Visvesvaraya Technological University, Belagavi.

PROJECT REPORT

On

"SOIL CLASSIFICATION USING MACHINE LEARNING METHODS AND CROP SUGGESTION"

Project Report submitted in partial fulfillment of the requirement for the award of the degree of

Bachelor of Engineering

In

Electronics and Communication Engineering

For the academic year 2019-20

Submitted by

1CR16EC053 Jahanavi G

1CR16EC175 Sushma U R

Under the guidance of

Prof. Mr. Rahul Tiwari

Assistant Professor

Department of ECE

CMRIT, Bengaluru

Department of Electronics and Communication Engineering

CMR Institute of Technology, Bengaluru – 560 037

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

CERTIFICATE

This is to certify that the dissertation work **"Soil Classification using Machine Learning Methods and Crop Suggestion"** carried out by Jahanavi G, Sushma U R with USN: 1CR16EC053, 1CR16EC175 bonafide students of **CMRIT** in partial fulfillment for the award of **Bachelor of Engineering** in **Electronics and Communication Engineering** of the **Visvesvaraya Technological University, Belagavi,** during the academic year **2019-20**. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said degree.

Signature of Guide	Signature of HOD	Signature of Principal
Mr. Rahul Tiwari Designation, Dept. of ECE., CMRIT, Bengaluru.	Dr. R. Elumalai Head of the Department, Dept. of ECE., CMRIT, Bengaluru.	Dr. Sanjay Jain Principal, CMRIT, Bengaluru.

External Viva

Name of Examiners — Signature & date

1.

2.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible, whose consistent guidance and encouragement crowned our efforts with success.

We consider it as our privilege to express the gratitude to all those who guided in the completion of the project.

We express my gratitude to Principal, **Dr. Sanjay Jain,** for having provided me the golden opportunity to undertake this project work in their esteemed organization.

We sincerely thank **Dr. R. Elumalai,** HOD**,** Department of Electronics and Communication Engineering, CMR Institute of Technology for the immense support given to me.

We express my gratitude to our project guide **Mr. Rahul Tiwari,** Assistant Professor, Department of Electronics and Communication Engineering, CMR Institute of Technology for their support, guidance and suggestions throughout the project work.

Last but not the least, heartful thanks to our parents and friends for their support.

Above all, we thank the Lord Almighty for His grace on us to succeed in this endeavor.

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

India is an agriculture based country in which 50% workforce is involved in agricultural activities. India accounts for 7.68% of total global agricultural output. Contribution of agriculture sector in Indian economy is much higher than world average (6.1%). In India, farmers follow a traditional method of farming. Farmers plant crops without the knowledge of the content and quality of the soil. As a result, farmers incur loss due to crop failure. Traditional farms in India still have some of the lowest per capital productivity and farmer incomes. This sector also requires a lot of human efforts to do different kinds of tasks like watering crop, cultivating crop, spreading pesticides etc.

Testing of soil is important because it helps in determining the soil fertility using which, crop suitable for that soil can be predicted. Right crop in the right soil returns a good yield. Soil is an important aspect of agriculture. There are several types of soil. Each soil type has its own features and is suitable only for few crops. We need to know the features and characteristics of the soil types to understand which crop is suitable for that soil type. Machine learning techniques can be helpful in this case. In recent years, it has progressed a lot. Machine learning is still an emerging and challenging research field in agricultural data analysis.

In this project, we have proposed a model that uses N, P, K, pH, moisture, and temperature values of the soil and suggests a suitable crop, estimates the amount of fertilizer, and predicts the yield using machine learning algorithms. Several machine learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest are used.

1.1 Existing System

In current scenario, government labs take a few days or weeks to test particular soil sample. There are chances of misplacement of soil samples and farmers getting wrong report.

Disadvantages of existing system are:

1. Manual approach.

2. Chances of loss.

3. Lack of farmer satisfaction.

4. Less efficient.

5. Comparatively long duration.

1.2 Proposed System

The main aim of our system is to partially automate current manual soil testing procedure. In our system we are proposing a method for crop suggestion, fertilizer estimation and yield prediction based on the attributes like nitrogen (N), phosphorous (P), potassium (K), pH, moisture content and temperature.

For our software model we will be training crop database and use machine learning algorithms to give the crop suitable for that particular soil sample along with the approximate yield and fertilizer estimation.

Chapter 2

LITERATURE SURVEY

New farming management systems were introduced which were to work over the internet. Ref [1] mentions a Farm Management System functional architecture that utilizes advanced Future Internet characteristics. Its main characteristics include apart from the support of the typical farming procedures, the seamless support and integration of different stakeholders and services, interworking with the networked infrastructures and the introduction of autonomic and cognitive elements in the overall management process.

Different methods have been proposed to analyze the fertility of the soil and to suggest the appropriate crop. In ref [2], a microcontroller based soil testing technique is proposed. An electronic device is used to measure the N(nitrogen), P(Phosphorous), K(potassium) and pH(potential of Hydrogen) values to estimate the fertility of the soil in the field of agriculture. This helps in predicting a suitable crop and type of fertilizer to be used. Sensors are used to send the ionic particles and its output is processed by signal conditioning unit. Microcontroller compares the pre-stored values with the actual values and displays the output on the LCD.

In Ref [3], a method for determining soil fertility by considering Ph and electrical conductivity parameter is presented. Ph is measured using Ph meter and electrical conductivity is measured using EC sensor. The reading of Ph meter gives the approximate ratio of various nutrient content present in soil and in what proportion. This approximation of soil nutrient will determine the suitable crop for the land.

Various agricultural parameters like light, soil moisture, temperature, and humidity etc. are monitored and controlled by monitoring and controlling units. Ref [4] reviews some of this monitoring system and proposes to add more parameters like wind speed, wind direction for monitoring as well as the automated control system for light, soil moisture, humidity, soil temperature.

Some methods use one or two prominent features as input to estimate other parameters on the basis of which actual output is predicted. In ref [5] location is the prominent feature. The estimated soil parameters using location and the previous crop data base are used to suggest the crops.

Ref [7], describes the SVM based classification and grading of soil samples using different scientific features. Different algorithms and filters are developed to acquire and process the colored images of the soil samples. These developed algorithms are used to extract different features like color, texture, etc. Different soil types like red, black, clay, alluvial, etc are considered. This project aims at combining both the techniques, where classification of crop for appropriate soil is a part of classification of soil.

In Ref [8], a system to test the soil fertility by using the principal of colorimetry is represented. An aqueous solution of the soil sample is prepared using extracting agents and is subjected to the photodiodes of the colour sensor. The solution develops a colour due to reaction of nutrients in the soil with chemicals. The output by the colour sensor is calibrated with standard values present in the database. To verify the results obtained by the colour sensor the Naive Bayes classification algorithm is used. This algorithm classifies the intensity values of the soil solutions into three class labels namely low, medium, and high. The intended system is thus beneficial to reduce the time required for testing the soil fertility and determining the accuracy of our results.

Ref [9], [10], [11] throws light on how information technology technique cloud computing could be deployed in farming for better data management and accessibility.

Machine learning technique includes examination of enormous dataset which helps in acquiring precise results. In this manner, it is additionally being utilized in the field of agribusiness. This works demonstrates an evaluation of modified k-Means clustering algorithm in crop prediction. The results and evaluation show comparison of modified k-Means over k-Means and k-Means++ clustering algorithm and modified k-Means has achieved the maximum number of high quality clusters, correct prediction of crop and maximum accuracy count [12].

Chapter 3

METHODOLOGY

3.1 Design

3.1.1 System Architecture

System Architecture design-identifies the overall hypermedia structure for the Web-App. Web-App architecture, addresses the manner in which the application is structure to manage user interaction, handle internal processing tasks, effect navigation, and present content. Web-App architecture is defined within the context of the development environment in which the application is to be implemented. Figure shows the System Architecture of the proposed system.

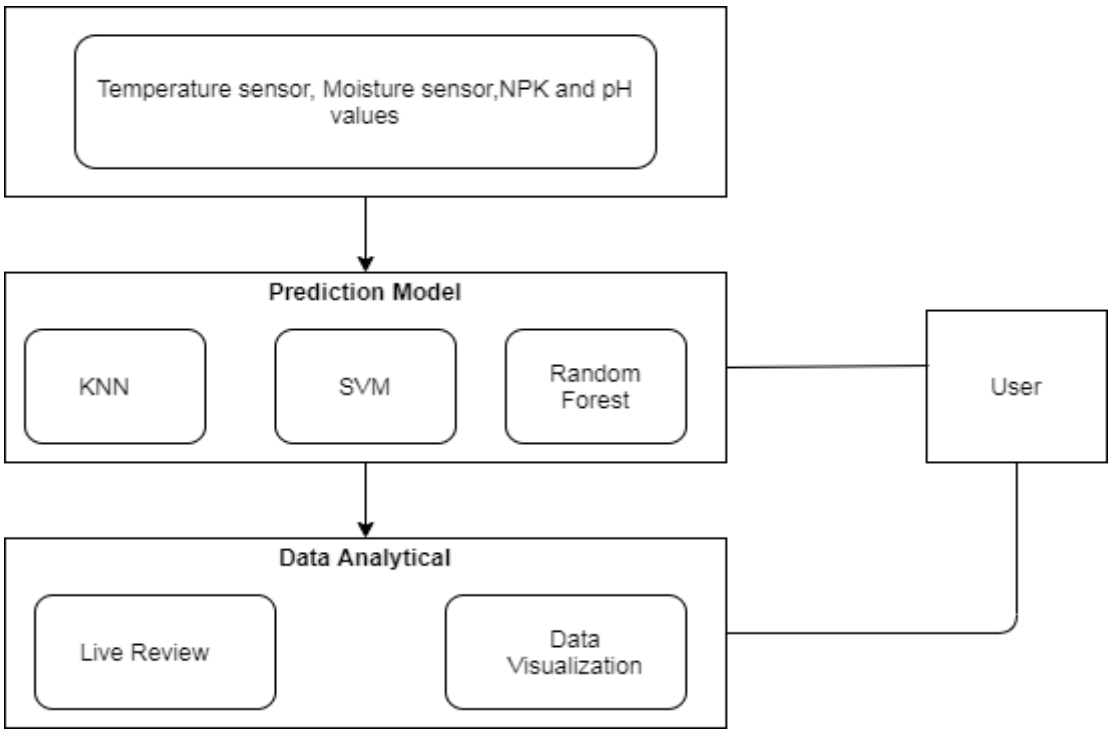

Fig.1: System Architecture

The framework utilizes soil properties such as N, P, K, pH, moisture and temperature values to predict the suitable crop, estimate the amount of fertilizer and predict the approximate yield of the soil using Machine Learning techniques. Machine Learning algorithms such as K- Nearest Neighbor (K-NN), Support Vector Machine (SVM),

Random Forest are used for prediction. The predicted crop, estimated yield and amount of fertilizer are displayed to the user.

