Ground-level ozone exceedances pose documented risks to human health and agricultural productivity, yet no operational early-warning system exists in South Africa. This study tests whether machine-learning models trained exclusively on South African Air Quality Information System (SAAQIS) hourly observations can forecast same-day ozone exceedances (>60 ppb) without atmospheric model output. We analysed 1.93 million hourly observations from 37 monitoring stations across five regions (Highveld, Vaal Triangle, Waterberg, Gauteng, Western Cape) for 2020–2025. Ozone missingness (~42%) was addressed with a three-tier imputation hierarchy, and 43 engineered features were derived per station-day across six groups (previous-day pollutants, previous-day meteorology, morning pollutants, morning meteorology, calendar/seasonal encodings, and episode persistence), with forecasts parameterised by cutoff c∈{prev_day,6,8,10,12}. We evaluated five model families (persistence baseline, ridge regression, logistic regression, random forest, XGBoost) on a temporally held-out test set (2024–2025) within the Predictability–Computability–Stability (PCS) framework. For regression, R-square improved from 0.40 (previous-day) to 0.71 (cutoff 12), with ridge explaining the most variance and random forest minimising mean absolute error. For classification, random forest was the best-balanced model (F1 = 0.83, accuracy = 0.95, AUC = 0.97 at cutoff 12), while XGBoost maximised recall (0.89 at cutoff 12), a key metric for public-health alerts. SHAP analysis identified a coherent set of dominant predictors including previous-day maximum ozone, morning maximum ozone, morning build-up slope, and ambient temperature and revealed a leadership inversion at cutoff 12, where same-day morning ozone surpassed previous-day maximum as the top predictor. Cutoff 10:00 emerged as the optimal operational setting, delivering R-square =0.53, F1 = 0.77, and recall up to 0.84 while preserving five hours of lead time before the typical 15:00 peak. All results were stable under PCS perturbations, demonstrating the feasibility of an observation-driven ozone early-warning prototype for South Africa.

Forecasting Ground-Level Ozone Exceedances Using Same-Day and Previous-Day Air Quality Measurements

Muofhe Masikhwa, **Clerence** Mathonsi,

INTRO

- South Africa currently lacks an operational ozone early-warning system, despite the Highveld being one of the world's most severe NOx pollution hotspots.
- Emissions from coal-fired power stations, biogenic VOCs and seasonal biomass burning, combined with persistent anticyclonic circulation, drive routine ozone exceedances that put communities at risk.

DATA

1. **Dataset:** 1.93 million hourly observations (2020–2025) from 37 SAAQIS stations across five regions (Highveld, Vaal Triangle, Waterberg, Gauteng, Western Cape).
2. **Station & Variables:** Greedy selection from 43 to 37 stations with common coverage; nine variables (O_3, NO, NO_2, NO_x, PM_{10}, SO_2, temperature, wind speed, wind direction).
3. **Feature Engineering:** 43 features per station-day across six groups (lagged pollutants, lagged meteorology, morning pollutants, morning meteorology, calendar/seasonal encodings, persistence).

METHODOLOGY

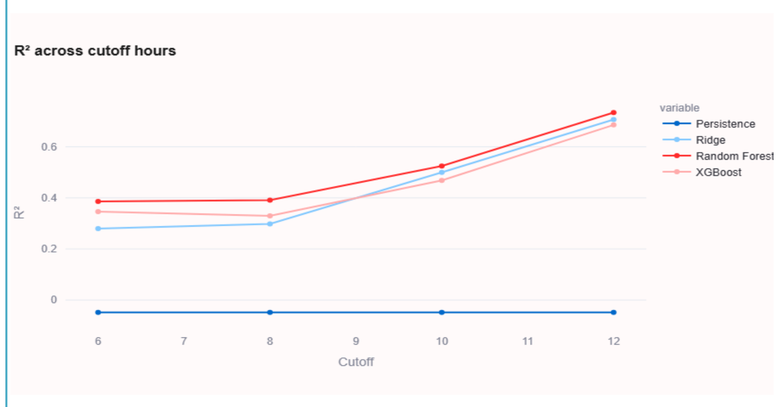
We tested five forecast and classification cutoffs (previous day, 06:00, 08:00, 10:00, 12:00), each denoting the final morning hour at which the forecast is produced.

