International Journal of Engineering and Management Research e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume-11, Issue-3 (June 2021)
www.ijemr.net https://doi.org/10.31033/ijemr.11.3.40

Novel Approach for Forecasting Sugarcane Crop Yield: A Real-Time Prediction

Ankit Kumar[1] and Anil Kumar Kapil[2]
[1]Faculty, Department of Mathematics and Computer Sciences, Motherhood University, Roorkee, INDIA
[2]Faculty, Department of Mathematics and Computer Sciences, Motherhood University, Roorkee, INDIA

[1]Corresponding Author: ankitsiet103@gmail.com

ABSTRACT

Agriculture is an environment in which there is considerable confusion. Crop development depends largely on several variables, including climate, temperature, genetics, politics and economics. In addition, a huge number of raw agriculture statistics are available, but study of those details for estimating crop yield is quite challenging. The most challenging job is therefore to include accurate details and awareness about the raw farm data. In order to evaluate cultivation yield, data mining could customize data expertise. The objective of this study was to predict crop yields through the use of data mining technological advances. Moreover, this paper compared various classification algorithms and it is expected that the results of the study may enhance the actual yields of sugarcane in a wide number of tropical fields. The specifications used in the forecast were plot (soil size, plant area, rain distance, previous year's plant yield), sugar-cane characteristics (cane class and sort), crop cultivation procedure (normal water resource size, cultivation technique, disease management process, sort / procedure of fertilizer) as well as rain quantity.

Keywords-- Sugarcane Yield, Crop Yield Prediction, Weather, Soil Data, Decision Tree

I. INTRODUCTION

Indian agriculture is known in different cultivation forms primarily because of resource shifts. India is a rural agriculture nation, and the forecasts are vital study areas that ensure food security worldwide as well as the Indian financial sector. India is a nation of the rural [10]. Indeed, India is a world farming force. Per year from the previous year, we get a large yield. The effect of various crop variables, evaluation and predicting performance in consecutive years is a major factor in this growth [5]. Crop predictions and projections are of great significance for the implementation of new policy practices that encourage crop growth. The cultivar of the nation is filled with rain for around 63 percent, while irrigation for 37 percent.

For its operational planning, sugar mills must know the expected sugar cane yields for each harvesting season. Data from each plot was gathered for a output year by the field survey workers at the sugar mill. In traditional words, expertise at sugar mills would use their expertise to predict every plot's sugar cane production, based on the data from the survey, on every farmer's historical production profile. The key drawback of using this approach for calculating the human performance is that the projected production and the real performance varies considerably [11].

Consequently, appropriate forecasting of crop production for productive land use planning as well as fiscal growth is also important. Recently, agricultural crop yield production prediction has improved at the inland field level. Soil quantity and climate patterns are identified the most important component in agricultural productivity. If the weather forecast is more reliable, peasants could be warned ahead of time to minimize the big losses as well as help economic and social development. The forecast also lets farmers determine, for example, whether to pick substitute crops or dump crops in crucial circumstances at an early point. In addition, forecasting crop yields will allow farmers to provide more insight into seasonal crop cultivation as well as its management [14]. For effective crop production and predictive yield results it is therefore important to measure and forecast crop yield prior crop harvesting. Since a non-linear relation is observed between crop requirement as well as crop yield, data mining approaches may help predict crop yields [11][3].

In this work, we propose data mining (DM) methods using machine learning (ML) for predicting the outputs of sugarcane. The matter is simplified into the conundrum of yield grades. Each projected element has its output grade (e.g. low yield quantity, medium yield quantity or even high yield quantity) as per its yield percentage. Then our aim is to forecast and plotting yield grades provided through the collected details. The attributes of the forecast, the sugar cane features, the schedule of planting and the rain intensity are taken as aspects of the yield forecasting [9]. The importance of this research is that we build a prediction dependent on data mining approach to forecast yield grades more reliably than it is based on human intelligence.

II. LITERATURE REVIEW

A sugar cane yield prediction approach using Random Forest [7] was suggested by [4]. In their study, the details on bio-agricultural-mass, environmental parameters (i.e. rainfall, radiation and regular

temperature, etc.) as well as yields of 2 previous years were included.

On the other hand [8] has presented two predictive tasks:

(i) the issue of identification to determine whether the output is higher or lower standard value measurements (t / ha) and
(ii) the issue of regression in 2 separate cycles to address forecast the yield figures.

