ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Analysis of Soil Behavior and Prediction of Crop Yield using Data Mining Approach

Supriya D M

M.Tech (CSE), Dept. of Computer Science and Engineering, PES College of Engineering, Mandya, Karnataka, India

ABSTRACT: Yield prediction is very popular among farmers these days, which particularly contributes to the proper selection of crops for sowing. This makes the problem of predicting the yielding of crops an interesting challenge. Earlier yield prediction was performed by considering the farmer's experience on a particular field and crop. This work presents a system, which uses data mining techniques in order to predict the category of the analyzed soil datasets. The category, thus predicted will indicate the yielding of crops. The problem of predicting the crop yield is formalized as a classification rule, where Naive Bayes and K-Nearest Neighbor methods are used.

KEYWORDS: Yield Prediction, Data Mining, Classification Rule, Soil Analysis.

I. INTRODUCTION

Crop forecasting or prediction is the art of predicting crop yields and production before the harvest actually takes place, typically a couple of months in advance. Crop forecasting relies on computer programs that describe the plant-environment interactions in quantitative terms. The soil testing program starts with the collection of a soil sample from a field. The first basic principle of soil testing is that a field can be sampled in such a way that chemical analysis of the soil sample will accurately reflect the field's true nutrient status. The purpose of soil testing in high-yield farming is to determine the relative ability of a soil to supply crop nutrients during a particular growing season, to determine the needs, and for diagnosing problems such as excessive salinity or alkalinity. Soil testing is also used to guide nutrient management decisions related to manure and sludge application with the objective of maximizing economic/agronomic benefits while minimizing the potential for negative impacts on water quality. Data Mining is a process of extracting hidden information from a database and transform it into an understandable structure for further use. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications [1].Throughout the years, many algorithms were created to extract knowledge from large sets of data. There are several different methodologies to approach this problem: classification, association rule, clustering, etc. Here we will focus on classification methodology. Classification techniques are designed for classifying unknown samples using information provided by a set of classified samples.

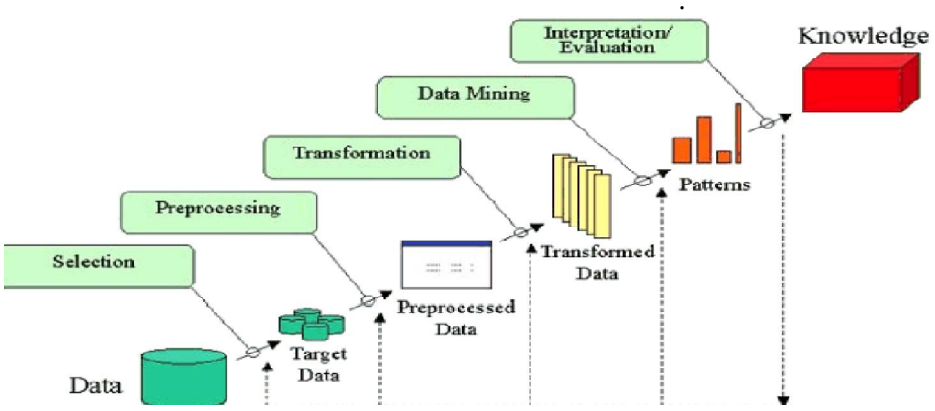

Figure 1: Data Mining Process

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

This set is usually referred to as a training set, because, in general, it is used to train the classification technique how to perform its classification. The classification task can be seen as a supervised technique where each instance belongs to a class, which is indicated by the value of a special goal attribute or simply the class attributes. Classification routines with data mining use a variety of algorithms and the particular algorithm used can affect the way records are classified. This work talks about Naive Bayes [5] classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. Despite of its naive design and most likely conspicuous assumptions, Naive Bayes work much better in many complex real world situations.

II. LITERATURE SURVEY

From the research article , the researcher express that large amount of data which is collected and stored for analysis. Making appropriate use of these data often leads to considerable gains in efficiency and therefore economic advantages.

There are several applications of Data Mining techniques in the field of agriculture. The researchers implemented K-Means algorithm to forecast the pollution in the atmosphere, the K Nearest Neighbour is applied for simulating daily precipitations and other weather variables and different possible changes of the weather scenarios are analyzed using Support Vector Machines. Soil profile descriptions were proposed by the researcher for classifying soils in combination with GPS based technologies. They were applied K-Means approach for the soil classification. In a similar approach, crop classifications using hyper spectral data was carried out by adopting one of the data mining approach i.e. Support Vector Machines.

