A Decision Support System for Increasing Agricultural Production

S. Revathi, M. P. Paulraj, P. Parameswari

***Abstract** : Agriculture acts an important and primary role in all countries. Especially in Indian economy, agriculture acts an important factor. So, farmers are always in need to increase their crop productivity, which depends on variety of factors present in soil. If a crop is planted on unsuitable soil, it leads to poor yield. So much more attention is needed while selecting the crop for planting. The huge amount of agricultural data that is available in many resources and data mining performs major role in agriculture. By using this data mining techniques, the hidden required pattern from the huge data can be identified and make them useful to the farmers and decision makers to obtain better yield performance. In our work, data mining technique is applied in the dataset of soil and crops, which belongs to Tamilnadu region. We have analyzed various physical properties and also chemical properties of soil like Soil type, Soil Texture, Color, Structure, WHC (Water Holding Capacity), Soil Moisture, pH, Electrical Conductivity (EC) and Temperature. We have considered Soil Type, pH, EC and Temperature in this work to find the suitable crop for best productivity. This research mainly concentrates on finding the correlation between soil properties and crops by applying clustering algorithms such as Simple K Means (SKM), Filtered Clusterer (FC) and Hierarchical Clusterer (HC) for finding the suitable crop according to the soil properties.*

***Key words**: Data mining, SimpleKMeans, Filteredclusterer and HierarchicalClusterer, Agriculture, Cluster*

I. INTRODUCTION

Data mining plays a major role in agricultural research. The various techniques such as clustering, classification, prediction and evaluation are used to identify and extract the useful patterns from vast amount of crop and soil database. Data mining which is the decision support process which mainly helps in agricultural management system for improving the crop yield. Apart from the agricultural field this decision support system plays an important role in various fields such as commerce, education and scientific applications. The field data mining contains lot of algorithms to extract the useful pattern from the large amount of data in many field but the following algorithms such as the K-Means algorithm, K-Nearest Neighbor algorithm, ID3 algorithms, ANN (Artificial Neural Network) and SVM (Support Vector Machines) algorithm are frequently used in agricultural researches. As an example K-means is helps to classify the soil data based on some constraints, which are collected by the GPS technology. Similarly the above mentioned algorithms are helps for classifying the soil and crop properties in agricultural field which is used for effective crop management system. Data warehousing, iunconvetional component analysis techniques are applied in Spatio-temporal data, weather data and other agricultural related data for identify the useful pattern. Statistical based techniques such as principle component analysis (PCA), regression models and biclustering techniques are also used in agriculture field for knowledge discovery. Agriculture mainly depends on soil. Soil types and their properties vary from region to region. In this work, soil attributes like soil type, soil pH, soil EC and soil temperature are taken into account because the soil properties are create a great impact on crop productivity so the requirement here is to identify the suitable soil while planting a crop. This approach helps to getting good quality and quantity in crop production. WEKA is a data mining tool, which contains number of machine learning algorithms. These algorithms are used for predicting unknown patterns from large database that are used for decision making. Following process are involved on soil database before applying the data mining algorithms. They are, Data collection, Data preprocessing, File conversion, Opening file from local file system, Loading data, Setting filters, Applying data mining algorithms and Identifying the useful pattern.

The proposed work investigates the performance and accuracy of Simple K Means, Filtered Clusterer and Hierarchical Clusterer based on clustering the soil and crop data. Through this process, the optimum suitability of soil for planting crops can be identified.

II. LITERATURE SURVEY

In this paper we made an attempt to bring out the work done by various researchers for reference, which helps to know the various data mining algorithms such as statistical data mining, MLR (Multi Linear Regression) used in agriculture for increasing crop productivity.

