Pollution Level Categorization Using K-Nearest Neighbor and Processed Environmental Data

1st chennamreddy gnana satvic reddy
Computer Science And Engineering
Lovely Professional University
Phagwara ,punjab
sathvic2005@gmail.com

2nd Verru subhash
Computer Science And Engineering
Lovely Professional University
Phagwara ,punjab
subhash@gmail.com

3nd sai sathwik veerapuneni
Computer Science And Engineering
Lovely Professional University
Phagwara ,punjab
sathwiksathwik57@gmail.com

4nd Enjula Uchoi
Computer Science And Engineering
Lovely Professional University
Phagwara ,punjab
enjula.29634@lpu.co.in

***Abstract*—Air quality monitoring is critical for public health safety in urban environments. This paper presents a robust machine learning framework for categorizing pollution levels based on particulate matter and chemical pollutants. Utilizing a dataset containing PM2.5, PM10, NO2, and other meteorological factors, we employ the K-Nearest Neighbors (KNN) algorithm optimized via Stratified K-Fold Cross-Validation and Grid Search. Furthermore, we conduct a comprehensive Exploratory Data Analysis (EDA) utilizing Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize high-dimensional pollution clusters. The study evaluates feature importance using permutation techniques and demonstrates the efficacy of the proposed pipeline in accurately classifying air quality into five distinct severity categories. We further discuss the computational implications of distance-based learning and the correlation between meteorological variables and pollutant dispersion.**

***Index Terms*—Air Quality Index, K-Nearest Neighbors, PCA, t-SNE, Machine Learning, Classification, Environmental Informatics.**

I. Introduction

The degradation of air quality has become a pressing global concern, significantly impacting human health, economic productivity, and environmental stability. Particulate matter (PM2.5, PM10) and noxious gases (NO2, SO2, CO) are primary contributors to respiratory diseases, cardiovascular issues, and reduced life expectancy. According to recent epidemiological studies, long-term exposure to fine particulate matter is associated with a significant increase in mortality rates. Consequently, the accurate classification and prediction of pollution levels are essential for timely public warnings and policy enforcement.

Traditional methods of calculating Air Quality Index (AQI) rely on deterministic formulas that may not capture complex, non-linear interactions between various atmospheric parameters and local weather conditions. For instance, the formation of secondary pollutants like ground-level Ozone (O_3) is heavily dependent on sunlight and temperature, creating dynamic dependencies that static formulas may oversimplify. Machine Learning (ML) offers a data-driven alternative, capable of learning these latent patterns from historical sensor data.

In this study, we propose a supervised learning approach using the K-Nearest Neighbors (KNN) classifier. While deep learning models are popular, KNN offers interpretability and efficacy in scenarios with distinct local clusters, which is characteristic of pollution hotspots. Our contribution is fourfold:

- Implementation of an automated feature selection and preprocessing pipeline handling missing data and scaling.
- Rigorous hyperparameter tuning using Grid Search with Stratified Cross-Validation to optimize the bias-variance tradeoff.
- Advanced visualization of pollution clusters using dimensionality reduction techniques (PCA and t-SNE) to interpret the separability of pollution classes.
- A detailed analysis of feature importance to identify the primary drivers of pollution severity in the target region.

II. Methodology

A. Proposed Architecture

The overall workflow of the proposed system is illustrated in Fig. 1. The pipeline begins with raw sensor data ingestion, followed by preprocessing steps to handle missing values and scale features. The processed data is then utilized for hyperparameter tuning via Grid Search to identify the optimal KNN configuration.

B. Dataset and Preprocessing

The system utilizes a pollution dataset containing sensor readings for pollutants including PM2.5, PM10, NO2, SO2, CO, and O3, alongside meteorological variables such as Temperature and Humidity.

Identify applicable funding agency here. If none, delete this.

