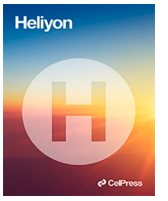
Heliyon 11 (2025) e42632

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Data science for pattern recognition in agricultural large time series data: A case study on sugarcane sucrose yield

Laura Valentina Bautista-Romero [a], Juan David Sánchez-Murcia [b], Joaquín Guillermo Ramírez-Gil [c,*]

[a] *Universidad Nacional de Colombia, sede Bogotá, Facultad de Ciencias Agrarias, Departamento de Agronomía, Colombia*
[b] *Universidad Nacional de Colombia, Sede Bogotá, Facultad de Ciencias, Departamento de Matemáticas, Colombia*
[c] *Universidad Nacional de Colombia, Sede Bogotá, Facultad de Ciencias Agrarias, Departamento de Agronomía, Laboratorio de Agrocomputación y Análisis Epidemiológico, Center of Excellence in Scientific Computing, Colombia*

ARTICLE INFO

Keywords:
Data cleaning
Visualization tools
Machine learning
Frequentist model
Digital tools

ABSTRACT

Data science (DS) is one of the areas with the greatest versatility for application in any field of knowledge. It allows the optimization of different processes of daily life and permits the analysis of massive amounts of data. DS combines computer programming with mathematics and statistics tools in multiple environments such as Python, Julia, R, among others. Here, a protocol was proposed to use DS tools applied to the organization, visualization, and analysis of historical data in sugarcane production systems in the tropics as a basis to identify patterns associated with sucrose. The protocol consisted of four phases: (i) data collection and organization, (ii) data management, cleaning and incorporation of new variables, (iii) visualization tools, and (iv) analysis and modeling based on a multiapproach using a frequentist model (generalized lineal model), a regularized regression model (Lasso) and machine learning models (AutoML). Each of the phases was implemented using multiple algorithms and techniques to automate processes such as queries, numerical calculations, sorting, grouping, dividing, pivoting, totalizing, concatenation, cleaning, visualization, and fitting to models using the free Python software and libraries including *Pandas*, *Numpy*, Plotly, Matplotlib, *SciPy*, *PySpark, Scikit-learn, Statsmode,* among others. Each of the phases allowed the elimination of variables that obscured the analysis process by considering parameters such as Pearson correlation, exploratory analysis, and modeling. Important variables that offered value in the analysis were obtained, considering those variables related to the soil as those of minor contribution, and climatic variables as the most informative. Our results present an alternative to traditional analyzes in the agricultural sector, based on a step-by-step protocol for the responsible use of DS in the search to understand the behavior and temporal historical patterns of sucrose in sugarcane.

1. Introduction

Data science (DS) also known as big data analytics (BDA) is an interdisciplinary field focused on the study of methods and techniques to extract knowledge from data, help understand multiple phenomena in human life, and optimize processes associated with

* Corresponding author.
E-mail address: jgramireg@unal.edu.co (J.G. Ramírez-Gil).

https://doi.org/10.1016/j.heliyon.2025.e42632
Received 7 July 2023; Received in revised form 7 February 2025; Accepted 10 February 2025
Available online 11 February 2025

decision-making [1–4]. DS combines elements of mathematics, statistics, artificial intelligence, data mining, programming, cloud computing environments, hardware platforms, application programming interfaces (APIs), web scraping, and other fields [2–6]. The DS process is divided into four stages, also called data science pipelines: (i) data collection and organization, (ii) data management, cleaning and incorporation of new variables, (iii) visualization tools and (iv) analysis and modeling [2,7]. Data acquisition involves collecting, preparing, and organizing data for further use. Data cleaning is the stage where data is validated, errors are corrected, missing data is completed, and irrelevant data is removed. Data visualization consists of illustrate results graphically, detecting visual patterns, communicating findings, analyzing, and modeling, the last of which implies building a mathematical model that explains and predicts the data behavior patterns [2,8].

Advancements in crop growth modeling, use of tools to monitor and collect information from farms in a less labor-intensive way, and global navigation systems lead to precision agriculture, in which precise measurements at local points and data-intensive approaches support decision-making on the field [6,9]. Aspects such as long production cycle times, short life cycles, climate issues, and variability in crop yields are challenges in the design and management of agricultural supply chains [10]. In this regard, under the concept of Agriculture 4.0, current agricultural systems aim to optimize the use of various technological tools for collecting direct and indirect data. This approach generates a large volume of information that must be efficiently processed to provide essential elements for informed decision-making [6,11].

Currently, DS tools are an emerging technology in plant modeling, used to improve decision-making processes, estimate results or better understand data behavior [2,9,12,13]. In addition, multiple sets of learning algorithms and statistical analyzes are applied to identify patterns and predict results in large data sets [2,14]. DS is prominent in a variety of fields, with increased application in several agricultural domains, including precision agriculture. Learning from these massive data collections is likely enable the identification of significant opportunities [2,4]. The great versatility of DS, combined with numerous algorithms, open-access software, and multiple data sources, has often led to its improper use [2,4]. This misuse can create false expectations and incorrect decisions, highlighting the need for careful application and interpretation to ensure reliable and meaningful insights in agriculture and beyond.

In this regard, DS holds considerable promise for application in agriculture. However, it is imperative to establish both theoretical and practical conceptualizations to ensure its responsible utilization [2,4,15]. Such conceptualizations are essential in order to promote informative outcomes and facilitate the integration of a set of elements for informed decision-making. The above underscores the need for comprehensive validation and responsible utilization of the entire array of elements covered by DS in its four previously delineated phases [2]. Particularly critical is the meticulous handling of Phase IV (analysis and modeling), which requires a systematic process to select the optimal model. This process must adhere to principles and involve the judicious selection of an appropriate approach based on data structure and user requirements.

DS in agriculture has the potential to revolutionize information analysis and visualization for users, but it must overcome multiple challenges [1,4,6]. These include integrating all actors in the value chain, providing informative support, aiding decision-making, achieving widespread adoption by farmers, ensuring access to digital tools and internet connectivity, among others [2,4,16]. For effective DS application, however, it is essential not only to have data accuracy and quality, but also to establish fundamental criteria that ensure these tools are used correctly [2]. DS faces gaps not only in the conceptual aspects related to its defined phases but also in its applications, especially in the agricultural sector, where it remains an emerging field [2]. Addressing these gaps is crucial to harnessing its full potential for informed decision-making and improved agricultural practices. Likewise, with this contribution, we aim to apply step-by-step strategies for the proper use of the entire set of methods, tools, algorithms, and software that can be used in DS under an approach of responsibility, transparency, and reproducibility, with the objective of ensuring the principles of DS framed within the concepts of the volume, velocity, variety, and veracity of the data [4].

Sugarcane (*Saccharum officinarum* L.) is one of the most important crops in the world [17], generating hundreds of thousands of jobs directly and indirectly, and has become a crucial source of income and development for several tropical countries [18]. It is known worldwide for its high productivity, its use in multiple technology processes, and its high-quality raw material [18]. In addition, it is an important industrial crop that is used for sugar, panela (unrefined sugar), paper, honey for animal feed, bioenergy and other products [17,18]. Brazil, India, China and Thailand are the countries with the highest production volume (~69 %) and the remaining production is produced and geographically distributed across a further 100 countries, including Colombia [17]. The combination of factors such as plant physiology (C4, highly productive), tropical climate, water availability, fertile valley soils, and few legal regulations has allowed Colombia to have one of the highest sugar productivity rates per hectare, represented by tons of sugar per hectare (TCH) [19].

The primary use of sugarcane is to produce sucrose, which can be used in a variety of products. Approximately 65–70 % of world sugar is derived from sugarcane [18,20,21]. The economic yield of sugarcane is determined by the accumulation of sucrose in the stalk [22]. The yield and quality of sugarcane crops depend on several factors and can be positively or negatively affected by climatic conditions [23]. Currently, agronomic and physiological areas of interest focus on improving sucrose yield simulation output accuracy, taking into account the effects of management, soils and weather, year-by-year and site by site [24]. Weather and climate-related events (i.e., growth environment, temperature, precipitation, and other forms of extreme weather) are the key factors for sugarcane production worldwide, especially in many developing countries [17,23,24]. There is great annual and regional variation in sugarcane yields, associated with varying rainfall and temperature due to low adaptive capacity, high vulnerability to natural hazards, and poor forecasting systems and mitigation strategies [25–27]. Gains in sugar content are economically more beneficial than the corresponding increase in cane yield [28]. The accumulation of sucrose has been studied more in sugarcane than in any other plant [29–32].

Modern agronomical research and practice are becoming increasingly data-intensive. Data are continuously collected, analyzed, and simulated to understand and predict crop growth and behavior under various circumstances in the sugarcane industry [33,34].

Conventionally, sugar mill and farm field experts will use their experience to estimate the sugarcane yield of each plot based on the survey data and the historical yield profile of each plot and farmer [35]. The main disadvantage of using this human-based yield estimation method is that there is a large discrepancy between the estimated and the actual yields [35].

It is well known that sugarcane grows under different weather conditions, directly affecting crop maturation, yield, and quality (% of sucrose) [24,28,32]. Raw material quality prediction models are important tools in sugarcane crop management and their goal is to provide productivity estimates during harvest, increasing the efficiency of strategic and administrative decisions [35,36]. In the sugar industry globally, simulation models have potential application to numerous sugarcane production decisions related to biochemistry, agronomy, physiology, pest management, milling processes, and off-site environmental damage. Such advances in improving the accuracy of sucrose yield prediction will greatly help improve production and grinding in the industry [24].

Agricultural systems, particularly commercial sugarcane production systems, and to a lesser extent those dedicated to panela (unrefined sugar) production, accumulate significant amounts of climatic, soil, yield, and quality data. Traditionally, this data is used according to specific and immediate objectives for each cycle. However, there exists a gap in both knowledge and utilization of these data due to the absence of longitudinal analyses that encompass a broad approach to identifying temporal patterns within the acquired historical data series. Adopting a longitudinal and historical approach can facilitate the identification of elements or criteria to improve the understanding of the yield and quality components in sugarcane such as sucrose. However, achieving this requires the application of modern data-analysis approaches, such as data science. To contribute to solving this problem and to give an approximation of the real behavior and temporal patterns of yields in the field determined from sucrose accumulation in the stalks, we apply data science tools with a particular structure as follows: (i) data collection and organization, (ii) data management, cleaning and incorporation of new variables, (iii) visualization tools, and (iv) analysis and modeling. Additionally, our work focused on implementing a step-by-step application of the four phases of data science, including steps, algorithms, and modeling approaches. Through this, we aim to contribute to the agricultural sector by promoting the responsible application of data science, enhancing productivity through mathematical, statistical, and other analytical approaches as a viable alternative to traditional methods.

2. Materials and methods

2.1. Methodological approach

Our methodological approach was based on the responsible and transparent use of data science (DS) as a powerful set of tools and strategies to identify, in detail and through a structured step-by-step process with logical sequencing, how a pattern identification process is developed in a time series of data in the agricultural sector. This is especially relevant when dealing with multiple variables, each with different acquisition and storage protocols, as well as the use of synthetic variables after a prior validation process (Fig. 1).

Our approach was based on the application of the four phases of DS: (i) data collection and organization, (ii) data management,

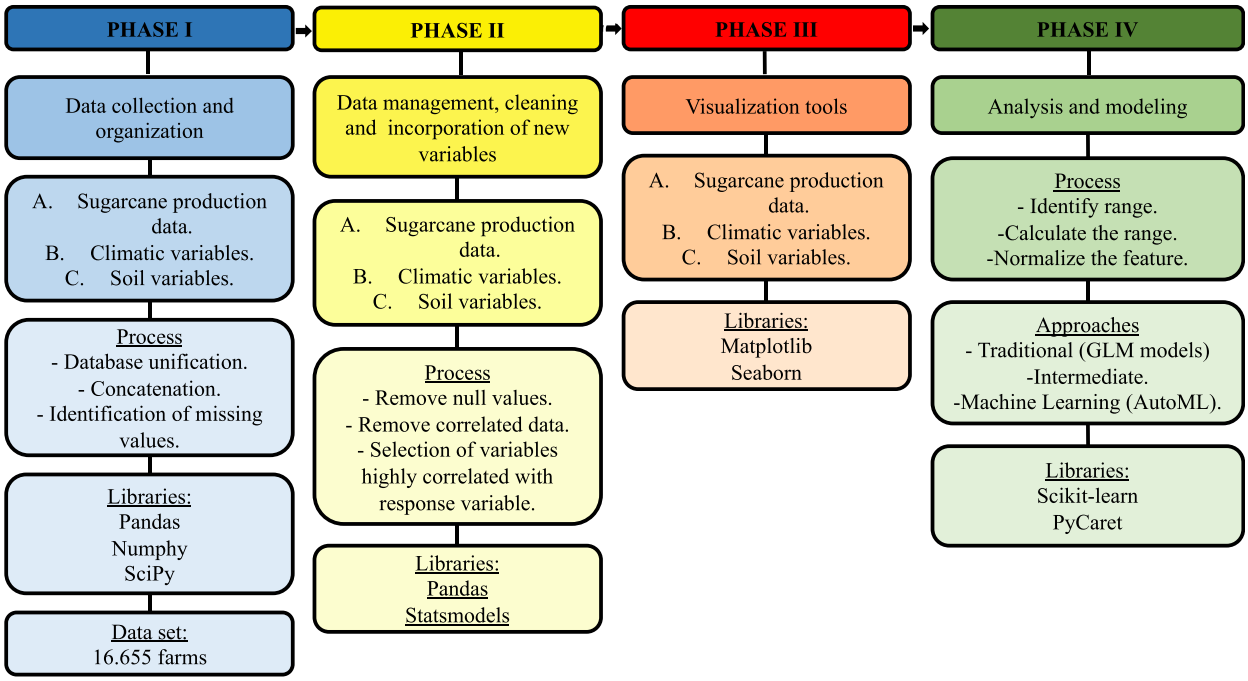


Fig. 1. Step-by-step process diagram for the implementation of the data science approach applied to the analysis of long data series in agriculture as a basis for ensuring the transparency and reproducibility of the group of methods, algorithms, libraries, and software.

cleaning, and incorporation of new variables, (iii) visualization tools, and (iv) analysis and modeling. The protocol proposed the use of DS tools in a responsible and reproducible way, using algorithms, methods, and models. Our objective is to provide a detailed step-by-step explanation of each of the phases and the tools that can be implemented in each of the four phases that we have considered under the DS approach. This process elaborates on the fundamentals and the necessary elements to automatization, improving not only data management processes, but also how secondary information can be obtained to enhance evidence-based decision-making. Furthermore, it illustrates how DS can lead to more accurate decisions in agricultural systems (Fig. 1).

