Faculty of Engineering & Technology

Electrical & Computer Engineering Department

ENCS5341 Machine Learning and Data Science

Assignment #3

Prepared by:

Amany Hmidan 1200255

Hiba Jaouni 1201154

Instructor: Dr. Ismail Khater

Section: 3

Date: Dec 28th, 2024

1. Abstract

This assignment investigates a range of machine learning techniques for classification, including K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machines (SVM) with various kernels, and ensemble methods such as Boosting and Bagging. The primary objective is to classify air quality levels based on environmental and demographic features while evaluating model performance using metrics such as precision, recall, and F1-score across all air quality categories. The study emphasizes the trade-offs between accuracy and computational complexity, assessing each model's capability to address challenges such as class imbalance and non-linear relationships within the dataset.

2. Dataset Description

The dataset is designed to analyse air pollution and predict air quality levels in a specific region. It encompasses a diverse set of environmental, industrial, and demographic variables, making it a comprehensive resource for studying pollution patterns and their impact on public health. The target variable, Air Quality Levels, is classified into four categories: Good, Moderate, Poor, and Hazardous. These classifications provide insights into the environmental conditions and potential health risks associated with varying pollution levels.

Key features include:

1. **Temperature (°C):** Regional average temperature affecting pollutant behavior.
2. **Humidity (%):** Relative humidity influencing particulate retention.
3. **PM2.5 (µg/m³):** Fine particulate matter levels.
4. **PM10 (µg/m³):** Coarse particulate matter levels.
5. **NO2 (ppb):** Nitrogen dioxide levels from emissions.
6. **SO2 (ppb):** Sulfur dioxide from industrial activities.
7. **CO (ppm):** Carbon monoxide from combustion sources.
8. **Industrial Proximity (km):** Distance to the nearest industrial zone.
9. **Population Density (people/km²):** Indicator of human activity impact on pollution.

2.1 Data Visualization and Analysis

The scatter plot shows the relationship between **Proximity to Industrial Areas** and **Population Density**, categorized by **Air Quality** into **Good**, **Moderate**, **Poor**, and **Hazardous**. **Good Air Quality** is linked to lower population densities and greater distances from industrial zones, while **Moderate** and **Poor Air Quality** are found closer to industrial areas with medium to high population densities. **Hazardous Air Quality** is scattered but often near industrial zones with high population density. This suggests that proximity to industrial areas significantly affects air quality, with closer distances leading to poorer conditions.

