International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Rapid Soil Testing and Fertilizer Recommendation Using Python

Mohd Imtiyaz Mohiuddin[1], Rabi Hassan[2], Zaki Ahmed Siddiqui[3], Mr. Syed Sultan Mahmood [4], Mrs.Zubeda Begum[5]

[1, 2,3] UG students, Department of ECE, ISL ENGINEERING COLLEGE, Bandlaguda, Chandrayangutta, Hyderabad, Telangana, India-500005.

[4,5] Asistant Professor, Department of ECE, ISL ENGINEERING COLLEGE, Bandlaguda, Chandrayangutta, Hyderabad, Telangana, India-500005.

ABSTRACT:

Machine learning is an emerging research field in crop yield analysis. Yield prediction is a very important issue in agriculture. Any farmer is interested in knowing how much yield he is about to expect. In the past, yield prediction was performed by considering farmer's experience on particular field and crop. The yield prediction is a major issue that remains to be solved based on available data. Machine learning techniques are the better choice for this purpose. Different Machine learning techniques are used and evaluated in agriculture for estimating the future year's crop production. This paper proposes and implements a system to predict crop yield from previous data. This is achieved by applying machine learning algorithms like Support Vector Machine and Random Forest on agriculture data and recommends fertilizer suitable for every particular crop. The paper focuses on creation of a prediction model which may be used for future prediction of crop yield. It presents a brief analysis of crop yield prediction using machine learning techniques.

INTRODUCTION

Low crop production is inappropriate soil fertility management [1]. The relatively low yield of common beans in Ethiopia including the study area is as a result of the low use of enhanced variety and low soil fertility status and poor management practice which are the major production constraints [2]; ATA [3]. Wossen [4] reported NPKSB nutrient ratios for common bean growth and yield. However, soil fertility mapping projects in Ethiopia recently reported the deficiency of K, S, Zn, B, and Cu in addition to N and P in major Ethiopian soils, and thus recommend the application of customized and balanced fertilizers [5]. Common bean N fertilizer requirement depends on soil fertility levels; for low soil nitrogen levels (below 34 kg $N \cdot ha^{-1}$) N fertilizer is generally recommended in order for deficiency symptoms not to manifest and for full development up to production, inorganic phosphorus fertilizer has a positive effect on the yield and yield components of common bean and sulfur is required in similar amount as that of phosphorus [6]. In Ethiopian soils, the deficiency of K, S, Zn, B, and Cu in addition to N and P have common bean recently reported by soil fertility mapping project, and thus recommend the application of customized and balanced fertilizers [11]. The fertilizer recommendation did not consider the existing soil nutrient supply and resulted in low crop yield response in the region [2].

Soils in agriculture are an important part of the ecological system that produces food and fiber for human consumption, but they are a limited and largely non-renewable resource [1, 2]. Soils are a key enabling resource and essential to the production of a wide range of goods and services integral to ecosystems and human well-being [3, 4]. Nonetheless, soil fertility depletion caused by a variety of factors (soil erosion, acidity, nutrient depletion, lack of soil fertility replenishment, nutrient mining, and lack of balanced fertilization) is a significant contributor to food insecurity [5, 6].

Soil quality (SQ), which is defined as the capacity of soil to function within the ecosystem and land use boundaries to sustain biological productivity, maintain environmental quality, and promote plant, animal, and human health, is now highly related to sustainable and productive agriculture [2, 7, 8]. Good-quality soils will preserve natural ecosystems by improving air and water quality for improved food and fiber production

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

while also protecting the environment and human health [9].

The SQ simultaneously addresses the issues of productivity and sustainability and makes it indispensable for developing countries such as Ethiopia [2, 4]. A better understanding of the SQ and the factors that degrade the SQ is necessary to fully exploit the potential benefits of soil resources. For example, poor soil physical and chemical health is very likely to result in poor aggregate stability, a decline in soil OM, nutrient-related plant stresses, crop yield stagnation, and exacerbate soil degradation [10, 11]. This suggests that SQ is linked to chemical properties, biophysical environments, and anthropogenic factors. Meanwhile, SQ cannot be measured directly in the field or laboratory; rather, it is inferred from measured soil physical, chemical, and biological properties and is thus expressed in terms of soil quality index (SQI) [2, 8, 12].

The SQI could be defined as a minimum set of parameters that provides numerical data about a soil's ability to perform one or more functions [13]. It aids in assessing overall soil condition and management response or resilience to natural and anthropogenic forces [1, 7, 14, 15]. Expert opinion (subjective) or mathematical and statistical (objective) methods are used to select a minimum soil data set (MDS) [13, 16]. The use of multivariate techniques of principal component analysis (PCA) (multiple correlations and factor analyses) to reduce statistical data has become more common [12, 17]. Thus, the SQI, which takes into account the physical, chemical, and biological properties of soils as well as their variability, is critical for long-term utilization and site-specific management of soil resources [2, 8, 12, 14, 15].

LITERATURE REVIW:

Among contemporary farmers, precision agriculture (PA) [1] is a well-known and improved approach of farm management. Crop health and output are monitored via the application of agricultural and information technologies in precision agriculture. PA aims to lower agricultural input costs while retaining the quality of the final output. Bulk applications of chemical fertilizers and pesticides have long been the norm, with the whole field being treated as a single unit.

The UN Council has advocated increasing the global supply of high-quality food as a means of achieving this goal. As a result, new approaches are needed to deal with the problem. One method of dealing with the issue is to forecast human population and agricultural yields.

