BIG DATA ANALYTICS IN CROP YIELD PREDICTION OF TAMIL NADU (RICE)

Abstract

Rice holds significant importance as a staple food and cultivated crop, ranking as the third most valuable food crop globally, following wheat and sorghum. This study focuses on leveraging data analytics and machine learning techniques to analyse rice-related data and establish correlations between fixed attributes to predict crop yield. The dataset pertains to Rice cultivation in Tamil Nadu over a span of 20 years and includes factors like area, production, yield, temperature, rainfall, humidity, and wind speed. The dataset underwent preprocessing to facilitate the application of data analytics and machine learning algorithms. The K-Means clustering algorithm was utilized to categorize rice productivity data, while the apriori algorithm was employed to extract association rules from the processed data. To predict yield, a spatial regression method was also utilized. Based on the data analysis results, employing a predefined k=3 clusters, the crop yield data from 28 districts were grouped into three clusters based on their proximity to the nearest centroid. Furthermore, it was observed that districts grouped together exhibited similar rice production levels.By applying the apriori method to the rice dataset, with a minimum support of 0.001 and a confidence level of 90%, numerous association rules was generated. Among these, 31 pertinent rules were identified for achieving "High Yield Production". The study optimized the districts' rice crop yield using both spatial and non-spatial regression models, validating the results using metrics like R2 and Root Mean Square Error. The study's primary outcome is a collection of well-defined and effective association rules that can facilitate yield prediction. These findings are anticipated to be valuable for

Author

Dr. R. Gangai Selvi
Associate Professor (Statistics)
Department of Physical Science & Information Technology
AEC&RI, TNAU
Coimbatore, Tamil Nadu, India.
ganga_agri@yahoo.com
gangastat@tnau.ac.in

researchers, farmers, and government authorities aiming to enhance rice crop productivity.

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

I. INTRODUCTION

To address the growing challenges in agricultural production, there's a crucial need to enhance our understanding of intricate agricultural ecosystems. Modern digital technologies play a pivotal role in achieving this understanding by continuously monitoring the physical environment and generating vast amounts of data at an unprecedented pace. In a time where technology's dominance is pervasive and data exchange is colossal, we encounter large datasets that conventional computing tools struggle to handle. This collection of substantial data is commonly referred to as big data. In the context of Indian agriculture, big data holds significant promise, yet its practical implementation might be uneven, intermittent, and require substantial time to yield significant advantages. The potential of big data in revolutionizing agriculture is substantial, potentially reshaping power dynamics along the agri-food value chain.

India boasts a substantial agrarian economy, with a significant portion of its rural populace engaged in farming. Particularly in Asia, rice stands as a vital staple food. Asia accounts for over 90% of global rice production and consumption, catering to 60% of the world's population. India, ranking second globally, is a major rice producer, following only China. Tamil Nadu, a prominent rice-growing state in India, has a rich history of rice cultivation owing to its favorable climatic conditions. The state cultivates rice across 2.2 million hectares, primarily encompassing irrigated and partially rainfed areas. The average state productivity stands at 2.8 tonnes per hectare. Successful yield largely hinges on meticulous planning and appropriate cultivation practices, as any deviations lead to losses for farmers. Government initiatives center around ensuring adequate storage of crops for long-term sustainability, especially during natural disasters. This study endeavors to forecast rice yield using the analytical potential of big data.

The Department of Agriculture has initiated numerous programs to educate farmers about cultivating suitable crops at the right time. The government offers various training sessions to impart the latest technical insights into production and productivity. Since rice cultivation is contingent on regional compatibility with factors like climate, humidity, rainfall, and soil characteristics, a lack of support in any of these areas translates to losses for farmers. Consequently, ensuring optimal rice yields during specific seasons becomes paramount. The accumulation of extensive data has paved the way for accurate processing and predictive models to estimate future rice yields.