3.1.2 Flow Chart

A flowchart is one of the seven basic quality tools used in project management and it displays the actions that are necessary to meet the goals of a particular task in the most practical sequence. Also called as process maps, this type of tool displays a series of steps with branching possibilities that depict one or more inputs and transforms them to outputs. The flow chart of the proposed system is as shown in the Fig.2.

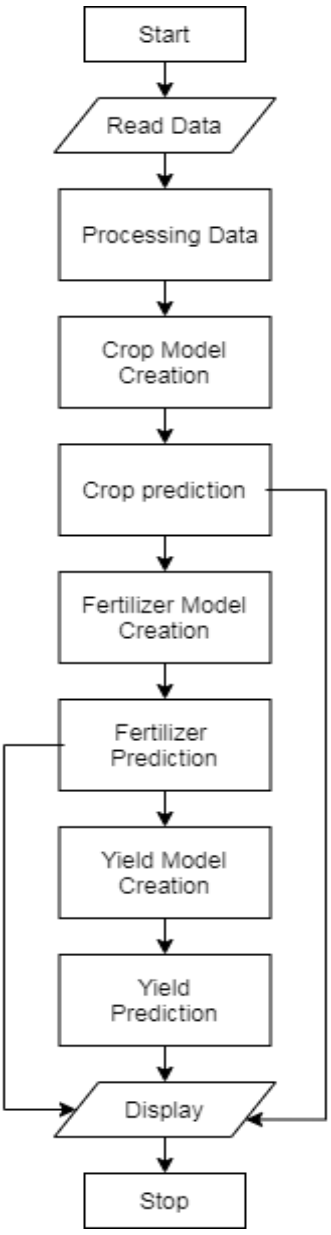

Fig.2: Flow chart

The N, P, K, pH, moisture and temperature values of the soil are taken from the user. The trained dataset comprises of preprocessed soil data. It is trained using machine learning algorithms such as K-NN, SVM, Random forest. The data read from the user is sent to this prediction model.

Crop prediction: The user enters the N, P, K, pH, moisture content and temperature values and chooses the machine learning algorithm to execute among the three. The entered data is processed to predict the suitable crop for the soil. The predicted suitable crop is displayed to the user.

Yield prediction: Along with the N, P, K, pH, moisture and temperature values and algorithm to be implemented user provides the crop for which yield is to be estimated. Taking into consideration all the entered conditions yield is predicted to be high, moderate or low.

Fertilizer Estimation: Along with the N, P, K, pH, moisture and temperature values and algorithm to be implemented user provides the crop for which amount of fertilizer is to be estimated. Taking into consideration all the entered conditions amount of fertilizer is predicted and displayed in the scale of 1 to 5 where 1 represents the least, 2, 3 represents moderate, 4 represents slightly more and 5 represents large amount of fertilizer.

3.1.3 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 are intended to model both computational and organizational processes as well as the data flows intersecting with the related activities. Although activity diagrams primarily show the overall flow of control, they can also include elements showing the flow of data between activities through one or more data stores. Fig.3 shows the activity diagram of the proposed system.

User first registers and then login using the register data. The register data is stored in data base. It checks from the data base whether the user is registered or not if not it says user does not exist. User uploads the soil data and prediction of crop, fertilizer and yield is done by using three algorithms by using the data set from csv file. Data preprocessing is done before the data set is used for prediction.

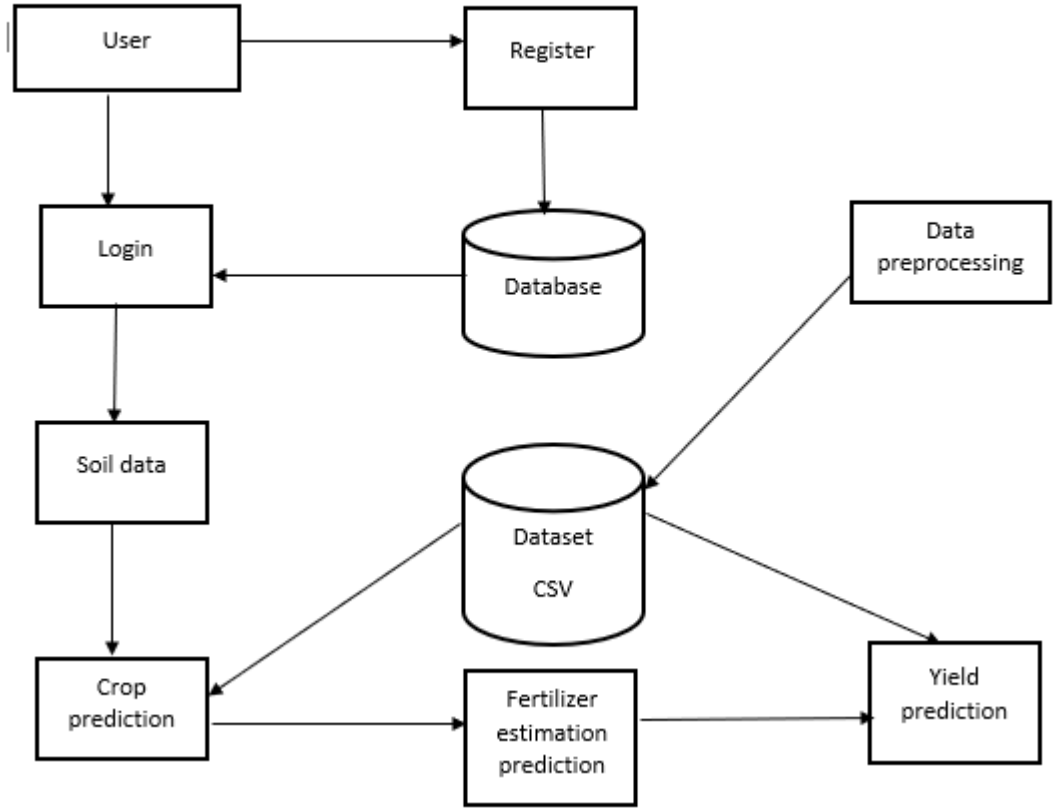

Fig.3: Activity Diagram

3.2 Algorithms

3.2.1 K-Nearest Neighbor (K-NN Algorithm):

It is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

The K-NN working can be explained as follows; initially the correct estimation of K is to be picked. It is ideally an odd number and bigger the K better is the exactness. The Euclidean distance to all the data points from the test data point is calculated. Suppose there are two points A(x1,y1) and B(x2,y2) the Euclidean distance between them is given by $\sqrt{(x2-x1)^2+(y2-y1)^2}$ as shown in Fig.4.

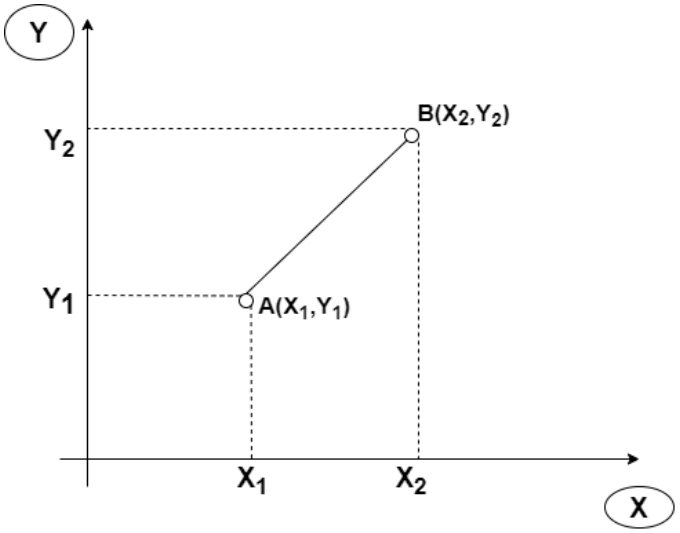

Fig.4: Calculation of Euclidean Distance

The K nearest neighbors as per the calculated Euclidean distance is considered. Among these K neighbours, the number of the data points in each category is calculated and the test data point is assigned to that category for which the number of the neighbor is maximum.

Example: Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. With the help of K-NN, we can easily identify the category or class of a particular dataset.

Consider the below diagram:

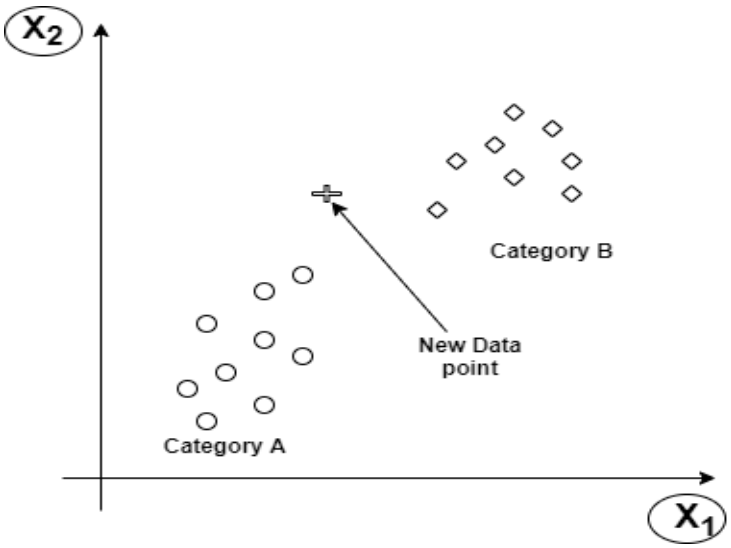

Fig.5: Example for K-NN Algorithm

[1] We will choose the number of neighbors, so we will choose the k=5.

