SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
Ramapuram, Chennai – 600 089
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

18CSP109L - MAJOR PROJECT

Scalable Approach for detecting air Quality inference using Ensemble Regerssors

BATCH NUMBER : D7

Team Members	Supervisor
RA2011003020206 - Samarendra T RA2011003020212 - Singa Selvamani S RA2011003020216 - Vimal Aditya Raj	Mrs. P. Preethy Jemima(Ast.Professor) SRMIST Ramapuram

Date:8/02/2024

ABSTRACT

It looks at how machine learning can predict air pollution in smart cities using sensor data. Instead of just using statistics, it uses machine learning to pick out important factors for predicting air quality. The goal is to see how different factors affect air quality predictions.Techniques like data normalization and correlation analysis to find important variables. It use optimization algorithms to choose the most important ones. The results using measures like correlation coefficient and mean absolute error.

Scope and Motivation

- Rapid industrialization and urbanization lead to increased emissions of industrial waste gases, raising air pollution levels. This pollution harms the respiratory system, affecting cardiopulmonary function and potentially causing lung cancer. This endeavor delivers precise air pollution forecasts and facilitates the issuance of early warnings, empowering authorities to implement timely remedial measures and individuals to plan activities accordingly. This proactive approach significantly minimizes the adverse impact of air pollution on health.

LITERATURE SURVEY

S.No	Title of the Paper	Year	Journal	Inferences
1	Dense Air Quality Sensor Networks: Validation, Analysis, and Benefits	2022	IEEE	Three validation methods assess reliability and accuracy, showing how such networks offer detailed pollution insights for better decision-making in cities
2	Ambient Air Monitoring System With Adaptive Performance Stability	2022	IEEE	An adaptive algorithm reduces packet loss in cellular transmissions, enhancing reliability under varying signal strengths.
3	A Deep Learning Approach Using Graph NeuralNetworks for Anomaly Detection in Air Quality Data	2020	IEEE	combines node information correlation to fuse features, constructing spatiotemporal graph structures for anomaly detection. The approach, employing Context Augmented Graph Autoencoder (Con-GAE), efficiently detects anomalies, as demonstrated on synthetic test sets from real-world data.

S.No	Title of the Paper	Year	Journal	Inferences
4	Revealing Influence of Meteorological Conditions on Air Quality Prediction Using Explainable Deep Learning	2022	IEEE	Employs the SHapley Additive exPlanation (SHAP) method to interpret Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models. Results indicate that considering meteorological conditions alone doesn't enhance prediction accuracy
5	A Novel Interpolation-SVT Approach for Recovering Missing Low-Rank Air Quality Data	2022	IEEE	The method employs low rank matrix completion and single value thresholding techniques to handle data sparsity
6	A Predictive Data Feature Exploration-Based Air Quality Prediction Approach	2020	IEEE	The superiority of the proposed approach over other methods, emphasizing the advantage of integrating forecasting data and conducting high-dimensional statistical analysis for air quality prediction.

S.No	Title of the Paper	Year	Journal	Inferences
7	Innovative Spatial-Temporal Network Modeling and Analysis Method of Air Quality	2022	IEEE	The resulting spatiotemporal network model is then used for community detection, revealing local similarities and regional interactions. Experimental results demonstrate the model's dynamic, reliable, and scalable nature, offering insights for air pollution prevention and prediction.
8	Sequence-to-Sequence Air Quality Predictor Based on the n-Step Recurrent Prediction	2022	IEEE	Experimental results demonstrate that the AAQP with n-step recurrent prediction outperforms existing models, offering faster training and more accurate predictions, especially for sudden air pollution events.

S.No	Title of the Paper	Year	Journal	Inferences
9	Data-Driven Air Quality Characterization for Urban Environments: A Case Study	2022	IEEE	Linear regression is recommended for AQI prediction. Overall, the approaches demonstrate feasibility for AQI prediction based on meteorological and historical pollutant data
10	An Agent Based Traffic Regulation System for the Roadside Air Quality Control	2023	IEEE	Machine learning and big data tools for air quality prediction and traffic regulation, operating on a Hadoop-based framework. The system aims to reduce vehicle emissions in polluted road sections while optimizing traffic flow, offering real-time recommendations based on air quality data

Objective

- To handle such complex and non-linear dynamic dependencies.
- To enhance the accuracy of air quality data imputation by jointly considering spatio-temporal correlations from both global and local perspectives.
- To learn useful knowledge from large historical data, which reduces the reliance on a great deal of computing powers and domain expertise.
- To capture complex relationships among observations of air quality and other pollutants.
- To solve the problem of complex linear prediction.

