ABSTRACT

One of the gases that causes health risk is ozone (O3). O3 is a greenhouse gas that contributes to climate change, produced when sunlight reacts with pollutants such as NOX, thereby affecting urban air quality. Since Secunda, a town in Mpumalanga, South Africa, is known for producing coal, using coal to oil plants, it is anticipated that high levels of emissions will emanate, including O3. Five classifier algorithms were applied to model the future ozone levels in this study: Random Forest, XGBoost, Logistic Regression, Support Vector Machine, and Decision Tree. The results of the study showed that Logistic Regression has the highest ROC-AUC (0.9280), followed closely by XGBoost (0.9053). XGBoost has the highest precision (0.1961), and F1-Score (0.2395), meaning when it predicts an exceedance, it's more likely to be correct than other models. The confusion matrix showed that XGBoost (2645) had the highest correct predictions compared to Logistic regression (2377). Therefore, XGBoost algorithm can be applied to the air quality dataset to forecast a 3-hour lead in O3. The web application is user-friendly and requires the following inputs: current 3-hour O3, NO2, NO, NOX, hour of the day, and day of the year. Since the XGBoost model was trained as a classification model, exceedance occurs when the probability is greater than 0.5. If the 3-hour exceedance exceeds 61 ppb, the prediction results are printed in red, otherwise, in green.

Deployment of a Three-Hour Lead Time for Ozone Levels Prediction in Secunda

Mpho Muloiwa and **Chale Justice Moferefere**

1. INTRODUCTION

O_3 is a greenhouse gas that contributes to climate change, produced when sunlight reacts with pollutants such as NO_X, thereby affecting urban air quality [1]. Because these gases have become major air pollutants in the atmosphere, several authors have attempted to develop models that can be used to predict O_3 levels in the atmosphere, to ensure that health risks are minimized. [2] applied long short-term memory (LSTM), support vector machine (SVM), recurrent neural network (RNN), and gated recurrent unit (GRU) for predicting O_3 levels in China. [3] applied kernel extreme learning machine (KELM) and SVM for predicting O_3 levels in Nanjing City. [4] applied random forest model for predicting O_3 levels in China. The challenge is that most models remain in the hands of authors and publishing journal houses, limiting access to the public, organizations, and decision makers. This means that even if the prediction models are useful, the public, organizations, and decision makers will not be able to use the prediction models to forecast the O_3 levels and issue warnings to the public. The available resources such as air quality application have been commercialized, which is a disadvantage to the general public. Therefore, there is a need for a free user-friendly O_3 prediction web application tool that is available for the public, organizations, and decision makers to use, in order to safeguard the public. This study will focus on deployment of a three-hour lead time for O_3 levels prediction, enabling the public, decision makers, and organizations to put in place action plans to avoid health risks.

2. METHODS AND MATERIALS

2.1 STUDY AREA AND DATA COLLECTION

Data was collected from the South African Air Quality Information System (SAAQIS) at the Secunda station. The data spans from January 2022 to December 2023. The data contain variables: date, time, NO_2, NO, NO_X, and O_3 **parts per billion (ppb).**

2.2 DATA VISUALIZATION AND MODELLING

The following visualizations were performed on Python software: Histograms, Time series plots, and Correlation analysis. Five classifier algorithms were applied to model the future ozone levels in this study: Random Forest, XGBoost, Logistic Regression, Support Vector Machine, and Decision Tree, and were evaluated using the following metrics: ROC-AUC, precision, recall, accuracy, training time, and confusion matrix.

- $Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$
- $Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$
- $F1 = \frac{2 \times True\ Positive}{2 \times True\ Positive + False\ Positive + False\ Negative}$
- $accuracy = \frac{Number\ of\ correct\ predictions}{Total\ Number\ of\ Predictions}$
- Training time
- Confusion matrix

2.3 DEPLOYMENT

The O_3 prediction model was deployed using the Streamlit library. Streamlit processes the user-entered data (NO_2, NO, NO_X, and O_3, hour of day, and day of the year) using the trained XGBoost model, and the output (3-hour lead time on O_3) is displayed on the front end. The web application is developed using Python programming language, Google Colab, VS Code, and notepad. After installing the Streamlit library in the Google Colab and VS Code, the application (app.py) was created and coded to enable it to use the trained model that is saved as a joblib file. After developing the app on VS code, it is deployed via localhost and opens in the default web browser, allowing users to interact with the web app.

