

PREDICTION OF CROP YIELD AND COST BY FINDING BEST ACCURACY USING MACHINE LEARNING APPROACH

Submitted in partial fulfillment of the requirements for the award of Bachelor
of Engineering degree in Computer science and
Engineering

BY

KARRI KRANTHI VARDHAN REDDY (Reg. No. 37110325)

KOVVURI BALAGOPAL REDDY (Reg. No. 37110379)



**DEPARTMENT OF COMPUTER SCIENCE AND
ENGINEERING**

SCHOOL OF COMPUTING

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

**Accredited with Grade "A" by NAAC I 12B Status
by UGC I Approved by AICTE**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600 119**

March - 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC I 12B Status
by UGC I Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai –
600 119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING **BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the bonafide work of **KARRI KRANTHI VARDHAN REDDY (Reg. No. 37110325), KOVVURI BALAGOPAL REDDY (Reg. No. 37110379)** who carried out the project entitled "**Prediction of crop yield and cost by finding best accuracy using machine learning approach**" under my supervision from November 2020 to March 2021.

Internal Guide
Dr. P. ASHA, M.E., Ph.D.,

Head of the Department
Dr. L. Lakshmanan M.E., Ph.D.,
Dr. S. Vigneshwari M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I KARRI KRANTHI VARDHAN REDDY (Reg. No. 37110325), KOVVURI BALAGOPAL REDDY (Reg. No. 37110379) hereby declare that the Project Report entitled **“Prediction of crop yield and cost by finding best accuracy using machine learning approach”** is done by me under the guidance of **Dr. P. ASHA, M.E., Ph.D.**, Department of Computer Science and Engineering at Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D.**, Dean, School of Computing and **Dr. L. Lakshmanan M.E., Ph.D.**, and **Dr. S. Vigneshwari M.E., Ph.D.**, Heads of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. P. ASHA, M.E., Ph.D.**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many Ways for the completion of the project.

ABSTRACT

Among worldwide, agriculture has the major responsibility for improving the economic contribution of the nation. However, still the most agricultural fields are under developed due to the lack of deployment of ecosystem control technologies. Due to these problems, the crop production is not improved which affects the agriculture economy. Hence a development of agricultural productivity is enhanced based on the plant yield prediction. To prevent this problem, Agricultural sectors have to predict the crop from given dataset using machine learning techniques. The analysis of dataset by supervised machine learning technique (SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments etc. A comparative study between machine learning algorithms had been carried out in order to determine which algorithm is the most accurate in predicting the best crop. The results show that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy with entropy calculation, precision, Recall, F1Score, Sensitivity, Specificity.

TABLE OF CONTENTS

ABSTRACT

v

LIST OF FIGURES

viii

CHAPTER NO.	TITLE	PAGE NO.
1.	INTRODUCTION	1
	1.1 INTRODUCTION	1
	1.2 DOMAIN OVERVIEW	1
	1.3 EXISTING SYSTEM	4
	1.4 PROPOSED SYSTEM	5
2.	LITERATURE REVIEW	6
3.	AIM AND SCOPE	11
	3.1 AIM OF THE PROJECT	11
	3.1.1 OBJECTIVES	11
	3.2 SCOPE	11
	3.3 PROJECT GOALS	11
4.	SYSTEM DESIGN & METHODOLOGY	12
	4.1 PROJECT REQUIREMENTS	12
	4.1.1 SOFTWARE REQUIREMENTS	12
	4.1.2 HARDWARE REQUIREMENTS	12
	4.2 SOFTWARE DESCRIPTION	12
	4.2.1 CONDA	13
	4.2.2. THE JUPYTER NOTE BOOK	13
	4.2.3 NOTE BOOK DOCUMENT	13
	4.2.4 JUPYTER NOTEBOOK APP	14
	4.2.5 KERNEL	14
	4.2.6 NOTEBOOK DASH BOARD	14

4.3	METHODOLOGY	15
4.4	MODEL SELECTION	16
4.5	MODULES	16
4.5.1	DATA VALIDATION AND PRE PROCESSING TECHNIQUE	17
4.5.2	EXPLORATION DATA ANALYSIS OF VISUALIATION AND TRAINING	19
4.5.3	PERFORMANCE MEASUREMENTS OF LOGISTIC REGRESSION & DECISION TREE	22
4.5.4	PERFORMANCE MEASUREMENTS OF SUPPORT VECTOR CLASSIFIER & RANDOM FOREST	23
4.5.5	PERFORMANCE MEASUREMENTS OF KNN & NAÏVE BAYES	25
4.5.6	GUI BASED PREDICTION OF CROP YIELD & YIELD COST	26
4.6	PARAMETER CALCULATIONS	28
4.6.1	ACCURACY CALCULATION	28
5.	RESULTS AND DISCUSSION	34
6.	CONCLUSION AND FUTURE WORK	37
	REFERENCES	38
	APPENDIX	
	A. SOURCE CODE	39
	B. SCREEN SHOTS	50
	C. JOURNAL PAPER	51

LIST OF FIGURES

FIGURE No.	FIGURE NAME	PAGE NO.
1.1	PROCESS OF MACHINE LEARNING	2
4.1	GIVEN DATA FRAME	18
4.2	SPLITTING THE GIVEN DATA SET	18
4.3	AFTER PROCESSING GIVEN DATA FRAME	19
4.4	COST PRODUCTION FOR 100 kg BY COUNTS	20
4.5	PERCENTAGE LEVEL OF CROP YIELD PRODUCTION BY STATE	21
4.6	CROP YIELD BY EACH SEASON	29
4.7	CROP YIELD COST PRODUCTION BY EACH SEASON	30
5.1	DATA SET	34
5.2	CONFUSION MATRICS RESULT	34
5.3	ACCURACY RESULTS OF CROP YIELD	35
5.4	ACCURACY RESULTS OF CROP YIELD COST	35
5.5	PREDICTION OUT PUT 1	36
5.6	PREDICTION OUT PUT 2	36

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

In developing countries, farming is considered as the major source of revenue for many people. In modern years, the agricultural growth is engaged by several innovations, environments, techniques and civilizations. In addition, the utilization of information technology may change the condition of decision making and thus farmers may yield the best way. For decision making process, data mining techniques related to the agriculture are used. Data mining is a process of extracting the most significant and useful information from the huge amount of datasets. Nowadays, we used machine learning approach with developed in crop or plant yield prediction since agriculture has different data like soil data, crop data, and weather data. Plant growth prediction is proposed for monitoring the plant yield effectively through the machine learning techniques.

It is also applicable for the automated process of farming is the beginning of a new era in Bangladesh that will be suitable for the farmers who seek experts to take suggestion about the appropriate crop on specific location of their land and don't want to forget any step of the cultivation throughout the process. Although, the opinion from experts is the most convenient way, this application is designed to give accurate solution in fastest manner possible. This research's main objective is to bring farming process a step closer to the digital platform.

1.2 DOMAIN OVERVIEW

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning

focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. It

feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated into three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It is provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based on the way they “learn” about data to make predictions: supervised and unsupervised learning. Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function from input variables(X) to discrete output variables(y). In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, biometric identification, document classification etc.



Fig 1.1 Process of Machine learning

Supervised Machine Learning is the majority of practical machine learning uses supervised learning. Supervised learning is where have input

variables (X) and an output variable (y) and use an algorithm to learn the mapping function from the input to the output is $y = f(X)$. The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (y) for that data. Techniques of Supervised Machine Learning algorithms include logistic regression, multi-class classification, Decision Trees and support vector machines etc. Supervised learning requires that the data used to train the algorithm is already labeled with correct answers. Supervised learning problems can be further grouped into Classification problems. This problem has as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for categorical for classification. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. A classification problem is when the output variable is a category, such as “red” or “blue”.

Agriculture is one of the most important occupations practiced in our country. It is the broadest economic sector and plays an important role in overall development of the country. About 60 % of the land in the country is used for agriculture in order to suffice the needs of 1.2 billion people. Thus, modernization of agriculture is very important and thus will lead the farmers of our country towards profit. Data analytic (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. Earlier yield prediction was performed by considering the farmer's experience on a particular field and crop. However, as the conditions change day by day very rapidly, farmers are forced to cultivate more and more crops. Being this as the current situation, many of them don't have enough knowledge about the new crops and are not completely aware of the benefits they get while farming them. Also, the farm productivity can be increased by understanding and

forecasting crop performance in a variety of environmental conditions. Thus, the proposed system takes the location

of the user as an input. From the location, the nutrients of the soil such as Nitrogen, Phosphorous, Potassium is obtained. This static data is the crop production and data related to demands of various crops obtained from various websites. It applies machine learning and prediction algorithm to identify the pattern among data and then process it as per input conditions.

1.3 EXISTING SYSTEM

It present a crop/weeds classification approach based on a three-steps procedure. The first step is a robust pixel-wise segmentation (i.e., soil/plant) and image patches containing plants are extracted in the second step. The third step, a deep CNN for crop/weed classification is used. The extracted blobs in the masked image containing plants information are fed to a CNN classifier based on a fine-tuned model of VGG-16 exploiting the ability of deep CNN in object classification and to reduce the limitations of CNNs in generalizing when a limited amount of data is available. The classification step can then be specialized to the types of plants needed by the application scenario. It evaluated the complete pipeline, including the first background removal phase and the subsequent classification stage. Experimental results demonstrate that can achieve good classification results on challenging data.

Precision agriculture is gaining increasing attention because of the possible reduction of agricultural inputs (e.g., fertilizers and pesticides) that can be obtained by using high-tech equipment, including robots. To focus on an agricultural robotics system that addresses the weeding problem by means of selective spraying or mechanical removal of the detected weeds. To describe a deep learning based method to allow a robot to perform an accurate weed/crop classification using a sequence of two Convolutional Neural Networks (CNNs) applied to RGB images. The first network, based on encoder-decoder segmentation architecture, performs a pixel wise, plant-type agnostic, segmentation between vegetation and soil that enables to extract a set of connected blobs representing plant instances.

Drawbacks

- It can't determine to improve the classification accuracy of our pipeline.
- Connecting the bridge manually and some corruption are happened.
- Private sectors domination high, profit low and credits not getting concern farmer.

1.4 PROPOSED SYSTEM

The designed system will recommend the most suitable crop for particular land. Based on weather parameter and soil content such as Rainfall, Temperature, Humidity and pH. The system takes the required input from the farmers or sensors such as Temperature, Humidity and pH. This all inputs data applies to machine learning predictive algorithms like Support Vector Machine (SVM) and Decision tree to identify the pattern among data and then process it as per input conditions. The system recommends the crop for the farmer and also recommends the amount of nutrients to be add for the predicted crop. The system has some other specification like displaying approximated yield in q/acre, and the production price of the crop.

Advantages

- Our goal is push for assisting farmers, government using our predictions. All these publications state they have done better than their competitors but there is no article or public mention of their work being used practically to assist the farmers. If there are some genuine problems in rolling out that work to next stage, then identify those problems and try solving them.
- It is targeted to those farmers who wish to professionally manage their farm by planning, monitoring and analyzing all farming activities.

CHAPTER 2 LITERATURE REVIEW

Title: Design and Characterization of a Fringing Field Capacitive Soil Moisture Sensor

Author: Manash Protim Goswami , Babak Montazer, and Utpal Sarma, Member, IEEE

Year: 2018

Description:

The optimization and implementation of a fringing field capacitive soil moisture sensor using the printed circuit board technology. It includes the analysis of a novel configuration of an interdigital sensor for measuring soil moisture with two existing configurations. The optimized designs were simulated by using a 3-D finite-element method and fabricated by using a copper clad board. The performance of the fabricated sensors was evaluated using four soil samples collected from different locations. The observations were compared with the standard gravimetric method to evaluate the soil water content of the samples. The characterization method and the results of the whole sensing system are discussed in terms of calibration, dynamic test, and repeatability.