Problem Statement

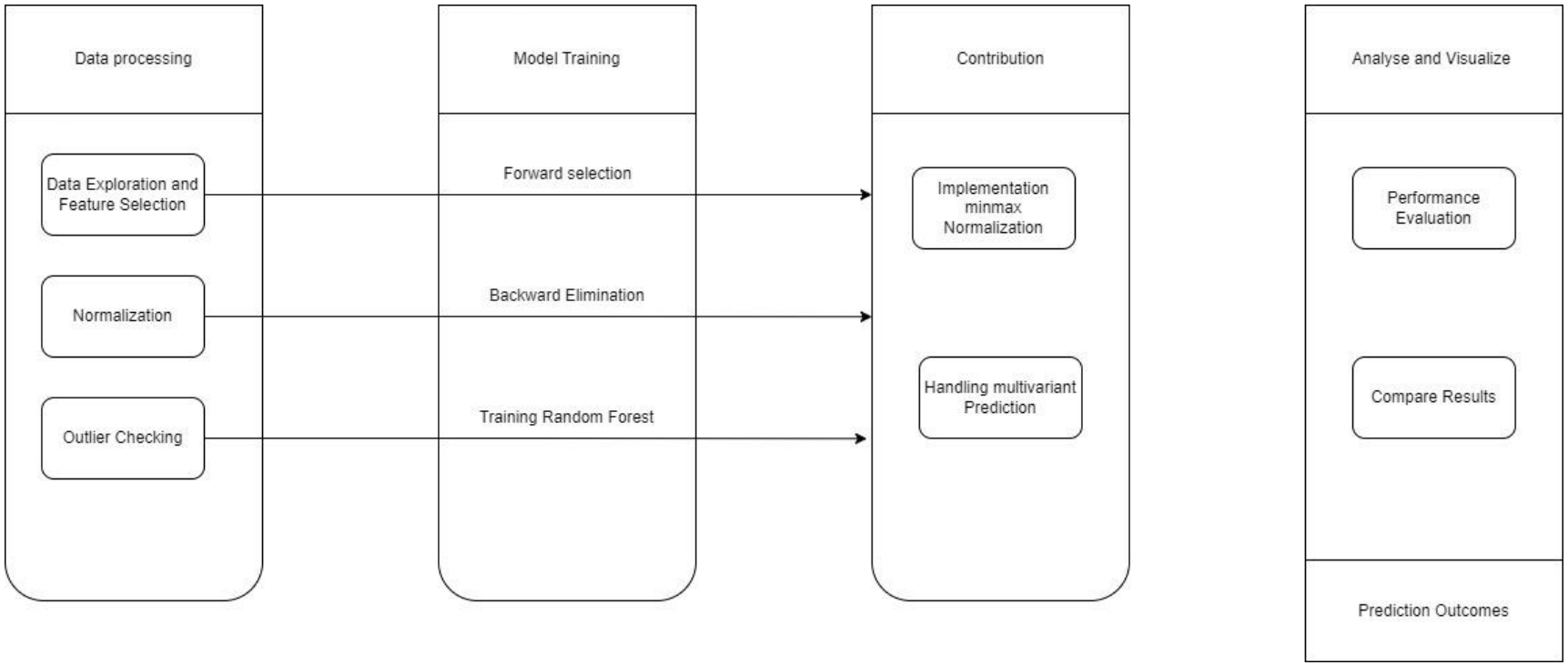
Currently, numerous sophisticated machine learning models have been proposed to enhance predictive performance. However, despite their ability to improve prediction accuracy, the ever increasing model complexity presents a huge obstacle for humans in interpreting the model results. Consequently, it has become a pressing issue to enhance models interpretability. On the one hand, interpretability provide supporting evidence to decision makers regarding model results, thereby enhancing trust in the models. On the other hand, interpretability empowers researchers and data scientists to gain profound insights into the strengths and weaknesses of the models, which shed light on model design and optimization.

