Soil Evaluation for Crops Cultivation

Asif Khan Shawon
Student Id: 011141138

Arif Faysal
Student Id: 011121090

Md. Shahadat Hossain
Student Id: 011132082

Md. Tahmid Chowdhury
Student Id: 011141060

A thesis in the Department of Computer Science and Engineering presented
in partial fulfillment of the requirements for the Degree of
Bachelor/Master of Science in Computer Science and Engineering

United International University
Dhaka, Bangladesh
January, 2019

Declaration

We, Asif Khan Shawon, Arif Faysal, Md. Shahadat Hossain and Md. Tahmid Chowdhury, declare that this thesis titled, Thesis Title and the work presented in it are our own. We confirm that:

- Where any part of this thesis has previously been submitted for a degree or any other qualification at United International University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

_____

Asif Khan Shawon
Student Id: 011141138
Department of Computer Science and Engineering
United International University
United City, Madani Ave, Dhaka 1212

_____

Arif Faysal
Student Id: 011121090
Department of Computer Science and Engineering
United International University
United City, Madani Ave, Dhaka 1212

_____

Md. Shahadat Hossain
Student Id: 011132082
Department of Computer Science and Engineering
United International University
United City, Madani Ave, Dhaka 1212

_____

Md. Tahmid Chowdhury
Student Id: 011141060
Department of Computer Science and Engineering
United International University
United City, Madani Ave, Dhaka 1212

Certificate

I do hereby declare that the research works embodied in this thesis entitled "**Soil Evaluation for Crops Cultivation**" is the outcome of an original work carried out by Asif Khan Shawon, Arif Faysal, Md. Shahadat Hossain and Md. Tahmid Chowdhury under my supervision.

I further certify that the dissertation meets the requirements and the standard for the degree of BSc in Computer Science and Engineering.

_____

Dr. Mohammad Nurul Huda
Professor and Director- MSCSE
Department of Computer Science and Engineering
United International University
Dhaka-1212, Bangladesh.

Abstract

Our thesis has focused on cultivation of 55(fifty five) crops in 2(two) districts of Bangladesh. The classification of crop was based on the elements of soil and some other soil states. Classification algorithms we used are Random Forest, J48 Decision Tree, KNN, Neural Network, Naïve Bayes, SVM and some other algorithms. These algorithms give different % of accuracy.

The main purpose of the thesis to make a proper way to find the crop which is suitable with the elements and state of a soil sample for maximum result. The main problem we faced the shortage of data and making the limited data suitable for a minimum level of accuracy for every algorithm we used. The data book was collected from Soil Research & Development Institute (SRDI), Khamarbari, Dhaka.

Acknowledgement

First of all,I wish to express my sincere thanks to our honorable supervisor Prof. Dr. Mohammad Nurul Huda, for providing the necessary facilities for the research.

I take this opportunity to express gratitude to all of the department faculty members for their help and support. I also thanks my parents for their unceasing encouragement, support and attention.

I also place on record, to sense of gratitude to one and all, who directly and indirectly, have helped us in their venture.

Our research is dedicated to Department of CSE, United International University and our honorable supervisor Prof. Dr. Mohammad Nurul Huda.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In general sense, farmers decide which crop they will cultivate and then after or just before sowing they apply fertilizer needed for that crop. A farm land already contains fertilizer of previous cultivation. And then, additional fertilizer makes farm land not suitable for the crop as we come to know that excess fertilizer doesn't mean more yield. It's a big problem for farmers as this decreases the yield of crop as well as wastage of fertilizer and money.

If the farmers can get the idea of their soil elements and which crops are suitable for the soil, they can get more yields. In addition, they can apply fertilizer if they get soil report in details for a particular farm land.

Now-a-days, data science can solve the problem by testing the elements of particular soil sample. Using classification, we use a data set for 51 crops. By using soil elements farmers can get the result of crops yield more crops. In some cases they don't even need to apply additional fertilizer. That will save their money and time.

Chapter 2

Classification

2.1. Introduction of Classification

Classification is a data analysis process to make a model to classify certain instance of data. It is a process of learning to make diction to classify instances and give accurate result as possible. It has two steps,

1. **Learning:** Construction of Classification Model Different Algorithms are used to build a classifier using the available training sets to making the model learn. Every model has to be trained for prediction of accurate results.
2. **Classification:** Models are used to predict class labels and testing the constructed model on test data and therefore estimate the accuracy of the classification rules.

Data requirement for classification algorithm:

1. Enough data instance to learning process draws many entities, relationships
2. Data arrangement should be normalized
3. Diverse Data
4. As low as possible missing data
5. Discrete data

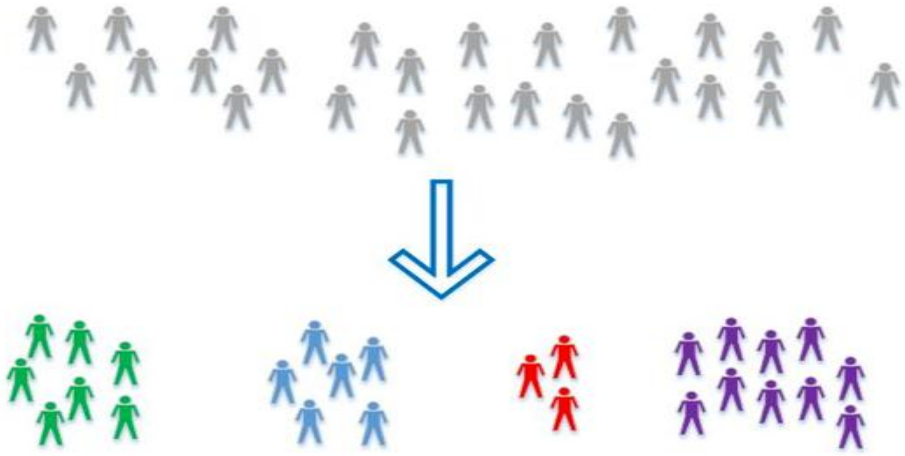

Figure 1: Classification Examples

2.2. Goals of Classification

Goal of classification is to settle the natural grouping from set of labeled data. In so many classification algorithm it can't be possible to say that a particular method is appropriate. It totally depends how the results a researcher wants to get and he needs to follow such way of different processes.