OBJECTIVES

- To develop modules for pre-processing of data
- To extract the crop yield parameters from bigdata
- To predict the crop yield using spatial panel models
- To suggest suitable policy measure for crop yield improvement

II. 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 was employed to categorize farms based on their conventional or system-oriented approaches in a study that focused on a collaborative project known as Agriculture in Concert with the Environment (ACE). The project aimed to comprehensively explore and evaluate various farm systems. To achieve this, the K-Means clustering technique was utilized to analyze existing farm system classifications, make comparisons with alternative approaches, and evaluate the socioeconomic status of these systems.

Urtubia et al. (2007) applied data mining techniques in forecasting problematic fermentations in the industrial wine production domain has demonstrated their effectiveness in identifying and understanding correlations for the early classification of winemaking fermentation issues. This approach involves analyzing datasets from 24 Cabernet sauvignon fermentations conducted in an industrial setting. The primary goal is to employ data mining tools to uncover unusual patterns and behaviors in these processes. The datasets encompass periodic measurements of 29 components, including sugar, alcohols, organic acids, and amino acids. By employing a two-stage classification process involving Principal Component Analysis (PCA) and K-Means Clustering, the study successfully identifies anomalies in fermentations, leading to the detection of over 70 percent of problematic instances within a 72-hour timeframe.

Kumar (2011) studied the influence of climate change on Indian agriculture while considering spatial factors that could affect agricultural climate sensitivity is a focus of this study. The analysis involves evaluating the potential impact of climate change on net revenue at the farm level in India, utilizing panel data spanning two decades and encompassing 271 districts. The findings indicate the presence of a favorable spatial autocorrelation pattern that enhances result precision. The study's outcomes highlight that the consequences of climate change on agricultural net revenue are comparatively less severe.

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) studied the likelihood and forecasting of cancer occurrences using the Apriori algorithm, coupled with transaction reduction, was the focal point of this investigation. The Apriori technique, renowned for its effectiveness in extracting prevalent item sets for Boolean association rules while employing transaction reduction, was employed to identify potential indications of cancer based on its symptoms. The research aimed to gain deeper insights into the specific type of cancer characterized by rapid spread and the specific symptoms indicative of its metastasis.

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.

Permaiet 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.

III. METHODOLOGY

The specifics of the research approach employed for this study are elaborated comprehensively in the subsequent sections.

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).

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. 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 (m^3/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

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 minimizing the total squared distances among 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 process involves selecting K points at random to act as initial 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 widely recognized method in the realm of unsupervised learning and vector quantization. The formulation of K-Means Clustering revolves around the

minimization of a formal objective function, specifically the mean-squared error distortion.

$$minimumMSE(P) = \sum_{i=1}^{N} ||x_i - C_{(i)}||^2$$

where

N is the number of data samples;
K is the number of clusters; d is
the dimension of the data vector;
$X = \{x_1, x_2, \ldots .x_N\}$ is a set of N data samples;
$P = \{p(i)|i = 1, \ldots .N\}$ is the class label of X;
$C = \{c_j|j = 1, \ldots .k\}$ are k cluster centroids.