[2] By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Hence the test point belongs to category A. Consider the below image:

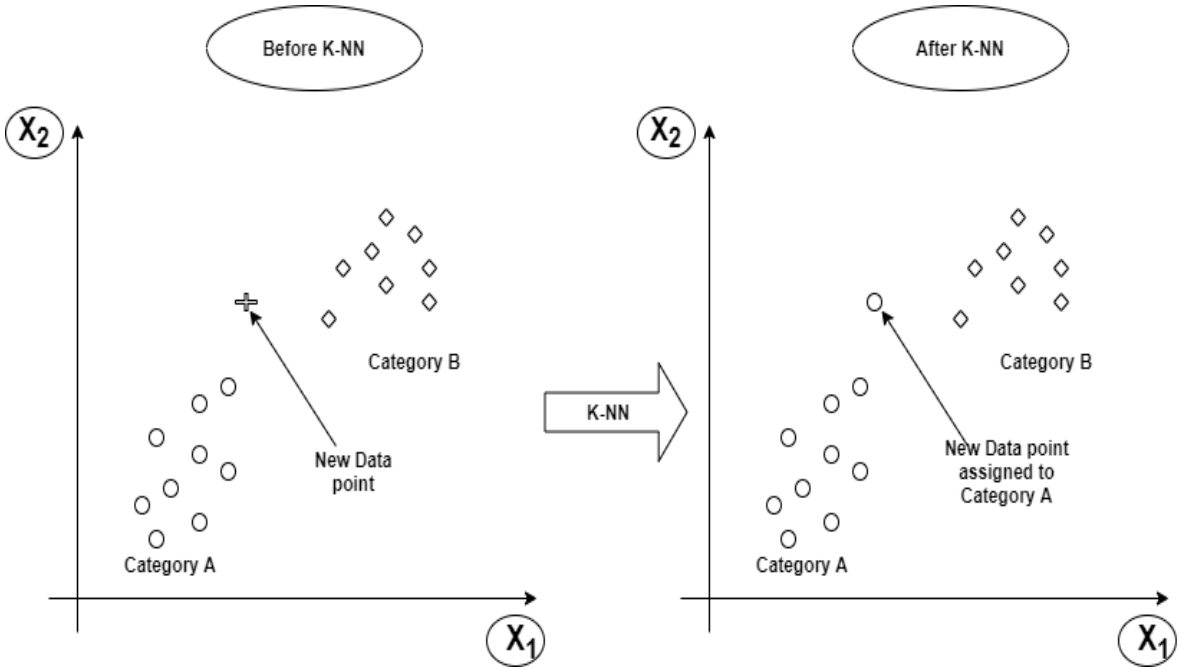

Fig.6: Before and After Execution of K-NN Algorithm

3.2.2 Random Forest Algorithm:

It is also called as Random decision forest algorithm. It is an ensemble machine learning method used for both classification and regression. It consists of a large number of individual decision trees that operate as a group. Each individual tree provides a class prediction and the class predicted by maximum number of trees becomes the model's prediction. For random forest to perform well, there needs to be some actual signal in the considered features so that models built using those features do better than random guessing and the predictions (and hence the errors) made by the individual trees need to have low correlations with each other. This is because the trees guard each other from their individual errors, i.e. even if some trees are wrong, many other trees will be right, so as an ensemble they are in the right direction.

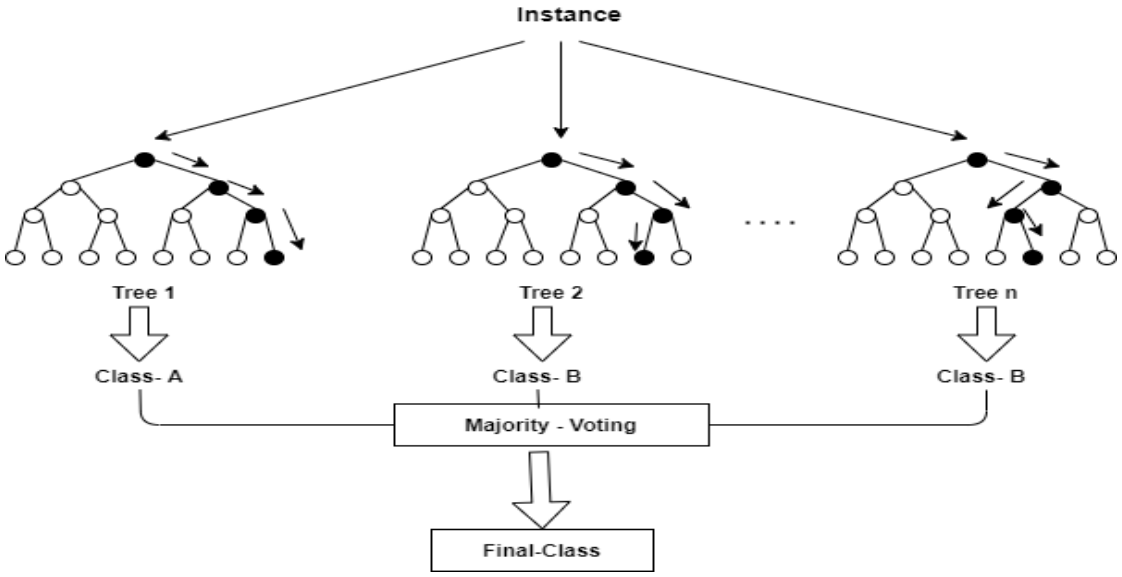

Fig.7: Random Forest Algorithm

3.2.3 Support Vector Machine (SVM) Algorithm:

It is a supervised machine learning algorithm used for both classification and regression challenges. It is mostly used in classification problems. In the SVM algorithm, each data is plotted as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well overall as shown in Fig.8.

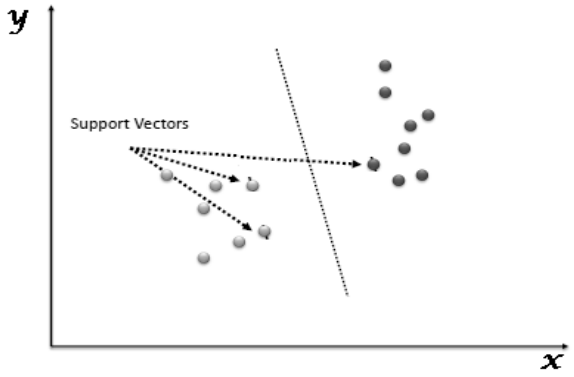

Fig.8: SVM Algorithm Hyper plane

When the data is non-linear, non-linear data is converted to linear data using kernel function. The SVM kernel is a function that takes low dimensional input space and transforms it to a higher dimensional space i.e. it converts not separable problem to separable problem for which in the SVM classifier, it is easy to have a linear hyper-plane.

Identifying the Right Hyper-plane:

Case-1: Here, we have three hyper-planes A, B and C. “Select the hyper-plane which segregates the two classes better”. Hence, hyper-plane “B” is the right choice.

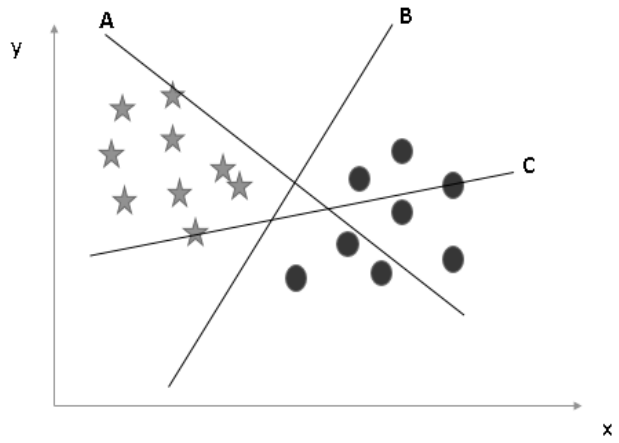

Fig.9: SVM, Hyper-plane Identification Case-1

Case-2: Here, we have three hyper-planes A, B and C and all are segregating the classes well. Here, maximizing the distances between nearest data points (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as Margin**.**

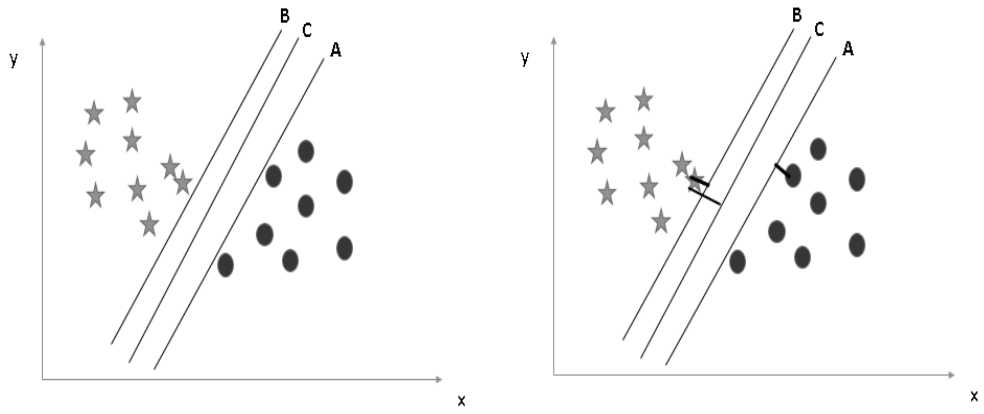

Fig.10: SVM, Hyper-plane Identification Case-2

The margin for hyper-plane C is high as compared to both A and B. Hence, the right hyper-plane is C. Another reason for selecting the hyper-plane with higher margin is robustness. If we select a hyper-plane having low margin then there is high chance of miss-classification.

Case-3: SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin. Here, hyper-plane B has a classification error and A has classified all correctly. Therefore, the right hyper-plane is A**.**

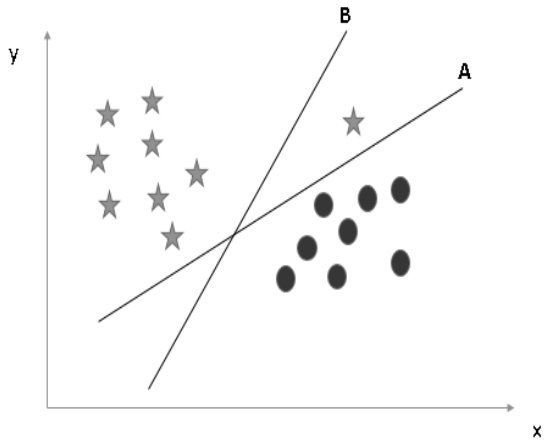

Fig.11: SVM, Hyper-plane Identification Case-3

Case-4: SVM classification is robust to outliers. One star at other end is like an outlier for star class. The SVM algorithm has a feature to ignore outliers and find the hyper-plane that has the maximum margin. Hence, the hyper-plane is as shown.

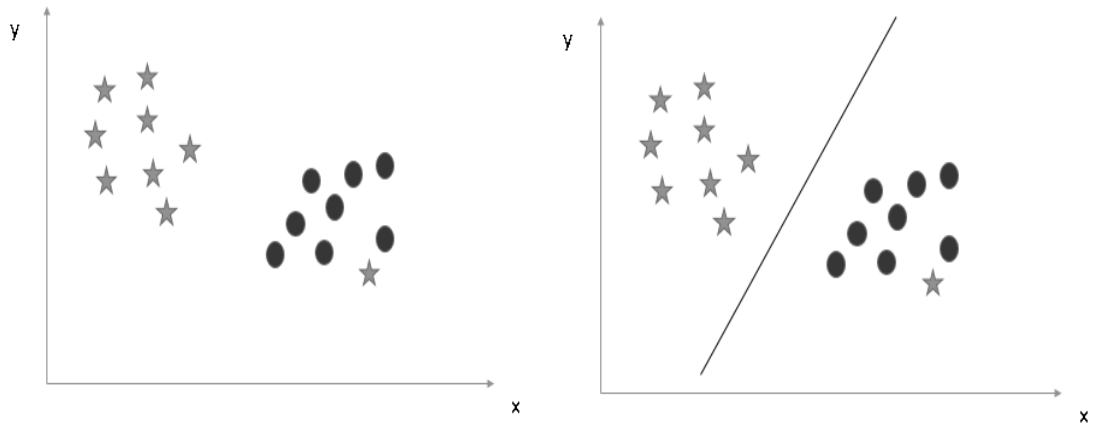

Fig.12: SVM, Hyper-plane Identification Case-4

Case-5: (Non-linear hyper-plane) By using z=x^2+y^2, the data points are plotted on axis x and z. Hence, non-linear data is converted to linear data that is the data points are clearly classified and in the SVM classifier, it is easy to have a linear hyper-plane.

