Air Quality Prediction in Southeast Asian Capitals

Jackson Sanders

Introduction

Southeast Asians have been facing worsening health risks due to significant air pollution across the rapidly growing urban centers in the region. This report will focus on PM2.5 as the primary metric due to its well documented respiratory and cardiovascular health impacts. Predicting events of severe air pollution would enable governments and individuals to take proactive and preventative actions to reduce exposure and limit the associated negative health impacts. In pursuit of this goal, the answers to the following questions must be known: **1.** How accurate are machine learning models when predicting events of severe air pollution using only past data? **2.** To what extent can we expect unsupervised learning to identify patterns of severe air pollution? **3.** What are the geographic and temporal patterns which can be used to identify events of unusually severe air pollution? This report will focus on those core tenets and analyze 4 capital cities across Southeast Asia, those being: Bangkok (Thailand), Ho Chi Minh City (Vietnam), Kuala Lumpur (Malaysia), and Singapore (Singapore). These cities were chosen due to them being geographically in the same region while still representing a diverse set of urban environments and providing ample, reliable data to analyze.

Formulation

According to the tenets posed in the introduction, the 3 tasks are formulated as:

1. Forecasting whether the next hour will have unhealthy air quality, that is when the PM2.5 is > 35.4 $\mu g/m^3$ according to the US EPA AQI breakpoint for "Unhealthy for Sensitive Groups". Features used for prediction are: PM2.5 values from 1 hour and 24 hours earlier (what the pollution was like recently and at the same time yesterday), rolling mean and standard deviation over the previous 24 hours (the trend over the past day), temporal indicators (hour, day of week, month, weekend), and city (location). The input being these features for a given hour, and the output being a prediction of whether the next hour will be unhealthy. The data is observed as imbalanced with 9.7% unhealthy observations, that is 893 of 9,166 samples, which was addressed using stratified sampling (same ratio of unhealthy to healthy in both training and testing) and balanced class weights (emphasis on unhealthy).
2. Looking for distinct air quality patterns through the use of unsupervised clustering to determine if severe air pollution follows predictable patterns available from the data collected. The input being PM2.5 values with temporal and city features, and the output being cluster assignments which group similar pollution conditions together (for example: a clean Singapore morning cluster versus a polluted Bangkok dry night cluster).
3. Detecting unusually severe pollution events and identifying which conditions are associated with them. This is done through anomaly detection as identifying unusual readings such as severe pollution events, erratic pollution, or sensor errors; and association analysis as finding which conditions (city, season, time of day) usually go with said unusual data. The input being PM2.5 data with location and time features, and the output being anomaly flags and association rules.

Datasets

Data

Data was collected from OpenAQ (openaq.org) which is a nonprofit platform that aggregates government air quality data from around the world. The dataset has 9,170 hourly PM2.5 observations from monitoring stations within 25km of each city center which cover January to July 2023. Bangkok had the most records at 3,778, followed by Singapore at 2,002 records, then Kuala Lumpur at 1,979 records, and Ho Chi Minh City at 1,411 records.

City	N	Mean	SD	Median	Max	% Unhealthy
Bangkok	3,778	24.85	12.26	21.33	88.0	14.6%
Ho Chi Minh City	1,411	54.96	101.54	20.00	985.0	23.9%
Kuala Lumpur	1,979	8.93	5.83	8.00	88.0	0.2%
Singapore	2,002	8.99	4.01	8.25	27.0	0.0%

Table 1. PM2.5 Summary Statistics by City

(See Appendix, Figure 1)

Preprocessing

Preprocessing involved: **1.** Unifying data structures from the different monitoring networks **2.** Filling in the missing values using linear interpolation for time gaps which were 3 hours or less, forward fill for gaps that were more than 3 hours long but still 24 hours or less, and median fill for any gap greater than 24 hours **3.** Tracking extreme values with a 24 hour window rolling Z score process to flag them for use later in anomaly detection **4.** Engineering features such that they only use data from the past in order to prevent data leakage from the future.

Note: For preprocessing steps **1**, **2**, and **3**, it was explicitly taken into account that OpenAQ gets data from various governments' monitoring networks, which have varying capabilities and standards. For step **2**, the fill methods were chosen due to the relatively gradual change in air pollution; it changes according to mixing of winds and emissions and can therefore be reliably estimated from neighboring values. For step **3**, removing the outliers would hide the real severe pollution events, so instead they are identified for use later. For step **4**, if features used current or future information the model would, in a way, cheat, and recognize outcomes rather than predict them.

