Entre Ciencia e Ingeniería, vol. 18, no. 35, enero-junio de 2024, páginas 41-45. DOI: https://doi.org/10.31908/19098367.2993
ISSN 1909-8367 (Impreso), ISSN 2539-4169 (En línea)

Formulation of a model to determine current and potential zones of cultivation for Hass avocado (*Persea americana* Mill) in the department of Risaralda based on edaphoclimatic and fruit quality variables[1]

Formulación de un modelo para determinar las zonas actuales y potenciales de cultivo de aguacate Hass (Persea americana Mill) en el departamento de Risaralda a partir de variables edafoclimáticas y de calidad del fruto

G.E. Guerrero, J.C. Chavarro, C.M. Castillo, C.A. Jaramillo, J.P. Arrubla y A.A. Patiño

Recibido: Octubre 6 de 2023 - Aceptado: mayo 19 de 2024

***Abstract*—Agriculture is one of the fundamental pillars of all societies worldwide, and the proper management of information allows timely decisions to be made about the advancement of companies. Government entities support areas of agriculture such as the cultivation of Hass avocado (*Persea americana* Mill). Among the challenges associated with this type of cultivation is the need to find potential areas for planting and suitable productivity and to contribute to technological developments in the agricultural sector, benefiting Hass avocado growers from the department of Risaralda. Therefore, in this study, a model that allows the determination of current and potential areas of cultivation for Hass avocado (*Persea americana* Mill) in this department based on edaphoclimatic variables and fruit quality is proposed. This model takes advantage of current trends in precision agriculture, including techniques derived from machine learning and supervised learning algorithms, among which is random forest.**

***Keywords*—Hass Avocado, Machine Learning, Random Forest, Potential Crop Zones.**

***Resumen*—La agricultura es uno de los pilares fundamentales de cualquier población a nivel mundial y el manejo adecuado de información permite tomar decisiones oportunas para el avance de cualquier empresa que el ser humano desarrolle. Las entidades gubernamentales apoyan renglones de la agricultura como es el cultivo del Aguacate (*Persea americana* Mill) variedad Hass. Entre los desafíos que tiene este tipo de cultivo, es encontrar zonas potenciales de siembra y productividad, con el fin de contribuir en desarrollos tecnológicos en el sector agrario, siendo beneficiarios los cultivadores de Aguacate Hass del departamento de Risaralda. Por lo tanto, con este estudio se propone formular un modelo que permita determinar zonas actuales y potenciales de cultivos de aguacate (*Persea americana* Mill) variedad Hass, en el departamento, con base en variables edafoclimáticas y de calidad del fruto, aprovechando las tendencias actuales de la agricultura de precisión, incluyendo técnicas derivadas del Aprendizaje Automático, como la utilización de algoritmos de Aprendizaje Supervisado, entre los cuales esta Random Forest.**

***Palabras clave*—Aguacate Hass, Aprendizaje Automático, Random Forest, Zonas Potenciales de Cultivos.**

[1]Producto apoyado por MinCiencias, Gobernación de Risaralda, y la Alcaldía de Pereira, trabajo realizado por los grupos de investigación GIA y Oleoquimica la Universidad Tecnológica de Pereira.

G.E. Guerrero, Universidad Tecnológica de Pereira, Pereira, Colombia, email: gguerrero@utp.edu.co
J.C. Chavarro, Universidad Tecnológica de Pereira, Pereira, Colombia, email: jchavar@utp.edu.co
C.M. Castillo, Universidad Tecnológica de Pereira, Pereira, Colombia, email: cesar.castillo@utp.edu.co
C.A. Jaramillo, Universidad Tecnológica de Pereira, Pereira, Colombia, email: swokosky@utp.edu.co
J.P. Arrubla, Universidad Tecnológica de Pereira, Pereira, Colombia, email: juanpablo77@utp.edu.co
A.A. Patiño, Pereira, Colombia, email: andrespatinomartinez@gmail.com

Como citar este artículo: C. M. Castillo, G. E. Guerrero, J. C. Chavarro, C. A. Jaramillo, J. P. Arrubla y A. A. Patiño. Formulation of a model to determine current and potential zones of cultivation for Hass avocado (Persea americana Mill) in the department of Risaralda based on edaphoclimatic and fruit quality variables, Entre Ciencia e Ingeniería, vol. 18, no. 35, pp. 41-45, enero-junio 2024. DOI: https://doi.org/10.31908/19098367.2993.

