MILESTONE RESEARCH FOUNDATION
Int. Jrl. of Computational Learning & Intelligence
Vol. 02, Issue No. 01, 2023

ORIGINAL RESEARCH

Bayesian Networks for Improved Estimation of Paddy Crop Production

Kadamba Pavani . Halavath Balaji . N Ch S N Iyengar

Sreenidhi Institute of Science and Technology
Yamnampet,Ghatkesar, Hyderabad (T.G), India

Received: 03 November 2022 / Revised: 26 November 2022 / Accepted: 06 December 2022

Abstract – In India's food security, Paddy crop production has a prominent role by dispensing more than 40% of production of crop. Depends on climatic situation, paddy crop gives better production. Changes in seasonal climatic situations like low temperature or rainfall have negative impact of crop yielding. For improving the decision making capacity of farmers and stakeholders by considering the agronomy and crop choice, effective methods are developing for predicting the crop productivity under different climatic situation. The aim of this paper is estimating the paddy crop yield of Anantapur district, India. Anantapur was selected for this report accordingly considering the information existing in the Indian administration chronicles with different atmospheric and production predications like area production, precipitation, rainfall, minimal temperature, intermediate temperature, maximal temperature, evapotranspiration of crop and production from 2007 to 2012 of the Kharif season which is from June to November are selected. The dataset utilizing was processed using alled the tool WEKA. Clustering is performed by using k-means cluster. Classifiers named naïve bayes and bayesnet are used in this report. Proposed methodology gives better performance using bayesnet instead of naivebayes classifier for the dataset.

Index Terms – Agriculture; Bayesian networks; clustering; classifiers; yield estimation; data mining

I. INTRODUCTION

Agriculture is treated as a backbone of Indian economy . India's economy depends on growth of agriculture yielding. Crop yield prediction is one of the big problem in agriculture. Most farmers are lack in getting crop yield they expected due to different reasons. Vast area of farming are not fulfilling satisfiable trim building due to climatic and economic challenges. Any deviation of monsoon brings large fluctuations in area and production. Multiple ways are available for improving and increasing the crop production and quality. Data mining also used for estimation of crop production and is a novel research field in terms of analysis of crop yield. Data mining is for extraction of significant information from the dataset and convert the information into understandable format for the future use.

Data mining is used for analyzing large datasets for establishing useful patterns and classifications. In data mining clustering and classification techniques gives a good prediction results. Bayesian Networks (BN) establish a great relevance in limitations which need discriminate reasoning for contingent dependencies. Those strategies are useful for assessing model structure and the factors uncertainty. Instead of deterministic comparisons, probabilistic factors are used by the Bayesian Networks (BN) for defining variable connections.

II. LITERATURE REVIEW

Bayesian Networks (BN) is a most popular technique for defining and modeling of undefined and typical areas. Many research works are carried out on benefits and challenges using BN for complex issues like

modeling environment. Using graphical model, encoding of many probabilistic relations between variables are carried out in BN. The model gain more benefits for data modeling when it is used by combining with statistical model. Some of the benefits noticed are; data entries missing while encoding dependencies among the variables are handled and helps in estimating the consequences of intervention.BN is used for future yield functions prediction.

Some other research is carried using BN to predict the occurrence of crop diseases. The study of developing probabilistic framework for drought forecast and results is reported. Another investigation is on various economic and agricultural forecast development analysis using BN modeling algorithms. Bayesian Network can also used for weed infestation risk modeling and reduction of usage of pesticides. The objective behind the research paper is to predict paddy crop yield in Anantapur district of Andhra Pradesh state, India. The important points need to notify in this research:

- Recognize whether the Bayesian Network(BN) potently estimating the paddy crop production for Anantapur district of Andhra Pradesh state.
- Examine performance of Bayesian network with different values.
- notice the effeciency of outcome

III. RESEARCH METHODOLOGY

In this section, research states about study area, methodologies used and datasets used.