It is advantageous for policymakers and farmers alike to be able to precisely estimate crop yields throughout the growing season since it allows them to anticipate market prices, plan import and export operations, and limit the social cost of crop losses. In addition to large agricultural companies and smallholders, agricultural enterprises and smallholders profit from such predictions since they are able to make educated decisions about the management and financing of their crops [2]. Because of the complexities of the data, crop production forecasting is a difficult assignment for policymakers to do successfully. Agricultural researchers and agroeconomists are always on the hunt for new mathematical strategies that might increase prediction accuracy while still making use of existing factors. In this field of study, the goal is to demonstrate how crop yields are related to the location of agriculture, while also taking into account environmental elements such as soil quality and irrigation systems. These models are built on the foundation of rule-based models with parameters. A solid understanding of the many linkages that may be established between agricultural methods and environmental circumstances is possessed by the professionals working in the project.

There is a problem with trying to build an empirical expert system using knowledge that cannot be characterized in such a way. Manual surveys and remote sensing data are used to forecast crop yield [3]. Mathematical studies based on previous years' observations and historical information may be useful for a certain region or country but cannot be applied universally. These problems have been handled by recent crop simulation model developments. Models of soil characteristics, climatic conditions, and crop management practices are used to simulate crop growth throughout the growing season in crop simulation. These modeling approaches need a massive data set in order to accurately estimate agricultural output [4] over large areas. It is common for researchers to use remote-sensing devices like satellites, planes, or even a simple camera.

Math, information theory, statistics, artificial intelligence, etc., all play a role in the field of machine learning because it is an interdisciplinary study. The primary goal of machine learning research is to develop fast and effective algorithms that can predict

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

data. In data analytics, machine learning is a technique used to build predictive models of the data collected. Reinforcement, unsupervised, and supervised learning are the three main categories of machine learning tasks.

It is possible for machines to learn behavior based on the input they receive from encounters with the outside world through reinforcement learning. Data sets that have not been labeled using traditional methods such as cluster analysis can benefit greatly from the use of unsupervised machine learning. There must be labeled data for supervised machine learning to work. Each set of labeled training data has an input value and a desired target output value. You can use this inferred function as a basis for mapping new values into the training data using a supervised learning technique. Reinforcement learning is preferred for solving decision-making problems, while unsupervised and supervised learning are both preferred for analyzing data from a data processing perspective.

In a study by Jasti et al. [5], ICT activities provided the most current information, advanced technology, and knowledge for farmers to improve their livelihoods and increase their productivity. The use of cutting-edge communication channels such as radio, mobile, and television for farmers' development and information evaluation is the primary focus of ICT's relevance. According to Hemamalini eta l. [6], farmers and individuals who have an interest in farming can benefit greatly from the use of information technology in the agricultural industry. Technology-enabled decision-making, increased productivity, and real-time communication are all critical for farmers. This is made possible by mobile communication tools used to provide market information, better weather forecasts, cross-market coordination, and a better understanding of agricultural market prices.

Individuals, businesses, and governments all rely on these data-driven models to make predictions. The food sector is presently developing machine learning approaches that can handle the complexity and unpredictability of input [7].

According to Mohamed et al. [8], several studies have examined the difficulties connected with the deployment and optimization of big data applications in various machine learning algorithms employed by cloud data centers or their networks. Big data applications and cloud services were developed using a MapReduce programming methodology with an open-source platform and Hadoop. A variety of innovative analyses and computations were made possible by combining Hadoop with MapReduce. Commodity clusters in geographically dispersed data centers are typically used by these services to deliver elastic and cost-effective solutions. However, as the number of people using data centers to move, store, and analyze big data has risen, this has created a variety of new problems that highlight the necessity of finding ways to lease resources more cheaply and efficiently. As a result, companies that have a large number of tenants requesting big data services have been challenged by the requirement to optimize the leasing of resources in order to minimize excess or under-utilization. To give a comprehensive overview of cloud computing's architecture, a new summary of big data programming paradigms and their applications was selected for this study. Any software-defined networking technology supporting big data systems and virtualization was included in this. The topologies and routing protocols as well as the traffic characteristics were also briefly reviewed to underline the consequences of big data, such as supporting networks and cloud data centers. A number of initiatives were undertaken to improve the performance and energy efficiency of big data systems for a variety of applications and measures of performance. It was the goal of the survey to compile all the relevant research and classify them by the level of data center, network, and application. There were also suggestions for future research.

Due to a lack of natural resources, information, knowledge, and data must be utilized to the fullest extent possible. The conversion of solar energy into chemical energy occurs, for example, during the process of photosynthesis. It is the soil's ability to store and distribute critical nutrients that will allow plants to thrive, and this process will be accountable for all forms of life on the planet. Because overexposure can lead to soil degradation, it is essential to use a fertilizer to preserve the quality of the soil. This makes soil analysis an excellent way to gauge the condition of the soil. If there are minimal or disorganized data, a soil analysis can assist provide a report by examining the soil in several laboratories. In order to provide fertilizer recommendations based on the existing composition of soil nutrition, many methods of machine learning analysis are being applied in this study. The results of soil testing at Tata's soil and water testing center were used in this investigation. The Hadoop Distributed File System (HDFS) was utilized to analyze the stochastic gradient descent (SGD) algorithm and the artificial neural network (ANN). Random forest (RF), SVM using RBF, K-nearest neighbors (KNNs), support vector

