Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 3, July 2021:3939- 3953

Air Pollution and Temperature in the Prediction of Covid-19

E.Gothai[1*] , P.Natesan[2], R.R.Rajalaxmi[3], S.Sakti[4], S.Sasi[5], P.Soundararajan[6]

[1] Associate Professor, Department of CSE, Kongu Engineering College, Perundurai, Tamilnadu, India

[2,3]Professor. Department of CSE, Kongu Engineering College, Perundurai, Tamilnadu, India

[4,5,6]UG Scholars, Kongu Engineering College, Perundurai, Tamilnadu, India

*Email: egothai@kongu.ac.in (corresponding author),

natesanp@kongu.ac.in, rrr@kongu.ac.in

ABSTRACT

Air pollution is the major problem all around the world, it can harm our air track and cause irritation, shortness of breath, coughing, wheezing and put our life risk at lung cancer, heart attack, and respiratory diseases. Temperature also associates with the lung problems which induces the changes in the characteristics of covid-19. Covid-19 is an ongoing pandemic spreads across the world which affects 50 million and causes 1 million deaths. It is caused by the strain of severe acute respiratory syndrome (SARS). Air pollutants can contribute in the affecting, recovery and mortality rate of covid-19.Collecting data's of air pollutant's like CO, NO, $PM_{2.5}$, PM_{10}, AQI and corona virus affected, recovery and death cases from various resources.We are using the Jupiter lab platform in anaconda software. Two types of predictions are made by each of the models, such as the number of deaths, and the number of recovery cases. The outcomes delivered by the analysis shows it a promising instrument to utilize these strategies for the current situation of the Corona virus pandemic. The results conclude that the Random Forest Regression performs fine among all the used models followed by Linear Regression and LASSO which performs well in forecasting the death rate as well as recovery.

Keywords: Random Forest Regression; Linear Regression; LASSO - Least Absolute Shrinkage and Selection Operator.

1. INTRODUCTION

1.1 MACHINE LEARNING

Machine learning is the branch of artificial intelligence, where it has a capability to learn itself from the gained experience. It has a view on development of algorithms that can self access the data and it has ability of self learning. It meets almost all real world domains like health sector, autopilot (AV), robotics, gaming, business applications etc.,. One of the main areas of Machine Learning is forecasting the future from the past learned data's, for that many standard Machine learning algorithm is used. Then using those algorithms learnt from the past forecasting the condition and the actions need in the future need to be taken, applications areas including weather forecasting, disease forecasting, stock market forecasting using different regression models. There are many studies done for forecasting various diseases like heart related diseases, lung related diseases etc.,

Air pollution is one of the most harmful to human being and other living beings which affects the environment severely. Releasing the excess or unwanted gases like NO_2, CO_2, O_3, SO_2 to the atmosphere is called air pollution which increases the density and size of particulate matter. It causes serious problems mainly respiratory related problems like irritation, shortness of breath, asthma, lung cancer etc.,

Temperature of our environment has also chances of increase in transmission. As it associates with the lung related diseases like asthma, shortness of breath etc., when the temperature of the environment increases the transmission range of the may decrease and when the temperature decreases the transmission range may increase, where it is a lung related disease.

Corona virus is the disease spreads all over the world in the recent times rapidly. It was found identified in Wuhan, china on December 2019. It affects over 50 million people and cause death around 1.5 million till September 2020. It is found that these was similar to the SARS virus caused in 2003.Covid-19 is known to be caused by human-to-human transmission .India currently has the huge number of confirmed cases across Asian countries and has the 2_{nd} highest number of confirmed cases in the world. It causes many problems across the world and all the countries announce the quarantine and lockdown which decreases the global economy hardly. The symptoms of these disease are fever, dry cough, difficulty in breathing etc., where it directly

attacks the lungs. On average it takes about 5 to 7 days to react in the body. Where it directly affects the lungs there is a association between the air pollution and covid-19 recovery cases and death cases.

