NZSE

School of Tech

Graduate Diploma in Data Analytics (Level 7)

Cover Sheet and Student Declaration

Course Title:	**Capstone Project (DA)**	**Course code:**	**GDDA713**
Student Name:	Aju Peter	**Student ID:**	764706847
Assessment No & Type:	Assessment 1[Project Proposal]	**Cohort:**	
Due Date:	10-05-2024	**Date Submitted:**	09-05-2024
Tutor's Name:	Mohammad Norouzifard		
Assessment Weighting	25%		
Total Marks	100		

Student Declaration:

I declare that:

- I have read the New Zealand School of Education Ltd policies and regulations on assessments and understand what plagiarism is.
- I am aware of the penalties for cheating and plagiarism as laid down by the New Zealand School of Education Ltd.
- This is an original assessment and is entirely my own work.
- Where I have quoted or made use of the ideas of other writers, I have acknowledged the source.
- This assessment has been prepared exclusively for this course and has not been or will not be submitted as assessed work in any other course.
- It has been explained to me that this assessment may be used by NZSE Ltd, for internal and/or external moderation.

Student signature: Aju Peter

Date: 09-05-2024

Tutor only to complete		
Assessment results:	**All Tasks except formatting /95 marks**	Proposal formatting **/5 marks**
	Total Marks: /100	

PM_{10} LEVEL FORECASTING USING MACHINE LEARNING IN PENROSE, AUCKLAND

A PROJECT PROPOSAL SUBMITTED TO NEW ZEALAND SCHOOL OF EDUCATION
IN FULFILMENT OF THE REQUIREMENTS FOR THE CAPSTONE PROJECT

GRADUATE DIPLOMA IN DATA ANALYTICS

SUPERVISORS

Dr. SARA ZANDI (NZSE)

LOUIS BOAMPONSEM (AUCKLAND COUNCIL)

May 2024

BY

AJU PETER

Table of Contents

List of Abbreviations

Ambient Temperature (AT)

Artificial Neural Networks (ANN)

Autoregressive Integrated Moving Average (ARIMA)

Bi-Directional Long Short-Term Memory (Bi-LSTM)

Chemical Transport Model (CTM)

Community Multi-Scale Air Quality (CMAQ)

Empirical Mode Decomposition (EMD)

Exploratory Data Analysis (EDA)

Fully Connected LSTM (FC-LSTM)

Grouped Pollutant LSTM (GP-LSTM)

Improved Complete Ensemble EMD with Adaptive Noise (ICEEMDAN)

Individual Group Pollutant LSTM (IGP-LSTM)

Light Gradient Boosting (LGB)

Local Data Assimilation and Prediction System (LDAPS)

Long Short-Term Memory networks (LSTM)

Mean Absolute Error (MAE)

Mean Absolute Relative Error (MARE)

Multilayer Perceptron (MLP)

Nash-Sutcliffe Efficiency (NSE)

Online Sequential Extreme Learning Machines (OS-ELM)

Particulate Matter with a diameter less than 10 micrometres (PM_{10})

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)

Rain Fall (RF)

Random Forest (RF)

Recurrent Neural Networks (RNNs)

Regression coefficient (R2)

Relative Humidity (RH)

Root Mean Squared Error (RMSE)

Self-Organizing Maps (SOMs)

Single Pollutant LSTM (SP-LSTM)

Solar Radiation (SR)

Support Vector Machines (SVM)

Vector Autoregressive Moving-Average (VARMA)

Wind Direction (WD)

Wind Speed WS

World Health Organization (WHO)

Abstract

Air pollution is one of the most dangerous threats to human health and a major cause of climate change. It is primarily caused by increasing industrial emissions, transportation, home heating, and other forms of fuel consumption. For this reason, predicting air pollution is extremely important. Air quality monitoring stations collect a large amount and variety of data, which has made air pollution forecasting a popular area of research. This extensive data collection helps in developing models that can predict and reduce the effects of air pollution.

This project aims to forecast air pollution in Auckland, particularly focusing on PM_{10} particulate matter in the Penrose area, a location significantly affected by industrial emissions, residential heating, and heavy traffic. This study proposes to develop a machine learning model that predicts PM_{10} concentrations a day ahead, thus supporting timely decision-making for pollution management. The methodology involves collecting historical air quality and meteorological data, followed by preprocessing, exploratory data analysis, and model training using advanced Machine Learning algorithms such as ANN, LSTM, RF, and SVM. Each model's performance will be evaluated based on RMSE and MAE to determine the most effective predictive approach.

By creating the most effective model, we will package it into a simple 'pip install' module. This will allow users to easily connect to and utilize live data from the Auckland Council's API, enabling the model to continuously update based on the latest air quality data. This approach ensures that our predictive model remains relevant and practical for real-world applications. This project not only seeks to enhance air quality management in Penrose but also provides a framework that could be adapted to other regions.

Keywords: Auckland Air Quality, PM_{10} Timeseries, PM_{10} Estimation, Machine Learning, Python

Introduction

Air pollution is a significant environmental health issue globally, leading to about seven million deaths annually according to the WHO. In Auckland, air pollution is linked to over 300 premature deaths annually, increasing medication use, hospital visits, and reducing active days for its citizens. The projected societal cost of air pollution in Auckland is approximately $1.07 billion annually. Despite Auckland's geographic advantage, located between the Tasman Sea and South Pacific Ocean which aids in maintaining cleaner air, local sources such as transportation, residential heating, and industrial activities often push pollution levels beyond safe standards (Boamponsem, n.d.).

I am working on the "Auckland Air Quality Forecasting" project in collaboration with the Auckland Council. This project aims to develop a dedicated air quality forecasting model for Auckland using machine learning algorithms. Given the serious health impacts of air pollution, precise and timely air quality forecasts are essential for effective decision-making and implementing pollution mitigation strategies. I will analyse historical air quality data and meteorological information, among other relevant factors, using data provided directly by the Auckland Council (Excel file) and supplemented by additional data from the Council's official website and NIWA, the National Institute of Water and Atmospheric Research. This rich combination of datasets will ensure a comprehensive understanding of both historical and current air quality conditions, which is crucial for developing an accurate forecasting model.

This project includes Section I: Problem definition - This section outlines specific real industry questions that the project aims to address, focusing on how machine learning can enhance air quality forecasting and its implications for environmental policy in Auckland. Section II: Literature Review - This section presents a detailed analysis of seven recent and relevant journal papers to provide a foundational understanding of the current advancements and methodologies in air quality prediction using machine learning. Section III: Discussion - This section explores potential machine learning solutions and data processing steps that will be employed to develop the air quality prediction model. Project Planning and Timeline Management - This section outlines the project planning process, including the use of the Monday.com. Methodology - This section details the selection of the four best predictive machine learning algorithms identified from the literature review. Tools Used for Project Development - This section includes a comprehensive list of all the tools used throughout the project, from start to finish. Section IV: Conclusion - This final section will summarize the key findings of the project, the efficient machine learning models identified. It will also suggest potential areas for further research and development to enhance future air quality prediction models.

I. Problem definition

Auckland air quality data becomes larger and more complex, it's getting harder to analyse. Currently, there is no specialized tool for predicting air quality, causing significant challenges for the Auckland Council in making informed decisions to manage pollution. This project aims to develop an advanced machine learning tool specifically designed to forecast air quality levels in Auckland. This tool will utilize historical air quality data, meteorological information, and possibly other relevant variables to train and validate machine learning algorithms capable of predicting air pollutant concentrations accurately and timely. The development of a predictive tool is crucial for implementing effective decision-making and mitigation strategies to address the growing concerns about air pollution and its harmful effects on public health. By applying this tool to real-time data, the project will enable the Auckland Council to enhance its pollution management practices and better protect public health in the region.