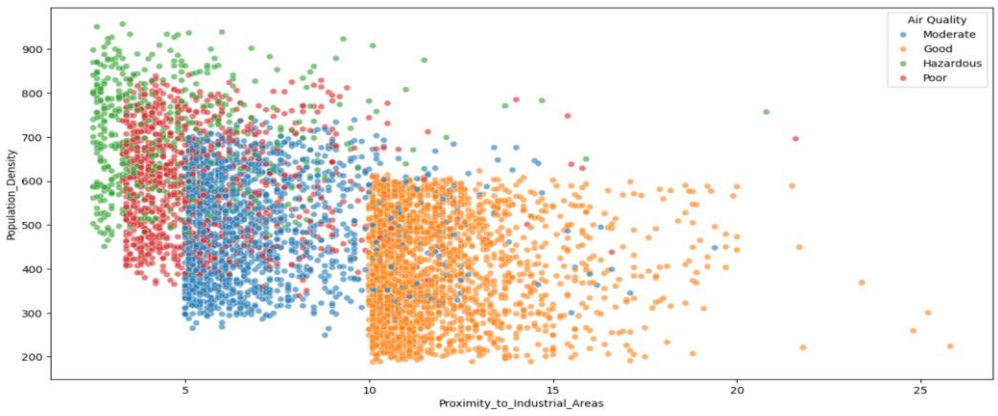
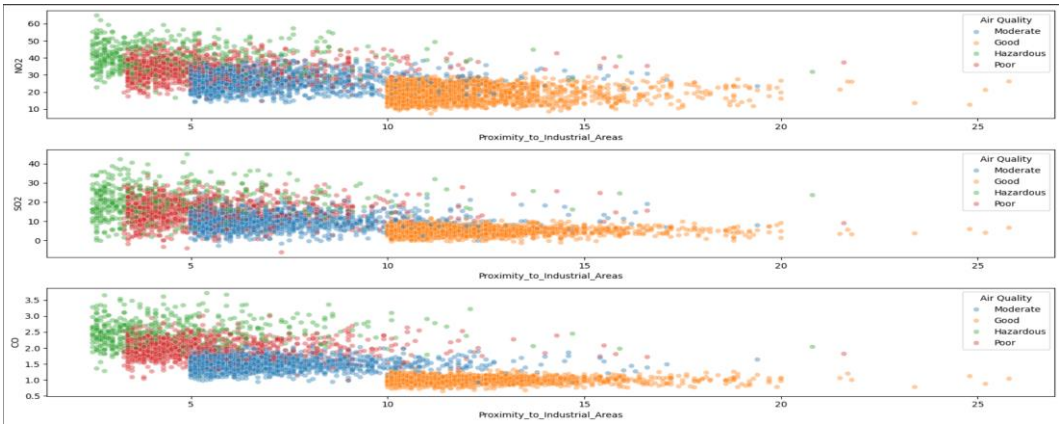
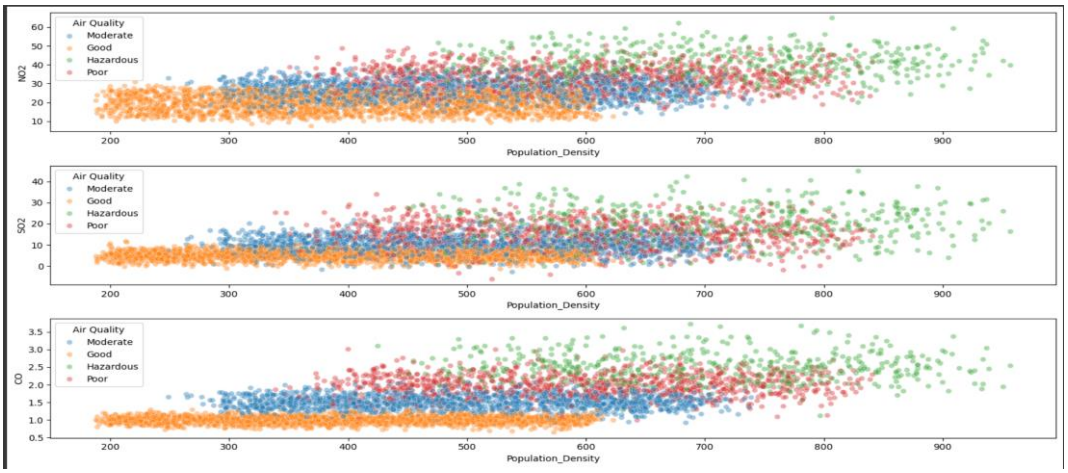


Figure 2 shows how NO_2, SO_2, and CO concentrations change with distance from industrial areas. Pollution levels are highest near industrial zones, where air quality is Poor or Hazardous. As the distance from these areas increases, pollutant levels decrease, and air quality improves to Moderate and Good. This highlights the significant impact of industrial emissions on nearby air quality.



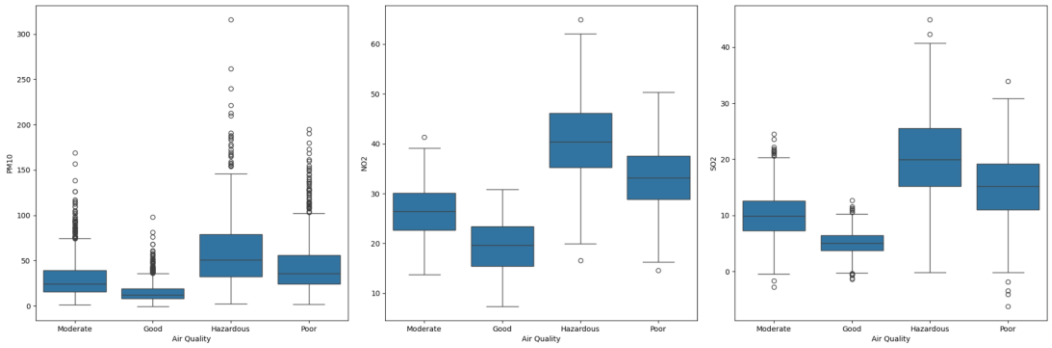
The plots show that as population density increases, concentrations of NO_2, SO_2, and CO rise due to intensified industrial, vehicular, and domestic emissions. Poor and Hazardous air quality is more common in high-density areas, while lower-density areas generally have better air quality. NO_2 and SO_2 steadily increase with density, and CO, though lower, follows a similar trend.



