Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10370103

OPTIMIZATION OF AGRICULTURE USING DATA SCIENCE AND MACHINE LEARNING

Dr. P. JOSEPHIN SHERMILA

Associate Professor, Department of Artificial Intelligence and Data Science, R.M.K. College of Engineering & Technology, Chennai. Email: blossomshermi@gmail.com

SYED ATHEEQUR RAHAMAN

UG Student, Department of AI & DS, R.M.K. College of Engineering and Technology. Chennai.
Email: syed21ad067@rmkcet.ac.in

NITHOSHBALAJI R

UG Student, Department of AI & DS, R.M.K. College of Engineering and Technology. Chennai.
Email: nith21ad044@rmkcet.ac.in

SAI MANOJ M

UG Student, Department of AI & DS, R.M.K. College of Engineering and Technology. Chennai.
Email: saim21ad084@rmkcet.ac.in

Abstract

The research aims to optimize agriculture using data science techniques. Agriculture is a critical sector for sustaining life on earth, and optimizing it can enhance food security and increase the profitability of farmers. Data science techniques such as machine learning, data analysis, and modeling to analyze agricultural data and derive insights are the key areas in this work. The project will focus on enhancing crop productivity, reducing wastage, and improving resource utilization. The findings of the project can help policymakers, farmers, and stakeholders to make informed decisions and take appropriate actions. The project aims to support the sustainable agriculture practices and promote the efficient use of resources for the benefit of both farmers and consumers.

Keywords: Agriculture, Crop Productivity, Data Analysis, Data Science, Machine Learning, Modeling, Optimization.

1. INTRODUCTION

In the current era, farming practices have evolved to incorporate intelligent methods. The choice of crops and their respective seasons depends on various factors like climate, soil type, and local preferences. For example, in spring, crops like tomatoes and cucumbers are often planted due to their ability to thrive in mild temperatures. In summer, heat-resistant crops such as corn and melons are preferred. During the fall, farmers may focus on harvesting wheat and pumpkins, which flourish in cooler conditions. Intelligent farming also utilizes data-driven approaches to optimize irrigation, pest control, and fertilization, enhancing crop yield and quality while minimizing environmental impact.

Due to the increase in population, skilled farmers know every detail of how crops should be planted and the interval needed for cultivation. We have collected many data on various crops. The data like humidity, Ph, rainfall, and temperature value of the soil and few more are collected from the different crops.

Data Science is largely used in our day-to-day activities. In Data science, one need to analyze data for actionable insights. Skills that are required for data science are deep learning, data visualization, mathematics and programming.

The model is a programming one. It is a calculation based model. It is based on many principles. Various principles and criteria can be used as a tool for the analysis and simulation of agricultural production plans as well as for the study of the impacts of the various rules in agriculture.

The motivation of this investigation is to find the best crop to be grown on the specific conditions, which gives the maximum yield and profit.

Various sections describe the work as follows. Section 2 describes about the literature related to the existing agriculture practices. Section 3 gives an idea about the data science process in agricultural yield followed by data visualization in section 4 which is followed by alternate forming in section 5. In section 6 the prediction methods used in this work has been described. Section 7 briefs about the classification.

2. RELATED WORK

An automated farming system that uses the Internet of Things is essential for optimal use of water and fertilizer.

These systems use sensors to measure humidity, temperature and soil moisture.

Few machine learning concepts with specific algorithms are used to calculate the amount of water needed for irrigation [1]. A control unit based on a microcontroller is used to control water flow and fertilizer application.

A mobile application is embedded for remote monitoring and control of the system. These systems also include animal detection and tree prevention.

Cloud servers are used for data analysis and decision making. Embedded systems, MPLAB and Proteus software are used to develop these systems. Smart irrigation systems using Internet of Things and cloud computing have also been proposed. All in all, these systems play an important role in the development of smart agriculture and the overall economic development of the country

Automation of agriculture is necessary to meet the increasing demand for food due to population growth and climate change. IoT and AI are used in agriculture to provide real-time monitoring, control and visualization of farm operations.

Digital technologies can be used in smart agricultural machine, irrigation systems, pest control, nutrition application, greenhouse cultivation, storage structures, drones for plant protection and crop health monitoring. The paper [2] provides an overview of recent research in digitally driven agriculture and identifies the most significant applications in agricultural engineering. Scientific databases including PubMed, Web of Science and Scopus were used to review research work carried out in the areas over the past 10 years.

