ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.10

An Analysis of Agricultural Soils by using Data Mining Techniques

Ramesh Babu Palepu[1], Rajesh Reddy Muley[2]
Associate Professor[1], Assistant Professor[2]
Department of CSE
Amrita Sai Institute of Science & Technology, Paritala, India

Abstract:
Agriculture is the most basic function to accomplish food demand all over the globe; it is a backbone particularly in the developing countries like India. The application of Data mining techniques in agriculture especially on soils can revise the situation of pledge making and improve cultivation yields in a better way. The analysis of soils plays an indispensable role for resolution making on several issues related to agriculture field. This paper presents about the role of data mining in perspective of soil analysis in the field of agriculture and also confers about several data mining techniques and their related work by several authors in context to soil analysis domain. The data mining techniques are of very up-to-the-minute in the area of soil analysis.

Keywords: soils, barren, Data mining Techniques

1. INTRODUCTION

In the current days of society, data mining is used in a massive areas and many off-the-shelf data mining tools, techniques and procedures are available and sphere of influence data mining application software's are reachable, but data mining in agricultural soil datasets is a comparatively a infantile research field. Now a day's data mining concept and techniques used to resolve the agriculture problems. In this paper it has been discussed about how data mining techniques are applied in agriculture field. Globally, day to day the requirement of food is escalating; hence the agricultural scientists, farmers, government, and researchers are tiresome to put extra attempt and use numerous techniques in agriculture for improvement in production. As an effect, the data generated in the field of agricultural data enhanced day by day. As the degree of data enlarges, it requires instinctive way for these data to be mined and analyzed when needed. Even at present, a very only some farmers are really using the new methods, tools and techniques in agriculture for better production. Data mining can be used for forecasting the future trends of agricultural processes [1].

Data mining is a guiding principle that results in the penetration of new prototypes in large data sets. The main objective of the data mining process is to extract knowledge from an existing data set and modernize it into a human understandable form for advance use. Data mining is the methodology of analyzing data from different viewpoints and makes it summation of useful information. Data mining can analyze versatile data there is no restriction on the type of data [2].

Data mining has been classified into two types such that one is descriptive another one is predictive. Descriptive data mining considers the existing data, that is raw data and then make it summarized. The descriptive mining represents the characteristics of the past events and allows us to learn how they influence the future. The foundation of predictive mining depends on probabilities, it is used to predict future based on the values considered from known results. Forecasting involves using the variables or field in the database to estimate anonymous results [3].

1.1 Data Relationships Data mining is a process of action for collecting data from heterogeneous sources, analyzing the data to generate useful information from it and then concise this information. Data mining software is considered as an analytical tool for analyzing data in many different dimensions, perspectives, or angles, and categorizing the analysis, and summarizing the relationships among these categories are identified. There are four types of data relationships:

1) Classes Data is stored in the form of classes. A data class consists of a data fields and some basic methods to access those data fields. The fields in a data class have similar characteristics.

2) Clusters A data cluster is a labeled portion of data containing similar items. It also refers as a method of partitioning a set of data into a set of sub-classes each of which has a specific meaning, called clusters. It helps users to understand the accepted alignment or structure of the data in set. Clustering comes under unsupervised learning.

3) Association It is one of the data mining techniques that determine the likelihood of the co-occurrence of data items in a data set. The relationships between co-occurring data items are expressed as association rules [4].

4) Sequential patterns It is used to finding statistically appropriate patterns between data elements where these data element values are distribute in a sequence [5].

Data mining provides the bridge between the different transaction and analytical systems. Using open-ended user queries data mining software analyzes relationships and patterns in stored data. Some of the analytical software include; machine learning, statistical, and neural networks.

1.2 Elements of Data Mining Data mining consists of five major essentials

a) Extract, transform, and load transactional data onto the data warehouse system.
b) Store and manage the data in a multidimensional database system.
c) Provide information access to business intelligence and technology professionals.
d) Analyze the data by application software.
e) Present the data in visual formats, such as a graph or table.

1.3 Levels of Analysis Data mining provides various levels of data analysis, some of them are;

1) Artificial neural networks: ANNs has huge potential to solve some problems fall under prediction, regression, pattern reorganization, and classification of data in data mining [6].

2) Genetic algorithms: Genetic algorithms are extensively used in data mining applications such as classification, feature selection, clustering etc. Some genetic algorithms such as Fuzzy classification and GA Tree is used for supervisor learning, some other such as Fuzzy C-Means algorithm is used classification of soil data based on soil properties. (Bhargavi and Jyothi 2011) [7].

3) Decision trees Decision trees represent sets of decisions; it is a predictive model of Data mining technique, which is a tree-shaped structure. It is used to classify data sets by mapping the observations of data item with that data item target values.

Decision trees are classified into two types, one is classification trees in which the target variable can take a finite set of values, another one is regression trees in which target variable can take continuous values (real numbers).Naturally the unclassified data sets are classified based on, Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART is used to split a dataset by creating two-way splits and CHAID splits a dataset using chi square tests to create multi-way splits. The advantage of CART is it requires less data preparation than CHAID [8].

4) Nearest neighbor method one of the methods used in Data mining for classification and regression is k-nearest neighbor algorithm (k-NN). KNN is a non parametric instance-based or lazy learning algorithm. Non parametric, means the data items are taken from the unspecified probability distribution. Hence it does not have assumptions on the underlying data distribution. In most of the experiments theoretical assumptions are not coincide with most of the practical data. In consequence KNN is most useful as in the real world. It is also a lazy algorithm. This means that it does not made any generalization until query is posed on training data points. In other words, there is no precise training phase or it is very nominal. This means the training phase is good fast, due to lack of generalization the KNN keeps all the training data. More accurately, all the training data is needed during the testing phase [9].

5) Rule induction Rule induction is one of the fundamental tools of data mining. It is used to find out regularities hidden in data are expressed in terms of rules. Naturally rules are expressions of the form if (attribute − 1, value − 1) and (attribute − 2, value − 2) and ⋯ and (attribute − n, value − n) then (decision, value). Some rule induction systems consists more complex rules, in which values of attributes may be expressed by reversal of some values or by a value subset of the attribute domain [10].

6) Data visualization Sometimes data relationships are represented with the help of graphical techniques such as geometric techniques, distortion techniques, hierarchical techniques, pixel oriented techniques etc. These techniques present visual interpretation of complex relationships in multidimensional data [11].

1.4 Evaluation Process of Data of Data Mining

Data mining is **"**The nontrivial process of identifying suitable, original, useful and eventually reasonable form of data". Generally the data is stored in various types of repositories such as files, databases, data warehouse etc, that stored data is progressively more important to build controlling mechanisms for exploration and rationalization. The data also used for the extraction of interesting knowledge that could help in decision-making. The bellow diagram shows data mining as a step in an iterative knowledge discovery process [12].

