VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belgaum-590018

A PROJECT REPORT **(15CSP85)** ON

"A MODEL TO PREDICT THE CROP BASED ON SOIL COMPOSITION USING MACHINE LEARNING TECHNIQUES"

Submitted in Partial fulfillment of the Requirements for the Degree of

Bachelor of Engineering in Computer Science & Engineering

By

PRAVALLIKA A.S.K.S (1CR17CS409)

SANIYA (1CR16CS147)

KOUSALYA R (1CR17CS405)

Under the Guidance of,

Mrs. RUBINI P.E.

Assistant Professor, Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR INSTITUTE OF TECHNOLOGY

CMR INSTITUTE OF TECHNOLOGY

#132, AECS LAYOUT, IT PARK ROAD, KUNDALAHALLI, BANGALORE-560037

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project work entitled **"A Model to predict the crop based on soil composition using Machine Learning Techniques"** carried out by **Ms. PRAVALLIKA A.S.K.S**, USN **1CR17CS409**, **Ms. SANIYA**, USN **1CR16CS147**, **Ms. KOUSALYA R**, USN **1CR17CS405,** bonafide students of CMR Institute of Technology, in partial fulfillment for the award of **Bachelor of Engineering** in Computer Science and Engineering of the Visveswaraiah Technological University, Belgaum during the year 2019-2020. 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.

(Mrs. Rubini P.E)	**Dr. Prem Kumar Ramesh**	**Dr. Sanjay Jain**
(Designation)	**Professor & Head**	**Principal**
Dept. of CSE, CMRIT	**Dept. of CSE, CMRIT**	**CMRIT**

External Viva

Name of the examiners — Signature with date

1. _____

2. _____

DECLARATION

We, the students of Computer Science and Engineering, CMR Institute of Technology, Bangalore declare that the work entitled "**A MODEL TO PREDICT THE CROP BASED ON SOIL COMPOSITION USING MACHINE LEARNING TECHNIQUES**" has been successfully completed under the guidance of Prof. Rubini P.E., Computer Science and Engineering Department, CMR Institute of technology, Bangalore. This dissertation work is submitted in partial fulfillment of the requirements for the award of Degree of Bachelor of Engineering in Computer Science and Engineering during the academic year 2019 - 2020. Further the matter embodied in the project report has not been submitted previously by anybody for the award of any degree or diploma to any university.

Place:

Date:

Team members:

PRAVALLIKA A.S.K.S (1CR17CS409) _____

SANIYA (1CR16CS147) _____

KOUSALYA R (1CR17CS405) _____

ABSTRACT

Agriculture planning plays a significant role in economic growth and food security of agro-based country. And still agriculture is being plagued by various problems. Farming practices and other activities of agriculture consume time, money as well as the efforts of a farmer. Farmers face a lot of issues with respect to the selection of the crop to be grown. If the wrong crop is grown then farmers have to deal with financial crisis. To feed the world population, an essential development in the agriculture production is important. Selection of crop is an important issue for agriculture planning. To overcome this problem, many researchers studied prediction of crop, soil classification and crop classification for agriculture planning using statistics methods or data mining techniques. But prediction of the appropriate crop for a soil is still a puzzle. There are many emerging techniques; the most prominent one is the machine learning technique. In this proposed work, the aim is to predict the right crop and appropriate fertilizer by analyzing the composition and other properties of the soil. For this machine learning algorithm is used, which will overcome the existing problems and increase the accuracy rate. Crop Prediction helps farmers in selecting proper crop for plantation and thus maximizing the yield rate and their earnings.

ACKNOWLEDGEMENT

We take this opportunity to express my sincere gratitude and respect to **CMR Institute of Technology, Bengaluru** for providing us a platform to pursue our studies and carry out our final year project

We have a great pleasure in expressing our deep sense of gratitude to **Dr. Sanjay Jain,** Principal, CMRIT, Bangalore, for his constant encouragement.

We would like to thank **Dr. Prem Kumar Ramesh,** Professor and Head, Department of Computer Science and Engineering, CMRIT, Bangalore, who has been a constant support and encouragement throughout the course of this project.

We consider it a privilege and honor to express our sincere gratitude to our guide **Mrs. Rubini P.E., Assistant Professor,** Department of Computer Science and Engineering, for the valuable guidance throughout the tenure of this review.

We also extend our thanks to all the faculty of Computer Science and Engineering who directly or indirectly encouraged us.

Finally, We would like to thank our parents and friends for all their moral support they have given us during the completion of this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CSM	**Crop Selection Method**
KNN	**K- Nearest Neighbor**
SVM	**Support Vector Machine**

CHAPTER 1

INTRODUCTION

This chapter gives a description about machine learning, tasks involved, models, advantages, disadvantages, objective of the proposed work and thesis overview.

India is the leading producer of few crops. Predominant occupation in India is agriculture. The soil quality of the agriculture includes the soil properties those related to organic matter such as N(Nitrogen), C(Carbon), Ph(Phosphorus), Mg(Magnesium), Ca(Calcium) and K(Potassium).

The main aim is to analyse and predict the crop using soil properties parameters. In this application the farmer provides the soil composition i.e., properties of the soil then the supervised machine learning algorithm will analyse and predict the right information as in what crops has to be grown with their appropriate fertilizer details.