For a list of the methods used, [1] defined and provided us facts. In Bangladesh, different crops are produced, and these crops are based on a multitude of factors, like atmospheric sciences, economy as well as geographical variables, covering these techniques as well as practices for anthropological output of varying crops. Information or data can be obtained to support farmers and governments to make good decisions. Their goal here was to apply data mining techniques to extract knowledge from agriculture data in Bangladesh in order to determine quality crop yields for important crops.

Researchers had used related, cost-effective and detailed remote sensed data to provide details on surface of the earth to estimate yield. A quantitative relationship between the remote - sensing data including crop yield was developed using two methods.

The analysis of the hyper-parameter optimization effects, characteristic engineering and functionality evaluation at plot level was proposed by [2] in sugar cane yield forecast. 65 attributes of soil composition, soil nutrients, amount of fertilizers, algorithm, etc. were integrated in the initial dataset. Writers implement other app engineering systems by adding new functionality or removing original apps. The Support Vector Mechanism (SVM) has been utilized as a forecasting algorithm wherein, R Relief [15] algorithms were utilized to pick a function, using Regression Tree (RT), Artificial Neural Networks (ANN), or Boosted Regression Trees (BRT). Grid quest is done via cross-validation for hyper-parameter optimization.

2.1 Problem Statement

The forecast of crop yield had been a major agricultural issue. Farmers increasingly curious and wishes to see how much he receives from current plantation. Mostly in past, an assessment of farmers' past experiences with a given crop was used to estimate the yield forecast. Recently computer-based forecasting models were applied for the crop yield prediction. Crop yield largely depends on climate, soil, water and harvesting preparation conditions. Precise knowledge on crop yield history is critical for decision-making in the context of agricultural strategic planning. Data mining (DM) and Machine learning (ML) has been a crucial means of achieving meaningful and successful solutions. The machine-based forecasting of crop yield consists of evaluating crop yield using available historic data such as soil and environmental factors, as well as historical crop yields.

III. MATERIALS AND METHOD

3.1 Data Collection

The data utilized in this research work were obtained from sugar mills in Muzaffarnagar district of Uttar Pradesh. The data included in this analysis was collected from different cane growing areas in the Muzaffarnagar district for a single cane phenotype over 10 years 1999 to 2019. Data collected from farmers' plots (e.g. cane class / type, soil type, surface area, fertilizer etc.) and from the production of information (e.g., actual yield) were obtained when farmers send the sugar-canes to the mills. The main data was collected for the evaluation and analysis purpose. Data collection and manipulation includes two separate application programs. The back-end of such applications is often MS-SQL for relational database applications (we used .csv files created in MS-Excel). Moreover, SPSS and WEKA tools were considered for data analysis and data validation.

Crop Yield Data

The crop yield of the i^{th} scenario, noted *yi*, is a single numeric value representing the total mass of sugarcane collected on the harvest day (in g/m^2).

The "Cane-management and accounting" based computer program was utilized to collect the actual production details for each plot when the farmers supply their sugar canes to the plant. The volume of rainfall figures was collected at approximately 10 rain sites across the farmers' plotlines. The rainfall size details for each parcel were then obtained from the closest weather station to the parcel. Using the GPS details that were already listed in the database above, the nearest rain facility to an entity was calculated. Table 1 shows data classification.

Table.1: Classifiers for cleaning and formatting of data

Variables	Details
SCType	4 different types of sugarcane varieties were considered for the study: CoS-08279, Co-86032, Co-98014, BO-91
WaterSrc	Filed irrigation sources: Natural Rain, Canals, and Groundwater
AvgRain	The average rainfall per area of the field (mm)
SoilType	Soil type: Clay, Silty, Loam, and Ferus soil
FertType	Fertilizers type: organic and chemical
GrooveWidth	Groove wide of the plot: 150, 140, 130, and 120 cm.
GradePrev	Grade yield of the field from previous season: Grade 1, 2, and 3.
GradeForecast	Target yield grade provided by the model: Grade 1, 2, and 3.
TargetGradePrev	Target yield grades from the previous season: Grade 1, 2, and 3.

CropYield	Amount of sugarcane produced per area (Tons/Hectare)
FarmerGrade	The farmer grade provided by the financial department: A: Good, B: Fair, and C: Poor
ContractsArea	The amount of sugarcanes that the farmer commits to deliver to the mill (Tons/Hectare).
AreaRemain	The actual plotting area used for the cultivation of sugarcane (in Hectares(Ha))

3.2 Models

The concept of the crop yield prediction, which contains the data element correlated with farming data. The input module includes the names of seeds, property, year of cultivation and tons. The pattern for function selection is sponsored. Choose the corresponding seed attribute values. The estimation formula for crop yield used to estimate production. If the feature is chosen, the data is categorized under the same material classification code. It is possible to forecast climate data as well as crop specifications utilized to estimate crop production. During the classification of crop data as regards crop names, season, and total crop yield information for forecasting rules would then apply.

