

Prediction of Sugar Production using Multivariate Linear Regression

Dr. Sachin Kumar¹ and Hirdesh Sharma²

¹Associate Professor in Dept. of MCA, Noida Institute of Engg. & Tech., Greater Noida, U.P., India
Email: sachinks.78@gmail.com

²Assistant Professor in Dept. of MCA, Noida Institute of Engg. & Tech., Greater Noida, U.P., India
Email: hirdesharma@gmail.com

Abstract—Sugar is a widely consumed commodity all over the world. India produced 33 million metric tons of sugar in 2018-2019, making it the world's largest sugar producer. Sugar production is affected by a variety of factors, including the region under sugar cane, sugar cane production, cane yield per hectare, the number of sugar factories in operation, and the total amount of sugar cane crushed. The current paper, titled "Prediction of Sugar Production Using Multivariate Linear Regression," focuses on sugar production forecasting. Forecasting allows you to keep track of your market by foreseeing challenges and opportunities. Forecasting accuracy is just as important for a company as it is for a person. To render the model using multivariate linear regression for predicting total sugar production, time-series data of sugar production from 1931 to 2018 was used. Cooperative Sugar, Vol. 51, No. 6, February 2020, is the source of the knowledge. The model was trained with 80% of the data and tested with 20% of the data. The association between different parameters is determined using a heatmap. For predicting total sugar output, a multivariate linear regression model is the best fit.

Index Terms— Multivariate Linear Regression, Heat map, Sugar, Sugar Cane.

I. INTRODUCTION

Sugar is one of the world's most important commodities, both in terms of production and consumption. Approximately 110 countries currently generate sugar from either cane or beet, with 8 countries producing sugar from both sources. Sugar cane accounts for approximately 80% of global sugar production, while sugar beet accounts for the remaining 20%. Indian sugar production increased as a result of a well-organized and well-planned sugar program [1-3]. Production, production requirements, capability requirements, and cane production have all been designed and closely monitored by planners. Furthermore, in order to achieve the objectives, set. From time to time, the government has formed committees and task forces to make policy changes in collaboration with industry, state agriculture departments, and others.

Sugar output prediction for agro-industry, or even small-scale farmers, is critical as a technique for determining the size of plantings and reducing the risks associated with this agricultural activity. Sugarcane sugar recovery is affected by water content, sucrose, and invert sugar (Sucipto et al., 2018). Total sugar production is one of the most important factors in sugar recovery, but it is dependent on a number of factors, including sugar cane location, sugar cane production, cane yield per hectare, the number of factories in operation, total cane crushed,

and so on. The aim of this study was to use multivariable linear regression to predict total sugar production [4-6]. Regression is a statistical technique for estimating relationships between variables that have a cause and effect relationship (Gulden et al., 2013). Multi-linear regression is a regression model with one dependent variable and multiple independent variables. Total sugar production is used as a dependent variable in this analysis, while other parameters are used as independent variables. The aim of this study was to create a multi-regression model that included sugarcane variables that could aid in the prediction of total sugar production [7-8].

II. RELATED WORKS

Various researchers are employing various machine learning algorithms to forecast sugarcane yield, sugar quality, and sugar quality based on various parameters.

Sucipto et al. (2018) used an artificial neural network (ANN) prediction model to predict the moisture content, sucrose, and invert sugar of sugarcane (*Saccharum officinarum* L.). The authors have noted that the sucrose level is inversely proportional to the amount of water and invert sugar present. The sucrose level is low when the water content and invert sugar are both high [9].

Xiao, Zeqing, et al. (2016), During the crushing season (December – March) in the Chinese province of Guangxi, the author examines the complex laws of five consistency indices (brix, purity, polarization, sucrose quality, and sugar reduction), as well as graphs over time. The findings demonstrate that the harvest is on schedule. Sugar cane has a number of advantages in terms of sugar production. The regression analysis results show that some of the quality indexes have a good correlation, and the regression models are extremely significant, implying that the prediction results are ideal [10].

S. Sucipto et al (2018), The possible use of bioelectrical properties as a non-destructive sugarcane sugar recovery technique was discovered in this research. The bioelectric properties of sugar cane in the lateral and longitudinal positions of the samples were calculated using a parallel plate capacitor. The sugar recovery as a function of bioelectrical properties has been predicted using an artificial neural network (ANN) [11].