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

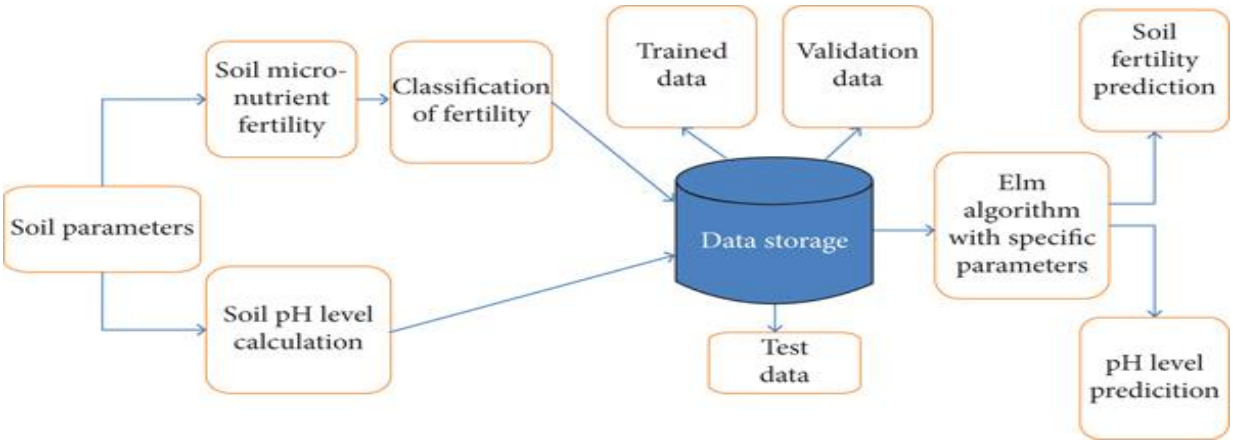
machine (SVM) utilizing polynomial function, and the regression tree (RT) were used to assess performance. Overall, the experimental analysis was performed correctly. Receiver operating characteristics with AUC (ROC) curve, coefficient of determination (R2), root mean square error (RMSE), and mean absolute percentage error (MAPE) measures of validation were used. SGD was found to outperform all other techniques in a study of diverse solution classes. The results also backed up the choice of the remedy and its recommendation.

CONCEPT:

Despite the importance of SQ assessment, very few studies have been conducted on smallholder arable lands in Ethiopia where traditional practices dominate soil management [2]. This emphasizes the importance of having adequate soil property information in order to intervene and prevent soil fertility degradation problems. Against this backdrop, the present study aimed to explore the soil quality status of farmlands belonging to different soil groups using different varied approaches. The process involved three main steps: (i) selecting appropriate indicators; (ii) converting indicators into scores; and (iii) combining the scores into an index [13, 25].where RSTC = assigned ranking values for soil textural class; RpH = assigned ranking values for soil pH; ROC = assigned ranking values for soil organic carbon; RNPK = assigned ranking values for nitrogen (N); phosphorus (P), and potassium (K) (Table 1). Furthermore, $a = 0.2$, $b = 0.1$, $c = 0.4$, and $d = 0.3$ refer to the weighted values corresponding to each of the four parameters. That is, out of 1(100%), the weighting value for soil textural class (a) = 0.2 (20%), soil pH (b) = 0.1(10%), soil organic carbon (c) = 0.4 (40%), and soil macronutrient contents (N, P, and K) (d) = 0.3 (30%).

DESIGN METHODLOGY:

BLOCK DIAGRAM



Soil Fertility/Nutrient/Index

The calculation is based on the number of samples classified as low, medium, or high and the rating classes of the measured soil parameters, which are multiplied by 1, 2, and 3, respectively. If the index value is less than 1.67, the fertility status is low; if the index value is between 1.67–2.33, the fertility status is medium; and if the index value is greater than 2.33, then the fertility status is high [25].where N_L = number of samples in low category; N_M = number of samples in the medium category; N_H = number of samples in high category, and N_T = total number of samples.

Principal Component Analysis (PCA) Based SQI (Statistical Model-Based SQI)

A statistics-based model is used to estimate SQI using PCA [17, 26]. The PCA method is more objective because it makes use of a variety of statistical tools (multiple correlation, factor, and analyses), which could prevent bias and data redundancy by selecting a minimal dataset (MDS) using formulas [12]. The PCA model included all the original observations of each soil parameter.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

The PCs with high eigenvalues represented the maximum variation in the dataset, while most studies have assumed to examine PCs only the variables having high factor loadings with eigenvalues >1.0 that explained at least 5% of the data variations were retained for indexing [12, 17].

Under a given PC, each variable had a corresponding eigenvector weight value or factor loading. Only the "highly weighted" variables were retained in the MDS. The "highly weighted" variables were defined as the highest weighted variable under a certain PC and absolute factor loading value within 10% of the highest values under the same PC [12, 23]. However, when more than one variable was retained under a particular PC, a multivariate correlation matrix is used to determine the correlation coefficients between the parameters. If the parameters were significantly correlated ($r > 0.70$), then the one with the highest loading factor was retained in the MDS and all others were eliminated from the MDS to avoid redundancy.

Still, the normalized PCA of SQI would be calculated if more than one highest eigenvectors were retained in the MDS [12, 23]. The noncorrelated and highly weighted parameters under a particular PC were considered important and retained in the data. Each PC explained a certain amount of variation in the dataset, which was divided by the maximum total variation of all the PCs selected for the MDS to get a certain weightage value under a particular PC [12, 26]. Thereafter, the SQI-3 (PCA) was computed using the following equation:

where PC Weight is the weightage factor determined from the ratio of the total percentage of variance from each factor to the maximum cumulative variance coefficients of the PC considered; individual soil parameter score is the score of each parameter in the MDS.

