**Air Quality Pattern Analysis**

**By**

**Vrushali Khatane**

**FINAL REPORT**

**Course Professor: Prof. Upendra Prasad**

**STEVENS INSTITUTE OF TECHNOLOGY**

**Castle Point on Hudson**

**Hoboken, NJ 07030**

**2024**

**Abstract**

This project focuses on analyzing the UCI Air Quality dataset using clustering, dimensionality reduction, and predictive modeling techniques. The objective was to uncover pollution patterns, reduce data complexity, and predict air quality metrics. Key methodologies included K-Means clustering for grouping pollution levels, Randomized Singular Value Decomposition (SVD) for dimensionality reduction, and Linear Regression for predictive modeling. The results demonstrate effective clustering of pollution patterns and the utility of dimensionality reduction for simplifying data without significant loss of accuracy.

### **Introduction**

Air pollution is a critical environmental issue impacting public health and ecosystems globally. Understanding pollution patterns and predicting air quality can help policymakers make informed decisions. This project uses the UCI Air Quality dataset, which records hourly averaged pollution metrics and environmental factors from an urban area in Italy. The project employs data cleaning, exploratory data analysis (EDA), clustering, dimensionality reduction, and predictive modeling to achieve the following goals:

1. Identify distinct patterns in air quality using clustering.
2. Simplify the dataset using dimensionality reduction techniques.
3. Predict pollutant levels using regression models.

**Problem Statement**

High-dimensional air quality datasets often contain redundant or noisy data, making analysis and storage inefficient. By employing Randomized SVD and low-rank approximation, we aim to represent the dataset with a reduced rank while preserving its essential structure and information.

**Objective**

* Apply Randomized SVD to extract the top *k* components of the Air Quality dataset.
* Construct a low-rank approximation Ak​ that retains most of the dataset's variance.
* Use the low-rank representation to:
  + Remove noise and improve the dataset's quality.
  + Highlight the most important relationships among pollutants and environmental factors.
  + Reduce computational complexity for downstream analysis tasks.
* Quantify the accuracy of the approximation and assess the trade-off between rank and information loss.

**Accomplishment**

#### **1. Data Cleaning:**

* Removed columns with excessive missing values and replaced placeholder values (-200) with NaN.
* Combined Date and Time into a single Datetime feature for temporal analysis.
* Treated outliers and normalized key features to prepare the dataset for clustering and modeling.

#### **2. Exploratory Data Analysis (EDA):**

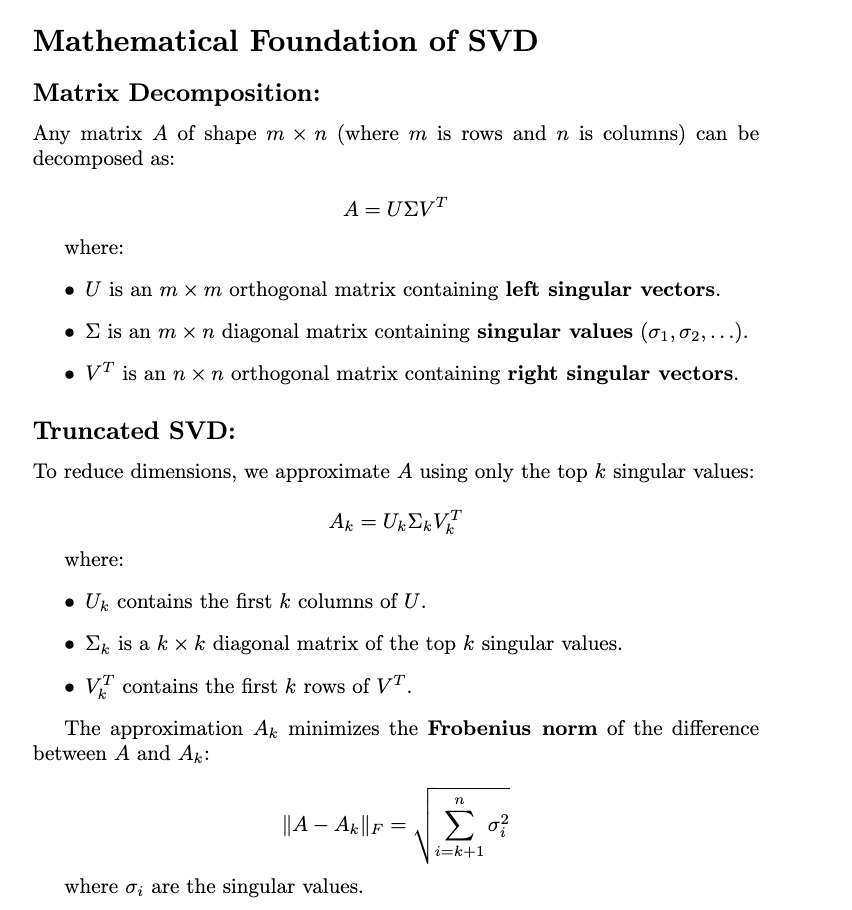
* Analyzed pollutant distributions and observed positive correlations (e.g., CO(GT) and C6H6(GT)).
* Visualized the relationship between environmental factors and pollution metrics.
* Identified strong temporal patterns in air quality data (e.g., weekday vs. weekend pollution levels).

