INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 08 ISSUE: 07 | JULY - 2024 SJIF RATING: 8.448 ISSN: 2582-3930

# Advanced Data and Prediction Model for Optimal Soil Fertility Management for Precision Agriculture

**[1]Priyanka MV, [2]Siddesh K T, [3]Mr.Kotru Swamy SM**

[1]Student, Department Of MCA, BIET, Davangere
[2]Assistant professor,Department Of MCA, BIET, Davangere
[3]Assistant professor,Department Of MCA, BIET, Davangere

**ABSTRACT:** Through the exploitation of scientific knowledge and technology, man has made wonderous strides in the field of automation, the focus of which lies in 'Robotics and Machine Learning'. All around us we see machines taking over work with accuracy and ease. Seen as a subset of artificial intelligence, machine learning relies on data, patterns in data and inference to aid technology in thinking for itself. This paper aims to apply the science of machine learning in the field of agriculture, by carrying soil fertility analysis using most accurate algorithm. The fertility of soil plays a principal role in determining the suitability of cultivating a particular crop on a given soil type. Analysis is carried out by the examination of various properties of the soil like the pH value, Electrical Conductivity, Moisture content, Temperature and (N)Nitrogen (P)Phosphorous (K) Potassium levels, followed up by soil type classification. Finally, a recommendation for the most suitable crop is provided in real time.

Keywords: Ensemble, Machine Learning, Artifical Intelligence

## 1.INTRODUCTION

Agriculture is the science and art of cultivating plants and live-stock. It constitutes one of the most important employment sectors of India as 43% of the workforce is involved in this activity spread over 60% of the land. Contribution in Indian Economy from this sector is about 15% of soil analysis is an important methodology towards a solution as it utilizes factual data concerning the contents present in the soil such as pH value, Electrical Conductivity value, moisture content, Temperature and (N)Nitrogen (P)Phosphorous (K) Potassium levels acquired by soil test sensors.

The data is then used to scientifically conclude the varieties of crops that are to be cultivated to achieve profitable harvest. Soil-to-crop pairing is a very important parameter in the Agricultural industry, as the right crop can mean exponentially higher per capita productivity.

Currently this process, if carried out is done manually in laboratories but by taking this project forward, we hand it over to automation by developing a software based on machine learning concepts like classification, clustering and inference. Analyzing the data in this manner reduces the human effort demanded by manual lab tests as well as plausible human error during analysis.

The developed software additionally generates a written report which documents the measured properties of the soil and the obtained result, altogether attaining effectiveness andaccuracy.

## 2. LITERATURE SURVEY

S. Panchamurthi. M. E, M. D. Perarulalan, A. Syed Hameeduddin, P. Yuvaraj In 4-step methodology is proposed to analyze the soil and predict a suitable crop for agriculture over a particular land. The first step is the procurement of soil moisture, soil temperature

Soil pH of the soil under test using sensors FC-28, DS18B20 and generic soil pH sensor respectively. The second step is the selection and development of a supervised machine learning algorithm that will collect training attributes and target attributes so as to form a relation between the two and classify entirely new inputs (without targets). During classification, the class labels for the given data are predicted. The third step is the feeding of measured data and collected data is fed into the system, compared and matched to a suitable crop variety. The fourth and final step is the recommendation of fertilizers based on the results of the soil test analysis, the nutrient requirement of the crop to be grown and moisture conditions of the field.[1]

Zeel Doshi, Subhash Nadkarni, Rashi Agrawal, Neepa Shah The authors of have developed an intelligent system called Agroconsultant using Big Data Analytics and Machine Learning. This system aims to assist farmers in making informed decisions on crop cultivation based on the sowing season, environmental factors, soil characteristics and geographical location. The crop suitability predictor sub-system acquires a training dataset with the attributes: Soil type, Aquifer thickness, Soil pH, Thickness of topsoil, Precipitation, Temperature and Location parameters.[2]