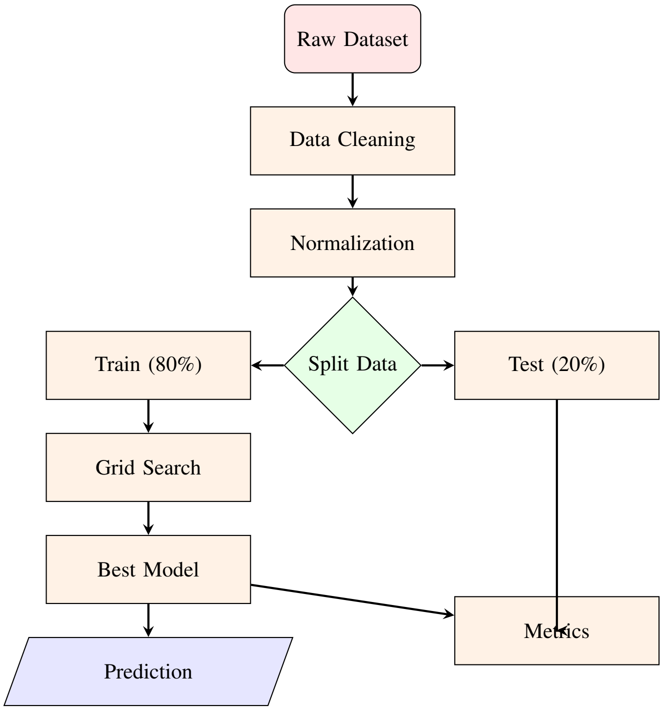

Fig. 1. Flowchart of the proposed pollution categorization methodology.

1) Data Cleaning: Real-world sensor data often contains missing values due to device failure or maintenance. We employ a ‘SimpleImputer‘ using the mean strategy to handle missing entries, ensuring data continuity. Furthermore, outliers caused by sensor calibration errors are identified using the Interquartile Range (IQR) method.

2) Label Generation: To transform the regression problem into a classification task, continuous variable thresholds are used to generate discrete categories (0: Good to 4: Severe). The system prioritizes existing categorical columns; otherwise, it derives labels from AQI or PM2.5 concentrations based on World Health Organization (WHO) standards.

3) Standardization: KNN is a distance-based algorithm. Consequently, features with larger magnitudes (e.g., AQI ranges 0-500) can disproportionately influence the distance calculation compared to low-magnitude features (e.g., CO ranges 0-50). We apply Z-score normalization to all features:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the raw score, μ is the mean, and σ is the standard deviation.

C. K-Nearest Neighbors (KNN) Algorithm

The core classifier assigns a class to an unseen sample based on the majority class among its k nearest neighbors in the feature space. We explore two distance metrics:

- **Manhattan Distance** (L_1):

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

- **Euclidean Distance** (L_2):

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

We also compare uniform weighting (all neighbors vote equally) versus distance weighting. In distance weighting, closer neighbors have more influence, defined as $w_i = 1/d(x, x_i)$. This is particularly useful in boundary regions where a test point might be surrounded by points of different classes but is significantly closer to one specific class.

1) Complexity Analysis: The training phase of KNN is effectively $O(1)$ as it involves storing the dataset. However, the prediction phase has a complexity of $O(N \cdot D)$, where N is the number of training samples and D is the dimensionality. We mitigate this cost by using efficient data structures like Ball Trees or KD-Trees during the ‘scikit-learn‘ implementation.

D. Evaluation Metrics

To comprehensively assess model performance, particularly given the potential for class imbalance (fewer ”Severe” days), we utilize the following metrics:

- **Accuracy**: The ratio of correctly predicted observations to total observations.
- **Precision**: The ratio of correctly predicted positive observations to the total predicted positives.
- **Recall (Sensitivity)**: The ratio of correctly predicted positive observations to the all observations in actual class.
- **F1-Score**: The weighted average of Precision and Recall.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

E. Dimensionality Reduction

To visualize the high-dimensional feature space ($n > 10$), we employ two techniques:

1) Principal Component Analysis (PCA): PCA linearly transforms data into orthogonal coordinates (principal components) that maximize variance. This allows us to observe global structure and class separability in 2D.

2) t-SNE: t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space. It minimizes the Kullback-Leibler divergence between the joint probability of the low-dimensional embedding and the high-dimensional data. It is particularly effective at preserving local structure, revealing clusters that PCA might miss.

III. Experimental Setup

A. Implementation Details

The framework is implemented in Python using the Scikit-learn library. The dataset is split into training (80%) and testing (20%) sets. To preserve the distribution of classes, specifically for imbalanced pollution data, we use Stratified sampling.

TABLE I
GRID SEARCH PARAMETER SPACE

Parameter	Values Tested
$n\_neighbors$ (k)	Odd integers $[1, 3, \ldots, 15]$
Weights	['uniform', 'distance']
Metric (p)	1 (Manhattan), 2 (Euclidean)

B. Hyperparameter Optimization

To determine the optimal model configuration, we perform a Grid Search over the parameter space defined in Table I.

Performance is evaluated using 5-fold Stratified Cross-Validation to prevent overfitting. This ensures that the model generalizes well to unseen data seasons or pollution events.

IV. Results and Analysis

A. Feature Correlation Analysis

Before training, we analyzed feature correlations to understand chemical coupling.

