International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 07 | July 2021 www.irjet.net p-ISSN: 2395-0072

Crop Recommendation System with Comparative Analysis of Different Machine Learning Algorithms

S Anusha[1], M Bindu[2], G Navya[3]

[1]*Student, Dept. of CSE Engineering, REVA University, Bangalore, Karnataka, India*
[2]*Student, Dept. of CSE Engineering, REVA University, Bangalore, Karnataka, India*
[3]*Student, Dept. of CSE Engineering, REVA University, Bangalore, Karnataka, India*

Abstract - *Agriculture plays a vital role in Indian economy. Many factors do effect the growth of crops. Temperature, humidity, pH, rainfall, amount of potassium, nitrogen, phosphorous in soil all of these are the factors on which the yield depends. Many farmers have no idea about what crop to be grown in which area that will lead to maximum yield as well as profit. Hence in this paper we are going to explain how machine learning algorithm can be used to predict the crop which is best suitable for the area with specific temperature , humidity, pH , rainfall, and potassium , nitrogen , phosphorus levels in soil. Along with that we are also doing comparative analysis of different machine learning algorithms as to which algorithm gives us the highest accuracy. We have carried out the same experiment using different machine learning algorithms and then found which algorithm is best.*

Key Words: **random forest, naive bayes, machine learning, support vector machine, logistic regression, Decision Tree, Gradient Boosting, K Neighbors.**

1.INTRODUCTION

Agriculture plays a certainly important role in the economy of India. As the population of our country is increasing the need for food is also tremendously increasing [7]. Hence there is a huge need for proper growth of crops. Farmers must be able to get maximum yield and profit. So we can help farmers by suggesting which crop to be grown at what time so that it will give them the maximum yield. And very few farmers have knowledge about what temperature, humidity, amount of rainfall, pH in soil, potassium, phosphorous, nitrogen level in soil is best suitable for specific crop. What is the need to monitor potassium , phosphorus and nitrogen level in soil. Potassium , phosphorus and nitrogen are the three primary macronutrients required for fortunate growth of crops[2].

The significant role of potassium in the growth of crop is to regulate the opening and closing of stomata, this in turn regulates the exchange of oxygen, water vapor and carbon dioxide**.** Potassium also helps to increase growth of root, improve drought resistance[1]. Deficiency of potassium will lead to following symptoms like brown scorching, curling of leaf tips, chlorosis which is yellowing between the leaf veins, plant growth, leaf, root, seed development will be reduced[1].The major role of phosphorus in crop growth is to help the plants to carry out photosynthesis[2]. Phosphorus deficiency in plants show following symptoms such as stunted growth, due to the accumulation of sugar fiber colors appear ranging from dark green to reddish purple[2]. The key role of nitrogen in crop growth is to provide energy for the plants to grow, produce fruits or vegetables. Lack of nitrogen leads to yellowing of the plants which is said to be chlorosis[3]. Hence all the three minerals play a vital role in crop growth. Different crop needs different amount of these three minerals. Hence while growing a specific crop the person has to take care of amount of these three minerals in soil along with other requirements like temperature, pH in soil, rainfall, humidity. So basically our project helps to predict the most suitable crop for specific area with certain temperature, humidity, pH, rainfall and three major mineral content in soil. From this the person can attain maximum yield and profit.

1.1 SYSTEM DESIGN AND ANALYSIS

The overall workflow has two parts . The first part consists of preprocessing and feature design, which are specific to data sources, and splitting data into training and test sets. The second part, focusing on machine learning, is independent of data sources. The data was split into training and test sets before designing features. Some data sources required feature design, others were directly used as features. Once we had features and labels, machine learning algorithms were trained and optimized on the training set and evaluated on the test set. Then the result is predicted for the unknown values and finally accuracy is calculated for all the algorithms.

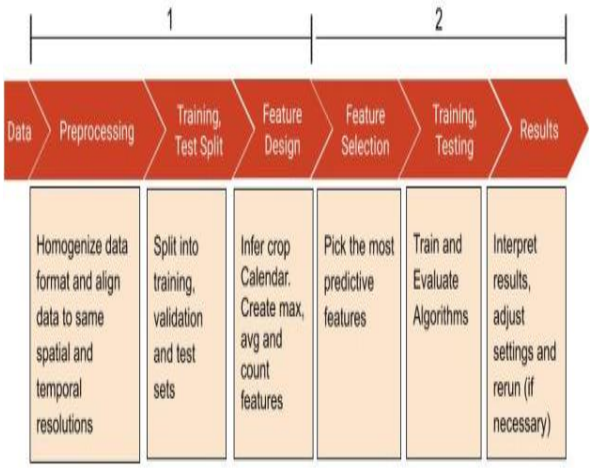

Fig-1: Work flow of the project [4]

2. METHODOLOGY

The main agenda of this project is to recommend a crop which is best suitable for area with specific temperature in degree celsius, humidity in percentage, pH value of the soil, rainfall in mm, and ratio of nitrogen , phosphorus ,potassium content in soil. The general procedure that we have followed has five steps.