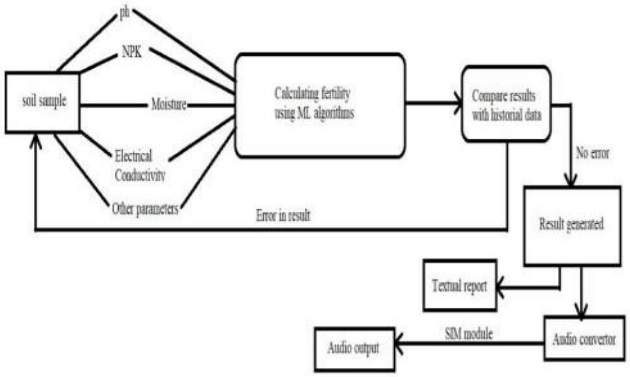
Mr.Ambarish G. Mohapatra, Dr. Bright Keswani, Dr.Saroj Kumar Lenka The data was preprocessed by first replacing missing dot values ('.') with large negative values. This was followed by the creation of class labels. The required labels were generated using production (in tons) and area under cultivation (in hectares) for each crop. Those whose production ÷ area value was greater than 0 were given label 1 and in all other cases, a class label of 0 was assigned. The system was trained using a Multi-labeled classification algorithm since more than one class can be assigned to a given instance. The rainfall predictor sub-system also works in the same way to predict the rainfall in a given state for each of the12 months of the year.

focuses on predicting the required N-P-K content in the soil using random forest algorithm. Additionally, it includes the development of a dynamic web interface to assist the farmers. The process begins by collecting the experimental N-P-K datasets, which are used to develop a random-forest based classification tree in the R server to predict required N-P-K for specific crop and soil type. The R Shiny functionality is adopted to test the prediction and classification model using two components namely ui.R and server.R. The ui.R is a client side user interface which includes as input values the available N-P-K content, soil type, crop type and yield target to produce a resultant output containing the required N-P-K levels to cultivate the inputted crop and its location on the map. The server. R is cloud computing-based data analytics server which comprises of the random forest algorithm required for the N-P-K content prediction, ICAR (Indian Council of Agricultural Research) data extraction from external web server and ggmap forlocation map generation.[3]

Jay Gholap Another approach to analyze the soil, more specifically the fertility of the soil is taken by the authors of A predictive model based on data-mining algorithms is created to process soil properties of a given sample. The concept applied is the decision tree method of classification. After accumulating sufficient data, various data-mining algorithms were applied and compared on the basis of accuracy and error rate leading to the adoption of J48 algorithm (91.90% accuracy). This J48 algorithm was further tuned to improve its accuracy through techniques like attribute selection and boosting, with the help of WEKA (Waikato Environment for Knowledge Analysis). Treating the base learner with attribute selection aids in removing irrelevant and redundant attributes, whose values do not affect the classification of the record. The attribute selection increased the accuracy to 93.20%. Boosting improves the performance of a weak learner by increasing the weights of incorrectly identified instances and decreasing the weights of correctly identified instances over its iterations. This enhanced the accuracy to a strong 96.73%.[4]

Keerthan Kumar T G, Shubha C, Sushma S A Chose on building two separate modules, each testing different characteristics of the soil to attain its fertility analysis and determine a suitable crop

for that land. Module 1 grades so as to provide a better understanding of the soil to the farmers. The soil's micro and macro nutrient content is chosen as the main criteria for testing and are called the feature variables. Regression algorithm is applied along with gradient descent and desired learning rate. Finally, the root mean square between predicted and true value is calculated. Module 2 makes the crop recommendations based on the soil type.[5]




Jayalakshmi, M. Savitha Devi Mainly deals with developing an efficient model to increase the accuracy of soil fertility prediction using Machine learning classifier algorithms. The main purpose of proposed work is to analyze the soil data using data mining techniques. Methodology begins by soil data collection followed by preprocessing of the datasets for screening missing attribute values and noisy data. Data mining techniques are applied on the preprocessed datasets. R tool is used to implement the classification algorithm. C5.0 model is considered as the best classifier technique for soil fertility analysis as its accuracy rate is 96%. C5.0 model divides the sample based on the field having maximum information gain. Each subsample is further divided based on the different field; this process continues till a state where subsamples cannot be further divided. The nodes then categorized together.[6]

## 3. METHODOLOGY

Figure 1 provides a flowchart-based visualization of the methodology of the system.

The process begins by gathering test data from the soil sample using various sensors. The sensors calculate the values such as pH, NPK, moisture, electric conductivity and other parameters. Datasets are used to train and test the system. Training is carried out using previously gathered data and the values observed by the farmers are used to test the system. The results obtained are compared with the historical data. If there is no error during comparison, result is generated as the final output. Further a textual report is generated along with the crop name in the audio format.