Digitization of agriculture with the help of artificial intelligence and the Internet of Things has matured from the nascent conceptual stage and reached the implementation stage.

The technical facts about the artificial intelligence, the Internet of Things and the challenges of adopting these digital technologies are discussed. Integrating digital technologies into farming practices can overlay the way for AI and IoT-based results on real farms.

Potential benefits of using digital technologies in agriculture include increased efficiency, reduced costs and increased yield.

Further research and implementation is needed to fully realize the potential of AI and IoT based typical results in agriculture.

Agricultural automation is a new challenge due to increasing population and food demand.

Traditional ways of farming methods are not enough to meet the demand and damage the soil. Automation methods like the IoT, wireless communication, machine learning, deep CNN, artificial intelligence and deep learning can help as given in [3].

Problems related to leaf disease, crop diseases, storage management, pesticide management, unwanted plant namely weed management and water management can be solved through automation.

Harmful pesticides, irrigation management, pollution and environmental impacts need immediate attention. Automation increases soil fertility and yield. This article reviews the work of many researchers related to different automation methodologies in agriculture.

This work [3] also proposes an IoT-based system for identifying and watering flowers and leaves in botanical gardens. IoT-based systems have been effective in automating agriculture. Further research and implementation are needed to fully give better solution to the potential of automation in agriculture.

The article [4] discusses the various steps involved in the implementation in the agents environment of artificial intelligence in various sectors, including agriculture, finance, industry, and security, with a focus on machine learning algorithms.

Machine learning approaches are used to gather statistical data and previously acquired information for specific tasks. The advancement of big data and data science is due to machine learning.

Many apps are now capable of complex tasks, such as analyzing historical data, gathering new data, reading faces, and forecasting weather. The use of artificial intelligence is highly dependent on the machine learning process, which utilizes a mathematical approach to create intelligent machines.

The paper concludes by highlighting the importance of artificial agent and IoT, and automation.

The application of automation and robotics in agriculture is becoming increasingly important for efficient irrigation and reducing water waste in horticulture, parks, gardens and golf courses.

Automation and smart systems also help to ensure food safety by choosing the right chemicals. Multitasking robots help you work faster and maintain quality. Smart agriculture uses sensors and applications to help maintain optimal moisture, temperature and irrigation processes.

The purpose of the work given in [5] is to clearly find the opportunities and scope of future automation and Internet of Things in the agricultural industry. These technologies have great potential to improve agricultural efficiency, sustainability and productivity while minimizing labor and resource costs.

This review paper [6] describes the development of a prototype of a smart farming system using the Internet of Things, designed to provide cost-effective solutions to farmers.

It is equipped with various sensors, including temperature, humidity and water level, which can be controlled through smart devices such as mobile phones.

This research also includes the development and testing of application software to successfully connect prototypes of smart farming systems. This study uses various technologies such as Node-Red, IBM Cloud and IBM Watson to achieve the desired results. This white paper provides insight into the use of IoT-based technologies for smart agriculture and their potential impact on agriculture.

3. DATA SCIENCE PROCESS IN AGRICULTURE YIELD

Data is required to build a model for good yield. In this step, data required for the model is identified and collected from various sources that help us to build our model. Data can be taken from various sources like Logs from webservers, Data from social medias, Census datasets, Data streamed from online sources using APIs

For this investigation the kaggle dataset is used. It consists of 2200 rows and 8 columns. The column value takes the different properties of the soil such as Potassium (K), Phosphorus (P), Nitrogen (N), Temperature, Humidity, Rainfall, Ph level and finally its label. Thus there are 8 columns in the dataset.

Each row of the dataset consists of different crops. A total of 22 crops were considered. The crops choosen are maize, rice, kidney beans, pigeon peas, mothbeans, chickpea, mungbeans, blackgram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, and coffee. 100 different entries per crops are considered and there is no null value present in the dataset. As mentioned there are 2200 rows.

4. DATA VISUALIZATION

The distribution of the properties of crops was visualized using Histogram plots.