ALGORITHMS:

Support Vector Machine

SVM develops a hyperplane or set of hyper planes in a high-or boundless dimensional space, which can be utilized for characterization, relapse, or different errands. Naturally, a great partition is accomplished by the hyperplane that has the biggest separation to the closest preparing information purpose of any class, since by and large the bigger the edge the lower the speculation blunder of the classifier. The computational burden have to be reasonable, the mappings are utilized by the SVM plan to guarantee the tiny items will be figured as far as the variable in the first degree, for that a bit capacity k(x, y) chose to get the ideal computational time.

Advantages

1) SVM calculation has a regularization parameter, which stays away from over-fitting.

2) SVM calculation utilizes the portion trap, so you can construct master learning about the issue.

Random Forest

Random forest is a supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithms of the same type. Random Forest algorithm can be used for classification and regression problems.

Advantages

1) The random forest algorithm is not biased, since, there are multiple trees and each tree is trained on a subset of data.

2) Random Forest algorithm is stable if a new data point is introduced in the dataset the overall algorithm is not affected.

XGBOOST:

XGBoost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XGBoost stands for "Extreme Gradient Boosting" and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression. One of the key features of XGBoost is its efficient handling of missing values, which allows it to handle real-world data with missing values without requiring significant pre-processing. Additionally, XGBoost has built-in

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

support for parallel processing, making it possible to train models on large datasets in a reasonable amount of time. XGBoost can be used in a variety of applications, including Kaggle competitions, recommendation systems, and click-through rate prediction, among others. It is also highly customizable and allows for fine-tuning of various model parameters to optimize performance.

Advantages:

1. Performance: XGBoost has a strong track record of producing high-quality results in various machine learning tasks, especially in Kaggle competitions, where it has been a popular choice for winning solutions.
2. Scalability: XGBoost is designed for efficient and scalable training of machine learning models, making it suitable for large datasets.
3. Customizability: XGBoost has a wide range of hyperparameters that can be adjusted to optimize performance, making it highly customizable.
4. Handling of Missing Values: XGBoost has built-in support for handling missing values, making it easy to work with real-world data that often has missing values.
5. Interpretability: Unlike some machine learning algorithms that can be difficult to interpret, XGBoost provides feature importances, allowing for a better understanding of which variables are most important in making predictions.

GBM

Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met. In contrast to AdaBoost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of the predecessor as labels. There is a technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees). The below diagram explains how gradient-boosted trees are trained for regression problems.

STACKING CLASSIFIER:

Ensemble learning is one of the most powerful machine learning techniques that use the combined output of two or more models/weak learners and solve a particular computational intelligence problem. E.g., a Random Forest algorithm is an ensemble of various decision trees combined.

Stacking

Stacking is one of the popular ensemble modeling techniques in machine learning. Various weak learners are ensembled in a parallel manner in such a way that by combining them with Meta learners, we can predict better predictions for the future.

This ensemble technique works by applying input of combined multiple weak learners' predictions and Meta learners so that a better output prediction model can be achieved.

In stacking, an algorithm takes the outputs of sub-models as input and attempts to learn how to best combine the input predictions to make a better output prediction.

Stacking is also known as **a stacked generalization** and is an extended form of the Model Averaging Ensemble technique in which all sub-models equally participate as per their performance weights and build a new model with better predictions. This new model is stacked up on top of the others; this is the reason why it is named stacking.

Architecture of Stacking

The architecture of the stacking model is designed in such as way that it consists of two or more base/learner's models and a meta-model that combines the predictions of the base models. These base models are called level 0 models, and the meta-model is known as the level 1 model. So, the Stacking ensemble method includes **original (training) data, primary level models, primary level prediction, secondary level model, and final prediction**. The basic

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

architecture of stacking can be represented as shown below the image.

Figure 1: Representing the Architecture of Stacking

- **Original data:** This data is divided into n-folds and is also considered test data or training data.
- **Base models:** These models are also referred to as level-0 models. These models use training data and provide compiled predictions (level-0) as an output.
- **Level-0 Predictions:** Each base model is triggered on some training data and provides different predictions, which are known as **level-0 predictions.**
- **Meta Model:** The architecture of the stacking model consists of one meta-model, which helps to best combine the predictions of the base models. The meta-model is also known as the **level-1 model**.
- **Level-1 Prediction:** The meta-model learns how to best combine the predictions of the base models and is trained on different predictions made by individual base models, i.e., data not used to train the base models are fed to the meta-model, predictions are made, and these predictions, along with the expected outputs, provide the input and output pairs of the training dataset used to fit the meta-model.

Steps to implement Stacking models:

There are some important steps to implementing stacking models in machine learning. These are as follows:

- Split training data sets into n-folds using the **RepeatedStratifiedKFold** as this is the most common approach to preparing training datasets for meta-models.
- Now the base model is fitted with the first fold, which is n-1, and it will make predictions for the nth folds.
- The prediction made in the above step is added to the x1_train list.
- Repeat steps 2 & 3 for remaining n-1folds, so it will give x1_train array of size n,
- Now, the model is trained on all the n parts, which will make predictions for the sample data.
- Add this prediction to the y1_test list.
- In the same way, we can find x2_train, y2_test, x3_train, and y3_test by using Model 2 and 3 for training, respectively, to get Level 2 predictions.
- Now train the Meta model on level 1 prediction, where these predictions will be used as features for the model.
- Finally, Meta learners can now be used to make a prediction on test data in the stacking model.

Stacking Ensemble Family

There are some other ensemble techniques that can be considered the forerunner of the stacking method. For better understanding, we have divided them into the different frameworks of essential stacking so that we can easily understand the differences between methods and the uniqueness of each technique. Let's discuss a few commonly used ensemble techniques related to stacking.