I. Introduction

FOOD security has become a pressing challenge due to rapid population growth, climate change, and water scarcity, especially in developing countries [1]. Globally, modifications and adaptations are made through agricultural practices to improve soil fertility and mediate climatic changes; therefore, the environmental impacts of agriculture must be evaluated in terms of water, nutrients (soil), and atmospheric components [2]. National and departmental government entities support emergent agricultural endeavors that provide a good opportunity to increase levels of production and commercialization of products [3].

Avocado (*Persea americana* Mill) of the Hass variety is the most common commercial avocado crop in the world due to its contents of essential nutrients and important phytochemicals [4]. This fruit is grown in Colombia, where the Hassavocado production system has expanded in recent years due to excellent economic opportunities and the high unmet domestic demand [5]. Departments in which Hass avocados are cultivated include Tolima, Antioquia, Caldas, Santander, Bolívar, Quindío, Cesar, Valle del Cauca, Risaralda and Cundinamarca [6]; in Risaralda, avocado is grown in 13 of the 14 municipalities. To address problems associated with avocado cultivation, approaches from information and communication technologies (ICT) are adopted, such as precision agriculture and crop data collection in real time [7]. Relevant topics in ICT for the agricultural sector include procedures for digital management of geographic information about crops, decision-making systems for mechanization of processes based on georeferencing, and information systems used for epidemiological early warnings [7] [8].

However, difficulties have been observed in managing Hass avocado given the variability in the edaphoclimatic conditions of the department, which has effects such as heterogeneity of fruit quality. In addition, the lack of scientific work limits the understanding of the specificities of cultivation in the region. This study proposes the development of a model to determine the current and potential cultivation areas for avocado (*Persea americana* Mill) var. Hass in the department of Risaralda based on edaphoclimatic and fruit quality variables using current trends in precision agriculture and machine learning. The results contribute to technological developments in the agricultural sector, with the department's Hass avocado growers being beneficiaries.

II. Background

Machine learning (ML) is an important decision-support tool in fields such as crop yield prediction, and it can support decisions about which agricultural products to grow and what to do during the growing season of crops. Therefore, several machine learning algorithms have been applied to support research aimed at predicting or estimating crop yield, where the most commonly used characteristics are temperature, precipitation, and soil type [9]. In this sense, the different applications of the data obtained from the soil, such as the selection of a dataset for training as well as the selection of soil environmental covariates, could boost the precision of machine learning techniques [10].

One of the instruments used within machine learning is crop-oriented recommendation systems. Based on the variables provided within these systems, a model is created that predicts or suggests which crop can be grown. The models use historical data, such as climatic data (temperature, humidity, pH, and precipitation) and fertilizer application (nitrogen, potassium, and phosphorus) [11]. The application of machine learning techniques supports processes associated with data analysis [12]. In [13], it is mentioned that different artificial intelligence techniques have been proposed. Among those techniques that have been integrated into precision agriculture, more specifically into the field of crop recommendation systems, algorithms such as k nearest neighbors (KNN), similarity-based models, ensemble-based models, and neural networks take into account various characteristics that are external in nature, such as meteorological data and the soil profile, to provide the best recommendations. The most important attributes of the data are obtained using techniques such as principal component analysis (PCA) and linear discrimination analysis (LDA), and the extracted attributes are used to train models such as the naïve Bayes classifier (NBC), random forest, and KNN. The selection of training data and performance evaluation are based on test data and rely on techniques such as cross-validation, RMSE, or precision statistics. Obtaining reliable crop yield predictions during cultivation is difficult, as crop production varies according to various climatic conditions, such as the dry period and temperature. It increases the need for analysis of crop production in different climatic conditions. Therefore, in [14], the automatic learning method was analyzed, and it is reported that random forest, a supervised algorithm, has the capacity to analyze crop growth in relation to the current climatic conditions and biophysical changes. Similarly, in [10], the ability of the random forest algorithm to predict soil classes from different training datasets and extrapolate this information to a similar area was evaluated.