Spreading of virus increases day by day so we need to predict the situation in the future whether the recovery and death cases increase or decreases and plan accordingly. There are different climatic conditions and pollution across India, under such conditions the characteristics of covid-19 differs in transmission and recovery. By analyzing manually it takes more time duration, while in machine learning the time will less such that precaution will be done quickly. By using the pollution and temperature data's we predict the recovery and mortality rate for both polluted and non polluted cities around India. Where real word data or the raw data is incomplete, inaccurate.

Regression is a finding the correlation between variables and identify the continuous output variable on the one or more variables. It is a supervised learning technique. This technique of learning the task of functions that maps an input to the output based on example input and output pairs. It gains the information from labeled training data consisting of a set of training examples.

To contribute to the current human crisis our attempt in this problem is to develop a prediction system for COVID-19. The prediction is done for the two important variables of the disease for the coming days: 1) the number of mortality cases 2) the number of recoveries. This problem of prediction has been considered as a regression problem in this study, so the study is based on some state-of-art supervised Machine Learning regression models such as linear regression ,Random Forest regression etc.

It is already lending a hand in diverse situations in healthcare. ML in healthcare helps to research multiple of various data points and suggest outcomes, provides timely risk scores, precise resource allocation, and has many other applications. The Advantages of Machine Learning are Data Input from Unlimited Resources. Machine learning can easily consume unlimited amounts of data with timely analysis and assessment and Fast Processing and Real-Time Predictions.

II. LITERATURE REVIEW

In the study of Yan-Jun Guand et al[5], the time series analysis of covid-19 incidence in Wuhan and Xiaogan provided by CDC of Hubei province. They used data set details like Covid-19 positive cases confirmed from all laboratories and pollutants details. Then they have done analysis using the Graphpad prism. Then the descriptive analysis done to provide the overview of air quality and covid-19 incidence. They used linear regression model in this study. They conclude that both $PM_{2.5}$ and NO_2 associated with the increase of Covid-19.

In the study of Ying-Chieh Chen et al[2], the metrological conditions and air pollution impact on Covid-19. In this mainly focused on the metropolitan areas in Italy on Milan and Florence .They have collected the data's like daily maximum, average and minimum temperature, dew point, humidity, wind speed and atmospheric pressure. They have used the Gaussian mixture model for analyzing the impact of covid-19 at metrological conditions. For SARS transmission the particulate matter($PM_{2.5}$) positively correlated. From those research it is possible to find that the air conditioned environment has the risk of higher transmission.

In the study by Yisi Liu et al[7],this is about the study of traffic related air pollution in the transmission of covid-19 in the time period of time. Road side air pollution like hourly BC ,$PM_{2.5}$, NO, NO_2, CO for a particular time from the WAQA(Washington Air Quality Advisory) and WSDOE(Washington State Department of Ecology). To analysis the association between covid-19 and each pollutant level. MAR(Multi variate Auto Regressive) model is used in the study. The weekly traffic pollutants concentrations like UFTS, BC, NO, CO, NO_2 and metrological conditions like wind speed, wind direction and temperature are collected for past two weeks. They find that BC and $PM_{2.5}$ it seems to be increase the transmission of Covid-19.

In the study of Furqan Rustam et al[6],in this study they have done prediction and future forecasting using machine learning models like LR(Linear Regression), SVM(Support Vector Machine), ES(Exponential Smoothing) and LASSO (Least absolute Shrinkage and Selection Operator) used in the study of forecasting of covid-19. In this they have done predictions in three types in each model 1. Positive cases 2. Mortality rate 3. Recovery rate from this models they forecast for next 10 days. They collected the data set from GitHub. The data contains daily collected case reports, directions, death and recovery cases. They built a supervised learning

model, for predictions. They take corresponding dataset as input, they trained the regressive models and the predictions are generated by the training models like LR, LASSO, SVM, ES. They uses the metrics to evaluate the model by R2 score, Mean Square Error and Mean Absolute Error among this models ES performs better than other models then LR and LASSO are equally performed. Then SVM is performed least than any other model in the conditions. The r2score of LR is 0.83. The r2score of LASSO is 0.98. The r2 score of ES is 0.98. The r2score of SVM is 0.59.

