Real-Time Air Quality Prediction Using Machine Learning Approaches

[1]Geetanjali Hota,[2] Samaleswari Prasad Nayak

[1PGT,] Computer Science, DAV Public School, Pokhariput, Bhubaneswar, India
[2] Silicon Institute of Technology, Bhubaneswar, India

Abstract—
Nowadays air pollution is a major problem in developing countries. It has several bad effects on human body. Humans are very sensitive to humidity, as the skin relies on the air to get rid of moisture. The process of sweating is to keep our body cool and maintain its current temperature. If the air is at 100-percent relative humidity, sweat will not evaporate into the air. As a result, we feel much hotter than the actual temperature when the relative humidity is high. If the relative humidity is low, we can feel much cooler than the actual temperature because our sweat evaporates easily. Through the work, air quality parameters are tackled by using machine learning approaches to predict the Relative Humidity in air. We have proposed a refined model to predict the hourly air Relative Humidity on the basis of meteorological data of previous days by formulating the prediction over 24 h as a multi-task learning (MTL) problem with the help of Linear Regression, Decision Tree Regression, Random Forest Regression and Support Vector Machine. This enables to select a good model with different regularization techniques. The results have been compared by using four algorithms for prediction of air quality.

***Index Terms* — Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression.**

I. INTRODUCTION

Air is a basic requirement for the survival and development of all lives on Earth. It affects health and influences the development of the economy. Today, due to the development of industrialization, the increase in the number of private cars, and the burning of fossil fuels, air quality is decreasing, with increasingly serious air pollution. There are many pollutants in the atmosphere, such as SO2, NO2, CO2, NO, CO. internationally, a large number of scholars have conducted research on air pollution and air quality forecasts, concentrating on the forecasting of contaminants. Air pollution affects the life of a society, and even endangers the survival of mankind. During the Industrial Revolution, there was a dramatic increase in coal use by factories and households, and the smog caused significant morbidity and mortality, particularly when combined with stagnant atmospheric conditions. Humans are very sensitive to humidity, as the skin relies on the air to get rid of moisture. The process of sweating is your body's attempt to keep cool and maintain its current temperature. If the air is at 100-percent relative humidity, sweat will not evaporate into the air. As a result, we feel much hotter than the actual temperature when the relative humidity is high. If the relative humidity is low, we can feel much cooler than the actual temperature because our sweat evaporates easily, cooling us off. For example, if the air temperature e is 75 degrees Fahrenheit (24 degrees Celsius) and the relative humidity is zero percent, the air temperature feels like 69 degrees Fahrenheit (21 C) to our bodies. If the air temperature is 75 degrees Fahrenheit (24 C) and the relative humidity is 100 percent, we feel like it's 80 degrees (27 C) out. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

II. RELATED WORK

Authors, in their work "Directed diffusion: A scalable and robust communication paradigm for sensor networks" evaluated the use of directed diffusion for a simple remote-surveillance sensor network. In such networks the nodes can co-ordinate using directed diffusion for distributed sensing of environmental phenomena. Researchers have proposed a geographical adaptive fidelity (GAF) algorithm in their work "Geography-Informed Energy Conservation for Ad Hoc Routing". The algorithm is supposed to reduce energy consumption in wireless ad hoc networks. Nodes are turned off to conserve energy by identifying equivalent nodes from a routing perspective. The algorithm is independent of the underlying ad hoc routing protocol. From the simulation over unmodified AODV & DSR, it was found that GAF 40-60 % less energy in comparison to unmodified ad hoc routing protocols. Moreover it extends the lifetime of self configuring systems and maintains the application fidelity at the same time. Authors, in their report on, "Geographical and Energy Aware Routing: A Recursive Data Dissemination Protocol for Wireless Sensor Networks," proposed the GEAR protocol(Geographic and Energy Aware Routing).The proposed algorithm routes a packet towards the target region using energy aware neighbor selection & disseminates the packet within the destination using Recursive Geographic Forwarding. The main thrust is to reduce flooding by propagating the query to the right geographical region. The protocol was found to exhibit efficient performance in non-uniform traffic distribution. To predict relative humidity in air the data set collected from UCI Machine learning repository. The data set have 9358 samples of data with 15 attributes. The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensory Device. For our work we have collected the dataset from the below details.