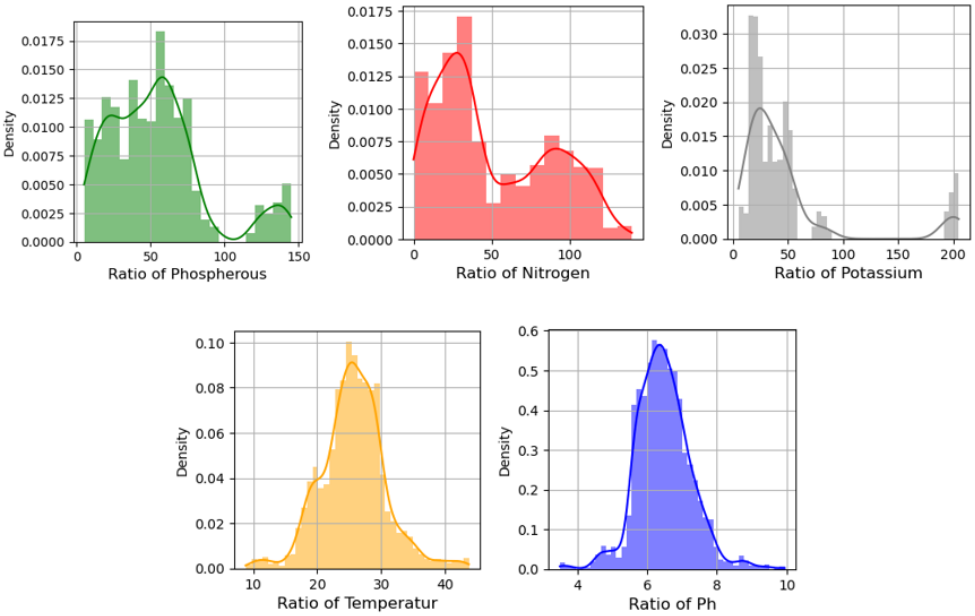
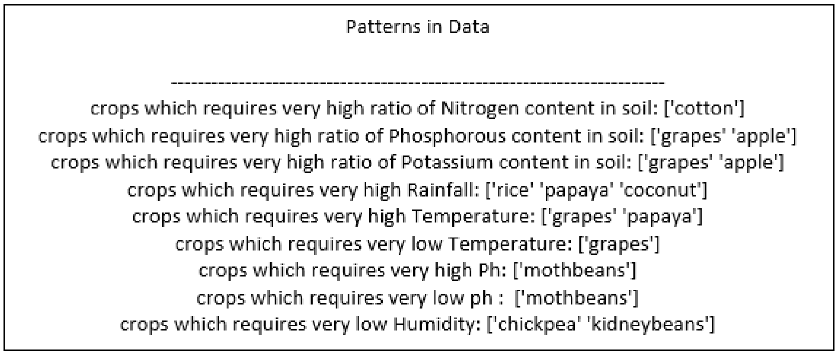


Figure 1: Visualization of properties of crops



Patterns in Data

----

crops which requires very high ratio of Nitrogen content in soil: ['cotton']
crops which requires very high ratio of Phosphorous content in soil: ['grapes' 'apple']
crops which requires very high ratio of Potassium content in soil: ['grapes' 'apple']
crops which requires very high Rainfall: ['rice' 'papaya' 'coconut']
crops which requires very high Temperature: ['grapes' 'papaya']
crops which requires very low Temperature: ['grapes']
crops which requires very high Ph: ['mothbeans']
crops which requires very low ph : ['mothbeans']
crops which requires very low Humidity: ['chickpea' 'kidneybeans']

Figure 2: Crops that requires specific nutrion content for its growth in the soil

Figure 1 displays a selection of visualizations representing the crops from the chosen dataset. The visualizations depict properties such as the ratios of phosphorous, nitrogen, and potassium, as well as temperature and pH levels associated with the crops.

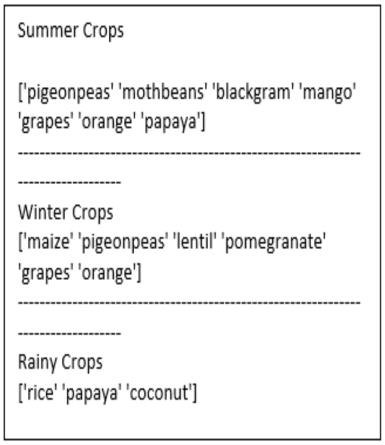
Research has been conducted with the primary objective of enabling a system to discern the specific soil properties necessary for cultivating a particular crop successfully. This investigative effort delves into understanding the precise characteristics and attributes of the soil that are essential to promote optimal growth and yield of the chosen crop variety.