Precision agriculture leverages technology to enhance crop production and soil management. Key components include soil fertility analysis and crop prediction. This methodology outlines the steps for conducting these analyses to improve agricultural practices.

**Figure 1. Flowchart-based visualization**

### 3.1 Dataset used:

A dataset containing a diverse dataset of medical imaging data containing pancreatic image, such as CT scansor MRI images, along with corresponding labels indicating cancerous or non-cancerous tissue.

**Figure 2: Data Set**

### 3.2 Data Preprocessing:

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | N | P | K | pH | EC | OC | S | Zn | Fe | Cu | Mn | B | Output |
| 2 | 138 | 8.6 | 560 | 7.46 | 0.62 | 0.7 | 5.9 | 0.24 | 0.31 | 0.77 | 8.71 | 0.11 | 0 |
| 3 | 213 | 7.5 | 338 | 7.62 | 0.75 | 1.06 | 25.4 | 0.3 | 0.86 | 1.54 | 2.89 | 2.29 | 0 |
| 4 | 163 | 9.6 | 718 | 7.59 | 0.51 | 1.11 | 14.3 | 0.3 | 0.86 | 1.57 | 2.7 | 2.03 | 0 |
| 5 | 157 | 6.8 | 475 | 7.64 | 0.58 | 0.94 | 26 | 0.34 | 0.54 | 1.53 | 2.65 | 1.82 | 0 |
| 6 | 270 | 9.9 | 444 | 7.63 | 0.4 | 0.86 | 11.8 | 0.25 | 0.76 | 1.69 | 2.43 | 2.26 | 1 |
| 7 | 220 | 8.6 | 444 | 7.43 | 0.65 | 0.72 | 11.7 | 0.37 | 0.66 | 0.9 | 2.19 | 1.82 | 0 |
| 8 | 220 | 7.2 | 222 | 7.62 | 0.43 | 0.81 | 7.4 | 0.34 | 0.69 | 1.05 | 2 | 1.88 | 0 |
| 9 | 207 | 7 | 401 | 7.63 | 0.59 | 0.69 | 7.6 | 0.32 | 0.68 | 0.62 | 2.43 | 1.68 | 0 |
| 10 | 289 | 8.6 | 560 | 7.58 | 0.44 | 0.67 | 7.3 | 0.63 | 0.66 | 0.94 | 2.43 | 1.79 | 1 |
| 11 | 138 | 8.1 | 739 | 7.55 | 0.33 | 0.78 | 9 | 0.69 | 0.41 | 1.15 | 2.75 | 2 | 0 |
| 12 | 151 | 8.1 | 549 | 7.59 | 0.45 | 0.97 | 9.6 | 0.71 | 0.38 | 1.33 | 2.79 | 2.41 | 1 |
| 13 | 144 | 7.2 | 306 | 7.53 | 0.73 | 0.89 | 9.2 | 0.63 | 0.47 | 1.03 | 2.79 | 2.38 | 0 |
| 14 | 138 | 5.3 | 444 | 7.68 | 0.6 | 0.78 | 9.7 | 0.73 | 0.36 | 1.32 | 3.32 | 2.12 | 2 |
| 15 | 144 | 8.3 | 549 | 7.45 | 0.53 | 0.81 | 10.2 | 0.51 | 0.56 | 1.26 | 2.9 | 2.29 | 0 |
| 16 | 201 | 7.7 | 676 | 7.39 | 0.77 | 0.72 | 9.7 | 0.58 | 0.47 | 1.02 | 3.77 | 2.56 | 1 |
| 17 | 182 | 9.2 | 718 | 7.47 | 0.34 | 0.67 | 10.6 | 0.77 | 0.41 | 1.28 | 3.04 | 2.79 | 0 |
| 18 | 238 | 7.5 | 771 | 7.38 | 0.88 | 0.75 | 11 | 0.46 | 0.38 | 1.16 | 2.96 | 1.32 | 0 |
| 19 | 270 | 8.1 | 655 | 7.45 | 0.55 | 0.67 | 10.2 | 0.28 | 0.44 | 1.26 | 7.75 | 2.56 | 1 |
| 20 | 213 | 6.1 | 803 | 7.38 | 0.78 | 0.61 | 10.5 | 0.3 | 0.49 | 0.66 | 7.74 | 1.85 | 0 |
| 21 | 245 | 8.1 | 560 | 7.31 | 0.63 | 0.78 | 11.6 | 0.29 | 0.43 | 0.57 | 7.73 | 0.74 | 0 |
| 22 | 282 | 8.3 | 454 | 7.43 | 0.62 | 0.75 | 11 | 0.32 | 0.5 | 0.81 | 4.99 | 2.65 | 1 |
| 23 | 213 | 8.3 | 676 | 7.36 | 0.51 | 0.69 | 23.6 | 0.28 | 0.93 | 1.04 | 2.17 | 1.97 | 0 |
| 24 | 207 | 9.4 | 127 | 7.62 | 0.68 | 0.75 | 27.2 | 0.28 | 0.72 | 1.04 | 2.43 | 0.94 | 0 |
| 25 | 307 | 8.3 | 11 | 7.49 | 0.84 | 0.69 | 15.8 | 0.26 | 0.57 | 0.73 | 2.13 | 2.47 | 1 |
| 26 | 207 | 7.2 | 496 | 7.55 | 0.61 | 0.64 | 26.2 | 0.24 | 0.47 | 1.15 | 1.99 | 1.94 | 0 |
| 27 | 201 | 5.5 | 359 | 7.44 | 0.76 | 0.78 | 26 | 0.18 | 0.83 | 0.71 | 2.24 | 1.91 | 0 |

