

# **PM2.5 PARTICULATE MATTER CONCENTRATION IN THE AIR**

**In Uganda**

**A PROJECT REPORT**

**Submitted by,**

**MITHUN K K**

## ABSTRACT

The term fine particles, or particulate matter 2.5 (PM<sub>2.5</sub>), refers to tiny particles or droplets in the air that are two and one half microns or less in width. Like inches, meters and miles, a micron is a unit of measurement for distance. There are about 25,000 microns in an inch. The widths of the larger particles in the PM<sub>2.5</sub> size range would be about thirty times smaller than that of a human hair. The smaller particles are so small that several thousand of them could fit on the period at the end of this sentence. In 2019, the capital city of Uganda, Kampala, was recorded as having a PM<sub>2.5</sub> reading of 29.1 µg/m<sup>3</sup> as its yearly average, a reading that placed it into the higher end of the 'moderate' pollution bracket, which requires a PM<sub>2.5</sub> reading of anywhere between 12.1 to 35.4 µg/m<sup>3</sup> to be classified as such. PM<sub>2.5</sub> refers to particulate matter that is 2.5 micrometers or less in diameter, going down to sizes as small as 0.001 microns and beyond. This incredibly small size, coupled with the constituents of what make up fine particulate matter, make it incredibly dangerous to human health when respired. This has caused PM<sub>2.5</sub> to be used as one of the major components in the calculation of the overall quality of air, along with other prominent pollutants such as PM<sub>10</sub>, Ozone and nitrogen dioxide.

This placed Kampala in 465<sup>th</sup> place out of all cities ranked worldwide, an extremely poor placing that shows that it is sitting in the upper echelons of most polluted cities around the world. Uganda itself as a country also came in with a PM<sub>2.5</sub> reading of 40.80 µg/m<sup>3</sup> in 2018, and then 29.10 µg/m<sup>3</sup> in 2019, a reading that placed it in 22<sup>nd</sup> place out of all countries ranked worldwide, coming in just behind other countries such as Ghana and Myanmar. This is highly indicative that Uganda could do much to improve the quality of its air, and to implement measures to drastically reduce pollution for the wellbeing of its citizens, which also requires people on an individual level to take responsibility for their actions. Likewise, the statement of this project is to determine the various levels of PM<sub>2.5</sub> in the air using AirQo's air quality sensing network which has more than 120 low-cost devices deployed across Uganda, and regularly sense and report on air quality using the PM<sub>2.5</sub> measure.

# INTRODUCTION







Using Kampala as the main example to go by, it can be seen that there are certain periods of the year where the air quality is particularly deteriorated and has higher readings of PM<sub>2.5</sub> than the rest of the year. Whilst there are many more locations throughout Uganda that also have pollution issues, some of the most up to date and concise data about pollution levels are currently located in Kampala, and hence it will be used to determine air quality fluctuations for the country. Over the course of 2019, there were some rather sporadic readings taken in Kampala, which differs from other cities around the world which often show a clear cut period of time in which the pollution level is at its worst, and when the air becomes cleaner (with winter colder periods often holding the title of being more polluted than their warmer counterparts). Some of the most polluted months on record were January to March, and then June through to August, and finally the month of December. All of these months came in above 30  $\mu\text{g}/\text{m}^3$ , with the worst offenders being February, July and August, which had PM<sub>2.5</sub> readings of 36.9  $\mu\text{g}/\text{m}^3$ , 39.9  $\mu\text{g}/\text{m}^3$  and 37.4  $\mu\text{g}/\text{m}^3$  respectively. These were all within the 'unhealthy for sensitive groups' bracket, with July being the most polluted month of the year with its reading of 39.9  $\mu\text{g}/\text{m}^3$ . These are the months of the year when the capital would have its air permeated by large amounts of smoke, haze and particulate matter, and whilst this is not truly indicative of the whole country, gives some insight into the sporadic nature of pollution spikes that occur in Uganda.

With a plethora of different pollutants available in the air, as well as subsequent high levels of PM<sub>2.5</sub> readings being recorded, there would thus be a large amount of health conditions and ailments that can occur to those who are exposed to excessive amounts of pollution in their day to day lives. These include shorter term ones that usually cease to be a problem when exposure is halted, and include problems such as dry coughs, chest pains, headaches, nausea and vomiting, as well as skin rash breakouts and irritation to the mucous membranes. In terms of more chronic health issues, ones such as higher rates of cancer become apparent amongst the general population, with many cases presenting themselves in people, usually those who work or live in environments where there is high exposure to carcinogenic materials. Other health issues would be ones such as ischemic heart disease, whereby the heart tissue fails to receive enough oxygen and sustains damage as a result. This can lead to higher rates of heart attacks, as well as other cardiac conditions such as angina and arrhythmias. Due to the tiny size of fine particulate matter, it has the insidious ability to enter the blood stream via the lungs, finding its way in via the tiny air sacs, or alveoli. Once in the blood stream, damage to the blood vessels can occur, as well as the material making its way to the far reaches of the body, with the hepatic and renal systems also being at risk for damage (liver and kidneys), as well as reproductive health being affected.

Whilst there are no groups that are fully immune to the negative side effects of air pollution present in Uganda, there are certain demographics that would be even more vulnerable or at risk for a number of different reasons, usually pertaining to age, health or individual disposition. These groups include people such as young children, the

elderly, those with preexisting health conditions, as well as those who have compromised immune systems, or hypersensitivity towards certain chemicals, resulting in serious allergic reactions. Pregnant mothers are also particularly vulnerable as well, with overexposure having the possibility to cause instances of miscarriage, as well as babies being born prematurely or with a low birth weight.

The image below provides the AQI (Air Quality Index) categories and ranges corresponding to different values of PM<sub>2.5</sub> and their resulting health impact for the reference.