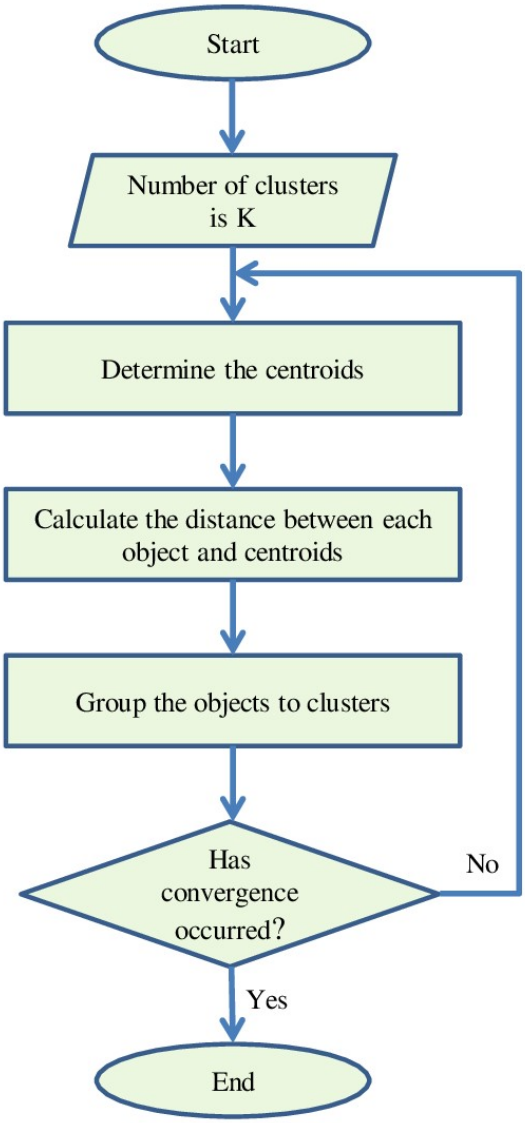

Figure 1: Flow Chart of K-Means Clustering

5. **Apriori Algorithm:** The apriori algorithm, developed by Agarwal and Srikant in 1994, serves as the foundational technique for mining frequent item sets through Boolean association rules. Frequent item sets represent subsets that occur frequently, while infrequent item sets encompass the supergroups of these frequent subsets. The algorithm is named "Apriori" due to its reliance on pre-existing knowledge about the properties of frequent item sets. This method employs an iterative strategy or a step-by-step exploration, where k-frequent item sets are utilized to discover sets with k+1 items.apriori Algorithm of data mining are the following

   - **Join Step:** This stepcreates sets with (K+1) items by combining each item with itself from the existing K-item sets.
   - **Prune Step:** During this phase, the database is examined to tally the occurrences of each item. If a candidate item fails to meet the necessary support threshold, indicating its infrequency, it is eliminated. This step is undertaken to streamline the candidate item sets by reducing their size.

   Steps in the Apriori Algorithm

   - During the initial iteration of the algorithm, individual items are treated as candidates for 1-itemsets. The algorithm proceeds to calculate the occurrences of each item.

   - Assuming a specified minimum support level, denoted as min sup, the algorithm identifies the collection of 1-item sets whose occurrences meet the min sup requirement. Only those candidates with counts equal to or greater than min sup are retained for further iterations, while the rest are discarded.

   - Subsequently, the algorithm proceeds to identify frequent items within 2-item sets that satisfy the min sup criterion. To achieve this, the join step involves creating 2-item sets by pairing items with themselves.

   - The 2-itemset candidates are filtered out based on the min-sup threshold value. Consequently, the table will exclusively contain 2-item sets that adhere to the min-sup requirement.

   - In the subsequent iteration, 3-item sets will be created through the join and prune process. This iteration adheres to the antimonotone property, where subsets of 3-itemsets – specifically, the 2-itemset subsets of each group – need to meet the min sup criteria. If all 2-itemset subsets are frequent, the superset becomes frequent; otherwise, it is eliminated.

   - Subsequently, the process advances to generate 4-itemsets through the combination of 3-itemsets with themselves, followed by the elimination of candidates if their subsets fail to satisfy the min sup requirement. The algorithm concludes when the most frequent itemset is attained.