Sankar Vani et al. are a group of people who work in the field of spirituality (2015) The author created an artificial neural network (ANN) model to predict sugar yields for various biomass loadings and substrate sizes while keeping other hydrolysis parameters constant. Yvette Everingham et al., (2016), Random data accuracy review Forests to explain annual sugar cane productivity changes and the suitability of the crop's predictor variables. Models in combination with observed and seasonal climate data Indexes of prediction are limited [12]. Gabriele Torino Pedroso et al. (2016) set out to develop a multiple regression model that targets sugarcane traits and could aid in the prediction of alcohol production. The multiple regression model used was one in which Pol and Fiber production traits were used as predictors of alcohol production [13].

Shen Yin et al. (2013) used multivariate statistical approaches to create Output prediction models. The models of the relationship between grape physicochemical indexes and wine quality are developed using multivariate statistical methods [14].

Katarzyna Wodarska et al. (2019) employs multivariate approaches to analyze the physicochemical and sensory profiles of apple juices and uses instrumental measurements to predict critical sensory properties. Color, turbidity, total soluble solids, sucrose, D-glucose, D-fructose content, pH, total acidity, total phenolic, total flavonoid content, and antioxidant potential are the physicochemical parameters [15].

Sanjay Sharma et al. are a group of researchers who have come up with a novel way to (2015) To predict rice output and consumption, the author establishes correlations between production input variables and consumption input variables. In order to forecast rice production and use, correlation analysis is used to construct the fuzzy inference method rule base [16].

Nasir, M.L. et al. (2000) interconnected neural networks, the researcher discovers a non-trivial novel approach to implementing artificial neural networks to predict corporate bankruptcy [17].

Repon Gama et al. discovered that ANN can accurately model and forecast non-linear data such as apple pomace hydrolysis and can be used to determine optimal hydrolysis conditions, as well as indicating that other hydrolysis conditions can be effectively modeled using an ANN [18].

III. PROPOSED MODEL

The proposed model's main objective is to estimate total sugar production. There are four sections to the proposed model. The proposed model [19-23] is represented in Fig. 1 as a flow diagram.

- a) Data Collection
- b) Data Preparation

- c) Features Selection
- d) Applying multi variant regression model

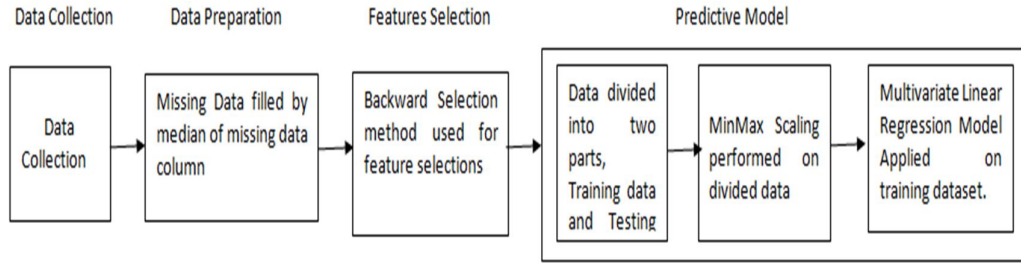


Fig. 1

IV. DATA COLLECTION

The data (SUGARCANE, SUGAR AND MOLASSES PRODUCTION AT A GLANCE) used for analysis consists of sixty records and each record contains twelve attributes collected from " ICAR –SUGARCANE BREEDING INSTITUTE, COIMBATORE ". Attributes of dataset and their description are given in Table I:

TABLE: I

S.No.	Variables	Description
1	Year	Year
2	SUGARAR	Area under sugarcane ('000 hectare)
3	PRCANE	Production of sugarcane ('000 tonnes)
4	YCPHT	Yield of cane per hectare (tonnes)
5	NOFACOP	No. of Factories in operation
7	AVCAPTONPD	Average capacity (tonnes per day)
8	TCANECRU	Total cane crushed ('000 tonnes)
9	TSUGPRO	Total sugar produced ('000 tonnes)
10	RESU	Recovery of sugar (% cane)
11	MOP	Molasses production ('000 tonnes)
12	REMO	Molasses (% cane)

V. DATA PREPARATION

Some columns of data were missing in the dataset. Missing data filled by median value of the missing data column. fillna() method of pandas used for fill the missing data. comma separator removed from the entire dataset [24].