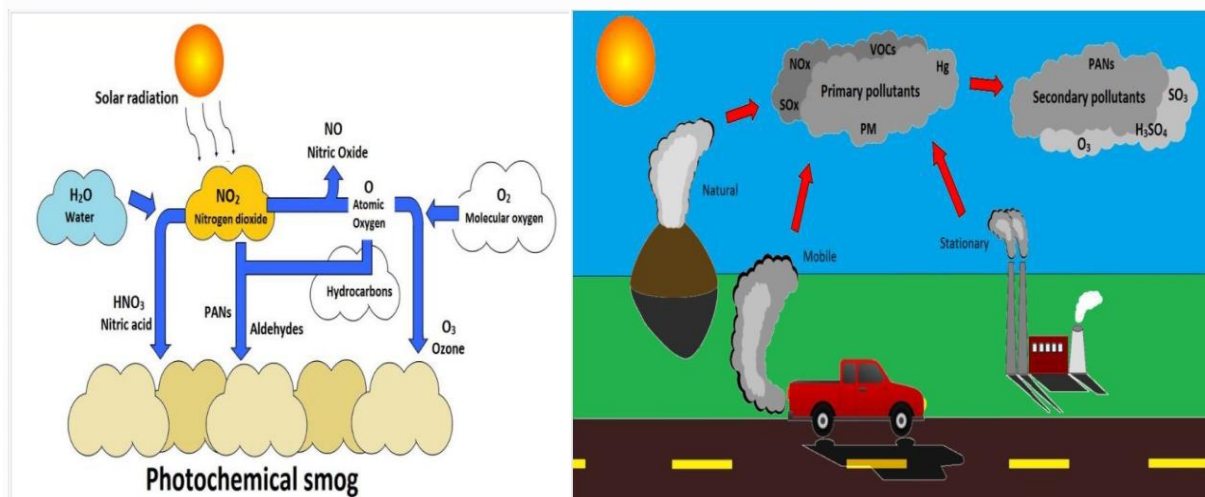
PM 2.5 concentrations (µg/m <sup>3</sup> )	Air Quality Index (EPA)		Possible health effects
0 to 12.0		<b>Good</b> 0 to 50	None.
12.1 to 35.4		<b>Moderate</b> 51 to 100	Unusually sensitive individuals may experience respiratory symptoms.
35.5 to 55.4		<b>Unhealthy for Sensitive Groups</b> 101 to 150	Increasing likelihood of respiratory symptoms in sensitive individuals, aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly.
55.5 to 150.4		<b>Unhealthy</b> 151 to 200	Increased aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; increased respiratory effects in general population.
150.5 to 250.4		<b>Very Unhealthy</b> 201 to 300	Significant aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; significant increase in respiratory effects in general population.
250.5 to 500.4		<b>Hazardous</b> 301 to 500	Serious aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; serious risk of respiratory effects in general population.

Our dataset is obtained from AirQo's air quality sensing network has more than 120 low-cost devices deployed across Uganda, which regularly sense and report on air quality using the PM2.5 measure. AirQo is a research project of the Makerere University College of Computing and Information Sciences committed to using technology to solve social problems across Africa with a current focus on Air Quality. With support from Google we have developed a network of low cost devices across Uganda and are using ML/AI to build forecasting, spatial and now calibration data. We explore using satellite radar data from Sentinel 5P to predict air quality in regions in Kampala. The use of satellite data could expand air quality predictions to areas without air quality sensor devices.

The sentinel 5p data provided was extracted from google earth engine. The following pollutants were extracted from their respective images;

- **SulphurDioxide**-[https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS\\_S5P\\_NRTI\\_L3\\_SO2?hl=en](https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_NRTI_L3_SO2?hl=en)
- **CarbonMonoxide**-[https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS\\_S5P\\_NRTI\\_L3\\_CO?hl=en](https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_NRTI_L3_CO?hl=en)
- **NitrogenDioxide**-[https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS\\_S5P\\_NRTI\\_L3\\_NO2?hl=en](https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_NRTI_L3_NO2?hl=en)
- **Formaldehyde**-[https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS\\_S5P\\_NRTI\\_L3\\_HCHO?hl=en](https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_NRTI_L3_HCHO?hl=en)
- **UvAerosolIndex**-[https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS\\_S5P\\_NRTI\\_L3\\_AER\\_AI?hl=en](https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_NRTI_L3_AER_AI?hl=en)
- **Ozone**-[https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS\\_S5P\\_NRTI\\_L3\\_O3?hl=en](https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_NRTI_L3_O3?hl=en)
- **Cloud**--[https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS\\_S5P\\_OFFL\\_L3\\_CLOUD?hl=en](https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_CLOUD?hl=en)

More info on Sentinel 5p data can be found here: <https://developers.google.com/earth-engine/datasets/catalog/sentinel-5p>



# OBJECTIVE OF THE PROJECT

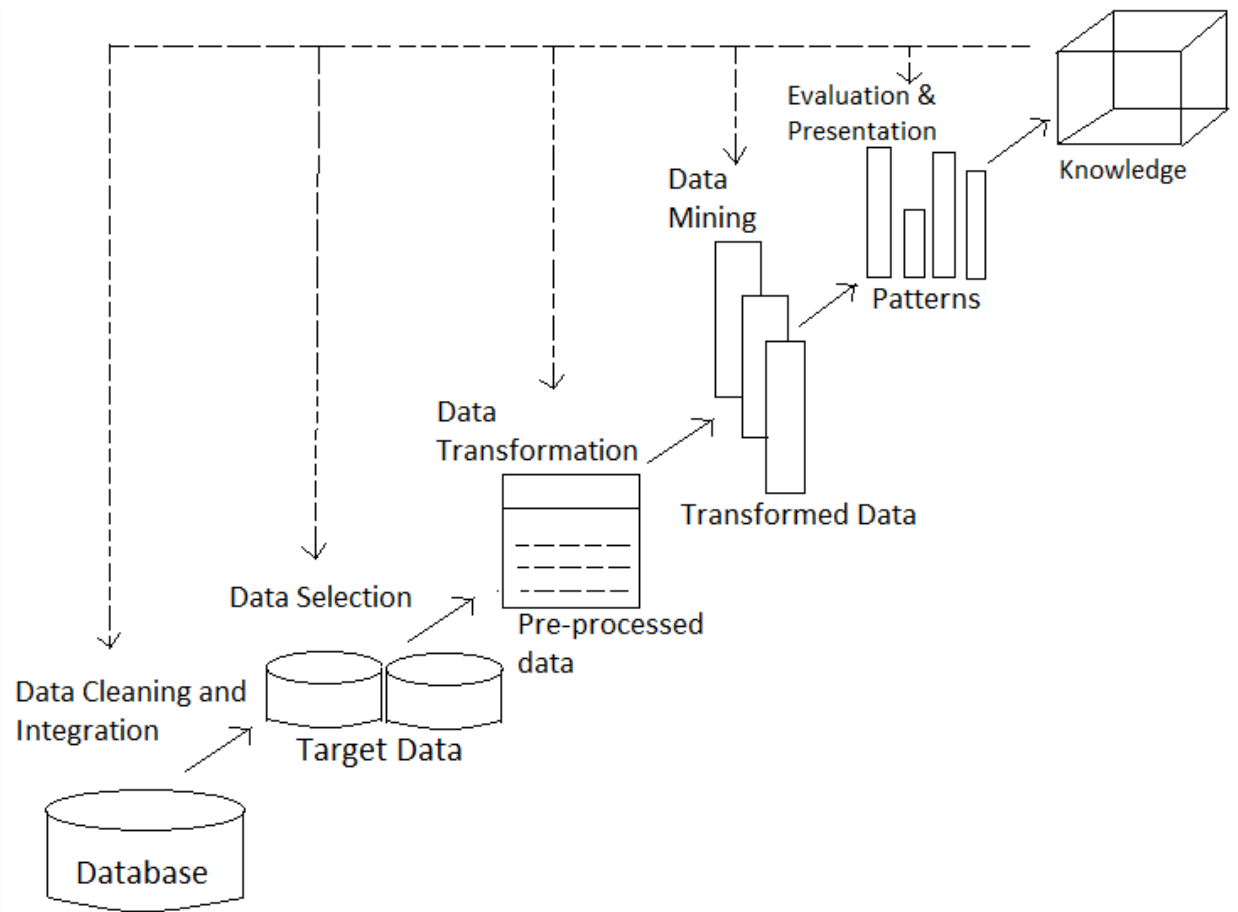
The objective of this project is to predict PM2.5 particulate matter concentration in the air for some locations in Uganda.

