IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Accurate Error Prediction of Sugarcane Yield Using a Regression Model

M. Mohanadevi[1*], V. Vinodhini[2]

[1*]Research Scholar, Dept. of Computer Science, Dr.N.G.P Arts and Science College, Coimbatore, Tamilnadu, India
[2]Associate Professor, Dept. of Information Technology, Dr.N.G.P Arts and Science College, Coimbatore, Tamilnadu, India

[*]*Corresponding Author: Monamoorthi04@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 13/Jul/2018, Published: 31/July/2018

***Abstract*—** In this paper an attempt has been made to review on application of data mining techniques in the field of agriculture. India is the largest producer and consumer of sugar in the world and its most efficient crops in converting solar energy into chemical energy. Sugar-cane is an important commercial crop of the world. About 45 million sugarcane farmers, their dependents and a large agricultural force, constituting 7.5 percent of the rural population, are involved in sugar-cane cultivation, harvesting and ancillary activities. Sugar industries development is backbone to economic development of the nation. In India, Sugar industry is the second largest agro-based industry and it contributes significantly to the socio economic development of the nation. The major Sugar-cane crop growing states in India are Uttar Pradesh, Bihar, Assam, Haryana, Gujarat, Maharashtra, Karnataka and Tamil Nadu. This paper presents state of Karnataka datasets to predict less error rate on better productivity and yield using regression metrics.

***Keywords*—** Agriculture data, Data mining Techniques, Weka tool,Regression Model

I. INTRODUCTION

Data mining can be viewed as a result of the natural evolution of information technology. An evolutionary path has been witnessed in the database industry in the development of the following functionalities: data collection, data management and data analysis. For instance, the early development of data collection and database creation mechanisms served as a prerequisite for later development of effective mechanisms for data storage and retrieval, and query and transaction processing. Become the next target systems opening query and transaction processing as common practice, data analysis and understanding has naturally. Data mining is the extraction of hidden predictive information from large databases. It's a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analysis offered by data mining move beyond the analysis of past events provided by retrospective tools typical of decision support systems.

Sugarcane is an important commercial crop of the world. Sugarcane cultivation is done in around 5 million ha of the land in India. Since sugarcane is annual crop and grows 12-18 months in the agriculture field in India. Hence this objective would provide the scientific information that what are the influences of various climatic factors in different growing time of sugarcane crop [1].

A process model for analysing data, and describes the support that Weka to Environment for Knowledge Analysis (WEKA) provides for this model [2]. The domain model learned by the data mining algorithm can then be readily incorporated into a software application. This WEKA based analysis and application construction process was illustrated through a case study in the agricultural domain.

This paper focus of study was to measure the growth in total factor productivity and Yield of the sugarcane crop in Karnataka state. The paper is organized as follows: Chapter 2 discusses the Literature Review. Chapter 3 discusses the Proposed Methodology 4 discusses the Results and Discussion. Chapter 5 discusses the conclusion.

II. RELATED WORK

Kumar et al. (2014) investigated the impact of climatic and non-climatic factors on productivity of major food grain crops in India using a Cobb-Douglas production at state level panel data in India. In this analysis, the authors

include average minimum temperature, average maximum temperature and actual rainfall as climatic factors in growing time of each crops (sowing time to harvesting time). Empirical result of the study reveals that productivity of wheat, barley, gram and rice crops are declined due to increase in actual average minimum temperature. The productivity of rice, maize, sorghum, and ragi crops are lead to decrease with increase in actual average maximum temperature in growing time of corresponding crops [3].

Sellam, *et al* explained various environmental parameters like Area under Cultivation (AUC), Annual Rainfall (AR) and Food Price Index (FPI) that influences the yield of crop and the relationship among these parameters was established. Using Regression Analysis (RA), Linear Regression (LR) the various environmental factors and their infliction on crop yield was analysed [4].