This process was applied to the management and identification of temporal patterns in a historical data set associated with sucrose yield (sucrose field, total sucrose, and % sucrose by cane, as described above) of sugarcane crops grown under tropical conditions (Colombia) based on agronomical and edaphoclimatic data. Similarly, we propose alternatives for the acquisition of data associated with predictive variables using web scraping techniques and their respective evaluation, validation and bias reduction processes (Fig. 1).

Our biological targets were sucrose (Suc), sucrose percentage in cane (Suc % cane), and sucrose measured in the field (Suc field). The Suc-field variable refers to the amount of extractable sugar contained in the cane grown in a plantation. It is a key indicator of quality and performance, used to assess the sugar production potential of a batch of cane. This variable is evaluated by taking juices extracted from a sample of 5 stalks taken in the field and measuring POL (apparent percentage of sucrose (mass/mass ratio)), which is usually done in the laboratory in a digital polarimeter, Brix and humidity of what is obtained in the first mill. The variable Suc represents the losses generated in the CMT (cut, moose, and transport) and the FM (foreign matter) of mineral and vegetable origin. The variable Suc % cane is generally calculated systematically from coefficients and mathematical formulas considering foreign matter and the dwell time. The analysis of the variable sucrose was carried out considering the variables: Suc-field, Suc, and Sac-%-cane.

2.2. Protocol for the data science phases applied to the analysis of patterns between edaphoclimatic and agronomical variables and sucrose on sugarcane

2.2.1. Phase I. Data collection and organization

One of the critical challenges in applying DS concepts is having enough data to identify patterns that enable informed decision making, given that traditional methods of data management consume significant time and economic resources.

For the above reason, we decided, in the first phase, to carry out an automated process for data management (easy selection and filtering of data tables based on position, value or labels), database unification (formatting CSV), concatenation (merging and joining data), identification of missing data as a basis for taming and understanding the data in matrices of easily identifiable rows and columns (data frame), series (lists) with a common ID (productive unit) and a timestamp (time series manipulation), as a basis for searching associated with analysis targets (parameters associated with sucrose). An open-source software as Python was used in Colab, Visual Studio Code, and Jupyter environments. Specific functions from libraries such as *Pandas* were applied for data management, *Numpy* for numerical calculations and data analysis, *SciPy* for optimization, integration, interpolation, and data modification using linear algebra, and *PySpark* for vertical scaling to meet the demands of big data rapidly. The databases obtained from the different sources were stored in Google Drive in order to link them to scripts with the respective codes for post-processing.

The data used in the work was provided by different sugarcane production industries under tropical conditions, specifically from Colombia. The main data sources were obtained at the farmer level. The variables analyzed correspond to the edaphic and agronomic characteristics and the yield components. Specifically, data was provided by sugar mills dedicated to sugar production and other production systems focused on panela production (unrefined sugar), located in the main producing regions of Colombia (departments of Valle del Cauca, Cauca, Meta, Risaralda, Boyacá, Cundinamarca, Santander, Nariño and Antioquia). In total, for the present analysis, 7674 production units were used, including soil, climatic, and crop management variables. Furthermore, 16,652 production units were analyzed without soil variables, but including climatic and agronomic management variables. Climate data was obtained from weather stations covering the sugarcane production zones analyzed. The linkage variables are related to agronomic and climatic characteristics. All data collected cover a period of 10 years (2012–2021).

A. Variables associated with the sugarcane production cycle

The analysis time was 2012–2021, the number of variables analyzed as predictors was 11, and the plot was used for the analysis as a correlation variable between the databases, whose extension varies from 1 to 95 ha on average. The variables were planting month, cutting date, weather season (rainy, dry, transition, based on interannual precipitation patterns), type of harvest (manual or mechanized), ripened application (yes or no), product applied during ripening (commercial maturate used by each production system), week of ripening application (moment in which the maturate was applied), variety (specific genotype), harvest age, TCH (tons of cane per hectare), and purity (percentage of sucrose in total juice solids). The protocols, forms of acquisition, and management of information are specific to each productive unit, but under strict industry guidelines so that they are homologous processes, guaranteeing the reproducibility of the results.

B. Climatic variables

Considering the area of influence of each of the stations in the geographic valley of the Cauca River, seven meteorological stations were selected for the analysis of the climatic variables. The chosen stations correspond from north to south to the following: Cartago (CAR), RUT District (RUT), Zarzal (ZAR), La Paila (PAI), Bugalagrande (BLG), Riofrío (RIO) and Tuluá (TUL). For each of the stations,

the CENICAÑA (Centro de Investigacion de la Caña de Azucar de Colombia) meteorological program provided information. Furthermore, to improve the data set in the Cauca valley region, and in other important production zones such as the Meta, Risaralda, Boyacá, Cundinamarca, Santander, Nariño and Antioquia departments, free data was used from the Weather Station of the Instituto de Hidrologia, Meteorologa y Estudios Ambientales-Ideam (http://dhime.ideam.gov.co/atencionciudadano/) with a buffer zone of 5 and 10 km for precipitation and temperature, since the evaluated zone does not present a change in elevation greater than 10 m. However, in areas with poor coverage by existing stations, "synthetic stations" were validated and implemented. In this context, data validation involved fitting a linear regression model using values from *in situ* stations and synthetic data, determining the bias based on the slope of the curve, and making the necessary corrections [2].

For the proposed use of synthetic data, the web scraping technique was implemented, using freely accessible climatic information data from the Nasa-Power API (https://power.larc.nasa.gov/data-access-viewer/). Web scraping is a technique for the automated consumption of freely accessible data, where websites are accessed with the aim of extracting and storing data in different formats to carry out the analytical process [37]. To automatized the extraction of structured data from the NASA-Power API, we utilized the *Scrapy* library, which is an open-source Python framework [38].

For the period 2012 to 2021, the following climatic variables were used: temperature (minimum, maximum and mean (°C)), relative humidity (minimum, maximum and mean (%)), solar radiation (kcal m^{-2}), precipitation (mm), calculated evapotranspiration (mm), wind speed (mean (m s^{-1})), wind variation (maximum and mean (m s^{-1})), season (dry, wet and transition), climatic variability condition ENSO (The El Niño-Southern Oscillation) (El Niño and La Niña), number of rainy days per month and growing degree days (GDD (equation)). The climatic variables were on a monthly time scale. For the definition and report on the time scale analyzed according to the phenomenon of climate variability (El Nino-La Nina), the information reported by CENICAÑA was consulted (https://www.cenicana.org/boletin-de-seguimiento-enos-condiciones-neutrales-25-abr-2023/). This information was validated with reports associated with the MEIv2 index obtained from the API of the National Office of Oceanic and Atmospheric Administration- NOAA (https://psl.noaa.gov/enso/mei/).

Equation (1). $GDD = \left(\frac{\text{Min T}+\text{Max T}}{2} - \textit{Base T}\right.$, where, GDD: growing degree days, MinT: Minimum temperature, MaxT: Maximum temperature, and Base T: Base temperature = 14 °C

C. Soil variables

Based on the original database and the information system generated by the mills (to produce sugar), and the panela production systems (unrefined sugar), which includes the results of the soil analyzes carried out in each of the plots, 13 variables were analyzed. The physico-chemical variables included were pH, EC (electrical conductivity), Ca (calcium), Na (sodium), camg (Ca/Mg ratio), Cu (copper), Fe (iron), Mn (manganese), Zn (zinc), sand (%), silt (%), clay (%), OM (organic matter). The dates of the results refer to the last analytical soil analysis for each temporal date (2012–2021). The analytical results were based on certified soil Laboratory, used by the sector or each productive unit included in this study.

2.2.2. Phase II. Data management, cleaning, and incorporation of new variables

In this phase of the DS protocol, the objective was to generate a clean data set free from null or atypical values, consistent across all production units, and with non-autocorrelated variables as a basis for identifying patterns and relationships in historical data series using multiple variables (agronomic management, soil and climate) regarding quality characteristics such as sucrose (Suc-field, Suc, and Sac-%-cane). For data analysis, the Pandas library was used. Pandas is a fundamental library in the field of data science. It provides flexible data structures, such as data frames, that allow efficient data storage and manipulation [39]. The use of Pandas makes it possible to perform cleaning and transformation operations, as well as statistical analysis and complex data manipulations [39]. In addition, Pandas offers a wide range of functions and methods that facilitate data exploration and processing [39]. Similarly, the libraries described in phase 1 were used as a complement. Additionally, the *Statsmodels* library was used to conduct statistical tests and statistical data exploration.

Once three datasets consisting of sugar productive variables, climate and soil were obtained, an exhaustive cleaning process was performed, removing null (values out of range, meaningless, and missing data) and correlated data. Initially, variables with many null values were observed. These values were removed from the data sets since they cannot be beneficial in pattern identification. The reason for this is that missing values can obscure general trends and relationships between variables, making it difficult to interpret the data [40]. The variables were then identified in the preprocessing code to analyze the quantitative and qualitative variables independently. Subsequently, each of the data sets was treated individually.

In the case of the soil database, with quantitative variables, once the data with the largest amount of null or zero-value data had been extracted, the dynamics of the variables with the same measurement scale was analyzed. Then, variables with different scales or with high levels of correlation were analyzed. With those variables that theoretically represent a single measurement, it was decided to apply a correlation analysis, more specifically using Pearson's correlation coefficient, with the objective of determining variables with high correlation and proceeding to eliminate them, as they generate noise when performing the analysis, and this can lead to overestimation [41]. The analysis carried out allows the influence of some variables on others to be determined and, with this, information redundancy and noise can be eliminated [42], ultimately leading to a predictive model of higher capacity. Specifically, variables with coefficients >0.80 were removed. The criterion for eliminating one of the autocorrelated variables was based on its low potential impact on the variables of interest (sucrose), as reported in the literature [24,28,31].

Once the cleaning and selection process was carried out, the variables were extracted according to the following criteria: missing values, zero values, different scales, high correlation level, and professional criteria (less potential meaning in relation to sucrose). With qualitative variables, the cleaning was performed considering typos.

A first group of variables was obtained, which were analyzed again by means of a dynamic analysis. A correlation analysis was performed and then a cleaning was performed in which the variables were extracted according to professional criteria, and the presence of null and missing data. In the case of the variables weeks of maturation and product related to the application, the data were replaced with 0 and not applicable, respectively. Then the database of final soil variables was generated.

For the climate database, quantitative variables, null data and data with a value of 0 were analyzed, followed by a visualization of the dynamics of the variables and a correlation analysis. With qualitative variables, the distribution of the variables was analyzed according to their percentage of participation in the database used. The consistency of the database was determined, and no modifications were made according to the results obtained previously.

Finally, for the sugar mill and farm systems an example database was used for implementation of data science processes, such as quantitative variables, null data and 0 values analysis. Additionally, variables dynamics after data extraction with some novelty, as mentioned above, and correlation analysis was performed. For the quantitative variables, a descriptive analysis was performed by evaluating the percentage of variable participation in the data. Therefore, a data analysis system was structured through the tools offered by data science to give an understanding and approximation of the dynamics of sucrose.

2.2.3. Phase III. Data visualization tools

In phase three of our approach to DS, the focus was on incorporating a set of visual tools. These tools not only help to understand the temporal dataset, its quality and structure, but also serve as the foundation for visually identifying potential temporal patterns, trends, and data characteristics (fit to normal distribution or other potential distributions), laying the groundwork for the subsequent phase (analysis).

Considering the process described in the data collection and organization section, we proceeded to perform an exploratory data analysis to address the sucrose production problem using our three data sets. To visualize and graphically analyze the data, *Matplotlib* and *Seaborn* libraries were used. *Matplotlib* is a visualization library that offers flexible and customizable options to create graphs [43]. *Seaborn* complements *Matplotlib*, providing a high-level interface for attractive statistical visualizations [44]. Using these tools, it can perform a visual analysis of the data, as well as an efficient presentation of the findings. It was used to create scatter plots, histograms, line plots, correlation matrices, and box plots.

Based on a visual exploratory analysis, it was agreed to evaluate the normality of the data based on the NQT methods as a statistical tool that can be used to transform non-normally distributed variables to a normal distribution. The above was achieved by rearranging the original data to fit the normal distribution and combining the quantiles of the original data and the normal distribution [45].

