A

Mini Project Report

On

CROP YIELD PREDICTION USING MACHINE LEARNING ALGORITHM

Submitted to the Faculty of Engineering of

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY

HYDERABAD

In partial fulfillment of the requirements of the award of Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

GATLA SAI KEERTHANA

(18X01A05D9)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NARSIMHA REDDY ENGINEERING COLLEGE

(UGC AUTONOMOUS)

(Approved by AICTE, Affiliated to JNTUH, Accredited by NBA & NAAC with A-Grade)

Maisammaguda (V), Kompally, Secunderabad, Telangana-500100.

2021-2022

NARSIMHA REDDY ENGINEERING COLLEGE

(UGC AUTONOMOUS)
(Approved by AICTE, Affiliated to JNTUH, Accredited by NBA & NAAC with A-Grade)
Maisammaguda (V), Kompally, Secunderabad, Telangana-500100.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Project Report entitled **“CROP YIELD PREDICTION USING MACHINE LEARNING ALGORITHM”** is a bonafide record of work carried out by **GATLA SAI KEERTHANA(18X01A05D9)** in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in **Computer Science and Engineering** of **Jawaharlal Nehru Technological University, Hyderabad** during the academic year 2021-2022.

PROJECT GUIDE

A.BARKATHULLA

Assitant Professor, Department of CSE

HEAD OF THEDEPARTMENT

Dr.U.M.FERNANDES DIMLO

Professor, Department Of CSE

ABSTRACT

Agriculture is one of the major and the least paid occupation in India. Machine learning can bring a boom in the agriculture field by changing the income scenario through growing the optimum crop. This paper focuses on predicting the yield of the crop by applying various machine learning techniques. The outcome of these techniques is compared on the basis of mean absolute error. The prediction made by machine learning algorithms will help the farmers to decide which crop to grow to get the maximum yield by considering factors like temperature, rainfall, area, etc.

CONTENTS

CHAPTER NAME **PAGE NUMBER**

CHAPTER NAME **PAGE NUMBER**

LIST OF FIGURES

1. INTRODUCTION

1.1 MOTIVATION

The history of agriculture in India[1] dates back to the Indus Valley Civilization Era. India ranks second in this sector. Agriculture and allied sectors like forestry and fisheries account for 15.4 percent of the GDP (gross domestic product) with about 31 percent of the workforce. India ranks first globally with the highest net cropped area followed by US and China. Agriculture is demographically the broadest economic sector and plays a significant role in the overall socio-economic fabric of India. Due to the revolution in industrialization, the economic contribution of agriculture to India's GDP is steadily declining with the country's broad-based economic growth.

1.2 PROBLEM DEFINITION

The problem that the Indian Agriculture sector is facing is the integration of technology to bring the desired outputs. With the advent of new technologies and overuse of non-renewable energy resources patterns of rainfall and temperature are disturbed. The inconsistent trends developed from the side effects of global warming make it cumbersome for the farmers to clearly predict the temperature and rainfall patterns thus affecting their crop yield productivity. In order to perform accurate prediction and handle inconsistent trends in temperature and rainfall various machine learning algorithms like RNN, LSTM, etc can be applied to get a pattern. It will complement the agricultural growth in India and all together augment the ease of living for farmers. In past, many researchers have applied machine learning techniques to enhance agricultural growth of the country

1.3 OBJECTIVE OF PROJECT

This paper focuses on predicting the yield of the crop by applying various machine learning techniques. The outcome of these techniques is compared on the basis of mean absolute error. The prediction made by machine learning algorithms will help the farmers to decide which crop to grow to get the maximum yield by considering factors like temperature, rainfall, area, etc.

2. LITERATURE SURVEY

2.1 PREDICTING YIELD OF THE CROP USING MACHINE LEARNING ALGORITHM

AUTHORS: P.Priya, U.Muthaiah & M.Balamurugan

The agriculture plays a dominant role in the growth of the country's economy. Climate and other environmental changes has become a major threat in the agriculture field. Machine learning (ML) is an essential approach for achieving practical and effective solutions for this problem. Crop Yield Prediction involves predicting yield of the crop from available historical available data like weather parameter, soil parameter and historic crop yield. This paper focus on predicting the yield of the crop based on the existing data by using Random Forest algorithm. Real data of Tamil Nadu were used for building the models and the models were tested with samples. The prediction will helps to the farmer to predict the yield of the crop before cultivating onto the agriculture field. To predict the crop yield in future accurately Random Forest, a most powerful and popular supervised machine learning algorithm is used.

2.2 APPLICATIONS OF MACHINE LEARNING TECHNIQUES IN AGRICULTURAL CROP PRODUCTION

AUTHORS: Mishra .S, Mishra .D and Santra .G. H

This paper has been prepared as an effort to reassess the research studies on the relevance of machine learning techniques in the domain of agricultural crop production. Methods/Statistical Analysis: This method is a new approach for production of agricultural crop management. Accurate and timely forecasts of crop production are necessary for important policy decisions like import-export, pricing marketing distribution etc. which are issued by the directorate of economics and statistics. However one has understand that these prior estimates are not the objective estimates as these estimate requires lots of descriptive assessment based on many different qualitative factors. Hence there is a requirement to develop statistically sound objective prediction of crop production. provided large amount of data.

2.3 A MODEL FOR PREDICTION OF CROP YIELD

AUTHORS: Manjula.E

Data Mining is emerging research field in crop yield analysis. Yield prediction is a very important issue in agricultural. Any farmer is interested in knowing how much yield he is about to expect. In the past, yield prediction was performed by considering farmer's experience on particular field and crop. The yield prediction is a major issue that remains to be solved based on available data. Data mining techniques are the better choice for this purpose. Different Data Mining techniques are used and evaluated in agriculture for estimating the future year's crop production. This research proposes and implements a system to predict crop yield from previous data. This is achieved by applying association rule mining on agriculture data. This research focuses on creation of a prediction model which may be used to future prediction of crop yield. This paper presents a brief analysis of crop yield prediction using data mining technique based on association rules for the selected region i.e. district of Tamil Nadu in India. The experimental results shows that the proposed work efficiently predict the crop yield production.

