­

|  |
| --- |
| A city with many tall buildings  Description automatically generated |
| Urban Air Quality  Predictive Modeling and Analysis of Air Quality in Mongolia |
| |  |  |  | | --- | --- | --- | | Nomin Lkhagvasukh | 12/15/24 | Strategic Thinking | |

A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Nomin Lkhagvasukh |
| *Student Number* | 2024461 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | CA 2 – Capstone Project Proposal |
| *Assessment Due Date* |  |
| *Date of Submission* | Sunday, 15th December 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Introduction 2](#_Toc185185775)

[Strategic Overview of the Business Problem 3](#_Toc185185776)

[Project Plan 3](#_Toc185185777)

[Timeline 3](#_Toc185185778)

[Business Understanding 3](#_Toc185185779)

[Data Understanding 4](#_Toc185185780)

[Data Preparation 6](#_Toc185185781)

[Machine Learning Implementation 7](#_Toc185185782)

[Linear Regression 7](#_Toc185185783)

[Random Forest Regressor 7](#_Toc185185784)

[Findings 8](#_Toc185185785)

[Conclusions 8](#_Toc185185786)

[Future Recommendations 9](#_Toc185185787)

[References 9](#_Toc185185788)

[Figures 9](#_Toc185185789)

[A GitHub link 9](#_Toc185185790)

# Introduction

The environment and public health are greatly impacted by air quality, particularly in metropolitan areas where pollution levels are frequently high. Air pollution is a serious problem in Mongolia, especially in Ulaanbaatar. **PM2.5 -** fine particulate matter, is the main pollutant of concern since it can enter the respiratory system deeply and provide serious health hazards. The aim of this project is to analyse air pollution data in Ulaanbaatar for 2023, focusing on PM2.5 levels.

This project involves collecting and preparing the dataset, performing exploratory data analysis to understand pollution trends, implementing machine learning models for forecasting, and presenting actionable insights. By focusing on Ulaanbaatar's air quality, this research addresses a pressing environmental and public health issue in Mongolia.

A screenshot of a weather report

Description automatically generated

Figure 1. Live most polluted major city ranking, Ulaanbaatar.

Ulaanbaatar is the capital city of Mongolia, which is top 5th most polluted city in the world.

Breathing air with PM2.5 levels going as high 224 μg/m³, which puts it directly into the ‘very unhealthy’ bracket, would have innumerable consequences, particularly on vulnerable parts of the population, with young children and pregnant mothers being the most at risk.

A table of health and medical information

Description automatically generated with medium confidence

Figure 2. Air pollution levels

# Strategic Overview of the Business Problem

**Problem Context:**

Ulaanbaatar, Mongolia's air quality is a serious problem, especially during the winter when pollution levels rise sharply as a result of coal burning for heating. Particulate matter of a diameter of 2.5 micrometres or less, or **PM2.5,** has a direct effect on population health by inducing cardiovascular and respiratory disorders. The management of public health and the formulation of public policy depend on the monitoring and forecasting of air quality.

**Business Problem:**

This project's goal is to use machine learning techniques to accurately forecast Ulaanbaatar's air quality. The impact of pollution on public health can be lessened by authorities issuing timely warnings and improving air pollution management methods with the use of accurate **AQI (Air Quality Index)** predictions.

# Project Plan

Create a machine learning model to forecast the air quality index (AQI) using environmental data, including temperature, humidity, PM2.5 levels, and other meteorological parameters.  
Analyse the various models' performances and offer suggestions for enhancements.

|  |  |
| --- | --- |
| Timeline | |
| **Week 1-2** | Data exploration, pre-processing, and cleaning |
| **Week 3-4** | Model development and tuning (Linear Regression and Random Forest) |
| **Week 5** | Model evaluation and comparison |
| **Week 6** | Final report preparation, insights, and conclusion writing |
| **Milestones:** | |
| **Data Preprocessing** | Completed cleaning and transforming data (Week 2) |
| **Model Development** | Linear Regression and Random Forest implemented (Week 4) |
| **Evaluation and Reporting** | Evaluation metrics gathered, final conclusions drawn (Week 6) |

Figure 3. Timeline of project

# Business Understanding

Air pollution is a major environmental and public health concern in Ulaanbaatar, Mongolia, particularly during the harsh winters. The primary sources of pollution include:

1. **Coal Burning**: Due to the extreme cold, households rely heavily on burning raw coal for heating, especially in **ger districts** (yurt settlements).
2. **Vehicle Emissions**: The growing number of vehicles, many of which are old or lack proper emission controls, contribute significantly to air pollution.
3. **Industrial Emissions**: Factories and power plants release pollutants into the atmosphere.
4. **Geographical Factors**: The city's location in a valley traps polluted air, especially during temperature inversions in winter, worsening air quality.

These factors lead to dangerously high levels of **PM2.5** (fine particulate matter) and **PM10**, which are linked to serious health issues such as respiratory diseases, cardiovascular problems, and premature death.