Preprocess the imaging data, including standardization, normalization and potentially augmentation techniques to enhance data quality and facilitate model training. Clean the collected

data by removing noise, irrelevant information, and formatting inconsistencies. Normalize and standardize the data to make it suitable for analysis.

### 3.3 Algorithm discription:

#### 1. SVM (Support Vector Machine)

Support Vector Machine is a supervised machine learning technique. It is used to find a hyperplane in a n-dimensional space that distinctly classifies the data points. The hyperplane should have the maximum margin i.e. plane having maximum distance between the different classification of data points [8]. Support vectors are the data points that are close to the hyperplane. Maximization of margin is done using these support vectors. The main aim of support vector machine is to get non-linear function from kernel function (linear function) [9]. Support vector machine used for crop yield prediction is called support vector regression. The main advantage of SVM is that, it is more effective in high dimensional spaces.

#### 2. K-NN (K- Nearest Neighbors)

K-NN (k-Nearest Neighbors) algorithm is a supervised learning algorithm. K-NN can be used for both classification and regression. It is deployed on learning technique where it considers all the previous sample space data while predicting the target value for a new input value. When there is a new input sample, it calculates the distance between the new input sample and all the training sample predictors. Euclidean distance is one of the popular methods to calculate the distance in this context [10]. Other distance functions like Minkowski and Manhattan can also be used. The value of "k" is non-parametric and is given as k=sqrt(n)/2, where n stands for the number of samples in the training dataset considered. It is suggested to keep the value of k preferably odd. k-NN is a lazy technique compared to all the machine learning techniques. The use of k-NN in the field of agriculture is given briefly in [11].

#### 3. Random Forest

Random forest method also called as random decision forest is an ensemble learning method [12] for classification, regression and other tasks. It operates by generating multiple tree of randomly sub-sampled features. The average of all the individual multiple trees are taken and given as the output. The advantage of this algorithm is that the predictive performance can compete with the best supervised learning algorithms. Random forest algorithm can be used for the prediction [13]. In case of crop prediction, Random Forest proves to be a better classifier as compared to Gaussian Naïve Bayes [14].

#### 4. Ensemble Technique

Ensemble methods are machine learning techniques that combine various base models aiming to obtain one optimal predictive model [15]. Ensembling uses two frameworks namely dependent and independent. In dependent frameworks the output of one base model depend on the other whereas, in independent framework the base models are independent as the base models work in parallelized manner. Independent framework is most commonly used because of its less execution time. The base models of independent framework will generate class labels these are subjected to majority voting technique to get the final ensembled class label [16]. Ensemble framework consists of basic components such as, a labelled training set of instances that is used in training the model. Base inducer is an algorithm that produces a base model considering the labelled training set as the input to the inducer. Diversity generator is used for the generation of diverse base models. Combiner is responsible for combining the class labels that are obtained from individual base models. The main advantage of ensemble technique is, it allows to produce better predictions compared to a single model.