IMPLEEMNTATION RESULTS:

The aim of proposed system is to help farmers to cultivate crop for better yield. The crops selected in this work are based on important crops from selected location. The selected crops are Rice, Jowar, Wheat, Soyabean, and Sunflower, Cotton, Sugarcane, Tobacco, Onion, Dry Chili etc. The dataset of crop yield is collected from last 5 years from different sources. There are 3 steps in proposed work. 1) Soil Classification: Soil classification can be done using

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

soil nutrients data. Two Machine learning algorithms used for soil classification are Random Forest and Support Vector Machine. The two algorithms will classify, and display confusion matrix, Precision, Recall, f1-score and average values, and at the end accuracy in percentage as output. 2) Crop Yield Prediction: Crop Yield Prediction can be done using crop yield data, nutrients and location data. These inputs are passed to Random Forest and Support Vector Machine algorithms. These algorithms will predict crop based on present inputs. 3) Fertilizer Recommendation: Fertilizer Recommendation can be done using fertilizer data, crop and location data. In this part suitable crops and required fertilizer for each crop is recommended. • Third Party applications are used to display Weather information, Temperature information as well as Humidity, Atmospheric Pressure and overall description.

Out[4]:

	Temparature	Humidity	Moisture	Soil Type	Crop Type	Nitrogen	Potassium	Phosphorous	Fertilizer Name
0	26	52	38	Sandy	Maize	37	0	0	Urea
1	29	52	45	Loamy	Sugarcane	12	0	36	DAP
2	34	65	62	Black	Cotton	7	9	30	14-35-14
3	32	62	34	Red	Tobacco	22	0	20	28-28
4	28	54	46	Clayey	Paddy	35	0	0	Urea

Figure 2: Representing the overall Dataset for fertilizer recommend based on soil

Out[6]:

	Temparature	Humidity	Moisture	Nitrogen	Potassium	Phosphorous
count	98.000000	98.000000	98.000000	98.000000	98.000000	98.000000
mean	30.204082	59.020408	43.102041	18.704082	3.418367	18.795918
std	3.431086	5.722049	11.301391	11.477642	5.834352	13.412534
min	25.000000	50.000000	25.000000	4.000000	0.000000	0.000000
25%	28.000000	54.000000	34.000000	10.000000	0.000000	9.250000
50%	30.000000	60.000000	41.000000	13.000000	0.000000	19.000000
75%	33.000000	64.000000	50.000000	24.000000	7.750000	30.000000
max	38.000000	70.000000	65.000000	42.000000	19.000000	42.000000

Figure 3: Representing the overall Description of Dataset based on describe function

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

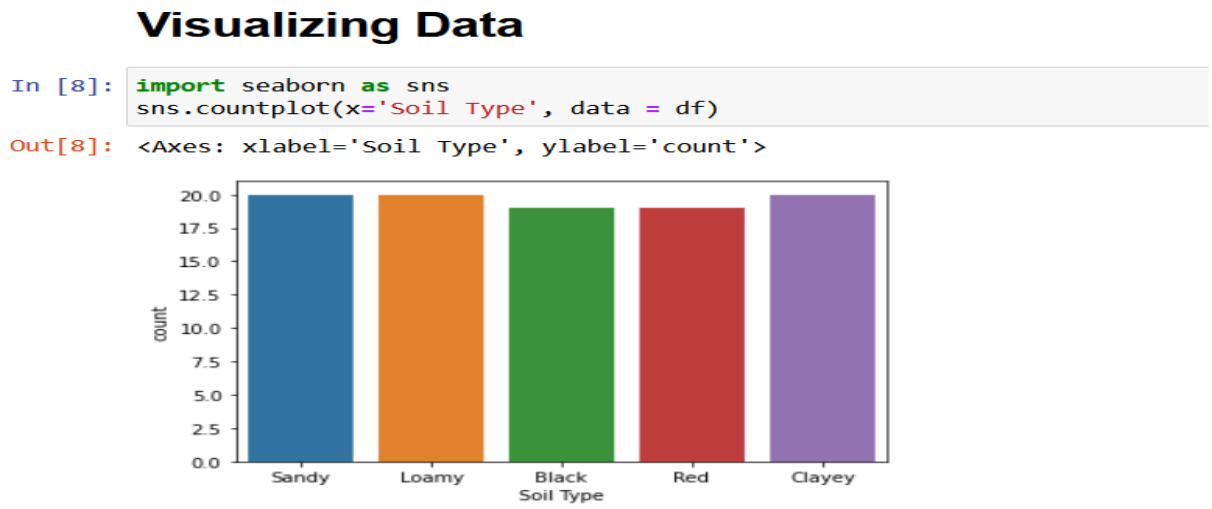


Figure 4: Representing the overall Description of soil based features based on seaborn bar plot function

