IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Air Pollution Prediction Using Genetic Programming

[1]Sahil
[1]PG Student – MBA,
[1]Advanced Construction Management,
[1]NICMAR University, Pune, India

Abstract: Air pollution is a pressing global problem that poses severe challenges to the environment and human health. It results from the release of harmful pollutants into the atmosphere, primarily through human activities. The burning of fossil fuels, industrial emissions, vehicular exhaust, and agricultural practices are among the major contributors to air pollution. As these pollutants accumulate in the air, they form a toxic mixture that degrades the air quality and poses a threat to ecosystems, wildlife, and human populations.

Air quality monitoring systems and early warning systems can be Implemented for extensive air quality monitoring and they can provide real-time information on pollution levels. Combined with advanced modeling techniques, this data can help predict future levels of pollution and provide communities with an early warning so they can take precautions.

Methods for predicting pollution levels include statistical models (e.g., linear regression, time series analysis), machine learning algorithms (e.g., support vector machines, random forests), Gaussian process regression, hybrid models, etc. Each method offers unique approaches to analyzing pollution data and making predictions.

GP is a data-driven approach that excels in predicting future pollution levels due to its ability to analyze large datasets, capture complex nonlinear relationships, automatically select relevant features, optimize the model structure, and generate interpretable models.GP's flexibility, adaptability, and ability to handle diverse data types make it a promising method for accurately predicting pollution levels and aiding in decision-making and mitigation strategies. However, the choice of prediction method may benefit from combining multiple approaches and domain knowledge for optimal results.

In our project, we have used GP for the prediction of NOx and SO2 for the Karve Road Ambient Air Quality Monitoring Station with promising results.

Index Terms - Air Pollution, Genetic Programming, Prediction Model, Particulate Matter (PM), Air Quality Index (AQI), Data Mining, Machine Learning, Meteorological Variables, Environmental Monitoring, Statistical Methods, Synthetic Minority Oversampling Technique (SMOTE), Decision Tree Algorithm, Artificial Neural Networks (ANN), Performance Evaluation, Environmental Parameters, MSE (Mean Square Error), MAE (Mean Absolute Error), R (Correlation Coefficient).

Chapter - 1

Introduction

1.1 Background

Global air pollution is a widespread environmental issue affecting countries and regions worldwide. It is caused by various factors, including industrial emissions, transportation, agricultural practices, and the burning of fossil fuels. The impacts of global air pollution extend beyond national borders, with pollutants being carried by wind currents and affecting air quality in distant locations. Global efforts, such as international agreements and collaborations, are crucial for addressing this issue and implementing strategies to reduce emissions, promote cleaner technologies, and mitigate the adverse effects of air pollution on human health and the environment.

Air pollution has been a significant problem for India as well, particularly in densely populated urban areas. The levels of pollutants such as sulfur dioxide and nitrogen oxides often exceed safe limits, leading to severe health issues and environmental degradation. The Indian government has implemented measures to combat air pollution, including introducing emission standards, promoting renewable energy, adopting cleaner transportation options, and implementing measures to control open burning.

The presence of nitrogen oxides (NOx) and sulfur dioxide (SO2) in the air can have significant ill effects on both human health and the environment. NOx is primarily released from vehicle emissions, power plants, and industrial processes. When inhaled, NOx reacts with other compounds in the atmosphere to form ground-level ozone, a major component of smog. Ozone can cause respiratory issues, worsen asthma symptoms, and lead to lung diseases. Additionally, NOx contributes to the formation of acid rain, which damages crops, forests, and aquatic ecosystems.

SO2 is mainly emitted from industrial processes and the burning of fossil fuels, particularly high-sulfur coal and oil. When released into the air, SO2 reacts with moisture to form sulfuric acid, a major component of acid rain. Acid rain can harm aquatic life, damage buildings and monuments, and harm crops and forests. Moreover, SO2 can irritate the respiratory system, leading to respiratory problems and exacerbating existing respiratory conditions like asthma

However, the prediction of the most approximate values of the pollutants is needed to address this complex issue and safeguard the health and well-being of the Indian population.

Genetic programming (GP) can be utilized in the development of air quality monitoring systems and early warning systems. GP can be applied to analyze large datasets collected from air quality monitoring stations, meteorological sensors, and other relevant sources. By evolving mathematical models through a process inspired by natural evolution, GP can identify complex patterns and relationships in the data, including nonlinear dependencies and interactions between different variables. This enables GP to automatically generate predictive models that can forecast future pollution levels based on current and historical data. By combining historical data with meteorological information and other relevant factors, these models can generate forecasts that help anticipate pollution spikes and issue early warnings. The use of GP in air quality monitoring and early warning systems enhances the accuracy of predictions, allowing for more effective decision-making and proactive measures to mitigate the impacts of air pollution on public health and the environment.

1.2 Problem Statement

Accurate prediction of air pollution levels is crucial for developing effective mitigation strategies, implementing appropriate regulations, and protecting public health. Traditional statistical and machine learning models have limitations in handling complex nonlinear relationships and capturing the dynamics of air pollution. Genetic Programming, a form of evolutionary computation, has emerged as a robust approach to addressing these challenges. Additionally, software tools provide an efficient platform for implementing GP-based models.

1.3 **Objectives**

1)Review the existing research on Air Pollution forecasting using Genetic Programming.
2)To Analyze Air Pollution data of Pune city one day ahead.
3)To develop GP(Genetic Programming) model for predicting NOx and SOx concentration.

Chapter - 2

Literature Review

Aditya (et al.2018) employed machine algorithms to detect and forecast the PM2.5 concentration level on the basis of a dataset containing atmospheric conditions in a specific city. They also predicted the PM2.5 concentration level for a particular date. First of all, they classify the air as polluted or not polluted by using the Logistic Regression algorithm, and then the Auto Regression algorithm was used to predict the future value of PM2.5 depending upon previous records.

Alkasassbeh (et al.2014) Monitoring and controlling air pollutants have been major environmental concerns so far. This paper aimed to put in hand a symbolic regression prediction model based on genetic programming. The main objective of the prediction model was to predict particulate matter in New York City, Jordan. This study analyzed the recording of five monitoring stations around the Al-Fahais cement plant. It incorporated and measured meteorological input variables such as relative humidity (R), atmospheric temperature, and wind speed.

Esfandani (et al.2017). In this paper, three models were proposed to predict Tehran air pollution based on information from the Mehrabad weather station. The accuracy and performance of the three models could be placed as BP-PSO, BP-GA, and BP. In other words, the error rate increased. These results mostly focused on PM10.

Gupta (et al.2018) VIT, Tamil Nadu. The present assessed the performance of the three best data mining models. (SVR, RFL, and CR) for predicting the accurate AQI data in India’s most populated cities. The synthetic minority oversampling technique (SMOTE), was used to equalize the class data to get better and consistent results, then using the statistical method for a like RMSE, MAE, MSE, and R2 to confirm the better result with higher accuracy.

Kumar (et al.2016). The present assessment evaluated the performance of the three best data mining models (SVR, RFL, and CR) for predicting accurate AQI data in India's most populated cities. The synthetic minority oversampling technique (SMOTE) was used to equalize the class data to obtain better and consistent results. Then, statistical methods such as RMSE, MAE, MSE, and R2 were used to confirm the better result with higher accuracy.

Gayatri (et al.2011). The proposed system would have definitely helped in improving the prediction of air pollution in Coimbatore city. Prediction using the decision tree-based C415 Algorithm technique improved the performance and reduced the complexity of the air pollution prediction model.

Gaganjyot (et al.2011). This report was a recent literature study review and compared current research work on air quality evaluation based on big data analysis. Machine learning models and techniques, along with advancements in IoT infrastructures and big data technologies, were discussed for real-time air quality monitoring in future smart cities.

Dr Dinde (et al.2020). According to this review, most researchers focused on forecasting the Air Quality Index (AQI) and pollutant concentration levels, which provide a better understanding of AQI. Artificial Neural Networks (ANN), linear regression, and logistic regression were found to be the preferred choices of many researchers for predicting air pollution concentrations.

Drewwill (et al.2019). A new approach was proposed to use Long Short-Term Memory (LSTM) to predict air quality. One of the challenges with the LSTM algorithm is the selection of parameters such as the window size and number of units in LSTM. In this study, Genetic Algorithm (GA) was utilized to address this issue. GA provides a more flexible performance as it allows for adapting the LSTM input sequence's fixed length, which is necessary for predicting pollution levels.

Lakhan (et al.2022) In this article, we conducted exploratory data analysis (EDA) and air quality index (AQI) prediction on a dataset of air pollution collected from various cities in India between 2015-2020. Among these cities, Delhi and Ahmedabad were found to be the worst affected by air pollution. The COVID-19 lockdown had a significant impact on reducing pollution levels across cities. Support Vector Regression (SVR) and Random Forest models were employed as machine learning techniques to make predictions on AQI, and both models performed well, achieving accuracies of 98% and 88% respectively.

