Analysis on Air Quality and its Effects on Agriculture

Saritha
Computer Science and Engineering
PES University
Bengaluru, India
saritha.k@pes.edu

Ritika Shetty
Computer Science and Engineering
PES University
Bengaluru, India
shettyritika07@gmail.com

Manasa Devi
Computer Science and Engineering
PES University
Bengaluru, India
manasa3a@gmail.com

Akash Dhotre
Computer Science and Engineering
PES University
Bengaluru, India
akashdhotre873@gmail.com

Prema R Hanchinal
Computer Science and Engineering
PES University
Bengaluru, India
premahanchinal17@gmail.com

2022 3rd International Conference for Emerging Technology (INCET) | 978-1-6654-9499-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/INCET54531.2022.9825185

Abstract—**Due to modernization and avid industrialization, there has been a significant improvement in the quality of life of people. But this desire for attaining a better quality of life comes at the cost of environmental pollution. One of the main effects of environmental pollution is air pollution, caused by increasing concentrations of toxic gases in the atmosphere leading to harmful diseases in human beings. It is not only detrimental to human beings but also has its ill effects on agriculture. Since the consequences of air pollution on plants are not materially visible, the relevant data should be considered, and the results are to be calculated. As the farmers' main concern is pests and plant diseases, the harmful effects of air pollution are often left unattended. Whilst a few plant species may resist critical levels of tainting due to suspended particulate matter and build up gases, others are prone to the damage. Hence, the response of plants to air pollution depends on the variety of toxic substances present, their concentration, and their range of receptiveness to it. With the machine learning approach, the effects of air pollution on agriculture in terms of patterns in crop yield over the years can be analyzed and predict the more resistant crop for the given pollution data. The software developed for this purpose can be helpful for farmers, as it will help them decide the crop to grow in their field so that the crop production has a minimal effect due to air pollution.**

Keywords—air quality, crop prediction, machine learning models, agriculture

I. Introduction

The impacts of air contamination on crop yield aren't visibly noticeable except if keenly observed and determined. More often than not, farmers address the difficulties of irritations and infections affecting the plantations. As a part of dealing with the above issue, the harmful effects caused by air pollution are left unattended. Though the changes in farming practices possibly lessen these effects, the information expected to distinguish and execute these progressions is neither viable nor accessible at a feasible level. Distinct plant species respond differently to pollution. Whilst a few plant species may resist critical levels of tainting due to suspended particulate matter and build up gases, others are prone to the damage. Hence, the response of plants to air pollution depends on the variety of toxic substances present, their concentration and their range of receptiveness to it.

We have designed and developed the software to solve the above problem. This software helps farmers with their decision in choosing the right crop. It assembles an analytical report based on the client's input, using satisfactory machine learning approaches. It details toxin information their after-effects on the crop yield. It also proposes an ideal crop for maximum production. These are the main functions of the system that should be satisfied. If a user enters the required data of the area under inspection, the software will analyse the data and recommend which crop is best suited. Thus, it will have less impact on the growth of plants due to the given factors.

II. Related Work

Shushing et al. [1] proposed an agricultural environment expectation and analysis model dependent on profound learning Long Short-Term Memory (LSTM). The experimental data comprises agrarian climate observing information of the 2018 yearly annual key research and development project. It incorporates factors like wind speed, pm2.5, temperature, humidity, etc., for analysis. Later, LSTM was streamlined with Gated Recurrent Unit (GRU) and an additional dropout layer to prevent overfitting. Comparative analysis was done dependent on the metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Square Error (RMSE) to choose the best model. In comparison with GRU, LSTM performed better.

Umer et al. [11] utilized four distinct methodologies with two different datasets to create Random Forest (RF) models for wheat yield estimating. The primary method used separate RFs for each district. The subsequent methodology utilized generic RFs for each district. In the third methodology, the RF depended on the average of each independent variable across the eight districts for each particular 8-day time span. The fourth method used the generic RF for each district. But in this case, it was applied using the averages of all parameters for Punjab province to predict yield for the year. The performance of the models was surveyed using RMSE, error and mean error rate and by contemplating the connection among reported and anticipated crop yields. The generic RF for the whole precisely predicted the crop yield within a 3.84% error.

