**AIR QUALITY ANALYSIS AND PREDICTION USING MACHINE LEARNING**

**A CAPSTONE PROJECT REPORT**

*Submitted in partial fulfillment of the*

*requirement for the award of the*

*Degree of*

**MASTER OF TECHNOLOGY**

**IN**

**INTEGRATED SOFTWARE ENGINEERING**

*by*

**NAGABHAIRU VINOD KUMAR (17MIS7149)**

*Under the Guidance of*

**DR. ANJALI GAUTHAM**



SCHOOL OF ELECTRONICS ENGINEERING

VIT-AP UNIVERSITY

AMARAVATI- 522237

*SEPTEMBER 2021*

**CERTIFICATE**

This is to certify that the Capstone Project work titled “AIR QUALITY ANALYSIS AND PREDICTION USING MACHINE LEARNING” that is being submitted by NAGABHAIRU VINOD KUMAR(17MIS7149) are in partial fulfilment of the requirements for the award of Master of Technology, is a record of bonafide work done under my guidance. The contents of this Project work, in full or in parts, have neither been taken from any other source nor have been submitted to any other Institute or University for the award of any degree or diploma and the same is certified.

Dr. Anjali Gautham

Guide

**The thesis is satisfactory / unsatisfactory**

**Internal Examiner External Examiner**

**Approved by**

**PROGRAM CHAIR DEAN**

M. Tech. SE School Of Computer Science and Engineering

**TABLE OF CONTENTS**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Chapter** | **Title** | **Page Number** |
| **0** |  | **Acknowledgement** | **4** |
| **1.** |  | **Abstract** | **5** |
| **2.** |  | **Introcution** | **6** |
| **3.** |  | **Literature Survey** | **7** |
| **4.** |  | **Problem Identification** | **12** |
| **5.** |  | **Objectives** | **14** |
| **6.** |  | **System Overview & Methodology of system** | **16** |
| **7.** |  | **Algorithm used in Air Quality prediction** | **17** |
| **8.** |  | **Implementation**     |  |  |  | | --- | --- | --- | |  | 8.1 Coding |  | |  | 8.2 Testing |  | |  | 8.3 Test Case |  | | **18** |
| **9.** |  | **Result** | **21** |
| **10)** |  | **CONCLUSION** | **22** |
| **11)** |  | **Reference** | **23** |

**ACKNOWLEDGEMENTS**

I wish to express our significant and true appreciation to DR. ANJALI GAUTHAM, Assistant Professor, Department of Computer Science, Vellore Institute of Information Technology,Amaravathi, who guided me into the complexities of this project non-chillingly with incomparable generosity. She always co-worked and Assisted with the examination work. She likewise revealed unmistakable fascination and significant help in the field of Machine Learning for advancement of our task work.she helped me to investigate multi dimensions in work for betterment in end output.

**ABSTRACT**

We find the air quality of India by using AQI to foresee the air quality list of secure Space. Air quality file of india can be a typical live accustomed demonstrate the waste product quantity the amount of chemical element oxide and live of oxide calculable respirable suspended material Estimated. (so2, no2, rpm, Spm). So forth levels over the total. We tended to create up an model to foresee the Air quality list upheld earlier years and foreseeing over a selected returning year as this text address for foreseeing and examination of Air quality. Angle nice helped multivariable Relapse downside. we tend to increase the portion of the model by applying esteem estimation for our fateful downside. Our model is virtuoso for with progress anticipating the air quality list of a complete region or any state or any delimited space supplied with the chronicled data on squander material fixation. In our model by finishing up the projected Boundary decreasing plans, we've associate degree inclination to accomplished higher than the standard relapse models. Our model has exactness on foreseeing the present out there dataset on anticipating the Air quality file, Additionally, we tend to utilize perceptive chain of importance strategy multi-Rules decision-creating methodology to appear faulty of inclination by similitude to ideal goal.