A. Features Selection

Aim of study is to predict the total sugar production. On observation of data shows that attribute Year, Recovery of sugar, Molasses production and Molasses (% cane) not contribute in production of sugar. Molasses is by-products of the sugar industry [25-26]. Heatmap is used for finding the co relation among the different attributes. The heatmap is a two-dimensional graphical representation of data where the individual values found in the matrix are displayed as colors. Fig 2 shows correlation among the attributes of data.

In this study backward elimination technique has used for features selection. In this technique firstly select the significant level (0.05) and fit the model with all the features. If highest P value feature has greater than the significant level, then remove the feature and model has ready otherwise repeat the process with remaining features [27]. Table 2 shows that list of significant features PRCANE, AVCAPTONPD and TCANECRU for prediction of total sugar production.

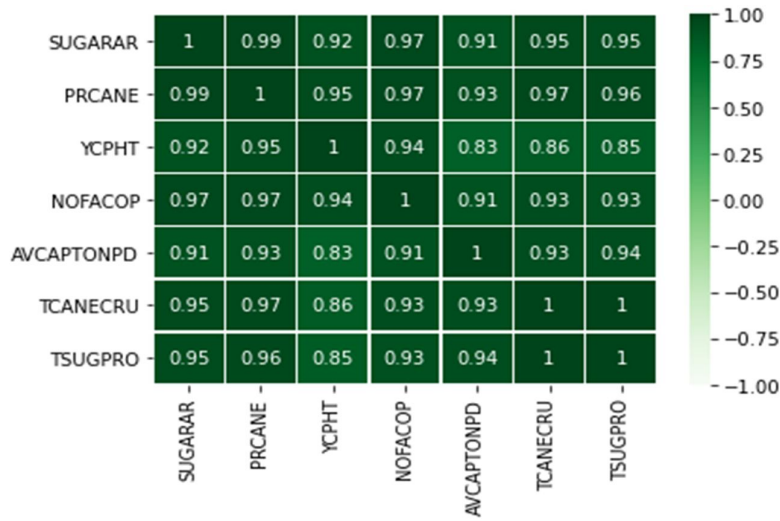


Fig: 2

TABLE: II

Features	P
PRCANE	0.001
AVCAPTONPD	0.02
TCANECRU	0

VI. APPLYING MULTI VARIANT REGRESSION MODEL

Regression analysis is one of the most studied techniques used in the analysis of data. It follows a supervised machine learning algorithm. Regression analysis is an important statistical method which helps us to analyze the relationship between two or more variables in the dataset. In Multivariate regression output is predicted based on the number of independent variables with one dependent variable [28]. The multivariate regression model formulates as estimate TSUGPRO as a function of 'TCANECRU', 'AVCAPTONPD', 'PRCANE'

$$TSUGPRO = F('TCANECRU', 'AVCAPTONPD', 'PRCANE')$$

The linear form of the model is:

$$TSUGPRO = \beta_1 * TCANECRU + \beta_2 * AVCAPTONPD + \beta_3 * PRCANE + \text{Intercept}$$

Table III shows the value of Coefficient/ Intercept at MinMaxScaler.

TABLE: III

Coefficient/ Intercept	Coefficient and Intercept value at MinMaxScaler
β_1	1.0149
β_2	0.0531
β_3	-0.0967
Intercept	0.0015

VII. RESULT

Aim of study is to build a model that predict the Total sugar production based on different parameter such as SUGARAR, PRCANE, YCPHT, NOFACOP, AVCAPTONPD and TCANECRU. But after feature selection

step only three independent variable left these are 'TCANECRU,' AVCAPTONPD 'and 'PRCANE'. These three independent variable having linear relationship with dependent variable TSUGPRO. In this study, the data set was divided into two sections Eighty percent for training and 20 percent for testing. Fig 3, Fig 4 and Fig 5 show the linear relationship between total sugar production and significant variables [29-31].

