Auckland ICT Graduate School

Internship Final Report

[23-02-2024]

[Forecasting Air Quality based on Traffic and Weather Data]

[Author: Mengzhe Zhao]

[Project Supervisor: Philipp Skavantzos]

[Company: ICT Green Future Challenge]

[Company Mentor: Sean Zeng]

Declaration of Originality

This report is my own unaided work and was not copied from nor written in collaboration with any other person.

Name: [Mengzhe Zhao]

Forecasting Air Quality based on Traffic and Weather Data

Mengzhe Zhao
Auckland ICT Graduate School
The University of Auckland
Auckland, New Zealand
mzha910@aucklanduni.ac.nz

***Abstract* — This report covers the implementation of Forecasting Air Quality based on Traffic and Weather Data project. The aim of this report is to present a brief overview of the project; to review the literature about machine learning models for predicting air quality; to highlight the requirements of the project based on the public organisations and personal customers; to detail the project implementation including data analysis, machine learning models, and web development; to specify the project results; and to exposit the reflective learning for the project, including the personal contribution to the project, the technical and soft skills learned from the project, the knowledge from the previous courses, the issues during the process of the project, and the future work for the project.**

Keywords — forecasting air quality, traffic, weather, machine learning, public organisations, personal customers, data analysis, web development, reflective learning, personal contribution, technical and soft skills, previous courses, issues, future work.

I. Introduction

Tiny particles floating in the air, known as particulate matter (PM), pose significant risks, particularly affecting young children, the elderly, and those with existing heart or lung conditions. These particles can lead to serious health issues like reduced lung function, heart attacks, and, in severe cases, premature death The user can select the output format, the model version, and the city on the user interface [1]. Research also links these tiny particles to cognitive issues, hindered brain development, and an increased risk of diabetes. Among these, PM2.5 particles are hazardous due to their ability to penetrate deeply into the lungs. In New Zealand, burning wood and coal for heat and vehicle emissions are the primary sources of PM2.5 [1]. Recent findings highlight a concerning trend in Auckland, where PM levels have risen by 2-3% within a year, surpassing the safety limits set by the World Health Organization. Even in areas traditionally considered cleaner, such as Queenstown, there has been a significant 28.7% increase in PM levels, negatively impacting respiratory health. A 2022 study highlighted the severity of the issue, revealing that PM2.5 from human activities led to nearly 1,292 premature deaths in New Zealand in 2016, with an additional 4,626 hospitalizations for heart and lung diseases [2]. This data underscores the grave health risks of fine particles in the air.

As a student in data science, I have a keen interest in diving deeper into a particular issue that touches on the well-being of the public and the state of our natural surroundings. This driving force is not only born out of personal interest but also from a genuine concern for the health of communities and the protection of the environment. The focus is exploring the relationships between traffic flow, changing weather conditions, and air quality. By gaining a clearer understanding of these connections, the goal is to craft a predictive model that can forecast air quality in the future. Such a model could enable governments and public organisations to make informed decisions that significantly reduce pollution levels.

A. Green Futures Sustainability Challenge

The Green Futures Sustainability Challenge is a noteworthy initiative that effectively demonstrates the commitment of the ICT Graduate School to fostering practical and impactful learning experiences. This initiative showcases the school's dedication to addressing sustainability issues and promoting sustainable practices through the utilisation of information and communication technology. By engaging students in this challenge, the school provides us with an opportunity to develop our skills and knowledge in a real-world context, enabling us to contribute meaningfully to the field of sustainability. Overall, the Green Futures Sustainability Challenge serves as a testament to the ICT Graduate School's commitment to equipping students with the necessary tools and expertise to tackle pressing environmental challenges. The challenge at hand entails engaging students in a series of projects that are specifically designed to develop sustainable solutions for various environmental issues. Our project involves the utilisation of advanced data analysis techniques to forecast air quality. This is achieved by examining and interpreting a combination of traffic and weather data. By leveraging this comprehensive approach, we can contribute towards the development of effective strategies to relieve environmental concerns and promote a more sustainable future. The concept being described encompasses the harmonious integration of theoretical knowledge derived from academic sources and the practical application of problem-solving

techniques in real-world scenarios. This approach aims to inspire and motivate students to utilise our technical aptitude in purposeful endeavours that effectively tackle urgent and significant global issues.

The Green Futures Challenge serves as a multilevel platform that extends beyond a mere academic exercise. It provides an avenue for students to actively participate in projects that yield tangible outcomes in the realm of environmental sustainability. By engaging in this initiative, students are empowered to contribute meaningfully towards addressing pressing ecological concerns. Through active engagement in this challenge, the students enrolled at the ICT Graduate School are not only utilising our technical proficiency but are also actively contributing to a broader endeavour aimed at fostering the development of a sustainable future.

B. Project Background

The project on air quality forecasting using traffic and weather data with machine learning is an important and timely initiative. The primary goal is to predict how clean or polluted the air will be in urban areas, considering factors like traffic and weather.

Traffic is a significant source of air pollution in cities. Vehicles emit gases and particles that can harm our health. By studying the data from traffic, we can learn how these emissions affect our air. For instance, more cars on the road usually mean more pollution. Weather also affects air quality. Things like wind, rain, and temperature can change how pollutants move and how concentrated they are.

Machine learning, a type of artificial intelligence, is a powerful tool for this project. It can analyse vast amounts of data from traffic and weather and find patterns that humans might not see. Using machine learning, we can better predict air quality by combining traffic and weather information. By predicting air quality, we can help relevant organisations make better decisions to protect our health and the environment.

II. Literature Review

The study aims to predict daily PM2.5 levels using a novel deep learning model called the weighted long short-term memory neural network extended model (WLSTME) [3]. It addresses spatial and temporal correlations in air pollution, considering the uneven distribution of monitoring stations and wind conditions. The model uses a multilayer perceptron (MLP) for generating weighted PM2.5 data and integrates this with meteorological data [3]. The study, tested on data from Beijing-Tianjin-Hebei, demonstrates that WLSTME performs better than the existing models, significantly improving PM2.5 prediction accuracy across different seasons and regions [3].

The study focuses on developing and evaluating machine learning models for predicting the Air Quality Index (AQI) [4]. The author utilises a comprehensive dataset collected over 11 years from Taiwan's Environmental Protection Administration [4]. It compares different machine learning algorithms like Random Forest, Adaptive Boosting (AdaBoost), Support Vector Machine (SVM), and Artificial Neural Networks (ANN), assessing AQI prediction. The results highlight the superior performance of ensemble models in predicting AQI with higher accuracy and reliability [4].

The study presents a comprehensive study on predicting PM2.5 levels in Taiwan using a gradient-boosting machine learning approach [5]. The research is based on a large dataset from Taiwan's Environmental Protection Administration and Weather Bureau, including data from numerous air and weather stations [5]. The study explores the geographical and meteorological variations in air quality across Taiwan and compares the forecasting performance between Taiwan and London [5]. It also examines the impact of industrial pollution on air quality predictions, highlighting the significance of including industry-related features in predictive models [5].

The study outlines a research project that introduces a novel methodology for accurately predicting the levels of airborne particulate matter (specifically PM10 and PM2.5) in real-time inside South Korea [6]. The Interpolated Convolutional Neural Network (ICNN) is a combination of interpolation methods with neural network technology. This tool converts non-uniform spatial data into a standardized grid format compatible with the model [6]. The accuracy and reliability of the ICNN model are assessed in forecasting elevated levels of particulate matter [6]. The findings show a notable level of accuracy, indicating the ability to provide timely alerts and contribute to protecting public health and the environment [6].