Title: Preliminary Study of Soil Available Nutrient Simulation Using a Modified WOFOST Model and Time-Series Remote Sensing Observations

Author: Zhiqiang Cheng 1,2 ID , Jihua Meng 1,*, Yanyou Qiao 1, Yiming Wang 1,2, Wenquan Dong 1 and Yanxin Han 1,2

Year: 2017

Description:

The approach of using multispectral remote sensing (RS) to estimate soil available nutrients (SANs) has been recently developed and shows promising results. This method overcomes the limitations of commonly

used methods by building a statistical model that connects RS- based crop growth and nutrient content. However, the stability and accuracy of this model require improvement. In this article, we replaced the

statistical model by integrating the World Food Studies (WOFOST) model and time series of remote sensing (T-RS) observations to ensure stability and accuracy. Time series of HJ-1 A/B data was assimilated into the WOFOST model to extrapolate crop growth simulations from a single point to a large area using a specific assimilation method. Because nutrient- limited growth within the growing season is required and the SAN parameters can only be used at the end of the growing season in the original model, the WOFOST model was modified. Notably, the calculation order was changed, and new soil nutrient uptake algorithms were implemented in the model for nutrient-limited growth estimation. Finally, experiments were conducted in the spring maize plots of Hongxing Farm to analyze the effects of nutrient stress on crop growth and the SAN simulation accuracy. The results confirm the differences in crop growth status caused by a lack of soil nutrients. The new approach can take advantage of these differences to provide better SAN estimates. In general, the new approach can overcome the limitations of existing methods and simulate the SAN status with reliable accuracy.

Title: Distinguishing Heavy-Metal Stress Levels in Rice Using Synthetic Spectral Index Responses to Physiological Function Variations

Author: Ming Jin, Xiangnan Liu, Ling Wu, and Meiling Liu

Year: 2016

Description:

Accurately assessing the heavy-metal contamination in crops is crucial to food security. This study provides a method to distinguish heavy-metal stress levels in rice using the variations of two physiological functions as discrimination indices, which are obtained by assimilation of remotely sensed data with a crop growth model. Two stress indices, which correspond to daily total CO₂ assimilation and dry-matter conversion coefficient were incorporated into the World Food Study (WOFOST) crop growth model and calculated by assimilating the model with leaf area index (LAI), which was

derived from time-series HJ1-CCD data. The stress levels are not constant with rice growth; thus, to improve the reliability, the two stress indices were obtained at both the first and the latter half periods of

rice growth. To compare the stress indices of different stress levels, a synthetic stress index was established by combining the two indices; then, three types of stress index discriminant spaces based on the synthetic index of different growth periods were constructed, in which the two-dimensional discriminant space based on two growth periods showed the highest accuracy, with a misjudgment rate of 4.5%. When the discrimination rules were applied at a regional scale, the average correct discrimination rate was 95.0%.

Title: Estimation of Organic Matter Content in Coastal Soil Using Reflectance Spectroscopy Research

Author: ZHENG Guanghui¹, Dongryeol RYU^{2,*}, JIAO Caixia¹ and HONG Changqiao¹

Year: 2015

Description:

Rapid determination of soil organic matter (SOM) using regression models based on soil reflectance spectral data serves an important function in precision agriculture. “Deviation of arch” (DOA)-based regression and partial least squares regression (PLSR) are two modeling approaches to predict SOM. However, few studies have explored the accuracy of the DOA-based regression and PLSR models. Therefore, the DOA-based regression and PLSR were applied to the visible near-infrared (VNIR) spectra to estimate SOM content in the case of various dataset divisions. A two-fold cross-validation scheme was adopted and repeated 10000 times for rigorous evaluation of the DOA-based models in comparison with the widely used PLSR model. Soil samples were collected for SOM analysis in the coastal area of northern Jiangsu Province, China. The results indicated that both modelling methods provided reasonable estimation of SOM, with PLSR outperforming DOA-based regression in general. However, the performance of PLSR for the validation dataset decreased more noticeably. Among the four DOA-based regression models, a linear model provided the best estimation of

SOM and a cutoff of SOM content (19.76 g kg^{-1}), and the performance for calibration and validation datasets was consistent. As the SOM content exceeded 19.76 g

kg⁻¹, SOM became more effective in masking the spectral features of other soil properties to a certain extent. This work confirmed that reflectance spectroscopy combined with PLSR could serve as a non-destructive and cost-efficient way for rapid determination of SOM when hyper spectral data were available. The DOA-based model, which requires only 3 bands in the visible spectra, also provided SOM estimation with acceptable accuracy.

Title: Remote Sensing and Geospatial Technological Applications for Site-specific Management of Fruit and Nut Crops: A Review

Author: Sudhanshu Sekhar Panda, Gerrit Hoogenboom, and Joel O. Paz

Year: 2010

Description:

Growth observation, impact assessment, and timely strategic response to small variations in crop production is known as precision agriculture. It has been used in a wide range of agricultural activities including field crop production, dairy farming, horticulture, and forest management. Site-specific crop management (SSCM) is one facet of precision agriculture which involves spatial referencing, crop and climate monitoring, attribute mapping, decision support systems, and differential action. SSCM is carried out with a greater degree of precision through the use of geospatial technologies. Geospatial technology is a combination of four essential tools: remote sensing, geographic information systems (GIS), global positioning systems (GPS), and information technology or data management. SSCM has become very common in management of field and row crops during recent years. Yet, its application for non-traditional horticultural crops has not been very widespread. Companies, such as Lanworth (<http://lanworth.com/>), which gather intelligence on natural resources, use geospatial technologies to map, analyze, and forecast agriculture and forest output. This is based on crop and forest growth parameters which is a large amount of data, terabytes in size. It helps in the SSCM of agricultural and forest products and increases the yield

in a sustainable manner. The application of aerial or satellite imaging, along

with GPS and GIS, is the first step towards the successful application of SSCM for fruit and nut crops. The use of SSCM for horticultural crops such as fruit and nut crops has potential for increasing net returns and optimizing resource. The delineation of orchards and spatial analysis using geospatial technology can provide additional information for management decision making, such as the determination of fruit yield, the quantification and scheduling of precise and proper fertilizer, irrigation needs, and the application of pesticides for pest and disease management. Ultimately, it will improve profits for producers. The use of remote sensing has become common for the general detection of the growth and health of orchards on a larger scale. However, digital imaging technology is increasingly being used for intensive site-specific management of orchards as well, for instance: estimating the amount of fruits on individual trees, fruit quality, and also leaf area index or crown cover. The overall goal to provide a review of studies that use geospatial technology and especially remote sensing for conducting SSCM in fruit and nut crops, including oranges, peaches, pecans, apples, grapes, blueberries, as well as other crops around the world.

It provides a comprehensive review on the development and application of SSCM for fruit and nut crops. The review also confirmed that not many studies (compared to row crop agriculture) have been conducted so far on fruit and nut crops' SSCM and one-stop of information on available technology for the fruit and nut crop SSCM. The study provides a detailed application of geospatial and information technology in the blueberry crop management with a case study in Georgia, U.S., and in a few locations in southeast Georgia.

CHAPTER 3 AIM AND SCOPE

3.1 AIM OF PROJECT

The primary aim of this proposed system is to assist the farmers in such a way that they can get all the necessary and detailed information of the cost of production and production yield based on their area of land, climate, water levels which will help the farmers to grow the crops in a much better way.

3.1.1 OBJECTIVES

The proposed model is based on the new technologies which makes the whole process efficient. It is going to be very important and profitable for the whole society especially for the farmers. This system will provide more accurate information about yield & cost of production which will be very helpful for the farmers to process the production.

3.2 SCOPE

The scope of this project is to investigate a dataset of crop records for agricultural sector using machine learning technique. To identifying crop predicting by farmer is more difficult. We try to reduce this risk factor behind selection of the crop.

3.3 PROJECT GOALS

- Exploration data analysis of variable identification
- Uni-variate data analysis
- Exploration data analysis of bi-variate and multi-variate
- Method of Outlier detection with feature engineering
- Comparing algorithm to predict the result

CHAPTER 4

SYSTEM DESIGN & METHODOLOGY

4.1 PROJECT REQUIREMENTS

4.1.1 Software Requirements

Operating System	: Windows
Tool	: Anaconda with Jupyter Notebook

4.1.2 Hardware requirements

Processor	: Pentium IV/III
Hard disk	: minimum 80 GB
RAM	: minimum 2 GB

4.2 SOFTWARE DESCRIPTION

Anaconda is a free and open distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Package versions are managed by the package management system “Conda”. So, Anaconda distribution comes with more than 1,400 packages as well as the Conda package and virtual environment manager called Anaconda Navigator.

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, install them in an environment, run the packages and update them. It is available for Windows, macOS and Linux.

The following applications are available by default in Navigator:

- JupyterLab
- Jupyter Notebook

- QtConsole

- Spyder
- Glueviz
- Orange
- Rstudio
- Visual Studio Code

4.2.1 Conda

Conda is an open source, cross-platform, language - agnostic package manager and environment management system that installs, runs and updates packages and their dependencies. It was created for Python programs, but it can package and distribute software for any language (e.g., R), including multi-languages. The Conda package and environment manager is included in all versions of Anaconda, Miniconda, and Anaconda Repository.

4.2.2 The Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

4.2.3 Notebook document

Notebook documents (or “notebooks”, all lower case) are documents produced by the Jupyter Notebook App, which contain both computer code (e.g. python) and rich text elements (paragraph, equations, figures, links, etc...). Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc.) as well as executable documents which can be run to perform data analysis.

4.2.4 Jupyter Notebook App

The *Jupyter Notebook App* is a server-client application that allows editing and running notebook documents via a web browser. The *Jupyter Notebook App* can be executed on a local desktop requiring no internet access (as described in this document) or can be installed on a remote server and accessed through the internet. In addition to displaying/editing/running notebook documents, the *Jupyter Notebook App* has a “Dashboard” (Notebook Dashboard), a “control panel” showing local files and allowing to open notebook documents or shutting down their kernels.

4.2.5 kernel

A notebook *kernel* is a “computational engine” that executes the code contained in a Notebook document. The *ipython kernel*, referenced in this guide, executes python code. Kernels for many other languages exist (official kernels). When you open a Notebook document, the associated *kernel* is automatically launched. When the notebook is *executed* (either cell-by-cell or with menu *Cell -> Run All*), the *kernel* performs the computation and produces the results. Depending on the type of computations, the *kernel* may consume significant CPU and RAM. Note that the RAM is not released until the *kernel* is shut-down.

4.2.6 Notebook Dashboard

The *Notebook Dashboard* is the component which is shown first when you launch Jupyter Notebook App. The *Notebook Dashboard* is mainly used to open notebook documents, and to manage the running kernels (visualize and shutdown). The *Notebook Dashboard* has other features similar to a file manager, namely navigating folders and renaming/deleting files.

4.3 METHODOLOGY

Download and install anaconda and get the most useful package for machine learning in Python.

Load a dataset and understand its structure using statistical summaries and data visualization.

machine learning models, pick the best and build confidence that the accuracy is reliable.

Python is a popular and powerful interpreted language. Unlike R, Python is a complete language and platform that you can use for both research and development and developing production systems. There are also a lot of modules and libraries to choose from, providing multiple ways to do each task. It can feel overwhelming.

The best way to get started using Python for machine learning is to complete a project.

It will force you to install and start the Python interpreter (at the very least).

It will give you a bird's eye view of how to step through a small project.

It will give you confidence, maybe to go on to your own small projects.

When you are applying machine learning to your own datasets, you are working on a project. A machine learning project may not be linear, but it has a number of well-known steps:

Define Problem.