1.1 Machine Learning

Machine Learning is a field of Computer Science, where new developments evolve at recent times, and also helps in automating the evaluation and processing done by the mankind, thus reducing the burden on the manual human power. According to techtarget, Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed.

Machine learning focuses on the development of computer programs that can change when exposed to new data. Finding out the suitable crops based on the soil's appearance becomes tedious for novice farmers. There also exists a need to prevent the agricultural decay.

1.1.1 Types of Problems and Tasks

Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available to a learning system. These are:

- ♣ Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs
- ♣ Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).
- ♣ Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal. Another example is learning to play a game by playing against an opponent.

Fig 1.1 represents the different tasks/types of Machine learning methods and their applications.

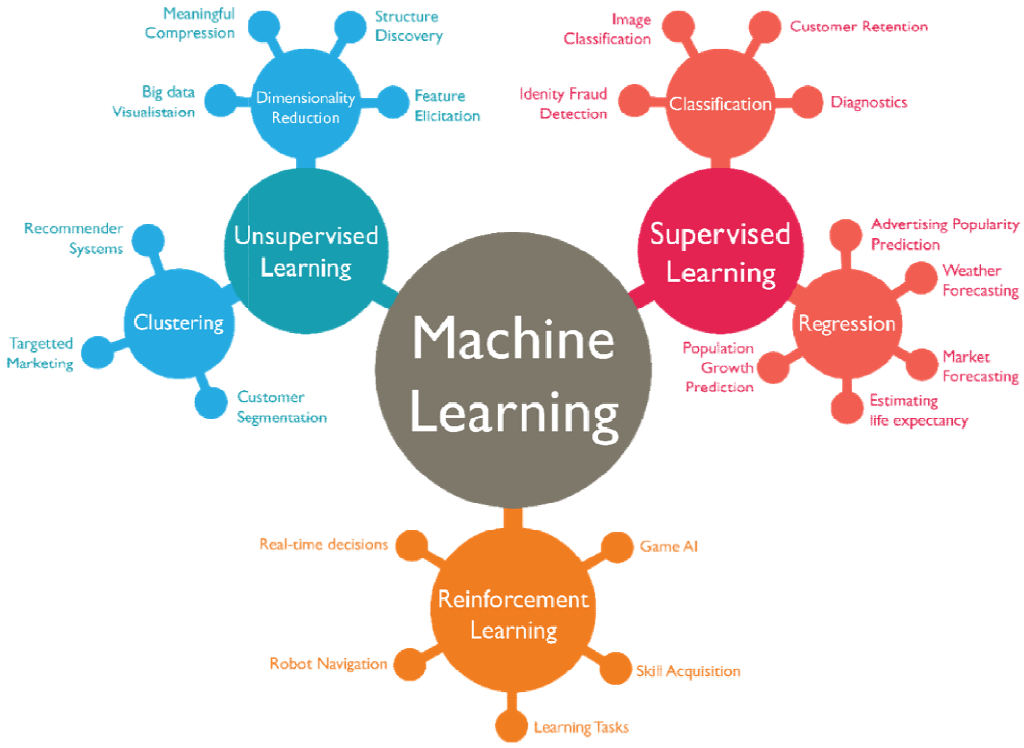

Fig 1.1: Machine Learning Tasks

1.2 Relevance of the Project

Indian agriculture is being plagued by various problems. The biggest problem with Indian farmers is the crop selection. The relevance of the ongoing project with the current research papers is that the crop prediction and selection of the right crop will help the farmer to increase their earnings.

1.3 Scope of the Project

The desired results from the proposed work can be met only when the right soil composition is provided by the farmer. The constraint in the proposed work is that farmer has to externally get the soil sample for the soil composition tested and provide the test results to get the details about the right crop to be predicted.

1.4 Chapter Wise Summary

In this chapter, the introduction to machine learning techniques is given. The importance of agriculture and farmer is also discussed in the above sections.

CHAPTER 2

LITERATURE SURVEY

The proposed work's main challenge is to get the soil composition which was already carried out by existing work. Those literature works will be discussed in the following sub-section.

2.1 An Efficient Analysis of Crop Yield Prediction

Authors: Shriya Sahu, Meenu Chawla and Nilay Khare

Title: An Efficient Analysis of Crop Yield Prediction using Hadoop Framework Based on Random Forest Approach

Published in: IEEE (2017)

Methodology: In this paper, various parameters are considered from soil to atmosphere for predicting the suitable crop. Soil parameters such as type, ph level, iron, copper, manganese, sulphur, organic carbon, potassium, phosphate, nitrogen are considered. The random forest algorithm is used to classify the dataset which provides result in good accuracy with poor error rate. Since this framework can handle large dataset by processing it in MapReduce programming model. The phases of the proposed work are: Data Collection, Data Classification(Random Forest Algorithm), Hadoop Framework – MapReduce programming model and Final Prediction. The implementation is carried out in ubuntu 14.04 LTS with Hadoop 2.6.0 and the dataset is collected from various online sources to predict the suitable crop.

Accuracy: The accuracy achieved by this methodology is 91.43%.

Future Work: The future research will be devoted to predicting the required fertilizer and pesticide ratio based on the atmospheric and soil parameters for the farm land.