2.2.4. Phase IV. Data analysis and modeling

Phase IV of our DS protocol is perhaps the most critical in the process, as there are currently multiple approaches to analyzing, modeling, and identifying relationships between a set of predictor variables and those to be predicted. In our case, based on the proposed objectives, our aim is to present different modeling methods, without implying that we propose them as "correct" methods, especially when our analysis example is so complex (sucrose in sugarcane). Thus, the intention is to demonstrate that there are multiple approaches for this phase and that a step-by-step process is required to adhere to the principles of each approach. Each specific biological target may have multiple modeling options, and what matters most is not necessarily finding a model with a perfect fit, but one that makes biological, ecological and productive sense. A less complex model that avoids overfitting would be preferable.

With the results obtained in the data visualization process (phase 3), the strategies to be used in the data pre-processing were identified. For numerical features, the MinMax normalization technique implemented in the *scikit-learn* library was used, in order to rescale our variables to the range 0–1 and obtain optimal and interpretable results [46]. This approach is particularly useful when dealing with features with varying scales or units, and we assume that the data follow a uniform distribution and outliers are not present (based on previous cleaning process) [46]. To normalize, we follow three steps according to equation (2): i) identify the range: determine the minimum and maximum values of the feature you want to normalize; ii) calculate the range: compute the range of the feature by subtracting the minimum value from the maximum value, and iii) normalize the feature: For each data point in the feature, subtract the minimum value, and then divide by the range. This ensures that the values are scaled between 0 and 1.

Equation (2). $X_{\text{norm}} = \frac{X - Xmin}{X_{\max - Xmin}}$, where: X is the original value of the feature, Xmin is the minimum value of feature, Xmax is the maximum value of the feature and Xnorm is the normalized value of feature

Regarding categorical features, one hot encoding process was used. This approach is a technique used to convert categorical variables into a numerical format to ensure that each category within a categorical variable is represented as a separate binary column, and only one of these columns has a value of 1 for each observation [47]. This approach is especially useful for algorithms that require numerical input data. It allows categorical variables to be included in models without assuming any ordinal relationship between categories [47].

For analysis and modeling, we used three approaches: (i) traditional (frequentist model), (ii) intermediate (regression with regularization models), and (iii) machine learning, which are described below.

A. Traditional Approach: Generalized Linear Model

The Generalized Linear Model (GLM) is an extension of the general linear model that allows the specification of models whose response variable follows different normal distributions [48–50]. GLM models allow for efficient prediction of agricultural crop yields, but in many cases they have been misused or misinterpreted [51]. Thus, the implementation of the GLM methodology is proposed with the objective of comparing strategies that enable an approximation in the process of predicting the dynamics of the variable in question (Suc-field, Suc, and Sac-%-cane) (Equations (3)–(5)). GLMs offer several advantages, such as the flexibility framework for modeling various types of data (continuous, binary, count, and categorical response variables). In addition, they incorporate the non-normal distributions, and data do not need to meet normality criteria. The other advantages are its high interpretability, great prediction, and low overdispersion [48–50]. Similarly, GMLs have some disadvantages. These include: i) assumption of linearity, ii) choice of link function and distribution, iii) limited scope, and iv) sensitivity to outliers [48–50].

Equation (3). Lineal predictor $\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi,}$ where η_i: variable to predict, β_0: intercepted with the Y axis, and β_1: regression coefficient of the predictor variable $x_{1i.}$

Equation (4). The link function describes how the mean, $E(Y_i) = \mu i$, depends on the linear predictor $g(\mu_i) = \eta_i$

Equation (5). a variance function that describes how the variance, $var(Y_i)$ depends on the mean

$var(Y_i) = \varphi V(\mu)$, where the dispersion parameter φ is a constant.

For the GLM methodology, the Google Colab environment was used to develop Python code using the *statsmodel* library. *Statsmodel* is a library designed for statistical data analysis and statistical modeling. Here, the variables Suc field, Suc, and Suc % Cane were fit to a GLM using a Gaussian distribution with link identity function. Data was loaded from the cleaned and consolidated databases described above, using firstly all the variables for the models. A split of the data set was performed for training and testing with a ratio of 80/20, respectively. From the models obtained, a p-value analysis was performed and new models were developed using only characteristics with p-values less than 0.05, thus reducing the number of variables used in the models and maintaining the accuracy of the models, as assessed by R^2 (determination coefficient) and MAE (mean absolute error). The method of variable elimination was a stepwise selection, in particular backward elimination.

The new set of models uses only variables whose p-value was less than 0.05 in the previous model. Using a k-Fold with 10 folds, we evaluated our models considering the MAE metric and the R^2 coefficient. The mean and standard deviation of the results obtained during the k-Fold iterations were evaluated.

B. *Intermediate Approach: least absolute shrinkage and selection operator*-LASSO

LASSO regression (least absolute shrinkage and selection operator) regression is a method used to identify important variables and their corresponding regression coefficients to minimize prediction errors in a linear regression model [52,53] (Equation (6)). This is achieved by applying a constraint on the coefficients, known as the regularization parameter (λ or α) [52,53] (Equation (6)). This constraint helps to control the complexity of the model and mitigate overfitting issues. Variables with a coefficient of zero are considered to be not important for making predictions [52,53]. The choice of the regularization parameter is typically determined by hyperparameter tuning using techniques such as k-Fold cross-validation [52,53]. LASSO regression has been shown to outperform standard methods in certain cases by effectively reducing overfitting while using the entire data set for validation [52,53].

Furthermore, LASSO presents advantages such as automatic variable selection methods; the regularization parameter (λ), that helps control the complexity of the model; sparse solutions, that mean it reduces the number of features used in the final model and handles multicollinearity that helps stabilize coefficient estimates and mitigate the impact of multicollinearity on model performance [52,53]. Regarding the disadvantages, the arbitrary selection of λ, can lead to under- or over-fitting of the model and setting coefficients to zero can lead to loss of information, especially if predictors with small coefficients are relevant for prediction and can therefore introduce bias, which is particularly relevant when the true coefficients are large [52,53].

Equation (6). $L\ lasso\ (\beta) = \sum_{i=1}^{n} (yi - xi^T \beta) + \lambda \sum_{J=1}^{M} |\beta j|$ where, β:_ the coefficients parameters, x_i^T: are the independent variables (features) and λ: the regularization parameter

For the implementation of the LASSO model, the *scikit-learn* library was used. From the results obtained when training the models with the two datasets used, an analysis was done based on the evaluated metrics (R^2 and MAE). Additionally, the regularization parameter was optimized. In this case, no feature selection is performed in addition to the one performed automatically by the LASSO model.

C. Advantage approach: Machine Learning

Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on the development of algorithms and models that enable computers to learn from predictions and make decisions based on data [54–56]. Instead of being explicitly programmed to perform specific tasks, ML algorithms use data to iteratively improve their performance on a given task [54–56]. ML presents multiple advantages, for example it can efficiently process large volumes of data, automate repetitive and rule-based tasks, and adapt to new data and generalize patterns; moreover, models can improve over time as they are fed with more data and receive feedback, and can identify complex and nonlinear patterns in data that may go unnoticed by traditional analytical methods [54–56]. On the other hand, ML presents disadvantages as it requires large amounts of training data to produce accurate results, it can be difficult to interpret the models and explain how they arrived at a particular prediction or decision, models require regular maintenance and updating to keep their performance optimal as data, and some ML algorithms, especially the more complex and computationally intensive ones, may require a large amount of computational resources and runtime [54–56].

In our specific case, we applied auto machine learning (Auto-ML) algorithms. Auto-ML refers to tools that automate the process of an ML workflow, offering the same capabilities as normal ML, without explicit programming knowledge [57]. The goal of AutoML is to reduce human intervention in data pre-processing, feature selection, and algorithm selection, thus automating machine learning [58]. AutoML is increasingly being used in scientific research areas [57,58]. This trend supports the objective of its implementation in the processing of sucrose data, which obeys patterns with various distributions such as weather and soil and, those that occur daily according to the area planted. The Auto-ML method evaluates over 100 models simultaneously, performing an automatic process of selection and optimization of hyperparameters. This results in the selection of a set of models based on multiple parameters evaluating the behavior of the models, thus avoiding overestimation and optimizing both time and processes [57,58].

In our case, we use the *PyCaret* library to create the models. PyCaret is a Python ML library that simplifies model development [59]. It provides an easy-to-use interface to perform tasks such as data loading and preparation, feature selection, model comparison, and optimization, all in a single line of code [59]. PyCaret automates ML workflow and allows users to leverage best practices for faster and more efficient results [59].

In this case, the same preprocessing was performed as that carried out in the other approaches with a 70/20/10 split of training, validation, and testing. A list of different models evaluated with different metrics was obtained, as well as a list of the 10 variables that each model considers most important in predicting the corresponding target variable. Based on the best three models obtained for each variable sorted by R^2 score, it was decided to create a mixed model [60] using an averaging rule, that is, the outputs of the three models were averaged to obtain a result. In addition, the assembly models were created and evaluated using 10k-fold with the best three models for the same approach (example decision tree approach). Mixed models were trained in the training and validation set to finally see their performance in the test set.

The importance of a feature to fit an ML model was calculated by averaging the impurity increments of all splits in which the feature was used as a decision variable. The impurity was calculated using the Gini impurity, which is based on the difference between the proportions of classes in a node [61]. In addition, the importance of a feature is calculated as the reduction in the average loss when that feature is used to split a tree. This reduction is calculated as the difference between the loss without the feature and the loss with the feature [62].

3. Results

In this manuscript, we present a protocol outlining the application of data science phases to analyze the relationships with sucrose levels in sugarcane crops to optimize sugar production. Our approach provides a structured framework for collecting, processing, analyzing, and interpreting data, ultimately enhancing decision-making processes in the agricultural industry. Using data science techniques, we aim to uncover insights that can improve the efficiency and yield of sugar production, contributing to the advancement of sustainable agricultural practices.

3.1. Phases I-II. Data acquisition, cleaning, and management phases and increase the size

In the first parts of the application of data science methodologies, a pre-processing process was performed, which consisted of reviewing the databases using graphic illustrations that allowed the filtering of inconsistent variables, ensuring the detection and removal of outliers or missing values (Fig. 2A and B).

In the climate database, the variables analyzed did not have missing data or zero values, so the cleaning process was continued (Fig. 2 A). Furthermore, the dynamics of the quantitative variables and missing data of the climate was observed (Fig. 2 E). As a particular case, the variation of the average wind presents atypical data (Fig. 2 E). This trend makes it possible to infer a source of error that may later affect the analysis of the variables in question. An indirect correlation of -0.73 was found for the variables minimum relative humidity and maximum temperature, and -0.77 for degree days and minimum relative humidity (Fig. 2C). In the case of direct correlation, a correlation of 0.94 was found for minimum relative humidity and average relative humidity and 0.86 for minimum wind speed and maximum wind speed (Fig. 2C). Based on the previous analysis, the mean relative humidity and mean temperature were removed due to high correlation, based on the criterion of eliminated autocorrelated variables with correlation coefficient >0.8 and understanding that the average climatic variables may have less impact on the physiology and biochemistry of sugar plants with respect to maximums and minimums.

In the case of variables associated with sugar mills and farms dedicated to panela production, it became evident that the variables related to sugarcane maturation, yield, and agroecological conditions had low null values (inconsistent or missing values) (Fig. 2 B). In Fig. 2 D, the correlations between these variables can be observed. When the behavior of the numerical variables related to mills and farms was visualized, it was found that the agronomic variables related to growth and development appear to have outliers (Fig. 3 B, C, D and E). Performing correlation analysis, it was seen that the variables related to the yield components, such as sucrose, presented a high correlation, as well as the variables related to the yield, therefore it was decided to eliminate redundant variables (Yield, Theorical yield and Brix) (Fig. 2C and 3 A). Finally, the target variables and the final sugar mill database were extracted. The link variables were increased to obtain a data set with variables of soil, climate, sugar mill and farms (Figs. 2 and 3 F and G, and Appendix 1). This data set included 16,655 samples. However, many of these samples did not contain relevant soil information (Fig. 2 and Appendix 1).