2.4 AGRICULTURAL CROP YIELD PREDICTION USING ARTIFICIAL NEURAL NETWORK APPROACH

AUTHORS: Dahikar, S. S, Rode and S. V.

By considering various situations of climatologically phenomena affecting local weather conditions in various parts of the world. These weather conditions have a direct effect on crop yield. Various researches have been done exploring the connections between large-scale climatologically phenomena and crop yield. Artificial neural networks have been demonstrated to be powerful tools for modeling and prediction, to increase their effectiveness. Crop prediction methodology is used to predict the suitable crop by sensing various parameter of soil and also parameter related to atmosphere. Parameters like type of soil, PH, nitrogen, phosphate, potassium, organic carbon, calcium, magnesium, sulphur, manganese, copper, iron, depth, temperature, rainfall, humidity. For that purpose we are used artificial neural network (ANN).

2.5 PREDICTIVE ABILITY OF MACHINE LEARNING METHODS FOR MASSIVE CROP YIELD PREDICTION.

AUTHORS: Gonzlez Snchez. A, Frausto Sols. J and Ojeda Bustamante. W

An important issue for agricultural planning purposes is the accurate yield estimation for the numerous crops involved in the planning. Machine learning (ML) is an essential approach for achieving practical and effective solutions for this problem. Many comparisons of ML methods for yield prediction have been made, seeking for the most accurate technique. Generally, the number of evaluated crops and techniques is too low and does not provide enough information for agricultural planning purposes. This paper compares the predictive accuracy of ML and linear regression techniques for crop yield prediction in ten crop datasets. Multiple linear regression, M5-Prime regression trees, perceptron multilayer neural networks, support vector regression and k-nearest neighbor methods were ranked. Four accuracy metrics were used to validate the models: the root mean square error (RMS), root relative square error (RRSE), normalized mean absolute error (MAE), and correlation factor (R). Real data of an irrigation zone of Mexico were used for building the models. Models were tested with samples of two consecutive years. The results show that M5- Prime and k-nearest neighbor techniques obtain the lowest average RMSE errors (5.14 and 4.91), the lowest RRSE errors (79.46% and 79.78%), the lowest average MAE errors (18.12% and 19.42%), and the highest average correlation factors (0.41 and 0.42). Since M5-Prime achieves the largest number of crop yield models with the lowest errors, it is a very suitable tool for massive crop yield prediction in agricultural planning.

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

Due to the revolution in industrialization, the economic contribution of agriculture to India’s GDP is steadily declining with the country’s broad-based economic growth. The problem that the Indian Agriculture sector is facing is the integration of technology to bring the desired outputs. With the advent of new technologies and overuse of non-renewable energy resources patterns of rainfall and temperature are disturbed. The inconsistent trends developed from the side effects of global warming make it cumbersome for the farmers to clearly predict the temperature and rainfall patterns thus affecting their crop yield productivity. In order to perform accurate prediction and handle inconsistent trends in temperature and rainfall various machine learning algorithms like RNN, LSTM, etc can be applied to get a pattern. It will complement the agricultural growth in India and all together augment the ease of living for farmers. In past, many researchers have applied machine learning techniques to enhance agricultural growth of the country.

3.2 PROPOSED SYSTEM

This paper focuses on the practical application of machine learning algorithms and its quantification. The work presented here also takes into account the inconsistent data from rainfall and temperature datasets to get a consistent trend. Crop yield prediction is determined by considering all the features in contrast with the usual trend of determining the prediction considering one feature at a time.

ADVANTAGES OF PROPOSED SYSTEM: Achieving the maximum crop at minimum yield is the ultimate Aim of the project. Early detection of problems and management of that problems can help the farmers for better crop yield. For the better understanding of the crop yield, we need to study of the huge data with the help of machine learning algorithm so it will give the accurate yield for that crop and suggest the farmer for a better crop.

3.3 SYSTEM STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate.

4. SYSTEM REQUIREMENTS

4.1 HARDWARE REQUIREMENTS

• System	: Pentium IV 2.4 GHz.
• Hard Disk	: 40 GB.
• Floppy Drive	: 1.44 Mb.
• Monitor	: 15 VGA Colour.
• Mouse	: Logitech.
• Ram	: 512 Mb.