Prepare Data.

Evaluate Algorithms.

Improve Results.

Present Results.

The best way to really come to terms with a new platform or tool is to work through a machine learning project end-to-end and cover the key steps.

Namely, from loading data, summarizing data, evaluating algorithms and making some predictions. Here is an overview of what we are going to cover:

Installing the Python anaconda platform.

Loading the dataset.

Summarizing the dataset.

Visualizing the dataset.

Evaluating some algorithms.

Making some predictions.

4.4 MODEL SELECTION

This is the most exciting phase in Applying Machine Learning to any Dataset. It is also known as Algorithm selection for Predicting the best results. Usually Data Scientists use different kinds of Machine Learning algorithms to the large data sets. But, at high level all those different algorithms can be classified in two groups : supervised learning and

unsupervised learning. Supervised learning : Supervised learning is a type of system in which both input and desired output data are provided. Input and output data are labeled for classification to provide a learning basis for future data processing. Supervised learning problems can be further grouped into **Regression** and **Classification** problems.

A **regression** problem is when the output variable is a real or continuous value, such as “salary” or “weight”. A **classification** problem is when the output variable is a category like filtering emails “spam” or “not spam”

Unsupervised Learning : Unsupervised learning is the algorithm using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. In our dataset we have the outcome variable or Dependent variable i.e Y having only two set of values, either M (Malign) or B(Benign). So we will use Classification algorithm of supervised learning.

4.5 Modules

- Data validation and pre-processing technique (Module-01)
- Exploration data analysis of visualization and training a model by

given attributes (Module-02)

- Performance measurements of logistic regression and decision tree algorithms (Module-03)
- Performance measurements of Support vector classifier and Random forest (Module-04)
- Performance measurements of KNN and Naive Bayes (Module-05)
- GUI based prediction of crop yield and yield cost (Module-06)

4.5.1 Data validation and pre-processing technique (Module-01)

Data validation process:

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers uses this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time- consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model. For example, time series data can be analyzed by regression algorithms; classification algorithms can be used to analyze discrete data. (For example to show the data type format of given dataset)

	State_Name	District_Name	Crop_Year	Season	Crop	Area	rainfall	Average Humidity	Mean Temp	Cost of Cultivation (/Hectare) C2	Cost of Production (/Quintal) C2	Yield (Quintal/Hectare)	cost of production per yield
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254.0	0.012360	57	62	23076.74	1941.55	9.83	19085.4365
1	Andaman and Nicobar Islands	NICOBARS	2001	Kharif	Arecanut	1254.0	0.084119	56	58	12610.85	1691.66	6.83	11554.0378
2	Andaman and Nicobar Islands	NICOBARS	2002	Whole Year	Arecanut	1258.0	0.080064	58	53	32683.46	3207.35	9.33	29924.5755
3	Andaman and Nicobar Islands	NICOBARS	2003	Whole Year	Arecanut	1261.0	0.181051	57	58	13209.32	2228.97	5.90	13150.9230
4	Andaman and Nicobar Islands	NICOBARS	2004	Whole Year	Arecanut	1264.7	0.035446	63	67	22560.30	1595.56	13.57	21651.7492

Fig 4.1 Given data frame

Data Validation/ Cleaning/Preparing Process:

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

```
#preprocessing, split test and dataset, split
X = df.drop(labels='CPPY', axis=1)
#Response variable
y = df.loc[:, 'CPPY']
```

```
#We'll use a test size of 30%. We also stratify
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, stratify=y)
print("Number of training dataset: ", len(X_train))
print("Number of testing dataset: ", len(X_test))
print("Total number of dataset: ", len(X_train) + len(X_test))
```

```
Number of training dataset: 163704
Number of testing dataset: 70160
Total number of dataset: 233864
```

Fig 4.2 Splitting the given dataset

Data Pre-processing:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw

format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format; for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set.

	State_Name	District_Name	Crop_Year	Season	Crop	Area	R	H	T	CC	CP	Y	CPPY
0	0	410	3	1	2	2025	33	45	10	21	30	13	126
1	0	410	4	1	2	2025	121	44	6	3	26	7	72
2	0	410	5	4	2	2030	118	46	1	33	45	11	200
3	0	410	6	4	2	2033	172	45	6	4	36	4	78
4	0	410	7	4	2	2037	75	51	15	20	24	23	144

Fig 4.3 After preprocessing given data frame

4.5.2 Exploration data analysis of visualization and training a model by given attributes (Module-02)

Exploration data analysis of visualization:

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts

that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and inapplied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

How to chart time series data with line plots and categorical quantities with bar charts.

How to summarize data distributions with histograms and box plots.

How to summarize the relationship between variables with scatter plots.

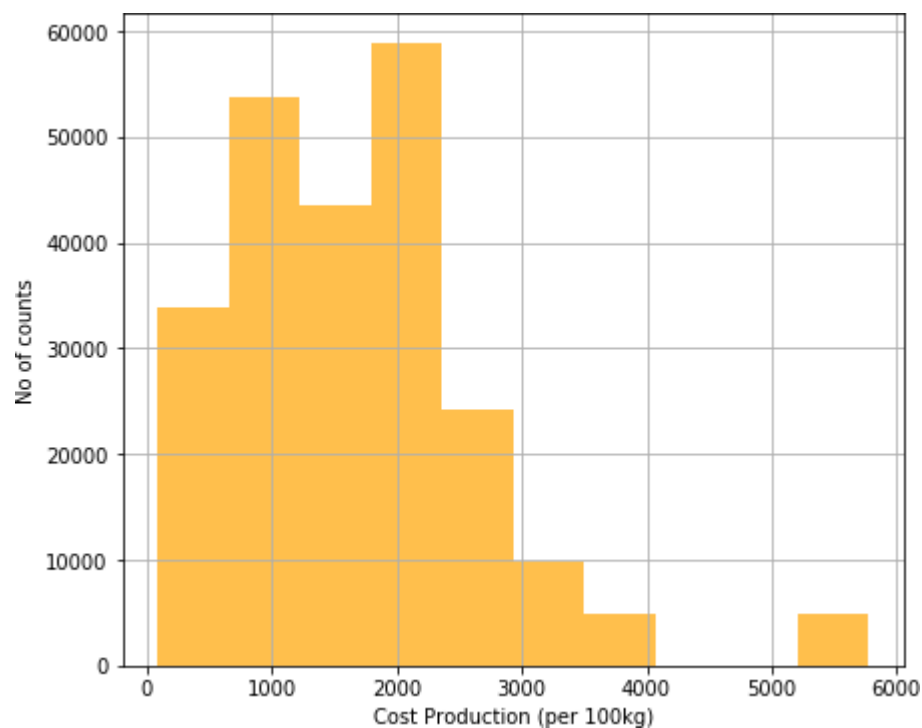


Fig 4.4 Cost production per 100kg by counts

Many machine learning algorithms are sensitive to the range and distribution of attribute values in the input data. Outliers in input data can skew

and mislead the training process of machine learning algorithms resulting in longer training times, less accurate models and ultimately poorer results.

Even before predictive models are prepared on training data, outliers can result in misleading representations and in turn misleading interpretations of collected data. Outliers can skew the summary distribution of attribute values in descriptive statistics like mean and standard deviation and in plots such as histograms and scatterplots, compressing the body of the data. Finally, outliers can represent examples of data instances that are relevant to the problem such as anomalies in the case of fraud detection and computer security.

It couldn't fit the model on the training data and can't say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise. Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

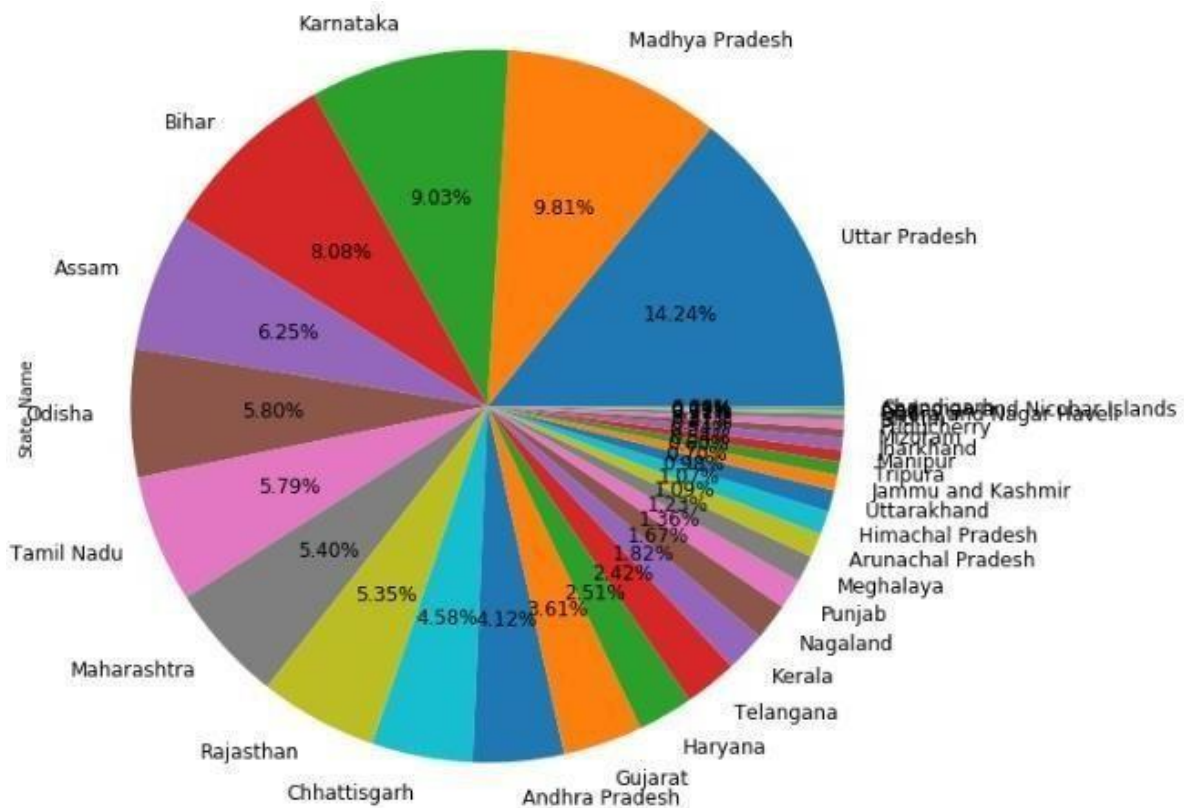


Fig 4.5 Percentage level of crop yield production by state

4.5.3 Performance measurements of logistic regression and decision tree algorithms (Module-03)

Logistic Regression:

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

In other words, the logistic regression model predicts $P(Y=1)$ as a function of X . Logistic regression Assumptions:

- Binary logistic regression requires the dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
- Only the meaningful variables should be included.
- The independent variables should be independent of each other. That is, the model should have little.
- The independent variables are linearly related to the log odds.
- Logistic regression requires quite large sample sizes.

Decision Tree:

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It

works for both continuous as well as categorical output variables.

Assumptions of Decision tree:

- At the beginning, we consider the whole training set as the root.
- Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.
- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or internal node.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed.

This process is continued on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes in the top of the tree have more impact towards in the classification and they are identified using the information gain concept. A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise or outliers.

4.5.4 Performance measurements of Support vector classifier and Random forest (Module-04)

Support Vector Machines (SVM)

A classifier that categorizes the data set by setting an optimal hyper plane

between data. I chose this classifier as it is incredibly versatile in the number of different kernelling functions that can be applied and this model can

yield a high predictability rate. Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms. They were extremely popular around the time they were developed in the 1990s and continue to be the go-to method for a high-performing algorithm with little tuning.