**.**

**INTRODUCTION**

Air quality transfer to the pollution into the air harmful to the world has filled with pollution in the air. Nowadays pollution has been a lot of trouble to change for any nation. In South Asia, it is ranked accordingly sixth many dangerous killers. One does not notice the harmful effects of a retardant if he/she has not practised it at intervals in The first place. Take Delhi, as an as proof tend all or any have practised what it looks like huffing deadly’ smog that remained for a count of a week, once Diwali. Voters were prompt not to leave there were asked to wear masks whenever going outside. Attempting Outside The window created American state needs wont to be living during a death chamber. Low visibility, a high vary of deaths, etc. were the implications of pollution. Being information an information analysis and Data Science enthusiast, created a call to associate air quality is considered to be unhealthy, own country to go looking out some underlying principles or patterns that might provide the American state with an insight into however severe the matter in which I ought to say the results were value sharing. So, here I’m penning this article to share my approach, and what I analysed from The data and to, in addition, produce people alert to the massive disadvantage our country is facing machine learning, Is comparatively sturdy to perturbations and doesn’t need the other physical variables for prediction. Therefore, Machine learning is a far higher chance of the evolution of air quality prediction. Before the Advancement of Technology, Air quality prediction was a tough unto crack. Information model’s part conditions with less accuracy. AQI and analysis have immensely augmented regarding the accuracy and the sure thing with the advancement of knowledge because the largest growing industrial nation, India is producing a record quantity of pollutants specifically greenhouse emission, pm2.5 etc. and different harmful aerial contaminants. Pollution of a particular state or a rustic maybe a life on the effect of pollution on the revered regions, as per the Indian air quality commonplace pollutants square measure indexed concerning their square measure numerous part gases that cause pollution on our environment. Every pollution has an individual index and scales at completely different levels. The major particles like (no2, so2, rpm, SPM) indexes AQI is with this individual AQI, The information will be categorized supported the limits. We do tend to collect information from The Indian Government Information that contains waste concentration occurring at numerous places across Asian nations. We do tend to begin by calculative the individual index of the waste for each available information point and realize their several AQI for the region. We've designed a model to predict the air Index of every accessible Information point within the dataset, our model is capable of statement India in any given space .by estimate the air quality, we to get back The key pollution in inflicting pollutants and also the location affected seriously by the waste across the Asian nation. With this statement model, Various Knowledge concerning the Information square measure extracted mistreatment numerous techniques to get heavily affected regions on a selected region (cluster). This provides additional info and information concerning the cause and seniority of the pollutants. Is international environmental downside that influences principally health of urban population? over the past few decades, epidemiologic studies have incontestable adverse health effects thanks to higher close stages of pollution. Studies have indicated that perennial exposures to close air pollutants over a chronic amount of your time increases the danger of being liable to airborne diseases like disorder, respiratory illness, and carcinoma (WHO, 2009). pollution has been systematically coupled to substantial.

**Literature Review**

Fine material (PM2.5) may be a crucial one as a result of it is a big concern to individuals health once its level among the air is relatively high. PM2.5 refers to very few particles within the air that reduce visibility and cause the air to seem hazy once levels area unit elevated. But within the planned system, we tend to calculate the air quality index of all the pollutants mistreatment The A QI formulae to understand the air quality level in a very specific town mistreatment gradient descent and Box-Plot analysis. Within the planned system the air quality index of the approaching years is often foretold mistreatment the current AQI values to understand the temporal variation and episodic rise of the pollution in the study region real-time air quality watching was administrated at the residential website within the gift study close Air quality was measured by surroundings for a fine material pm10.pm2.5, so2 no2 the fine particulate monitor. We have a retardant statement that comes below the class of classification. It's a multi-class classification within which the categories were given to US area unitNO2: Nitrogen Dioxide and is emitted mostly from combustion from power sources or transport.SO2: Sulphur Dioxide and is emitted mostly from coal burning, oil burning, manufacturing of Sulphuric acid.( spm) suspended particulate matter and are known to be the deadliest form of air pollution. They are microscopic in nature and are found to be suspended in the earth's atmosphere. Rspm: Respirable suspended particulate matter. A subform of rspm and are responsible for respiratory diseases.pm2\_5: Suspended particulate matter with diameters less than 2.5 micrometres. They tend to remain suspended for longer durations and are potentially very harmful. The aim is to classify the given information into higher than given categories. So as to try to, therefore, we've got to analyses initially the info given to the US. For analyzing the options, we tend to area unit mistreatment totally different techniques. The coaching of a model is often wiped out in many ways. It depends on, however, the information is ready for more process. The info is often used directly betting on true or the info is often accustomed kinda bar graph. When these modifications, we elect a specific model on that we are going to train our information. This model will be statistical regression, supply Regression, Random Forest, etc. Parameter standardization may also be wiped out in order to extend our accuracy. Once the model is trained. We will take a look at our information by applying our algorithms on the take a look at with assist.

PROBLEM IDENTIFICATION

Observing and retentive high normal air have become the crucial challenge in metropolitan areas that have More industries, firms and populations. As there's an increase in population, there's a rise within the transportation, Usage of electricity and fuels. There are tons to waste drop within the land that we tend to area unit cognizant of. The air is additionally extremely contaminated which causes an additional serious threat to any or all Sorts of living organisms within the earth. This provides rise to the necessity for the observance and assessing the standard of air and consequently, the government should run attentive to take necessary actions. This analysis Work concentrates on activity a good analysis on all the foremost works wiped out this facet victimization machine learning algorithms. The essential objective of any Air quality index is to remodel the measured concentrations of individual air waste into one numerical index victimization appropriate aggregation mechanism. Ideally, each index ought to mirror each measured and publically perceived quality of the close air for the fundamental quantity it covers. As a result, air quality indices conceive to standardize and synthesize pollution info and allow comparisons to be without delay undertaken and to satisfy public demands for correct, simple to interpret Knowledge. In the style of air quality Indices, the subsequent criteria should be used: be without delay perceivable by the public; embrace the foremost criteria pollutants, and their synergism; be expandable for different pollutants and averaging times; be associated with national close Air Quality Standards employed in Individual provinces; avoid “eclipsing” (eclipsing happens once a pollution index doesn’t Indicate poor air quality despite the very fact that concentrations of 1 or additional air pollutants could have reached unacceptably (High values); Avoid, “ambiguity” (ambiguity happens once a pollution index offers falls alarmed despite the very Fact that concentrations of all the pollutant's area unit among the permissible limit except one); be usable as an alert system; be supported valid air quality knowledge obtained from observance stations that area unit set, therefore, on representing the final air quality within the community.