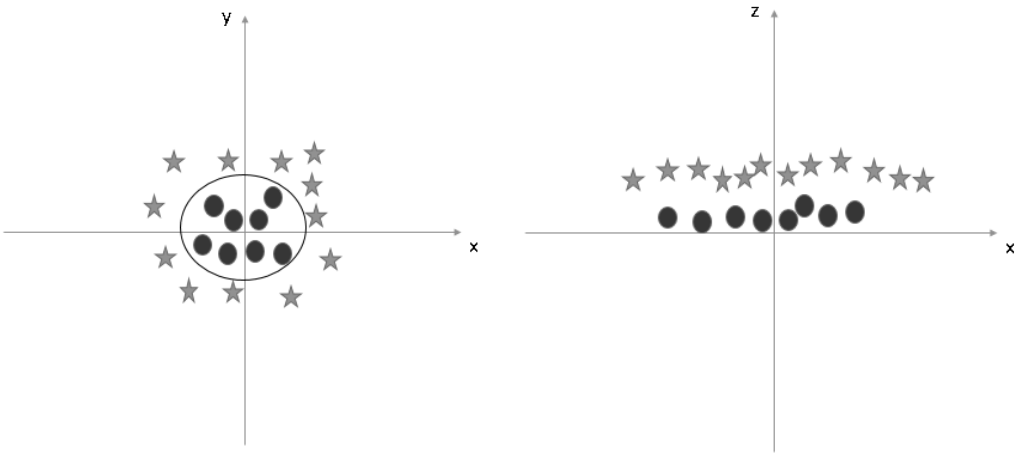

Fig.13: SVM, Hyper-plane Identification Case-5

The SVM algorithm has a technique called the **kernel trick**. The SVM kernel is a function that takes low dimensional input space and transforms it to a higher dimensional space i.e. it converts not separable problem to separable problem. It is mostly useful in non-linear separation problem. Hence when the algorithm is implemented it does some extremely complex data transformations, then finds out the process to separate the data based on the labels or outputs defined.

3.3 Machine Learning Implementation Overview

Machine learning (ML) is a branch of artificial intelligence in which computers learn to discern and act on subtle patterns in data without being explicitly programmed to do so. The ML algorithm learns by using large amounts of training data to adjust its internal parameters until it can reliably discriminate similar patterns in data it has not seen before.

ML is heavily dependent on data, without data, it is impossible for an “AI” to learn. It is the most crucial aspect that makes algorithm training possible

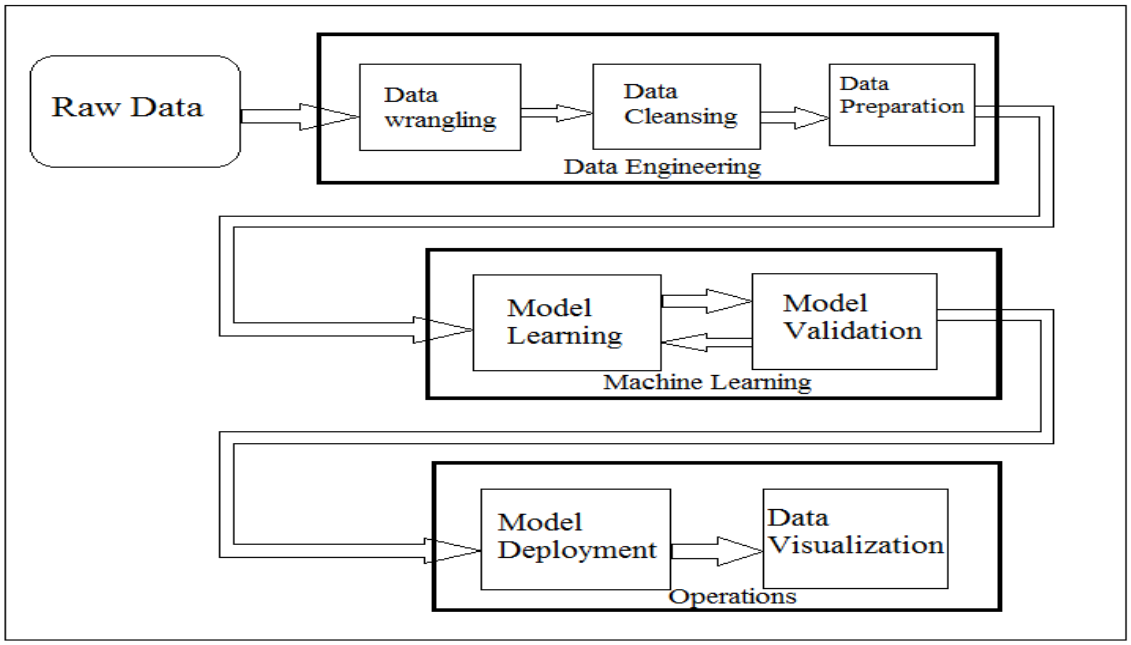

Fig.14: Machine Learning Implementation Overview

3.3.1 Dataset

A data set is a collection of data, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question.

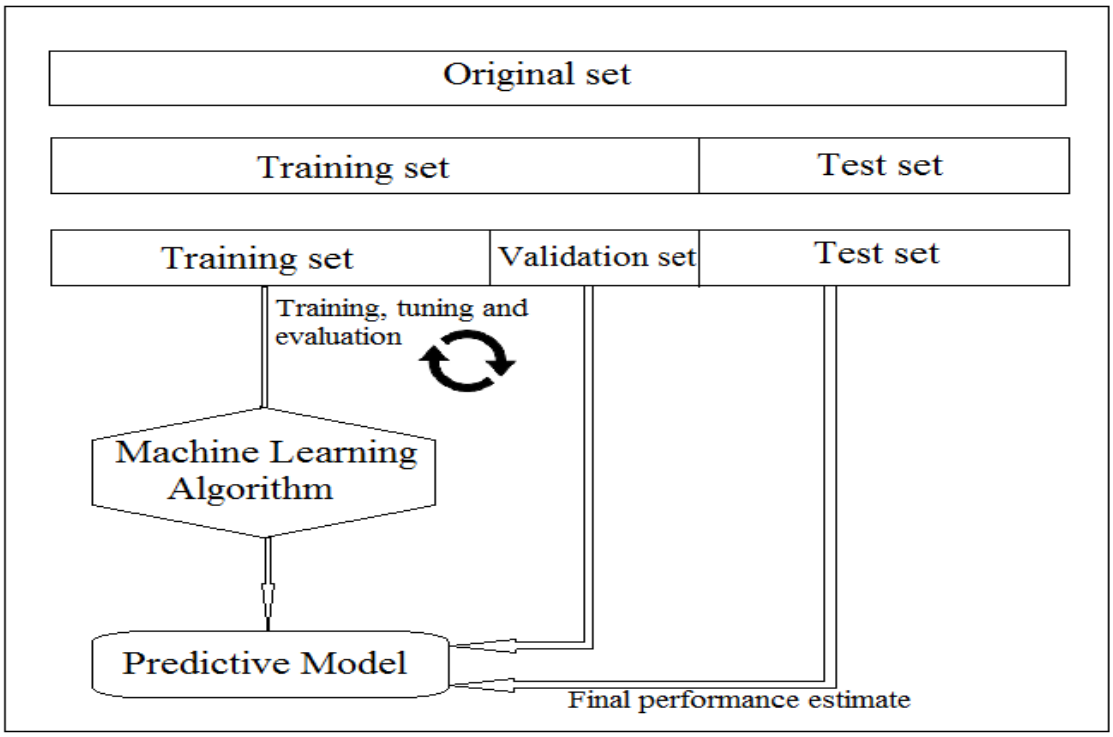

Fig.15 Dataset in ML

From training, tuning, model selection to testing, there are three different data sets: the training set, the validation set, and the testing set.

Training Data set: The training data set is the one used to train an algorithm to understand how to apply concepts such as neural networks, to learn and produce results. It includes both input data and the expected output. Training sets make up the majority of the total data, around 60 %.

Validation Data set: The Validation data set is used to select and tune the final ML model.

Test Data set: The test data set is used to evaluate how well the algorithm was trained with the training data set. Testing sets represent 20% of the data. The test set is ensured to be the input data grouped together with verified correct outputs, generally by human verification.

3.3.2 Data Pre-Processing

Pre-processing includes selection of the right data from the complete data set and building a training set. The process of putting together the data in this optimal format is known as feature transformation.

Data Pre-processing includes:

Data Wrangling: The data spread in different files, is gathered together to form a data set.

Data Cleansing: The main focus is to deal with missing values and remove unwanted characters from the data.

Data Preparation: In this step, main focus is on analysis and optimisation of the number of features. Features those are important for prediction are chosen so that there are faster computations and low memory consumption.

3.4 System Implementation

3.4.1 Model Description

- Registration - Registration is the first step before login. User need to register by giving the required information like user name email phone number and password.
- Login - After registration, user needs to login by submitting the username and password. Here user need to give username and password same as the registration form.

- Data Pre-processing - It is a data mining technique that transforms raw data into an understandable format. Raw data (real world data) is always incomplete and that data cannot be sent through a model. That would cause certain errors. That is why we need to pre-process data before sending data through a model.
- Model Fitting - By using three algorithms namely Naïve Bayes, Random forest, and KNN which predicts the live data using the training data set. Data analytical gives the live review and data visualization.
- Crop prediction - Using NPK, moisture, temperature and Ph attributes the three algorithms mainly Naïve Bayes, Random forest, and KNN which predicts the live data using the training data set. The algorithm is based on the accuracy. By this we predict the crop prediction.
- Fertilizer estimation prediction - Using NPK, moisture, temperature, Ph, and crop attributes the three algorithms mainly Naïve Bayes, Random forest, and KNN which predicts the live data using the training data set. The algorithm is based on the accuracy, by this we predict the fertilizer estimation.
- Yield Prediction - Using NPK, moisture, temperature, Ph, and crop attributes the three algorithms mainly Naïve Bayes, Random forest, and KNN which predicts the live data using the training data set. The algorithm is based on the accuracy. By this we predict the yield prediction.
- Comparative analysis - Three algorithms mainly Naïve Bayes, Random forest, and KNN we use all three algorithms and compare these three algorithms and use the algorithm based on the accuracy. By this we can predict the yield, crop prediction and fertilizer estimation.

3.5 System Testing

Software testing is the process of analyzing a software item to detect the differences between the existing and required conditions and to evaluate the features of software item. Software testing is an activity that should be done throughout the development process. Software testing is a task intended to detect defects in software by contrasting a computer program's expected results with its actual results for given set of inputs.