One of the researcher used an intensified fuzzy cluster analysis for classifying plants, soil and residue regions of interest from GPS based color images. In the agricultural science, clustering techniques are found in grading apples before marketing. Weeds were detected on precision agriculture. The researchers worked on rainfall variability analysis and its impact on crop productivity. The effect of observed seasonal climatic conditions such as rainfall and temperature variability on crop yield prediction was considered through an empirical crop model. Furthermore, there are two approaches to investigate the impact of climate change on crop production which include the crop suitability approach and the production function approach.

III. METHODOLOGY

In this work the experiments are performed using Rapid Miner 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. There accuracy is obtained by evaluating the datasets. Each algorithm has been run over the training dataset and their performance in terms of accuracy is evaluated along with the prediction done in the testing dataset. RapidMiner 5.3 is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is one of the world's most popular and most used open source data mining solutions. RapidMiner has a comfortable user interface, where in a process view analyses are configured. It uses a modular concept, where respective operators are used in the analysis process. These operators have input and output ports through which the operators can communicate with the other operators in order to receive input data or pass the data and generated models over to the following operator. In this way, the entire analysis process creates a data flow. K-Nearest Neighbor [10] makes predictions based on the outcome of the *K* neighbors closest to that point. Therefore, to make predictions with *KNN*, we need to define a metric for measuring the distance between the query point and cases from the examples sample. One of the most popular choices to measure this distance is known as Euclidean (1).

$$D(x,p) = \sqrt{(x-p)^2} \qquad (1)$$

Where *x* and *p* are the query point and a case of the examples sample, respectively. Since *KNN* predictions are based on the intuitive assumption that objects close in distance are potentially similar, it makes good sense to discriminate between the *K* nearest neighbors when making predictions. Let the closest points among the *K* nearest neighbors have

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

more say in affecting the outcome of the query point. This can be achieved by introducing a set of weights *W (2)*, one for each nearest neighbor, defined by the relative closeness of each neighbor with respect to the query point.

$$W(x, p_i) = \frac{\exp(-D(x,p_i))}{\sum_{i=1}^{k} \exp(-D(x,p_i))} \quad (2)$$

Where *D(x, pi)* is the distance between the query point *x* and the *i*th case *pi* of the example sample. The weights defined in this manner above will satisfy :

$$\sum_{i=1}^{k} W(x_0, x_i) = 1 \quad (3)$$

Thus, for classification problems, the maximum of y is taken for each class variables, as shown:

$$\max\left(y = \sum_{i=1}^{k} W(x_0, x_i) y_i\right) \quad (4)$$

Naive Bayes [10] classifiers can handle an arbitrary number of independent variables, whether continuous or categorical. Given a set of variables, X = {x1, x2, x3..., xd}, we want to construct the posterior probability for the event Cj among a set of possible outcomes C = {c1, c2, c3..., cd}. In a more familiar language, X is the predictors and C is the set of categorical levels present in the dependent variable. Using Bayes' rule:

$$p(C_j | x_1, x_2, \ldots, x_d) \propto p(x_1, x_2, \ldots, x_d) | C_j) \, p(C_j) \quad (5)$$

Where p(Cj | x1, x2, x3..., xd) is the posterior probability of class membership, i.e., the probability that X belongs to Cj. Since Naive Bayes assumes that the conditional probabilities of the independent variables are statistically independent we can decompose the likelihood of a product of terms:

$$p(X \mid C_j) \propto \prod_{k=1}^{d} p(x_k \mid C_j) \quad (6)$$

And rewrite the posterior as:

$$p(C_j \mid X) \propto p(C_j) \prod_{k=1}^{d} p(x_k \mid C_j) \quad (7)$$

Using Bayes' rule above, we label a new case X with a class level Cj that achieves the highest posterior probability.

