PREDICTING THE IMPACT OF AIR POLLUTANTS ON RICE AND WHEAT USING MACHINE LEARNING

[1]Ayush Raj Singh, [2]Ian Sequeira, [3]Daksh Ramchandani,[4]Mirudhula Nadar, [5]Sunita Sahu
[1, 2, 3, 4] Student, [5] Assistant Professor
Department of Computer Engineering,
Vivekanand Education Society's Institute of Technology, Chembur, Mumbai-400074, India

Abstract: Agricultural crop production and quality can deteriorate when exposed to high concentrations of various air pollutants. Deterioration can range from visible markings on the foliage, to reduced growth and yield, to the premature death of the plant. Accurate prediction of crop development stages plays an important role in crop production management. Such predictions will also support the allied industries in strategizing the logistics of their business. An increase of atmospheric CO_2 works as carbon fertilizer, improves plant growth and productivity of crops and negatively impacts the nutrients such as iron, zinc and crude protein contents in the grains. In this paper, we are proposing a web-based system for predicting the impact of air pollutants on crop quality and production using machine learning tools.

We are considering parameters like SO_2, NO_X, Suspended Particulate Matter (SPM). These factors will help find the impact and complications of air pollutants present in our environment on crop yield. We are using Gradient Boosted Regressor for our work.

Index Terms: **Air Pollutants, Crop Production, Air Quality Index (AQI), Gradient Boosted Regressor, Random Forest, Linear Regression, Statistical Analysis.**

I. INTRODUCTION

Since the start of the industrial revolution in the 19th-century environmental pollution has grown into a global transboundary problem that affects air, water, soil and ecosystems, and is linked directly to human health and well-being. Pollution is linked to three main human activities: fossil-fuel combustion, primarily by industry and transport; the application of synthetic fertilizers and pesticides in agriculture; and the growing use and complexity of chemicals [11].

Air pollution trends over Indian megacities and their local-to-global implications were observed. Air pollutants like SO_2, NO_X, Suspended Particulate Matter (SPM), Respirable suspended particulate matter (RSPM) in three megacities of India: Delhi, Mumbai and Kolkata. Statistical analysis shows the gradually declining concentration trend of SO_2. However, between 1992 and 1994, the concentration has increased because of growing emissions from power plants and the transport sector. It then proceeds to illustrate that there are large variations in ambient NO_X concentration after 2000 (with two peaks during 2002 and 2009 and a sharp dip in 2005) in Kolkata. This could reflect the unavailability of appropriate monitoring data and use of high sulfur diesel as fuel, older vehicles, and unavailability of clean coal. The figure shows that in Delhi, the concentration of NO_X gradually increased by 164%, i.e., 523 from 22 µgm-3 in 1991 to 58 µgm-3 in 2012. Finally, the literature shows that there were large variations in SPM concentration before 2000 (with two peaks during 1993 and 1996 and a sharp dip in 1997) in Kolkata. This, again, could reflect the unavailability of adequate monitoring data and usage of high sulfur diesel as fuel, older vehicles, and the unavailability of clean coal. [7]

Air pollutants have a negative impact on plant growth, primarily through interfering with resource accumulation. Once leaves are in close contact with the atmosphere, many air pollutants, such as O_3, SO_2 and NO_X, affect the metabolic function of the leaves and interfere with net carbon fixation by the plant canopy. Air pollutants that are first deposited on the soil, such as heavy metals, first affect the functioning of roots and interfere with soil resource capture by the plant.

Based on our collected dataset, we plotted the variation of air pollutants.