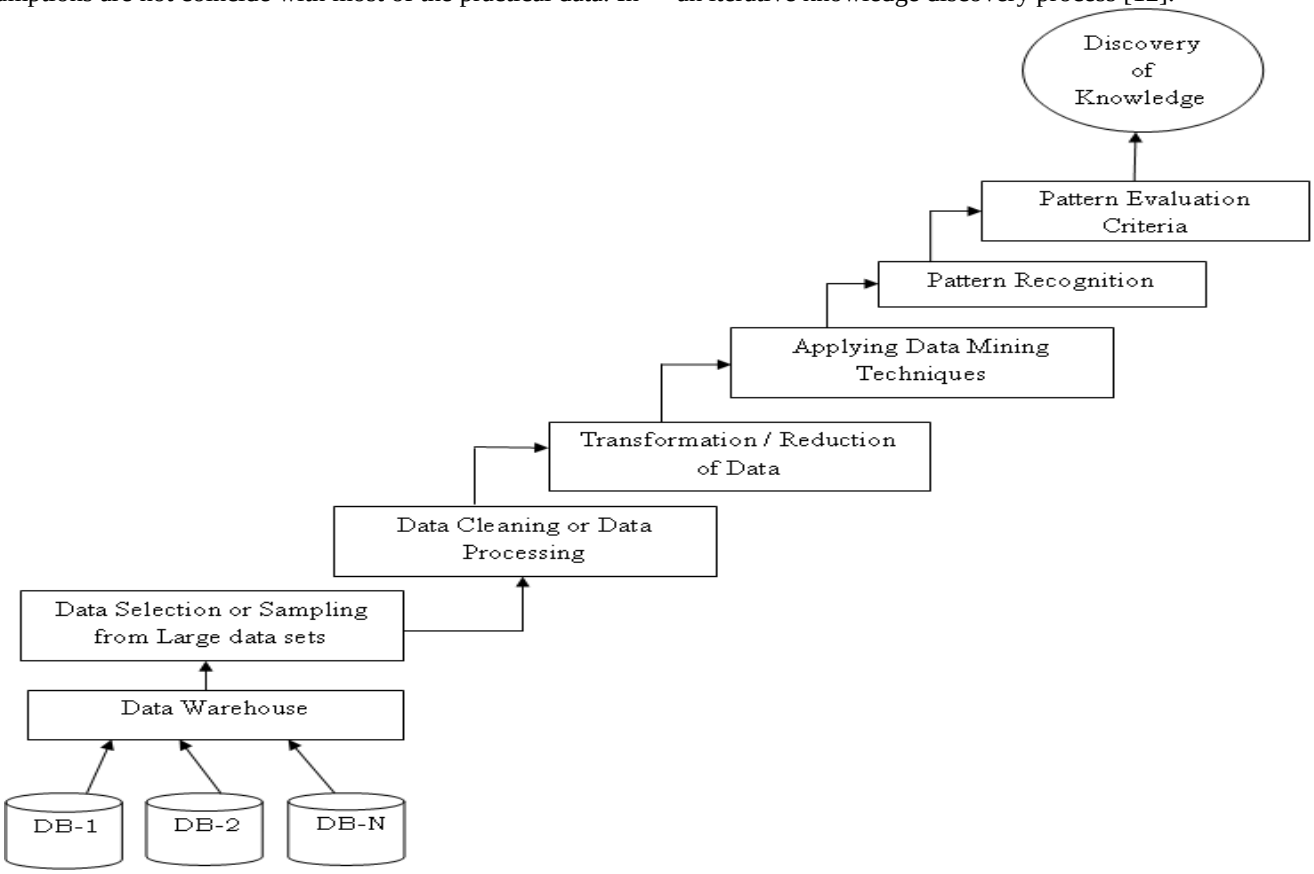

Figure 1. 4 Iterative Knowledge Discovery Process

1.4.1 Data Collection Methods

Data collection is a fundamental activity to discover useful patterns from the data by applying the dissimilar machine learning techniques. The following are some natural the data collection methods;

(i) Observations
(ii) Surveys and Questionnaires
(iii) Document reviews
(iv) Interviews
(v) Focus groups
(vi) Case studies
(vii) Illustrated presentations
(viii) Collection Methods
(ix) Other visual representations

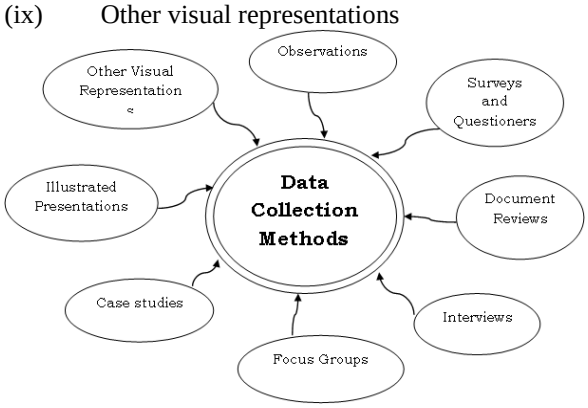

Figure .2.1.4.1 Data

The data gathered from heterogeneous sources are categorized, integrated, and interrelated then stored in information repositories such as Files, Data bases, data marts etc., for further processing or mining.

1.4.2 Knowledge discovery in data mining

Data mining is used dissimilar methods to obtain the significant information from source data. Significant patterns and hidden associations without human intervention were determined by data mining from massive amount of data. Stated by Mukesh Kumar and Arvind Kalia (2006) [13]. As stated by Abdulsalam S. O., Adewole, K. S., Bashir, S.A., Jimoh, R.G. & Olagunju, M. (2012) [13], KDD is a process of applying appropriate techniques and methods of machine learning to acquire the constructive information from enormous collection of data. In this process low-level data is transformed into other forms, which are very compact, abstract and ready to useful. Creating short reports, modeling the process of data and predict the data models that to be used in future are generated by using the KDD.

Data mining affords you with insights and correspondence between data items, that had officially gone unrecognized or been ignored because it had not been considered possible to investigate them. The following steps are in the process of data mining.

Data cleaning Is the data preprocessing step, in this phase we make the data which has accuracy, completeness, consistency, timeliness, etc, by removing irrelevant data from the collection. It simply a data smoothing process.

Data integration In this preprocessing step data is accessed from disparate sources which may be in inconsistent form and then combined into a common source.

Data selection In this step, the relevant data required for task analysis is taken out from unified data source. In this process, sometimes data consolidation and transformation may be takes place.

Data transformation Here by performing summary or aggregation operations, we select the data that is in appropriate form for the mining process.

Data mining This is the crucial step in which intellectual techniques are applied to extract patterns which are potentially valuable.

Pattern evaluation This step based on given measures, purely interesting patterns representing knowledge are identified.