In the last part of Phases I and II, it was decided to create two new datasets from the above-mentioned ones (edaphic, climatic, and agronomic). One of these new datasets contained all variables but eliminated specific samples with null data (inconsistent or missing values), and the other eliminated all the variables with null data (Figs. 2 and 3). The size of these data sets were 7674 samples with 41 characteristics and 16,655 samples with 25 characteristics, respectively. The missing data related to the maturation process were

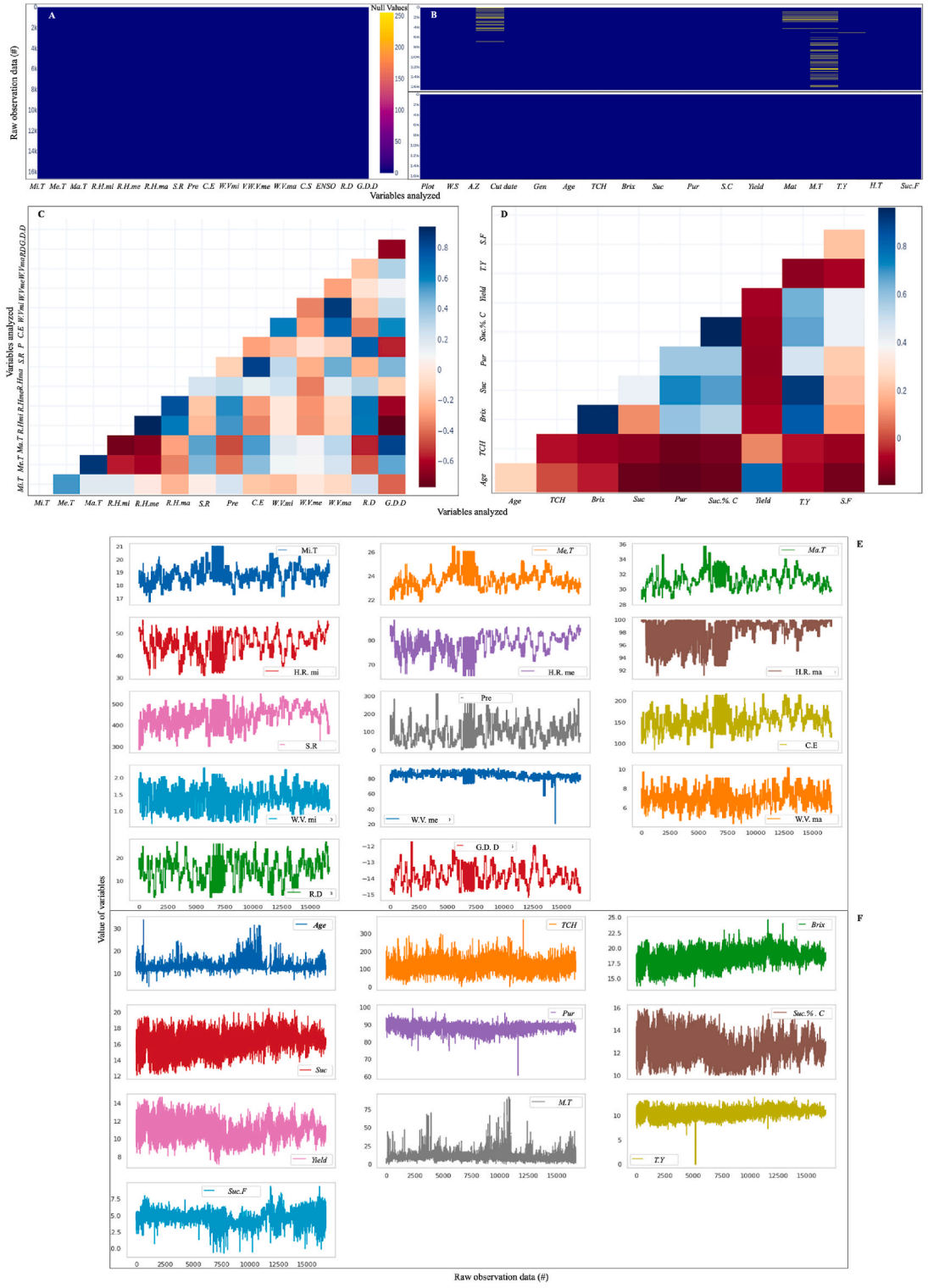


Fig. 2. Data science graph representation of variables associated with climatic and sugarcane production cycle in data acquisition, cleaning, gestion and basic tools to visualization.

A-B: Graph corresponding to data cleaning analysis by variable taking into account the number of rows with null or zero data to climate and sugar mills variables. C-D: Graph corresponding to correlation analysis of numerical climate and sugar mills variables. E-F: Graph corresponding to the analysis of climate and sugar mills variable dynamics. Mi.T: minimum temperature. Me.T: mean temperature. Ma.T: maximum temperature. R.H.mi: minimum relative humidity. R.H.me: mean relative humidity. R.H.ma: maximum relative humidity. S.R: solar radiation. Pre: precipitation. C.E:

cyclical evapotranspiration. W.V.mi: minimum wind velocity. W.V.V. me: variation mean wind velocity. W.V.ma: maximum wind velocity. C.S: climatic season. ENSO: ENSO El Nino. R.D: rains days. G.D.D: growing degree days. W.S: weather station. A.Z: agroecological zone. Gen: genotype. TCG: Tons sugarcane per hectare. Suc: sucrose. Pur: purity. S.C: Sucrose cane. Mat: ripening. M.T: ripening time. T.Y: Theorical yield. H.T: harvest type. Suc.F: sucrose field.

analyzed more closely and it was determined that the null values correspond to the fact that for these samples no maturation product was applied; therefore, we made the appropriate correction (Fig. 2 B, C, and F).

In the soil database, it was found that some variables have a huge number of zeros (some were inconsequential and others 0 real) (Appendix 1). As for the number of null data, it was observed that the variables prof and texture have many of these. It was decided to eliminate some variables related to the characteristics of the soil. Other variables eliminated professional criteria related to some aspects of the soil, as well as variables corresponding to the measurements to be performed to obtain soil data, because these do not contain important information for the data analysis (Appendix 1, and Fig. 3 G).

It was evident that the dynamics of the variables related to the soil measurements were similar (Appendix 1). In the case of the variables of chemical properties, the scale differs, but the dynamics of the variables are similar (Appendix 1). Later, we decided to make a correlation graph evaluating the relevance of excluding (based on the high autocorrelation values and the agronomic sense about its potential relationship with sucrose) one of these two variables (Appendix 1). From this, it was determined that the variable related to phosphorus (available at P) is eliminated because of the inconveniences that may be presented for its later interpretation because of the absence of bibliographic support for the variable. The same procedure as described above was carried out for variables related to the physicochemical properties of the soil, finding that those related to chemical properties present a similar dynamic (Appendix 1).

Following the approach used for the variables analyzed previously, a correlation graph was created to determine the highly correlated variables that could generate redundancy in subsequent analyses (Appendix 1). Qualitative variables were analyzed, and variables were extracted according to the analysis performed so far. A new analysis of the dynamics of the variables and the correlation between the variables was performed.

Using the previous data set as the basis, additional analysis of the variable dynamics was performed. It was observed that certain variables may contain outliers (Appendix 1), and variations in scale were observed across features. Furthermore, correlation analysis did not reveal significant correlations among variables (Appendix 1). Subsequently, the target variables were identified, and the final soil database was compiled.

3.2. Phase III. Data visualization tools

Given that three approaches were applied for data analysis, a crucial strategy to initiate the modeling process is to prepare and statistically review the data beforehand. An exploratory analysis was performed to determine the type of data available (continuous, discrete, nominal, or multinomial) and therefore what modifications should be made to enable the use of the data in the execution of the three approaches to data analysis in phase four (Fig. 3A–G and Appendix 1). A visualization of the variables was initially performed using histograms to see if any contained outliers and to observe their distribution (Fig. 3A–G, and Appendix 1). There were many zeros in characteristics related to agronomic characteristics and chemical properties, and it appeared that our target variables were normally distributed (Fig. 3F and G).

After conducting a correlation analysis and examining the target variables, it was observed that a variable related to performance components exhibited a positive correlation with the target variables (see Fig. 2C and D). Consequently, it was decided to plot each of these variables (see Fig. 2 E). The analysis revealed that the purity of the variable shows a weak positive correlation with the variables Sucrose and Sucrose % Cane, although its relationship with the variable Sucrose field is less evident (Fig. 3 A)

An analysis of the distribution of categorical variables was carried out (Fig. 3B and C and D). This analysis revealed that there is a higher amount of data available for the first dry and third transitional seasons. In addition, manual harvesting was found to have the highest amount of data of the different types of harvesting methods. Furthermore, regarding the application of ripening agents, it was observed that the application of maturates accounted for much of the data, indicating that most samples included in the data set had a ripening agent applied, with maturates being the most commonly used type (Fig. 3B and C and D).

Box and whisker plots were made for the variables analyzed above with respect to the target variables to determine their effect. In the case of the sucrose variable, the variables with the greatest effect on the data were found to be the third transition season, manual harvest, application of ripening agents, and the El Niño phenomenon (Fig. 3 B, C, and D). Similarly, for the Sucrose Field variable, the variables with the greatest effect are the third transition season, application of ripening agents, type of ripening agent used, and the Niño phenomenon (Fig. 3, B, C, and D). Regarding the percentage of sucrose cane, the variables with the greatest impact are the third transitional season, manual harvesting, application of ripening agents and the El Niño phenomenon (Fig. 3, B, C, and D).

After the data analysis process was completed, a normality test was performed (Fig. 3 E). It was observed that the distribution of our target variables was not Gaussian. Therefore, a normalization method known as Normal Quantile Transformation (NQT) was implemented using the *scikit-learn* library. The comparison between the distributions can be seen in Fig. 3, panel E, where it became evident that the variables achieved a distribution that passed the normality test and were then scaled to a range of -5 and 5, with a mean of 0 and a standard deviation of 1.

The analysis revealed patterns that guided us to improve the treatment of the data. We noticed that the data exhibited different scales, prompting us to standardize them (see Fig. 2, panel E). Additionally, there was a class imbalance in the use of ripening agents for categorical variables (see Fig. 3, panels F and G). However, the box and whisker plots indicated that the crops treated with ripening

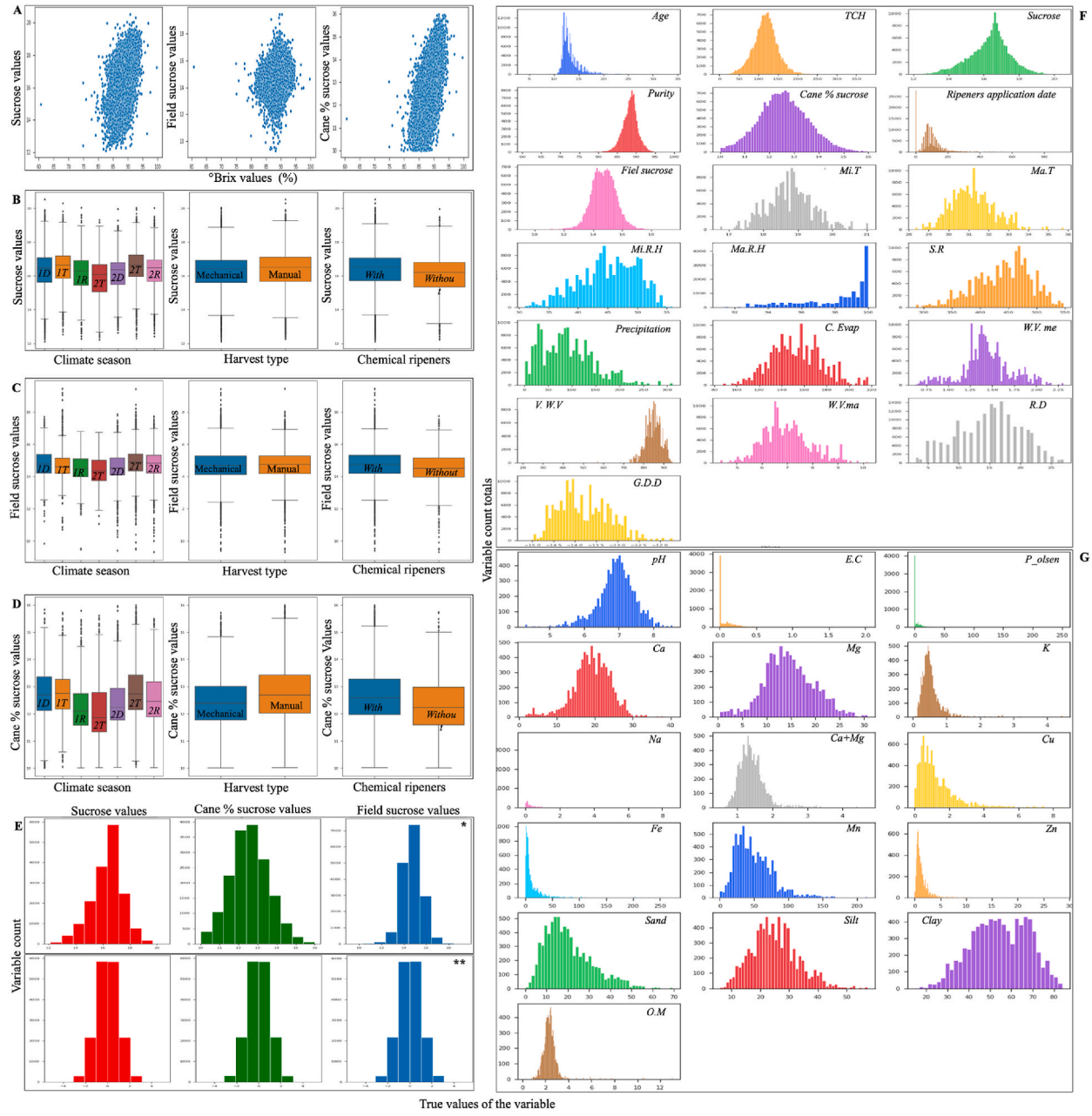


Fig. 3. Data science visualization and basics analysis tools associated with climatic, sugarcane production cycle and soil variables.
A: Dispersion plot of autocorrelated yield sugar cane components variables. B, C y D: Box-and-whisker plot of the categorical variables associated with sucrose variables. F-G: Histogram of climatic, sugarcane production cycle and soil variables selected to analysis (phase IV)) for their quality (cleanliness), number of data (absence of null data), and not autocorrelated. E: Histograms of our target variables before and after of normal transformation. * Raw data. ** Fit to the normal distribution of variables. nD: dry weather condition. nT: transition weather condition. nR: rainy weather condition. n: represents the number of the climatic season of the year, indicating whether it falls within the first period (January–June) or the second period (July–December). P: phosphorus. TCH: Tons sugarcane per hectare. E.C: electric conductivity. O.M: organic matter. Mi.T: minimum temperature. Ma.T: maximum temperature. Mi. R.H: minimum relative humidity. Ma.R.H: maximum relative humidity. S.R: solar radiation. C.Evap: cyclical evapotranspiration. V. W.V: variation mean wind velocity. Ma.W.V maximum wind velocity. R.D: rains days. G.D.D: growing degree days.

agents had higher levels of sucrose (see Fig. 3, panels F and G). Furthermore, we observed that there were no categorical variables with high cardinality, suggesting that a hot encoding could be used. After conducting the exploratory analysis and making the necessary adjustments, the three approaches were executed and their suitability in the modeling process was evaluated (see Fig. 3A–G).

