The International Society of Precision Agriculture presents the
16th International Conference on
Precision Agriculture
21–24 July 2024 | Manhattan, Kansas USA

Paula Marcela Valencia Ramírez[1]

A paper from the Proceedings of the
16th International Conference on Precision Agriculture
21-24 July 2024
Manhattan, Kansas, United States

Abstract.

Sucrose is one of the most important indicators in the final profitability of Colombian sugar mills, therefore, its understanding and forecast are fundamental for the business. In this work, a proposal is formulated for an analysis model that allows predicting the percentage of sucrose based on historical data from mechanically harvested farms with the objective of knowing the numerical value of sucrose for each month of milling and be able to plan monthly and annual sugar production.

Regarding the selection of the predictor variables for the construction of the model, the Lasso regularization method was applied. The most important variables for the creation of the model were the accumulated rainfall from ten months of age of the crop, the weeks of maturation, the foreign matter, and the percentage of Diatraea.

Regarding the creation of the model, the XGBoost algorithm was used to perform the numerical prediction of the percentage sucrose variable. To validate the model, 20% of the data was used using the metric: mean square error.

It was found that the theoretical mean square error of the model was 0.25 in terms of the response variable percentage of sucrose, obtaining successful behavior in the monthly predictions of the total cane ground at Ingenio Providencia. For the year 2023, an average difference between the predicted data versus reality of 0.11% of sucrose was obtained which is very close to reality.

Keywords. *machine learning, sucrose, XGBoost, predictor, variables*

[1] Ingenio Providencia. Quality Director. Colombia.pvalencia@providenciaco.com.

The authors are solely responsible for the content of this paper, which is not a refereed publication. Citation of this work should state that it is from the Proceedings of the 16th International Conference on Precision Agriculture. EXAMPLE: Last Name, A. B. & Coauthor, C. D. (2024). Title of paper. In Proceedings of the 16th International Conference on Precision Agriculture (unpaginated, online). Monticello, IL: International Society of Precision Agriculture.

Introduction

Sucrose is the final product of the sugar cane agroindustrial process and is a fundamental indicator for the profitability of sugar mills. Your final result can be affected by multiple variables that have a positive or negative impact. (Maldonado and Arteaga,2017). In particular, since sugarcane processing to obtain sucrose begins in the field, characteristics such as cane variety, soil, management practices and climate affect the final sucrose response. (Larrahondo, 1995).

The Colombian sugar sector in recent years has suffered a decrease in sucrose content, which is worrying because this decrease directly affects the profitability of the sector.One of the reasons for the decrease in sucrose is the atypical and adverse climate of recent years, which largely explains this panorama (Cenicaña, 2016).

It is essential that the sugar sector continues to search for information and analysis tools that allow us to better understand the dynamics of sucrose. One of these tools may be the use of machine learning that allows predicting the sucrose variable and identifying what improvement actions should be implemented to improve this indicator and therefore contribute to increasing the profitability of the sector.

Machine learning (ML) covers a wide range of algorithms and modeling tools that are used for a wide variety of tasks and data processing, and has been applied in most scientific disciplines in recent years. (Carleo et al, 2019). Machine learning was proposed by Samuel in 1952 and has been widely applied in computer vision, economics and data mining among other areas. (Wei et al,2019).

According to Shinde and Shah (2018) a lot of research has been carried out to make machines intelligent. Learning is a natural human behavior that has also become an essential aspect of machines. Traditional machine learning algorithms have been applied in many areas of knowledge achieving excellent results.

Zhao and Della (2020) developed empirical models built using machine learning algorithms to predict the sucrose content of sugarcane based on multi-temporal spectral data with a combination of agrometeorological data. The performance of the built models was evaluated by comparing the statistical model and the machine learning model with only spectral data. The result showed that the machine learning model has better performance compared to the statistical model.

Shendryk and Davi (2021) used Sentinel-1 and Sentinel-2 satellite imagery in combination with climate, soil, and elevation data to predict field-level sugarcane yields across multiple sugar mill areas in the humid tropics. from Australia. They compared the predictive performance of models based solely on satellite images and a fusion of satellite images with climate, soil and topographic information. Random hyperparameter search was the method used to optimize and identify the most accurate decision tree-based machine model. Overall, gradient boosting was the most accurate method for predicting sugarcane attributes.

The algorithm used in this study was XGBoost, which stands for extreme Gradient Boosting, it is an implementation of gradient boosting algorithms that has gained popularity due to its speed and performance.

This algorithm has proven to be very effective in solving machine learning problems; it is based on the gradient boosting algorithm that optimizes the precision of the models and computational efficiency. It has a great advantage with the presence of missing data, this makes it possible to build models even if all the data is not available.