Sujatha, *et al* described about the purpose of various classification techniques that could be utilized for crop yield prediction. A few of the data mining methods, such as the Naïve Bayes, J48, random forests, SVM, artificial neural networks were presented. A system using climate data and crop parameters used to predict crop growth has been proposed [5].

Ankalaki, *et al* presented a comparative study on DBSCAN and AGNES algorithm for clustering. Crop yield was forecasted using MLR (Multiple Linear Regression) and a formula was derived for each crops. From the proposed work, we can conclude that DBSCAN was more time consuming than the optimal and efficient number of clusters. Regression analysis performed for the forecasting that showed a highly dependency on the dataset. Proper data collection will make the model significant, otherwise it can lead to inaccurate results [6].

Kushwaha, *et al.* Predicted the suitability of a crop for a particular climatic condition and the possibilities of improving the crops quality by using weather and disease related data sets. They have proposed an analysis, classification and prediction algorithm that helps in building a decision support system for precision farming. It was based on the Hadoop file system [7].

Rub, *et al.* Presented a comparative study on the regression models that could be used for predicting yield. The algorithms discussed were Multilayer perception Model (MLP), Reg-tree (Regression tree), RBF (Radial Basis Function Network and SVM. They have concluded that SVM serves as a better model as far as yield prediction was concerned [8].

Raorane, *et al.* discussed about the various data mining techniques for improving the crop production in agriculture. A few of Data mining methods, such as ANN, Decision Tree algorithm, Regression Tree, Bayesian network, SVM, k means were used for classification [9].

A. Nagarajan (2013) was conducted a study on "sustainable farming practices in sugarcane cultivation". The objective was to minimize the cost of production and maximize the productivity without affecting the environment and certain steps need to be wakened for sugarcane cultivation such as land preparation practice, planting sets practice, water management practice, inter cropping management practice, ratoon management practice, harvesting management practice and from these practices they concluded that it was a great help to evaluate the adoption of different sustainable sugarcane farming practices [10].

Nazir et al. (2013) found that the costs of inputs of sugarcane i.e. urea, DAP, FYM, land preparation, seed and its application, weeding and cost of irrigation were the important factors which influenced on the returns of sugarcane growers. The effectiveness was examined by using the Cobb-Douglas production function, MVP and allocative efficiency were also calculated. They also found that the high prices of inputs, low price of output, delay in payments and lack of scientific knowledge were the major problems in sugarcane production. In order to enhance the productivity of sugarcane in the country, government should solve the identified problems to increase the income of sugarcane growers [11].

The cost and returns analysis was used to assess the profitability, whilst multiple linear regression analysis was used in identifying the determinants of profitability [12].

Gupta et al. (2012) analyzed the climatic impact on crop productivity of rice, sorghum and millet at macro level. The authors included average temperature and actual rainfall in growing time of these crops. The empirical findings of this study showed that climate change is likely to reduce the yields of rice, sorghum and millet crop in 16 major agriculture intensive states of India [13].

Srivastava & Rai (2012) also mentioned in their review article that there is need a research to identify the climatic effect on cane productivity in India [14].

Kumar et al. (2011) concluded that climate change has shifted the weather condition which affected the seasonal crops and reduced the available growing time of rice and sugarcane crops in Uttarakhand and Uttar Pradesh (India) [15].

Geethalakshmi et al. (2011) also showed that the productivity of rice has declined by 41% with 40C increase in temperature in Tamil Nadu (India) [16].

Masters et al. (2010) mentioned that climate change has a significant negative effect on agriculture production that occupies around 40% of the land globally [17].

Kalra et al. (2008) undertook a state wise analysis for four states of India, namely Punjab, Haryana, Rajasthan and Uttar Pradesh. The study also concluded that wheat, mustard, barley and chickpea production has decreased due increase in seasonal temperature [18].

Study by Kapur et al. (2009) mentioned that rainfall may decrease crops yields by 30% by the mid-21st century. This study also justified that there would be

reduction in arable land that could be results in more pressures on agriculture production In India [19].

