Smart City Air Quality Prediction UsingMachine Learning

Mrs.Sugur Swathi
Conmputer Science and Engineering
(JNTUH)
Sphoorthy Engineering College
(JNTUH)
Hyderabad, India
swathi.tekumal@gmail.com

CH. Prathima
Computer Science and Engineering
(JNTUH)
Sphoorthy Engineering College
(JNTUH)
Hyderabad, India
prathimareddychinthala@gmail.com

M. Manisha Reddy
Computer Science and Engineering
(JNTUH)
Sphoorthy Engineering College
(JNTUH)
Hyderabad, India
mavurammanisha606@gmail.com

N. Sai Roshan
Computer Science and Engineering
(JNTUH)
Sphoorthy Engineering College
(JNTUH)
Hyderabad, India
buntynani590@gmail.com

***Abstract*— Air pollution is one of the major hazards among the environmental pollution. Due to human activities, industrialization and urbanization air is getting polluted. The major air pollutants are CO, NO, C6H6 etc. The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. Earlier techniques such as Probability, Statistics etc were used to predict the quality of air, but those methods are very complex to predict. The Machine Learning (ML) is the better approach to predict the air quality. With the need to predict air relative humidity by considering various parameters such as CO, Tin oxide, nonmetallic hydrocarbons, Benzene, Titanium, NO, Tungsten, Indium oxide, Temperature etc approach uses Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest Method (RF) to predict the Relative humidity of air and uses Root Mean Square Error to predict the accuracy. The tools used in Machine Learning Algorithms, Python, Feature Engineering, Pandas, Numpy, Seaborn, Flask, HTML, CSS etc.**

This model can be used by several government organizations and can help them in making the right decisions related to approval or rejection of any industrial project to control pollution level of our country by using Machine Learning algorithms.

Keywords – Air Pollution, meteorological parameters, Machine Learning, Feature Engineering, Feature Scaling, Machine Learning algorithms.

Introduction

While the use of machine learning algorithms is an effective tool for predicting air quality, the quality and quantity of data used to train the models are critical factors in ensuring accurate predictions. Therefore, it is essential to gather data from various sources and ensure that it is of high quality. Combining the use of technology with policy measures is crucial in mitigating the negative impact of air pollution on human health and the environment.

The Environment is nothing but everything that encircles us. The environment is getting polluted due to human activities and natural disaster, very severe among them is air pollution. The concentration of air pollutants in ambient air is governed by the meteorological parameters such as atmospheric wind speed, wind direction, relative humidity, and temperature. Urbanization is one of the main reasons for air pollution because, increase in the transportation facilities emits more pollutants into the atmosphere and another main reason for air pollution is Industrialization. The major pollutants are Nitrogen Oxide (NO), Carbon Monoxide (CO), Particulate matter (PM), SO2 etc. PM2.5 is the most harmful and global killer among the air pollutants. The weight of PM2.5 is a fine and tiny particle in the air where its diameter is 2.5 micrometers as visualized in Fig.1. PM2.5 is a very tiny particle that able to diffuse into respiratory systems and gives a bad impact on the human lung.

Air Quality Index(AQI), is used to measure the quality of air. Earlier classical methods such as probability, statistics were used to predict the quality of air, but those methods are very complex to predict the quality of air. Due to advancement of technology, now it is very easy to fetch the data about the pollutants of air using sensors. Assessment of raw data to detect the pollutants needs vigorous analysis.

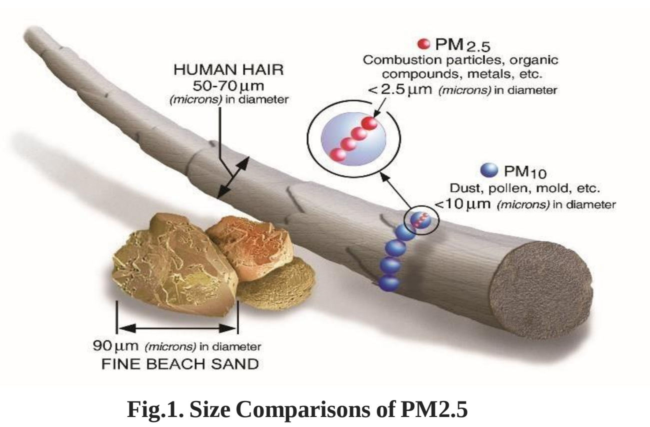

HUMAN HAIR
50-70µm
(microns) in diameter
PM2.5
Combustion particles, organic compounds, metals, etc.
< 2.5µm (microns) in diameter
PM10
Dust, pollen, mold, etc.
<10µm (microns) in diameter
90µm (microns) in diameter
FINE BEACH SAND

