**Summer Training (MC300) report on “Data Science & Machine Learning Internship”**

*A Minor Report*

*Submitted in partial fulfilment of the requirements for the award of the Degree of*

***Bachelor of Technology***

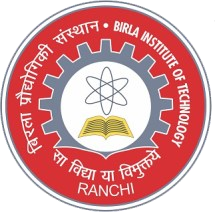
*in*

***Computer Science & Engineering***

*By*

***Krishna kant***

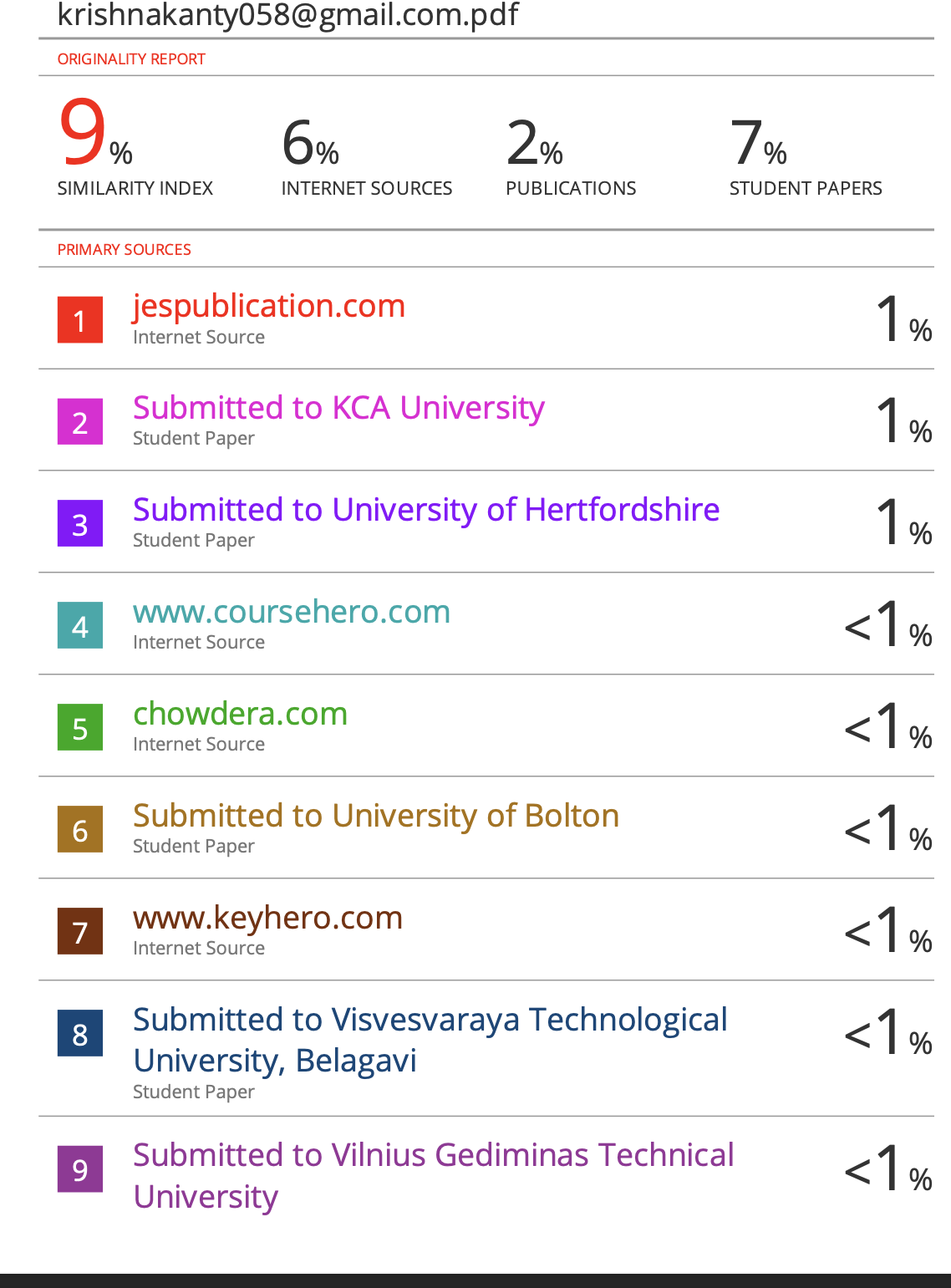
BTECH/15051/20



**Birla Institute of Technology, Mesra, Patna campus**

**Patna-800014**

**Attach Plagiarism Report**

****

**APPROVAL OF THE GUIDE**

Recommended that the B.Tech. Summer Training/Internship titled “**Prediction of Agriculture Crop Production in India using Machine Learning”** submitted by **BTECH/15051/20 and Krishna Kant** is approved by me for submission. This should be accepted as fulfilling the partial requirements for the award of Degree of Bachelor of Technology in **Computer Science**. To the best of my knowledge, the report represents work carried out by the student in **Upskill Campus & UniConverge Technologies Private Limited** and the content of this report is not form a basis for the award of any previous degree to anyone else.

**Date: 23 July 2023 ADNAN MAHMOOD**

**Associate Lecturer**

**Department of Computer Science and Engineering**

**Birla Institute of Technology, Mesra, Patna campus**

# DECLARATION CERTIFICATE

I certify that

1. The work contained in the report is original and has been done by myself under the general supervision of my supervisor.
2. The work has not been submitted to any other Institute for any other degree or diploma.
3. I have followed the guidelines provided by the Institute in writing the report.
4. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
5. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the report and giving their details in the references.
6. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

**Date:22 july 2023 Krishna Kant**

**BTECH/15051/20**

**Computer Science & Engineering**

**Birla Institute of Technology, Mesra, Patna campus**

# CERTIFICATE OF APPROVAL

This is to certify that the work embodied in this Summer Training Report entitled “ **Prediction of** A**griculture Crop Production in India using Machine Learning”**, is carried out by **Krishna Kant(BTECH/15051/20)** has been approved for the degree of Bachelor of Technology in **Computer Science & Engineering** of Birla Institute of Technology, Mesra, Patna campus.

Date:

Place:

(Chairman)

**S CHAUDHURI** (Panel Coordinator)

Head of the Department Examiner

Dept. of Comp. Sc. & Engg. Dept. of Comp. Sc. & Engg.

# *ABSTRACT*

Machines Learning(ML) is used in various field arounds all over the world . The Prediction of Agriculture Crop Production in India using Machine Learning Models Crop production plays a crucial role in the agricultural sector, impacting the economy and food security of a nation. Accurate prediction of crop production can aid policymakers, farmers, and other stakeholders in making informed decisions regarding resource allocation and planning. This report presents a study on the application of machines learnings(ml) models to Predicts Crops Yield Productions in India ,prediction of crop name for a particular region of the country, a country heavily reliant on agriculture.The objective of particular internship project was to develop and evaluate machines learnings model(Ml) capable of accurately predicting crop Yield production in India and predict crops name that are suitable and favorable for particular region of the country. To achieve this, historical data on crop production, state name, cultivation cost,production cost,crop varities and other relevant factors datasets were collected from Indian government website.Preprocessing techniques were applied to clean and transform the data into a suitable format for model training and testing.The implementation and comparisons of a different numbers of machine learnings(ml) algorithms, including linear Regression, Logistic Regression,support vector Machine(SVM), Decision trees, Random forests,Naive Bayes ,etc. The most pertinent variables for model training were chosen using feature engineering techniques, resulting in the best possible prediction performance. The models were tested using suitable metrics including mean absolute(ma)errors,rootmeansquared(rms)errors,Rsquared(R2),Precision,Recall,Accuracy,etc after being trained on a various attributes of the dataset.The findings of that study showed that crop yield production in India and crop name prediction may be accurately predicted by machine learning algorithms. The most accurate model had low error rates, high accuracy, and a great correlation with actual crop production values. Cross-validation techniques were used to further validate the models' prediction ability and gauge their resilience.Accurate crop yield production forecasts can help farmers choose the right crops, manage resources effectively, and manage risk. Theseforecasts can be used by policymakers to develop effective agricultural policies, address probable yield changes, and plan for food security.

**Keywords:** SVM’s; Naive Bayes(NB); Decision Trees;Randoms Forest; Logistics Regression; Linear Regressions;Accuracy;Precision;Recall;R-squared;absolute error.

# *ACKNOWLEDGEMENT*

I want to convey my sincere thanks to my project guide **ADNAN MAHMOOD** Associate Lecturer, Department of Computer Science and Engineering, BIT Mesra,Patna Campus, for his guidance ,expert assistance and unwavering support. I am grateful to **S CHAUDHURI**, In charge of the Department, Computer Science and Engineering, BIT Mesra, Patna Campus, for continuous support and cooperation in my summer internship.

I am extremely thankful to the **upskill campus, UniConverge Technologies Pvt LTD(UCT), KAUSHLENDRA SINGH SISODIA(Founder of UCT),Coordinators and mentors of this internship** for their constant support through out the internship period, Encouragement and Cooperation to carry out this work and for helping me directly or indirectly in this **internship** journey of 42 days. I was lucky to got support from you all.

I would like to thanks **Dr. Ayan Kumar Das**, Assistant Professor in the Department of Computer Science and Engineering at BIT Mesra, Patna Campus, for serving as the summer training coordinator. I would like to thank the whole **teaching staff** at the Department of CSE whose advice during evaluations enabled me to successfully complete my internship project. I want to express my gratitude to my seniors also for their invaluable help in completing my project.

I want to express my gratitude to my parents, friends, and classmates for their support over the 42 days that I spent working on my internship project. Last but not least, I want to express my gratitude to everyone who helped us, directly or indirectly, complete this internship.

Date:22 july 2023

(Signature)

**KRISHNA KANT**

**BTECH/15051/20**

***CONTENTS***

**COMPANY CERTIFICATE ........................................................................................ii**

**Attach Plagiarism Report ..............................................................................................iii**

**ABSTRACT ……………………………………………………………………………vii**

**ACKNOWLEDGEMENT ………………………………………………………….....viii**

**LIST OF FIGURES ……..…………………………………………………………..... xi**

**LIST OF TABLES …………………………………………………………………..... xii**

**CHAPTER 1 INTRODUCTION .....................................................................................01**

**1.1 Introduction.................................................................................................................1-2**

**1.2 Motivation For The Work..........................................................................................2-3**

**1.3PROBLEM STATEMENT..........................................................................................3-4**

**1.3.1 Context.....................................................................................................................03**

**1.3.2 Content.......................................................................................................04**

**1.3.3 Acknowledgements...................................................................................04**

**1.3.4Columns .....................................................................................................04**

**1.3.5 Data Sets link............................................................................................04**

**CHAPTER 2 LITERATURE SURVEY.......................................................................... 5-** **7**

**CHAPTER 3 METHODOLOGY.....................................................................................7-13**

**3.1Existing System...............................................................................................................07**

**3.2Proposed System...........................................................................................................7-9**

**3.2.1 Collection of dataset...........................................................................................8-9**

**3.2.2 Selection of attributes.......................................................................................9-10**

**3.2.3 Pre-processing of Data........................................................................................10**

**3.2.4 Prediction of Crop name and crop Yield .........................................................11**

**3.3 Machine Learning Algorithms..................................................................................11-12**

**3.4 Machine Learning Interfaces .................................................................................12-13**

**3.5 UseCase Diagram *.........................................................................................................*13**

**CHAPTER 4 EXPERIMENTAL ANALYSIS .............................................................14-33**

**4.1 Systems Configurations ...............................................................................................14**

**4.1.1 Hardwares Requirements.........................................................................................14**

**4.1.2 Software requirements..............................................................................................14**

**4.2 Sample Code ..............................................................................................................14-26**

**4.2.1import the libraries .............................................................................................14-16**

**4.2.2 Importing Dataset....................................................................................................16**

**4.2.3 Data preprocessing..............................................................................................16-17**

**4.2.4 Data Visualization ..............................................................................................18-20**

**4.2.5 Cross-Validation .....................................................................................................20**

**4.2.6 Accurate Algorithm Selection............................................................................21-22**

**4.2.7 Model training and testing ................................................................................22-24**

**4.2.8 Making a Predictive System...............................................................................25-26**

**4.3 PERFORMANCE ANALYSIS ................................................................................26-28**

**4.4 PERFORMANCES MEASURES.............................................................................28-29**

**4.5 INPUT AND OUTPUT.............................................................................................29-31**

**4.5.1 Input ................................................................................................................29-30**

**4.5.2 Output..............................................................................................................30-31**

**4.6 Result......................................................................................................................31-33**

**5 CONCLUSIONS AND FUTURE WORKS ...............................................................34-35**

**APPENDIX A ACCEPTED / COMMUNICATED PAPERS ……………………... 36-27**

**REFERENCES ……………………………………………………………………….......38**

***LIST OF FIGURES***

Figure 3.2.1 ***Collection of Data* 9**

Figure 3.2.2 ***Correlation matrix* 10**

Figure 3.2.3 ***Data Pre-processing*  10**

Figure 3.4 ***Block diagram of machine learning based system*  13**

Figure 3.5 ***UseCase diagram of crop yield prediction*  13**

Figure 4.2.3 ***Screenshot of code of data preprocessing Steps*** **17**

Figure 4.2.4 ***Data Visualization***  **18**

Figure 4.2.7 ***performance measure parameters*  24**

Figure4.3  ***Confusion matrix*  27**

Figure4.3  ***Correlation matrix* 27**

Figure 4.4 ***Accuracy comparison*  28**

Figure 4.4 ***algorithms comparison*  29**

Figure 4.5.1 ***input for crop name prediction*  30**

Figure 4.5.1 ***input for yield prediction***  **30**

Figure 4.5.2 ***output for crop name prediction*  31**

Figure 4.5.2 ***output of crop yield prediction*  31**

8

***LIST OF TABLES***

**TABLE 1: Accuracy Table........................................................................33**

**TABLE 2: R2\_Score Table........................................................................................33**

***CHAPTER 1***

**INTRODUCTION**

* 1. **Introduction**

For millennia, agriculture has served as the foundation of India's economy, supporting millions of people's livelihoods, maintaining food security, and considerably boosting the country's economic progress. Since the majority of the population relies on agriculture directly or indirectly, the performance of this industry is crucial to the overall health of the country. India's agricultural sector has seen significant technological and data availability developments in recent years, making it primed for the use of machine learning techniques to address issues and improve agricultural practices.The unpredictable nature of crop output is one of the main problems the Indian agricultural sector faces.