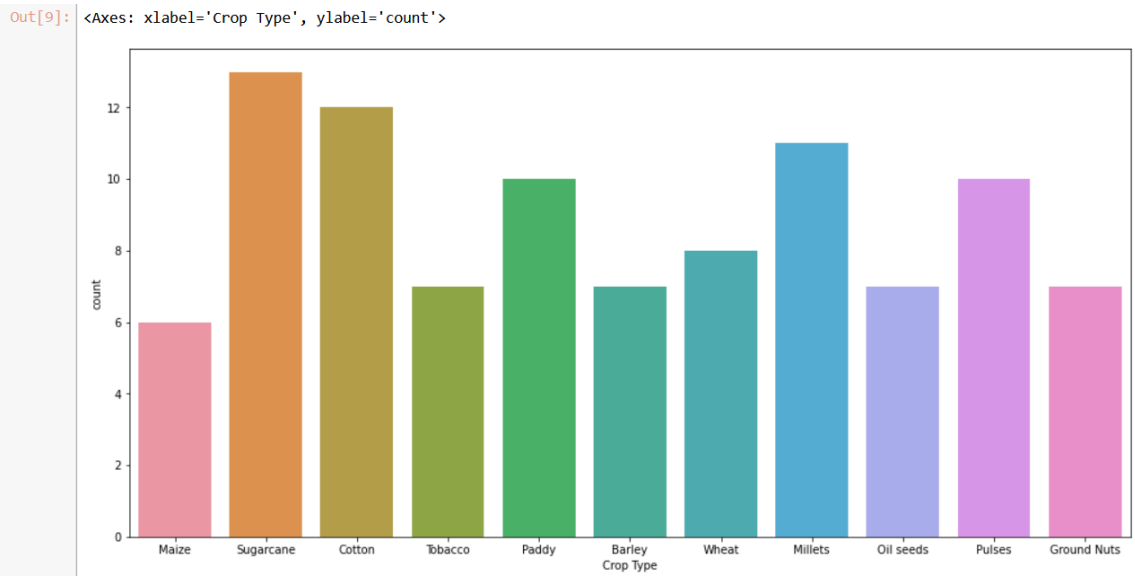


Figure 5: Representing the overall Description of crop based features based on seaborn bar plot function

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Out[16]: <Axes: xlabel='Soil Type', ylabel='Temparature'>

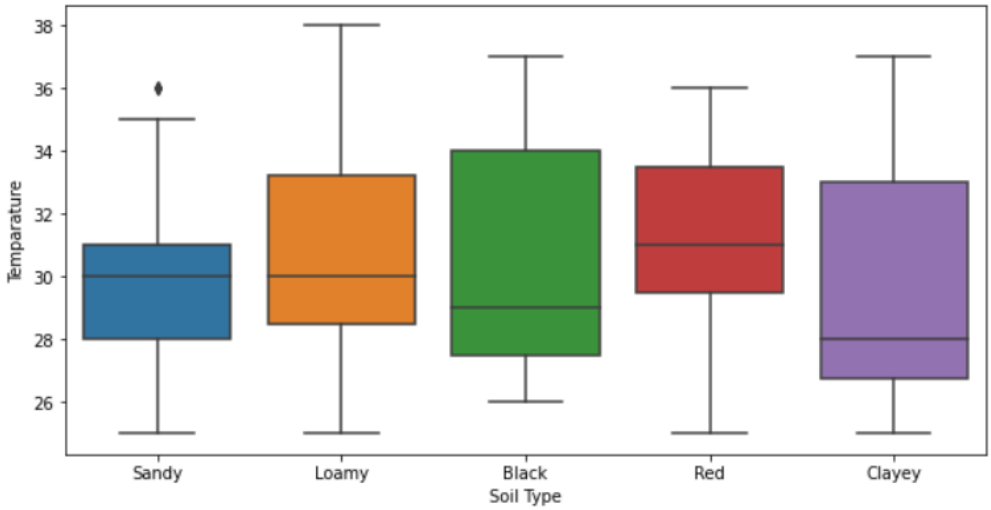
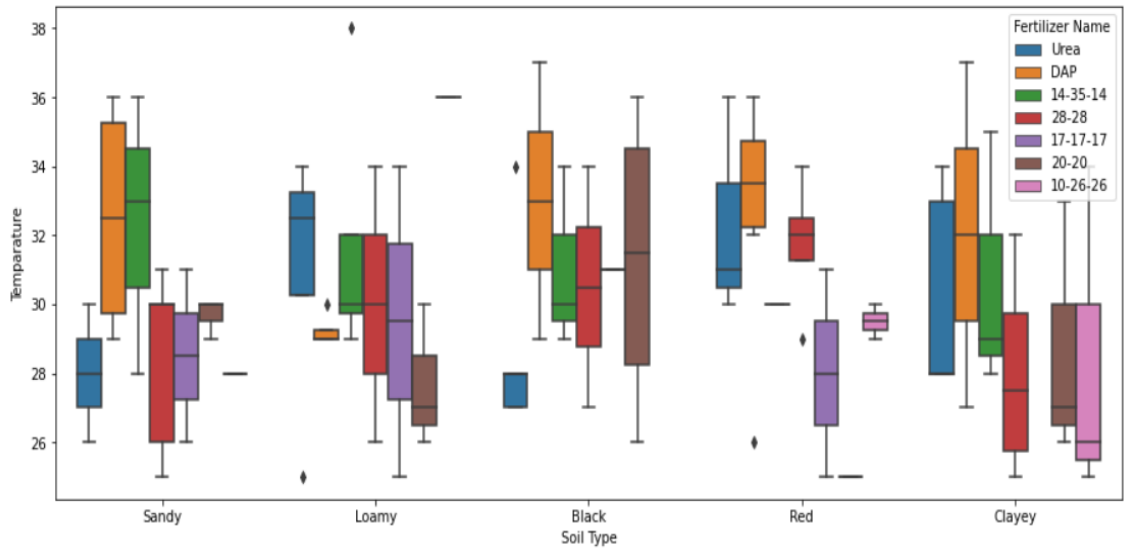


Figure 6: Representing the overall Description of soil based features based on seaborn bar plot function with temperature

<Axes: xlabel='Soil Type', ylabel='Temparature'>



International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Figure 7: Representing the overall Description of soil based features based on seaborn bar plot function with fertilizer types