First step is collect statistical data. Second step is preprocessing [8] the data which involves data cleaning, dimensionality reduction.

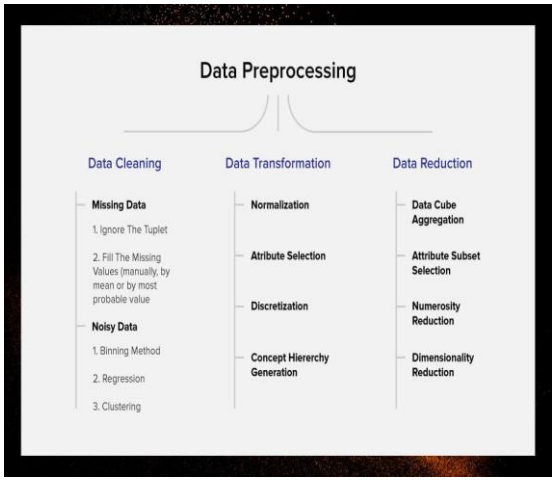

Fig-2: shows steps involved in data preprocessing [5]

Then comes the third step which is building a model using suitable machine learning algorithms like logistic regression, naive bayes algorithm, k nearest neighbor algorithm, random forest algorithm, support vector machine algorithm .

The fourth step is training the model. Fifth step is testing if our model is trained properly or not.

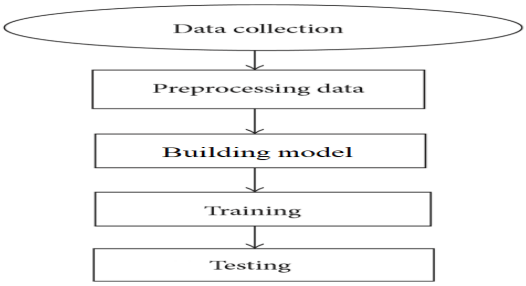

Fig-3: flowchart of steps involved in implementation [6]

The above steps are explained slightly in detail in the way we have implemented. First step is we have imported all the required packages. Second step is to input the statistical data which consists of descriptive features like temperature, humidity, pH, rainfall ,and nitrogen , phosphorus, potassium levels in soil and also consists of target feature which is crop nothing but label . The dataset looks like the picture given below.

Table-1: shows first fifteen columns of dataset

```
In [26]: data = pd.read_csv("Crop_recommendation.csv")
         data.head(15)
```

Out[26]:

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice
5	69	37	42	23.058049	83.370118	7.073454	251.055000	rice
6	69	55	38	22.708838	82.639414	5.700806	271.324860	rice
7	94	53	40	20.277744	82.894086	5.718627	241.974195	rice
8	89	54	38	24.515881	83.535216	6.685346	230.446236	rice
9	68	58	38	23.223974	83.033227	6.336254	221.209196	rice
10	91	53	40	26.527235	81.417538	5.386168	264.614870	rice
11	90	46	42	23.978982	81.450616	7.502834	250.083234	rice
12	78	58	44	26.800796	80.886848	5.108682	284.436457	rice
13	93	56	36	24.014976	82.056872	6.984354	185.277339	rice
14	94	50	37	25.665852	80.663850	6.948020	209.586971	rice

The dataset we have taken is very balanced. In order to show this we implemented a graph of crop on x axis versus count of the crop on y axis which shows how balanced our dataset is.

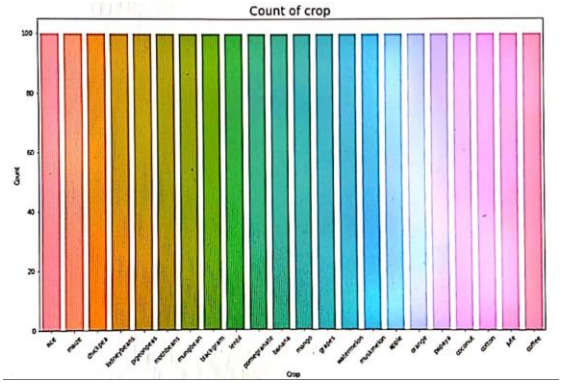

Fig-4: graph displaying count of crop

Since these descriptive features and target feature are all together we need to separate them.