Fig.1. Size Comparisons of PM2.5

Machine Learning:

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn gradually improving its accuracy.

Over the last couple of decades, the technological advances in storage and processing power have enabled some innovative products based on machine learning, such as Netflix’s recommendation engine and self-driving cars. Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects.

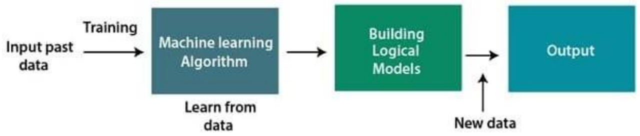


Fig.2. Machine Learning

I. LITERATURE REVIEW

Since last few years, most major cities around the world have experienced pollution levels that exceed all international norms, resulting in a slew of life-threatening issues. The life threatening effects of PM2.5 have prompted research to establish a reliable model for predicting PM2.5 levels in contaminated air. Machine learning techniques have been applied to predict air pollution levels, including in context of smart cities. In this literature review, we will discuss recent research on smart city air pollution prediction using machine learning.

A. Approaches of Air Pollution Prediction in Smart Cities

A study has been proposed Air Quality Prediction using Machine Learning Model in smart cities of Delhi [5]. This method is to outcome the challenge of predicting the Air Quality Index (AQI), to minimize the urban air pollution before it gets adverse in Delhi. The Machine Learning methods that were applied are Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms. The result demonstrates that by using these two Machine Learning algorithms, AQI was performed well and predicted successfully, with an accuracy of 91.62% for ANN while 97.3% for SVM. Medium Gaussian SVM gave the maximum accuracy from the six functions of SVM were used to predict the accuracy.

B. Air Quality Prediction in Smart Cities Using MachineLearning Technologies Based on Sensor Data: A Review: Iskandaryan et. Al, 2022

This paper provides a comprehensive review of various machine learning techniques and sensor technologies used to predict air quality in smart cities. The authors discuss the challenges and opportunities of air quality prediction in urban environments and the use of machine learning technologies to address these challenges. The authors highlight the potential of machine learning algorithms in improving the *accuracy* of air quality prediction and the need for further research in this field. The implementation of such systems can help reduce health issues caused by poor air quality and improve the overall quality of life in urban areas.

C. Air pollution prediction with machine learning: a case study of Indian cities, 2022

It mainly focuses on predicting air pollution levels in Indian cities using machine learning algorithms. The study uses data from various sources to build models for predicting air pollution levels. Machine learning algorithms can analyze large amounts of data and identify patterns that can help in predicting air pollution levels with high accuracy.

D. Chatbot for construction firms using scalable blockchain network by Kareem Adel, Ahmed keema, & Mohamed Marzouk b, 2022

The paper describes the use of different machine learning algorithms, such as decision trees, artificial neural networks, and support vector machines, to predict the AQI and various air pollutants, such as sulfur dioxide, nitrogen dioxide, and particulate matter. The use of machine learning algorithms can help issue timely warnings to the public about potential spikes in pollution levels, enabling people to take precautionary measures to protect their health. The use of machine learning algorithms for air quality forecasting and pollution prediction has the potential to improve public health and the environment by providing accurate and timely information about air quality and pollution levels.