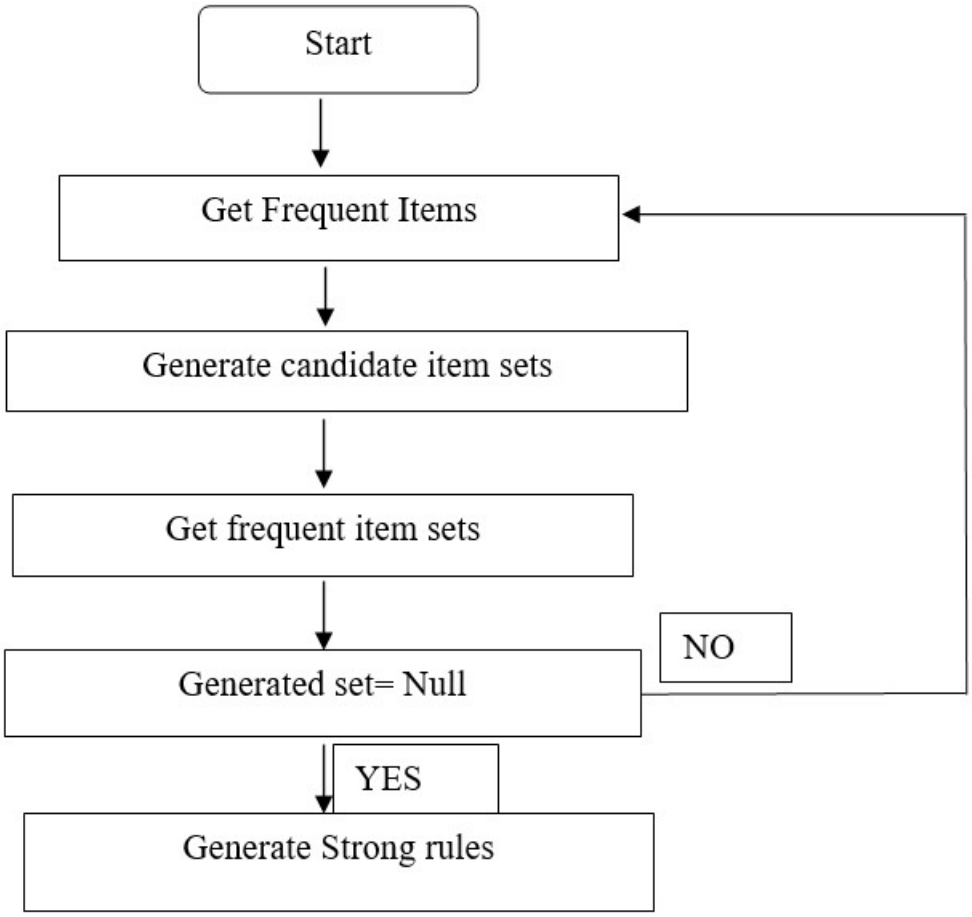

Figure 2: Flow Chart of Apriori

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 'x_{it}',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.

$$y_{it} = \rho \sum_j w_{ij} y_{ij} + \alpha_i + x_{it}\beta + u_{it} \quad \text{for i = 1, 2, ... , R;} \quad t = 1, 2, \ldots, T \quad \ldots\ldots(1)$$

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 Anselinet 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,

$y_{it} = x_{it}\beta + u_{it}$ ..... (2)

$u_{it} = \alpha_i + \varepsilon_{it}$ ..... (3)

$\varepsilon_{it} = \lambda W\varepsilon_{it} + \upsilon_{it}$ ..... (4)

$\upsilon_{it} = \rho\upsilon_{it-1} + e_{it}$ where $e_{it} \sim IIDN(0, \sigma_e^2)$ ..... (5)

Where

y_{it} = observation for the i^{th} district/individual at the t^{th} time period,

x_{it} = k x 1 vector of observations on the non-stochastic regressors,

u_{it}= 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 (I_R – ρW) is non-singular, where I_R is an identity matrix of dimension 'R'.

On the one hand, the spatial weights matrix expresses the spatial connectivity of the system: each element [w_{ij}] 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 (Elhorstet al., 2006; Ertur and Koch, 2007).

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

$$y_{it} = \rho\sum_{j=1}^{N} w_{ij} y_{jt} + x_{it}\beta + \sum_{j=1}^{N} w_{ij} x_{ijt}\theta + \mu_i + \varepsilon_{it} \quad .....(6)$$

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

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

$$Y = X\beta + u \quad \text{.....(7)}$$

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.

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

$$Y = \rho WY + X\beta + u \quad \text{.....(8)}$$

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 x1 regression coefficient vector β and u = vector of the error term.

IV. RESULTS AND DISCUSSION

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.

