**Big Data Analytics in Crop yield prediction of Tamil Nadu (Rice)**

Dr.R.Gangai Selvi,

Associate Professor (Statistics)

Department of Physical Sciences and Information Technology

Agricultural Engineering and Research Institute,

Tamil Nadu Agricultural University,

Coimbatore-641003.

ganga\_agri@yahoo.com

**Abstract**

Rice is one of the important, main staple food and cultivated crop. After wheat and sorghum, rice is the third most valuable food crop worldwide.The present work is carried out to employ data analytics, machine learning methodologies for rice data and to create association rules on fixed attributes and their correlations for yield prediction of crops. Tamil Nadu's secondary data of Rice crop for 20 years based on the following factors area, production, yield, temperature, rainfall, humidity, and wind speed were used in this study. On a prepared dataset, the pre-processing processes are conducted to run data analytics and machine learning algorithms. K-Means cluster algorithm is the method employed in this study to categorize data on rice productivity and Apriori algorithm is used to generate association rules using the processed data. For the prediction of yield, spatial regression method is also applied. According to the results of the data analysis using the pre-defined k=3 clusters, the crop yield for rice collected from twenty-eight (28) districts was divided into three clusters based on the distance to the nearest centroid. Additionally, it was also revealed that when districts are grouped together, the amount of production of the rice harvest is similar. Using the Apriori method on a generated rice dataset, minimum support and confidence levels equal to 0.001 and 90% respectively were selected, while this algorithm produced many association rules of variables. Among them, results of 31 related rules were suggested for "High Yield Production". The districts of rice crop yield are optimised using the spatial and non-spatial regression model and validated using R2 and Root Mean Square Error. The main result of this study is a set of efficient and well-constructed association rules for yield prediction, which will be useful for scholars, farmers, and government officials to increase the productivity of rice crops.

*Keywords: K-means clustering, Apriori algorithm, Rice data, Spatial and Non- Spatial regression model*

1.Introduction

To tackle the increasing challenges of agriculture production, the complex agricultural ecosystems need to be better understood. This can happen by means of modern digital technologies that monitor continuously the physical environment, producing large quantities of data in an unprecedented pace. In an era where technology has reached the pinnacle of its use and has completely overpowered our lives, the amount of data exchanged is enormous. The high volumes of data sets, that a traditional computing tool cannot process, are being collected daily. We refer to these high volumes of data as big data. Big data offers tremendous hope in Indian agriculture. However, the implementation is likely to be “bumpy”, sporadic and may take quite long to realize substantial benefits. Big data has the potential to create the next major technological “sea change” in agriculture. The technology may change the “balance of power” in the agri-food value chain.

India has a large agrarian economy with majority of its rural population existing on farming. Rice is the most important staple food in Asia. More than 90 per cent of the world’s rice is grown and consumed in Asia, where 60 per cent of the world’s population lives. India is the second leading producer of rice in the entire world, preceding only by china. Tamil Nadu one of the leading rice growing states in India, has been cultivating rice from time immemorial as this state is endowed with all favourable climatic conditions suitable for rice growing. The state has 2.2 million ha. under rice cultivation, which covers mainly irrigated and partly rainfed areas. The state average productivity is 2.8 tonnes/ha. the proper planning and proper sowing only give the expected yield otherwise there is a loss to the farmers. The main aspect of the government is to fulfill enough storing of crop for long-term, mainly at a time of natural disaster. In this research, an attempt made to predict the yield of rice using big data analytics.

The Department of agriculture has been created many programmers to educate the farmers to grow proper crop in proper time. The government has introduced many classes to the farmers to transfer the latest technical knowledge in production and productivity. Since rice cannot be grown in all the regions of the state, to grow the crop rice the region must support to the climate, humidity, rainfall, soil features. If any of the above is not support their will be loss to the farmers. Hence taking necessary measures to gain maximum yield of rice during the respective seasons become the main priority of the farmers. Storing huge amount of data has given to processing and predicting the yield of rice accurately in the future.

**OBJECTIVES**

1. To develop modules for pre-processing of data

2. To extract the crop yield parameters from bigdata

3. To predict the crop yield using spatial panel models

4. To suggest suitable policy measure for crop yield improvement

**2.REVIEW OF LITERATURE**

James MacQueen (1967), initially applied the idea of K-Means examined a procedure whose primary objective is to divide a basic sample into k sets to have an effective within-class variance and the K-Means method is the ideal approach to clustering or similarity grouping issues, enabling any researcher to gain qualitative knowledge of massive volumes of data by supplying him with reasonably accurate similarity groupings.