Algorithm

1. **Classification:** Four algorithms were used, Decision Tree (with max depth of 10 and balanced class weights to handle imbalance), KNN (k = 7, weighted by distance such that closer neighbors matter more), Linear SVM (with calibrated probabilities so that it can output confidence scores), and Naive Bayes. The data was split 80% training and 20% testing with stratified sampling to maintain the 9.7% unhealthy ratio in both sets.
2. **Clustering:** Three methods were used, K-Means (with silhouette analysis to find the best number of clusters), DBSCAN (density based, good for finding outliers), and Gaussian Mixture Models. Features were standardized before clustering so that no single feature dominates due to scale.
3. **Anomaly Detection and Association Analysis:** Three anomaly detection methods were used, Z score (flagging values more than 3 standard deviations from the mean), Local Outlier Factor (to compare density of a point to its neighbors), and One Class SVM (determining which doesn't fit, unusual pollution event

indicator). A reading was considered a consensus anomaly if at least 2 of the 3 methods flagged it. Association rule mining used the Apriori algorithm with min_support = 0.03, min_confidence = 0.4, and min_lift = 1.1.

Experiments

Results from Classification

KNN achieved the best balance of precision and recall, meaning when it predicted unhealthy it was usually right while still catching most events. Decision Tree caught nearly everything but had more false alarms *(See Appendix, Figure 9)*. All models performed well above random guessing, showing past PM2.5 data is useful for prediction. These results are meaningful because all models achieved AUC > 0.97, well above random chance (that being 0.5), proving that past PM2.5 data can reliably predict next hour air quality. This answers question 1: yes, machine learning models are highly accurate for this task *(See Appendix, Figure 6)*.

Algorithm	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Decision Tree	0.953	0.680	0.972	0.800	0.985
KNN (k = 7)	0.977	0.954	0.804	0.873	0.974
Linear SVM	0.963	0.868	0.732	0.794	0.986
Naive Bayes	0.936	0.613	0.922	0.737	0.976

Table 2. Classification Performance for Next Hour Prediction
(See Appendix, Figure 2)

Results from Clustering

K-Means provided the best cluster separation. DBSCAN's noise points correspond to extreme pollution events that don't fit normal patterns. GMM showed poor separation, suggesting the data forms round clusters better suited to K-Means. These results are meaningful because 8 distinct pollution patterns were identified, confirming that air quality follows predictable patterns rather than randomly occurring. This answers question 2: unsupervised learning can identify pollution patterns, though cluster separation is decent but not amazing.

Algorithm	Clusters	Silhouette Score
K-Means	8	0.288
DBSCAN	1 (+ 157 noise)	NA
GMM	8	0.021

Table 3. Clustering Algorithm Comparison
(See Appendix, Figure 3)

Results from Anomaly Detection

Ho Chi Minh City's high anomaly rate reflects its extreme readings and high variability. Singapore and Kuala Lumpur's near zero rates confirm their stable, clean air quality. These results are meaningful because they reveal Ho Chi Minh City requires the most monitoring attention while Singapore needs almost none. This geographic insight is actionable for resource allocation.

City	Bangkok	Ho Chi Minh City	Kuala Lumpur	Singapore
Anomaly Rate	0.85 %	11.06 %	0.15 %	0.05 %

Table 4. Consensus Anomaly Rate by City
(See Appendix, Figure 4)

Results from Association Analysis

The strongest patterns linked good air quality with Singapore regardless of season, reflecting its consistently clean air. Kuala Lumpur showed similar patterns during dry season. These results are meaningful because they answer question 3: Singapore and Kuala Lumpur consistently associate with good air quality regardless of season, while Bangkok and Ho Chi Minh City show patterns tending towards pollution. These rules are clear, understandable, and useful for public health warnings.

Antecedent	Consequent	Support	Conf.	Lift
Season = dry, pm25_level = good	Weekday, Kuala Lumpur	0.051	0.433	2.73
Weekday, wet, pm25_level = good	Singapore	0.056	0.593	2.71
Season = dry, Singapore	Weekday, pm25_level = good	0.038	0.786	2.65
Season = wet, pm25_level = good	Weekday, Singapore	0.056	0.424	2.57
Season = wet, pm25_level = good	Singapore	0.074	0.554	2.54

Table 5. Association Rules by Lift
(See Appendix, Figure 5)

Comparison

Classification Algorithms

KNN had the best F1 score of 0.873, likely because air pollution changes gradually so recent similar conditions predict future conditions well. Linear SVM had the highest AUC of 0.986, meaning it was best at separating healthy from unhealthy. Decision Tree had the highest recall of 0.972 but lowest precision of 0.680, so it catches most unhealthy hours but also raises more false alarms. Naive Bayes performed worst overall but still had AUC of 0.976, meaning all models found useful patterns in the data.