Source: UCI machine learning repository- Air Quality data set
Data Set Characteristics: Multivariate, Time-Series
Number of Instances: 9358
Attribute Characteristics: Real
Number of Attributes: 15
Associated Tasks: Regression
Missing Values?: Yes

Sl.NO		Attribute		Description
0		Date		Date (DD/MM/YYYY)
1		Time		Time (HH.MM.SS)
2		CO(GT)		True hourly averaged concentration CO in mg/m^3 (reference analyzer)
3		PT08.S1(CO)		PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
4		NMHC(GT)		True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)
5		C6H6(GT)		True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
6		PT08.S2(NMHC)		PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
7		NOx(GT)		True hourly averaged NOx concentration in ppb (reference analyzer)
8		PT08.S3(NOx)		PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
9		NO2(GT)		True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
10		PT08.S4(NO2)		PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
11		PT08.S5(O3)		PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
12		T		Temperature in Â°C
13		RH		Relative Humidity (%)
14		AH		AH Absolute Humidity

Table 1- Air Attribute Information

Date	Time	CO	PTO8.S1	NMHC	C6H6	PT08.S2	NO_x	PT08.S3	NO2	PT08.S4	PT08.S5	T	RH	AH
10-03-2022	**18:00**	**2.6**	**1360**	**150**	**11.9**	**1046**	**166**	**1056**	**113**	**1692**	**1268**	**13.6**	**48.9**	**0.7578**
10-03-2022	**19:00**	**2**	**1292**	**112**	**9.4**	**955**	**103**	**1174**	**92**	**1559**	**972**	**13.3**	**47.7**	**0.7255**
10-03-2022	**20:00**	**2.2**	**1402**	**88**	**9.0**	**939**	**131**	**1140**	**114**	**1555**	**1074**	**11.9**	**54.0**	**0.7502**
10-03-2022	**21:00:**	**2.2**	**1376**	**80**	**9.2**	**948**	**172**	**1092**	**122**	**1584**	**1203**	**11.0**	**60.0**	**0.7867**
10-03-2022	**22:00:**	**1.6**	**1272**	**51**	**6.5**	**836**	**131**	**1205**	**116**	**1490**	**1110**	**11.2**	**59.6**	**0.7888**

Table 2- Air Content in the Sample Dataset

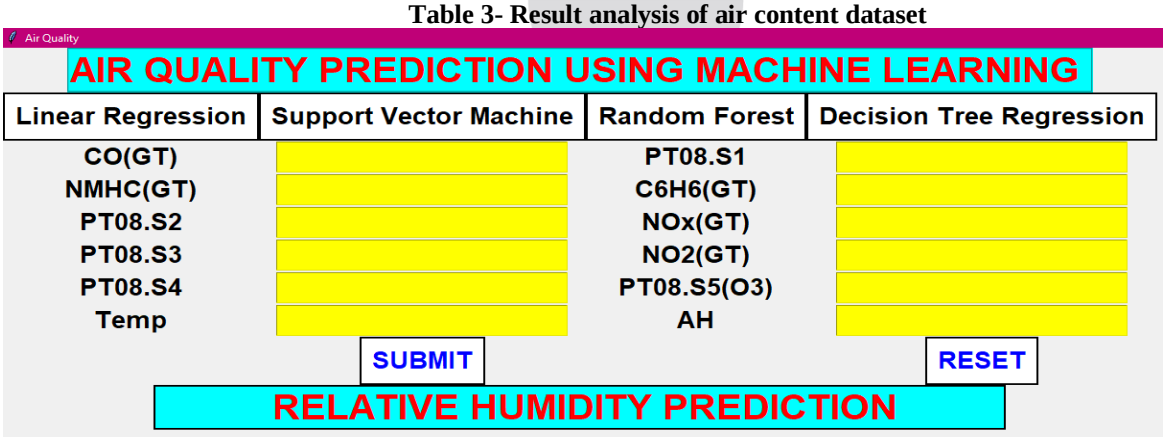
III. Purposed Methodology and Result analysis

As stated earlier, Air Quality Index Prediction involved three main processes, which are based on four main algorithms – Decision Tree Regression, Random Forest Regression, Linear Regression, Support Vector Regression. However, the main objective of this work was to evaluate the new AQI performance in terms of the capacity of distinguishing different air pollution