**Importance of Air Quality Prediction**

Accurate prediction of air quality has several strategic benefits for different stakeholders:

* **Government and Policy Makers**:

Helps authorities in putting policies into place that lower pollution levels, such as traffic control measures, cleaner heating system subsidies, or limitations on the burning of coal. Makes it possible to plan public health campaigns or emergency medical responses more effectively during times of excessive pollution.

* **Health Organizations**:

During times of extreme air pollution, hospitals and clinics can plan for an increase in admissions.  
When there is poor air quality, public health organisations might issue advice to reduce outdoor activity.

* **Residents and Businesses**:

By wearing masks, using air purifiers, or scheduling activities during times when the air quality is better, residents can take preventative measures.

Companies can modify their operations to reduce the amount of hazardous contaminants that workers, particularly those who work outside, are exposed to.

# Data Understanding

**Dataset Overview:**

* The dataset contains multiple features, including **PM2.5**, **temperature**, **humidity**, **wind speed**, **AQI Category**, and **Now Cast Concentrations**, among others.
* **Target Variable**: The **AQI** value (dependent variable) is the primary focus for prediction.
* **Data Quality**: The dataset might have missing values, outliers, or categorical variables that need transformation for use in machine learning models.

**Exploratory Data Analysis (EDA):**

* Descriptive statistics such as mean, median, standard deviation, and distribution of the key features.
* Visualizations such as **histograms**, **scatter plots**, **heatmaps** to show correlations, and **box plots** to detect outliers.

A graph of different colored bars

Description automatically generated

Figure 4. Stacked Bar Plot of Categories by Month in Ulaanbaatar

This stacked bar plot displays the **Air Quality Index (AQI) categories** by month.

**Winter months (January, February, November, December)**: Show a significant increase in "Unhealthy," "Very Unhealthy," and even "Hazardous" AQI levels. This indicates higher pollution levels during colder months.

**Spring and summer months (March to August)**: Show a higher proportion of "Good" and "Moderate" AQI levels, suggesting improved air quality during warmer seasons.

**September and October**: Show a mixed trend with a reduction in unhealthy AQI categories.

**A graph showing a number of data

Description automatically generated with medium confidence**

Figure 5. Air Quality Index Over Time

A graph of a box plot

Description automatically generated

Figure 6. Boxplot AQI by Month

**Conclusion:**

The graph highlights a clear **seasonal trend** in air quality:

* **Poor air quality** during winter, likely due to increased coal/wood burning and lower atmospheric dispersion.
* **Better air quality** in spring and summer months, possibly due to favourable weather conditions and reduced emissions.

# Data Preparation

**Pre-Processing Steps**

* **Data Cleaning**: Removing rows with missing or erroneous data.
* **Encoding Categorical Variables**: One-hot encoding or label encoding applied to categorical variables (AQI Categories).
* **Feature Engineering**: Adding new features, normalizing, and scaling data as necessary. Converted the "Date (LT)" column to extract time-based features like month and hour.
* **Splitting Data**: The data was split into **training** (80%) and **testing** (20%) sets to ensure unbiased evaluation of the model.

# Machine Learning Implementation

Building and evaluating multiple regression models to predict the air quality in Ulaanbaatar in 2023. Understand the relationship between PM2.5 levels and AQI: Explore how PM2.5 concentrations influence the overall air quality in Ulaanbaatar.

Two machine learning models were implemented to predict AQI:

**Models Used:**

1. **Linear Regression**: A simple approach that helps establish a baseline performance.
2. **Random Forest Regressor**: A robust tree-based ensemble method that captures non-linear relationships and improves predictive accuracy.

**Model Evaluation Metrics:**

* **Mean Absolute Error (MAE)**: Measures the average magnitude of errors.
* **Root Mean Squared Error (RMSE)**: Measures the square root of the average squared differences between actual and predicted values.
* **R²**: Represents how well the model explains the variance in the target variable.

## Linear Regression

Linear Regression is a baseline model used for understanding the relationship between predictors and AQI.

**Results:**

* **Mean Absolute Error (MAE)**: 11.30
* **Root Mean Squared Error (RMSE)**: 14.68
* **R²**: 0.9713

These findings show that while linear regression is capable of capturing broad AQI trends, it may not be able to handle intricate patterns or non-linear connections in the data.

## Random Forest Regressor

Random Forest is a robust ensemble learning method that captures non-linear relationships and interactions between features.

**Tuning the Model:**

The following hyperparameters were optimized:

* **Number of Trees (n\_estimators)**: 100
* **Max Depth**: 15
* **Min Samples Split**: 2

**Results:**

* **Mean Absolute Error (MAE)**: 0.0715
* **Root Mean Squared Error (RMSE)**: 0.6519
* **R²**: 0.9999

Random Forest significantly outperformed Linear Regression, with near-perfect accuracy. Its ability to model non-linear patterns made it particularly suitable for this dataset.

Findings

**Seasonal Patterns:** Because more coal is used throughout the winter, AQI levels are continuously higher at this time.

**Importance of the Feature:** The most important predictor of AQI was PM2.5 concentration.  
AQI was also impacted by meteorological factors such as humidity and temperature.