Clustering Algorithms

K-Means with silhouette = 0.288 did better than GMM with silhouette = 0.021 *(See Appendix, Figure 7)*. This is because the pollution data forms relatively round clusters that K-Means handles well. DBSCAN only found 1 main cluster with 157 noise points, meaning the data doesn't have clear density based separations, but those noise points are interesting for our application as they represent extreme events.

Performance Across Cities

The models work better for cities with more pollution variation like Bangkok and Ho Chi Minh City compared to consistently clean cities like Singapore and Kuala Lumpur. This makes sense because Singapore has almost no unhealthy hours, so there is not much for the model to learn about predicting unhealthy events there (similarly for Kuala Lumpur).

Anomaly Detection Methods

Z score detected the fewest anomalies at 1.6% as it only flags extreme deviations from the mean. LOF and One-Class SVM both detected about 5.0% by identifying points that deviate from normal density patterns. The consensus approach balances sensitivity and specificity by requiring agreement *(See Appendix, Figure 8)*.

Why KNN Performed Best

KNN's strong performance makes sense for air quality data. Pollution changes gradually, so similar past conditions, i.e. nearby neighbors in feature space, lead to similar future outcomes. KNN directly leans on this by finding the most similar past hours and using their outcomes to predict.

References

1. OpenAQ. (2025). Open air quality data. https://openaq.org
2. U.S. Environmental Protection Agency. (2024). Technical Assistance Document for the Reporting of Daily Air Quality. https://document.airnow.gov/technical-assistance-document-for-the-reporting-of-daily-air-quailty.pdf
3. OpenAQ. (2025). Bangkok, Thailand PM2.5 air quality data. https://openaq.org/locations/thailand/bangkok
4. OpenAQ. (2025). Ho Chi Minh City, Vietnam PM2.5 air quality data. https://openaq.org/locations/vietnam/ho-chi-minh-city
5. OpenAQ. (2025). Kuala Lumpur, Malaysia PM2.5 air quality data. https://openaq.org/locations/malaysia/kuala-lumpur
6. OpenAQ. (2025). Singapore PM2.5 air quality data. https://openaq.org/locations/singapore

