Agriculture Analysis Using Data Mining And Machine Learning Techniques

Vanitha CN[1], Archana N[2], Sowmiya R[3]
[1]*Professor, Department of Computer Science and Engineering, Kongu Engineering College, Erode, Tamil Nadu, India, rushtovanitha@gmail.com*
[2 & 3] *Student, Department of Computer Technology, Kongu Engineering College, Erode, Tamil Nadu, India, [2]16bsr005archana@gmail.com [3]mathisowmir@gmail.com*

Abstract

Agriculture is an important application in India. The modern technologies can change the situation of farmers and descision making in agricultural field in a better way. Python is used as a front end for analysing the agricultural data set. Jupyter Notebook is the data mining tool used to predict the crop production. The parameter includes in the dataset are precipitation, temperature, reference crop, evapotranspiration, area, production and yield for the season from January to December for the years 2000 to 2018. The data mining techniques like K-Means Clustering, KNN, SVM, and Bayesian network algorithm where high accuracy can be achieved.

Keywords: Bayesian Network, Support Vector Machine, K- Nearest Neighbour, K- Means Clustering.

1. Introduction

Agriculture is the major source of the Indian Economy. Day by day, the population increases. So the demand of food increases. To get rid of these situations farmers, agricultural scientists, and researchers are trying for better crop yield.

The analyzing process of hidden patterns according to various perspectives for classification and converted into relevant information is called as data mining in which data is arranged in particular areas like data repository. The efficient analysis using data mining techniques help farmers to take decisions. These informations help them to reduction of costs and increasing the production rate. The data mining process includes following steps: Extracting, transforming and loading data in a repository and managing the data in multidimensional databases. Data mining provide data access to analysts using application software. The analyzed data is easily represented using graphs.

2. Classification Techniques

2.1 SVM Classification

Support Vector Machine algorithm is prominent data analysis methodology and it is used for classification and regression techniques. Here the data points have been plotted using n-dimensional space with the value of particular characteristics as the value of a specific coordinate. The classification is done by finding the hyper-plane line that differentiate the classes separately.

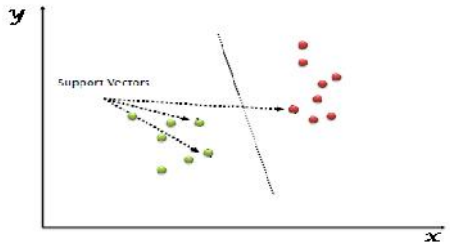

Figure 1 SVM Implementation diagram

In figure 1, the SVM Implementation diagram contains Support Vectors and hyper-plane. Support Vectors are The co-ordinates of a partcular class is known as suppot vectors. Hyperplane line is used to separate the two classes.

In figure 2, the suggestions of crop with respect to climate, rainfall, decisions, etc., are considered to alert the farmers to predict the crop yield.