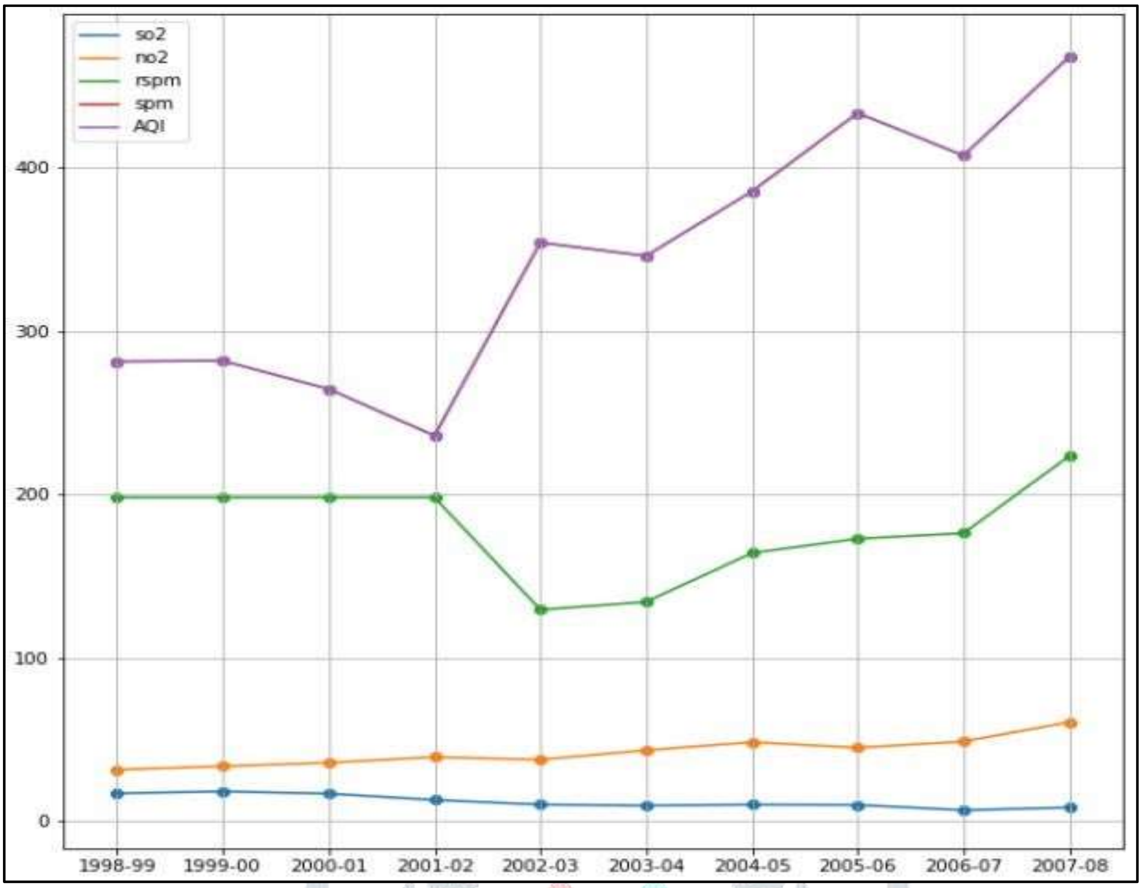


Fig. 1 Air Pollutants Variation

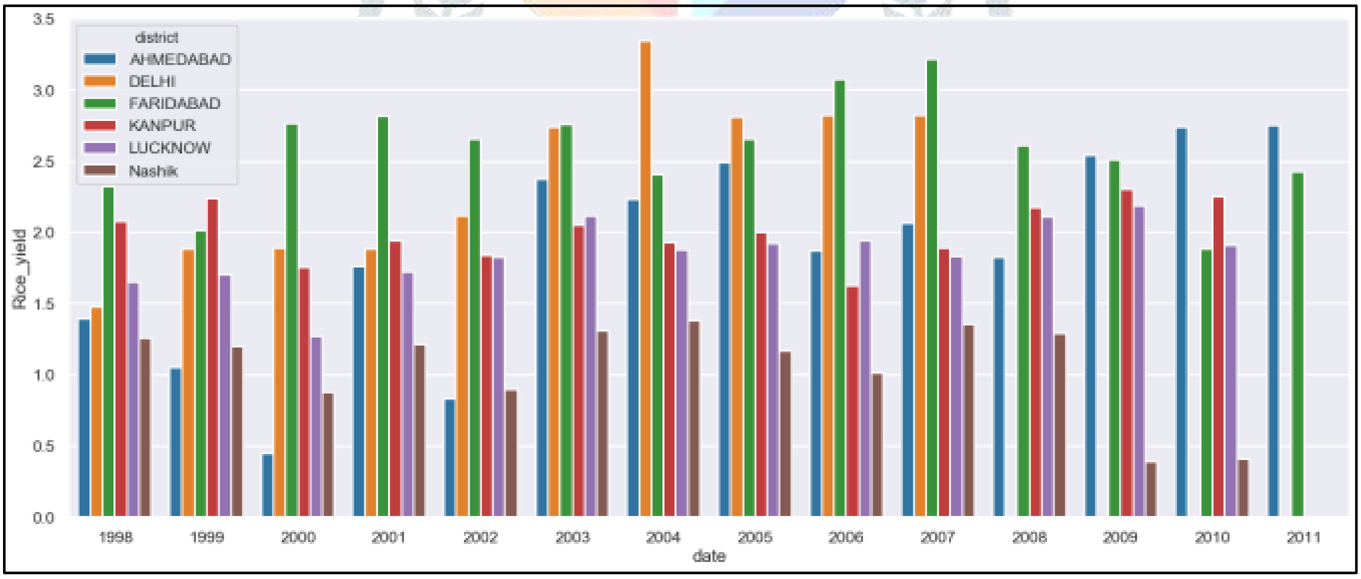


Fig. 2(a) Rice Yield Variation

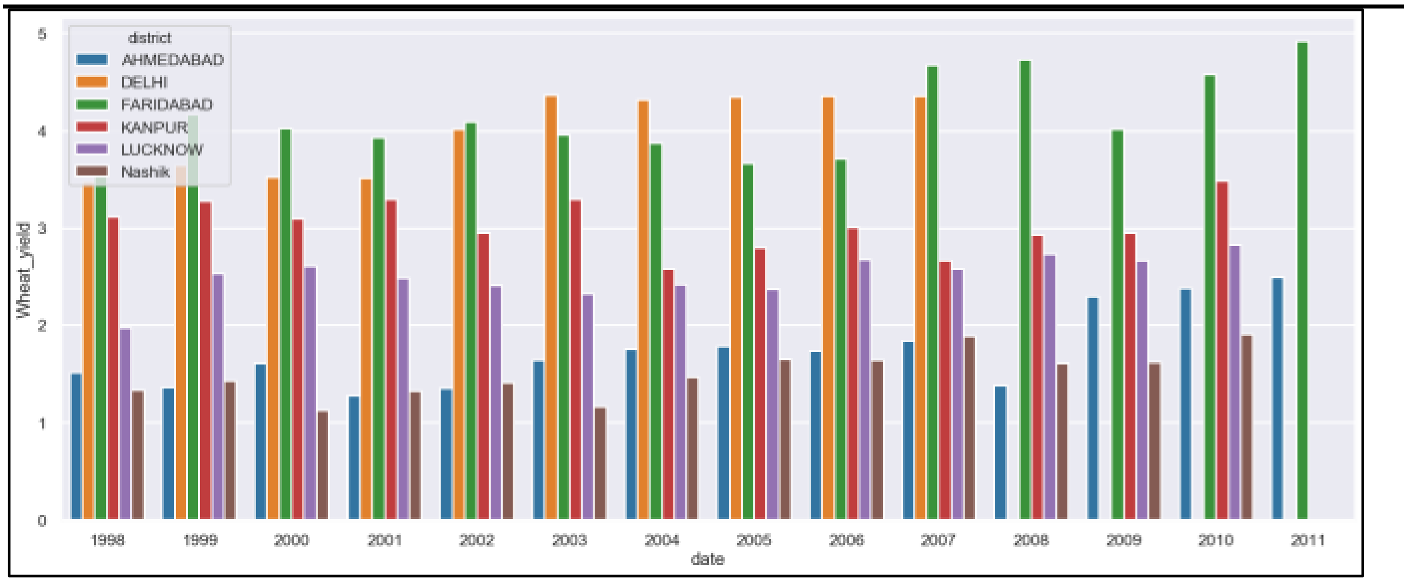


Fig. 2(b) Wheat Yield Variation

Figure 1 depicts the variation in yield of rice when it is affected by SO_2, NO_2, RSPM, and SPM, over 10 years from 1998-99 to 2007-08. As the trends suggest, the Air Quality Index started to rise significantly between 2001-02 and 2002-03. Figure 2(a) & Fig. 2(b) depict the growth of Rice and Wheat respectively, over time in different districts of India.

Through our research, we realized that models using a correlation between the aforementioned pollutants and the yield of crops (Rice & Wheat) are yet to be developed and implemented. Hence, we obtained the datasets with specifically these features and carried out exploratory analysis on them.