Liang (et al.2020). To predict air pollution, data related to the day of the week, month of the year, topography, meteorology, and pollutant rate were utilized. Several methods were employed, including regression support vector machine, geographically weighted regression, Artificial Neural Networks (ANN), and autoregressive non-linear neural networks. A prediction model was proposed to enhance the mentioned methods, resulting in a reduction of the error percentage by 57%, 47%, 47%, and 94% respectively. Among the algorithms used, the autoregressive nonlinear neural network was found to be the most reliable for predicting air pollution.

V.M. (et al.2020). The concentration of air pollutants in ambient air is influenced by meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. The Air Quality Index (AQI) is utilized to measure the quality of air. The proposed work employed a supervised learning approach using various algorithms such as Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF). The results demonstrated that the AQI prediction achieved through RF is promising, and a thorough analysis of the results was performed.

Goyal (et al.2011). A study was presented that aimed to forecast the daily Air Quality Index (AQI) value for the city of Delhi, India. The researchers utilized previous records of AQI and meteorological parameters, employing Principal Component Regression (PCR) and Multiple Linear Regression techniques. The prediction of the daily AQI for the year 2006 was carried out using data from the years 2000-2005, and different equations were - -employed. Subsequently, the predicted AQI values were compared with the observed AQI values of 2006 for different seasons: summer, monsoon, post-monsoon, and winter. Multiple Linear Regression was used for this comparison. Principal Component Analysis (PCA) was applied to identify collinearity among the independent variables. The principal components derived from PCA were then used in Multiple Linear Regression to address collinearity issues and reduce the number of predictors. The results showed that Principal Component Regression performed better in predicting the AQI during the winter season compared to other seasons. It is important to note that this study only considered meteorological parameters for forecasting future AQI and did not include ambient air pollutants that may have adverse health effects.

Liu (et al.2019). The study focused on two different cities, Beijing and Italian cities, to forecast the Air Quality Index (AQI) and predict the concentration of NOx. Two publicly available datasets were utilized for this purpose. The first dataset, obtained from the Beijing Municipal Environmental Centre, covered the period from December 2013 to August 2018. It consisted of 1738 instances and included fields such as hourly averaged AQI, as well as concentrations of PM2.5, O3, SO2, PM10, and NO2 in Beijing. The second dataset comprised data from Italian cities and spanned from March 2004 to February 2005, with 9358 instances. This dataset included attributes such as the hourly averaged concentration of CO, Nonmethane Hydrocarbons, Benzene, NOx, and NO2. However, the study primarily focused on predicting NOx concentration as it is an important predictor for air quality evaluation. To make predictions for AQI and NOx concentration, the study employed Support Vector Regression (SVR) and Random Forest Regression (RFR) techniques. SVR demonstrated better performance in predicting AQI, while RFR showed better performance in predicting NOx concentration.

Guan (et al.2018). Various machine learning algorithms were utilized to predict the PM2.5 concentration. Data were collected from the official website of the Environment Protection Agency (EPA) for the city of Melbourne, including PM2.5 air parameters. Additionally, unofficial data from the Airbeam mobile device used to measure PM2.5 values was collected. The machine learning algorithms employed were Artificial Neural Network (ANN), Linear Regression (LR), and Long Short Term Memory (LSTM) recurrent neural network. Among these algorithms, LSTM demonstrated the best performance and accurately predicted high PM2.5 values.

HeidarMaleki (et al.2019).The researchers predicted the hourly concentration values for ambient air pollutants, including NO2, SO2, PM10, PM2.5, CO, and O3, for four air quality monitoring stations in Ahvaz, Iran. Ahvaz is known as one of the most polluted cities in the world. They also calculated and predicted the Air Quality Index (AQI) and Air Quality Health Index (AQHI) for these stations. For the prediction of air pollutant concentrations and air quality indices from August 2009 to August 2010, they employed the Artificial Neural Network (ANN) machine learning algorithm. The inputs to the ANN algorithm consisted of meteorological parameters, air pollutant concentrations, time, and date.

Sharma (et al.2018). The researcher conducted a detailed data analysis of air pollutants from 2009 to 2017 and specifically focused on the observation of air pollutant trends in Delhi, India during 2016-2017. She predicted the future trends of various pollutants including Sulfur Dioxide (SO2), Nitrogen Dioxide (NO2), Suspended Particulate Matter (PM), Ozone (O3), Carbon Monoxide (CO), and Benzene using data analytics and time-series regression forecasting techniques based on previous records. The study primarily examined the AnandVihar and Shadipur monitoring stations in Delhi. The results indicated a significant increase in PM10 concentration levels, with evident increases in NO2 and PM2.5, indicating heightened pollution levels in Delhi. CO was predicted to decrease by 0.169mg/m3, while NO2 concentration levels were projected to increase by 16.77 µg/m3 in the coming years. Ozone was predicted to increase by 6.11mg/m3, Benzene was expected to decrease by 1.33mg/m3, and SO2 was forecasted to increase by 1.24µg/m3.

Shakir (et al.2018). The researchers conducted an analysis of the proportions of various air pollutants, including NO, NO2, CO, PM10, and SO2, based on the time of day and day of the week. They also examined the impact of environmental parameters such as temperature, wind speed, and humidity on these air pollutants. The data for the study was collected from the pollution control board of Karnataka. By utilizing the WEKA tool and employing the ZeroR algorithm, the study found that the concentration levels of air pollutants were higher on working days, particularly during peak hours, and lower on weekends or holidays. Furthermore, by employing the Simple K-means Clustering algorithm, the study revealed the relationships and dependencies between environmental factors like temperature, wind speed, humidity, and the concentrations of air pollutants (NO, NO2, PM10, CO, and SO2).

TikheShruti (et al.2013). The research utilized two soft computing algorithms, namely Artificial Neural Network (ANN) and Genetic Programming (GP), for predicting future concentration levels of air pollutants, including Oxides of Sulfur (SOx), Oxides of Nitrogen (NOx), and Respirable Suspended Particulate Matter (RSPM), in Pune city, Maharashtra. Pune is listed as the second most polluted city in India. The study developed a total of six models, three for each algorithm (ANN and GP), based on hourly average data of pollutant concentrations spanning over a period of more than seven years (2005-2011). The results showed that the GP algorithms outperformed the ANN models in terms of predictive performance for the given pollutants.

Chaloulakou (et al.2003) The researchers in this study implemented the Artificial Neural Network (ANN) and Multiple Linear Regression (MLR) algorithms to forecast the concentration of PM10 over a two-year period in Athens, Greece. Prior to inputting the data into the ANN, the dataset was divided into three subsets: the training dataset, which contained two-thirds of the available records, and the remaining cases were equally divided into the validation and test sets. A comparison between ANN and MLR was conducted, indicating that ANN outperformed MLR in terms of performance. The study concluded that if properly trained, ANN can provide adequate prediction solutions or results as per the requirements.

Gunasekaran (et al.2012) The main objective of this study was to monitor the air quality in the Salem Swadeswari College area of Tamil Nadu, covering the period from April 2011 to March 2011. The findings indicated that the area did not have any significant pollution issues related to pollutants such as Sulfur Dioxide, Oxides of Nitrogen, and Suspended Particulate Matter, as their annual average concentrations fell within the range of national standards. However, the annual average concentration of the pollutant PM10 was slightly higher than the national standard levels. Additionally, the monthly 24-hour average concentration of PM10 during the same year exceeded the national standard level, except for the months of July to October.