978-1-5386-9533-3/19/$31.00 ©2019 IEEE

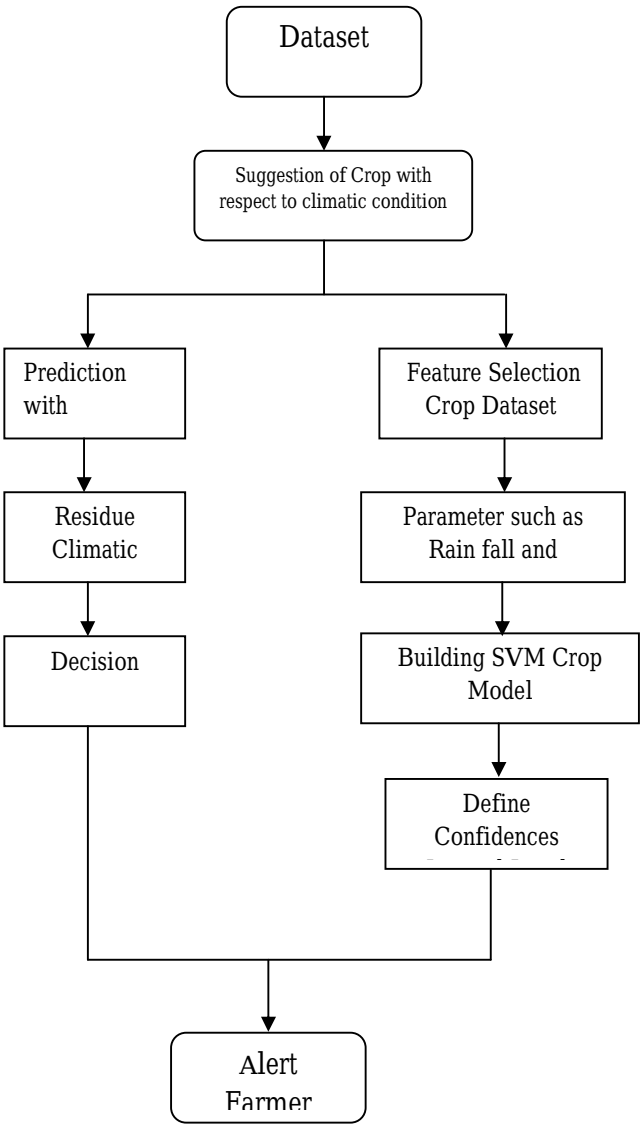

Figure 2 SVM Classification system flow diagram

2.2 SVM Algorithm

- Step1: Read the crop dataset.
- Step2: Create the data frame and extraction feature of Crop production, year, temperature, mean rainfall and mean temperature dataset.
- Step 3: Create the SVM class using e107 package and linear, non-linear and kernel sequences model.
- Step4: Predication crop for temperature: The first phase read data frame and set list of crop per year, area and temperature. The second phase applied support vector matrices to prediction state for crop dataset. The SVM results calculate on regression format dataset.
- Step 5: Predication crop for year of production: The first phase read data frame and set list of crop per season, rainfall and temperature. The second phase applied support vector matrices to prediction state for crop dataset. The SVM results calculate on regression format dataset.
- Step 6: Predication crop for Rainfall mean value: The first phase read data frame and set list of crop per season, rainfall and crop. The second phase applied support vector matrices to prediction state for crop dataset. The SVM results calculated on regression format dataset and linear search.

2.3 Bayesian Network

A Bayesian network is an **acyclic graph consists of edges and nodes** with directions in which each edge represents to a conditional dependency, and each node represents to a unique random variable. The probabilistic graphical model that uses Bayesian inference for computations are called as Bayesian networks.

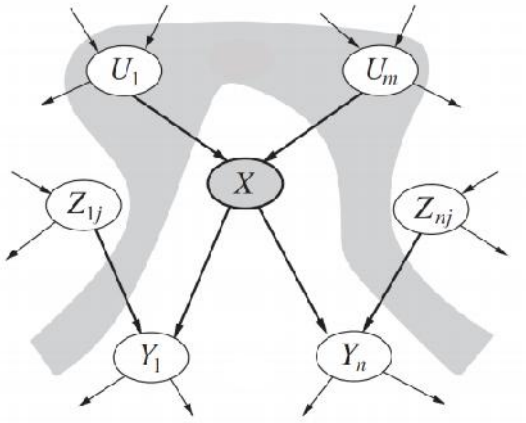

Figure 3 Bayesian Network

In Figure 3, Bayesian Network is represented using an acyclic graph. It consists of nodes and edges.

A Selective Attribute network is a compact, graphical model of a probability distribution which assigns a probability to every event of interest. For example, in the crop detection, a Selective Attribute

network can be used to calculate the probability value of any particular crop prediction given the season displayed by a rice crop dataset.

Selective Attribute Classification is a basic task in data analysis and pattern recognition that requires the construction of a classifier, that is, a function that assigns a class label to instances described by a set of attributes. The induction of classifiers from data sets of pre-classified instances is a central problem in machine learning. Numerous approaches to this problem are based on various functional representations such as decision trees, decision lists, neural networks, decision graphs, and rules. given C whenever $\Pr(A|B, C) = \Pr(A|C)$ for all possible values of A, B and C, whenever $\Pr(C) > 0$.