Proposed work

Outlier and Missing Data Processing: It is necessary to process the original datas outliers and missing values. As the data came from the official platform and no illogical values were found after the screening, for example, no negative values for atmospheric concentration and air pressure are within reasonable limits, so the data are considered to be true and valid, and no outliers need to be processed. The data were normalized by implementing a Min-Max normalization technique. This helps in removing the units in the acquired data or the impact of differing scales . The Min-Max normalization technique is used for scaling the data values within a fixed range (zero to one). Initially, the Min-Max normalization technique subtracts the minimum value from data points and further divides by its range. Feature selection is the process of minimizing the number of input variables when building a predictive model. Forward selection is a type of stepwise regression that begins with a null model. The approach initiates with no variables in the model and step by step adds variables to the model until no variable not included in the model can make a significant contribution to the models conclusion. The variable with the highest test statistic that is more than the cut-off value or the lowest p-value with less than the cut-off value is chosen and added to the model.

Architecture

.



Modules

Module 1 : Data Preprocessing
Module 2 : Correlation Analysis
Module 3 : Random Forest Regressor
Module 4 : Support Vector Regression (SVR)
Module 5 : Performance Measures

Module Discription

Data Preprocessing:

Data preprocessing is essential for machine learning and deep learning projects. Its objective is to transform raw data into a format suitable for model training, cleaning and reducing noise to enhance model performance. Data preprocessing consists in transforming the data values of a certain dataset, aiming to optimize the information acquisition and process. Normally, there is a very large contrast between the maximum and minimum values of the dataset, so normalizing the data minimizes the complexity of the algorithm for its corresponding processing. The normalization of the data allows an adequate benefit for the classification of algorithms related to neural networks. In this case, if the back-propagation technique is used in neural networks, the normalization of the input values will speed up the training phase, turning it into a more efficient neural network.

Module Discription

Correlation Analysis:

The correlation between all potential pairs of values in a table is shown in the matrix. It is an effective tool for compiling a sizable dataset and for locating and displaying data patterns. A correlation matrix simplifies the process of selecting different assets by tabulating their correlation with one another. It is vital to identify the correlations between PM concentrations and influencing factors for developing a good prediction model. It guarantees that the proposed regression model utilizes the efficient features for AQP. PM2.5 is affected by several factors, but all the factors are important in effective AQP. On the other hand, the irrelevant/inactive factors affect the proposed models performance by means of time complexity. Therefore, it is important to compute the correlation coefficients (CCs) for every factor that helps in selecting the optimal features for effectiveforecasting of air pollution. Let us consider characteristic time series data as x = (x , x , x , ..., x) and other dataas y = (y , y , y , ..., y) . The CC between the factors is computed as described in Equation

Module Discription

Random Forest Regressor:

The RF algorithm incorporates growing classification and regression trees (CARTs). Each CART is built using random vectors. For the RF-based classifier model, the main parameters were the number of decision trees, as well as the number of features ( ) in the random subset at each node in the growing trees. During model training, the number of decision trees was determined first. For the number of trees, a larger number is better, but takes longer to compute. A lower leads to a greater reduction in variance, but a larger increase in bias. can be defined using the empirical formula , where M denotes the total number of feature RF can be applied to classification and regression problems, depending on whether the trees are classification or regression trees. Assuming that the model includes T regression trees (learners) for regression prediction, the final output of the regression model is

$$H(x) = \frac{1}{T} \sum_{i=1}^{T} h_i(x)$$

Module Discription

Support Vector Regression (SVR):

SVM is a machine learning algorithm that constructs hyperplanes for separating different classes and is generally used for analyzing data with a categorical output variable. In the case of the continuous numeric output variable, regression analysis is used, namely SVR . All of the SVR kernels, including linear, poly, rbf, sigmoid, and precomputed, were considered in this study and the linear kernel function showed the best results. Therefore, the model parameter of SVR used in this study was the linear kernel function. The advantages of SVR include being robust to outliers, having high prediction accuracy, and easy implementation SVR has been applied to overcome non-linear limitations and uncertainties in order to achieve better prediction accuracy . SVR has been successfully applied to forecast the levels of PM10 concentration in Bangkok, Thailand, with air quality data and meteorological variables.