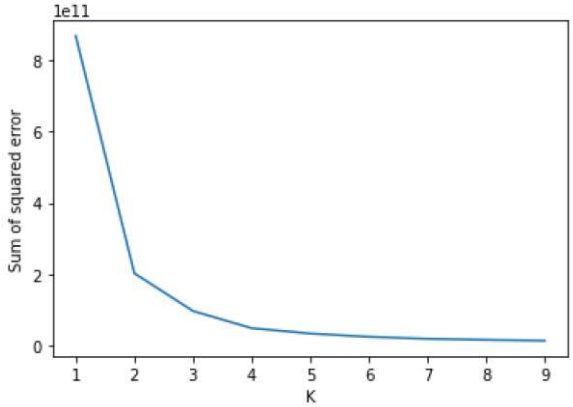

Figure 3: Elbow Technique

The value of K is established based on the Sum of Squared Error (SSE), which tends to exhibit an "elbow effect" in the graphical representation. In the provided figure, an elbow point around 3 to 4 indicates a suitable choice for the number of clusters. SSE is commonly employed as a benchmark in research to identify the optimal cluster count.

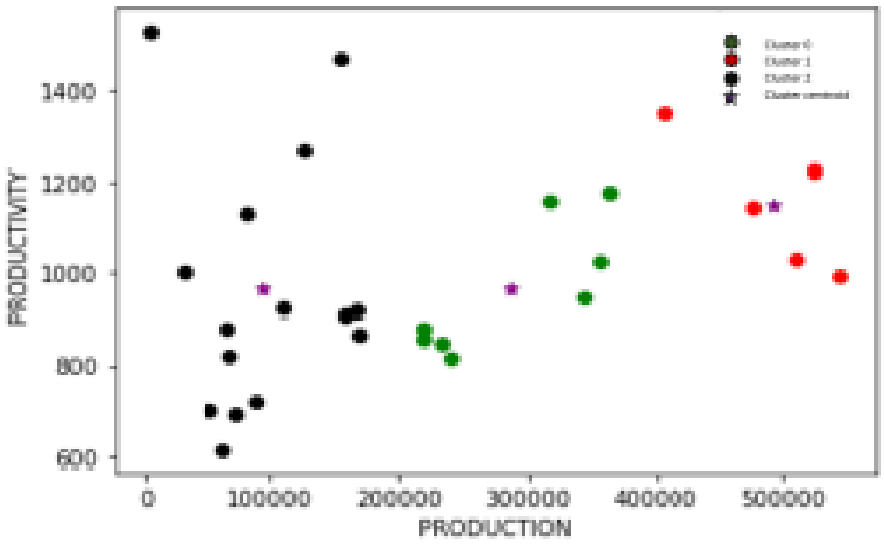

Figure 4: Scatter plot of Production vs Productivity

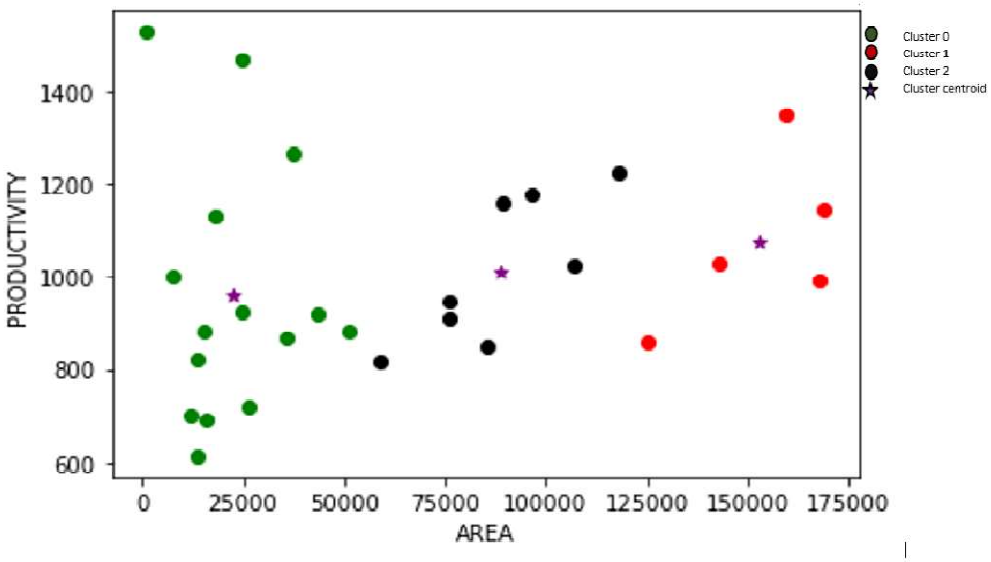

Figure 5: Scatter Plot of Area vs Productivity

Fig 4 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 5, the cluster 0 has given the major productivity which gained the high cluster. The area which has low productivity according to cluster 2.