How to disentangle the many names used to refer to support vector machines.

The representation used by SVM when the model is actually stored on disk.

How a learned SVM model representation can be used to make predictions for new data.

How to learn an SVM model from training data.

How to best prepare your data for the SVM algorithm.

Where you might look to get more information on SVM.

Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of 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 algorithm of the same type i.e. multiple decision *trees*, resulting in a *forest of trees*, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

The following are the basic steps involved in performing the random forest algorithm:

Pick N random records from the dataset.

Build a decision tree based on these N records.

Choose the number of trees you want in your algorithm and repeat steps 1 and 2.

In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

4.5.5 Performance measurements of KNN and Naive Bayes (Module-05)

K-Nearest Neighbor (KNN):

K-Nearest Neighbor is a supervised machine learning algorithm which stores all instances correspond to training data points in n-dimensional space. When an unknown discrete data is received, it analyzes the closest k number of instances saved (nearest neighbors) and returns the most common class as the prediction and for real-valued data it returns the mean of k nearest neighbors. In the distance-weighted nearest neighbor algorithm, it weights the contribution of each of the k neighbors according to their distance using the following query giving greater weight to the closest neighbors.

Usually KNN is robust to noisy data since it is averaging the k-nearest neighbors. The k-nearest-neighbors algorithm is a classification algorithm, and it is supervised: it takes a bunch of labeled points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point (those are its nearest neighbors), and has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point (the “k” is the number of neighbors it checks). Makes predictions about the validation set using the entire training set. KNN makes a prediction about a new instance by

searching through the entire set to find the k “closest” instances. “Closeness” is determined using a proximity measurement (Euclidean) across all features.

Naive Bayes algorithm:

The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach you would come up with if you wanted to model a predictive modeling problem probabilistically. Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method. The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class. To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability.

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location. Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

4.5.6 GUI based prediction of crop yield and yield cost (Module-06)

Tkinter is a python library for developing GUI (Graphical User Interfaces). We use the tkinter library for creating an application of UI (User Interface), to create windows and all other graphical user interface and Tkinter will come with Python as

a standard package, it can be used for security purpose of each users or accountants. There will be two kinds of pages like registration user purpose and login entry purpose of user.

4.6 Parameter calculations

4.6.1 Accuracy calculation

False Positives (FP): A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN): A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

True Positives (TP): A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN): A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

It achieved precision, recall, true positive rate (TPR), and false positive rate (FPR) for each classification techniques as it is shown in the above tables and also achieved different interesting confusion matrix for each classification techniques and we can see the classification performance of each classifiers by the help of confusion matrix. We use a confusion matrix to compute the accuracy rate of each severity class. For each class, it demonstrates how instances from that class receive the various classifications. Here in the next table we have shown instances that are correctly classified and incorrectly classified in accordance with overall accuracy of each classification techniques. All classifiers perform similarly well with respect to the number of correctly classified instances.

Comparing Algorithm with prediction in the form of best accuracy result:

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning

problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution .

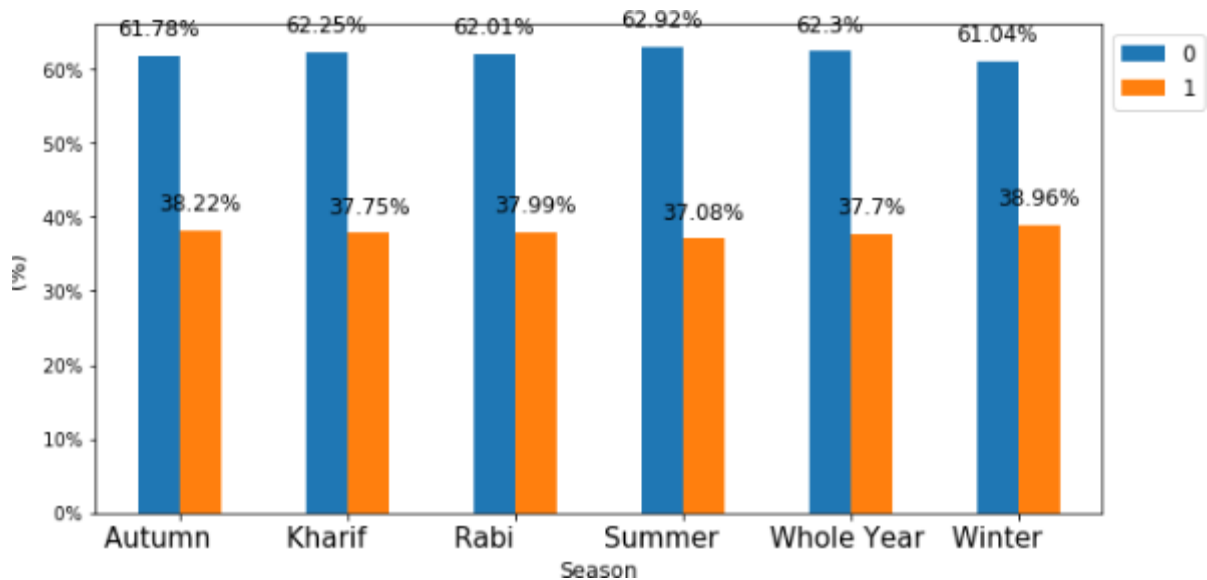


Fig 4.6 Crop yield prediction by each season

In the next section you will discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test

harness.

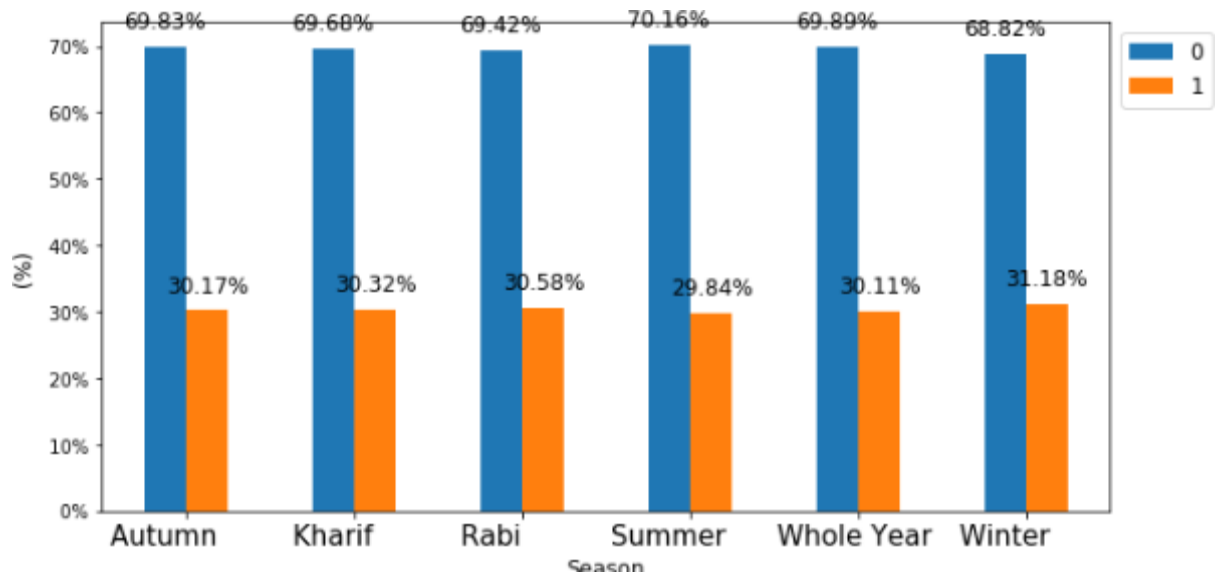


Fig 4.7 Crop yield cost production by each season

- Now, the dimensions of new features in a numpy array called 'n' and it want to predict the species of this features and to do using the predict method which takes this array as input and spits out predicted target value as output.
- So, the predicted target value comes out to be 0. Finally to find the test score which is the ratio of no. of predictions found correct and total predictions made and finding accuracy score method which basically compares the actual values of the test set with the predicted values.

Sensitivity

Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive). Sensitivity is also termed as Recall. This implies that there will be another proportion of actual positive cases, which would get predicted incorrectly as negative (and, thus, could also be termed as the false negative). This can also be represented in the form of a false negative rate. The sum of sensitivity and false negative rate would be 1. Let's try and understand this with the model used for predicting whether a person is suffering from the disease. Sensitivity is a measure of the proportion of people suffering from the disease who

got predicted correctly as the ones suffering from the disease. In other words, the person who is unhealthy actually got predicted as unhealthy.

Mathematically, sensitivity can be calculated as the following:

$$\text{Sensitivity} = (\text{True Positive}) / (\text{True Positive} + \text{False Negative})$$

The following is the details in relation to True Positive and False Negative used in the above equation.

- True Positive = Persons predicted as suffering from the disease (or unhealthy) are actually suffering from the disease (unhealthy); In other words, the true positive represents the number of persons who are unhealthy and are predicted as unhealthy.
- False Negative = Persons who are actually suffering from the disease (or unhealthy) are actually predicted to be not suffering from the disease (healthy). In other words, the false negative represents the number of persons who are unhealthy and got predicted as healthy. Ideally, we would seek the model to have low false negatives as it might prove to be life-threatening or business threatening.

The higher value of sensitivity would mean higher value of true positive and lower value of false negative. The lower value of sensitivity would mean lower value of true positive and higher value of false negative. For healthcare and financial domain, models with high sensitivity will be desired.

Specificity

Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative). This implies that there will be another proportion of actual negative, which got predicted as positive and could be termed as false positives. This proportion could also be called a false positive rate. The sum of specificity and false positive rate would always be 1. Let's try and understand this with the model used for predicting whether a person is suffering from the disease. Specificity is a measure of the proportion of people not suffering from the disease who got predicted correctly as the ones who are not suffering

from the disease. In other words, the person who is healthy actually got predicted as healthy is specificity.

Mathematically, specificity can be calculated as the following:

$$\text{Specificity} = (\text{True Negative}) / (\text{True Negative} + \text{False Positive})$$

The following is the details in relation to True Negative and False Positive used in the above equation.

- True Negative = Persons predicted as not suffering from the disease (or healthy) are actually found to be not suffering from the disease (healthy); In other words, the true negative represents the number of persons who are healthy and are predicted as healthy.
- False Positive = Persons predicted as suffering from the disease (or unhealthy) are actually found to be not suffering from the disease (healthy). In other words, the false positive represents the number of persons who are healthy and got predicted as unhealthy.

The higher value of specificity would mean higher value of true negative and lower false positive rate. The lower value of specificity would mean lower value of true negative and higher value of false positive.

Prediction result by accuracy

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

$$\text{True Positive Rate(TPR)} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{False Positive rate(FPR)} = \text{FP} / (\text{FP} + \text{TN})$$

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-

defaulters.

Accuracy calculation:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Accuracy is the most intuitive performance measure and it is simply a ratio

of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

Precision: The proportion of positive predictions that are actually correct. (When the model predicts default: how often is correct?)

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict)

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

General Formula

$$\text{F-Measure} = 2\text{TP} / (2\text{TP} + \text{FP} + \text{FN})$$

F1-Score Formula

$$\text{F1 Score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

CHAPTER 5 RESULTS AND DISCUSSION

The data obtain from online resources and then we save it as csv format and process it then after we train some data and test it so that it will learn itself and we can predict the output using this data by applying the ml algorithms and create a graph to represent it.

DATASET:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	State_Name	District_Name	Crop_Year	Season	Crop	Area	rainfall	Average H	Mean Tem	Cost of Cu	Cost of Pr	Yield (Quin	cost of production per yelc		
2	Tamil Nadu	CUDDALORE	2002	Kharif	Jowar	199	0.006214	58	58	15371.45	1595.56	32.42	51728.06		
3	Tamil Nadu	CUDDALORE	2003	Kharif	Jowar	1784	0.128503	59	64	17705.93	1279.6	12.94	16558.02		
4	Tamil Nadu	CUDDALORE	2004	Kharif	Jowar	1327	0.089727	34	75	22560.3	1279.6	56	71657.6		
5	Tamil Nadu	CUDDALORE	2005	Kharif	Jowar	891	0.01813	18	90	45239.51	2277.68	10.29	23437.33		
6	Tamil Nadu	CUDDALORE	2006	Whole Year	Jowar	661	0	23	93	15371.45	683.58	39.83	27226.99		
7	Tamil Nadu	CUDDALORE	2007	Whole Year	Jowar	1085	0.019572	27	98	33046.12	2068.67	13.7	28340.78		
8	Tamil Nadu	CUDDALORE	2008	Whole Year	Jowar	878	0	59	84	19083.55	2509.99	6.42	16114.14		
9	Tamil Nadu	CUDDALORE	2009	Whole Year	Jowar	798	0.025685	72	86	85801.95	1279.6	32.42	41484.63		
10	Tamil Nadu	CUDDALORE	2011	Kharif	Jowar	30	0	65	84	29876.36	1559.04	6.83	10648.24		
11	Tamil Nadu	CUDDALORE	2011	Rabi	Jowar	1	0	63	86	33046.12	581.69	11.98	6968.646		

Fig 5.1 Dataset

The above Fig 5.1 represent the dataset which contains data like state name, District name, Crop year, Season, Crop, Area, Rainfall, Average Humidity, Mean Temperature, cost of production per yield.

Data it will be like 0.7 as training data and 0.3 as testing data where by using this data it will learn to do as for the new data and predict it.

CONFUSION MATRIX RESULT:

```

Classification report of Random Forest Results:

              precision    recall  f1-score   support

     0               1.00      1.00      1.00       1776
     1               1.00      1.00      1.00       1087

 accuracy               1.00
 macro avg              1.00
 weighted avg           1.00

Accuracy result of Random Forest is: 100.0

Confusion Matrix result of Random Forest is:
[[1776  0]
 [  0 1087]]

Sensitivity : 1.0
Specificity : 1.0

```

Fig 5.2 CONFUSION MATRIX RESULT

The above Fig 5.2 represent the result of the confusion matrix like it is the output for every algorithm you used in your machine learning where it gives results of accuracy, specificity, precision, recall and averages for every factor there will be a specified formula by using the formula all the other algorithms also can predicted.

Parameters	LR	DT	RF	SVC	KNN	NB
Precision	0.96	1	1	0.62	0.91	0.95
Recall	0.96	1	1	1	0.92	1
F1-Score	0.96	1	1	0.77	0.92	0.97
Sensitivity	0.95	1	1	1	0.92	1
Specificity	0.92	1	1	0	0.85	0.90
Accuracy (%)	94.73	100	100	62.06	89.59	96.50

Fig 5.3 accuracy results of crop yield

The above figures describe the comparison of accuracy results of crop yield prediction.

Parameters	LR	DT	RF	SVC	KNN	NB
Precision	0.70	1	1	0.69	0.92	0.87
Recall	0.88	1	1	1	0.90	0.95
F1-Score	0.78	1	1	0.82	0.91	0.91
Sensitivity	0.87	1	1	1	0.90	0.94
Specificity	0.13	1	1	0	0.82	0.67
Accuracy (%)	65.17	100	100	69.36	88.08	86.48

Fig 5.4 accuracy results of crop yield cost

The above figures describe the comparison of accuracy results of crop yield cost prediction.

To find out the best algorithm which gives more accuracy we have compared the six machine algorithms

From the analysis after the comparison of machine learning algorithms we got decision

Tree and Randomn forest gave more accuracy compared to other. We use Random forest algorithm for prediction as it is collection of decision trees.

Prediction of Yield and Yield cost by Crop using Machine Learning

Tamilnadu Agricultural Department

Farmer Name: kranthi

Village Name: dindigul

District Name: DINDIGUL

Season Details: Whole Year

Temperature Level: TA60

Humidity Level: HA60

Crop Details: Tapioca

Rainfall Value: 60

RandomForest Cost Yield is two tons and Cost Production is fourty thousand

Check the result

Fig 5.5 Prediction output 1

Prediction of Yield and Yield cost by Crop using Machine Learning

Tamilnadu Agricultural Department

Farmer Name: bala

Village Name: navalur

District Name: KANCHIPURAM

Season Details: Autumn

Temperature Level: TA50

Humidity Level: HA25

Crop Details: Coconut

Rainfall Value: 20

RandomForest Cost Yield is one ton and Cost Production is twenty thousand

Check the result

Fig 5.6 Prediction output 2

In the above images we have given inputs Farmer name, village name, district and season details. We have mentioned temperatu0re level, humidity level, crop details which we want to grow and rainfall.

After using machine learning it gives how much yield is grown and cost production (i.e. how much money farmer will earn from crop).

CHAPTER 6 CONCLUSION AND FUTURE WORK

CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. Finally we predict the crop using machine learning algorithm with different results. This brings some of the following insights about crop prediction. As maximum types of crops will be covered under this system, farmer may get to know about the crop which may never have been cultivated and lists out all possible crops, it helps the farmer in decision making of which crop to cultivate. Also, this system takes into consideration the past production of data which will help the farmer get insight into the demand and the cost of various crops in market.

FUTURE WORK

- Remaining SMLT algorithms will be involve to finding the best accuracy with applying to predict the crop yield and cost.
- Agricultural department wants to automate the detecting the yield crops from eligibility process (real time).
- To automate this process by show the prediction result in web application or desktop application.
- To optimize the work to implement in Artificial Intelligence environment.

REFERENCES

- [1] Bendre, M. R., Thool, R.C., Thool, V. R., “Big Data in Precision Agriculture : Weather Forecasting for Future Agriculture”, 1 st International Conference on Next Generation Computing Technologies, pp.744-750, 2015.
- [2] Grajales, D.F.P., Mosquera, G.J.A, Mejia, F., Piedrahita, L.C., Basurto, C., “Crop-Planning, Making Smarter Agriculture With Climate Data”,Fourth International Conference on Agro-Geoinformatics, pp.240-244, 2015.
- [3] Hemageetha, N., “A survey on application of data mining techniques to analyze the soil for agricultural purpose”, 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp.3112-3117, 2016.
- [4] Mayank Champaneri, Chaitanya Chandvidkar , Darpan Chachpara, Mansing Rathod, “Crop yield prediction using machine learning” International Journal of Science and Research ,April 2020.
- [5] Pavan Patil, Virendra Panpatil, Prof. Shrikant Kokate, “Crop Prediction System using Machine Learning Algorithms”, International Research Journal of Engineering and Technology, Feb 2020.
- [6] Raval Agrawal, H., Agrawal, P., “Review on Data Mining Tools”, International Journal of Innovative Science, Engineering & Technology, Vol. 1, Issue 2, pp.52- 56, 2014.
- [7] Sabri Arik, Tingwen Huang, Weng Kin Lai, Qingshan Liu , “Soil Property Prediction: An Extreme Learning Machine Approach” Springer, vol. 3, Issue 4,666-680,2015.
- [8] Shivnath Ghosh,Santanu Koley, “Machine Learning for Soil Fertility and Plant Nutrient Management using Back Propagation Neural Networks” IJRITCC, vol. 2, Issue 2,292-297,2014.
- [9] Tan, L., “Cloud-based Decision Support and Automation for Precision Agriculture in Orchards”, ScienceDirect, pp. 330-335, IFAC–Papers OnLine 49-16, pp.330-335, 2016.
- [10] Zhihao Hong,Z. Kalbarczyk,R. K. Iyer, “A Data-Driven Approach to Soil

Moisture Collection and Prediction” IEEE Xplore, vol. 2, Issue 2, 292-297, 2016.

APPENDIX

A. SOURCE CODE

```
#import libraries for access and functional purpose
import pandas as p
import numpy as n
import matplotlib.pyplot as plt
import seaborn as s

#read the given dataset df
= p.read_csv("df.csv")
listcrops = p.Categorical(df['Crop'])
listcrops
df['Crop'].value_counts()
df['Crop'].nunique()

#To describe the dataframe
df.describe()

#Checking datatype and information about dataset
df.info()

df[df.dtypes[df.dtypes == 'float64'].index].describe()
p.Categorical(df['State_Name']).describe()
p.Categorical(df['District_Name']).describe()

#find sum of duplicate data
sum(df.duplicated())

#Correlation
df.corr()

#Checking minimum or maximum yields (100kg/2.47 acre)
print("Minimum yield of crops is (100kg/2.47 acre):", df["Yield
(Quintal/ Hectare) "].min())

print("Maximun yield of crops is (100kg/2.47 acre):", df["Yield
(Quintal/ Hectare) "].max())

#Checking minimum or maximum cost production for c2 scheme
(per 2.47 acre)
```

```
print("Minimum cost production for c2 scheme(per 2.47 acre):",  
df["Cost of Production (`/Quintal) C2"].min())  
  
print("Maximun cost production for c2 scheme(per 2.47 acre):",  
df["Cost of Production (`/Quintal) C2"].max())
```

```

#Rename the data
df.rename(columns={'Cost of Cultivation ( ` /Hectare) C2':'CC'},

inplace=True)

df.rename(columns={'Cost of Production ( ` /Quintal) C2':'CP'},
inplace=True)

df.rename(columns={'Yield (Quintal/ Hectare) ':'Y'}, inplace=True)

#show the dataframe
df.head()

p.crosstab(df.State_Name,df.Crop)

df.rename(columns={'Mean Temp':'T'}, inplace=True)
df.rename(columns={'Average Humidity':'H'}, inplace=True)
df.rename(columns={'rainfall':'R'}, inplace=True)

#Rename the data
df.rename(columns={'Cost of Cultivation ( ` /Hectare) C2':'CC'},
inplace=True)

df.rename(columns={'Cost of Production ( ` /Quintal) C2':'CP'},
inplace=True)

df.rename(columns={'Yield (Quintal/ Hectare) ':'Y'}, inplace=True)

df.rename(columns={'cost of production per yield':'CPPY'},
inplace=True)

from sklearn.preprocessing import LabelEncoder

var_mod = ['State_Name', 'District_Name', 'Crop_Year', 'Season',
'Crop', 'Area',

'R', 'H', 'T', 'CC', 'CP', 'Y', 'CPPY']

le = LabelEncoder()

for i in var_mod:

    df[i] = le.fit_transform(df[i]).astype(str)

#preprocessing, split test and dataset, split response variable X
= df.drop(labels='CPPY', axis=1)

#Response variable y
= df.loc[:, 'CPPY']

#We'll use a test size of 30%. We also stratify the split on the
response variable, which is very important to do because there are so

```

few fraudulent transactions.