The study discusses air quality prediction, highlighting three primary methodologies: numerical simulation, statistical methods, and machine learning [7]. It emphasises the evolution from complex numerical simulations, limited by assumptions on pollutant emissions and computational demands, towards statistical methods that explore data patterns without relying on meteorological theories [7]. Further, the review notes a shift towards machine learning due to its proficiency in handling nonlinear relationships, which is common in air quality data [7]. Deep learning, a subset of machine learning, is identified as particularly effective for predicting air quality because of its capability to model long-term dependencies, as demonstrated by Long-Short-Term-Memory (LSTM) networks. The review also mentions feature selection techniques to enhance model efficiency, pointing out the role of metaheuristic algorithms in optimising these selections [7]. Finally, it identifies a gap in addressing the spatial characteristics of air quality data, suggesting that incorporating spatial correlations could improve prediction accuracy [7].

This study investigates the effectiveness of hybrid deep learning methods, Convolutional Neural Network-Multilayer Perceptron (CNN-MLP) and Convolutional Neural Network-Multilayer Perceptron (CNN-LSTM), in detecting ARP Man in the Middle (MitM) attacks using the Kitsune Network Attack Dataset [8]. It compares the performance of these models across different feature scaling techniques: StandardScaler, MinMaxScaler, and MaxAbsScaler [8]. The findings indicate that the CNN-MLP model consistently outperforms the CNN-LSTM model regarding accuracy, precision, recall, and F1-Score across all feature scaling methods [8]. The study suggests that deep learning, particularly CNN-MLP with Standard

Scaler, can significantly enhance the detection of ARP MitM attacks, achieving high accuracy rates [8].

The study explores the prediction of the Air Quality Index (AQI) using machine learning techniques [9]. It focuses explicitly on comparing three distinct methods: Support Vector Regression (SVR), Random Forest Regression (RFR), and CatBoost Regression (CR), using datasets from four major Indian cities. The study demonstrates that RFR generally offers the lowest root mean square error (RMSE) values across multiple cities, indicating higher AQI prediction accuracy than SVR and CR [9]. The use of the Synthetic Minority Oversampling Technique (SMOTE) for data balancing is highlighted, showing its effectiveness in enhancing model performance [9]. This investigation provides a comprehensive analysis and comparison of machine learning techniques for AQI prediction, contributing valuable insights for environmental monitoring and public health initiatives [9].

This study introduces a new spatial-temporal deep learning model, the Pollution-Predicting Net (PPN), designed to predict PM2.5 levels over the Beijing-Tianjin-Hebei region in China [10]. It leverages an encoder-decoder architecture, integrating preceding PM2.5 observations and numerical weather prediction data [10]. The model uses a weighted loss function to enhance forecasting accuracy, especially during extreme pollution events [10]. The results demonstrate the PPN model's superior performance in predicting air quality, offering a promising tool for early warning systems and regional pollution management compared to traditional models like WRF-Chem [10].

The study presents a comprehensive approach to traffic and pollution modelling in Zaragoza, Spain, to enhance air quality awareness. It emphasises the integration of various data sources, including traffic flow, vehicle types, meteorological conditions, and pollutant emissions, to simulate traffic patterns and their environmental impact. The study utilises SUMO for traffic simulation and Vehicular Emissions Inventories (VEIN) for emissions estimation, highlighting the importance of accurate and dynamic models in predicting air quality. This methodology aims to inform citizens and policymakers, supporting more intelligent urban planning and health-conscious decision-making.

III. Project Requirements

There are several business goals for our project, including the following:

- Based on the prediction model, governments and relevant organisations can publish policy to control traffic, and to provide accurate warnings about air pollution, which could affect human health and other social activities.
- Based on the prediction model, the developed web application allows everyday users to predict the air quality, including air quality index (AQI), fine particulate matter (PM2.5) and particulate matter 10 (PM10).
- Based on the prediction model, the developed web application allows professional users to get the back-end machine learning model and the collected data in Auckland, Wellington, and Christchurch.

IV. Project Implementation

A. Data Collection

based on the research work we have done; we made the decision to combine Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) to build the prediction model. Secondly, we collected the traffic, weather, and air quality data from several sources. The traffic data is from the New Zealand Transport Agency (NZTA). The traffic data source website is https://opendata-nzta.opendata.arcgis.com/datasets/NZTA::tms-daily-traffic-counts-api/explore. The weather data is from the National Institute of Water and Atmospheric Research (NIWA). The weather data source website is https://cliflo.niwa.co.nz/. The air quality data is from the regional council, including Auckland Council, Greater Wellington Regional Council, and Environment Canterbury. The air quality data source websites are https://environmentauckland.org.nz/Data/DataSet/Interval/Latest, https://graphs.gw.govt.nz/#dataViewer, and https://data.ecan.govt.nz/Catalogue/Search?Query=air&CollectionId=0.

The dataset's time range is determined from the beginning of 2020 to the end of 2023. We structure three different datasets for three cities: Auckland, Wellington, and Christchurch. The following statement explains why three single datasets exist for three cities, respectively. Firstly, the time range of Auckland data is limited up to August 2023. Fig. 1 is from the Environmental Data Portal of Auckland Council. As it shows, the air quality monitoring station at Queen Street stopped monitoring the air quality on 23rd August 2023, and there has yet to be air quality data after that date. This is why I can only consider the time range up to 23rd August 2023 for Auckland. For the datasets of Wellington and Christchurch, I can collect the data up to the end of 2023. Secondly, the air quality relevant features collected from three different councils differ. Three different cities may have different focused air quality features for the respective environmental conditions. Only Auckland Council considers the Air Quality Index (AQI). Only Auckland Council and Greater Wellington Regional Council monitor ozone (O_3). Only Auckland Council and Environment Canterbury monitor sulphur dioxide (SO_2). Only Greater Wellington Regional Council and Environment Canterbury monitor carbon monoxide (CO). Thus, combining three datasets into one whole dataset is not appropriate. For example, there will be many missing values of the AQI feature on the rows labelled as 'Wellington' or 'Christchurch'. So, three different machine learning models are built based on three separate datasets for accurate prediction.

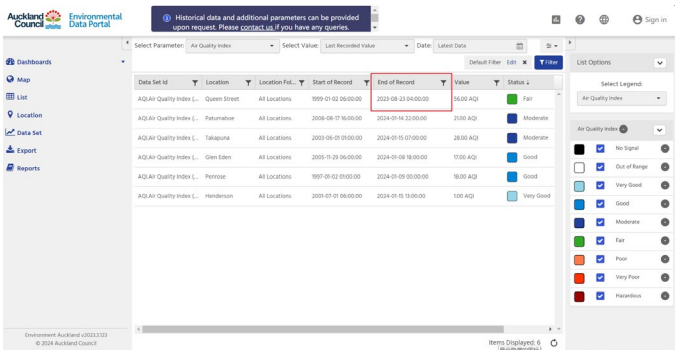

Fig. 1. Screenshot from Auckland Council

There two features about traffic; one is light traffic count (such as personal cars), another one is heavy traffic count (such as industrial trucks). Fig. 2 demonstrates the features relevant to weather. The air quality features for Auckland contain AQI, PM2.5, PM10, O_3, SO_2, and NO_x (Fig. 3). The air quality features for Wellington include PM2.5, PM10, NO_2, O_3, and CO (Fig. 4). The air quality features for Christchurch include PM2.5, PM10, CO, NO, NO_2, and SO_2 (Fig. 5). Each row of the data frame means the daily average value of each feature.

