PREDICTIVE ANALYTICS PROJECT REPORT

TITLE

Air Quality Index

Submitted by:-

Adarsh kamal

Registration No:- 12306714

Program and Section:- B.Tech (CSE) and K23CD

Course Code: INT-234

Under the Guidance of

(Maneet Kaur)

Discipline of CSE/IT

School of Computer Science and Engineering

Lovely Professional University, Phagwara

Machine Learning–Based Analysis and Prediction of Air Quality Levels

Adarsh Kamal
Department of Computer Science
University Name, India

Abstract—

Air pollution poses a serious threat to public health and environmental sustainability, particularly in rapidly urbanizing regions. Accurate prediction and analysis of air quality indicators are therefore essential for effective environmental monitoring and policy formulation. This study presents a machine learning–based framework for analyzing and predicting air quality levels using real-world air pollution monitoring data. The dataset includes pollutant statistics (minimum, maximum, and average concentrations), spatial attributes (latitude and longitude), temporal attributes, and categorical information related to monitoring locations. A comprehensive workflow was followed, including data preprocessing, exploration data analysis (EDA), feature engineering, and the development of multiple predictive models. For classification, air quality levels were categorized into four classes—Good, Moderate, Poor, and Hazardous—and evaluated using Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Support Vector Classifier (SVC). Additionally, a regression model using Random Forest Regressor was implemented to predict the continuous average pollutant concentration. Model performance was assessed using accuracy, precision, recall, F1-score, confusion matrices for classification, and RMSE, MAE, and R^2 for regression. The results demonstrate that ensemble-based models, particularly Random Forest, provide superior predictive performance and robust generalization capabilities for air quality analysis.

Index Terms—Air quality prediction, machine learning, environmental data analysis, classification, regression, random forest.

I. Introduction

Air pollution has emerged as one of the most critical environmental challenges of the modern era. Rapid industrialization increased vehicular emissions, population growth, and urban expansion have significantly deteriorated air quality in many regions. Prolonged exposure to polluted air is associated with severe health issues such as respiratory diseases, cardiovascular disorders, and reduced life expectancy. Consequently, accurate monitoring and prediction of air pollution levels are vital for public health protection and sustainable development. With the increasing availability of environmental data from air quality monitoring stations, machine learning techniques have become powerful tools for analyzing complex pollution patterns. These techniques can capture non-linear relationships, handle high-dimensional datasets, and provide improved predictive accuracy compared to traditional statistical methods.

II. Literature Review

Recent research highlights the growing role of machine learning in environmental monitoring and air quality prediction. Traditional statistical techniques often struggle to model complex, non-linear interactions among pollution sources, meteorological factors, and geographical characteristics. Tree-based models, including Decision Trees and Random Forests, have gained significant attention due to their ability to handle non-linearity and reduce overfitting through ensemble learning.

III. Dataset Description

The dataset consists of air pollution measurements collected from multiple monitoring stations. Key attributes include pollutant statistics, geographical coordinates, temporal features derived from timestamps, and categorical variables such as city, state, station, and pollutant type.

IV. Methodology

The methodology includes data preprocessing, exploratory data analysis, feature selection, and model development. Missing values were handled using median imputation, categorical variables were label-encoded, and features were standardized prior to model training.

Model Development

Two types of predictive tasks were implemented in this study: air quality classification and pollutant concentration regression. Multiple machine learning models were selected to ensure a fair and comprehensive comparison between linear, distance-based, tree-based, ensemble, and margin-based learning approaches.

Logistic Regression: Logistic Regression was used as a baseline classification model due to its simplicity and interpretability. It models the probability of each air quality class as a function of a linear combination of input features. Despite its linear nature, Logistic Regression provides valuable insights into the effectiveness of basic feature relationships.

K-Nearest Neighbors (KNN): KNN is a distance-based classification algorithm that assigns a class label based on the majority class among the K closest data points in the feature space. Since KNN relies on distance calculations, feature scaling was applied to ensure unbiased neighbor selection. Different values of K were tested to analyze model sensitivity and performance.