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,  
random_state=1, stratify=y)
```

```

print("Number of training dataset: ", len(X_train))

print("Number of testing dataset: ", len(X_test)) print("Total
number of dataset: ", len(X_train)+len(X_test))
count_classes = p.value_counts(df['Crop'], sort = True).sort_index()
count_classes.plot(kind = 'bar', figsize=(20,15))
plt.title("Crop details")
plt.xlabel("Catogeries")
plt.ylabel("Strength values")
no=sum(df['CPPYPr']==0)
yes=sum(df['CPPYPr']==1)
colors=['orange','black']
locations=[1,2]
heights=[no,yes]
labels=['Unexpected Cost Production','Expected Cost Production']
plt.bar(locations,heights,color=colors,tick_label=labels,alpha=0.7)
plt.xlabel('Yield of Crost Production')
plt.ylabel('No. of each crop')
plt.title('Prediction results expecting from farmer by yield of crost
production amount')
no=sum(df['YPr']==0)
yes=sum(df['YPr']==1)
colors=['orange','black']
locations=[1,2]
heights=[no,yes]
labels=['No Yield','Yield']
plt.bar(locations,heights,color=colors,tick_label=labels,alpha=0.7)
plt.xlabel('Yield of Crop')
plt.ylabel('No. of each crop')
plt.title('Prediction results expecting from farmer by yield of crop')
df['CP'].hist(figsize=(7,6), color='orange', alpha=0.7) plt.xlabel('Cost

```

Production (per 100kg)')

plt.ylabel('No of counts')

plt.title('Cost Production per 100kg by counts')

df['CPPY'].hist(figsize=(7,6), color='black', alpha=0.7)

plt.xlabel('Cost Production of crop')

```

plt.ylabel('No of counts')
plt.title('Cost Production of crop by counts') #
Heatmap plot diagram
fig, ax = plt.subplots(figsize=(15,10)) s.heatmap(df.corr(),
ax=ax, annot=True) df.boxplot(column="CP",
by="Season", figsize=(15,10)) #Propagation by variable
def PropByVar(df, variable):
    dataframe_pie = df[variable].value_counts()
    ax = dataframe_pie.plot.pie(figsize=(10,10), autopct='%1.2f%%',
    fontsize = 12)
    ax.set_title(variable + ' (in Percentage)', fontsize = 15)
    return n.round(dataframe_pie/df.shape[0]*100,2)
PropByVar(df, 'Crop')
count_classes = p.value_counts(df['State_Name'], sort =
True).sort_index()
count_classes.plot(kind = 'bar', figsize=(20,15))
plt.title("Dataset by each states")
plt.xlabel("Number of States")
plt.ylabel("Given data counts")
#Density Plots
plt = df.plot(kind= 'density', subplots=True, layout=(4,3),
sharex=False,
sharey=False,fontsize=12, figsize=(15,10))
def y_No_y_bar_plot(df, bygroup):
    dataframe_by_Group = p.crosstab(df[bygroup],
columns=df["YPr"], normalize = 'index')
    dataframe_by_Group = n.round((dataframe_by_Group * 100),
decimals=2)
    ax = dataframe_by_Group.plot.bar(figsize=(10,5));
    vals = ax.get_yticks()
    ax.set_yticklabels(['{:.3.0f}%'.format(x) for x in vals]);
    ax.set_xticklabels(dataframe_by_Group.index,rotation = 0,

```

```
fontsize = 15);  
    ax.set_title('Crop Yield Prediction Vs No Crop Yield Prediction (%)  
(by ' + dataframe_by_Group.index.name + ')\n', fontsize = 15)
```

```

ax.set_xlabel(dataframe_by_Group.index.name, fontsize = 12)
ax.set_ylabel('%', fontsize = 12)
ax.legend(loc = 'upper left',bbox_to_anchor=(1.0,1.0), fontsize=
12)

```

```

rects = ax.patches #

```

Add Data Labels for

rect in rects:

```

    height      =      rect.get_height()
    ax.text(rect.get_x() + rect.get_width()/2,
            height + 2,
            str(height)+'%',
            ha='center',
            va='bottom',
            fontsize = 12)

```

```

    return dataframe_by_Group

```

```

y_No_y_bar_plot(df, 'Season')

```

#According to the cross-validated MCC scores, the random forest is the best-performing model, so now let's evaluate its performance on the test set.

```

from sklearn.metrics import confusion_matrix, classification_report,
matthews_corrcoef, cohen_kappa_score, accuracy_score,
average_precision_score, roc_auc_score

```

```

X = df.drop(labels='YPr', axis=1)

```

#Response variable

```

y = df.loc[:, 'YPr']

```

#We'll use a test size of 30%. We also stratify the split on the response variable, which is very important to do because there are so few fraudulent transactions.

```

from sklearn.model_selection import train_test_split

```

```

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=1, stratify=y)

```

```

from sklearn.linear_model import LogisticRegression

```

```

logR= LogisticRegression()

```

```
logR.fit(X_train,y_train) predictR =  
logR.predict(X_test)  
print("")
```

```

print('Classification report of Logistic Regression Results:')
print("")
print(classification_report(y_test,predictR)) x
= (accuracy_score(y_test,predictR)*100)
print('Accuracy result of Logistic Regression is:', x)
print("")
cm1=confusion_matrix(y_test,predictR)
print('Confusion Matrix result of Logistic Regression is:\n',cm1)
print("")
sensitivity1 = cm1[0,0]/(cm1[0,0]+cm1[0,1])
print('Sensitivity : ', sensitivity1 )
print("")
specificity1 = cm1[1,1]/(cm1[1,0]+cm1[1,1])
print('Specificity : ', specificity1)
print("")
from sklearn.tree import DecisionTreeClassifier
dtree = DecisionTreeClassifier() dtree.fit(X_train,
y_train)

predictDT = dtree.predict(X_test)
print("")
print('Classification report of Decision Tree Classifier Results:')
print("")
print(classification_report(y_test,predictDT)) x
= (accuracy_score(y_test,predictDT)*100)
print('Accuracy result of Decision Tree Classifier is', x)
print("")
cm2=confusion_matrix(y_test,predictDT)
print('Confusion Matrix result of Decision Tree Classifier is:\n',
confusion_matrix(y_test,predictDT))
print("")

```

```
sensitivity1 = cm2[0,0]/(cm2[0,0]+cm2[0,1])  
print('Sensitivity : ', sensitivity1 )  
print("")
```

```

specificity1 = cm2[1,1]/(cm2[1,0]+cm2[1,1])
print('Specificity : ', specificity1)
from sklearn.ensemble import RandomForestClassifier rfc
= RandomForestClassifier()
rfc.fit(X_train,y_train) predictR =
rfc.predict(X_test) print("")
print('Classification report of Random Forest Results:')
print("")
print(classification_report(y_test,predictR)) x
= (accuracy_score(y_test,predictR)*100)
print('Accuracy result of Random Forest is:', x)
print("") cm1=confusion_matrix(y_test,predictR)
print('Confusion Matrix result of Random Forest is:\n',cm1)
print("")
sensitivity1 = cm1[0,0]/(cm1[0,0]+cm1[0,1])
print('Sensitivity : ', sensitivity1 )
print("")
specificity1 = cm1[1,1]/(cm1[1,0]+cm1[1,1])
print('Specificity : ', specificity1)
print("")
from sklearn.svm import SVC s
= SVC()
s.fit(X_train,y_train) predicts =
s.predict(X_test) print("")
print('Classification report of Support Vector Machines Results:')
print("")

print(classification_report(y_test,predicts)) x
= (accuracy_score(y_test,predicts)*100)
print('Accuracy result of Support Vector Machines is:', x) print("")

```

```

cm2=confusion_matrix(y_test,predicts)
print('Confusion Matrix result of Support Vector Machines is:\n',cm2) print("")
sensitivity1 = cm2[0,0]/(cm2[0,0]+cm2[0,1])
print('Sensitivity : ', sensitivity1 )
print("")
specificity1 = cm2[1,1]/(cm2[1,0]+cm2[1,1])
print('Specificity : ', specificity1)
print("") root
= Tk()
root.title("Indian Crop Prediction")
#root.configure(background='black')
canvas = Canvas(root,width=1880,height=1253)
canvas.pack()
photo = PhotoImage(file ='im3.png')
canvas.create_image(0,0,image=photo,anchor=NW)
# gui_stuff_____
-----

# entry variables State
= StringVar()
State.set(None) Season
= StringVar()
Season.set(None)
TI = StringVar()
TI.set(None)
HI = StringVar()
HI.set(None)
YI = StringVar()
YI.set(None)
Yc = StringVar()
Yc.set(None) fn=
StringVar()

```

vn= StringVar()

dn=StringVar()

```

rl= IntVar() #
Heading
w2 = Label(canvas, justify=LEFT, text="Crop Prediction using
Machine Learning", fg="white", bg="red")
w2.config(font=("Elephant", 20))
w2.grid(row=1, column=0, columnspan=2, padx=100)
w2 = Label(canvas, justify=LEFT, text="Indian Agricultural
Department", fg="green")
w2.config(font=("Times Roman", 15))
w2.grid(row=2, column=0, columnspan=2, padx=100) #
labels
FNLb = Label(canvas, text="Farmer Name",fg="yellow", bg="black")
FNLb.grid(row=5, column=0, pady=15, sticky=W)
VNLb = Label(canvas, text="Village Name",fg="yellow", bg="black")
VNLb.grid(row=6, column=0, pady=15, sticky=W)
DNLb = Label(canvas, text="District Name",fg="yellow", bg="black")
DNLb.grid(row=7, column=0, pady=15, sticky=W)
# labels
stLb = Label(canvas, text="State Name",fg="yellow", bg="black")
stLb.grid(row=8, column=0, pady=15, sticky=W)
seLb = Label(canvas, text="Season Details",fg="yellow",
bg="black")
seLb.grid(row=9, column=0, pady=15, sticky=W)
tlLb = Label(canvas, text="Temperature Level",fg="yellow",
bg="black")
tlLb.grid(row=5, column=1, pady=15, sticky=W)
hlLb = Label(canvas, text="Humidity Level",fg="yellow", bg="black")
hlLb.grid(row=6, column=1, pady=15, sticky=W)
ylLb = Label(canvas, text="Yield level (per Hectare)",fg="yellow",
bg="black")
ylLb.grid(row=7, column=1, pady=15, sticky=W)
ycLb = Label(canvas, text="Yield cost amount (per
Hectare)",fg="yellow", bg="black")

```

```
ycLb.grid(row=8, column=1, pady=15, sticky=W)
destreeLb = Label(canvas, text="DecisionTree", fg="white",
bg="red")
```

```

destreeLb.grid(row=14, column=0, pady=10, sticky=W)
ranfLb = Label(canvas, text="RandomForest", fg="white", bg="red")
ranfLb.grid(row=15, column=0, pady=10, sticky=W)
lrLb = Label(canvas, text="LogisticRegression", fg="white",
bg="red")
lrLb.grid(row=16, column=0, pady=10, sticky=W)
svLb = Label(canvas, text="SVC", fg="white", bg="red")
svLb.grid(row=17, column=0, pady=10, sticky=W)

# entries
OPTIONS1 = sorted(I1)
OPTIONS2 = sorted(I2)
OPTIONS3 = sorted(I3)
OPTIONS4 = sorted(I4)
OPTIONS5 = sorted(I5)
OPTIONS6 = sorted(I6)
FNEn = Entry(canvas, textvariable=fn)
FNEn.grid(row=5, column=0)
VNEEn = Entry(canvas, textvariable=vn)
VNEEn.grid(row=6, column=0)
DNEEn = Entry(canvas, textvariable=dn)
DNEEn.grid(row=7, column=0)
stEn = OptionMenu(canvas, State,*OPTIONS3)
stEn.grid(row=8, column=0)
seEn = OptionMenu(canvas, Season,*OPTIONS4)
seEn.grid(row=9, column=0)
tlEn = OptionMenu(canvas, TI,*OPTIONS5)
tlEn.grid(row=5, column=1)
hlEn = OptionMenu(canvas, HI,*OPTIONS6)
hlEn.grid(row=6, column=1)
ylEn = OptionMenu(canvas, YI,*OPTIONS1)
ylEn.grid(row=7, column=1)

```

```
ycEn = OptionMenu(canvas, Yc,*OPTIONS2)
ycEn.grid(row=8, column=1)
dst = Button(canvas, text="DecisionTree Algorithm",
```