Fig. 2. Weather Features

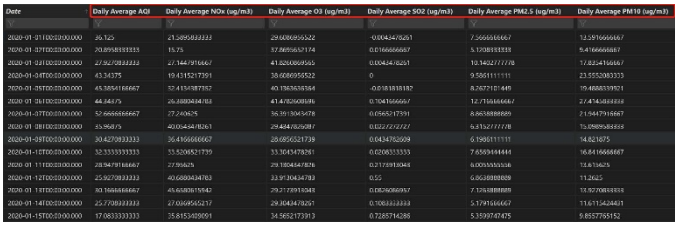

Fig. 3. Air Quality Features for Auckland

Fig. 4. Air Quality Features for Wellington

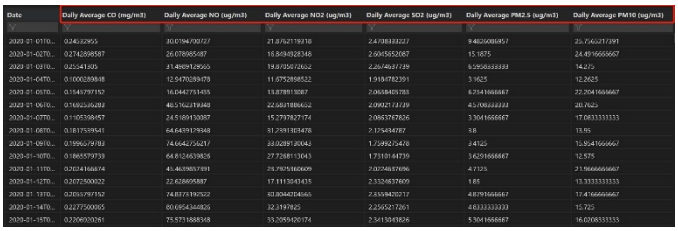

Fig. 5. Air Quality Features for Christchurch

B. *Exploratory Data Analysis (EDA)*

The day-of-week average values for PM10 show that more human activities during the weekday mean worse air quality, and fewer activities during the weekend mean better air quality (Fig. 6).

There is seasonal periodicity for the monthly average value of PM2.5 in Christchurch, which could be caused by burning for heating in the winter (Fig. 7). Christchurch is on the Southern Island, colder than the Northern Island in winter.

The geospatial analysis for Auckland exposes that the region with more traffic has the worst air quality (Fig. 8).

Fig. 6. Day of Week Average PM10 in Wellington

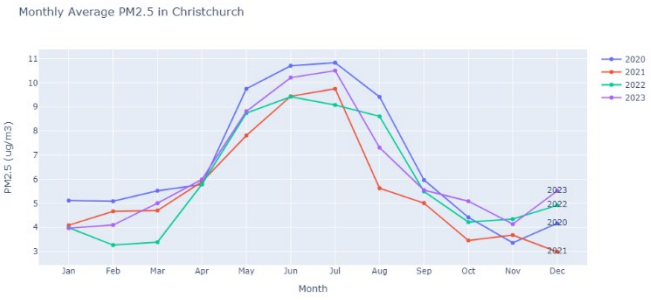

Fig. 7. Monthly Average PM2.5 in Christchurch

Fig. 8. Geospatial Analysis for Auckland

C. Data Preprocessing

1) Missing Values

There are two types of features: seasonal features and non-seasonal features. The seasonal mean is used to fill in the missing values of seasonal features. The median is utilised to fill the missing values of non-seasonal features. For the seasonal mean function, the parameter 'n' is set as '365' (Fig. 9). Fig. 10 compares the different imputing methods for a seasonal feature (O_3 ug/m3 in Auckland).

```
def seasonal_mean(ts, n, lr=0.7):
    """
    Compute the mean of corresponding seasonal periods
    ts: 1D array-like of the time series
    n: Seasonal window length of the time series
    """
    out = np.copy(ts)
    for i, val in enumerate(ts):
        if np.isnan(val):
            ts_seas = ts[i-1::-n]  # previous seasons only
            if np.isnan(np.nanmean(ts_seas)):
                ts_seas = np.concatenate([ts[i-1::-n], ts[i::n]])  # previous and forward
            out[i] = np.nanmean(ts_seas) * lr
    return out
```

Fig. 9. Seasonal Mean Function

2) Outliers

The outliers are detected using the interquartile range and replaced using the median.

3) Data Scaling

There are three types of data distribution: normal distribution, skewed normal distribution, and non-normal distribution (Fig. 11, Fig. 12, & Fig. 13). Min-Max Scaler is used so that the data can be scaled into the range between 0 and 1 for the CNN algorithm, and the original different distribution can be maintained.

