Random Forest effectively forecasts next-day ozone exceedance for public health warnings.

Will Air Quality Have an Exceedance Today?

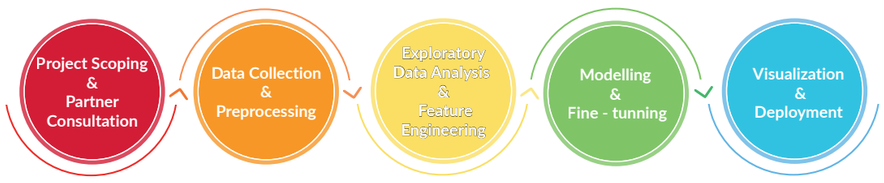
Using ensemble models to predict ground-level ozone exceedances in the South African Highveld

Rivoningo Mageza, Kamogelo Moukangwe

INTRODUCTION

Despite having some of the most robust air quality legislation on the continent – including provisions for government-mandated action to improve quality in high-priority areas like the Highveld, there has been limited research and development of forecasting tools and public health warning systems for ground-level ozone, which is named as a critical pollutant in the Air Quality Act (Act 39 of 2004). This project employed ensemble models (XGBoost, Random Forest) to develop a ground-level ozone prediction model, made accessible through a web app for potential use in air quality improvement research and public health alerting.

METHODOLOGY



Data Sources

- Pollutants: Hourly NO_x & O_3 from 36 stations (2 years)
- Meteorology: Temperature, humidity, wind speed (Open-Meteo)
- Real-time inputs: WAQI API (SAAQIS) + Open-Meteo forecasts

Modeling

- 10 models tested: tree-based (Random Forest, XGBoost), kernel-based (SVM) and deep learning (LSTM)
- Top 3 Best performers underwent hyperparameter tuning
- Final model selection was based on comparison of 1-week of forecast against actual SAAQIS readings

Validation

- Offline validation: Forecasts from the three candidate models were compared against actual SAAQIS readings over a one-week period. The best-performing model (RF) was selected for deployment.
- User evaluation (post-deployment): A Small tester group completed a survey on perceived accuracy & usefulness

Visualization and Deployment

- An Interactive Streamlit app with ozone forecasts, exceedance alert (>61 ppb), and optional daily subscription was built.

RESULTS

Candidate models were compared using standard regression metrics: R^2 measures explained variance (higher = better), while RMSE and MAE quantify prediction error (lower = better). Training time (in seconds) reflects computational efficiency.

- Random Forest achieved the highest R^2 (0.781) with low prediction error (RMSE = 5.94 ppb, MAE = 4.30 ppb).
- XGBoost performed nearly as well but trained 20× faster making it more suitable for real-time retraining.
- Other models underperformed: SVM had similar error but slower training; BD-LSTM showed poor accuracy).
- Feature importance ranked lagged ozone and temperature as the top predictors, followed by NO_2 and hour-of-day features

Table 1: Performance comparison of four candidate models.

Model	R^2	RMSE (ppb)	MAE (ppb)	Train Time (s)
Random Forest	0.781	5.94	4.3	8.67
SVM	0.778	5.98	4.29	5.54
XGBoost	0.776	6.01	4.34	0.43
BD-LSTM	0.407	9.73	7.28	37.87

MODEL DEPLOYMENT

Tool Stack:

Deployment followed a local execution architecture model, with files hosted on GitHub. The visualization app with the forecasting tool runs locally on user's machine. This implementation prioritizes security while maintaining reproducibility and accessibility to users with no programming experience.

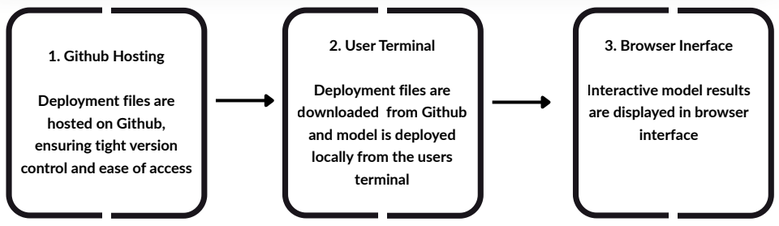


Figure 1: Summary of Deployment Flow

DISCUSSION

- EDA revealed strong diurnal (afternoon peak) and seasonal (warmer months) ozone patterns → informed feature design (hour sin/cos, day of week)
- Feature importance (dominated by lagged ozone and temperature) confirms ozone persistence and photochemical drivers as top predictors – consistent with atmospheric chemistry
- Tree-based ensembles (RF, XGBoost) outperformed other models, aligning with literature on data with mixed features
- XGBoost trained 20× faster than Random Forest with comparable accuracy
- Performance varied by station and time of day – morning hours and low-pollution sites.