2.2 Crop Selection Method to Maximize Crop yield rate

Authors: Rakesh Kumar, M.P. Singh, Prabhat Kumar and J.P. Singh.

Title: Crop Selection Method to Maximize Crop yield rate using Machine Learning Technique

Published in: IEEE (2015)

Methodology: This work presents a technique named CSM to select sequence of crops to be planted over season. CSM method may improve net yield rate of crops to be planted over season. The proposed method resolves selection of crop (s) based on prediction yield rate influenced by parameters (e.g. weather, soil type, water density, crop type). The crop sowing table data considered are gathered from farmer of Patna District, Bihar (India). It takes crop, their sowing time, plantation days and predicted yield rate for the season as input and finds a sequence of crops whose production per day are maximum over season.

Accuracy: Performance and accuracy of CSM method depends on predicted value of influenced parameters.

Future Work: There is need to adopt a prediction method with more accuracy and high performance.

2.3 Analysis of Soil Behaviour and Prediction of Crop Yield

Authors: Monali Paul, Santosh K. Vishwakarma and Ashok Verma

Title: Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach

Published in: IEEE (2015)

Methodology: In this work the experiments are performed using RapidMiner 5.3. Two important and well known classification algorithms K-Nearest Neighbor (KNN) and Naive Bayes (NB) are applied to the soil dataset which is taken from the soil testing laboratory Jabalpur, M.P. And classification of soil into low, medium and high categories are done in order to predict the crop yield using available dataset. This

study can help the soil analysts and farmers to decide sowing in which land may result in better crop production.

Accuracy: Accuracy is different from both classification methods. By experimenting with KNN algorithm, 30 lands are classified as Low category soil, 45 lands of Medium category soil and 25 lands of High category soil. By experimenting with Naive Bayes algorithm, 15 lands are classified as Low category soil, 40 under Medium category soil and 45 lands of High category soil.

Future Work: In this work, a small dataset is used due to occurrence of some complexities. Thus a larger dataset of 1GB or more may be used in the later work.

2.4 Comparison

	Methods used	Advantages	Disadvantages	Dataset	Result
Approach 1	Random forest,Hadoop framework and MapReduce Programming model.	Large dataset can be easily stored and classified	The performance and accuracy can be increased by reducing work on the classification method.	The dataset is gathered in two ways: Online data and sensor node data.	The proposed work, gives better accuracy and works faster.
Approach 2	CSM method, decision tree, Artificial Neural Networks, Regression.	Sequence of crops that can be grown on the same land is predicted.	If the prediction is not wisely and correctly choosen then there are chances of predicting the wrong sequence of crops.	The crop sowing table, data are gathered from Patna District,Bihar.	The proposed method resolve selection of crops based on yield rate.
Approach 3	Data mining Methods (Naïve Bayes and K-Nearest Neighbors).	Patterns are discovered from the large data set (data regarding the previous years yield, weather etc).	This approach uses small dataset due to some complexities.	The soil dataset is taken from Soil Testing laboratory Jabalpur, MP.	The proposed method helps farmers to decide sowing in which land may result in better crop production

Table 2.1 Comparison of Approaches

CHAPTER 3

SYSTEM REQUIREMENTS SPECIFICATION

As a conglomeration of math, data science and software engineering, for working and understanding of machine learning algorithms three fields are required to be proficient with. They are: Probability, Statistics and Data Modelling.

3.1 Hardware Requirements

- Intel core i3 or more
- Hard disk: 256GB
- RAM: 4GB
- 1-2 NVidia Cards

3.2 Software Requirements

- Operating System: Windows 8 and above.
- Anaconda Navigator
- Python libraries: sklearn, pandas
- Python framework: Flask
- PyCharm

CHAPTER 4

SYSTEM ANALYSIS AND DESIGN

System design is the process of the defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements .Systems design could be seen as the application of systems theory to product development. Object-oriented analysis and methods are becoming the most widely used methods for computer systems design. Systems design is therefore the process of defining and developing systems to satisfy specified requirements of the user. The UML has become the standard language in object oriented analysis and design.