II. Literature Review

Prediction of Agriculture considering the impact of air pollutants is an essential task for the decision-makers at national and regional levels for rapid decision- making. An accurate Crop Yield Prediction model can help farmers to decide on what kind of precautions need to be taken while growing crops. There are different approaches to Agriculture prediction. This review article has investigated what has been done on the impact of air pollutants on agriculture in the literature.

Ghude Sachin D., et al. in [1] has proposed a system to calculate yield reduction and crop production losses based on AOT40 exposure metrics and its concentration-response (CR) relationships for predicting the yield reduction of a selected crop at different ozone exposure levels. Results of this calculation showed the significant yield loss on the crops like rice, wheat, cotton and soybean.

In [2],authors Yadav Achchhelal et al. have conducted research on an area of north India to understand the impact of elevated O_3 and CO_2 exposure on wheat for that they divided the region in four parts with different exposure conditions viz. ambient conditions, elevated CO_2, elevated O_3 and combined effects of elevated CO_2 and O_3. They have concluded that the crops reacted in a very similar manner but the extents of responses or sensitivity of every wheat cultivar were different with the change in climate and cultivars.

Reference [3] states that the authors were considering the expanding air contaminants inside the past years which drove them to lead the examination on the effect of air pollutants on farming yields. Air contaminants like sulphur dioxide (SO2), nitrogen oxides (NO_X), ozone (O_3), fluorine (F) and suspended particulate matter (SPM) makes direct harm to the harvests like noticeable markings on the foliage, decreased development and yield and premature death of the plant and furthermore the joined impacts of those pollutants is generally more prominent than the amount of their individual impacts.

Authors Auffhammer et al. in paper [4] have extended the study of the consequences of greenhouse gases. In previous studies, they found that atmospheric brown clouds are partially the offset of the warming effects of greenhouse gases. They have studied the statistical model of historical rice harvests in India when coupled with regional climate scenarios from a parallel climate model indicating the joint reductions in brown clouds and greenhouse gases would have negative impacts on harvests that also contributed to the slowdown in harvest growth that occurred during the past twenty years.

Reference [5] states that the authors have observed the pattern of air pollutant emissions of the last two decades in rural and more remote areas. Authors have conducted an on-field experiment on crops that have shown deleterious effects on the physiology and metabolism of plants. Responses of air pollutants on plants vary between different species and their cultivars; it also depends on the type of pollutant, concentration, duration and magnitude. They have suggested screening out sensitive and tolerant cultivars in India and establishing the exposure indices of all the important crops to reduce crop loss.

Reference [6] contains a study on artificial neural networks (ANN) and decided to check its ability of approximation and prediction by making a prediction model. They made a model to understand which crop is best suited to a given area considering parameters like soil type, pH of soil, temperature, rainfall and mineral contents in soil.

III. Proposed System

In this paper, we are proposing a web-based system for predicting an approximate estimation of the yield for a crop type in a location, based primarily on air-pollution factors but also rainfall, soil type, etc.

The proposed system predicts the yield of the crop for a particular land, based on emission of air pollutants like SO_2, NO_X, Suspended Particulate Matter (SPM), Ozone emission and the Air Quality Index (AQI) for that area.

Our web application is developed using Flask, HTML, CSS and JavaScript.

Our project's block diagram (Fig. 3) can be bifurcated into 3 levels:

1. Preprocessing the data using 10-Fold Cross Evaluation.
2. Hyperparameter optimization using 10-Fold Cross Evaluation for the following parameters:
   2.1. n_estimators (No.of trees).
   2.2. learning_rate (Scaling of each residual tree).
   2.3. max_depth (Maximum depth of the individual residual trees).
   2.4. alpha (Parameter in Huber Loss function).
   2.5. min_sample_leaf_split (Minimum no. of samples to split the leaf).
3. Getting user input from the web application and returning a prediction of:
   3.1. Yield Losses due to gases with highest feature importance.
   3.2. Approximation of the yield for the specified area along with a mean squared error.