Scope of the project:

Developing a machine learning tool for forecasting air quality across Auckland is a huge project, Auckland has ten different monitoring stations, each tracking a wide variety of pollutants including PM_{10}, PM2.5, toxic gases (NO2, CO, O3, SO2), and black carbon. Given the project's 16-week time constraint and following the advice of my internal supervisor, Dr. Sara Zandi, I have decided to specifically focus on PM_{10}, a major pollutant, in the Penrose area. This part of Auckland is notably impacted by industrial emissions, residential heating, and heavy traffic, all of which contribute to high PM10 levels. PM_{10} particles are particularly hazardous as they can deeply penetrate the respiratory system, potentially leading to serious respiratory and cardiovascular diseases. Vulnerable groups such as individuals with existing respiratory and heart conditions, those with diabetes, as well as the young and elderly, are at an increased risk from the detrimental effects of air pollution. By focusing on PM_{10} in Penrose, this project allows for the effective application of advanced machine learning techniques, thereby making a substantial contribution to improving public health in one of Auckland's most pollution-affected areas.

II. Literature review

Air pollution, particularly PM with a diameter less than 10 micrometers, poses significant health risks and environmental challenges. Forecasting PM_{10} concentrations is crucial for implementing effective air quality management strategies. This literature review explores how machine learning methods, particularly their ability to process complex data, are used to improve air quality assessments. We focus on recent studies that demonstrate the effectiveness of these technologies in monitoring and predicting air quality trends.

I have used the PRISMA flow diagram (See Figure 1) to guide the literature review process, ensuring a systematic approach to reviewing research articles. For my search, I selected five well-known publishers: Elsevier, EBSCO, ACM, Springer, and IEEE, to find advanced research papers related to air quality prediction, especially those focusing on predicting PM10 pollutants in recent years. Since, air quality is a hot research topic in globally, I was able to successfully gather a significant number of PM10-related research reports, which will substantially aid in this capstone project.

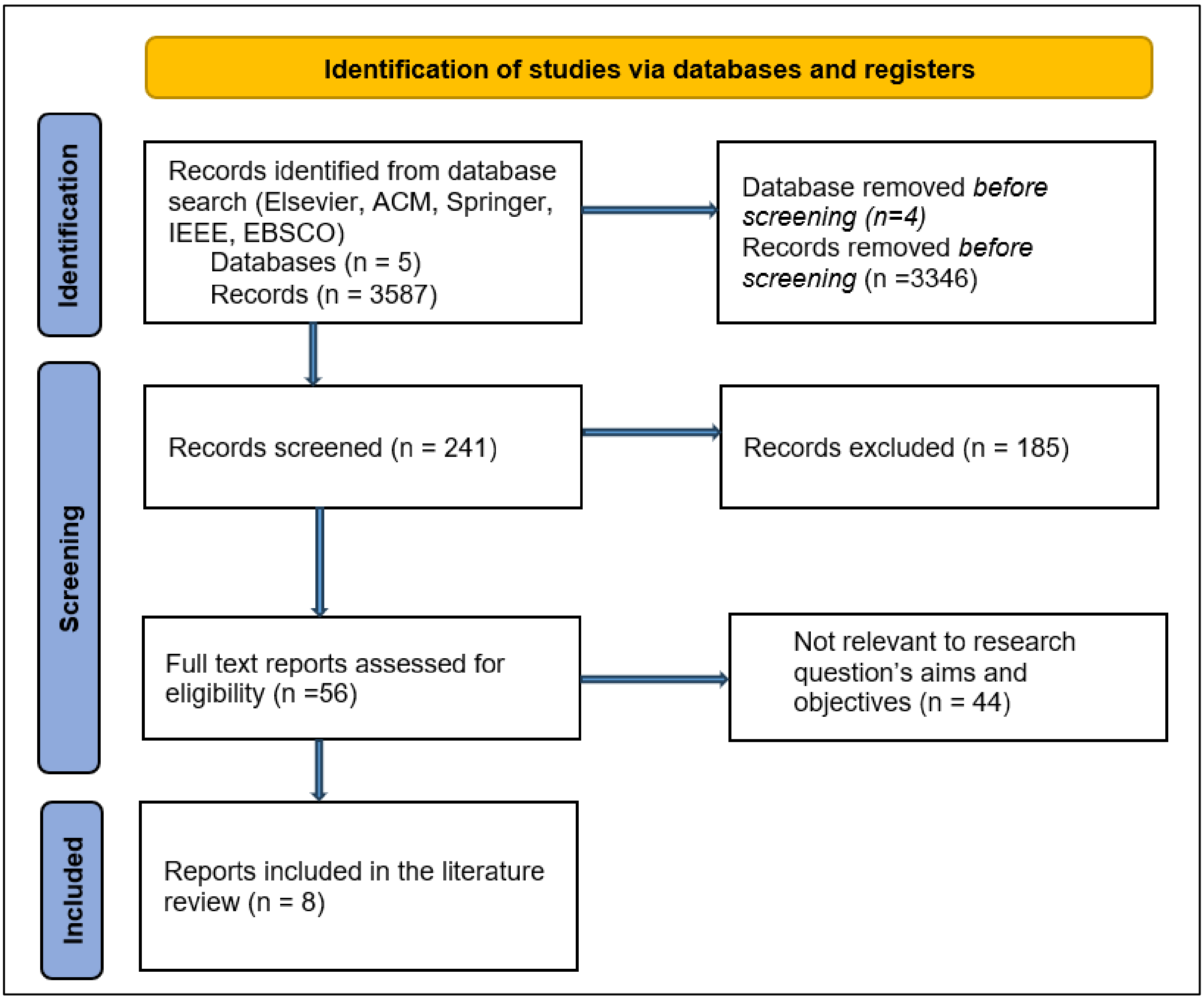

Identification of studies via databases and registers
Identification
Screening
Included
Records identified from database search (Elsevier, ACM, Springer, IEEE, EBSCO)
Databases (n = 5)
Records (n = 3587)
Database removed *before screening (n=4)*
Records removed *before screening* (n =3346)
Records screened (n = 241)
Records excluded (n = 185)
Full text reports assessed for eligibility (n =56)
Not relevant to research question's aims and objectives (n = 44)
Reports included in the literature review (n = 8)

Figure 1.Prisma Chart

Gualtieri aimed to compare the effectiveness of linear models and ANN in forecasting PM_{10} concentrations in Brescia, Italy. Traditional statistical techniques were utilized in the linear model to predict PM_{10} concentrations from historical data, while the ANN model applied machine learning to capture more complex non-linear relationships in the data. The findings indicated that the ANN model generally outperformed the linear model, showcasing its superior ability in handling complex atmospheric datasets. The SOMs showed that both models predictions exhibit the same clustering as the observations, However, it is noted that the ANN, while generally effective, tended to underestimate the highest clustered PM_{10} concentrations. Specifically, the worst case of underestimation by the ANN was by 5.8 µg/m^3. This means that in the instances where PM_{10} levels peaked, the ANN model's predictions were, at most, 5.8 µg/m^3 lower than the actual observed value (Gualtieri et al., 2018).

Garcia Nieto evaluated the effectiveness of four different models namely SVM, MLP, VARMA, and ARIMA, in predicting PM_{10} concentration in Northern, Spain. The SVM, especially the RBF-SVM, was highlighted for its robust handling of non-linear data, proving more accurate than the others. MLP was also effective but less precise, while VARMA and ARIMA models showed good performance with ARIMA noted for its simplicity and efficiency in handling non-stationary data. The study concluded that SVMs, particularly RBF-SVM, are superior for environmental predictions due to their ability to manage complex patterns (García Nieto et al., 2018).