Chapter - 3

Materials and Methods

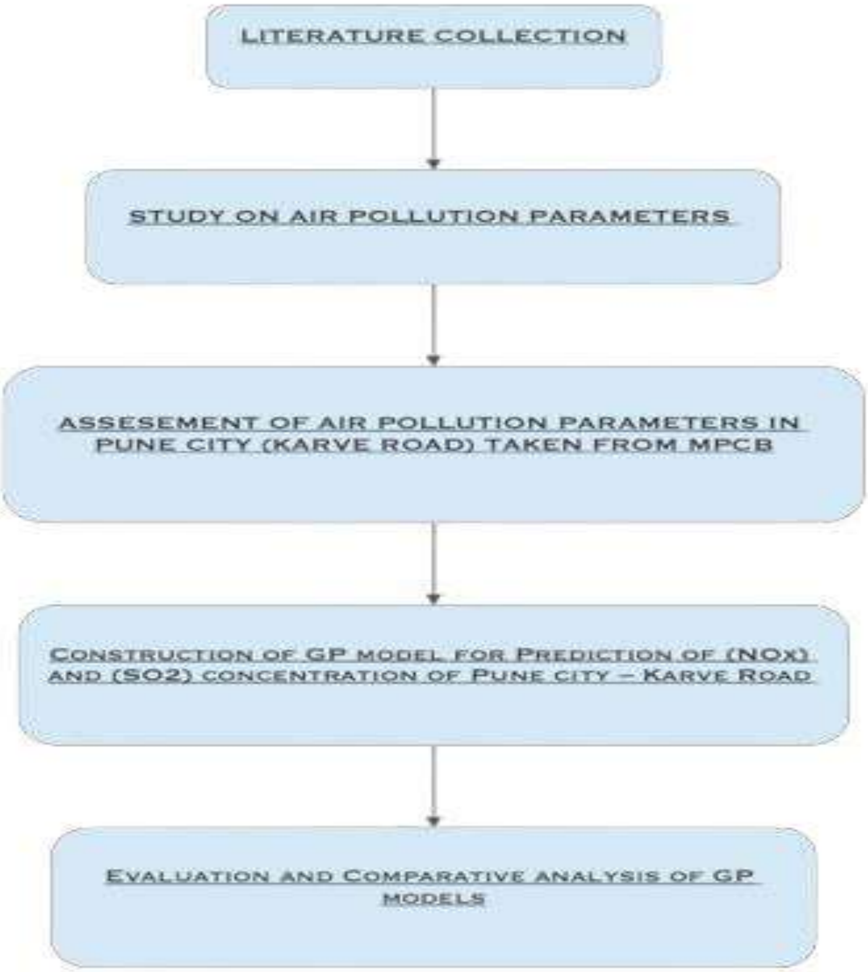


Fig: 3.1: Flowchart of methodology

3.1 Data Collection and Pre-processing

3.1.1 **Central Pollution Control Board (CPCB) Data**

The Central Air Pollution Control Board (CPCB) plays an important role in monitoring and controlling air pollution in India. It collects extensive air quality data from various monitoring stations across the country. The CPCB data includes information on pollutants such as PM, NOx, and SO2 measured at different locations and time intervals. These data serve as a valuable resource for researchers studying air pollution and developing prediction models.

3.1.2 **Pune Ambient Air Monitoring Station Data**

The ambient air monitoring station in Pune, India, provides localized air quality data specific to the region. These data include measurements of various pollutants, meteorological parameters, and other relevant factors. Researchers often utilize this data to assess the air pollution levels in Pune and develop models tailored to the specific environmental conditions of the area.
We have collected the data for the duration of 2138 calendar days (during the interval of 01/01/2017 - 09/11/2022), to assess the air pollution levels and have got a total of 2128 data points.

3.1.3 **Data Pre-processing Techniques**

Data pre-processing is a crucial step in air pollution prediction, aimed at ensuring data quality and preparing it for model development. Common pre-processing techniques include data cleaning to remove outliers and inconsistencies, data normalization to scale variables, and feature selection to identify the most relevant predictors. Additionally, missing data imputation and handling of temporal dependencies may be necessary for time series data.

This dataset includes 35064 records with multi-features in each station. The period of recording is from March 1st, 2013, to February 28th, 2017. The data are composed of: date, the concentration of $PM2.5$, $PM10$, Sulfur dioxide $SO2$, Nitrogen dioxide $NO2$, carbon monoxide CO, ozone $O3$, dew point, temperature, atmospheric pressure, combined wind direction, cumulated wind speed, cumulated hours of snow, and rain. However, Air quality and meteorological monitoring equipment will cause leakage in data collection due to machine failure, due to some uncontrollable reasons. The existence of such missing values will have some impact on data mining.

The equation of the spline linear interpolation function is:

$$SL(x) = f(x_{i-1})\frac{x - x_i}{x_{i-1} - x_i} + f(x_i)\frac{x - x_{i-1}}{x_i - x_{i-1}} \quad x \in [x_{i-1}, x_i], i = 1, 2, 3, \ldots, n$$

In order to improve the prediction accuracy, we normalize the values of $PM2.5$,2.5 concentration using the Min-Max normalization, the method is given in the equation :

$$x = \frac{x - min}{max - min}$$

In machine learning applications, feature selection is an essential step that can be done in several ways. Most of the previous work has applied a mathematical correlation to find the relationship between the input and output variables. When there are many features to enter the network for training, finding the correlation between the target output value and those features reduces the complexity of training and improves performance.

The Pearson correlation is the most popular method used to find the correlation between two variables. The following equation can calculate its coefficient r:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x and y represent variables, and x and y represent the mean of the variables.

Evaluation index of the models:

Once the structure of the model is determined, the training set is used to train the network until convergence. In order to assess the efficiency of the model, three indicators are used in this article, including the mean absolute error (MAE), the mean squared error (RMSE), and the coefficient of determination ($R2$ square).

MAE

MAE (Mean Absolute Error) is the arithmetic mean of the absolute values of the deviations between the true value and the model prediction value of all samples, which can better reflect the real prediction error situation. The calculation formula is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

RMSE

RMSE (Root Mean Square Error) is the square root of the mean of the square of all of the errors. It may well reflect the accuracy of the prediction error. The calculation formula is shown below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$$

R2

The coefficient of determination reflects the proportion of all variations of the dependent variable that can be explained by the independent variable through the regression relationship. The closer the value of *R*2�2 is to 1 becomes, the better the independent variable can explain the dependent variable. See the calculation formula below:

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

3.2 Genetic Programming (GP) and Discipulus Software

3.2.1 Genetic Programming: Overview and Concepts

Genetic Programming is a machine learning technique that mimics the process of natural evolution to evolve computer programs capable of solving complex problems. GP starts with an initial population of randomly generated programs and applies genetic operators such as selection, crossover, and mutation to iteratively evolve and improve the programs. The fitness of each program is evaluated based on how well it solves the problem at hand.

3.2.2 Introduction to Discipulus Software

Discipulus is a software tool specifically designed for Genetic Programming, providing a user-friendly interface and advanced features for developing and analyzing GP-based models. It offers functionalities for data pre-processing, model representation, fitness function design, evolutionary operators, and performance evaluation. Discipulus simplifies the implementation of GP-based models, making them accessible to researchers and practitioners in the field of air pollution prediction.

3.2.3 Advantages of Genetic Programming and Discipulus in Air Pollution Prediction

Genetic Programming and Discipulus software offer several advantages in the domain of air pollution prediction:

- Ability to handle complex relationships: Genetic Programming excels in capturing nonlinear and complex relationships between air pollution variables, enabling the modeling of intricate patterns in the data.

- Automatic feature selection: Genetic Programming can automatically select relevant features from the input data, reducing the dimensionality and focusing on the most informative variables.

- Adaptability and flexibility: Genetic Programming's evolutionary nature allows it to adapt to changing environmental conditions and adjust the model structure accordingly.

- Interpretability: GP-based models offer interpretability, as the evolved programs can be analyzed and understood to gain insights into the factors influencing air pollution levels.

- Efficiency and usability: Discipulus software provides a user-friendly interface and streamlined workflows, simplifying the development and deployment of GP-based air pollution prediction models.

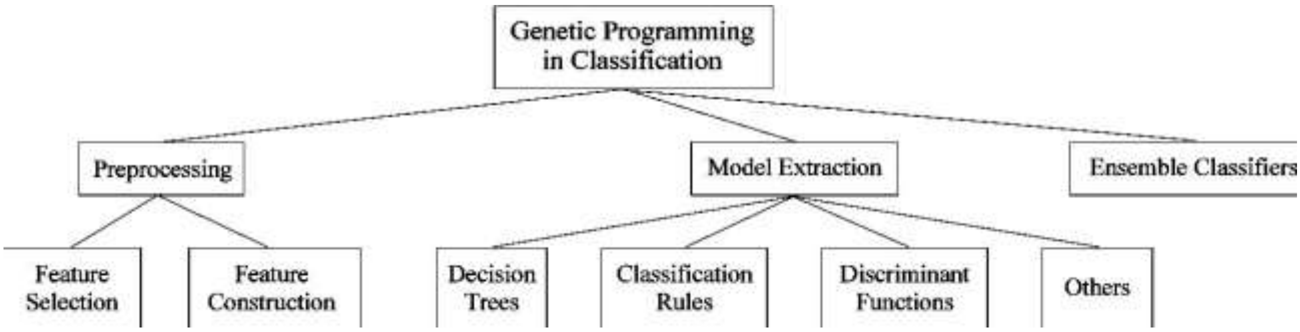


Fig: 3.2 - Classification of genetic programming

3.3 Methodologies for Air Pollution Prediction

3.3.1 **GP-Based Air Pollution Prediction Models**

GP-based models for air pollution prediction involve the evolutionary search for mathematical equations or programs that can accurately estimate pollutant levels based on input variables. These models typically use historical air pollution data, meteorological parameters, and other relevant features as input. The GP process evolves a population of candidate programs, applying genetic operators to improve their fitness and predictive capabilities.

3.3.2 **Fitness Function Design**

The fitness function in GP defines the objective or goal that the evolved programs aim to optimize. In air pollution prediction, the fitness function assesses how well a program predicts the pollutant levels compared to the actual values. Various fitness functions, such as mean squared error or mean absolute error, can be used to quantify the fitness of the evolved programs.

3.3.3 **Evolutionary Operators**

Evolutionary operators including selection, crossover, and mutation, drive the evolution process in GP. Selection determines which programs are selected as parents for reproduction, crossover combines genetic material from two parents to create offspring, and mutation introduces random changes in the offspring's genetic material to promote diversity and exploration.

3.3.4 **Model Complexity Control**

Controlling the complexity of evolved programs is essential to prevent overfitting and ensure generalization. Techniques such as size constraints, depth limits, and parsimony pressure can be employed to control the complexity of GP-based air pollution prediction models.