4.2 SOFTWARE REQUIREMENTS

• Operating System	: Windows
• Coding Language	:Python 3.7

5. SYSTEM DESIGN

5.1 SYSTEM ARCHITECTURE

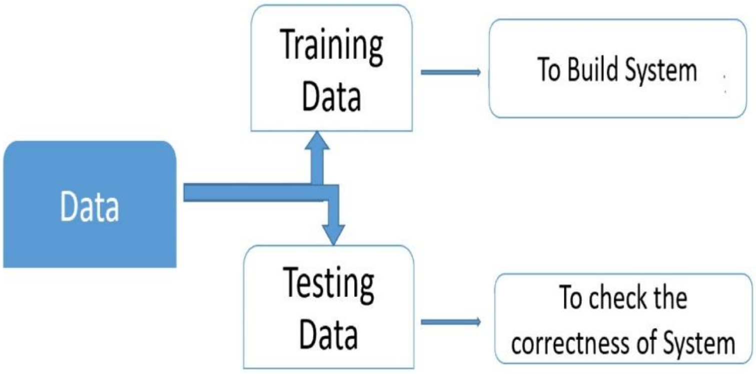


Fig 5.1System Architecture

5.2 DATA FLOW DIAGRAM

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

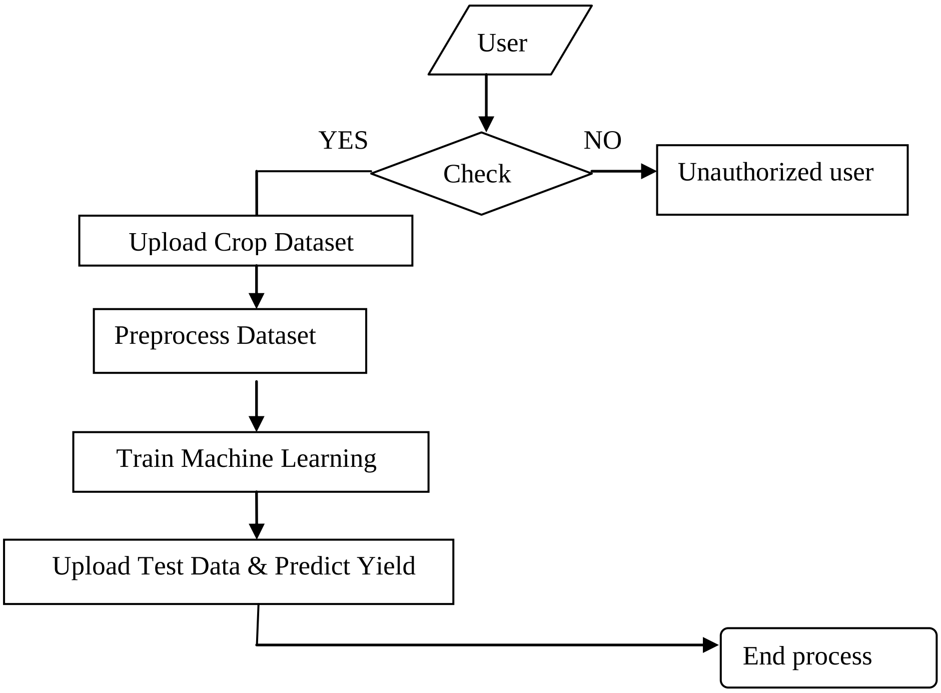


Fig 5.2 Data Flow Diagram

5.3 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group. The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems. The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems. The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

5.4 USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

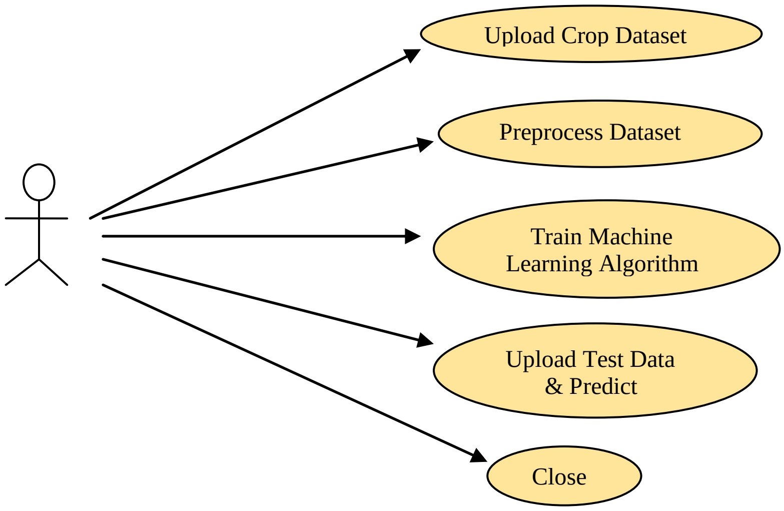


Fig 5.4 Use Case Diagram

5.5 CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

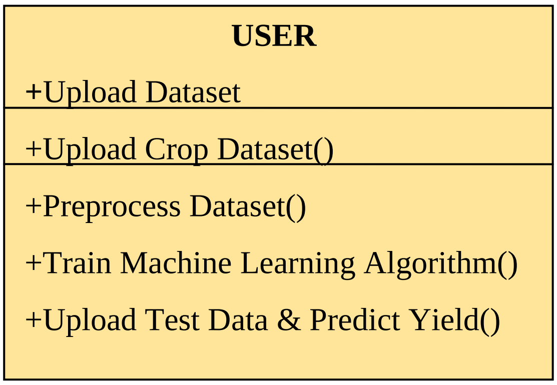

USER
+Upload Dataset
+Upload Crop Dataset()
+Preprocess Dataset()
+Train Machine Learning Algorithm()
+Upload Test Data & Predict Yield()