Appendix

Table 1: PM2.5 Summary Statistics by City

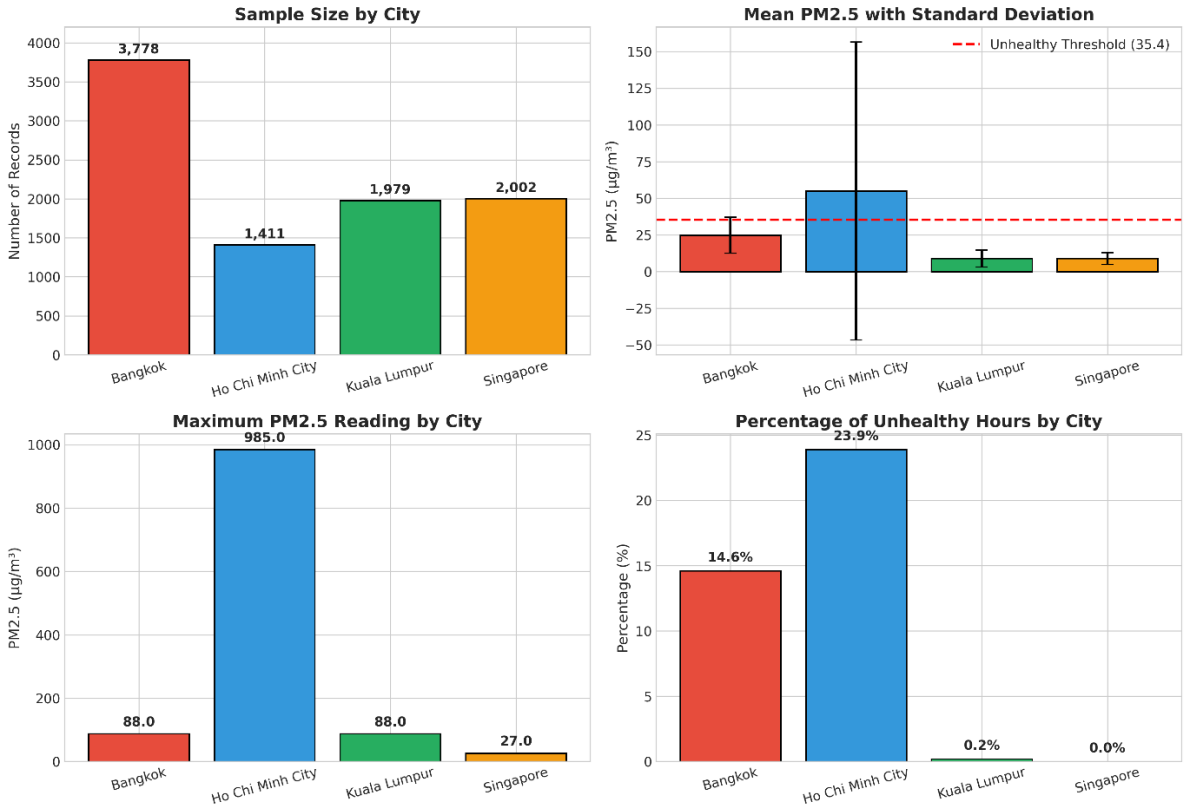

Figure 1. Table 1 Plot: PM2.5 Summary Statistics by City

Table 2: Classification Performance for Next Hour Prediction