Kurnaz and Demir utilized RNNs to model and predict SO_2 and PM_{10} levels from data collected during the COVID-19 pandemic period. RNNs were chosen for their capability to capture temporal dependencies, which are crucial in environmental monitoring. RNN are shown to be superior for capturing temporal dependencies in air quality data compared to other models like SVMs and traditional neural networks, especially in handling time series data relevant to environmental monitoring. The study highlighted challenges such as data quality and training complexity but also emphasized the significant potential of RNNs in improving air quality monitoring and management in industrial regions. The models achieved high accuracy, demonstrating the effectiveness of RNNs in supporting environmental policy-making and public health initiatives (Kurnaz & Demir, 2022) .

Ke was developed an automated air quality forecasting system and applied in China, incorporating multiple machine learning models (MLR, MLP, RF, GBDT, SVR) and an ensemble model. This system was designed to automatically find the best model and hyperparameters without human intervention by using random parameter CV or grid search CV. supported by a knowledge base containing the meteorological observed data, pollutant concentrations, pollutant emissions, and model reanalysis data, utilizing a comprehensive dataset from 2015 to 2019 across several major cities. The system's performance, evaluated against seven criteria and pollution level forecasts, combined with the forecasting results for the next 3-days, indicated that it could deliver forecasts surpassing most traditional numerical models, showing a promising application prospect in environmental meteorology (Ke et al., 2022).

Navares and Aznarte studied on air quality prediction in Madrid utilized LSTM networks to forecast various pollutants. It compared different LSTM configurations: GP-LSTM grouped by pollutant class and IGP-LSTM with individual groups of pollutants as inputs. The findings suggested that LSTMs, especially GP-LSTM, provided more accurate forecasts with smaller biases. GP-LSTM outperforms not only traditional models like Linear Regression and Random Forest but also other LSTM configurations including SP-LSTM (Single Pollutant LSTM), suggesting that the method of grouping pollutants enhances the predictive accuracy specifically for PM_{10}. The research underscored the potential of integrating LSTM networks into air quality forecasting, which could become a critical tool in environmental management and public health as data and computational methods improve (Navares & Aznarte, 2020).

Sharma et al. developed a hybrid AI framework that combined OS-ELM with EMD algorithms to predict hourly air quality levels across various Australian cities. The integration of these advanced AI techniques addressed the non-linear and non-stationary nature of air quality data effectively. The study found that the hybrid model, ICEEMDAN coupled with OS-ELM, outperformed traditional models, providing accurate and reliable forecasts crucial for public health advisories and environmental management. This approach highlighted the significance of hybrid AI models in tackling the complexities of atmospheric data (Sharma et al., 2020).

Lakindu Mampitiya, Namal Rathnayake, Yukinobu Hoshino, and Upaka Rathnayake focused on forecasting PM_{10} concentrations using machine learning models in Sri Lanka. Their study conducted a comparative analysis of eight machine learning models (ANN, Bi-LSTM, Ensemble, XGBoost, CatBoost, LightGBM, LSTM, and GRU). The findings revealed that an ensemble model, integrating state-of-the-art methodologies, outperformed the other seven models. The machine learning approaches considered air quality and meteorological factors, including O_3, CO, NO_2, SO_2, $PM_{2.5}$, AT, RH, SR, rainfall (RF), WS, and WD. Min-Max Normalization was applied to scale the dataset to a predefined range of 0 to 1. The results demonstrated that ensemble models, which combine multiple predictive techniques, can perform better than individual models such as ANN and Bi-LSTM networks, particularly in forecasting PM_{10} levels (Mampitiya et al., 2024).

Kim, Lim, and Cha conducted a recent study in Seoul, South Korea, researchers employed tree-based machine learning algorithms to predict short-term concentrations of particulate matter (PM_{10} and $PM_{2.5}$). Utilizing meteorological data from the LDAPS and PM data from 40 observation stations, the study compared the performance of the LGB algorithm against the CMAQ based CTM and ADAM model. Meteorological inputs included temperature, dew point, wind speed, and other relevant variables. The LGB model demonstrated superior performance, evidenced by lower bias and RMSE values, and higher R^2 values in predicting PM concentrations, showcasing its potential over traditional CTM models, particularly in handling dynamic urban pollution scenarios. This highlights LGB's applicability in environmental policy and public health monitoring due to its high accuracy and computational efficiency (Kim et al., 2022).

III. Discussion

1. Proposed Solutions for PM_{10} Prediction in Penrose

This project, I am working on creating a machine learning model that can predict daily PM_{10} pollution levels in Penrose accurately. The aim is to provide forecasts a day in advance to help the community and local authorities prepare for and manage air quality better.

The project will have two main deliverables. By week 15, I plan to have the machine learning model ready and packaged as a pip-installable application. This package will use real-time data from the Auckland Council API to predict the average 24-hour PM_{10} concentration, making it easier for local environmental agencies and health departments to use.

If time persist, the next step would be to develop a web platform by week 16. This platform would integrate the model with the Auckland Council API for live updates and offer a more interactive way for the public to view and understand air quality data.

However, if we run into time constraints and can't finish the web platform, we will conclude the project with the pip-installable package. Looking ahead, we could consider adding a web platform interface and possibly live streaming features as future developments. This plan ensures we still provide a useful tool for monitoring air pollution and lays the groundwork for potential enhancements.

Ultimately, the goal of this project is to support informed decision-making, contributing to better management and mitigation of air pollution in the Penrose area. To achieve this, I reviewed recent scientific literatures and identified four machine learning algorithms known for their effectiveness in air quality forecasting. These include:

- ANN- Known for their ability to model complex non-linear relationships between inputs and outputs.
- LSTM - A type of recurrent neural network particularly suited for sequence prediction problems like time-series forecasting.
- RF - An ensemble learning method that is robust to overfitting and effective at handling large datasets with multiple input variables.
- SVM - Effective in high-dimensional spaces and capable of performing both classification and regression tasks.

These models were chosen for their proven success in environmental studies, suggesting they could accurately predict PM_{10} levels in Penrose. Moving forward, I will implement these algorithms to analyze local PM_{10} data and evaluate each model's performance using metrics like RMSE, and MAE. This evaluation will help identify the most suitable machine learning model to predict air pollution and enhance public health efforts in Penrose.

2. Project Management

This table (Table 1) the 16-week timeline for my capstone project, "Air Quality Prediction using Machine Learning," in collaboration with Auckland Council. It details weekly tasks from the initial setup and data collection to final model refinement and presentation.

Table 1.Time and Project Management.

Week	Task Description
1-3	• Selected "Air Quality Prediction using Machine Learning" from Auckland Council as my project. • Chose Dr. Sara as my internal supervisor; had our initial meeting with Mr. Louis Boamponsem from Auckland Council, who introduced the project details. • Began collecting and reviewing recent research papers relevant to this project. • Weekly meeting with Dr. Sara on Mondays
4	• Met again with Dr. Sara, received historical sensor data from Auckland Council. • Decided to focus on predicting PM_{10} values in Penrose due to the 16-week time constraint for the project. • Started using the PRISMA chart for organizing the literature review. • Weekly meeting with Dr. Sara on Mondays
5	• Currently fine-tuning the project proposal with Dr. Sara; plan to submit the final draft by the end of this week. • Began using MONDAY.com for project planning and time management. • Weekly meeting with Dr. Sara on Mondays
6	• Upcoming: Finalize and submit the project proposal, including a video presentation. • Weekly meeting with Dr. Sara on Mondays
7	• Data Collection: Gather meteorology and transport data from the Auckland Council website and NIWA. • Weekly meeting with Dr. Sara on Mondays
8	• Data Preprocessing: Clean, transform, engineer features, integrate, and encode data using Python and its libraries in Jupyter Notebook. • Weekly meeting with Dr. Sara on Mondays
9	• Exploratory Data Analysis: Analyze data to identify trends, patterns, and anomalies. Create visualizations to help recognize patterns or issues using Python. • Weekly meeting with Dr. Sara on Mondays
10	• Skill Development: Learn new ML algorithms (ANN, LSTM) and review others studied last semester. • Weekly meeting with Dr. Sara on Mondays
11	• Apply Machine Learning: Engage in feature engineering, model selection, and initial training of different machine learning models. • Weekly meeting with Dr. Sara on Mondays
12-13	• Model Refinement: Fine-tune the models, adjust parameters, and utilize cross-validation to ensure effectiveness and avoid overfitting. • Weekly meeting with Dr. Sara on Mondays
14	• Evaluation and Testing: Assess the model using RMSE, MAE, if possible, to check real-time applicability. • Weekly meeting with Dr. Sara on Mondays
15	• Finalization and Report Submission: Make final adjustments to the model, compile findings, methodologies, and evaluations into a final report. • Submission of final model pip installation package to Auckland Council for real-time prediction verification. • Weekly meeting with Dr. Sara on Mondays