## GENERAL BACKGROUND

### Data Mining

Data mining is a technique for transforming unstructured converting data into useful knowledge. Organizations may gain a better understanding of their customers by using software to look for trends in large amounts of data. This enables them to create more effective advertising strategies, increase sales, and save costs. Efficient data gathering, warehousing, and computer interpretation are all required for data mining. Data mining is the process of autonomously examining enormous amounts of data for patterns and trends which go further than basic comparison. Data mining estimates the likelihood of upcoming occurrences by utilizing advanced mathematical algorithms for data segments. Data mining is also known as data knowledge discovery (KDD). Data Mining is related to Data Science, which is done by a professional in a given circumstance, on a given data collection, and with a certain goal in mind. Text mining, online mining, audio and video mining, graphical data mining, and social media mining are only some of the services available. It's done using either basic or extremely specialized software. By outsourcing data mining, all of the work may be completed more quickly and at a lower cost. Specialized businesses can also take use of new technology to acquire data that would otherwise be hard to locate manually. Although there is a wealth of material available on multiple platforms, there is a scarcity of expertise. The most difficult task is to evaluate the data in order to extract significant information that may be utilized to solve an issue or advance the firm. There are a plethora of strong tools and approaches for mining data and extracting more information from it.

Data mining and KDD (Knowledge discovery in Database) are often connected to each other. The process of KDD is used to form high-level knowledge from low level data. That is using KDD unknown, unwanted and implicit data could be removed from the large volume of data. The image below shows the process of KDD and the steps involved in it.



After collecting the raw data the following steps in KDD are performed:

1. **Data cleaning:** it is the process of removal of unwanted and noisy data from the huge volume of data, cleansing is the other name of this process.
2. **Data integration:** it is the process of combining certain data which are showing the same property.
3. **Data selection:** In this phase the data needed for the particular work is identified and fetched out from the huge volume of data.
4. **Data transformation:** The process of translating data from one format to another format which is useful for data analysis purpose, data consolidation is another name of data transformation.
5. **Data mining:** In the data mining step the whole dataset is traversed deeply to collect useful information for the work.
6. **Pattern evaluation:** It is the process of finding out the pattern and relations associated with each attribute and its instances.
7. **Knowledge representation:** The mining data is visually represented in this step and also the interpretations of results are being done.

## **Machine Learning**

Machine learning is a form of artificial intelligence (AI) that enables systems to understand and grow under their self without the need for programming. The construction of computer programs that can collect data and adjust on their own is what machine learning is all about. The training process begins with observations or data, such as examples, firsthand experiences, or instructions, so that we may look for patterns in the data and make better decisions in the future based on the examples we provide. The overarching objective is for computers to learn on their own, without the need for human intervention, and to adapt their behavior as a result.

## **SCOPE OF THE PROJECT**

- This will help researchers to develop low-cost air quality monitoring sensors that work in extreme conditions to tackle rising air pollution.
- Using artificial intelligence technology and machine learning, this data is then processed before it is uploaded onto a cloud-based service accessible to consumers and the public via a smart phone application.
- The equipment, which required expensive maintenance, broke down frequently because they were not designed specifically for the local environment, city officials say. So this will help to maintain low cost sensors and with less cost.
- The real-time data would be used to influence decisions at a personal, local council and at country level regarding policies to be enacted to reduce air pollution.



# IMPLEMENTATION

## DATA AND FEATURES

Our dataset is obtained from AirQo's air quality sensing network has more than 120 low-cost devices deployed across Uganda, which regularly sense and report on air quality using the PM2.5 measure.

We're trying to predict the ordinal variable pm2.5, which represents a level of pm2.5 in the air. There are 6 grades as the following:

- 0 represents the 'very unhealthy'
- 1 represents the 'good'
- 2 represents the 'hazardous'
- 3 represents the 'moderate'
- 4 represents the 'sensitive'
- 5 represents the 'unhealthy'

There are 9923 rows 71 columns in this dataset, where 'ID', 'date', 'device' columns are unique and random identifier (which represents the device details and when its used). Remaining 68 columns are taken for processing and analysis which helps to identify and related actions. And each row represents the data obtained from a single unit of low-cost air quality device by AirQo's survey as it's measured.