3.3.5 **Discipulus Software for Air Pollution Prediction**

Discipulus software provides a comprehensive platform for implementing GP-based air pollution prediction models. It offers functionalities for data pre-processing, model representation, fitness function design, evolutionary operators, and performance evaluation. The software simplifies the implementation process, allowing researchers to focus on model development and analysis.

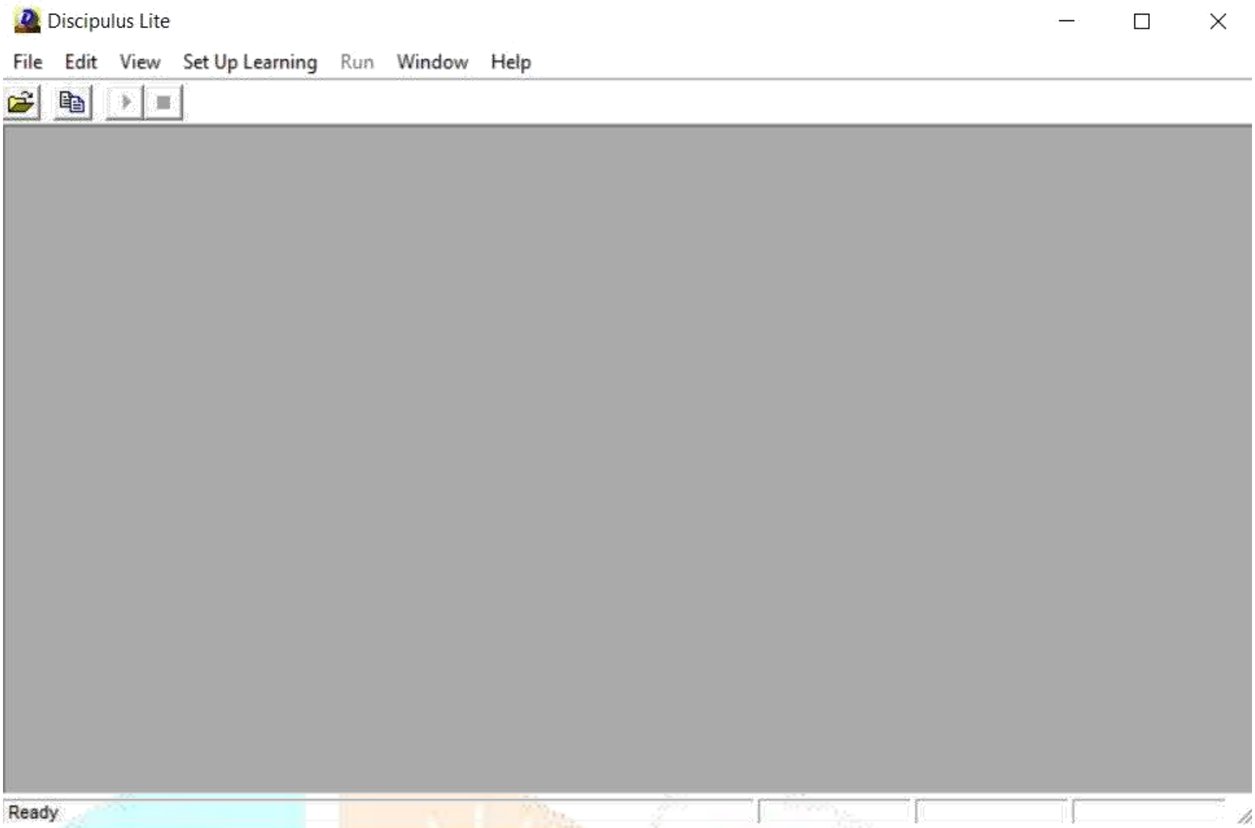


Fig: 3.3 - Discipulus software

3.3.6 **Feature Selection and Extraction**

Discipulus software incorporate feature selection techniques to identify the most informative variables for air pollution prediction. It offers a range of feature selection algorithms, including genetic-based approaches, to automatically select relevant features from the input data. Additionally, feature extraction methods, such as principal component analysis, can be utilized to transform the input variables into a lower-dimensional space.

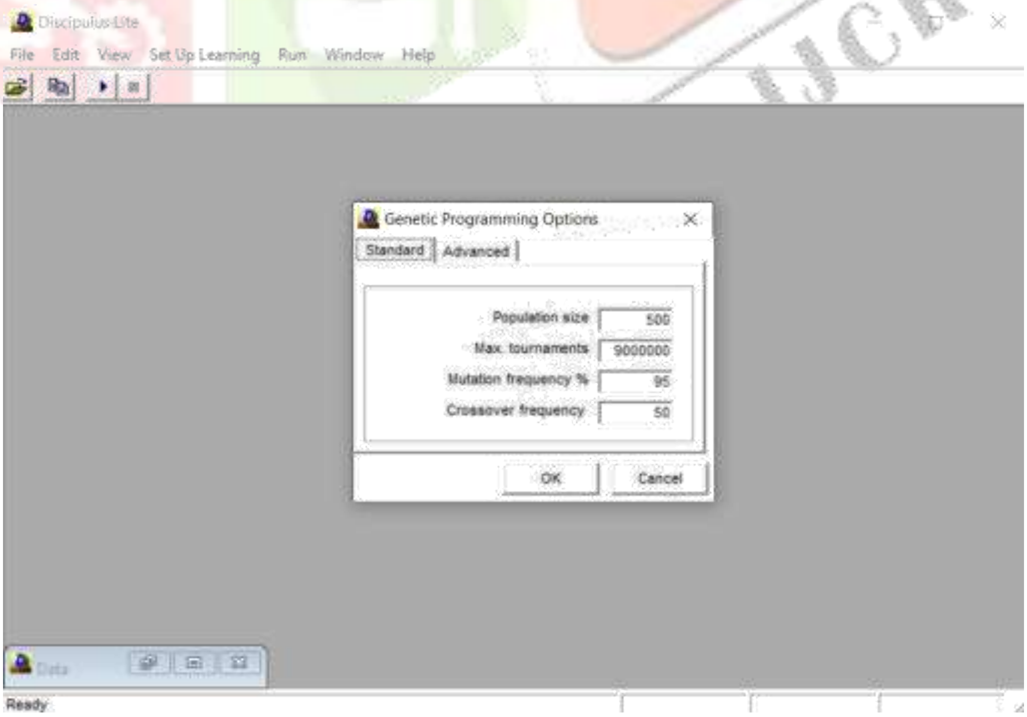


Fig: 3.4 - Genetic programming settings in discipulus

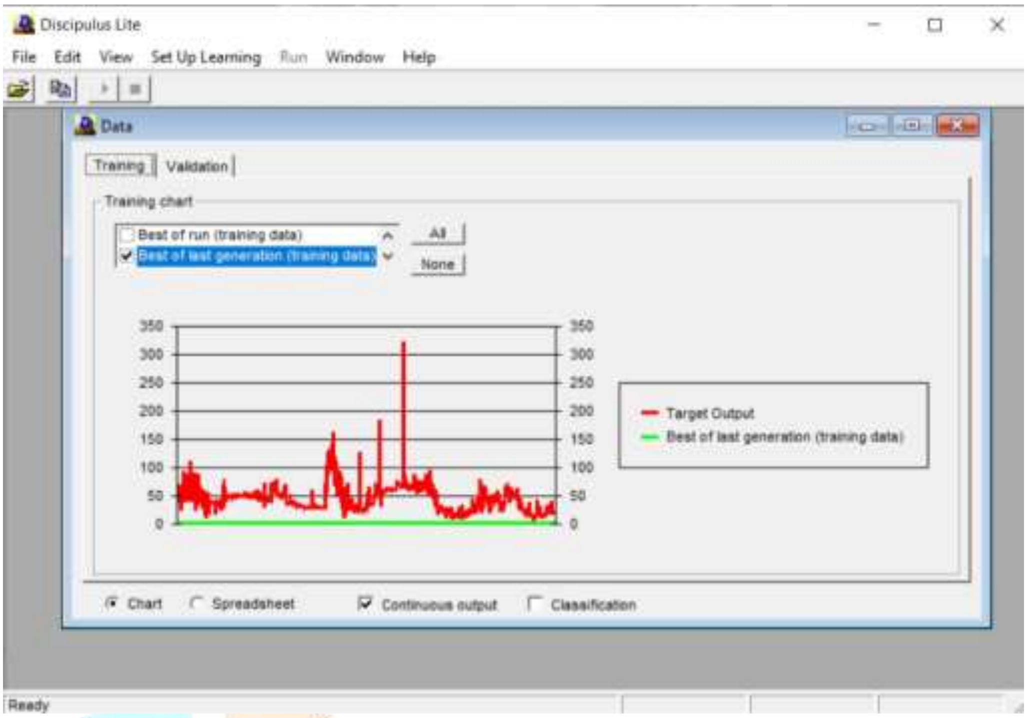


Fig: 3.5 - Data for training

3.3.7 **Model Building and Training**

Discipulus facilitates the development and training of GP-based air pollution prediction models. It provides tools for model representation, allowing researchers to define the program structure and search space. The software supports the application of evolutionary operators, enabling the evolution of candidate programs through generations. Model training involves optimizing the program structures and their associated parameters to achieve accurate predictions.