II. EXISTING SYSTEM

Big data and data mining techniques were utilized to retrieve and examine information on air quality in a prior attempt to anticipate air pollution. Predicting pollutant concentration levels was the project’s main goal, in particular for ozone (O3), Sulphur dioxide (SO2), and nitrogen oxide (NO2). In order to examine historical data and estimate future air pollutant concentrations, the current system of air quality prediction using machine learning uses a variety of approaches and models. This helps to safeguard public health and improve environmental conditions Several methods and models are used in the current system of air quality prediction employing machine learning to analyse air pollutant concentrations and forecast future conditions. Artificial neural networks (ANNs), which are created to mock the activity of the human brain and can learn from the past data to generate predictions about the future, are one common approach. To forecast future pollutant concentrations, ANNs can be trained using historical air quality data and meteorological data. Support vector machines (SVMs), supervised learning models that examine data and identify patterns, are a different strategy that can be used to forecast air pollution concentrations based on past data. SVMs can learn from data and produce precise predictions. Moreover, there are ensemble models, which mix various machine learning.

There are several existing methods for air pollution prediction using machine learning, and the choice of method depends on the specific requirements and characteristics of the data and the use case.

Here are some common methods:

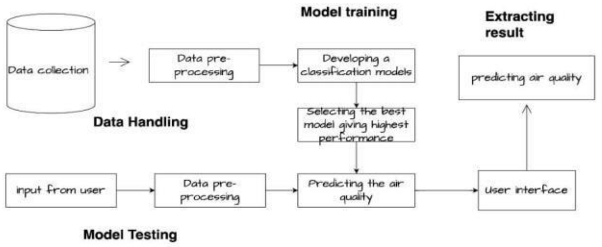
Regression-based models: These models use regression algorithms such as linear regression or polynomial regression to predict air pollution levels based on historical data.

Decision tree-based models: These models use decision trees to represent the complex relationships between the input variables and the output variable, which can be air pollution levels.

Deep learning-based models: These models use deep learning algorithms such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to predict air pollution levels, which can handle complex spatio-temporal relationships and time-series data.

Overall, the choice of method depends on the specific requirements and characteristics of the data and the use case. It's essential to evaluate the performance of the different methods and choose the most appropriate one for a given use case.

III. ARCHITECTURE



IV. PROPSED SYSTEM

Our system will use the Linear regression , Support vector machine and Random forest algorithm for prediction of the pollution of next day. Air quality prediction using machine learning is a highly useful and important application of data science that can help individuals, organizations, and governments to take proactive measures to mitigate air pollution and its associated health risks. Here is a proposed system architecture for air quality prediction using machine learning:

Data Collection - Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems. In this step, we need to identify the different data sources, as data can be collected from various sources. Fetching some components like SO2, NO2, PM, Ozone, Air Quality etc.

Data Analysis: Now the cleaned and prepared data is passed on to the analysis step. This step involves: Selection of analytical techniques, Building models and Review the result.

Data Pre-processing -After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.