Another opportunity to consider is machine learning models/algorithms and their possible applications to geospatial data. Special attention is given to the models that are based on artificial neural networks (multilayer perceptron, general regression neural networks, self-organized maps), statistical learning theory (support vector machines) [15], geostatistical techniques such as ordinary kriging [16], or algorithms such as random forest and random forest spatial interpolation (RFSI) [17].

To visualize different crops and their associated characteristics, high-resolution yield maps are used. These maps are an essential tool in modern agriculture and are obtained through spatial interpolation; however, spatial interpolation is generally performed using methods that can be computationally demanding [18]. To this end, some work has been carried out to explicitly consider spatial analyses in machine learning approaches. It includes observations made at

the prediction location using random forest and comparing these predictions with those based on deterministic interpolation methods, such as ordinary kriging, regression kriging, and random forest for spatial prediction (RFsp). For studies focused on precipitation and temperature, RF generally outperformed regression kriging, inverse distance weighting, and RFsp; in addition, RF was substantially faster than RFsp [17].

III. Methodology

This study seeks to formulate a recommendation for a model that will use the random forest algorithm and integrate three data sources: climatic, building, and fruit quality variables. Following the construction of the model, the development of an information system is proposed to determine current and potential areas of Hass avocado cultivation in the department of Risaralda based on edaphoclimatic and fruit quality variables. This work will be performed under the following project: "Development of an information system to determine current and potential areas of cultivation for Hass avocado (*Persea americana* Mill) in the department of Risaralda based on edaphoclimatic and fruit quality variables (Research Project contract 424-201 MinCiencias)". To construct a suitable and precise machine learning model for this project, relevant datasets will be obtained; then, the data will be preprocessed. The data will be cleaned, and spatial data in the form of geographical coordinates will be added to each data point. Data creation will be performed according to the selected machine learning algorithm, and evaluations of the results obtained from the test and training phases of the model will be carried out. Finally, predictions of the potential for land cultivation will be created.

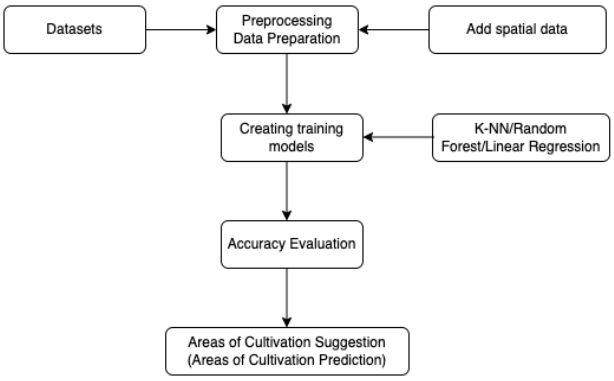

Fig. 1. Methodology for developing the model

A. Description of the Dataset.

Field sampling will be carried out to collect data on soils, fruit quality, and climatic conditions from the Hass avocado-producing farms of the municipalities of Pereira, Dosquebradas, Santa Rosa de Cabal, Marsella, Apía, Belén de Umbría, Gúatica and Quinchía, department of Risaralda. The first dataset will include data from seven (7) LynkBOX CLIMA PLUS climatic stations that have been installed on the different farms located in the municipalities selected for the study (Fig. 2), allowing the data of climatic variables to be recorded. Temperature (°C), relative humidity (%), precipitation (mm), solar radiation (W/m2), and wind speed (m/s) data will be recorded for a year. The second set of data will be based on a soil fertility analysis, and pH and contents of organic matter, potassium (K), calcium (Ca), magnesium (Mg), sodium (Na), and phosphorus (P) will be evaluated according to Colombian technical standards NTC 5264, 5403, 5349 and 5350. Additionally, the third dataset will be derived from the analysis of fruit quality samples, and variables such as fruit dry matter, quality, moisture, and calcium content will be determined.