16	• Presentation: Prepare and deliver a presentation summarizing your project findings, methodology, results, and implications. • If time persist, we need to develop a user-friendly web platform that integrate Auckland Council API for real-time prediction. • Weekly meeting with Dr. Sara on Mondays

3. Methodology

Data Collection:

The data for this project includes historical PM_{10} 24-hour average concentration levels obtained from the Auckland Council for the years 2014 to 2023. To enhance the accuracy of our predictions, additional data on meteorological conditions and traffic volumes will be collected. This supplementary data collection is planned post-approval of this project proposal.

Machine Learning Models:

The project will utilize advanced machine learning algorithms, including ANN, LSTM, RF, and SVM. These models were chosen based on their frequent mention in recent studies for their effectiveness in similar applications. The performance of these models will be compared to identify the most suitable algorithm for predicting PM_{10} levels in the Penrose area.

Development of Technical Skills:

An integral part of this project involves the development of technical skills, particularly in implementing newer algorithms like ANN and LSTM.

4. Tools required in project development.

In the development of this project, I have utilized and plan to continue using a variety of tools that support different aspects of the research and implementation process. These tools have been chosen based on their functionality with the project's requirements.

Documentation and Presentation:

- Microsoft Word: Utilized for drafting and formatting the project proposal.
- Microsoft PowerPoint: Used for creating visually engaging presentations to communicate project plans and findings effectively.
- Microsoft Excel: Employed for tracking project status and updates in a structured format.

Research and Literature Review:

- Elsevier & EBSCO Database: A primary source for accessing scientific research papers relevant to air quality prediction and machine learning applications.
- PRISMA: This framework guides the literature review process, ensuring a systematic approach to reviewing research articles.
- Zotero: An essential tool for managing and organizing research papers. It also facilitates proper citation and referencing, which are crucial for academic integrity.

Project Planning and Management:

- MONDAY.com: This project management tool helps in planning, organizing, and tracking progress through a Gantt chart, enabling effective time management and ensuring project milestones are met on schedule.

Data Analysis and Model Development:

- Python: The programming language used for EDA and machine learning model development due to its robust libraries and community support.
- Jupyter Notebook: The IDE for writing, testing, and presenting Python code.

These tools support from initial research to data analysis, ensuring the project is organized and maintains high standards of quality.

IV. Conclusions

This project proposal outlines the development of a machine learning model to predict daily PM_{10} pollution levels in Auckland's Penrose area, with the aim of assessing its potential to improve public health by reducing air pollution. The approach will involve evaluating four advanced machine learning algorithms: ANN, LSTM, RF, and SVM. These algorithms were selected based on their proven efficacy in environmental studies and their potential to accurately forecast pollution levels. The model development process, including data preparation and analysis, will be carried out using Python, which is well-known language in data science for its extensive libraries and support. After a thorough evaluation of each model's performance, the best-performing model will be identified.

Our initial deliverable by week 15 will be a pip-installable package that utilizes real-time data from the Auckland Council API to accurately predict daily PM_{10} concentrations. This will provide local environmental agencies and health departments with a crucial tool for air quality management. If time persist, we aim to further expand this project by developing an interactive web platform by week 16. This platform will integrate with the model and the Auckland Council API for live updates, offering the public an engaging and informative way to track air quality. However, if time constraints prevent the completion of the web platform, the project will still achieve its core objective with the pip-installable package.

Moreover, the flexible design of our machine learning model could be scaled up and applied to other major areas. We could also expand the model to track more types of air pollutants, providing a complete picture of Auckland's air quality. This expansion would not only extend our monitoring reach but also improve the data we collect, which can lead to better environmental policies and health advice. Overall, this project isn't just about tackling current air quality issues; it's about building groundwork for ongoing improvements in environmental monitoring.

V. References

Boamponsem, L. (n.d.). *Auckland air quality – 2021 annual data report*.

García Nieto, P. J., Sánchez Lasheras, F., García-Gonzalo, E., & De Cos Juez, F. J. (2018). PM10 concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: A case study. *Science of The Total Environment*, *621*, 753–761. https://doi.org/10.1016/j.scitotenv.2017.11.291

Gualtieri, G., Carotenuto, F., Finardi, S., Tartaglia, M., Toscano, P., & Gioli, B. (2018). Forecasting PM10 hourly concentrations in northern Italy: Insights on models performance and PM10 drivers through self-organizing maps. *Atmospheric Pollution Research*, *9*(6), 1204–1213. https://doi.org/10.1016/j.apr.2018.05.006

Ke, H., Gong, S., He, J., Zhang, L., Cui, B., Wang, Y., Mo, J., Zhou, Y., & Zhang, H. (2022). Development and application of an automated air quality forecasting system based on machine learning. *Science of The Total Environment*, *806*, 151204. https://doi.org/10.1016/j.scitotenv.2021.151204

Kim, B.-Y., Lim, Y.-K., & Cha, J. W. (2022). Short-term prediction of particulate matter (PM10 and PM2.5) in Seoul, South Korea using tree-based machine learning algorithms. *Atmospheric Pollution Research*, *13*(10), 101547. https://doi.org/10.1016/j.apr.2022.101547

Kurnaz, G., & Demir, A. S. (2022). Prediction of SO2 and PM10 air pollutants using a deep learning-based recurrent neural network: Case of industrial city Sakarya. *Urban Climate*, *41*, 101051. https://doi.org/10.1016/j.uclim.2021.101051

Mampitiya, L., Rathnayake, N., Hoshino, Y., & Rathnayake, U. (2024). Performance of machine learning models to forecast PM10 levels. *MethodsX*, *12*, 102557. https://doi.org/10.1016/j.mex.2024.102557

Navares, R., & Aznarte, J. L. (2020). Predicting air quality with deep learning LSTM: Towards comprehensive models. *Ecological Informatics*, *55*, 101019. https://doi.org/10.1016/j.ecoinf.2019.101019

Sharma, E., Deo, R. C., Prasad, R., & Parisi, A. V. (2020). A hybrid air quality early-warning framework: An hourly forecasting model with online sequential extreme learning machines and empirical mode decomposition algorithms. *Science of The Total Environment*, *709*, 135934. https://doi.org/10.1016/j.scitotenv.2019.135934