Hence the third step is to separate the descriptive and target features, where descriptive features act as input whereas target feature act as output. The model will be later trained to predict the target feature based on the specific pattern of descriptive feature.

Fourth step is to split the data into train and test data , once we train the model later rest of the data which is reserved as test data can be used to evaluate if our model is trained perfectly or its facing problems like underfitting or overfitting. The problem underfitting arises due to reason such as insufficient data. It mainly occurs when the model is very simple with few features or highly regularized and overfitting is caused when the model predicts right output only for the data we have trained, it predicts wrong output for the unseen data.

In the next step we are supposed to create an object of specific algorithm. There are many algorithms in machine learning that can be used to train the model. In all the algorithms we must select an algorithm with highest accuracy. Hence when we used different machine learning algorithms [9] to train the model and then found accuracy of each model we noticed that for our dataset random forest algorithm and naive bayes algorithm are giving highest accuracy whereas logistic regression and support vector machine algorithm were giving comparatively lesser accuracy . The output appeared as shown in the image below.

```
LogisticRegression 0.9504132231404959
DecisionTreeClassifier 0.9724517906336089
GradientBoostingClassifier 0.9834710743801653
KNeighborsClassifier 0.9752066115702479
RandomForestClassifier 0.9931129476584022
NaiveBayes 0.9944903581267218
Support vector machine 0.9641873278236914
```

Fig-5: accuracy of various algorithms executed in jupyter notebook

Later we even implemented a graph for visualization. In the graph we can clearly notice that the accuracy of random forest and naive bayes algorithm are highest and the accuracy of logistic regression and support vector machine algorithm are less when compared with accuracy of other algorithms. Hence it is better to train our model using random forest algorithm or naive bayes algorithm. The graph with accuracy on x axis and algorithms on y axis is as follows:

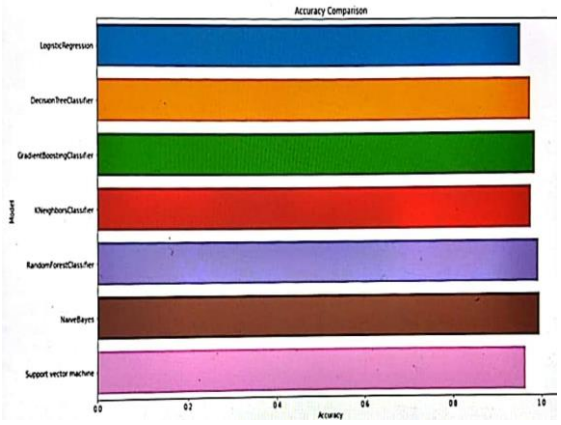

Fig-6: graph representing comparison of accuracy

Fifth step is to create an object of specific algorithm which will be used to train the model.

Sixth step is to train the model using fit method.

Seventh step is to check if the model is trained accurately or not with the help of predict method. This prediction is carried out on the test data. If the actual output and the predicted output are same then the model is trained suitably .

Eighth step is to carry out evaluation by finding accuracy and confusion matrix.

Since we found random forest algorithm is having suitable accuracy we concluded that training the model using random forest algorithm would be the best. Accuracy is a measure of how well our model is trained. It is a kind of evaluation. Confusion matrix is another method of evaluation. When we implemented the confusion matrix for the model trained using random forest algorithm we found that no much errors were there.

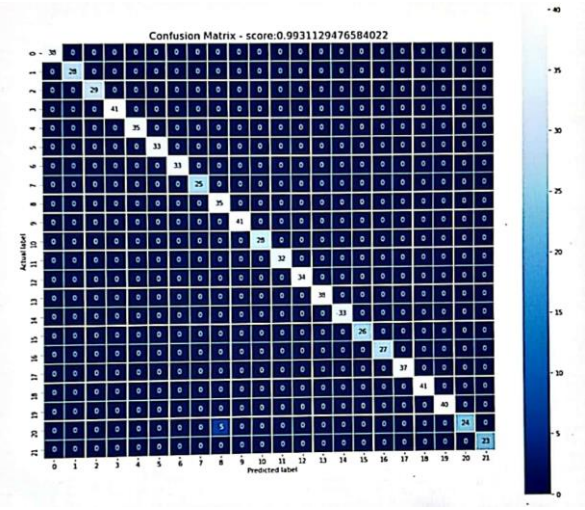

Fig-7: confusion matrix plot

Only one label is wrongly predicted . Rest of the labels are predicted perfectly.