Knowledge representation This is the final phase discovered knowledge is visually represented to the user. This essential step uses revelation techniques to help users understand and interpret the data mining results. The following are concision of KDD process;

Step-1 Data is gathered from heterogeneous sources is integrated into a single data store.

Step-2 the data is pre-processed and turned into standard format.

Step-3 To generate patterns or rules as a output, data mining algorithms used.

Step-4 The generated rules and patterns are transformed to useful knowledge or information.

1.4.3 Data Mining Algorithms

To extract the knowledge, data mining algorithms first analyze the data then generate the meticulous types of patterns. The data mining model can use this result to select and define the optimal parameters. These parameters are used to extort the significance full information. The mining model of an algorithm produces from our data, it can take various forms, such as in the form of clusters, decision trees, mathematical models, and as a set of regulations [14]. The following are some mining and machine learning algorithms in brief.

1.4.3.1 Classification: Any data mining or Machine learning algorithms can follow three learning approaches: that are, supervised learning, unsupervised learning, and Semi-supervised learning. The classification algorithms are falling into the category of supervised learning. The classification is a two step process. In the first step, by analyzing the data tuples described by their attributes within the predetermined data set we build a model. In this case class labels are predetermined. In the second step we estimate the analytical accuracy of the built-in model used for classification. Classification and prediction are two forms of data analysis that can be used to make data models for important data classes or to predict future trends of data. We also use prediction to evaluate the unlabeled data samples through construction and use of a data model. Either to apply classification or prediction, we can perform preprocessing of data such as data cleaning, data transformation and relevance analysis of data. The results of the classification and prediction models are evaluated according to theirs scalability, speed, predictive accuracy, robustness and interpretability. The different classification techniques for discovering knowledge are Rule Based Classifiers, Bayesian Networks (BN), Decision Tree (DT), Nearest Neighbor (NN), Artificial Neural Network (ANN), Support Vector Machine (SVM), Rough Sets, Fuzzy Logic, and Genetic Algorithms [17].

1.4.3.2 Regression is a data mining function which is used to predict a numeric or continues value. Both of the techniques classification and regression exhibits the similar functionality (predictive analysis), but classification classify the data into discrete sets. The regression is used to establish the relationship between dependent variables versus independent variables. It maps a data item to a real-valued prediction variable. Regression also represents the variations occurred in one variable in accordance with variations in another variable. One important issues regarding regression is it describe the relationship among the variables in detail than correlation – it represents the strength of the relation between the variables. Regression tasks are often treated as classification tasks with quantitative class labels. The methods for prediction are Linear Regression (LR) and Nonlinear Regression (NLR) [25]. The following is an example for linear regression equation.

Regression Equation (y) = a+bx

Slope (B) =($\sum XY - (\sum X)(\sum Y) / (N\sum X^2 - (\sum X)^2$
Intercept(a) = ($\sum Y - b(\sum X)) /N$

1.4.3.3 Association Rule Mining Agrawal, Imielinski, & Swami in 1993[22] discovered the association rule technique. which finds the correlation and association relationship among massive data items. The association rule mining is represented as, a set of transactions, where each transaction contain set of data items, now an association expression between the items A and B is in the form A => B, the meaning of such a rule is that transactions which contain A also contain B within the database [23]. When we use association rules, we should consider two important basic rule measures such as **Support** and **Confidence.** These two measures have some threshold value to mine interesting patterns from huge collection of data. Naturally the support is the realistic value and confidence is an estimated value. The rules which are mined are not satisfying the threshold value is called uninteresting. To measure the interestingness of an association rule, the rule must be simple, certainty, and utility. The rule basic measure support and confidence between the items A and B is defined as

$$Confidence\ (A \Rightarrow B) = \frac{\#Transactions\ containing\ both\ A\ \&\ B}{\#Transactions\ containing\ A}$$

$$Suppot\ \left(A \Rightarrow B\right) = \frac{\#Transaction\ containing\ both\ A\ \&\ B}{\#Total\ number\ of\ Transactions}$$

In association rule mining, the confidence value is 100% then the analyzed data is always correct; such a rules are called **exact.** Strong association rules have the minimum support threshold and minimum confidence threshold. The association rules are categorized depending on the following principles:

a) Data dimensions involved in the association rule.
b) Type of data values in the rule.
c) Level of abstraction of data incurred in the rule.
d) Depending on other extensions of data.

An association rules exposed in various areas like, market basket analysis, customer segmentation, catalog design, store layout and telecommunication alarm prediction. The different association rule mining algorithm are Apriority Algorithm (AA), Partition, Dynamic Hashing and Pruning(DHP), Dynamic Item set Counting (DIC), FP Growth (FPG), SEAR, Spear, Eclat & Declat, MaxEclat [24].

1.4.3.4 Clustering Is an unsupervised learning technique, to find the likelihood between unlabeled data elements into groups. In clustering, the data elements with the groups are more relevant to each other and differ with data elements in other groups. Basically, clustering is two kinds; one is conventional clustering and other is conceptual clustering. The following are some characteristics should have a good clustering algorithm.

a) A cluster algorithm should easily interpretable and usable.
b) It should have high dimensionality.
c) It is able to partition constrained oriented data.
d) It should be scalable.
e) It is able to produce uninformed shaped clusters.
f) It is capable to deal the data which has different kinds of attributes.
g) It should take less domain knowledge.
h) It should handle noise data.

The clustering is used to determine a new set of groups, these groups are of significance in themselves, and their assessment is intrinsic [18]. The different clustering methods are Hierarchical Methods(HM), Partitioning Methods (PM), Density-based Methods(DBM), Model-based Clustering Methods(MBCM), Grid-based Methods and Soft-computing Methods [fuzzy, neural network based], Squared Error—Based Clustering (Vector Quantization), Clustering graph and network data etc.. [19,20,21].

1.4.3.5 Sequence analysis algorithms Today the most challenging area in machine learning is to extracting hidden patterns from the sequential data. This means, patterns are shared among data objects. We can perform sequential mining either on single data sequences or multiple data sequences. Sequential mining is more relevant to the temporal databases. The main limitation of sequential patterns is, there is no probability assessment followed by a pattern.

Data mining represents a stated removal of hidden information from large files. This is a new technology with great possible to help companies focusing on the most important information in their large data [3]. Data mining uses different methods for the purpose of acquiring necessary information. Different methods are used for different purposes, where each method has its merits and limitations. Data mining tasks can be divided into descriptive and predictive. An illustrative tasks aim to find human understandable patterns and associations, after considering data as well as the entire model establishment, predicting tasks aim to predict some response of interest [15]. The main techniques for data mining include Association rules, Classification, Clustering and Regression [3]. The different data mining techniques used for solving different agricultural problem has been discussed [16]. The graphical representation of different data mining techniques is shown bellow.