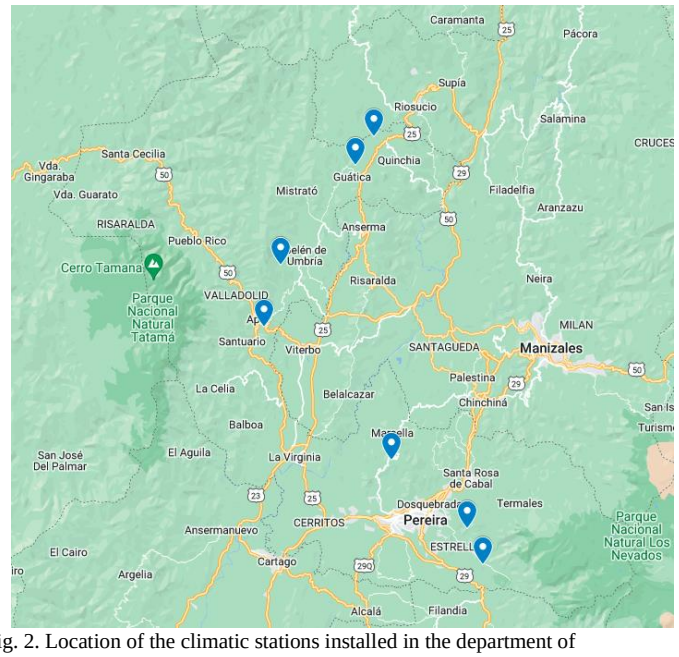

Fig. 2. Location of the climatic stations installed in the department of Risaralda

B. Data Preparation.

The data will have different formats within the datasets generated, as observed in the datasets described above; therefore, it will be imperative to clean and normalize the data for later use in the model. To address missing data, the random forest algorithm [19] will be used. A process for scaling the data or normalization will also be carried out to convert the dates and times of the different data sources.

C. Random Forest.

When the datasets are configured, the prediction process will be carried out using the random forest algorithm and local spatial information, that is, data on the spatial dependencies and complex spatial patterns that arise [20]. The initial data training will be performed using the random forest algorithm; this algorithm is based on decision trees and generates a prediction through a series of division rules. The spatial correlation between the data obtained is not included in the standard random forest output, and it will be taken into account that the nearby data contain information about a prediction location. Therefore, additional spatial variables will

be incorporated into the random forest model.

D. Accuracy Assessment.

Accuracy metrics, such as accuracy and root mean square error (RMSE), will be used to verify predictions [17].

Accuracy is a metric that delivers the total percentage of elements that are classified correctly, where the percentage is denoted as a value between 0 and 1; the higher the value is, the more accurate the model.

The root mean square error (RMSE) is a periodically realistic quantity between the statistical value of the population and the samples predicted by the model. RMSE refers to an anomaly between the expected values and the observations. These individual changes are detected as anomalies when the calculations are estimated as prediction errors, and the calculations are performed using data samples known as prediction errors. The square is then quantified to obtain the RMS value of a set of data values [21].

IV. Results

The model is developed and implemented on a two-core MacBook Pro Intel Core i5 computer, and the programming language used is Python with the scikit-learn library. The model is run using the random forest algorithm. A test of the model is established with the independent variables being relative humidity, precipitation, solar radiation, wind speed, wind direction, and the variable to be predicted being ambient temperature.

E. Model Parameterization.

To tune the model, cross-validation or k-fold cross-validation is chosen, where two separate datasets were created from the original data: a training set (and test set) and a validation set, where k-fold = 100.

F. Hyperparameters and Metrics for the Evaluation of the Decision Tree Model.

In the process of constructing the random forest model, the number of trees is set to 100 (n_estimators = 100), and the default values are retained for the rest of the hyperparameters. The number of trees was selected by using k-fold validation. The evaluation metrics yielded the following results: accuracy = 97.9536827 and RMSE = 0.6800923.

G. Evaluation of the Model.

After parameterization and training of the model, a prediction of ambient temperature based on climatic variables is obtained from the random forest model. Table I shows the importance of the explanatory variables (in percentage) in the model.

TABLE I
IMPORTANCE OF EXPLANATORY VARIABLES IN THE MODEL.

No.	Variable	(%)
1	Relative humidity	74.9
2	Precipitation	10.1
3	Solar radiation	8.3
4	Wind speed	4.2
5	Wind Direction	2.5

H. Model Test.

The model is tested with a dataset constructed of the previously described variables. The prediction of ambient temperature obtained is in the range of 14 to 40 degrees Celsius, which is considered a required temperature for the cultivation of this fruit [22].

V. Conclusions

Generating a prediction of potential areas for crop cultivation based on machine learning algorithms allows Hass avocado producers to make decisions based on factors such as temperature, rainfall, and soil conditions. The information system proposed under the execution of this project (contract 424-201 MinCiencias) will make it possible to determine, based on predictions, if the current production areas of the avocado crop are the most appropriate and identify the potential areas of cultivation.

References

[1] El-Bendary, N., Elhariri, E., Hazman, M., Saleh, S.M., Hassanien, A.E.: Cultivation-time recommender system based on climatic conditions for newly reclaimed lands in egypt. Procedia Computer Science 96, 110–119 (2016)

[2] Singh, R., Kumari, T., Verma, P., Singh, B.P., Raghubanshi, A.S.: Compatible package-based agriculture systems: an urgent need for agro-ecological balance and climate change adaptation. Soil Ecology Letters 4(3), 187–212 (2022)

[3] Perfetti, J.J., Bravo-Ureta, B.E., García, A., Pantoja, J., Delgado, M., Blanco, J., Jara, R., Moraga, C., Paredes, G., Naranjo, J., et al.: Adecuación de tierras y el desarrollo de la agricultura colombiana: políticas e instituciones. Fedesarrollo 447, 456 (2019)

[4] Dreher, M.L., Davenport, A.J.: Hass avocado composition and potential health effects. Critical reviews in food science and nutrition 53(7), 738–750 (2013)

[5] Ramírez-Gil, J.G., Ramelli, E.G., Osorio, J.G.M.: Economic impact of the avocado (cv. hass) wilt disease complex in Antioquia, Colombia, crops under different technological management levels. Crop protection 101, 103–115 (2017)

[6] MADR, M.d.A.y.D.R.: CADENA DE AGUACATE, Indicadores e instrumentos. Lect Econ. 2019;52(52):165–94 (2019)

[7] Martínez, D.H.F., Galvis, C.P.U.: Tic para la investigación, desarrollo e innovación del sector agropecuario (2018)

[8] Orozco, O. A., Llano Ramírez, G.: Sistemas de información enfocados en tecnologías de agricultura de precisión y aplicables a la caña de azúcar, una revisión. Revista Ingenierías Universidad de Medellín 15(28), 103–124 (2016)

[9] Van Klompenburg, T., Kassahun, A., Catal, C.: Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture 177, 105709 (2020)

[10] Machado, D.F.T., Silva, S.H.G., Curi, N., Menezes, M.D.d.: Soil type spatial prediction from random forest: different training datasets,

transferability, accuracy and uncertainty assessment. Scientia Agricola 76, 243–254 (2019)
[11] Jyothika, P., Ramana, K.V., Narayana, L.: Crop recommendation system to maximize crop yield using deep neural network.
[12] Vite Cevallos, H., Carvajal Romero, H., Barrezueta Unda, S.: Aplicación de algoritmos de aprendizaje automático para clasificar la fertilidad de un suelo bananero. Conrado 16(72), 15–19 (2020)
[13] Katarya, R., Raturi, A., Mehndiratta, A., Thapper, A.: Impact of machine learning techniques in precision agriculture. In: 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), pp. 1–6 (2020). IEEE
[14] Geetha, V., Punitha, A., Abarna, M., Akshaya, M., Illakiya, S., Janani, A.: An effective crop prediction using random forest algorithm. In: 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), pp. 1–5 (2020). IEEE
[15] Pozdnoukhov, A., Kanevski, M.: Machine learning algorithms for analysis and modeling of geospatial data. In: Annual Conference of International Association for Mathematical Geology (IAMG 07), Beijing, China, 25-31 August (2007)
[16] Carranza, J.P., Salomón, M.J., Piumetto, M.A., Monzani, F., MONTENEGRO CALVIMONTE, M., Córdoba, M.A.: Random forest como técnica de valuación masiva del valor del suelo urbano: una aplicación para la ciudad de río cuarto, córdoba, argentina. In: Congresso Brasileiro de Cadastro Técnico Multifinalitário-COBRAC (2018)
[17] Sekulic, A., Kilibarda, M., Heuvelink, G.B., Nikoli´c, M., Bajat, B.: Random Forest Spatial Interpolation. Remote Sensing 12(10), 1687 (2020)
[18] Mariano, C., Monica, B.: A random forest-based algorithm for data-intensive spatial interpolation in crop yield mapping. Computers and Electronics in Agriculture 184, 106094 (2021)
[19] Kuhn, M. Johnson, K. Kuhn, M, Johnson, K. Classification trees and rule-based models. Applied predictive modeling, 369–413 (2013)
[20] Talebi, H., Peeters, L.J., Otto, A., Tolosana-Delgado, R.: A truly spatial random forests algorithm for geoscience data analysis and modeling. Mathematical Geosciences 54, 1–22 (2022)
[21] Rajeshwari, M., Shunmuganathan, N., Sankarasubramanian, R.: Performance of soil prediction using machine learning for data clustering methods. Journal of Algebraic Statistics 13(2), 858–864 (2022).
[22] FAO - GAEZ Data Portal.: Data sheet Persea americana. Available from: https://gaez.fao.org/pages/ecocrop-find-plant (2023)