The aim of testing phase is to discover defects or errors by testing individual program components. During a system testing, these components are integrated to form a complete system. At this stage, testing was focused on establishing that the system met

its functional requirements, and does not behave in an unexpected way. Test data were inputs which had been devised to test the system and the outputs were predicted from these inputs if the system operates according to its specification. Testing was done to examine the behavior in a cohesive system. The test cases were selected to ensure that the system behavior can be examined in all possible combination of conditions.

Accordingly, the expected behavior of the system under different combinations was given. Therefore, test cases were selected which had inputs and the outputs were on expected lines. Inputs that were not valid and for which suitable messages had to be given and the inputs that did not occur frequently were regarded as special cases.

Test Environment

A testing environment is a setup of software and hardware on which the testing team is going to perform the testing of the newly built software product. This setup consists of the physical setup which includes hardware, and logical setup that includes Server Operating system, client operating system, database server, front end running environment, browser (if web application), or any other software components required to run this software product. This testing setup is to be built on both the ends.

Test Case

Set of test inputs, execution conditions, and expected results were developed for a particular objective, such as to exercise a particular program path or to verify compliance with a specific requirement. It included the following.

- Features to be tested
- Items to be tested
- Purpose of testing
- Pass/Fail criteria

Testing in Machine Learning

A Data Science/Machine Learning career has primarily been associated with building models that could do numerical or class-related predictions. This is unlike conventional software development, which is associated with both development and "testing" the software. And the related career profiles are software developer/engineers and test engineers/QA professionals. However, in the case of Machine Learning, the career profile is a data scientist. The usage of the word "testing" in relation to Machine Learning models is primarily used for testing the model performance in terms of accuracy/precision of the model. It can be noted that the word, "testing" means

different for conventional software development and Machine Learning models development. Hence as mentioned above the traditional unit/integration testing would not work on machine learning models hence it is tested based on its accuracy and prediction.

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

Accuracy=Number of correct predictions/Total number of predictions

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

Accuracy=TP+TN/TP+TN+FP+FN

Where *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.

When it comes to forecasting the models are evaluated based on the expected results they predict, In case of stock market forecasting, we have divided the data into training set and testing set again it is split into training dataset and validation dataset in the training set. We train our model using the training dataset and validation dataset is used to test the trained data. 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 hyper parameters. A test dataset is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset. If a model fit to the training dataset also fits the test dataset well. As you can see in the above graph dotted vertical line passing through the y axis is the point from which our prediction starts and the prices depicted in blue line is our predicted stocks values and the black line is the observed value. Hence by observing the predicted versus observed value we can tell how well our model works.

Chapter 4

HARDWARE AND SOFTWARE

4.1 Hardware Requirements

Hardware requirements specifications list the necessary hardware for the proper functioning of the project.

- System : Pentium IV 2.4 GHz.
- Hard Disk : 500 GB.
- Ram : 4 GB
- Any desktop / Laptop system with above configuration or higher level**.**

4.2 Software Requirements

Software requirements specification is a description of a software system to be developed, laying out functional and non-functional requirements, and may include a set of use cases that describe interactions the users will have the software.

- Operating system : Windows XP / 10
- Coding Language : Python
- Frame work : Flask
- IDE : VS code
- Database : MySQL

4.2.1 Python Introduction

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms.

The Python interpreter and the extensive standard library are freely available in source or binary form for all major platforms from the Python Web site and may be freely distributed. The same site also contains distributions of and pointers to many free third party Python modules, programs and tools, and additional documentation.

The Python interpreter is easily extended with new functions and data types implemented in C or C++ (or other languages callable from C). Python is also suitable as an extension language for customizable applications.

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

- Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

- Python is Interactive − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

- Python is Object-Oriented − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.

- Python is a Beginner's Language − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

Following are the features of python:

- Easy-to-learn − Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.

- Easy-to-read − Python code is more clearly defined and visible to the eyes.

- Easy-to-maintain − Python's source code is fairly easy-to-maintain.

- A broad standard library − Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.

- Interactive Mode − Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.

- Portable − Python can run on a wide variety of hardware platforms and has the same interface on all platforms.

- Extendable − you can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.

- Databases − Python provides interfaces to all major commercial databases.

- GUI Programming − Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.

- Scalable − Python provides a better structure and support for large programs than shell scripting.

4.2.2 MySQL

MySQL is a relational database management system, which organizes data in the form of tables. MySQL is one of many databases servers based on RDBMS model, which manages a seer of data that attends three specific things-data structures, data integrity and data manipulation. With MySQL cooperative server technology we can realize the benefits of open, relational systems for all the applications. MySQL makes efficient use of all systems resources, on all hardware architecture; to deliver unmatched performance, price performance and scalability. Any DBMS to be called as RDBMS has to satisfy Dr.E.F.Codd's rules.

Following are the Distinct Features of MySQL:

- MySQL is portable

  The MySQL RDBMS is available on wide range of platforms ranging from PCs to super computers and as a multi user loadable module for Novel NetWare, if you develop application on system you can run the same application on other systems without any modifications.

- MySQL is compatible

  MySQL commands can be used for communicating with IBM DB2 mainframe RDBMS that is different from MySQL , that is MySQL compatible with DB2 .MySQL RDBMS is a high performance fault tolerant DBMS , which is specially designed for online transaction processing and for handling large database applications.

- Multithreaded server architecture

  MySQL adaptable multithreaded server architecture delivers scalable high performance for very large number of users on all hardware architecture including symmetric multiprocessors (sumps) and loosely coupled multiprocessors. Performance is achieved by eliminating CPU, I/O, memory and operating system bottlenecks and by optimizing the MySQL DBMS server code to eliminate all internal bottlenecks.

Dr.E.F.OCDD's rules are as follows:

These rules are used for valuating a product to be called as relational database management systems. Out of 12 rules, a RDBMS product should satisfy at least 8 rules plus rule called rule 0 that must be satisfied.

RULE 0.FOUNDATION RULE

For any system that is to be advertised as, or claimed to be relational DBMS. That system should manage database with in self, without using an external language.

RULE 1.INFORMATION RULE

All information in relational database is represented at logical level in only one way as values in tables.

RULE 2.GUARANTEED ACCESS

Each and every data in a relational database is guaranteed to be logically accessibility by using to a combination of table name, primary key value and column name.

RULE 3.SYSTEMATIC TREATMENT OF NULL VALUE

Null values are supported for representing missing information and inapplicable information. They must be handled in systematic way, independent of data types.

RULE 4.DYNAMIC ONLINE CATALOG BASED RELATION MODEL

The database description is represented at the logical level in the same way as ordinary data so that authorized users can apply the same relational language to its interrogation as they do to the regular data.

RULE 5.COMPRHENSIVE DATA SUB LANGUAGE

A relational system may support several languages and various models of terminal use. However there must be one language whose statement can express all of the following: Data Definitions, View Definitions, Data Manipulations, Integrity, Constraints, Authorization and transaction boundaries.

RULE 6.VIEW UPDATING

All views that are theoretically updatable are also updatable by the system.

RULE 7.HIGH LEVEL UPDATE, INSERT and DELETE

The capability of handling a base relational or derived relational as a single operand applies not only retrieval of data also to its insertion, updating, and deletion.

RULE 8.PHYSICAL DATA INDEPENDENCE

Application program and terminal activities remain logically unimpaired whenever any changes are made in either storage representation or access method.

RULE 9.LOGICAL DATA INDEPENDENCE

Application programs and terminal activities remain logically unimpaired whenever any changes are made in either storage repres3entation or access methods.

RULE 10: INTEGRITY INDEPENDENCE:

Integrity constraints specific to particular database must be definable in the relational data stored in the catalog, not in application program.

RULE 11: DISTRIBUTED INDEPENDENCE:

Whether or not system support data base distribution, it must have a data sub-language that can support distributed databases without changing the application program.

RULE 12: NON SUB-VERSION:

If a relational system has low level language, that low language cannot use to subversion or by pass the integrity rules and constraints expressed in the higher level relational language.

4.2.3 Flask

Flask is a web framework. This means flask provides you with tools, libraries and technologies that allow you to build a web application. This web application can be some web pages, a blog, a wiki or go as big as a web-based calendar application or a commercial website.

Flask is part of the categories of the micro-framework. Micro-framework are normally framework with little to no dependencies to external libraries. This has pros and cons. Pros would be that the framework is light, there are little dependency to update and watch for security bugs, cons is that some time you will have to do more work by yourself or increase yourself the list of dependencies by adding plugins. In the case of Flask, its dependencies are: Werkzeug a WSGI utility library and jinja2 which is its template engine.

In a given web application, you may want to be able to express relationships between objects. In the To-Do List example, users own multiple tasks, and each task is owned by only one user. This is an example of a "many-to-one" relationship, also known as a foreign key relationship, where the tasks are the "many" and the user owning those tasks is the "one."

In Flask, a many-to-one relationship can be specified using the db.relationship function.

4.2.4 Visual studio code

Visual Studio Code is a source-code editor that can be used with a variety of programming languages, including Java, JavaScript, Go, Node.js and C++. It is based on the Electron framework, which is used to develop Node.js Web applications that run on the Blink layout engine. Visual Studio Code employs the same editor component (codenamed "Monaco") used in Azure DevOps (formerly called Visual Studio Online and Visual Studio Team Services).Instead of a project system, it allows users to open one or more directories, which can then be saved in workspaces for future reuse. This allows it to operate as a language-agnostic code editor for any language. It supports a number of programming languages and a set of features that differs per language. Unwanted files and folders can be excluded from the project tree via the settings. Many Visual Studio Code features are not exposed through menus or the user interface, but can be accessed via the command palette.

Visual Studio Code can be extended via extensions, available through a central repository. This includes additions to the editor and language support. A notable feature is the ability to create extensions that add support for new languages, themes, and debuggers, perform static code analysis, and add code linters using the Language Server Protocol.

Visual Studio Code includes multiple extensions for FTP, allowing the software to be used as a free alternative for web development. Code can be synced between the editor and the server, without downloading any extra software.

Visual Studio Code allows users to set the code page in which the active document is saved, the newline character, and the programming language of the active document. This allows it to be used on any platform, in any locale, and for any given programming language.

Language support

Out-of-the-box, Visual Studio Code includes basic support for most common programming languages. This basic support includes syntax highlighting, bracket matching, code folding, and configurable snippets. Visual Studio Code also ships with IntelliSense for JavaScript, TypeScript, JSON, CSS, and HTML, as well as debugging support for Node.js. Support for additional languages can be provided by freely available extensions on the VS Code Marketplace.

Data collection

Visual Studio Code collects usage data and sends it to Microsoft, although this can be disabled. In addition, because of the open-source nature of the application, the telemetry code is accessible to the public, who can see exactly what is collected. According to Microsoft, the data is shared with Microsoft-controlled affiliates and subsidiaries, although law enforcement may request it as part of a legal process.

Chapter 5

RESULTS

The Web-application is as shown in Figures.