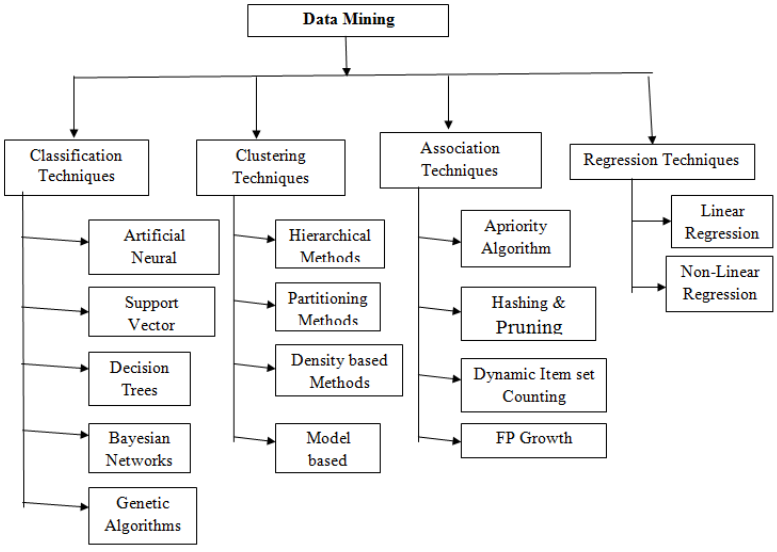

Figure.3.43 Data Mining Techniques

2. LITERATURE REVIEW

2.1 Introduction We know that data mining is applied to analyze huge data sets and found useful prototypes in the data. Numerous studies belongs to different fields has used a variety of techniques of data mining and machine learning to predict, to rectify, to enhance or to gain knowledge in those fields. At now we analyze agricultural data especially soil information by using classification, regression, correlation, clustering, natural trees, and statistical machine learning and other analysis methods. The results of soil analysis on different data sets with a range of data mining techniques may useful to farmers to get right insight to perform their activities with less cost and to improve the crop yields, such as by measuring soil properties the formers decide what kind of crops to be adopted and use of fertilizers etc. By analyzing the previous data the formers may also has knowledge about market basket analysis. Different mining and machine learning software applications includes in different methodologies that has been developed either by commercial or by research institutes. Some of the developed techniques are in used by the industry, by business people, or by the formers in multiple dimensions. Now a day the researchers, data analysts and scientists has more concentrated on how mining and machine learning techniques are used to analyze various soil profiles to enrich the field of agriculture [26]. The soil analysis may useful in many dimensions such as, to protect environment, diagnosis of crop culture troubles, to identify nutrient deficiencies, energy conversation, and so on. In soil analysis we can test different properties of soils like pH, organic matter, ammonium N, calcium carbonate equivalency, etc.

2.2 Crop Productivity Mr. Narsi Reddy Gayam stated in his research learning "A study of crop yield distribution and crop insurance" which takes the input data from INDIA relating sugarcane and Soybean. He discovered that proposition of predictability of crop yields. The intensive data qualitatively analyzed by using Lilliefore method, here he considered unfounded hypothesis are normally distributed. The actual results indicate the considerations of the hypothesis in all cases are not true. Hence he concludes crop yield are not normally distributed [27]. The result found by Mr.NR Reddy is very much useful to estimate risk management implicated in sugarcane and soybean crops.

2.3 Application of DM Techniques in Agriculture Dr. Bharat Misra, et al., [28] observed the research studies on application of data mining techniques in the field of agriculture. Some of the techniques, such as ID3 algorithms, the k-means, and the k-nearest neighbor, artificial neural networks and support vector machines applied in the field of agriculture were presented. Data mining in application in agriculture is a relatively new approach for forecasting / predicting of agricultural crop/animal management. The study explores the applications of data mining techniques in the field of agriculture and allied sciences. Historical crop yield information is important for supply chain operation of companies engaged in industries that use agricultural produce as raw material. Livestock, food, animal feed, chemical, poultry, fertilizer pesticides, seed, paper and many other industries use agricultural products as intergradient in their production processes. An accurate estimate of crop size and risk helps these companies in planning supply chain decision like production scheduling. Business such as seed, fertilizer, agrochemical and agricultural machinery industries plan production and marketing activities based on crop production estimates.

2.4 Spatial DM yashovardhankelkar, et al,[29]. Surveyed and says that data selection is the data relevant to the analysis is decided and retrieved from the various data locations. Data preprocessing is the process of data cleaning and data integration is done. Data cleaning is also known as data cleansing; in this phase noise data and irrelevant data are removed from the collected data. Data integration is multiple data sources, often heterogeneous, are combined in a common source. In Data transformation the selected data is transformed into forms appropriate for the mining procedure. Data Mining is the crucial step in which clever techniques are applied to extract potentially useful patterns. The decision is made about the data mining technique to be used. Interpretation and Evaluation is interesting patterns representing knowledge are identified based on given measures. The discovered knowledge is visually presented to the user [30].

2.5 Soils Clustering A. Banumathi, A. Petalakshmi (2012) has explained how data mining can be applied on large datasets to discover patterns using the method of clustering. They have analyzed the Fuzzy C-Means algorithm and have made inferences that initial seed value selected either sequentially or randomly have effect on the value of the ensuing cluster. Amrender kumar (2004) [31] explains the techniques for forecasting of crops. The data mining application agriculture use more than one application for implementing the results. The various data mining and neural network techniques are used for forecasting the result. These techniques are decision tree, rule induction, navie-bayes, neural network, ANN, radial basis function, recurrent network, multilayer perceptron and RBF techniques and predict the result. Anwiti Jain, Anad Rajavat, Rupali Bhartiya (2012) [32] has explained about the clustering mechanism which is an unsupervised technique of learning. The clustering enables to identify groups based on attributes values. The groups are uniform in terms of objects they contain in them. In their research they have pointed out the use of K means clustering algorithm with modification to find the cluster's centre. They have applied the algorithm to very large data sets. They have shown that clustering has many ways of application in data mining except the way they have used. Use of iterative clustering mechanism has influence of the cluster centre chosen for each iteration. They have used the approach of optimization formulation of problem in designing the algorithm together with novel iterative method. The research paper shows improvement in cluster centre detection when tested on large random datasets. Dr. Rajesh (2011) explains the application of data mining in agriculture. Data mining is the great technique currently used in agriculture and industries. In this research the researcher explain the k-means clustering to classify the patterns. In this research they discuss the particular area and analysis on this areas agriculture patterns and obtain the required result. Association technique is used for the clustering A number of studies have applied data mining techniques to extract meaning from data collected from natural systems research. For example, the collection of data from natural systems is challenging, with most of the data sets incomplete due to the difficulty and methods of data collection. Missing data sets can be problematic and may limit the analysis and extraction of new knowledge. The problem of missing values was analyzed by Ragel and Cremilleux (1999, p.1): "To complete missing values a solution is to use relevant associations between the attributes of the data. A number of studies have been carried out on the application of data mining techniques for agricultural data sets. For example, a study by Ibrahim (1999) on a sample data set applied six classification algorithms to 59 data sets and then six clustering algorithms were subsequently applied to the data generated. The results