The process entails a comprehensive analysis of various factors that contribute to a conducive soil environment for the desired cropas shown in figure 2. This encompasses a thorough examination of elements such as nutrient content, pH levels, moisture retention capacity, and physical structure of the soil. By meticulously studying these essential properties, the system aims to develop a deep understanding of the intricate relationship between soil composition and crop cultivation requirements.

Furthermore, the investigation strives to establish a robust framework where the system can effectively interpret the gathered data and apply intelligent algorithms to determine the specific soil conditions most conducive to the healthy growth and development of the selected crop. This entails creating a sophisticated model that can assimilate and process a myriad of soil-related information, ultimately leading to accurate recommendations for farmers and cultivators.

The ultimate goal of this research endeavor is to equip the system with the ability to offer tailored and data-driven insights to individuals engaged in agriculture. By discerning the precise soil attributes necessary for successful crop cultivation, the system seeks to optimize agricultural practices, enhance crop yields, and contribute to sustainable and efficient agricultural management.

Seasonal crops dominate agricultural practices worldwide. Consequently, this study aims to develop a system capable of discerning the ideal crops for cultivation during distinct seasons. By doing so, the system's ability to recommend suitable crops aligns with the changing environmental conditions, potentially maximizing yield. This investigation seeks to optimize agricultural productivity by ensuring that the crops chosen for each season are well-suited to prevailing weather patterns, soil characteristics, and other contextual factors, ultimately contributing to more effective and sustainable cultivation practices. Investigation done in this aspect is shown in figure 3.



```
Summer Crops

['pigeonpeas' 'mothbeans' 'blackgram' 'mango'
'grapes' 'orange' 'papaya']
----------------------------------------------------------------
------------------
Winter Crops
['maize' 'pigeonpeas' 'lentil' 'pomegranate'
'grapes' 'orange']
----------------------------------------------------------------
------------------
Rainy Crops
['rice' 'papaya' 'coconut']
```

Figure 3: Crops that grows in specific season

5. ALTERNATE FARMING USING CLUSTERS

Another way of investigation has been done, by clustering similar crops. Alternate farming using clustering refers to the application of clustering techniques in agriculture to identify groups or clusters of similar crops, which can be grown alternatively in a particular area. Clustering is a machine learning technique that involves grouping data points based on their similarities or differences. In the context of agriculture for the benefit of farmers, clusters are to identify crops that have similar growth patterns, nutrient requirements, and climatic preferences.

The idea behind alternate farming using clustering is to diversify crop production in a particular area by growing different crops in a cyclic manner. This approach shows betterment for the improved progress in soil health, reduced pest and disease incidence, and increase crop yields. Clustering can also be used to identify complementary crops that can be grown together to enhance their productivity.

Overall, alternate farming using clustering is a promising approach to sustainable agriculture that can help to optimize crop production [7] and reduce the environmental impact of agriculture.

5.1 Elbow method in cluster analysis

The elbow method is a widely utilized approach in research, particularly in data science and machine learning, to identify the most suitable number of clusters in a dataset during cluster analysis.

To implement the elbow method, researchers plot the within-cluster sum of squares (WCSS) against the number of clusters. WCSS quantifies the total squared distance between data points and their corresponding cluster centers. As the number of clusters increases, the WCSS tends to decrease because each data point can be assigned to a more specific cluster.

The key idea of the elbow method is to choose the inflection point on the plot where adding more clusters doesn't result in a significant reduction in WCSS. This point is referred to as the "elbow," and it signifies the optimum number of clusters for the dataset. By selecting the elbow point, researchers can avoid overfitting or underfitting the data, ensuring a reasonably balanced and interpretable clustering solution.

Here in this investigation the elbow [8] method is used to identify the effective number of clusters.

The “elbow” in the plot refers to the point where the decrease in variance explained begins to level off, creating a bend in the plot that resembles an elbow. This point is often taken as the optimal number of clusters to use for the dataset, as it represents a balance between fitting the data well and avoiding overfitting. Figure 4 represents the number of clusters vs crops.