Bayesian network B, that encodes a distribution PB (Crop_Year, Crop_Production, Crop_Temperature, Crop_Rainfail, Crop_n), from a given training set. We can then use the resulting model so that given a set of attributes a1. . . an, the classifier based on B returns the label c that maximizes the posterior probability PB(c|Crop_Yea_a1, Crop_Production_a2, . . . , Crop_an). Note that, by inducing classifiers in this manner, we are addressing the main concern expressed in the introduction: remove the bias by the independence assumptions embedded in the naïve Bayesian classifier**.**

Network with a relatively good MDL score that performs poorly as a classifier and to understand the possible discrepancy between good predictive accuracy and good MDL score, we must re-examine the MDL score. Recall that the log likelihood term in Equation 2 is the one that measures the quality of the learned model, and that D = {Crop_u1,. . , Crop_uN} denotes the training set. In a classification task, each Crop_ui is a tuple of the form Crop_hai1, . . . ,Crop_ ain,Crop_ cii that assigns values to the attributes Crop_A1, . . . ,Crop_ An and to the class variable Crop_C. We can rewrite the log likelihood function as

LL(Crop_B|Crop_D) = N Xi =1 log PB(Crop_ci|Crop_ai1, . . . ,Crop_ ain) + N Xi =1 log PB(Crop_ai1, . . . , Crop_ain)

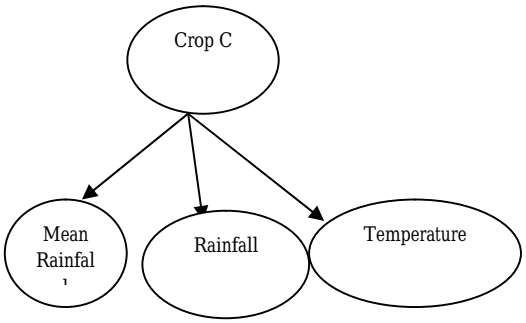

Figure 4 Structure of the Selective Attribute

In Figure 4, The attributes are independent from one another. Ex: Mean rainfall, rainfall, temperature are does not depaend upon each other.These three are depend on the crop C.

2.4 Selective Attribute Networks Algorithm

- Step1: Read the Crop dataset.
- Step2: Create the data list from Crop dataset and feature extraction for prediction crop details.
- Step3: Create the Selective Attribute (Bayesian) net using Neuralnet package.
- Step 4: To create a Rice, Coconunt, Arecanut, Black pepper and Dry ginger crop base net using model2network function in Jupyter Notebook.
- Step5: To read the dataset and assign the data into CA, CS, Ck, Cw, CB, CL and CE object variable. The object contains Crop year, Crop production, Crop Area, Crop mean Temperature, Crop mean Rainfall, mean Crop Temperature and Rainfall values, and districts details is connected the Bayesian network.
- Step 6: The crop classification rule and probability values assign the Bayesian net.
- Step7: To create custom Bayesian net using Bayesian theory in Rice, Coconunt, Arecanut, Black pepper and Dry ginger crop, etc.,

- Step 8: To check the Bayesian Rule for Rice, Coconunt, Arecanut, Black pepper, Dry ginger crop, etc.., Return the accuracy values.
- Step 9: Repeat the process Step 3 to Step 8.
- Step 10: To accuracy calculate the TP, TN, FP and FN values.

2.5 KNN (K Nearest Neighbours) Classifier

K-Nearest Neighbors is one of the classification algorithms in Machine Learning. It is otherwise called as supervised learning model and lazy algorithm because of instance learning. It is used for different applications like pattern recognition, data mining and intrusion detection. Implementation is simple for tiny data sets. The training data does not need any knowlegde about the structure of data before the analysis. Drawback of this classifier is finding out the nearest neighbour for each sample.Lot of space is needed when the training data is large. The distance between test data and the training data should be calculated for every test data. So, the testing needs a lot of time. There are two phases in this classifier: Training phase: Save the examples, Prediction phase: Get the test instance and find the training set.