The data mining algorithms were compared in this work. They are: (Multiple Linear Regression), PLSR (Partial Least Square Regression), Multivariate Adaptive Regression Spliness (MARS), ANN, RF (Random Forest), BT (Boosted Trees) and SVM (Support Vector Machine) to determine the organic content in soil, Clay Content (CC) and pH measured in water. The distinct wavelet transform is used to perform the comparison which was achieved by the nominated list of wavelet co-efficient.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author

S. Revathi*, Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, India. revathi.cse@srit.org

M.P.Paulraj, Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, India. revathi.cse@srit.org

P.Parameswari, Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, India. revathi.cse@srit.org

PCA (Principle Component Analysis) was used for compressing the spectra then this result is processed through K-Means clustering algorithm to classify the different types of spectra in the spectral library. The comparison of various data mining algorithms suggests the best predictions for three soil chemical properties (SOC, CC, and pH in water) [1]. The salinity and moisture in the soil are important factors which are also a part of crop yield rate. [2] Evaluated the cause of moisture and salinity in soil on crop yield by developing ANN and MLR Models which is important for the irrigation in saline soil. This model contains the input variables (Soil properties) as bulk density, sand (%), silt (%), clay (%) including soil moisture and salinity at different growth stages of crop. The sunflower crop was taken into account for the research. The soil salinity varied with different ranges such as low, average and high saline soils. Correlations between crop yield, moisture and Salinity in soil are understood by the ANN models at different stages of crop growth. Finally the author concludes the ANN model produces the most accurate result than the MLR. ANN methodology is an alternative modeling and simulation tool which is especially designed for dynamic non linear system. The effective irrigation system provides the well crop growth. [3] Collected the samples from each plot such as pre irrigation, Well Water irrigation (WW), Site and Sugar Beet Rinse irrigation (SBRW) Water for testing the cause of land with pond treatment on chemical attributes of soil and growth of the plant and yield. The following chemical properties of were taken into account they are: pH, EC (Electrical conductivity), total phosphorus and total nitrogen. Simple linear regression analysis was used to evaluate the relationship between the response variables. Pearson's correlation coefficients were helps to test the relationships among chemical properties of soil which is appeared in the soil profile and also evaluated by using (PCA) Principle Component Analysis. This research is not only suitable for test the growth of the crop and its yield. It also applicable for monitoring the changes in soil properties and other crop's quality. The tillage helps to make the smoothening soil qualities. [4] Mainly focused on to determining the long term effect in the tillage systems (t1, t2, t3, t4). The properties of soil and yield in a green pea to find the cause of tillage methods in soil. The following soil properties were taken into accounts for this research they are, soil's bulk density, pH, organic carbon, penetration resistance and hydraulic conductivity. The linear regression algorithm was used to identify relations among cone index (CI), Bulk density (BD) and organic content (OC). Finally the result figure out, there were no consistent difference in four tillage system and yield performance was increased using conservation tillage because there is no consistent different among tillage treatments. Classifying the soil based on their quality will improve the higher crop productivity. [5] Focused on characterizing the quality of soils in semiarid Mediterranean eco system. So they developed a suitable index for characterizing soil quality. Preliminary soil attributes are considered. These attributes were based on different physical properties such as granularity, aggregate stability, WHC (Water Holding Capacity), available water, bulk density, particle density, plant cover and also bio chemical properties such as organic carbon, nitrogen, pH, kcl and water, Electrical Conductivity (EC), CEC (Cation Exchange Capacity), available ions such as K, Mg, Fe, copper and zinc. Discrimination between the initial variables was considered by the principle component analysis (PCA). In order to select these variables that may be included in an index. The index reflects the quality of soils in Mediterranean agricultural areas. The variation that occurs in the agricultural land potentially reflects by the index which may be considered a suitable tool for the early detection of changes in the soil. This research concluded with result as, limiting indicators for each soil type can be related to the different soil function.

III. PROPOSED METHODOLOGY

This Part is organized as Block diagram, data construction and Clustering concepts (Simple K-means, Filtered Clusterer and Hierarchical Clusterer)

A. Block Diagram

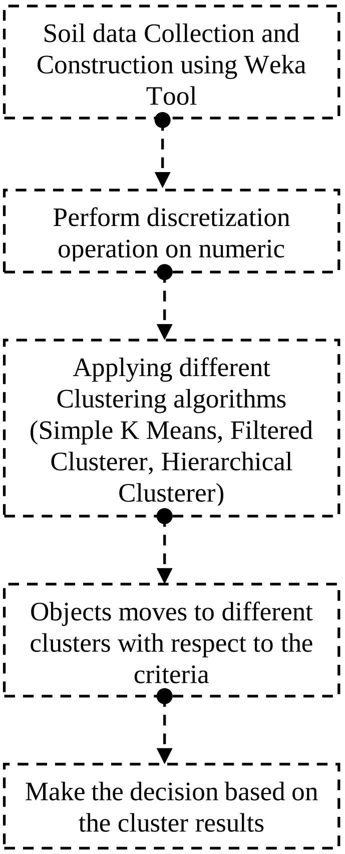

Soil data Collection and Construction using Weka Tool
Perform discretization operation on numeric
Applying different Clustering algorithms (Simple K Means, Filtered Clusterer, Hierarchical Clusterer)
Objects moves to different clusters with respect to the criteria
Make the decision based on the cluster results