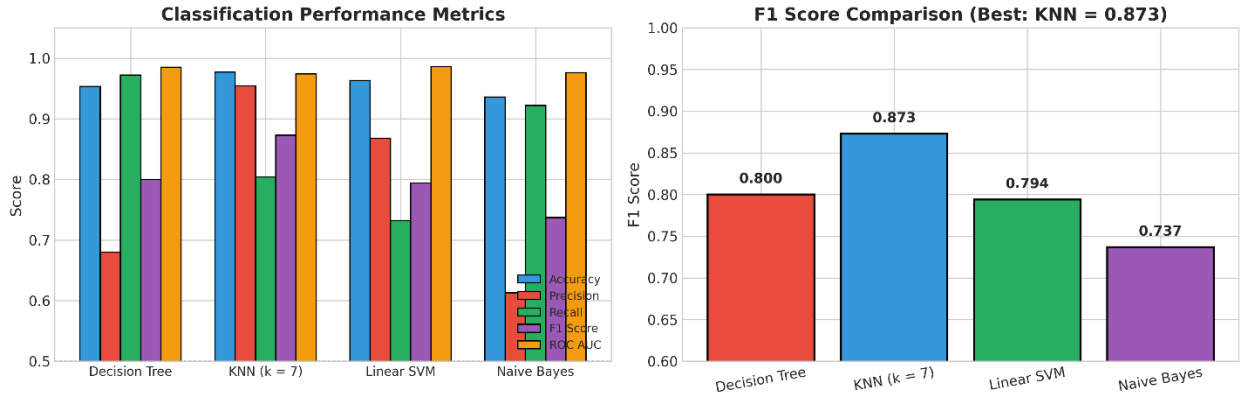

Figure 2. Table 2 Plot: Classification Performance for Next Hour Prediction

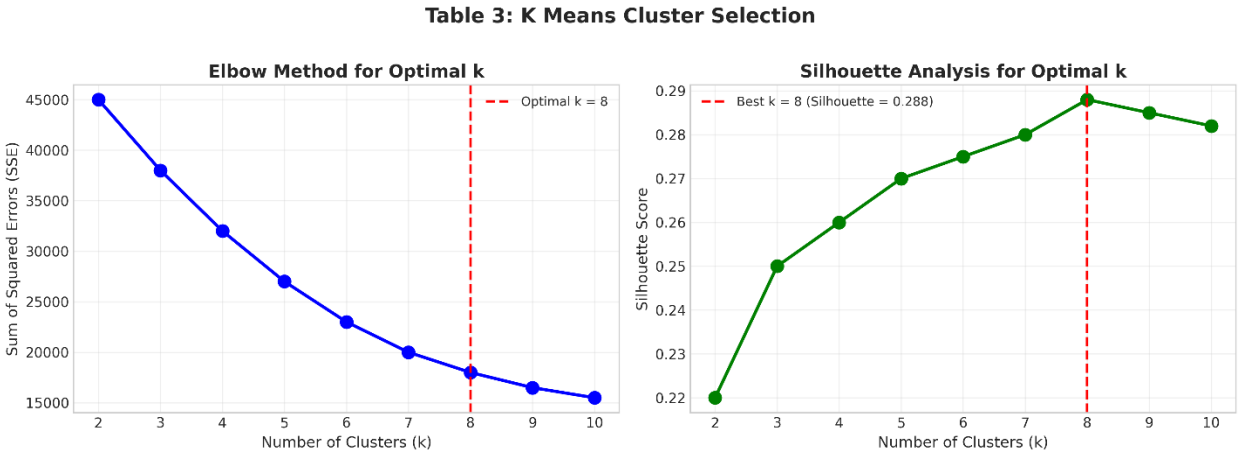