Bharat et al. [9] performed a C4.5 calculation to find the most affecting climatic parameter on the harvest yields of chosen crops in chosen areas of Madhya Pradesh. The model had an 82% precision even though some of the other agro-input parameters responsible for crop yield didn't consider.

Yumiao et al. [10] talked about using advanced machine learning algorithms to build within season yield prediction models for winter wheat using multi-source data. In particular, yield driving factors had been secluded from four diverse information sources, including satellite pictures, climate data, soil maps, and historical yield records. The outcome of the various models had been analysed for assessing the area level winter wheat yield in the Conterminous United States within the growing season. The machine learning models' performance was better compared to the linear regression models. With the best outcome accomplished by utilizing the AdaBoost model (R squared = 0.86, RMSE = 0.51 t/ha, MAE = 0.39 t/ha). Moreover, the outcomes show that combining data from numerous sources beat single-source satellite data. The AdaBoost model showed high accuracy somewhere in the range of 0.8 and 0.9 for various combinations of datasets.

III. Dataset

Three types of data are collected to analyze crop yield patterns amidst air pollution over time: crop yield, atmospheric gases, climate and soil data.

Air Quality Data is a combined and clean version of the Historical Daily Ambient Air Quality Data released by the Ministry of Environment and Forests and Central Pollution Control Board of India under the National Data Sharing and Accessibility Policy. It contains the data of different pollutants like Sulphur dioxide (SO2), Nitrogen dioxide (NO2), Respirable suspended particulate matter (RSPM), and suspended particulate matter (SPM) levels recorded at location monitoring stations in various places of India.

Crop Yield Data is the second dataset collected from the joint office of the Department of Agriculture and Cooperation, which is called the Directorate of Economics and Statistics. The department collects data and publishes statistics on various aspects of agriculture and related areas. It contains the details about the Area, Production and Yield information of principal crops in different places of India in a given period.

Crop Recommend Data, an augmentation of rainfall, climate and fertilizer data available for India. It is specifically prepared for building prediction models to recommend the most suitable crop to grow in the given conditions.

IV. Proposed Methodology

A. Architecture

The main objective is to recommend the crop based on the soil constituents, atmospheric gases and climatic parameters and suggest suitable fertilizer for farming to improve crop yield. For this purpose, the system is built based on client-server architecture. The system is maintained such that agriculturists and students can see the analysis and get the result or extend the output for further analysis.

The software is implemented based on the agile software development lifecycle. An agile approach is an iterative approach that develops software from the beginning of the project to deliver it properly up to the end. As the project is of small scale, being developed by a small team, this approach is well suited and helpful as it can adapt to ambiguous requirements.

Fig. 1 shows the machine learning approach used to build the solution. The data pre-processing is the first step where the datasets are converted to such a state that the machine can understand and learn the data. Data visualization and prediction models use this pre-processed data. Prediction models are built based on multiclass classification algorithms since there are more than two classes to predict as output for each input. In this case, classes represent the various crops present in the dataset. Different machine learning models are built and trained on historical data of crop yield and atmospheric conditions of past years. Further, using performance metrics models' performance is evaluated. The web application uses the high performing model in the end.

B. Data Preparation and Data Pre-processing

First, the Air Quality data and Crop Yield data are merged (merged_data) based on spatiotemporal data to examine the effect of pollutants on crop yield. The observations from the analysis were that some crops showed less production in the period when the pollutant concentration in the atmosphere was high. Some crops like Sugarcane showed no ill effect in their yield due to any pollutants.

In the next part, the pre-processing of merged_data is as follows: For each crop in the crops list, evaluate the ideal values of SO2, NO2, SPM and RSPM to get a high crop yield. These results merged with crop recommendation data which had crop labels for only climate and soil data, but now it also contains pollutant data. TABLE I shows the newly formed dataset with features. The dataset is ready to be utilized in training a model and predictions on the crop.