Table-2: output of random forest classifier

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	38
banana	1.00	1.00	1.00	28
blackgram	1.00	1.00	1.00	29
chickpea	1.00	1.00	1.00	41
coconut	1.00	1.00	1.00	35
coffee	1.00	1.00	1.00	33
cotton	1.00	1.00	1.00	33
grapes	1.00	1.00	1.00	25
jute	0.88	1.00	0.93	35
kidneybeans	1.00	1.00	1.00	41
lentil	1.00	1.00	1.00	28
maize	1.00	1.00	1.00	32
mango	1.00	1.00	1.00	34
mothbeans	1.00	1.00	1.00	38
mungbean	1.00	1.00	1.00	33
muskmelon	1.00	1.00	1.00	26
orange	1.00	1.00	1.00	27
papaya	1.00	1.00	1.00	37
pigeonpeas	1.00	1.00	1.00	41
pomegranate	1.00	1.00	1.00	40
rice	1.00	0.83	0.91	29
watermelon	1.00	1.00	1.00	23
accuracy			0.99	726
macro avg	0.99	0.99	0.99	726
weighted avg	0.99	0.99	0.99	726

Ninth step is visualization. We can even represent the data in the form of graph for better understanding.

The main logic used here is , the relation between descriptive and target feature is found and then trained the model that for the descriptive feature with specific pattern this specific target feature must be predicted. Then accordingly when a new data is given based on the relation between descriptive and target feature the model predicts the output. We can even plot graph to check the correlation between label and other features. Label which is target feature on x axis and descriptive features on y axis. By observing this graph we get to know about how each descriptive feature is correlated to target feature. It basically shows which crop need what amount of descriptive features. The below graph with label on x axis and rainfall in mm on y axis shows how these two features are correlated.

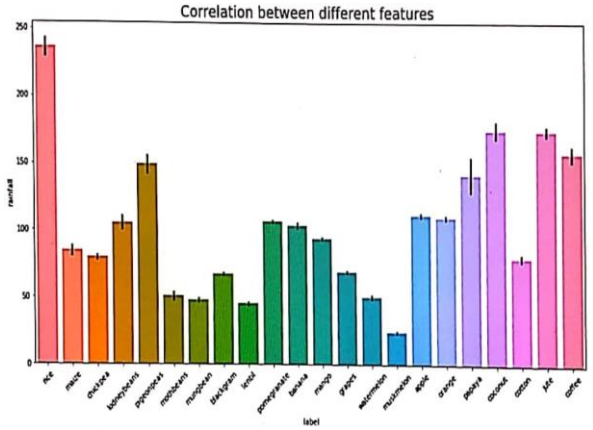

Fig-8: correlation between rainfall and various crops

The below graph with label on x axis and pH in soil on y axis shows us which crop needs what amount of pH.

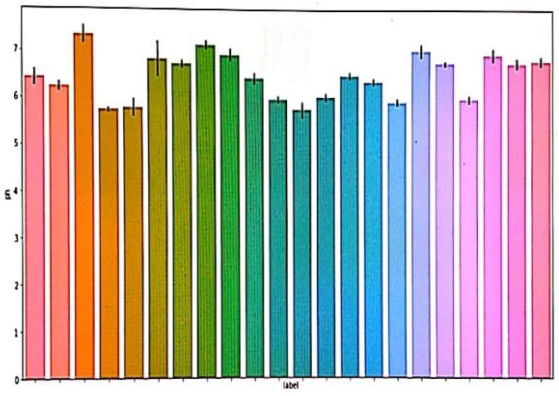

Fig-9: correlation between pH and various crops

The below graph with label on x axis and humidity in percentage on y axis shows us which crop needs what amount of humidity.