3. RESULTS AND DISCUSSION

3.1 VISUALIZATION

In Figure 1, NO_2 concentrations were significantly higher in 2023 than in 2022, with peaks reaching 78.49 ppb. For NO, a gradual increase from January to July 2023 is observed, reaching a maximum of 207.982 ppb. The trend observed for NO, NO_2, and NO_X is that peaks occur more frequently in the spring season than in the summer, autumn, and winter seasons. O_3 levels peak in summer, autumn, and spring, while drops are observed in winter. The maximum value observed was 102.424 ppb, while the minimum value was 0.652 ppb. Similar trends in South Korea, with the values reported in spring, autumn, winter, and summer as 46.9, 38.2, 33, and 32.1 ppb, respectively [2]. Overall, O_3 concentrations tend to be higher in the afternoon, while NO_2, NO, and NOX peak during rush hour traffic emissions.

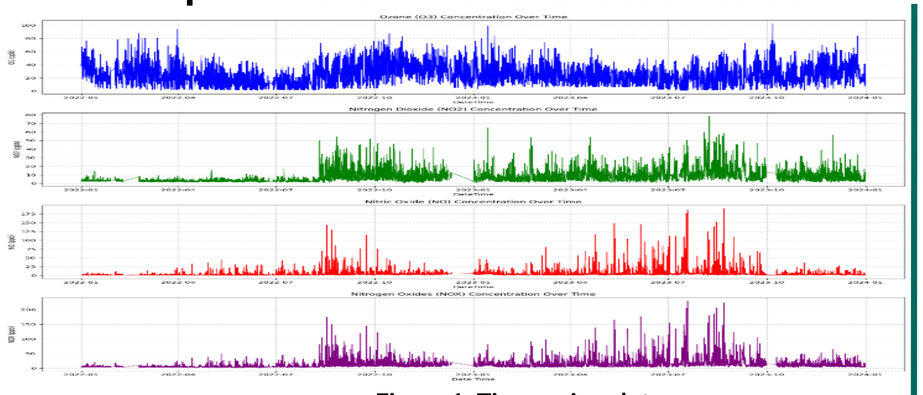

Figure 1: Time series plots

In Figure 2, the histograms for NO_2, NO, and NO_X are strongly left-skewed, indicating more instances of lower values than higher. The O_3 histogram showed a more spread-out distribution than NO and NO_X, with a peak in the mid-range and tails extending to both lower and higher concentrations. In Figure 3, the correlation matrix, the variable hour was positively correlated to O_3 with a value of 0.33, and was the highest amongst NO_2, NO, and NO_X, the day of week, month, and year variables. This means that as the day progresses, O_3 levels in the atmosphere increase, which is a health concern. Similar results were reported by [2], who confirmed that O_3 value peaked midday, reaching maximum values in the afternoon, in South Africa. NO_X, NO, and NO_2 are negatively correlated with O_3, with values of -0.25, -0.21, and -0.22, respectively. This means that when O_3 increases, NO_x, NO, and NO_2 decrease in the atmosphere.