C. Machine Learning Models

We have used five different machine learning algorithms to build and train models on the prepared dataset. For training purposes, the crop label parameter is considered the dependent value and the remaining parameters independent

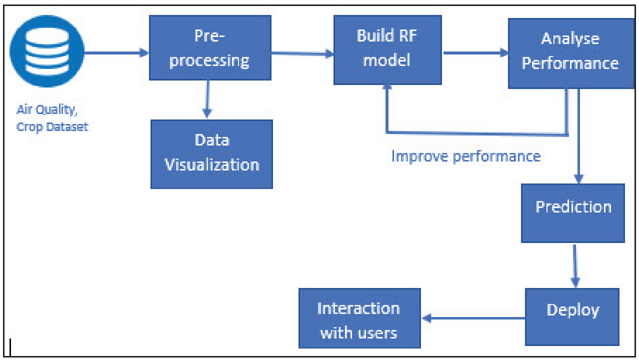

Fig. 1. The architecture of the approach

TABLE I. LIST OF FEATURES

Sl. No.	Features	Data Type	Description
1	SO2	Numeric	Sulphur dioxide concentration in ug/m3
2	NO2	Numeric	NO2 gas concentration in ug/m3
3	RSPM	Numeric	RSPM gas concentration in ug/m3
4	SPM	Numeric	SPM gas concentration in ug/m3
5	N	Numeric	The ratio of nitrogen content in the soil
6	P	Numeric	The ratio of phosphorous content in the soil
7	K	Numeric	The ratio of potassium content in the soil
8	temperature	Numeric	The temperature in degree celsius
9	humidity	Numeric	Relative humidity in percentage
10	soil ph.	Numeric	Ph. value of the soil
11	rainfall	Numeric	Rainfall in mm
12	label	String	Crop name

variables. The accuracy of each model is computed by comparing actual test set values and predicted values. After testing, these models are fine-tuned for better performance before performing comparative analysis.

1) *Adaboost Model:*

AdaBoost is an ensemble learning method. This boosting algorithm combines multiple low accuracy (or weak) models to create a high accuracy (or Strong) model. First, the AdaBoost algorithm builds a model and makes predictions. It assigns higher weights to miss-classified points. Again builds a model on this new data. In each iteration, the algorithm tries to reduce the training errors to provide an excellent fit for the model. Thus, the final model uses the weighted average of individual models. The AdaBoost model accuracy is 31.52%, the precision is 21%, recall is 32%, and F1-score is 22% and support value of 368. The accuracy of 31.52% tells that it is a poor model. Fig. 2 shows the frequency of prediction errors. So, to improve the performance, the model is optimised using parameter tuning. The important parameters used here are base_estimators, n_estimators, and learning_rate. The model uses Support Vector Classifier as a Base Estimator. Fig. 3 shows the frequency of prediction errors after parameter tuning. After the Parameters Tuning, we got an accuracy of 96.73%, considered as a good model, precision was 95%, recall of 97%, and F1- score of 96% and also, we found an average bias of 30.725, an average variance of 6.459.

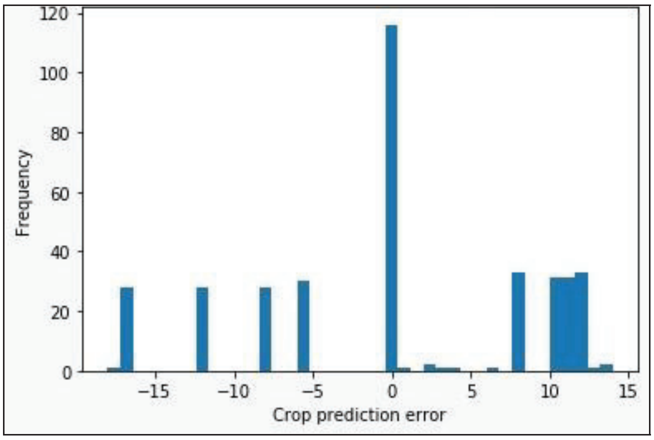

Fig. 2. Prediction error plot for AdaBoost model before Parameter tuning.