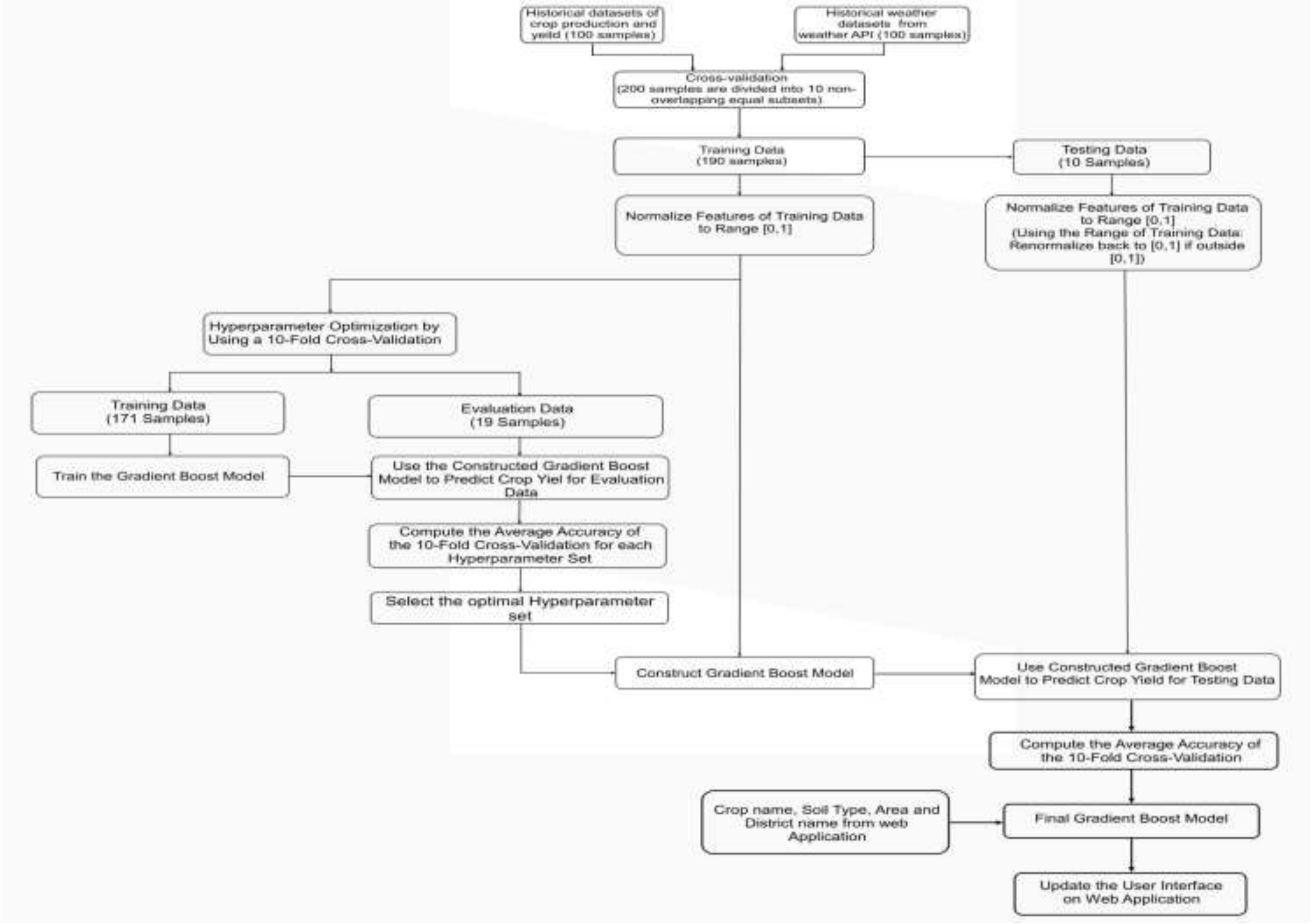
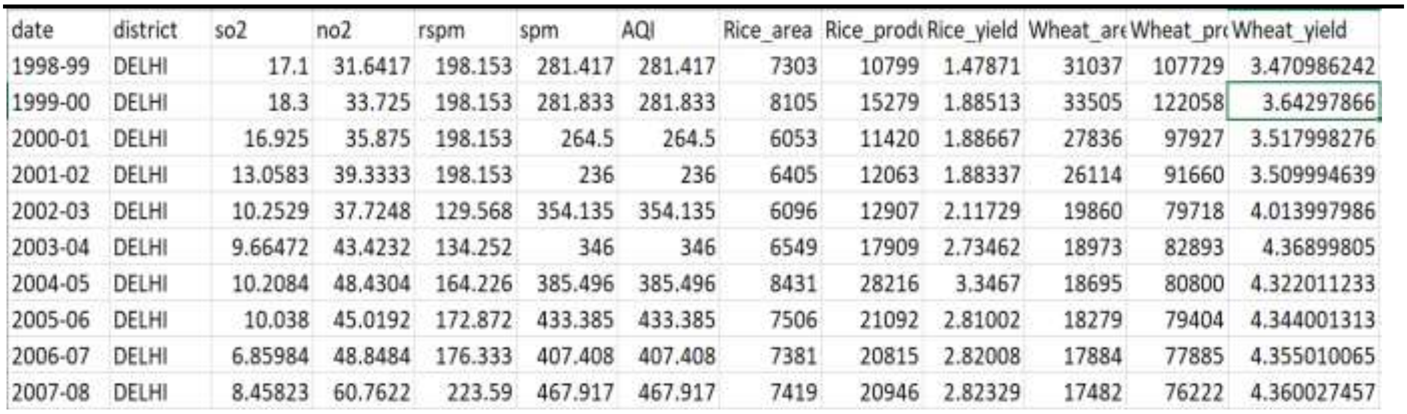

Historical datasets of crop production and yeild (100 samples)
Historical weather datasets from weather API (100 samples)
Cross-validation (200 samples are divided into 10 non-overlapping equal subsets)
Training Data (190 samples)
Testing Data (10 Samples)
Normalize Features of Training Data to Range [0,1]
Normalize Features of Training Data to Range [0,1] (Using the Range of Training Data: Renormalize back to [0,1] if outside [0,1])
Hyperparameter Optimization by Using a 10-Fold Cross-Validation
Training Data (171 Samples)
Evaluation Data (19 Samples)
Train the Gradient Boost Model
Use the Constructed Gradient Boost Model to Predict Crop Yiel for Evaluation Data
Compute the Average Accuracy of the 10-Fold Cross-Validation for each Hyperparameter Set
Select the optimal Hyperparameter set
Construct Gradient Boost Model
Use Constructed Gradient Boost Model to Predict Crop Yield for Testing Data
Compute the Average Accuracy of the 10-Fold Cross-Validation
Crop name, Soil Type, Area and District name from web Application
Final Gradient Boost Model
Update the User Interface on Web Application

Fig. 3 Block Diagram

a) Dataset used

Our dataset consists of the pre-processed datasets of the Air Quality and Crop Yield/Production for certain states of India. For air quality, we have considered the features like Sulphur-dioxide, Nitrogen-dioxide,, Suspended Particulate Matter and Respirable Suspended Particulate Matter. (Air Quality Index is dropped as it has a high correlation with SPM)



date	district	so2	no2	rspm	spm	AQI	Rice_area	Rice_prod	Rice_yield	Wheat_are	Wheat_pr	Wheat_yield
1998-99	DELHI	17.1	31.6417	198.153	281.417	281.417	7303	10799	1.47871	31037	107729	3.470986242
1999-00	DELHI	18.3	33.725	198.153	281.833	281.833	8105	15279	1.88513	33505	122058	3.64297866
2000-01	DELHI	16.925	35.875	198.153	264.5	264.5	6053	11420	1.88667	27836	97927	3.517998276
2001-02	DELHI	13.0583	39.3333	198.153	236	236	6405	12063	1.88337	26114	91660	3.509994639
2002-03	DELHI	10.2529	37.7248	129.568	354.135	354.135	6096	12907	2.11729	19860	79718	4.013997986
2003-04	DELHI	9.66472	43.4232	134.252	346	346	6549	17909	2.73462	18973	82893	4.36899805
2004-05	DELHI	10.2084	48.4304	164.226	385.496	385.496	8431	28216	3.3467	18695	80800	4.322011233
2005-06	DELHI	10.038	45.0192	172.872	433.385	433.385	7506	21092	2.81002	18279	79404	4.344001313
2006-07	DELHI	6.85984	48.8484	176.333	407.408	407.408	7381	20815	2.82008	17884	77885	4.355010065
2007-08	DELHI	8.45823	60.7622	223.59	467.917	467.917	7419	20946	2.82329	17482	76222	4.360027457