were studied and the patterns and properties of the clusters were formed to provide a basis for the research. The research provided a comparison of performance for the 6 classification algorithms set to their default parameter settings. It was found that Kernel Density, C4.5 and Naïve Bayes followed by rule learner, IBK and OneR were the most accurate. The study utilized the WEKA data mining benchmark program. The main objectives of the research conducted by Ibrahim (1999) was to apply unsupervised clustering to the file built in step 1 to analyze the generated clusters and determine whether there are any significant patterns. Ibrahim (1999, p. 2) outlined a number of findings: It was discovered that number of instances was not useful in clustering the data sets, as it was the only significant variables in clustering the data sets before it was excluded from the generated data set. This prevented analysis based on other variables including the variables that contain values for the accuracy of each classification algorithm. The research conducted by Ibrahim (1999) has provided a platform from which further work in this field might be undertaken. The scope of the research was limited and the investigation revealed a number of interesting clusters in machine learning performance data. It can be concluded that a larger investigation is required which uses more data sets and data set characteristics. In another study WEKA was used to develop a classification system for the sorting and grading of mushrooms (Cunningham and Holmes, 1999). The system developed a classification system that could sort mushrooms into grades and attained a level of accuracy equal to or greater than the human inspectors. The process involved the pre-processing of the data, not just cleaning the data, but also creating a test dataset in conjunction with agricultural researchers. The attributes used to create the set included both objective and subjective measurement. The total dataset used a total of 282 mushroom types, criteria and attributes. The objective attributes were weight, firmness and percentage of cap opening. The subjective attributes were used to estimate the degree of dirt, stalk damage brushing, shrivel and bacterial blotch. The above data was collected and then compared with the grading of the three human inspectors and allocated a grade 1st, 2nd or 3rd. The data, a total of 68 attributes including photo images, was used by the j4.8 algorithm classifier within WEKA to create a model for the human inspectors and the automated system. The model created using the human rules showed that each inspector used different combinations of attributes when assigning grades to mushrooms (Cunningham and Holmes, 1999). The application of data mining techniques provided within the WEKA software application created a model that analyzed all attributes and created a model that was faster and more accurate than the human system. The decision tree analysis method has been used in the prediction of natural datasets in agriculture and was found to be useful in prediction of soil depth for a dataset. In Mckenzie and Ryan (1999) the uses of slope angle, elevation, temperature and other factors were analyzed. and models created for prediction of soil depth across a sample area. The model was tested through the use of random data sets. "at each level, trees with increasing numbers of terminal nodes were fitted 20 times with 5% of the data randomly selected and withheld to provide a test of the predictive strength of the model" (Mckenzie and Ryan, 1999). [33]

Table.1. Description of DM Methods and used in Agriculture

S.No	Worked done on	Outcome	DM Techniques Used	Name of Biographer
1	Classification and Prediction of Future Weather by using Back Propagation Algorithm An Approach	Focuses on weather forecasts	Neural Networks	SanjayD.Sawaitul, Prof. K.P.Wagh & Dr.P.N.Chatur
2	Modeling and prediction of rainfall using artificial neural network and ARIMA techniques	Prediction of rainfall	Neural Networks	V.K.Somvanshi, et al.,
3	An investigation into the application of neural networks, fuzzy logic, genetic algorithms, and rough sets to automated knowledge acquisition for classification problems	Classifying soil in combination with GPS	K-means	I. Jagielska C. Mattehews, T. Whitfort
4	A vision-based hybrid classifier for weeds detection in precision agriculture through the Bayesian and Fuzzy k-Means paradigms	Wine fermentation	K-means	Tellaeche, X. P., Pajares, A.,BurgosArtizzu, G., & Ribeiro
5	High resolution continuous soil classification with morphological soil profile descriptions	Prediction of yield in Agriculture	Fuzzy set	K.Verheyen, D. Adriaens, M.Hermy, and S. Deckers
6	Using data mining techniques to predict industrial wine problem fermentations	weeds precision detection in agriculture	Fuzzy set	Urtubia, A., Pérez Correa, J. R., Soto, A., & Pszczolkowski, P.
7	Data mining Techniques for Predicting Crop Productivity	kharif and rabi crops production effected by climatic factors	Decision Trees	S.Veenadhari, Dr. Bharat Misra, Dr. CD Singh
8	Downscaling of precipitation for climate change scenarios: a support vector machine approach	Using Bayesian network learning method developing the model for agriculture	Bayesian network	Tripathi, S., Srinivas, V. V., & Nanjundiah, R. S.
9	Crop productivity mapping based on decision tree and Bayesian classification	Scale back procedure on crop yielding using k-nearest neighbor algorithms	KNN	Veenadhari, S.

10	Unsupervised neural network approach to medical data mining techniques	Daily precipitations simulation of weather and other conditions	KNN	Shalvi D & De Claris N,
11	Data Mining Techniques in Agricultural and Environmental Sciences	Classifying the weather sample information into linearly severable	Support Vector Machine	Altannar Chinchulunn, Petros Xanthopoulos, Vera Tomaino, P.M.Pardalos
12	A K-nearest neighbor simulator for daily precipitation and other weather variable	Conducting climate impact studies	Support Vector Machine	B. Rajagopalan & U. Lal

3. ANALYSIS OF SOILS:

Soils are formed by the combination of weathered rock materials with humus. We also know that soils supplies water and nutrients to flora. Soils also protect and purify rain water, pests and wildlife habitation. A little bit of hydroponics (soilless agriculture) are developed throughout the globe, but still their percentage is only 10.6 of total soils under conservation. The country prosperity is depending upon soils of that country [34]. Soil analysis and their classification is very critical because in the entire globe the types of the soils are same, but their analysis results may vary from location to location depending on various characteristics of soils. When analyzing the soils, we should consider the fundamental substantial, organic and compound properties of soils. The classification of soils deals with the methodical cataloging of soils depends on their individual characteristics, as well as decisive factor that dictate choice in use. Classification of Soils is one of the challenging areas in data mining and machine learning, In classification of soils can be started from the viewpoint of soils as a matter and soil as a resource [33]. The following are various types of soils in India.

3.1 Alluvial Soils These soils are produced by the deposition of sediments carried by rivers. These soils are highly loaded with humus and very much productive. They are found in Great Northern plain, lower valleys of Narmada and Tapti and Northern Gujarat. These soils are transformed year by year.

