Affordable Air Quality Prediction for Resource-Constrained Regions in Africa.

Authors:

- Yiga Gibert
- Mubahood Mubaraka

Presented by:

Group N

Makerere University, Kampala Uganda.

Contents of Presentation

- Introduction and Problem Area
- Research Problem Definition
- Research Aims & Objectives
- Research Approach/Methodology
- Results and Discussion
- Model Implementation
- Performance Evaluation
- XAI
- Limitations
- Recommendations
- Future Works

Introduction and Problem Area

- Air quality monitoring and prediction are essential for public health and environmental planning, especially in developing regions.
- PM2.5 and other pollutants significantly impact health outcomes, but traditional air quality sensors are expensive and not widely accessible in low-resource areas.
- This research aims to develop an affordable, accurate air quality prediction model for resource-constrained regions in Africa using alternative data sources.

Research Problem Definition

Problem Statement: How can machine learning be leveraged to accurately predict air pollutant levels (specifically PM2.5) in resource-constrained regions of Africa using cost-effective, alternative data sources?

Key Challenges:

- Lack of comprehensive air quality monitoring infrastructure, leading to high-dimensional data with potentially missing values.
- Identifying the most relevant features from alternative data sources (e.g., meteorological conditions, urban density) to accurately predict air pollutant levels.
- Developing a robust and computationally efficient predictive model that can be easily deployed in low-resource settings.

Research Aims & Objectives

Aim: Develop an accurate and affordable machine learning-based model to predict air pollutant levels, specifically PM2.5, in resource-constrained regions of Africa using alternative data sources.

Objectives:

- Conduct comprehensive exploratory data analysis (EDA) to identify key patterns and features from alternative data sources (e.g., weather conditions, urban characteristics, satellite imagery).
- Apply advanced feature engineering techniques to create robust and informative predictors from the available data.
- Evaluate and optimize multiple machine learning models, including ensemble methods, to determine the best-performing predictor of air pollution levels.
- Assess the model's performance, identify limitations, and provide recommendations for future improvements and real-world deployment.

Research Approach/Methodology

- Data Collection: Gather data from various sources, including meteorological conditions, urban features, and satellite imagery, to compensate for the lack of comprehensive air quality monitoring infrastructure.
- Exploratory Data Analysis (EDA): Analyze the data to identify key patterns, correlations, and seasonal trends that can inform feature selection and engineering.
- Feature Engineering: Create advanced features, such as lagged, cyclical, and interaction terms, to capture the complex relationships between air pollutant levels and the available predictors.

Research Approach/Methodology

- Methodology covers data processing steps, including loading, merging, and feature engineering (e.g., time-based features, interactions).
- Uses models like RandomForest, GradientBoosting, XGBoost, and LightGBM, with standardization and custom ensemble methods for robustness.

Results and Discussion:

- Data is split using time-series validation (TimeSeriesSplit), and performance metrics are computed for each model (RMSE, MAE, R2).
- Ensemble results, especially from stacking and voting regressors, are likely among the findings.

Model Implementation:

- Includes classes for stacking and ensemble model creation, with models like VotingRegressor and custom TimeSeriesStackingRegressor.
- SHAP and LIME explainability tools are incorporated for enhanced interpretability.

Performance Evaluation:

- Metrics such as RMSE, MAE, and R2 are computed across models to gauge effectiveness, including cross-validation for reliability.

Performance Evaluation:

	RMSE Mean	RMSE Std	MAE Mean	MAE Std	R2 Mean	R2 Std
rf	20.244591	12.121096	0.253646	2.116780	1.506905	0.062062
gbm	19.519225	11.613928	0.307073	2.230457	1.594060	0.064196
xgb	19.643412	11.579694	0.295052	1.957154	1.583866	0.074259
lgbm	19.911628	12.395239	0.273158	1.904234	1.013207	0.090711
voting	19.315583	11.490802	0.320426	2.081509	1.448573	0.062366
stacking	20.653994	12.672833	0.221088	2.402298	1.113261	0.102814

XAI (Explainable AI):

- SHAP and LIME explanations are generated for models, providing interpretability of feature importance and individual predictions.

XAI (Explainable AI):

XAI (Explainable AI):

XAI (Explainable AI):

XAI (Explainable AI):

XAI (Explainable AI):