Crop yields can be strongly impacted by variables like geographical region, soil fertility, pest infestations, and market dynamics,area of cultivating land, production of crops which can cause changes in agricultural production. The ability of farmers, decision-makers, and stakeholders to make informed decisions and put into practice efficient methods to increase agricultural productivity and resilience depends critically on accurate and timely estimates of crop production.

This report's main goal is to investigate how machine learning may be used to forecast Indian agricultural crop production. I want to find out if different machine learning models can reliably predict crop yields and favorable crops for particular area if they can do so. Through this internship project, I hope to better understand how machine learning may enhance conventional agricultural methods and promote efficient and sustainable crop production. To accomplish my goals, I will gather and evaluate sizeable datasets that include historical crop production records, production costs, cultivation costs, yields, state names, crop types, crop name,soil characteristics, geographic data, and socio-economic aspects.

The data will be cleaned up and put into a format that is suitable for training machine learning models using preprocessing techniques. I will next put Various Machine Learning

Algorithms, including ensemble approaches, Support Vector Machines, Decision Trees, and regression models, into practice and assess their predictive power.

I'll also look at how feature engineering and selection can be improved the Performances and interpretability of Models. To ensure the Accuracy and generalizability of forecasts, I will also go through the importance of model selection and validation. Accuracy, Precision, Recall, and F1-score,r2 scoreis justs a few of the performance metrics that will be used to assess these machine learning models.

The results of this study should provide useful information about how machine learning might be used to predict crop productivity in Indian agriculture. The adoption of such cutting-edge technology in the agricultural industry has the potential to transform established procedures, reduce resource waste, increase agricultural productivity, and ultimately support food security and sustainable economic growth.

I want to raise more awareness and knowledge of the revolutionary potential of machine learning in Indian agriculture through this paper. I believe that the conclusions drawn from this study will make a substantial contribution to the development and modernization of India's agriculture sector as I go forward into an era of data-driven decision-making, benefiting farmers, decision-makers, and the country as a whole.

* 1. **MOTIVATION FOR THE WORK**

In the current agricultural landscape, the prediction of crop yield in India using machine learning is of enormous value and relevance. The agriculture industry in India has a variety of difficulties, such as changing weather patterns, a lack of resources, unpredictability in the market, and an ever-increasing need for food due to population growth. To solve these issues and promote sustainable agriculture practices, utilizing the power of machine learning to forecast crop yields offers a possible answer. The following essential elements are the driving forces behind this work's motivation:

1.**Ensuring Food Security**:As a country with a sizable population wholly dependent on agriculture for a living, India continues to place a high focus on maintaining food security. Accurate agricultural production forecasting enables more effective resource allocation and planning, empowering the nation to proactively manage food shortages and distribution issues.

2.**Optimizing Resource Allocation**: Resources used in agriculture, like water, fertilizer, and pesticides, are limited and frequently expensive. Farmers may allocate resources more efficiently, minimizing waste, and lessen the impact of agricultural activities on the environment by using crop yield predictions.

3.**Empowering Farmers:**Farmers are now more capable thanks to the development of machine learning technologies, who now have access to data-driven insights and decision-making tools. Accurate crop production estimates enable farmers to make knowledgeable decisions about crop selection, planting times, and market trends, enhancing productivity and profitability.

4.**Enhancing Policy Formulation**: To create efficient policies and programs that help the agriculture sector, policymakers and agricultural authorities need trustworthy data. Crop production forecasts based on machine learning offer a solid basis for evidence-based policies, promoting sustainable growth and development.

5.**Modernizing Agriculture**: The use of machine learning into agriculture is consistent with the overarching objective of contemporary farming methods. We can bridge the gap between traditional and intelligent agriculture by utilizing cutting-edge technologies, increasing productivity while preserving resources.

In conclusion, the Purposes of my works is to implements Machine Learning Techniques for crop production prediction, which has the potential to change India's agricultural industry. We can improve agricultural practices, encourage sustainability, and strengthen food security by leveraging the power of data-driven insights, which will help farmers, policymakers, and the country as a whole.

**1.3 PROBLEM STATEMENT**

My project/problem statement was  **“Prediction of Agriculture Crop Production in India ”.**

**1.3.1 Context**

Agricuture Production in India from 2001-2014

**1.3.2 Content**

This dataset provides information on the agriculture crops grown and produced In our Country. it is completely licensed data from Https://datas.gov.in.

**1.3.3Acknowledgements**  
This dataset can help India's varied agricultural cultivation and production issues.For example, it may forecast crop yield output and recommend crops for a certain location.

**1.3.4** **Columns**

Crop: String, Crop\_name Variety:String,Crop Subsidary\_name

state: Stings,Crops Place of Productions/Cultivation Quantity: Int, Number Of Quint../Hect.. int, number Of Years of manufacturing Date\_&Time, ..Medium (number Of Days), HowLong (duration in Days), Production Season Unit: Ton.., String

Cost:Int, Production Costs plus Cultivation Cost Recommended

**1.3.5 Data Sets link:**

*Google Drive URL: d/1zfqvs8-mAO6E0JpgvhBdueNx8Th03pUp/view?usp=sharing*

***CHAPTER 2***

**LITERATURE SURVEY/REVIEWS**

Literature Survey/Review on my Machine Learning project Prediction of Agriculture Crops Production in India:

[1] The article's title is "Machine Learning Techniques for Crop Yield Prediction: A Review".The writers are Singh, R., Sajwan, & Aggarwal.

published in the 2018 issue of the International Journal of Computer Applications.

This in-depth analysis covers various machine learning methods used to forecast crop yields. It gives an overview of various techniques used for crop production forecasting, including Ensemble Methods, Decision\_Trees, Support\_Vector\_Machines, and Regressions. The paper explores each technique's applicability in the context of Indian agriculture and identifies its advantages and disadvantages.

[2] The article's title is "A Review of Machine Learning Approaches for Crop Yield Prediction".Authors include Sharma, S., and Bhatia, G.

Published(written) in the International Journal(World level)of Emerging Trends & Technologies in ComputerScience(CS) fields issue for 2019.

This review paper explores how machine learning is used to forecast crop yields and evaluates the prospects and difficulties associated with its use in India. It talks about how crucial data pretreatment, feature engineering, and model choice are to creating precise prediction models. The potential for using satellite imagery and remote sensing data to enhance agricultural yield projections is also explored in this research.

[3] The title of the study is "Crop Yield Prediction Using Machine Learning: A Comparative Study".

The writers are A. Agarwal and A. Saxena.

published in the International Journal of Computer Applications' 2017 edition.

This comparative research evaluates the efficacy of several machine learning algorithms for agricultural production prediction using historical data from many sites in India. The authors assess the precision and stability of several models, including linear regression, Random Forests, and i can Gradient Boosting machines. .The outcomes highlight the advantages of each tactic and how well-suited they are to certain regions and crops.

[4] Title:"Predicting Crop Yields under Climate Change: MachineLearning(ML) subModels and Simulation-Based Approaches" is the heading of the study.

Authors include A. Jaiswal and V. Deora.

2019 IEEE Access publication

This study examines the use of simulation-based methods and machine learning algorithms to forecast crop yields in India under various climate change scenarios. For more accuracy, the authors investigate how to incorporate climatic data, soil details, and crop-specific parameters into the models. The study highlights how critical it is to modify forecast models to account for shifting climate circumstances.

[5] The title of the study is "Agricultural Crop Yield Prediction Using Machine Learning Techniques: A Review".N. Soni and N. Patel are the writers.

ICCES 2019, the 2019 International Conference on Communication and Electronics Systems

This review article provides an overview of the machine learning techniques utilized in India to predict agricultural yields. The authors examine the issues with data availability and quality as well as how these issues might impact model performance. The focus of the research is on the benefits of using ensemble techniques to increase forecast accuracy.

These studies of the literature offer important new perspectives on the status of the field today and the potential of machine learning to forecast agricultural output in India. For the country's agricultural and food security, they emphasize the necessity for precise crop output projections. By analyzing diverse machine learning approaches and their applications and identifying the challenges and opportunities for more study and advancement, these papers contribute to the growing body of knowledge on the subject.

***CHAPTER 3***

***METHODOLOGY***

**3.1 EXISTING SYSTEM**

The existing system for forecasting crop production in India typically uses specialized expertise and conventional statistical models. The ability of these traditional methods to handle big and complicated datasets, nonlinear relationships, and dynamic changes in environmental factors is frequently constrained. Key features of the current system include:

[1] Historical Data Analysis:Agronomists(Agriculture experts) and statisticians examine historical crop production data gathered from a variety of sources, such as census data, surveys, and agricultural databases. To comprehend historical trends of crop yield, they might do trend analysis and time series forecasting.

[2]. Expert Opinion and Domain Knowledge: Agricultural experts provide insights based on their experience and domain knowledge, which are often used in conjunction with statistical models to make predictions. However, this reliance on human expertise can lead to subjective interpretations and potential biases.

[3].Manual Data Preprocessing:Data cleansing, feature selection, and normalization are all steps in the preprocessing of data that are frequently carried out manually. This procedure can take a while and be prone to human mistake.

[4].Limited Scalability: It may be difficult to adapt to various geographies and crops because conventional methodologies are not always able to manage large-scale and diversified datasets.

The flaws in the current system serve as a reminder of the need for more sophisticated, data-driven systems that make use of machine learning. By utilizing machine learning, we can overcome these difficulties and create predictive models that are more precise, scalable, and able to handle enormous volumes of data, resulting in enhanced predictions of crop yield and well-informed agricultural decision-making.

**3.2 PROPOSED SYSTEM**

First I have done the Data cleaning and data preprocessing part means EDA parts.then after i have done the visualization part in which i have shows the different different scope of agriculture things like leading producer state,crop yield ,crop production in various year,state wise crop production,cultivation cost,recommended zone for various crops and its variety ,year wise productions of different different crops,etc and then after i have divided data into training and target data and then after i have done machine learning model designing ,training and testing parts(implementation parts) of different different model and check how efficiently they work and also checked accuracy,precision,recall,f1 score ,R2 score,etc to evaluate a performances of differents different machines learnings(ml) models.in this project i have used various supervised learning Algorithms (classification and regression both) for designing the different different machine learning sub models..

sub models are as follows:

***MACHINE LEARNING MODEL NO 1*:**Create Machines Learning model for Predicting crops(upaj) Based on the state name, the costs of cultivations (/Hectare) A2s+FLs, the costs of cultivations (/Hectares) C2s, and the yield (Quintals/Hectares). dataset used:datafile (1).csv

***MACHINE LEARNING MODEL NO 2*:**Predict crops yield by the crops name,state name,Costs of Cultivations (/Hectares) A2+FL,Cost of Cultivations (/Hectares) C,Costs of Productions (`/Quintals) C2s.i have used dataset datafile (1).csv

some of the basics step of machine learnings:

***3.2.1 Collection of dataset***

Data collecting is the first step in the development of my Crop name and Crop yield prediction system. I partition the dataset once it has been assembled to provide separate training and testing datasets. Then, using the testing dataset, the performance of the prediction model is assessed after it has been trained using the training dataset. We set aside 70–80% of the data for this project for training and 20–30% for testing.

The website devoted to Indian agriculture kindly donated the dataset required for this study. We used the "datafile (1).csv" specifically as the main dataset for our investigation.