3.2 Black Soils The basic material of black **s**oils volcanic rocks and lava-flow. These soils are spread over Deccan Lava Tract which includes parts of Maharashtra, Chhattisgarh, Madhya Pradesh, Gujarat, Andhra Pradesh and Tamil Nadu. These soils have huge amount of clay, it may be more than 62%. These soils have chemical composition of alumina, iron oxide, magnesium carbonates and lime and also Potash but are short of in Phosphorus, Nitrogen and Organic matter.

3.3 Red Soils The color of these soils is due to heavy presentation of iron oxide. Acidic content in these soils are high, and phosphates and nitrogen is less. These soils are unable to retain moisture and water. These soils are formed due to slow breaking of metamorphic rocks and crystalline. They spread over the whole of Tamil Nadu, Andhra Pradesh, Chhattisgarh, Karnataka, Maharashtra and parts of Orissa.

3.4 Literate Soils These soils are formed where heavy temperatures and heavy rainfalls occurred alternatively, normally these soils are final creation of weather. These soils contain good composition of bauxite or ferric oxides and less composition of potash and nitrogen. These are spread around in Kerala, Tamil Nadu, Maharashtra, Chhattisgarh and hilly areas of Orissa and Assam.

3.5 Mountain Soils These soils are formed at hill slopes by gathering of organic matter derived from forest and woodlands. They are found in Himalayan region and also appeared in other regions according to altitude. These soils characteristics are based on climate, ground configuration and on parent rocks.

3.6 Desert Soils In the desert regions of Rajasthan, soils are not well developed. As evaporation is in excess of rainfall, the soil has a high salt, alkaline content. These soils are naturally sandy and have less organic matter.

Table .2. Brief details of soils in India

Nature Of Soil	Formation	Characteristics	Section of Country	%	Crops grown
Black Cotton Soil (or) Regur Soil	a) Is of volcanic origin b) Lava soil due to disintegration of basalt, formed in the area where it has formed. c) It is also classified as Cheen ozem	a) Deep, fine grained b) Varying in colour from black to chestnut brown. c) Rich in Iron, Potash, Lime, Calcium, Alumina, Carbonates & Humus. d) Moisture retentive, very sticky when wet. e) Forms deep cracks when dry.	Occurs mainly in Deccan trap covering large areas in Maharashtra, Gujarat, M.P, Karnataka, Andhra Pradesh & Tamil Nadu.	29.6 %	Cotton, Jowar, Wheat, Sugarcane, Linseed, Gram, Fruit & vegetable.
Red Soil .	a) Formed by weathering of crystalline and metamorphic rocks. b) Mixture of clay and sand	a) Red in colour because of its high Iron-Oxide (FeO) content. b) Deficient in nitrogen, lime, phosphoric acid and humus. c) Rich in Potash.	Large parts of T.N. Karnataka, North East Andhra, M.P & Orissa.	28%	Wheat, Rice, Millets, Pulses, (needs fertilizer and irrigation)
Laterite Soil	Formed due to weathering of lateritic rocks in low temperatures and heavy rainfall with alternating dry & wet period.	a) Red in colour because of its high Iron-Oxide (FeO) content. b) Poor in Nitrogen & Lime, rich in Iron. c) High content of acidity and inability of retain moisture.	Karnataka, Summits of the Western and Eastern Ghats Malwa Plateau, Goa & Kerala.	2.62 %	Unsuitable for agriculture due to high content of acidity and inability to retain moisture. Cashew and tropical grow well on it.

Arid & Desert Soil	Formed due to accumulations of calcium carbonate, gypsum, and/or salts.	Rich in Phosphates but poor in Nitrogen.	NW India. Covers entire area west of the Aravalis in Rajasthan & parts of Haryana, Punjab & Gujarat.	6.13 %	Fertile if irrigated. e.g: Ganga Nagar area of Rajasthan (Wheat basket of Rajasthan).
Saline & Alkaline Soil	Formed by an excess of sodium salts and high proportion sodium in redeemable complex.	a) Soils have effervescence of Sodium, Magnesium, Calcium. b) Salinity is usually confined to the upper layers and the soil can be reclaimed by improving drainage. c) Alkalinity is removed by application of Gypsum.	Arid and Semi-Arid areas of Rajasthan, Punjab, Haryana, Bihar and Uttar Pradesh.	1.29 %	Infertile, requires Soil-reclamation.
Forest Soil	The formation of soils depends upon type of vegetation grows.	Rich in organic matter. a) In some places it shows sign of Podzolisation. b) Deficient in Potash, Phosphorus & Lime. c) Needs continued use of fertilizers for good yields.	Himalayan Range, Southern hills of Peninsula.	7.94	Plantation Crops like tea, coffee, spices and tropical fruits.
Peaty and other Organic Soil	Formed with Bitumen, Humic acids, Lignins, cutin etc.	a) High accumulation or Organic matter & small amount of soluble salts. b) Deficient in Phosphorus & Potash.	Peaty Soil - Found in Kottayam and Alleppey district of Kerala. Marshy Soil - Coastal areas of Orissa, W.B, T.N, North Bihar & Almora (U.P).	2.17	Not conductive to cultivation.

4. BARREN SOILS

Those ecosystems in which less than one third of the area has vegetation or other cover. In general, Barren soil has thin soil, sand, or rocks. Barren soils include deserts, dry salt flats, beaches, sand dunes, exposed rock, strip mines, quarries, and gravel pits. The following are categories of barren soils [35].

4.1.1 Bare Exposed Rock Those ecosystems characterized by areas of bedrock exposure, desert pavement, scarps, talus, slides, volcanic material, rock glaciers, and other accumulations of rock without vegetative cover. This does not include rock exposures in tundra regions.

4.1.2 Beaches Those ecosystems along shorelines characterized by smooth sloping accumulations of sand and gravel. The surface is stable inland, but the shoreward part is subject to erosion by wind and water and to deposition in protected areas.

4.1.3 Dry Salt Flats Those ecosystems occurring on the flat-floored bottoms of interior desert basins that do not qualify as wetlands.

4.1.4 Mixed Barren Land Those regions in which a mixture of barren land features occurs and the dominant land use occupies less than two-thirds of the area. This includes, for example, a desert region where combinations of salt flats, sandy areas, bare rock, surface extraction, and transitional activities could occur in close proximity.

4.1.5 Sandy Areas Other Than Beaches Those ecosystems composed primarily of dunes -- accumulations of sand transported by wind. These accumulations most commonly are found in deserts although they also occur on coastal plains, river flood plains, and deltas and in periglacial environments. This does not include sand accumulations in tundra areas.

4.1.6 Strip Mines, Quarries, and Gravel Pits Those regions where vegetative cover and overburden are removed to expose such deposits as coal, iron ore, limestone, and copper. This includes inactive, unreclaimed, and active strip mines, quarries, borrow pits, and gravel pits until other cover or use has been established. This does not include unused pits or quarries that have been flooded.