4.1 Architectural Design

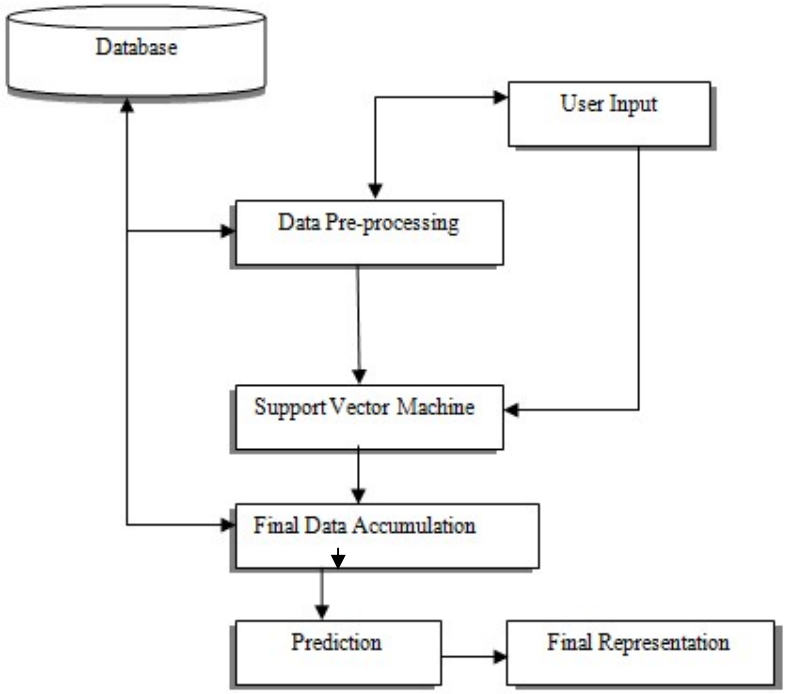

Fig 4.1 Architectural Design

Fig 4.1 represents the architectural design of the proposed work. System architecture is a conceptual model that defines the structure and behaviour of the system. It comprises of the system components and the relationship describing how they work together to implement the overall system.

4.2 Dataflow Diagram

A dataflow diagram is a graphical representation of the "flow" of data through an information system, modelling its process aspects. A DFD is often used as a preliminary step to create an overview of the system without going into great detail, which can later be elaborated. DFDs can also be used for the visualization of data processing. A DFD shows what kind of information will be input to and output from the system, how the data will advance through the system, and where the data will be stored. Fig 4.2 represents the data flow between datasets, user and prediction model and Fig 4.3 shows different steps that are carried out during the process of prediction.

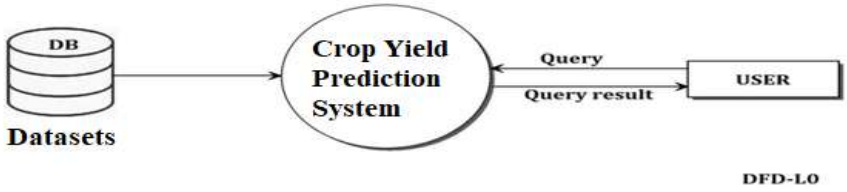

Fig 4.2 Data Flow Diagram (Level zero)

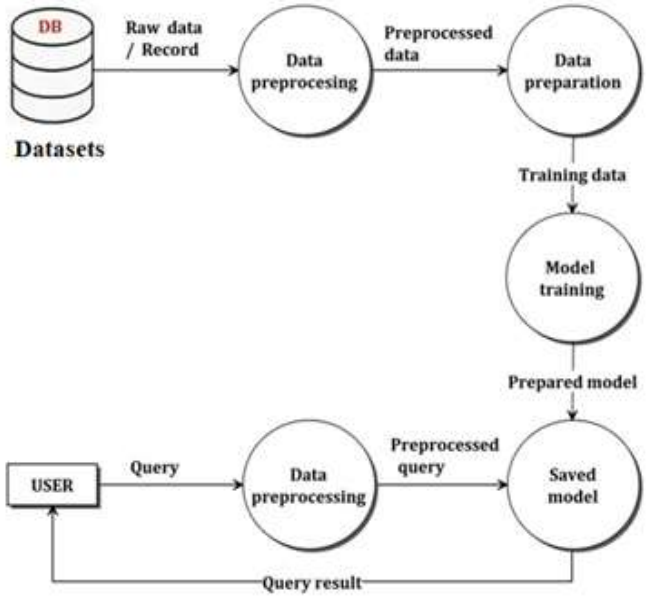

Fig 4.3 Data Flow Diagram (Level 1)

4.3 Use Case Diagram

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. A use case diagram can identify the different types of users of a system and the different use cases and will often be accompanied by other types of diagrams as well. While a use case itself might drill into a lot of detail about every possibility, a use case diagram can help provide a higher-level view of the system. It has been said before that "Use case diagrams are the blueprints for your system". They provide the simplified and graphical representation of what the system must actually do. Fig 4.4 is the Use case diagram, which tells what a user will do in the proposed work