Bernhardt et al. (1996), used cluster analysis to classify farms for conventional/system approach research addressed that a multidisciplinary project named Agriculture in Concert with the Environment (ACE) proposed the clustering K-Means approach was developed to examine current whole-farm system groupings, compare them to alternatives and assess their socioeconomic position.

 Urtubia et al. (2007) applied data mining techniques to predict industrial wine problem fermentations had shown data mining tools to detect and spot relationships for classifying winemaking fermentations early on. It uses the datasets of 24 industrial fermentations of Cabernet sauvignon to explore data mining to detect anomalous behaviors and held periodic measurements of 29 components that included sugar, alcohols, organic acids and amino acids. It was detected by two-stage classification procedures such as Principal Component Analysis (PCA) and K-Means Clustering process which concludes that Detection of over 70 percent of the problematic fermentations within 72 hours.

Kumar (2011) studied the impact of climate change on Indian agriculture by accounting for spatial feature that may influence the climate sensitivity of agriculture. An estimate of climate change's impact on farm level net revenue in India was based on panel data over a period of 20 years covering 271 districts. It was found that there was a positive spatial autocorrelation which increased the accuracy of the results. The results revealed that climate change had a less severe impact on agricultural net revenue.

Chakraborty et al. (2012). Weather forecasting using incremental K-Means Clustering explained that clustering is a powerful tool which used in various forecasting tools. The generic incremental K-mean clustering algorithm is proposed in this study as a method for weather forecasting. The primary air pollution database will be used in this study's usual K-Means Clustering and a list of weather categories will be created using the clusters' peak mean values. Whenever new data are coming, the incremental K-Means is used to group data into those clusters where the weather category has been already defined. Thus, it can predict weather information in the future.

Singh et al. (2012) Finding the chances and prediction of cancer through Apriori algorithm with transaction reduction suggested that the Apriori method, which is regarded as influential for mining popular item sets for Boolean association rules with transaction reduction that might be used to detect cancer from its symptoms, was examined. They researched to learn more about the cancer kind that spreads the fastest as well as the signs by which cancer spreads.

Chakir and Le Gallo (2013) predicted the future land use allocation in France using spatial panel data. In this study secondary data of variables like area under Agriculture, Forest, Urban and other uses were taken from Teruti survey from 1992 to 2003 all over the country.

Elhorst and Vega (2013) pointed out an issue related to spatial econometric modelling, spillover effects, and spatial weight matrix W with the findings of their research and offered strategies for selecting a model specification, which were interesting and promising steps for applied research involving spatial econometrics.

Charliepaul and Gnanadurai (2014) Comparison of K-mean algorithm and apriori algorithm- An analysis had been implemented in the software, python. They studied the method of K-Means Clustering and apriori algorithm briefly with a suitable real-time example. Also deliberately given the comparison between the clustering and association rules.

 Gangai Selvi and Mani (2015) studied Land Use dynamics in Tamil Nadu through a Spatial Econometric modeling approach analyzed the temporal and spatial changes in land use categories.The major factors which are influenced for agricultural land use changes were growth in human population, irrigation, rainfall, temperature, wage rate, fertilizer, demand for non-agricultural uses such as industries, housing, roads and other development infrastructure such as education institutions, health and other rural and urban amenities.

 Permai*et al.,* (2019) analysed the average expenditure of Papua province using linear regression method with OLS and Spatial Autoregressive (SAR) method. In this study average expenditure as independent variable and the eight dependent variable which affects the expenditure of Papua province were taken from 28 districts of Papua. Based on the smaller RMSE and AIC it was concluded that the SAR model was better than OLS.

**3.METHODOLOGY**

The details of the research methodology used for this study are discussed in detail under the following heads.

**3.1. Description of the Study area**

 Therefore, a brief description of the research area's location, size, climate, soil type, irrigation coverage and other factors that could have an impact on rice yield, either directly or indirectly is provided. The study was related to the Rice crop in the aspect of overall districts of Tamil Nadu (2000-01 to 2019-20).

**3.2. Nature and Collection of Data**

The study was primarily based on secondary data. Secondary data on Rice were collected for the entire state, collected for the period of 20 years from 2000-01 to 2019-20 of 28 districts of Tamil Nadu.