3. Experimental Approach

3.1 Data Preprocessing

We used the **Interquartile Range (IQR)** to remove outliers because the data features contain extreme values that could skew analysis and affect model performance. IQR is a robust statistical method for identifying and eliminating outliers based on the spread of the data.



Outliers were present in almost all features, including temperature, humidity, PM2.5, PM10, NO_2, SO_2, CO, proximity to industrial areas, and population density. These outliers could distort relationships between features and air quality categories, as seen in their broad ranges and extreme values. By removing values outside, we ensure that the dataset retains the essential patterns without being unduly influenced by rare or extreme values.

The target variable represents different air quality levels, encoded with label encoding: Good (0), Hazardous (1), Moderate (2), and Poor (3). This method is appropriate because the categories have a natural order, enabling models to process the values while maintaining their ordinal relationship.

3.2 Data Preparation

The dataset is divided into features and the target variable in data preparation. The data is then split into training (75%) and testing (25%) sets, ensuring the models are trained on most of the data and evaluated on a smaller, unseen portion.

3.3 Applying Classification Algorithms

3.3.1 KNN

We applied the KNN (K-Nearest Neighbours) algorithm to a dataset, experimenting with three different distance metrics: Euclidean, Manhattan, and Cosine distance. For each metric, we tested various values of K (number of neighbours) from 1 to 9 to determine the optimal configuration based on cross-validation accuracy and performance metrics (Accuracy, Precision, Recall, F1-Score, ROC-AUC).

Here is the table summarizing the results obtained from the KNN classification, based on different distance metrics (Euclidean, Manhattan, and Cosine), and the optimal value of KKK for each metric:

Metric	K	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Euclidean	1	0.925962	0.92619	0.925962	0.925236	0.860627
Euclidean	2	0.911538	0.918832	0.911538	0.909486	0.883027
Euclidean	3	0.939423	0.937425	0.939423	0.936979	0.888995
Euclidean	4	0.928846	0.928278	0.928846	0.925569	0.922248
Euclidean	5	0.935577	0.931736	0.935577	0.932013	0.932876
Euclidean	6	0.932692	0.93081	0.932692	0.929563	0.941711
Euclidean	7	0.936538	0.932	0.936538	0.93223	0.951272
Euclidean	8	0.934615	0.931604	0.934615	0.930747	0.9534
Euclidean	9	0.944231	0.941404	0.944231	0.940693	0.953807

Manhattan	1	0.927885	0.927032	0.927885	0.92636	0.879813
Manhattan	2	0.915385	0.923311	0.915385	0.912952	0.900525
Manhattan	3	0.935577	0.933156	0.935577	0.932711	0.92776
Manhattan	4	0.936538	0.935506	0.936538	0.932811	0.943736
Manhattan	5	0.945192	0.942534	0.945192	0.941566	0.944303
Manhattan	6	0.936538	0.935284	0.936538	0.933223	0.954711
Manhattan	7	0.946154	0.944378	0.946154	0.943133	0.955761
Manhattan	8	0.938462	0.938352	0.938462	0.935331	0.965819
Manhattan	9	0.941346	0.939974	0.941346	0.938215	0.96575
Cosine	1	0.861538	0.881508	0.861538	0.869259	0.816602
Cosine	2	0.848077	0.872983	0.848077	0.855321	0.85427
Cosine	3	0.863462	0.887917	0.863462	0.87329	0.893031
Cosine	4	0.873077	0.882738	0.873077	0.876961	0.908369
Cosine	5	0.871154	0.883262	0.871154	0.875986	0.913209
Cosine	6	0.877885	0.887554	0.877885	0.881832	0.933956
Cosine	7	0.885577	0.892213	0.885577	0.887587	0.932843
Cosine	8	0.883654	0.889687	0.883654	0.886039	0.943226
Cosine	9	0.878846	0.886086	0.878846	0.881221	0.943053