### 3.4 Techniques:

- **Bootstrap Aggregating (Bagging):** Randomly sample subsets of the training data (with replacement) to train multiple decision trees independently.

- **Hyperparameter Tuning:** Optimize hyperparameters such as the number of trees (n_estimators), maximum depth of each tree

(max_depth), minimum number of samples required to split a node (min_samples_split).

- **Tree Pruning:** Implement pruning techniques such as cost-complexity pruning (or post-pruning) to prevent overfitting and improve the generalization ability of the decision tree model.

- **Ensemble Techniques:** Utilize techniques such as stacking or blending to combine predictions from multiple base models effectively.

- **Feature Randomization:** Randomly select a subset of features for consideration at each node of the decision tree within the Random Forest.

## 4. RESULT

From the above research carried out on survey of various papers as well as analysis of various machine learning algorithms for prediction, we are able to develop the software for crop prediction through soil fertility analysis. The resultant outcome will take a decision for the user regarding a suitable crop-soil pairing. For instance a general expectation can be, that soil with high moisture content would cause the software to select crops like rice and wheat and millet like plants which require clayey soil or soils that lean towards acidic might result in the software to suggest crops like gram and other pulses. The software is being developed keeping farmers in mind. The software will help to test various factors of the soil and suggest the best crop that can be grown by analyzing the properties using Machine Learning algorithms. Aim is to overcome the drawbacks of manual soil testing process by replacing the process with the proposed solution that provides result in real time. The project also replaces the traditional method of soil testing in a effective way, So that the farmers can get to know about the soil quickly for better productivity.