Decision Tree Classifier: The Decision Tree classifier builds a hierarchical tree structure by recursively splitting the dataset based on feature values that maximize information gain. This model is easy to interpret and can capture non-linear relationships but is prone to overfitting when used alone.

Random Forest Classifier: Random Forest is an ensemble learning technique that constructs multiple decision trees using bootstrapped samples and aggregates their predictions. This approach reduces overfitting, improves generalization, and provides feature importance measures. In this study, Random Forest served as the primary classification model due to its robustness and strong performance.

Support Vector Classifier (SVC): Support Vector Classifier aims to find an optimal hyperplane that maximizes the margin between different air quality classes. By using kernel functions and standardized features, SVC effectively handles non-linear decision boundaries in high-dimensional feature spaces.

Random Forest Regressor: In addition to classification, a Random Forest Regressor was implemented to predict the continuous average pollutant concentration. Similar to its classification counterpart, this ensemble-based regressor captures complex non-linear relationships and provides stable predictions with reduced variance.

```
models = {
    "LogisticRegression": LogisticRegression(max_iter=200),
    "KNN": KNeighborsClassifier(n_neighbors=5),
    "DecisionTree": DecisionTreeClassifier(),
    "RandomForest": RandomForestClassifier(n_estimators=200),
    "SVC": SVC()
}

model_accuracies = {}
```

The dataset was divided into training and testing subsets to ensure unbiased model evaluation, and standardized feature scaling was applied where required. Model performance was evaluated using appropriate classification and regression metrics.

Comparing Models:-

```
# ---------------- Compare Model Accuracies ----------------
plt.figure(figsize=(7,4))
names = list(model_accuracies.keys())
scores = list(model_accuracies.values())
sns.barplot(x=names, y=scores)
plt.ylim(0, 1)
plt.title("Model Accuracy Comparison")
plt.ylabel("Accuracy")
plt.xlabel("Model")
plt.tight_layout()
plt.show()

print("\nModel Accuracies:")
for n, s in model_accuracies.items():
    print(f"{n}: {s:.4f}")
```

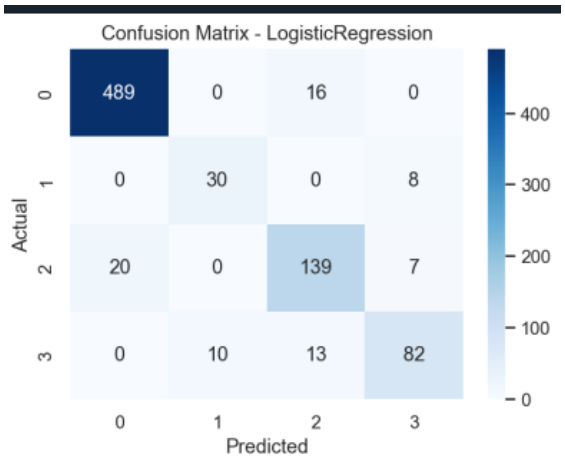
V. Results and Discussion

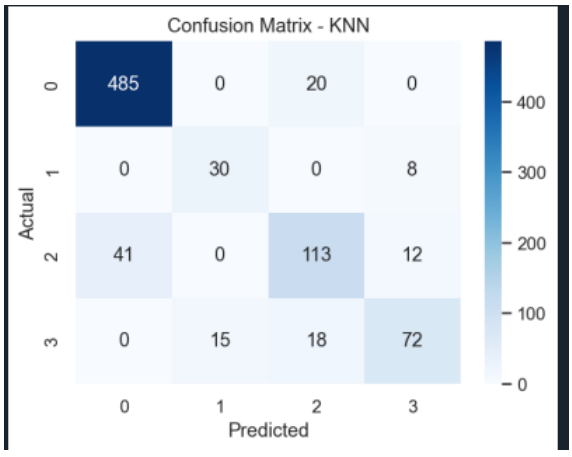
Experimental results show that Random Forest achieves the highest classification accuracy and robust generalization. Feature importance analysis indicates that pollutant statistics are the most influential predictors. The regression model demonstrates low error and high R^2, indicating strong predictive capability.