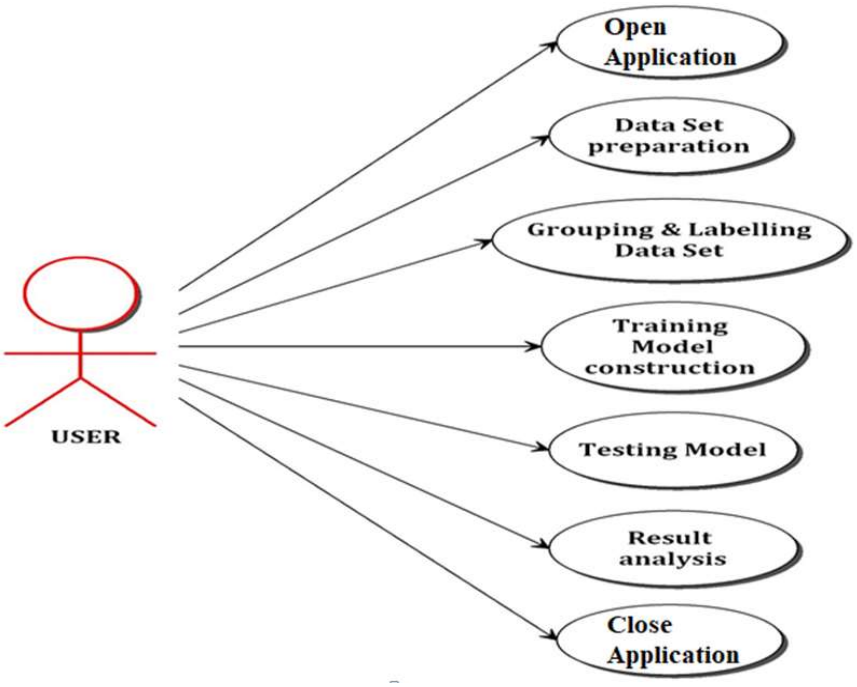

Fig 4.4 Use Case Diagram

Fig 4.5 is the sequence diagram which gives the information about the communication that happens between the user and the model.

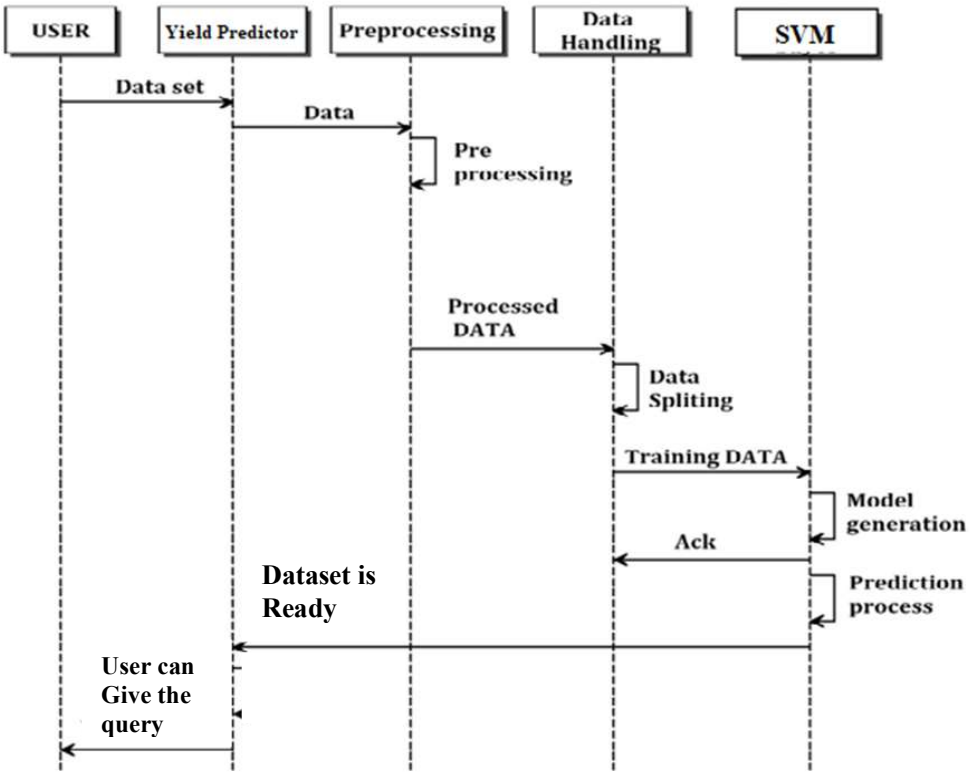

Fig 4.5 Sequence Diagram

CHAPTER 5

IMPLEMENTATION

5.1 Methodology

The implementation can be done in 2 steps:

- Dataset collection and pre-processing
- Building the Model

In the first step, data is collected from kaggle. There are 1600 datasets. The features considered in the dataset are Calcium (Ca), Magnesium (Mg), Potassium (K), Sulphur (S), Nitrogen (N), Lime, Carbon (C), Phosphorus (P), Moisture and the Target (class). Depending on the soil type, the crops are divided into 4 classes/Targets. After data collection, data pre-processing is done by removing redundant values, missing values and replacing the null values. Fig 5.1 gives the detailed view of the steps carried during the process.

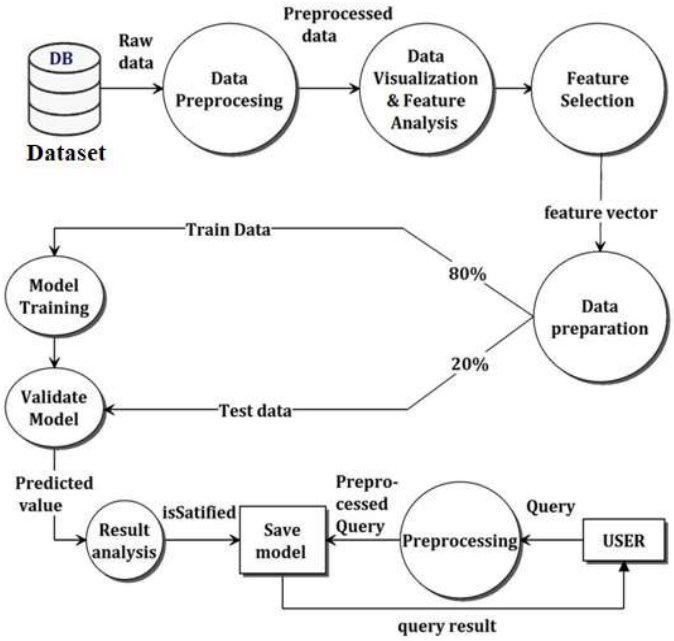

Fig: 5.1: Data Flow diagram

In the second step, the model is built. The model considered in the proposed work is SVM. The dataset is then split into training and testing data i.e. 80% and 20%. The model is trained and tested using the present dataset. The model will now be ready to take new queries and predict crops that can be grown for the particular soil details provided and also print the fertilizer details that can be used for the better growth of the crop.