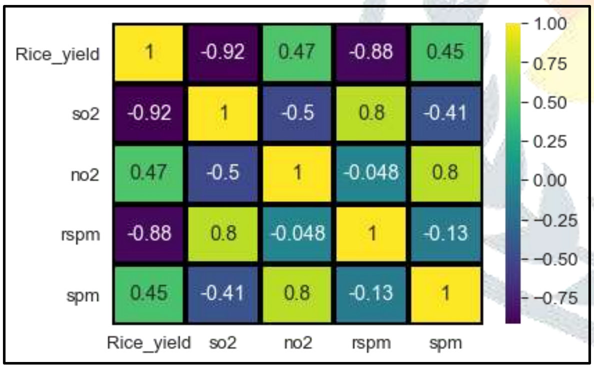
Fig 4: Snapshot of the Delhi dataset used for analysis and training of the model

For the prediction model, we have considered the crops majorly consumed in India, like, Rice and Wheat. The weather dataset from weather API. It allowed us to query the AQICN's database for the user-queried State and return the current weather data in the region which is used as the input to Gradient Boosted Regressor Model.

In Figure 5(a), from the given heatmap we can see a good negative correlation between Rice Yield and Sulphur-dioxide (-0.92) and also between Rice Yield and Respirable Suspended Particulate Matter (-0.88), hence, these features will play an important role in the prediction of yield losses in Rice due to SO2 and RSPM.

In Figure 5(b), from the given heatmap we can see a good negative correlation between Wheat Yield and Sulphur-dioxide (-0.89) and also between Wheat Yield and Respirable Suspended Particulate Matter (-0.89), hence, these features will also play an important role in the prediction of yield losses in Wheat due to SO2 and RSPM.

In Figure 6, 7, 8 and 9, we can see a regression plot for the aforementioned features in accordance with their correlation along with the Kernel Density Estimation plot and histograms for the individual features.



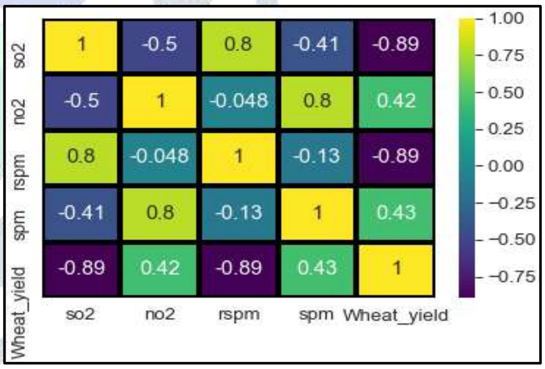


Fig. 5(a): Correlation Coefficient with Yield of Rice Fig. 5(b): Correlation Coefficient with Yield of Wheat

Figure 5(a) and Fig. 5(b) depict the relationship between every feature of our air pollutant dataset (SO_2, NO_2, RSPM & SPM).

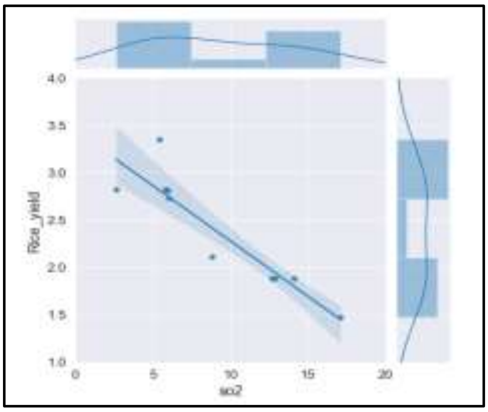


Fig. 6: Rice Yield Correlation with SO_2

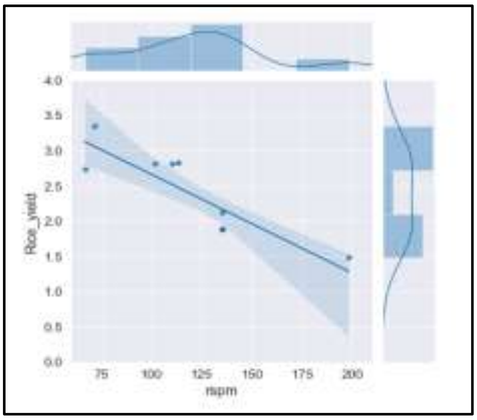


Fig. 7: Rice Yield Correlation with RSPM

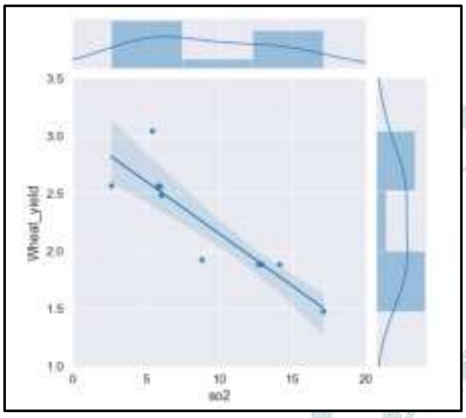


Fig. 8: Wheat Yield Correlation with SO_2

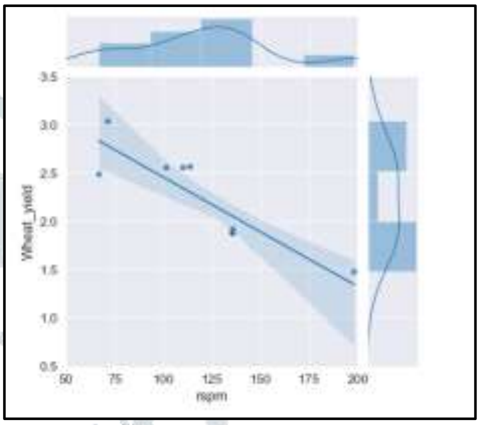


Fig. 9: Wheat Yield Correlation with RSPM

Figure 8 & Fig. 9 depict the correlation of Wheat yield under the influence of SO_2 & RSPM respectively.

3.2 Data Cleaning

This is regarded as one of the most important steps as it directly affects the accuracy of the model. For Air Quality, there were features like SPM, NO_2 that showed lesser correlation than those of SO_2 and RSPM hence those features were dropped and the machine learning model accuracy increased to 95%. For Crops, the features of area and production over a period of 10 years were considered in the model as they also directly affect the yield prediction. All null values and unnecessary columns are dropped and this dataset is split into training sets, testing sets and validation sets using 10-Fold Cross Validation.