**Fig 4.1: Fertility Classification**

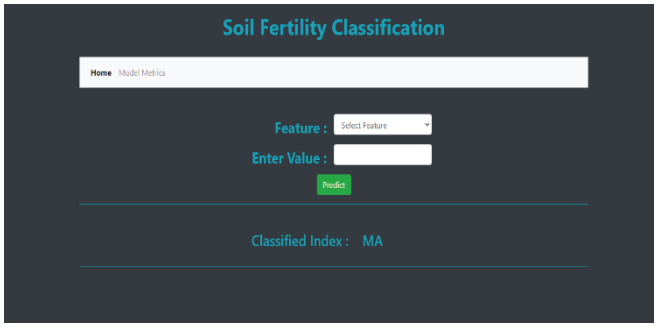



**Figure 4.2: predicted result**

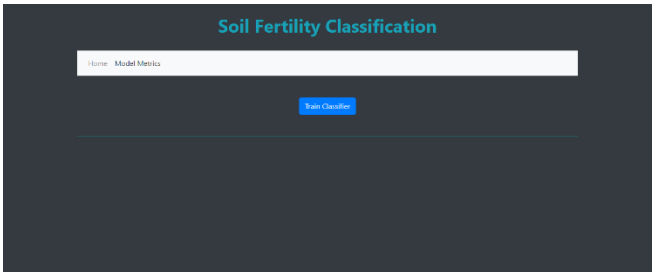



**Fig 4.3: Model Metrics Page**

**Fig 4.4(a): B Feature Score**

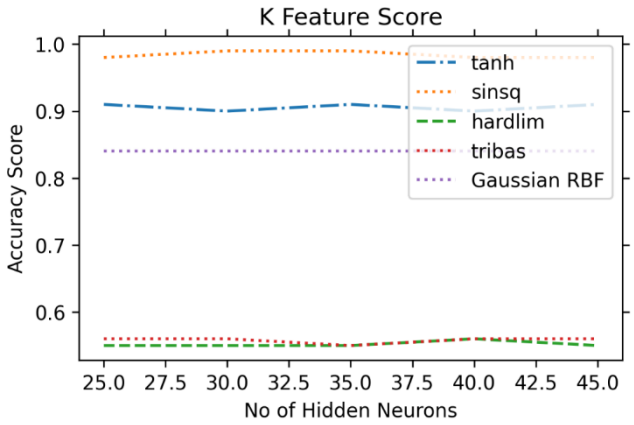



**Fig 4.4(b):K Feature Score**

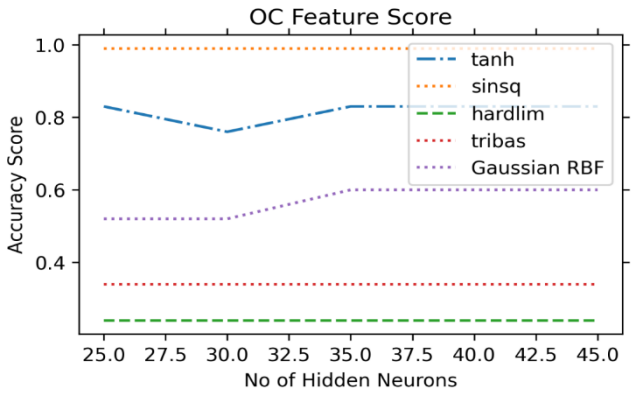
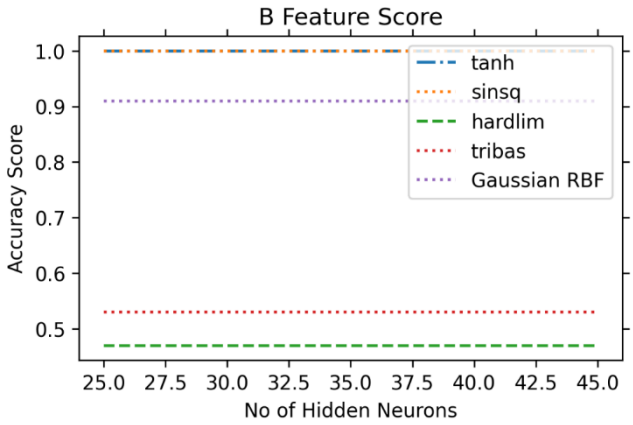



**Fig 4.4(c):OC Feature Score**




## 5. CONCLUSION

In this paper, aim to perform the analysis of agriculture-fit soil, to predict a crop suitable for cultivation in the given soil so as to achieve high and profitable yield. With the intention of inculcating the advancements in technology and computation into this project, the analysis will be done by a system developed based on the concept of machine learning. Machine learning algorithms are applied to datasets collected by various laboratories and research centers. These datasets consist of statistical values of the various properties of the soil along with appropriate crop pairing. On completion of the training, the system will be able to accurately predict the appropriate crop to pair with unknown value attributes. Concurrently a report will be generated giving the user an idea of their soil so that there is a clear understanding about the reason for the selected crop. This report will help users to better understand the soil on which they cultivate crops which in turn leads to the informed irrigation and use of fertilizers in correct amounts. Traditional methods are replaced for accurate and reliable results.

## 6. REFERENCES

1. https://data.worldbank.org/indicator/SL.AGR.EMPL.ZS
2. S. Panchamurthi. M. E, M. D. Perarulalan, A. Syed Hameeduddin, P. Yuvaraj, "Soil Analysis and Prediction of Suitable Crop for Agriculture using Machine Learning", International Journal for Research in Applied Science & Engineering Technology (IJRASET)7(2), 45-48.
3. Zeel Doshi, Subhash Nadkarni, Rashi Agrawal, Neepa Shah, "AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms"6(1), 112-125
4. Mr.Ambarish G. Mohapatra, Dr. Bright Keswani, Dr.Saroj Kumar Lenka, "15(3), 178-192.
5. Jay Gholap, "PERFORMANCE TUNING OF J48 ALGORITHM FOR PREDICTION

OF SOIL FERTILITY"9(4), 231-245.

6. Keerthan Kumar T G, Shubha C, Sushma S A, “Random Forest Algorithm for Soil Fertility Prediction and Grading Using Machine Learning”31(6):39-43.
7. R. Jayalakshmi, M. Savitha Devi “Relevance of Machine Learning
Algorithms on Soil Fertility Prediction Using R”293(10):1239-44.
8. https://towardsdatascience.com/support-vector-machine-introductio n-to-machine-learning-algorithms-934a444fca47
9. Rakesh Kumar, M.P.Singh, Prabhat Kumar, J.P.Singh “Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique”27(7):1-7.
10. https://towardsdatascience.com/how-to-measure-distances-in-machi ne-learning-13a396aa34ce