Module Discription

Performance Measures:

The proposed models efficacy was evaluated using different loss functions, such as MAE,SMAPE,RMSE and MSE . The MAE performance measure effectively reflected the actual situation of the forecasting error. In addition, the other performance measures, such as and SMAPE , effectively evaluate the degree of data change and measure the prediction quality of the proposed model. On the other hand, the is determined as the average or mean square difference between the estimated and actual values.

Software and Hardware Requirements

Hardware:

Processor: Minimum i5 Dual Core
Ethernet connection (LAN) OR a wireless adapter (Wi-Fi)
Hard Drive: Minimum 200 GB; Recommended 200 GB or more
Memory (RAM): Minimum 16 GB; Recommended 32 GB or above

Software:

Python For AI/ML/DL Programming
Jupyter Notebook IDE (Integrated Development Environment) for Development.
PyTorch or TensorFlow for Deep Learning Coding.
Sklearn for Machine Learning/Feature Extraction/Evaluation Metrics Coding.
Numpy for implementing Linear Algebra.
Plotly for Data Visualization (For Graphs).
Matplotlib for Data Visualization (For Graphs).
Seaborn for Data Visualization (For Graphs).
Pandas for dealing with Tabular Data.

IMPLEMENTATION

```
[1]
import warnings
warnings.filterwarnings('ignore')
```
Python

```
[2]
import pandas as pd
```
Python

```
[3]
import numpy as np
```
Python

```
[4]
import seaborn as sns
```
Python

```
[5]
import plotly.express as px
```
Python

```
[6]
import matplotlib.pyplot as plt
```
Python

```
[7]
import matplotlib
```
Python

```
[8]
plt.style.use('fivethirtyeight')
```
Python

```
[9]
from sklearn.model_selection import train_test_split
```
Python

IMPLEMENTATION

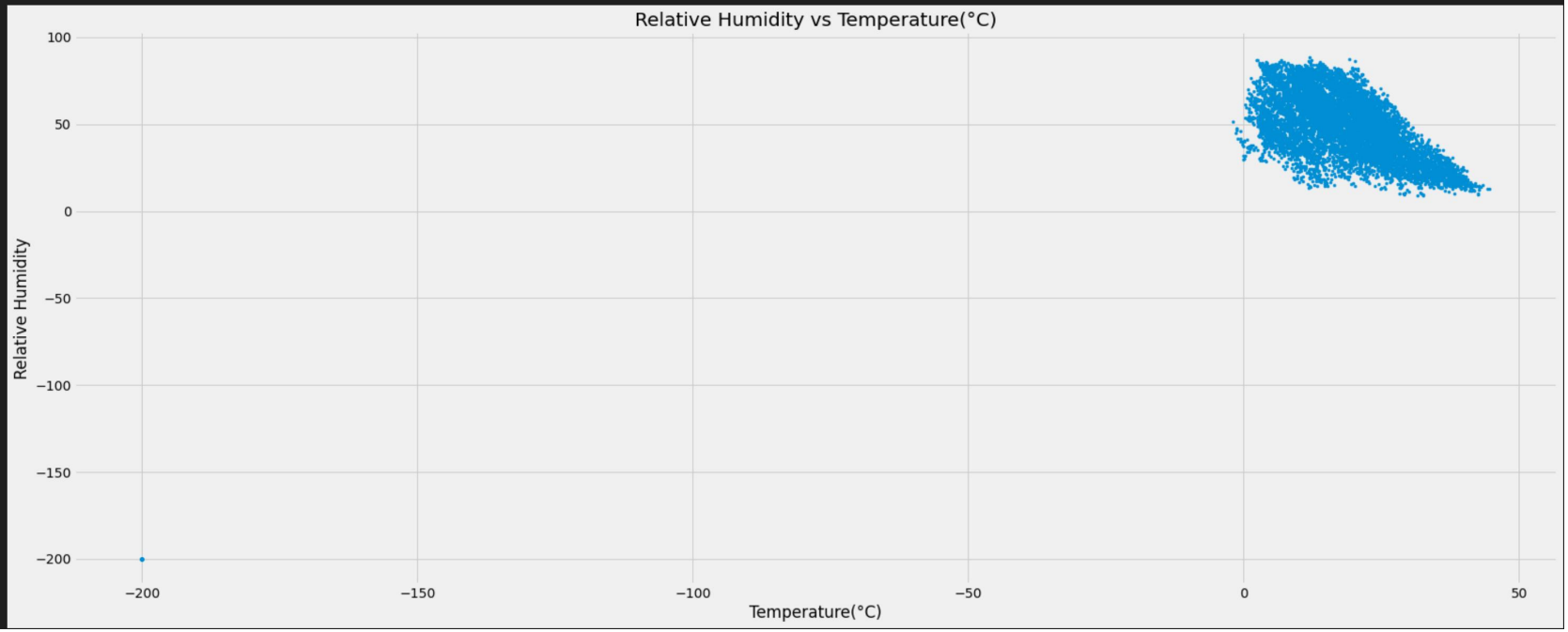
```
[10]  from sklearn.preprocessing import StandardScaler                    Python

[11]  from sklearn.ensemble import RandomForestRegressor                  Python

[12]  from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score   Python

[13]  from sklearn.svm import SVR                                         Python

      ▷                                                                   Python

                                                                          Python

[14]  df = pd.read_csv("Dataset/AirQualityUCI.csv")                       Python

[15]  df.head().style.background_gradient(cmap='Blues').set_properties(**{'font-family':'Segoe UI'})   Python

...
```

