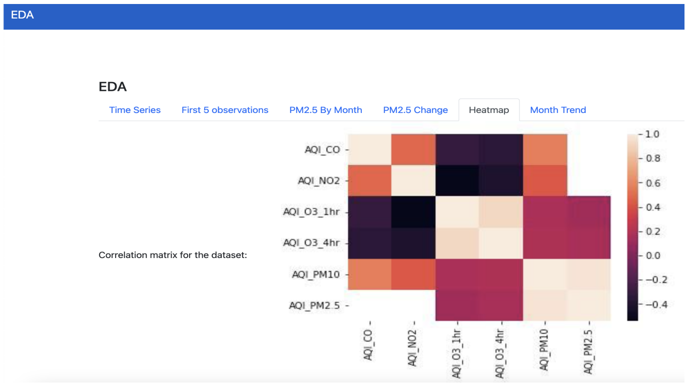
Model	Decision Tree	Random Forest
Accuracy	0.993	0.995

TABLE III: Accuracy with Two Classification Models

When looking further into the confusion matrix of the Random Forest, the model could achieve in predicting each category of PM2.5 with around the accuracy of 0.9, as shown in the following table. However, this model did very well in predicting the good category, while a little bit worse in predicting unhealthy category.

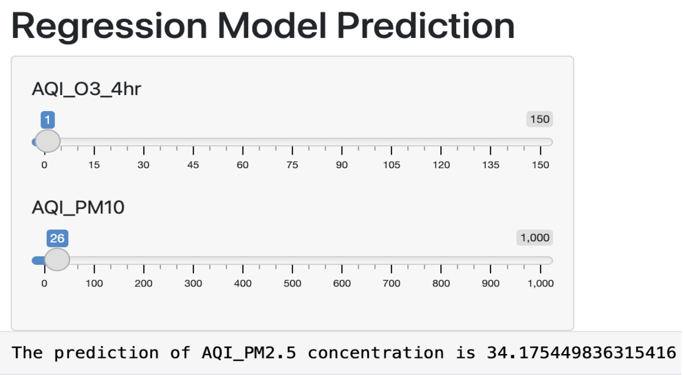
Categories	Predicting Accuracy
Good	1.00
Moderate	0.94
Unhealthy for Sensitive People	0.97
Unhealthy	0.92

TABLE IV: Accuracy of Random Forest in Predicting Each Level of PM2.5



Correlation matrix for the dataset:

Fig. 10: Shiny App for EDA graphs.



Regression Model Prediction

```
The prediction of AQI_PM2.5 concentration is 34.175449836315416
```

Fig. 11: Shiny App for Regression Model.

As in the previous two figures, this study has used the results to create two different shiny apps that could show the results of exploratory data analysis and predict the concentration of PM2.5 with the regression model. The shiny app for exploratory data analysis would showcase six different options for images from exploratory data analysis. And the shiny app for modelling could take the inputs as AQI_O3_4hr and AQI_PM10 and predict the final AQI_PM2.5 as the output.

VII. Conclusion and Future Work

This study has investigated the dataset collected by Australia government, including conducting exploratory data analysis, fitting regression models and classification models. Exploratory data analysis revealed that there may be linear relationship between AQI_O3_1hr and AQI_O3_4hr, as well as indicating a linear relationship between AQI_PM10 and AQI_PM2.5.

After the exploratory data analysis, this study also tried to fit the multiple linear regression model with different possible predictor variables. The result suggested that the model with AQI_O3_4hr and AQI_PM10 should be the one that have great RMSE. For this part, after checking the assumptions for multiple linear regression, this model seemed to offend the assumption of homogeneity of variance, leading to the possible future work of fitting non-parametric model or some linear mixed effect models.

As for the classification model, this study tried two different ones, including the decision tree and random forest. After comparing between those two models, the random forest is believed to be better in predicting different levels of PM2.5, especially in predicting good categories.

In the future, as mentioned above, this study could also try some different non-parametric models. Also, this study could compare the models fitted by the Australia dataset with the other frequently used dataset.

VIII. Appendix

A. Timeline of Work

Our work timeline and distributions so far are summarized below as the Figure 12 Gantt chart below.

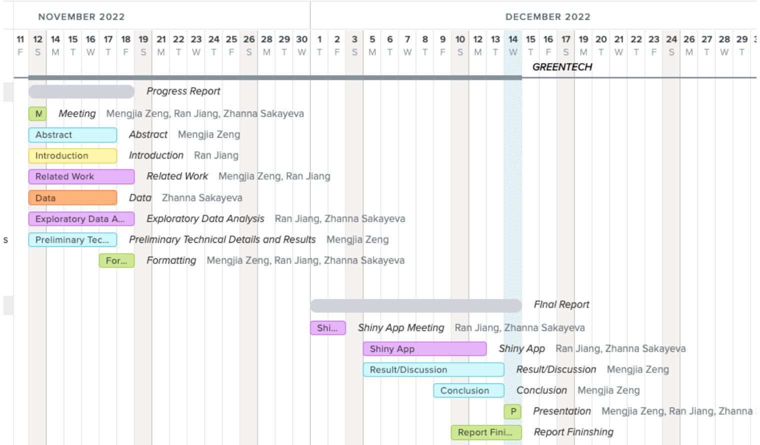


Fig. 12: Timeline of Work.

B. Code

The codes for the project can be accessed through these links Code and ShinyApp.

IX. Contributions

Ran Jiang (33%): Wrote the introduction, related work and some part of the exploratory data analysis. Wrote code for the shiny app.

Mengjia Zeng (33%): Wrote the abstract, related work methods, results, conclusion and future work. Wrote codes for data wrangling and data modeling.

Zhanna Sakayeva (33%): Wrote the paragraphs about data, exploratory data analysis and some part of related work, wrote codes for data visualization.

References

[1] Alléon, A., Jauvion, G., Quennehen, B., Cassard T., and Lissmyr, D. Plumenet: Large-scale air quality forecasting using a convolutional lstm network, 2020.

[2] Jiang, Wenrong. ”The data analysis of Shanghai air quality index based on linear regression analysis.” Journal of Physics: Conference Series. Vol. 1813. No. 1. IOP Publishing, 2021.

[3] Kang, Gaganjot Kaur, et al. ”Air quality prediction: Big data and machine learning approaches.” Int. J. Environ. Sci. Dev 9.1 (2018): 8-16.

[4] Ketu, Shwet. ”Spatial Air Quality Index and Air Pollutant Concentration prediction using Linear Regression based Recursive Feature Elimination with Random Forest Regression (RFERF): a case study in India.” Natural Hazards 114.2 (2022): 2109-2138.

[5] Li, Xiang, et al. ”Deep learning architecture for air quality predictions.” Environmental Science and Pollution Research 23.22 (2016): 22408-22417.

[6] Yi, Xiuwen, et al. ”Deep distributed fusion network for air quality prediction.” Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018.

[7] Xiangyu Zou, Jinjin Zhao, Duan Zhao, Bin Sun, Yongxin He, Stelios Fuentes, ”Air Quality Prediction Based on a Spatiotemporal Attention Mechanism”, Mobile Information Systems, vol. 2021, Article ID 6630944, 12 pages, 2021.