Area of study

The study area for this report is carried on selected mandals of Anantapur district in Andhra Pradesh. Anantapur district comes under the region of Rayalaseema of Andhra Pradesh and the headquarters of district is stationed at the city, Anantapur having the population rate of 4,083,315 in which the urban population is 28.09% and the rate of literacy is 64.28% according to 2011 census. In terms of area, it is a largest district in Andhra Pradesh state with 19,130 square kilometers and is a seventh largest district in India. The north side of Anantapur district is bounded by Kurnool district, east side by Kadapa district, Chittoor district on southeast, on the west and southwest by Karnataka state.

The central and northern sides are high plateau having surface covered with low hill ranges or large granite rocks. The surface of southern side of district is most hilly, the ranging of plateau is upto 2,000 ft (610 m). Papagni, Penna, Thadakaleru, Swarnamukhi and Chithravathi are the five rivers flow in the district. The average annual rainfall of the district is 381 millimeters. The major agricultural crops in Anantapur district are Paddy and Groundnut. The Indian government stated that Anantapur is one of the country's 250 backward districts in the year 2006. And is the one receiving Background Regions Grant Fund(BRGF) among 13 districts.

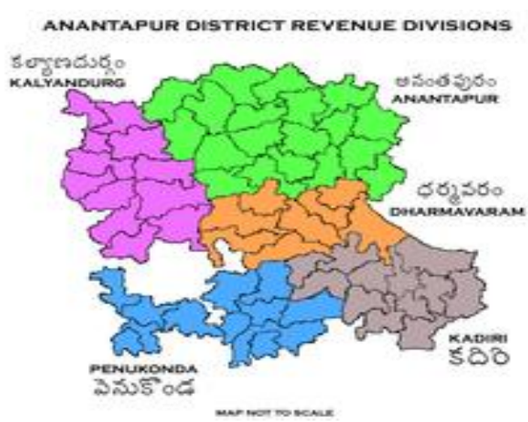

Fig.1. Anantapur district Revenue division

As the Anantapur is located in rain shadow area of peninsula of India, it receives very less amount of rainfall. Anantapur district is divided into 5 revenue divisions namely Anantapur, kadiri, Dharmavaram, Penukonda, Kalyandurg divisions which are further divided into 63 mandals.

Dataset Used

All the used datasets in this study were taken from the freely accessible chronicles of Government of India . The datasets of Paddy crop production are available from 2007 to 2012 for the Kharif season. From the complete dataset, only a needed prominent parameters having more impact on agricultural production were taken in the current study.

- Rainfall / Precipitation (mm): Calculation of total rainfall of every mandal in the district for each year of Kharif season(June to November) on the account of monthly mean precipitation of that particular year.
- Minimal, normal, maximal temperature(in degree Celsius): Temperature variations will have more impact on the yielding of crop. That is the reason why maximal , average and minimal temperature in every mandal of the district for each year was taken in to consideration in current study. From monthly mean temperature of maximal, average and minimal temperatures for every mandal of the district for that particular year, average temperatures of Kharif season (June-November) was calculated.
- Evapotranspiration of crop(mm) : It was computed by considering the monthly mean of Kharif season (from June to November) of each and every mandal in the district for that particular year.
- Area (Hectares): The cultivated zone of paddy in Kharif season (June to November) in each mandal selected of the district for year is taken into account in this study.
- Production (Tonnes): The production of paddy for the Kharif season (June to November) for area cultivated is to be designed for Anantapur district is considered for new analysis.

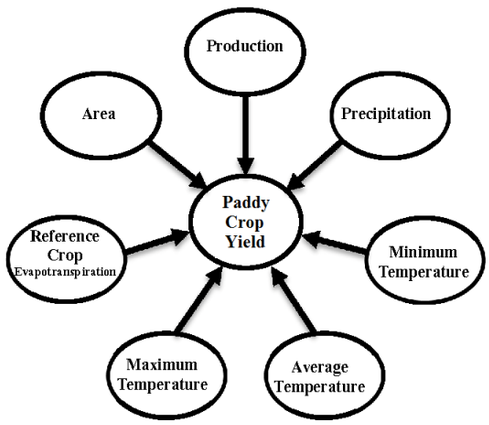

Fig.2. Study of climatic parameters on Anantapur district, Andhra Pradesh

Used Methodology

The various stages of this research paper include:

(1) preprocessing of dataset
(2) Establishing the estimation replica utilizing WEKA
(3) Examine the consequences

Preprocessing of dataset

Preprocessing is performed for proper arrangement of datasets in Microsoft Office Excel (also known as Ms-Excel) to control data inconsistency and handling the missing values. It contains fields serial number, name, mandal, rate of precipitation, low temperature, normal temperature, high temperature, evapotranspiration of crop, area and production. Sorting of records/files take place based on the cultivation area and unused fields was deleted and file was saved with .csv format.

Establishing the estimation replica utilizing WEKA

WEKA (abbreviated as the Waikato Environment for Knowledge Analysis*)* is a data mining tool which open / unpaid source i.e.; freely available software under General public License GNU developed at the university named Waikato, in New Zealand. WEKA consists of group of data mining algorithms which are useful in solving real-time data mining problems. Workbench of WEKA tool is a group of data analysis algorithms, predictive techniques; graphical user interfaces visualization and feature extraction tools.

WEKA is written in Java language so it is platform independent and the algorithms are called from one's own code of java or implemented directly on suitable dataset. Modeling algorithms used in programming languages other than java and data preprocessing utilities in C language use the actual non-Java related WEKA version is a TCL/TK front-end is designed/made tool for data analyzing in the domain of agriculture, but Weka 3, the latest version which was fully Java-based whose development started in the year 1997 is used for different areas of application now particularly for educational and research fields.

Following are the steps designed for data mining in WEKA:

- Preprocess the data and visualization
- Selecting attributes
- Clustering (Cobweb, K-means)
- Classification (j48 and other decision trees)
- Prediction (Nearest neighbor, Bayesian Network)
- Evaluating model

For the current research, a data set on Paddy particulars of Anantapur district across mandals and across 5 years (2007-2012) was considered for analysis with simple k-means clustering having 3 attributes named Mandal, year and paddy. It has 315 Instances which are clustered accordingly and produce results. After applying K-means algorithm the result is subjected to the J48 algorithm on the clustered data sets and the result confusion matrix is generated and BayesNet algorithm in the Bayes submenu was executed by selecting parameters and give effective outcome.

Examine the consequences

Rotation estimation also known as Cross validation, is used for analysis of performance of the data mining model predicted. For the dataset used in current study, cross validation of 10-folds was given for training of data and testing of data. Randomly the data was divide into 10 parts where 9 parts are utilized for training the data and 1 part for testing the data by repeating the process for 10 times and the outcomes was evaluated by considering performance metrics.

IV. PERFORMANCE EVALUATION

Classifiers performance was identified by the Precision , recall and accuracy values. Each instance is classified as two classes namely true and false which in turn gives four classifications as follows:

True Positive (TP): Perfectly classified that a sample is positive.

False Positive (FP): Imperfectly classified that a sample is positive.

False Negative (FN): Imperfectly classified that a sample is negative.

True Negative (TN): Perfectly classified that a sample is negative.

These values can be depicted as a confusion matrix also called contingency table as shown below:

TABLE 1 CONFUSION MATRIX

	True	**False**
True	True Positive (TP)	False Positive (FP)
False	False Negative (FN)	True Negative (TN)

Major diagonal of the contingency table are perfect classifications known as true positives and true negatives. The fields remaining are known as errors will assigned as zero. The confusion matrix gives performance metrics , precision ,recall and F-Measure were computed as:

Precision Value : It gives number of true positive predictions to the total number of positive predictions.

$$\text{Precision / Specificity} = TP / (TP+FP)$$

Recall Value: Recall gives percentage of total results which are relevant and classified correctly by algorithm.

$$\text{Recall / Sensitivity} = TP / (TP+FN)$$

F-Measure Value: The F-Measure (F) defines accuracy of test and it take values of specificity and recall for evaluation. F-Measure is nothing but the harmonic of the precision and sensitivity.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

F-measure can be calculated, using below equation

$$F1 = (2TP)/ (2TP + FP + FN)$$

The weighted values of the specificity and sensitivity gives the F-measure where it acquires its incentive very good least at 0. For effective evaluation, error rates MAE, RMSE, RAE, RRSE are computed.