3.3. Phase *IV. Analysis and modeling phase*

In this section, as explained in the Materials and Methods section, our objective was not to generate a forecasting model for variables associated with sucrose in sugarcane or to compare the three evaluated approaches. The objective of this section was to exemplify the fourth phase of DS, for which we selected three methodological approaches with a contrast in moderation from their assumptions, fitting methods, and model evaluation (Fig. 1). Similarly, this approach to modeling based on a responsible step-by-step process using approaches and algorithms allows us to approximate the behavior of sucrose in sugarcane based on predictor variables such as soil, climate, and agronomic factors.

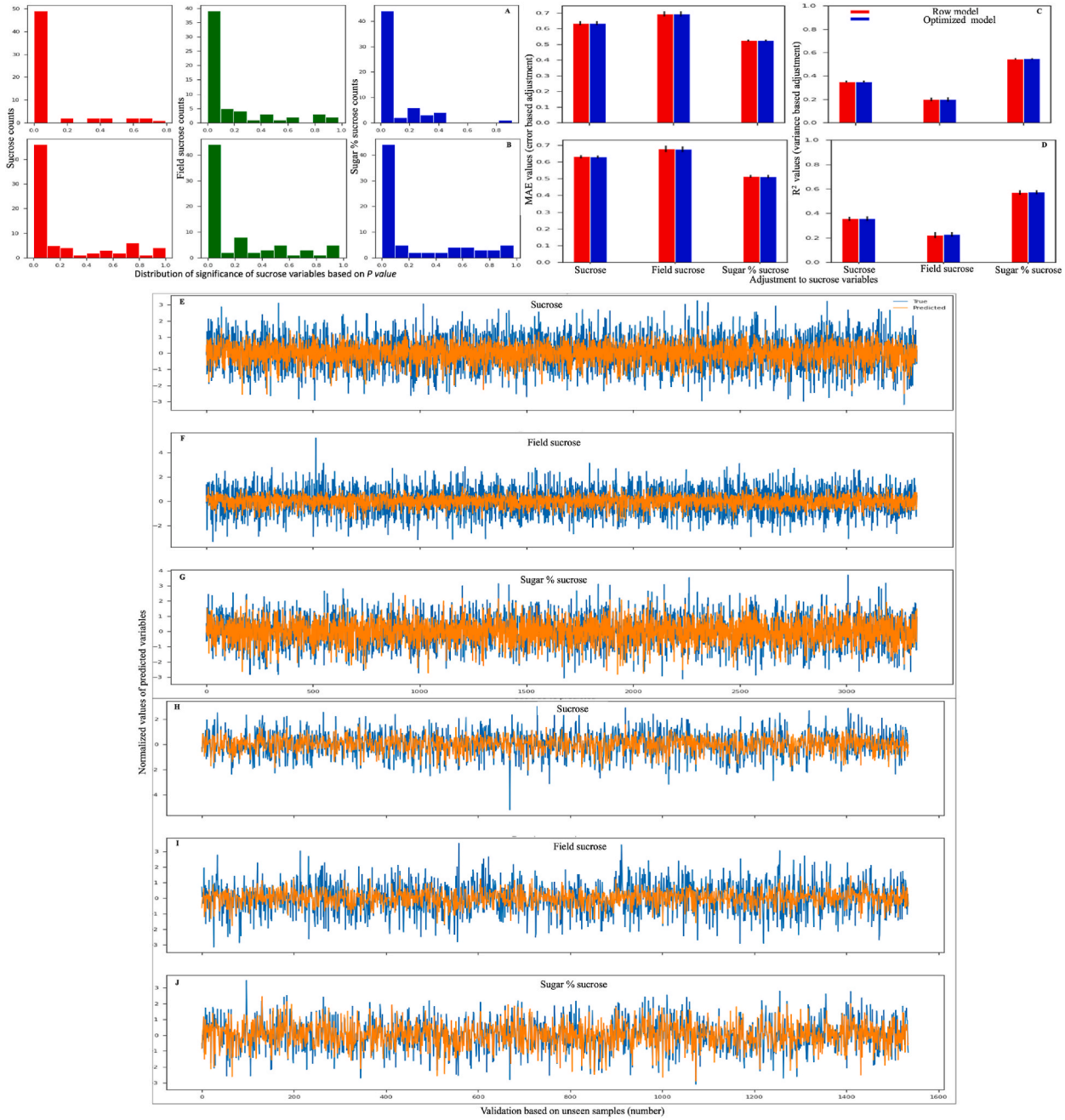


Fig. 4. Statistical and fit parameters associated with data analysis under the traditional approach using generalized linear models due to their ability to predict variables associated with sucrose in sugarcane.
A-B: *P-value* histogram for non soil and soil dataset. C-D: Statistical Metrics for non soil and soil dataset. **F-G-H:** True vs Predicted values for non soil test set to fit the sucrose variables. H-I-J: True vs Predicted values for soil test set to fit the sucrose variables.

A. Traditional Approach

Using the data to fit and train the GLM models, it was possible to observe that, in the data set without soil variables, there were fewer features with *p-values* greater than 0.05 compared to the data set with soil variables (Fig. 4A and B).

Similarly, models that only used variables with p-values less than 0.05 achieved similar metrics (Fig. 4C and D). For the models

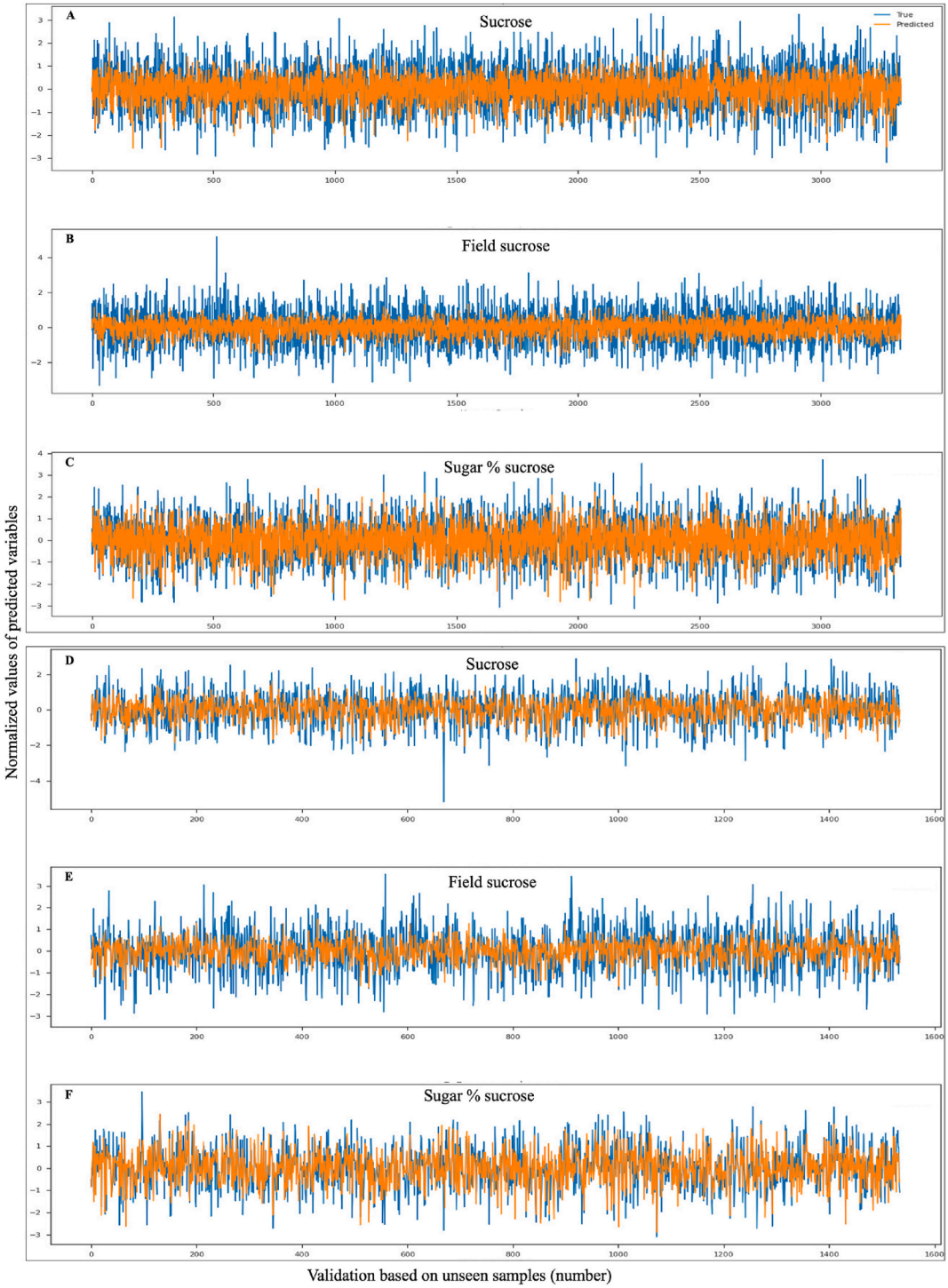


Fig. 5. Fit parameters associated with data analysis under the intermediate approach using LASSO models due to their ability to predict variables associated with sucrose in sugarcane.

applied to the database without soil variables, an MAE of 0.63-0.69-0.52 and an R^2 of 0.34-0.22-0.54 were obtained for the variables Sucrose, Sucrose field and Sucrose% Cane, respectively (Fig. 4C). In particular, the 10 % cane sucrose variable yielded the best results among these predictions.

On the contrary, for the models that use the database with soil variables, an MAE of 0.64-0.68-0.50 and an R^2 of 0.33-0.21-0.56 were obtained for the variables Sucrose, Sucrose field, and Sucrose% Cane, respectively (Fig. 3 D). Models incorporating soil variables were found to perform better for the Sucrose field and Sucrose % Cane variables, while models excluding soil variables performed better for the Sucrose variable (Fig. 3C and D). These evaluations were conducted on the test set.

Fig. 4E–J depict the values obtained with the GLM models compared to the actual values of the test set. It was noted that the models for the sucrose field variable did not fit the data as well as expected. Using GLM, which identifies the variables with the greatest impact on our objectives, it was observed that the soil variables did not exert a significant influence. However, it is essential not to discard them entirely, as in some cases they worsened the predictive results of the models (Fig. 4E–J).

B. Intermediate Approach

The LASSO models were performed, and the hyperparameters were fitted. It was observed that, for all models, the best regularization parameter was 0, indicating that the trained data presented a behavior equivalent to a linear regression, and no feature selection was performed. The results of the intermediate approach were not very helpful since the regularization parameter resulted in 0, so no feature selection was performed. However, the models' metrics were acceptable compared to those obtained in the previous approach (Figs. 4 and 5).

Performance evaluation was carried out on the two test data sets. In the database without soil variables, an MAE of 0.63-0.69-0.52 and R^2 of 0.34-0.22-0.54 were obtained for the variables Sucrose, Sucrose field and Sucrose % Cane, respectively (Fig. 5 A, B, and C). For models using the database that include soil variables, the MAE was 0.64-0.68-0.50, and the R^2 values were 0.33-0.21-0.56 for the Sucrose variables, the Sucrose field, and the Sucrose% Cane variables, respectively (Fig. 5, A, B, C, D, E, and F). The results obtained by implementing this model were similar to those obtained in the GLM, as regularization did not prove to be an effective strategy (Figs. 4 and 5).

C. Machine Learning Approach

The Auto-ML models implemented by the *Pycaret* library could be trained and the results could be compared. In the case of the data analyzed, the models were ranked based on their coefficient R^2 and other multiple metrics. The best performing models for the dataset both with and without soil variables were found to be the Light Gradient Boosting Machine and the Extra Trees Regressor (refer to Tables 1–3). Furthermore, a ranking of the variables of the highest importance was obtained for each of the best performing models (Fig. 6, panels A–F). Furthermore, it was observed that the ensemble models showed a slight improvement over the individual models.

In the Extra Trees Regressor models, the importance of a feature was determined by averaging the impurity increments of all splits in which the feature was used as a decision variable. Conversely, in the Light Gradient Boosting Machine model, feature importance was calculated based on the reduction in average loss when that feature was used to split a tree. This reduction was computed as the difference between the loss without the feature and the loss using the feature. Across all models, the variable 'Purity' emerged as the most important, followed by maximum relative humidity, TCH, maximum and minimum temperature, age of sugar, solar radiation, wind velocity, ENSO El Niño phenomenon, and application of ripeners, among others. Interestingly, soil variables were found to be unimportant in the construction of any model (Fig. 6A–F).