IMPLEMENTATION

```
plt.figure(figsize=(25,10))
plt.xlabel('Temperature(°C)')
plt.ylabel('Relative Humidity')
plt.title("Relative Humidity vs Temperature(°C)")
plt.scatter(df['T'], df['RH'], marker='.', aa=True)
```

[25] Python

```
<matplotlib.collections.PathCollection at 0x23a359f6970>
```



IMPLEMENTATION

```python
plt.figure(figsize=(5,3))
sns.distplot(reg_results['error'])
```

Python

```python
plt.figure(figsize=(10,10))
x=np.linspace(0,5,5)
plt.plot(reg_results['y_test'], reg_results['y_pred'], 'b.')
plt.plot(x, x, 'r-')
plt.xlim(2.25,4)
plt.ylim(2.25,4)
plt.title("Results", fontsize=16)
plt.xlabel("Real", fontsize=14)
plt.ylabel("Predicted", fontsize=14)
plt.savefig("Results.png")
plt.show()
```

Python

Python

```python
svr = SVR()
```

Python

```python
svr.fit(X_train,y_train)
```

Python

```python
svr.score(X_test,y_test)
```

Python

Results and Discussion

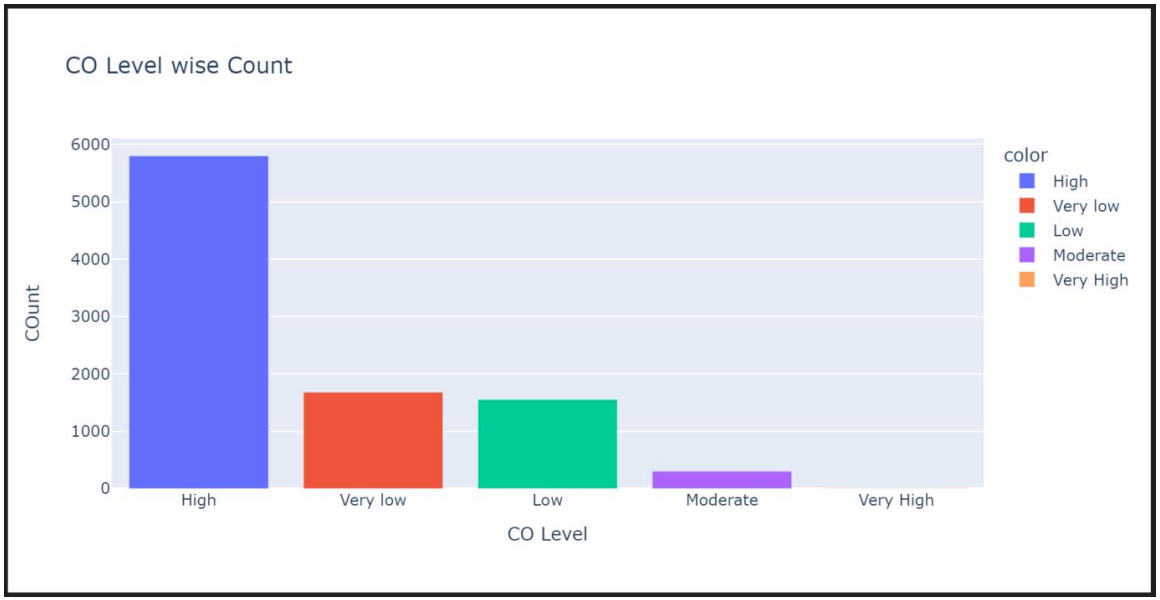
	CO_GT	PT08_S1_CO	NMHC_GT	C6H6_GT	PT08_S2_NMHC	Nox_GT	PT08_S3_Nox	NO2_GT	PT08_S4_NO2	PT08_S5_O3	T	RH	AH
count	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000	9357.000000
mean	-34.207524	1048.990061	-159.090093	1.865683	894.595276	168.616971	794.990168	58.148873	1391.479641	975.072032	9.778305	39.485380	-6.837604
std	77.657170	329.832710	139.789093	41.380206	342.333252	257.433866	321.993552	126.940455	467.210125	456.938184	43.203623	51.216145	38.976670
min	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000	-200.000000
25%	0.600000	921.000000	-200.000000	4.000000	711.000000	50.000000	637.000000	53.000000	1185.000000	700.000000	10.900000	34.100000	0.692300
50%	1.500000	1053.000000	-200.000000	7.900000	895.000000	141.000000	794.000000	96.000000	1446.000000	942.000000	17.200000	48.600000	0.976800
75%	2.600000	1221.000000	-200.000000	13.600000	1105.000000	284.000000	960.000000	133.000000	1662.000000	1255.000000	24.100000	61.900000	1.296200
max	11.900000	2040.000000	1189.000000	63.700000	2214.000000	1479.000000	2683.000000	340.000000	2775.000000	2523.000000	44.600000	88.700000	2.231000