3.3 Data Mining

The data mining approaches can be divided into two areas: statistical methods, such as the regressions of Ridge and Lasso, the key regression variable and the partial least square regression; the machine-learning techniques, like Artificial neural network (ANN), Decision trees, random forests (compound decision trees), k-nearest neighbors (KNN), Regression trees, Support vector machine (SVM) regressions, etc.

In data mining, 2 approaches are used a) clustering and b) classification. Data analysis could be carried out based on classification and forecasting. Classification method can improve predictive performance. The algorithm for data mining, including supervised, unsupervised, and semi-supervised computing, outlines three different ways. Artificial Neural Networks (ANNs), Naive Bays, Decision Tree, C4.5, C5.0, Random Forest, SVMs are classification methods for finding information.

3.3.1 Decision Tree

Decision tree (DT) is an ensemble based on labeled input data. Trees generated can be used for classification and for this reason is called a statistical classifier. However, DTs are prone to over fitting.

3.4 Methodology

We show the methods used in our predictive framework for the yield of sugar cane. The machine input is a database of data with a pre-selected collection of variables, as well as the performance is the expected outcomes of the related information. Three predictive models for data mining have been developed:

(i) Random Forest (RF) (multiple Decision trees classification),
(ii) Gradient Boosting tree analysis (e.g. XGB)[17]
(iii) Regression Analysis

3.4.1 Random Forest Algorithm

Random forest algorithms are a powerful and far more stable optimization methodology than just a singleton Decision tree. RFs add many decision-making bodies to reduce overfitting as well as errors that may occur based on biased and, ultimately, have valuable outcomes (Verikas, Gelzinis & Bacauskiene, 2011). Random forests can be used to find a subset of variables to increase classification precision for the collection of variables. It has also been found that both classification accuracies, outer identification as well as data visualization can be enhanced by the projected matrix of proximity of data through Random Forest trained utilizing such variables [19].

3.4.2 Regression Analysis

The analysis of regression is used for analysis and determination of the associations among proper reaction as well as response variables. Yearly forecasting, cultivated area, are perceived factors involved for assessment during this study. Crop output that rely on a parameter depending on both of these environmental conditions.

$$y = a + bx + \epsilon$$

where y is the dependent variable of linear function x, where, $x = (x_1, x_2, \ldots, x_n)^t \in R^n$, ϵ is the error term where, $\epsilon = (\epsilon_1, \epsilon_2, \ldots, y_n)^t \in R^n$ and a is line intercept (mean value of y-axis at the given value of x) and b is slope of change of regression line, trained by data available as well as specific modelling techniques. Regression coefficient can be calculated using below formula:

$$R_{coff} = \Sigma\ [\ (x_i - x`)(y_i - y`)\ ]\ /\ \Sigma\ [\ (x_i - x`)^2]$$

where $y`$ is the predicted value of the dependent variable, $x_i$ is the value of the independent variable for observation $i$, $y_i$ is the value of the dependent variable for observation $i$, $x`$ is the mean x score, and $y`$ is the mean y score.

3.5 Implementation

Finally, the yield grade of each plot is assigned according to the value of yield amount by using the criteria considered the crop cutting and the criteria set by mill operation management into 3 different grades. The criteria used to assign the yield grade can be shown in Table 1.

Table 2: Conditions used for assignment of yield grades

YieldGrade	Conditions
Grade I (Low)	Sugar-cane yields < 7 T/ha
Grade II (Medium)	7 T/ha < Sugar-cane yields < 12 T/ha
Grade III (High)	Sugar-cane yields > 12 Tons/ha

We added the single-hot encoding for the categorical variables after data pre-processing. We alter

the dataset arbitrarily to 70:30 on the train and the test results. There are 8764 registers in the train and 3752 registers in the check datasets. For each tree, the data not used for training, the out-of-bag (OOB) data can be used to test the generalization performance (OOB error) and it may also be used to estimate variable importance. The number of records in the training range is 3936, 3758 and 1074 respectively for Grades I, II and III. The research collection includes 1712, 1607, and 437 accordingly, documents in Grade I, Grade II, as well as Grade III. In the basis of this, in the training and evaluation datasets we can demonstrate the distribution of Yield-Grade concentrations in Figure 1.