For the models derived from the database without soil variables, MAE values of 0.49-0.55-0.41 and R^2 values of 0.59, 0.44, and 0.7 were obtained for the variables Sucrose, Sucrose Field, and Sucrose % Cane, respectively (Tables 1–3, and Fig. 6 A, B, C, G, H and I).

Table 1
Statistical and fit parameters associated with model fit to sucrose under the machine learning approach based on AutoML.

Model	MAE	MSE	RMSE	R^2	RMSLE	MAPE	TT (Sec)
Light Gradient Boosting Machine	0.5215	0.4532	0.6727	0.5521	0.3181	2.3360	0.510
Extra Trees Regressor	0.5304	0.4758	0.6891	0.5298	0.3193	2.3954	5.094
Random Forest Regressor	0.5350	0.4777	0.6907	0.5279	0.3244	2.4435	9.894
Gradient Forest Regressor	0.5700	0.5298	0.7274	0.4764	0.3436	2.4347	2.756
Ridge Regression	0.6382	0.6507	0.8062	0.3569	0.3727	2.6424	0.0.28
Bayesian Ridge	0.6383	0.6508	0.8062	0.3568	0.3731	2.6346	0.052
Huber Regressor	0.6378	0.6534	0.8079	0.3540	0.3699	2.6085	0.291
Linear Regression	0.6394	0.6536	0.8081	0.3538	0.3717	2.7247	0.330
AdaBoost Regressor	0.6614	0.6844	0.8266	0.3241	0.3955	2.3432	1.231
Orthogonal Matching Pursuit	0.6761	0.7279	0.8526	0.2808	0.3943	2.3957	0.0018
K Neighbors Regressor	0.6955	0.8004	0.8941	0.2075	0.3800	2.4391	0.318
Decision Tree Regressor	0.7500	0.9636	0.9813	0.0445	0.3707	4.0763	0.175
Lasso Regression	0.8031	1.0136	1.0064	−0.0019	0.6211	1.0066	0.030
Elastic Net	0.8031	1.0136	1.0064	−0.0019	0.6211	1.0066	0.029

MAE: Mean absolute error. MSE: Mean squared error. RMSE: The root mean squared error. R^2: Coefficient of determination. RMSLE: Logarithmic error of root mean squared. MAPE: The mean absolute percentage error. TT: Computational time.

Table 2
Table 1. Statistical and fit parameters associated with model fit to Field sucrose under the machine learning approach based on AutoML.

Model	MAE	MSE	RMSE	R^2	RMSLE	MAPE	TT (Sec)
Extra Trees Regressor	0.5705	6.149000e^{-01}	0.7835	3.871000e^{-01}	0.3437	3.0653	4.355
Random Forest Regressor	0.5963	6.397000e^{-01}	0.7995	3.623000e^{-01}	0.3657	2.8187	8.799
Light Gradient Boosting Machine	0.6007	6.484000e^{-01}	0.8048	3.538000e^{-01}	0.3644	3.1671	0.507
Gradient Forest Regressor	0.6300	6.882000e^{-01}	0.8292	3.141000e^{-01}	0.3883	2.7619	3.257
Ridge Regression	0.6791	7.847000e^{-01}	0.8853	2.183000e^{-01}	0.4121	2.6653	0.029
Bayesian Ridge	0.6793	7.85000e^{-01}	0.8855	2.180000e^{-01}	0.4147	2.6271	0.081
Linear Regression	0.6798	7.866000e^{-01}	0.8864	2.164000e^{-01}	0.4111	2.6609	0.374
Huber Regressor	0.6776	7.877000e^{-01}	0.8870	2.154000e^{-01}	0.4075	2.8279	0.645
K Neighbors Regressor	0.6770	8.007000e^{-01}	0.8939	2.031000e^{-01}	0.3773	3.8644	0.345
Orthogonal Matching Pursuit	0.7068	8.322000e^{-01}	0.9119	1.707000e^{-01}	0.4282	2.6215	0,036
AdaBoost Regressor	0.7268	8.526000e^{-01}	0.9231	1.503000e^{-01}	0.4695	2.0359	1.172
Lasso Regression	0.7974	1.006100+e^{00}	1.0028	−2.800000e^{-03}	0.6167	1.0133	0.033
Elastic Net	0.7974	1.006100+e^{00}	1.0028	−2.800000e^{-03}	0.6167	1.0133	0.032
Lasso Least Angle Regression	0.7974	1.006100+e^{00}	1.0028	−2.800000e^{-03}	0.6167	1.0133	0.047

MAE: Mean absolute error. MSE: Mean squared error. RMSE: The root mean squared error. R^2: Coefficient of determination. RMSLE: Logarithmic error of root mean squared. MAPE: The mean absolute percentage error. TT: Computational time.

Table 3
Statistical and fit parameters associated with model fit to sugar % sucrose under the machine learning approach based on AutoML.

Model	MAE	MSE	RMSE	R^2	RMSLE	MAPE	TT (Sec)
Extra Trees Regressor	0.4200	3.0920000 e^{-01}	0.5555	6.906000 e^{-01}	0.2703	2.440	4.439
Light Gradient Boosting Machine	0.4278	3.150000 e^{-01}	0.5608	6.846000 e^{-01}	0.2728	2.491	0.523
Random Forest Regressor	0.4363	3.318000 e^{-01}	0.5757	6.677000 e^{-01}	0.2788	2.416	9.462
Gradient Forest Regressor	0.4676	3.640000 e^{-01}	0.6028	6.358000 e^{-01}	0.2949	2.437	3.269
Ridge Regression	0.5167	4.358000 e^{-01}	0.6597	5.634000 e^{-01}	0.3154	2.820	0.028
Bayesian Ridge	0.5166	4.358000 e^{-01}	0.6597	5.634000 e^{-01}	0.3153	2.831	0.142
Linear Regression	0.5169	4.364000 e^{-01}	0.6602	5.627000 e−01	0.3151	2.881	0.354
Huber Regressor	0.5165	4.369000 e^{-01}	0.6606	5.621000 e^{-01}	0.3145	2.914	0.429
Orthogonal Matching Pursuit	0.5465	4.812000 e^{-01}	0.6936	5.175000 e^{-01}	0.3291	2.993	0.078
AdaBoost Regressor	0.5595	4.910000 e^{-01}	0.7003	5.086000 e^{-01}	0.3553	2.476	1.193
K Neighbors Regressor	0.5997	6.160000e^{-01}	0.7846	3.827000 e^{-01}	0.3487	2.721	3.28
Decision Tree Regressor	0.6230	6.749000 e^{-01}	0.8209	3.244000 e^{-01}	0.3488	3.642	0.169
Passive Aggressive Regressor	0.7103	8.020000e^{-01}	0.8802	1.837000 e^{-01}	0.3758	4.314	0.123
Lasso Regression	0.7970	1.004100e^{+00}	1.0013	−2.300000e^{-03}	0.6177	1.008700 e+00	0.028

MAE: Mean absolute error. MSE: Mean squared error. RMSE: The root mean squared error. R^2: Coefficient of determination. RMSLE: Logarithmic error of root mean squared. MAPE: The mean absolute percentage error. TT: Computational time.

These results indicate that the ML approach model yielded better results than traditional and intermediate approaches (Fig. 6C). Similarly, for models using soil variables, MAE values of 0.52-0.54-0.4 and R^2 values of 0.55-0.45-0.69 were obtained for the variables Sucrose, Sucrose Field and Sucrose % Cane, respectively (Fig. 6 D, E, F, J, K and L). Fig. 6C and D illustrate that the ML models fit better to the data; however, the Sucrose field variable remains the most complex. Furthermore, our results indicate that, in general, soil variables were not highly predictive of Sucrose, Sucrose Field, and Sucrose % Cane.

4. Discussion

The execution of all the processes related to data science, consisting of a long process of adjustment, modification, and reorganization of the primary databases, were necessary, these being the basis of the processing. The main sources of errors were analyzed. Upon examining the database, it became evident that its consolidation had not been proposed with the intention of post-processing to derive actionable insights based on historical trends. Instead, the primary objective appeared to be to store information for record-keeping purposes. It was concluded that daily improvements to databases are necessary to improve agricultural production yield and profits, aligning with technological advances. It is well recognized that the sheer volume of collected information can pose challenges in terms of storage and processing, due to the storage and processing capacity of the servers typically available to these companies. However, there are alternatives in the market for acquiring high-capacity storage servers that can address these issues.

As mentioned in previous works [11], the use of large datasets and digital tools to collect, aggregate, and analyze information (big data) has the potential to redefine long-standing relationships between stakeholders in food and agriculture, such as between farmers and agricultural corporations. Proponents of big data promise unprecedented levels of precision, information storage, processing, and analysis that were previously unattainable due to technological constraints. Consequently, an effective organization of information is crucial to maximize the potential of data processing and analysis.

Big data presents significant opportunities to address the challenges faced by modern societies, including the diverse needs of consumers, financial analysts, marketing agents, producers, and decision-makers [2,4]. This suggests a transition to an information

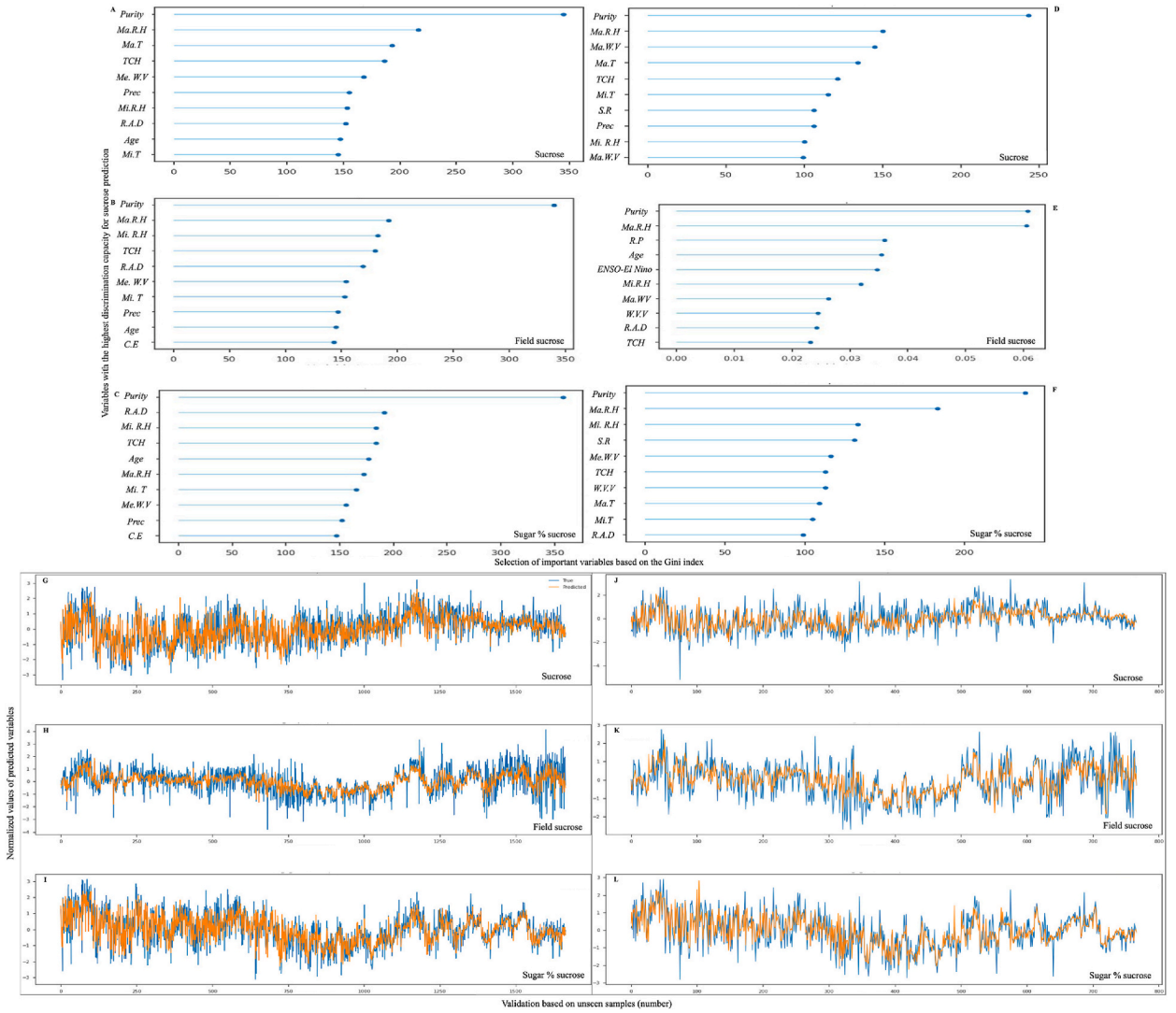


Fig. 6. Statistical and fit parameters associated with data analysis under the machine learning approach using AutoML due to their ability to predict variables associated with sucrose in sugarcane.
A-B-C-G-H-I: variable importance and True vs Predicted values for non-soil test set to fit the sucrose variables. D-E-F-J-K-L: variable importance and true vs. predicted values for soil test set to fit the sucrose variables. Model fit to sucrose: Light Gradient Boosting Machine (lightgbm). Model fit to field sucrose: Extra Trees Regressor. Model fit to sugar % sucrose: Extra Trees Regressor.

revolution within the agricultural sector, where sensor technology and data analytics from other industries are increasingly being applied to agricultural contexts.