**Prepared data**

Training data Test data

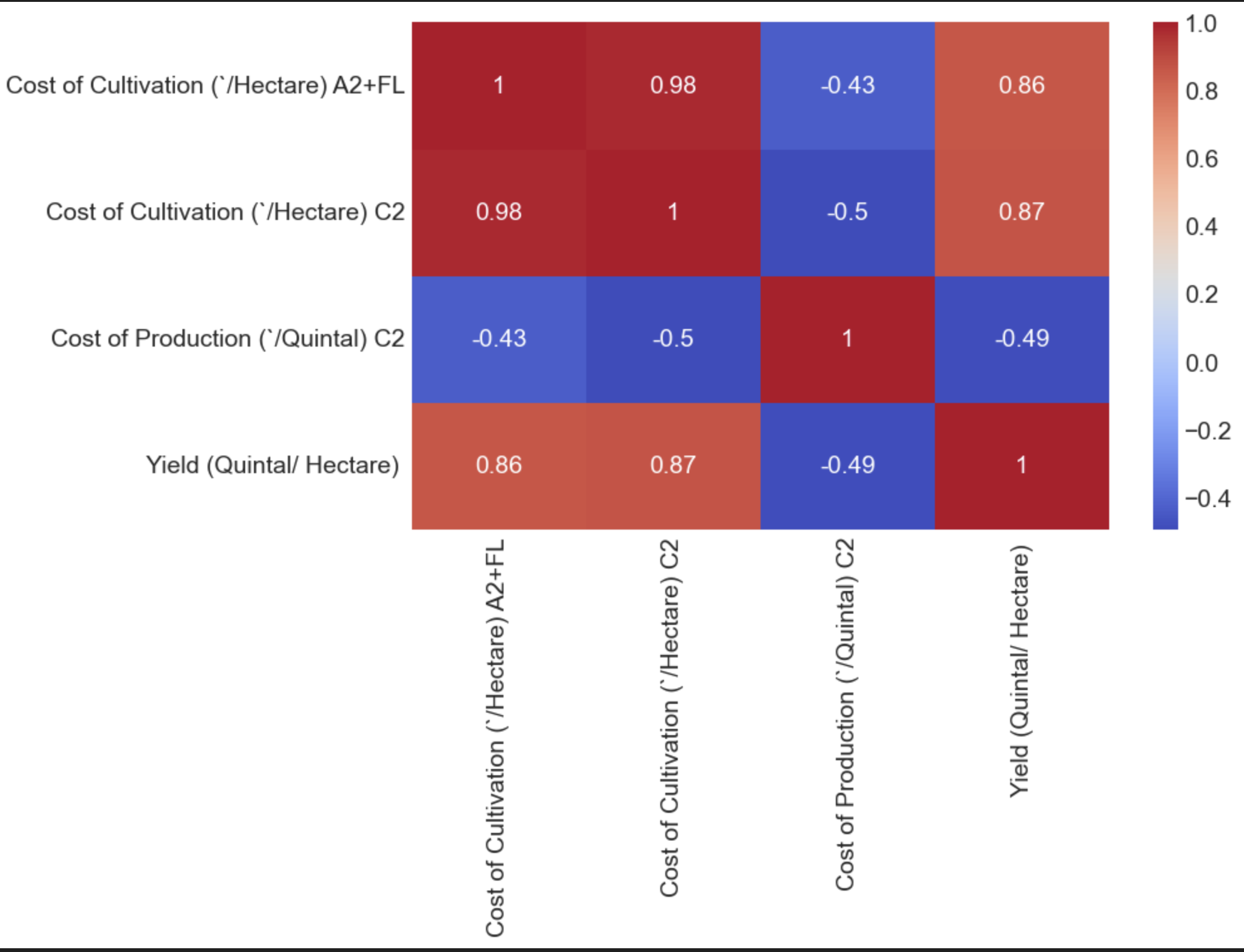
***Figure: Collection of Data***

**3.2.2 *Selection of attributes***

The selection of pertinent qualities for the prediction system is an essential phase in the feature selection process since it has a direct impact on the efficiency and performance of the system. We improve the forecast's accuracy and effectiveness by selecting essential attributes with care.

The name of the crop, the state, the cost of cultivation per hectare (A2+FL), the cost of cultivation per hectare (C2), the yield per hectare (QH), and the cost of production are all key crop data that we take into account in our prediction model. These qualities have been found to be important indicators of crop performance.

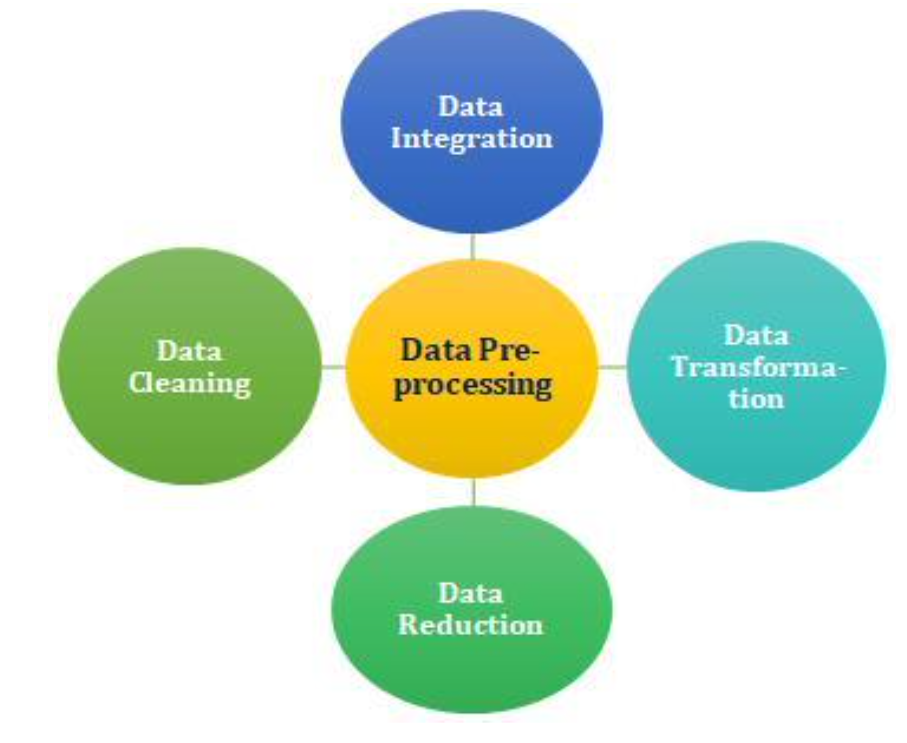
We use the correlation matrix to identify the factors that are most important for our model. The interrelationships between the various variables and their effects on the predicted crop production are identified with the use of this analytical approach. We can optimize the performance and accuracy of our prediction system by using the correlation matrix to help us decide which attributes to include.

****

***Figure: Correlation matrix***

**3.2.3 *Pre-processing of Data***

In order to build a machine learning model that is reliable and accurate, data pre-processing is essential. Results may be unreliable if the initial data isn't accurate or in the right format. As a result, pre-processing is done to change the data's format so that it meets the needs of the model.Various methods are used at this stage to deal with problems with the dataset, such as noise, duplication, and missing values. The data pre-processing pipeline must include the import of datasets, splitting them for training and testing, and attribute scaling.We ensure that the model can perform efficiently and produce more accurate predictions by carefully pre-processing the data. This preliminary phase considerably raises the model's Accuracy, making it more precise for Real-World Applications.

****

***Figure: Data Pre-processing***

**3.2.4 *Prediction of Crop name and crop yield***

We use a wide range of machine learning techniques to pursue classification and regression tasks in order to attain the best outcomes. These include Logistic Regression, Naive Bayes, AdaBoost, Random Trees, Linear Regression, and XGBoost as well as Decision Trees, Support Vector Machines (SVM), and Logistic Regression. Every algorithm brings unique advantages and traits to the table.We conduct a thorough comparison examination of these algorithms to assure the best performance. We identify the approach with the highest accuracy for both submodels—crop name prediction and crop yield prediction—through this evaluation.We can safely design a prediction system that excels in accuracy and dependability and is well-equipped to address real-world difficulties in agriculture and beyond by carefully choosing the most appropriate algorithm for each submodel.

**3.3 *Machine Learning Algorithm***

***Various Machine Learning Algorithms that i have used in this project are as follows:***

**1.*Linear Regression***:Common and basic statistical methods for examining the relationship between a dependent variable (also known as the response variable) and one or more independent variables (also known as predictor variables) include predictive modeling and linear regression. In evaluating the relationship between crop yields (the dependent variable) and different parameters, such as crop name, soil conditions, and agricultural techniques (the independent variables), linear regression shows useful when applied to forecasting agriculture crop production in India.We can learn more about how these important agricultural factors affect crop yields by using predictive modeling approaches, particularly linear regression. Making educated judgments and optimizing agricultural production techniques are made easier thanks to this study, which advances and sustains Indian agriculture.

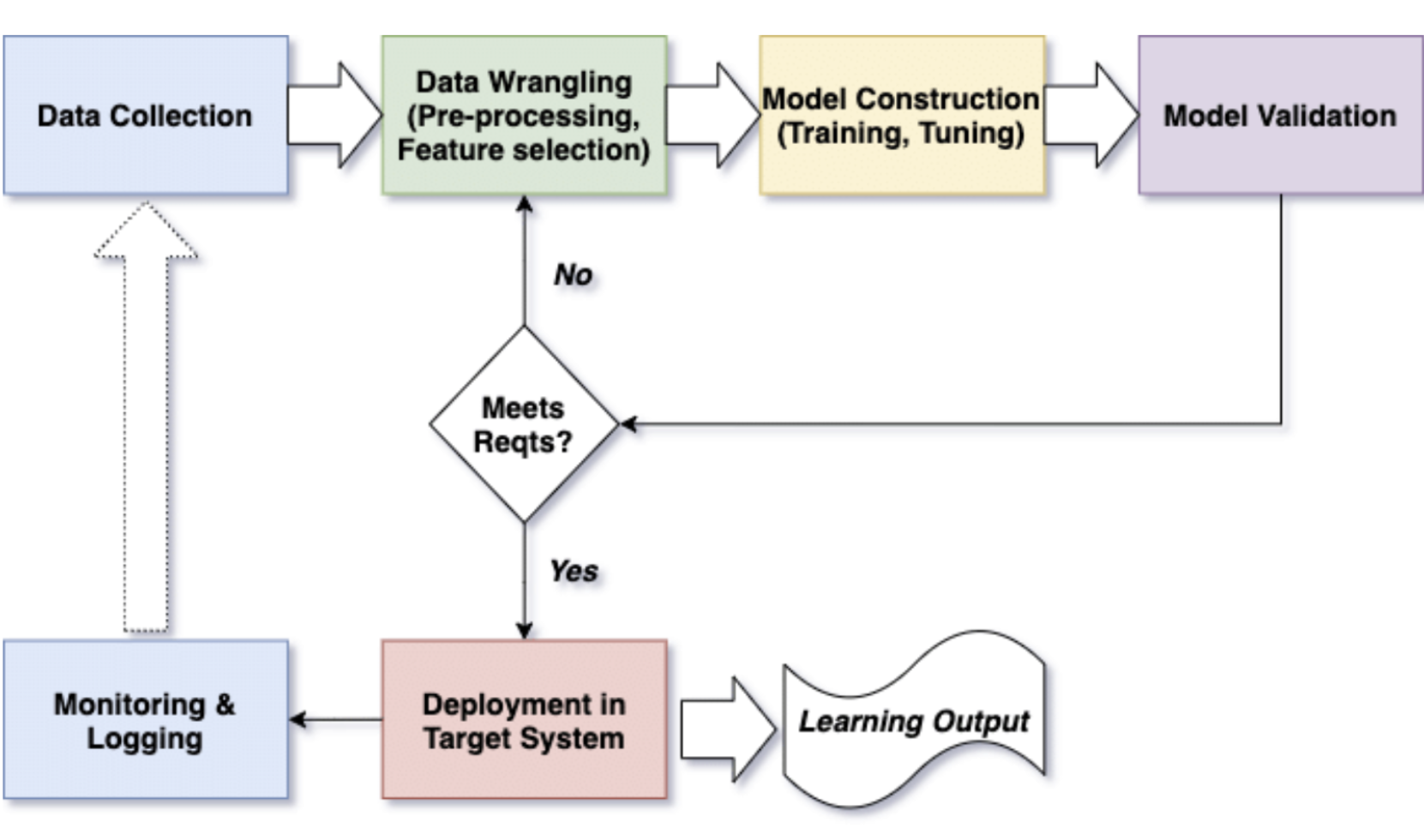
**2.*Logistic Regression*:**When categorical variables with two possible outcomes (often written as 0 and 1) are present in the dependent variable, Logistics regressions is a statistical approaches used for Binary Classification(BC) tasks. Logistic regression can be used when the job contains binary outcomes, such as forecasting whether a crop yield will be above or below a specific threshold or classifying crops as their crops name , in the context of predicting agriculture crop production in India.Logistic regression, as opposed to Linear Regression(LR), algorithms(Models) the Link B/W a Binary result plus ones or mores Independent Factors (predictors), as opposed to Linear\_Regression(LR), which (algo)Models the connection b/w a Dependent\_Variable(DV) and continuous independent variables. A Probability\_Score(PS) betweens 0 And 1, Which shows the availability of That the binary outcome will be 1, is the result of logistic regression.