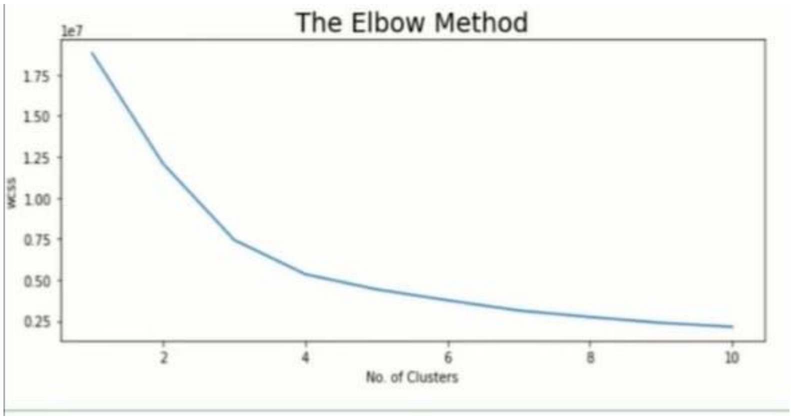


Figure 4: Cluster identification using elbow method

5.2 K-Means Algorithm

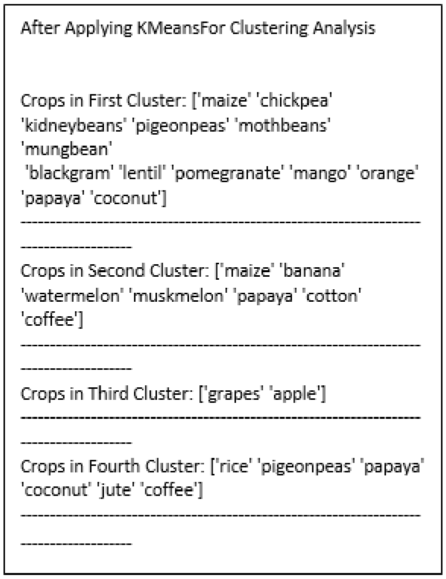
Clustering can be done in different ways. It can be graph based method, grid based method, model based method hierarchical based method, partitioning methods and density based methods. Thus there are six methods in which clustering can be classified. In these six methods, density based approach is the method, where the K-means is derived.

The K means algorithm takes the input parameter K from the user and partitions the dataset containing N objects into K clusters so that the resulting similarity among the data objects inside the group (intra-cluster) is high but the similarity of data objects with the data objects from outside the cluster is low (inter-cluster).

In this investigation the dataset containing N number of objects is divided into k clusters. The s clustering method is done based on the below steps.

1) Assign K objects randomly from the dataset (DS) as cluster centers (CC)
2) Based on the mean values (Re) assign each object to which the particular object is most similar.
3) Now cluster mean value is updated. i.e., The mean for each cluster is calculated again and the new cluster value is updated with this new mean value.
4) Repeat Step 1 to 4 until no change occurs.

Figure 5 gives the information about the similar crops in each cluster and figure 6 represents the k means methodology.



```
After Applying KMeansFor Clustering Analysis


Crops in First Cluster: ['maize' 'chickpea'
'kidneybeans' 'pigeonpeas' 'mothbeans'
'mungbean'
 'blackgram' 'lentil' 'pomegranate' 'mango' 'orange'
'papaya' 'coconut']
---------------------------------------------------------------------
-------------------
Crops in Second Cluster: ['maize' 'banana'
'watermelon' 'muskmelon' 'papaya' 'cotton'
'coffee']
---------------------------------------------------------------------
-------------------
Crops in Third Cluster: ['grapes' 'apple']
---------------------------------------------------------------------
-------------------
Crops in Fourth Cluster: ['rice' 'pigeonpeas' 'papaya'
'coconut' 'jute' 'coffee']
---------------------------------------------------------------------
-------------------
```