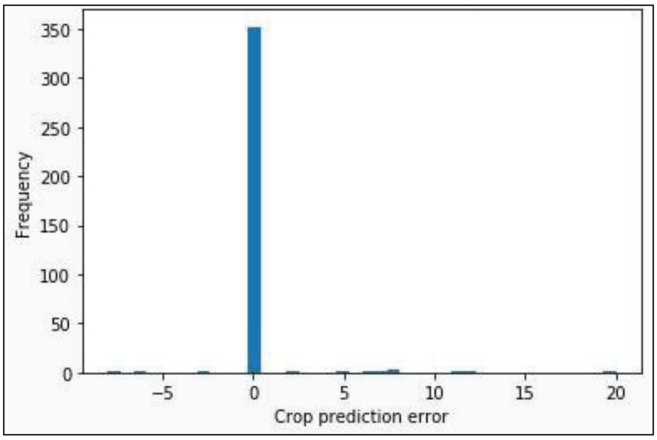

Fig. 3. Prediction error plot for AdaBoost model after Parameter tuning.

2) *Random Forest Model*

A random forest is a meta-estimator that uses averaging to improve predictive accuracy and control over-fitting by fitting several decision tree classifiers on different sub-samples of the input data. The data are trained and tested against different hyper-parameters to achieve the best accuracy with the least possible error. The model is optimized using hyper-parameter tuning. The hyper-parameters tuning helps get the optimal combination of values in the algorithm for the prediction. The change in n_estimators, max_features and max_depth gave the accuracy with least error < 0.9 and accuracy of 94%. The Randomized Search CV and Grid Search CV gave the best optimization for the model with an average error of 0.26 degrees, an improvement of 0.13 and accuracy of 97.63%. The model depicted an accuracy of 95% with the overfitting and cross-validation of 20 folds on the dataset. After hyper-parameter tuning using Randomized Search CV, the model achieved 97.63% with 0.56% improvement. With Grid search CV, improvisation of 0.33% and an accuracy of 97.63%.

Fig. 4 depicts the out of bag error rate concerning n_estimators for the hyperparameter max_features sqrt, log2 and None. The model's mean square error is 0.046, RMSE is 0.677, values are less than 1. The model has a variance of 0.89 is high with a bias of 2.127 and the loss expected being 3.125. All these values indicate that the model is a good solution for the prediction. The calculated roc_auc score (area under the curve) is 0.68 is greater than the random guess value of 0.5. It shows that after optimization, the model performance improves in a better way for prediction with less error rate.