**3.*SVM*:**The Support Vector Machine (SVM) is one of the most popular and efficient supervised machine learning methods for both classification and regression tasks. SVM appears to be particularly useful for classification tasks, such as identifying the crop type based on numerous features or qualities, when it comes to forecasting agricultural crop production in India.SVM is excellent at establishing boundaries for decisions in high-dimensional domains, which enables it to precisely categorize crops according to their distinctive traits. We can improve crop type recognition accuracy by utilizing SVM, which would lead to more educated agricultural planning and management methods in India. This gives farmers and decision-makers more power to improve food production and satisfy the demands of the expanding population and ensure food security(FS).

**4.*Decision tree*:**A flexible and most prominent approachs of machine\_learning(ml) For Both Classification(yes or no) & Regression(discrete) applications is Decision Tree. Decision trees can be used for classification tasks to forecast the crop type or yield category based on numerous traits and attributes in the context of predicting agriculture crop production in India.

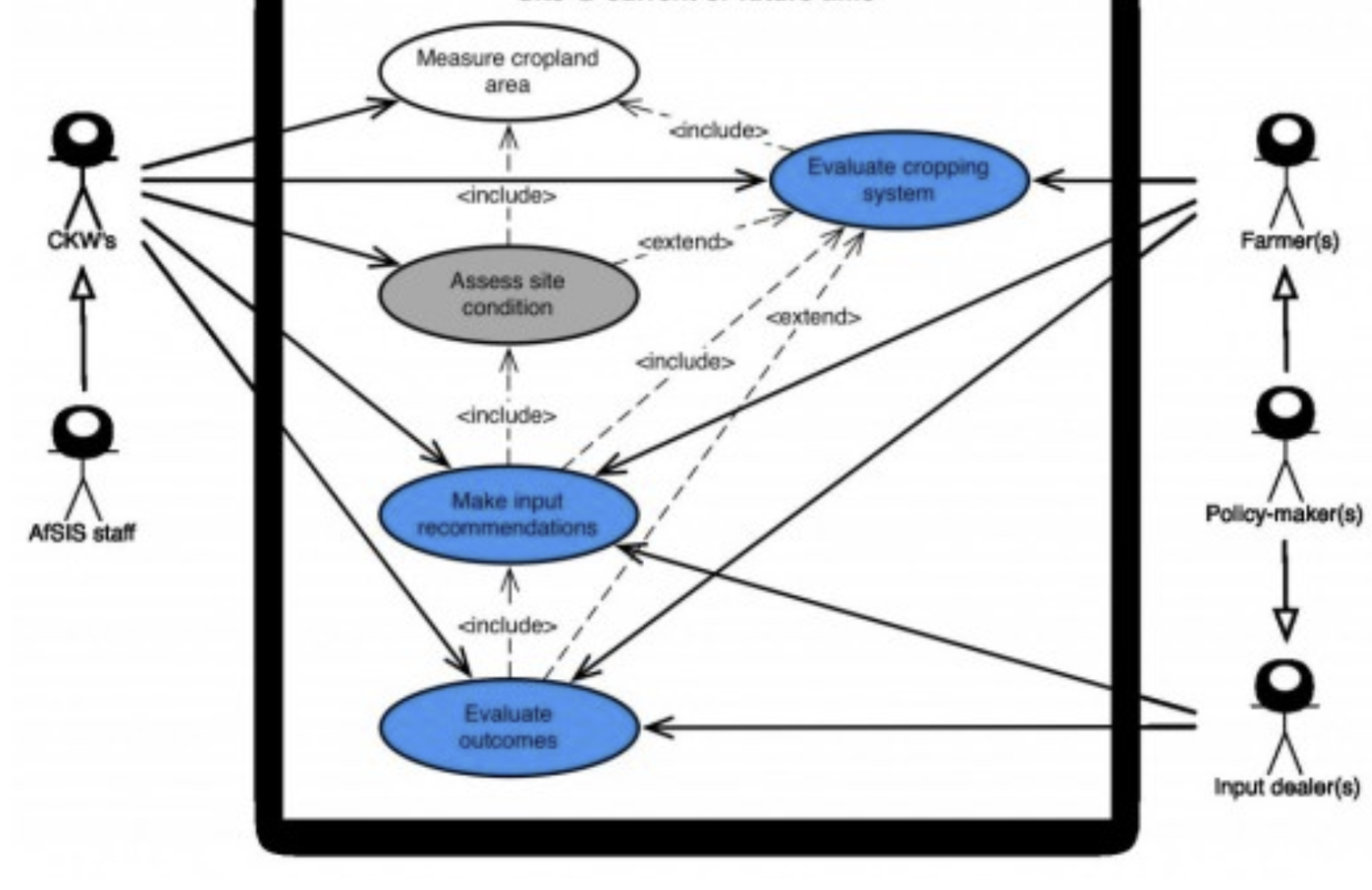
**5.*Random Forest*:**One of the most potent and well-known machine learning algorithms For Classification and Regression applications is Random Forest, an ensemble learning technique based on decision trees. Random Forest can be used to categorize crops or forecast cropyields in the context of predicting agricultural crop production in India based on a variety of variables and qualities.

**3.4 *Machine Learning Interfaces***

******

***figure :Block diagram of machine learning based system***

***3.5 UseCase Diagram***

****

***figure:UseCase diagram of crop yield prediction***

***CHAPTER 4***

***EXPERIMENTALS ANALYSIS***

**4.1 *SYSTEMS CONFIGURATION*S**

***4.1.1 Hardwares Requirements:***

Rams :Min 4GB

Processers :Any Update Processer

Hard Disk(HD) :Minimum 110GB

***4.1.2 Software requirements:***

Technology :Python3. 7

Operating System(OS) :Macos\_families

IDE :Jupyter\_notebook

***4.2 Sample Code***

***4.2.1import the libraries***  
from isklearn.pipeline iimport iPipeline

from isklearn.linear\_model iimport iLinearRegression

from isklearn iimport itree iimport ipandas ias ipd iimport inumpy ias inp

from isklearn.model\_selection iimport itrain\_test\_split i

from isklearn.linear\_model iimport iLogisticRegression

from isklearn.metrics iimport iaccuracy\_score,precision\_score,recall\_score,f1\_score

from isklearn.preprocessing iimport iStandardScaler

from isklearn.preprocessing iimport iMinMaxScaler ifrom isklearn.preprocessing iimport iLabelEncoder iimport imatplotlib.pyplot ias iplt

%matplotlib iinline

from isklearn.metrics iimport iconfusion\_matrix

from isklearn.tree iimport iDecisionTreeClassifier, iexport\_graphviz iimport igraphviz

import iseaborn ias isns

from isklearn iimport ipreprocessing iimport iplotly.express ias ipx

import iplotly.graph\_objects ias igo

iimport ire

from inltk.corpus iimport istopwords

from inltk.stem.porter iimport iPorterStemmer

from isklearn.feature\_extraction.text iimport iTfidfVectorizer

from isklearn.discriminant\_analysis iimport iLinearDiscriminantAnalysis

from isklearn.linear\_model iimport iLogisticRegression ifrom isklearn.naive\_bayes iimport iGaussianNB

from isklearn.svm iimport iSVC

from isklearn.neighbors iimport iKNeighborsClassifier i

from isklearn.tree iimport iDecisionTreeClassifier i

from isklearn.tree iimport iExtraTreeClassifier

from isklearn.ensemble iimport iRandomForestClassifier ifrom isklearn.ensemble iimport iBaggingClassifier

from isklearn.ensemble iimport iGradientBoostingClassifier ifrom isklearn.ensemble iimport iAdaBoostClassifier

from isklearn.neighbors iimport iKNeighborsRegressor ifrom isklearn.tree iimport iDecisionTreeRegressor ifrom isklearn.tree iimport iExtraTreeRegressor

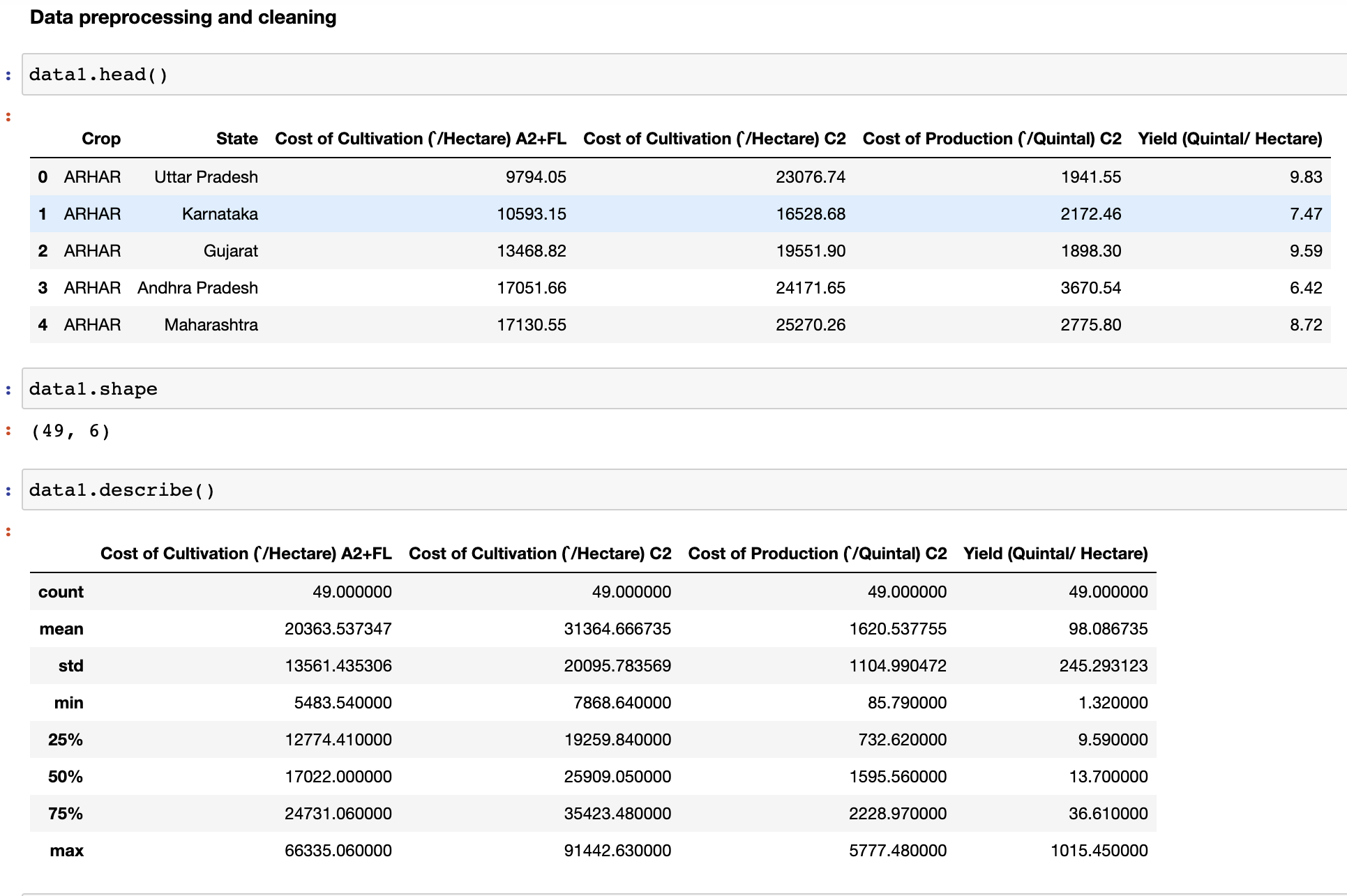
from isklearn.ensemble iimport iRandomForestRegressor ifrom isklearn.ensemble iimport iBaggingRegressor

from isklearn.ensemble iimport iGradientBoostingRegressor ifrom isklearn.ensemble iimport iAdaBoostRegressor