2.3. Classification Algorithms

Classification algorithm describe as listed below:

2.3.1 Naïve Bayes

Naive Bayes classifier is based on Bayes (1702-1761) Theorem. It's a family of algorithm where all of the algorithm render a basic common principle. The Bayes classification represent a supervised learning method as well as statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve predictive and diagnostic problems.

Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Using Bayes theorem we can find the probability of A happening, given that B has occurred. Here B is the evidence and A is the hypothesis. The assumption made here is that the features are independent, which is presence of one particular feature does not affect the other. Hence it's called naïve.

2.3.2 KNN

KNN means K number of nearest neighbors. It is a very simple algorithm which can classify by measuring distance, takes the class occurs maximum number among the classes in N nearest neighbors.

How it works:
An instance is classified by KNN counting the maximum class occurred in K number of nearest neighbor. The value of K can be any number depending on the dataset. Lowest value, k=1 means the most nearest neighbor. There are 3 types of function to determine nearest neighbors.

Euclidean: $\sqrt{\sum_{i=1}^{K}(x_{i-}\, xy_i)^2}$

Equation 1: Euclidean distance

Manhattan: $\sum_{i=1}^{k} |x_i - y_i|$

Equation 2: Manhattan distance

2.3.3 Random Forest

Random forest is a very easy and flexible algorithm to use in machine learning. It gives high accurate result without some requirements needed in other algorithms and it's simplicity made it the most used algorithm.

How it works:
Random forest creates a forest like random decision tree. It most of the time trained with the bagging method, the method is a combination of learning models increase the overall result. It builds many decision tree and merge them to get an accurate prediction. It adds additional randomness while growing the tree. Instead of searching the most frequent result it search for the best feature.

2.3.4 J48 Decision Tree

J48 is the extension of ID3 with more feature of fulfill the continuous attribute value ranges, decision tree pruning, derivation of rules and missing values.

How it works:
ID3(**Iterative Dichotomiser 3**) creates a decision tree of a given dataset. It starts on the original datasets root node and do some iteration. On each iteration it checks the unsolved attributes and calculate the entropy or IG(Information Gain). Then it selects the attributes in ascending order. Then the data set is split by the selected attribute to produce subset of data. This process will be continued till all every element in same class.

2.3.5 Support Vector Machine

SMV(Support Vector Machine) is a machine learning algorithm that can be used for both regression and classification. But mostly it's used in classification. This algorithm places data in dimensional place. Number of dimension is dependent on feature number. Then it performs classification by finding the hyper-plane that differentiate the two classes very well.

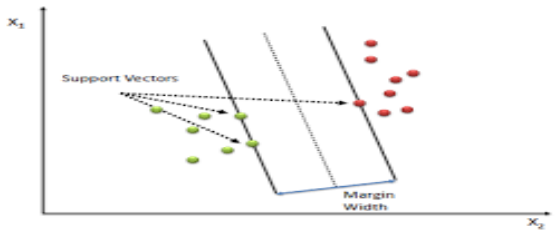

Figure 2: Support Vector

The line in the middle of two lines separates the two spaces of classes indicated by red and green dots. These dots are called support vectors.

2.3.6 Deep Neural Network

Neural Network is mainly electronic network based on the neural structure of human brain. This algorithm process data and learn comparing previous known data classes.

How it works:
A neuron is the unit of the neural network. It is set of input values (xi) and associated weights(wi) A function (g) sums the weights and maps the result to the output(y).

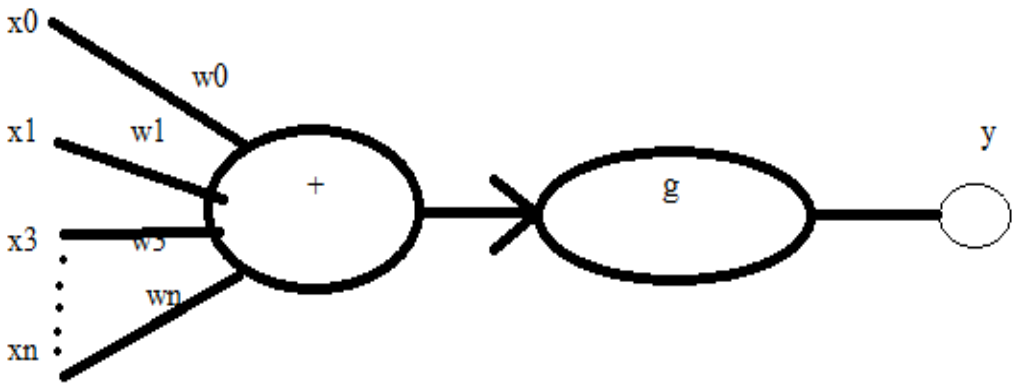

Figure 3: Deep Neural Network

Neurons are organized in input layer, hidden layer and output layer. The input layer contains the data. The very next layer is hidden layer. Hidden layer may be many in number. The next layer is the output layers that is single node for each classes.

Chapter 3

Data Collection and Preprocessing

3.1. Data Collection

We collected data from “Upozila Nirdeshika”, Land and soil resource utilization guide book. The books were collected from Soil Resource Development Institute. We took two books of two Upazila- Hijla and Srinagar as these two books contain more data than others. The guide boos discusses about all characteristic, soil property, elements. It also contains the measurements of elements which determines the crop for cultivation.