Therefore, the objective of this work is to build and validate a predictive analytics model for the variable percentage of sucrose in sugarcane using machine learning models, specifically the XGBoost algorithm.

Methodology

The methodology implemented to carry out this project was CRISP DM (Cross-Industry Standard Process for Data Mining). CRISP–DM [CRISP-DM, 2000], is the most widely used reference guide in the development of Data Mining projects (CRISP-DM). This methodology has the following phases: Understanding the business, understanding the data, data preparation, modeling, evaluation, and deployment.

The creation of a machine learning model is proposed that allows predicting the percentage of sucrose from variables found in the harvest history of the Siagri information system of Ingenio Providencia.

Data collection

The Providencia mill has a data information system called SIAGRI that provides all the necessary data required for the analysis. The SIAGRI information system is fed daily with production data from each farm harvested by Ingenio Providencia, information that was required for the execution of the project, where daily data on the percentage of sucrose and other variables associated with the crop are recorded. The database analyzed included the periods between 2013 -2021.

Cleaning and preparing data

Regarding the quality of the data, the number of null values, duplicate data and verification of outliers was verified. Records with data considered atypical were eliminated, such as a dose per hectare of ripening agent greater than two liters and weeks of ripening greater than 20. In this case, an expert in the crop was consulted.

Lasso regularization model

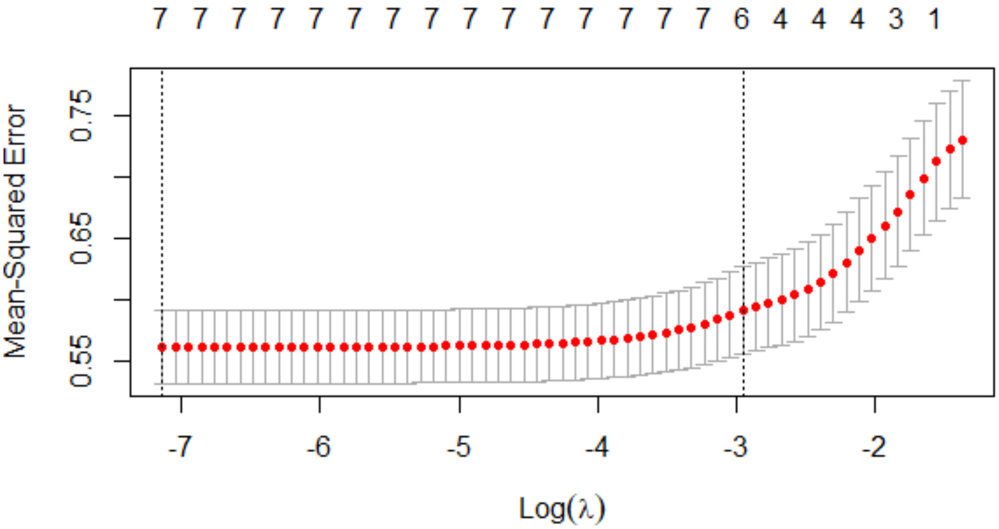
It may happen that many of the variables included within the prediction model are not associated with the response, these unnecessary variables can incur complexity of the resulting model, by regulating these variables we can obtain a model with greater precision. Regularization methods are used for model selection and to avoid overfitting in predictions. When estimating regression coefficients by least squares, we may encounter the problem of collinearity. This problem prevents obtaining adequate estimates and predictions, so regularized regression methods such as Ridge, Lasso and Elastic Net must be used. For this work, the Lasso regression method was used.

The analysis was developed using the glmnet package of the R program. A train data set was defined with 80% of the observations in the database and a test data set with the remaining 20% of observations, these were used to train and test the model.

The k-fold Cross Validation was performed in order to find the best lambda that minimizes the

RMSE mean square error, below is the graph where each RMSE value was contrasted with the log lambda:

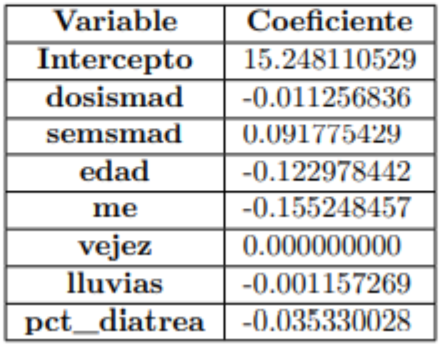
Figure 2. Behavior of the RMSE versus the Log (ƛ)



In the previous figure it is observed that the smallest error occurs for a Log(\lambda) = -5.344152, which means a \(\lambda = 0.0007958893\), with this value of lambda we proceed to train a model with the train data set and after this make a predict with the test data set.