Count: This refers to the number of times a particular value or event appears in a dataset. In the context of the table you provided, the "count" feature might represent the number of times a specific air quality measurement falls within a certain range. For example, it could indicate how many times the CO level falls between 10 and 20 ppm.

Mean: Also known as the average, the mean is the sum of all the values in a dataset divided by the number of values. It provides a central tendency of the data.

Min: The minimum value in a dataset is the single lowest value present.

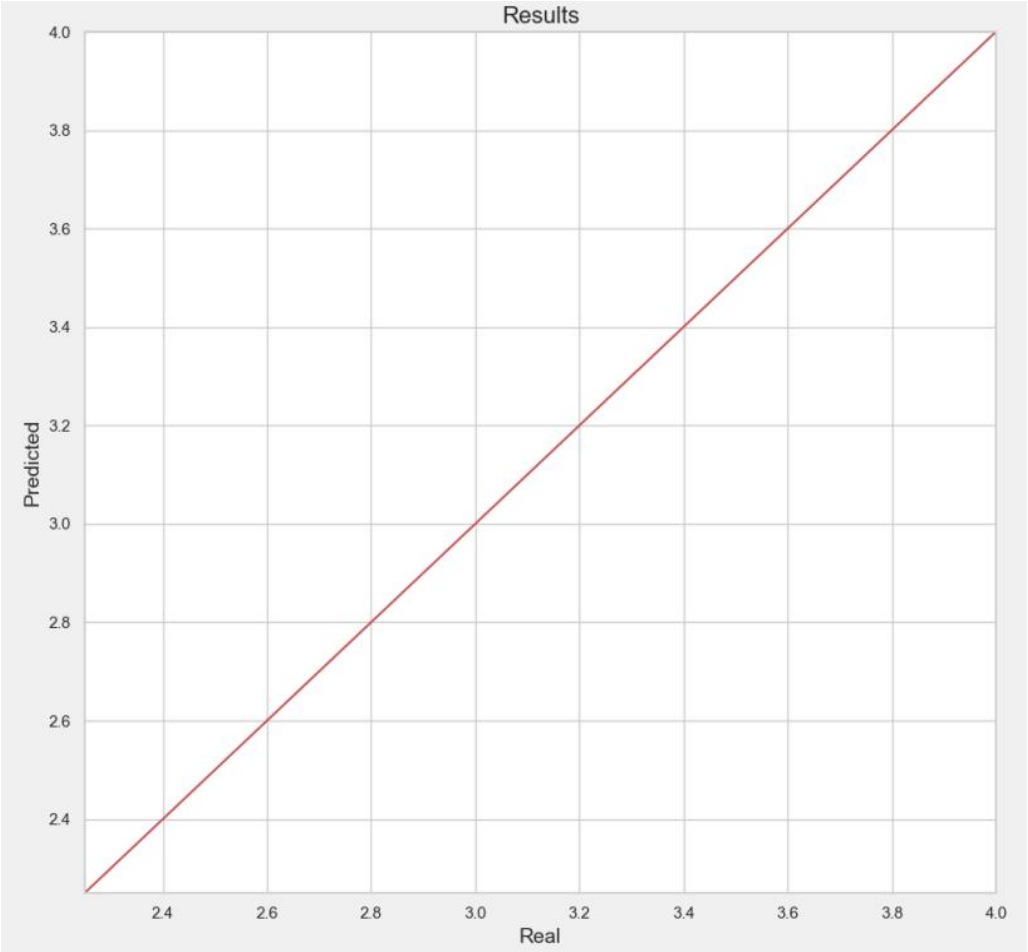
Results and Discussion



This bar graph represents the CO levels with respect to count from the dataset which varies from very low to very high. The count of "high" is in the lead with 5801 inputs. Followed by "very low" with 1683 inputs. The "low" bar has a count of 1556 inputs and "moderate" with 305 inputs. The "very high" bar is has the least count of 12 inputs. The inputs are the pre-defined values from the given dataset.

Results and Discussion

This graph represents the final results of the output. The x-axis (Real) represents the real accuate data. The y-axis(Predicted) represents the predicted values of the trained set. The values of both trained set and real data match, thus forming a straight line.



Conclusion

The contribution extends to the program facilitating data transformation, linear regression, mathematical model fitting, and the processing of prediction functions. It also enables the graphical presentation of results in the form of diagrams that illustrate the correlation between pollutants. This contributes to the evaluation of linear regression functions for their accuracy. It has been shown that the resulting functions enable predictions with high precision; the predicted values correlate very well with the obtained data. Accurate air quality forecasting has important theoretical and practical value for the public; without it, neither the government nor the public can effectively avoid the health damage caused by air pollution or improve the emergency response capability of heavy pollution days.The regression models to predict air indicators based on machine learning algorithms. The experimental results show that both the SVR-based model and the RFR-based model can achieve good results, but the RFR model performs better in experiments.