Fig.1. Block Diagram

B. Data Collection

The major and the minor crops are considered in this work that is cultivated especially in Tamilnadu. Crops and Soil properties such as soil type, soil pH, soil electrical conductivity and soil temperature values are collected from http://www.tn.gov.in/crop/AreaProduction.htm which is exclusively providing information about soil properties and suitable crops for their healthy growth and obtain high yield. The dataset is being collected and it needs to be converted or downloaded as Microsoft Excel file (.xls). Now we have to bring it to the Comma Separated Values (CSV).

C. Dataset Construction

Weka has the ability to read '.csv' extension files. The many applications from different databases or spread sheets save as .csv format and export into the software (Weka) as .csv or .arff file. The first row contains the attribute name and other row contains the attribute value. In this work the soil dataset contains crop name and their appropriate soil properties such as soil type, temperature, pH and EC (Electrical Conductivity) as attributes. These data belongs to agricultural lands of the Tamilnadu. The details are collected in Microsoft Excel format and are saved as .csv or .arff format. Once the file is exported, WEKA will identify the attributes and performs some primary statistics on the attributes. By using Weka tool these data can be preprocessed. Some algorithms, such as ARM (Association Rule Mining) and clustering are only executed on categorical data so needs to perform discretization operation on numeric attributes. Three such attributes are there in our dataset: 'temperature', 'pH', 'EC'. In this case we performs discretization process by removing the keyword 'numeric' for the attributes 'temperature', 'pH', 'EC' in the .csv file, and change it as the discrete values.

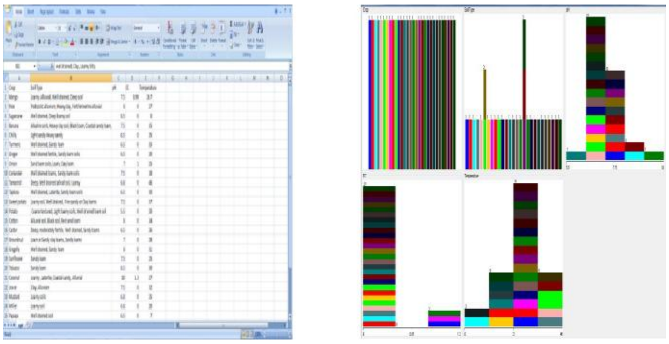


Fig. 2. Sample dataset in .csv format and Data Preprocessing

Figure2. Data preprocessing

D. Clustering Concepts

Clustering is a portion of data or group of data with similar characteristics. Clustering plays an important part in various applications of data mining such as Medical, wireless sensor network, search engine algorithms etc. According to the weka, there is a more number of clustering algorithms that can be enforced to a data set in order to segregating the information. In this research we have applied three important clustering algorithms namely Simple K Means, Filtered Clusterer and Hierarchical Clusterer.

i. Simple K Means

In WEKA SimpleKMeans algorithm automatically manages a combination of categorical and numerical attributes. Furthermore, this algorithms performs distance calculation when normalizes numerical attributes. In this algorithm number of clusters (Seeds) has to be specified at initial stage. Distance between each instance has been calculated by using Euclidean distance (by default) for clustering similar instances by using following formula:

$$\text{Dist}\ ((x, y), (a, b)) = \sqrt{}(x\text{-}a)\ ^{2}\text{+}(y\text{-}b)^{2}\ldots\ldots\ldots\ (1)$$