Table 2: Features ranked by importance

Rank	Feature
1	O_3 lag
2	Temperature
3	NO_2
4	Hour (sin)
5	Hour (cos)
6	Day of week
7	Humidity
8	Wind speed

NEXT STEPS

- Improve low-pollution & morning forecast accuracy
- Add missing meteorological drivers such as Include boundary layer height and solar radiation
- Replace estimated NO with direct measurements
- Stabilise recursive forecasts
- Deploy XGBoost for fast retraining

VISUALISATIONS

- The web interface design prioritises clarity and actionability. Key features:
- Colour-coded gauge – Shows forecasted ozone (ppb) with green-yellow-red zones. Red means exceedance (>61 ppb).
- Tabbed navigation – Four tabs (Ozone Forecast, Model Performance, Geographical, About) keep things organised.
- Map toggle – Settings switch lets users switch to map view for station locations.
- Table + line graph – 24-hour forecast shown both as numbers (precise) and a trend line.
- Exceedance alert – Prominent warning appears when ozone >61 ppb.
- This makes the tool useful for both technical and non-technical users

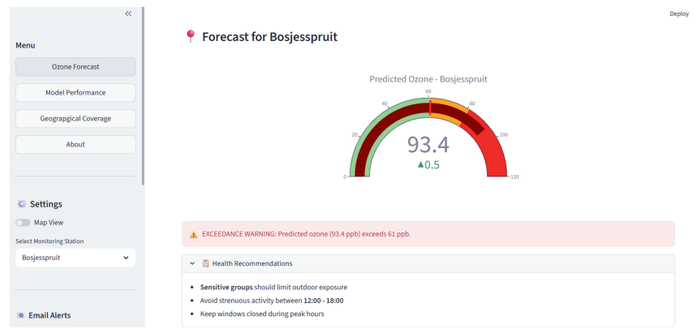


Figure 1: Screenshot of the Streamlit app interface. The main panel displays a colour-coded gauge showing the forecasted ozone concentration (ppb). The navigation bar (top) provides four tabs: Ozone Forecast, Model Performance, Geographical, About, and a Settings option (Switch to Map View).

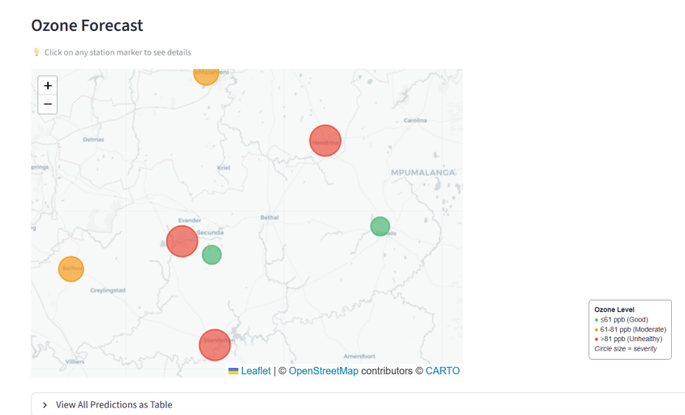


Figure 2: Interactive Map as Deployed in the Web Interface. This gives an overview of the status for all the stations the model has prediction capability for, with the option to get detailed information for a specific station by simply clicking on the respective bubble.

USER TESTING

User Acceptance Testing (UAT) involved 11 participants. Installation success, forecast generation, visual clarity, and overall helpfulness were assessed via survey that users were asked to complete after installing and navigating the app. Readme files were updated following feedback provided on installation difficulty.

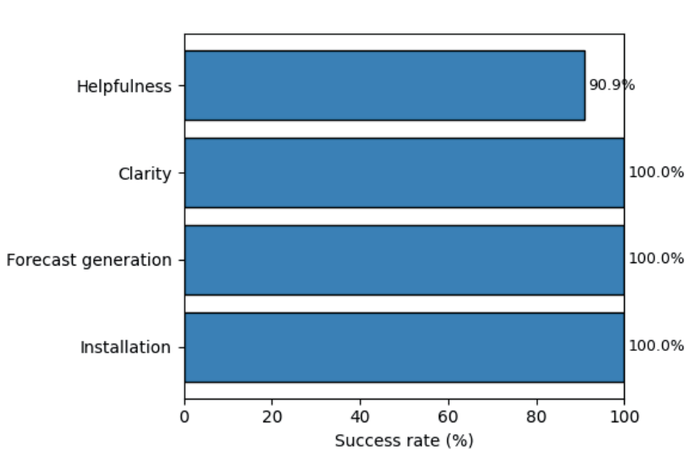


Figure 3: Bar chart Summary of feedback from UAT. All users successfully installed the app, generated a forecast, and found the visualisations clear. 10 of 11 users (90.9%) found the app helpful.

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