In addition, with the increasing number of samples, the time complexity of the SVR model increased cubically. Therefore, the SVR model is not suitable for processing a large number of samples. In summary, this establishes two prediction models based on different prediction scenarios, which improved the prediction accuracy of air indicators and provided guidance for modeling and analyzing urban air quality.

Future Work

Future work includes comparing additional expandable graph learning models and exploring transfer learning and node alignment techniques to reduce re-training effort in industrial scenarios. Future work could be related to the utilization of the created R program to make predictions based on a wider set of parameters, larger data sets taken over longer periods of time, a diversity of monitoring locations, adaptation to other application domains, and improving the program to use other statistical methods supported by the Python language, with a special emphasis on supporting further chemical analysis of complex interactions and processes with gasses in the atmosphere.

Outcome

Machine learning to predict Relative Humidity (RH) in an air quality dataset. It trains two models (Random Forest Regressor and Support Vector Regressor) and evaluates their performance.

Key Outcomes:

- **Model Performance:** Metrics (MAE, MSE, RMSE, R-squared) and scatter plots will reveal which model predicts RH more accurately.
- **Generalizability:** The program checks if the model performs well on unseen data.
- Feature Importance (Random Forest only): This identifies which air quality measurements were most important for the model's predictions.
- If machine learning is suitable for predicting RH in this specific data.
- Potential improvements for the models.
- Important air quality measurements for accurate RH prediction.
- It lays the groundwork for further exploration and refinement of machine learning approaches for predicting RH in air quality data.

Proof of Publication

SRM INSTITUTE OF SCIENCE & TECHNOLOGY

SINGA SELVAMANI S (RA2011003020212) <ss2360@srmist.edu.in>

Notification of Acceptance -IJRPR

Editor-International Journal of Research Publication and Reviews <editor@ijrpr.com>
Reply-To: Editor-International Journal of Research Publication and Reviews <editor@ijrpr.com>
To: Singa selvamani <ss2360@srmist.edu.in>

Fri, May 10, 2024 at 10:33 AM

Dear **Singa selvamani ,**

This is to inform you that your paper has been accepted for publication in **IJRPR,** which will be published in current issue. The Unique ID of your paper is **IJRPR-83696.**

Paper Title: SCALABLE APPROACH FOR DETECTING AIR QUALITY INFERENCE USING ENSEMBLE REGRESSORS

You are advised to complete the process for the publication of your research paper.