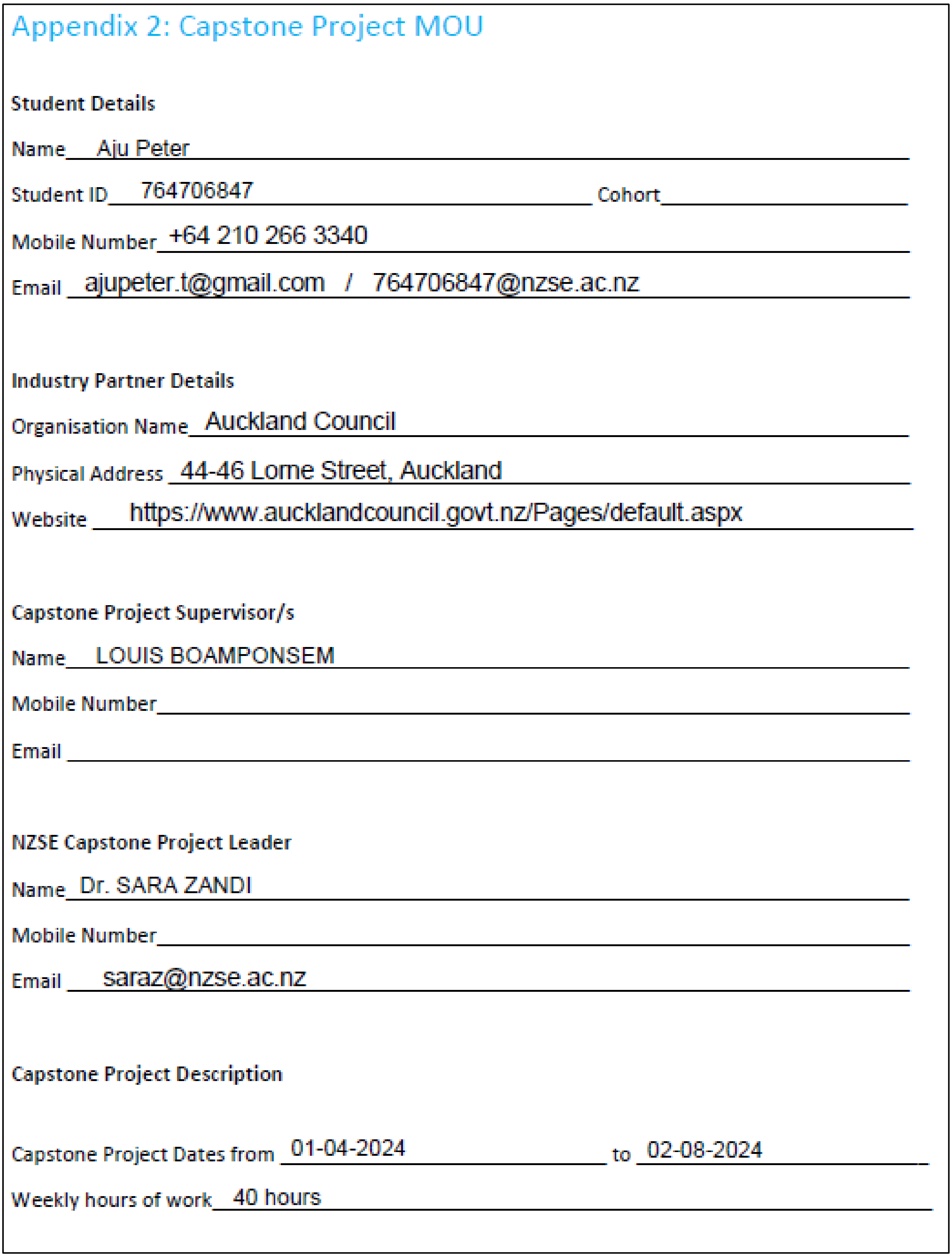
Appendix 1: Literature Review Summary

The primary aim of this project is to develop an accurate machine learning model that can forecast a day ahead concentrations of PM_{10} pollutants in the Penrose area. To find out the best predicting machine learning algorithm, I reffered recent scientific research papers and identified four best performed machine learning algorithms and its evaluation metrics in air quality prediction (See Table 2). It includes ANN, LSTM, RF, and SVM. These models were chosen for their proven success in environmental studies, suggesting they could accurately predict PM_{10} concentration.

Table 2.Research outputs

Sl No.	Research Name	Models Used	Type of Data used	Evaluation Method	Best performed model
1	Forecasting PM_{10} hourly concentrations in northern Italy: Insights on models' performance and PM_{10} drivers through self-organizing maps	Artificial Neural Network (ANN)	Air quality Environmental data SOMs		ANN
2	PM_{10} concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA	SVM, MLP, VARMA, and ARIMA	Air quality Environmental data		RBF-SVM
3	Prediction of SO_2 and PM_{10} air pollutants using a deep learning-based recurrent neural network: Case of industrial city Sakarya	RNNs	Air quality Environmental data	RMSE	RNN
4	Development and application of an automated air quality forecasting system based on machine learning	MLR, MLP, RF, GBDT, SVR	Air quality Environmental data		Ensemble model
5	A hybrid air quality early-warning framework: An hourly forecasting model with online sequential extreme learning machines and empirical mode decomposition algorithms	OS-ELM: ICEEMDAN: Feature Extraction and Processing: Model Training and Validation	Air Quality Variables: Geographical Scope	RMSE, MAE	
6	Predicting air quality with deep learning LSTM: Towards comprehensive models	FC-LSTM GP-LSTM IGP-LSTM SP-LSTM	Air Pollutants Biotic Factors Meteorological Data	RMSE	GP-LSTM
7	Performance of machine learning models to forecast PM_{10} levels	$(LSTM) < (XGBoost) < (GRU) < (Bi-LSTM) < (ANN) < (CatBoost) < (LightGBM) < (Ensemble)$	O3, CO, NO2, SO2, PM2.5, AT, RH, SR, RF-rainfall, WS, WD	R^2, RMSE, MSE, MAE, MARE, and NSE	$Ensemble$
8	Short-term prediction of particulate matter (PM_{10} and $PM_{2.5}$) in Seoul	LGB LDAPS CMAQ CTM	O_3, CO, NO_2, SO_2, $PM_{2.5}$, AT, RH, SR, RF, WS, WD	RMSE, R^2,	LGB