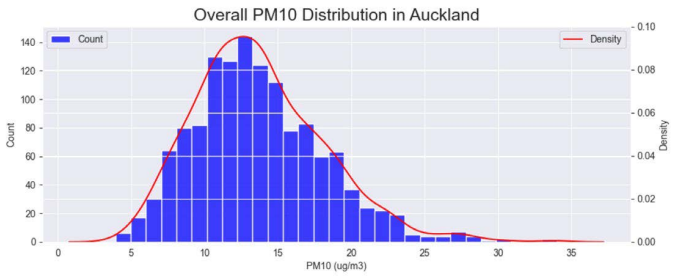

Fig. 11. Normal Distribution

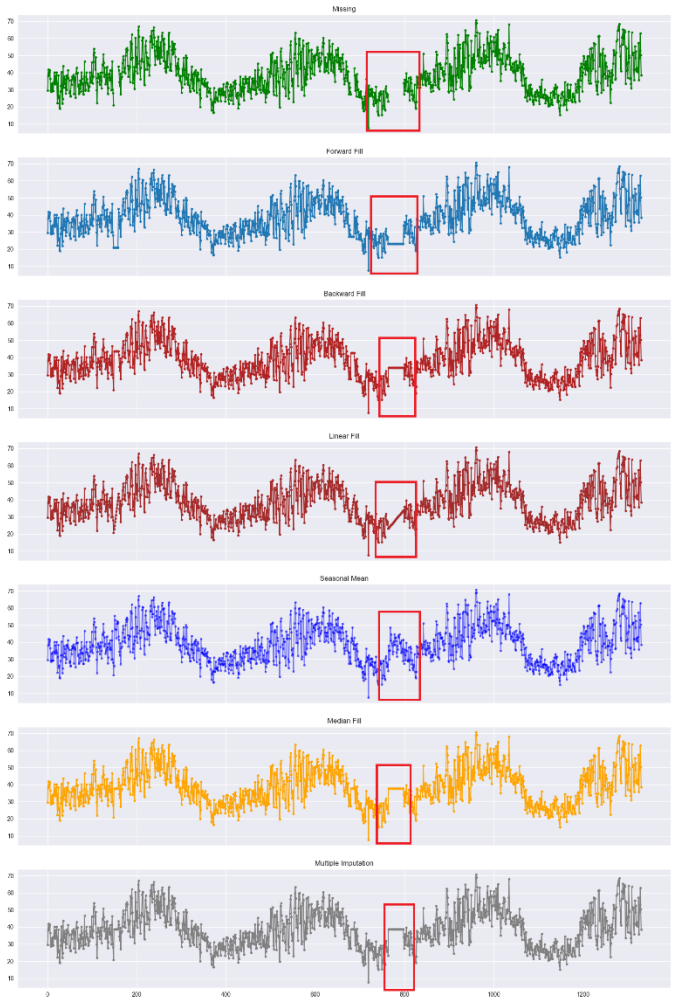

Fig. 10. Different Imputing Methods Comparison

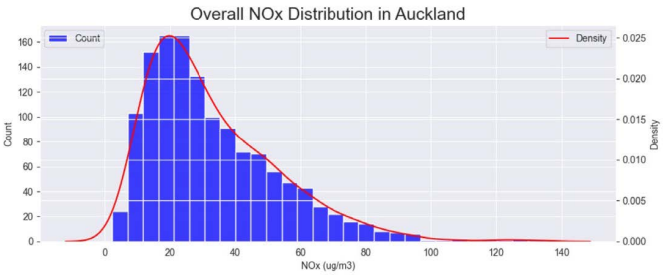

Fig. 12. Skewed Normal Distribution

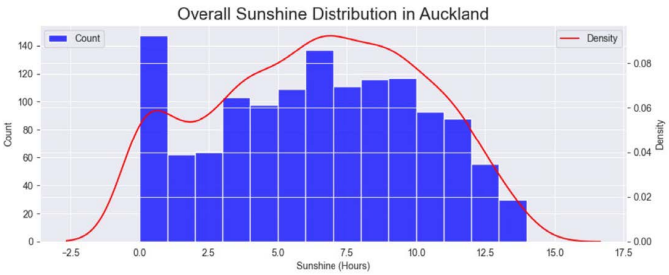

Fig. 13. Non-Normal Distribution

D. Feature Engineering and Feature Selection

1) Feature Engineering

'Month' and 'Year' are built by extracting each date's month and year. 'Day of Week' and 'Is Weekend' are built by extracting each day of the week and judging if it is a weekend. 'Average Temperature' is established by calculating the average of the maximum and minimum temperatures. 'Total Traffic' is established by calculating the sum of the light and heavy traffic counts.

2) Feature Selection

- For Auckland, relative humidity is the top feature correlative to PM2.5 and PM10 (Fig. 14 & Fig. 15).
- For Wellington, the top feature correlative to PM2.5 is wind run, and the top feature correlative to PM10 is gust speed (Fig. 16 & Fig. 17).
- For Christchurch, the top feature correlative to PM2.5 is CO, and the top feature correlative to PM10 is NO (Fig. 18 & Fig. 19).