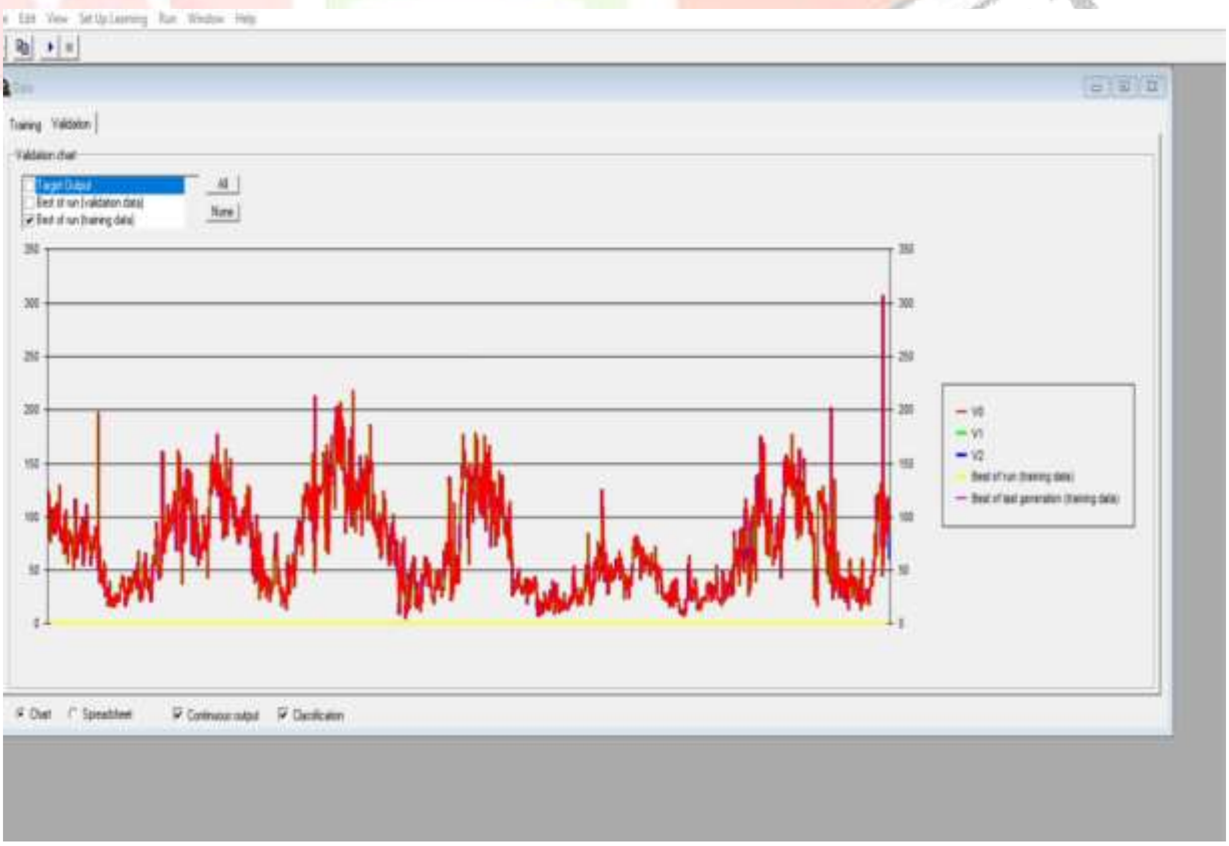
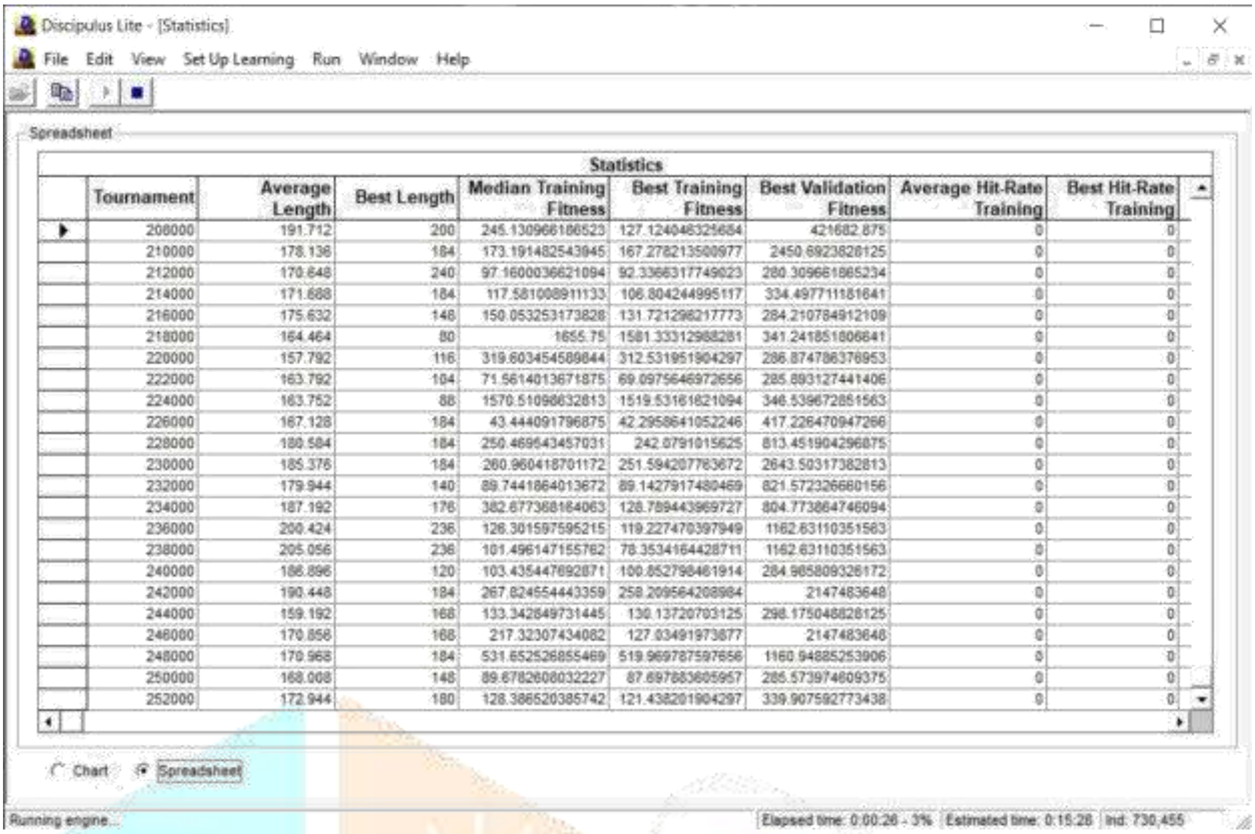


Fig: 3.6 - Best of run training graph



Statistics							
Tournament	Average Length	Best Length	Median Training Fitness	Best Training Fitness	Best Validation Fitness	Average Hit-Rate Training	Best Hit-Rate Training
208000	191.712	200	245.130966186523	127.124046325684	421682.875	0	0
210000	178.136	184	173.191482543945	167.278213500977	2450.6923828125	0	0
212000	170.648	240	97.1600036621094	92.3366317749023	280.309661865234	0	0
214000	171.688	184	117.581008911133	106.804244995117	334.497711181641	0	0
216000	175.632	148	150.053253173828	131.721298217773	284.210784912109	0	0
218000	164.464	80	1655.75	1581.33312988281	341.241851806641	0	0
220000	157.792	116	319.603454589844	312.531951904297	286.874786376953	0	0
222000	163.792	104	71.5614013671875	69.0975646972656	285.893127441406	0	0
224000	163.752	88	1570.51098632813	1519.53161621094	346.539672851563	0	0
226000	167.128	184	43.444091796875	42.2958641052246	417.226470947266	0	0
228000	180.584	184	250.469543457031	242.0791015625	813.451904296875	0	0
230000	185.376	184	260.960418701172	251.594207763672	2643.50317382813	0	0
232000	179.944	140	89.7441864013672	89.1427917480469	821.572326660156	0	0
234000	187.192	176	382.677368164063	128.789443969727	804.773864746094	0	0
236000	200.424	236	126.301597595215	119.227470397949	1162.63110351563	0	0
238000	205.056	236	101.496147155762	78.3534164428711	1162.63110351563	0	0
240000	186.896	120	103.435447692871	100.852798461914	284.985809326172	0	0
242000	190.448	184	267.824554443359	258.209564208984	2147483648	0	0
244000	159.192	168	133.342849731445	130.13720703125	298.175048828125	0	0
246000	170.856	168	217.32307434082	127.03491973877	2147483648	0	0
248000	170.968	184	531.652526855469	519.969787597656	1160.94885253906	0	0
250000	168.008	148	89.6782608032227	87.697883605957	285.573974609375	0	0
252000	172.944	180	128.386520385742	121.438201904297	339.907592773438	0	0

Fig: 3.7 - Real time spreadsheet

3.3.8 **Model Scrutiny and Validation**

Discipulus offers capabilities for evaluating and validating the performance of GP-based air pollution prediction models. Researchers can assess the model's accuracy using various performance evaluation metrics, such as root mean squared error, mean absolute percentage error, or the coefficient of determination (R-squared). Cross-validation techniques, such as k-fold cross-validation, can be employed to assess the model's generalization ability.