Where, MAE - Mean Absolute Error
RMSE - Root Mean Squared Error
RAE - Relative Absolute error
RRSE - Root Relative Squared Error

V. RESULTS

The results acquired after running clustering and classification strategies on the data set of paddy crop production in Anantapur district of Andhra Pradesh using WEKA tool are shown.

a) Outcomes of proposed methodology on Paddy particulars 2007-08 data set is given as:

Precision = 0.938, Recall = 0.968, F-measure = 0.953

TPR (True Positive Rate) = 0.968

FPR (False Positive Rate) = 0.016

b*) Outcomes of proposed methodology on Paddy particulars 2008-09 data set is given as:*

Precision = 0.968, Recall = 0.952, F-measure = 0.96

TPR (True Positive Rate) = 0.952

FPR (False Positive Rate) = 0.008

c) Outcomes of proposed methodology on Paddy particulars 2009-10 data set is given as:

Precision = 0.967, Recall = 0.937, F-measure = 0.952

TPR (True Positive Rate) = 0.937

FPR (False Positive Rate) = 0.008

d) Outcomes of proposed methodology on Paddy particulars 2010-11 data set is given as:

Precision = 0.967, Recall = 0.937, F-measure = 0.952

TPR (True Positive Rate) = 0.937

FPR (False Positive Rate) = 0.008

e) *Outcomes of proposed methodology on Paddy particulars 2011-12 data set is given as:*

Precision = 0.939, Recall = 0.984, F-measure = 0.961

TPR (True Positive Rate) = 0.984

FPR (False Positive Rate) = 0.016

The time to implement proposed model is 0.02 seconds. Instances correctly classified are 301(9505550%), incorrectly classified are 14(4.4444%), kappa statistic 0.9444 and it gives error rates, 0.0344 MAE, 0.1312 RMSE, 10.7523% RAE and 32.7907% RRSE.

VI. CONCLUSION

Paddy is nurtured all through the country of India which generates over 40% of overall production of food grains. This made paddy as one of the top prominent crop in Andhra Pradesh. After onset of monsoon in the month June of Kharif season, paddy is planted. Production of paddy is mostly depends on the rainfall rate of the year and reports shows that there is a frequent increase of paddy production every year. So that farmers have more interest and scope in paddy crop[planting and improving the production of crop. Distinct programmes and technologies are proposed and conducted by the Indian government which gradually increase the paddy production in the past few years. Farmers experienced some difficulties mentioned will reduced by using best estimation model which can estimate crop productivity under distinct Climatic conditions.

In the current study, a substitute technique used is Bayesian networks for estimation of paddy production. The BayesNet classifier on the clustered data with J48 classification technique produced high performance and better outcomes for paddy dataset of Anantapur district, Andhra Pradesh.

TABLE-2 CLUSTER CENTROIDS FOR PADDY PARTICULARS

Attribute	Mandal	Year	Paddy
Full Data (315)	Anantapur	2007-08	338.287
Cluster 0(61)	Roddam	2010-11	407.114
Cluster 1(66)	Beluguppa	2011-12	141.136
Cluster 2(61)	Vajrakarur	2009-10	287.312
Cluster 3(65)	Garladinne	2007-08	588.251
Cluster 4(62)	Anantapur	2008-09	268.532

TABLE-3 CONFUSION MATRIX

A	B	C	D	E	Year
61	0	1	0	1	2007-08
1	60	0	1	1	2008-09
1	1	59	1	1	2009-10
1	1	1	59	1	2010-11
1	0	0	0	62	2011-12

TABLE-4 PERFORMANCE METRIC-1 OF EVERY YEAR

YEAR OF CLASS	TP RATE (TPR)	FP RATE (FPR)
2007-2008	0.968	0.016
2008-2009	0.952	0.008
2009-2010	0.937	0.008
2010-2011	0.937	0.008
2011-2012	0.984	0.016