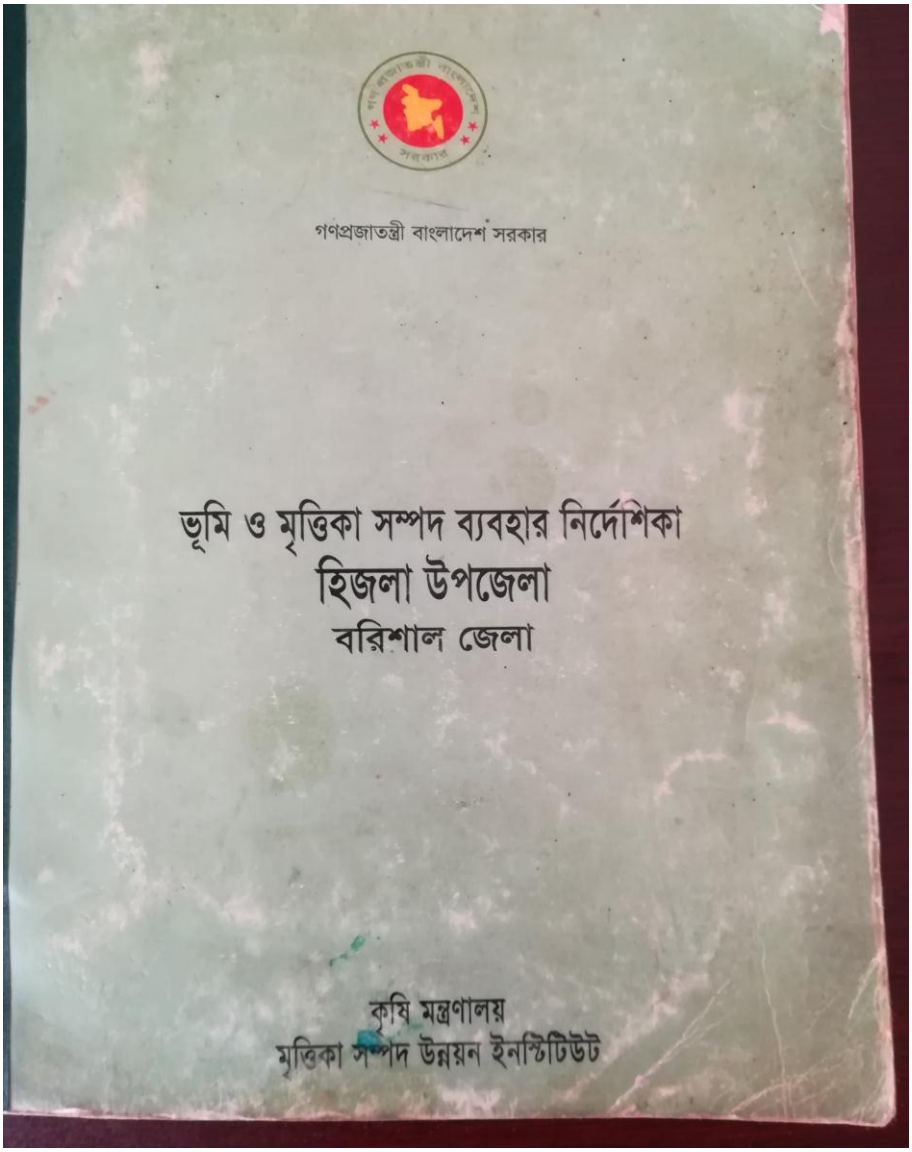

Figure 4: Upozila Nirdeshika (Hizla)

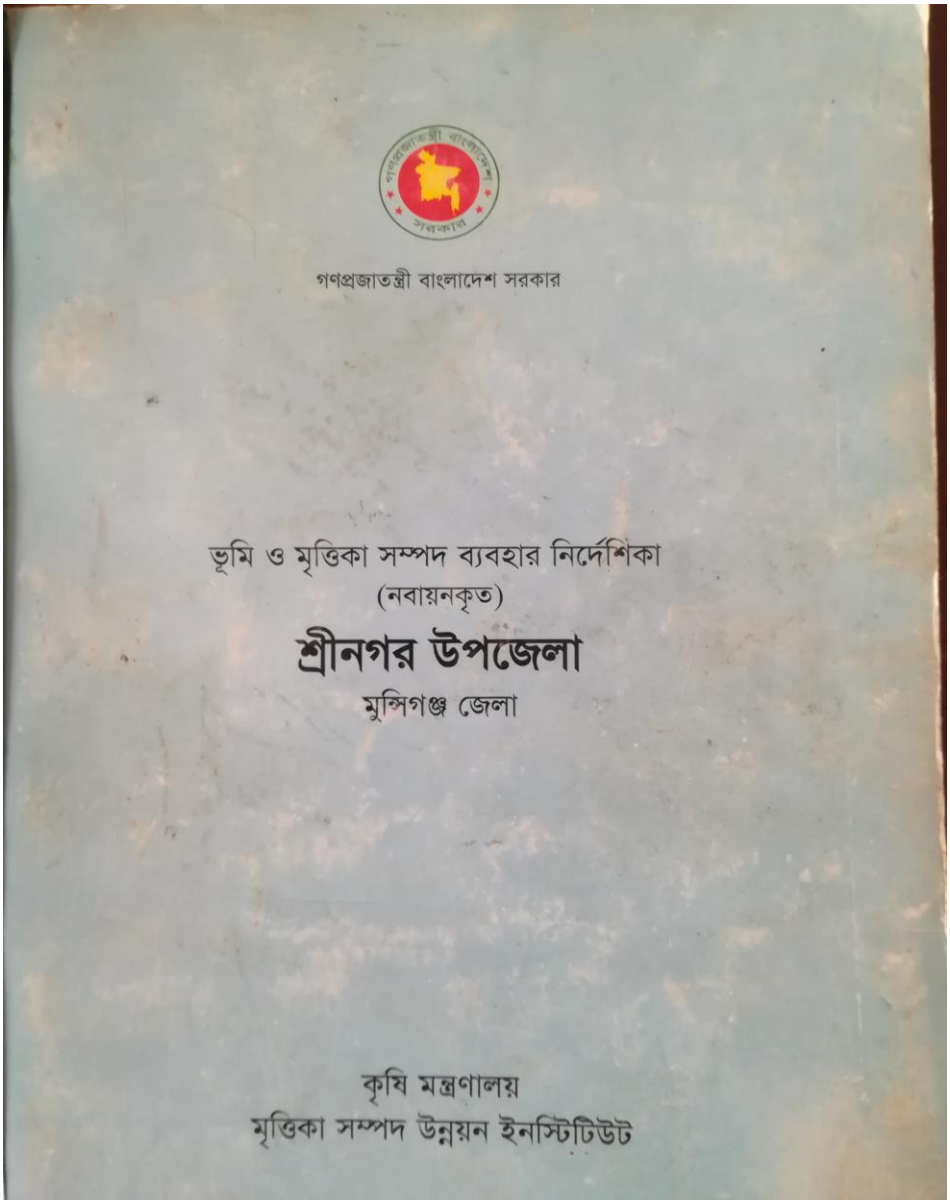

Figure 5: Upozila Nirdeshika (Srinagar)

3.2. Preprocessing

The guide books of Upazila Hijla and Srinagar had many data classified in various fields. These guide books also provide a Map with it for easy understanding of data and their origins. But these data are not in the desired form that can be classified so easily. For that reason we had to preprocess the data by following preprocessing methods in order to create our desired dataset so that the classifying algorithms work finely. Firstly it is being find out, how many unions is there on a particular Sub-District. Different types of soil can be available in each single Union. According to soil series the areas of the map is shown in different color. We marked first 4 features as it can be differentiate only for different areas. We took the non-zero values and didn't take the zero values which was unnecessary. But it is necessary to consider the missing data while we will consider a big dataset from the soils of different areas of Bangladesh.