As demonstrated in this work, a variety of technological advances have created opportunities for big data and its use in decision making [1]. In many cases, computational capacity both in terms of speed and volume allows for novel analyses that were previously not possible. It is now possible to conduct an analysis of large volumes of data (such as weather data) and use them for actionable decision making. Interestingly, data from multiple sources, including public data, machine and sensor data, and other privately held data, are often integrated. Consequently, it should be mentioned that the important objective of this type of analysis is to strengthen the databases in order to have as much information as possible related to climate, soils, production, and other factors, allowing decisions to be made with greater certainty and risk reduction.

Another aspect worthy of mention is related to access to databases. As presented by Coble et al. [4] much of the useful big data produced today is in the hands of the private sector. They suggest that the landscape of public and private farm data is likely to change and that access to data for research will be a critical issue. In many cases, the data are held by a variety of firms ranging from small individual farms to large corporate input suppliers. Meanwhile, many authors argue that agricultural research must go beyond Fisher's experimental design and use analytical techniques capable of learning from machine and sensor data, that is, it must rely on observation of farm production data along with data from controlled experiments, thus promoting the continuous development of the

agricultural sector together with technology and data science, as evidenced by the work presented here [4,6,11].

The results obtained in our study indicate that DS, using the correct approach, can generate informed solutions. As is recognized, sugarcane production is a complex process that involves a multitude of factors, including weather conditions, soil quality, and production variables. To effectively manage and optimize sugarcane production, it is crucial to have a clear understanding of these variables and how they interact with each other. As noted, data visualization plays a crucial role in this process, helping to identify patterns, trends, and relationships that would otherwise be difficult to detect.

Data visualization allowed us to identify highly correlated variables that helped reduce the number of features used in the models and choose the best approaches to perform data pre-processing. Using data visualization tools and techniques, it could be possible to gain a better understanding of the data, leading to more informed decision making and better modeling results. We shall now discuss the results obtained for the different types of approach.

The models trained using the traditional approach allowed us to obtain the variables related to our target variables. From these results, it is important to highlight the absence of most of the soil variables, potassium and zinc being those that are relevant for the three target variables. Other variables of importance are the use of ripening agents, age, TCH, purity, and variables related to temperature and humidity. Although the results in terms of MAE and R^2 are not optimal, they serve as a baseline for the other approaches. On the other hand, the intermediate approach did not obtain additional information with the LASSO model since, when performing the hyperparameter tuning, the optimal regularization parameter was zero, which prevented us from performing feature selection. Moreover, the MAE and R^2 results are equal to those of the generalized linear models.

Finally, the use of the models obtained with the machine learning approach shows significant advantages compared to other approaches. The reason for this may be that machine learning tends to be more effective when dealing with complex datasets [63]. It is important to note the ease that Auto-ML provides for the creation and comparison of multiple models without the need to write large extensions of code. In addition to providing tools to analyze and optimize the models in a simple way, this allowed us to focus on the results and facilitate the machine learning workflow. An important point regarding the use of different ML models is the interpretability of the results obtained, although multiple techniques are being developed to improve this field [54,63]. In the present work, a classical approach was used by means of metrics obtained by the models based on decision trees (e.g. GINI), which allowed us to make a comparison with the results of the traditional approach of generalized linear models, noting that in most cases the results are very similar.

Based on the results obtained and the techniques and approaches applied, it was possible to conclude that ML techniques have proven to be a valuable tool in exploring sugarcane relationships, but the quality of the analysis is highly dependent on the size and quality of the data set. With a larger dataset, ML models can identify complex patterns and interactions that may not be apparent with a smaller dataset. This not only improves the accuracy and reliability of the analysis, but also improves the predictive power of the models. As a result, more information can be gleaned from the analysis, providing useful information to the sugarcane industry, and advancing the field of sugarcane production.

In addition to this, ML can help improve sugarcane production by leveraging weather, soil and milling variables to make more accurate predictions and more informed decisions. Our approach was to analyze causal relationships using a set of multiple variables that can help improve production, identify optimal planting times, and determine appropriate fertilization and irrigation strategies. It can also be applied to optimize the grinding process, improve efficiency, and reduce waste. By analyzing mill variables such as sugar extraction rates, throughput, and energy consumption, machine learning algorithms can identify areas for improvement and suggest optimal operating conditions. In general, ML has the potential to revolutionize the sugarcane industry by providing more accurate and informed decision-making tools.

As seen in this paper, ML has great potential to revolutionize the sugarcane sector. Data can be analyzed more efficiently and accurately, allowing researchers and farmers to gain valuable information that can improve crop yields, optimize production processes, and minimize environmental impact. One way to improve the work done is with the use of neural networks. Neural networks are an essential subset of ML algorithms that have gained great popularity in recent years for their ability to model complex relationships between variables. These networks can be trained to recognize patterns and make predictions based on input data, and their potential applications within the sugarcane industry are vast.

Our study focused on understanding and implementing the four fundamental phases of data science applied to agriculture, using the specific example of the components of sugarcane yield and saccharose. Our aim was to provide a practical implementation and validation of the data science toolset, which has great potential for agricultural applications. Through our work, we identified critical aspects that must be addressed for the successful implementation of data science tools in agriculture. These include the absence of standardized protocols for data acquisition, management, and database design, leading to high data cleaning costs and time. Furthermore, there is a lack of specific needs or questions from the agricultural sector regarding database queries and the absence of quantifiable indicators for evaluation, making real-time decision-making challenging. Furthermore, the lack of availability of large amounts of data, particularly freely accessible climate data, presents a significant limitation.

Therefore, we suggest that future research should focus on improving data acquisition and management protocols, and on incorporating highly informative variables obtained from freely accessible sources, such as synthetic and reanalysis climate data, spectral variables from remote sensors, and others. Finally, we emphasize the need for agricultural production systems to incorporate data science platforms and applications to visualize results, which can be used as evidence-based decision-making tools. By addressing these limitations and incorporating data science tools into agricultural practices, we can improve the efficiency and use of resourses in agricultural production.

5. Conclusions

The utility of this study lay in the use of data science tools to establish a protocol for processing the information generated daily in sugar mills, making it possible to understand the factors that influence their production. The protocol for the exploration and analysis of large databases, such as those found in the sugarcane industry, aims to reduce the number of potential variables that could introduce errors during processing. This is achieved through the application of cleaning processes and transformation operations provided by the Pandas library. Subsequently, exploratory analysis helps identify data types and guides the selection of appropriate methodologies. During the analysis and modeling phase, the implementation of three approaches proved particularly useful in determining the relevance of employing strategies such as ML. For the data set examined, AutoML emerged as a suitable data processing strategy for the sugarcane industry database, based on the results obtained from target variables such as Sucrose, Sucrose field, and Sucrose % Cane. Notably, the variable Sucrose field presented complexities in analysis.

CRediT authorship contribution statement

Laura Valentina Bautista-Romero: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Juan David Sánchez-Murcia:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation. **Joaquín Guillermo Ramírez-Gil:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Data availability statement

Data will be made available on request. For requesting data, please write to the corresponding author**.** For the specific case of the developed codes, and in pursuit of reproducibility and transparency of our work, the codes are deposited in an open-access repository (https://github.com/agrocompuepidemlab/Data-Science-for-Pattern-Recognition-in-Agricultural-)**.**

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

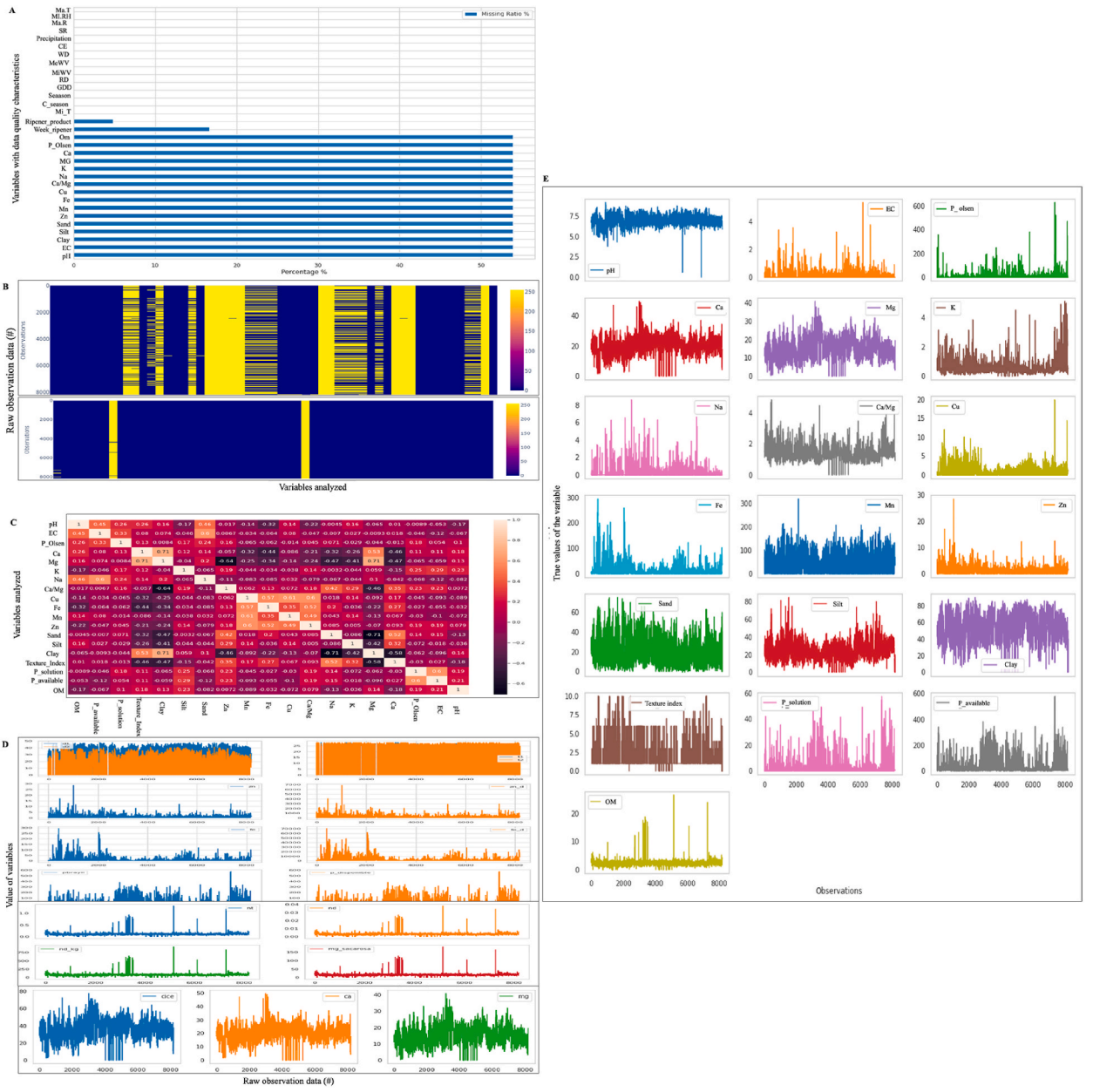
Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Edison David Serrano Cardenas for supporting the design and first version of the code to apply data analysis. Also, we want to thank the Geomatic and Meteorology pro-grams of Cenicaña program for helping and supper with part of the climatic data set. In addition, we thank the cane sector in Colombia for the support associated with the data with which the present investigation was carried out. In addition, the authors would like to thank the anonymous reviewers and editors for their valuable comments and suggestions.

Appendix 1. Data science tools for the management of soil variables associated with sugarcane production systems



A: Importance of variables for sucrose prediction. **B:** Data quality based on missing versus valid data. **C:** Correlogram for identifying autocorrelated variables. **D:** Temporal distribution of analyzed variables as a basis for assessing data quality. **E:** Temporal dynamics of variables and their relationship with the values of each variable.s

References

[1] F. Provost, T. Fawcett, Data science and its relationship to big data and data-driven decision making, Big Data 1 (2013) 51–59, https://doi.org/10.1089/big.2013.1508.

[2] D.V. Rodríguez-Almonacid, J.G. Ramírez-Gil, O.L. Higuera, F. Hernández, E. Díaz-Almanza, A comprehensive step-by-step guide to using data science tools in the gestion of epidemiological and climatological data in rice production systems, Agronomy 13 (2023) 2844, https://doi.org/10.3390/agronomy13112844.

[3] L. Igual, S. Seguí, Introduction to data science, in: L. Igual, S. Seguí (Eds.), Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications, Springer International Publishing, Cham, 2024, pp. 1–4, https://doi.org/10.1007/978-3-031-48956-3_1.

[4] K.H. Coble, A.K. Mishra, S. Ferrell, T. Griffin, Big data in agriculture: a challenge for the future, Appl. Econ. Perspect. Pol. 40 (2018) 79–96, https://doi.org/10.1093/aepp/ppx056.

[5] W. van der Aalst, Data science in action, in: W. van der Aalst (Ed.), Process Mining, Springer, Berlin, Heidelberg, 2016, pp. 3–23, https://doi.org/10.1007/978-3-662-49851-4_1.

[6] A. Kamilaris, A. Kartakoullis, F.X. Prenafeta-Boldú, A review on the practice of big data analysis in agriculture, Comput. Electron. Agric. 143 (2017) 23–37, https://doi.org/10.1016/j.compag.2017.09.037.