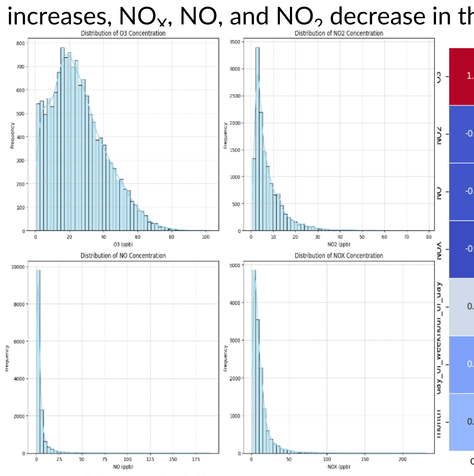

Figure 2: Histograms

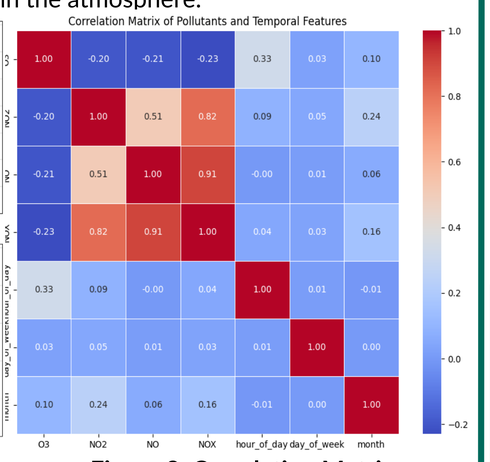

Figure 3: Correlation Matrix

3.2 MODELLING

3.2.1 MODELLING PERFORMANCE EVALUATION

In Figure 4, logistic regression has the highest ROC-AUC (0.9280), followed closely by XGBoost (0.9053), indicating that both perform best at distinguishing between exceedance and non-exceedance events. XGBoost has the highest precision (0.1961), meaning when it predicts an exceedance, it's more likely to be correct than other models. The values are generally low across all models, indicating many false positives. XGBoost has the highest F1-Score (0.2395), suggesting the best balance between precision and recall among the models.

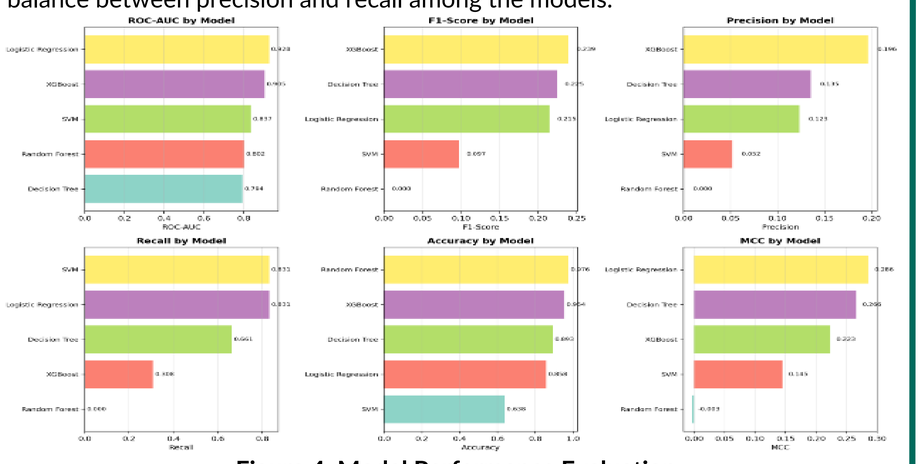

Figure 4: Model Performance Evaluation

3.2.2 CONFUSION MATRIX AND MODEL SENSITIVITY ANALYSIS

In Figure 5, XGBoost correctly predicted no exceedance (2625) and exceedance (20), but incorrectly predicted no exceedance (45) and exceedance (82). Logistic regression correctly predicted no exceedance (2323) and exceedance (54), but incorrectly predicted no exceedance (11) and exceedance (384). This is a much higher number of false alarms than with XGBoost. Overall, XGBoost (2645) had more correct predictions compared to Logistic regression (2377). In Figure 6, the variable that appears to be the highest contributor is current O_3 levels, which are vital for predicting O_3 levels over the next three hours. This makes sense because the current O_3 level should affect future O_3 levels.