Fig 5.5 Class Diagram

5.6 SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

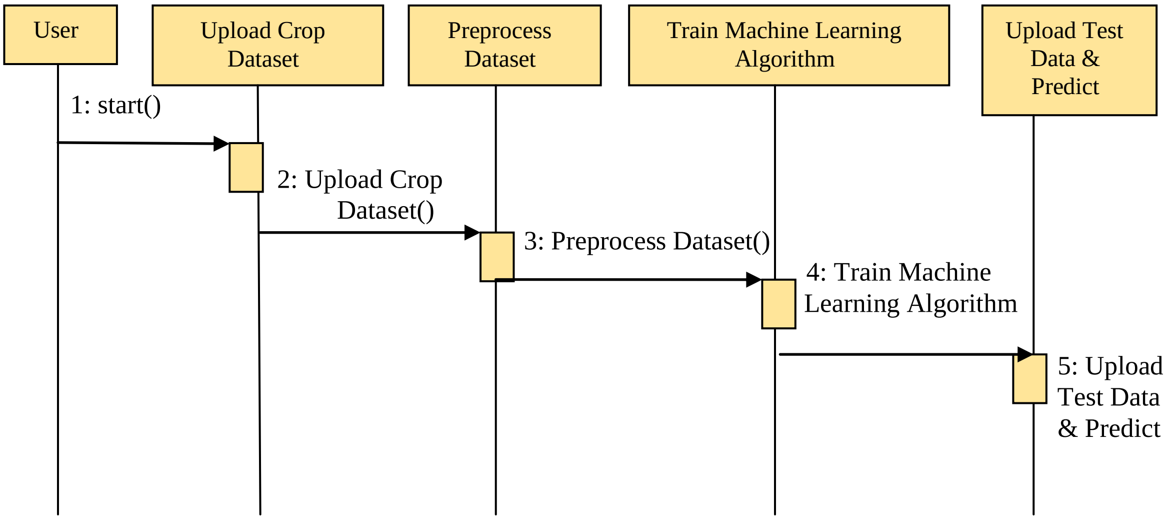

User
Upload Crop Dataset
Preprocess Dataset
Train Machine Learning Algorithm
Upload Test Data & Predict
1: start()
2: Upload Crop Dataset()
3: Preprocess Dataset()
4: Train Machine Learning Algorithm
5: Upload Test Data & Predict

Fig 5.6 Sequence Diagram

5.7 ACTIVITY DIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

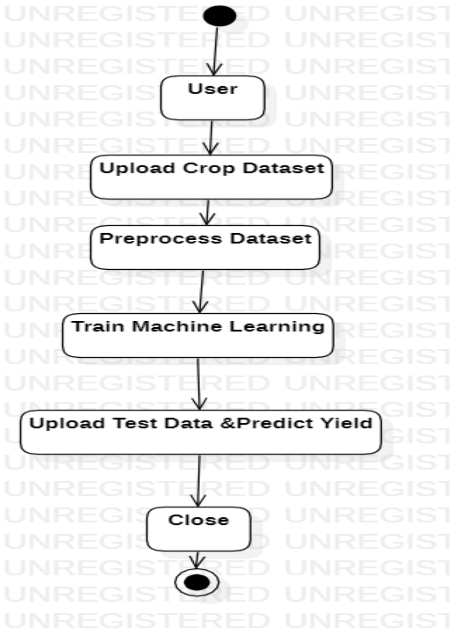


Fig 5.7 Activity Diagram

6. IMPLEMENTATION:

6.1 MODULES

Upload Agriculture Dataset

Preprocess Dataset

Train Machine Learning Algorithm

Upload Test Data & Predict Yield

6.2 MODULES DESCRIPTION

Upload Crop Dataset

The crop production dataset that is used to predict the name and yield of the crop is fed into classification and regression algorithms.

Preprocess Dataset

Experiments were conducted on Indian government dataset and it has been established that Random Forest Regressor gives the highest yield prediction accuracy. Sequential model that is Simple Recurrent Neural Network performs better on rainfall prediction while LSTM is good for temperature prediction. By combining rainfall, temperature along with other parameters like season and area, yield prediction for a certain district can be made.

Train Machine Learning Algorithm

This focuses on district wise yield prediction according to the crop sown in the district. Yield is being predicted for given crops district wise and crops with best yield.

Upload Test Data &Predict Yield

Results reveals that Random Forest is the best classifier when all parameters are combined. This will not only help farmers in choosing the right crop to grow in the next season but also bridge the gap between technology and the agriculture sector.

7. SOFTWARE ENVIRONMENT

What is Python:-

Python is currently the most widely used multi-purpose, high-level programming language. Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java. Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time. Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc.

The biggest strength of Python is huge collection of standard library which can be used for the following –

- Machine Learning
- GUI Applications (like Kivy, Tkinter, PyQt etc.)
- Web frameworks like Django (used by YouTube, Instagram, Dropbox)
- Image processing (like Opencv, Pillow)
- Web scraping (like Scrapy, BeautifulSoup, Selenium)
- Test frameworks
- Multimedia

7.1 HISTORY OF PYTHON

What do the alphabet and the programming language Python have in common? Right, both start with ABC. If we are talking about ABC in the Python context, it's clear that the programming language ABC is meant. ABC is a general-purpose programming language and programming environment, which had been developed in the Netherlands, Amsterdam, at the CWI (Centrum Wiskunde &Informatica). The greatest achievement of ABC was to influence the design of Python. Python was conceptualized in the late 1980s. Guido van Rossum worked that time in a project at the CWI, called Amoeba, a distributed operating system. In an interview with Bill Venners[1], Guido van Rossum said: "In the early 1980s, I worked as an implementer on a team building a language called ABC at Centrum voor Wiskunde en Informatica (CWI). I don't know how well people know ABC's influence on Python

7.2 ADVANTAGES OF PYTHON

- Extensive Libraries: Python downloads with an extensive library and it contain code for various purposes like regular expressions, documentation-generation, unit-testing, web browsers, threading, databases, CGI, email, image manipulation, and more. So, we don’t have to write the complete code for that manually.

- Extensible: As we have seen earlier, Python can be extended to other languages. You can write some of your code in languages like C++ or C. This comes in handy, especially in projects.

- Embeddable: Complimentary to extensibility, Python is embeddable as well. You can put your Python code in your source code of a different language, like C++. This lets us add scripting capabilities to our code in the other language.

- Improved Productivity: The language’s simplicity and extensive libraries render programmers more productive than languages like Java and C++ do. Also, the fact that you need to write less and get more things done.

- IOT Opportunities: Since Python forms the basis of new platforms like Raspberry Pi, it finds the future bright for the Internet Of Things. This is a way to connect the language with the real world.

- Simple and Easy: When working with Java, you may have to create a class to print ‘Hello World’. But in Python, just a print statement will do. It is also quite easy to learn**,** understand, and code**.** This is why when people pick up Python; they have a hard time adjusting to other more verbose languages like Java.

- Readable: Because it is not such a verbose language, reading Python is much like reading English. This is the reason why it is so easy to learn, understand, and code. It also does not need curly braces to define blocks, and indentation is mandatory. This further aids the readability of the code.