Land Class	Roughness	Water Removal Condition	Drainage Class	Kint (Surface/ Mid level)	Soil Strength (Surface)	Dry Condition on the Ground	Feedback (Surface)	Nitrogen	Phosphorus	Potassium	Sulfur	Zinc	Boron	Malibedenam	Magnesium	Matter (Ton/hectare)	Crops
			Oppressed														Sugercane (Moderate)
High Land	Rugged	Very Early	Poorly Oppressed	Doash	Frosty	Much	Neutral	83	28	91	23				16		Sugercane (Middle)
High Land	Rugged	Very Early	Poorly Oppressed	Doash	Frosty	Much	Neutral	50	15	100	5	-	0.25				Coconut (Moderate)
High Land	Rugged	Very Early	Poorly Oppressed	Doash	Frosty	Much	Neutral	151	41	251	16	0.3	0.8				Coconut (Middle)

Figure 6: First look data sheet

3.2.1 Feature Extraction

Our dataset has in total 17 features which are the elements of the soil. These features are given below

->Land Type
->Water Recession
->Drainage Status
->Soil Components
->Soil Rigidity
->Soil Juice
->Soil Reaction
->Soil Series
->Chemical Elements
- -> Nitrogen
- ->Phosphorus
- ->Potassium
- ->Sulfur
- ->Zinc
- ->Boron
- -> Molybdenum
- ->Magnesium
- ->Organic Matter

We mainly worked with 9 elements among of them. Because the other elements were not effective for the results. But if we can take data from all the Upazila's then the rest of the elements we ignored can be a major part of prediction.

Nirogen	Phosphor	Potassium	Sulfur	Zinc	Boron	Malibedei	Magnesiu	Organic M	Crops
29	9	27	7	0	0	0	0	2	Tomato
86	26	82	20	0	0.65	0.15	0	4	Tomato
20	6	13	3	0	0	0	0	0	Tomato
61	19	38	8	0.5	0.15	0	0	3	Tomato
33	7	27	7	0	0	0	0	2	Cabbage
98	22	82	20	0	0.3	0.15	0	4	Cabbage
30	18	10	5	0	0	0	0	0	Cabbage
91	53	31	16	1.5	0.2	0.15	0	8	Cabbage
61	23	31	10	1.55	0.2	0.15	0	4	Cabbage
25	9	20	5	0	0	0	3	2	Cauliflower
78	26	61	16	0	0.3	0.15	10	4	Cauliflower
20	10	20	4	0	0	0	0	0	Cauliflower
61	31	61	13	1.5	0.3	0.15	0	4	Cauliflower
38	14	38	5	0	0	0	0	2	Pineapple

Figure 7: After Preprocessing

We used 51 crops as our class value from different season and types. Table of all the crops given below.

Crops list in different season:

Crops

Seasons	Summer	Winter	Rainy	Spring	Fall	All
	Asparagus bean		Cotton	Red Amarnath	Sweet Potato	Coconut
	Ladies Finger			Amarnath	Pineapple	Eggplant
	Pointed Gourd			Cucumber		Soybean
	Pani Kochu			Sesame(2)	Sesame(2)	Rice Pumpkin
	Chili (2)	Chili(2)			Sugarcane	Banana
	Jackfruit	Khesari			Onion	
	Mango	Chinese Almond				
	Cheena (2)		Cheena(2)			
	Corn(3)	Corn(3)		Corn(3)		
			Jute(2)	Jute(2)		
			Wheat			
	Sunflower(3)	Sunflower(3)	AUS	Sunflower(3)		
		Turmeric	Kangkong			
		Cauliflower	basella alba			
		Gram				
		Mustard				
	Sweet Pumpkin(2)	Sweet Pumpkin(2)				
		Tomato				
		Cabbage				
	Ground Nut(3)	Ground Nut(3)		Ground Nut(3)		
	Bean	Potato		Boro		
	Ginger					
	Knitted aus					

Table 1: Class Values in Different Season

3.2.2 Data Visualization

Table of attribute type:

Data Set

Name	Type	Description
Nitrogen	Numeric	Amount of Nitrogen in soil
Phosphorus	Numeric	Amount of Phosphorus in soil
Potassium	Numeric	Amount of Potassium in soil
Sulfur	Numeric	Amount of Sulfur in soil
Zinc	Numeric	Amount of Zinc in soil
Boron	Numeric	Amount of Boron in soil
Molybdenum	Numeric	Amount of Molybdenum in soil
Magnesium	Numeric	Amount of Magnesium in soil
Ogranic Matter	Numeric	Amount of Organic Matter in soil

Table 2: Attribute Type

We used WEKA to apply different data visualization method to visualize our dataset. Here some pie chart model is shown below. Pie charts with a few of crops among with their required soil elements.