Dr.S.D.Sundar singh and R. Veeraputhiranhas had conducted a study on "irrigation management in sugarcane" (2000) and concluded that Tamil Nadu was the leading producer of sugarcane was compared to other states. But, the scarcity of water was a limiting factor. Water was vital in certain stages of growth of sugarcane. Irrigation water was essential yet a constraint in sugarcane production, efficient supply of water, considering the soil, climate, crop, environment conditions was important. The various strategies include selection of varieties, mulching, and gradual widening of furrows, alternate furrow method of irrigation, drip irrigation, and an innovative method called surge irrigation. The authors stressed in the fact that an optimum soil moisture environment was a pre-requisite to reduce the adverse of shoot borer in sugarcane [20].

III. METHODOLOGY

A. Block Diagram

The outline of the work illustrated by using a figure in this section. Figure 1: describe the overview of the proposed block diagram. At the beginning stage, this large data set is carried out into pre-processing and is called as Data pre-processing. In the next stage, the model are generated using Machine learning algorithm. In the final stage, validating the model is done by comparing the result of various algorithm.

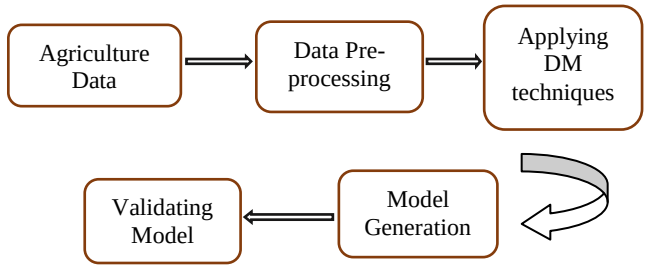

Figure 1. Block diagram

B. Data Description

The present study focus on past 15 years data from the period 2001 to 2015.The data was obtained from the database of Department of Agricultural Cooperation and Farmers Welfare, Government of India, Directorate of Economic and Statistic in state of Karnataka State. This raw data set is then pre-processed and analysed by using Weka tool. In this work uses, 29 districts data are considered for prototype analysis of the tools. The given Agriculture data base has different attributes like District Name, Year, Crop type, Area, Production and Yield class with their respective values .The dataset used by us contains 6 attributes and 332 instances for Agriculture data to measure the regression error metrics. We have applied different algorithms using WEKA data mining tool for our analysis purpose.

C. Explorer Interface

It first pre-processes the data and then filters the data. Users can then load the data file in CSV (Comma Separated Value) format and then analyse the Regression error metrics result by selecting the following algorithms using 10 cross validation: Linear Regression, REP Tree, and Random Tree. In Weka we measured the class value is numeric, the correlation coefficient is given. An Evaluating a machine learning algorithm on a regression problem and a number of different performance measures to review. Of note the performance summary for regression algorithms are two things:

- **Correlation Coefficient**. This is how well the predictions are correlated or change with the actual output value. A value of 0 is the worst and a value of 1 is a perfectly correlated set of predictions.

- **Root Mean Squared Error**. This is the average amount of error made on the test set in the units of the output variable. This measure helps you get an idea on the amount a given prediction may be wrong on average.

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

- **Mean Absolute Error:** This scale the error to the mean.
- **Root relative squared Error**: The regression line predicts the average y value associated with a given x value. Note that is also necessary to get a measure of the spread of the y values around that average. To do this, we use the root-mean-square error.

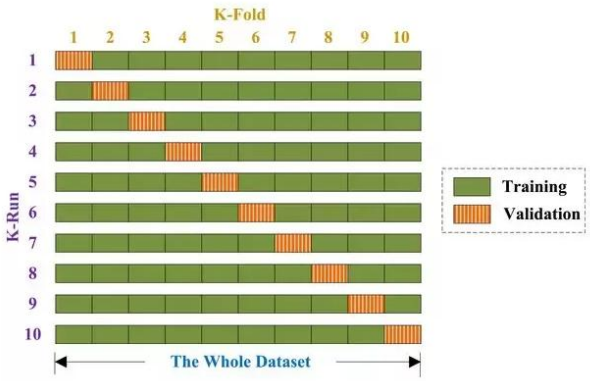

Figure 2: Basic Diagram of 10-fold Cross Validation

D. Random Forest

The basic idea is that you divide your training dataset into *k* subsets. You then train the Random Forest on *k - 1* subsets, setting aside the other subset. Test the model on