The mean square error of the Lasso model for a lambda of 0.0007958893 is 0.6952. Now we proceed to create a model with all the observations and with the best lambda, to see the coefficients of the different regressor variables.

Table 1. Coefficients of the LASSO model for a λ = = 0.0007958893



Variable	Coeficiente
Intercepto	15.248110529
dosismad	-0.011256836
semsmad	0.091775429
edad	-0.122978442
me	-0.155248457
vejez	0.000000000
lluvias	-0.001157269
pct_diatrea	-0.035330028

The previous table shows the resulting variables after carrying out the regularization method. Lasso reduces the coefficient estimates towards zero, penalizing the least important variables for the model.

Final data set

The variables chosen after applying the Lasso regularization method are the following:

- Weeks of maduration
- Dose of maduration agent
- Total Foreign Matter
- % Diatraea infestation
- Rains
- Age

Definition of variables

Sucrose percentage: This is the response variable and is essential for the profitability of the sugar mills because in the end it translates into kilos of sugar produced.

Ripening dose: Chemical ripening agents increase the sucrose content in the upper internodes of the cane. The dose of the ripening agent corresponds to the amount in liters per hectare applied.

Weeks from the date of application: Ripeners such as the bonus must be applied in the final stage of cane growth and a few weeks must pass after application for them to achieve their sucrose concentration effect.

Foreign matter: Foreign material, especially that made up of buds, has a high incidence on the levels of color and impurities such as soluble polysaccharides, phenols and amino-nitrogens. These chemical constituents affect the crystallization process and the final quality of the sugar in relation to its color and also affect the extraction of sucrose.

Rainfall: These are the millimeters of rain accumulated in the final periods of sugarcane growth. Rainfall increases the moisture content in the soil and stems and this affects the concentration of sucrose in the crop. Furthermore, the rainfall regime can influence the contents of foreign materials, since a greater amount of soil adhered to the cane is expected during the rainy season or greater precipitation.

Diatraea Infestation Percentage: It is the percentage of damage caused by the Diatraea pest insect on the crop. The estimated decrease in tonnage of harvested cane is 0.826% and for sucrose it is 0.038% for each percentage unit of infestation intensity (Cenicaña, 2013).

Predictive modeling

For this study where the variable we want to predict is numerical, we chose the XGBoost algorithm, which as noted in the literature review has proven to be an exceptionally powerful and versatile machine learning tool. Its superior performance, flexibility, built-in regularization, intelligent handling of missing data, and scalability make it an ideal choice for a wide range of modeling tasks.

The software used to create the model was Google colab, open source software that allows writing and executing Python code.

It was initialized with a random seed, we worked with a 70-30 partition, that is, 70% of the data to train the model and 30% of the data for testing. The hyperparameters used for model calibration were: colsample_bylevel, eta and gamma, to obtain better performance. Cross-validation was performed with a CV of 10 to facilitate optimization of hyperparameters and evaluation of model performance.

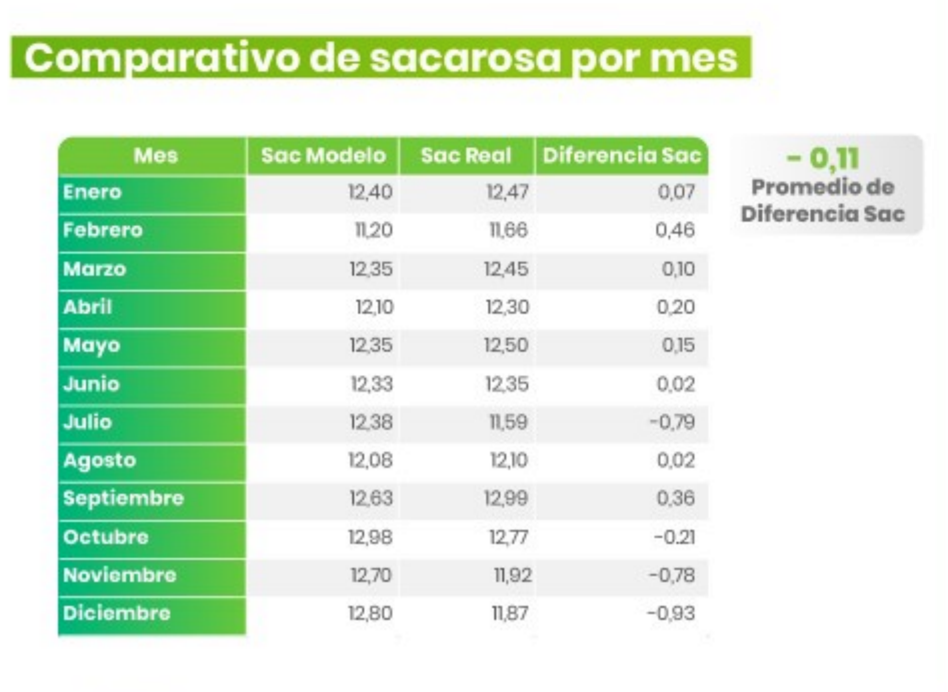
The metric used on the testing data set to evaluate the performance of the created model was the root mean square error (RMSE), which can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Results