In the study of Cindy Feng et al (2014) find out the impact of ambient fine particle matter ($PM_{2.5}$) exposure on the risk of influenza. They investigated the effect of PM2.5 by flu season across different age groups. PM2.5 had only minimum effect on influenza incidence across all age groups at the non flu season. They analyzed the data sources and description statistically and the data was collected from the surveillance system of government. They have used Gaussian regression model and linear regression.

Claire M. Midgley et al (2015) determined the seasonality of respiratory virus and the impact of increased molecular testing. Descriptive and quantitative analyses were conducted by statistical computing software. They assessed various data-smoothing techniques, including the use of moving averages or polynomial regression. We determined season characteristics nationally and by census region, including season onset, duration, peak, offset, and percentage of annual detections that occur within the season.

2.1 DATASET COLLECTION

Dataset have been collected from pollution data's like $PM_{2.5}$, PM_{10}, NO, NO_2, CO, AQI, and positive cases, recovery cases and mortality rate of covid-19 in the period of April to September of 2020 for the both polluted and non polluted cities across India. Totally 186 rows and 16 columns of data are used. The above dataset contains 32 districts air pollutant contents and covid-19 cases of India for 6 months. In our dataset there are three data type columns are object so we check the unique values for that.

Table 2.1 Dataset Collection

	City	Date	PM2.5	PM10	NO	NO2	NOx	CO	AQI	AQI_Bucket	High	Low	Average	cases	recovery	death
0	Ahmedabad	01-04-2020	24.79	86.06	4.50	21.29	13.70	4.50	153	Moderate	45	20	33	64	3	0
1	Ahmedabad	02-04-2020	25.52	88.12	4.46	24.24	15.15	4.46	164	Moderate	45	27	36	47	0	2
2	Ahmedabad	03-04-2020	33.48	87.96	4.51	41.44	23.58	4.51	173	Moderate	42	23	32	11	0	1
3	Ahmedabad	04-04-2020	39.44	84.83	3.82	39.88	22.28	3.82	182	Moderate	39	25	30	34	1	0
4	Ahmedabad	05-04-2020	43.77	87.96	3.35	40.99	22.49	3.35	152	Moderate	36	24	28	15	14	2
5	Ahmedabad	06-04-2020	46.64	95.30	3.75	29.60	17.24	3.75	145	Moderate	41	25	31	18	0	0
6	Ahmedabad	07-04-2020	45.80	104.50	3.87	16.80	11.04	3.87	159	Moderate	37	20	31	24	0	0
7	Ahmedabad	08-04-2020	36.30	90.07	5.19	15.65	11.51	5.19	183	Moderate	36	21	30	15	2	1
8	Ahmedabad	09-04-2020	32.12	89.74	5.07	21.61	14.33	5.07	168	Moderate	32	20	26	19	0	0
9	Ahmedabad	10-04-2020	30.73	78.56	4.36	28.65	17.23	0.59	113	Moderate	31	21	26	45	4	4
10	Ahmedabad	11-04-2020	30.18	74.99	3.06	34.81	19.26	0.45	89	Satisfactory	32	20	25	31	4	3
11	Ahmedabad	12-04-2020	26.70	72.83	2.72	27.00	15.18	0.36	101	Moderate	32	20	25	9	0	0
12	Ahmedabad	13-04-2020	18.77	61.54	2.81	23.01	13.01	0.39	95	Satisfactory	36	11	24	38	15	0
13	Ahmedabad	14-04-2020	31.16	72.09	3.22	25.90	14.73	0.37	89	Satisfactory	43	18	30	31	7	2
14	Ahmedabad	15-04-2020	23.69	64.81	2.28	17.18	10.04	0.43	88	Satisfactory	42	20	31	44	23	1

In the above the given Table 1, Totally 186 rows and 12 columns of data is collected.