The final pre-processed dataset was used to create a hyper parameterized model which is in turn stored in a pickle file to be deployed in our application.

3.3 User Input

Our web application require user to enter the State, Crop and Area, as an input, and it fetches the current levels of NO_2, SPM, RSPM, SO_2 for the requested state from the weather database which in turn gives us an input to our prediction model to calculate the effect of individual gases with the yield losses of the most important features along with the prediction for the amount of crop production for the required area.

3.4 Machine Learning Model

Gradient Boosted Regression is the model we chose for our proposed system as it can optimize on different loss functions and provides several hyperparameter tuning options that make the function fit very flexible, the model also handles missing data (imputation not required) and lastly, it provided us with the highest accuracy among the models we considered (Multiple Linear Regression and Random Forest Regressor).

IV. Methodology Employed

4.1 Algorithm Used

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

Weak learners are decision trees constructed in a greedy manner with split points based on purity scores (i.e., Gini, minimise loss). Thus, larger trees can be used with around 4 to 8 levels. Learners should still remain weak and so they should be constrained (i.e., the maximum number of layers, nodes, splits, leaf nodes). When a decision tree is the weak learner, the resulting algorithm is called gradient boosted trees, which usually outperforms random forest. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

The algorithm starts off by taking the average of the existing features as the initial prediction for the model and then builds a tree using the residuals of the values.The prediction of the residual tree is then multiplied by the learning rate and added to previous prediction and hence it moves a small step in the right direction.

Basically, the input test sample is split into several decision trees and the value of the prediction of all those trees built using residuals scaled by the learning rate will give the predicted answer of decreasing variance until it cannot decrease any further or maximum depth is reached.

4.2 Model Training

For training our model, we have considered features like yield of rice and wheat, across different states for the whole year. The features of air pollution we have used include Nitrogen dioxide (NO_2), Sulphur Dioxide (SO_2), Respirable Suspended Particulate Matter (RSPM) and Suspended Particulate Matter (SPM). We have considered multiple districts in different states of India.

The dataset obtained is made to undergo preprocessing to make it suitable for model training. After preprocessing, the dataset is split into 70% training and 30% test dataset. Different machine learning algorithms such as Gradient Boost Regressor, Random Forest Classifier (RFC) and Linear Regression (LR) are used to determine which algorithm shows the best correlation between different features, and mean squared error (MSE) should be close to zero. Depending on these criterias, an algorithm is chosen. After choosing the appropriate algorithm the dataset is trained and later the data modelling is performed.

This prediction has led to an accuracy of 95-98% and an R^2 score close to 0.01 (0.007183). This predicted value is then displayed in the redirected web-page with the mean absolute error.

V. Result and Discussion

This paper demonstrates the use of various regression algorithms: Gradient Boosted Regressor, Random Forest Regression and Linear Regression, for predicting the crop yield in different districts of India.

A comparative study was performed using three different regression algorithms to enhance accuracy. After training all models, the accuracies of the different models were compared. Gradient Boosted Regressor Algorithm tops the list with a R^2 score of 0.96, followed by Random Forest Regressor at 0.83.

Table 5.1: Performance Comparison of Different Algorithms

Model	R2 Score	MSE Score
Gradient Boosted Regressor	0.96	0.007184
Random Forest Regressor	0.83	0.063630
Multiple Linear Regression	0.79	0.063698

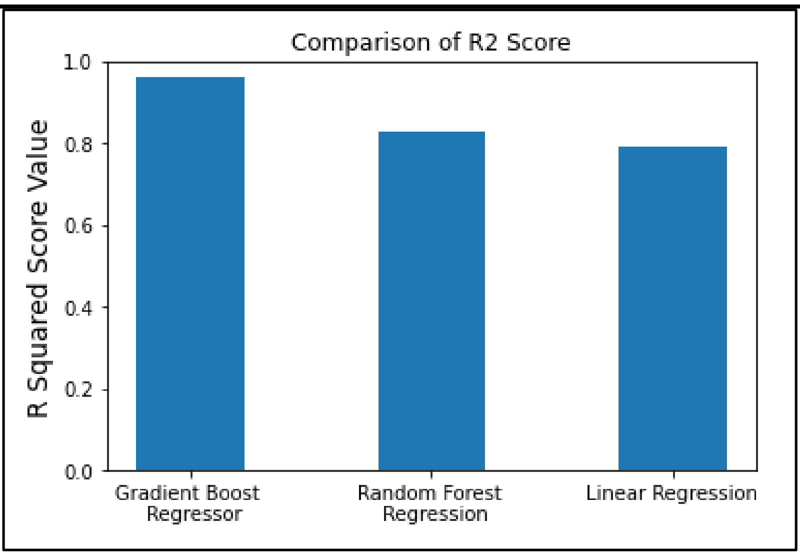


Fig 10: Comparison of R Squared Score Values

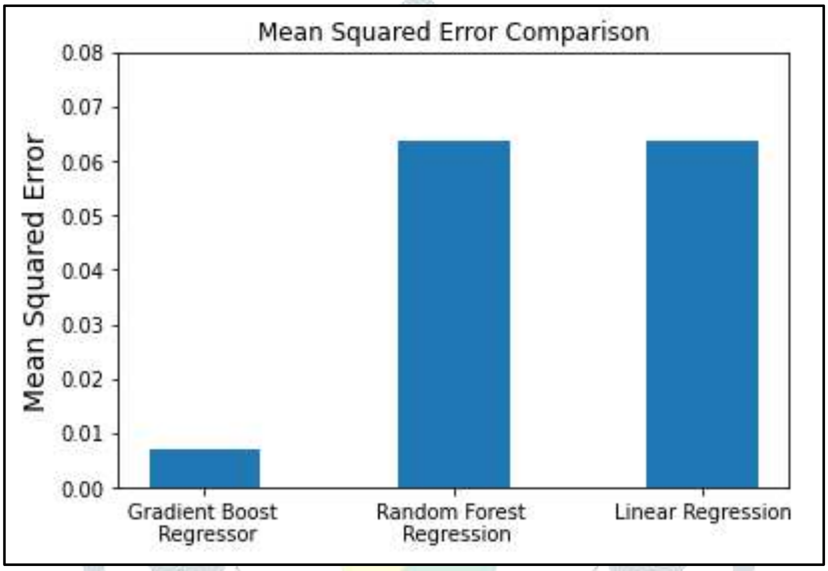


Fig 11: Comparison of Mean Squared Error Values

Figure 10 & 11 show the plots of the models we considered (after hyper parameterization) V/S the metric used to measure the accuracy as stated in Table 5.1. We observed that the metrics can further be improved only through extrapolation and no further hyper parameterization can improve the accuracy.