IV. SYSTEM ARCHITECTURE

System Architecture is the conceptual design that defines the structure/behavior of a system. It defines the system components or building blocks and provides a plan which includes production and system development, that will work together to implement the overall system. The overall system architecture of our project along with interactions between system components is as shown in figure(2). Crop prediction is the art of predicting crop yields and production before the harvest actually takes place. Earlier yield prediction was performed by considering the farmer's experience on a particular field and crop. This work presents a system, which uses data mining techniques in order to predict the category of the analyzed soil datasets. The predicted category will indicate the yielding of crops. Crop forecasting or prediction is the art of predicting crop yields and production before the harvest actually takes place, typically a couple of months in advance.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

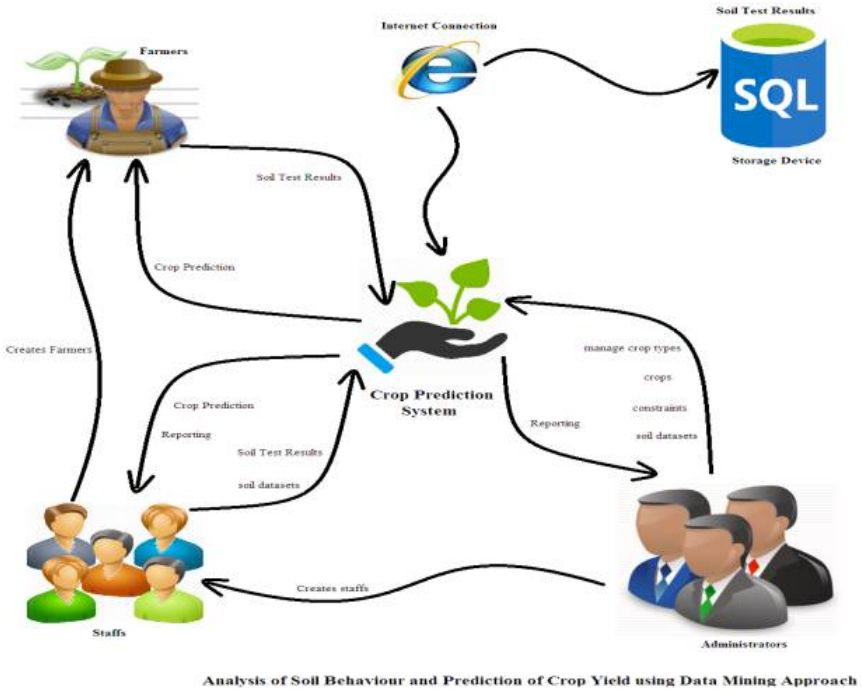

Figure (2): System Architecture.

Crop forecasting relies on computer programs that describe the plant-environment interactions in quantitative terms. The soil testing program starts with the collection of a soil sample from a field. The first basic principle of soil testing is that a field can be sampled in such a way that chemical analysis of the soil sample will accurately reflect the field's true nutrient status. The purpose of soil testing in high-yield farming is to determine the relative ability of a soil to supply crop nutrients during a particular growing season, to determine the needs, and for diagnosing problems such as excessive salinity. This work presents a system, which uses data mining techniques in order to predict the category of the analyzed soil datasets. The category, thus predicted will indicate the yielding of crops. The problem of predicting the crop yield is formalized as a classification rule, where Naive Bayes and K-Nearest Neighbor methods are used.

V. SYSTEM DESIGN

The purpose of the design phase is to plan a solution of the problem specified by the requirements document. This phase is the first step in moving from the problem domain to the solution domain. In other words, starting with what is needed; design takes us toward how to satisfy the needs. The design of a system is perhaps the most critical factor affecting the quality of the software; it has a major impact on the later phases particularly testing and maintenance.

The design activity often results in three separate outputs –

- Architecture design.
- High level design.
- Detailed design

Architecture Design:

Architecture focuses on looking at a system as a combination of many different components, and how they interact with each other to produce the desired result. The focus is on identifying components or subsystems and how they connect. In other words, the focus is on what major components are needed.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

High Level Design:

In high level design identifies the modules that should be built for developing the system and the specifications of these modules. At the end of system design all major data structures, file format, output formats, etc., are also fixed. The focus is on identifying the modules. In other words, the attention is on what modules are needed.

Detailed Design:

In the detailed design the internal logic of each of the modules is specified. The focus is on designing the logic for each of the modules. In other words how modules can be implemented in software is the issue. A design methodology is a systematic approach to creating a design by application of a set of techniques and guidelines. Most methodologies focus on high level design.