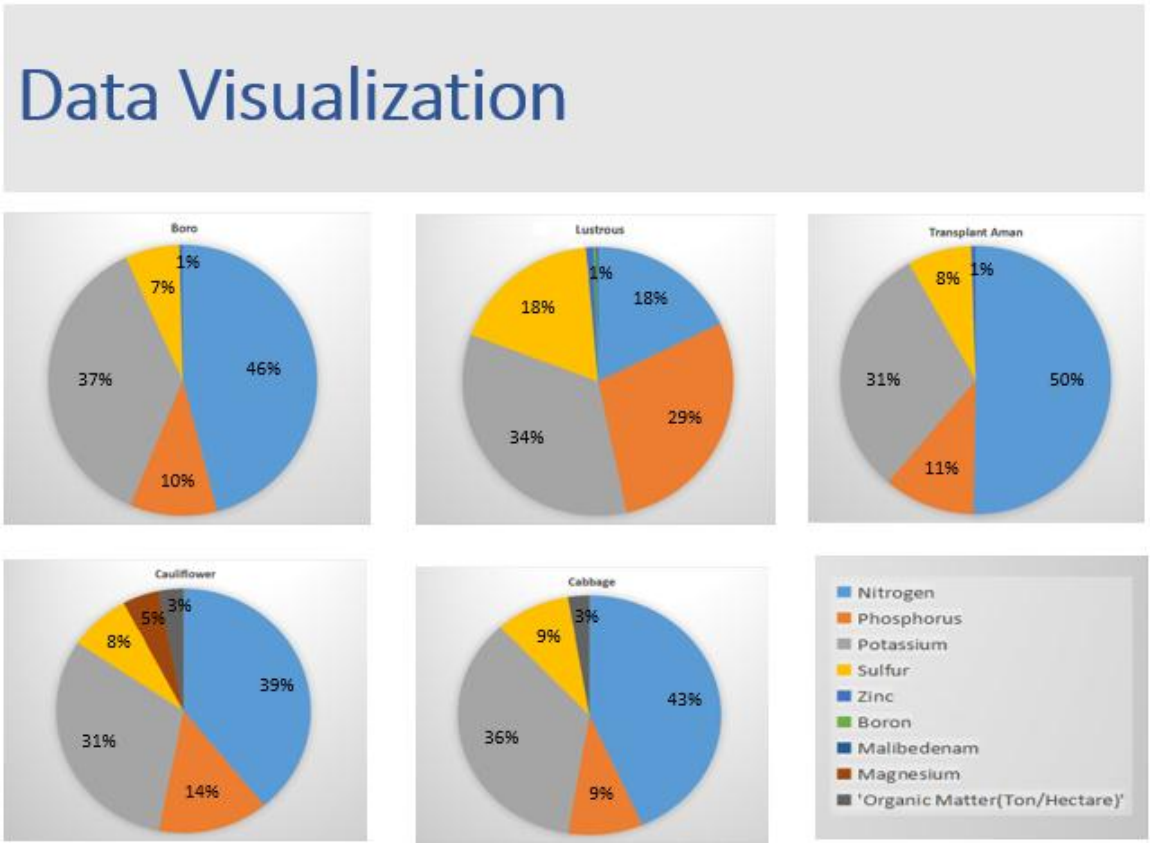

Figure 8: Pie Chart of Crops Split into Soil Elements

Among of the attribute value we took the Sulfur and compared with the crops in a graph. In this graph we can see the effect of sulfur on crops.

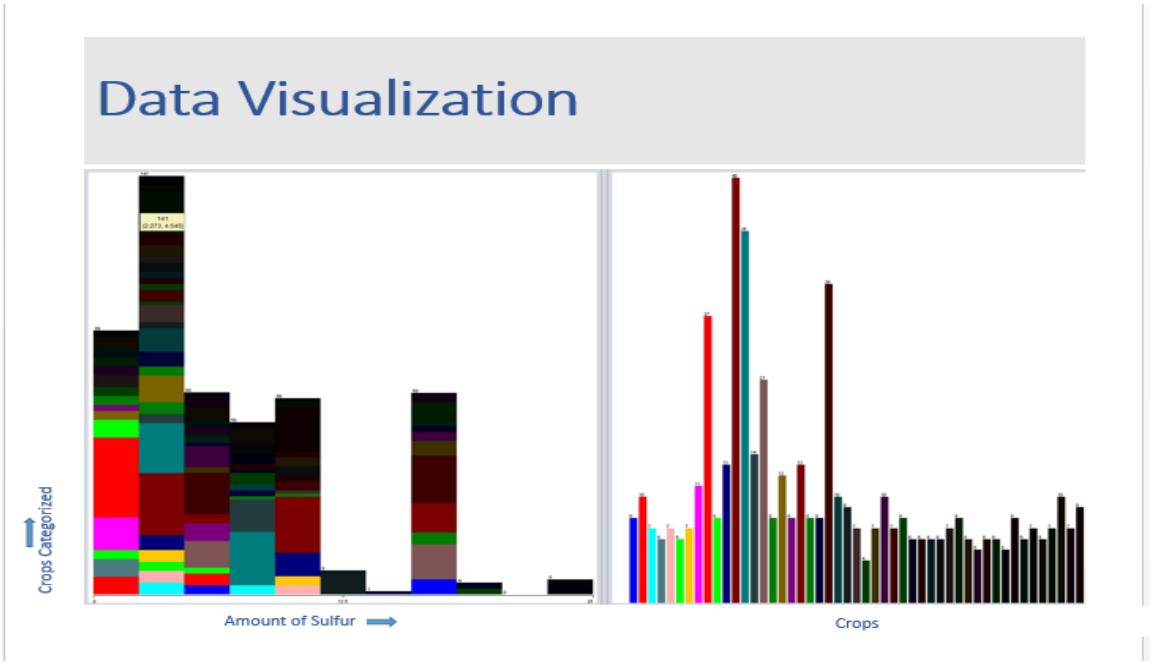

Figure 9: Quantity of Sulfur by Crops Graph

And then we took all the crops and made a bar chart for each crop what is the quantity of attributes.