**OBJECTIVES**

The overall goal of this project the air quality of Bharat by mistreatment machine learning algorithms to predicts the air quality index (the AQI). An air quality Index could be normal live to work out the standard of air. A concentration of gases is the asso2, a no2, the co2, and rspm, SPM. Etc. rerecorded by the agencies. A model to predict the air quality index had supported historical information of previous years and predicting over specific approaching the year as a multivariable regression drawback. They have improved the potency of the model by applying the price estimation for the prognostic drawback we are able to say that this model is capable of successfully predicting the air a quality index shall of a year to be the increase region supplied with the historical information of a waste product concentration this system has used the linear regression (LR) and multilayer perceptron (ANN) the protocol for a prediction of the pollution of the next day. This plan the system stem would be.

Two necessary tasks

(i). Detects the degree had of a pm2.5 supported given region values.

(ii) Predicts the amount of a pm2.5 for the selected date. A logistic Regression has employed the sigh tight whether an information sample is either Increased. Motorcar regression is utilized to predict future value pm2 of PM2.5 supported the Previous PM2.5 Reading. The first goal is to predict pollution la town in town with the bottom info section set!

**SYSTEM METHODOLOGY**

Nowadays it's necessary to society to seem for awareness of daily levels of pollution AQI may be a tool that is employed to report the air quality standing and trends supported a specific customary in Bharat we tend to area unit victimization cpcb stands for shrewd air quality index or environment pollution index this index provides a plan regarding the atmospheres standing as an air quality and additionally tells the final public to grasp, however clean overall this index may be used to offer significant analysis of pollution to policies manage men in industries will cut back level of domination Pollution AQI is illustration the accumulative impact of all the pollution to indicate overall Air quality standing in higher manner the AQI of specific pollution springs primarily from the physical measuring of pollution file likepm10,pm2.5, no2, so2 etc. within the represent study, six totally different ways accustomed to calculate quality index to create solicitations and method The came back data we'll create utilization of one or two normal libraries, and a few well-known outsider libraries. When putting in libraries, we'll outline knowledge into knowledge variables and set up them from the primary day of the year currently we tend to area unit heading toward knowledge retrieval and putting in place our panda's knowledge frame. Since we've got a large records summary of daily Summary. We'll utilize it to figure out a panda's knowledge frame. The Panda's knowledge Frame is an extremely useful data structure for a few, programming errands that area unit most prominently notable for cleansing and handling data to be utilized in machine learning undertakings. At that time we'll derive the options and classify the info into categories as we tend to have explicit higher than the foremost necessary factor is knowledge cleansing. The infamous speech communication, “refuse in, waste out”, is as fitting as ever regarding machine learning. Be that because it could, the Information cleansing some portion of an examination venture is not solely a standout amongst the most vital elements it's likewise the foremost tedious and relentless.

**SYSTEM DESIGN**

The records have just been separated into a plaything and test set with a splitter. Each information has just been labelled accordingly. First, we take the Trainset organizer. we'll let train our model with the assistance of histograms graphs. The feature so extracted is stored during a histogram graph. This process is completed for each data within the plaything . Now we'll build the model of our classifiers (Algorithms). The classifiers which we'll take under consideration are rectilinear regression , Polynomial Regression, Random Forest and Neural Networks. With the assistance of our histogram, we'll train our model. the foremost important thing during this process is to tune three parameters accordingly, such we get the foremost accurate results. Once the training is fully complete, we'll take the test set we split before. Now for every data variable of the test set, we'll extract the features using feature extraction techniques then compare its values with the values present within the histogram formed by the plaything . The output is then predicted for every test day. Now so as to calculate accuracy, we'll compare the anticipated value with the labelled value. the various metrics that we'll use are our confusion matrix, accuracy score, f1 score etc.

**MACHINE LEARNING**

The information the knowledge we've for this project was air quality dataset obtained from

ministry of surroundings and forests, and central pollution panel of INDIA| India| Republic

Of Bharat| Bharat| asian country land beneath the national information sharing and accessibility policy electronic computer the air quality dataset from the India air quality information to predict the air quality index of a particular region, we would like the material concentration of all he gases which may be out there among the online website that holds all the knowledge that pollutes the cities every year. The air formula or progressing to Be applied therefore, on calculate the AQI by exploitation the regression toward the mean formula for a particular year. Several information sets or progressing to be foreign among the directory and null values or progressing to be set to the infinite information. The expected and actual values or progressing to be delineate

Victimization the Box-Plot analysis therefore, on get eliminate the outliers. Electronic computer provide the Information set transfer Tool which enable the user to transfer information to predict the air quality index of a Specific Region, we'd like the waste concentration of all the gases which can be on the market within the Cpcb.NIC.in website that holds all the info that pollutes the cities per annum. The AQI formulae are applied to calculate.The AQI by victimization. The statistical regression algorithmic program for a specific year. Many Information sets are foreign within the directory and null values are set to the infinite data. The anticipated and actual values are pictured victimization the Box-Plot analysis to get rid of the outliers.