the remaining subset. Then repeat that process *k* times, such that by the end, each of the subsets has acted as the 'test set' once. This can be achieved through a simple for loop. You can use any number of subsets (k), though I would personally recommend 10 as a good measure. So as you can see, the dataset is split into 10 parts. The model is then trained-and-tested 10 times, each time on a different 9 parts, and tested on the tenth. By the end you can aggregate all of your test results into a dataset, and compare the predictions with the true values. This works for both classification and regression.

```
RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.3 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.7856
Mean absolute error                      8.9651
Root mean squared error                 12.4506
Relative absolute error                 64.6729 %
Root relative squared error             67.1612 %
Total Number of Instances              332
Ignored Class Unknown Instances                1
```

Figure 3: Screenshot view of Random Forest Output

E. REP tree

The basic idea is that you divide your training dataset into *k* subsets. You then train the REP tree on *k - 1* subsets, setting aside the other subset. Test the model on the remaining subset. Then repeat that process *k* times, such that by the end, each of the subsets has acted as the 'test set' once. This can be achieved through a simple for loop. You can use any number of subsets (k), though I would personally recommend 10 as a good measure. So as you can see, the dataset is split into 10 parts. The model is then trained-and-tested 10 times, each time on a different 9 parts, and tested on the tenth. By the end you can aggregate all of your test results into a dataset, and compare the predictions with the true values. This works for both classification and regression.

```
Time taken to build model: 0.02 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.7107
Mean absolute error                      9.5964
Root mean squared error                 13.0577
Relative absolute error                 69.2267 %
Root relative squared error             70.436  %
Total Number of Instances              332
Ignored Class Unknown Instances                1
```

Figure 4: Screenshot view of REP tree Output

F. Linear Regression

The basic idea is that you divide your training dataset into *k* subsets. You then train the linear regression on *k - 1* subsets, setting aside the other subset. Test the model on the remaining subset. Then repeat that process *k* times, such that by the end, each of the subsets has acted as the 'test set' once. This can be achieved through a simple for loop. You can use any number of subsets (k), though I would personally recommend 10 as a good measure. So as you can see, the dataset is split into 10 parts. The model is then trained-and-tested 10 times, each time on a different 9 parts, and tested on the tenth. By the end you can aggregate all of your test results into a dataset, and compare the predictions with the true values. This works for both classification and regression.

```
Time taken to build model: 0.39 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8009
Mean absolute error                      7.7852
Root mean squared error                 11.092
Relative absolute error                 56.161  %
Root relative squared error             59.8325 %
Total Number of Instances              332
Ignored Class Unknown Instances                1
```

Figure 5: Screenshot view of Linear Regression Output

IV. EXPERIMENTS AND DISCUSSION

An Explorer the data mining techniques that have been used by us using different algorithms Linear Regression, Random Forest and REP tree. Through these techniques we trained out results on the basis of time taken to build model, Correlation coefficient, Mean absolute error, Root Mean Squared Error, Relative Absolute Error and Root Relative Squared Error and Algorithm scoring Error prediction is shown in Table 1 and Figure 6.

Linear Regression classified Best Correlation Coefficient rate is 0.8009 compare with Reptree and Random forest algorithm and this algorithm achieved Minimum Error rate on Mean Absolute Error = 7.7852, and time taken to build model=0.39 seconds. So from Explorer Interface data mining technique we can deduce that Linear Regression have predict least error rate and it takes less time to build model it.