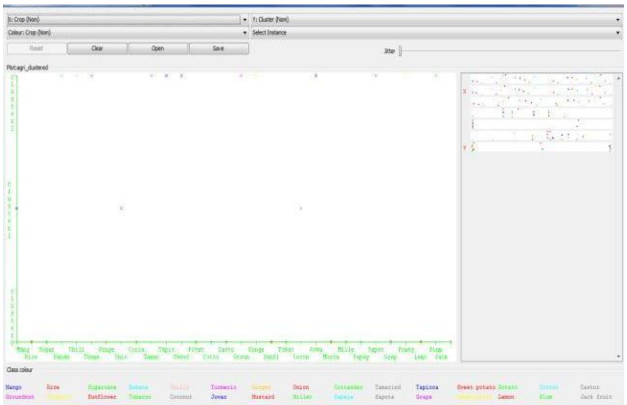


Fig. 3. SKM Output

In Figure 2, cluster number as the x-axis, the soil properties (assigned by WEKA) as the y-axis and 'crop' attribute as the color dimension. For an instance the Simple K Means forms three clusters. Most suitable crop in each cluster is Rice, Mango Chilly. This crop suitability can be determined by soil properties.

The table I shows, crop and soil type are the nominal attributes remaining are numeric attributes. For each crop, the numeric attributes may contain certain range of numeric values. In table 1, the pH value from 7.3-7.75 is most suitable for the growth of Mango.

Table - I: Output clusters in SimpleKMeans algorithm

Attributes	**Cluster#**	**Cluster#**	**Cluster#**
	0	**1**	**2**
Crop	Rice	Mango	Chilly
Soil type	Sandy loam	Loamy, Alluvial, Well drained, Deep soil	Well Drained, Sandy loam
pH	7.3-7.75	6.85-7.3	6.4-6.85
EC	-inf-0.13	0.91-1.0 4	-inf-0.13
Temperature	24	25-85	28

ii. Filtered Clusterer

The training and test data can be processed by filter without modifying their structure. Here the number of clusters is two, which has been assigned default

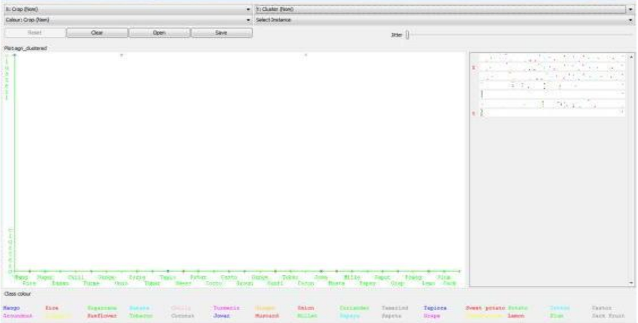


Fig. 4. FC Output

The Filtered Clusterer forms two numbers of clusters (by default). Here also cluster 0 and cluster 1 represent that most suitable crops are rice and mango respectively.

Table - II: Output Clusters in FilteredClusterer

Attributes	Cluster #	Cluster #
	0	1
Crop	Rice	Mango
Soil type	Sandy loam	Loamy, Alluvial, Well drained, Deep soil
pH	6.4-6.85	7.3-7.75
EC	-inf-0.13	-inf-0.13
Temperature	24.574	30.633

iii. Hierarchical Clusterer

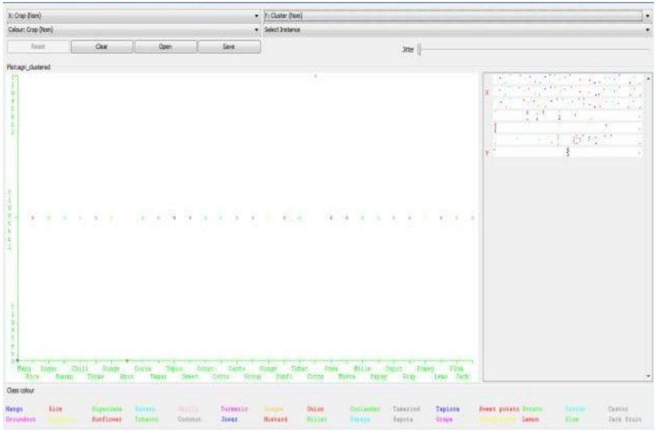
Generally clustering techniques are classified as Hierarchical clustering and partitioning clustering. This algorithm works based on either agglomerative or divisive clustering methods. In this algorithm number of clusters has to be specified.