Appendix 2: Capstone Project MOU



Appendix 2: Capstone Project MOU

Student Details

Name Aju Peter

Student ID 764706847 Cohort

Mobile Number +64 210 266 3340

Email ajupeter.t@gmail.com / 764706847@nzse.ac.nz

Industry Partner Details

Organisation Name Auckland Council

Physical Address 44-46 Lorne Street, Auckland

Website https://www.aucklandcouncil.govt.nz/Pages/default.aspx

Capstone Project Supervisor/s

Name LOUIS BOAMPONSEM

Mobile Number

Email

NZSE Capstone Project Leader

Name Dr. SARA ZANDI

Mobile Number

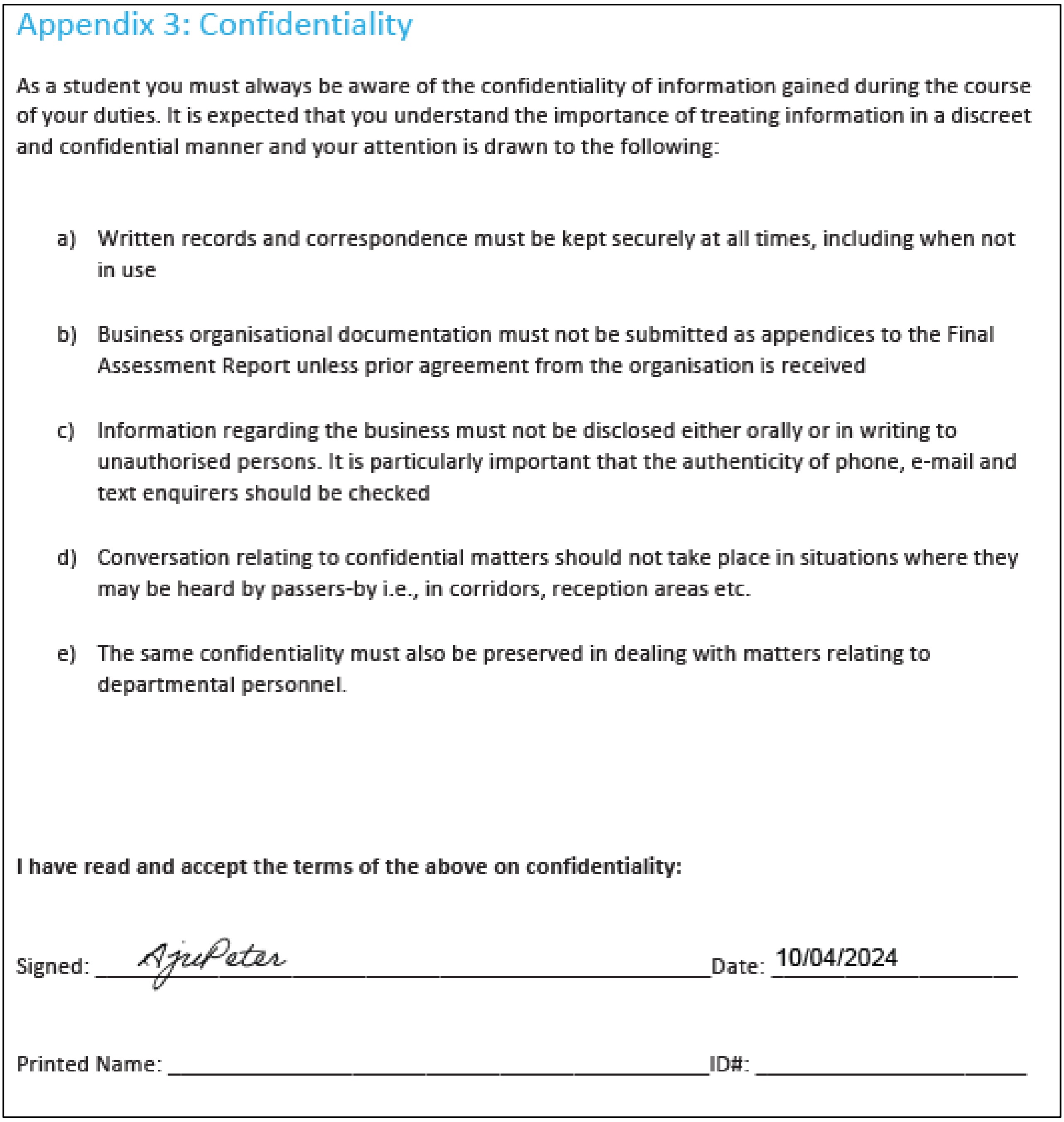
Email saraz@nzse.ac.nz

Capstone Project Description

Capstone Project Dates from 01-04-2024 to 02-08-2024

Weekly hours of work 40 hours

Appendix 3: Confidentiality Acceptance



Appendix 3: Confidentiality

As a student you must always be aware of the confidentiality of information gained during the course of your duties. It is expected that you understand the importance of treating information in a discreet and confidential manner and your attention is drawn to the following:

a) Written records and correspondence must be kept securely at all times, including when not in use

b) Business organisational documentation must not be submitted as appendices to the Final Assessment Report unless prior agreement from the organisation is received

c) Information regarding the business must not be disclosed either orally or in writing to unauthorised persons. It is particularly important that the authenticity of phone, e-mail and text enquirers should be checked

d) Conversation relating to confidential matters should not take place in situations where they may be heard by passers-by i.e., in corridors, reception areas etc.

e) The same confidentiality must also be preserved in dealing with matters relating to departmental personnel.

I have read and accept the terms of the above on confidentiality:

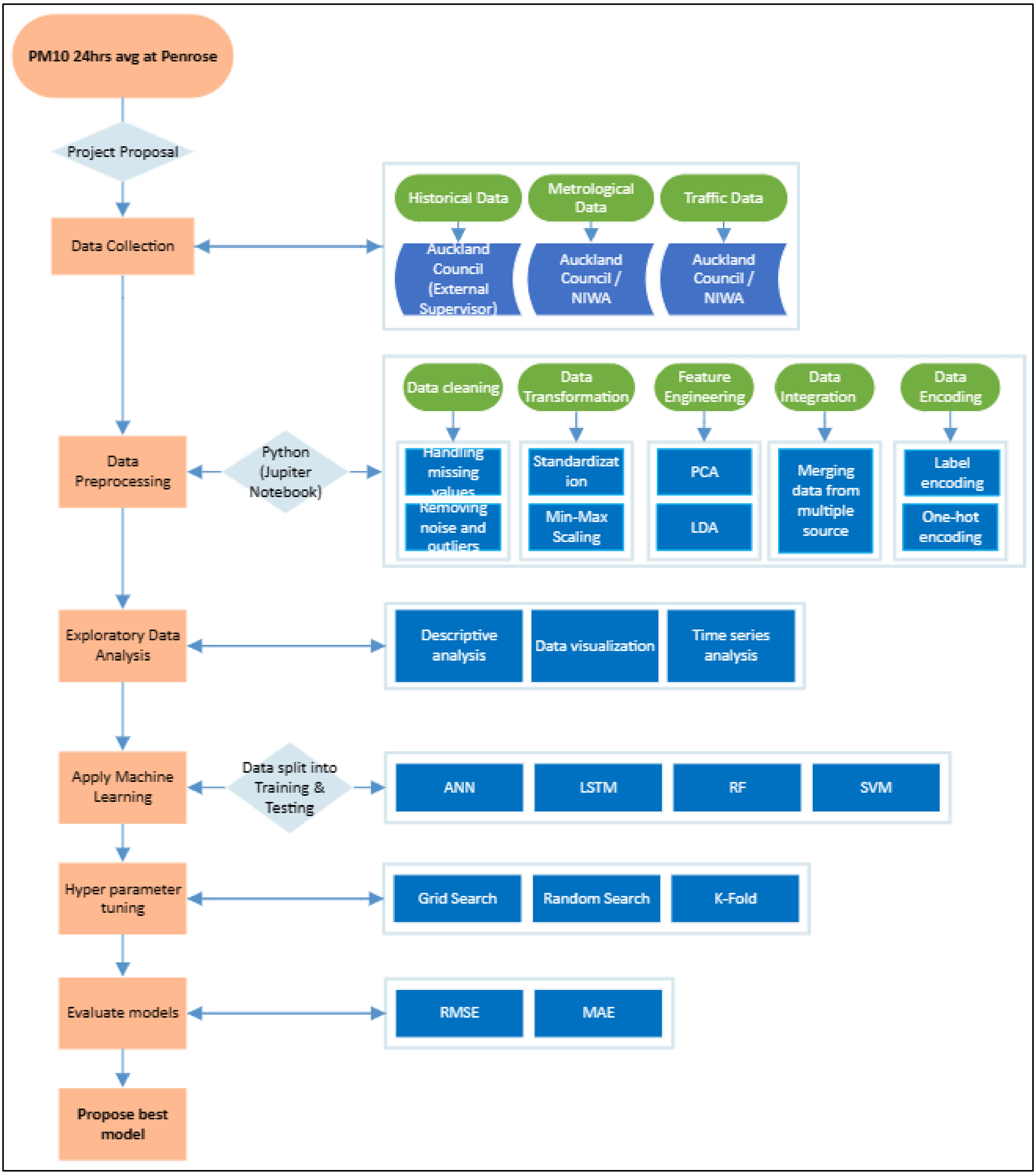
Signed: AjuPeter ______________________________ Date: 10/04/2024 ____________