Algorithm	Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error
REP tree	0.71074	9.5964	13.057	69.226
Linear Regression	**0.8009**	**7.7852**	**11.092**	**56.161**
Random Forest	0.7856	8.9651	12.45	64.672

Table 1: Explorer Interface results

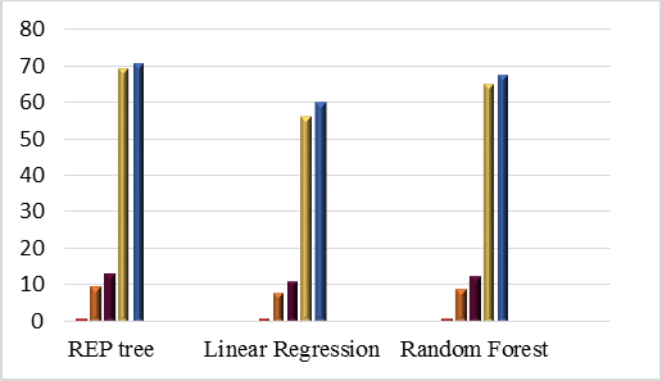

Figure 6: Error rate Analysis

V. CONCLUSION

Agriculture is the most significant application area particularly in the developing countries like India. In 2010-11 Karnataka state was best annual growth in production. Use of information technology in agriculture can hang the scenario of decision making and farmers can yield in better way. There is a growing number of applications of data mining techniques in agriculture and also a growing amount of data that are currently available from many resources in sugar industry to increase profitability we should reduce the cost of cultivation and improving the productivity per unit. It is possible through new research innovations, technological interventions and mechanization [21]. It would be effective, if an efficient research and development effort on sugarcane is taken. So, the government should initiate to improve the sugar cultivation where cane yield and sugar need to be improved substantially. In this paper we have discussed about the role of data mining in outlook of agriculture field. This is relatively a novel research field and it is expected to grow in the future. There is a lot of work to be done on this emerging and interesting research field. The multidisciplinary approach of integrating computer science with agriculture will help in forecasting/managing agricultural crops effectively. In this study made less error rate prediction using Linear Regression Algorithm and its classified Best Correlation Coefficient rate is 0.8009 compare with Reptree and Random forest algorithm and this algorithm achieved Minimum Error rate on Mean Absolute Error = 7.7852, and time taken to build model=0.39 seconds.

REFERENCES

[1] Ajaykumar ,pritee sharma*,"Climate change and sugarcane Productivity in india : An Econometric Analysis",*Journal of social and development sciences vol5.No.2 pp-111-122,Jun 2015 .

[2] Camps-Valls G, Gomez-Chova L, Calpe-Maravilla J, Soria-Olivas E, Martin-Guerrero JD, Moreno J., 2003, *"Support vector machines for crop classification using hyperspectral data"*, Lect Notes Comp Sci 2652: pp. 134–141.

[3] Kumar, A., Sharma, P. & Ambrammal, S. K. (2014). "*Climatic Effects on Food Grain Productivity in India: A Crop Wise Analysis",* Journal of Studies in Dynamics and Change, 1(1), 38-48.

[4] Sellam,V, Poovammal, E*., "Prediction of Crop Yield using Regression Analysis",* Indian Journal of Science and Technology, Vol. 9, issue.38, pp.1- 5, 2016.

[5] Sujatha, R., Isakki, P., *"A study on crop yield forecasting using classification techniques",* International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE), pp.1-4, 2016.

[6] Ankalaki, S., Chandra, N., Majumdar, J*., "Applying Data Mining Approach and Regression Model to Forecast Annual Yield of Major Crops in Different District of Karnataka",* International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, issue 2, pp.25-29, 2016.

[7] Kushwaha, A.K., Sweta Bhattachrya, *"Crop yield prediction using Agro Algorithm in Hadoop",* International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 5, issue.2, pp.271-274, 2015.

[8] Rub, G., *"Data Mining of Agricultural Yield Data: A Comparison of Regression Models*", 9th Industrial Conference, Vol.5633, pp.24-37, 2009..