Gloria Edith Guerrero Alvarez. Chemistry graduated on December 15, 1992 from the National University of Colombia, Bogotá, as a doctor in chemical sciences. The title was awarded on April 6, 2001 by the National University of Colombia, Bogotá. researcher at Cenipalma headquarters in Bogotá and currently a professor at the Technological University of Pereira assigned to the Chemical Technology program of the Faculty of Technology, linked since February 6, 2004. The areas of interest are: agrochemistry, particularly studies of plant protection plants, use of secondary metabolites for the selection of promising materials, valorization of agricultural and agroindustrial byproducts, and applications of precision agriculture in commercial crops of national interest.
ORCID: https://orcid.org/0000-0002-0529-5835

Julio César Chavarro Porras. Systems and computing engineer, from Universidad Distrital Francisco José de Caldas Bogotá, Colombia. Specialist in physical instrumentation from Universidad Tecnológica de Pereira. Ph.D. engineering, emphasis in computer science. Director Group Artificial Intelligence (GIA) from Universidad Tecnológica de Pereira He has been a professor-researcher of Universidad Tecnológica de Pereira for over 26 years. He has been active in research groups whose areas of interest and teaching are related to software engineering and AI.
ORCID: https://orcid.org/0000-0001-8876-8855.

Cesar Manuel Castillo Rodriguez. Systems Engineer, M.Sc. Computer and Systems Engineering. Professor of the Universidad Tecnológica de Pereira, Faculty of Engineering, with experience in in-person higher education systems and virtual learning environments. Researcher Group Artificial Intelligence (GIA) from Universidad Tecnológica de Pereira. Areas of interest and teaching related to Data Science and Precision Agriculture. Studies are oriented toward the development of projects, the maximization of technological, physical, and environmental resources, and the consolidation of work groups.
ORCID: https://orcid.org/0000-0001-6561-8096.

César Augusto Jaramillo Acevedo. Systems and computing engineer, MSc in Systems and Computing Engineering from Universidad Tecnológica de Pereira. Director and researcher in projects related to Industry 4.0, precision agriculture, education, and business development. He has been a professor-researcher of Universidad Tecnológica de Pereira for over 12 years. He has been active in research groups whose areas of interest and teaching are related to software engineering, compilers, AI, IoT systems, the cloud, distributed systems, and Industry 4.0.
ORCID: https://orcid.org/0000-0002-0529-5835.

Juan Pablo Arrubla Vélez. Chemist (Universidad del Quindío, 1999), MSc. in Chemistry (Universidad Industrial de Santander, 2003), Doctor in Environmental Sciences (Universidad Tecnológica de Pereira, 2016); is a professor and researcher at the School of Chemical Technology of the Technological University from Pereira, Colombia since 2006. Has skills and experience in analytical method development, gas chromatography and mass spectrometry, and sample preparation. He has also worked on research in the fields of environmental pollution, natural products, and bioenergy.
ORCID: https://orcid.org/0000-0003-3572-4247.

Andrés Alfonso Patiño Martínez. Engineer agronomist from the Universidad de Santa Rosa de Cabal - UNISARC in the department of Risaralda - Colombia (2002), Master's degree in Agricultural Production Systems from the University of Caldas in the department of Caldas - Colombia (2018), and Junior Researcher categorized by Minciencias. He has worked in the research area in fruits, especially Avocado, Blackberry, and Banana. Currently, he serves as the dean of the Faculty of Agricultural Sciences at UNISARC.
ORCID: https://orcid.org/0000-0003-0602-2422.