# EXPLORATORY DATA ANALYSIS

The image below gives info regarding the dataset (fig.1):

Data columns (total 71 columns):

#	Column	Non-Null Count	Dtype
0	ID	9923 non-null	object
1	date	9923 non-null	object
2	device	9923 non-null	object
3	site_latitude	9923 non-null	float64
4	site_longitude	9923 non-null	float64
5	humidity	9923 non-null	float64
6	temp_mean	9903 non-null	float64
7	SulphurDioxide_SO2_column_number_density	4291 non-null	float64
8	SulphurDioxide_SO2_column_number_density_amf	4291 non-null	float64
9	SulphurDioxide_SO2_slant_column_number_density	4291 non-null	float64
10	SulphurDioxide_cloud_fraction	4291 non-null	float64
11	SulphurDioxide_sensor_azimuth_angle	4291 non-null	float64
12	SulphurDioxide_sensor_zenith_angle	4291 non-null	float64
13	SulphurDioxide_solar_azimuth_angle	4291 non-null	float64
14	SulphurDioxide_solar_zenith_angle	4291 non-null	float64
15	SulphurDioxide_SO2_column_number_density_15km	4291 non-null	float64
16	CarbonMonoxide_CO_column_number_density	5463 non-null	float64
17	CarbonMonoxide_H2O_column_number_density	5463 non-null	float64
18	CarbonMonoxide_cloud_height	5463 non-null	float64
19	CarbonMonoxide_sensor_altitude	5463 non-null	float64
20	CarbonMonoxide_sensor_azimuth_angle	5463 non-null	float64
21	CarbonMonoxide_sensor_zenith_angle	5463 non-null	float64
22	CarbonMonoxide_solar_azimuth_angle	5463 non-null	float64
23	CarbonMonoxide_solar_zenith_angle	5463 non-null	float64
24	NitrogenDioxide_NO2_column_number_density	3005 non-null	float64
25	NitrogenDioxide_tropospheric_NO2_column_number_density	3005 non-null	float64
26	NitrogenDioxide_stratospheric_NO2_column_number_density	3005 non-null	float64
27	NitrogenDioxide_NO2_slant_column_number_density	3005 non-null	float64
28	NitrogenDioxide_tropopause_pressure	3005 non-null	float64
29	NitrogenDioxide_absorbing_aerosol_index	3005 non-null	float64
30	NitrogenDioxide_cloud_fraction	3005 non-null	float64
31	NitrogenDioxide_sensor_altitude	3005 non-null	float64
32	NitrogenDioxide_sensor_azimuth_angle	3005 non-null	float64
33	NitrogenDioxide_sensor_zenith_angle	3005 non-null	float64
34	NitrogenDioxide_solar_azimuth_angle	3005 non-null	float64
35	NitrogenDioxide_solar_zenith_angle	3005 non-null	float64

36	Formaldehyde_tropospheric_HCHO_column_number_density	5277	non-null	float64
37	Formaldehyde_tropospheric_HCHO_column_number_density_amf	5277	non-null	float64
38	Formaldehyde_HCHO_slant_column_number_density	5277	non-null	float64
39	Formaldehyde_cloud_fraction	5277	non-null	float64
40	Formaldehyde_solar_zenith_angle	5277	non-null	float64
41	Formaldehyde_solar_azimuth_angle	5277	non-null	float64
42	Formaldehyde_sensor_zenith_angle	5277	non-null	float64
43	Formaldehyde_sensor_azimuth_angle	5277	non-null	float64
44	UvAerosolIndex_absorbing_aerosol_index	9588	non-null	float64
45	UvAerosolIndex_sensor_altitude	9588	non-null	float64
46	UvAerosolIndex_sensor_azimuth_angle	9588	non-null	float64
47	UvAerosolIndex_sensor_zenith_angle	9588	non-null	float64
48	UvAerosolIndex_solar_azimuth_angle	9588	non-null	float64
49	UvAerosolIndex_solar_zenith_angle	9588	non-null	float64
50	Ozone_O3_column_number_density	9387	non-null	float64
51	Ozone_O3_column_number_density_amf	9387	non-null	float64
52	Ozone_O3_slant_column_number_density	9387	non-null	float64
53	Ozone_O3_effective_temperature	9387	non-null	float64
54	Ozone_cloud_fraction	9387	non-null	float64
55	Ozone_sensor_azimuth_angle	9387	non-null	float64
56	Ozone_sensor_zenith_angle	9387	non-null	float64
57	Ozone_solar_azimuth_angle	9387	non-null	float64
58	Ozone_solar_zenith_angle	9387	non-null	float64
59	Cloud_cloud_fraction	8414	non-null	float64
60	Cloud_cloud_top_pressure	8414	non-null	float64
61	Cloud_cloud_top_height	8414	non-null	float64
62	Cloud_cloud_base_pressure	8414	non-null	float64
63	Cloud_cloud_base_height	8414	non-null	float64
64	Cloud_cloud_optical_depth	8414	non-null	float64
65	Cloud_surface_albedo	8414	non-null	float64
66	Cloud_sensor_azimuth_angle	8414	non-null	float64
67	Cloud_sensor_zenith_angle	8414	non-null	float64
68	Cloud_solar_azimuth_angle	8414	non-null	float64
69	Cloud_solar_zenith_angle	8414	non-null	float64
70	pm2_5	9923	non-null	object