The highlighted results are the best models in KNN across different distance metrix , values of KKK, and evaluation metrics like Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

The results of using three different distance metrics—**Euclidean**, **Manhattan**, and **Cosine** shed light on how distance measures affect the classification performance. For the air quality dataset, **Euclidean Distance** emerged as the most effective metric, producing the highest accuracy and cross-validation score. This result suggests that the air quality features in the dataset exhibit a linear relationship, making Euclidean distance suitable for measuring the similarity between data points. **Manhattan Distance**, which measures the sum of absolute differences, also performed similarly to Euclidean, indicating that the dataset may have some discrete or grid-like relationships between features that this metric can capture effectively. However, **Cosine Distance** performed the worst, which is likely due to the dataset's features being continuous and not vector orientations or angles, where Cosine similarity would typically excel. The lower accuracy with Cosine distance suggests it is less suited for datasets with environmental variables such as temperature, humidity, and pollutant concentrations.

The optimal value of K was determined for each distance metric based on cross-validation accuracy. For both **Euclidean** and **Manhattan distances**, the best K was 5, yielding the highest accuracy (94.13% for Euclidean and 94.32% for Manhattan). This suggests that a smaller value of K, where each prediction is based on the five nearest neighbours, offers a good balance between model complexity and accuracy. In contrast, for **Cosine Distance**, the optimal K was 8, but the overall accuracy was lower at 89.22%, indicating that a higher K did not improve the model's performance. This trend emphasizes that the best value of K depends not only on the distance metric but also on the nature of the dataset. In this case, the air quality dataset benefits from a lower K with Euclidean and Manhattan distances, while a larger K did not produce optimal results for Cosine distance.

3.3.2 Logistic Regression

The logistic regression model was trained using both L1 (Lasso) and L2 (Ridge) regularization techniques, with varying values of the regularization parameter C.

Regularisation	C	Accuracy	Precision	Recall	F1-Score	ROC-AUC
L1	0.01	0.910577	0.900206	0.910577	0.900616	0.878669
L1	0.1	0.951923	0.94874	0.951923	0.949024	0.986895
L1	1	0.952885	0.951536	0.952885	0.951625	0.989039
L1	10	0.952885	0.952068	0.952885	0.951886	0.989006
L1	100	0.952885	0.952068	0.952885	0.951886	0.988992
L2	0,01	0.927885	0.915599	0.927885	0.920449	0.97991
L2	0.1	0.950962	0.948739	0.950962	0.947809	0.987158
L2	1	0.952885	0.950542	0.952885	0.951095	0.988908
L2	10	0.953846	0.953081	0.953846	0.952856	0.989055
L2	100	0.952885	0.952068	0.952885	0.951886	0.988999

The comparison between L1 and L2 regularization in logistic regression reveals distinct behaviours due to their mathematical properties. L1 regularization achieved its best performance at C = 1.0, where the model had an accuracy of 95.29% and an F1-score of 95.16%. This indicates that L1 effectively reduces overfitting by shrinking less important features to zero, acting as a form of feature selection. However, as C increased beyond 1.0, performance plateaued, suggesting that reducing regularization strength further had little impact. Conversely, lowering C (increasing regularization) to 0.01 led to a significant drop in accuracy (91.05%), precision, and recall. This demonstrates that overly strong regularization limits the model's ability to capture critical patterns, emphasizing the importance of balancing regularization strength.