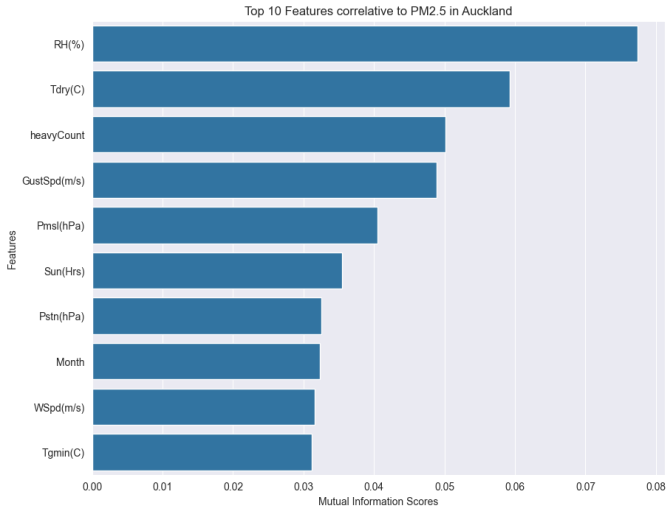

Fig. 14. Top 10 Features Correlative to PM2.5 in Auckland

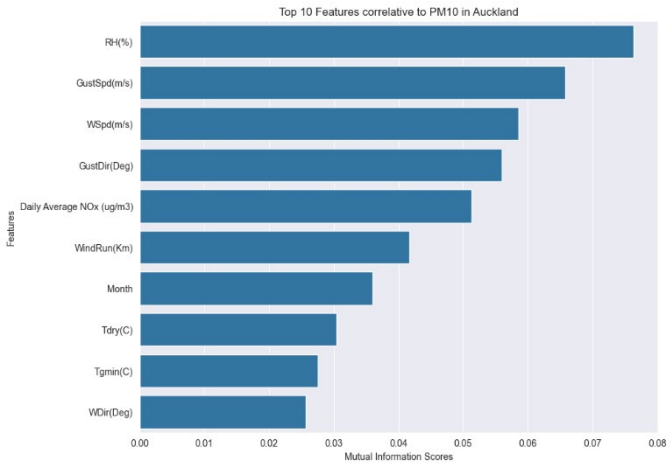

Fig. 15. Top 10 Features Correlative to PM10 in Auckland

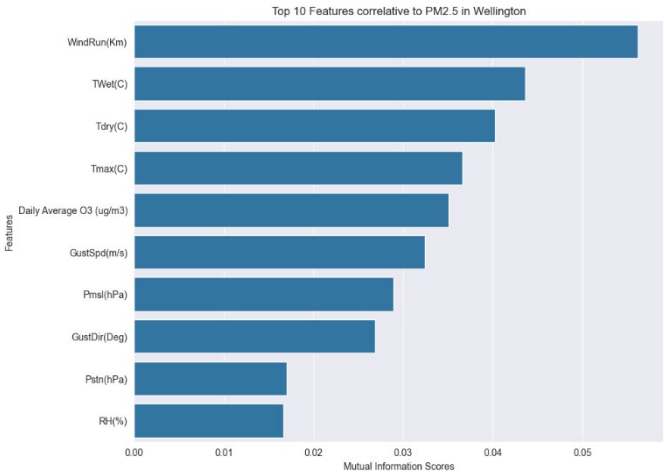

Fig. 16. Top 10 Features Correlative PM2.5 in Wellington

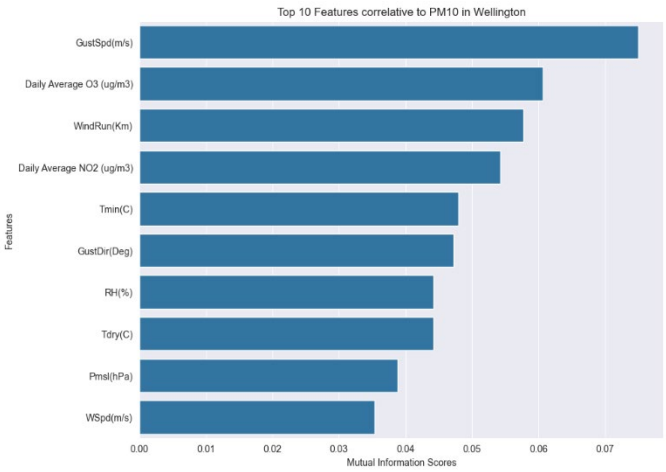

Fig. 17. Top 10 Features Correlative PM10 in Wellington

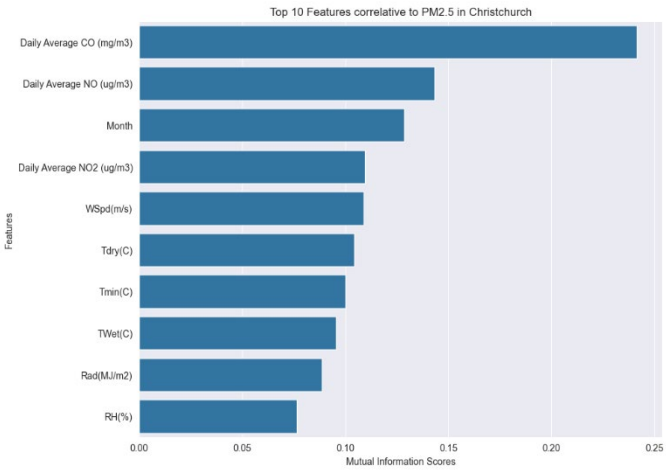

Fig. 18. Top 10 Features Correlative PM2.5 in Christchurch

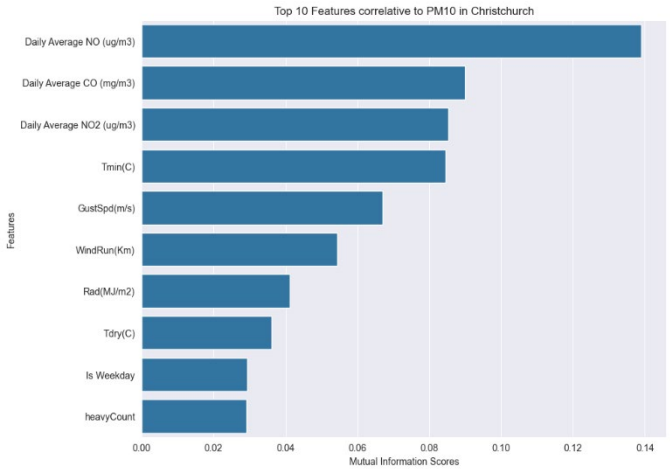

Fig. 19. Top 10 Features Correlative PM10 in Christchurch

E. Professional Version Model

Several algorithms can be utilised for time-series prediction. Traditional machine learning algorithms include Linear Regression, Decision Tree, Random Forest, and XGBoost. For deep learning algorithms, there are Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU). Fig. 20 shows the model comparison. Mean Squared Error (MSE) is used to evaluate the model accuracy; the smaller MSE means the prediction is closer to the actual situation.

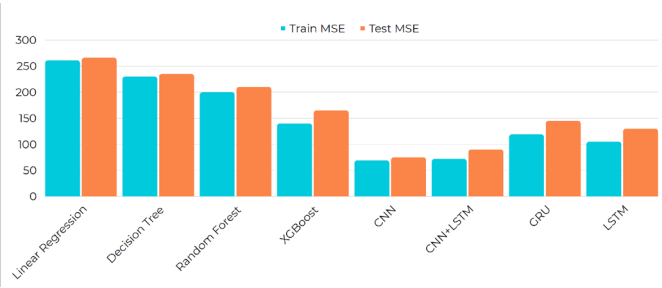

Fig. 20. Model Comparison

It is obvious that the CNN integrated with LSTM has the best performance. Thus, we decided to use the CNN integrated with LSTM. For the primary version model, only one CNN layer is built, which is used as the popular version model (Fig. 21). For the final version model, there are two CNN layers, and the model is optimised by many methods, which is the professional version model (Fig. 22).

The following statements are the model optimisation techniques.

- Model Architecture Tuning: The CNN and LSTM layers, including the number of layers, neurons, and kernel sizes, are adjusted to balance model complexity and performance.
- Regularisation Techniques: The dropout and L1/L2 regularisation are implemented to prevent overfitting and improve the model's generalisation ability.
- Hyperparameter Optimisation: To enhance prediction accuracy, techniques like grid search or random search are used to find the optimal set of model parameters, including learning rate, batch size, and epoch number.

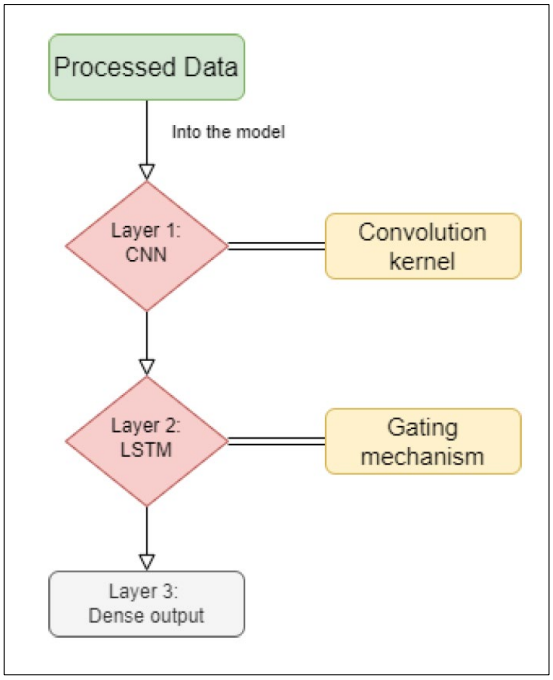

Fig. 21. Popular Model Structure

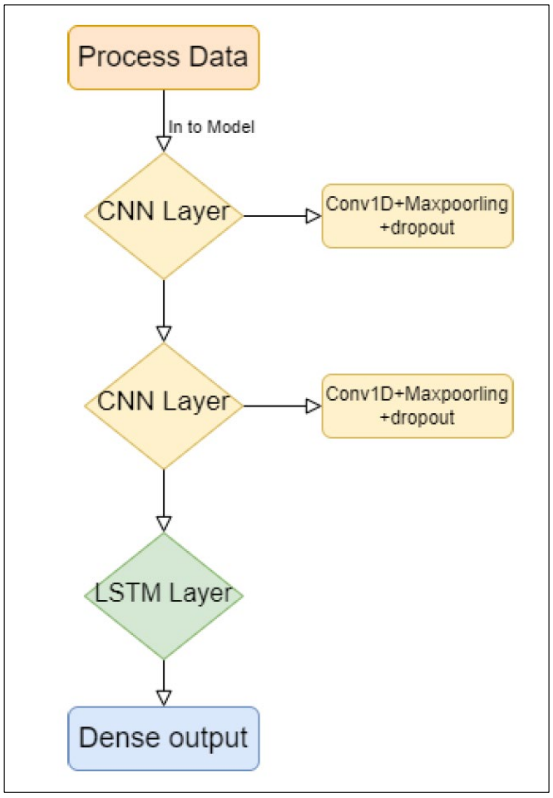

Fig. 22. Professional Model Structure

Mean Squared Error (MSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) are the indicators evaluating the model's performance. The smaller MSE and MAE mean the prediction is closer to the actual situation. The R^2 ranges from -∞ to 1. An R^2 value closer to 1 indicates that the model explains a higher proportion of the variance in the dependent variable, mean the model fits the data better. TABLE 1 and TABLE 2 show the indicators of Auckland's model and Christchurch's model. Fig. 23 and Fig. 24 show the actual values versus the predictive values for Auckland's model. Fig. 25 and Fig. 26 show the actual values versus the predictive values for Christchurch's model. Christchurch's model has the best performance compared to the other cities.