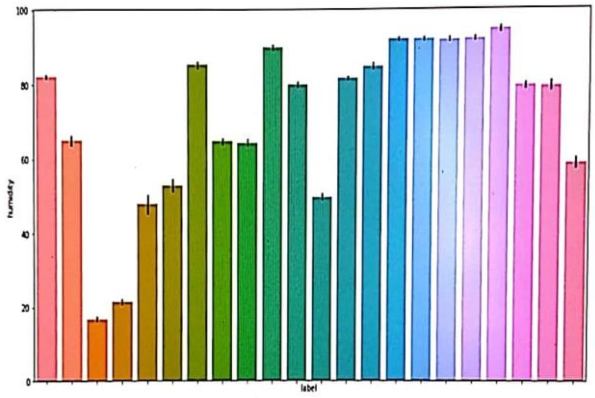

Fig-10: correlation between humidity and various crops

The below graph with label on x axis and temperature in degree Celsius on y axis shows us which crop needs what amount of temperature.

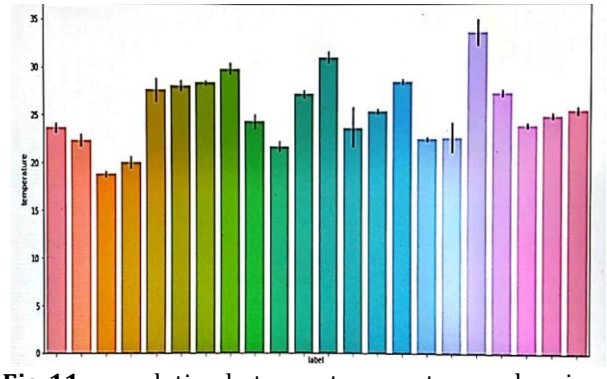

Fig-11: correlation between temperature and various crops

The below graph with label on x axis and ratio of potassium content in soil on y axis shows us how these two are correlated.

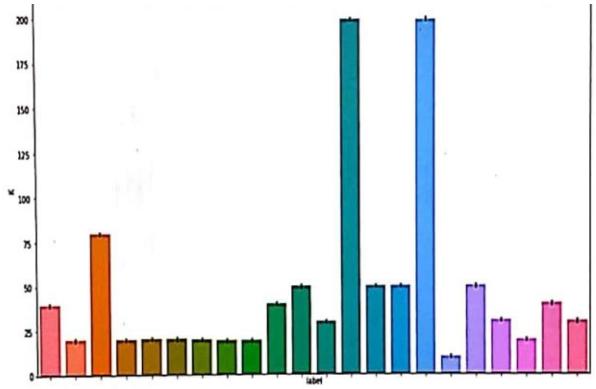

Fig-12: correlation between potassium content in soil and various crops

The below graph with label on x axis and ratio of phosphorous content in soil on y axis shows which crop needs what amount of phosphorus.

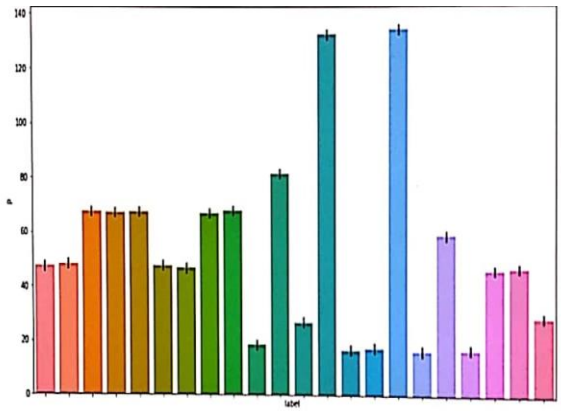

Fig-13: correlation between phosphorus content in soil and various crops

The below graph with label on x axis and ratio on nitrogen content in soil on y axis shows us how these two are correlated.

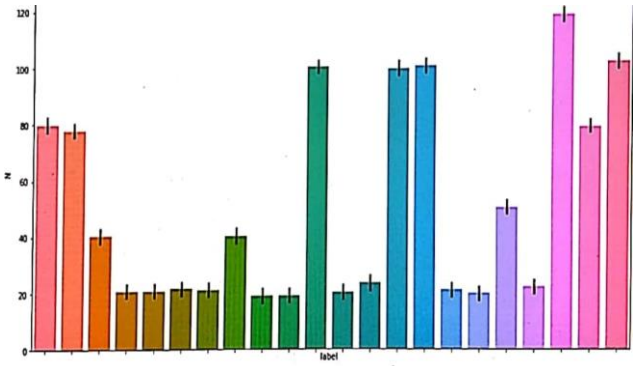

Fig-14: correlation between nitrogen content in soil and various crops

If the model clearly understands the relation between descriptive features and target feature then the model is trained perfectly . Hence the model can now predict the output for different unseen data. Since our model was trained properly it was able to predict the right output for different data. Output was as shown in the figure below.