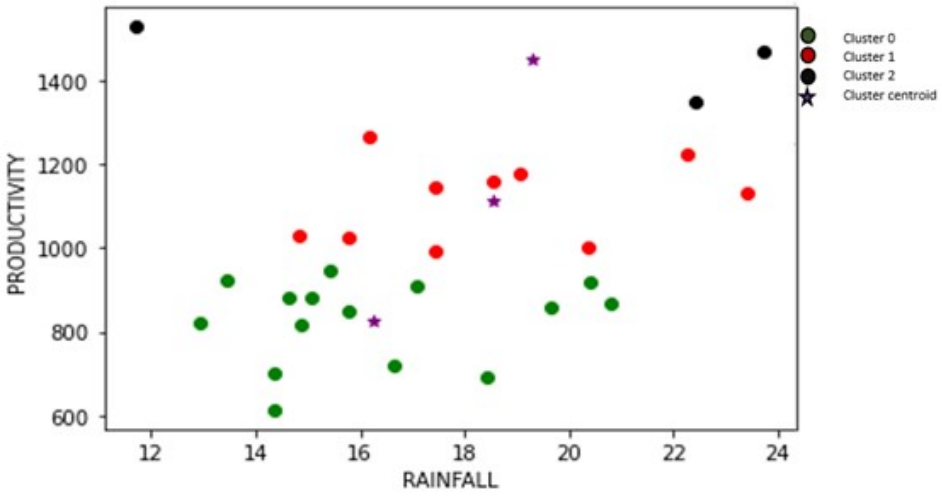

Figure 6: Scatter Plot of Rainfall vs Productivity

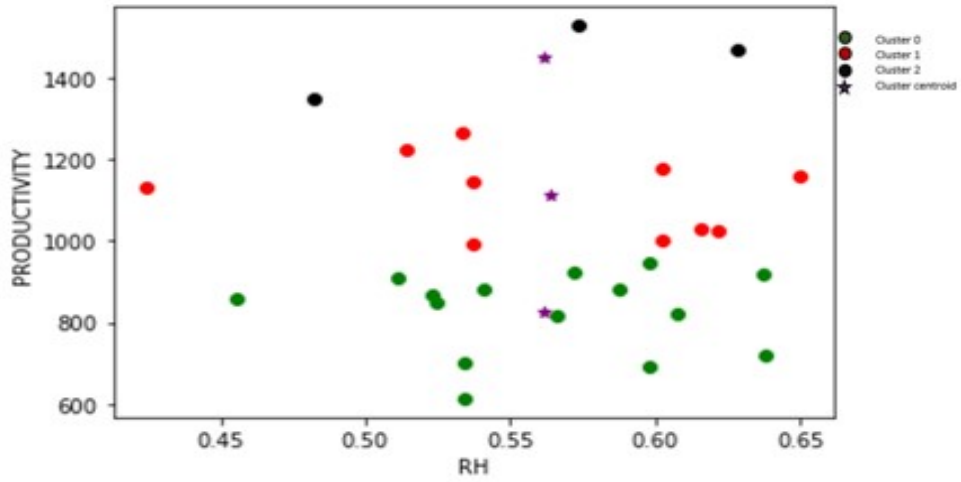

Figure 7: Scatter Plot of RH vs Productivity

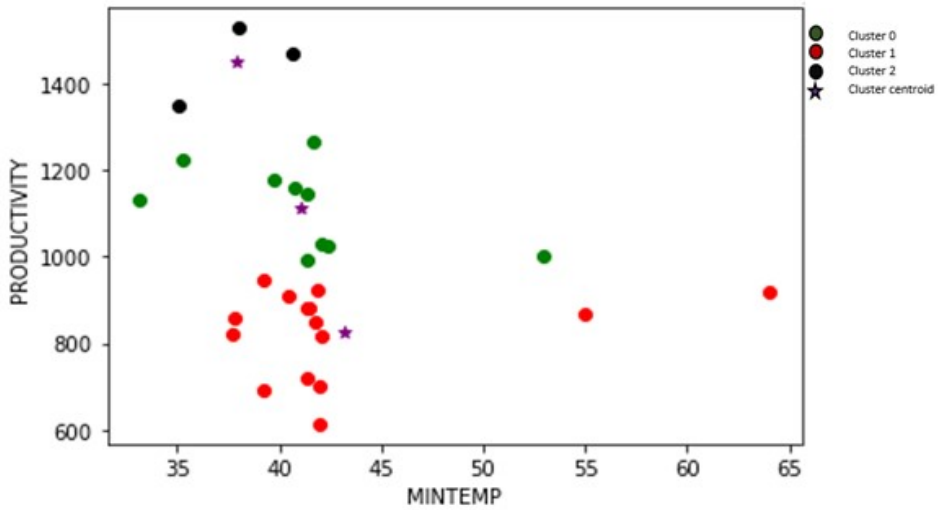

Figure 8: Scatter Plot of Min Temp vs Productivity

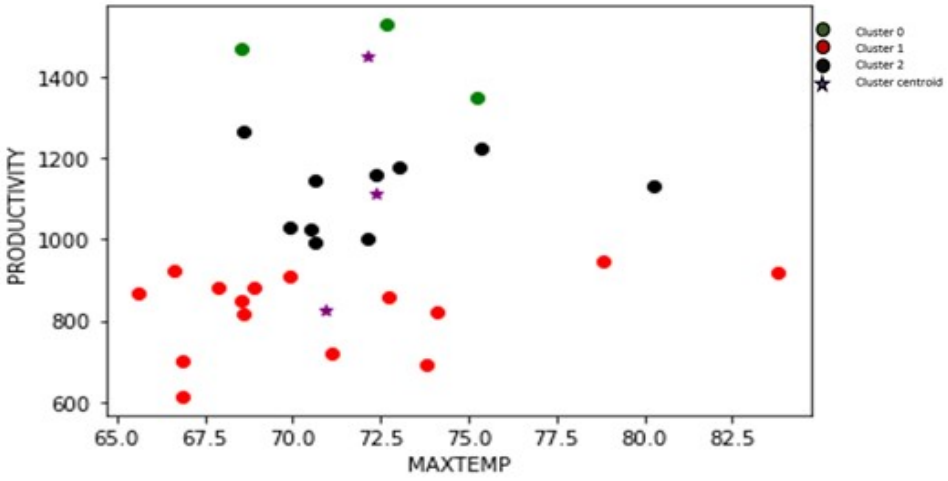

Figure 9: Scatter Plot of Max Temp vs Productivity

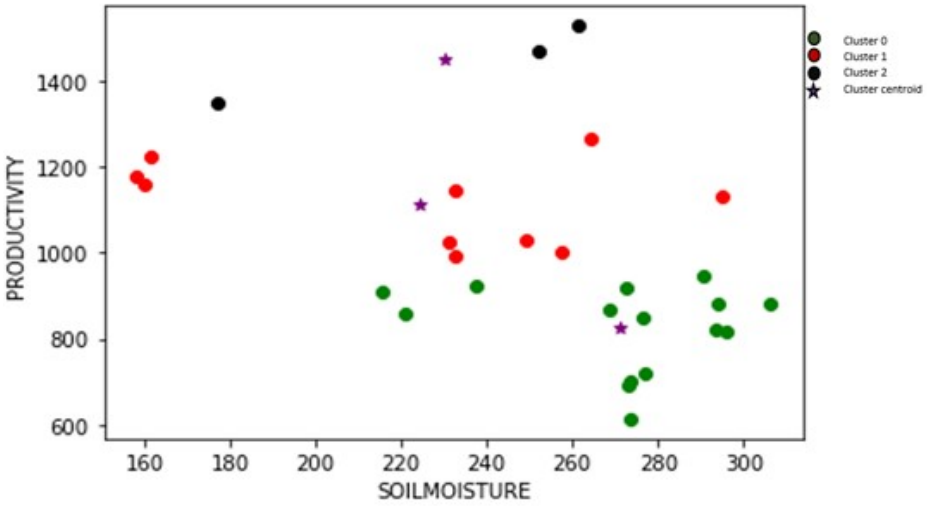

Figure 10: Scatter Plot of Soil Moisture and Productivity

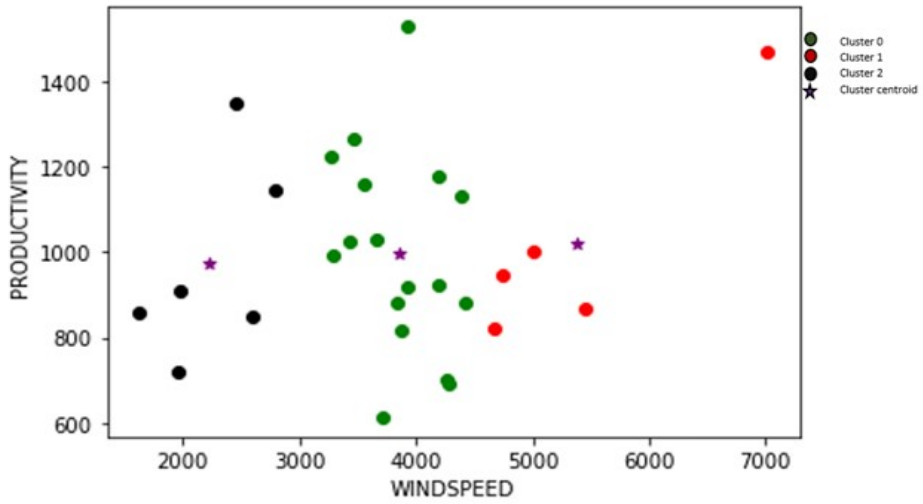

Figure 11: Scatter Plot of Wind Speed and Productivity

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.

2. **Apriori Algorithm:** Employing apriori rules on a dataset encompassing 11 variables led to the generation of 432 rules. However, a subset of only 31 rules was chosen based on specific criteria: minimum support and the highest counts within pairwise combinations. The Support and coverage values ranged from 0.214 to 0.321, while maintaining a 1% confidence level. The counts for the best combinations varied from 9 to 6, exhibiting a lift ranging between 