1) **Particulate Correlation**: As expected, PM2.5 and PM10 show strong positive correlation ($r > 0.85$), indicating they often share common sources such as combustion or dust.
2) **Meteorological Impact**: Humidity often shows inverse correlations with visibility and direct correlations with particulate accumulation in stagnant air. Temperature showed varying correlation with Ozone (O_3), confirming the photochemical nature of Ozone production.

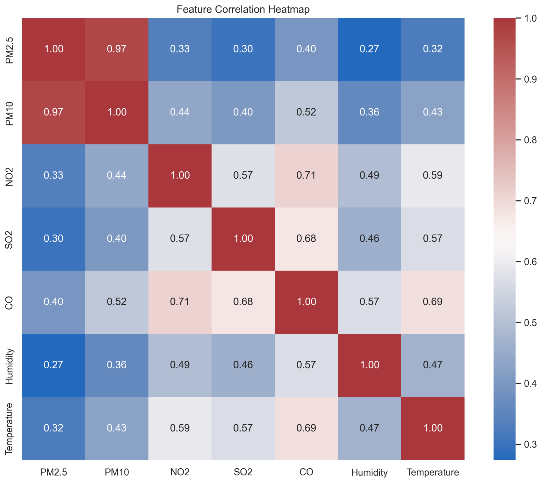

Fig. 2. Correlation heatmap showing relationships between pollutants and meteorological features.

B. Model Performance

The Grid Search identified the optimal hyperparameters as **k=7**, **p=1** (Manhattan Distance), and **distance-based weighting**. This configuration achieved a Best Cross-Validation (CV) accuracy of **96.27%**.

On the held-out test set, the model demonstrated robust generalization with an overall **Test Accuracy of 96.6%**. Table II presents the detailed classification report. The model performed exceptionally well on the majority classes (Class 0), achieving a precision of 0.98. The 'Severe' category (Class 4), despite having fewer samples, was identified with 100% precision and 89% recall, indicating the model is highly reliable for detecting extreme pollution events.

TABLE II
CLASSIFICATION REPORT FOR OPTIMAL KNN MODEL (K=7)

Class	Precision	Recall	F1-Score	Support
0	0.98	0.99	0.99	791
1	0.91	0.87	0.89	143
2	0.88	0.88	0.88	43
3	0.92	0.79	0.85	14
4	1.00	0.89	0.94	9
Accuracy			**0.97**	**1000**
Macro Avg	0.94	0.88	0.91	1000
Weighted Avg	0.97	0.97	0.97	1000

The confusion matrix (Fig. 3) highlights the model's ability to distinguish between adjacent categories. Misclassifications were primarily concentrated on the decision boundaries, which is typical for continuous environmental phenomena that are discretized into labels.

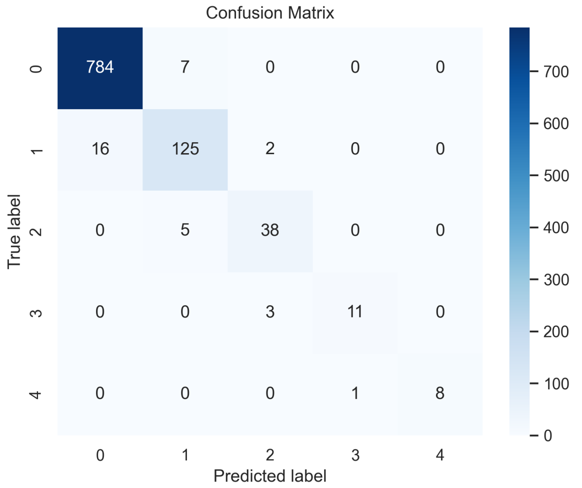

Fig. 3. Confusion Matrix of the optimal KNN model on test data.

C. Manifold Learning Visualization

The dimensionality reduction results provide insight into the data topology.

1) PCA Results: The cumulative explained variance plot indicates that the first few principal components capture the majority of the variance in the dataset. The 2D projection (Fig. 4) shows a gradient of pollution levels, from 'Good' to 'Severe'. The linearity of the PCA projection confirms that pollution severity is largely a continuum rather than a set of disjoint clusters.