L2 regularization, in contrast, showed its best performance at C = 10.0, with an accuracy of 95.38% and an F1-score of 95.28%. Unlike L1, which zeroes out features, L2 penalizes all coefficients uniformly, resulting in a more balanced model. Increasing C (weakening regularization) to 100.0 did not significantly improve performance, as the model had already reached optimal complexity. However, reducing C to 0.01 caused accuracy to drop to 92.79%, reflecting underfitting due to overly strong regularization. This trend highlights L2's stability and its ability to generalize well across different CC values, making it more robust in scenarios with subtle feature contributions.

Comparing logistic regression to KNN, the former's superior accuracy and stability can be attributed to its inherent ability to model linear decision boundaries and its flexibility with regularization. Logistic regression's performance improves with optimized C, unlike KNN, whose accuracy is highly dependent on the choice of K and distance metric. Furthermore, the high ROC-AUC scores in logistic regression indicate better discrimination between air quality levels, as it adapts to feature importance through regularization. In summary, L1 is ideal for sparse datasets, while L2 excels in capturing balanced feature contributions, making logistic regression more adaptable and consistent than KNN for this air quality dataset.

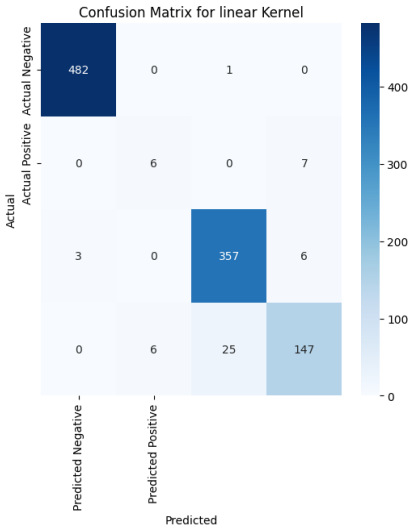
3.3.3 Support Vector Machines (SVM)

Support Vector Machines (SVM) are supervised learning models used for classification and regression tasks. They work by finding the hyperplane that best separates data into classes in a high-dimensional feature space. The goal of the SVM is to maximize the margin between data points from different classes while minimizing classification errors.

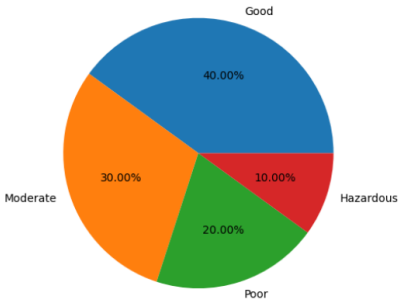
SVMs use different types of kernel functions to map data into higher-dimensional spaces where it becomes easier to find a linear separating hyperplane. In this experiment, we will apply three common kernels as follows:

1- **Linear Kernel:** Assumes data is linearly separable and directly finds a hyperplane in the original feature space.
Results of applying SVM with a linear kernel:
Accuracy: 0.9538

Class	Precision	Recall	F1-Score
0	0.99	1	1
1	0.5	0.46	0.48
2	0.93	0.98	0.95
3	0.92	0.83	0.87



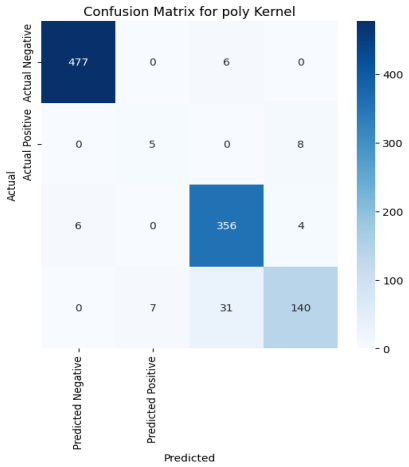
From the evaluation metrics and the confusion matrix, we observe that classes Good (0), Moderate (2), and Poor (3) have high precision, recall and F1-score, with slightly lower F1-score and recall for calls Poor(3). However, for class Hazardous (1), all metrics have low values, reflecting difficulty in classifying this class correctly. This poor performance is likely due to class imbalance, as Hazardous (1) has only 13 samples.



Although the model achieves high accuracy (95.38%), it might be the best metrics due to data imbalance shown in the above figure, where it is primarily driven by the dominance of the major classes (0 and 2).
To conclude, the linear kernel performs well for the majority classes (0 and 2), indicating that the data is likely linearly separable for these classes. However, for minority classes (1 and 3), the linear kernel might not capture the underlying patterns.

2- **Polynomial Kernel:** Maps data into a higher-dimensional space using polynomial transformations, making it possible to classify data that is not linearly separable.
Results of applying SVM with a poly kernel of degree 3:
Accuracy: 0.9404

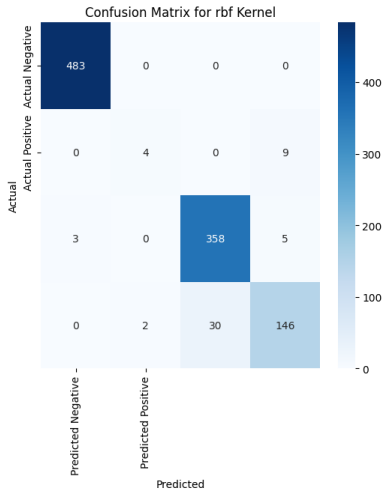
Class	Precision	Recall	F1-Score
0	0.99	0.99	0.99
1	0.42	0.38	0.4
2	0.91	0.97	0.94
3	0.92	0.79	0.85



We observe that for dominant classes (0 and 2), the performance is high and similar to the linear kernel performance. Additionally, class 3 is also similar to linear kernel with a slightly lower recall. For class 1, the performance is slightly lower than the already poor performance with the linear kernel, reflecting the class imbalance issue.
Overall, the linear kernel achieves better accuracy and recall for minority classes with fewer misclassifications overall. The polynomial kernel may overfit slightly or struggle with the limited data for minority classes (Hazardous and Poor).

3- **Radial Basis Function (RBF):** It computes the similarity as a Gaussian function of the distance between two data points, it is non-linear and very flexible, capable of modeling complex decision boundaries, since it maps data into an infinite-dimensional space to separate classes effectively.
Results:
Accuracy: 0.9529

Class	Precision	Recall	F1-Score	Support
0	0.99	1	1	483
1	0.67	0.31	0.42	13
2	0.92	0.98	0.95	366
3	0.91	0.82	0.86	178



We observe that performance is high for classes 0 and 2, just like linear and poly kernels. However, for class 1, while recall is still low, precision is significantly higher than with linear (0.50) and

polynomial (0.42) kernels, showing better handling of this minority class. And for class 3, rbf gave a slightly better recall than the polynomial kernel (0.79), and similar performance overall. Moreover, the accuracy is a little higher than poly and almost like linear kernel.

After experimenting with three different kernels, we learned that the choice of the kernel affects the model's performance as follows:

Impact on Accuracy:

The **linear kernel** achieves the highest accuracy due to its simplicity and strong performance for the dominant classes. However, accuracy is not always a reliable metric, especially in imbalanced datasets.

The **RBF kernel** achieves similar accuracy but provides better performance on minority classes by modeling complex non-linear relationships.

Impact on Minority Classes:

The **RBF kernel** was the best for minority classes like Hazardous (1) and Poor (3), providing better precision and F1-scores. This demonstrates its flexibility in handling non-linear class boundaries.

The **polynomial kernel** struggles the most with minority classes due to its complexity and sensitivity to parameter tuning (e.g., degree).

Impact on Model Complexity:

The **linear Kernel** is computationally efficient and easier to train, making it a good choice for linearly separable data or high-dimensional datasets.

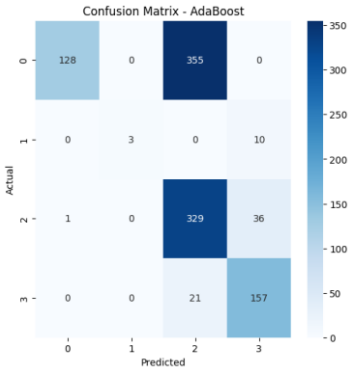
The **RBF kernel** is more computationally expensive but provides the best balance between precision and recall, especially for non-linear relationships.