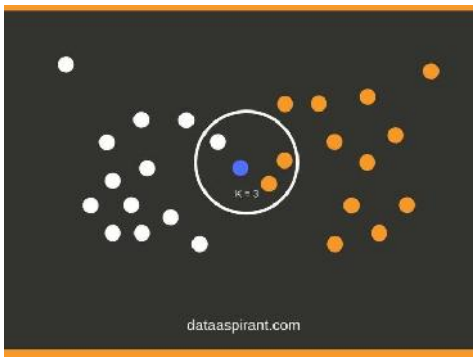

Figure.5 KNN Classifier Diagram

In Figure.5 The **white and orange** circles are two different classes.There are 26 circles. The blue circle is the target circle used for prediction. Here, the value of k is three. The Euclidean distance is used to calculate equal distance. The classes are close when the similarity score is less. In this image, we have calculated the minimum distance.

3. Clustering Technologies

3.1 K-Means Clustering

K-Means Clustering is one of the clustering method that process a group of data points into a small number of clusters. For example, the items in a mall are clustered in different categories (medium, large, X-L are grouped as the size of the dress). This is a qualitative method of splitting a group of data. A quantitative approach is used to measure unique characteristics of the products. In *k* clusters, the number of data points have to be partitioned. The aim of this methodology is to assign a cluster to each data point. K-means algorithm aims to find out the clusters positions that minimize the *distance* from the data points to the cluster.

3.2 ALGORITHM

Step 1 The k cluster centroids are used to label new data

Step 2 Labels are need for the training data .

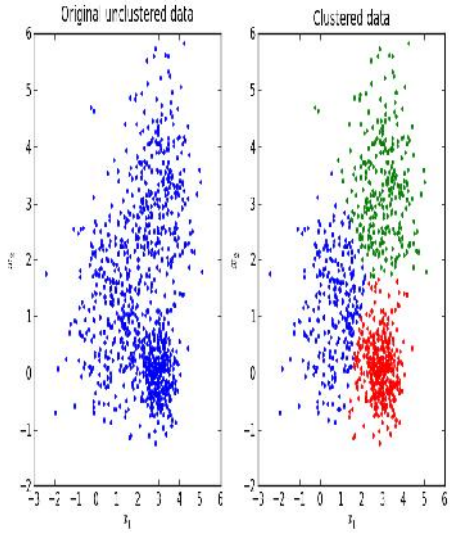

Figure.6 Unclustered data and Clustered data

In Figure.6 The unclustered data consists of overall dataset points together and the clustered data consists of different groupings of overall dataset.

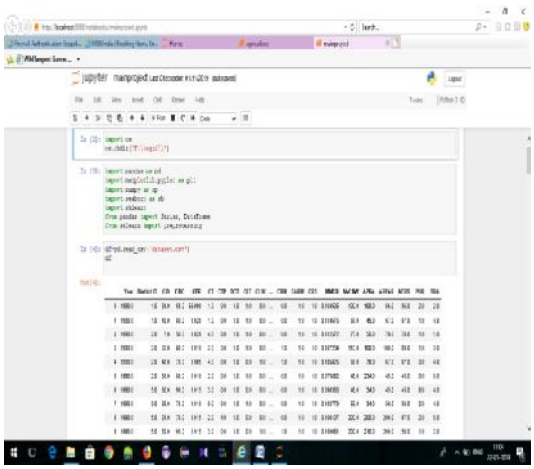

Figure.7 Importing the dataset

In Figure.7, Dataset in excel sheet is imported as csv file using python language in Jupyter Notebook.

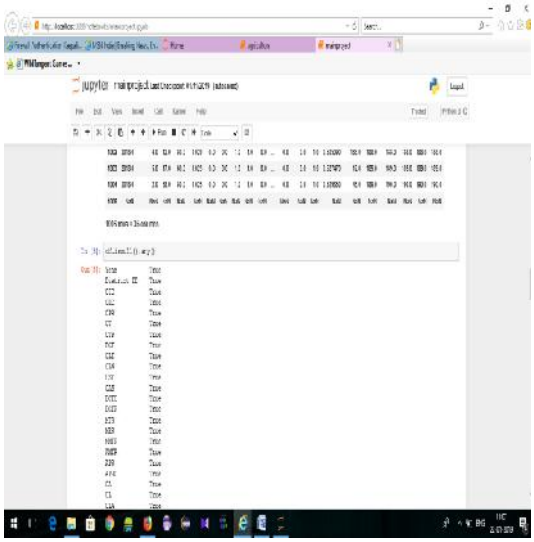

Figure.8 Checking null values

In Figure.8 Null values are checked using boolean values.

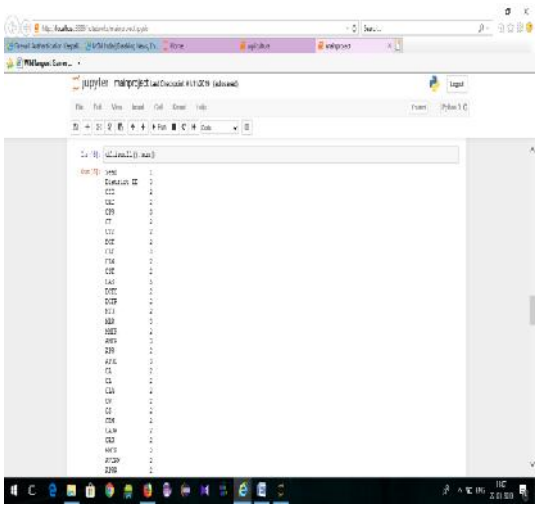

Figure.9 Sum of null values

In Figure.9 Sum of null values in each row is checked.

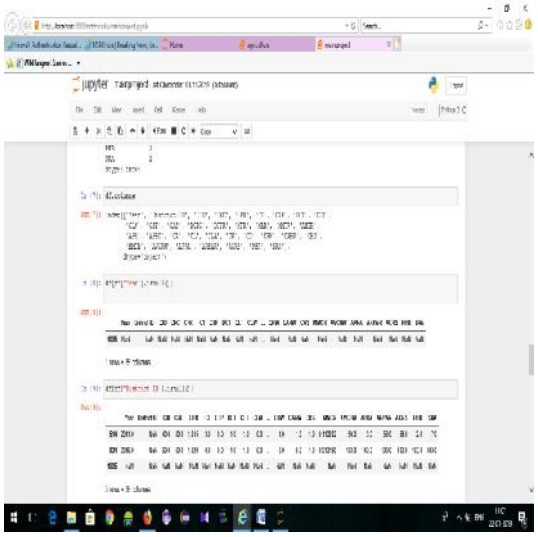

Figure.10 Column names and Place of null values in each row

In Figure.10 Column names are displayed and place of null values in each row is represented as NaN.

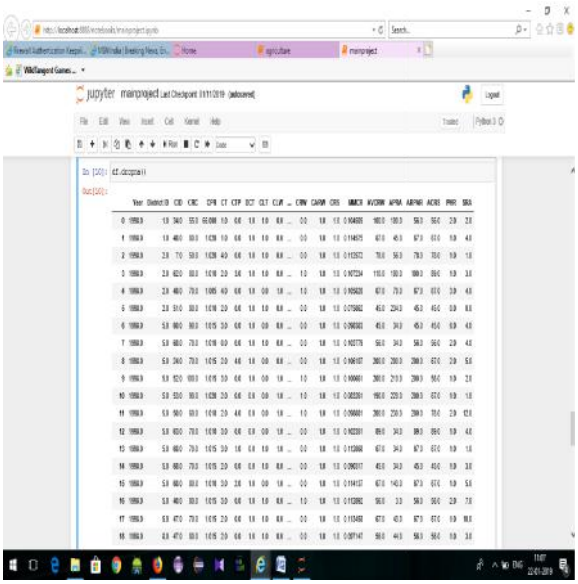

Figure.11 Dropping of null values

In Figure.11 Null values are dropped in each row.

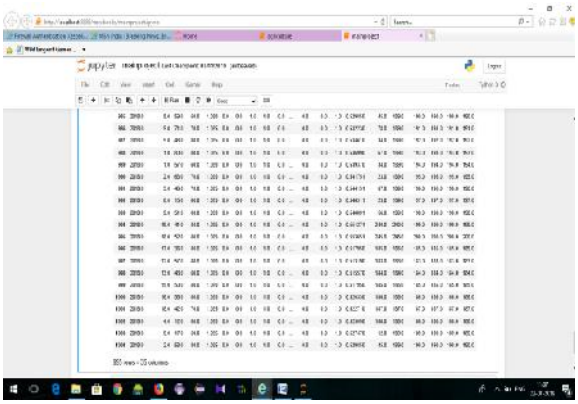

Figure.12 After Dropping null values

In Figure.12 After dropping of the null values the actual number of rows in the dataset is reduced.

4. Conclusion

In this analysis, we used some of the common data mining techniques in the field of agriculture. Some of these techniques, such as the k-means, k nearest neighbor, SVM, and bayesian network are discussed and an application in agriculture for each of these techniques is presented. Data mining in agriculture is an upcomming research field. Efficient techniques can be developed and used for solving complex agricultural problems using data mining. Future enhancement of this agriculture analysis is to predict the crop yield using these techniques. It is useful for making crop descisions for farmers and government organizations. In future, the ANN and NN classification approach can be used for the better classification and improve the classification performance of the crop yield prediction. It could understanding of the high dimensional between complex yearly and seasonal climatic patterns which determine crop yield helps both farmers and other decision makers to be able to predict the effects of drought and other climatic conditions.

6. References

[1] Veenadhari S, Misra B, Singh CD. Data mining techniques for predicting crop productivity—A review article. In: IJCST. 2011; 2(1).

[2] Jain A, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv. 1999;31(3):264–323.

[3] Berkhin P. A survey of clustering data mining technique. In: Kogan J, Nicholas C, Teboulle M, editors. Grouping multidimensional data. Berlin: Springer; 2006. p. 25–72

[4] V. Arulkumar. "An Intelligent Technique for Uniquely Recognising Face and Finger Image Using Learning Vector Quantisation (LVQ)-based Template Key Generation," International Journal of Biomedical Engineering and Technology 26, no. 3/4 (February 2, 2018): 237-49.

[5] Han J, Kamber M. Data mining: concepts and techniques. Massachusetts: Morgan Kaufmann Publishers; 2001.

[6] C.V. Arulkumar, G. Selvayinayagam and J. Vasuki, "Enhancement in face recognition using PFS using Matlab," International Journal of Computer Science & Management Research, vol. 1(1), pp. 282-288, 2012

[7] H. Anandakumar and K. Umamaheswari, "A bio-inspired swarm intelligence technique for social aware cognitive radio handovers," Computers & Electrical Engineering, vol. 71, pp. 925–937, Oct. 2018.

[8] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise.

In: Paper presented at International conference on knowledge discovery and data mining. 1996

[9] C.V. Arulkumar et al., "Secure Communication in Unstructured P2P Networks based on Reputation Management and Self Certification", International Journal of Computer Applications, vol. 15, pp. 1-3, 2012.

[10] Ramesh D, Vishnu Vardhan B. Data mining techniques and applications to agricultural yield data. In: International journal of advanced research in computer and communication engineering. 2013; 2(9).

[11] ability of machine learning methods for massive crop yield prediction. Span J Agric Res. 2014;12(2):313–28