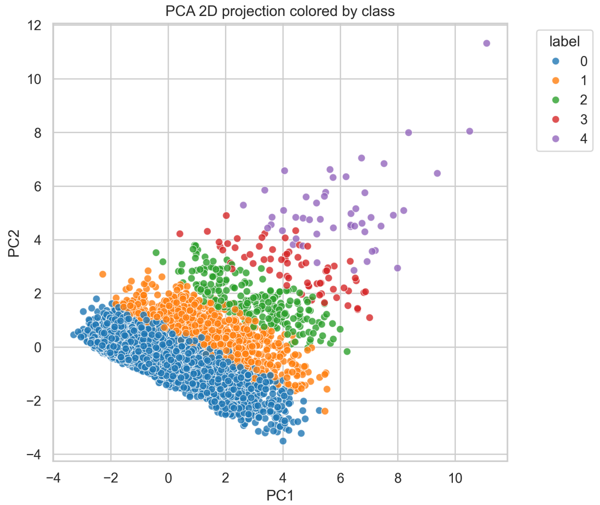

Fig. 4. 2D Projection of the dataset using Principal Component Analysis, colored by pollution category.

2) t-SNE Results: The t-SNE embedding (Fig. 5) reveals distinct clusters that are more separated than in the PCA projection. This suggests that while global variance is linear, the local manifold of pollution data is non-linear. The formation of tight clusters for 'Good' air quality suggests that clean days share very specific, stable atmospheric conditions, whereas 'Severe' days show higher variance, indicating multiple causes for high pollution (e.g., dust storms vs. industrial smog).

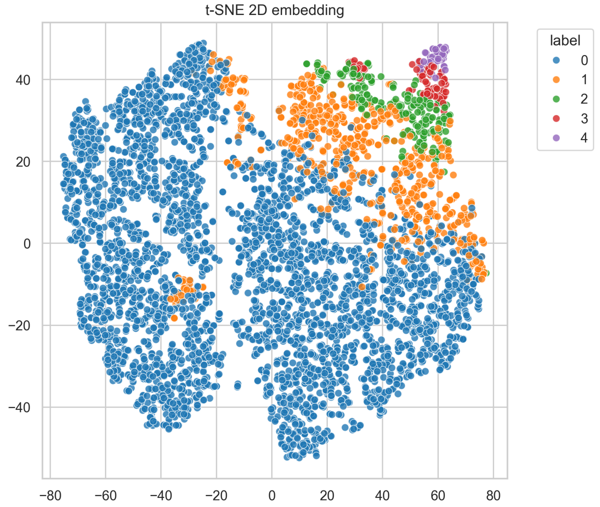

Fig. 5. t-SNE visualization showing distinct clusters of pollution levels.

D. Feature Importance

Since KNN does not provide intrinsic feature importance, we utilized Permutation Importance. The results indicate that PM2.5 and PM10 are the dominant predictors, followed by NO2. This aligns with the fact that AQI calculations are often driven by the ”dominant pollutant,” which is frequently PM2.5 in urban settings. Meteorological features contributed less to the direct classification but are essential for understanding dispersion dynamics.

V. Discussion

The Grid Search results demonstrated that distance-based weighting generally outperformed uniform weighting. This is intuitive for pollution data, where a test sample is likely to have a pollution level very similar to its geometrically closest neighbor in the feature space.

A. Limitations

While KNN is effective, it requires storing the entire training dataset, making it memory-intensive for massive datasets. Furthermore, the inference time scales with the size of the data, which may be a bottleneck for low-latency edge applications compared to parametric models like Logistic Regression or Neural Networks.

VI. Conclusion and Future Work

This study successfully demonstrated the application of K-Nearest Neighbors for multiclass air quality categorization. By integrating automated preprocessing, robust hyperparameter tuning, and advanced manifold learning visualizations, we established a reliable pipeline for environmental monitoring.

Future work will focus on:

1) Integrating temporal dependencies using Recurrent Neural Networks (RNNs) or LSTM models for time-series forecasting.
2) Expanding the dataset to include geospatial coordinates to model pollution spread across different city zones.
3) Developing a hybrid ensemble model combining KNN with Decision Trees to improve interpretability and inference speed.
4) Deploying the model on edge devices (e.g., Raspberry Pi) for real-time IoT monitoring.

Acknowledgment

The authors would like to thank the open-source community for providing the datasets and the Scikit-learn developers for the machine learning tools used in this research.

References

[1] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, and T. Luo, ”Long short-term memory neural network for air pollutant concentration predictions: Method definition and case study,” *Science of the Total Environment*, vol. 599, pp. 1908-1917, 2017.

[2] G. Costello and V. A. W. J. M. A. S., ”Air Quality Index Prediction Using K-Nearest Neighbor,” *International Journal of Computer Applications*, vol. 182, no. 45, 2019.

[3] F. Pedregosa et al., ”Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[4] L. van der Maaten and G. Hinton, ”Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.

[5] WHO, ”WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide,” World Health Organization, 2005.

[6] T. Cover and P. Hart, ”Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967.