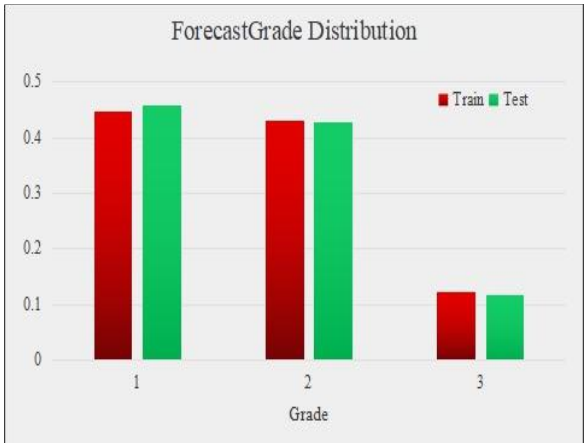

Figure 1: Distribution of Forecast Yield-Grade data for training and testing

$$Adjusted\_R^2 = \left[1 - \frac{(1-R^2)\times(n-1)}{n-k-1}\right]$$

where, n is the number of data-points, and k is the number of independent variables.

IV. RESULTS

By assessing the consistency of the research data collection, we determine our assessment techniques. Only if the forecast in the outcome rate is similar to the real return amount in the set of data that is known as the right identification. We equate our two approaches of random forest as well as regression against various non-machine benchmark methods. Outcomes of the study are represented in Tables and Figures (see Figure 2 and Figure 3, also Table 3 through Table 5). Note: figures in brackets (Table 4) are the normalized values of individual grades.

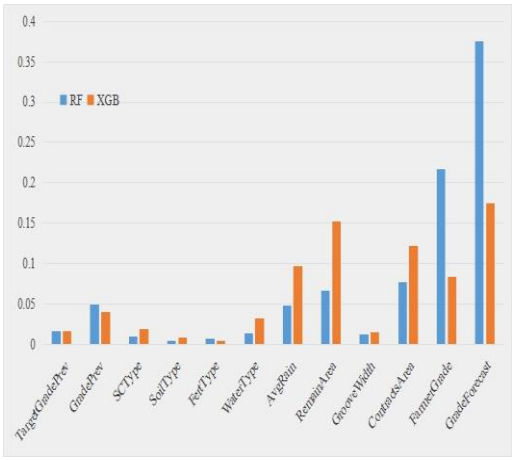

Figure 2: Outcomes of RF and XGB models

Table 3: Comparison of Model outputs

Variables	Classifiers significance value	
	RF	XGB
Target Grade Prev	0.01589	0.01605
Grade Prev	0.04928	0.04012
SC Type	0.00936	0.01821
Soil Type	0.0046	0.00802
Fert Type	0.00719	0.00472
Water Type	0.01304	0.0321
Avg Rain	0.04755	0.09692
Remain Area	0.06624	0.15156
Groove Width	0.01278	0.01512
Contracts Area	0.07629	0.12192
Farmer Grade	0.21663	0.08398
GradeForecast	0.37564	0.17476

Table 4: RF model-based Confusion matrix

		Predicted Value		
		Grade I	Grade II	Grade III
Classification value	Grade I	1380 (0.772)	204 (0.246)	2 (0.001)
	Grade II	201 (0.300)	1196(0.721)	24 (0.019)
	Grade III	14 (0.032)	212 (0.597)	204 (0.519)

Table 5: Correlation of regression analysis

Regression Analysis	
Multiple R	0.769370058
R^2	0.46149855
Adjusted R^2	0.426094201
Standard Error	634709.746
Observations	126
Intercept (y-constant)	**Co-efficient**
2075	2816.115392
	29.76863859

The estimation is based on the regression function R^2 (coefficient of determination), which reflects the proportion of crop variation in the statistical analysis. In addition, adjusted R^2 is determined, which shows how well a line or curve is fitted. In addition, the Adjusted R^2 always will be less than or equal to R^2 on adding additional variables.