The **polynomial kernel** can introduce unnecessary complexity, leading to overfitting or degraded performance, especially with small datasets.

3.3.4 Ensemble Methods

1- **Boosting:** is an ensemble learning technique that aims to improve the performance of weak learners by combining them in a sequential manner.
Accuracy: 0.5933

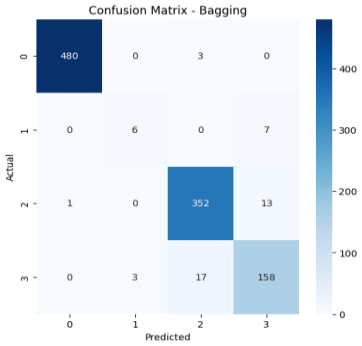
Class	Precision	Recall	F1-Score
0	0.99	0.27	0.42
1	1	0.23	0.38
2	0.47	0.9	0.61
3	0.77	0.88	0.82



We observe that boosting achieved low accuracy (59.33%) due to poor handling of the dominant classes, as we see that recall is very low in class 0, though it performed relatively better on minority classes (Class 3 and 2). And for class 1, precision is high, but recall is extremely low, failing to identify most instances.

2- **Bagging:** is an ensemble learning technique that reduces variance and overfitting by training multiple base learners independently on random subsets of the data and combining their predictions to improve stability and accuracy.
Accuracy: 0.9577

Class	Precision	Recall	F1-Score
0	1	0.99	1
1	0.67	0.46	0.55
2	0.95	0.96	0.95
3	0.89	0.89	0.89

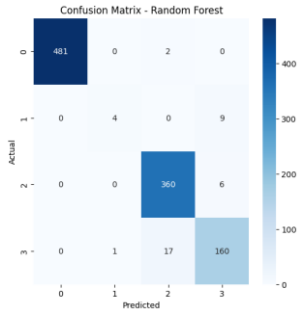


We observe that bagging provides excellent handling of the dominant class, since the performance is perfect for class 0. The performance is also high for classes 2 and 3, and moderate precision but low recall for minority class 1.

3- **Random Forest:** a specialized form of Bagging designed for decision trees, with added randomness in feature selection to improve performance and reduce overfitting.

Accuracy: 0.9663

Class	Precision	Recall	F1-Score
0	1	1	1
1	0.8	0.31	0.44
2	0.95	0.98	0.97
3	0.91	0.9	0.91



Random forest outperformed both boosting and bagging, it provided excellent performance on dominant classes 0,2, and 3 , and for monitory class 1, the recall is still low but the precision is relatively high.

To conclude, Bagging and Random Forest outperformed Boosting, as they maintain consistent performance across all classes and handle class imbalances effectively.

Boosting struggled because it focused excessively on misclassified samples, which led to suboptimal generalization for dominant classes.

Comparing to methods discussed earlier, Ensemble methods (Bagging, Random Forest) outperform individual models like KNN, Logistic Regression, or SVM, as they combine multiple learners to reduce overfitting and variance while improving robustness and accuracy. Boosting, despite lower performance in this case, is often more effective than individual models for datasets with complex decision boundaries or high imbalance.

4. Conclusion

In this assignment, we explored and compared four machine learning approaches—K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machines (SVM), and Ensemble Methods (Bagging, Boosting, and Random Forest)—to classify air quality levels. The findings highlight that Random Forest emerged as the most robust and accurate method, excelling across all classes, particularly in managing class imbalances and providing reliable predictions. Bagging followed closely, offering consistent performance with reduced variance. Logistic Regression, particularly with L2 regularization, demonstrated stability and efficiency for linearly separable data, while SVM with an RBF kernel excelled in capturing non-linear relationships and improving minority class classification. Although KNN showed promise for simpler, linear datasets, it was less effective for complex, imbalanced data. Overall, ensemble methods, particularly Random Forest, provided the best balance of accuracy, robustness, and adaptability, making them the most suitable choice for the air quality classification problem.

Appendix

Link to colab: https://colab.research.google.com/drive/10ltVhW-J_aif-cktwAYZSkiB0-TtLy_W?usp=sharing