**3.3. Selection of variables in the crop yield prediction model**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Variable** | **Definition** |
|  | **Dependent variable** |
| 1 | *PRODVTY* | Productivity of Rice |
|  | **Independent variables** |
| 2 | *AREA* | Area of Paddy |
| 3 | *PRODN* | Production of Rice |
| 4 | *AUI* | Area under irrigation (m3/ha) |
| 5 | *MAX\_TEMP* | Maximum Temperature (°C) |
| 6 | *MIN\_TEMP* | Minimum Temperature (°C) |
| 7 | *RAIN* | Rainfall ( mm) |
| 8 | *HUMD* | Relative Humidity |
| 9 | *SOIL* | Soil Moisture |
| 10 | *WS* | Wind speed ( m/sec) |
| 11 | *WD* | Wind Direction |

**3.4.K-Means Clustering**

An unsupervised clustering procedure called K-Means Clustering divides the input data points into different classes based on how similar they are to one another. By reducing the sum of squared distances between the data points, the grouping is accomplished. Distance measures and similarity measures are the two primary categories of measurements. The similarity or dissimilarity of the pair of items is determined using distance measurements. Since K-Means is the most basic type of clustering, it only groups the data as a crisp set and has limits when dealing with high-dimensional and constrained data. Today, clustering very big-scale data is a difficult issue because in the real world, with the advancement of information technology, the volumes of data processed by many applications are crossing the Peta scale threshold. This study enhances the performance of the fundamental K-Means Clustering algorithm.

**Procedure**

Step 1: The procedure randomly chooses K points to serve as the first cluster centers (or "means").

Step 2: Based on the Euclidean distance between each point and each cluster centre, each point in the dataset is assigned to a closed cluster.

Step 3: The average of the points within each cluster is recomputed for each cluster centre.

Step 4: Repetition of steps 2 and 3 causes the clusters gradually converge.

K-Means is a well-known technique in unsupervised learning and vector quantization. The K-Means Clustering is formulated by minimizing a formal objective function, meansquared-error distortion.

𝑁

𝑚𝑖𝑛𝑖𝑚𝑢𝑚𝑀𝑆𝐸(𝑃) = ∑ ||𝑥𝑖− 𝐶(𝑖)||2

𝑖=1

where

N is the number of data samples; K is the number of clusters; d is the dimension of the data vector;

𝑋 = {𝑥1, 𝑥2, … .𝑥𝑁} is a set of N data samples;

𝑃 = {(𝑖)|𝑖 = 1, … .𝑁} is the class label of X; 𝐶 = {𝑐𝑗|𝑗 = 1, … .𝑘} are k cluster centroids.



**Fig 1.Flow chart of K-Means Clustering**

**3.5.Apriori algorithm**

The Apriori algorithm is the original algorithm of Boolean association rules of mining frequent item sets, raised by Agarwal and Srikant in 1994. Frequent item sets are subsets of frequent item sets, whereas infrequent item sets are the supersets of frequent item sets. The name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent item sets are used to find k+1 item sets.

Apriori Algorithm of data mining are the following

1. Join Step: This step generates (K+1) item sets from K-item sets by joining each item with itself.
2. Prune Step: This step scans the count of each item in the database. The candidate item is deleted, if it does not receive the required amount of support since it is deemed to be uncommon. To make the candidate item sets smaller, this step is carried out.

**Steps in the Apriori algorithm**

1. In the first iteration of the algorithm, each item is taken as a 1-itemsets candidate. Each item's occurrences will be counted by the algorithm.
2. Let there be some minimum support, min sup. The set of 1 item sets whose occurrence is satisfying the min sup are determined. Only those candidates which count more than or equal to min sup, are taken ahead for the next iteration and the others are pruned.
3. Next, 2-itemset frequent items with min sup are discovered. For this in the join step, the 2-itemset is generated by forming a group of 2 by combining items with itself.
4. The 2-itemset candidates are pruned using the min-sup threshold value. Now the table will have 2 item sets with min-sup only.
5. The next iteration will form 3 –item sets using the join and prune step. This iteration will follow the antimonotone property where the subsets of 3-itemsets, that is 2 itemset subsets of each group fall in min sup. If all 2-itemset subsets are frequent then the superset will be frequent otherwise it is pruned.
6. Next step will follow making 4-itemset by joining 3-itemset with itself and pruning if its subset does not meet the min sup criteria. The algorithm is stopped when the most frequent itemset is achieved.