[7] S. Biswas, M. Wardat, H. Rajan, The art and practice of data science pipelines: a comprehensive study of data science pipelines in theory, in-the-small, and in-the-large, in: 2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE), 2022, pp. 2091–2103, https://doi.org/10.1145/3510003.3510057.

[8] M. Muller, I. Lange, D. Wang, D. Piorkowski, J. Tsay, Q.V. Liao, C. Dugan, T. Erickson, How data science workers work with data: discovery, capture, curation, design, creation, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–15, https://doi.org/10.1145/3290605.3300356.

[9] R. Parvez, T. Ahmed, M. Ahsan, S. Arefin, N.H. Khan Chowdhury, F. Sumaiya, M. Hasan, Integrating multinomial logit and machine learning algorithms to detect crop choice decision making, in: 2024 IEEE International Conference on Electro Information Technology (eIT), 2024, pp. 525–531, https://doi.org/10.1109/eIT60633.2024.10609925.

[10] O. Ahumada, J.R. Villalobos, Application of planning models in the agri-food supply chain: a review, Eur. J. Oper. Res. 196 (2009) 1–20, https://doi.org/10.1016/j.ejor.2008.02.014.

[11] K. Bronson, I. Knezevic, Big Data in food and agriculture, Big Data & Society 3 (2016) 2053951716648174, https://doi.org/10.1177/2053951716648174.

[12] A. Scuderi, G. La Via, G. Timpanaro, L. Sturiale, The digital applications of "agriculture 4.0": strategic opportunity for the development of the Italian citrus chain, Agriculture 12 (2022) 400, https://doi.org/10.3390/agriculture12030400.

[13] MdF.B. Alam, S.R. Tushar, S.Md Zaman, E.D.R.S. Gonzalez, A.B.M.M. Bari, C.L. Karmaker, Analysis of the drivers of Agriculture 4.0 implementation in the emerging economies: implications towards sustainability and food security, Green Technologies and Sustainability 1 (2023) 100021, https://doi.org/10.1016/j.grets.2023.100021.

[14] L. Cao, Data science: a comprehensive overview, ACM Comput. Surv. 50 (43) (2017) 1–43:42, https://doi.org/10.1145/3076253.

[15] O. Elijah, T.A. Rahman, I. Orikumhi, C.Y. Leow, M.N. Hindia, An overview of internet of things (IoT) and data analytics in agriculture: benefits and challenges, IEEE Internet Things J. 5 (2018) 3758–3773, https://doi.org/10.1109/JIOT.2018.2844296.

[16] S. Wolfert, L. Ge, C. Verdouw, M.-J. Bogaardt, Big data in smart farming – a review, Agric. Syst. 153 (2017) 69–80, https://doi.org/10.1016/j.agsy.2017.01.023.

[17] K.A. Figueroa-Rodríguez, F. Hernández-Rosas, B. Figueroa-Sandoval, J. Velasco-Velasco, N. Aguilar Rivera, What has been the focus of sugarcane research? A bibliometric overview, Int J Environ Res Public Health 16 (2019) 3326, https://doi.org/10.3390/ijerph16183326.

[18] M. de Matos, F. Santos, P. Eichler, Chapter 1 - sugarcane world scenario, in: F. Santos, S.C. Rabelo, M. De Matos, P. Eichler (Eds.), Sugarcane Biorefinery, Technology and Perspectives, Academic Press, 2020, pp. 1–19, https://doi.org/10.1016/B978-0-12-814236-3.00001-9.

[19] Asocaña, Sector Agroindustrial de la Caña, Asocaña - Sector Agroindustrial de La Caña. https://www.asocana.org/publico/info.aspx?Cid=215, 2023. (Accessed 14 July 2023).

[20] N. Aguilar-Rivera, A framework for the analysis of socioeconomic and geographic sugarcane agro industry sustainability, Soc. Econ. Plann. Sci. 66 (2019) 149–160, https://doi.org/10.1016/j.seps.2018.07.006.

[21] D. Carson, F. Botha, Genes expressed in sugarcane maturing internodal tissue, Plant Cell Rep. 20 (2002) 1075–1081, https://doi.org/10.1007/s00299-002-0444-1.

[22] F. Santos, P. Eichler, G. Machado, J. De Mattia, G. De Souza, Chapter 2 - by-products of the sugarcane industry, in: F. Santos, S.C. Rabelo, M. De Matos, P. Eichler (Eds.), Sugarcane Biorefinery, Technology and Perspectives, Academic Press, 2020, pp. 21–48, https://doi.org/10.1016/B978-0-12-814236-3.00002-0.

[23] C.N. Bezuidenhout, A. Singels, Operational forecasting of South African sugarcane production: Part 1 – system description, Agric. Syst. 92 (2007) 23–38, https://doi.org/10.1016/j.agsy.2006.02.001.

[24] G.J. O'Leary, A review of three sugarcane simulation models with respect to their prediction of sucrose yield, Field Crops Res. 68 (2000) 97–111, https://doi.org/10.1016/S0378-4290(00)00112-X.

[25] M. Christina, M.-R. Jones, A. Versini, M. Mézino, L. Le Mézo, S. Auzoux, J.C. Soulié, C. Poser, E. Gérardeaux, Impact of climate variability and extreme rainfall events on sugarcane yield gap in a tropical Island, Field Crops Res. 274 (2021) 108326, https://doi.org/10.1016/j.fcr.2021.108326.

[26] D. Greenland, Climate variability and sugarcane yield in Louisiana, J. Appl. Meteorol. Climatol. 44 (2005) 1655–1666, https://doi.org/10.1175/JAM2299.1.

[27] Y. Zhao, L.-X. Yu, J. Ai, Z.-F. Zhang, J. Deng, Y.-B. Zhang, Climate variations in the low-latitude plateau contribute to different sugarcane (saccharum spp.) yields and sugar contents in China, Plants 12 (2023) 2712, https://doi.org/10.3390/plants12142712.

[28] M. Sachdeva, S. Bhatia, S.K. Batta, Sucrose accumulation in sugarcane: a potential target for crop improvement, Acta Physiol. Plant. 33 (2011) 1571–1583, https://doi.org/10.1007/s11738-011-0741-9.

[29] M.D. Hatch, J.A. Sacher, K.T. Glasziou, Sugar accumulation cycle in sugar cane. I. Studies on enzymes of the cycle, Plant Physiol 38 (1963) 338–343.

[30] M.E. Wagih, A. Ala, Y. Musa, Evaluation of sugarcane varieties for maturity earliness and selection for efficient sugar accumulation, Sugar Tech 6 (2004) 297–304, https://doi.org/10.1007/BF02942512.

[31] P.H. Moore, Temporal and spatial regulation of sucrose accumulation in the sugarcane stem, Functional Plant Biol 22 (1995) 661–679, https://doi.org/10.1071/pp9950661.

[32] Q. Khan, Y. Qin, D.-J. Guo, L.-T. Yang, X.-P. Song, Y.-X. Xing, Y.-R. Li, A review of the diverse genes and molecules involved in sucrose metabolism and innovative approaches to improve sucrose content in sugarcane, Agronomy 13 (2023) 2957, https://doi.org/10.3390/agronomy13122957.

[33] C. Driemeier, L.Y. Ling, G.M. Sanches, A.O. Pontes, P.S.G. Magalhães, J.E. Ferreira, A computational environment to support research in sugarcane agriculture, Comput. Electron. Agric. 130 (2016) 13–19, https://doi.org/10.1016/j.compag.2016.10.002.

[34] L.F. Maldaner, J.P. Molin, Data processing within rows for sugarcane yield mapping, Sci. Agric. 77 (2019) e20180391, https://doi.org/10.1590/1678-992X-2018-0391.

[35] B.F.T. Rudorff, G.T. Batista, Yield estimation of sugarcane based on agrometeorological-spectral models, Remote Sensing of Environment 33 (1990) 183–192, https://doi.org/10.1016/0034-4257(90)90029-L.

[36] M.S. Scarpari, E.G.F. de Beauclair, Sugarcane maturity estimation through edaphic-climatic parameters, Sci. Agric. 61 (2004) 486–491, https://doi.org/10.1590/S0103-90162004000500004.

[37] E. Uzun, A novel web scraping approach using the additional information obtained from web pages, IEEE Access 8 (2020) 61726–61740, https://doi.org/10.1109/ACCESS.2020.2984503.

[38] Zyte and many other contributors, Scrapy, A fast and powerful scraping and web crawling framework, Scrapy (2024). https://scrapy.org/. (Accessed 25 April 2024).

[39] W. Mckinney, Pandas: a foundational Python library for data analysis and statistics. https://www.dlr.de/sc/en/Portaldata/15/Resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf, 2011. (Accessed 7 June 2023).

[40] Statistical Analysis with Missing Data, third ed., Wiley, 2019. https://www.jstor.org/stable/1165119. (Accessed 26 April 2024).

[41] K.K. Nicodemus, J.D. Malley, C. Strobl, A. Ziegler, The behaviour of random forest permutation-based variable importance measures under predictor correlation, BMC Bioinf. 11 (2010) 110, https://doi.org/10.1186/1471-2105-11-110.

[42] D.V. Lindley, Regression and correlation analysis, in: J. Eatwell, M. Milgate, P. Newman (Eds.), Time Series and Statistics, Palgrave Macmillan UK, London, 1990, pp. 237–243, https://doi.org/10.1007/978-1-349-20865-4_30.

[43] J.D. Hunter, Matplotlib: a 2D graphics environment, Comput. Sci. Eng. 9 (2007) 90–95, https://doi.org/10.1109/MCSE.2007.55.

[44] M.L. Waskom, seaborn: statistical data visualization, J. Open Source Softw. 6 (2021) 3021, https://doi.org/10.21105/joss.03021.

[45] R. Krzysztofowicz, Transformation and normalization of variates with specified distributions, J. Hydrol. 197 (1997) 286–292, https://doi.org/10.1016/S0022-1694(96)03276-3.

[46] S.G.K. Patro, K.K. Sahu, Normalization: a preprocessing stage. https://doi.org/10.48550/arXiv.1503.06462, 2015.

[47] P. Cerda, G. Varoquaux, B. Kégl, Similarity encoding for learning with dirty categorical variables, Mach. Learn. 107 (2018) 1477–1494, https://doi.org/10.1007/s10994-018-5724-2.

[48] A.J. Dobson, A.G. Barnett, An Introduction to Generalized Linear Models, fourth ed., Chapman and Hall/CRC, New York, 2018 https://doi.org/10.1201/9781315182780.
[49] T.J.H. Pregibon Daryl, Generalized linear models, in: Statistical Models in S, Routledge, 1992.
[50] R.H. Myers, D.C. Montgomery, A tutorial on generalized linear models, J. Qual. Technol. 29 (1997) 274–291, https://doi.org/10.1080/00224065.1997.11979769.
[51] R.-C. Yang, Towards understanding and use of mixed-model analysis of agricultural experiments, Can. J. Plant Sci. 90 (2010) 605–627, https://doi.org/10.4141/CJPS10049.
[52] J. Ranstam, J.A. Cook, LASSO regression, Br. J. Surg. 105 (2018) 1348, https://doi.org/10.1002/bjs.10895.
[53] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Stat. Soc. B 58 (1996) 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
[54] J.G. Carbonell, R.S. Michalski, T.M. Mitchell, 1 - an overview of machine learning, in: R.S. Michalski, J.G. Carbonell, T.M. Mitchell (Eds.), Machine Learning, Morgan Kaufmann, San Francisco (CA), 1983, pp. 3–23, https://doi.org/10.1016/B978-0-08-051054-5.50005-4.
[55] H.M.E. Misilmani, T. Naous, Machine learning in antenna design: an overview on machine learning concept and algorithms, in: 2019 International Conference on High Performance Computing & Simulation (HPCS), 2019, pp. 600–607, https://doi.org/10.1109/HPCS48598.2019.9188224.
[56] P.P. Shinde, S. Shah, A review of machine learning and deep learning applications, in: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1–6, https://doi.org/10.1109/ICCUBEA.2018.8697857.
[57] K.M. Lee, J. Yoo, S.-W. Kim, J.-H. Lee, J. Hong, Autonomic machine learning platform, Int. J. Inf. Manag. 49 (2019) 491–501, https://doi.org/10.1016/j.ijinfomgt.2019.07.003.
[58] B. Liu, A very brief and critical discussion on AutoML, ArXiv, https://www.semanticscholar.org/paper/A-Very-Brief-and-Critical-Discussion-on-AutoML-Liu/b50959f1f80db92e383ec6b14fdac871d9399b53, 2018. (Accessed 26 April 2024).
[59] N. Sarangpure, V. Dhamde, A. Roge, J. Doye, S. Patle, S. Tamboli, Automating the machine learning process using PyCaret and streamlit, in: 2023 2nd International Conference for Innovation in Technology (INOCON), 2023, pp. 1–5, https://doi.org/10.1109/INOCON57975.2023.10101357.
[60] J. Mendes-Moreira, C. Soares, A.M. Jorge, J.F.D. Sousa, Ensemble approaches for regression: a survey, ACM Comput. Surv. 45 (10) (2012) 1–10, https://doi.org/10.1145/2379776.2379786, 40.
[61] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42, https://doi.org/10.1007/s10994-006-6226-1.
[62] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: a highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html. (Accessed 26 April 2024).
[63] J. Alzubi, A. Nayyar, A. Kumar, Machine learning from theory to algorithms: an overview, J. Phys.: Conf. Ser. 1142 (2018) 012012, https://doi.org/10.1088/1742-6596/1142/1/012012.