Figure 5: Similar crops grouped in specific clusters

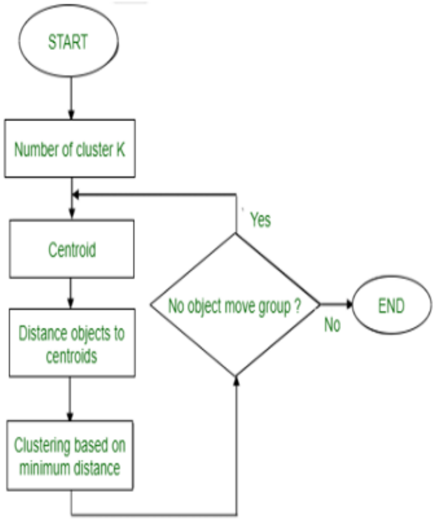

START
Number of cluster K
Centroid
Distance objects to centroids
Clustering based on minimum distance
No object move group ?
Yes
No
END

Figure 6: Flowchart representing the Methodology in K- Means

6. PREDICTIVE MODEL

In this section the predictive models which have been investigated is described. Regression is the method by which the value of the classified one can be obtained. One of the regression analysis is the logistic regression which does the prediction in a statistical way. Hence this analysis is commonly used in predictive modeling. This technique allows modeling the probability of an event, given a set of predictor variables. Logistic regression is a popular choice for predictive modeling because it is simple to implement and easy to interpret the results.

In a logistic regression model, the dependent variable is binary, meaning it can only take on two values (for example, 0 or 1). The purpose of the analysis is to model the probability that the dependent variable is equal to 1, given the set of predictor variables. The dependent variable can be continuous or categorical and is used to estimate the probability of an event.

Once a logistic regression model is fit to the data, it can be used to predict new data. The model can be used to estimate the probability of an event of interest, given the values of the predictor variables. These probabilities can be converted to binary predictions (eg, yes or no) using a threshold value. Regression model works efficient for more than 2 labels. Here we have 20 labels.

6.1 Creating Training and Testing set

Using Scikit library the training and testing has been done.

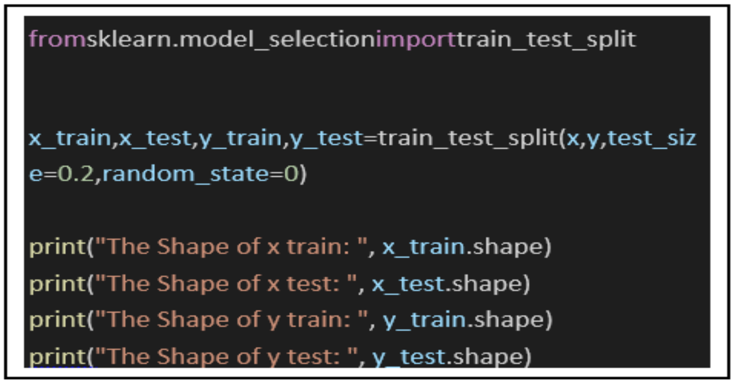
Output:

The Shape of x train: (1760, 7)

The Shape of x test: (440, 7)

The Shape of y train: (1760,)

The Shape of y test: (440,)

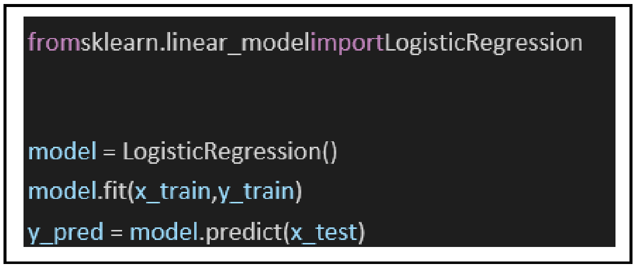


```
fromsklearn.model_selectionimporttrain_test_split


x_train,x_test,y_train,y_test=train_test_split(x,y,test_siz
e=0.2,random_state=0)

print("The Shape of x train: ", x_train.shape)
print("The Shape of x test: ", x_test.shape)
print("The Shape of y train: ", y_train.shape)
print("The Shape of y test: ", y_test.shape)
```

Figure 7: Creating Predictive model



```
fromsklearn.linear_modelimportLogisticRegression


model = LogisticRegression()
model.fit(x_train,y_train)
y_pred = model.predict(x_test)
```

Figure 8: Prediction using logistic regression

Figures 7 and Figure 8 shows the prediction models performed in the optimization in agriculture. Figure 9 and Figure 10 shows the performance evaluation and the right prediction checking investigations.