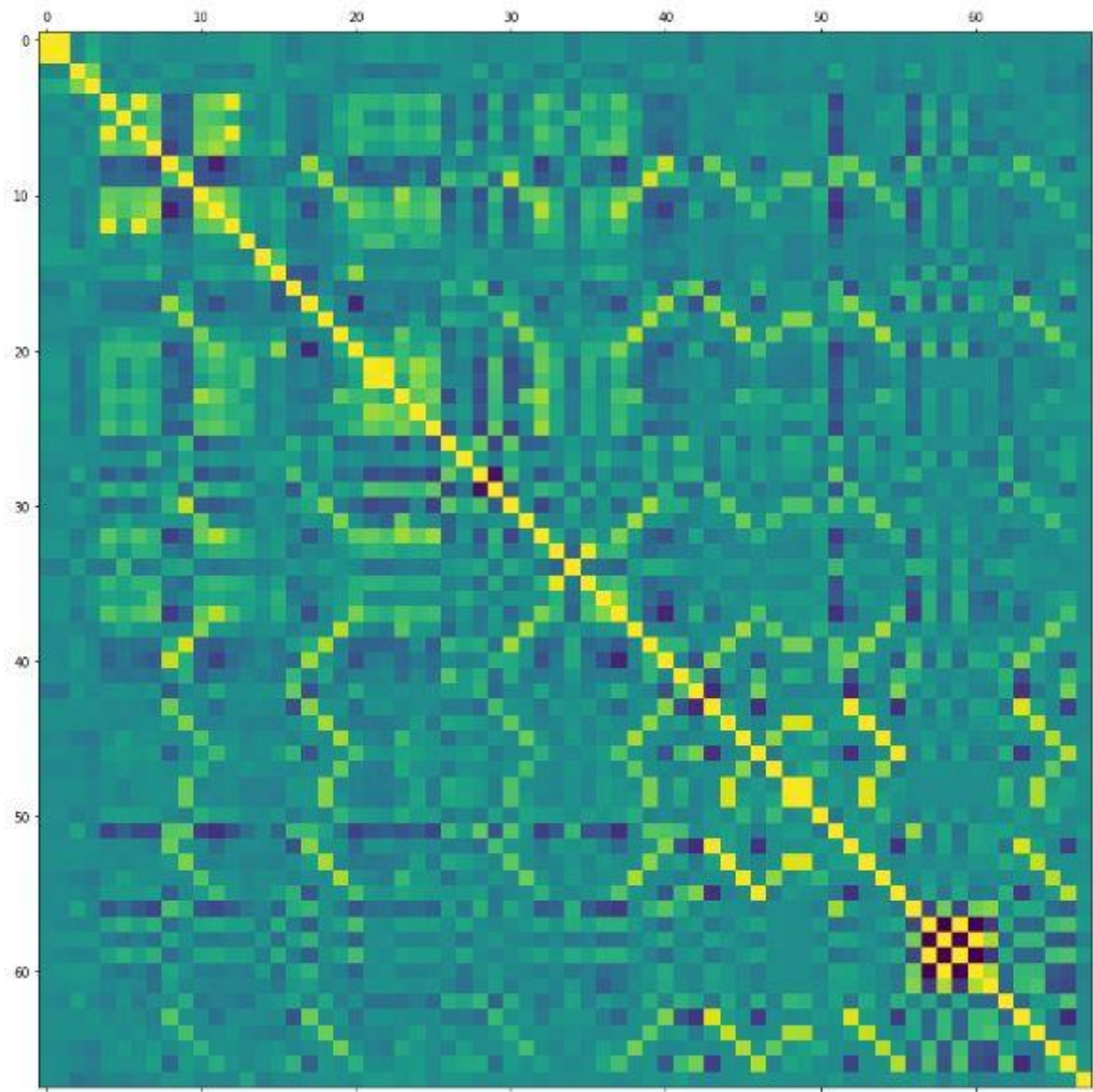
dtypes: float64(67), object(4)

memory usage: 5.4+ MB



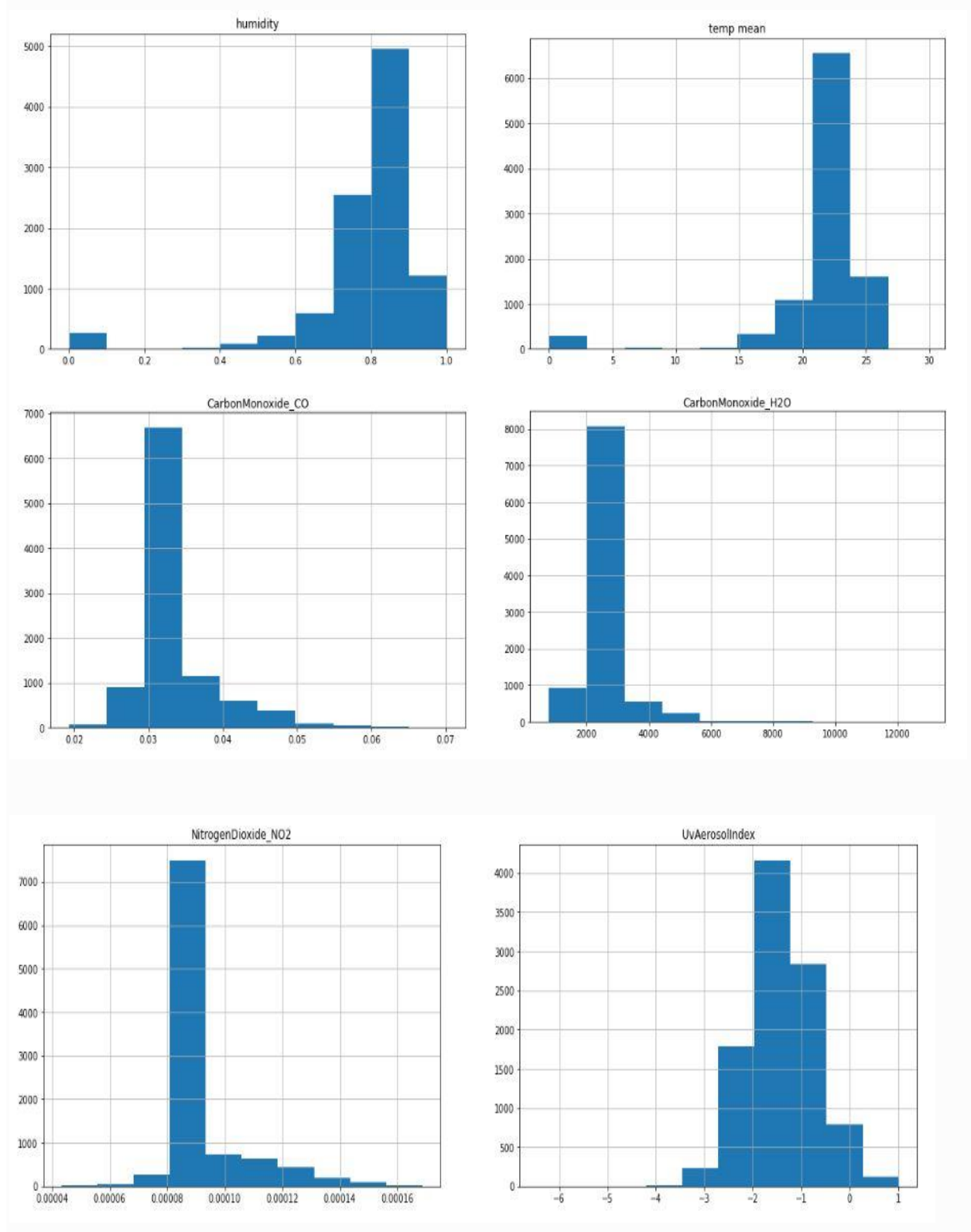
## CORRELATION

The below explains how one or more variables are related to each other (**fig.2**).