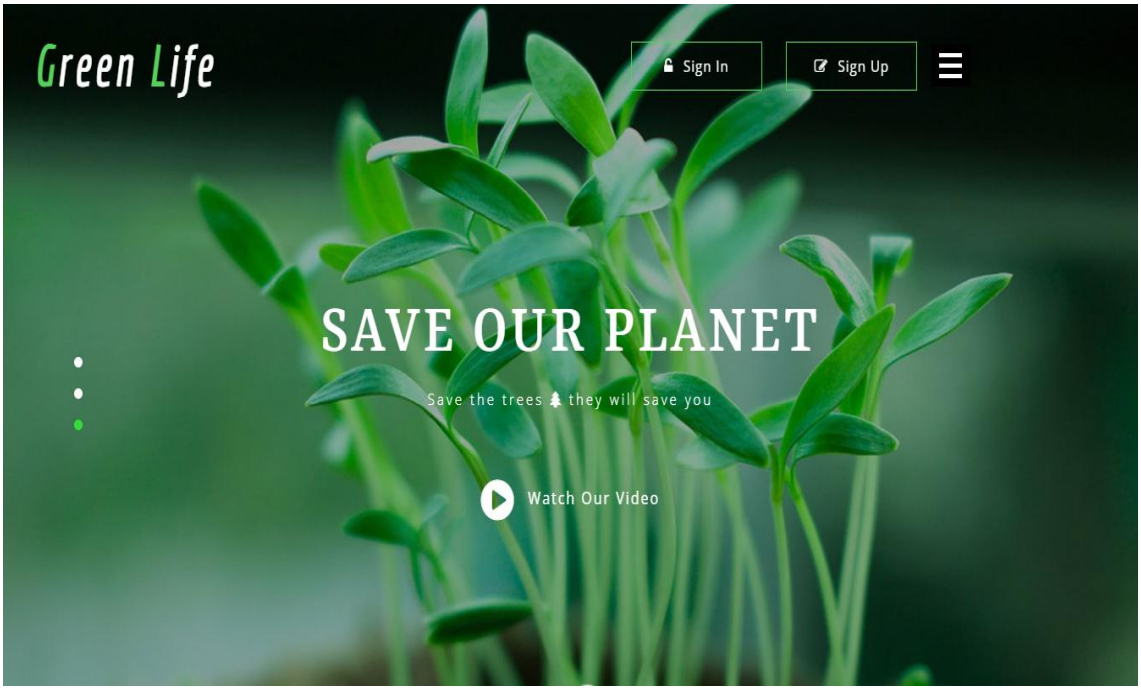

Fig.16: Web-Application

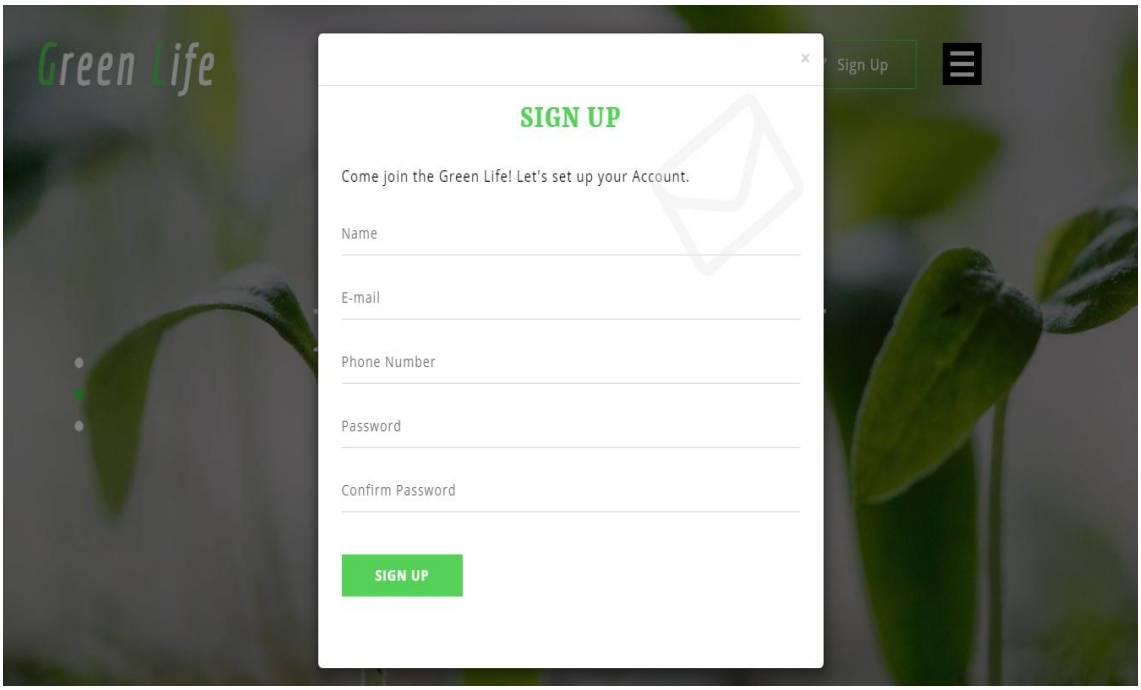

Fig.17: Web-Application, Sign Up box

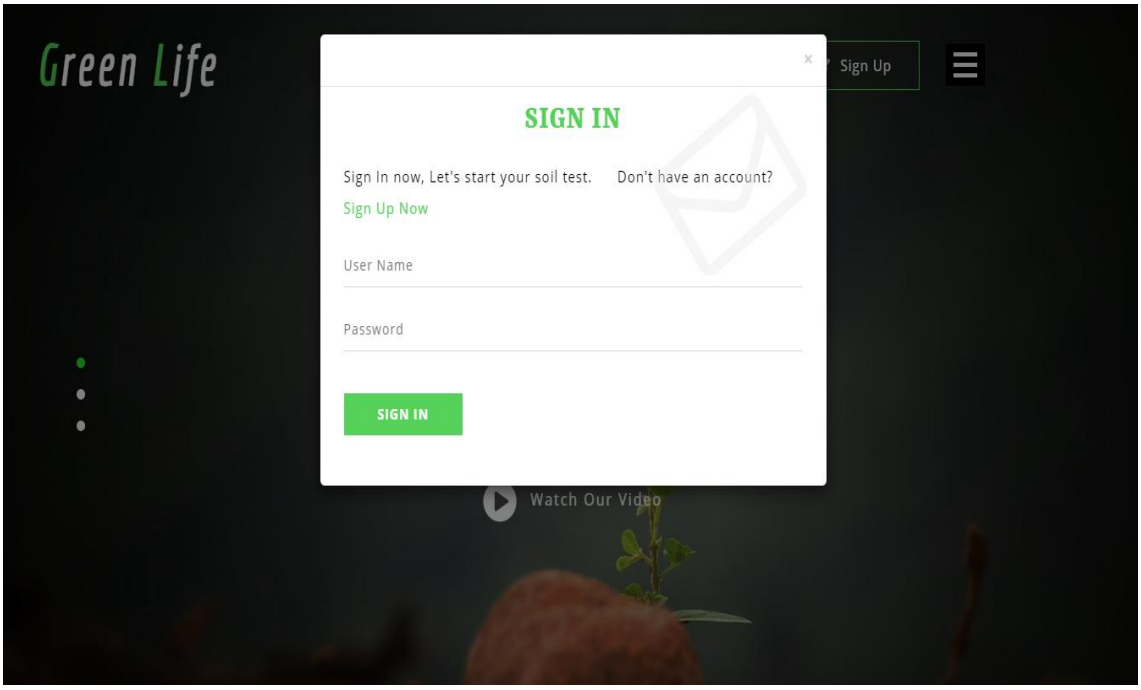

Fig.18: Web-Application, Sign In box

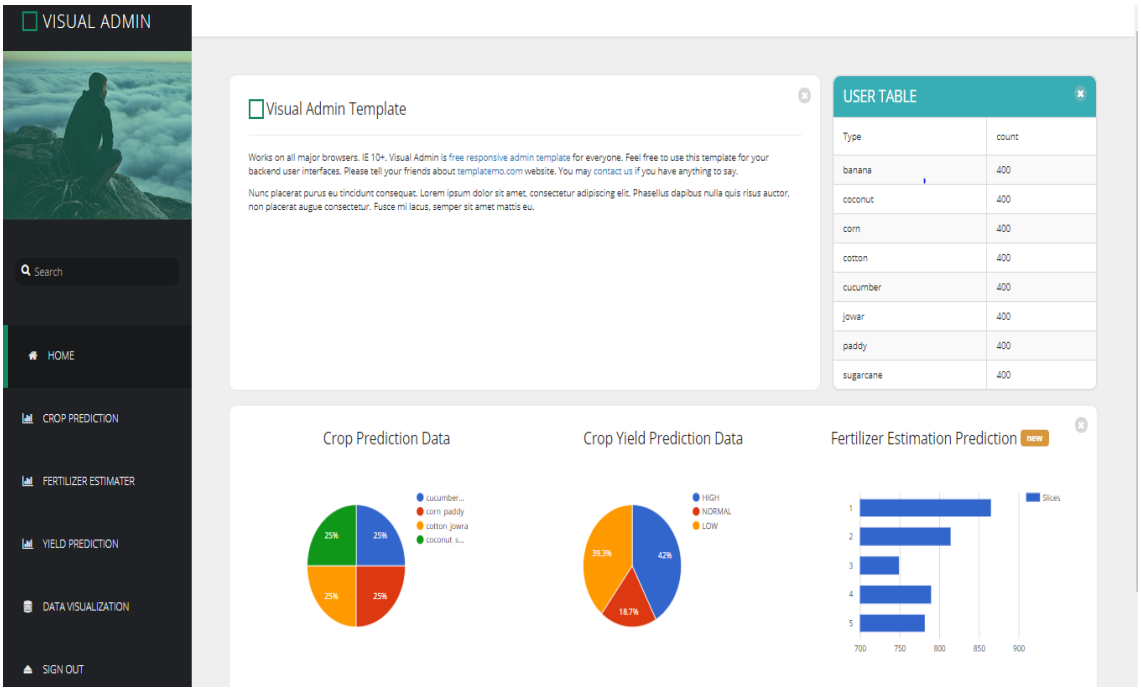

Fig.19: Web-Application Home page

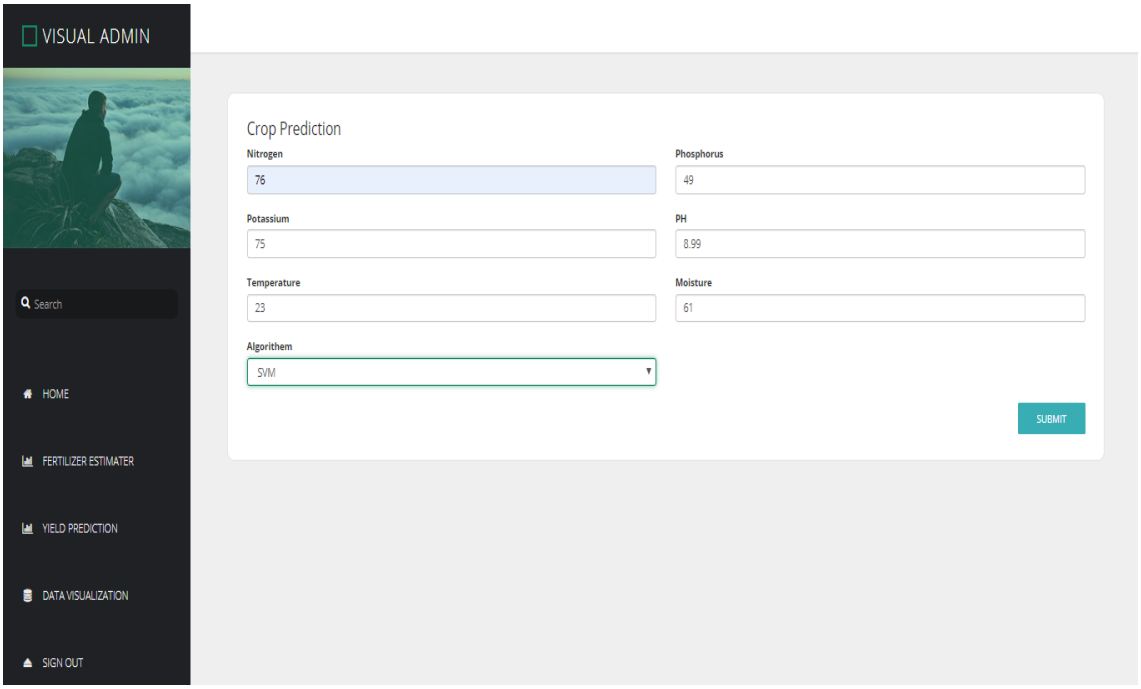

Fig.20: Web-Application, Crop Prediction

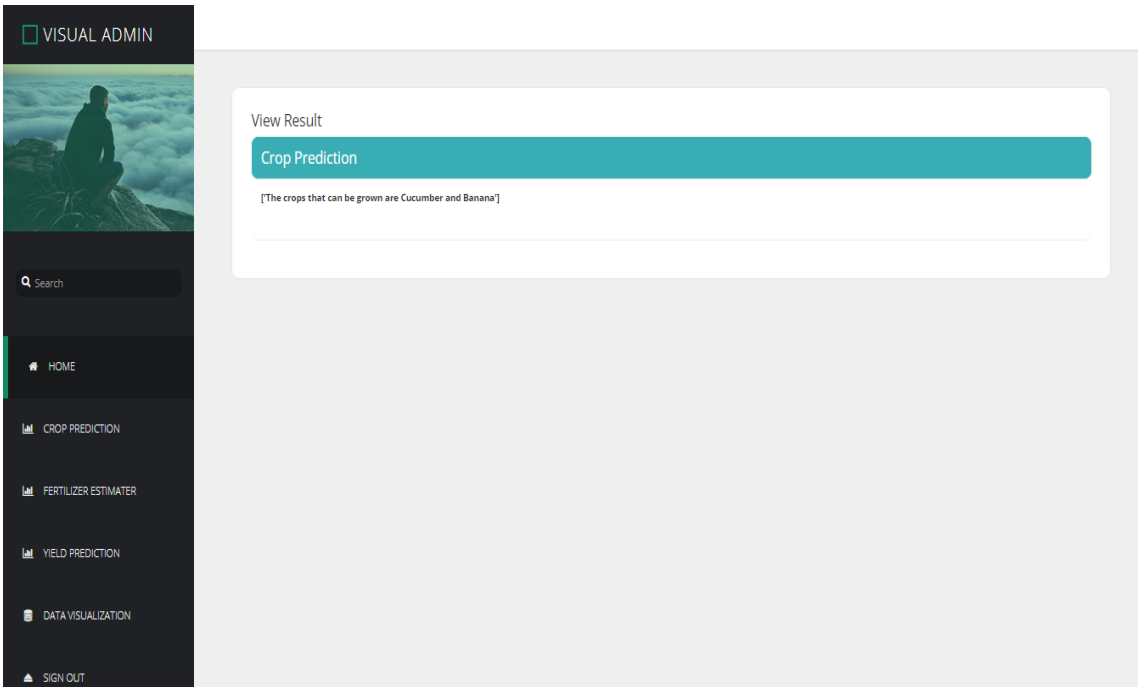

Fig.21: Web-Application, Crop Prediction Result

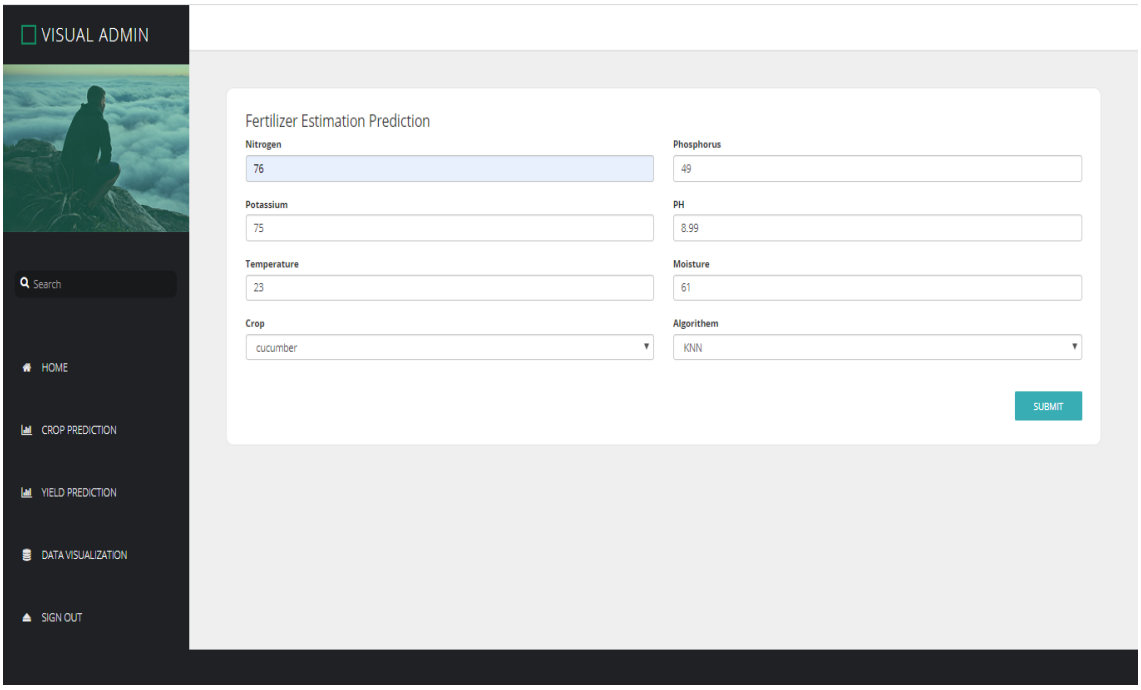

Fig.22: Web-Application, Fertilizer Estimation

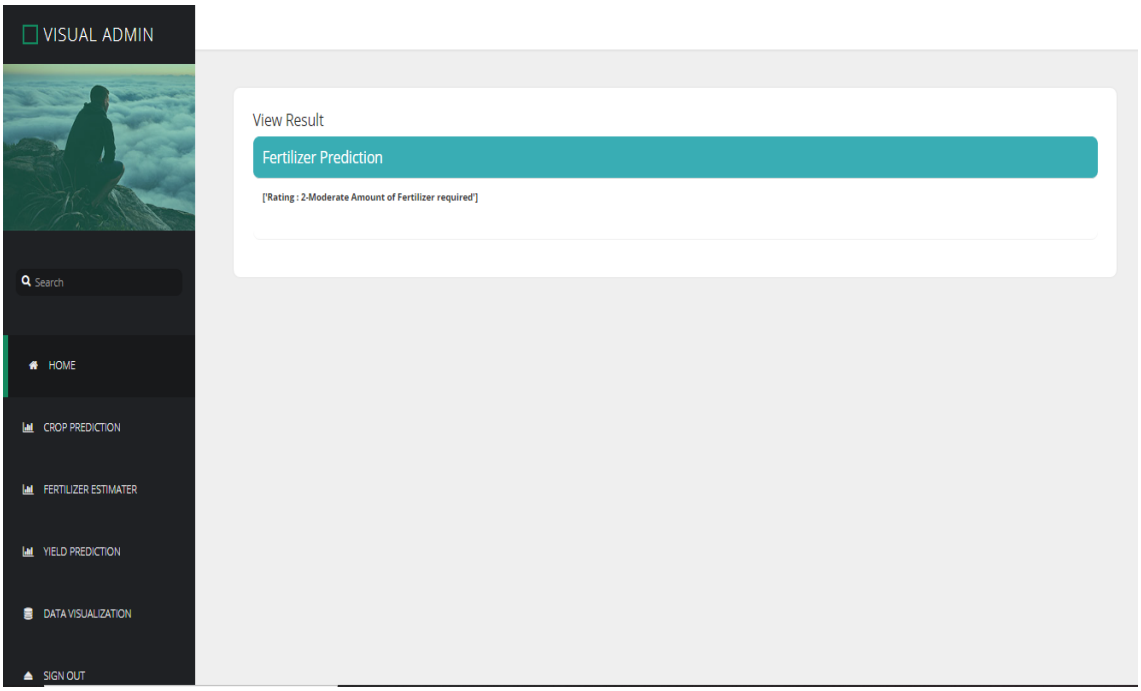

Fig.23: Web-Application, Fertilizer Estimation Result

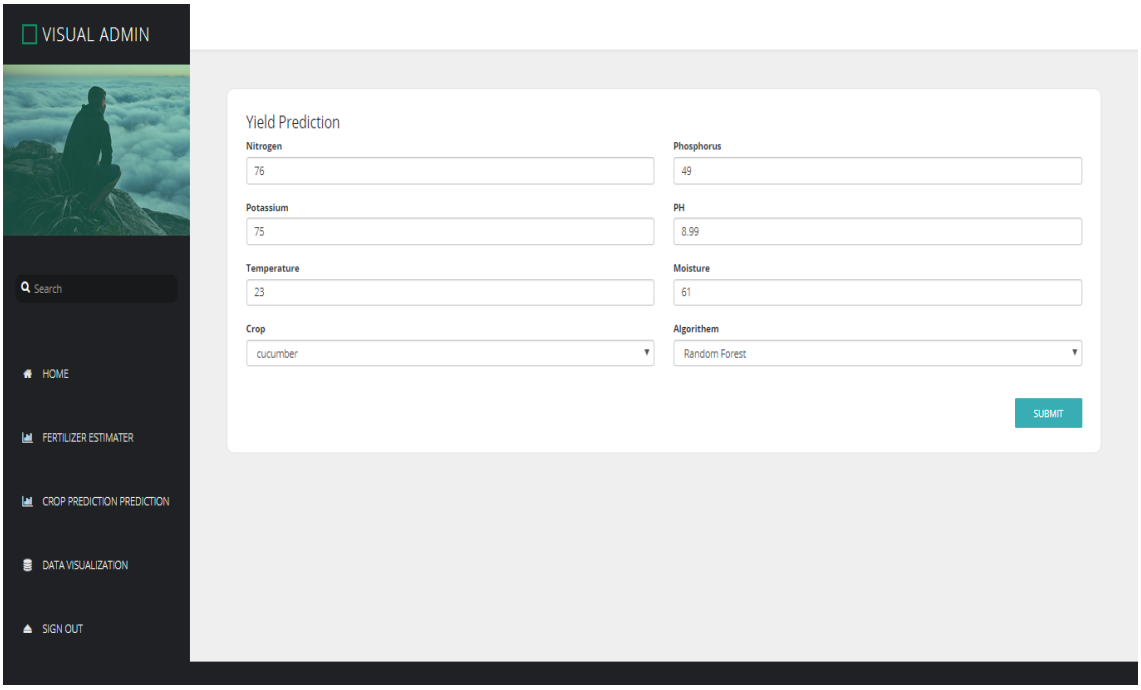

Fig.24: Web-Application, Yield Prediction

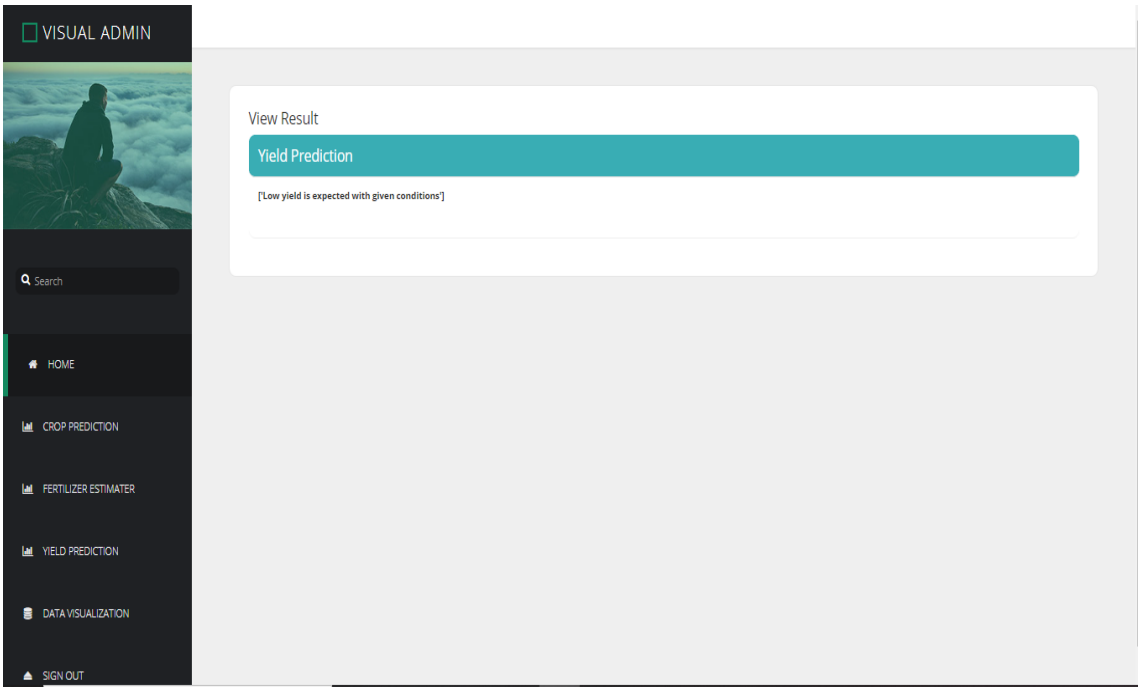

Fig.25: Web-Application, Yield Prediction Result

[1] Eight crops cucumber, banana, corn, paddy, cotton, jowar, coconut, sugarcane are considered. There are around 4000 data in the training dataset with 500 data for each crop.

[2] For any given values of N, P, K, pH, temperature and moisture content, two suitable crops, approximate yield and the approximate amount of fertilizer needed are being predicted with the assistance of any one chosen algorithm among the three accessible algorithms.

It is seen that for N=73, P=50, K=74, pH=5.76, Moisture=62, Temperature=24, and the algorithm selected being K-NN, the crop suggested is cucumber or banana. For the same values of N, P, K, pH, moisture and temperature when the crop chosen is Jowar and selected algorithm is SVM, the fertilizer estimation is "2-moderate amount of fertilizer". For the same values of N, P, K, pH, moisture and temperature when the crop chosen is paddy and selected algorithm is Random Forest the yield prediction is "low amount of yield in the specified conditions".

[3] Implementation of three algorithms facilitated the comparative analysis of their performance.

The accuracy of the SVM, K-NN, Random Forest algorithms in crop prediction, fertilizer estimation and yield prediction is as shown in Table 1.

	SVM	KNN	Random Forest
Crop Prediction	100	100	100
Fertilizer Estimation	100	70	98
Yield Prediction	91.5	94.4	97.72

Table 1: Comparative Analysis of Algorithms

All the three algorithms work equally well for crop prediction, SVM is found to have higher accuracy in fertilizer estimation and Random forest has higher accuracy in yield prediction.

Table 2 shows few examples of the application prediction for various input values:

N	P	K	pH	Moisture	Temperature	Crop Predicted (K-NN)	Crop Chosen	Fertilizer Estimation (SVM)	Crop Chosen	Yield Prediction (RF)
73	50	74	5.76	62	24	cucumber banana	jowar	2-moderate	paddy	low
49	47	75	7.11	21	13	corn paddy	corn	1-least	corn	high
38	41	76	8.16	82	29	cotton jowar	sugarcane	4-slightly more	cucumber	low
71	109	75	8.28	34	33	coconut sugarcane	banana	5-large	sugarcane	moderate
58	57	73	6.76	42	20	corn paddy	coconut	1-least	jowar	high

Table 2: Sample result

Chapter 6

APPLICATIONS AND ADVANTAGES

Proposed system considers N, P, K, pH, temperature and moisture values which play a vital role in determining the soil characteristic for suitable crop prediction, fertilizer estimation and yield prediction. Machine Learning techniques are proved to be efficient and accurate over other methods of prediction. Proposed system involves implementation of data mining techniques for prediction which is an added advantage. Manual analysis is not required and all records will be systematically stored and can be easily accessed. Hence the proposed system is a useful tool to the farmers. It can also be used by the agriculture department.

Chapter 7

CONCLUSIONS AND SCOPE FOR FUTURE WORK

A web application which uses soil parameters such as N, P, K, pH, Moisture, Temperature values to predict the suitable crop and estimates the amount of fertilizer, to predict the approximate yield considering the soil parameters along with the crop using machine learning techniques is proposed. Use of right soil parameters and machine learning algorithms such as K-Nearest Neighbor, Support Vector Machine, Random Forest makes the application more reliable and accurate. Providing comparative analysis helps in using the right algorithm for accurate prediction and estimation. For the dataset considered all the three algorithms worked similarly well for crop prediction, SVM is found to have higher accuracy in fertilizer estimation and Random forest has higher precision in yield prediction.

Later on, by expanding the quantity of information in the dataset, by considering more crops and furthermore by including other soil parameters, better outcome can be expected. Including sensors and an IoT model to gauge the soil parameters in the field can be the one stop answer for testing soil fertility and anticipating the yield.

REFERENCES

[1] Alexandros Kaloxylos, Robert Eigenmann, Frederick Teye,"Farm management systems and the Future Internet era", Kaloxylos et al. / Computers and Electronics in Agriculture 89 (2012) 130–144.

[2] D S Suresh, Jyothi Prakash K V & Rajendra, "Automated Soil Testing Device", (ITSI-TEEE)2320-8945,Volume 1,Issue 5,2013.