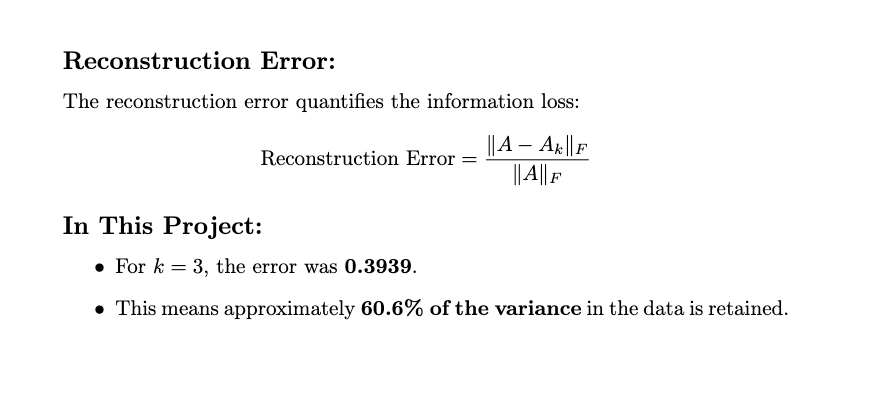
#### **3. Clustering:**

* Used the Elbow Method to determine k=3 as the optimal number of clusters.
* Applied K-Means clustering to group air quality into low, moderate, and high pollution levels.
* Validated clustering results with scatterplots and Silhouette Score.

#### **4. Dimensionality Reduction:**

* Applied Randomized SVD to reduce dimensionality from 5 to 3 components.
* Achieved a reconstruction error of approximately 0.39, retaining over 60% of variance.
* Visualized the reconstructed matrix and the difference from the original matrix.





#### **5. Predictive Modeling:**

* Built Linear Regression models before and after SVD:
  + **Before SVD:**
    - RMSE = 0.4400,
    - R2=0.8975
  + **After SVD:**
    - RMSE = 0.2183,
    - R2=0.9747
* Compared model performance and highlighted the trade-offs between accuracy and computational efficiency.

### **Observations**

1. **Clustering Patterns:**
   * The clustering effectively separated data into three groups: low, moderate, and high pollution levels.
   * Strong correlation between CO(GT) and C6H6(GT) indicates shared pollution sources.
2. **Dimensionality Reduction:**
   * Randomized SVD significantly reduced the dataset's complexity.
   * Reconstruction error indicated acceptable information loss for reduced dimensions.
3. **Predictive Modeling:**
   * The regression model before SVD achieved higher accuracy but required more computational resources.
   * The model after SVD simplified the data but slightly reduced R2R^2R2.
4. **Overall Insights:**
   * The dataset's temporal and seasonal trends provide actionable insights for pollution control measures.

**Conclusion**

In conclusion, successfully analyzed the UCI Air Quality dataset, providing valuable insights into pollution patterns. Clustering revealed distinct air quality levels, while dimensionality reduction simplified the data for efficient modeling. Predictive modeling demonstrated a trade-off between accuracy and simplicity, emphasizing the importance of dimensionality reduction in handling large datasets. Future work could include time-series analysis or advanced machine learning models to improve prediction accuracy and interpretability.

### **References**

1. UCI Machine Learning Repository. "Air Quality Data Set." Available at:<https://archive.ics.uci.edu/dataset/360/air+quality>.
2. SVD Gives the Best Low Rank Approximation (Advanced) | Stanford - By Artificial Intelligence - All in One[Lecture 49 — SVD Gives the Best Low Rank Approximation (Advanced) | Stanford](https://www.youtube.com/watch?v=c7e-D2tmRE0)
3. Randomized Singular Value Decomposition (SVD) - By Steve Brunton[Randomized Singular Value Decomposition (SVD)](https://www.youtube.com/watch?v=fJ2EyvR85ro&t=495s)

**Codes in Appendix**