7.3 DISADVANTAGES OF PYTHON

- Speed Limitations: We have seen that Python code is executed line by line. But since Python is interpreted, it often results in slow execution. This, however, isn’t a problem unless speed is a focal point for the project. In other words, unless high speed is a requirement, the benefits offered by Python are enough to distract us from its speed limitations.

- Weak in Mobile Computing and Browsers: While it serves as an excellent server-side language, Python is much rarely seen on the client-side. Besides that, it is rarely ever used to implement smartphone-based applications. One such application is called Carbonnelle. The reason it is not so famous despite the existence of Brython is that it isn’t that secure.

- Design Restrictions: As you know, Python is dynamically-typed. This means that you don’t need to declare the type of variable while writing the code. It uses duck-typing. But wait, what’s that? Well, it just means that if it looks like a duck, it must be a duck. While this is easy on the programmers during coding, it can raise run-time errors.

- Underdeveloped Database Access Layers: Compared to more widely used technologies like JDBC (Java DataBase Connectivity) and ODBC (Open DataBase Connectivity), Python’s database access layers are a bit underdeveloped. Consequently, it is less often applied in huge enterprises.

- Simple: No, we’re not kidding. Python’s simplicity can indeed be a problem. Take my example. I don’t do Java; I’m more of a Python person. To me, its syntax is so simple that the verbosity of Java code seems unnecessary.

7.4 WHAT IS MACHINE LEARNING?

Before we take a look at the details of various machine learning methods, let's start by looking at what machine learning is, and what it isn't. Machine learning is often categorized as a subfield of artificial intelligence, but I find that categorization can often be misleading at first brush. The study of machine learning certainly arose from research in this context, but in the data science application of machine learning methods, it's more helpful to think of machine learning as a means of building models of data.

7.5 CATEGORIES OF MACHINE LEARNING

Supervised learning involves somehow modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data. This is further subdivided into classification tasks and regression tasks: in classification, the labels are discrete categories, while in regression, the labels are continuous quantities. We will see examples of both types of supervised learning in the following section.

Unsupervised learning involves modeling the features of a dataset without reference to any label, and is often described as "letting the dataset speak for itself." These models include tasks such as clustering and dimensionality reduction. Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data. We will see examples of both types of unsupervised learning in the following section.

7.6 APPLICATIONS OF MACHINE LEARNING

- Emotion analysis
- Sentiment analysis
- Error detection and prevention
- Weather forecasting and prediction
- Stock market analysis and forecasting
- Speech synthesis

7.7 ADVANTAGES OF MACHINE LEARNING

- Easily identifies trends and patterns: Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent.

- No human intervention needed (automation): With ML, you don’t need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own.

- Continuous Improvement: As ML algorithms gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model.

- Handling multi-dimensional and multi-variety data: Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

- Wide Applications: You could be an e-tailer or a healthcare provider and make ML work for you. Where it does apply, it holds the capability to help deliver a much more personal experience to customers while also targeting the right customers.

7.8 DISADVANTAGES OF MACHINE LEARNING

- Data Acquisition*:* Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. There can also be times where they must wait for new data to be generated.

- Time and Resources*:* ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function.

- Interpretation of Results*:* Another major challenge is the ability to accurately interpret results generated by the algorithms. You must also carefully choose the algorithms for your purpose.

- High error-susceptibility: Machine Learning is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enough to not be inclusive. You end up with biased predictions coming from a biased training set.

7.9 MODULES USED IN PROJECT

- Tensorflow: Tensorflow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google. TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 open-source license on November 9, 2015.
- Numpy: Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. .It is the fundamental package for scientific computing with Python. It contains various features including these important ones:
    - A powerful N-dimensional array object
    - Sophisticated (broadcasting) functions
    - Tools for integrating C/C++ and Fortran code
    - Useful linear algebra, Fourier transform, and random number capabilities
- Pandas: Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis.
- Matplotlib: Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible.
- Scikit-learn: Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.

8. SAMPLE CODE

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import normalize
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import r2_score,mean_squared_error

crop_dataset = pd.read_csv ('Dataset/Dataset.csv')
crop_dataset.fillna(0, inplace = True)
crop_dataset ['Production'] = crop_dataset ['Production'].as type (np.int64)

le = LabelEncoder ()
crop_dataset['State_Name'] = pd.Series(le.fit_transform(crop_dataset['State_Name']))
crop_dataset['District_Name'] = pd.Series(le.fit_transform(crop_dataset['District_Name']))
crop_dataset ['Season'] = pd.Series (le.fit_transform (crop_dataset ['Season']))
crop_dataset ['Crop'] = pd.Series (le.fit_transform (crop_dataset ['Crop']))
crop_datasets = crop_dataset.values
cols = crop_datasets.shape[1]-1
X = crop_datasets [:,0:cols]
Y = crop_datasets [:,cols]
Y = Y.astype ('uint8')

X = normalize(X)

#X = X.reshape(-1, 1)
#Y = Y.reshape(-1, 1)
```

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=0)

print(X)
print(Y)
clf=DecisionTreeRegressor(max_depth=100,random_state=0,max_leaf_nodes=20,max_features=5,splitter="random")
clf.fit(X, Y)
predict = clf.predict (X_test)
print (predict)
#score = clf.score (predict, Y)
#print (score)

test = pd.read_csv ('Dataset/test.csv')
test.fillna(0, inplace = True)
test['State_Name'] = pd.Series(le.fit_transform(test['State_Name']))
test['District_Name'] = pd.Series(le.fit_transform(test['District_Name']))
test['Season'] = pd.Series(le.fit_transform(test['Season']))
test['Crop'] = pd.Series(le.fit_transform(test['Crop']))
test = test.values
test = normalize(test)
cols = test.shape[1]
test = test[:,0:cols]
print(clf.predict(test))

mse = mean_squared_error(predict,y_test)
rmse = np.sqrt(mse);
print(rmse)
```

8.1 SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner.

8.2 TYPES OF TESTS

- Unit testing: Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive.