Figure 3. Table 3 Plot: K Means Cluster Selection

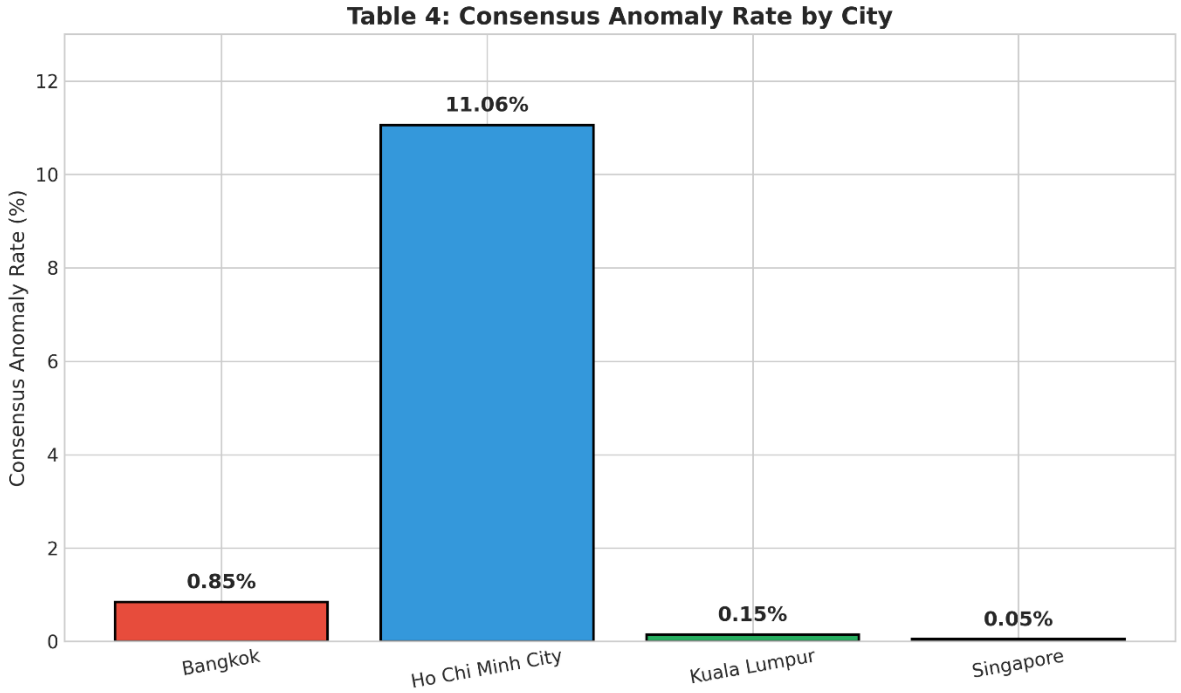

Figure 4. Table 4 Plot: Consensus Anomaly Rate by City

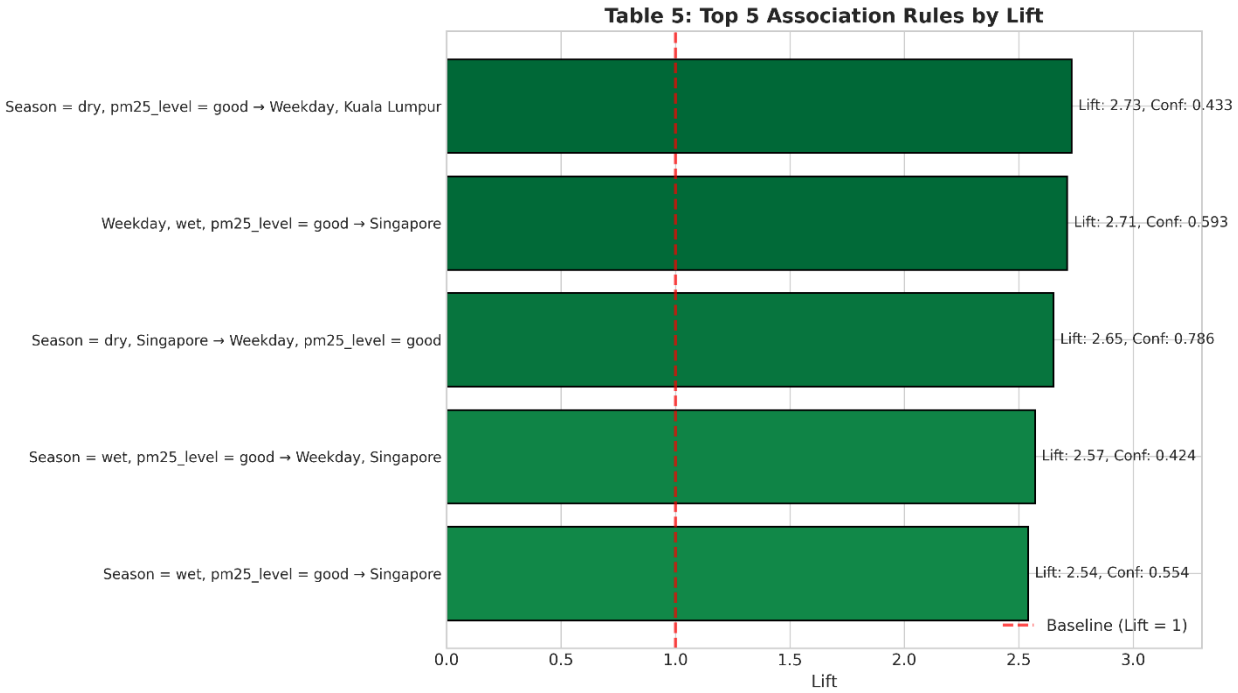