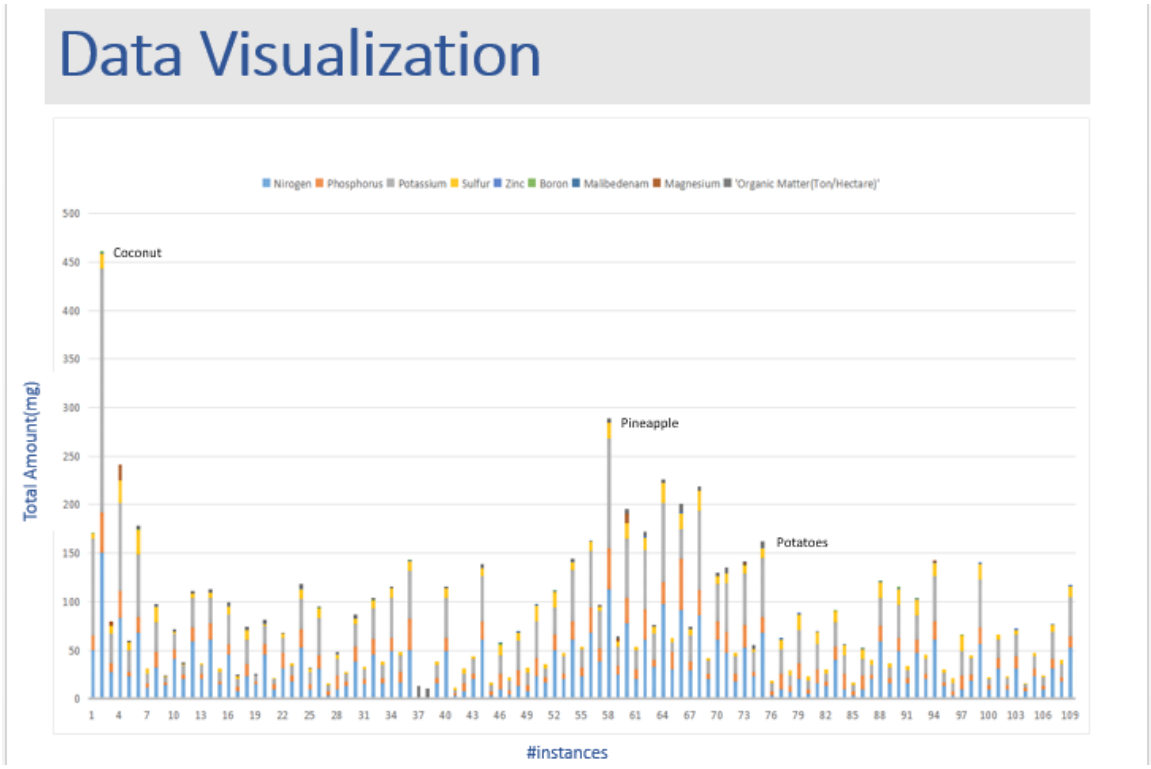

Figure 10: Bar Chart (Crops vs Attribute)

Chapter 4

Experiments and Results

4.1. Experiments

We implemented different machine learning algorithms on our dataset and got different accuracy.

We developed a JAVA based Soil Evaluation tool to compare all the machine learning algorithms. Some screenshots given below as for example.

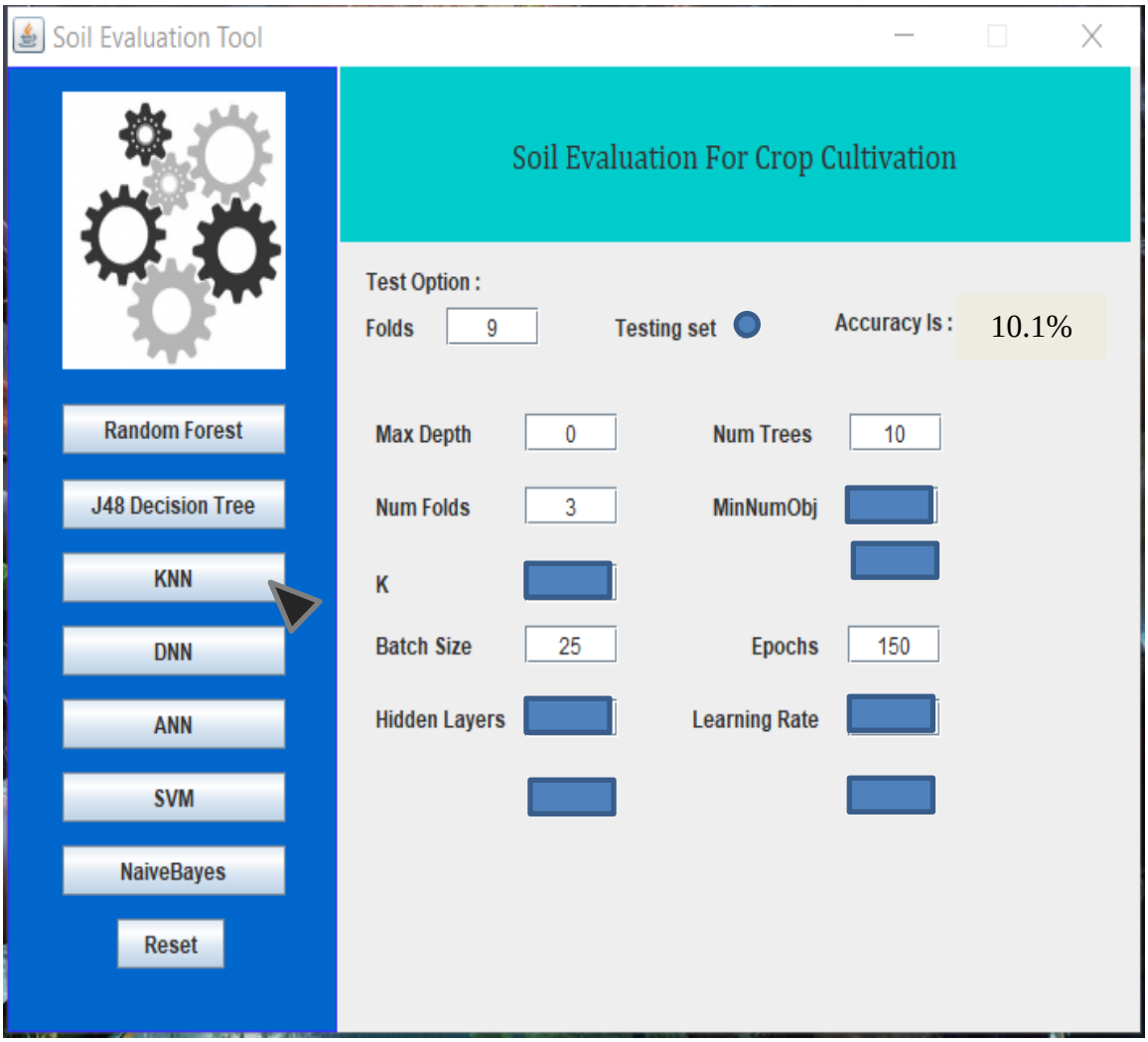

Figure 11: KNN

For KNN we can see the value of K is 2, number of fold is 9 and we get the accuracy for testing data set is 10.1%.