- Integration testing: Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent.

- Functional test: Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:

  Valid Input : identified classes of valid input must be accepted.
  Invalid Input : identified classes of invalid input must be rejected.
  Functions : identified functions must be exercised.
  Output : identified classes of application outputs must be exercised.
  Systems/Procedures : interfacing systems or procedures must be invoked.

- System Test: System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

- White Box Testing: White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

- Black Box Testing: Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

9. OUTPUT SCREENS

To run project double click on ‘run.bat’ file to get below screen

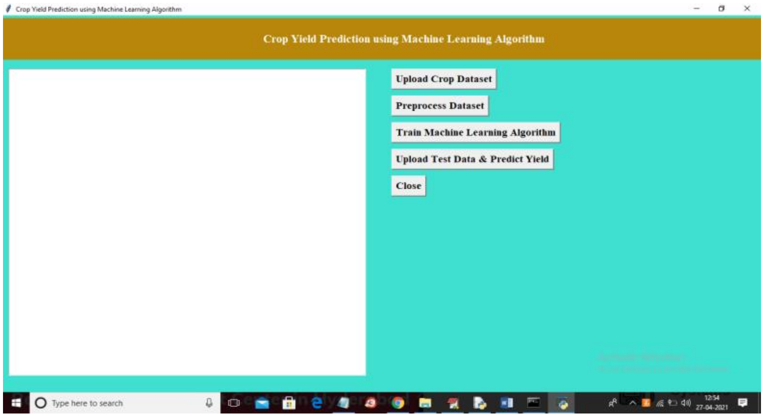

Crop Yield Prediction using Machine Learning Algorithm
Crop Yield Prediction using Machine Learning Algorithm
Upload Crop Dataset
Preprocess Dataset
Train Machine Learning Algorithm
Upload Test Data & Predict Yield
Close
Type here to search

Fig 9.1 Upload Crop Dataset

In above screen click on ‘Upload Crop Dataset’ button to upload dataset

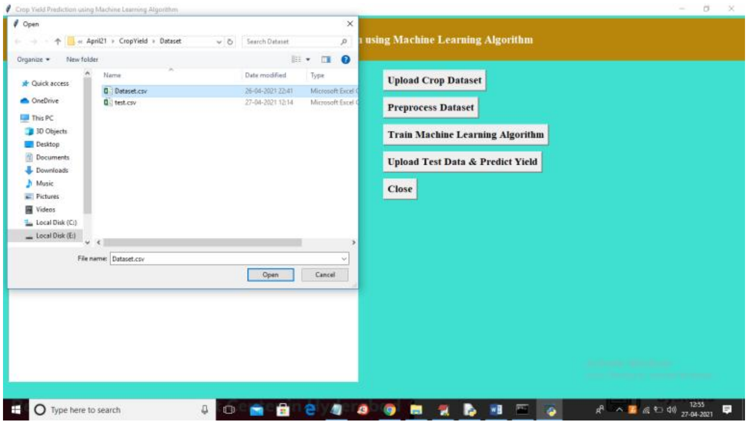


Fig 9.2 Dataset.csv File

In above screen selecting and uploading ‘Dataset.csv’ file and then click on ‘Open’ button to load dataset and to get below screen

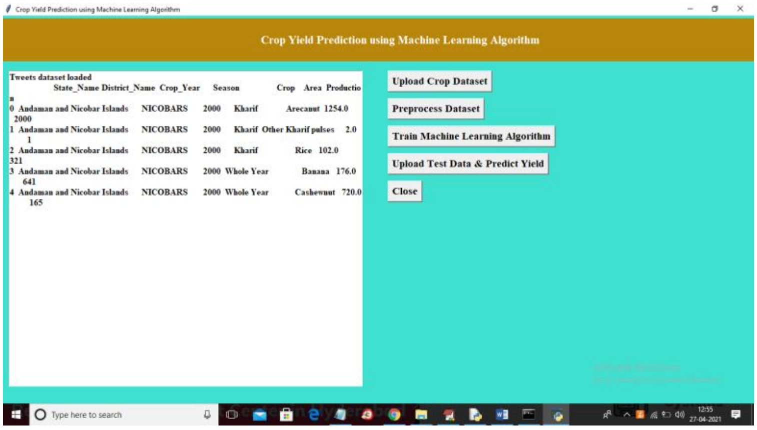
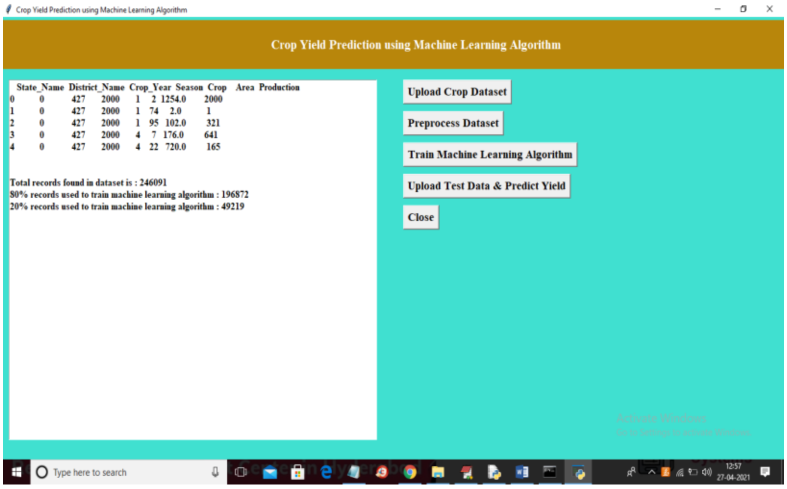

```
Tweets dataset loaded
              State_Name District_Name  Crop_Year     Season         Crop   Area  Productio
n
0  Andaman and Nicobar Islands     NICOBARS     2000     Kharif        Arecanut  1254.0
 2000
1  Andaman and Nicobar Islands     NICOBARS     2000     Kharif  Other Kharif pulses     2.0
    1
2  Andaman and Nicobar Islands     NICOBARS     2000     Kharif            Rice   102.0
321
3  Andaman and Nicobar Islands     NICOBARS     2000  Whole Year         Banana   176.0
   641
4  Andaman and Nicobar Islands     NICOBARS     2000  Whole Year       Cashewnut   720.0
    165
```
Crop Yield Prediction using Machine Learning Algorithm
Crop Yield Prediction using Machine Learning Algorithm
Upload Crop Dataset
Preprocess Dataset
Train Machine Learning Algorithm
Upload Test Data & Predict Yield
Close
Type here to search