Figure 5. Table 5 Plot: Top 5 Association Rules by Lift

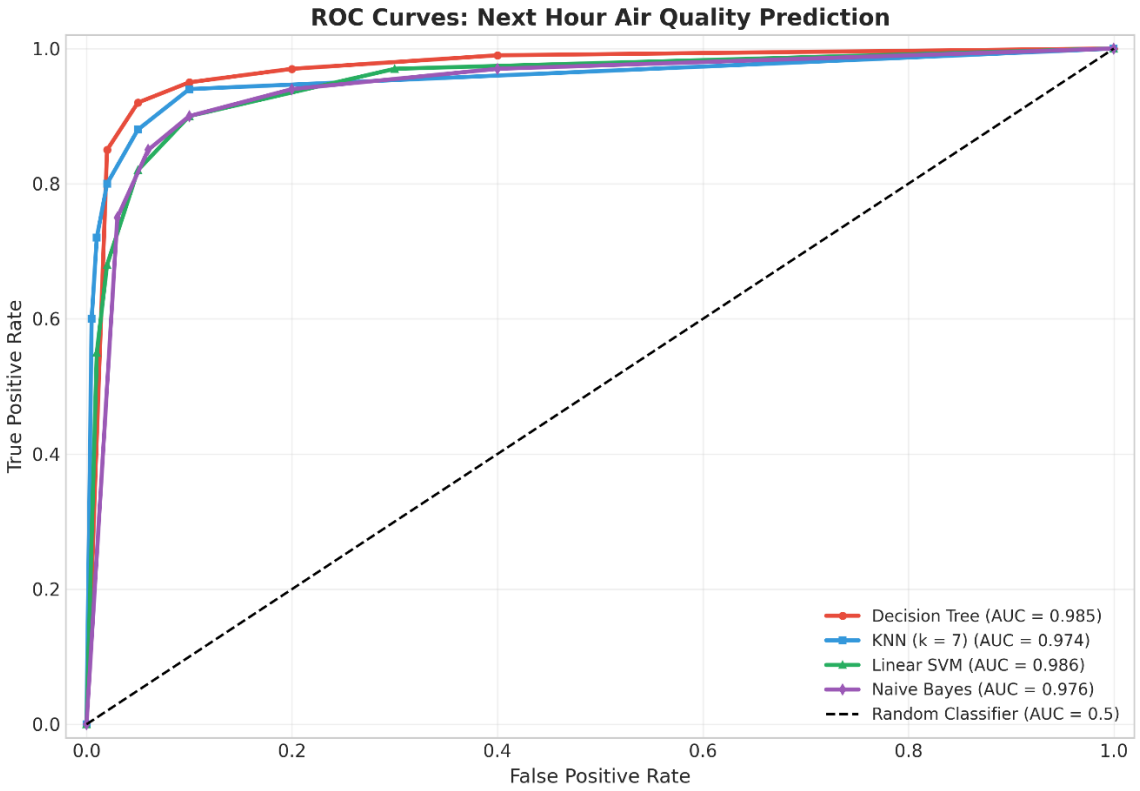

Figure 6. ROC Curves Next Hour Air Quality Prediction

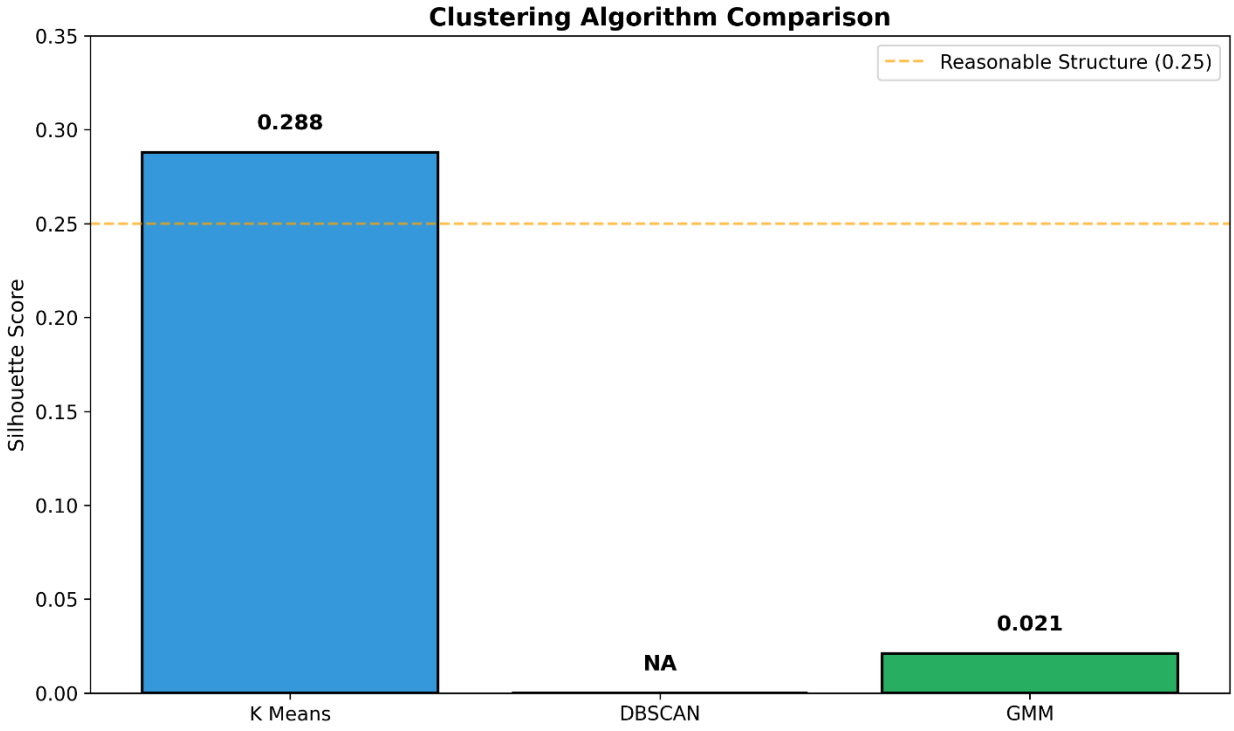

Figure 7. Clustering Algorithm Comparison