TABLE 1

Indicators for Auckland's Model

Indicators	Primary	Optimised
MSE	34.179	25.500
MAE	3.990	3.340
Train R^2	0.380	0.426
Test R^2	0.183	0.200

TABLE 2

Indicators for Christchurch's Model

Indicators	Primary	Optimised
MSE	13.889	12.110
MAE	3.190	2.508
Train R^2	0.720	0.770
Test R^2	0.400	0.470

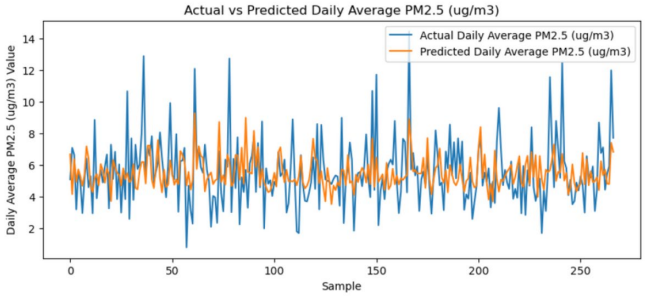

Fig. 23. Actual Values vs. Predictive Values for PM2.5 in Auckland

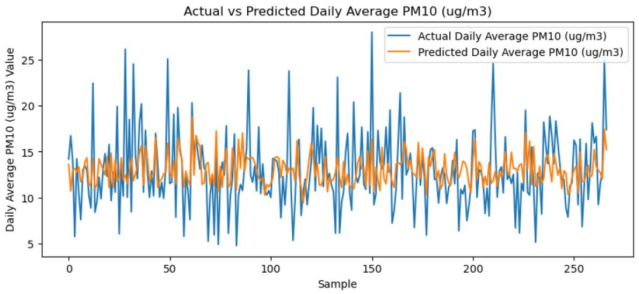

Fig. 24. Actual Values vs. Predictive Values for PM10 in Auckland

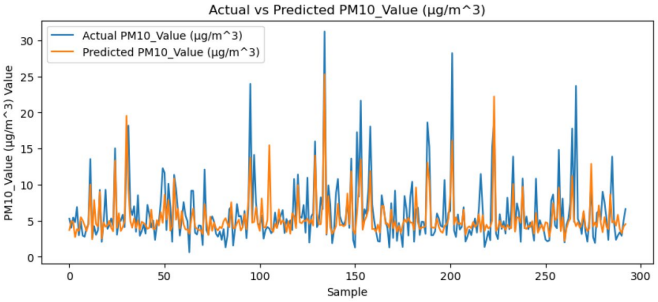

Fig. 25. Actual Values vs. Predictive Values for PM10 in Christchurch

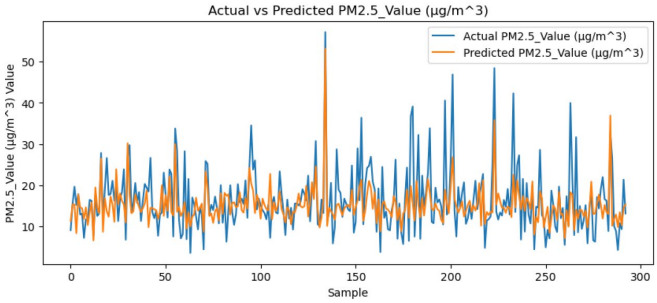

Fig. 26. Actual Values vs. Predictive Values for PM2.5 in Christchurch

F. *Popular Version Model*

The popular version model has fewer features compared to the professional version model, which are easily understandable to the general customers. The features considered in the popular version model contain wind direction, wind speed, rainfall, relative humidity, maximum temperature, minimum temperature, light traffic count, and heavy traffic count. The popular version model runs 1.39 times faster than the professional version model. TABLE 3 shows the indicators for three cities' models. Fig. 27 shows the actual values versus the predictive values for Auckland's model. Fig. 28 shows the actual values versus the predictive values for Wellington's model. Fig. 29 shows the actual values versus the predictive values for Christchurch's model. Auckland's model has the best performance compared to the other cities.

TABLE 3

Indicators for Popular Models

City	MSE	MAE	Test Loss
Auckland	11.333	2.426	0.012
Wellington	17.008	2.992	0.012
Christchurch	34.179	3.990	0.008

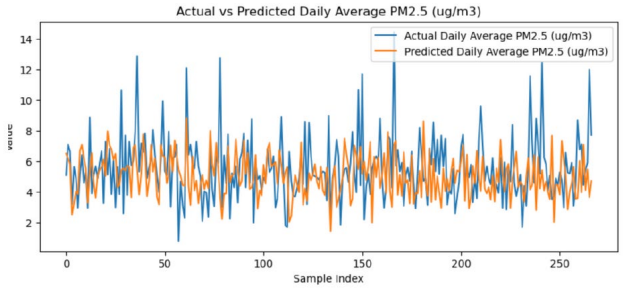

Fig. 27. Actual Values vs. Predictive Values for PM2.5 in Auckland

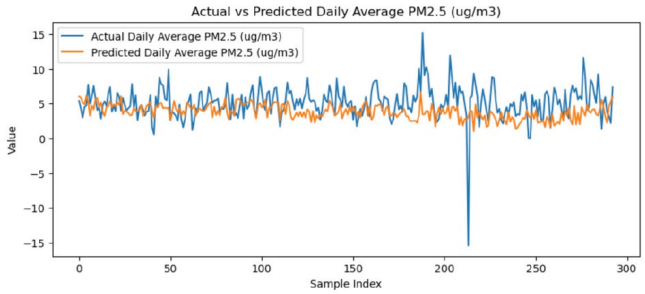

Fig. 28. Actual Values vs. Predictive Values for PM2.5 in Wellington

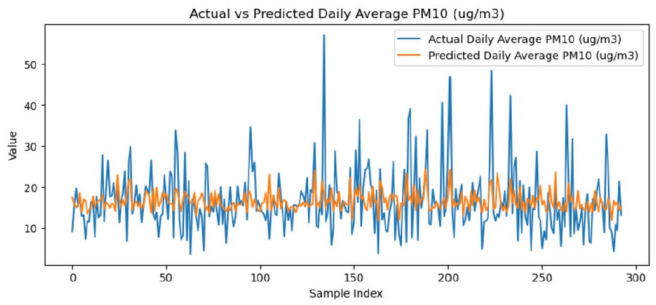

Fig. 29. Actual Values vs. Predictive Values for PM2.5 in Christchurch

G. *Web Development*

There are two functions of the web application. Firstly, the web is developed to allow users to predict air quality, including PM2.5 and PM10. Secondly, the customer can use the web to access the historical data of our project and export it as a file.

1) Air Quality Prediction

The user can select the output format, the model version, and the city on the user interface (Fig. 30).

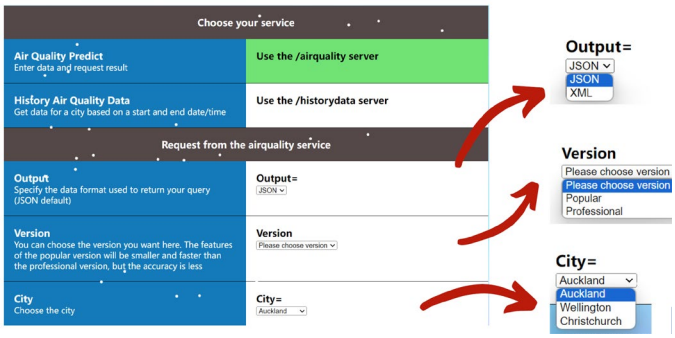

Fig. 30. Web User Interface

a) Popular Version Model

Fig. 31 shows the user interface of the popular version model. If the user does not have data for some features, our website can provide some recommended data from NZTA's real-time data API. Fig. 32 shows an output result example.