3.4 Performance Evaluation Metrics

3.4.1 **Accuracy Measures**

Accuracy measures quantify how closely the predicted air pollution levels align with the actual observed values. Common accuracy measures used in air pollution prediction include mean squared error (MSE), mean absolute error (MAE), and correlation coefficient (R).

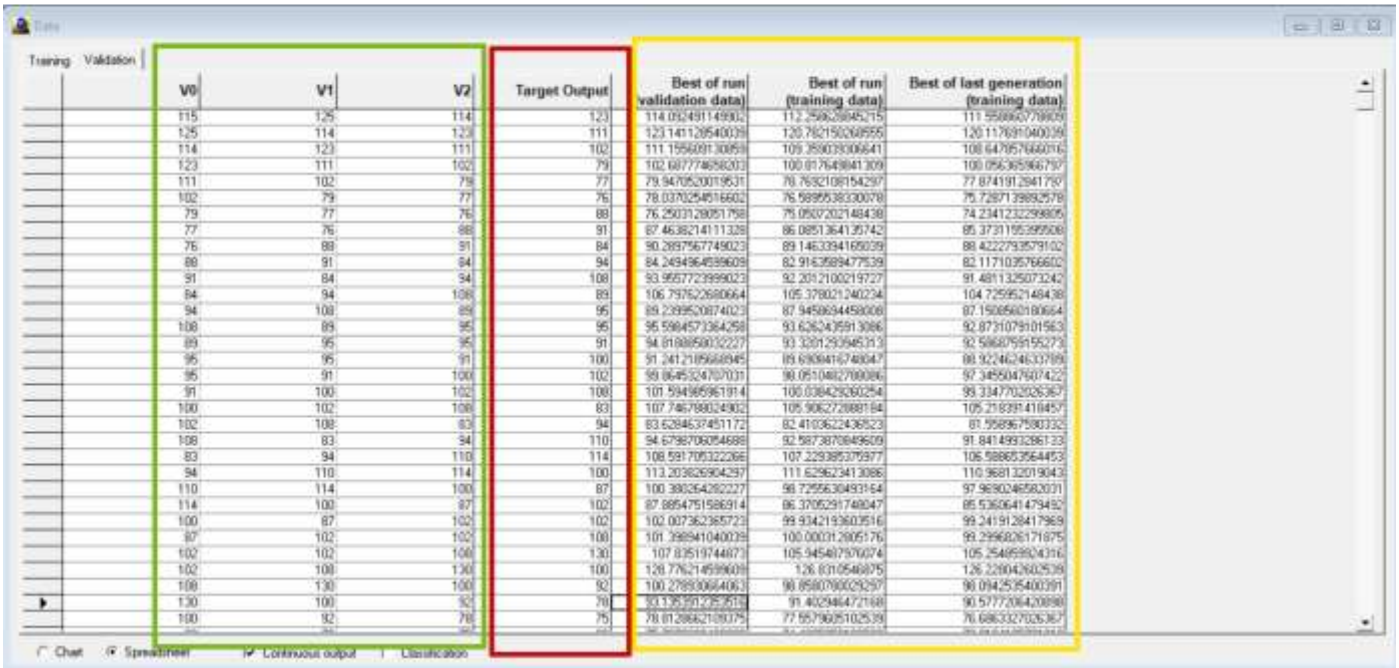


Fig: 3.8 – High efficiency GP Model Prediction

Chapter - 4

Result and Discussion

Table no. 4.1: Statistical analysis of dataset for NOx concentration

Descriptive Statistics	
NOx	
Mean:	43.40443605
Median:	41
Mode:	27
Range:	314
Standard Deviation:	21.82573697
Sample Size:	2119
Maximum Value:	323
Minimum Value:	9
Most Frequent Value:	27
Least Frequent Value:	112
Unique Values:	115

From the above Table no.4.1, observations are as follows:

During the interval from January 1, 2017, to September 11, 2022, the average concentration of NOx on Karve Road AQM station was found to be 43.40443605. This value represents the mean, which is the arithmetic average of the dataset, whereas the median, which represents the middle value of the NOx concentration dataset, was determined to be 41. This indicates that half of the observed values were above 41 and the other half were below 41.

Furthermore, the mode, which refers to the concentration value with the highest frequency, revealed that the most common or prevailing NOx concentration recorded was 27. This implies that 27 appeared more frequently in the dataset compared to other values and to understand the spread or variability of the NOx concentration well, the range was calculated. The range is a simple measure that involves finding the difference between the maximum and minimum values.The maximum and the minimum values as observed, were 323 and 9 respectively.

In this case, the range of the NOx concentrations was determined to be 314, indicating the span between the highest and lowest values in the dataset.

In addition to that, the standard deviation was calculated to measure the degree of variability or dispersion of the NOx concentrations within the dataset. The standard deviation was found to be 21.82573697, indicating the extent to which the concentrations deviate from the mean. A higher standard deviation suggests a greater dispersion of values, while a lower standard deviation signifies a more concentrated dataset.

Among the observed NOx concentrations in the dataset for Karve Road AQM station between January 1, 2017, and September 11, 2022, the most frequently occurring value was 27 wherein, the least frequent value was 112.

4.1 NOx Prediction Model

Table no. 4.2: Result of NOx model

POPULATION SIZE	PROPORTION TRA: VAL	MSE (AVG)	MAE (AVG)	R
500	80:20	181.5509185	7.128774561	0.821082458
	70:30	1728.099131	29.27853986	0.775886619
	60:40	81.10954363	3.985189964	0.790142552
1000	80:20	164.7602333	6.305312685	0.83644989 ↑
	70:30	1728.980428	29.24415986	0.775469946

	60:40	80.33125249	3.950132869	0.793874039
1500	80:20	183.6592449	7.834191618	0.826547291
	70:30	1727.582006	29.28660638	0.777568907
	60:40	84.77771724	4.02754064	0.78398406

Table no. 4.2 presents the maximum R value obtained from a model that was prepared for a population size of 1000, a mutation frequency of 80%, and a crossover frequency of 60%. This R value serves as an indicator of the model's performance in predicting the NOx concentration.

The maximum R value is a measure of the correlation or goodness-of-fit between the predicted and observed values in the model. It represents the highest level of agreement between the model's predictions and the actual measurements. A higher R value indicates a stronger correlation and suggests that the model has captured the patterns and variations in the NOx concentration well.

In addition to the maximum R value, the table also includes the predicted and observed values of the NOx concentration. These values represent the estimates generated by the model and the actual measurements recorded, respectively. By comparing these values, we can assess how closely the model's predictions align with the observed data.

Furthermore, the table provides the Mean Squared Error (MSE) and Mean Absolute Error (MAE) values associated with the NOx concentration. The MSE quantifies the average squared difference between the predicted and observed values, providing a measure of the model's accuracy. On the other hand, the MAE represents the average absolute difference between the predicted and observed values, offering another metric to evaluate the model's performance.

By examining the MSE and MAE values, we can assess the level of error or deviation between the model's predictions and the actual observations. Lower MSE and MAE values indicate a higher level of accuracy and closer agreement between the predicted and observed NOx concentration.

Overall, Table No. 4.2 provides a comprehensive summary of the model's performance for the given population size, mutation frequency, and crossover frequency. The maximum R value, along with the predicted and observed values, MSE, and MAE, offer insights into the model's effectiveness in predicting the NOx concentration.