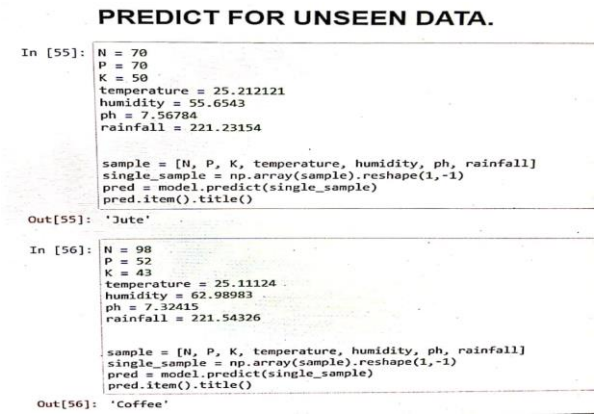

Fig-15: prediction for unseen data

RESULT

We implemented about seven machine learning algorithms in the most efficient manner to recommend crop which gives maximum yield at specific conditions**.** We also plotted graphs to check the relation between label and other features. And we also plotted graph of algorithms with their accuracy. In the graph we could notice that the accuracy of random forest and naive bayes algorithms are highest and the accuracy of logistic regression and support vector machine algorithm are less when compared with accuracy of other algorithms. We were able to train the model using different machine learning algorithms and hence the model can predict the crop which is best suitable for particular conditions of temperature, humidity, rainfall, pH , potassium, phosphorus, nitrogen levels in soil.

3. CONCLUSIONS

Our project can help many farmers in deciding which crop to be grown that will give them the maximum yield. In current system farmers are not connected with any technology and analysis. So there is a huge chance of loss of money. Sometimes wrong selection of crop will effect their income. To reduce this we can develop a website, which will recommend crop . The prediction which the application will make will be more precise if several parameters are taken into consideration and the algorithm to be used for the prediction is supposed to be supervised learning algorithm so that there will be minor or no chance of error as the training is guided by the training model. In future we can further enhance this project by creating a smart chat bot using Artificial Intelligence and Natural language processing technologies. So the farmers need not know computer operation to predict the crop . We can include voice recognition in the smart chat bot so that the farmers who are

illiterate can also comfortably use this project. We can also include all regional languages so that farmers from any corner of our country could benefit from the project.

REFERENCES

[1] PHOSLAB. (n.d.). HOW DOES POTASSIUM HELP PLANTS GROW? Retrieved from PHOSLAB TESTING LABORATORIES: https://www.phoslab.com/how-does-potassium-help-plants-grow/#

[2] AGRICULTURAL NUTRIENT PROFILE: PHOSPHORUS-PART 1. (n.d.). Retrieved from TAURUS: https://taurus.ag/importance-of-phosphorus-to-crops/

[3] PHOSLAB. (n.d.). HOW DOES NITROGEN HELP PLANTS GROW? Retrieved from PHOSLAB TESTING LABORATORIES: https://www.phoslab.com/how-does-nitrogen-help-plants-grow/#

[4]Paudel, D. (n.d.). Machine learning for large-scale crop yield forecasting. Retrieved from ScienceDirect: https://www.sciencedirect.com/science/article/pii/S0308521X20308775

[5] Yulia Gavrilova, HYPERLINK "https://serokell.io/team" \l "olgabolgurtseva" Olga Bolgurtseva (September 23rd, 2020). What Is Data Preprocessing in ML? Retrieved from Serokell: https://serokell.io/blog/data-preprocessing

[6]Mosavi, A. (n.d.). Basic flow for building the machine learning (ML) model. Retrieved from ResearchGate: https://www.researchgate.net/figure/Basic-flow-for-building-the-machine-learning-ML-model_fig2_328609059

[7] B.Manjula Josephine, K. R. (ISSUE 02, FEBRUARY 2020). Crop Yield Prediction Using Machine Learning. INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9.

[8] Kumara, B.A., Kodabagi, M.M., Choudhury, T. et al. Improved email classification through enhanced data preprocessing approach. Spat. Inf. Res. 29, 247–255 (2021). HYPERLINK "https://doi.org/10.1007/s41324-020-00378-y" \t "_blank" https://doi.org/10.1007/s41324-020-00378-y

[9] A. K. B and M. M. Kodabagi, "Efficient Data Preprocessing approach for Imbalanced Data in Email Classification System," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), 2020, pp. 338-341, doi: 10.1109/ICSTCEE49637.2020.9277221.