The subsequent sections will give deeper insight on our web application.

Figure 12 contains the feature importance plots of the F-Score of the features considered from the dataset out of which the highest scores of 699 and 613 belong to RSPM and SO_2 respectively and they are considered in the analysis described in Fig. 18 and Fig. 19.

In Fig. 13, we have to enter the crop, area and district we want the prediction for upon which we get the results as described in Fig. 18, Fig. 19, Fig. 20 which contains the analysis of the gases with the 2 most highest F-Scores in Fig. 16.

Figure 14 contains the Wheat Yield values as predicted by the GBM if SO_2 was present in the atmosphere (Predicted) V/S if it were absent (Real) thus, showing us the effect of the gas on the prediction for that district.

Figure 15 contains the Wheat Yield values as predicted by the GBM if RSPM was present in the atmosphere (Predicted) V/S if it were absent (Real) thus, showing us the effect of the particulate matter on the prediction for that district.

Figure 16 contains the Wheat Yield values as predicted by the GBM if both the aforementioned gases were present in the atmosphere (Predicted) V/S if it were absent (Real) thus, showing us the effect of the gases & particulate matter on the prediction for that district.

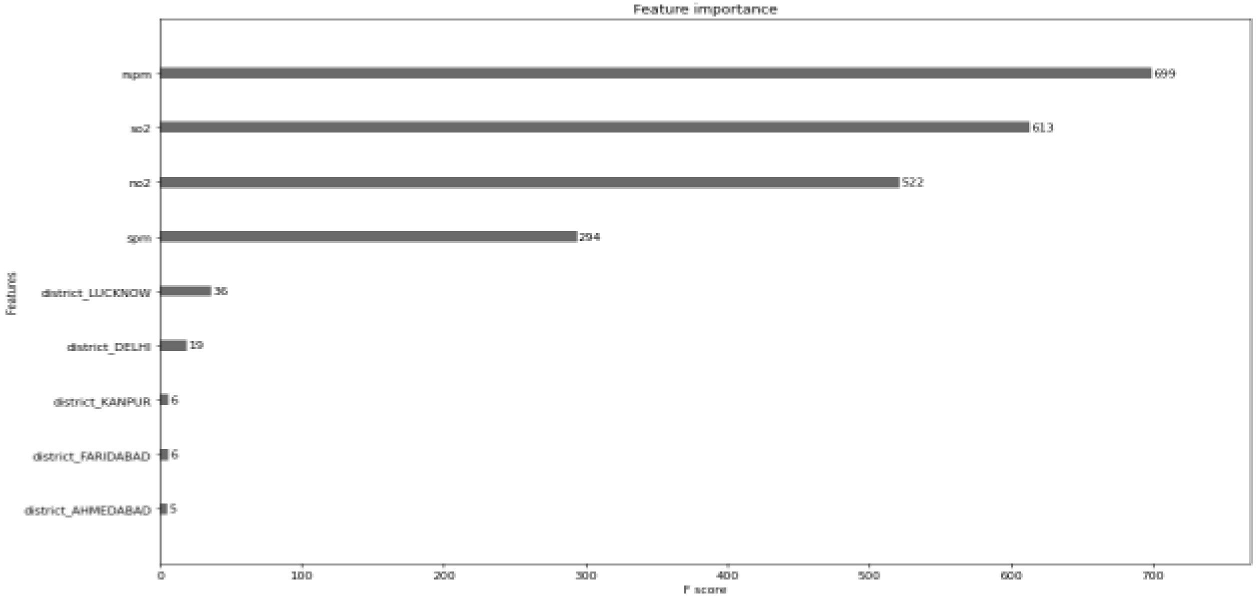
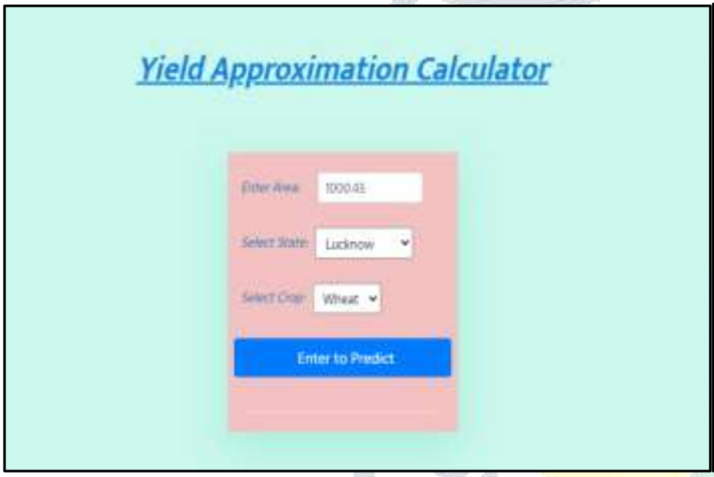