1. Fill & Send Copyright Transfer Form (download from our website **https://www.ijrpr.com/download/COPY-RIGHT-FORM.pdf**)

   (Note: Take a print of copyright form, fill it, scan it & send us via mail to ijrpr4@gmail.com or send it in word format with digital/scanned signatures. Do not forget to mention your paper ID in subject of mail.)

   Alternatively you can submit the Online Copyright Form by clicking the following link

   https://forms.gle/s95xHMivEBYgmmSy8

2. Submit your Publication fee

   (Note: Send the Payment receipt in image/word/pdf format along with copyright form to ijrpr4@gmail.com Do not forget to mention your paper ID in subject of mail.)

3. Paper will be published within 24 to 36 working Hours after the completion of step 1 & 2. We will provide the Soft Individual copy of the certificates to all authors of a research paper.

   Do not forget to mention your paper ID in subject.

   Publication Fee:

International Authors	**14 US Dollars (Click Here to Pay)**
Indian Authors (Author Affiliation in Paper Must be in Indian Territory)	**Rs. 499 (Click Here to Pay)**
E-Certificate	**Free**

Publication Charge includes:

- Publication of one entire research paperOnline
- Individual Soft Copy of Certificate to all author of paper.
- Editorial Fee
- Indexing, maintenance of link resolvers and journal infrastructures.

References

[1] Farzaneh Farhadi,Roberto Palacin,Phil Blythe Machine Learning for Transport Policy Interventions on Air Quality IEEE Access, 2023

[2] Guyu Zhao,Guoyan Huang,Hongdou He,Qian Wang Innovative Spatial-Temporal Network Modeling and Analysis Method of Air Quality IEEE Access, 2019 .

[3] Xiaoling Lin,Hongzhang Wang,Jing Guo,Gang Mei A Deep Learning Approach Using Graph Neural Networks for (c) Wisen IT Solutions Page 27 of 30 Anomaly Detection in Air Quality Data Considering Spatiotemporal Correlations IEEE Access, 2022

[4] Bo Liu,Shuo Yan,Jianqiang Li,Guangzhi Qu,Yong Li,Jianlei Lang,Rentao Gu A Sequence-to-Sequence Air Quality Predictor Based on the n-Step Recurrent Prediction IEEE Access, 2019

[5] Martha Arbayani Zaidan,Yuning Xie,Naser Hossein Motlagh,Bo Wang,Wei Nie,Petteri Nurmi,Sasu Tarkoma,Tuukka PetÃ¤jÃ¤,Aijun Ding,Markku Kulmala Dense Air Quality Sensor Networks: Validation, Analysis, and Benefits IEEE Sensors Journal, 2022

[6] Rady Purbakawaca,Arief Sabdo Yuwono,I. Dewa Made Subrata,Supandi,Husin Alatas Ambient Air Monitoring System With Adaptive Performance Stability IEEE Access, 2022

[7] Yangwen Yu,James J. Q. Yu,Victor O. K. Li,Jacqueline C. K. Lam A Novel Interpolation-SVT Approach for Recovering Missing Low-Rank Air Quality Data IEEE Access, 2020

[8] Ying Zhang,Yanhao Wang,Minghe Gao,Qunfei Ma,Jing Zhao,Rongrong Zhang,Qingqing Wang,Linyan Huang A Predictive Data Feature Exploration-Based Air Quality Prediction Approach IEEE Access, 2019

[9] Yuchao Zhou,Suparna De,Gideon Ewa,Charith Perera,Klaus Moessner Data-Driven Air Quality Characterization for Urban Environments: A Case Study IEEE Access, 2018.

[10] Abdelaziz El Fazziki,Djamal Benslimane,Abderrahmane Sadiq,Jamal Ouarzazi,Mohamed Sadgal An Agent Based Traffic Regulation System for the Roadside Air Quality Control IEEE Access, 2017