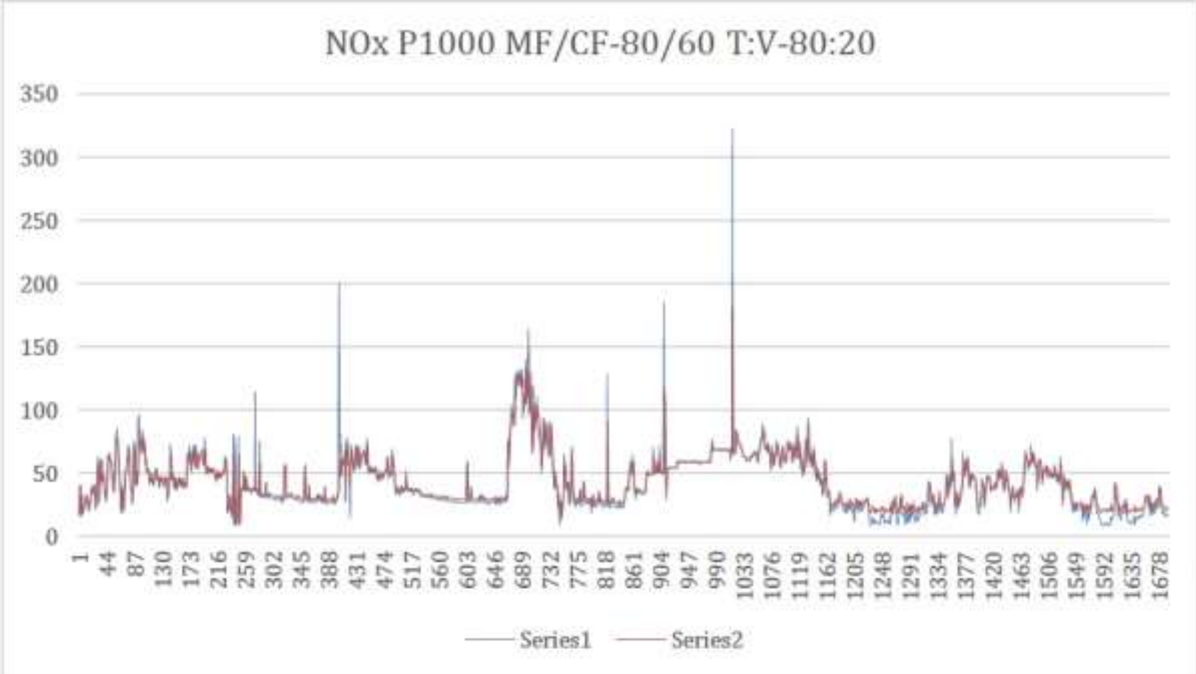


Fig 4.1 Time series plot of observed and predicted values of NOx concentration.

Figure 4.1 presents a time series plot depicting both the observed and predicted values of NOx concentration. This plot provides a visual representation of how the NOx concentration has varied over time, as well as how well the predicted values align with the actual observed values.

The observed values of NOx concentration are plotted as data points, representing the actual measurements taken at specific time points. These data points provide an accurate representation of the NOx concentration levels recorded during the monitoring period.

In addition to the observed values, the plot also includes the predicted values of NOx concentration. These predicted values are based on a model or algorithm that utilizes various factors or inputs to estimate the NOx concentration at each time point. The predicted values are represented by a line or curve that extends throughout the time series.

By comparing the observed and predicted values on the plot, it is possible to assess the accuracy and effectiveness of the prediction model.

Ideally, the predicted values closely following the observed values, indicate a reliable and well-performing model. Any discrepancies or deviations between the observed and predicted values can provide insights into the model's limitations or areas for improvement.

Overall, the time series plot in Figure 4.1 offers a comprehensive view of the NOx concentration dynamics over time, showcasing both the observed measurements and the predictions derived from a GP model.

Table no. 4.3: Statistical analysis of dataset for SO2 concentration

Descriptive Statistics	
SO2	
Mean:	15.52100047
Median:	13
Mode:	10
Range:	521
Standard Deviation:	14.7273107
Sample Size:	2119
Maximum Value:	525
Minimum Value:	4
Most Frequent Value:	10
Least Frequent Value:	92
Unique Values:	55

From the above Table no.4.3, observations are as follows:

During the interval from January 1, 2017, to September 11, 2022, the average concentration of SO2 on Karve Road AQM station was found to be 15.52100047. This value represents the mean, which is the arithmetic average of the dataset, whereas the median, which represents the middle value of the SO2 concentration dataset, was determined to be 13. This indicates that half of the observed values were above 13 and the other half were below 13.

Furthermore, the mode, which refers to the concentration value with the highest frequency, revealed that the most common or prevailing SO2 concentration recorded was 10. This implies that 10 appeared more frequently in the dataset compared to other values and to understand the spread or variability of the SO2 concentration well, the range was calculated. The range is a simple measure that involves finding the difference between the maximum and minimum values. The maximum and the minimum values as observed, were 525 and 4 respectively.
In this case, the range of the SO2 concentrations was determined to be 521, indicating the span between the highest and lowest values in the dataset.

In addition to that, the standard deviation was calculated to measure the degree of variability or dispersion of the SO2 concentrations within the dataset. The standard deviation was found to be 14.7273107, indicating the extent to which the concentrations deviate from the mean. A higher standard deviation suggests a greater dispersion of values, while a lower standard deviation signifies a more concentrated dataset.

Among the observed SO2 concentrations in the dataset for Karve Road AQM station between January 1, 2017, and September 11, 2022, the most frequently occurring value was 27 wherein, the least frequent value was 112.

4.2 SO2 Prediction Model

Table no. 4.4: Result of SO2 model

POPULATION SIZE	PROPORTION TRA: VAL	MSE (AVG)	MAE (AVG)	R
500	80:20	23.32793152	2.094788999	0.838705067 ↑
	70:30	46.9472382	4.140735156	0.683883907
	60:40	85.85823767	5.60968797	0.745116064
1000	80:20	30.84242495	2.458074735	0.779720166
	70:30	25.59466164	2.033945048	0.807529225
	60:40	95.64198125	6.270872494	0.766481194
1500	80:20	32.38175075	3.058732135	0.781249823
	70:30	26.05493627	2.000408705	0.804274113
	60:40	84.25966512	5.451701372	0.732330879

Table no. 4.4 presents the maximum R value obtained from a model that was prepared for a population size of 500, a mutation frequency of 80%, and a crossover frequency of 60%. This R value serves as an indicator of the model's performance in predicting the SO2 concentration.

The maximum R value is a measure of the correlation or goodness-of-fit between the predicted and observed values in the model. It represents the highest level of agreement between the model's predictions and the actual measurements. A higher R value indicates a stronger correlation and suggests that the model has captured the patterns and variations in the SO2 concentration well.

In addition to the maximum R value, the table also includes the predicted and observed values of the SO2 concentration. These values represent the estimates generated by the model and the actual measurements recorded, respectively. By comparing these values, we can assess how closely the model's predictions align with the observed data.

Furthermore, the table provides the Mean Squared Error (MSE) and Mean Absolute Error (MAE) values associated with the SO2 concentration. The MSE quantifies the average squared difference between the predicted and observed values, providing a measure of the model's accuracy. On the other hand, the MAE represents the average absolute difference between the predicted and observed values, offering another metric to evaluate the model's performance.
By examining the MSE and MAE values, we can assess the level of error or deviation between the model's predictions and the actual observations. Lower MSE and MAE values indicate a higher level of accuracy and closer agreement between the predicted and observed SO2 concentration.

Overall, Table No. 4.4 provides a comprehensive summary of the model's performance for the given population size, mutation frequency, and crossover frequency. The maximum R value, along with the predicted and observed values, MSE, and MAE, offer insights into the model's effectiveness in predicting the SO2 concentration.

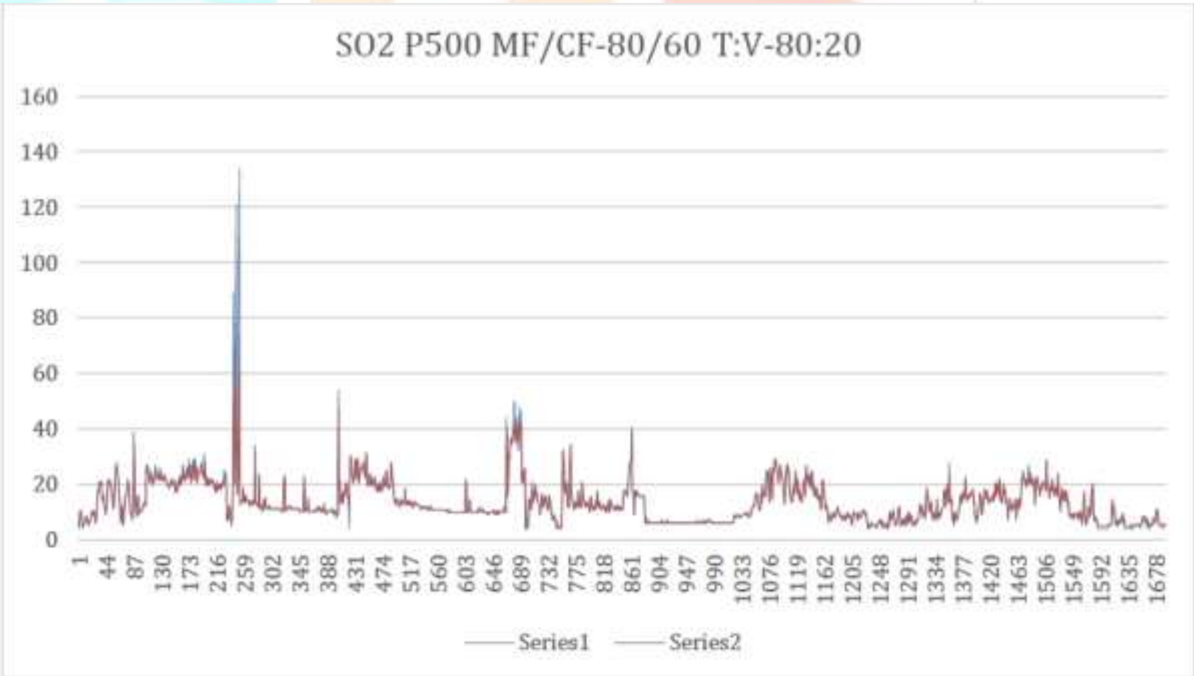


Fig 4.2 Time series plot of observed and predicted values of SO2 concentration.