Table 2.2 $PM_{2.5}$ levels

$PM_{2.5}$	**Condition**
0-60	GOOD
61-90	MODERATE
91-210	POOR
211-252	VERY POOR
253 & above	UNHEALTHY

Table 2.3 PM_{10} levels

PM_{10}	**Condition**
0-100	GOOD
101-150	MODERATE
151-350	POOR
351-420	VERY POOR
421 & above	UNHEALTHY

Table 2.4 NO_2 levels

Table 2.5 CO levels

NO_2	Condition
0-42	GOOD
43-94	MODERATE
95-295	POOR
296-667	VERY POOR
668 & above	UNHEALTHY

CO	Condition
0-1.7	GOOD
1.8-10.3	MODERATE
10.3-14.7	POOR
14.8-30.2	VERY POOR
30.3 & above	UNHEALTHY

Table 1.5 AQI levels

AQI	Condition
0-100	GOOD
101-200	MODERATE
201-300	POOR
301-400	VERY POOR
401& above	UNHEALTHY

The above tables 2.1, 2.2, 2.3, 2.4 and 2.5 contains the air pollutants and its levels. By these levels we know the conditions of $PM_{2.5,}$ PM_{10}, NO_2, CO, AQI levels like good, moderate, poor , very poor and unhealthy. For example in table 1.1 the city Chennai on April month the levels of $PM_{2.5}$, PM_{10}, NO_2, CO, AQI levels are 122.88, 168.34, 20.24, 1.1 ,230 which are very poor, poor, good , good and poor where the average level of air pollution on April month in Chennai was poor condition. These tables show the levels of air pollutants.

2.3 EXPERIMENTAL SETUP

For our study we implement in machine learning by using python. We use Jupiter note book in Anaconda navigator in windows 10.

III. METHODOLOGY

Data collection

Initially we collected the pollution data's ($PM_{2.5}$, PM_{10}, NO, NO_2, CO, AQI) then we collected the covid-19 dataset (positive, recovery and death cases) and atmospheric temperature of the cities for the month is collected from the sources then it was merged into single CSV file format.

Data preprocessing

By reading the CSV file we started the preprocessing and data Visualization by analyzing shape of the dataset. Then we check the null value and data type from the collected dataset. We check the unique value for object data type. We analyze the covid-19 data and air pollution. We use category encoding to encode city name and drop the unwanted column and fix the final dataset. Then we change the date into date and time format and then we encode it into a ordinal integer by using to-ordinal. We check the correlation between the data's. After analyzing it we plot final the correlation graph of the dataset. We use SK learn train split model is used to split the data set for the training and testing.

Regression model

We use regression models like Linear Regression, Random Forest Regression, Lasso to train our dataset. For 811 data's in each model there are two types predictions are made 1.Recovery rate of covid-19 2.Mortality rate of covid-19. We evaluate the model by the Metrics like R^2 score, mean square error and mean absolute error. Now we can able to visualize the best model among these metrics.

R_2Score=1-(sum(actual-predicted)^^2)/sum((actual-mean(actual))^^2))

$$\mathbf{MSE}=\frac{\mathbf{1}}{\mathbf{N}}\sum^{\mathbf{N}}(\mathbf{Y}\mathrm{i}-\mathrm{Yi})^{\mathbf{2}}$$