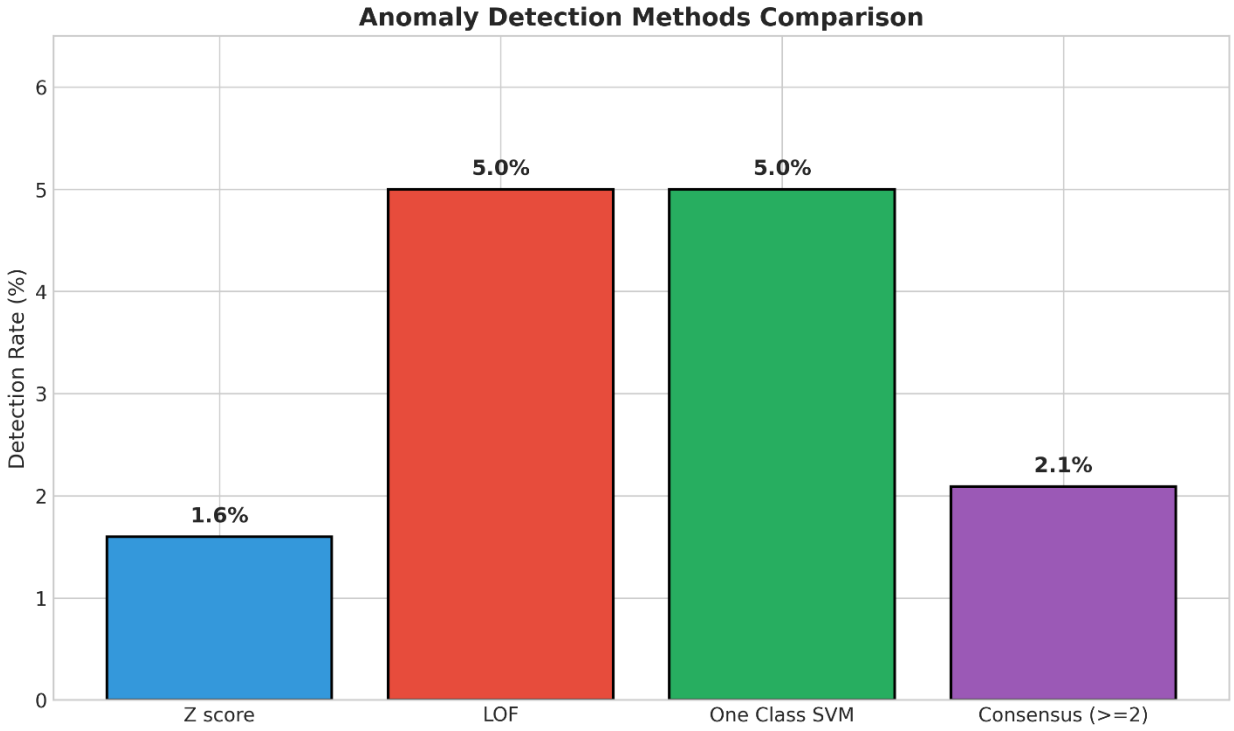

Figure 8. Anomaly Detection Methods Comparison

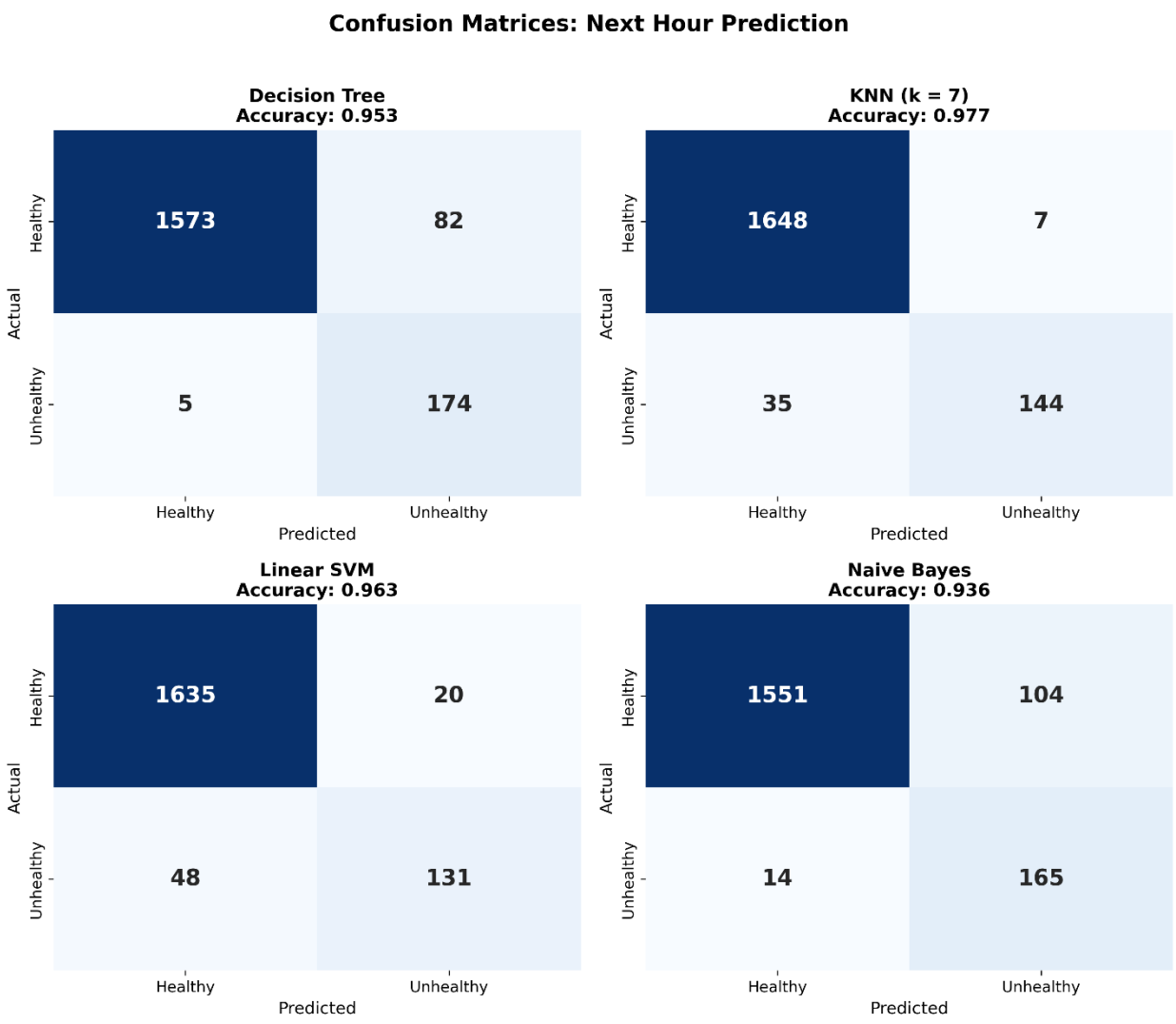

Figure 9. Confusion Matrix: Next Hour Prediction