5.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) 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. Fig 5.5 represents the classification of two classes done by Support Vector Machine.

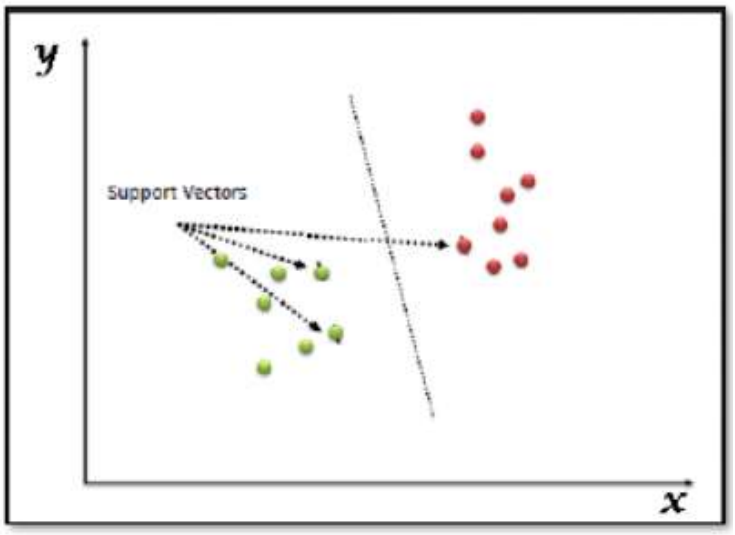

Fig 5.2 Super Vector Method

Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).

CHAPTER 6

RESULTS AND DISCUSSION

There are 9 columns (the 10th column is the target) and 1600 rows in the dataset. The dataset is divided into 80% for training and 20% for testing. The Model is trained using the training data. After that model is given test data and the crops with the fertilizer required are predicted.

Precision measures how accurate are the predictions i.e. the percentage of the prediction are correct. True Positive (TP), True Negative (TN) , False Positive (FP) and False Negative (FN). The formula of Precision is as follows:

$$Precision = \frac{TP}{TP + FP}$$

Recall measures how many good positives are found i.e. how many possible positive cases are found.

$$Recall = \frac{TP}{TP + FN}$$

F1 score is the measure of a test's accuracy. It considers both the prediction p and recall r of the test to compute the score.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one. Table 6.1 represents the confusion matrix.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Table 6.1: Confusion Matrix

Fig 6.1 gives the precision, recall and f1-score values calculated by the proposed model.

	precision	recall	f1-score	support
1	1.00	1.00	1.00	82
2	1.00	1.00	1.00	87
3	1.00	1.00	1.00	73
4	1.00	1.00	1.00	78
accuracy			1.00	320
macro avg	1.00	1.00	1.00	320
weighted avg	1.00	1.00	1.00	320

Fig 6.1: Precision, Recall and f1-score values

Fig 6.2 gives the confusion matrix calculated by the proposed model.

```
Confusion Matrix :
[[82  0  0  0]
 [ 0 87  0  0]
 [ 0  0 73  0]
 [ 0  0  0 78]]
```

Fig 6.2: Confusion matrix

Fig 6.3 is the Home Page of the proposed work application.

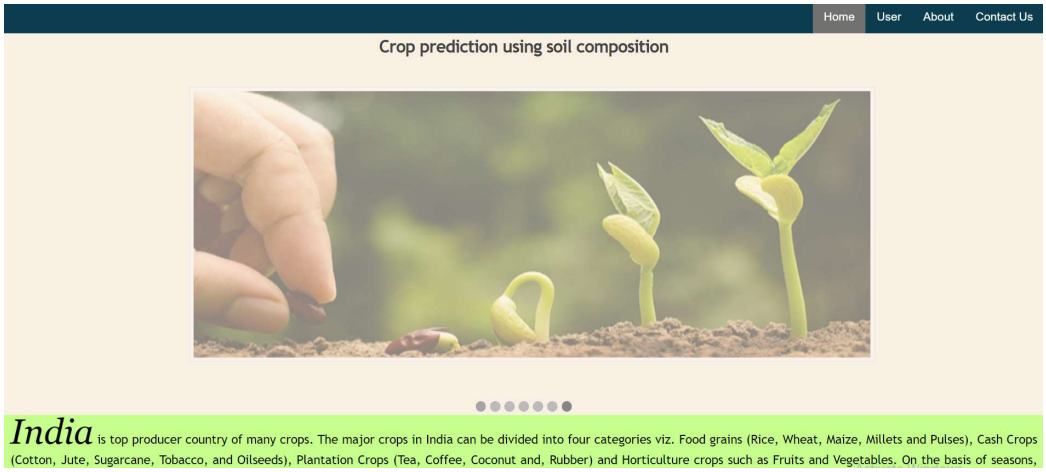

Fig 6.3: Home page

Fig 6.4 shows the registration page of the proposed work application where user registers.