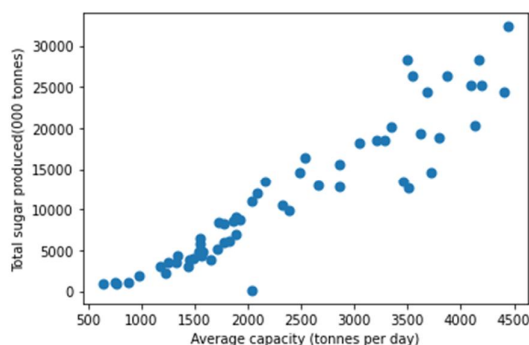


Fig: 3

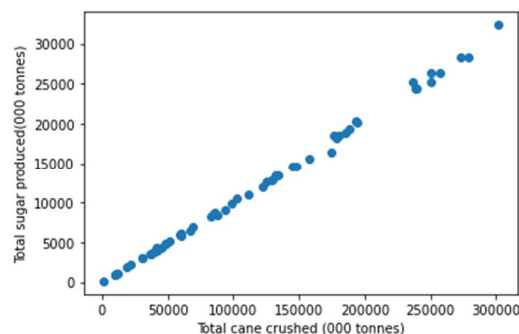


Fig: 4

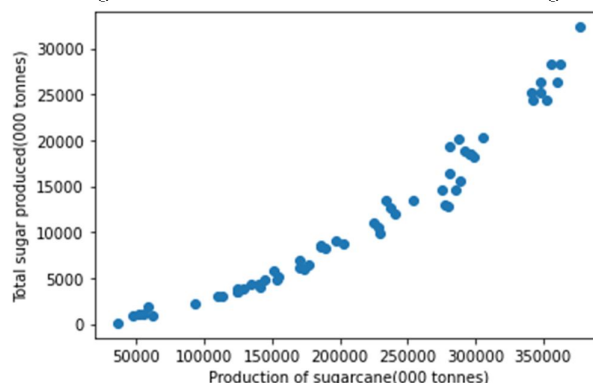


Fig: 5

VIII. CONCLUSION

A model for predicting total sugar production based on various variables is proposed in this study. Data was obtained from the "ICAR –SUGARCANE BREEDING INSTITUTE, COIMBATORE" for research. The backward exclusion approach was used to select significant features from the dataset for the model [32]. The multivariate linear regression technique was used to test selected features. Here, 80% of the data was used to train the model and 20% was used to test it. MinMaxScaler transforms features by scaling them from zero to one in this paper. For predicting total sugar production, a multivariate linear regression model was used after scaling. This model provided the best fit for the forecast, with a model accuracy of 97.23%[33].

REFERENCES

- [1] Sucipto, S., Niami, M. W., Hendrawan, Y., Al-Riza, D. F., Yuliatun, S., Supriyanto, S. and Somantri, A. S.(2018) Prediction of water content, sucrose and invert sugar of sugarcane using bioelectrical properties and artificial neural network . International Food Research Journal 25(6): 2674-2680
- [2] Pedroso G., Silva-Mann R., Camargo M., Russo S. (2014) Applied multiple regression for autocorrelated sugarcane data . African Journal of Agricultural Research Vol. 9(10), pp. 914-920
- [3] Yin S., Zhu X., Karimi H. (2013) Quality Evaluation Based on Multivariate Statistical Methods. Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2013, Article ID 639652, 10 pages
- [4] Xiao Z., Liao X., Guo S. (2017) Analysis of Sugarcane Juice Quality Indexes. Hindawi Journal of Food Quality Volume 2017, Article ID 1746982, 6 pages
- [5] Everingham Y., Sexton J., Skocaj D., Inman-Bamber G. Accurate prediction of sugarcane yield using a random forest algorithm, Agron. Sustain. Dev. (2016) 36: 27