MSE = Mean Square Error

N = Number of samples

Y_i = Predicted value

Y_i^2 = Actual value

3.1 IMPLEMENTATION METHODOLOG

The pollution and the covid–19 data is collected and convert it into CSV file. Then started preprocessing. First, check the shape of the data set and then info to clarify the data types available in datasets. In our dataset there are three data type columns are object so we check the unique values for that. Then we check the any null value in our dataset. First we encode the city name using category encoding. Then changed the date to-ordinal using to-ordinal. In our dataset there is no active cases column, so we calculate the active cases. Then we set the final dataset as $PM_{2.5}$, PM_{10}, NO, CO, AQI, cases, recovery and death. After completing the preprocessing for our dataset the splitting of data to train and test the algorithm is done by train split model. We split our data in the ratio of 70:30. In this 70% is for training and remaining 30% for testing our dataset. Then the training and testing is done by regression models to predict the mortality rate and recovery rate. Then we find the best one among this regression models. In this we done two thing first we train using pollution and evaluate accuracy for trained models linear regression and random forest regression. Then we train using pollution and atmospheric temperature (high, low, average) and we evaluate accuracy for trained models linear regression, random forest regression, lasso and then we done hyper parameter tuning for developed model using grid search CV and randomized search CV and then we evaluate the tuned accuracy for the models.

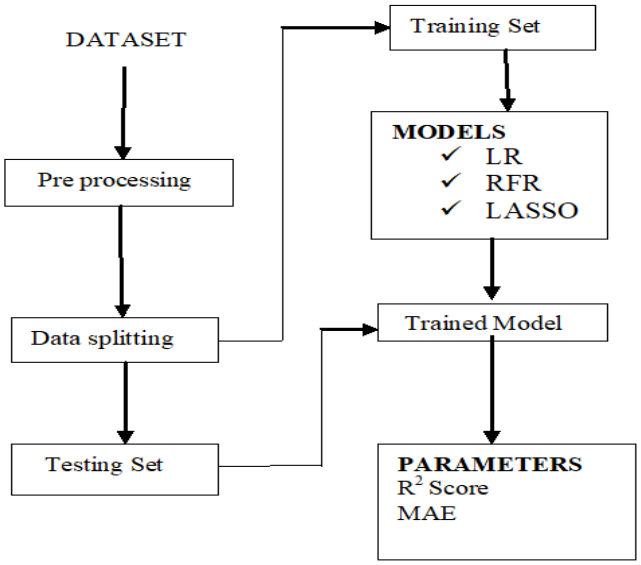

Figure 1.1Work flow

IV. RESULTS AND DISCUSSION

In this paper many regression models like Linear Regression, Random Forest Regression for predicting the mortality rate and recovery rate of covid-19 by using the pollution dataset. We test and train about 811 data's of sample. We used R^2 Score metrics for evaluating the model. The R^2 score was calculated for three models. Random forest regression gives high accuracy. Followed by Linear Regression. In linear Regression for mortality rate prediction it gives 71.19% accuracy for recovery rate it gives 69.19% accuracy. Random forest regression it gives 79.29% for mortality rate and 80.86% accuracy for recovery rate. Among these algorithm Random Forest Regression gives low MAE value.

Table4.1 R^2 Score for various Regression models of mortality rate

Regression models	R^2 Score
Linear Regression	0.7119
Random Forest Regression	0.7929

Table 4.2 R^2 Score for various Regression models of Recovery rate

REGRESSION MODELS	R^2 Score
Linear Regression	0.772
Random Forest Regression	0.8156
LASSO Regression	0.7568

REGRESSION MODELS R^2 SCORE
MODELS
Linear Regression
Random Forest
Regression
0.772
0.6919
0.8156
Linear Regression
Random Forest
0.8086
LASSO
Regression
0.7568

Table 4.3 R^2 Score for various Regression models of recovery rate with atmospheric temperature

Temperature and pollution is used to predict covid-19 mortality and recovery rate using random forest regression, linear regression and lasso among this random forest regression performed well followed by linear regression and lasso. After adding atmospheric temperature the accuracy of random forest for predicting mortality is 0.8584 then for recovery is 0.8156 for linear regression accuracy for mortality rate is 0.8073 for recovery is 0.772. And for lasso accuracy rate for mortality is 0.8095 and for recovery is 0.7568.