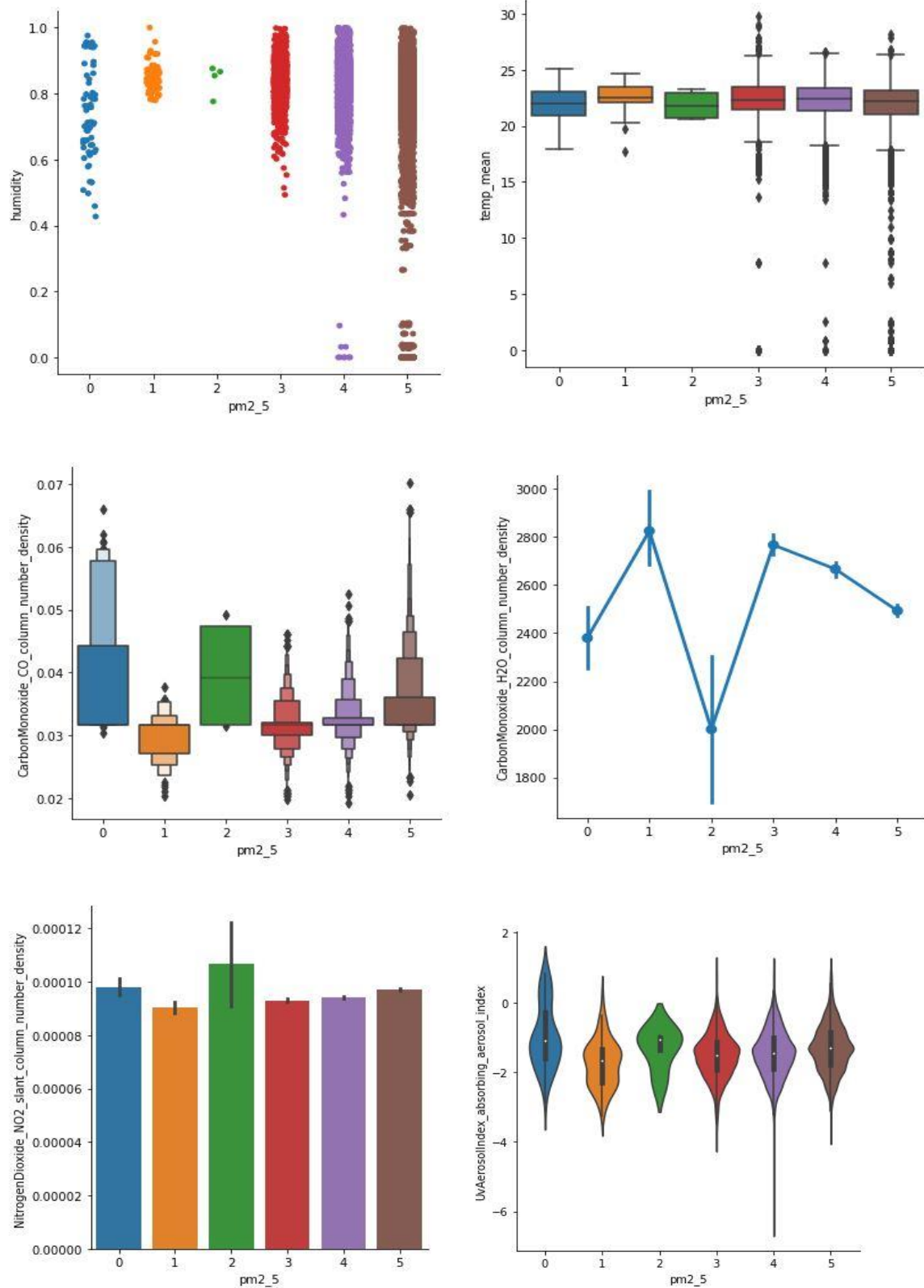


Understanding the above image, there are many columns which is highly correlated meaning they can be predicted by other independent variables in the dataset.

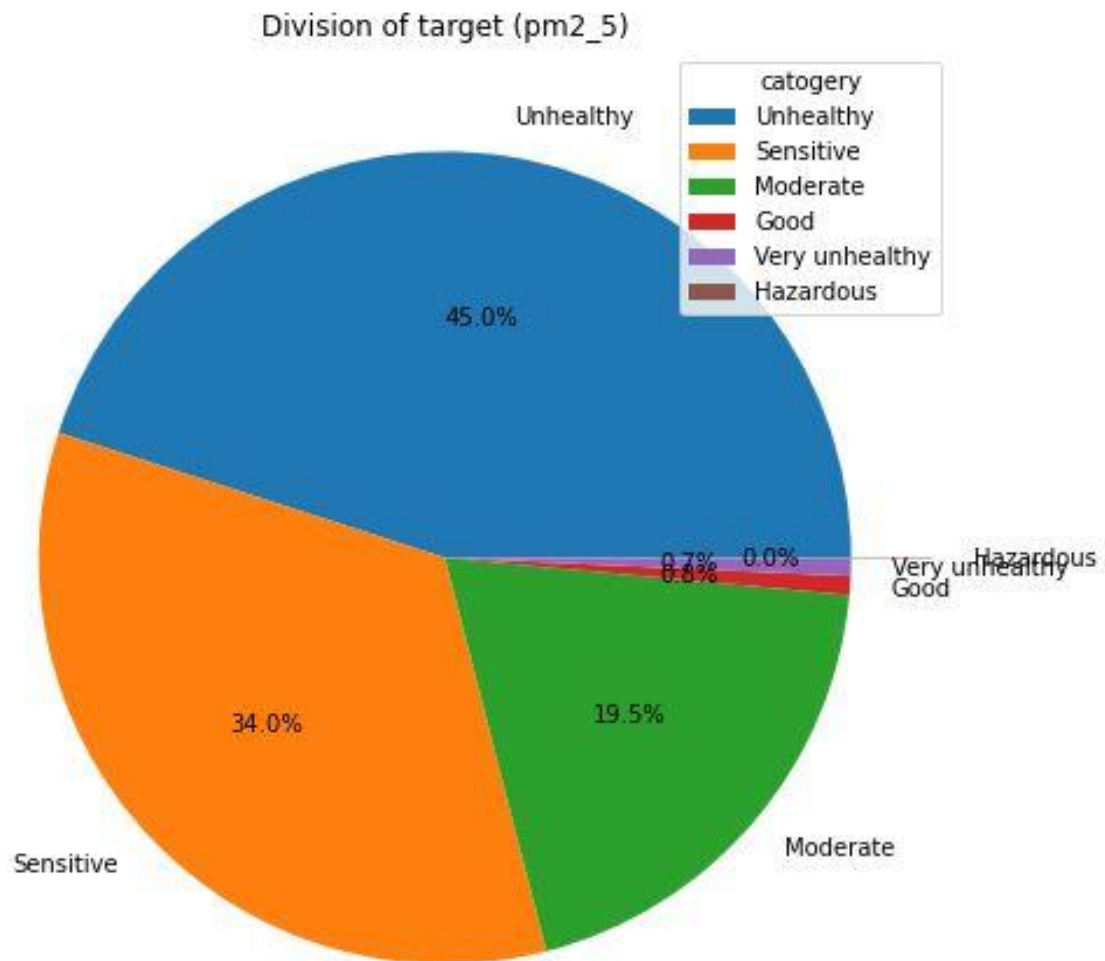
The following diagrams show the most related columns to the target (**fig.3**):