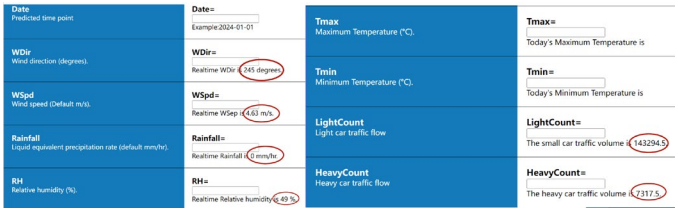

Fig. 31. Popular Version Interface

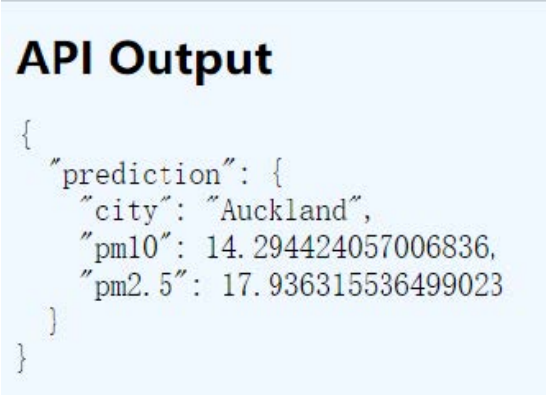

Fig. 32. Popular Version Output

b) Professional Version Model

For the professional version model, 19 features need to be imputed. Moreover, it needs more time than the popular version. Fig. 33 and Fig. 34 show the interface and the output example of the professional model.

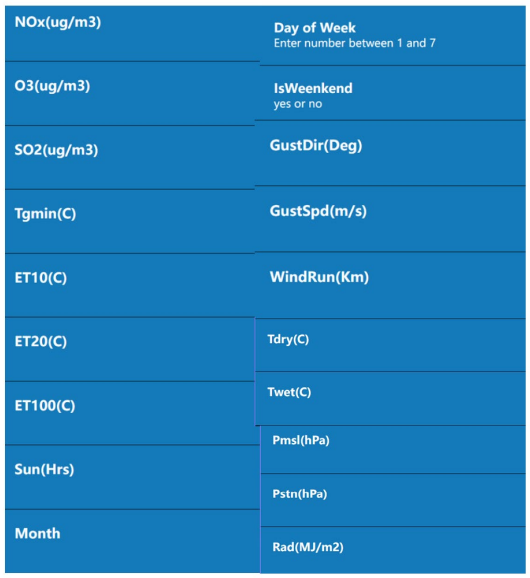

Fig. 33. Professional Version Interface

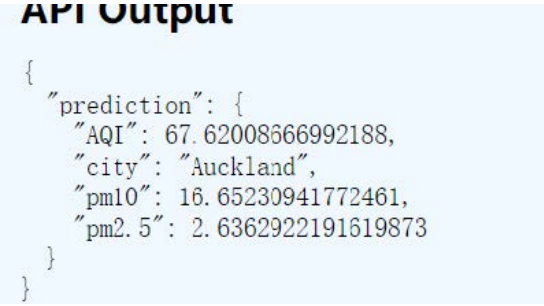

Fig. 34. Professional Version Output

2) Historical Data Access

The user can select the output format, the city, the start time, and the end time to get the specific data from the project database (Fig. 35 & Fig. 36). The user can select the specific features and get the customised dataset (Fig. 37).

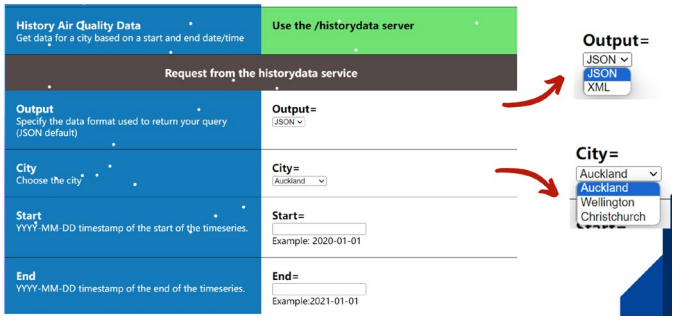

Fig. 35. Historical Data Access Interface

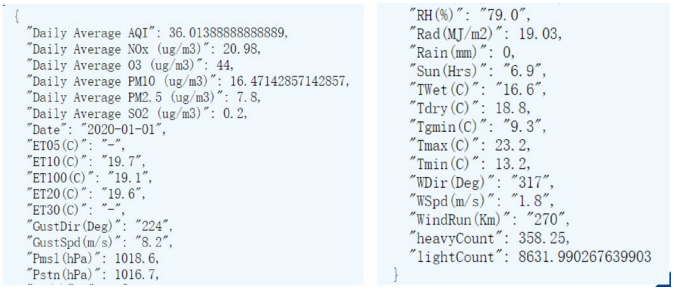

Fig. 36. Historical Data Output

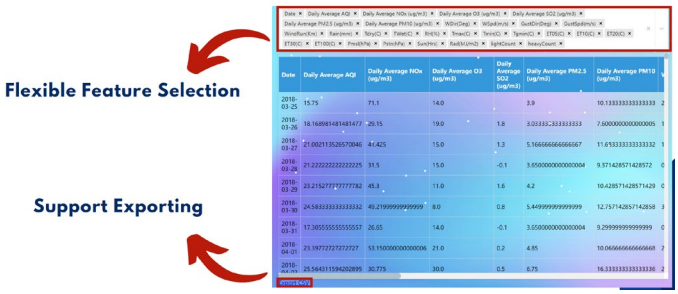

Fig. 37. Customised Feature Selection

V. Reflection

A. Personal Contribution

My contributions to the project include data collection, EDA, data preprocessing, feature engineering, and feature selection. Therefore, these sections are more detailed than the other sections.

B. Professional Attributes

Over the previous weeks, I have acquired several valuable skills through this Green Future Challenge internship. I will use the following bullet points to elaborate the benefits for my professional growth.

1) Technical Skills

a) Data Collection

I have learned how to collect data from several different sources and combine and restructure them into one whole dataset so that a machine learning model can be established based on it. For example, some features are monitored hourly at several monitoring stations, and I need to calculate the mean values of different stations and calculate the mean values of hourly values as the daily average value.

b) EDA and Data Prepprocessing

I have learned what should be completed at the EDA stage and the purpose of EDA for the project. Especially for the data preprocessing, which is important to the data quality, different imputing missing values methods are used for different features respectively. For example, the distribution of PM2.5 shows the apparent outliers, and the distribution of PM10 is a right-skewed normal distribution, so we use the median to input the missing values robustly.

2) Soft Skills

a) Teamwork

I have learned how to complete teamwork and communicate with other group members. Because I seldom work as a team member in the previous courses, this project is an excellent chance for me to improve my teamwork skills effectively. For example, the data quality often affects the machine learning model performance, and I need to communicate with the group member who works on the model frequently, restructure the datasets and try different data cleaning methods based on the model performance.

b) Networking

I have improved my networking skills through this project. In this project, the academic supervisor, the mentor, and the tutor instructed and assisted us. By communicating with them, I not only gained technical suggestions but also learned about industrial experience, which is crucial to me in preparing for future industrial work.

C. Application of Knowledge from Previous Courses

1) INFOSYS 722 – Data Mining and Big Data

In this course, I acquired a robust foundation in decision-making, machine learning, and data mining. Through practical application, I have learned to utilise various methods and techniques by SPSS, Python, and PySpark to analyse large datasets and extract valuable findings. This includes understanding how corporations effectively employ such data to achieve significant success. Through this course, I learned about data mining concepts and put them into action by developing a prototype through multiple iterations.