4.1.7 Transitional Areas Those regions that are in transition from one land use activity to another. This transitional phase occurs when, for example, forest lands are cleared for agriculture, wetlands are drained for development, or when any type of land use ceases as areas become temporarily bare as con- struction is planned for such future uses as residences, shopping centers, industrial sites, or suburban and rural residential subdivisions. This also includes land being altered by filling, such as occurs in spoil dumps or sanitary landfills.

4.2 Reasons for soils Barren

4.2.1 Nature, location and magnitude of soil degradation related to agriculture Soil is defined as the top layer of the earth's crust and is composed of mineral particles, water, air and organic matter, including living organisms. It is a complex, mutable, living resource which performs many vital functions: food and other biomass production, storage, filtration and transformation of substances including water, carbon and nitrogen. Soil further serves as a habitat and a gene pool, and provides a basis for human activities, landscape and heritage, and the supply of raw materials. Six of the soil degradation processes recognized by the Commission (water, wind and tillage erosion; decline of soil organic carbon; compaction; salinisation and sodification; contamination; and declining soil biodiversity) are closely linked to agriculture. The degree of

risk of soil degradation is established as a function of the underlying predisposing factors, and does not indicate the actual occurrence of degradation processes in particular areas. The major drivers for water erosion are intense rainfall, topography, low soil organic matter content, percentage and type of vegetation cover, inappropriate farming practices and land marginalization or abandonment. Apart from soil characteristics (such as soil texture), and soil type, the soil organic carbon content is determined by land use, climate (mainly temperature and precipitation) and soil hydrology. Risk related to soil organic carbon decline is defined in terms of the potential of soils to lose organic carbon (removal of carbon from the soil) compared to rates of accumulation of soil organic carbon. Maintaining and optimizing organic carbon levels (as a specific objective of land management) is important in contributing to climate change mitigation. The natural susceptibility of soils to compaction mainly depends on soil texture, with sandy soils being least and clayey soils most susceptible. Human-induced compaction is caused by soil use and land management.

4.2.2 The main natural factors which are influencing soil Stalinization and sodification Climate, the salt contents of the parent material and groundwater, land cover and topography. The most influential human-induced factors are land use, farming systems, and land management, such as the use of salt-rich irrigation water and/or insufficient drainage.

4.2.3 Deforestation The lost of forest in India is 1.3 million hectares per year. One of the major causes of desertification is the cutting down of trees. According to the National Remote Sensing Agency (NRSA), India had less than 11.4% of area under forest as per the 1992 observation. But the more recent satellite pictures show that the forest cover is now less than 10%.

4.2.4 Erosion Loss of vegetative cover has made land more susceptible to erosion. Agents of erosion like wind and water have left vast tracts of land barren. Water erodes top soil to an extent of around 12,000 million tons (mt) per annum. The loss of top soil represents a permanent depletion of the resource base. The annual loss caused by the erosion of top soil through water comes to Rs.12, 000 Cores.

4.2.5 Over-Irrigation Big irrigation projects no doubt have brought prosperity to millions of farmers. But, due to over-enthusiasm, many farmers have resorted to successive cropping and over-irrigation, thereby leading to water-logging and consequent Stalinization and alkalinisation. This situation mainly arises due to poor drainage.

4.2.6 Floods and Droughts It is ironical that in India both floods and droughts occur regularly and alternately. According to the National Commission on Agriculture (1976), there are three types of drought:

(i) Meteorological drought caused by a marked decrease in rainfall.

(ii) Hydrological drought caused by prolonged meteorological drought and its consequent effects on water sources.

(iii) Agricultural drought caused by insufficient rainfall to support crops. 35% of the land is drought-prone and receives rainfall of less than 750 mm. Another 18.5% of the land receiving 750-1000 mm. falls in the transitional zone. The remaining 46.5% receiving rainfall of over 1000 mm. falls under the humid zone. The impact of drought leads to shortage of fodder, shortage of drinking water, loss in agricultural production, and a general decline in living standards. Drought is both man-made and environment-induced. Man has major role in the creation of drought-prone areas due to his over-exploitation of natural resources like forests, degradation for grazing lands, excessive withdrawal of ground water, silting of tanks, rivers, etc. Floods, on the other hand, are caused by heavy rains in a very short period. Each situation could have been altered had there been good vegetal cover. Vegetation helps in reducing run-off, increasing infiltration and reducing soil erosion. The land area prone to floods has doubled from 20 million hectares to above 40 million hectares in the last ten years.

4.2.7 Grazing India possesses an area which is just a fortieth of the total land area of the world supporting 197 million cattle, and ranking first in the world for cattle population. To support such an immense cattle population we have only 13 mha as pasture land. This has led to serious problems as animals have encroached into forest lands and even agricultural lands. Due to lack of green fodder, animals are pushed to the fringes of reserve forests and are thus destabilizing the forest vegetation. Soil dreadful circumstances are occurred due to overgrazing leads to desert like situations which, may also cause, reduction in animal productivity and enlarge the economic stress on human beings and also effect the livelihood of the people who depend on animals. Grazing would not be a problem if the dung of the animals is left as fertilizer. Unfortunately, it is removed to be used as fuel, to be sold to intensively farmed areas, etc.

5. THE SCOPE OF DATA MINING TOOLS AND TECHNIQUES

Data mining has wide scope in the field of agriculture, especially on soils to derive knowledge. Various techniques of mining and machine learning are potentially used in agriculture for estimation and prediction of forming issues. The fuzzy algorithms are applied to managing crops, K-means algorithm used to classify soils, SVM technique applied to prediction of yields, ANN techniques are recommended to find crop diseases, ICA-Integrated Component Analysis plays vital role in weather forecasting, interpolation techniques explores soil suitability for maze. The natural classification of the soils is proposed by Unified Soil Classification System (USCS) [38]. According to USCS soils are divided into three categories; soils contain sand and gravels, soils have high organic material and those having silts and clays. Data mining regression models are used to estimate soil-water-retention, and percentage of organic matter. The scope of the mining and machine learning is huge; the following areas have certain scope on soil analysis.

a) Predicting trends and behavior of soils depending on ingredients and climate conditions.
b) Discovery of soil patterns which are previously unknown.
c) Decision trees for crop and soil management.
d) Artificial Neural Networks to sensing the soils for adoption of crops.
e) Genetic algorithms used for soil allocation strategies.
f) Nearest neighbor methods for soils classification.
g) Rule based induction for yield prediction.

The following are applications of Data mining techniques on soils used for agriculture;

(i) Testing of soil fertility - Cluster analysis and statistical methods are most effective techniques of data mining was used to assess the soil fertility.