```
command=DecisionTree,bg="cyan",fg="green")
dst.grid(row=14, column=1,padx=10)
rnf = Button(canvas, text="Randomforest Algorithm",
command=randomforest,bg="cyan",fg="green")
rnf.grid(row=15, column=1,padx=10)
lr = Button(canvas, text="LogisticRegression Algorithm",
command=LogisticRegression,bg="cyan",fg="green")
lr.grid(row=16, column=1,padx=10)
sv = Button(canvas, text="SVC Algorithm",
command=SVC,bg="cyan",fg="green")
sv.grid(row=17, column=1,padx=10)
```

B. SCREENSHOTS

The screenshot shows a Tkinter window titled "tk" with a black background. At the top, there is a red banner with the text "Prediction of Yield and Yield cost by Crop using Machine Learning" in white. Below this is a green banner with the text "Tamilnadu Agricultural Department" in white. The interface contains several input fields and dropdown menus arranged in two columns. On the left, there are labels for "Farmer Name", "Village Name", "District Name", and "Season Details", each followed by an empty input field or a dropdown menu showing "None". On the right, there are labels for "Temperature Level", "Humidity Level", "Crop Details", and "Rainfall Value", each followed by a dropdown menu showing "None" or an empty input field. At the bottom left, there is a red label "RandomForest" followed by a long orange bar. At the bottom right, there is a cyan button labeled "Check the result".

Prediction of Yield and Yield cost by Crop using Machine Learning

Tamilnadu Agricultural Department

Farmer Name

Village Name

District Name

Season Details

Temperature Level

Humidity Level

Crop Details

Rainfall Value

RandomForest

Check the result

The screenshot shows the same Tkinter window as the first one, but with the input fields filled with data. The "Farmer Name" field contains "kranthi", the "Village Name" field contains "dindigul", the "District Name" dropdown shows "DINDIGUL", and the "Season Details" dropdown shows "Whole Year". The "Temperature Level" dropdown shows "TA60", the "Humidity Level" dropdown shows "HA60", and the "Crop Details" dropdown shows "Tapioca". The "Rainfall Value" field contains "60". The orange bar at the bottom left now displays the text "Cost Yield is two tons and Cost Production is fourty thousand". The cyan button "Check the result" remains at the bottom right.

Prediction of Yield and Yield cost by Crop using Machine Learning

Tamilnadu Agricultural Department

Farmer Name

Village Name

District Name

Season Details

Temperature Level

Humidity Level

Crop Details

Rainfall Value

RandomForest Cost Yield is two tons and Cost Production is fourty thousand

Check the result

C. JOURNAL PAPER

PREDICTION OF CROP YIELD AND PRODUCTION VALUE BY FINDING BEST ACCURACY USING MACHINE LEARNING APPROACH

**KOVVURI BALAGOPAL REDDY [1], KARRI KRANTHI VARDHAN REDDY [2] ,
ASHA P[3]**

**[1][2] UG Student, Dept. of CSE, Sathyabama Institute of Science and
Technology, Chennai, India**

**[3]Associate professor, Dept. of CSE, Sathyabama Institute of Science and
Technology, Chennai, India**

ABSTRACT: Among around the globe, agribusiness has the huge commitment in regards to improving the financial responsibility of the country. In any case, regardless the most cultivating fields are youthful in light of the shortfall of association of organic framework control developments. Due to these issues, the yield creation isn't improved which impacts the cultivating economy. From this time forward an improvement of cultivating benefit is redesigned reliant on the plant yield assumption. To thwart this issue, Agricultural zones need to anticipate the yield from given dataset using AI methodology. The examination of dataset by coordinated AI technique(SMLT) to get a couple of information takes after, factor unmistakable confirmation, uni-variate assessment, bi-variate and multi- variate examination, missing worth prescriptions, etc An overall report between AI counts had been finished to sort out which estimation is the most careful in predicting the best gather. The results show that the reasonability of the proposed AI computation technique can match best precision with entropy figuring, exactness, Recall, F1 Score, Sensitivity, Specificity.

Keywords: dataset, Machine learning-Classification method

crop yield [1]. Agribusiness is

I INTRODUCTION

Crop creation might be a convoluted improvement that is impacted by soil and natural condition input boundaries. Agriculture is the foundation of each economy. Agricultural scientists in elective nations have shown that attempts of crop yield augmentation through favorable to pesticide state approaches have LED to unsafely high substance use. These examinations have announced a connection between's substance use and

partner exchange area that is profiting intensely from the occasion of finder innovation, information science, and AI (ML) methods inside the most recent years. These improvements get back to fulfill ecological and populace pressures round-looked by our general public, any place reports show a necessity for powerful worldwide agribusiness yield increment to create nourishment for a developing populace on a more sweltering planet. The vast majority of the work tired the area of yield predicting through cubic centimeter

utilizes some sort of far-off detecting information over the homestead. Agribusiness looks to broaden and enhance the crop yield and hence the nature of the yields to support human existence. Nonetheless, inside the current time, people will in general require a great deal of like a shot appreciated positions. There are less, and less people worried in crop development. moreover, the consistent increment of human populace makes the development of the yields at the legitimate time and opportune spot even a great deal of crucial, in light of the fact that the environment is dynamic and accordingly the movements from conventional climate design are a ton of continuous than before make. The data hole between antiquated ways that of developing and new agrarian advancements might be survived if the PC code might be intended to show the intelligent effect of environment factors, especially the effect of greatest occasions (for example warmth, rainfalls and overabundance water) happening at totally extraordinary developing periods of crops. The temperature change without a doubt influences the local and world food creation, along these lines arranging PC code to show crop forecasts needs new strategy for temperature change examines, circumstances for temperature change transformation, and policymakers which will restrict the overwhelming impacts of climate on food give. The dirt sort will adjustment after some time due to climate and vermin, consequently crop the executives should deal with an extravagant amount of data, straightforwardly or by implication related with each other. It will along these lines by thinking about a worked on the real world, to allow a brisk appraisal of the effect of temperature change in agribusiness.

Farming ought to adjust to those environment changes, and it will do accordingly by creating models

which will in principle enhance the executives rehearses, augment the turns of the new yield to deal with the progressions of soil, novel rearing projects. By boosting the value of anticipating, the occasional environment changes might be discovered and recorded in an incredibly convenient way. Later on, by exploitation PC code upheld AI, one will conveniently evaluate the temperature change effect and check achievable circumstances that consolidate found out changes in climatic conditions and water dispersion. information {processing} is that the way toward dissecting the test information gathered over a sum and changed areas from totally various perspectives, separate patterns or examples {of information of information on info} and switch them into supportive data for clients. Clients will at that point also reason as well as sum up the connections found out from the gathered information, and normally anticipate what information to anticipate. Machinelearning procedures are a piece of information handling and information investigation and spotlight only on trademark connections or examples among monstrous datasets or huge relative data sets. The examples, affiliations, or connections among this information will extra be reawakened into data that is offered to the client as recorded examples and future patterns. This data given by AI will encourage ranchers with crop development by foreseeing probabilities of crop misfortunes

or stop misfortunes out and out.

For the most part, information mining is the way toward examining information from alternate points of view and summing up it into helpful data. Yield forecast benefits the ranchers in diminishing their misfortunes and to get best costs for

their harvests. Agriculture is most significant occupation drilled in our country. It is the biggest financial area and assumes a significant part in by and large improvement of the country. Determination of harvest is significant issue in horticultural arranging. Yield creation rate relies upon topography of an area, climate condition, soil type and gathering techniques. Various subsets of these impacting boundaries are utilized for various yields by various expectation models. AI strategy is utilized for government approaches. Harvest selector strategy could be appropriate for limit misfortunes of harvest when troublesome conditions happen, and this selector could be utilized to expand the yield rate when likely exists for ideal developing conditions. Independent of the capital put as far as soil, water and nature of seeds of the yield rate creation the harvest may bomb carrying shocking misfortunes to the rancher and his family.

In this pre-owned AI approach with created in harvest or plant yield expectation since horticulture has diverse information like soil information, crop information, and climate information. Plant development forecast is proposed for checking the plant yield viably through the AI methods. It is likewise material for the robotized cycle of cultivating is the start of another time in Bangladesh that will be reasonable for the ranchers who look for specialists to take recommendation about the suitable harvest on explicit area of their territory and don't have any desire to fail to remember any progression of the development all through the interaction. Albeit, the assessment from specialists is the most advantageous way, this application is intended to give exact arrangement in quickest way conceivable.

This present

expectation of harvest yield utilizing various calculations. AI strategies which are broadly utilized in expectation strategy are relapse tree, irregular timberland, convolution neural organization and K-closest calculation. There is consistently a huge danger factor for the ranchers to choose which specific yield ought to develop during season, on specific land asset. It relies upon various boundaries, for example, creation rate, cost and diverse

exploration's fundamental target is to bring cultivating measure a bit nearer to the advanced stage.

II RELATED WORK

Virendra Panpatil et al[1]

had accomplished gigantic work for Indian ranchers by making productive yield proposal framework. They created framework utilizing classifier models like KNN, DT, CNN. The proposed framework can be utilized to figure out best season of planting, development of plant and Plant reaping. They utilized distinctive classifier for accomplishing better exactness for instance. The best favorable position of framework that it can without much of a stretch versatile all things considered/be utilized to test on various yields.

Mayank et al[2] have presumed that this paper fabricate extemporized framework for crop yield utilizing administered AI calculations and with objective to give simple to utilize User Interface, increment the precision of crop yield forecast, investigate distinctive

climatic boundaries. The primary favourable position of proposed framework is precision rate is more than 75 percent on the whole the yields and areas chose in the examination.

Zhihao et al[3] two relapse administered AI strategies SVM, RVM to show viability in soil quality

forecast. a shrewd remote gadget for detecting soil dampness and meteorological information. **Sabri Arik et al[4]** includes a check for Soil Fertility and Plant Nutrient by utilizing back spread calculation. The outcomes are exact and empowers improvement in soil properties. It performs better when contrasted with conventional techniques. Be that as it may, framework is moderate wasteful and not steady. **Shivnath et al[5]** proposed about BPN to assess the test informational collection. Back Propagation Network utilizes a concealed layer which helps in better execution in foreseeing soil properties. BPN here, is utilized to build up a self-prepared capacity to foresee soil properties with boundaries. This gives more exactness and performs better compared to the customarily utilized strategies ,in any case, at times the framework turns out to be moderate and irregularity is found in the yield.

Raval et al[6] examine about the Knowledge Discovery Process and the rudiments of different Data Mining Techniques, for example, Association rules, Classification, Clustering, Prediction and Sequential Patterns. Agrawal et al examine about different Data Mining devices, for example, Dashboards, Text-Mining instruments. They give an outline about these apparatuses and the different situations where they can be conveyed. **Grajales et al[7]** have proposed a web application that uses open dataset like verifiable creation, land cover, nearby environment conditions and coordinates them to give simple admittance to the ranchers. The projected engineering basically centers around free devices for the advancement of the application. The client will choose area from map that subtleties are accessible at a single tick.

Bendre et al[8] gathers information about Global Information

System, GPS, VRT and RS are controlled utilizing Map Reduce calculation and direct relapse calculation to figure the climate information that can be utilized in accuracy agribusiness. The motivation behind this examination was to research the powerful model to improve the exactness of precipitation guaging.