situations. This index is a air quality index prediction by using relative humidity system based on the national ambient air quality index prediction standards.
The methodology applied in the calculation of the index can be applied for any country with replacing specific standards.
This project is using a dataset downloaded from UCI, Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/air+quality).
Data
Monitoring and ambient air-quality studies show that some of the air pollutants in several
Large cities
Data
Monitoring and ambient air-quality studies show that some of the air pollutants in several
Large cities
Data
Monitoring and ambient air-quality studies show that some of the air pollutants in several
Large cities
Data
Monitoring and ambient air-quality studies show that some of the air pollutants in several
Large cities
Data
Monitoring and ambient air-quality studies show that some of the air pollutants in several
Large cities
Data
Monitoring and ambient air-quality studies show that some of the air pollutants in several Large cities The main objective of this work was to evaluate the new AQI Performances in terms of the capacity of distinguishing different air pollution Situations. This index is a national air quality rating system based on the national ambient Air quality standards. It is based on the new project of Tunisian ambient air quality Standards. The methodology applied in the calculation of this index can be applied for any
country with replacing specific standards. the main objective of this work was to evaluate the new AQI performances in terms of the capacity of distinguishing different air pollution situations. This index is a national air quality rating system based on the national ambient air quality standards. It is based on the new project of Tunisian ambient air quality standards. The methodology applied in the calculation of this index can be applied for any country with replacing specific standards. The main objective of this work was to evaluate the new AQI Performances in terms of the capacity of distinguishing different air pollution Situations. This index is a national air quality rating system based on the national ambient air quality standards. It is based on the new project of Tunisian ambient air quality Standards. The methodology applied in the calculation of this index can be applied for any
country with replacing specific standards. The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensory Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city. Data were recorded from March 2004 to February 2005. Here the dataset is split in ratio 70:30 for training and testing purpose respectively. The dataset was split through the code using train_test_split () which is present in model selection module of sklearn package. In this step root mean square is calculated from Y_test and Y_pred by using mean squared error function. The lesser is root mean squared error better is the model. The following code snippet shows the calculation of root mean squared error for linear regression. Similarly root mean squared error for other 3 models is calculated. For prediction of relative humidity in air 4 machine learning algorithms are applied. And root mean squared error for all models are calculated as follows.

Model	Root mean square error
Linear Regression	7.877001354843912
Decision Tree	1.3439746300685167
Random Forest	0.806269945533892
Support Vector Machine	51.599403446952785

Table 3- Result analysis of air content dataset



Result Prediction

IV. Conclusion

In this project four different machine learning algorithms are applied for prediction of relative humidity. They are Linear Regression, Decision Tree, Random Forest and Support Vector Machine. Root Mean Square Error of Linear Regression is 7.877001354843912,Decision Tree is 1.3439746300685167, Random Forest is 0.806269945533892, Support Vector Machine is 51.599403446952785.From the above it is conclude that Random Forest Regression is the best model for relative humidity prediction in Air.

References

[1] Pandey, Gaurav, Bin Zhang, and Le Jian. " Predicting sub-micron air pollution indicators: a machine learning approach." ; Environmental Science: Processes & amp; Impacts 15.5 (2013): 996-1005.

[2] Dan wei: Predicting air pollution level in a specific city [2014]

[3] Dixian Zhu, Changjie Cai, Tianbao Yang and Xun Zhou: A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. Big data and cognitive computing [2018].

[4] José Juan Carbajal-Hernándezab Luis P.Sánchez-Fernándeza Jesús A.Carrasco-OchoabJosé Fco.Martínez-Trinidadb: Assessment and prediction of air quality using fuzzy logic and autoregressive models: Center of Computer Research – National Polytechnic Institute, Av. Juan de Dios Bátiz S/N, Gustavo A. Madero, Col. Nueva. Industrial Vallejo, 07738 México D.F., Mexico1. (2012) Doi :https://doi.org/10.1016/j.atmosenv.2012.06.004

[5] Sachit Mahajan, Ling-Jyh Chen, Tzu-Chieh Tsai : An Empirical Study of PM2.5 Forecasting Using neural network. IEEE Smart World Congress, At San Francisco, USA [2017]

[6] Athanasiadis, Ioannis N., et al. "Applying machine learning techniques on air quality data for real-time decision support." First international NAISO symposium on information technologies in environmental engineering (ITEE'2003), Gdansk, Poland. 2003.

[7] Ioannis N. Athanasiadis, Kostas D. Karatzas and Pericles A. Mitkas. "Classification techniques for air quality forecasting." Fifth ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, 17th European Conference on Artificial Intelligence, Riva del Garda, Italy, August 2006.

[8] M. Caselli & L. Trizio & G. de Gennaro & P. Ielpo. "A Simple Feedforward Neural Network for the PM10 Forecasting: Comparison with a Radial Basis Function Network and a Multivariate Linear Regression Model." Water Air Soil Pollut (2009) 201:365–377.

[9] S.Bordignon, C. Gaetan and F. Lisi, "Nonlinear models for ground-level ozone forecasting." Statistical Methods and Applications, 11, 227-246, (2002).

[10] K.Chidananda Gowda and Edwin Diday. Symbolic clustering using a new dissimilarity measure. pattern recognition, 24(6):567–578, 1991.

[11] K.Chidananda Gowda and Edwin Diday. Symbolic clustering using a new similarity measure. IEEE Transactions on Systems, Man, and Cybernetics, 22(2):368–378, 1992.

[12] Edwin Diday. Symbolic data analysis: a mathematical framework and tool for data mining. In Advances in Data Science and Classification, pages 409–416. Springer, 1998.