V. SOFTWARE AND HARDWARE

1.HARDWARE

Hardware	Types
RAM	4GB
Processor	Intel i3 and above
HardDisk	500GB minimum

Table 5.1. Hardware Description

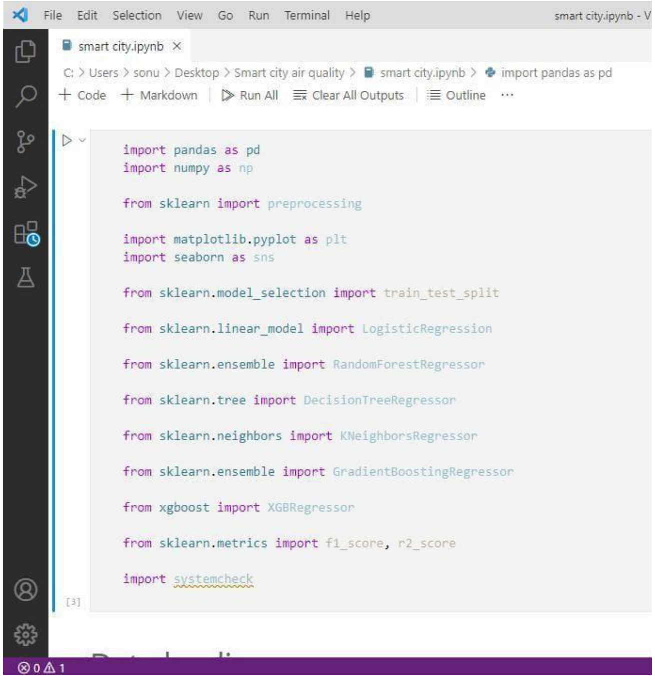
2.SOFTWARE

Software	Types
Platform	Python 3.7
IDE	Jupyter

Table 5.2. Software Description

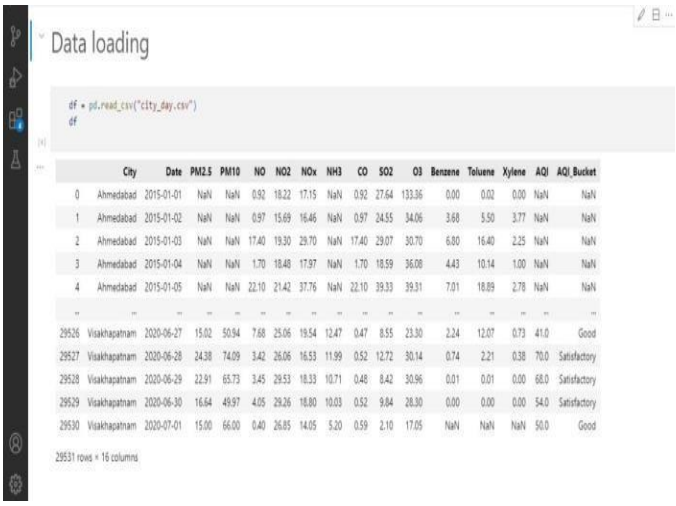
VI. CODING AND IMPLEMENTATION

1.IMPORTING LIBRARIES


File Edit Selection View Go Run Terminal Help smart city.ipynb - V
smart city.ipynb
C: > Users > sonu > Desktop > Smart city air quality > smart city.ipynb > import pandas as pd
+ Code + Markdown Run All Clear All Outputs Outline

```
import pandas as pd
import numpy as np

from sklearn import preprocessing

import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LogisticRegression

from sklearn.ensemble import RandomForestRegressor

from sklearn.tree import DecisionTreeRegressor

from sklearn.neighbors import KNeighborsRegressor

from sklearn.ensemble import GradientBoostingRegressor

from xgboost import XGBRegressor

from sklearn.metrics import f1_score, r2_score

import systemcheck
```

2.DATA LOADING

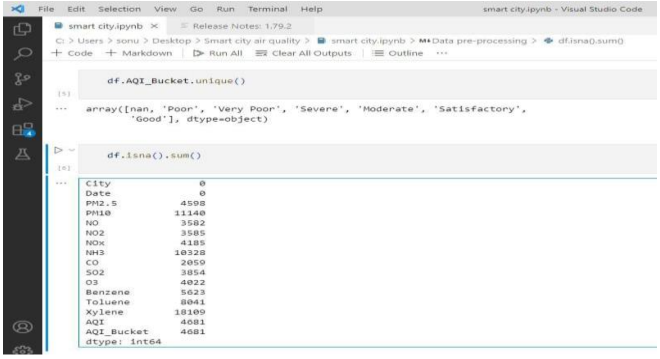


```
df = pd.read_csv("city_day.csv")
df
```