TABLE-5 PERFORMANCE METRIC-2 OF EVERY YEAR

YEAR OF CLASS	Precision/ specificity	Recall/ sensitivity	F-measure
2007-2008	0.938	0.968	0.953
2008-2009	0.968	0.952	0.96
2009-2010	0.967	0.937	0.952
2010-2011	0.967	0.937	0.952
2011-2012	0.939	0.984	0.961

TABLE-6 PERFORMANCE METRIC-3 OF PADDY PARTICULARS 2007-12 DATA SET

CLASS YEAR	MAE	RMSE	RAE	RRSE
2007-2001	0.0344	0.1312	10.7523%	32.7907%

Timely removal and organizing of problems affecting with production of crop will useful for farmers to make decisions in reducing the occurrence of the loss in bad situations. In other case, existence of good growing conditions, there is a drastic increase of productivity of crop. Developing the Decision Support Systems (DSS) for specific crop by considering efficient climatic conditions makes farmers to take the decisions very easily and

timely manner. One that kind of DSS was implemented on the dataset used in current study which provides estimation of paddy crop production, visualization of historic data and Geographic Information System (GIS). The current study also exhibit the utilization of the Bayesian Networks (BN) for improving the developed DSS for Estimation of crop production.

VII. FUTURE WORK

Now-a-days, diseases are very frequently affecting the agriculture in this world. We are not able to predict what kind of diseases are affecting for crops. The future study will develop on the basis which is helpful for prediction of disease affected by particular crop and also provide information of the pesticide to use for overcoming from diseases are suggested.

REFERENCE

1. Devika, B., & Ananthi, B. (2018). Analysis of crop yield prediction using data mining technique to predict annual yield of major crops. *International Research Journal of Engineering and Technology*, *5*(12), 1460-1465.
2. Scott, L. M., & Janikas, M. V. (2010). Spatial statistics in ArcGIS. In *Handbook of applied spatial analysis* (pp. 27-41). Springer, Berlin, Heidelberg.
3. Ramesh, D., & Vardhan, B. V. (2015). Analysis of crop yield prediction using data mining techniques. *International Journal of research in engineering and technology*, *4*(1), 47-473.
4. Manjula, E., & Djodiltachoumy, S. (2017). A model for prediction of crop yield. *International Journal of Computational Intelligence and Informatics*, *6*(4), 298-305.
5. LK, S. S., Ahmed, S. T., Anitha, K., & Pushpa, M. K. (2021, November). COVID-19 Outbreak Based Coronary Heart Diseases (CHD) Prediction Using SVM and Risk Factor Validation. In *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)* (pp. 1-5). IEEE.
6. Surya, P., & Aroquiaraj, I. L. (2018). Crop yield prediction in agriculture using data mining predictive analytic techniques. *International Journal of Research and Analytical Reviews*, *5*(4), 783-787.
7. Gibson, D., Kumar, R., & Tomkins, A. (2005, August). Discovering large dense subgraphs in massive graphs. In *Proceedings of the 31st international conference on Very large data bases* (pp. 721-732).
8. Kadamba Pavani, D., & Balaji, H. (2019). BAYESIAN NETWORKS FOR IMPROVED ESTIMATION OF PADDY CROP PRODUCTION.
9. Krishna, B. L., Lakshmi, P. J., & Prakash, P. S. (2012). Combination of Density Based and Partition Based Clustering Algorithm-DBK Means. *IJCSIT) International Journal of Computer Science and Information Technologies*, 4491.
10. Ahmed, S. T., Ashwini, S., Divya, C., Shetty, M., Anderi, P., & Singh, A. K. (2018). A hybrid and optimized resource scheduling technique using map reduce for larger instruction sets. *International Journal of Engineering & Technology*, *7*(2.33), 843-846.
11. Ahmed, S. T., & Basha, S. M. (2022). *Information and Communication Theory-Source Coding Techniques-Part II*. MileStone Research Publications.