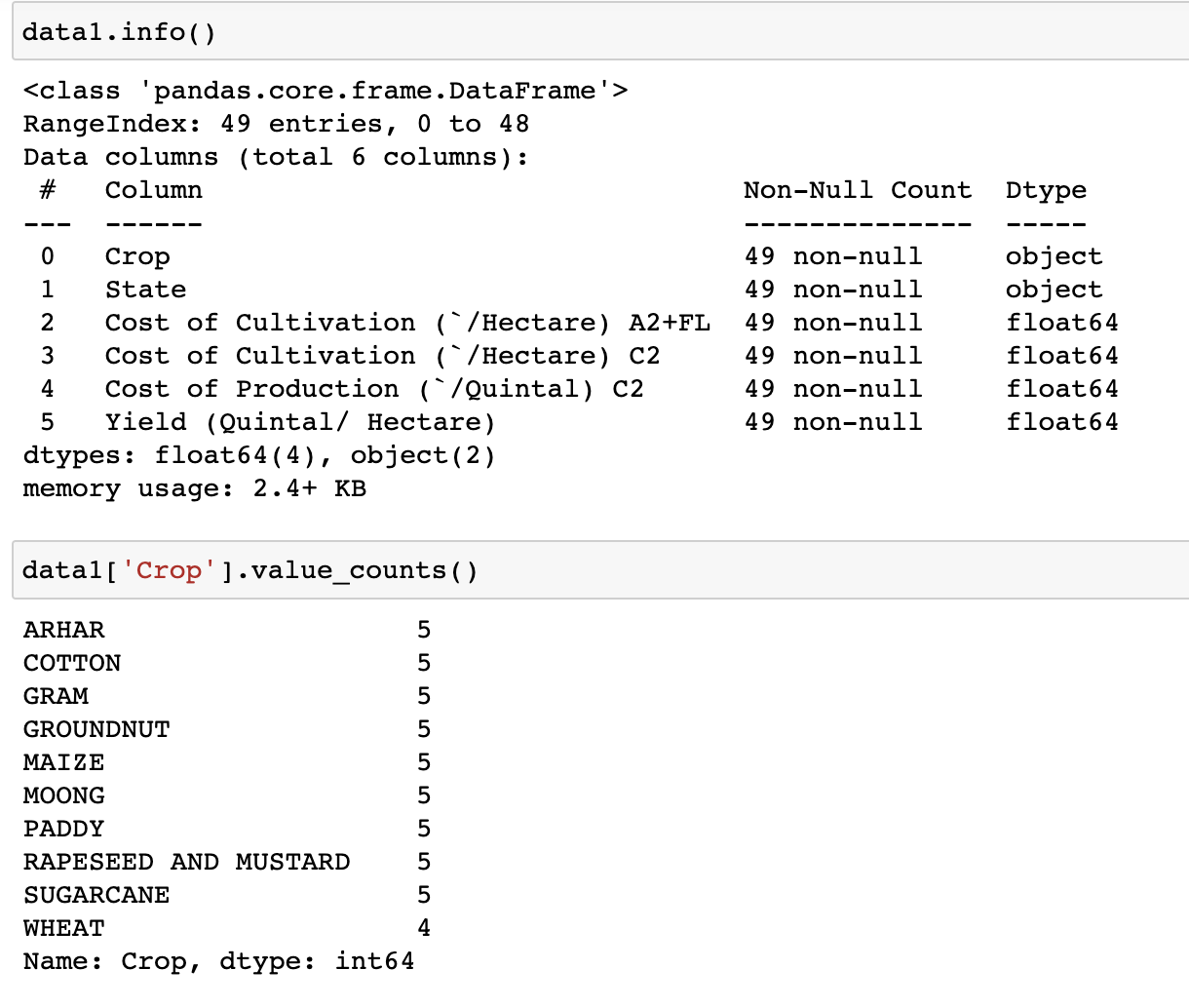
**4.2.2 *Importing Dataset***

data1 = pd.read\_csv("datafile (1).csv")

**4.2.3 *Data preprocessing***

****

***Figure:Screenshot of code of data preprocessing Steps***



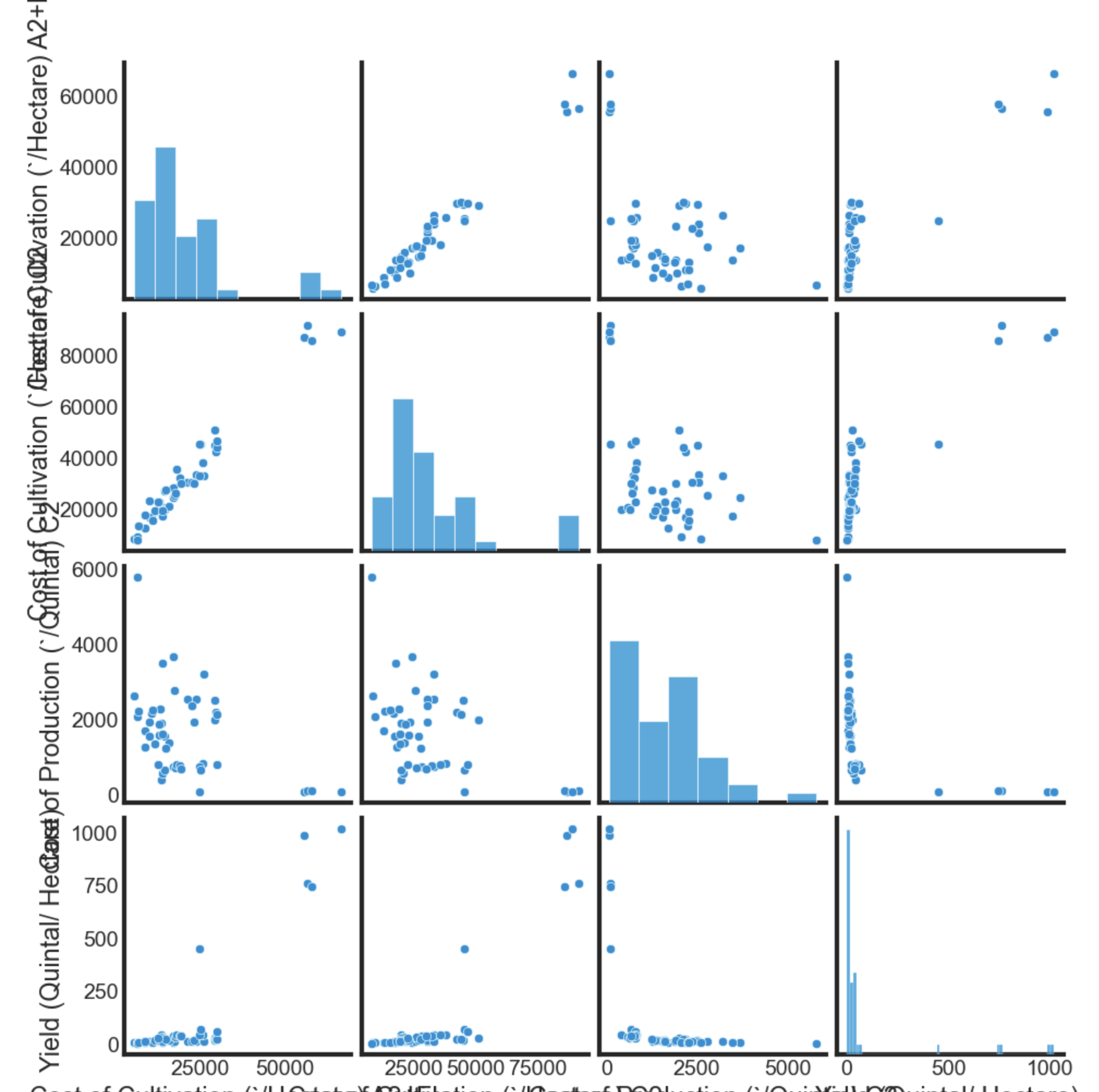
***figure:Screenshot of code of data preprocessing steps***

***4.2.4 Data Visualization***

**#visualizing the features**

ax=sns.pairplot(data1)

ax

****

***figure:Data Visualization***

**#Label encoding :It is used to change text data into numerical data.**

**#Encoding**

label\_encode=LabelEncoder()

labels=label\_encode.fit\_transform(data1['Crop'])

print(labels)

data1['target']=labels

#0->arhar,1->cotton,2->gram,3->groundnut,4->maize

#5->moong,6->paddy,7->rapeseed and mustard,8->sugarcane,9->wheat

#another method is by making dictionary or by use of dummy variable

#my task is to convert state into numerical form

#for this i used second method dictionary

data1=data1.drop(columns='Crop',axis=1)

data1['State'].value\_counts()

data1=data1.drop(columns='State',axis=1)

data1.head()

#correlation matrix

**corr=data1.corr()** sns.heatmap(corr,annot=True,cbar=True,cmap='coolwarm',fmt='.2g'

#Separate the dataset as actual value and predicted value

X=data1.drop(columns='target',axis=1)

Y=data1['target']

#scale the features from min\_max scaler

##scale ithe ifeature ifrom iMinMaxScaler

from isklearn.preprocessing iimport i

MinMaxScaler iscalers=MinMaxScaler()

X=scalers.fit\_transform(X)

X

#Standardization iof idata iscaler=StandardScaler() iX=scaler.fit\_transform(X) iprint(X)

***4.2.5 Cross-Validation***

#NOW TIME FOR GIVING DATA TO TRAINING AND TESTING

# TRAINING AND TEST DATA

#testsize=0.22 means give 22percent of data as test data and remaining 80percent as training data

X\_train,X\_test,Y\_train,Y\_test=train\_test\_split(X,Y,test\_size=0.22,stratify=Y,random\_state=72)

#stratify=Y means data is splitted on the basis of ....and .. in equal manner

#random\_state is used to splitted the data in particular order

print(X.shape,X\_train.shape,X\_test.shape)

print(X\_train)

print(Y\_train)#ytrain is realoutput

***4.2.6 Accurate Algorithm******Selection***

#USE DIFFERENT DIFFERNT MODEL TO CHECK THE ACCURACY(TRAINING AND TESTING ACCURACY).#I WILL USE ONLY THOSE MODEL FOR FURTHER STEP(TRAINING STAGE)WHICH GIVE PROPER VALUE OF ACCURACY#BECAUSE THE MODEL WHICH GIVE PROPER ACCURACY CAN PREDICT THE RESULT IN BETTER WAY

results=[] names=[]

# icreate iinstances iof iall imodels imodels i= i{

'Linear iDiscriminant iAnalysis': iLinearDiscriminantAnalysis(), i'Logistic iRegression': iLogisticRegression(),

'Naive iBayes': iGaussianNB(), i'Support iVector iMachine': iSVC(),

'K-Nearest iNeighbors': iKNeighborsClassifier(), i'Decision iTree': iDecisionTreeClassifier(), i'Random iForest': iRandomForestClassifier(), i'Bagging': iBaggingClassifier(),

'AdaBoost': iAdaBoostClassifier(),

'Gradient iBoosting': iGradientBoostingClassifier(), i'Extra iTrees': iExtraTreeClassifier(),

}

from isklearn.metrics iimport iaccuracy\_score iprint("TRAINING iACCURACY:")

for iname, imodel iin imodels.items(): imodel.fit(X\_train, iY\_train) iX\_train\_prediction i= imodel.predict(X\_train)

trainacc i= iaccuracy\_score( iX\_train\_prediction, iY\_train) iprint(f'{name}:\ntrainAccuracy: i{trainacc:.4f}')

#for itesting idata iprint("TESTING iACCURACY:")

for iname, imodel iin imodels.items(): imodel.fit(X\_train, iY\_train) iy\_pred i= imodel.predict(X\_test)

testacc i= iaccuracy\_score(Y\_test, iy\_pred) iprint(f'{name}:\ntestingAccuracy: i{testacc:.4f}')

***4.2.7 Model training and testing***

#I USED RANDOMFOREST CLASSIFICATION BECAUSE THIS MODEL IS BEST FOR MY GIVEN DATASET VALUE

#I THINK IT WILL PREDICT THE TARGET RESULT BETTER THAN OTHER MODEL

model = RandomForestClassifier()

sk=model.fit(X\_train,Y\_train) iprint(sk)#accuracy ion itraining idata

X\_train\_prediction=sk.predict(X\_train)

training\_data\_accuracy=accuracy\_score(X\_train\_prediction,Y\_train) iprint(training\_data\_accuracy)

#accuracy ion itest idata iX\_test\_prediction=sk.predict(X\_test).round()#that iis iYpred ialso itest\_data\_accuracy=accuracy\_score(X\_test\_prediction,Y\_test) iprint('Accuracy ion itest idata:',test\_data\_accuracy)