## Model Development

Our strategy for model improvement is exploratory. the target of our undertaking is to form sure the conclusion of malignancy with the sole accuracy. This must be accomplished by exploring different avenues regarding distinctive systems from a particular field. we've considered that programmed learning descriptors and algorithms. The Machine Learning Algorithms that we are using are simple regression, Random Forest, Logistic Regression Subsequently our point is to locate the only mix which might furnish us with the sole precision. Along those lines, this task is totally test based. additionally, parameter tuning could even be an interesting piece of any Machine Learning Algorithm. Then regardless of whether that calculation works exceptionally solid in a specific condition, at that point too thanks to terrible determination of parameters, the precision may be below. during this way , we likewise needed to centre around the right arrangement of parameters. Hence parameter tuning must be exhausted whichever show we pick. Then Parameter tuning technique must should either has to be possible physically or by utilize the lattice seek of technique. Network looking is that the procedure during which information is checked with the best possible goal to urge ideal parameters for some of random model. Contingent upon the kind of model that's utilizing, tuning of the precise parameters is critical . Framework seeking applies to a solitary model sort also as reasonably models. Network looking are often connected in machine learning with the absolute best goal to work out the only real parameters for its utilization in some random model. It alright could even be computationally greatly costly and might forgot an extended opportunity to remain running on the machine. Matrix Search constructs a model on every conceivable that parameter mix. At last point , it repeats it through every parameter is blend lastly stores a model for each mix..

**ALGORITHMS UESD IN THE PREDICTION**

**SIMPLE LINEAR REGRESSION**

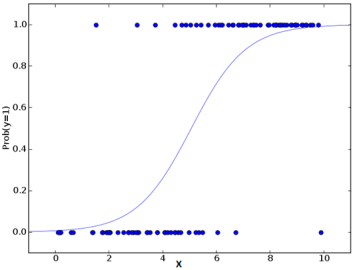
Here Simple Statical regression is often a factual strategy that allows the US to foreshorten and consider connections Between 2 perpetual (quantitative) factors: One variable, meaning x, is viewed due to the indicator, logical, or Free issue. the opposite variable, indicated y, is viewed due to the reaction, result, or ward variable. Since Alternate terms are utilized fewer many times nowadays, we'll use the “predictor “and “reaction” terms to direct to the factors skilled during this course. Alternate terms or documented simply to allow your data of them. That you just should expertise they are in numerous fields. Easy easy regression gets its descriptor, "Basic,” Since it issues the investigation of simply one indicator variable. Conversely, totally different straight Relapse that we do tend to seem at later during this course, gets its descriptor “numerous,” on the grounds that it Issues the Investigation of a minimum of two construction factors. The previous information is given in Table One. we are ready to observe a positive relationship between X and Y. there's also a relation ascertained between X and Y, Higher the estimation of X many correct are visiting be the prediction of Y. regression Incorporates exploring the simplest line through the points additionally cited as curve. The dark Line during Figure two is that the regression curve and consists of the expected score on Y for each estimation of X. The vertical lines from the focuses to the regression curve speak to the blunders of expectation. As we are getting to see Within the graph, the red purpose terribly is extremely is incredible} very preparing to the regression curve, Its Accuracy is best. Conversely, the yellow purpose is incredibly distant from the road. Then the curve and afterwards it..

**RANDOM FOREST REGRESSION**

Random Forest may be a refillable, easy to utilise machine learning algorithm that produces, even while not Hyper-parameter standardization, a consistent outcome tons of (usually) than not. It's likewise a standout amongst. The foremost utilized algorithms, because it's easy and should o.k. be utilized for every classification and regression Tasks. Random Forest might be a sort of supervised learning algorithm . As is clear from its name, it makes a forest and makes it by a random method of choice. The “forest” it assembles, is an outfit of call trees, Ready with the “bagging” strategy tons of (usually) than not. the general thought of the material technique is that a mix of learning models expands the general outcome. To state it's in easy words: It sorts of numerous call Trees and clubs all of them is along to induce tons of correct prediction. Random Forest (RF) may be a very fillable and straightforward to utilise Machine Learning (ML) calculation. It creates exceptionally precise outcomes Even while not the High degree of hyper-parameter standardization. Irregular Forest (RF) is additionally a standout Among the Foremost utilized Machine Learning (ML) calculations. this will get on the grounds that it's Extraordinarily basic and should likewise be utilized for every characterization and relapse tests. Random Forest as 2 Tree. Fundamentally there or 2 phases in Random Forest (RF) calculation. 1s tis irregular dry creation. The second is to play out AN expectation from the recently created Irregular woods' classifier. the entire procedure could also be given as: :i) Select every which way, “K” options from the general , “m” options. Here k << m. ii) Now Among these “K” options, mistreatment the only split purpose calculate the node, “d.” iii) Then split the node into its girl nodes by mistreatment the only form.