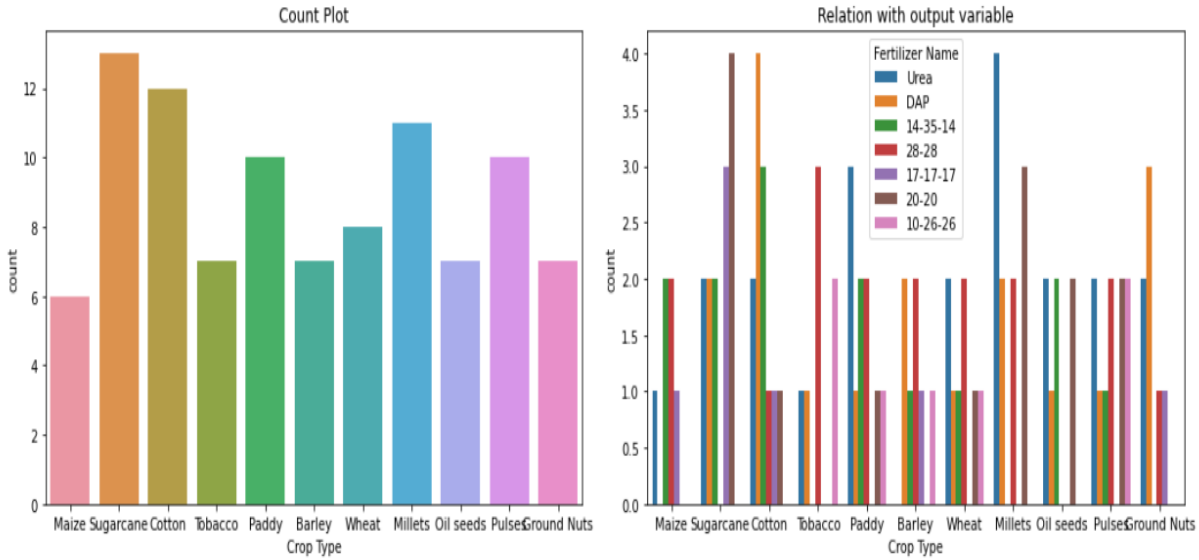


Figure 8: Representing the overall Description of count and bar plot for crops type

Preprocessing using One-Hot Encoder ¶

```
In [25]: y = df['Fertilizer Name'].copy()
         X = df.drop('Fertilizer Name', axis=1).copy()

In [26]: from sklearn.compose import ColumnTransformer
         from sklearn.preprocessing import OneHotEncoder
         ct = ColumnTransformer(transformers=[('encoder', OneHotEncoder(), [3,4])],
         X = np.array(ct.fit_transform(X))

In [27]: X[0]

Out[27]: array([ 0.,  0.,  0.,  0.,  1.,  0.,  0.,  0.,  1.,  0.,  0.,  0.,  0.,
                 0.,  0.,  0., 26., 52., 38., 37.,  0.,  0.])
```

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Figure 9: Representing the One hot function for labeling of dataset with label types

Train-test split

```
In [28]: from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8, shu
```

Feature Scaling

```
In [29]: from sklearn.preprocessing import StandardScaler
         sc = StandardScaler()
         X_train = sc.fit_transform(X_train)
         X_test = sc.transform(X_test)

In [30]: X_train[0]

Out[30]: array([-0.46770717, -0.46770717, -0.52790958, -0.50800051,  1.89426379,
                -0.28867513, -0.3380617 , -0.31399291,  3.82099463, -0.38348249,
                -0.23249528, -0.31399291, -0.31399291, -0.40519021, -0.26171196,
                -0.3380617 ,  0.72127432,  0.79246323,  0.72066886, -1.15755769,
                 0.8197544 ,  0.74890405])
```

Figure 10: Representing the training and Testing features with Feature scalling

```
encode_ferti = LabelEncoder()
df['Fertilizer Name'] = encode_ferti.fit_transform(df['Fertilizer Name'])

#creating the DataFrame
Fertilizer = pd.DataFrame(zip(encode_ferti.classes_,encode_ferti.transform
Fertilizer = Fertilizer.set_index('Original')
Fertilizer
```

Original	Encoded
10-26-26	0
14-35-14	1
17-17-17	2
20-20	3
28-28	4
DAP	5
Urea	6

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Figure 11: Representing the overall Description of fertilizer types and its encoded values

	Temparature	Humidity	Moisture	Soil Type	Crop Type	Nitrogen	Potassium	Phosphorous	Fertilizer Name
0	26	52	38	4	3	37	0	0	6
1	29	52	45	2	8	12	0	36	5
2	34	65	62	0	1	7	9	30	1
3	32	62	34	3	9	22	0	20	4
4	28	54	46	1	6	35	0	0	6

```
model = pickle.load(open('randomgirdcv.pkl','rb'))
ans = model.predict([[34.0, 65.0, 62.0, 1, 1, 7.0, 9.0, 30.0]])
if ans[0] == 0:
    print("10-26-26")
elif ans[0] ==1:
    print("14-35-14")
elif ans[0] == 2:
    print("17-17-17*")
elif ans[0] == 3:
    print("20-20")
elif ans[0] == 4:
    print("28-28")
elif ans[0] == 5:
    print("DAP")
else:
    print("Urea")
```

```
14-35-14
```

Figure 12: Representing the overall recommended fertilizer with soil type classification

CONCLUSIONS

Agribusiness crop yields may be increased by carefully selecting the right crops and putting in place supportive infrastructure. Weather, soil fertility, water availability, water quality, crop pricing, and other factors are taken into consideration while making agricultural predictions. Machine learning is critical in crop production prediction because it can anticipate crop output based on factors such as location, meteorological conditions, and season. The use of this tool assists farmers in making informed decisions about which crops to grow on their land. It is stated in this paper that a machine learning framework for agricultural yield prediction may be used. The data set

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

for an experiment contains information about the crop. The proposed classifiers with Stacking and XGBOOST classifiers are being used for classification. It is clear that the feature selection and feature extraction have improved the performance of machine learning algorithms. When compared with other classifiers, XGBOOST with Grid Search has best training accuracy of 100 and test at 90%.