[9] Raorane, A.A., Kulkarni R.V., *"Data Mining: An effective tool for yield estimation in the agricultural sector*", International Journal of Emerging Trends & Technology in Computer Science(IJETTCS), Vol. 1, issue 2, pp.75-79, 2012.

[10] Nagarajan*,"Sustainable farming practices in sugarcane cultivation",* Kisan world, A journal of agriculture and Rural Development, Vol 40, Jan 2013, pp. 28 – 31.

[11] Nazir A,Jariko,G.A. and Junejo,M.A. (2013) *"Factor Affecting Sugarcane Production in Pakistan"*Munich Personal RePEc Archive.http://mpra.ub.Uni-menchen.de/50359/.

[12] Dlamini,S Rugambisa,J.I.Masuku,M.B.and Belete,A.(2010),"*Technical Efficiency of the small scale sugarcane faramers in Swaziland: A case study of Vuvulane and Big Bed Farmers",* African Journal of Agricultural Research,Vol.5(9).935-940.

[13] Gupta, S., Sen, P. & Srinivasan, S. (2012). *"Impact of Climate Change on Indian Economy: Evidence from Food Grain Yields",* Centre for Development Economics Working Paper 218, Delhi.

[14] Srivastava, A. K. & Rai, M. K. (2012). "*Sugarcane Production: Impacts of Climate Change and its Mitigation"* ,*Biodiversitas*, 13(4), 214-227.

[15] Kumar, V., Sharma, Y. & Chauhan, S. (2011a). *"Impact of Climate Change on the Growth and Production of Saccharum Offcinarum and Magnifera Indica"* International Journal of Science Technology and Management, 2(1), 42-47.

[16] Geethalakshmi, V., Lakshmanan, A., Rajalakshmi, D., Jagannathan, R., Sridhar, G., Ramara, Bhuvaneswari, A. P., Gurusamy, K. L. & Anbhazhagan, R. (2011*)"Climate Change Impact Assessment and Adaptation Strategies to Sustain Rice Production in Cauvery Basin of Tamil Nadu"*, Current Science, 101(03). 342-347.

[17] Masters, G., Baker, P. & Flood, J. (2010)*"Climate Change and Agricultural Commodities"* CABI Working Paper, 02.

[18] Kalra, N., Chakraborty, D., Sharma, A., Rai, J., Monica, H. K., Subhash, C., Kumar, P., Ramesh, B. S., Barman, D., Mittal, R. B., Lal, M. & Sehgal, M. (2008).*"Effect of Increasing Temperature on Yield of Some Winter Crops in Northwest India",* Current Science, 94(1), 82-88.

[19] Kapur, D., Khosla, R. & Mehta, P. B. (2009). "*Climate Change: India's Options",* Economic and Political Weekly, 36(31), 34-42.

[20] Sundar singh and Veeraputhiran,"*Enhancing sugarcane productivity",* Kisanworld, A journal of Agriculture and Rural Development', June 2000, pp. 18-19.

[21] D.Venkatesh,M.Venkateswars :"*An overview of the indian sugar industry",* BIMS International Journal of social science research ISSN 2455-4839.

Authors Profile

Dr.V.Vinodhini has completed her Master of degreee from Sri Krishna Arts and Science College Coimbatore and Mphil Computer Science degreee from Dr.N.G.P Arts and Science College Coimbatore and Ph.D Computer Science Degree from Karpagam University. She is currently working as Associate Professor in Department of Information Technology at Dr.N.G.P Arts and Science College Coimbatore .She has published more than 10 research papers in reputed international Journal and her main research work focuses on Data Mining and Warehousing. She has 10 years of teaching Experience.

Miss M.Mohanadevi has completed her Bachelor of Science from Dr.N.G.P Arts and Science College Coimbatore and her Master of Science degree completed from Bharathiar University Coimbatore. Currently she is doing MPhil in Computer Science at Dr.N.G.P Arts and Science College Coimbatore. She is doing Research in the Area of Data Mining and Warehousing.