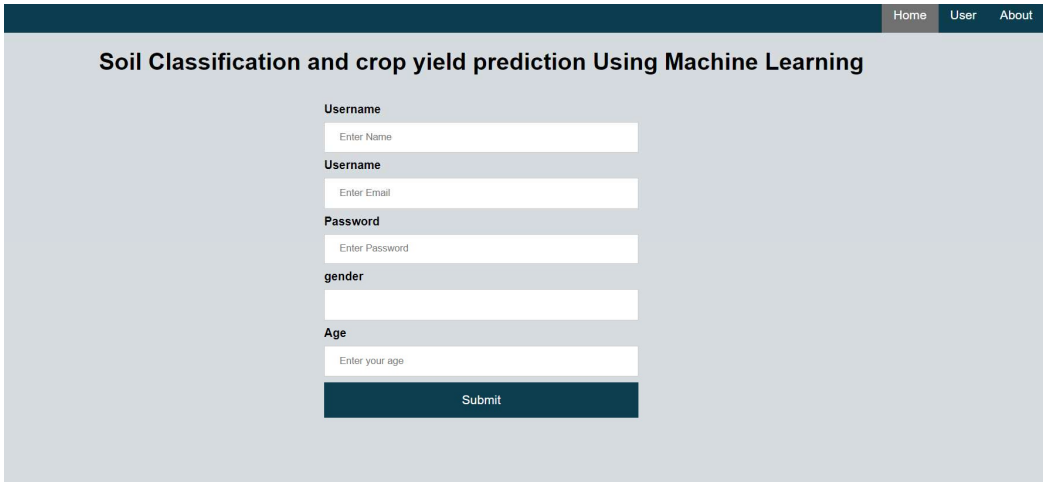

Fig 6.4: Registration Page

Fig 6.5 shows the login page of the application where user can login with the registered username and password.

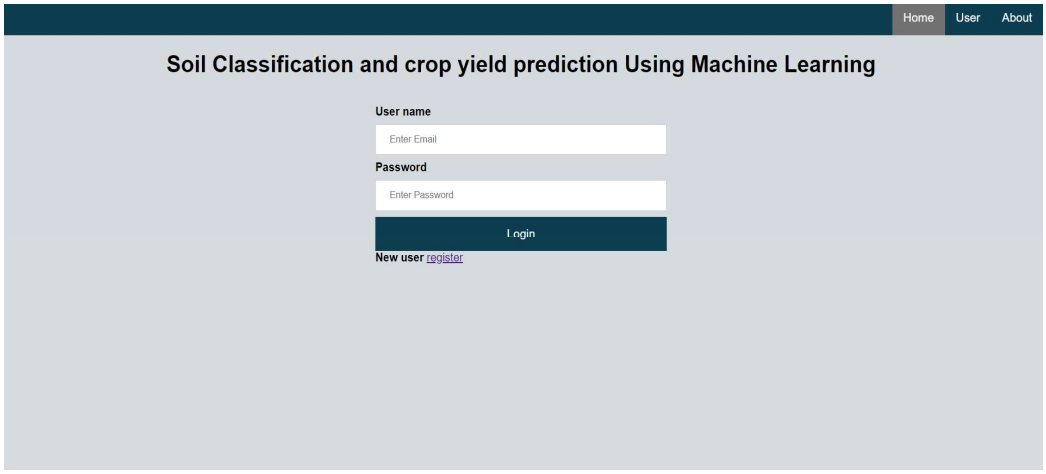

Fig 6.5: Login Page

Fig 6.6 shows the windows that appears when login is successful. The user has to enter the soil composition details asked in the form and click on "Predict" button.

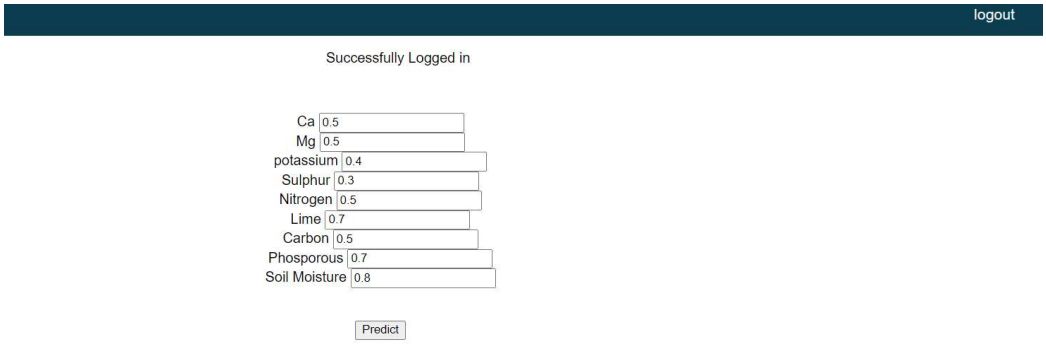

Fig 6.6: Soil Details Form(User enters soil details here)

Fig 6.7 shows the results i.e. the crops and the fertilizers appropriate to the crops are predicted.

