Seasonal Pollution Episode Prediction

Contributors: Nishal Save • Shantanu Ghaisas • Sanmati Sawalwade

🎯 Problem Statement

North Indian cities experience severe air pollution during winter months (Oct-Feb) due to stubble burning, meteorological conditions, and festival emissions. This project develops an ML-based early warning system to predict severe pollution episodes.

Delhi | Lucknow | Patna | Chandigarh | Gurugram | Jaipur | Amritsar

📊 Dataset Overview

Samples: 200 records | **Features:** 15 predictive features

Target: Binary classification (Normal: 55.5%, Severe: 44.5%)

Period: Winter seasons 2019-2024

temperature_c	humidity_pct	wind_speed
wind_direction	fire_count_punjab	fire_count_haryana
previous_day_aqi	days_from_diwali	pollution_momentum

🤖 Model Comparison & Selection

Model	Train Accuracy	Test Accuracy	Recall (Severe)	F1-Score	Overfitting Gap	Status
Decision Tree (Original)	93.75%	42.50%	39%	0.378	51.25%	High Overfit
Decision Tree (Tuned) ✓	**76.88%**	**55.00%**	**56%**	**0.526**	**21.88%**	Best Model
Random Forest (Tuned)	90.62%	37.50%	28%	0.286	53.12%	Poor
XGBoost (Tuned)	100.00%	40.00%	33%	0.333	60.00%	Severe Overfit

📈 Best Model Performance (Tuned Decision Tree)

55%	56%	50%	0.526	0.521
Test Accuracy	Recall (Severe)	Precision	F1-Score	AUC-ROC

Hyperparameters: max_depth=7, min_samples_split=30, min_samples_leaf=5

🔍 Key Observations & Findings

1. Feature Importance

Wind direction (15.4%), pollution momentum (10.8%), and wind speed (9.5%) are the top predictors. Weather conditions dominate over fire counts.

2. Model Selection

Simpler models (Decision Tree) outperformed complex ensembles (RF, XGBoost) due to limited dataset size (200 samples).

3. Hyperparameter Impact

Tuning improved Decision Tree from 42.5% → 55% accuracy and reduced overfitting gap from 51% → 22%.

4. Recall Priority

Model detects 56% of severe episodes (10/18). Missing severe episodes is critical for public health warnings.

5. Engineered Features

Pollution momentum, stubble impact score, and inversion risk add predictive value beyond raw measurements.

6. Limitations

Small dataset constrains model learning. More data would improve ensemble methods significantly.

📦 Project Deliverables

- **01_Data_Collection.ipynb** - Data preprocessing & feature engineering
- **02_Model_Training.ipynb** - Model training, tuning & evaluation
- **03_Application.py** - Interactive Streamlit web app
- **04_Prediction_Demo.ipynb** - Prediction demonstration
- **models/best_model.pkl** - Trained model file

💡 Recommendations

- Collect more training data (target 500+ samples)
- Integrate real-time weather & fire APIs
- Add satellite imagery features
- Combine ML with meteorologist expertise
- Deploy as public health warning system
- Implement threshold tuning for higher recall