It was found that the mean square error of the model was 0.25 in terms of the response variable percentage of sucrose, obtaining successful behavior in the monthly predictions of the total cane in the Ingenio.

Table 2. Comparison of sucrose by month



Comparativo de sacarosa por mes

Mes	Sac Modelo	Sac Real	Diferencia Sac
Enero	12,40	12,47	0,07
Febrero	11,20	11,66	0,46
Marzo	12,35	12,45	0,10
Abril	12,10	12,30	0,20
Mayo	12,35	12,50	0,15
Junio	12,33	12,35	0,02
Julio	12,38	11,59	-0,79
Agosto	12,08	12,10	0,02
Septiembre	12,63	12,99	0,36
Octubre	12,98	12,77	-0.21
Noviembre	12,70	11,92	-0,78
Diciembre	12,80	11,87	-0,93

- 0,11
Promedio de
Diferencia Sac

In the previous illustration, the average of the mean square error of the months from January to December 2023 is observed, observing an average RMSE of 0.11 in the model prediction with respect to the reality of the % sucrose in Ingenio Providencia, data which is very close to reality.

The model has a good predictive capacity, regarding the farms, since it is anticipating the final results, which makes it possible to plan the amount of sugar that will be produced month by month.

Discussion and Conclusions

The results obtained in the Lasso regularization model have complete logic from an agronomic point of view, that is, sucrose is a variable that is affected by previous existing climatic conditions, such as precipitation. The reduction in moisture content in the stems induces the conversion of reducing sugars to sucrose; On the contrary, if there is previous humidity or precipitation, the formation of sugar is reduced. The accumulated rains from 10 months until the harvest date largely explain the final result of sucrose.

The application of the ripener is essential to increase the sucrose content in the sugarcane, but the number of weeks in which the ripener acts after being applied is also critical.

The contents of foreign matter, as mentioned above, affect the crystallization process and the final quality of the sugar in relation to its color and also affect the extraction of sucrose.

The percentage of Diatraea, which is one of the insect pests that affect sugar cane in Colombia, negatively affects the sucrose contents, since due to its perforations in the stem, considerable decreases in the contents of this response variable occur. .

Regarding the prediction model, the mean square error was 0.1% during the year 2023, this result allows us to demonstrate the successful operation of the model with respect to its performance on real data.

It is evident that the algorithm used allows us to get very close to the reality of the variable monthly and annual percentage of sucrose, allowing the Ingenio to plan the amount of sugar it will produce per month.

References

Maldonado y Arteaga, M., 2017. RECUPERACIÓN DE SACAROSA EN LA COSECHA MECÁNICA EN VERDE DE CAÑA DE AZÚCAR. ESAI Business School de la Universidad Espíritu Santo (UEES), Km. 2.5 Via a Samborondon, Samborondon Ecuador.

Larrahondo, J., 1995. Cenicaña 1995. Calidad de la Caña de Azúcar. En: El Cultivo de la Caña de la caña en la Zona azucarera de Colombia. Cali.412 p.

Cenicaña, . 2016. Carta Informativa. Etanol 10 años de producción en

Colomba Año 4. Número 3. Cali, Colombia.

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., ... & Zdeborová, L. (2019). Machine learning and the physical sciences. Reviews of Modern Physics, 91(4), 045002.

Wei, J., Chu, X., Sun, X. Y., Xu, K., Deng, H. X., Chen, J., ... & Lei, M. (2019). Machine learning in materials science. InfoMat, 1(3), 338-358.

Shinde, P. P., & Shah, S. (2018, August). A review of machine learning and deep learning applications. In 2018 Fourth international conference on computing communication control and automation (ICCUBEA) (pp. 1-6). IEEE.

Zhao, Y., & Della Justina, D. (2020). Machine learning approaches for crop growth monitoring using multi-temporal and multi-variety remotely sensed data. In IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium (pp. 4890-4893). IEEE

Shendryk, Y., Davy, R., & Thorburn, P. (2021). Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. Field Crops Research, 260, 107984.

https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf