PARSHWANATH CHARITABLE TRUST'S
A.P. SHAH INSTITUTE OF TECHNOLOGY
Department of Computer Science and Engineering
Data Science

CSE DATA SCIENCE

AQI FORECAST

Khushi Kadam - 21107031

Shravani Kulkarni - 21107063

Krish Jaswal – 21107039

Vedant Mayekar - 22207007

Project Guide
Prof. Vaibhav Yavalkar

Outline

- Introduction
- Literature Survey of the existing systems
- Limitations of the existing systems
- Problem statement
- System Design
- Technologies and methodologies
- Implementation
- Conclusion
- References

Introduction

- Motivation :

- Urban air pollution is escalating, causing health complications like respiratory and cardiovascular diseases. Lack of public awareness contributes to inadequate preventive measures against these health risks, necessitating the integration of an AQI tracker.

- Inadequate air quality monitoring in some regions leads to a lack of real-time data, hindering timely interventions to address pollution. Consequently, rising respiratory illnesses emphasize the pressing need for effective solutions, including the implementation of an AQI tracker for better public health management.

Introduction

- <u>Objectives:</u>

- To develop an accurate AQI forecasting model for targeted regions or cities using KNN and Linear Regression algorithm

- To increase understanding of air quality monitoring importance among the public by providing a user-friendly platform

- To provide insights to aid in effective pollution control measures for environmental authorities

- To enable proactive health precautions based on forecasted AQI levels

- To assist authorities and public in mitigating adverse health effects during pollution episodes

Literature Survey of the existing system

Title and Year of publication	Author Name	Methodology	OUTCOME	Conclusion
[1] “Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis” 2023	N. Srinivasa Gupta,Yashvi Mohta,Khyati Heda,Raahil Armaan,B. Valarmathi,and G. Arulkumaran	A study examined air indicators (e.g., AQI, PM2.5, NOx) and used SVR for prediction. Results highlight pollutants like sulfur dioxide, nitrogen dioxide, ozone, and PM2.5.	Balancing the dataset through SMOTE improved model accuracies across support vector regression, random forest regression, and CatBoost regression, as evidenced by increased R-SQUARE and decreased MSE, RMSE, and MAE, particularly notable in New Delhi, Bangalore, Kolkata, and Hyderabad.	Balancing datasets with SMOTE significantly improves accuracy in predicting AQI using machine learning models, with random forest regression and CatBoost regression consistently outperforming SVR in India's major cities.

Title and Year of publication	Author Name	Methodology	OUTCOME	Conclusion
[2] “A Study and Analysis of Air Quality Index and Related Health Impact on Public Health” 2020	Pranav Shrirama and Srinivas Malladib	The methodology involves deploying gas sensors along with an ADuC812 device to measure concentrations of CO, NO2, SO2, and O3 for calculating the air quality index, as well as utilizing statistics and AI to monitor air pollution, particularly focusing on PM2.5 and PM10 levels near schools, with a smart system generating alerts to parents, teachers, and medical personnel when air quality exceeds specified standards.	The literature study identifies research gaps in understanding the relationship between air pollution, air quality index, and public health, emphasizing the need for a comprehensive approach addressing both direct and indirect health effects, highlighting the importance of real-time monitoring and preventive measures for urban commuters.	Existing systems provide air quality index information but lack in addressing health impacts adequately, necessitating a more nuanced understanding of specific air pollutants' effects on different parts of the body for effective preventive measures.

Title and Year of publication	Author Name	Methodology	OUTCOME	Conclusion
3] “Air Quality ndex – A Comparative Study for Assessing the Status of Air Quality” 2015	Shivangi Nigam, B.P.S. Rao, N. Kumar, V. A. Mhaisalkar	Real-time air quality monitoring was conducted at a residential site in NEERI, Nagpur, using Environment S.A CAAMS Analyzer to measure PM10, PM2.5, SO2, and NO2 concentrations, employing Beta Attenuation Method for PM10 and PM2.5, UV fluorescence method for SO2, and Chemiluminescence Analyzer for NO2 measurement, calibrated via traceable standard reference gas method.	The outcome highlights the dominance of particulate matter, especially PM10, in contributing to poor air quality in the residential site NEERI, Nagpur, emphasizing the urgent need for effective pollution control and management strategies to address public health concerns and promote civic well-being.	The conclusion underscores the urgent need for robust pollution control measures, particularly focusing on particulate matter like PM10, to mitigate the severe public health risks associated with air pollution in residential areas like NEERI, Nagpur.

Limitations of existing systems

From the literature review of existing systems, we find that,

- Generalizability: SMOTE-balanced dataset‘s effectiveness may vary outside major Indian cities due to diverse pollution and population dynamics.
- Health Focus: Existing systems often lack specificity in addressing health outcomes, hindering tailored preventive measures against air pollution.
- Implementation Challenges: Urgent pollution control measures face hurdles like resource availability and community engagement, especially in residential areas like NEERI, Nagpur.

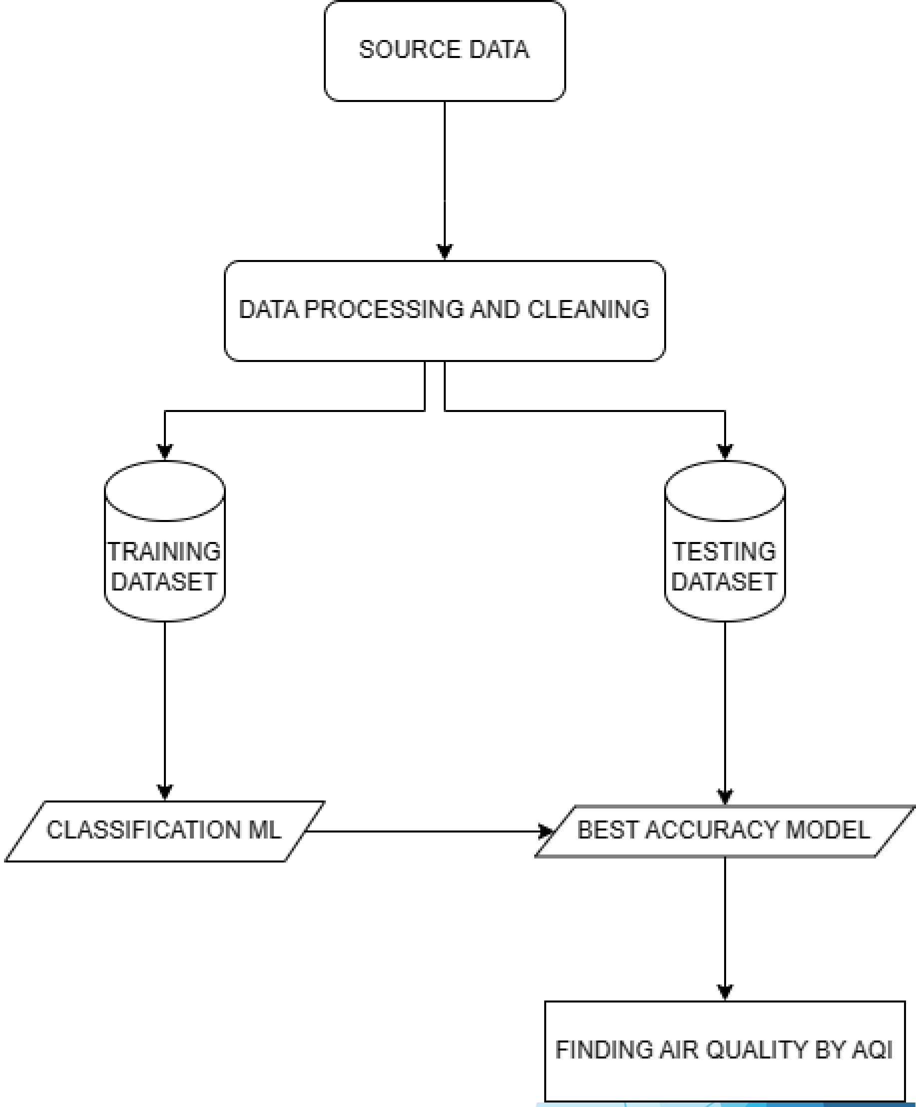
Problem Statement

- Air pollution poses significant health risks and environmental challenges in many regions worldwide. However, the lack of accurate and timely information on air quality often hinders effective pollution control measures and proactive health precaution.

- The gap in information availability and accuracy underscores the need for a robust AQI tracker that can deliver precise forecasts, increase public awareness, and support authorities in implementing targeted pollution control strategies.

System Design

- **Source Data Identification**: Obtain relevant data including historical air quality measurements and meteorological data.
- **Data Preprocessing**: Clean and preprocess the data to address missing values, outliers, and inconsistencies.
- **Model Training and Testing**: Train various ML algorithms like linear regression and KNN on the preprocessed data.
- **Model Selection**: Evaluate and compare the performance of the trained models to select the best-performing one.
- **Model Deployment**: Deploy the selected model for AQI forecasting in a production environment.
- **Monitoring and Evaluation**: Continuously monitor and evaluate the model's performance for accuracy and reliability.



Technologies and Methodologies

- **Linear Regression**

Linear regression is a fundamental algorithm used in statistics and machine learning for understanding the relationship between a dependent variable and one or more independent variables. Essentially, it seeks to find the best-fitting straight line that represents the linear relationship between the variables

- **KNN**

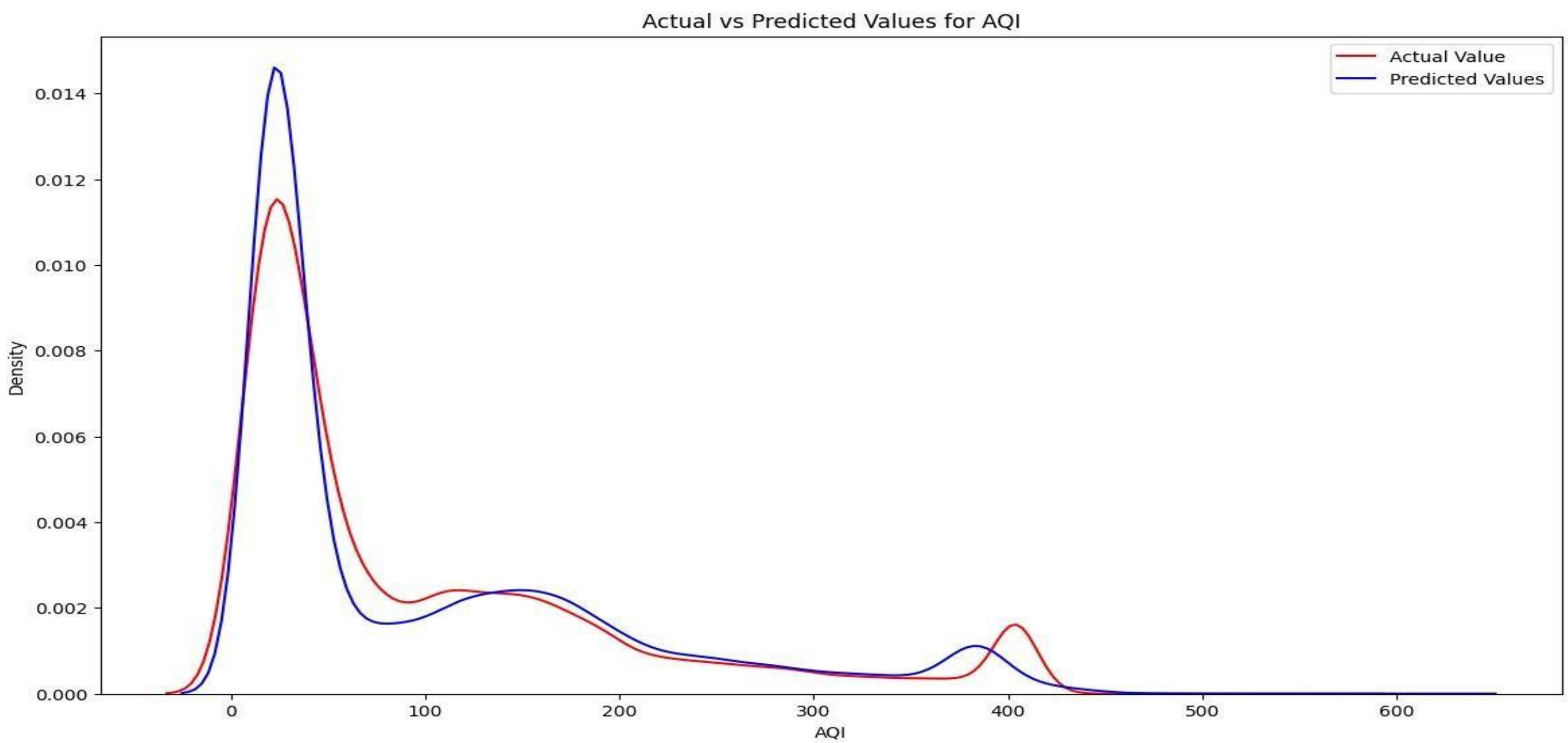
The K-Nearest Neighbors (KNN) algorithm is a simple yet powerful supervised machine learning technique used for classification and regression tasks. The "nearest" neighbors are identified based on a distance metric, typically Euclidean distance, in the feature space. However, its performance can be sensitive to the choice of the distance metric and the value of K.

Implementation

Linear Regression accuracy:

```
RMSE_train=(np.sqrt(metrics.mean_squared_error(Y_train,train_pred)))
RMSE_test=(np.sqrt(metrics.mean_squared_error(Y_test,test_pred)))
print("RMSE TrainingData = ",str(RMSE_train))
print("RMSE TestData = ",str(RMSE_test))
print('-'*50)
print('RSquared value on train:',model.score(X_train,Y_train))
print('RSquared value on test:',model.score(X_test,Y_test))
```

```
RMSE TrainingData =  13.456303399066831
RMSE TestData =  13.553602851130973
--------------------------------------------------
RSquared value on train: 0.9843578096524475
RSquared value on test: 0.984069092328833
```



KNN accuracy:

```
#fit the model on train data
KNN = KNeighborsClassifier().fit(X_train2, Y_train2)

#predict on train
train_preds5 = KNN.predict(X_train2)
#accuracy on train
print("Model accuracy on train is: ", accuracy_score(Y_train2, train_preds5))

#predict on test
test_preds5 = KNN.predict(X_test2)
#accuracy on train
print("Model accuracy on test is: ", accuracy_score(Y_test2, test_preds5))
print('-'*50)

#Kappa Score
print('KappaScore is: ', metrics.cohen_kappa_score(Y_test2,test_preds5))
```

```
Model accuracy on train is:  0.9983079004607032
Model accuracy on test is:  0.9968218423578175
--------------------------------------------------
KappaScore is:  0.9952893818649885
```

Conclusion

- In conclusion, the AQI forecast project aims to provide accurate and reliable predictions of Air Quality Index (AQI) levels by leveraging machine learning algorithms.
- By preprocessing and analyzing relevant data, training various models, and selecting the best-performing one, the project seeks to address challenges in air quality prediction.
- Through continuous monitoring and evaluation, the deployed model will contribute to informed decision-making for public health and environmental management, ultimately improving air quality and safeguarding community well-being.

References

- “Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis” 2023[1]
- “A Study and Analysis of Air Quality Index and Related Health Impact on Public Health” 2020[2]
- “Air Quality Index – A Comparative Study for Assessing the Status of Air Quality” 2015[3]

Thank You...!!