SCOPE:

The work can be extended further to add following functionality. Mobile application can be build to help farmers by uploading image of farms. Crop diseases detection using image processing in which user get pesticides based on disease images. Implement Smart Irrigation System for farms to get higher yield.

REFERENCES:

1. W. Haoxiang and S. Smys, "Big data analysis and perturbation using data mining algorithm," Journal of Soft Computing Paradigm, vol. 3, no. 1, pp. 19–28, 2021.

2. M. Sivakami and P. Prabhu, "Classification of algorithms supported factual knowledge recovery from cardiac data set," International Journal of Current Research and Review, vol. 13, pp. 160–165, 2021.

3. A. Bora, N. Vasantha Gowri, M. Naved, and P. S. Pandey, "An utilization of robot for irrigation using artificial intelligence," International Journal of Future Generation Communication and Networking, vol. 14, no. 1, 2021.

4. A. Raghuvanshi, U. K. Singh, G. S. Sajja et al., "Intrusion detection using machine learning for risk mitigation in IoT-enabled Smart irrigation in Smart farming," Journal of Food Quality, vol. 2022, Article ID 3955514, 8 pages, 2022.

5. V. D. P. Jasti, A. S. Zamani, K. Arumugam et al., "Computational technique based on machine learning and image processing for medical image analysis of breast cancer diagnosis," Security and Communication Networks, vol. 2022, Article ID 1918379, 7 pages, 2022.

6. V. Hemamalini, S. Rajarajeswari, S. Nachiyappan et al., "Food quality inspection and grading using efficient image segmentation and machine learning-based system," Journal of Food Quality, vol. 2022, Article ID 5262294, 6 pages, 2022.

7. A. Raghuvanshi, U. Singh, and C. Joshi, "A review of various security and privacy innovations for IoT applications in healthcare," Advanced Healthcare Systems, Wiley, Hoboken, NJ, USA, pp. 43–58, 2022.

8. S. H. Mohamed, T. E. H. El-Gorashi, and J. M. Elmirghani, "A survey of big data machine learning applications optimization in cloud data centers and networks," 2019, https://arxiv.org/abs/1910.00731.

9. S. Jankatti, B. K. Raghavendra, S. Raghavendra, and M. Meenakshi, "Performance evaluation of Map-reduce jar pig hive and spark with machine learning using big data," International Journal of Electrical and Computer Engineering, vol. 10, no. 4, p. 3811, 2020.

10. A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review," Computers and Electronics in Agriculture, vol. 151, pp. 61–69, 2018.

11. M. Omid, A. Farjam, A. Akram, and Z. F. Niari, "A neural network based modeling and sensitivity analysis of energy inputs for predicting seed and grain corn yields," Journal of Agriculture, Science and Technology, vol. 16, 2018.

12. S. K. S. Fan, C. J. Su, H. T. Nien, P. F. Tsai, and C. Y. Cheng, "Using machine learning and big data approaches to predict travel time based on historical and real-time data from Taiwan electronic toll collection," Soft Computing, vol. 22, no. 17, pp. 5707–5718, 2018.

13. T. K. Fegade and B. V. Pawar, "Crop prediction using artificial neural network and support vector machine," Data Management, Analytics and Innovation, Springer, Berlin, Germany, pp. 311–324, 2020.

14. E. Khosla, R. Dharavath, and R. Priya, "Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression," Environment, Development and Sustainability, vol. 22, pp. 1–22, 2019.

15. P. Tiwari and P. K. Shukla, "A hybrid approach of TLBO and EBPNN for crop yield

prediction using spatial feature vectors," Journal of Artificial Intelligence, vol. 1, no. 2, pp. 45–59, 2019.

16. A. Li, S. Liang, A. Wang, and J. Qin, "Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques," Photogrammetric Engineering and Remote Sensing, vol. 73, no. 10, pp. 1149–1157, 2007.

17. C. H. Mamatha, B. P. Reddy, R. Kumar, and S. Kumar, "Analysis of big data with neural network," International Journal of Civil Engineering and Technology, vol. 8, no. 12, pp. 211–215, 2017.

18. K. Kira and L. A. Rendell, "Feature selection problem: traditional methods and a new algorithm," in Proceedings of the Tenth National Conference on Artificial Intelligence, pp. 129–134, San Jose, CA, USA, July 1992.

19. Z. Lu and Z. Liang, "A complete subspace analysis of linear discriminant analysis and its robust implementation," Journal of Electrical and Computer Engineering, vol. 2016, Article ID 3919472, 10 pages, 2016.

20. A. Gupta and L. K. Awasthi, P4P: Ensuring Fault-Tolerance For Cycle-Stealing P2P Applications, GCA, Ahmedabad, India, 2007.

21. A. Gupta, "Performance insight 360: A cloud-based quality management framework for educational institutions in India," in Proceedings of the IEEE 15th Conference on Business Informatics, Vienna, Austria, July 2013.

22. M. Rakhra, R. Singh, T. K. Lohani, and M. Shabaz, "Metaheuristic and machine learning-based Smart engine for renting and sharing of agriculture equipment," Mathematical Problems in Engineering, vol. 2021, Article ID 5561065, 13 pages, 2021.