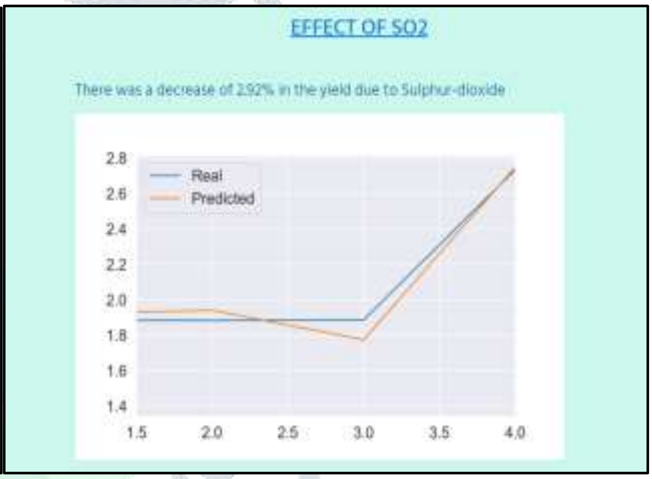
Fig 12: Feature Importance Plot


Yield Approximation Calculator

Enter Area: 100045

Select State: Lucknow

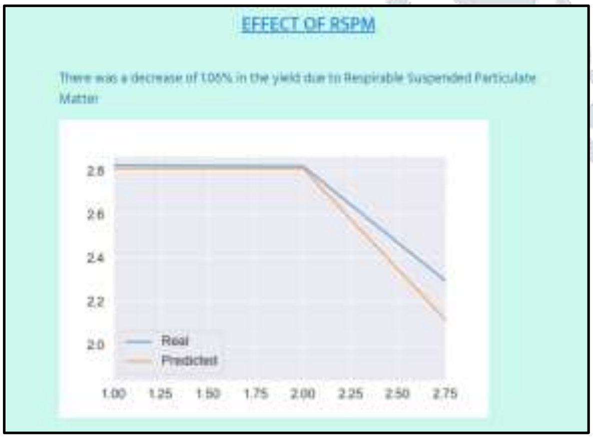
Select Crop: Wheat


EFFECT OF SO2

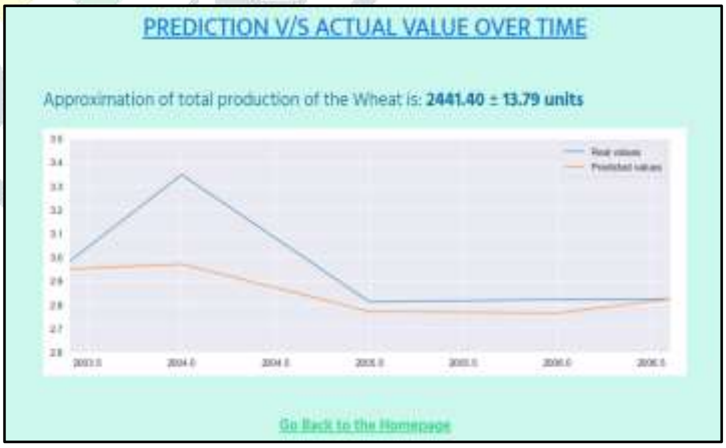
There was a decrease of 2.92% in the yield due to Sulphur-dioxide

Fig 13: Crop Yield Approximation Calculator

Fig 14: Effect of SO_2


EFFECT OF RSPM

There was a decrease of 1.06% in the yield due to Respirable Suspended Particulate Matter


PREDICTION V/S ACTUAL VALUE OVER TIME

Approximation of total production of the Wheat is: **2441.40 ± 13.79 units**

Fig 15: Effect of RSPM

Fig 16: Approximation of total production with graph of Real & Predicted Values

VI. Conclusion

Due to increasing pollution in the atmosphere in the past few years, we have considered the following features upon which we have built a model which predicts the yield losses and approximation of the production with an accuracy 96%.

The purpose of our proposed work is to provide farmers, researchers and government entities with an open-source web-based application that can predict a rough estimation of the yield losses of crops due to various air pollutants and also a rough estimation of how the yields are going to be. The idea behind this website is to be able to enter a location and find out the rough yield for a crop in that location based primarily on air-pollution factors but also on factors such as soil type, rainfall, etc. Right now, our scope is limited to only a few districts of India but we aspire to incorporate as many districts as we can find data for.

VII. Future Scope

Further developments can allow us to predict how farming in the same location along with the air pollution factors will affect the crop yield since repeated farming depletes nutrients of soil. Eventually, we can train the model to learn from its own predictions to find out the pattern of mistakes we are making that are leading to decreasing yield so that it can suggest necessary steps to prevent them.

References

[1] Ghude Sachin D., et al. "Reductions in India's crop yield due to ozone." Geophysical Research Letters 41.15 (2014): 5685-5691.

[2] Yadav Achchhela, et al. "The effects of elevated CO2 and elevated O3 exposure on plant growth, yield and quality of grains of two wheat cultivars grown in north India." Heliyon 5.8 (2019): e02317.

[3] U. Mina, R. Sigh, B. Chakrabarti "Agricultural Production and Air Quality: An Emerging Challenge" International Journal of Environmental Science: Development and Monitoring (IJESDM) ISSN No. 2231-1289, Volume 4 No. 2 (2013)

[4] Auffhammer, Maximilian, V. Ramanathan, and Jeffrey R. Vincent. "Integrated models show that atmospheric brown clouds and greenhouse gases have reduced rice harvests in India." Proceedings of the National Academy of Sciences 103.52 (2006): 19668-19672.

[5] Richa Rai, Madhu Rajput, Madhoolika Agrawal* and S.B. Agrawal "GASEOUS AIR POLLUTANTS : A REVIEW ON CURRENT AND FUTURE TRENDS OF EMISSIONS AND IMPACT ON AGRICULTURE" Journal of Scientific Research Vol. 55, 2011 : 77-102, Banaras Hindu University, Varanasi ISSN : 0447-9483 (2011)

[6] Dahikar, Snehal S., and Sandeep V. Rode. "Agricultural crop yield prediction using artificial neural network approach." International journal of innovative research in electrical, electronics, instrumentation and control engineering 2.1 (2014): 683-686.

[7] Gurjar, B.R., Ravindra, K. and Nagpure. "Air pollution trends over Indian megacities and their local-to-global implications". Atmospheric Environment (2016), 142, pp.475-495.

[8] Sahu, Shriya, Meenu Chawla, and Nilay Khare. "An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach." 2017 International Conference on Computing, Communication and Automation (ICCCA). IEEE, 2017.

[9] Ravikumar Hoogar, CP Mansur, LS Gajanana and SB Manjunath. "Effect of methane emission, mechanisms and management options in rice field" Journal of Pharmacognosy and Phytochemistry 2020; 9(3): 790-793

[10] Thomas van Klompenburg, Ayalew Kassahun, Cagatay Catal. "Crop yield prediction using machine learning: A systematic literature review" Computers and Electronics in Agriculture Volume 177, October 2020, 105709

[11] European Environment Agency. "Increasing environmental pollution (GMT 10)" SOER 2015

[12] Burney, Jennifer, and V. Ramanathan. "Recent climate and air pollution impacts on Indian agriculture." Proceedings of the National Academy of Sciences 111.46 (2014): 16319-16324.

[13] Aggarwal, P. K. "Global climate change and Indian agriculture: impacts, adaptation and mitigation." Indian Journal of Agricultural Sciences 78.11 (2008): 911.