7. CLASSIFICATION METRICS

Sklearn provides a number of metrics to evaluate the performance of classification models. These include accuracy_score, precision_score, reminiscence_score, f1_score, roc_auc_score, and confusion_matrix.

```
#Evaluation of the model performance
fromsklearn.metricsimportclassification_report


cr=classification_report(y_test,y_pred)
print(cr)
```

Figure 9: Model performance evaluation

This code calculates the accuracy score for a binary classification problem where y_pred are the predicted values and y_true are the actual values. The resulting accuracy score is printed on the console. Overall, sklearn provides a number of metrics to evaluate model performance, making it a powerful tool for machine learning and predictive modeling.

7.1Checking for Right Prediction

To check the correctness of the model predictions, we usually compare the predicted values with the actual values of the dependent variable. If the predicted values are close to the actual values, the predictions of a model are considered accurate.

To check for correct predictions is to use a scatter or line graph to visualize [9] the predicted values against the actual values. When the predicted values are close to the actual values, the plot shows tight clustering around the diagonal. If the predicted values are far from the actual values, the plot shows a scatter pattern with large deviations from the diameter. In summary, verifying correct predictions requires comparing predicted values with actual values and evaluating model performance using a variety of metrics and visualization techniques.

```
prediction = model.predict((np.array([[90,40,40,20,80,7,200]])))


print("The Suggested Crop from Given Condition is : ", prediction)
```

Figure 10: Checking for prediction

Output:

The Suggested Crop from Given Condition is: ['rice']

Hence the model successfully predicts the crops for the input entered into the model using the dataset.

The input array in the above code represents each properties of the soil. Hence each values maps to the column of the dataset and predicts the crops as the output.

8. CONCLUSION

With the passage of time, the population of the world increases drastically. Therefore, it is necessary to take precautionary measures in advance. Agricultural optimization is the first and most important way to provide results for growing the best crops in the soil according to the conditions. We included factors affecting plant growth. Rainfall, Temperature, Nitrogen in soil, Potassium in soil, Phosphorus in soil, Humidity, Ph of the soil are the seven factors which are considered here since they are very important for plant growth.

Concluding this we can say that this method will be very helpful for farmers in analyzing weather and soil characteristics used for specific crops. Continuous farming on agricultural land is made possible by optimization of agriculture.

References

1) S. A. Sivakumar, G. Mohanapriya, A. Rashini, R. Vignesh,” Agriculture Automation using Internet of Things”, International Journal of Advance Engineering and Research Development, Vol.5, No.2, Feb 2018.\

2) A. Subeesh, C.R. Mehta,” Automation and digitization of agriculture using artificial intelligence and internet of things”, Artificial Intelligence in Agriculture, Vol.5, 2021.

3) Kirtan Jha, Aalap Doshi, Poojan Patel, Manan Shah,” A comprehensive review on automation in agriculture using artificial intelligence”, Artificial Intelligence in Agriculture, Vol.2, 2019.

4) Amit Kumar, “Agriculture and Automation: a Review”, Journal of Emerging Technologies and Innovative Research, Vol. 5, No. 4, pp. 1249-1254, Apr 2018.

5) Amith A Kulkarni et al., “Applications of Automation and Robotics in Agriculture Industries; A Review”, IOP Conference Series: Materials Science and Engineering, 2020.

6) Subham Patra, Arnab Samanta, Suman Paira, “Automation in Agriculture”, International Journal of Engineering Research & Technology, Vol. 9, No.11, pp.162-164, 2021.

7) Vashishth, Tarun& Sharma, Mr. Vikas& Chaudhary, Sachin& Panwar, Rajneesh & Sharma, Shashank& Kumar, Prashant, “Advanced Technologies and AI-Enabled IoT Applications in High-Tech Agriculture”, 2023, DOI: 10.4018/978-1-6684-9231-4.ch008

8) Shahin, Mahtab& Saeidi, Soheila& Shah, Syed &Kaushik, Minakshi& Sharma, Rahul & Pious, Sijo&Draheim, Dirk., "Cluster-Based Association Rule Mining for an Intersection Accident Dataset,*"International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, uetta, Pakistan, pp. 1-6, 2021. doi:10.1109/ICECube53880.2021.9628206.

9) Maksymilian Mądziel,, “Future Cities Carbon Emission Models: Hybrid Vehicle Emission Modelling for Low-Emission Zones” *Energies 16*(19), pp. 6928; 2023, doi.org/10.3390/en16196928