The following plots shows how related these columns with target (**fig.4**):



The following representation of data is relative to the whole. Each portion in the circle represents an element of the collected data **(fig.5)**.

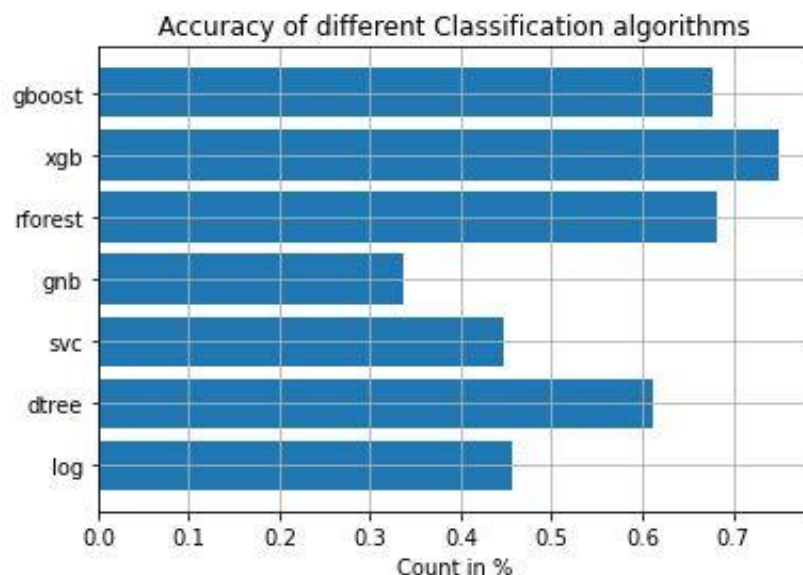


## FEATURE SELECTION

In feature selection first a function is created with data and threshold as the parameters to check for multicollinearity among the input variables. Then the function is called with the training dataset as data and 0.9 as threshold. The result is then stored in the variable 'multicol'. The output of the function suggests that the columns; 'Cloud\_cloud\_base\_height', 'Cloud\_cloud\_base\_pressure', 'Cloud\_cloud\_top\_height', 'Cloud\_sensor\_azimuth\_angle', 'Formaldehyde\_HCHO\_slant\_column\_number\_density', 'NitrogenDioxide\_tropospheric\_NO2\_column\_number\_density', 'Ozone\_O3\_slant\_column\_number\_density', 'Ozone\_sensor\_azimuth\_angle', 'Ozone\_sensor\_zenith\_angle', 'Ozone\_solar\_azimuth\_angle', 'Ozone\_solar\_zenith\_angle', 'SulphurDioxide\_SO2\_slant\_column\_number\_density\_15km', 'SulphurDioxide\_SO2\_slant\_column\_number\_density' and 'site\_longitude' are highly correlated. These columns are then dropped from the training data. Further the columns 'ID', 'date' and 'device' does not correlate with our output and would not have any significant effect on our training model, so it is also dropped from our dataset. After feature selection the categorical columns are label encoded and the dataset is split in test data and train data using train test split with train data having 75% and test data having 25%. After splitting data we move on to model creation.

## BUILDING THE MODEL

In the initial stage of building the model, I have created a variable called 'models' is created with dictionary which containing the different classification algorithms i.e. Logistic regression, Decision tree classifier, SVC, GaussianNB, Random forest classifier where, n\_estimators given 500, XGBclassifier and Gradient boosting classifier. Then the train data which is fit to different algorithm models with the help of a loop and the accuracy and f1 score of different models are compared and mentioned below **(fig.6)**.





Where, y axis refers to;

- 'gboost' – Gradient boosting algorithm
- 'xgb' – XGBoosting algorithm
- 'rforest' – Random forest classifier algorithm
- 'gnb' – GaussianNB algorithm
- 'svc' – SVC algorithm
- 'dtree' – Decision tree classifier algorithm
- 'log' – Logistic regression algorithm

Understanding the different models, XGB classifier gives more accuracy as compared. The classification report as below (**fig.7**):

	precision	recall	f1-score	support
0	0.20	0.08	0.11	13
1	1.00	0.09	0.17	22
2	0.00	0.00	0.00	1
3	0.72	0.70	0.71	475
4	0.67	0.68	0.68	859
5	0.82	0.85	0.84	1111
accuracy			0.75	2481
macro avg	0.57	0.40	0.42	2481
weighted avg	0.75	0.75	0.75	2481

## CONCLUSION

The project successfully predicted different levels of PM2.5 particulate matter concentration in the air for some locations in Uganda. Out of the various machine learning models used in the project XGB classifier algorithm showed significant improvement in accuracy of 75%.

This model may assist stakeholders, decision-makers and researchers in rapid seismic risk assessment in order to formulate and implement new plans and policies in PM2.5 concentration in the air in Uganda. Further investigation should be carried out for a better understanding of the applicability of the machine learning model in PM2.5 concentration prediction based on the need and interests of the decision-makers and researchers.