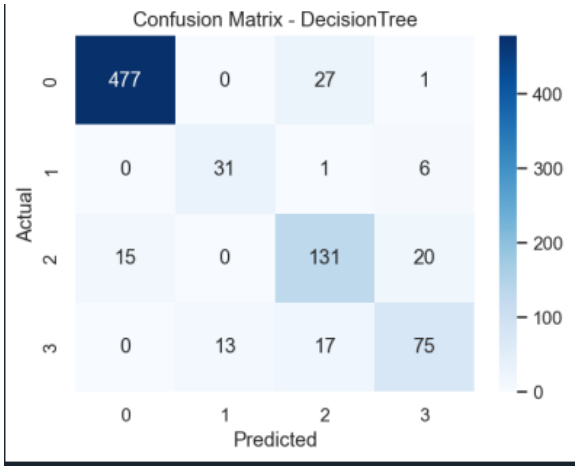
Classification Results

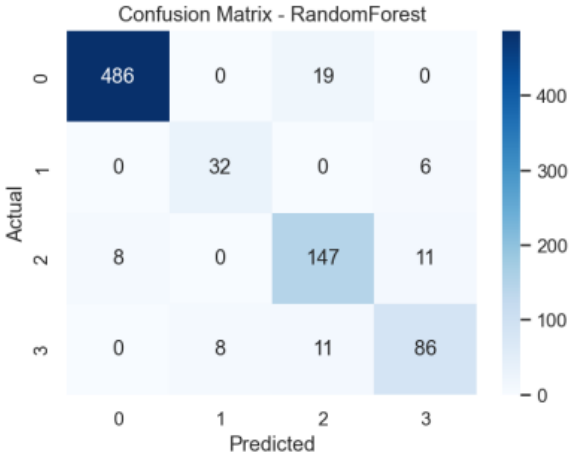
The classification models were evaluated using accuracy, precision, recall, F1-score, and confusion matrices. Among the evaluated models, Random Forest consistently achieved the highest accuracy and F1-score, indicating strong predictive performance and balanced classification across air quality categories. Logistic Regression provided a reasonable baseline, while KNN and Decision Tree showed moderate performance. SVC demonstrated competitive results but required careful feature scaling.

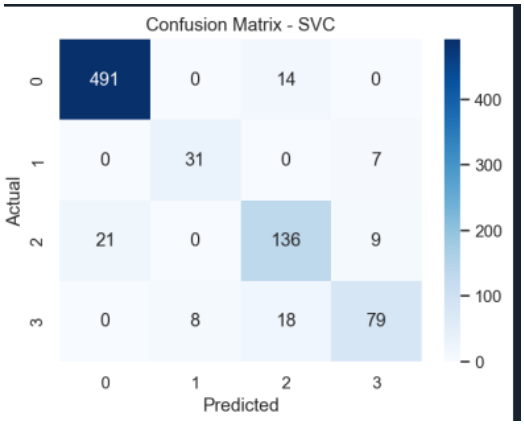
The confusion matrix analysis revealed that ensemble-based models reduced misclassification between adjacent air quality categories, which is crucial for reliable air quality assessment.











VI. Conclusion

This study confirms the effectiveness of machine learning techniques for air quality analysis and prediction. Ensemble-based models, particularly Random Forest, outperform other approaches in handling complex and non-linear pollution patterns. Future work may incorporate meteorological variables and real-time deployment.

References

1. World Health Organization. Air Pollution and Health.

2. L. Breman, Random Forests. Machine Learning.

3. C. M. Bishop, Pattern Recognition and Machine