VI. EXPERIMENTS AND RESULTS

The Experiments are performed on the real world data obtained from the Soil Testing Laboratory in Jabalpur, Madhya Pradesh. Datasets considered in this work have sufficient amount of readings of nutrients and micronutrients taken from different lands of Jabalpur area. The dataset used in this experiment consists of 100 instances with 12 attributes. The tuples of dataset thus define the availability of nutrients and micronutrients in soil. Availability of these nutrients and micronutrients can be divided into different categories which can be used to decide its effects on the yielding capability of crops. The details of these categories are summarized in Table1. According to the soil science department of JNKVV Jabalpur soils falling under medium (M) category shows a good yielding capability. Soils of high (H) and very high category shows moderate yielding while the soils under low (L) and very low category shows poor yielding capability. A Training dataset is used in this experiment which an additional attribute along with all 12 attributes has called category (pre-defined by the laboratory). Category defines the quality of the particular soil with respect to the readings of nutrients and micronutrients of that soil.

Table 1
CATEGORIES OF NUTRIENTS AND MICRONUTRIENTES

Elements	Very low	Low	Medium	High	Very high
pH	<5.0	5.1 - 6.5	6.6 -7.5	7.6 - 8.0	>8.0
Organic carbon(OC) in %	<0.25	0.50-0	0.51-0.75	0.76 - 1.00	>1.00
Nitrogen (N) in kg/ha	<150	151 - 250	251 - 400	401 - 600	>600
Phosphorus (P) in kg/ha	<5	6 -10	11 -20	21 - 40	>40
Potassium (K) in kg/ha	<200	201 - 250	251 - 400	401 - 600	>600
Sulphur (S) in kg/ha	<10	11 - 20	21 -30	31 - 40	>40
Zinc (Zn) in mg/kg	<0.30	0.31 -0.60	0.61-1.20	>1.20	Not Defined
Iron (Fe) in mg/kg	Not Defined	<4.50	4.51 - 9.0	>9.0	Not Defined
Copper (Cu) in mg/kg	Not Defined	<0.20	0.21-0.40	>0.40	Not Defined
Marganese (Mn) in mg/kg	<1.0	1.0 - 2.0	2 -4	>4.0	Not Defined

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

VII. CONCLUSION AND FUTURE ENHANCEMENT

Conclusion: The system "crop prediction using data mining technology" is developed and tested successfully and satisfies all the requirement of the client. The goals that have been achieved by the developed system are:

- Simplified and reduced the manual work.
- Large volumes of data can be stored.
- It provides Smooth workflow.

Future Enhancements: We can add device to get values directly from soil testing lab to server. We can add module if any queries is there, the staff can directly interact with the administrator very easily.

REFERENCES

[1] Mucherino, P. Papajorgji, P.M. Pardalos, "Data Mining in Agriculture", Springer, 2009.
[2] Mucherino, Petraq Papajorgji, P. M. Pardalos ,"A survey of data mining techniques applied to agriculture", 25 May 2009 Springer-Verlag 2009 .
[3] Sally Jo Cunningham and Geoffrey Holmes,"Developing innovative applications in agriculture using data mining", Department of Computer Science, University of Waikato Hamilton, New Zealand.
[4] Cover TM, Hart PE,"K Nearest Neighbor pattern classification", IEEE Trans Info Theory 13(1) : 21-27, 1967.
[5] P.Bhargavi, Dr.S.Jyothi, "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.8, August 2009.
[6] Vishnu Kumar Goyal, "A Comparative Study of Classification Methods in Data Mining using RapidMiner Studio", (IJIRSE) International Journal of Innovative Research in Science & Engineering.
[7] Georg Rub, Rudolf Kruse, Martin Schneider and Peter Wagner, "Data Mining with Neural Networks for Wheat Yield Prediction".
[8] Shweta Taneja, Rashmi Arora, Savneet Kaur, "Mining of Soil Data Using Unsupervised Learning Technique", International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 7 No.11, 2012.
[9] M.C.S.Geetha," Implementation of Association Rule Mining for different soil types in Agriculture", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 4, April 2015.
[10] M.Soundarya, R.Balakrishnan," Survey on Classification Techniques in Data mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014.
[11] D Ramesh , B Vishnu Vardhan, "Data mining technique and applications to agriculture yield data", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013 .
[12] Gideon O Adeoye, Akinola A Agboola, "Critical levels for soil pH, available P, K, Zn and Mn and maize ear-leaf content of P, Cu and Mn in sedimentary soils of South- Western Nigeria", Nutrient Cycling in Agroecosystems, Volume 6, Issue 1, pp 65-71, February 1985.
[13] D. Almaliotis, D. Velemis, S. Bladenopoulou, N. Karapetsas, "Appricot yield in relation to leaf nutrient levels in Northern Greece", ISHS Acta Horticulturae 701: XII International Symposium on Apricot Culture and Decline .