**Fig 2.Flow Chart of Apriori**

**3.6.Spatial Panel Data Models**

In panel data models, observations are indexed by *i*(district) and *t* (time). Under the assumption of balanced panels, most formulations of the model adopt an unobserved heterogeneity perspective and in the present study, an unobserved time-invariant covariate ‘*i*’ exists. If ‘α*i*’is correlated with the observed covariates ‘*xit*’,then the disturbance term cannot absorb ‘α*i*’. Thus, in the case of the SAR model, the (spatial) fixed-effects model (Elhorst, 2003) is indicated as follows.

  for i = 1, 2, ... , R; t = 1, 2, …,T ..….. (1)

Fortunately, the problem can be solved in the spatial context in much the same way as in the non-spatial context: by de-meaning the data, district-wise. However, as Anselin*et al.*, (2008) observed, the computation of the means is complicated by the spatial dependencies (the *W* matrix), and must be done carefully. But given a correct de-meaning, then just as in the non-spatial context, a regression equation without the fixed effects (the α’s) can be obtained.

Returning to the case of a temporally invariant covariate (α*i*) if it can be assumed notto be correlated with the observed *x*’s, then in principle it could be absorbed into the disturbance term, resulting in the spatial random effects model. For example, Elhorst (2009) describes a SEM modelin which,

*yit =xitβ +uit* ...….. (2)

*uit=*+ ε*it* ...….. (3)

ε*it= λW*ε*it+υit* ...….. (4)

*υit= ρυit-1+eit* where*eit~IIDN(0, σe2)*...….. (5)

Where, = observation for the *ith* district/individual at the *tth* time period,

*xit* = k x 1 vector of observations on the non-stochastic regressors,

*uit*= regression disturbance,

*ρ*=spatial lag or spatial auto regressive parameter in (1),

*λ* = spatial error dependence or spatial auto correlation parameter,

*ρ*=(time-series) first-order correlation coefficient in (5) and

W = R x R spatial row-standardized weight matrix whose diagonal elements are zero, such that (IR – *ρ*W) is non-singular, where IR is an identity matrix of dimension ‘R’.

On the one hand, the spatial weights matrix expresses the spatial connectivity of the system: each element [wij] of the matrix indicates how observation ‘i’ is spatially connected to observation ‘j’. For instance, two observations may be considered as spatially connected, if they share a common border or if they are located within a certain distance of one another. On the other hand, given the definition of the spatial weights matrix, each spatial autoregressive coefficient ‘*ρ*’ indicates the intensity of spatial error auto-correlation. In this model, both the parameters ‘*β*’ and the spatial autoregressive coefficient ‘*ρ*’ are allowed to vary across equation, but, they are assumed to be constant over time. Clearly, this is a strong assumption that the model makes (Chakir and Gallo, 2013).

Alternatively, one may first test whether spatially lagged independent variables must be included and then whether the model should be extended to include a spatially lagged dependent variable or a spatially auto-correlated error term (Florax and Folmer 1992, Elhorst and Freret 2007) or adopt an unconstrained spatial Durbin model and then, test whether this model can be simplified (Elhorst*et al.,* 2006; Ertur and Koch, 2007).

An unconstrained spatial Durbin model (SDM) with spatial fixed effects takes the form

 ...…..(6)

Where, θ, just as *β*, is an (*k*,1) vector of fixed but unknown parameters.

**3.7.Non-Spatial linear regression Model (OLS Model)**

The typical strategy in most empirical research is to begin with a non-spatial linear regression model and then examine if the model needs to be extended with spatial interaction effects. The particular to general method is the name given to this strategy. The form of the non-spatial linear regression model is

*Y = Xβ + u* ...…..(7)

Where

Y = *R* x1 vector of observations on the dependent variable,

*R* = No. of districts,

X = *R* x *k* matrix of observations on the exogenous variables, with associated *k* x 1 regression coefficient vector β and

u = vector of the error term.