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
0	Ahmedabad	2015-01-01	NaN	NaN	0.92	18.22	17.15	NaN	0.92	27.64	133.36	0.00	0.02	0.00	NaN	NaN
1	Ahmedabad	2015-01-02	NaN	NaN	0.97	15.69	16.46	NaN	0.97	24.55	34.06	3.68	5.50	3.77	NaN	NaN
2	Ahmedabad	2015-01-03	NaN	NaN	17.40	19.30	29.70	NaN	17.40	29.07	30.70	6.80	16.40	2.25	NaN	NaN
3	Ahmedabad	2015-01-04	NaN	NaN	1.70	18.48	17.97	NaN	1.70	18.59	36.08	4.43	10.14	1.00	NaN	NaN
4	Ahmedabad	2015-01-05	NaN	NaN	22.10	21.42	37.76	NaN	22.10	39.33	39.31	7.01	18.89	2.78	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
29526	Visakhapatnam	2020-06-27	15.02	50.94	7.68	25.06	19.54	12.47	0.47	8.55	23.30	2.24	12.07	0.73	41.0	Good
29527	Visakhapatnam	2020-06-28	24.38	74.09	3.42	26.06	16.53	11.99	0.52	12.72	30.14	0.74	2.21	0.38	70.0	Satisfactory
29528	Visakhapatnam	2020-06-29	22.91	65.73	3.45	29.53	18.33	10.71	0.48	8.42	30.96	0.01	0.01	0.00	68.0	Satisfactory
29529	Visakhapatnam	2020-06-30	16.64	49.97	4.05	29.26	18.80	10.03	0.52	9.84	28.30	0.00	0.00	0.00	54.0	Satisfactory
29530	Visakhapatnam	2020-07-01	15.00	66.00	0.40	26.85	14.05	5.20	0.59	2.10	17.05	NaN	NaN	NaN	50.0	Good

29531 rows × 16 columns

3.DATA PREPROCESSING


File Edit Selection View Go Run Terminal Help smart city.ipynb - Visual Studio Code
smart city.ipynb Release Notes: 1.79.2
C: > Users > sonu > Desktop > Smart city air quality > smart city.ipynb > M↓Data pre-processing > df.isna().sum()
+ Code + Markdown Run All Clear All Outputs Outline

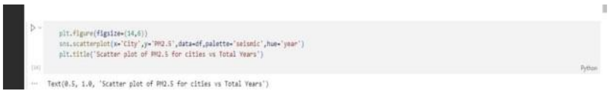
```
df.AQI_Bucket.unique()
```

```
array([nan, 'Poor', 'Very Poor', 'Severe', 'Moderate', 'Satisfactory',
       'Good'], dtype=object)
```

```
df.isna().sum()
```

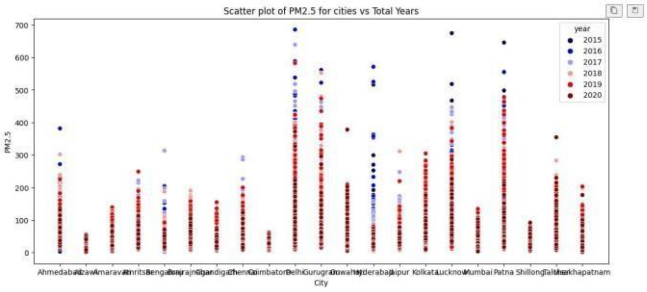
```
City              0
Date              0
PM2.5          4598
PM10          11140
NO             3582
NO2            3585
NOx            4185
NH3           10328
CO             2059
SO2            3854
O3             4022
Benzene        5623
Toluene        8041
Xylene        18109
AQI            4681
AQI_Bucket     4681
dtype: int64
```

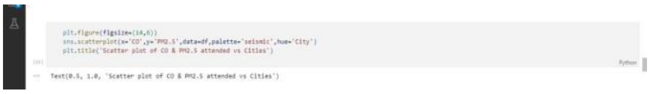
4. DATA VISUALIZATION



```
plt.figure(figsize=(14,6))
sns.scatterplot(x='City',y='PM2.5',data=df,palette='seismic',hue='year')
plt.title('Scatter plot of PM2.5 for cities vs Total Years')
```

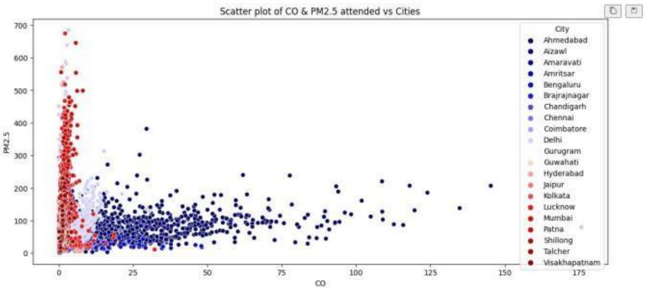
```
Text(0.5, 1.0, 'Scatter plot of PM2.5 for cities vs Total Years')
```