[3] DhareshVadalia, Minal Vaity, Krutika Tawate, Dynaneshwar Kapse, "Real Time soil fertility analyzer and crop prediction", IRJET,Volume 4,Issue 3,March 2017.

[4] Apurva C. Pusatkar, Vijay S. Gulhane, "Implementation Of Wireless Sensor Network For Real Time Monitoring Of Agriculture", IRJET, Volume 3,Issue 5,May 2016.

[5] Prof.D.S.Zingade, Omkar Buchade, Nilesh Mehta, Shubham Ghodekar, Chandan Mehta, "Crop Prediction System using Machine Learning", IJETCS, Volume: 3 Issue: 2 April – 2018.

[6] J.Jayaprahas, S.Sivachandran, K.Navin, K.Balakrishnan, "Real Time Embedded Based Soil Analyzer", International Research Journal of Engineering and Technology (IRJET), Volume: 3 Issue 3, March 2014.

[7] Ashwini Rao, Janhavi U, Abhishek Gowda N S, Manjunatha, Mrs.Rafega Beham, "Machine Learning in Soil Classification and Crop Detection", IJSRD - International Journal for Scientific Research & Development, Vol. 4, Issue 01, 2016.

[8] Surili Agarwal, Neha Bhangale, Kameya Dhanure, Shreeya Gavhane, V.A.Chakkarwar, Dr. M.B.Nagori, "Application of Colorimetry to determine Soil Fertility through Naive Bayes Classification Algorithm", 9th ICCCNT, July 10-12, 2018, IISC, Bengaluru, India.

[9] Yanxin Zhu , Di Wu and Sujian Li, "Cloud Computing and Agricultural Development of China", IJCSI Issues,Vol 10,Issue 1,No 1,January 2013.

[10] Rakesh Patel , Mili. Patel, "Application of Cloud Computing in Agricultural Development of Rural India", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013, 922-926.

[11] Seena Kalghatgi, Kuldeep P. Sambrekar,"Using Cloud Computing Technology in Agricultural Development", IJISET Vol. 2 Issue 3, March 2015.

[12] Utkarsha P. Narkhede, K.P.Adhiya, "Evaluation of Modified K-Means Clustering Algorithm in Crop Prediction", International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-4 Number-3 Issue-16 September-2014.

[13] Miss.Snehal S.Dahikar, Dr.Sandeep V.Rode, "Agricultural Crop Yield Prediction Using Artificial Neural Network Approach", International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, Vol. 2, Issue 1, January 2014.

[14] Meenakshi Malik, Mamta Bansal, R.P Agarwal,A.K Kanojia R.V.Singh, "Crop selection Algorithm-Technique for Price prediction", International Journal of Research in Economics and Social Sciences (IJRESS)**,** Vol. 7 Issue 3, March- 2017.

[15] Md.Tahmid Shakoor, Karishma Rahman, Sumaiya Nasrin Rayta, "Intelligent Agriculture Information Monitoring Using data mining techniques", BRAC University, 18th April, 2017**.**