- [6] Sucipto, S, Arwani, M, Hendrawan, Y, Widaningtyas, S, Al Riza, D F, Yuliatun, S, Supriyanto, S, Somantri, A S Bioelectrical measurement for sugar recovery of sugarcane prediction using artificial neural network . EECSI 2018, Malang – Indonesia
- [7] Włodarska K., Pawlak-Lemańska K., Sikorska E. Prediction of key sensory attributes of apple juices by multivariate analysis of their physicochemical profiles. *British Food Journal*, vol. 121 no. 10
- [8] Sharma S. , Patil S. Key indicators of rice production and consumption, correlation between them and supply-demand prediction. *International Journal of Productivity and Performance Management*, vol. 64 no. 8
- [9] Nasir M., John R., Bennett S., Russell D., A Patel Predicting corporate bankruptcy using artificial neural networks . *Journal of Applied Accounting Research*, vol. 5 no. 3
- [10] Vani S, Sukumaran RK and Savithri S. Prediction of Sugar yields during hydrolysis of lignocellulosic biomass using artificial neural network modeling. *Bioresour Technol* 2015; 188:128-135.
- [11] Gama R., Dyk J., Burton M., Pletschke B. Using an artificial neural network to predict the optimal conditions for enzymatic hydrolysis of apple pomace . Springer-Verlag Berlin Heidelberg 2017
- [12] Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S (2016) Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Trans Med Imaging* 35(5):1207–1216
- [13] Barakat N, Bradley AP, Barakat MNH (2010) Intelligent support vector machines for diagnosis of diabetes mellitus. *IEEE Trans Inf Technol Biomed* 14(4):1114–1120
- [14] Becker C, Gather U (2001) The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules. *Comput Stat Data Anal* 36(1):119–127
- [15] Black N, Payne M (2003) Directory of clinical databases: improving and promoting their use. *BMJ Qual Saf* 12(5):348–352
- [16] Burges CJ (1998) A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc* 2(2):121–167
- [17] Chen M, Hao Y, Hwang K, Wang L, Wang L (2017) Disease prediction by machine learning over big data from healthcare communities. *IEEE Access* 5:8869–8879
- [18] Çomak E, Arslan A, Türkoğlu İ (2007) A decision support system based on support vector machines for diagnosis of the heart valve diseases. *Comput Biol Med* 37(1):21–27
- [19] Devarajan M, Subramaniaswamy V, Vijayakumar V, Ravi L (2019) Fog-assisted personalized healthcare-support system for remote patients with diabetes. *J Ambient Intell Hum Comput* 1:1–14
- [20] Finney JM, Walker AS, Peto TE, Wyllie DH (2011) An efficient record linkage scheme using graphical analysis for identifier error detection. *BMC Med Inform Decis Mak* 11(1):10–12
- [21] Imhof M, Bauer M, Gather U, Löhlein D (1998) Statistical pattern detection in univariate time series of intensive care on-line monitoring data. *Intensive Care Med* 24(12):1305–1314
- [22] Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD (2016) Machine learning and decision support in critical care. *Proc IEEE Inst Electr Electron Eng* 104(2):444–446
- [23] Li J, Xie X, Song J, Yang H, Faraut G (2019) Guest editorial special issue on automation science and engineering for smart and interconnected healthcare delivery systems. *IEEE Trans Autom Sci Eng* 16(1):2–5
- [24] Maglogiannis I, Loukis E, Zafropoulos E, Stasis A (2009) Support vectors machine-based identification of heart valve diseases using heart sounds. *Comput Methods Programs Biomed* 95(1):47–61
- [25] Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S (2016) Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Trans Med Imaging* 35(5):1207–1216
- [26] Barakat N, Bradley AP, Barakat MNH (2010) Intelligent support vector machines for diagnosis of diabetes mellitus. *IEEE Trans Inf Technol Biomed* 14(4):1114–1120
- [27] Becker C, Gather U (2001) The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules. *Comput Stat Data Anal* 36(1):119–127
- [28] Noura K, Trabelsi A (2012) Intelligent monitoring system for intensive care units. *J Med Syst* 36(4):2309–2318
- [29] Shen L, Chen H, Yu Z, Kang W, Zhang B, Li H, Liu D (2016) Evolving support vector machines using fruit fly optimization for medical data classification. *Knowl-Based Syst* 96:61–75
- [30] Sinha P, Sinha P (2015) Comparative study of chronic kidney disease prediction using KNN and SVM. *Int J Eng Res Technol* 4(12):608–612
- [31] Verplancke T, Van Looy S, Benoit D, Vansteelandt S, Depuydt P, De Turck F, Decruyenaere J (2008) Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. *BMC Med Inform Decis Mak* 8(1):1–8
- [32] Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ (2010) Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med Inform Decis Mak* 10(1):1–7
- [33] Yu Z, Luo P, You J, Wong HS, Leung H, Wu S, Han G (2015) Incremental semi-supervised clustering ensemble for high dimensional data clustering. *IEEE Trans Knowl Data Eng* 28(3):701–714.