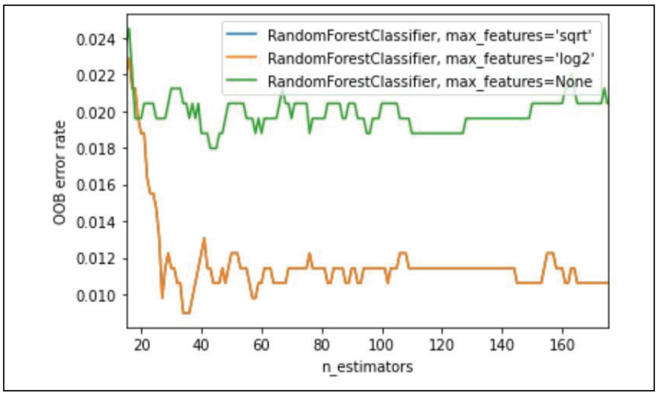

Fig. 4. The OOB error rate for n_estimators of Random Forest model

3) *KNN Model*

The KNN classification model is used for our analysis because it gives the nearest neighbour in the data points for the old point as the data grows. The optimal k value is taken from the n number of samples. The square root of n samples gives k-nearest. KNN is best for an accurate predictive model as k value increases to locate the data points becomes hard. We have considered five n_neighbors and 20 k-fold cross-validations for both testing and training the data. The model's error rate is less than 0.05, the accuracy of 95.91%, precision and recall are 0.96 and 0.81 with an f1-score of 0.71. The AUC(Area under curve) value obtained is 0.58, a low variance of 0.58 and a bias of 0.23. It is a good fit for the data and a good performance for prediction. Fig. 5 shows the error rate curve. As the k value of the model increases, the mean error value that ranges from 0.03 to 0.055 (<1) also increases. It concludes that the model is a good model for classification of crops.

4) *Decision Tree Model*

The Decision Tree classifier is the number of decisions put together to form a single tree for the entire dataset. The model built has a mean square error, RMSE, variance, bias, loss of 2.67, 4.94, 14.00, 10.83, 24.83, respectively. First, the model had 61.41% accuracy. Later, it is optimized by increasing the max depth of the decision tree. After optimization, the obtained MSE, RMSE, variance, bias, loss values are 0.43, 2.1, 2.83, 2.2, 5.03 respectively. It showed accuracy up to 94%.

5) *Naïve Bayes Model*

The Naïve Bayes algorithm is a supervised learning algorithm based on the Bayes theorem and used for solving classification problems. It is a probabilistic classifier, which means it predicts based on the probability of an object. It converts a given dataset into a frequency table and generates a likelihood table by finding the probabilities of the features. And then, it uses the Bayes theorem to calculate posterior probability. The accuracy achieved up to 98.57%. The model's mean square error is 0.233, RMSE value is 1.74, variance is 0.26, bias is 2.58, loss is 2.84.

D. *Evaluation of Models*

The models are evaluated based on performance metrics. TABLE II depicts the performance metrics of each model. Each metric is defined as follows.

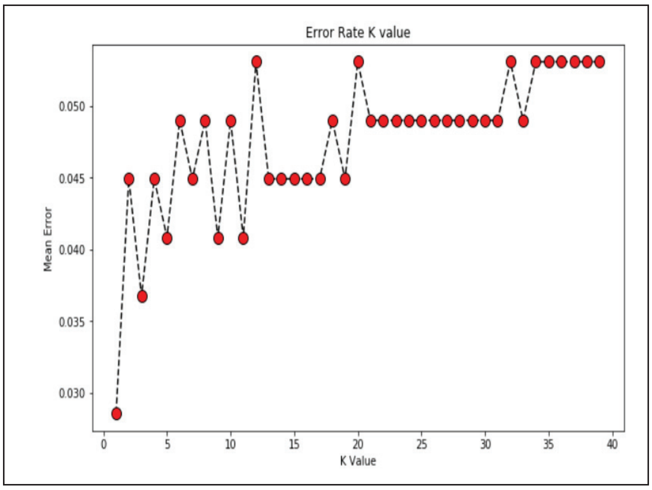

Fig. 5. The Error rate of KNN model

Accuracy is the ratio of the number of classifications a model correctly predicts to the total number of predictions made. Among the first four models, Random Forest achieved the highest accuracy of 97.63% followed by the AdaBoost model with 96.73%. Even though Naïve Bayes has higher accuracy than Random Forest, it is prone to overfitting.

MSE is the measure of how well a regression line fits the data points. Out of all the models, Random Forest has the least average error.

Precision is the quality of a prediction made by a model. It is the ratio of true positives to the total number of positive predictions.

Recall is the measure of identifying true positives in the model. It is the ratio of actual true positive to the total predictions.

F1-score is the harmonic mean of precision and recall and is a better measure than accuracy. The Naïve Bayes model has a high f1-score followed by AdaBoost and KNN.

AUC (Area under the curve) score is the measure of separability. It is equal to the probability that a classifier will rank a randomly chosen positive instance higher than the negative one.

ROC (Receiver Operating Characteristic)-AUC Curve is the probability curve that measures the performance of the classification model by depicting the rate of true positive with respect to the false positive. Higher the AUC better the curve. Compared to other machine learning models,

TABLE II. PERFORMANCE METRICS OF MODELS

Model Name	**Accuracy (%)**	**RMSE**	**MSE**	**Variance**	**AUC score**	**F1 score**	**Precision**	**Recall**
Random Forest	97.63	0.678	0.046	0.999	0.681	0.944	0.94	0.95
AdaBoost	96.73	1.8274	0.2798	6.459	0.68	0.959	0.95	0.97
KNN	95.91	0.858	0.387	0.893	0.580	0.949	0.94	0.96
Decision Tree	94.00	2.1	0.43	2.83	0.6	0.94	0.96	0.95
Naïve Bayes	98.57	1.739	0.233	0.261	0.61	0.98	0.99	0.99

Random Forest has a high AUC value which concludes that the model performance is better for multiclass classification.

Fig. 6 shows ROC-AUC of the Random Forest. Each curve represents the AUC score for different classes. The dashed line called random guess with an AUC score of 0.5 acts as a reference line to depict the performance of the model. Higher the area under the curve better the model.

Out of five models, the Random Forest model proved to be better at prediction with the least error and 0.13% increase in improvisation after hyper-parameter tuning.

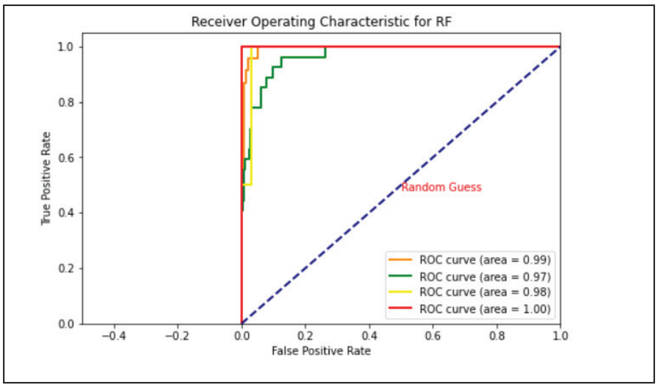

Fig. 6. ROC curve for Random Forest model

V. Results

The Random Forest model has been used for the final prediction and in the deployed software. Crop yields are analyzed based on parameters i.e., SO2, NO2, RSPM, SPM. From the obtained results, it has been observed that Kharif crops like Rice, Maize, Millet, etc., are being affected strongly by the above-mentioned pollutants whereas Rabi crops like Wheat, Barley, etc., have shown an aversion to SO2. Fig. 7 shows the variation between wheat yield and SO2, NO2 concentration. Pearson Correlation shows that yields of Rice, Wheat, Maize, Barley are strongly and negatively correlated to RSPM, SPM, SO2, NO2 affecting the yield and production of the crops. On the other hand, Sugarcane and Cotton crops are positively correlated with SO2, NO2 as shown in Fig. 8. It concludes that crops grown in different seasons have an adverse effect due to the harmful gases and particulate matter present in the air. India is a major producer of Rabi, Kharif and cash crops with seasonal and annual spices. The air quality variation shows that farming will become challenging in future days.

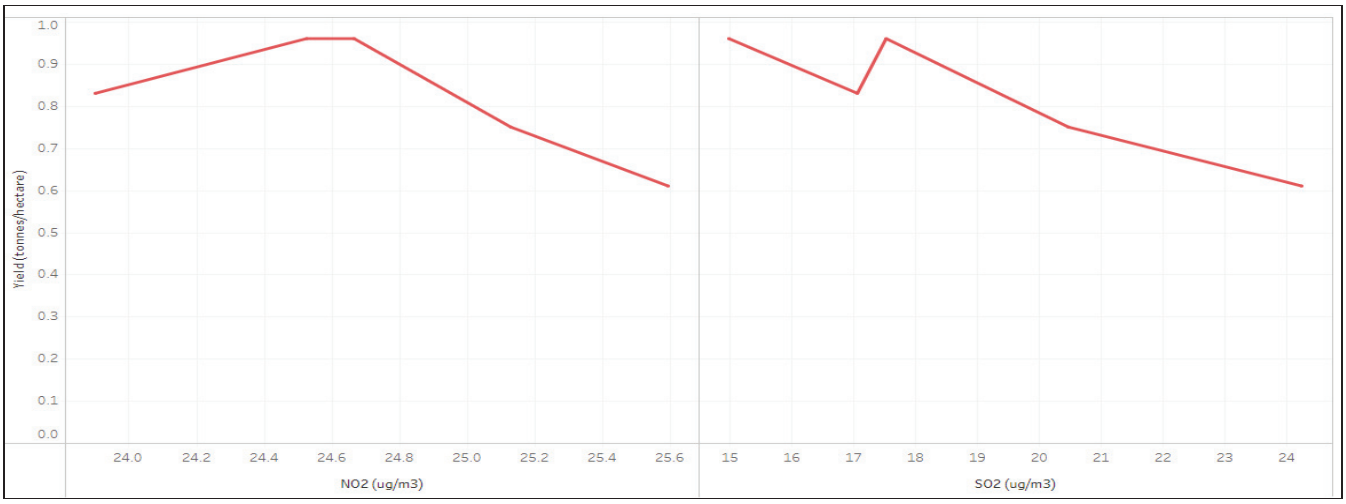

Fig. 7. Variation between Wheat yield and NO2 and SO2

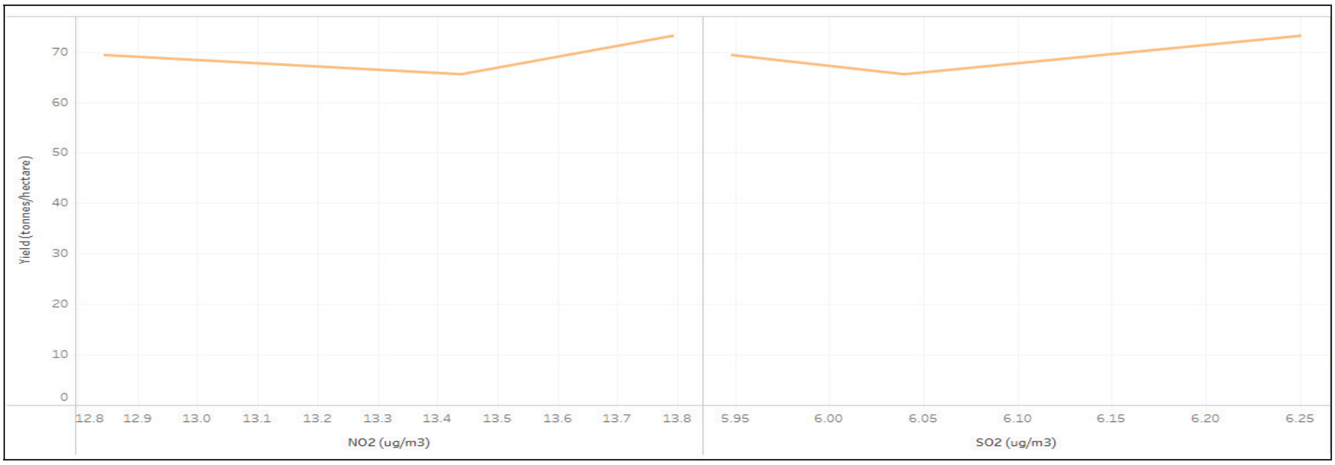

Fig. 8. Variation between Sugar cane yield and NO2 and SO2

VI. CONCLUSION

As industrialization has been increasing by leaps and bounds, the adverse effects of air pollution on agricultural lands have become an alarming issue. So, we have decided to develop an application that helps in minimizing the ill effects by predicting the optimal crop using the available pollution data. Hence, the hypothesized crop will have the potential to adapt to the ever-changing environment. Following the crop output, our application also suggests beneficial fertilizers to enhance the crop yield.