# CODE

## IMPORTING LIBRARIES:

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
```

## IMPORTING DATASET:

```
train=pd.read_csv('train_AQI.csv')
train
```

## SHAPE OF DATASET:

```
train.shape
```

## DATASET INFO:

```
train.info()
```

## DATASET DESCRIPTION:

```
train.describe()
```

## CHECKING DATATYPE:

```
print(train.dtypes.to_string())
```

## CHECKING NULL VALUES:

```
print(train.isna().sum().to_string())
```

## DROPPING UNNECESSARY COLUMNS:

```
train.drop('ID',axis=1,inplace=True)
train.drop('date',axis=1,inplace=True)
train.drop('device',axis=1,inplace=True)
```

## CHANGING NULL VALUES:

```
for i in train.columns:
    train[i].fillna(train[i].mode()[0],inplace=True)
```

## TARGET COLUMN COUNTS:

```
train['pm2_5'].value_counts()
```

## CONVERTING LABEL INTO NUMERIC FORM:

```
from sklearn.preprocessing import LabelEncoder

encode=LabelEncoder()

train['pm2_5']=encode.fit_transform(train.pm2_5)
```

## FEATURE SELECTION:

```
corr=train.corr()

plt.figure(figsize=(25,25))
sns.heatmap(corr,cmap="Blues",linewidths=1,annot=True)

size=plt.figure(figsize=(15,15))
plt.matshow(corr,fignum=size.number)
plt.show()
```

Output in (fig.1)

## SPLITTING TARGET COLUMN:

```
x_data=pd.DataFrame(train.drop('pm2_5',axis=1))

target_data=pd.DataFrame(train['pm2_5'])
```

## DETECTING MULTICOLINEARITY AND REMOVEING:

```
def multicolliniarity(data,threshold):
    output=set()
    corr=data.corr()
    for i in range(len(corr.columns)):
        for j in range(i):
            if abs(corr.iloc[i,j])>threshold:
                b=corr.columns[i]
                output.add(b)

    return(output)

multicol=multicolliniarity(x_data,0.90)

data=x_data.drop(multicol,axis=1)
data
```

## DATA VISUALISATION OF MOST RELATED:

### FINDING MOST RELATED COLUMNS:

```
forgraph=train.drop(multicol,axis=1)
```

```
corr2=forgraph.corr()['pm2_5']  
corr2
```

```
highcorr=abs(corr2)>=0.1  
print(highcorr[highcorr==True])
```

```
plt.figure(figsize=(20,20))  
plt.subplot(3,2,1)  
plt.title('humidity')  
train.humidity.hist()
```

```
plt.subplot(3,2,2)  
plt.title('temp mean')  
train.temp_mean.hist()
```

```
plt.subplot(3,2,3)  
plt.title('CarbonMonoxide_CO')  
train.CarbonMonoxide_CO_column_number_density.hist()
```

```
plt.subplot(3,2,4)  
plt.title('CarbonMonoxide_H2O')  
train.CarbonMonoxide_H2O_column_number_density.hist()
```

```
plt.subplot(3,2,5)  
plt.title('NitrogenDioxide_NO2')  
train.NitrogenDioxide_NO2_slant_column_number_density.hist()
```

```
plt.subplot(3,2,6)  
plt.title('UvAerosolIndex')  
train.UvAerosolIndex_absorbing_aerosol_index.hist()
```

Output in (fig.2)

### COMPARISON OF THOSE COLUMNS WITH TARGET:

```
sns.catplot(data=forgraph,x='pm2_5',y='humidity')
```

```
sns.catplot(data=forgraph,x='pm2_5',y='temp_mean',kind='box')
```

```
sns.catplot(data=forgraph,x='pm2_5',y='CarbonMonoxide_CO_column_number_density',kind='boxen')
```

```
sns.catplot(data=forgraph,x='pm2_5',y='CarbonMonoxide_H2O_column_number_density',kind='point')
```

```
sns.catplot(data=forgraph,x='pm2_5',y='NitrogenDioxide_NO2_slant_column_number_density',kind='bar')
```

```
sns.catplot(data=forgraph,x='pm2_5',y='UvAerosolIndex_absorbing_aerosol_index',kind='violin')
```

Output in (fig.3)

## WHOLE DATA REPRESENTATION:

```
label=['Unhealthy','Sensitive','Moderate','Good','Very unhealthy','Hazardous']
```

```
plt.figure(figsize=(8,8))
plt.title('Division of target (pm2_5)')
hazardous=[0,0,0,0,0,.2]
plt.pie(forgraph.pm2_5.value_counts(),autopct='%1.1f%%',labels=label,explode=hazardous)
plt.legend(title='catogery')
plt.show()
```

Output in (fig.4)

## SPLITTING DATASET INTO TRAIN AND TEST:

```
from sklearn.model_selection import train_test_split
```

```
xtrain,xtest,ytrain,ytest=train_test_split(data,target_data,train_size=.75)
```

## BUILDINNG MODELS:

### IMPORTING ALGORITHMS:

```
from sklearn.linear_model import LogisticRegression
```

```
from sklearn.tree import DecisionTreeClassifier
```

```
from sklearn.svm import SVC
```

```
from sklearn.naive_bayes import GaussianNB
```

```
from sklearn.ensemble import RandomForestClassifier
```

```
from xgboost import XGBClassifier
```

```
from sklearn.ensemble import GradientBoostingClassifier
```



## FITTING TRAIN DATA INTO MODELS AND CHECKING ACCURACY:

```
result=[]
```

```
models={'log':{'model':LogisticRegression()},
        'dtree':{'model':DecisionTreeClassifier()},
        'svc':{'model':SVC()},
        'gnb':{'model':GaussianNB()},
        'rforest':{'model':RandomForestClassifier(n_estimators=500)},
        'xgb':{'model':XGBClassifier()},
        'gboost':{'model':GradientBoostingClassifier()}}
```

```
for i in models:
    a=models.get(i)
    model=a.get('model')
    model.fit(xtrain,ytrain)
    score=model.score(xtest,ytest)
    out={'model':i,'score':score}
    result.append(out)
```

## THE RESULTS SAVED TO VARIABLE:

```
final=pd.DataFrame(result)
final
```

```
plt.barh(final.model,final.score)
plt.title('Accuracy of different Classification algorithms')
plt.xlabel('Count in %')
plt.grid()
```

Output in (fig.5)

## ACCURATE ALGORITHM AND PREDICTION:

```
model=XGBClassifier()
```

```
model.fit(xtrain,ytrain)
```

```
model.score(xtest,ytest)
```

```
predict=model.predict(xtest)
```

## CLASSIFICATION REPORT:

```
from sklearn.metrics import classification_report
```

```
print(classification_report(ytest,predict))
```

Output in (fig.6)