Steps in agglomerative method:

a. Process starts with single instance (from leaf)
b. Merges the similar instance
c. Finally form a cluster

Steps in divisive method:

a. Process starts with single cluster (From root)
b. Then recursively split each instances from the cluster



X: Crop (Nom)
Y: Cluster (Nom)
Colour: Crop (Nom)
Select Instance

Fig. 5. HC Output

The Hierarchical Clusterer forms three numbers of clusters. Most suitable crop in each cluster is mango, rice and coconut. For these crops, suitable soil type can be sandy loam, pH value can be 6.5-10.0, EC value can be 1.3-1.5 and value of temperature can be 23.0-30.0.

IV. RESULT ANALYSIS

The experiment helps us to predicting preferable crop to the soil based on soil properties. The clustering analysis forms a group of crops which are provide higher yield depends on soil properties such as soil type, EC, pH and temperature. The following tables are demonstrating the SKM, HC and FL results with respect to soil properties.

Table – III: Correlation between Soil Types and Suitable Crops

Clustering Algorithms	Soil types	Suitable Crops
SKM	Well drained, Sandy loam, Alluvaial, Deep soil	Rice, Mango, Banana, Cotton, Tapioca, Jowar, Onion, Mustard, Sweet, Potato, Lemon, Sunflower
FC	Well drained, Sandy loam, Deep soil	Rice, Mango, Jowar, Tapioca, Sweet, Potato, Lemon, Sunflower
HC	Loamy, Sandy loam	Rice, Mango, Jowar, Tapioca, Sweet potato, Lemon, Coconut Sunflower, Castor, Jack fruit, Sapota Tamarind

Table – IV: Correlation between Soil pH and Suitable Crops

Clustering Algorithm	pH Range	Suitable Crops
SKM	6.5 to 7.5	Rice, Mango, Banana, Cotton, Tapioca, Jowar, Onion, Mustard, Sweet potato, Lemon, Sunflower
FC	6.5 to 8.0	Rice, Mango, Jowar, Mango, Tapioca, Sweet potato, Lemon, Sunflower
HC	6.5 to 10.0	Rice, Mango, Jowar, Mango, Tapioca, Sweet potato, Lemon, Sunflower, Castor, Jack fruit, Tamarind, Sapota, Coconut

Table – V: Correlation between Soil EC and Suitable Crops

Clustering Algorithms	Soil EC	Suitable Crop
SKM	0 to 1.096	Rice, Mango, Banana, Cotton, Tapioca, Jowar, Onion, Mustard, Sweet potato, Lemon, Sunflower
FC	0 to 1.096	Rice, Mango, Jowar, Mango, Tapioca, Sweet potato, Lemon, Sunflower
HC	1.3 to1.5	Jowar, Mango, Tapioca

Table – VI: Correlation between Soil Temperature and Suitable Crops

Clustering Algorithms	Soil Temperature Range	Suitable Crops
SKM	26.23 to 22.52	Rice, Mango, Banana, Cotton, Tapioca, Jowar, Onion, Mustard, Sweet potato, Lemon, Sunflower
FC	56.4076 to 26.23	Rice, Mango, Jowar, Mango, Tapioca, Sweet potato, Lemon, Sunflower
HC	23.0 to 30	Mango, Jowar, Tapioca, Tamarind, Sapota, coconut, Jack fruit, Castor

The SimpleKMeans and FilteredClusterer produce moreover similar and accurate result with sum of squared errors 65.1055 and 70.0469 respectively but the computation time of each algorithm is same (0.02 seconds). The following table shows different clustering algorithms with optimal crop which is based on the different soil properties.

In SimpleKMeans (SKM) Clustering, the cluster 0, cluster 1and cluster 2 represents the crop as Rice, Mango and Banana based on their similar pH values. In FilteredClusterer (FC), the cluster 0 and cluster 1 represents the crop as Rice and Mango based on the pH values. In HierarchicalClusterer (HC), the cluster 0, cluster 1and cluster 2 represents the crop as Mango, Rice and coconut. According to the above result the SimpleKMeans (SKM) Clustering and FilteredCluster (FC), provides the similar result.

Table - VII: Optimal Crop Suitability

Clustering Algorithms	Optimal Crop Suitability
SKM	Rice, Mango, Banana (3 Clusters)
FC	Rice, Mango (2 Clusters)
HC	Mango, Rice, Coconut (3 Clusters)

This clustering analysis gives the better predictability from the large database which helps to the researchers to work on soil and agricultural field. These results may also helps to the farmers to do cultivation depends on the properties in soil, which leads to the better yield.

Performance Evaluation of each Clustering Algorithms

The SKM, FC and HC clustering algorithms provide results which are different from one another even though somewhat it provides partially similar results. The performance of these algorithms was different in terms of a. Time taken to build the model and b. Number of incorrectly classified Instances

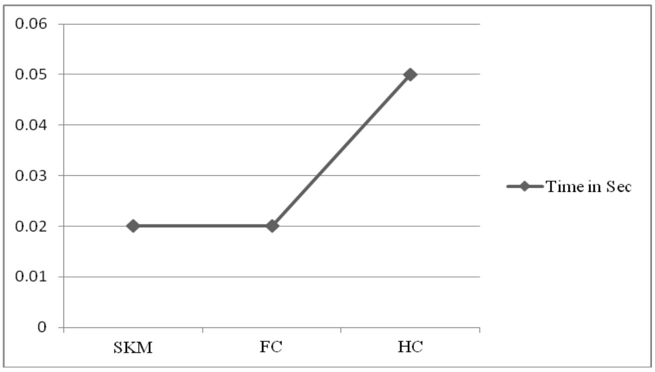


Fig. 6. Time taken to build the model

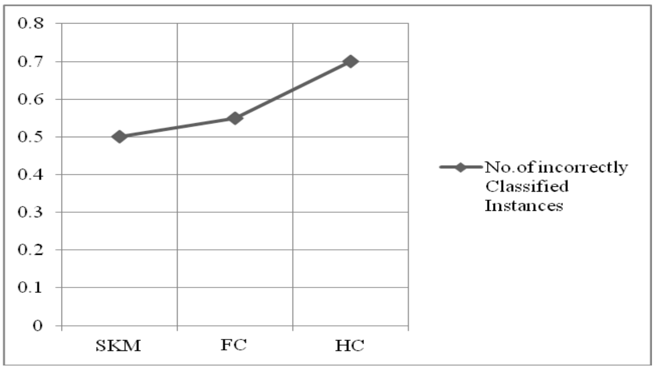


Fig.7. Number of incorrectly classified Instances

V. CONCLUSION

The soil dataset contains the various soil properties and crop details. By means of Weka tool the various clustering algorithms was applied on the soil dataset and each algorithms shows the suitable crop for different soil properties. This clustering analysis gives the better predictability from the large database which helps to the researchers to work on soil and agricultural field. These results may also helps to the farmers to do cultivation depends on the properties in soil, which leads to the better yield.

FUTURE WORK

In the future work climatic data set and improvised soil dataset will be considered. The climatic dataset will be interrelated with the soil dataset. Furthermore the regions can be classified based on the identified result to provide better understanding.

REFERENCES

1. R.A. Viscarra Rossel, T. Behrens. “Using data mining to model and interpret soil diffuse reflectance spectra”, Geoderma, 2010, Vol. 158(1), pp.46-54 .
2. Xiaoqin Dai, Zailin Huo, Huimin Wang, “Simulation for response of crop yield to soil moisture and salinity with artificial neural network”, Field crop research, 2011, Vol. 121(3), pp.441-449.
3. Ying Zhagang, Xin Li, Zhigang Wang, Haijing Liang, Maio Hu, Qingjuan Meng, “Study on the response of soil chemical properties and corn (Zea mays L.) to the land application with sugar beet rinse water”, Agricultural water management, 2012, Vol. 115, pp.38-46.
4. J.L. Pikul Jr, R.E. Raming, D.E. Wilkins, “Soil properties and crop yield among four tillage systems in a wheat-pea rotation”, Soil and tillage research, 1993, Vol. 26(2), pp.151-162.