Fig 9.3 Preprocess Dataset

In above screen dataset loaded and we can see dataset contains some non-numeric values and ML will not take non-numeric values so we need to preprocess dataset to convert non-numeric values to numeric values by assigning ID to each non-numeric value. So click on ‘Preprocess Dataset’ button to process dataset.

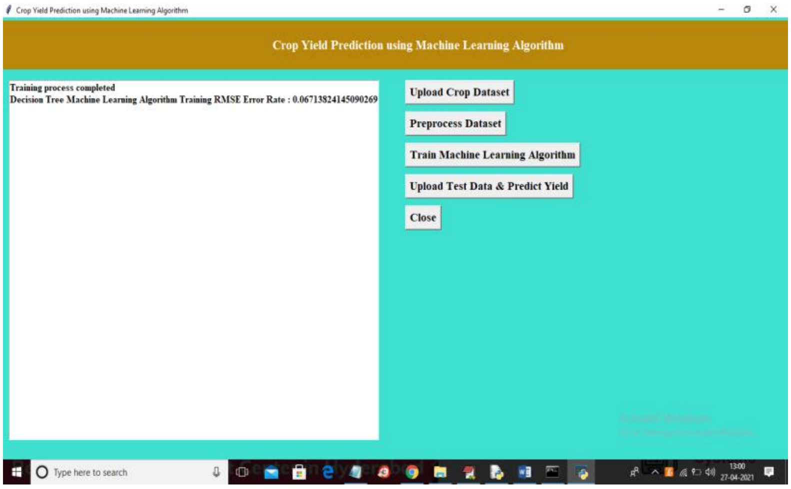

Crop Yield Prediction using Machine Learning Algorithm
Crop Yield Prediction using Machine Learning Algorithm
Upload Crop Dataset
Preprocess Dataset
Train Machine Learning Algorithm
Upload Test Data & Predict Yield
Close
Type here to search
12:57
27-04-2021

```
  State_Name  District_Name  Crop_Year  Season  Crop    Area  Production
0          0            427       2000       1     2  1254.0        2000
1          0            427       2000       1    74     2.0           1
2          0            427       2000       1    95   102.0         321
3          0            427       2000       4     7   176.0         641
4          0            427       2000       4    22   720.0         165

Total records found in dataset is : 246091
80% records used to train machine learning algorithm : 196872
20% records used to train machine learning algorithm : 49219
```

Fig 9.4 Train Machine Learning Algorithm

In above screen all non-numeric values converted to numeric format and in below lines we can see dataset contains total 246091 records and application using (80%) 196872 records to train ML and using (20%) 49219 records to test ML prediction error rate (RMSE (root mean square error)). Now click on ‘Train Machine Learning Algorithm’ button to train Decision Tree Machine learning algorithm on above dataset and then calculate prediction error rate


Crop Yield Prediction using Machine Learning Algorithm
Crop Yield Prediction using Machine Learning Algorithm
Upload Crop Dataset
Preprocess Dataset
Train Machine Learning Algorithm
Upload Test Data & Predict Yield
Close
Type here to search
13:00
27-04-2021

```
Training process completed
Decision Tree Machine Learning Algorithm Training RMSE Error Rate : 0.06713824145090269
```

Fig 9.5 Upload Test Data & Predict Yield

In above screen ML is trained and we got prediction error rate as 0.067% and now Decision Tree model is ready and now click on ‘Upload Test Data & Predict Yield’ button to upload test data and then application will predict production

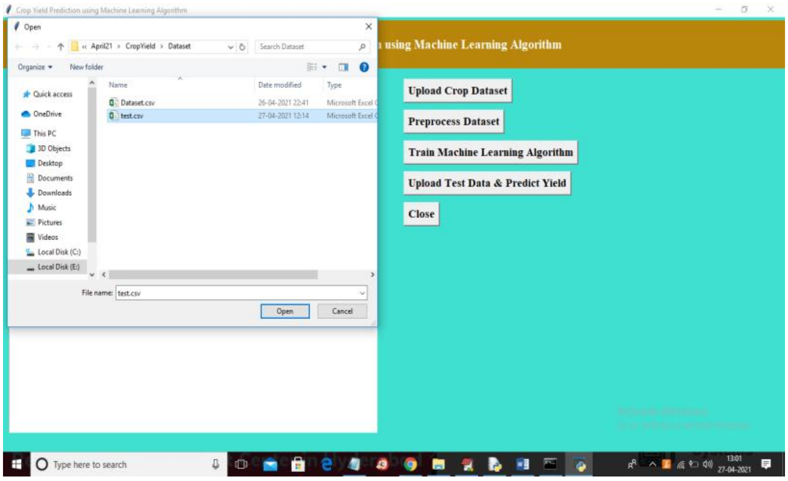
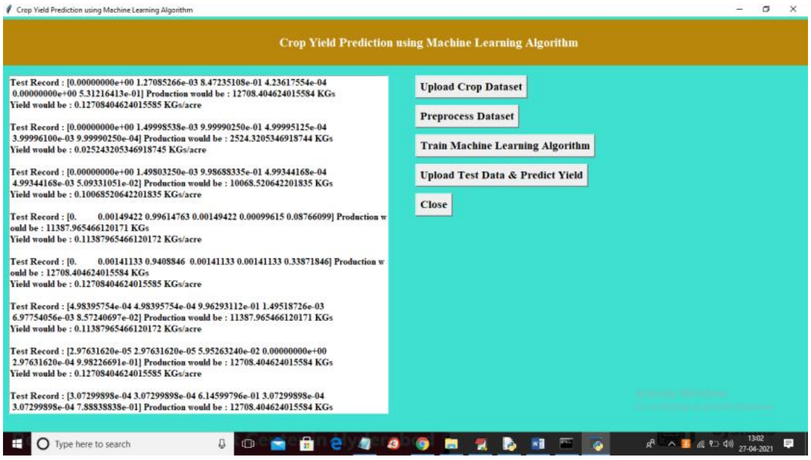