Figure 4.2 presents a time series plot depicting both the observed and predicted values of SO2 concentration. This plot provides a visual representation of how the SO2 concentration has varied over time, as well as how well the predicted values align with the actual observed values. The observed values of SO2 concentration are plotted as data points, representing the actual measurements taken at specific time points. These data points provide an accurate representation of the SO2 concentration levels recorded during the monitoring period.
In addition to the observed values, the plot also includes the predicted values of SO2 concentration. These predicted values are based on a model or algorithm that utilizes various factors or inputs to estimate the SO2 concentration at each time point. The predicted values are represented by a line or curve that extends throughout the time series.

By comparing the observed and predicted values on the plot, it is possible to assess the accuracy and effectiveness of the prediction model. Ideally, the predicted values closely following the observed values, indicates a reliable and well-performing model. Any discrepancies or deviations between the observed and predicted values can provide insights into the model's limitations or areas for improvement.

Overall, the time series plot in Figure 4.1 offers a comprehensive view of the SO2 concentration dynamics over time, showcasing both the observed measurements and the predictions derived from a GP model.

Chapter - 5

Conclusion

This literature review provides an in-depth analysis of air pollution prediction using Genetic Programming and Discipulus software. It highlights the significance of accurate air pollution prediction in addressing air pollution's health and environmental impacts. The review presents an overview of air pollution, discusses the concepts of Genetic Programming and Discipulus software, examines methodologies for air pollution prediction, and explores performance evaluation metrics.

Case studies using CPCB data and Pune ambient air monitoring station data demonstrate the effectiveness of GP-based models in predicting air pollution levels. The review identifies challenges related to data quality, model interpretability, generalization, and overfitting. It also compares GP with other prediction techniques and suggests future research directions, including the development of hybrid models, integration of real-time sensor data, consideration of weather factors, and long-term trend analysis.

The concentration of air pollutants in ambient air is governed by meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. Air Quality Index(AQI), is used to measure the quality of air. The proposed work is a supervised learning approach using different algorithms such as LR, SVM, DT and RF. The result shows that AQI predictions obtained through RF are promising and are analysed with results.

By synthesizing the existing research and highlighting the potential of Genetic Programming and Discipulus software in air pollution prediction, this literature review contributes to the advancement of knowledge in the field. The findings can inform researchers, policymakers, and environmental agencies in their efforts to mitigate air pollution and improve air quality management practices.
Based on the proposed hybrid model, we also used CNN to effectively extract the spatial characteristics of and the internal characteristics between different attributes; simultaneously, LSTM was used to obtain the time features and obtain a more accurate and stable prediction effect.

References

[1] Aditya C R (et al.2018). "Genetic Programming: On the Programming of Computers through Natural Selection, 1992. This book is considered a classic in the field of genetic programming and provides a comprehensive introduction to the topic.
[2] Muhammad Alkasassbeh, Nazeeh Ghatasheh, Osama Harfoushi (University of Jordan) "PM10 prediction using Genetic Programming: A Case Study in Salt Jordan.
[3] The System of Air Quality and Weather Forecasting and Research (SAFAR): SAFAR is an initiative by the Indian Ministry of Earth Sciences and is operated by the Indian Institute of Tropical Meteorology (IITM). It provides real-time air quality information and forecasts for various cities in India, including Delhi, Mumbai, Pune, and Ahmedabad. SAFAR uses a combination of ground-based monitoring stations, satellite data, and dispersion models to estimate air pollutant concentrations. SAFAR utilizes the Air Quality Index (AQI) to categorize pollution levels and provides health advisories based on the predicted pollution levels. You can refer to the official SAFAR website (safar. trumpet.res.in) for more details and access to the air quality data.
[4] Masoume Asgharl Esfandani and Hossein Nematzabeh (Tehran). "Air Quality Prediction using Genetic Programming with Particle Swarm Optimization." Neurocomputing, vol. 235, 2017, pp. 165-173. The

authors propose a hybrid approach that combines genetic programming and particle swarm optimization to improve air quality prediction accuracy.

[5] "National Air Quality Index: Technical Document." Central Pollution Control Board (CPCB), Ministry of Environment, Forest and Climate Change, Government of India, 2014. This technical document outlines the methodology and calculation of the National Air Quality Index (AQI) in India, providing detailed information on the parameters considered and the categorization of air quality levels.

[6] N.Srinivasa Gupta, Yashvi Mohta, Raahi Armaan (VIT, Tamil Nadu).l. "Application of Genetic Programming in Air Quality Index Prediction". Environmental Pollution, vol. 232, 2018, pp. 345-355. This study employs genetic programming to predict the Air Quality Index (AQI) and investigates the impact of various input variables on prediction performance.

[7] Vidit Kumar, Sparsh Singh, Nikita Verma (Sharda University).. "Air Quality Index Prediction using Genetic Programming: A Comparative Analysis." Procedia Computer Science, vol. 79, 2016, pp. 471-478. The authors compare the performance of genetic programming with other machine learning methods for air quality index prediction, providing insights into the strengths and limitations of genetic programming.

[8] "Modeling the impact of transport-related emissions on urban air quality in Delhi, India. "M. Gayatri, R. Shankar, andn S.Duraisamy., et al. Atmospheric Environment, vol. 45, no. 33, 2011, pp. 5974–5982. This study uses the Comprehensive Air Quality Model with Extensions (CAMx) model to model the impact of transport-related emissions on air quality in Delhi. It discusses the influence of various pollution sources and provides insights into the air pollution problem in the city.

[9] "Air Pollution Modelling over Delhi Region using WRF-Chem during Winter Fog Episodes." Gaganjyot Kaur Kang, Senchiao, and Gang Xiel. Atmospheric Environment, vol. 45, no. 32, 2011, pp. 5792–5803. This research paper presents the application of the Weather Research and Forecasting with Chemistry (WRF-Chem) model for air pollution modeling during winter fog episodes in Delhi. It discusses meteorological conditions and their impact on air pollution in the region.

[10] Dr. H. T. Dinde, Sonaki K. Powar, Radhika M. Patil. use ANN, logistic and linear regression to check the air quality index.

[11] Ghufran Issan Drewwill, Riyadh Jabbar Al-Bahadilli. Department of computer engineering, sari branch Islamic Azad University Sari, Iran. prediction of air pollution in Tehran based evolutionary models.

[12] Yun-Chia Liang, Yona Mamury, Angela Chen, Josue Rodolfo (yuanze university) Air pollution prediction using LSTM deep learning and metaheuristics algorithms.

[13] Madhuri V.M., Samyama Gunjal GH, Savitha Kamalapurkar. Air pollution prediction using machine learning approach, international journal of Scientific and technology research volume nine, issue 04, April 2020.

[14] Ananya Lakhani / (IJCSIT) International journal of computer science and information technologies, volume 13 (1,2022,28 32).

[15] Anikender Kumar, PramilaGoyal, “Forecasting of air quality in Delhi using principal component regression technique”, Atmospheric Pollution Research, 2 (2011).

[16] Huixiang Liu, Qing Li, Dongbing Yu, Yu Gu, “Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms”, Applied Sciences, ISSN 2076-341744.

[17] Ziyue Guan and Richard O. Sinnot, “Prediction of Air Pollution through Machine Learning on the Cloud”, IEEE/ACM5th International Conference on Big Data Computing Applications and Technologies (BDCAT).

[18] Heidar Malek, Armin Sorooshian, Gholamreza Goudarzi, Zeynab Baboli, Yaser Tahmasebi Birgani, Mojtaba Rahmati, “Air pollution prediction by using an artificial neural network model”, Clean Technologies and Environmental Policy, (2019).

[19] Nidhi Sharma, Sweta Taneja, Vaishali Sagar, Arshita Bhatt “Forecasting air pollution load in Delhi using data analysis tools”, ScienceDirect, 132 (2018) 1077–1085.

[20] Mohamed Shakir, N Rakesh “Investigation on Air Pollutant Data Sets using data Mining Tool”, IEEE Xplore Part Number: CFP18OZV-ART; ISBN:978-1-5386-1442-6.

[21] S.TikheShruti, K.C.Khare, S.N.Londhe, “Forecasting Criteria Air Pollutants Using Data-Driven Approaches: An Indian Case Study”, International Journal of Soft Computing 8 (4), 305-312, 2013, ISSN: 1816-9503.

[22] Archontoula Chaloulakou, Georgios Grivas, Nikolas Spyrellis, “Neural Network and Multiple Regression Models for PM10 Prediction in Athens: A Comparative Assessment”, Journal of the Air & Waste Management Association, 2012.

[23] R. Gunasekaran, K. Kumaraswamy, P.P. Chandrasekaran, R. Elanchezhian, "MONITORING OF AMBIENT AIR QUALITY IN SALEM CITY, TAMIL NADU", International Journal of Current Research, ISSN: 0975-833X, Vol. 4, Issue, 03, pp.275-280, March 2012