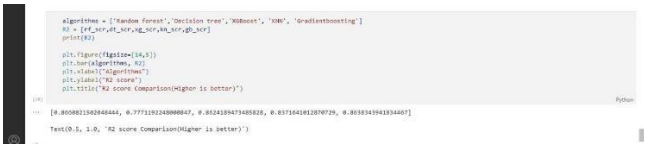


```
plt.figure(figsize=(14,6))
sns.scatterplot(x='CO',y='PM2.5',data=df,palette='seismic',hue='City')
plt.title('Scatter plot of CO & PM2.5 attended vs Cities')
```

```
Text(0.5, 1.0, 'Scatter plot of CO & PM2.5 attended vs Cities')
```

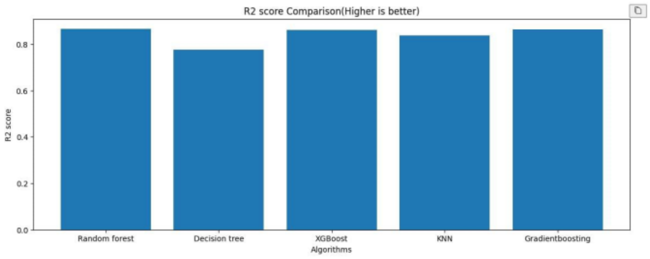


5. COMPARISION



```
algorithms = ['Random forest','Decision tree','XGBoost', 'KNN', 'Gradientboosting']
R2 = [rf_scr,dt_scr,xg_scr,kn_scr,gb_scr]
print(R2)

plt.figure(figsize=[14,5])
plt.bar(algorithms, R2)
plt.xlabel("Algorithms")
plt.ylabel("R2 score")
plt.title("R2 score Comparison(Higher is better)")
```

```
[0.8600821502048444, 0.7771192248000847, 0.8624189473485828, 0.8371641012870729, 0.8638343941834467]

Text(0.5, 1.0, 'R2 score Comparison(Higher is better)')
```



RANDOM FOREST REGRESSION HAS MORE ACCURACY

VII. CONCLUSION

The experience of working on the project was fantastic. Planning, designing, and implementation are crucial, as we had learned from our textbooks, which became clearer to us as a result. It helped us collaborate while allowing us to express our creativity.

VIII. REFERENCES:

1] Verma, Ishan, Rahul Ahuja, Hardik Meisheri, and LipikaDey. ”Air pollutant severity rediction using Bi-directional LSTM Network.” In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 651-654. IEEE, 2018.

[2] Ayele, TemeseganWalelign, and RutvikMehta.”Air pollution monitoring and prediction using IoT.” In 2018 Second International Conference on Inventive Communication 6 Fig. 12. RH w.r.t Temperature Fig. 13. RH w.r.t CO and Computational Technologies (ICICCT), pp. 1741-1745. IEEE,2018.

[3] S. Ameer, M. Ali Shah, A. Khan et al., “Comparative analysis of machine learning techniques for predicting air quality in smart cities,” IEEE Access, vol. 7, pp. 128325–128338, 2019

[4] Kumar, Dinesh. ”Evolving Differential evolution method with random forest for prediction of Air Pollution.” Procedia computer science 132 (2018): 824-833.

[5] Jiang, Ningbo, and Matthew L. Riley. ”Exploring the utility of the random forest method for forecasting ozone pollution in SYDNEY.” Journal of Environment Protection and Sustainable Development 1.5 (2015): 245-254.

[6] A. GnanaSoundari, J. GnanaJeslin and A.C. Akshaya, "Indian Air Quality Prediction And Analysis Using Machine Learning", International Journal of Applied Engineering Research ISSN 0973-4562, vol. 14, no. 11, 2019.

[7] Bhalgat P, Bhoite S, Pitare S (2019) Air Quality Prediction using Machine Learning Algorithms. Int J Comput Appl Technol Res 8(9):367–370. https://doi.org/10.7753/IJCATR0809.1006

[8]Castelli M, Clemente FM, Popoviˇc A, Silva S, Vanneschi L (2020) A machine learning approach to predict air quality in California. Complexity 2020(8049504):1–23. https://doi.org/10.1155/2020/8049504