Hyper parameter tuning is done for these models for linear regression we use normalization as hyper parameter. For random forest regression we use n estimators, max features. For lasso we use alpha and normalization as hyper parameter. In this we use two types of cross validation for hyper parameter tuning. Grid search CV and randomized search CV among this for our project grid search CV performance well for tuning a model performance. For mortality grid search CV tuned linear regression gives 0.08149,randomized search CV gives 0.8079.for recovery grid

search CV tuned random forest algorithm provides 0.7808,randomized search CV provides 0.7780. For mortality grid search CV tuned random forest regression gives 0.8805,randomized search CV gives 0.8608.for recovery grid search CV tuned random forest regression algorithm provides 0.8422,randomized search CV provides 0.0.8347. For mortality grid search CV tuned LASSO gives 0.8113, randomized search CV gives 0.8077.for recovery grid search CV tuned random forest regression algorithm provides 0.7709,randomized search CV provides 0.7699. Among these random forest regression performs well for well it retrieves low mean absolute error then linear regression and lasso.

Table 4.4 Tuned accuracy for mortality

Regression Models	Hyper parameter tuning with grid search CV	Hyper parameter tuning with random search CV
Linear Regression	0.8149	0.8079
Random Forest Regression	0.8805	0.8608
LASSO	0.8113	0.8077

Table 4.5 Tuned accuracy for recovery

Regression Models	Hyper parameter tuning with grid search CV	Hyper parameter tuning with random search CV
Linear Regression	0.7808	0.778
Random Forest Regression	0.8422	0.8347
LASSO	0.7709	0.7669