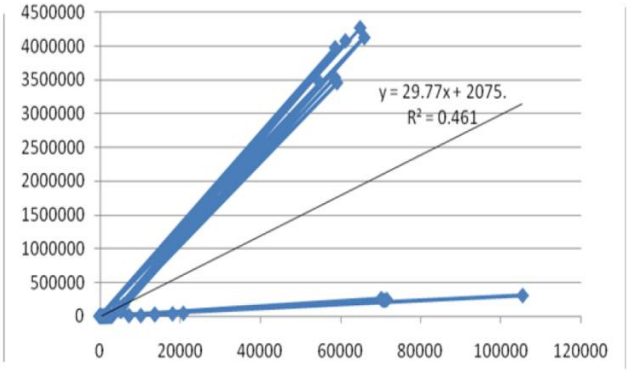

Figure 3: Regression output

V. CONCLUSION

It was observed that previous studies have been trying to estimate sugarcane yield in smaller areas where the conditions and sowing periods are the same. In this study, we effectively modelled sugarcane yield estimates in the Muzaffarnagar district of Uttar Pradesh with a total accuracy of 80.86 percent considering various crop harvesting parameters under various circumstances. We have created a model that can in real time predict crop returns by using time series projecting as well as forecasting crop yield utilizing data mining methods to analyze parameters including years, season, crop size, region and output, etc. The findings obtained were tested and evaluated using IBM-SPSS statistic tools. Thus, efficient and effective strategies that can be built to handle complicated agricultural problems utilizing data mining and machine learning tools have been examined. This in-turn could provide valuable information for crop production enabling farmers as well as aid in socio-economic development of a nation.

REFERENCES

[1] Ahamed, A. M. S., Mahmood, N. T., Hossain, N., Kabir, M. T., Das, K., Rahman, F., & Rahman, R. M. (2015). Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh. In: *IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp. 1-6.

[2] Bocca, F. F. & Rodrigues, L. H. A. (2016). The effect of tuning, feature engineering, and feature selection in data mining applied to rain fed sugarcane yield modeling. *Computers and Electronics in Agriculture, 128*, 67-76.

[3] Chekole, A. & Beshah, T. (2019). Application of data mining tools for identifying determinant factors for crop productivity. *International Journal of Computer Applications, 181*, 16-21.

[4] Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development, 36*(2), 27.

[5] Gandhi, N., Armstrong, L. J., & Nandawadekar, M. (2017). Application of data mining techniques for predicting rice crop yield in semi-arid climatic zone of India. In: *IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, pp. 116-120.

[6] Hammer, R. G., Sentelhas, P. C., & Mariano, J. C. (2020). Sugarcane yield prediction through data mining and crop simulation models. *Sugar Tech, 22*(2), 216-225.

[7] Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. *PLoS One, 11*(6).

[8] Kumar, R., Singh, M. P., Kumar, P., & Singh, J. P. (2015). Crop selection method to maximize crop yield rate using machine learning technique. In: *International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, pp. 138-145.

[9] Manjula, E. & Djodiltachoumy, S. (2017). A model for prediction of crop yield. *International Journal of Computational Intelligence and Informatics, 6*(4), 2349-6363.

[10] Medar, R. A. & Rajpurohit, V. S. (2014). A survey on data mining techniques for crop yield prediction. *International Journal of Advance Research in Computer Science and Management Studies, 2*(9), 59-64.

[11] Raorane, A. A. & Kulkarni, R. V. (2012). Data mining: An effective tool for yield estimation in the agricultural sector. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 1*(2), 1-4.

[12] Raorane, A. A. & Kulkarni, R. V. (2013). Role of data mining in Agriculture. *International Journal of Computer Science and Information Technologies, 4*(2), 270-272.

[13] Singh, P., Singh, S. N., Tiwari, A. K., Pathak, S. K., Singh, A. K., Srivastava, S., & Mohan, N. (2019). Integration of sugarcane production technologies for enhanced cane and sugar productivity targeting to increase farmers' income: strategies and prospects. *3 Biotech, 9*(2), 48.

[14] Srinivas, P. & Santhuja, P. (2019 Sep). Utilization of data mining methods to investigate crop yield forecast. In: *International Conference on Emerging Trends in Science and Engineering (ICESE), 1*, 1-4.

[15] Thuankaewsing, S., Khamjan, S., Piewthongngam, K., & Pathumnakul, S. (2015). Harvest scheduling

algorithm to equalize supplier benefits: A case study from the Thai sugar cane industry. *Computers and Electronics in Agriculture, 110*, 42-55.
[16] Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition, 44*(2), 330-349.
[17] Wang, J., Li, P., Ran, R., Che, Y., & Zhou, Y. (2018). A short-term photovoltaic power prediction model based on the gradient boost decision tree. *Applied Sciences, 8*(5), 689.
[18] Zheng, Z., Lu, P., & Lantz, B. (2018). Commercial truck crash injury severity analysis using gradient boosting data mining model. *Journal of Safety Research, 65*, 115-124.
[19] Ziegler, A. & König, I. R. (2014). Mining data with random forests: current options for real world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4*(1), 55-63.