In the process of developing the model, we have considered the pollutants which are accessible. By inputting the geologically affiliated data, the application will have the ability to analyze and predict the output accurately. With the above-suggested improvisations, our application will have a significant impact on the agricultural sector. The proposed work might act as a blueprint for the agriculturalists in the near future for the study of the impact of pollution on agriculture, to analyze the related trends in crop production and air quality and also to predict the crop yield for the coming years.

REFERENCES

[1] Shushing Chen, Bingham Li, Jibe Cao and Bo Mao, "Research on Agricultural Environment Prediction Based on Deep Learning", Procedia Computer Science 139 (2018) Journal, pp 33–40, Elsevier B.V, doi: 10.1016/j.procs.2018.10.214.

[2] Vanda Eva Molnar, Edina Simon, Sarawut NInsawat, Bela Tothmeresz and Szilard Szabo, "Pollution Assessment Based on Element Concentration of Tree Leaves and Topsoil in Ayutthaya Province, Thailand", *Int. J. Environ. Res. Public Health* 2020 Journal, vol. 17, issue 14, July 2020, doi:10.3390/ijerph17145165.

[3] Heck W.W, Taylor O.C and Tingey D.T, "Effects on Photosynthesis, Carbon Allocation, and Plant Growth Associated With Air Pollutant Stress", Springer, Dordrecht, 1988, doi:10.1007/978-94-009-1367-7_13.

[4] Wang, Yunqi,Li, Bai,Liu and Yanju.Liu, Xuan, "Application of a coupled model of photosynthesis and stomatal conductance for estimating plant physiological response to pollution by fine particulate matter (PM2.5)", Environmental Science and Pollution Research, doi: 10.1007/s11356018-2128-6, 2018.

[5] Feifei Sun, Yun DAI and Xiaohua Yu, "Air pollution, food production and food security: A review from the perspective of food system, Journal of Integrative Agriculture", Journal of Integrative Agriculture, Volume 16, Issue 12, pp 2945-2962, 2017, doi: 10.1016/S2095-3119(17)61814-8.

[6] Abdulwaheed Tella, Abdul-Lateef Balogun and Ibrahima Faye, "Spatio-temporal modelling of the influence of climatic variables and seasonal variation on PM_{10} in Malaysia using multivariate regression (MVR) and GIS", Geomatics, Natural Hazards and Risk Journal, vol. 12, issue 1, pp 443-468, 2021, Feb 2021, doi: 10.1080/19475705.2021.1879942.

[7] J. K. Jaiswal and R. Samikannu, "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression", 2017 WCCCT, IEEE, pp. 65-68, 2017, doi: 10.1109/WCCCT.2016.25.

[8] T. Doan and J. Kalita, "Selecting Machine Learning Algorithms Using Regression Models", 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 2015, pp. 1498-1505, doi: 10.1109/ICDMW.2015.43.

[9] S.Veenadhari, Bharat Misra, Bharat and CD Singh, "Machine learning approach for forecasting crop yield based on climatic parameters", 2014 International Conference on Computer Communication and Informatics (ICCCI-2014), Jan 2014, Coimbatore, INDIA, doi:10.1109/ICCCI.2014.6921718.

[10] Yumiao Wang, Zhou Zhang, Luwei Feng, Qingyun Du and Troy Runge, "Combining Multi-Source Data and Machine Learning Approaches to Predict Winter Wheat Yield in the Conterminous United States", *Remote Sens.* 2020, April 2020, doi: 10.3390/rs12081232.

[11] Umer Saeed, Jan Dempewolf, Inbal Becker-Reshef, Ahmad Khan, Ashfaq Ahmad and Syed Aftab Wajid, "Forecasting wheat yield from weather data and MODIS NDVI using Random Forests for Punjab province, Pakistan", International Journal of Remote Sensing, Sept 2017, vol 38, issue 17, pp 4831-4854, doi: 10.1080/01431161.2017.1323282.