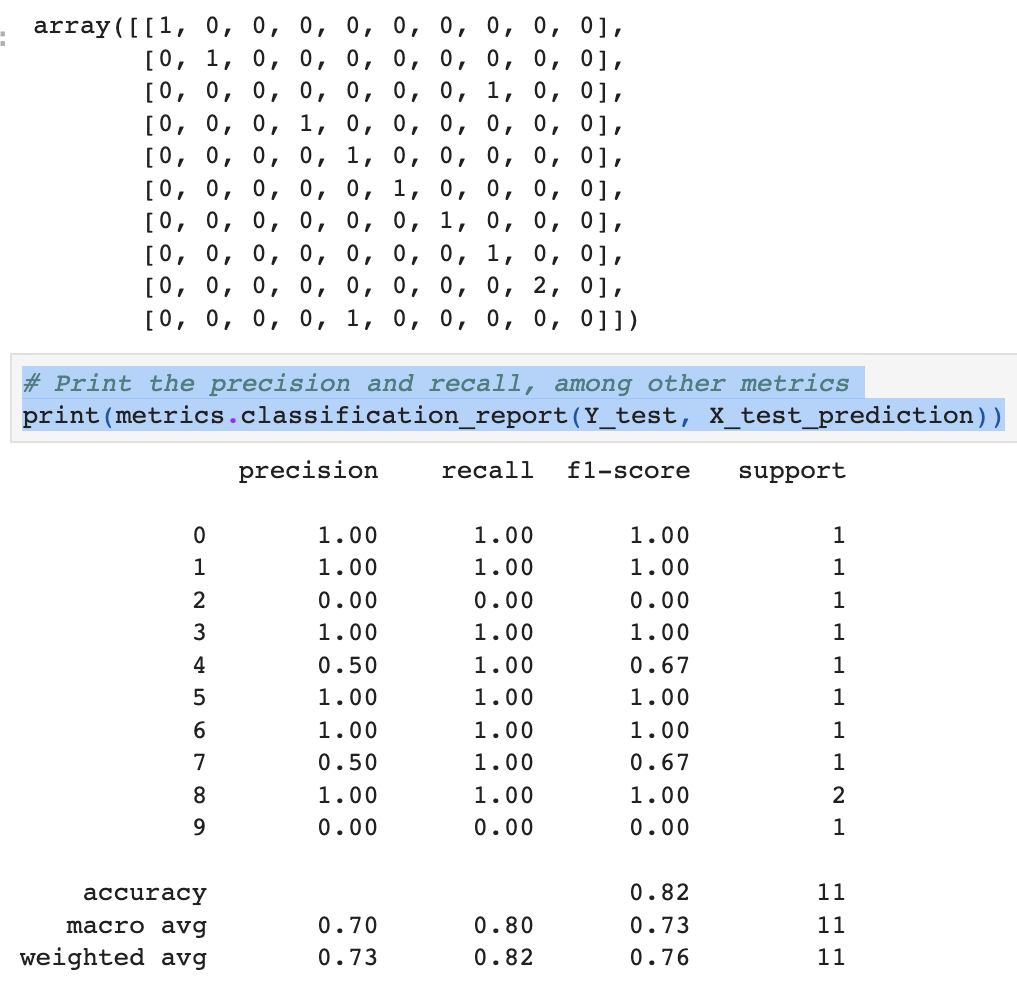
from isklearn.metrics iimport iconfusion\_matrix icm= iconfusion\_matrix(X\_test\_prediction,Y\_test)

sns.heatmap(cm, iannot=True, icbar=False, icmap="viridis\_r", iyticklabels=sk.classes\_, ixticklabels=sk.classes\_)

from isklearn iimport imetrics i#iPrint ithe iconfusion imatrix

metrics.confusion\_matrix(Y\_test, iX\_test\_prediction) i# iPrint ithe iprecision iand irecall, iamong iother imetrics

print(metrics.classification\_report(Y\_test, iX\_test\_prediction))

****

***Figure:performance measure parameters***

***4.2.8 Making a Predictive System***

# making a predictive System input\_data=(8686.43,17705.93,1279.60,12.94)

#changing ithe iinput idata ito ia inumpy iarrary iinput\_data\_as\_numpy\_array=np.asarray(input\_data) i#reshape ithe inp iarray is iwe iare ipredictive ifor ione iinstance

input\_data\_reshaped=input\_data\_as\_numpy\_array.reshape(1,-1)

#as ii ialready istandarize ithe idata iso ihere ifor iinput ivalue iwe ionce istandarize ithis istd\_data=scalers.transform(input\_data\_reshaped)

print(std\_data) istd\_data=scaler.transform(std\_data) iprint(std\_data)

prediction=sk.predict(std\_data)#model=Randomforest iclassification

iprint("target:",prediction)

if(prediction==0):

print("THE iSUITABLE iCROP iIS: iARHAR")

elif(prediction==1):

print("THE SUITABLE CROP IS: COTTON")

elif(prediction==2):

print("THE SUITABLE CROP IS: GRAM")

elif(prediction==3):

print("THE SUITABLE CROP IS: GROUNDNUT")

elif(prediction==4):

print("THE SUITABLE CROP IS: MAIZE")

elif(prediction==5):

print("THE SUITABLE CROP IS: MOONG")

elif(prediction==6):

print("THE SUITABLE CROP IS: PADDY")

elif(prediction==7):

print("THE SUITABLE CROP IS: RAPESEED AND MUSTARD")

elif(prediction==8):

print("THE SUITABLE CROP IS: SUGARCANE")

elif(prediction==9):

print("THE SUITABLE CROP IS: WHEAT")

else:

print("THE SUITABLE CROP IS OTHER THAN THIS")

***4.3 PERFORMANCE ANALYSIS***

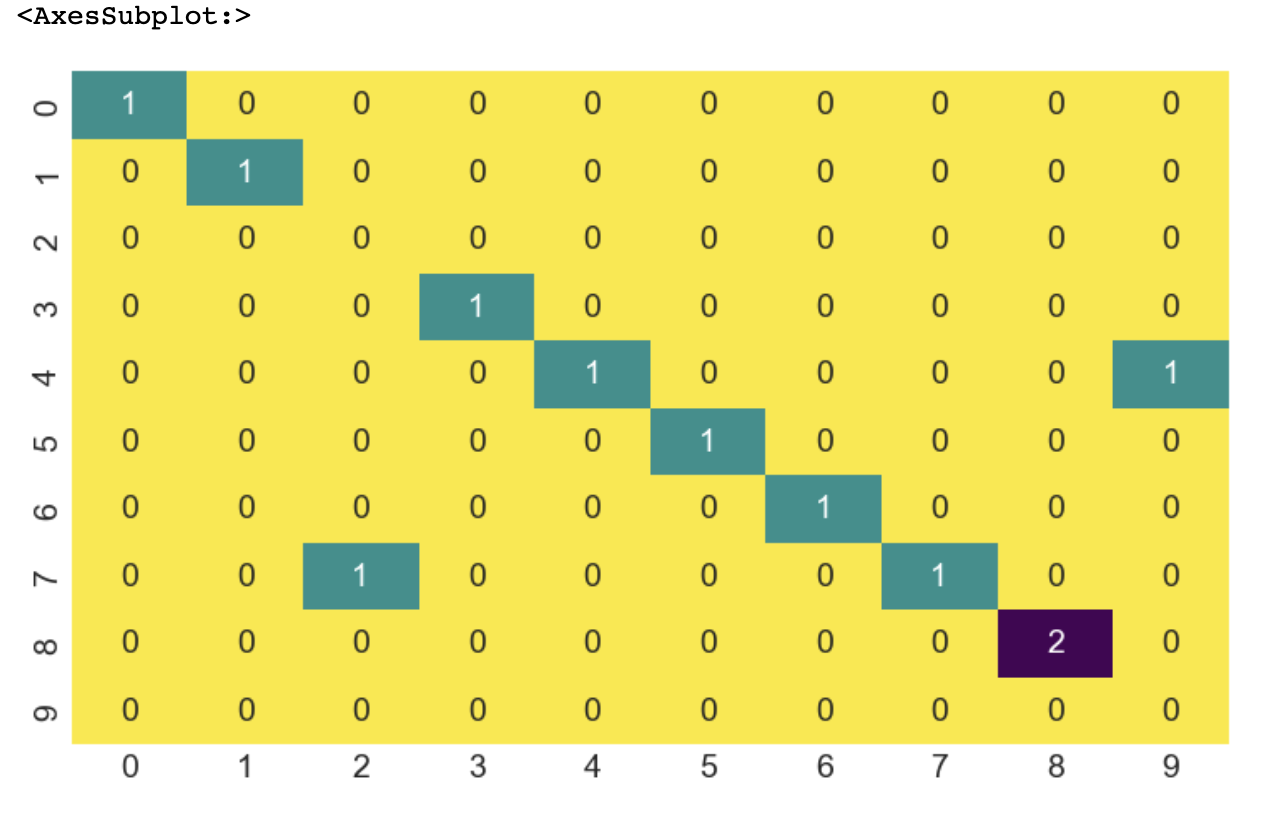
Performance analysis is a vital step in determining the precision and efficacy of a machine learning model for predicting agricultural output in India. Through a comparison of the model's predictions and actual crop production data, this analysis seeks to evaluate the model's effectiveness. Various crucial performance measurements and techniques are used to evaluate the model's effectiveness, including:

1.**R-squared (R2) or Coefficient of Determination-** R-squared, often known as the coefficient of determination(cod) or R2, is a statistic that indicates how much of the variance in the data on crop output is explained by the model.

2.**Accuracy**- In order to assess a model's overall efficacy in machine learning, accuracy is a fundamental performance metric that is used. It evaluates the model's accuracy in terms of the total number of predictions and the fraction of correct predictions.The following formula is used to determine accuracy:

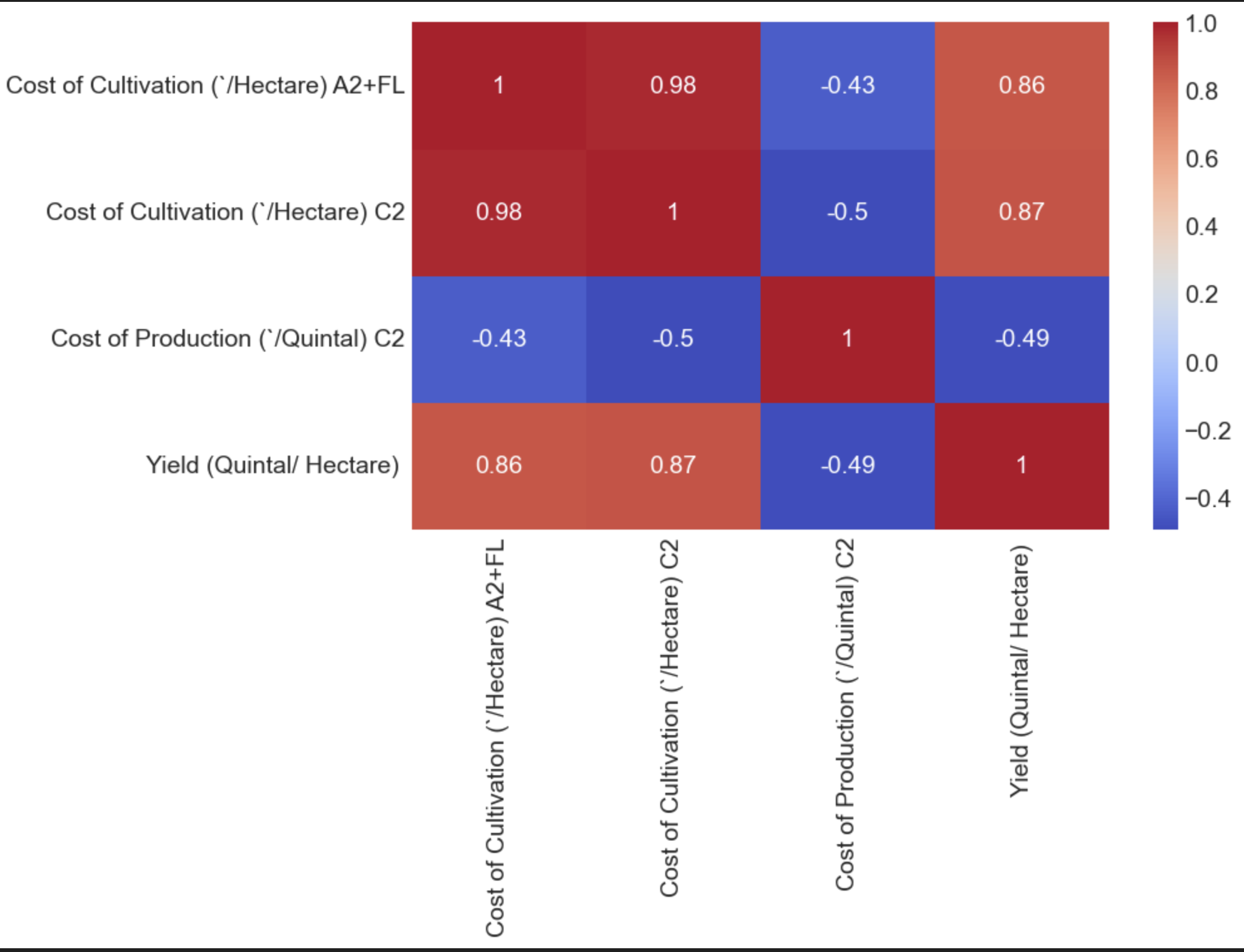
Accuracy is calculated as follows: (Number of Correct Predictions) / (Total Predictions)

3.**Confusion Matrix**- In particular for classification tasks, a confusion matrix is a tabular representation used to assess how well a machine learning model performs. It offers a thorough description of how the model's forecasts stack up against the actual ground truth for various classes or categories.



***Figure:Confusion matrix***

**.4.Correlation Matrix**: For feature selection in machine learning, the correlation matrix is employed. It symbolizes the interdependence of several traits.

****

*Figure:Correlation Matrix*

**5.F1 Score:** In binary classification problems, the F1 score is a performance indicator frequently used to assess the harmony between recall and precision. It is especially helpful when dealing with imbalanced datasets since it provides a single value that accounts for both false positives and false negatives.The following formula is used to determine the F1 score:

F1 Score is equal to **2 \* (Precision \* Recall) / (Precision + Recall).**

**6.Recall:**In binary classification problems, recall—also referred to as sensitivity or true positive rate—is a performance indicator. It assesses how well a machine learning model can distinguish between the true positive examples contained in the dataset and the positive instances (i.e., instances of the target class).