**Model Comparison:** Linear Regression was a solid baseline, but Random Forest performed better, with an almost perfect R2 score (0.9999).

## Visualizations

This figure shows two scatter plots comparing **Actual AQI (Air Quality Index)** vs **Predicted AQI** values for two models:

A graph of different types of graphs

Description automatically generated with medium confidence

Figure 7. Actual vs Predicted AQI

**Linear Regression**

For a perfect prediction, the points primarily line up with the red dashed line.  
The little variations, however, imply that while Linear Regression performs well generally, it is not entirely correct.

**Random Forest**

The red dashed line is nearly exactly where the points line up.  
This indicates that Random Forest forecasts have a very good accuracy and are very near to the real AQI values.

A graph with a line and a red line

Description automatically generated

Figure 8. Residual Plot: Actual AQI vs Residuals

A graph with lines and numbers

Description automatically generated

Figure 9. Random Forest Distribution

This plot shows the **Residual Distribution** for the **Random Forest model**.

The discrepancy between actual and expected values is represented by residuals. The majority of residuals are concentrated close to zero, as the plot shows.

This suggests that there are very few errors in the Random Forest model's predictions. The model consistently predicts AQI values that are extremely close to the actual ones, as evidenced by the very tall and narrow peak near zero that exhibits low variability in errors.

**Summary:**

* **Random Forest** performs significantly better than **Linear Regression**, as it predicts the AQI values more accurately.
* This aligns with the earlier metrics ( R² close to 1 for Random Forest), confirming its superior performance.

# Conclusions

This assignment used machine learning techniques, especially Random Forest and Linear Regression models, to analyse Ulaanbaatar air quality data. Evaluation of the models' performance and prediction of Air Quality Index (AQI) values based on important parameters were the objectives. The following findings were obtained by data processing, model implementation, and analysis:

## **Air Quality Analysis**

* **AQI Distribution by Month**:

Ulaanbaatar's air quality varies significantly throughout the seasons.  
The "Unhealthy," "Very Unhealthy," and even "Hazardous" AQI scores are greater throughout the winter months (January, November, and December). This suggests that there is significant air pollution during the winter months, most likely as a result of rising coal consumption, increased demand for heating, and weather-related factors like temperature inversion.

* **Monthly Trends**: Higher numbers of "Good" and "Moderate" AQI values indicate better air quality during warmer months (May through September). This demonstrates how air pollution varies with the seasons, becoming milder in the summer.
* **Significance of Parameters**:

Raw Concentration and other pollutants are among the parameters that greatly influence AQI fluctuations, indicating that monitoring these parameters is essential for pollution control.

## **Modelling Results and Evaluation**

Two machine learning models were implemented: **Linear Regression** and **Random Forest**. The findings are as follows:

**Linear Regression**:

* Linear Regression showed a relatively good fit for predicting AQI values, but it struggled to handle more complex, non-linear relationships in the dataset.
* Performance Metrics:
  + **Mean Absolute Error (MAE):** Moderate error observed.
  + **R² Score:** High, but not perfect, indicating scope for improvement.

**Random Forest**:

* The Random Forest model outperformed Linear Regression significantly. It captured complex patterns in the dataset and provided highly accurate predictions.
* Performance Metrics:
  + **MAE:** 0.071 (very low, indicating minimal errors).
  + **RMSE:** 0.65 (small errors overall).
  + **R² Score:** 0.9999 (indicating nearly perfect predictions).

**Residual Distribution**:

* The Random Forest model showed residuals tightly clustered around zero, highlighting its ability to make predictions with minimal error.
* The narrow distribution suggests consistency and reliability in the model’s performance.

**Comparison of Results**:

* The scatter plots of "Actual vs Predicted AQI" showed that Random Forest predictions align almost perfectly with the actual AQI values, whereas Linear Regression deviated slightly.
* This emphasizes the superiority of Random Forest for complex datasets like air quality.

## **Crucial Information**

Air Quality Issues: The air in Ulaanbaatar is filthy, especially during the winter. Addressing this issue requires policies that promote cleaner energy sources, increase heating efficiency, and decrease the use of coal.

# Future Recommendations

# References

# Figures

[Figure 1. Live most polluted major city ranking, Ulaanbaatar. 2](#_Toc185189171)

[Figure 2. Air pollution levels 2](#_Toc185189172)

[Figure 3. Timeline of project 3](#_Toc185189173)

[Figure 4. Stacked Bar Plot of Categories by Month in Ulaanbaatar 5](#_Toc185189174)

[Figure 5. Air Quality Index Over Time 5](#_Toc185189175)

[Figure 6. Boxplot AQI by Month 6](#_Toc185189176)

[Figure 7. Actual vs Predicted AQI 8](#_Toc185189177)

[Figure 8. Residual Plot: Actual AQI vs Residuals 9](#_Toc185189178)

[Figure 9. Random Forest Distribution 9](#_Toc185189179)

# A GitHub link

<https://github.com/CCT-Dublin/capstone-project-Nomin1016>