**3.8.Spatial Lag Model**

This model is also known as the spatial autoregressive model. The dependent variable 'Y' levels are said to be dependent on the 'Y' levels in neighbouring locations. Thus, it is an expression of the notion of a geographical overflow. SAR, or the Spatial Auto-Regressive Model, is

*Y =* ρ*WY + X*β + u ...…..(8)

Where

Y = *R* x1 vector of observations on the dependent variable,

*R =* No. of districts,

W = *R* x *R* spatial weights matrix (with 0 diagonal elements),

ρ = spatial autoregressive coefficient or the spatial lag parameter,

WY = spatially lagged dependent variable representing an average of spatially neighbouring Y values,

X = *R* x *k* matrix of observations on the exogenous variables, with associated *k* x 1

regression coefficient vector β and u = vector of the error term.

**4.RESULTS AND DISCUSSION**

**4.1.K-Means clustering performance**

The K-Means clustering performs all the variables which have been taken into account. The target/ dependent variable is taken as productivity and all the other variables were taken the independent listed as Area, Area under irrigation, Production, Rainfall, Minimum and Maximum Temperature, Relative humidity, Soil moisture, Wind direction and Wind speed.



**Fig 3. Elbow Technique**

 The number of K is determined by the Sum of squared error (SSE). This sort of produces an “elbow effect” in the graph. In the above Figure an elbow occurring with 3 to 4 is a good number to choose this plot. SSE is often used as a research reference in determining optimal clusters.

****

**Fig 4.1 Scatter plot of Production vsProductivityFig 4.2 Scatter plot of Area vsProductivity**

Fig 2.1 The groups were depicted on the graph in different colours green, red and black based on the centroids of cluster 0,1,2 respectively showing the production versus productivity which was considered as the target variable. By the obtained plot Fig 2.2, the cluster 0 has given the major productivity which gained the high cluster. The area which has low productivity according to cluster 2.

****

**Fig 4.3 Scatter plot of Rainfall vs Productivity Fig 4.4 Scatter plot of RH vs Productivity**

****

**Fig 4.5 Scatter plot of Min Temp vs Productivity Fig 4.6 Scatter plot of Max Temp vs Productivity**



**Fig 4.7Scatter plot of Soil Moisture and Productivity Fig 4.8Scatter plot of Wind speed and Productivity**

During the period of SW monsoon there is a significant increase in the frequency of rainy days in the districts Thanjavur, Thiruvallur, Thiruvarur, Kanchipuram, Villupuram, Cuddalore, Perambalur and Kanniyakumari which have higher productivity of rice cop-up with the needed rainfall pattern of paddy.

**4.2 Apriori algorithm**

By applying Apriori rules in the data of 11 variables, it resulted in generation of 432 rules. But only 31 rules were selected based on minimum support and the number of highest counts in the pairwise combination. The value of Support and coverage ranges from 0.214 to

0.321 with confidence level of 1%. Number of counts considered for best combination ranges from 9-6 with lift of ranging between 2.800 to 3.111. So, we can reduce the rules number and increase the trust value by filtering according to the support. The result changed from 7412 to 432 and we can reduce it more and more.

**4.3 Spatial Regression Model for Rice yield (Productivity)**

When the comparison of OLS (Non-Spatial regression model) and Spatial regression model made to predict Rice yield. The variables Area, Production, Area under irrigation, Rainfall were positively influenced the target variable Productivity with the level of significance at 1% and 5% respectively. Lagged variable productivity shows significance at 10% level. Also higher R2 shows good fit of the model and has minimum RMSE value.

**Regression Model**

 **Dependent Variable: Productivity**

|  |  |  |
| --- | --- | --- |
| **Variables** | **Regression without Spatial Effect - OLS****(Non-Spatial Regression Model)** | **Regression with Spatial Effect****SAR Model** |
| Constant | 40.1145  | 35.4471  |
| Production | 0.0127\*\* | 0.0122\*\*  |
| Area | 0.0016\*\*\* | 0.0015\*\*\*  |
| AUI | 0.0117\*\*  | 0.0017\*\*\*  |
| Rainfall | 0.0166\*\*  | 0.0016\*\*\*  |
| Min Temp | -0.0328  | -0.0368  |
| Max Temp | -0.0460  | -0.0547  |
| RH | 0.0333\*\*  | -0.0553  |
| Soil Moisture | 1.9950 (NS) | 1.9708 (NS) |
| Wind direction | -0.0032  | -0.0003  |
| Wind speed | 0.1972 (NS) | 0.2070 (NS) |
| *WX*\_Productivity | -  | 0.0324\*  |
| F  | 74.0699  | 70.5227  |
| R2  | 0.8287  | **0.8394**  |
| Adjusted R2  | 0.8176  | 0.8179  |
| AIC  | 984.13  | 983.85  |
| LL  | -334.09  | **-332.20**  |
| RMSE  | 0.2752  | **0.2733**  |