(ii) Soils Classification- A variety of mining and machine learning techniques such as SVM, C4.5, ANN etc., are used for

soil classification depending on various parameters of soils.
(iii) Yield prediction – Multilayer perception, radial basis function, and RBF techniques of data mining are used for yield prediction.
(iv) Association rule Apriori is used to Chemical Speciation and Fractionation in Soil and Sediment Heavy Metals, growth of agriculture, Rainfall and weather effects on soils, Pesticides and their effects.
(v) Prediction techniques of mining are exercise on real-time determination of various nutrients of soils such as nitrogen, organic carbon etc.
(vi) Sequence patterns are used to study of crop sequences.
(vii) Decision tree analysis is used to check the soil salinity sodification, soil erosion, soil degradation etc.

6. CONCLUSION

Agriculture is the most significant application area particularly in the developing countries like India. Use of information technology in agriculture can change the situation of decision making and farmers can yield in better way. Data mining plays a crucial role for decision making on several issues related to agriculture field. It discusses about the role of data mining in the agriculture field and their related work by several authors in context to agriculture domain. It also discusses on different data mining applications in solving the different agricultural problems. In this survey it came to know that the k-means algorithm is used for soil classifications using GPS-based technologies, Classification of plant, soil, and residue regions of interest by color images, Detecting weeds in precision agriculture.

7. REFERENCES

[1] Hetal Patel Research Scholar Charusat, Dharmendra Patel Assistant Professor Charusat, Changa in (2014) A Brief survey of Data Mining Techniques Applied to Agricultural Data.

[2]. M.C.S.Geetha, "A Survey on Data Mining Techniques in Agriculture" Vol. 3, Issue 2

[3]. N.Neelaveni, Ms. S. Rajeswari "DATA MINING IN AGRICULTURE- a Survey" Volume 4, Issue 4 [4] https:// docs.oracle.com/cd/B28359_01/datamine.111/b28129/market_ basket.htm#BABFDDCG

[5] Mabroukeh, N. R.; Ezeife, C. I. (2010). Taxonomy of sequential pattern mining algorithms". ACM Computing Surveys.

[6]. www.iasri.res.in/sscnars/data_mining/4-Artificial %20 Neural % 20Networks_Amrender.pdf

[7] Genetic algorithms and its applications in data mining – winter school on "Data mining techniques & tools for knowledge discovery in agriculture data sets".

[8]. https://en.wikipedia.org/wiki/Decision_tree_learning

[9]. https://saravananthirumuruganathan. wordpress.com/ 2010 /05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm

[10]. Rule induction jerzy w. Grzymala-busse university of Kansas.

[11]. Komatineni. Divya, Dr. Y.Sangeetha "Paddy Seeds Categorizing Based on Morphological Feature Using Data Mining Algorithms" Vol 5, Issue- 8.

[12]. Kulwant Kaur, Maninderpal Singh "Knowledge Discovery and Data Mining to Identify Agricultural Patterns"

[13]. Knowledge Discovery and Data Mining to Identify Agricultural Patterns, Kulwant Kaur, Maninderpal Singh, IJESRT [1337-1345], March, 2014

[14]. https://msdn.microsoft.com/en-us/library/ms175595.aspx

[15]. Abdullah, A. and A. Hussain, 2006. Data Mining a New Pilot Agriculture, Extension Data Warehouse. Journal of Research and Practice in Information Tehnology, 3 (3): 229-240

[16]. Mucherino, A., Papajorgji, P., & Pardalos, P. (2009), "Data mining in agriculture" (Vol. 34), Springer

[17]. Beniwal, S., & Arora, J. (2012). Classification and feature selection techniques in data mining. International Journal of Engineering Research & Technology (IJERT), 1(6).

[18]. Lior Rokach, Oded Maimon. Clustering Methods. Chap-15

[19]. Han, J, Kamber, M., & Pei, J. (2006). Data mining: concepts and techniques. Morgan kaufmann.

[20]. Xu, R & Wunsch, D (2005). Survey of clustering algorithms. Neural Networks, IEEE Transactions on, 16(3), 645-678

[21]. Periklis Andritsos Data Clustering Techniques. University of Toronto, Department of Computer Science. ftp://ftp.cs.toronto.edu/csrg-technicalreports/443/depth.pdf

[22]. Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In ACM SIGMOD Record (Vol. 22, No. 2, pp. 207-216). ACM.

[23]. Srikant, R V Q & Agrawal, R (1997, August). Mining Association Rules with Item Constraints. In KDD (Vol. 97, pp. 67-73).

[24]. Zaki, M J (1999). Parallel and distributed association mining: A survey. IEEE concurrency, 7(4), 14-25.

[25]. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

[26]. V. Ramesh and K. Ramr Classification of agricultural land soils: A data mining approach International Journal on Computer Science and Engineering (IJCSE) ISSN : 0975-3397 Vol. 3 No. 1 Jan 2011379

[27]. Rainfall variability analysis and its impact on crop productivity Indian agriculture research journal 2002 29,33.,8) SPRS Archives XXXVI-8/W48 Workshop proceedings: Remote sensing support to crop yield forecast and area

estimates GENERALIZED SOFTWARE TOOLS FOR CROP AREA ESTIMATES AND YIELD FORECAST by Roberto Benedetti A, Remo Catenaro A, Federica Piersimoni B

[28]. S.Veenadhari, Dr. Bharat Misra, Dr. CD Singh, "Data mining Techniques for Predicting Crop Productivity – A review article", International Journal of Computer Science and technology, march 2011.

[29].D.Rajesh, "Application of Spatial Data mining in Agriculture" International Journal of Computer Applications, Volume 15, February 2011.

[30]. International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 2, February 2012) 275 "Survey on Data Mining "VibhaMaduskar and Prof. yashovardhankelkar

[31]. Amrender Kumar, "WEATHER BASED CROP FORECASTING TECHNIQUES", Agricultural Statistics Research Institute.

[32]. Anwiti Jain, Anad Rajavat, Rupali Bhartiya "Design, Analysis and Implementation of Modified K-Mean Algorithm for Large Data-Set to Increase Scalability and Efficiency", IEEE Explore December 2012

[33]. Leisa J. Armstrong, Dean Diepeveen, Rowan Maddern "The application of data mining techniques to characterize agricultural soil profiles"

[34]. http://www.yourarticlelibrary.com/soil/soils-of-india-six-different-types-of-soils-found-in-india/12779/

[35]. https://www.hq.nasa.gov/iwgsdi/Barren_Land.html

[36]. Jay Gholap, Anurag Ingole, Jayesh Gohil, Shailesh Gargade, Vahida Attar "Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction" Dept. of Computer Engineering and IT, College of Engineering, Pune, Maharashtra-411005, India

[37]. A. Kumar & N. Kannathasan, (2011), "A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining ", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3

[38]. Isbell, R. F. (1996). The Australian Soil Classification. Australian soil and land survey handbook. (Vol. 4). Collingwood, Victoria, Australia: CSIRO Publishing.