**LOGISTIC REGRESSION:**

Could be a classification rule acquainted with assign perceptions to a separate arrangement of Categories. Dissimilar to regression that yields persistent range qualities, logistical regression changes Its yield utilizing the logistical sigmoid capability to revive opportunity esteem which could then be ready to be Mapped to a minimum of two separate categories. Let's assume we're given info on understudy take a glance at Results and our objective is to foresee whether an understudy can pass or founder obsessed with the range of hours Dozed, and hours spent considering. We've 2 highlights (hours refreshed, hours contemplated) an2 classes: passed (1) and fizzed

.

Fig(a) Represents the range and curve of algorithm

**8. IMPLEMENTATION**

**CODING**

**Importing Dataset:**

**import Numpy as np**

**Import SeaBorn as sns**

**Import pandas as Pd**

**Import matplotlib.pyplot as plt**

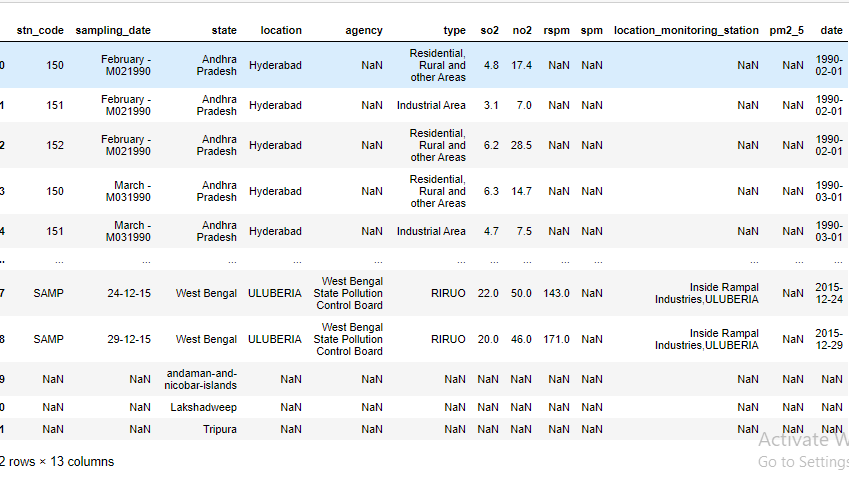
**%Matplotlib inline**

**Import warnings**

**warnings.filterwarnings ('ignore')**

**Load the dataset**

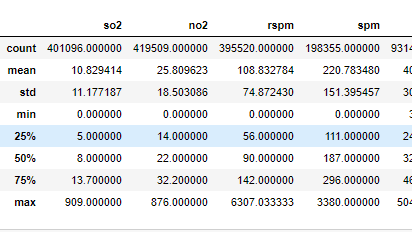
**Data = pd.read\_csv ('data.csv', encoding='cp1252')**

****

Fig(b) Displaying Loaded Dataframe

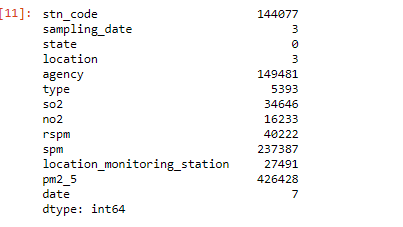
**Descriptive Analysis:**

DF. Describe ()

****

Fig(c) Statistical info about data

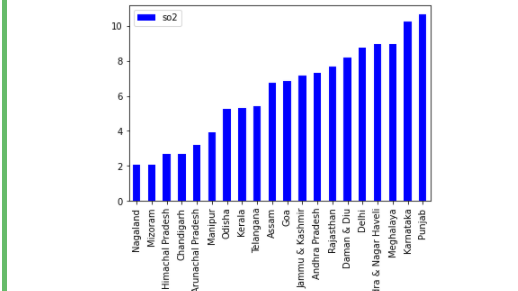
data.isnull ().sum ()



# Visualization for states with highest pollutants

**Data [['so2','state']].group by (["state"]).mean ().sort values (by='so2').head (20).plot. Bar (colour='b')**

**Plt. Show ()**

****

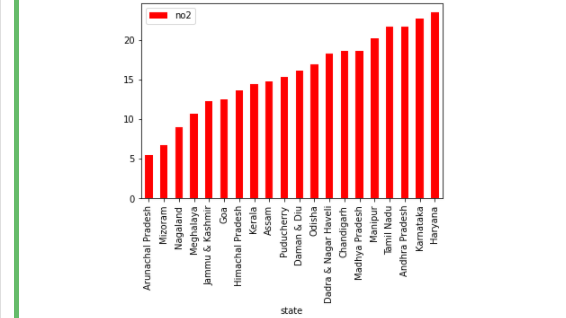
Fig(d) **states with highest so2 pollutants**

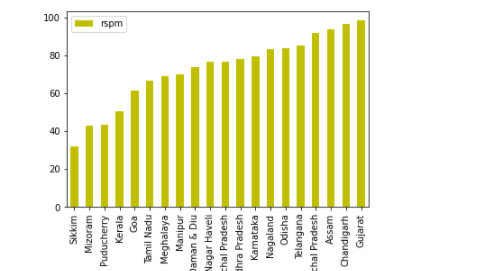
Data [['no2','state']].group by (["state"]).mean ().sort values (by='no2').head (20).plot. Bar (colour='r')

Plt show ()

Data [['rspm','state']].group by (["state"]).mean ().sort values (by='rspm').head (20).plot. Bar (colour=’y’)

Plt. Show()

****  Fig(e) **states with highest no2 pollutants**

****

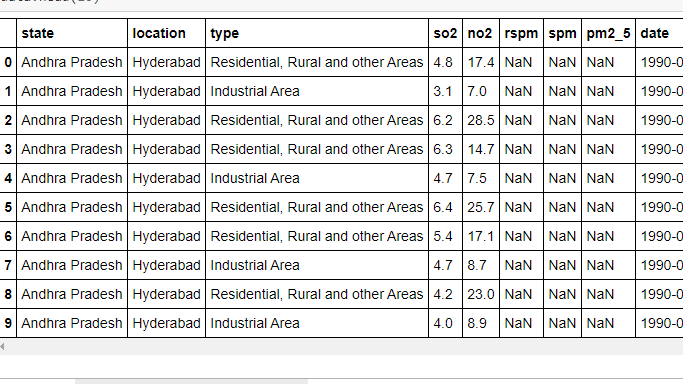
Fig(f) **states with highest rspm pollutants**

### 

### DROP UNNECESSARY COLUMN

Data. Drop (['stn\_code','agency','sampling\_date','location\_monitoring\_station'], axis=1, in place=**True**)

Data. Head (10)

****

Fig(g) **droping coloumns**

# FILL MISSING VALUES BY MEAN

# grp\_state = data.groupby ('state')

# def. impute\_mean\_by\_state (series):

# Return series.fillna (series. Mean ())

# Data ['rspm'] =grp\_state ['rspm'].transform (impute\_mean\_by\_state) #fill value with mean value group by state

# Data ['so2'] =grp\_state ['so2'].transform (impute\_mean\_by\_state)

# Data ['no2'] =grp\_state ['no2'].transform (impute\_mean\_by\_state)

# Data ['Spm'] =grp\_state['Spm'].transform(impute\_mean\_by\_state)

# Data ['pm2\_5']=grp\_state['pm2\_5'].transform(impute\_mean\_by\_state)

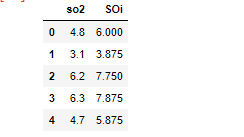
# data.describe()

# 

Fig(h) **Statistical info after Feature engineering**

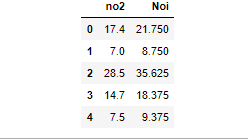
### CALCULATE AIR QUALITY INDEX FOR SO2 BASED ON FORMULA¶

|  |
| --- |
| Def cal\_SOi(so2): Si=0 if (so2≤40): Si= so2\*(50/40) elf (so2>40 and so2≤80): Si= 50+(so2-40)\*(50/40) elf (so2>80 and so2≤380): Si= 100+(so2-80)\*(100/300) elf (so2>380 and so2≤800): Si= 200+(so2-380)\*(100/420) elf (so2>800 and so2≤1600): Si= 300+(so2-800)\*(100/800) elf (so2>1600): SI= 400+(so2-1600)\*(100/800) return Si data['SOI']=data['so2'].apply(cal\_SOi) df= data[['so2','SOI']] df. Head() |



### CALCULATE AIR QUALITY INDEX FOR no2 BASED ON FORMULA

|  |
| --- |
| def cal\_Noi(no2): |
| ni=0 |
| if(no2<=40): |
| ni= no2\*50/40 |
| elif(no2>40 and no2<=80): |
| ni= 50+(no2-40)\*(50/40) |
| elif(no2>80 and no2<=180): |
| ni= 100+(no2-80)\*(100/100) |
| elif(no2>180 and no2<=280): |
| ni= 200+(no2-180)\*(100/100) |
| elif(no2>280 and no2<=400):ni= 300+(no2-280)\*(100/120) |
| else: |
| ni= 400+(no2-400)\*(100/120) |
| return ni |
| data['Noi']=data['no2'].apply(cal\_Noi) |
| df= data[['no2','Noi']]df.head() |



### CALCULATE AIR QUALITY INDEX FOR Spm BASED ON FORMULA

def. cal\_SPMi (Spm):

Sip=0

If (Spm<=50):

Spi=Spm\*50/50

Elif (Spm>50 and Spm<=100):

Spi=50+ (spm-50)\*(50/50)

Elif (Spm>100 and Spm<=250):

Spi= 100+ (spm-100)\*(100/150)

Elif (Spm>250 and Spm<=350):

Spi=200+ (spm-250)\*(100/100)

Elif (Spm>350 and spm<=430):

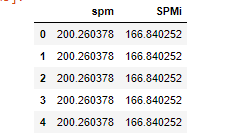
Spi=300+ (spm-350)\*(100/80)