 \* - Significance at 5% level

 \*\* - Significance at 10% level

 \*\*\* - Significance at 0.1% level

 (NS) – Not Significant

**5. Outcome**

This research essentially deals with effective recommendation system for the agriculture for analyzing the data and identifying the most suitable pairwise variables of rice crop using clustering and association rule technique respectively. These proposed clustering and Apriori and spatial model approach provide the best result which can be helpful for predicting the paddy yield accurately.

**6.References**

Aishwarya, S.P., S. Pramod and K. Anita. 2019. ‘Yield Prediction of Paddy based on Temperature and Rain Fall Using Data Mining Techniques’ *International Journal of Recent Technology and Engineering* 8 (2S11):65-70.

Anselin, Luc, 2003. ‘Spatial externalities, spatial multipliers, and spatial econometrics’ *International regional science review* 26 (2):153-166.

Baltagi, Badi H, SeuckHeun Song, Byoung Cheol Jung, and Won Koh. 2007. "Testing for serial correlation, spatial autocorrelation and random effects using panel data." *Journal of Econometrics* 140 (1):5-51.

Bansal, A., M. Sharma, and S. Goel. 2017. ‘Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining’ *International Journal of Computer Applications* 157 (6):0975-8887.

Brossette, S.E., A.P. Sprague, J.M. Hardin, K.B. Waites, W.T. Jones, and S.A. Moser. 1998. ‘Association rules and data mining in hospital infection control and public health surveillance’ *Journal of the American medical informatics association* 5 (4):373-381.

Celik, A. 2020. ‘Using Apriori Data Mining Method in COVID-19 Diagnosis’ *Journal of Engineering Technology and Applied Sciences* 5 (3):121-131.

Chakraborty, S., N. Nagwani, and L. Dey. 2012. ‘Weather Forecasting using Incremental K-Means Clustering’ *International Journal of Biometrics and Bioinformatics.* 4 (3).

Charliepaul, C.K., and G.I. Gnanadurai. 2014. ‘Comparison of K-mean algorithm and Apriori algorithm–An analysis’ *International Journal On Engineering Technology and Sciences–IJETS™* 1 (III).

Dharshinni, N., F. Azmi, I. Fawwaz, A. Husein, and S.D. Siregar. 2019. ‘Analysis of accuracy K-Means and apriori algorithms for patient data clusters’ *Journal of Physics: Conference Series*.

Gangai Selvi. R and Mani. K (2015), Ph.D thesis titled ‘Land Use Dynamics in Tamil Nadu – A Spatial Econometric Analysis’ TNAU, Coimbatore.

Hayatu, I.H., A. Mohammed, B.A. Ismaâ, and S.Y. Ali. 2020. ‘K-Means clustering algorithm based classification of soil fertility in north west Nigeria’ *FUDMA JOURNAL OF SCIENCES* 4 (2):780-787.

Mirmozaffari, M., A. Alinezhad, and A. Gilanpour. 2017. ‘Data Mining Apriori Algorithm for Heart Disease Prediction’ *Int'l Journal of Computing, Communications & Instrumentation Engg*4 (1):20-23.

Muñoz, P., R. Barco, E. Cruz, A. Gómez-Andrades, E.J. Khatib, and N. Faour. 2017. ‘A method for identifying faulty cells using a classification tree-based UE diagnosis in LTE’ *EURASIP Journal on Wireless Communications and Networking* 2017 (1):1-20.

Narkhede, U.P., and K.P. Adhiya. 2014. ‘Evaluation of modified K-Means clustering algorithm in crop prediction’ *International Journal of Advanced Computer Research* 4 (3):799.

Nikita, S.S., and S.S. Sambare. 2021. ‘Crop yield Prediction Using Apriori Algorithm And Machine Learning Technique’ *International Journal of Current Engineering and Technology* (8):916-920.

Palanivel, K., and C. Surianarayanan. 2019. ‘An Approach for Prediction of Crop Yield using Machine Learning and Big Data Techniques’ *International Journal of Computer Engineering and Technology* 10 (3).