2) CIVIL 763 – Smart Infrastructure Analytics

In this course, I developed skills in data analytics for the real-world challenges of traffic engineering. I have studied most Python extension libraries in data science, including EDA, data preprocessing, geo-spatial analysis, and several machine learning algorithms. Through a data science project based on real-world traffic data, I implemented all theories to practical working, which are helpful to me in this Green Future internship.

3) COMPSCI 762 – Foundations of Machine Learning

I have gained a solid understanding of machine learning, including the theoretical and practical skills needed to tackle complex problems. I have learned to appreciate the computational and statistical aspects of learning from data, which has equipped me to implement and evaluate machine learning algorithms effectively. The course has broadened my perspective on the potential of machine learning in solving real-world challenges, enhancing my critical thinking and problem-solving skills. Additionally, mastering programming tools like Python for algorithm implementation has been a valuable skill set I have acquired.

D. Issue

During the project, I faced a tricky issue due to the data monitoring time of the regional council. At first, I used daily aggregate data, shown in Fig. 38. Air quality on Monday and Sunday is the best of the whole week, which differs from general cognition. After checking the data source, the reason is that the daily aggregate data is monitored at midnight each day. This means the data monitored at midnight on Monday reveals the actual air condition on Sunday. Thus, I collected the hourly aggregate data for each day and calculated the average value of the hourly aggregate as the daily average value of each day. The result is close to general cognition: Saturday and Sunday have the best air quality of the week (Fig. 39).

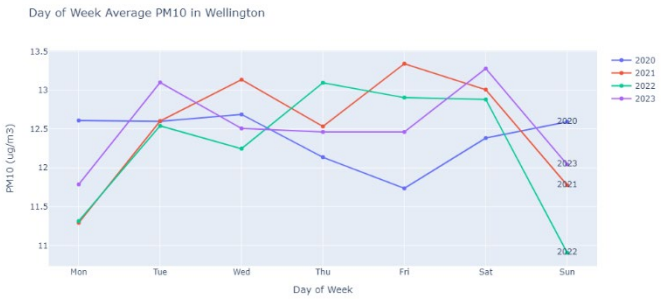

Fig. 38. Daily Aggregate Data

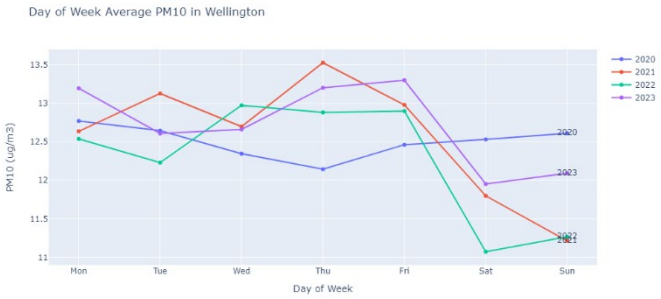

Fig. 39. Average Hourly Aggregate Data

E. Future Work

I use the following bullet points to state the future work, which could improve the performance of this project.

- Trying the methods of Transformer model [11].
- Further works on the feature engineering.
- Filtering for the data with better quality.
- Trying to integrate three cities' data into one whole database and build one whole machine learning model.

VI. Conclusion

In concluding this report on our project "Forecasting Air Quality based on Traffic and Weather Data," we have successfully developed a machine learning system to predict air quality. Our journey began with a thorough review of existing literature on similar models, laying the groundwork for our approach. We then identified the specific needs of public organizations and personal users, ensuring our project met these requirements.

The implementation phase involved analysing traffic and weather data, selecting and training suitable machine learning models, and developing a user-friendly web interface. Our

efforts resulted in a tool that can provide accurate air quality forecasts, a significant step forward in environmental monitoring and public health.

Reflecting on the project, I realized that my contribution spanned technical development and project management, enhancing both my technical abilities and soft skills. I leveraged knowledge from previous coursework and learned new skills, particularly in data analysis and machine learning. Despite facing challenges, such as data quality issues and technical setbacks, the project taught me valuable problem-solving techniques and the importance of perseverance.

There is potential to refine our models and expand the project's scope. This experience has been immensely educational, laying a solid foundation for my future data science technology.

ACKNOWLEDGMENT

I am grateful to Auckland ICT Graduate School for offering me this valuable internship experience. Working in the industry has shown me exactly what skills are in demand and has guided me towards what I should focus on in my future studies.

I want to thank our academic supervisor, Philipp Skavantzos; our industry mentors, Andrew Meads and Sean Zeng; and our project tutor, Xiaoheng Ji, for their support and guidance during the internship.

REFERENCES

[1] 'PM$_{2.5}$ concentrations | Stats NZ'. Accessed: Feb. 21, 2024. [Online]. Available: https://www.stats.govt.nz/indicators/pm2-5-concentrations

[2] 'EHINZ'. Accessed: Feb. 21, 2024. [Online]. Available: https://www.ehinz.ac.nz/projects/hapinz3/key-findings-from-hapinz/

[3] F. Xiao, M. Yang, H. Fan, G. Fan, and M. A. A. Al-qaness, 'An improved deep learning model for predicting daily PM2.5 concentration', *Sci. Rep.*, vol. 10, no. 1, p. 20988, Dec. 2020, doi: 10.1038/s41598-020-77757-w.

[4] Y.-C. Liang, Y. Maimury, A. H.-L. Chen, and J. R. C. Juarez, 'Machine Learning-Based Prediction of Air Quality', *Appl. Sci.*, vol. 10, no. 24, p. 9151, Dec. 2020, doi: 10.3390/app10249151.

[5] M. Lee *et al.*, 'Forecasting Air Quality in Taiwan by Using Machine Learning', *Sci. Rep.*, vol. 10, no. 1, p. 4153, Mar. 2020, doi: 10.1038/s41598-020-61151-7.

[6] S. Chae, J. Shin, S. Kwon, S. Lee, S. Kang, and D. Lee, 'PM10 and PM2.5 real-time prediction models using an interpolated convolutional neural network', *Sci. Rep.*, vol. 11, no. 1, p. 11952, Jun. 2021, doi: 10.1038/s41598-021-91253-9.

[7] Z. Zhao, J. Wu, F. Cai, S. Zhang, and Y.-G. Wang, 'A hybrid deep learning framework for air quality prediction with spatial autocorrelation during the COVID-19 pandemic', *Sci. Rep.*, vol. 13, no. 1, p. 1015, Jan. 2023, doi: 10.1038/s41598-023-28287-8.

[8] H. H. Satyanegara and K. Ramli, 'Implementation of CNN-MLP and CNN-LSTM for MitM Attack Detection System', *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 6, no. 3, pp. 387–396, Jun. 2022, doi: 10.29207/resti.v6i3.4035.

[9] N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, 'Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis', *J. Environ. Public Health*, vol. 2023, pp. 1–26, Jan. 2023, doi: 10.1155/2023/4916267.

[10] Y. Qiu *et al.*, 'Regional aerosol forecasts based on deep learning and numerical weather prediction', *Npj Clim. Atmospheric Sci.*, vol. 6, no. 1, p. 71, Jun. 2023, doi: 10.1038/s41612-023-00397-0.

[11] R. Merritt, 'What Is a Transformer Model?', NVIDIA Blog. Accessed: Feb. 22, 2024. [Online]. Available: https://blogs.nvidia.com/blog/what-is-a-transformer-model/