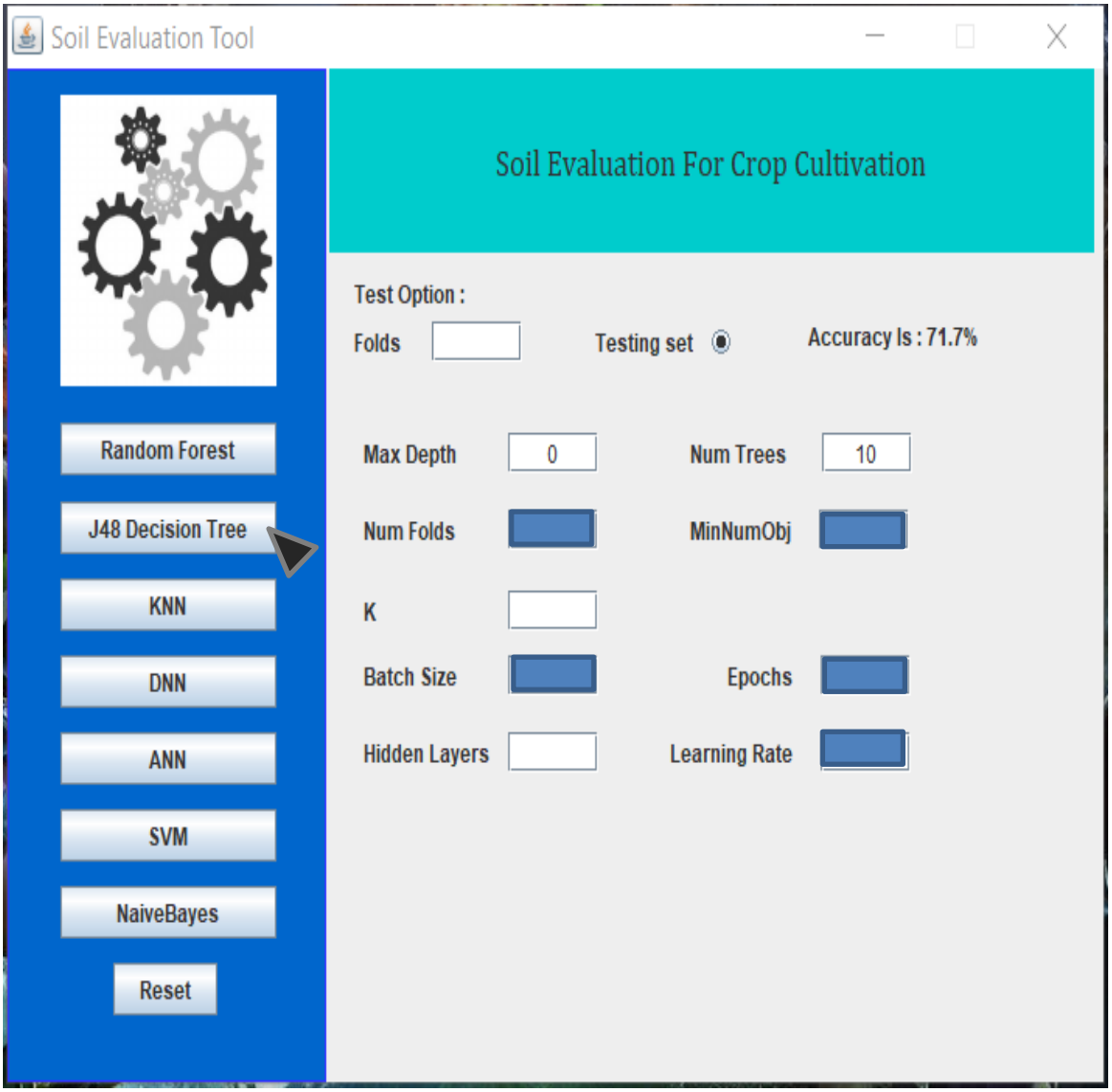

Figure 12: J48 Decision tree

For J48 Decision tree it shows the accuracy for testing data set is 71.7%.

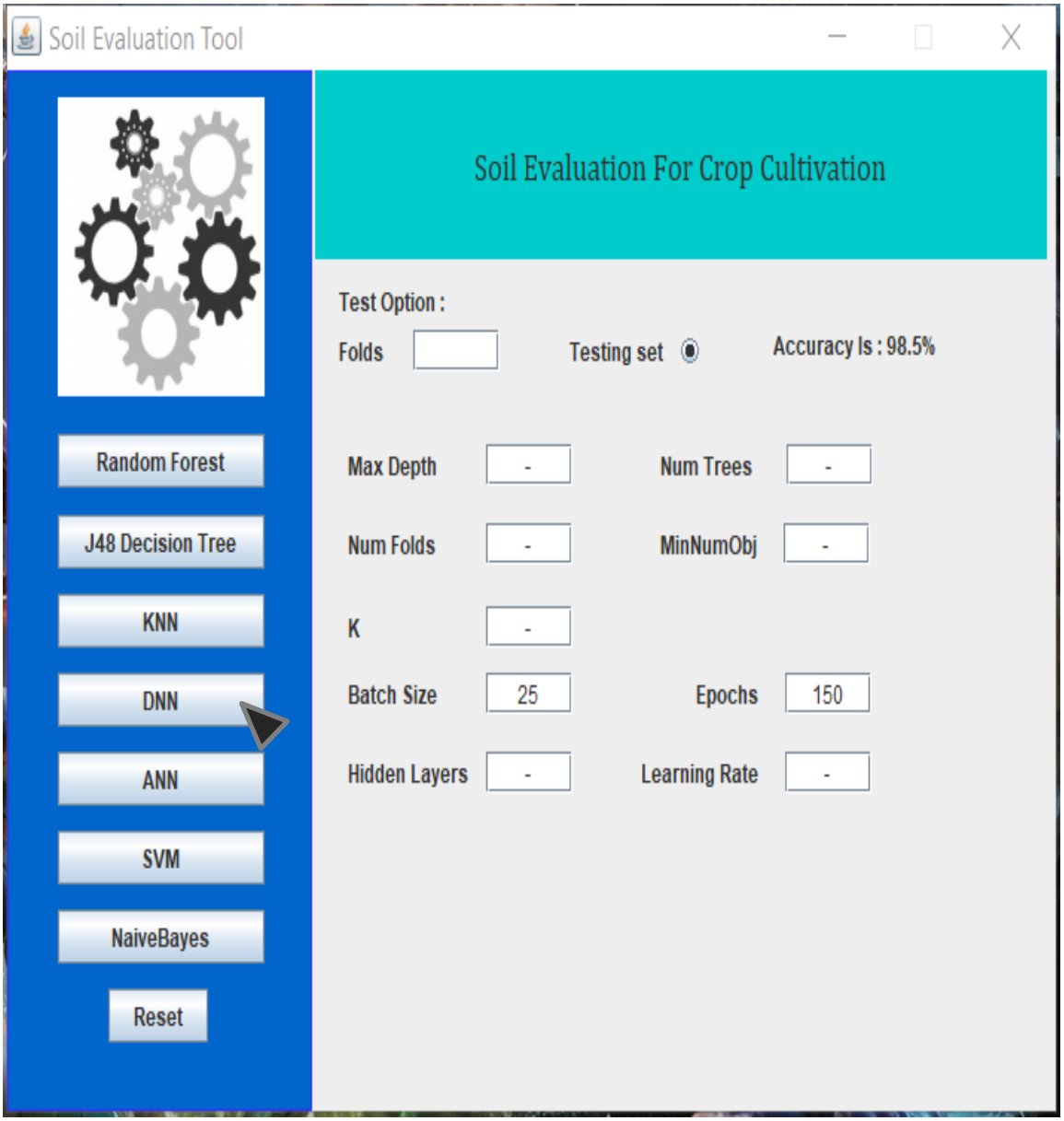

Figure 13: Deep Neural Network

For Deep Neural Network we got so far the best result which is 98.5%. When the batch size was 25 and the epochs was 150.