Printed Name: ______________________________ ID#: ______________________

Appendix 4: Complete Project Life Cycle

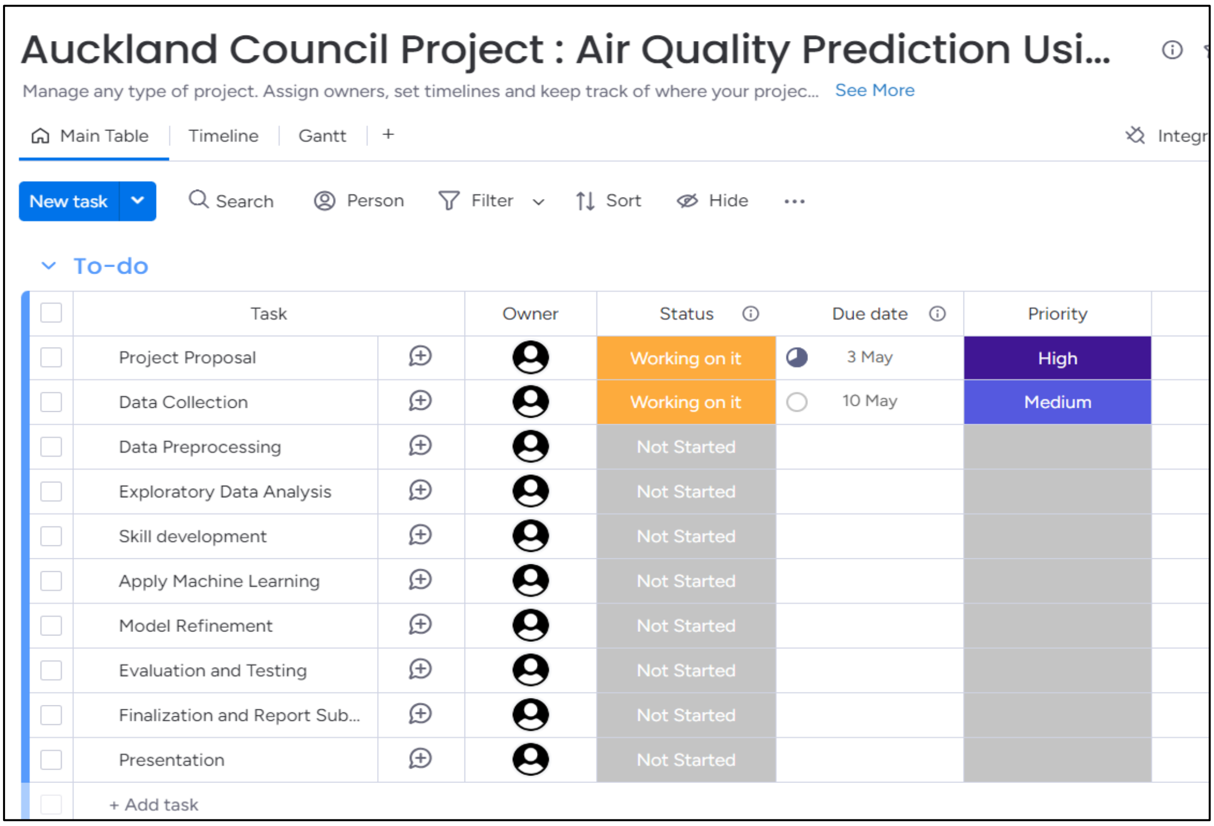
This project aims to predict the 24-hour average concentration of PM_{10} in Auckland's Penrose area using machine learning. The project life cycle includes planning, data collection, preprocessing, analysis, applying machine learning models and selecting the best ML algorithm. (See Figure 1.)

Figure 2.Project Life Cycle for PM_{10} 24hrs Average Concentration Prediction Using Machine Learning



Appendix 5: Project Management Tool (MONDAY.com)

I am utilizing MONDAY.com (see Figure 2), a well-known project management tool, to plan, organize, and track my project's progress using a Gantt chart. This helps in effective time management and ensures that project milestones are met according to the schedule. Additionally, I have invited both my internal and external supervisors to join this platform. This allows them to stay updated on the project's progress and contribute their valuable insights or make necessary adjustments as needed.



Auckland Council Project : Air Quality Prediction Usi...

Manage any type of project. Assign owners, set timelines and keep track of where your projec... See More

To-do

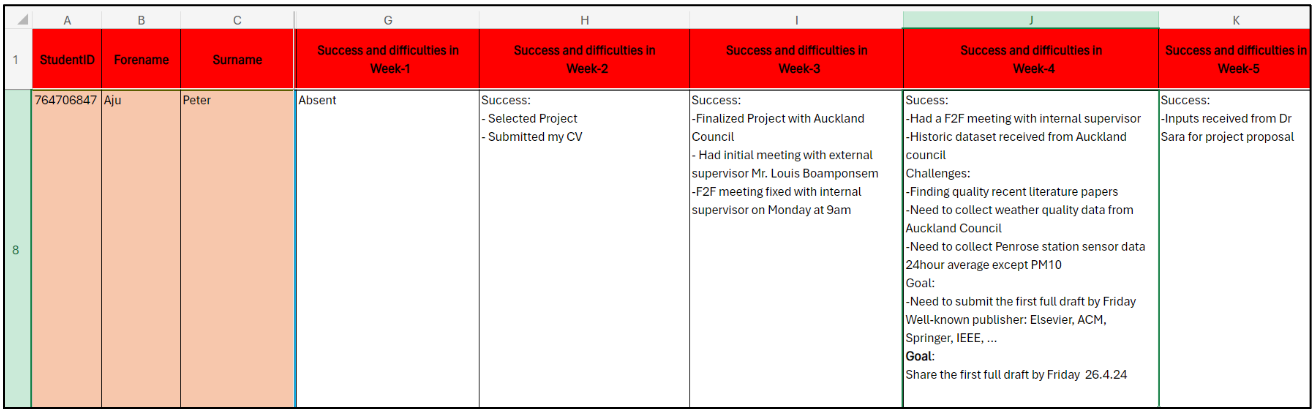
	Task	Owner	Status	Due date	Priority
□	Project Proposal		Working on it	3 May	High
□	Data Collection		Working on it	10 May	Medium
□	Data Preprocessing		Not Started		
□	Exploratory Data Analysis		Not Started		
□	Skill development		Not Started		
□	Apply Machine Learning		Not Started		
□	Model Refinement		Not Started		
□	Evaluation and Testing		Not Started		
□	Finalization and Report Sub...		Not Started		
□	Presentation		Not Started		

+ Add task

Figure 3. Project management tool-MONDAY.com.

Appendix 6: Weekly Progress Chart

Screenshot of Capstone Project progress week wise (See Figure 3).