Retrieval Number: D1133029420/2020©BEIESP
DOI: 10.35940/ijitee.D1133.029420

5. A. Sanchez-Navarro, J.M. Gil-Vazquez, M.J. Delgado-Iniesta et al., “Establishing an index and identification of limiting parameters for characterizing soil quality in Mediterranean Ecosystem”, Catena, 2015, Vol. 131, pp.35-45.
6. Oswalt Manoj. S, S. Revathi, Adrijovin. J. J, Ramalakshmi. K, “Survey on Soil Properties and Crop Management based on Data mining Concepts-”, 2015, Seventh International Conference on Advanced Computing 2015(ICoAC2K15).
7. D. Casanova, J. Goudriaan, JC.M. Withagen, “Rice yield prediction from yield components and limiting factors”, European journal of agronomy, 2002, Vol. 17(1), pp.41-61.
8. M.E. Holzman, R. Rivas, M.C. Piccolo, “Estimating soil moisture and the relationship with crop yield using surface temperature and vegetation index” International Journal of Applied Earth Observation and Geoinformation, 2014, Vol.28, pp.181-192.
9. Yi-Ping Wang, Yuvan Shen, “Identifying and characterizing yield limiting soil factors with the aid of remote sensing and data mining techniques”, Precision Agriculture, 2015, Vol. 16(1), pp. 99-118
10. H.Borman, “Assessing the soil texture –specific sensitivity of simulated soil moisture to projected climate change by SVAT modeling”, Geoderma, 2012, Vol. 185-186, pp. 73-83.
11. Cecile Gomez, Phillipe Lagacherce, Guillaume Coulouma, “Regional predictions of eight common soil properties and their spatial structures from hyper spectral Vis-NIR data”, Geoderma, 2012, Vol.189-190, pp.176-185.
12. S.M. Bateni, D.-S. Jeng, S.M. Mortazavi Naeini, “Estimating soil thermal properties from sequences of land surface temperature using hybrid genetic algorithm-finite difference method”, Engineering Applications of Artificial Intelligence, 2012, Vol. 25, pp. 1425-1436.
13. Johan Arvidsson, Aron Westlin, Fredrik Sörensson, “Working depth in non-inversion tillage-effects on soil physical properties and crop yield in Swedish field experiments”, Soil and Tillage Research, 2013, Vol.126, pp. 259-266.
14. N. Nassi o Di Nasso, M.V. Lasorella, N. Roncucci, E. Bonar , “Soil texture and crop management affect switchgrass (Panicum virgatum L.) productivity in the Mediterranean”, Industrial Crops and Products 2015, Vol. 65, pp. 21-26 .
15. Lotfollah Abdollahi, Hansen. E.M., Rickson. R.J, Munkholma. L.J. “Overall assessment of soil quality on humid sandy loams: location, rotation and tillage” Soil & tillage research, 2015, Vol.145, pp.29-36.
16. Sara Marinari, Roberto Mancinelli, Paola Brunetti, Enio Campiglia,“Soil quality, microbial functions and tomato yield under cover crop mulching in the Mediterranean environment”, Soil and Tillage Research, 2015, Vol. 145, pp.20-28

ACKNOWLEDGEMENT

The authors express their heart-felt gratitude to the Management of Sri Ramakrishna Institute of Technology for providing support in carrying out this project.

AUTHORS PROFILE

S.Revathi, is currently working as Assistant Professor at Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu, India. She holds Master degree in Computer Science and presently doing her research in Anna University, Chennai. She has research interest on Data Science and Machine learning.

Dr. Paulraj Murugesa Pandiyan, is currently working as Principal at Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu, India. Previously he has worked as Professor in School of Mechatronics Engineering, Universiti Malaysia Perlis, Malaysia. He holds a Ph.D in Computer Science and carries 32 years of Teaching Experience and more than a decade of Research and Guiding Experience in the field of Neural Networks. He has also worked in Universiti Malaysia Sabah, Kota Kinabalu and Govt College of Technology. He has published more than 400 technical papers in referred journals, national and international conferences in the field of Neural Networks. He has guided and currently guiding many Masters and PhD students.

P.Parameswari, currently working as an Assistant professor at Sri Ramakrishna Institute of Technology from January 2019. She has completed M.Tech-IT (Information and Cyber Warfare) from Kongu Engineering College, Erode and completed BE-Computer Science and Engineering at EBET Group of Institutions, Tirupur. Her area of interest are networks and cyber security.