2.800 and 3.111. Thus, filtering based on support not only reduced the number of rules from 7412 to 432 but also enhanced the trustworthiness. Further reduction is possible, allowing for a more streamlined rule set.**4.3Spatial 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 R^2 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
R^2	0.8287	**0.8394**
Adjusted R^2	0.8176	0.8179
AIC	984.13	983.85
LL	-334.09	**-332.20**
RMSE	0.2752	**0.2733**

* - Significance at 10% level
** - Significance at 5% level
*** - Significance at 1% level
(NS) – Not Significant

V. OUTCOME THE PROJECT

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.

VI. REFERENCES

[1] 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.
[2] Anselin, Luc, 2003. 'Spatial externalities, spatial multipliers, and spatial econometrics' International regional science review 26 (2):153-166.
[3] Baltagi, Badi H, Seuck Heun 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.
[4] 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.
[5] 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.
[6] Celik, A. 2020. 'Using Apriori Data Mining Method in COVID-19 Diagnosis' Journal of Engineering Technology and Applied Sciences 5 (3):121-131.
[7] Chakraborty, S., N. Nagwani, and L. Dey. 2012. 'Weather Forecasting using Incremental K-Means Clustering' International Journal of Biometrics and Bioinformatics. 4 (3).
[8] 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).
[9] 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.
[10] Gangai Selvi. R and Mani. K (2015), Ph.D thesis titled 'Land Use Dynamics in Tamil Nadu – A Spatial Econometric Analysis' TNAU, Coimbatore.
[11] 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.
[12] Mirmozaffari, M., A. Alinezhad, and A. Gilanpour. 2017. 'Data Mining Apriori Algorithm for Heart Disease Prediction' Int'l Journal of Computing, Communications & Instrumentation Engg4 (1):20-23.
[13] 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.
[14] 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.
[15] 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.
[16] 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).