III EXISTING SYSTEM

It present a harvest/weeds characterization approach dependent on a three-steps methodology. The initial step is a vigorous pixel-wise division (i.e., soil/plant) and picture patches containing plants are removed in the subsequent advance. The third step, a profound CNN for crop/weed order is utilized. The removed masses in the covered picture containing plants data are taken care of to a CNN classifier dependent on a tweaked model of VGG-16 abusing the capacity of profound CNN in article grouping and to lessen the constraints of CNNs in summing up when a restricted measure of information is accessible. The arrangement step would then be able to be specific to the sorts of plants required by the application situation. It assessed the total pipeline, including the primary foundation expulsion stage and the ensuing grouping stage. Trial results show that can accomplish great grouping results on testing information. Accuracy horticulture is acquiring expanding consideration in light of the conceivable decrease of

Agriculture sources of info (e.g., manures and pesticides) that can be gotten by utilizing innovative gear, including robots. To zero in on a Agriculture mechanical technology framework that tends to the weeding issue by methods for particular splashing or mechanical evacuation

of the recognized weeds. To depict a profound learning based technique to permit a robot to play out an exact weed/crop grouping utilizing a succession of two Convolutional Neural Networks (CNNs) applied to RGB pictures. The main organization, in view of encoder-decoder division engineering, plays out a pixel savvy, plant-type skeptic, division among vegetation and soil that empowers to remove a bunch of associated masses addressing plant cases.

IV PROPOSED SYSTEM

In this proposed work need load the information, check for tidiness, and afterward trim and clean your dataset for examination. Ensure that you record your means cautiously and legitimize your cleaning choices. This is the most energizing stage in Applying Machine Learning to any Dataset. It is otherwise called Algorithm choice for Predicting the best outcomes. Normally Data Scientists utilize various types of Machine Learning calculations to the enormous informational collections. Yet, at significant level every one of those various calculations can be ordered in two gatherings: regulated learning and solo learning. Managed learning: Supervised learning is a kind of framework wherein both information and wanted yield information are given. Info and yield information are marked for order to give a learning premise to future information handling. Administered learning issues can be additionally gathered into Regression and Classification issues

4.1 TRAIN THE DATASET

The principal line imports iris informational collection which is as of now predefined in sklearn module. Iris informational index is essentially a table which contains data about different assortments of iris blossoms. For instance, to import any calculation and train_test_split class from sklearn and numpy module to be used during this program. At that point we typify load_data() strategy in data_dataset variable. Further we partition the dataset into preparing information and test information utilizing train_test_split technique. The X prefix in factor means the element esteems and y prefix indicates target esteems. This technique partitions dataset into preparing and test information haphazardly in proportion of 67:33. At that point we exemplify any calculation. In the following line, we fit our preparation information into this calculation so PC can get prepared utilizing this information. Presently the preparation part is finished.

4.2 TESTING THE DATASET

Presently we have measurements of another bloom in a numpy cluster called 'n' and we need to anticipate the types of this blossom. We do this utilizing the

foresee strategy which accepts this exhibit as information and lets out anticipated objective incentive as yield. So the anticipated objective worth comes out to be 0. At last we discover the grade which was the proportion of no. of expectations discovered right and complete forecasts created. We do that utilizing the aggregate technique that fundamentally looks at the genuine estimations of the test set with the anticipated qualities.

V MODULES DESCRIPTION

5.1 VARIABLE IDENTIFICATION PROCESS/DATA VALIDATION PROCESS

Endorsement methodology in AI are used to get the goof speed of the Machine Learning (ML) model, which can be considered as close to the certifiable error speed of the dataset. If the data volume is adequately enormous to be illustrative of the general population, you may not need the endorsement strategies. In any case, in certified circumstances, to work with trial of data that may not be a veritable specialist of the quantity of occupants in given dataset. To finding the missing worth, duplicate worth and depiction of data type whether it is coast variable or entire number. The case of data accustomed provides a reasonable appraisal of a model fit on the readiness dataset while tuning model hyper limits. The appraisal ends up being more uneven as capacity on the endorsement dataset is combined into the model plan.

The endorsement set is used to survey a given model, anyway this

is for progressive evaluation. It as AI engineers uses this data to align the model hyper limits. Data grouping, data examination, and the route toward watching out for data substance, quality, and configuration can add up to a drawn-out arrangement for the afternoon. During the collaboration of data conspicuous verification, it helps with understanding your data and its properties; this data will help you with picking which figuring to use to create your model. For example, time course of action data can be poor somewhere around backslide figurings; portrayal estimations can be used to look at discrete data.

that are more instinctual and accomplices than extents of alliance or significance. Data portrayal and exploratory data examination are whole fields themselves and it will recommend a more significant bounce into some the books referred to at the end.

5.2 EXPLORATION INFORMATION INVESTIGATION OF PERCEPTION

Data discernment is a huge ability in applied estimations and AI. Estimations does to be certain focus in on quantitative portrayals and evaluations of data. Data insight surrenders a huge arrangement of contraptions for gaining an emotional understanding. This can be valuable when researching and getting familiar with a dataset and can help with recognizing plans, degenerate data, exemptions, and essentially more. With a little space data, data insights can be used to discuss and show key associations in plots and graphs

5.3 LOGISTIC REGRESSION

It is a real method for analyzing an instructive file where there are in any event one free factors that choose an effect. The effect is assessed with a bipartite variable (where there are only two possible effects). The aim of key fall away is to search out the most effective fitting model to portray the association between the bipartite attribute of premium (subordinate variable = respond or result variable) and a lot of free (pointer or enlightening) factors. Determined backslide is a Machine Learning portrayal computation that is used to envision the probability of a full scale ward variable. In essential backslide, the penniless variable is an equal variable that contains data coded as 1 (surely, accomplishment, etc) or 0 (no, failure, etc)

5.4 SUPPORT VECTOR MACHINES (SVM)

A classifier that masterminds the instructive assortment by setting an ideal hyper plane between data. I picked this classifier as it is amazingly versatile in the amount of different kernelling limits that can be applied and this model can yield a high consistency rate. Sponsorship Vector Machines are perhaps conceivably the most notable and examined AI computations. They were very standard around the time they were made during the 1990s and continue being the go-to strategy for a high- performing computation with little tuning.

5.5 K-NEAREST NEIGHBORS (KNN)

K-Nearest Neighbors is a regulated AI figuring which stores all

events identify with planning data centers in n-dimensional space. Right when a dark discrete data is gotten, it separates the closest k number of cases saved (nearest neighbors) and returns the most broadly perceived class as the assumption and for authentic regarded data it reestablishes the mean of k nearest neighbors. In the distance weighted nearest neighbor figuring, it stacks the responsibility of all of the k neighbors according to their distance using the going with request giving more critical burden to the closest neighbors. By and large KNN is solid to noisy data since it is averaging the k-nearest neighbors. The k-nearest neighbors computation is a plan count, and it is guided: it takes a ton of named centers and uses them to sort out some way to name distinctive core interests. To name another point, it looks at the stamped concentrates closest to that new point (those are its nearest neighbors), and has those neighbors vote, so whichever name most of the neighbors have is the name for the new point (the "k" is the amount of neighbors it checks). Makes assumptions regarding the endorsement set using the entire getting ready set. KNN makes an estimate about another case by means of glancing through the entire set to find the k "closest" events. "Closeness" is settled using a proximity assessment (Euclidean) across all features.

VI RESULT AND DISCUSSION

VII CONCLUSION

The logical connection started

Parameters	LR	DT	RF	SVC	KNN	NB
Precision	0.96	1	1	0.62	0.91	0.95
Recall	0.96	1	1	1	0.92	1
F1-Score	0.96	1	1	0.77	0.92	0.97
Sensitivity	0.95	1	1	1	0.92	1
Specificity	0.92	1	1	0	0.85	0.90
Accuracy (%)	94.73	100	100	62.06	89.59	96.50

**FIG 2 ANALYSIS OF PRECISION
OUTCOMES OF CROP YIELD
PREDICTION**

Parameters	LR	DT	RF	SVC	KNN	NB
Precision	0.70	1	1	0.69	0.92	0.87
Recall	0.88	1	1	1	0.90	0.95
F1-Score	0.78	1	1	0.82	0.91	0.91
Sensitivity	0.87	1	1	1	0.90	0.94
Specificity	0.13	1	1	0	0.82	0.67
Accuracy (%)	65.17	100	100	69.36	88.08	86.48

**FIG 3 ANALYSIS OF PRECISION
OUTCOMES OF CROP YIELD
PREDICTION**

**FIG 4 CROP PREDICTION USING
MACHINE LEARNING**

**FIG 5 PREDICTION OF YIELD
COST BY CROP USING MACHINE
LEARNING**

from data cleaning and planning, missing worth, exploratory examination finally model design and evaluation. Finally we anticipate the yield using AI computation with different results. This brings a bit of the going with encounters about yield gauge. As most prominent sorts of yields will be covered under this structure, farmer may turn out to be more familiar with about the collect which may never have been created and penetrates down each possible yield, it helps the farmer in unique of which respect create. In like manner, this system examines the past production of data which will help the farmer with getting information into the premium and the cost of various harvests in market.

Remaining SMLT counts will be incorporated to finding the best accuracy with applying to anticipate the gather yield and cost. Agricultural division needs to motorize the distinguishing the yield crops from capability measure (progressing). To robotize this communication by show the conjecture achieve web application or work zone application. To update the work to execute in Artificial Intelligence environment.

REFERENCES

- [1] Mayank Champaneri, Chaitanya Chandvidkar , Darpan Chachpara, Mansing Rathod, "Crop yield prediction using machine learning" International Journal of Science and Research ,April 2020.
- [2] Pavan Patil, Virendra Panpatil, Prof. Shrikant Kokate, "Crop Prediction System using

Machine Learning Algorithms", International Research Journal of Engineering and Technology, Feb 2020.

- [3] Zhihao Hong,Z. Kalbarczyk,R. K. Iyer, "A Data-Driven Approach to Soil Moisture Collection and Prediction" IEEE Xplore,vol. 2, Issue 2,292-

297,2016

- [4] Sabri Arik, Tingwen Huang, Weng Kin Lai, Qingshan Liu , "Soil Property Prediction: An Extreme Learning Machine Approach" Springer, vol. 3, Issue 4,666-680,2015.
- [5] Shivnath Ghosh,Santanu Koley, "Machine Learning for Soil Fertility and Plant Nutrient Management using Back Propagation Neural Networks" IJRITCC, vol. 2, Issue 2,292- 297,2014.
- [6] Raval Agrawal, H., Agrawal, P., "Review on Data Mining Tools", International Journal of Innovative Science, Engineering & Technology, Vol. 1, Issue 2, pp.52-56, 2014.
- [7] Grajales, D.F.P., Mosquera, G.J.A, Mejia, F., Piedrahita, L.C., Basurto, C., "Crop-Planning, Making Smarter Agriculture With Climate Data",Fourth International Conference on Agro-GeoInformatics, pp.240-244, 2015.
- [8] Bendre, M. R., Thool, R.C., Thool, V. R., "Big Data in Precision Agriculture : Weather Forecasting for Future Agriculture", 1 st International Conference on Next Generation Computing Technologies, pp.744- 750, 2015.
- [9] Tan, L., "Cloud-based Decision Support and Automation for Precision Agriculture in Orchards", ScienceDirect, pp. 330-335, IFAC–Papers OnLine 49-16, pp.330-335, 2016.
- [10] HemaGeetha, N., "A survey on application of data mining techniques to analyze the soil for agricultural purpose", 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp.3112-3117, 2016.
- [11] Rub, G., "Data Mining of Agricultural Yield Data:A Comparison of Regression Models", 9th Industrial Conference, Vol.5633, pp.24-37, 2009.
- [12] Sujatha, R., Isakki, P., "A study on crop yield forecasting using

classification techniques", International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE), pp.1-4, 2016.

[13] Kushwaha, A.K., SwetaBhattachrya, "Crop yield prediction using AgroAlgorithm in Hadoop", International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 5- No2, pp.271-274, 2015.

[14] Fathima, G.N., Geetha, R., "Agriculture Crop Pattern Using Data Mining Techniques", International Journal of Advanced Research in Computer Science and Engineering, Vol. 4, Issue 5, pp.781-786, 2014.