Here is the table of comparison of all machine learning algorithms:

Classifier	Training Data	Test Set
KNN(k=2)	71.7%	10.1%
Decision Tree	87.2%	71.7%
Random Forest	91.5%	97.9%
Naïve Bayes	67.3%	24.2%
SVM	27.9%	24.2%
Neural Network	67.3%	21.2%
DNN	98.5%	98.5%

Table 3: Comparison of all Machine Learning Algorithms We Implement

4.2. Results

In case of prediction of crops Deep Neural Network gives 99.5% of accuracy. Which is the best so far. Our research was about to found the best classifier among the machine learning algorithms we implemented. We followed the process while implementing is given below:

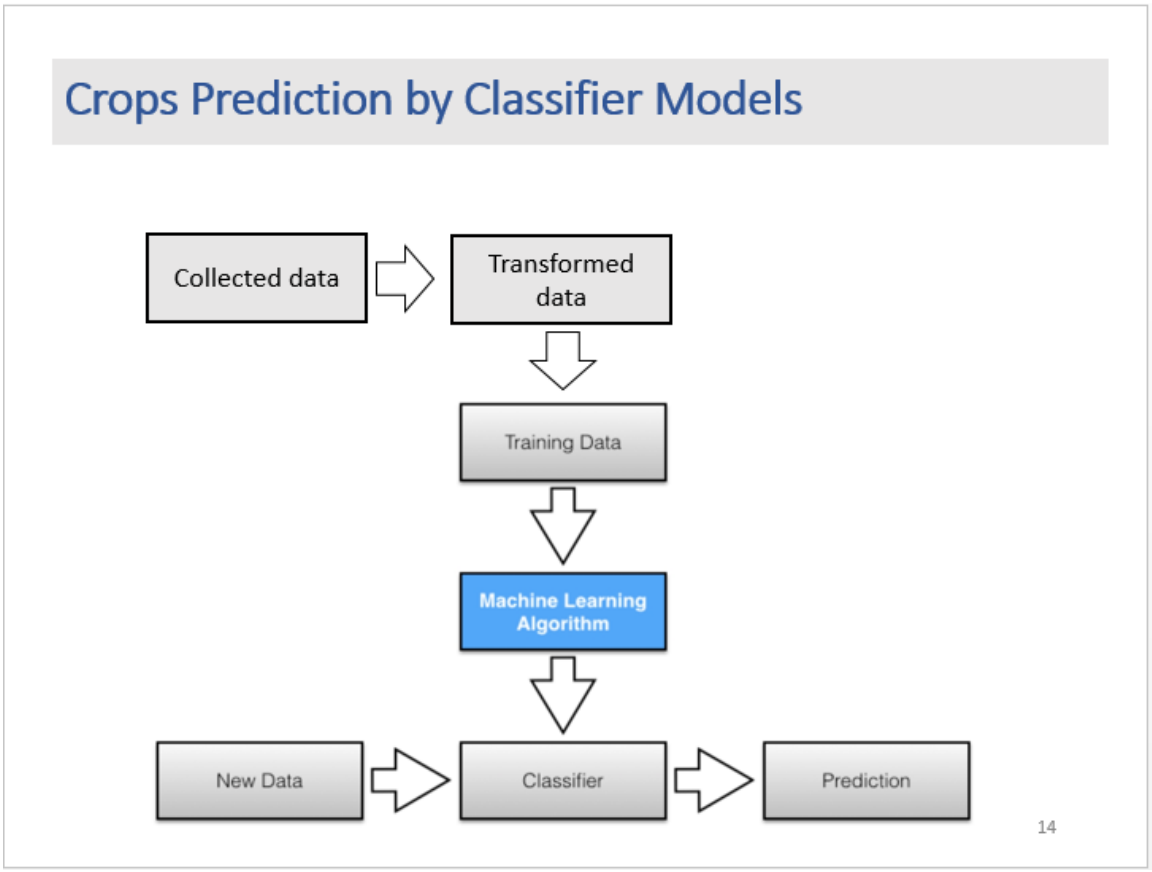

Figure 14: Prediction Procedure

Chapter 5

Conclusion and Future Works

This research has proposed a system of soil evaluation for crops cultivation. We have experimented on various machine learning algorithms. The study concludes the following:

i) The Deep Neural Network provided the best accuracy of 98.5% on test set.
ii) Experimentation was done on limited data.
iii) KNN provides lowest accuracy.
iv) We did not consider missing data.
v) We did not consider land type, climate data and irrigation period in our dataset.

In future we would like to concentrate on the limitations described above.

References

[1] 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.

[2] F.Sarmadian and R. Taghizadeh Mehrjardi, “Modeling of Some Soil Properties Using Artificial Neural Network and Multivariant Regression in Gorgan Province, North of Iran” Global Journal of Environmental Research 2(1): 30-35, 2008 “ISSN 1990-925X” ”IDOSI Publications, 2008.

[3] Akash Bhuiyan, Afrida Islam, Junaeid Ahmed, Md. Aminul Islam, Mohammad Nurul Huda, “Automated Soil Clustering for Crops Cultivation” 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT) [4] G. Gehlen et al., “Mobile P2P Web Services using SIP,” Mobile Information Systems, vol. 3, pp. 165–185, 2007.

[5] R Isbell, "The Australian Soil Classification, " National Committee on Soil and Terrain (NCST),1996.

[6] T Hari Eswaran, Thomas Rice, Robert Ahrens, Bobby A. Stewart, "Soil Classification- A global desk refernce," December 26, 2002 by CRC Press, 280 Pages.