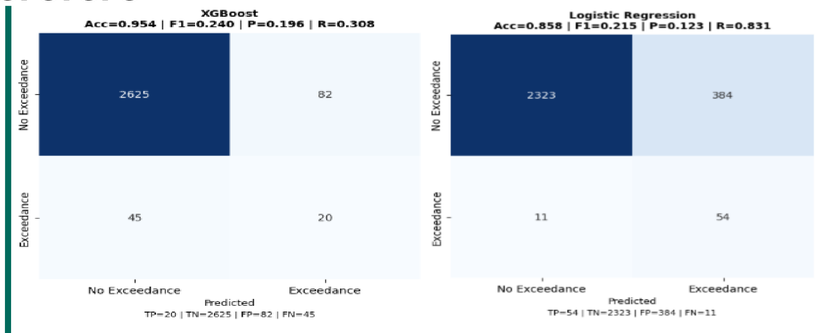

Figure 5: Confusion Matrix

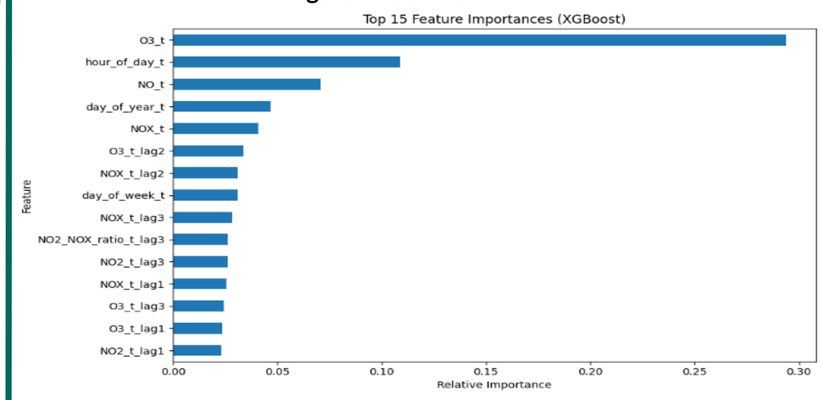

Figure 6: Model Sensitivity Analysis

3.3 DEPLOYMENT

Figure 7 presents the web application interface, built on the XGBoost 3-hour O_3 exceedance trained model. On the left-hand side, input features required for 3-hour O_3 exceedance are: current 3-hour O_3, NO_2, NO, NO_X, hour of the day, and day of the year. If the 3-hour exceedance exceeds 61ppb, the prediction results are printed in red. In this case, the exceedance was not reached, hence, the probability is less than 0.5, which is displayed by the blue bar graph.

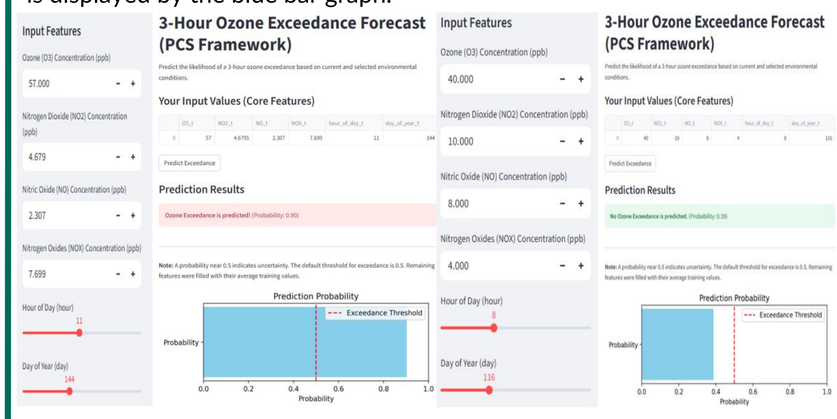

Figure 7: Web Application interface

4. CONCLUSION

This research study focused on deployment of a three-hour lead time for O_3 levels prediction using the trained XGBoost model. Three visualizations were performed: time series graphs, correlation matrix analysis, and Histograms. The visualizations produced insightful information such as O_3 concentrations being lower during the early morning hours and peaks in the afternoon. Both XGBoost and Logistic regression produced superior results, however, XGBoost had a slight edge, making it the best model. The wep application was successfully deployed using Streamlit. The web application is user-friendly and requires the following inputs: current 3-hour O_3, NO_2, NO, NO_X, hour of the day, and day of the year. Since the XGBoost model was trained as a classification model, exceedance occurs when the probability is greater than 0.5. If the 3-hour exceedance exceeds 61 ppb, the prediction results are printed in red, otherwise, in green.

5. REFERENCES

[1] Yan, B., Luo, J., Zhang, M., Zhang, Y., Xiao, T., Wang, L., Liu, B., Han, Y., He, G., Yang, L. and Huang, Z., 2024. Analysis of Ozone Pollution Characteristics, Meteorological Effects, and Transport Sources in Zhuzhou, China. Atmosphere, 15(5), p.559.

[2] Ma, W., Yuan, Z., Lau, A.K., Wang, L., Liao, C. and Zhang, Y., 2022. Optimized neural network for daily-scale ozone prediction based on transfer learning. Science of the Total Environment, 827, p.154279.

[3] Su, X., An, J., Zhang, Y., Zhu, P. and Zhu, B., 2020. Prediction of ozone hourly concentrations by support vector machine and kernel extreme learning machine using wavelet transformation and partial least squares methods. Atmospheric Pollution Research, 11(6), pp.51-60.

[4] Zhan, Y., Luo, Y., Deng, X., Grieneisen, M.L., Zhang, M. and Di, B., 2018. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environmental Pollution, 233, pp.464-473

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Capstone Project - MIT 808

Course Coordinators:
Dr. Vukosi Marivate (vukosi.marivate@cs.up.ac.za)
Abiodun Modupe (abiodun.modupe@cs.up.ac.za)

Scan me