Else:

Spi=400+ (spm-430)\*(100/430)

Return spi data ['Spm']=data['Spm'].apply(cal\_SPMi)df= data[['Spm’, ‘Spm']]

DF. Head ()



### The worst sub-index reflects overall AQI.If multiple pollutants are measured at a monitoring site, then the largest or "dominant" AQI value is reported for the location

def. cal\_aqi (si, Ni, rspmi, spmi): aql=0

If (si>ni and si>rspmi and si>spmi):

Aqi=si

if(ni>si and in>rspm and ni>spmi ):aqi=ni if(rspmi>si and rspmi>ni and rspmi>spmi ):

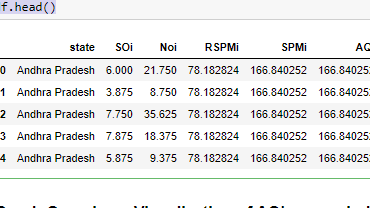
Aqi=rspmi

if (spmi>si and spmi>ni and spmi>rspmi):aqi=spmi

Return aqi

data ['AQI']=data. Apply(lambda x:cal\_aqi(x['SOi'],x['Noi'],x['RSPMi'],x['SPMi']),axis=1)

df= data[['state','SOi','Noi','RSPMi','SPMi','AQI']]



### AQI RANGE for corresponding AQI value

def. AQI\_Range(x):

If x<=50: return "Good"

Elif x>50 and x<=100:

Return "Moderate"

Elif x>100 and x<=200:

Return "Poor"

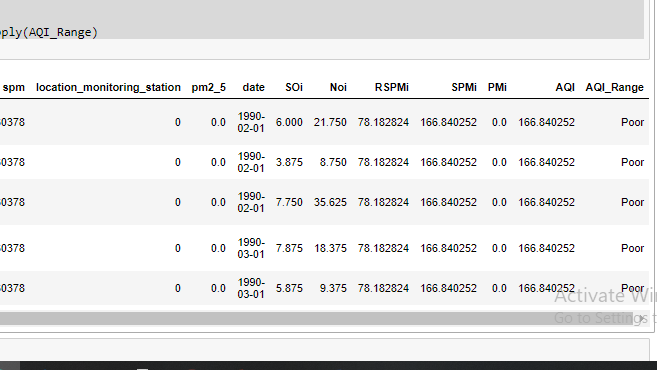
Elif x>200 and x<=300:

Return "Unhealthy"

Elif x>300 and x<=400:

Return "Very unhealthy"

elif x>400: return "Hazardous “data['AQI\_Range'] = data['AQI'] .apply(AQI\_Range)df. Head ()



#### **Train/Test Preparation:**

From sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y, test\_size=0.2,random\_state=101)

from sklearn.metrics import mean\_squared\_error

From sklearn.metrics import r2\_score, mean\_squared\_error

**Linear Model**

From sklearn.linear\_model import LinearRegression

From sklearn.model\_selection import KFold

From sklearn.model\_selection import cross\_val\_score

0 166.840252

1 166.840252

2 166.840252

3 166.840252

4 166.840252

Name: AQI, dtype: float64

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y, test size=0.2,random\_state=101)

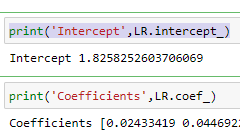
Model = LinearRegression ()

Model. Fit(X\_train, y\_train)

LinearRegression ()

print ('Intercept', LR.intercept\_)

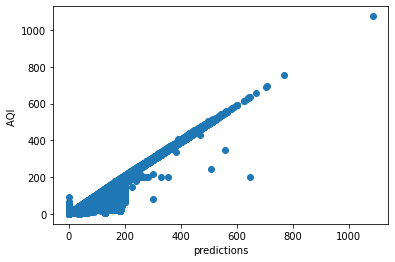
print ('Coefficients',LR.coef\_)



plt. Scatter (y\_test, predictions)

plt.xlabel ('predictions')

plt.ylabel ('AQI ')

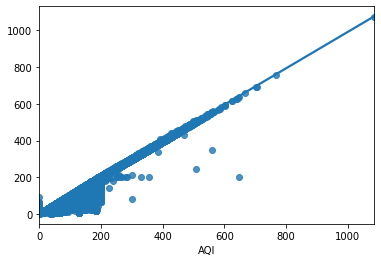


LR.score (X\_test, y\_test)

LR.predict ([[4.8, 21.75, 78.18,100]])

LR.predict ([[5.2, 7.625, 76.53,75.0]])

sns.regplot (y\_test, predictions)



print ('R^2\_Square:%.2f '% r2\_score(y\_test, predictions))

print('MSE:%.2f '% np.sqrt(mean\_squared\_error(y\_test, predictions)))

R^2\_Square:0.99

MSE:9.76

**Linear Regression Model 2**

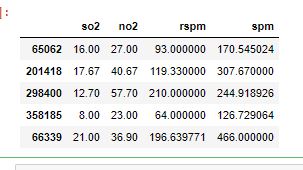
X1= data[['so2','no2','rspm','spm']]

y1 = data['AQI']

y.tail()

X\_train1, X\_test1, y\_train1, y\_test1 = train\_test\_split(X1,y1, test\_size=0.2,random\_state=101)

X\_train1.head()



LR1 = LinearRegression ()

LR1.fit (X\_train1, y\_train1)

prediction1 = LR1.predict (X\_test1)

Plt. Scatter (y\_test1, prediction1) #scatter plot for actual and predicted values

plt.xlabel ('Y Test')

plt.ylabel ('Predicted Y')

LR1.predict ([[9.1, 16.3, 67,179]])

y\_test1\_np= np.array (y\_test1)

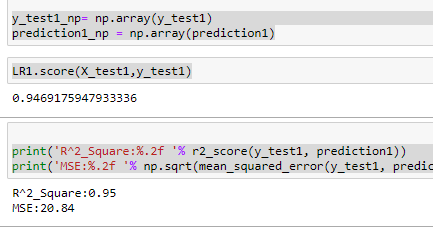
prediction1\_np = np.array (prediction1)

0.9469175947933336

LR1.score (X\_test1,y\_test1)

print ('R^2\_Square:%.2f '% r2\_score(y\_test1, prediction1))

print('MSE:%.2f '% np.sqrt(mean\_squared\_error(y\_test1, prediction1))



**CONCLUSION**

We can say that on the premise of supply, linear and R^2 worth for statistical regression

Model one is that the best, since its low values overallAQI is related with all the freelance

Variables (so2, no2, rspm, Spm) Loss information of knowledge of information whereas,

Modelling data, and low worth denotes less knowledge is lost overall. AQI is related with all

The freelance variables (so2, no2, rspm, Spm and pm2\_5).AQI has been increasing over the

Years. As for supply regression, solely model one provides North American nation with

Correct results additional thus as a result of AQI\_Range\_Binary is the variable quantity we

Tend to use. Except for model a pair of and model three variable quantity is sort label (sort

Of Ara) and also the accuracy results are relatively lower thus we are able to say that

Although the factors are associated with sort label during a manner, however there relations

Aren’t enough to be used as for prediction and estimation functions? As for vital variables

Involved we tend to cannot use regular regression coefficients and p-value to calculate a

Similar (Explained above). Our dataset additionally contains multiple regression, all the

Freelance Variables are somewhat associated with one another as we are able to see in our

Results. After closing that multiple regression do exist in our dataset, and once we try to take

Away extremely multiple regression variables the worth of R^2 drops, so creating North

American nation conclude that our dataset isn’t acceptable each statistical regression and

Supply regression (because assumptions for same are violated). After stepwise regression, we

Tend to conclude that the foremost vital variables that ought to be used for regression with

Our Variable quantity are 'so2', 'Spm,” 'no2,” 'pm2\_5' (with p-value <

zero.05). Regularization per Seas no result on the model, although there's a small increase

Within the accuracy and Cross\_val\_score however it's not that massive that we must always

**REFERENCES**

[1]Dragon, essayist Georgiana. “Air quality index prediction exploitation K-nearest neighbour technique no. one (2010): 103-108.

[2] Carbajal-Hernández, José Juan, "Assessment, and prediction of air quality exploitation formal logic and autoregressive models.” part surroundings sixty (2012): 37-50...

[3] Kumar, Unkinder and P. Goyal, “Forecasting of daily air quality index in Delhi”, Science of the entire surroundings 409, no. 24(2011): 5517 – 5523...

[4] International Journal of Applied Engineering analysis ISSN 0973-4562 Volume fourteen, Number 11, 2019 (Special Issue) © analysis Asian country Publications. http://www.ripublication.com

[5]https://www.researchgate.net/publication/292643438\_Air\_Quality\_Index\_\_A\_Comparative\_Study\_for\_Assessing\_the\_Status\_of\_Air\_Quality

[6] Hutchison Keith D., star Smith and Sharia J. Fauria, “Correlating MODIS aerosol optical thickness knowledge with ground-based PM2.5 observations across TX to be used in an exceedingly real time air quality prediction system, “Atmospheric surroundings thirty-nine.37 (2005):7190 – 7203.

 [7] V. M. Niharika and P. S. Rao, “A survey on air quality foretelling techniques, “International Journal of computing and data Technologies, vol. 5, no. 1, pp.103-107, 2014.

[8] D. J. Novak, D. E. Crane, and J. C. Stevens, “Air pollution removal by urban trees and shrubs within the, “Urban biology urban Greening, vol. 4, no. 3, pp. 115-123, 2014`. 6. T. Chiwewe and J. Distill, “Machine learns.

**BIODATA**



Name : N.VINOD KUMAR

Mobile Number : 7036620809

E-mail : vinodnagabairu.k@vitap.a.cin

Permanaent Address : Vijayawada

**NOTE:** Its **MANDATORY** for a student to attach all the PPT’s, Sample Materials, Specification Sheets, Programming Codes and a 5-10 minutes demo Video of the Project Digitally In CD . Stick the Compact Disk (CD) in the final page of the Thesis after binding it.