Fig 9.6 Test.csv File

In above screen selecting and uploading ‘test.csv’ file and then click on ‘Open’ button to load test data and then application will give below prediction result


Crop Yield Prediction using Machine Learning Algorithm
Crop Yield Prediction using Machine Learning Algorithm
Upload Crop Dataset
Preprocess Dataset
Train Machine Learning Algorithm
Upload Test Data & Predict Yield
Close
Type here to search
13:02
27-04-2021

```
Test Record : [0.00000000e+00 1.27085266e-03 8.47235108e-01 4.23617554e-04
 0.00000000e+00 5.31216413e-01] Production would be : 12708.404624015584 KGs
Yield would be : 0.12708404624015585 KGs/acre

Test Record : [0.00000000e+00 1.49998538e-03 9.99990250e-01 4.99995125e-04
 3.99996100e-03 9.99990250e-04] Production would be : 2524.3205346918744 KGs
Yield would be : 0.025243205346918745 KGs/acre

Test Record : [0.00000000e+00 1.49803250e-03 9.98688335e-01 4.99344168e-04
 4.99344168e-03 5.09331051e-02] Production would be : 10068.520642201835 KGs
Yield would be : 0.10068520642201835 KGs/acre

Test Record : [0.         0.00149422 0.99614763 0.00149422 0.00099615 0.08766099] Production w
ould be : 11387.965466120171 KGs
Yield would be : 0.11387965466120172 KGs/acre

Test Record : [0.         0.00141133 0.9408846  0.00141133 0.00141133 0.33871846] Production w
ould be : 12708.404624015584 KGs
Yield would be : 0.12708404624015585 KGs/acre

Test Record : [4.98395754e-04 4.98395754e-04 9.96293112e-01 1.49518726e-03
 6.97754056e-03 8.57240697e-02] Production would be : 11387.965466120171 KGs
Yield would be : 0.11387965466120172 KGs/acre

Test Record : [2.97631620e-05 2.97631620e-05 5.95263240e-02 0.00000000e+00
 2.97631620e-04 9.98226691e-01] Production would be : 12708.404624015584 KGs
Yield would be : 0.12708404624015585 KGs/acre

Test Record : [3.07299898e-04 3.07299898e-04 6.14599796e-01 3.07299898e-04
 3.07299898e-04 7.88838838e-01] Production would be : 12708.404624015584 KGs
```

Fig 9.7 OUTPUT

In above screen each test record is separated with newline and in above screen in square bracket we can see test data values and after square bracket we can see predicted production and after that we can see predicted YIELD per acre. So each test record and its prediction is separated with newline.

10. CONCLUSION

The paper presented the various machine learning algorithms for predicting the yield of the crop on the basis of temperature, rainfall, season and area. Experiments were conducted on Indian government dataset and it has been established that Random Forest Regressor gives the highest yield prediction accuracy. Sequential model that is Simple Recurrent Neural Network performs better on rainfall prediction while LSTM is good for temperature prediction. By combining rainfall, temperature along with other parameters like season and area, yield prediction for a certain district can be made. Results reveals that Random Forest is the best classifier when all parameters are combined. This will not only help farmers in choosing the right crop to grow in the next season but also bridge the gap between technology and the agriculture sector.

FUTURE SCOPE

- Machine Learning is a crucial perspective for acquiring real-world and operative solution for crop yield issue. From a given set of predictors, ML can predict a target/outcome by using Supervised Learning.
- Crop yield prediction incorporates forecasting the yield of crop per acre from past historical data which gives an overview what crop is harvested according the season in different states.

11. REFERENCES

1. Agriculture Role on Indian Economy Madhusudhan L - https://www.omicsonline.org/open-access/agriculture-role-on-indianeconomy-2151-6219-1000176.php?aid=62176
2. Priya P., Muthaiah U., Balamurugan M. International Journal of Engineering Sciences Research Technology Predicting Yield of the Crop Using Machine Learning Algorithm.
3. Mishra S., Mishra, D., Santra, G. H. (2016). Applications of machine learning techniques in agricultural crop production: a review paper.Indian J. Sci. Technol, 9(38), 1-14.
4. Manjul E., Djodiltachoumy S. (2017). A Model for Prediction of Crop Yield. International Journal of Computational Intelligence and Informatics,6(4), 2349-6363.
5. Dahikar S. S., Rode, S. V. (2014). Agricultural crop yield prediction using artificial neural network approach. International journal of innovative research in electrical, electronics, instrumentation and control engineering, 2(1), 683-686.
6. Gonzlez Snchez A., Frausto Sols J., Ojeda Bustamante W. (2014). Predictive ability of machine learning methods for massive crop yield prediction.
7. Mandic D. P., Chambers, J. (2001). Recurrent neural networks for prediction: learning algorithms, architectures and stability. John Wiley Sons, Inc.
8. Hochreiter S., Schmidhuber J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
9. Sak H., Senior A., Beaufays F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Fifteenth annual conference of the international speech communication association.
10. Liaw A., Wiener M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.