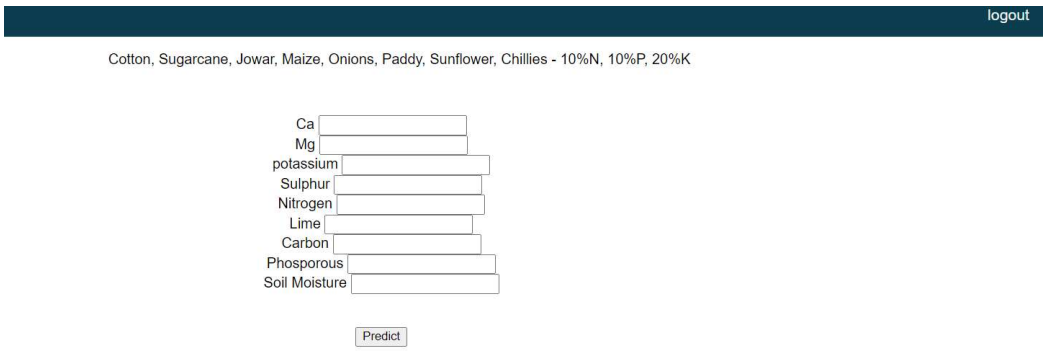

Fig 6.7: Crops and fertilizer predicted

CHAPTER 7

TESTING

After training the model, the model is tested with new set of data. There are 9 features considered in the dataset. They are Calcium, Magnesium, Potassium, Phosphorus, Nitrogen, Carbon and Moisture. Table 7.1 represents the dataset that is used in the proposed work.

Ca	Mg	K	S	N	Lime	C	P	Moisture	class
0.7	0.6	0.8	0.8	0.7	0.8	0.3	0.1	0.9	4
0.5	0.5	0.4	0.3	0.5	0.7	0.5	0.7	0.8	2
0.6	0.8	0.1	0.3	0.7	0.5	0.5	0.6	0.6	2
0.7	0.7	0.7	0.5	0.8	0.7	0.4	0.1	0.7	4
0.8	0.8	0.2	0.3	0.5	0.5	0.7	0.8	0.5	2
0.7	0.5	0.6	0.7	0.8	0.8	0.3	0.4	0.8	4
0.7	0.5	0.8	0.8	0.2	0.4	0.7	0.5	0.7	3
0.1	0.4	0.7	0.5	0.6	0.7	0.5	0.6	0.9	1
0.5	0.8	0.3	0.4	0.7	0.7	0.8	0.6	0.7	2
0.4	0.2	0.6	0.6	0.6	0.5	0.8	0.7	0.5	1
0.8	0.7	0.5	0.6	0.1	0.4	0.7	0.6	0.7	3
0.6	0.6	0.7	0.5	0.2	0.4	0.7	0.8	0.9	3
0.6	0.5	0.8	0.5	0.1	0.4	0.5	0.8	0.5	3
0.4	0.2	0.5	0.7	0.7	0.6	0.8	0.5	0.5	1
0.5	0.7	0.8	0.7	0.7	0.8	0.4	0.2	0.5	4

Table 7.1: Dataset

Accuracy, is one metric for evaluating classification models. It is the ratio of number of correct predictions to the total number of input samples. The formula for accuracy is as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

The accuracy achieved using the SVM model in the proposed work is 100%. As the number of classes the machine has to predict is only 4, and there are 1600 rows. So machine predicts the right crops for the appropriate soil composition values.

CHAPTER 8

CONCLUSION AND FUTURE SCOPE

8.1 Conclusion

Agriculture is the backbone of many countries including India. Since integrating the information technology with the agriculture will guide the farmer to improve the productivity. In this proposed work the system described works faster and gives better accuracy in prediction to predict the suitable crops and fertilizers for the field. It includes various parameters of soil to analyse the crop. This prediction makes the farmers to improve the productivity, growth, and quality of the plants.

8.2 Future Scope

In future this project can be modified by adding few more things like:

- Predicting the pesticide ratio based on the atmospheric and soil parameters for the farm land.
- Database can be created to store the soil details which given by the farmer and also the predicted crop name.
- New CSV file can be created where the new datasets can be added and this can be used for training the model in the future.

REFERENCES

1) Karthikeyan R, Gowthami M, Abhishek A and Karthikeyan P, "Implementation of Effective Crop Selection By using the Random Forest Algorithm", published by International Journal of Engineering & Technology in 2018.
2) Monali Paul, Santosh K. Vishwakarma and Ashok Verma, "Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining", published by International Conference on Computational Intelligence and Communication Networks in 2015.
3) Rakesh Kumar, M.P Singh, Prabhat Kumar and J.P.Singh, "Crop Selection Method to Maximize Crop Yield rate using Machine Learning Technique", published by International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials(ICSTM) in 2015.
4) Shriya Sahu, Meenu Chawla and Nilay Khare, "An Efficient Analysis of Crop Yield Prediction Using Hadoop Framework Based On Random Forest Approach" published by International Conference on Computing, Communication and Automation (ICCA) in 2017.
5) Tom M. Mitchell, Machine Learning, McGraw Hill Education, Indian Edition 2013.