Regression
1. Persistence Model(Baseline)
2. Ridge Regression
3. Random Forest
4. XGBoost

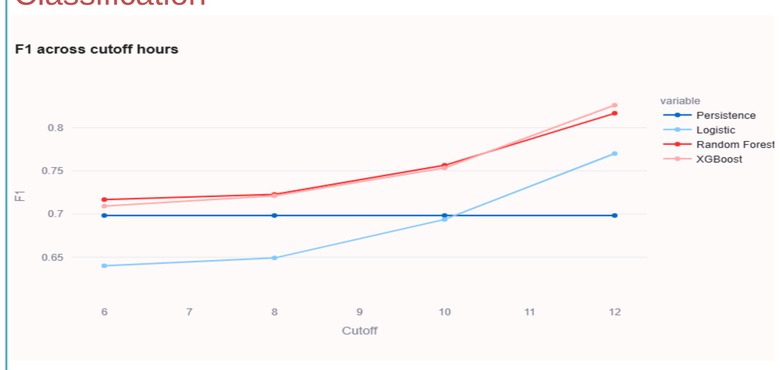
Classification
1. Persistence Model(Baseline)
2. Logistic Regression
3. Random Forest
4. XGBoost

RESULTS

Regression



Classification



DISCUSSION

Regression results: All models beat the persistence baseline. Performance improved with later cut-offs with best R-square rising from 0.40 (previous-day) to 0.71 (cut-off 12).

Classification results: Random forest was the most balanced model, achieving F1 = 0.83, accuracy = 0.95, precision = 0.85, and AUC = 0.97 at cutoff 12.

Model trade-offs: XGBoost consistently delivered the highest recall (0.81 to 0.89), making it preferable when avoiding missed exceedances is more important than limiting false alarms.

PCS ANALYSIS

Predictability was enforced by evaluating all five model families on a temporally held-out test set (2024–2025) and using Time Series Split cross-validation during training to prevent information leakage.

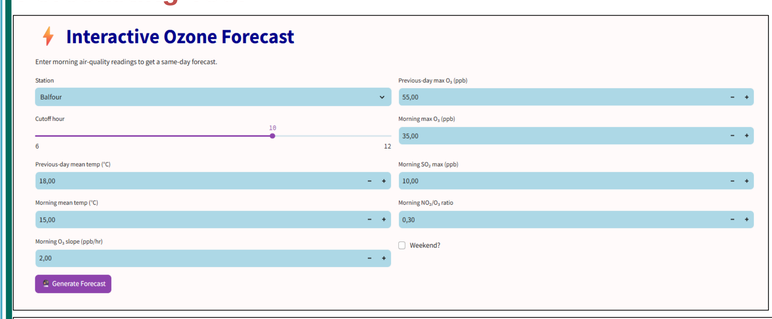
Computability was observed by ensuring the pipeline is tractable for full analysis and repeated testing. The complete workflow including data cleaning, three-tier imputation, feature engineering across five cutoffs, training five model families, and computing SHAP values executes in under 30 minutes on commodity hardware, making comprehensive stability experiments and routine re-runs feasible.

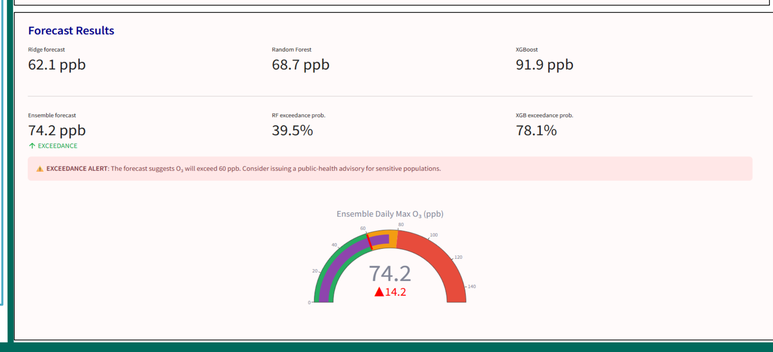
Stability was tested by assessing whether results hold under reasonable perturbations to data, models, and analyst choices. We applied four perturbation axes:

- data perturbation via bootstrap resampling and station subsampling;
- imputation sensitivity by comparing observed-only versus fully imputed datasets;
- algorithm switching across all five model families; and
- human-judgment sensitivity by varying the exceedance threshold (50, 60, 80, 100 ppb).

Deployment Artefacts

Forecasting Tool





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