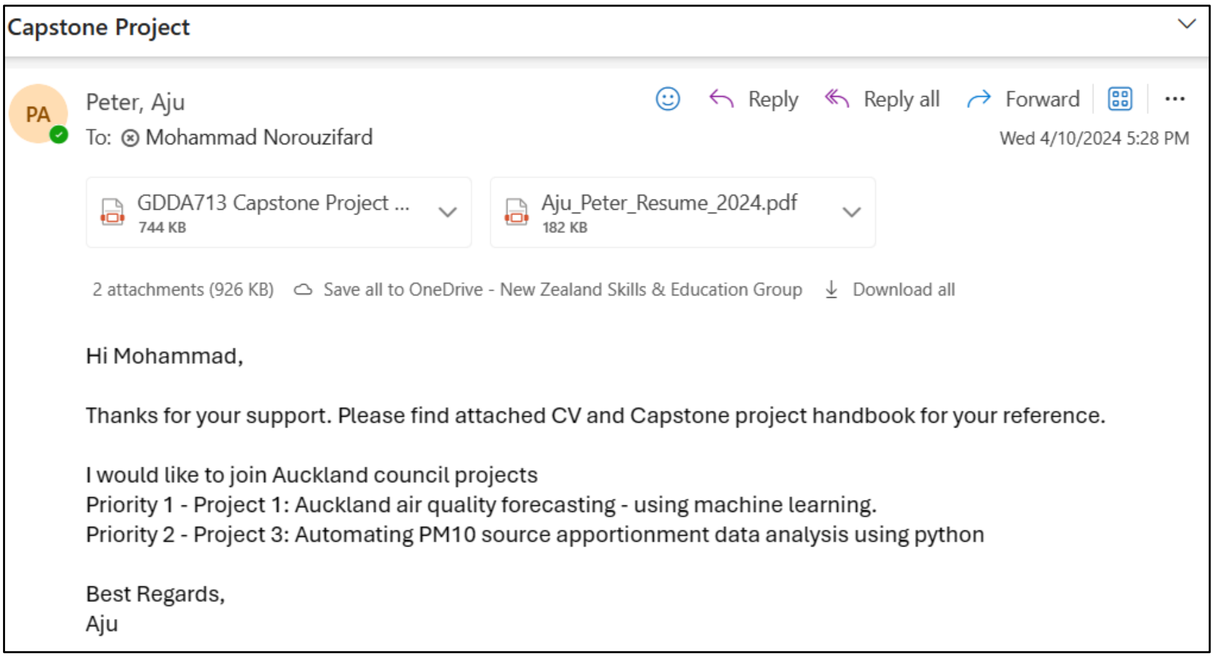


	A	B	C	G	H	I	J	K
1	StudentID	Forename	Surname	Success and difficulties in Week-1	Success and difficulties in Week-2	Success and difficulties in Week-3	Success and difficulties in Week-4	Success and difficulties in Week-5
8	764706847	Aju	Peter	Absent	Success: - Selected Project - Submitted my CV	Success: -Finalized Project with Auckland Council - Had initial meeting with external supervisor Mr. Louis Boamponsem -F2F meeting fixed with internal supervisor on Monday at 9am	Sucess: -Had a F2F meeting with internal supervisor -Historic dataset received from Auckland council Challenges: -Finding quality recent literature papers -Need to collect weather quality data from Auckland Council -Need to collect Penrose station sensor data 24hour average except PM10 Goal: -Need to submit the first full draft by Friday Well-known publisher: Elsevier, ACM, Springer, IEEE, ... **Goal**: Share the first full draft by Friday 26.4.24	Success: -Inputs received from Dr Sara for project proposal

Figure 4.Capstone project progress.

Appendix 7: Communication History

I have sent email expressing my interest in the Auckland Council machine learning project (See Figure 4).


Capstone Project
PA
Reply
Reply all
Forward
GDDA713 Capstone Project ...
744 KB
Aju_Peter_Resume_2024.pdf
182 KB
2 attachments (926 KB)
Save all to OneDrive - New Zealand Skills & Education Group
Download all

Peter, Aju

To: Mohammad Norouzifard

Wed 4/10/2024 5:28 PM

Hi Mohammad,

Thanks for your support. Please find attached CV and Capstone project handbook for your reference.

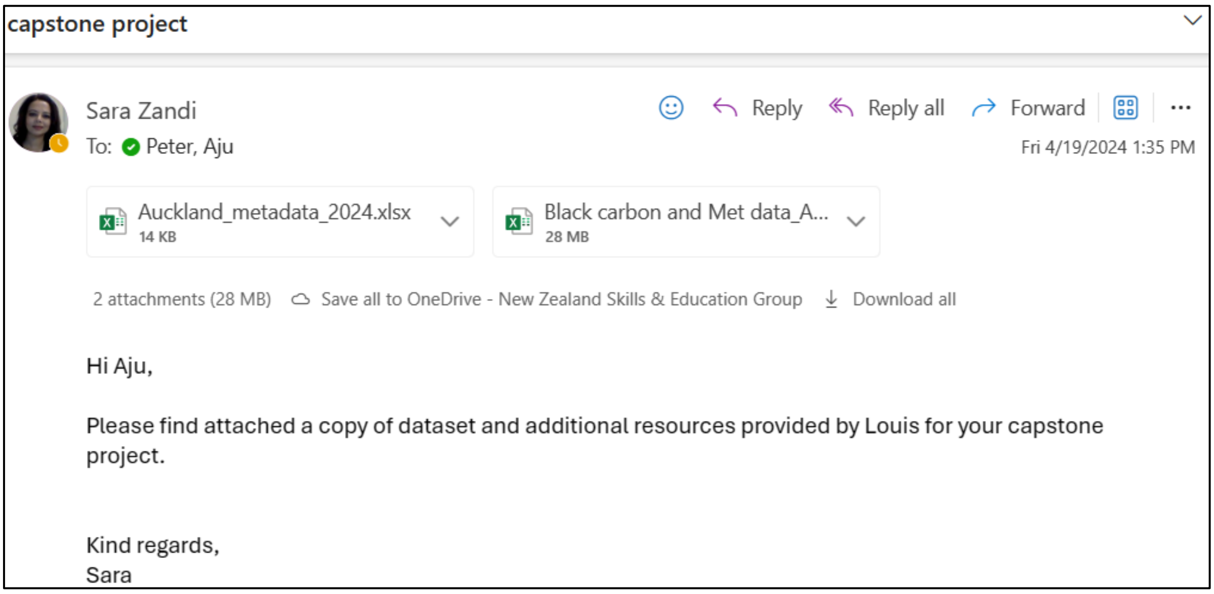
I would like to join Auckland council projects
Priority 1 - Project 1: Auckland air quality forecasting - using machine learning.
Priority 2 - Project 3: Automating PM10 source apportionment data analysis using python

Best Regards,
Aju

Figure 5.Expression of interest.

Appendix 8: Data Collection History

Dr. Sara Zandi shared Historical datasets and additional information's from Auckland Council (See Figure 5).


capstone project
Reply
Reply all
Forward
Auckland_metadata_2024.xlsx
14 KB
Black carbon and Met data_A...
28 MB
2 attachments (28 MB)
Save all to OneDrive - New Zealand Skills & Education Group
Download all

Sara Zandi

To: Peter, Aju

Fri 4/19/2024 1:35 PM

Hi Aju,

Please find attached a copy of dataset and additional resources provided by Louis for your capstone project.

Kind regards,
Sara

Figure 6.Historic dataset.

Appendix 9: Minutes of meetings

Minutes of External Meeting with Mr. Louis Boamponsem, Auckland Council

1. Date: 16-04-2024

- Project Introduction: Overview of the air quality forecast project using machine learning models.
- Expectation of project: Building a highly predictive machine learning model
- Discussion on Air Quality in Auckland: Highlighted the importance of air quality management in Auckland.
- Existing Tools: Discussed the current lack of air quality forecasting tools at Auckland Council.
- Data Sharing: Agreed to share 10 years historic air quality datasets with the project team.

Weekly Meeting with Dr. Sara Zandi

2. Date: 22-04-2024

- Research Question: Finalization of the research question.
- Selection of Pollutants and Location: Chose PM10 as the primary pollutant for study due to its significance in New Zealand. Selected the monitoring station location at Penrose, discussing its relevance.
- Health Impact Discussion: Discussed health issues related to PM10 and PM2.5 and its sources.
- Data Review: Reviewed the historic data shared by Auckland Council.

3. Date: 29-04-2024

- Dataset Analysis: Discussed the dataset shared by Auckland Council.
- Resource Sharing: Shared recent research papers and Auckland Council annual report on air quality predictions.
- Additional Data Acquisition: Discussed acquiring supplementary datasets like weather data from the Auckland Council website and NIWA. Decided to collect daily meteorological data including wind speed, wind direction, temperature, humidity, and solar radiation.
- Prediction Focus: Decided to focus on predicting the 24-hour average of PM10 concentration a day ahead.
- Exploratory Data Analysis (EDA): Discussed EDA processes, methods for handling missing data, detecting outliers, feature selection, and feature engineering.
- Project Proposal: Discussed about the overall project proposal formatting and requirements.

4. Date: 06-05-2024

- Project Proposal: Reviewed the project proposal and provided suggestions for improvement.
- Final Approval: Reviewed and approved the final draft of the project proposal.