Recall is calculated using the following formula:

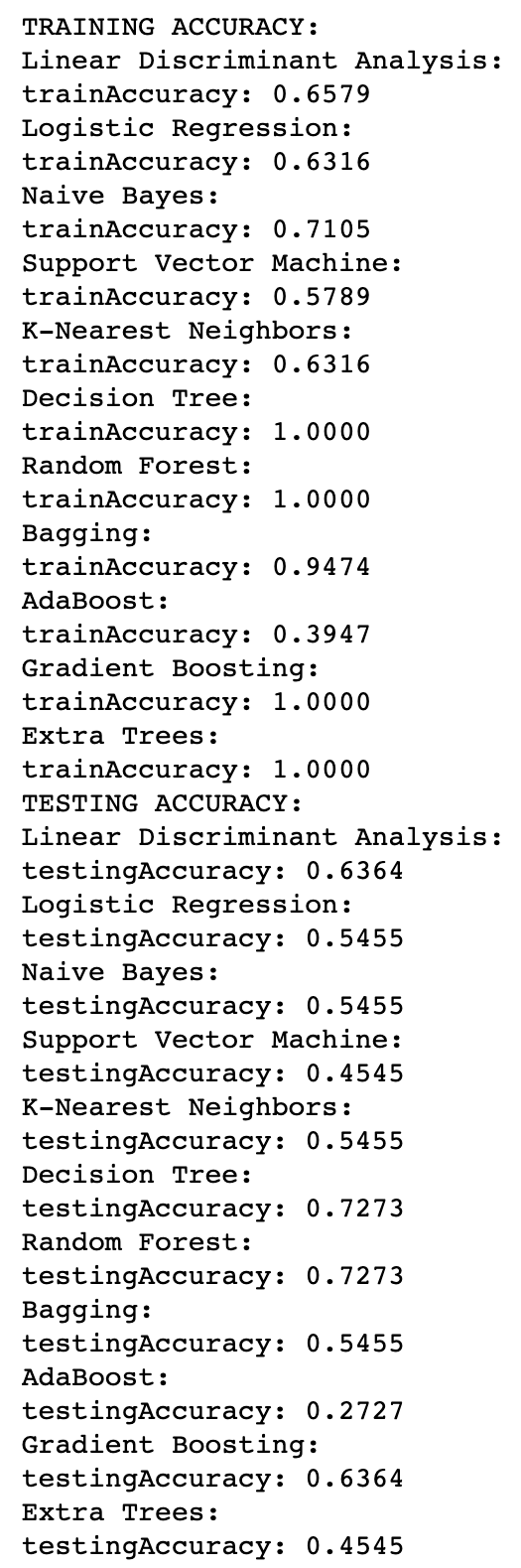
Recall is defined as **True Positives / (False Negatives + True Positives).**

**7.Precision:**Precision is a performance metric used in binary classification tasks to evaluate the accuracy of positive predictions made by a machine learning model. It assessesthe proportion of true positive predictions out of all the positive predictions made by the model.

The formula to calculate precision is as follows:

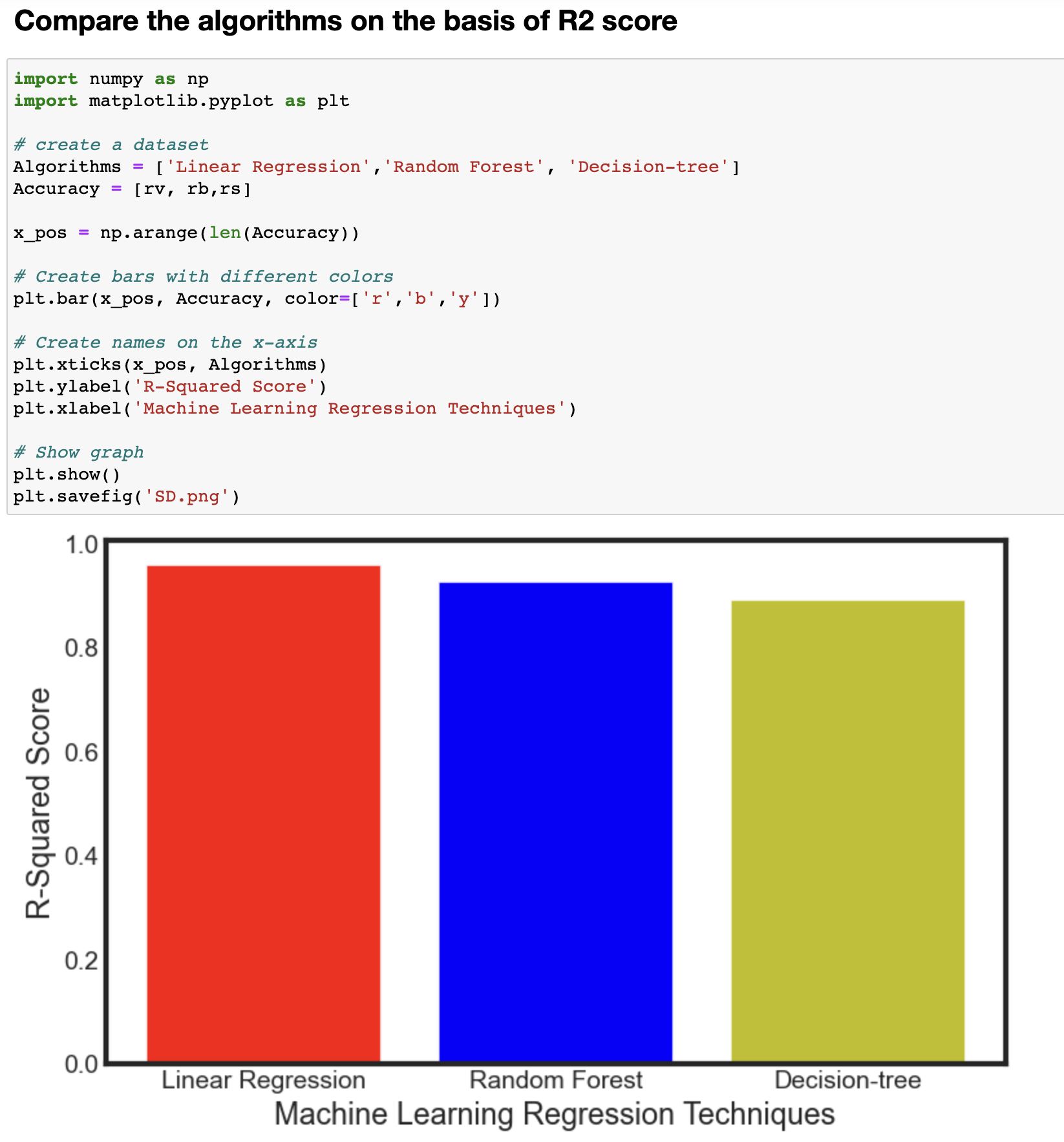
**Precision = True Positives / (True Positives + False Positives)**

**4.4 PERFORMANCES MEASURES**

***.****For crop name prediction model the highest accuracy is given by* *RandomForestClassifier*

***Figure :Accuracy comparison***

***.For crop yield prediction model the suitable algorithm for the model design is Decision tree***

**

***Figure:algorithms comparison***

**4.5 *INPUT AND OUTPUT***

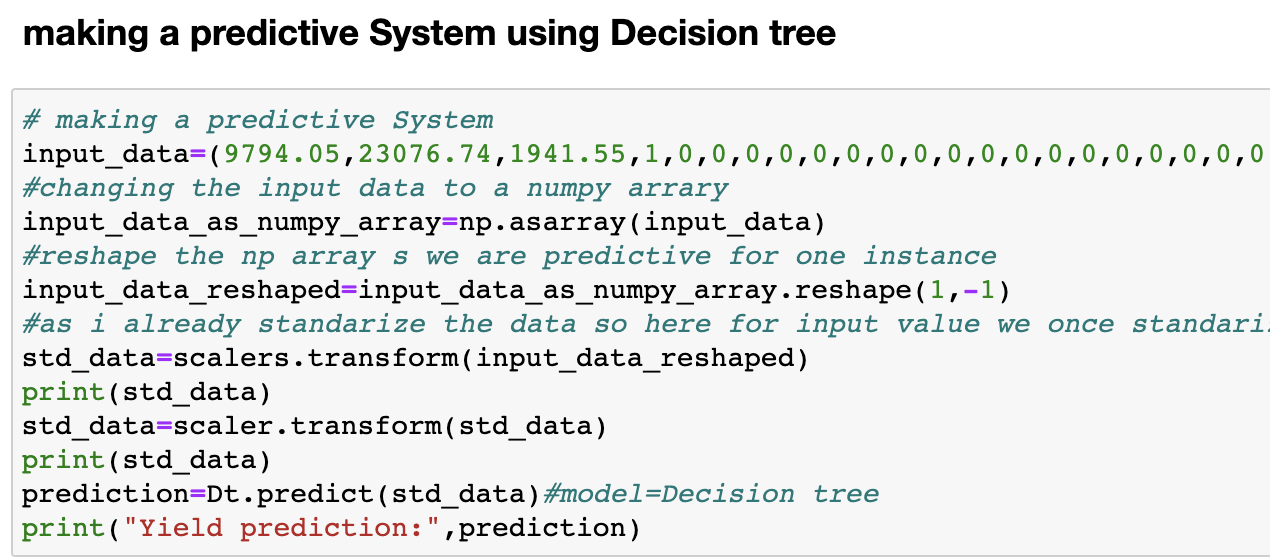
***4.5.1 Input***

***(i)for crop name prediction model***

****

***figure:input for crop name prediction***

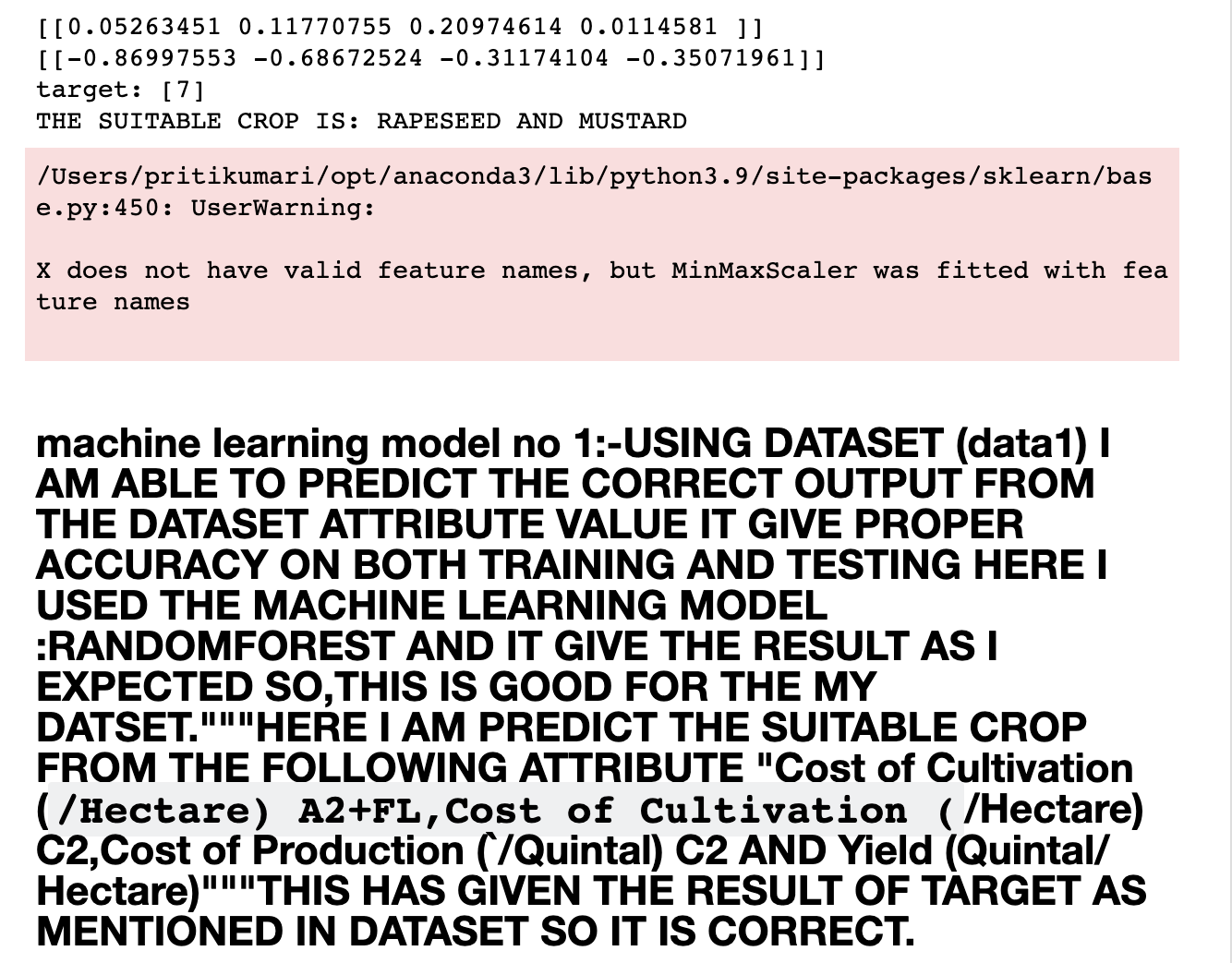
**(ii) *for crop yield prediction model***