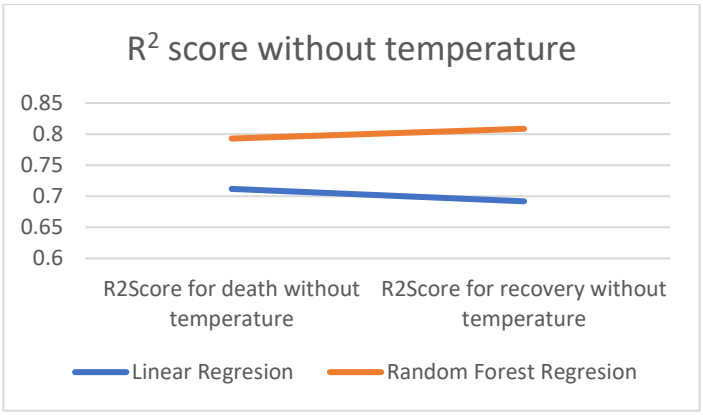

Figure 4.1 R^2 Score of regression algorithms without temperature

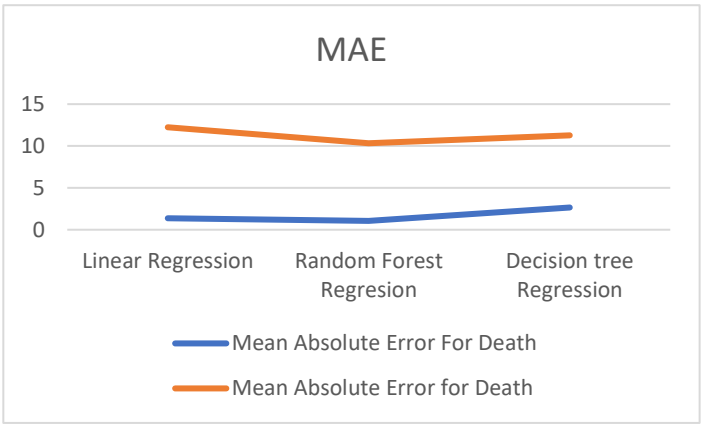

Figure 4.2 illustrate the R^2 Score of regression algorithms without temperature

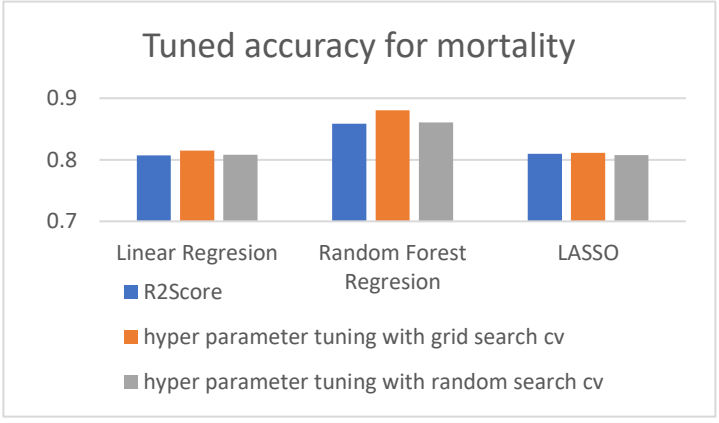

Figure 4.3 Tuned accuracy for mortality

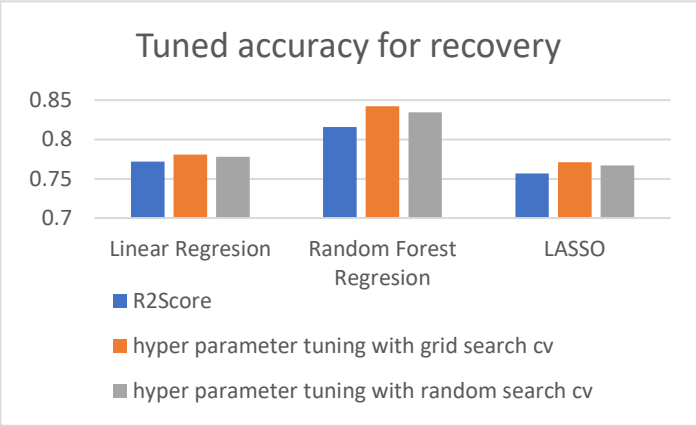

Figure 4.4 Tuned accuracy for recovery

V. CONCLUSION

Air pollution contributes the major problem in recovery rate and mortality rate of covid-19. The dataset is collected from the various sources and preprocessing is done. The dataset is created from the different sources and pre processing is done, then splitting the dataset. Then train and test the dataset using algorithms. We are using Random Forest Regression, Linear Regression, using these algorithms we are predicting and comparing the accuracy. Using these algorithms we are comparing the results in the graph. So we built a model to predict the recovery rate and mortality rate using the positive cases and air pollution data's. In the proposed work, we have used many models to find accurate one. We have used regression models in that Random Forest Regression algorithm gives the better accuracy rates. Then we used temperature for predicting and evaluate model by using R^2 score, MAE, MSE. Then we done hyper parameter tuning for this algorithm.

References

1. [1] Casanova, L. M., Rutala, J. S., Weber, W. A. &Sobsey, M. D. Efects of air temperature and relative humidity on coronavirus survival on surfaces. Appl. Environ. Microbiol. 76, 2712–2717 (2020).
2. [2] Pani, S. K., Lin, N.-H. & Babu, S. R. Association of COVID-19 pandemic with meteorological parameters over Singapore. Sci. Total Environ. 740, 2 (2020).

3. [3] Xie, J. & Zhu, Y. Association between ambient temperature and COVID-19 infection in 122 cities from China. Sci. Total Environ. 724, 138201 (2020).
4. [4] Wu, X., Nethery, R. C., Sabath, B. M., Braun, D. &Dominici, F. Exposure to air pollution and COVID-19 mortality in the United States. MedRxiv https://doi.org/10.1101/2020.04.05.20054502 (2020).
5. [5] Fattorini, D. & Regoli, F. Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy. Environ. Pollut. 2, 114732 (2020).
6. [6] Chen, K., Wang, M., Huang, C., Kinney, P.L., Anastas, P.T., 2020. Air pollution reduction and mortality benefit during the COVID-19 outbreak in China. Lancet Planet. Health 4,e210–e212.
7. [7] Clement J, Dumoulin B, Gubbelmans R, Hendriks S, van de Woestijne KP. Reference values of total respiratory resistance and reactance between 4 and 26 Hz in children and adolescents aged 4–20 years.
8. [8] Neuberger M, Moshammer H, Kundi M. Declining ambient air pollution and lung function improvement in Austrian children. Atmospheric Environment, 2002.