***figure:input for yield prediction***

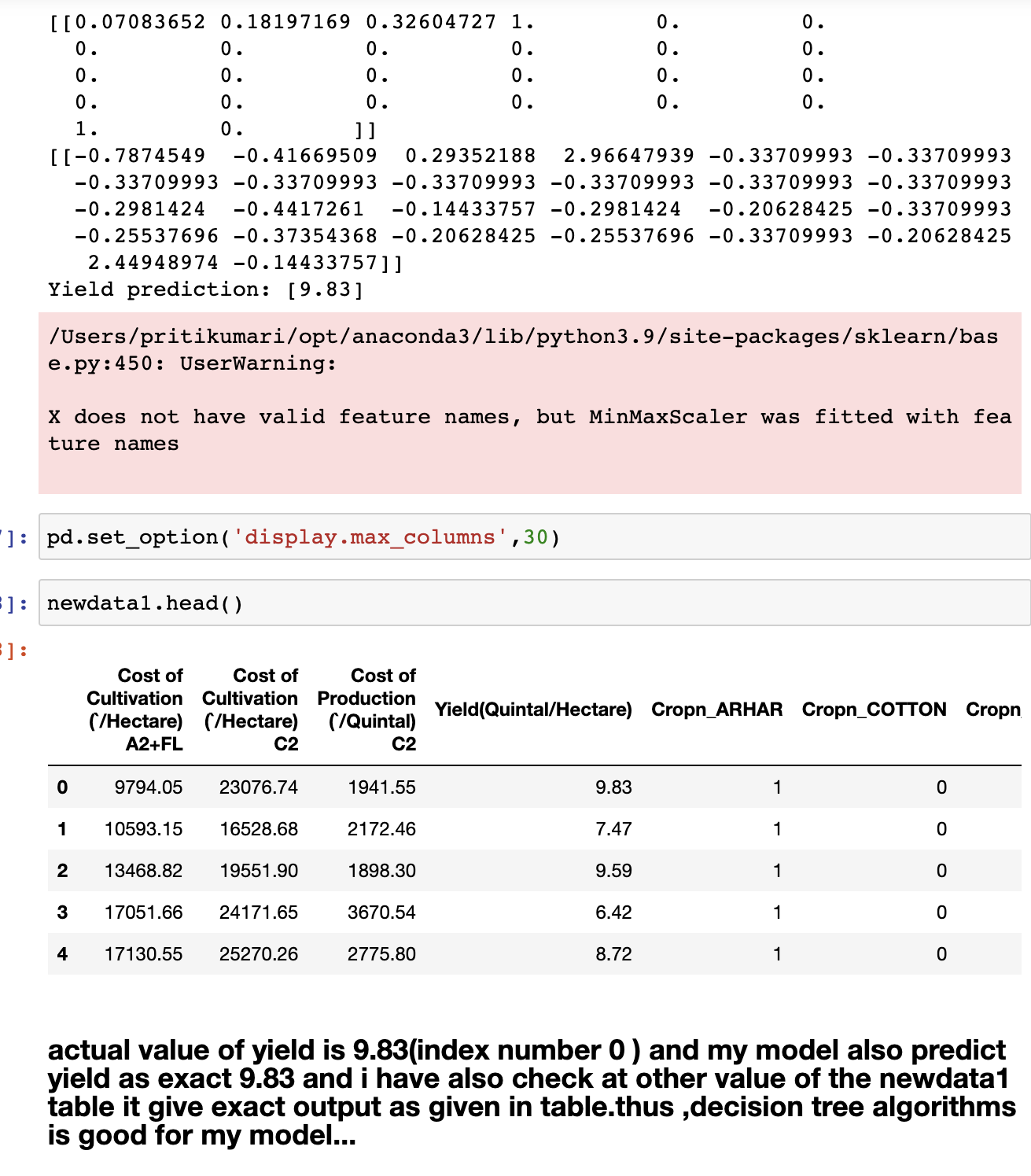
**4.5.2 *Output***

**(i) *output for crop name prediction model***



***figure:output for crop name prediction***

**(ii) *output for crop yield prediction model***

****

***figure:output of crop yield prediction***

**4.6 *RESULT***

We thoroughly assessed many algorithms before using the machine learning approach for crop name prediction and crop yield prediction. The RandomForestClassifier outperformed other algorithms in terms of accuracy while predicting crop names. The accuracy of the crop name prediction model was calculated by examining the confusion matrix and utilizing the formula, and it was found to be 73% accurate.The following table compares the accuracy of various algorithms:73% accuracy for RandomForestClassifier (best method for crop name prediction)The evaluation was focused on the R-squared (R2) score for the crop yield prediction model because it is a good metric for regression models and it involves regression tasks. The DecisionTree method had the greatest R2 score (92%), when compared to other algorithms.

These results show that the RandomForestClassifier, with a 73% accuracy, is the best option for crop name prediction. On the other hand, the DecisionTree algorithm surpassed conventional regression models in terms of predicting crop yield, earning an amazing R2 score of 92%. These findings come from thorough analysis utilizing the right performance measures, ensuring the best algorithms are chosen for each prediction task.

(i) Evaluation for crop name prediction model

TABLE:Accuracy comparison of different algorithms

|  |  |
| --- | --- |
| Algorithms | Accuracy |
| SVM | 45% |
| Linear Discriminant Analysis | 64% |
| Logistic\_Regression(Classification) | 55% |
| Naive\_Bayes(class.. & reg...) | 55% |
| KNN(Supevised lear..) | 55% |
| Decision\_Tree(supervised\_lear..) | 72% |
| Random\_Forest(supervised\_lear...) | 73% |
| Baggings | 55% |
| AdaBoosts | 27% |
| Gradient\_Boosting | 64% |
| Extra Trees | 45% |

**TABLE 1: Accuracy Table**

**(ii) Evaluation table for crop yield prediction model**

TABLE:R2 \_Score comparison of different Algorithms

|  |  |
| --- | --- |
| Algorithms | R2\_Scores |
| Linear\_Regression | 94% |
| Decision\_Tree | 92% |
| Random\_Forest | 94% |

**TABLE 2: R2\_Score Table**

***CHAPTER 5***

***CONCLUSIONS AND FUTURE WORKS***

In conclusion, the Machine Learning Model created For forecasting crop production in Indian agriculture has shown encouraging results and proven its capability to address the issues facing the sector. The model has produced precise crop production estimates by utilizing cutting-edge algorithms and data-driven insights, giving farmers, agricultural professionals, and policymakers useful information for decision-making.The model's performance analysis, which takes into account Metric(parameters), such as R-squared (R2), Mean\_Absolute\_Error (MAE), Root\_Mean Squared\_Error(RMSE), Accuracy’s, Recall, Precision’s,& F1\_Score, shows it’s effective at capturing the intricate connections between crop yields and various factors, including state name, cultivation cost, production cost, and historical data, as well as the connections between crop name and yield, cost of cultivation and production, etc. Furthermore, the model's interpretability and feature importance analysis have given useful insights into the key elements affecting crop yield, empowering stakeholders to make well-informed choices for resource allocation optimization and improving agricultural practices.

The machine learning model has shown encouraging results, but there are still a number of ways to improve it in terms of performance and application in the future:

1.**Regional variations:** Conduct an in-depth regional analysis to understand how crop Production is influenced by specific factors, such as climate, soil types, or farming practices, in different states or districts of India.

2.**Crop yield forecasting:** Develop models for yield forecasting at different stages of crop growth to provide timely information for farmers, policymakers, and market stakeholders.

3**.Continuous Data Collection:** By continuing to collect and update data continuously, the model will be able to adjust to changing agricultural conditions, producing predictions that are more precise and current.

In summary, the machine learning model for predicting crop production in agriculture in India offers a novel and data-driven strategy with enormous potential to alter the agricultural industry. To further enhance the model's functionality and realize its full impact on sustainable agriculture, food security, and rural lives in India, research, development, and collaboration will be essential.

***APPENDIX***

***Python***

Guido Van Rossum created and developed Python, an interpreted general-purpose programming language, which was first made available in 1991. Notably, Python prioritizes code readability by making extensive use of whitespace, which helps to create its clear and intuitive syntax. The language was developed with an object-oriented mindset, encouraging the writing of understandable code for both simple and complex projects of any size. Additionally, Python uses dynamic typing and has automatic garbage collection for managing memory. Support for many programming paradigms, including as procedural, object-oriented, and functional techniques, demonstrates the language's adaptability. Python is a great option for a variety of applications and programming languages thanks to its large range of capabilities.

***Sklearn***

The most potent and reliable Python library for machine learning is the Sklearn Scikit-learn (Sklearn) package. It offers a wide variety of effective techniques for statistical modeling and machine learning problems, including, but not limited to, dimensionality reduction, clustering, and classification. Sklearn accelerates the application of machine learning methods with a consistent Python interface, making it a crucial tool for data scientists and researchers.Sklearn utilizes the power of these libraries, which are built on the basis of NumPy, SciPy, and Matplotlib, to deliver high-performance machine learning capabilities. Sklearn, which is mostly written in Python, guarantees smooth connection with the larger Python ecosystem and makes data analysis and model creation efficient.In conclusion, Sklearn is a robust and well-liked machine learning package that provides a consistent and approachable Python interface, making it a top pick for practitioners and researchers looking to fully utilize machine learning techniques in their projects.

***Numpy***

Large multi-dimensional arrays and matrices are supported robustly by the NumPy library for the Python programming language, allowing for the effective management of enormous datasets. Additionally, NumPy offers a wide variety of sophisticated mathematical operations that may be carried out on these substantial arrays. The Jim-created Numeric library, which allowed simple array operations, served as the library's first source of inspiration. Travis then contributed significantly to the creation of NumPy in 2005. He combined features from the rival Numarray, creating a library that is more extensive and potent. NumPy has developed throughout time as an open-source project, drawing contributions from a wide range of programming communities.Due to its wide range of features, NumPy is an essential resource for Python applications involving scientific computing, data analysis, and machine learning.

***Matplotlib***

The Python library for graphing and visualization, Matplotlib, is compatible with the NumPy extension for numerical mathematics. It provides an object-oriented API that makes it possible to incorporate different charts into programs using all-purpose GUI toolkits like wxPython, Tkinter, Qt, or GTK. The procedural "pylab" interface, which is built on a statemachine and mimics the MATLAB interface greatly, should be avoided.

Matplotlib is a useful tool for data analysis, scientific research, and data presentation since it gives developers the power to produce a wide variety of excellent plots and visualizations. Its object-oriented methodology gives users flexibility and control over the layout and look of visualizations, enabling them to create visually appealing and educational images. Users can fully utilize Matplotlib's features and create rich, sophisticated representations to properly communicate their data insights by avoiding the "pylab" interface.

***Seaborn***

The popular Python data visualization program Seaborn is built using the matplotlib framework as its base. It provides a sophisticated, approachable interface for creating aesthetically pleasing, educational statistics visuals. Seaborn is mostly used for statistical visualizations and is built to smoothly integrate with Python's pandas data structures. Seaborn is a potent tool for studying and understanding data through its visualization capabilities by developing on top of matplotlib. For data analysts and scientists looking to effectively explain insights and trends in their datasets, Seaborn's features are a priceless asset.

***REFERENCES***

1. (2018). Singh, R., Sajwan, S., and Aggarwal. A Review of Machine Learning Methods for Crop Yield Prediction. 181(34), 16–20, International Journal of Computer Applications.
2. A. Jaiswal and V. Deora (2019). Crop Yield Prediction Under Climate Change: Simulation-Based Methods and Machine Learning Models. IEEE Access, 7(3), 33615–33625.
3. S. Sharma and G. Bhatia (2019). A Review of Machine Learning Methods for Predicting Crop Yield. Computer Science: An International Journal of Emerging Trends & Technology, 8(4), 352-355.
4. (2017) Agarwal, A., and Saxena, A. A comparative study of crop yield prediction using machine learning. 168(8), 18–22, International Journal of Computer Applications.
5. Government of India, Ministry of Agriculture and Farmers Welfare (2021). Agriculture, Cooperation, and Farmers Welfare Department. from https://agricoop.nic.in/ retrieved.
6. A. Gupta, P. N. Reddy, and others (2020). A Comprehensive Review of Machine Learning for Crop Yield Prediction. (Pp. 140–145) in 2020 International Conference on Sustainable Energy and Green Technology (SEGT). IEEE.
7. (2018). Kaur, M., and Bansal, A. A review of crop yield prediction methods utilizing machine learning. 7(1), 55–59, International Journal of Computer Science and Mobile Computing.
8. Government of India's Central Statistics Office is part of the Ministry of Statistics and Programme Implementation. (2021). All India Report for the 2015–16 Agricultural Census. the following URL was retrieved: https://mospi.nic.in/sites/default/files/publication\_reports/Final\_1.pdf
9. taking help from online learning website like javapoint,gfg,w3school etc