World Journal of Information Systems 2(1) 45-56
DOI: https://doi.org/10.17013/wjis.v2i1.32

Article

Forecasting Model for Improving Agricultural Yield Using Recurrent Neural Networks

Hernán Julio Parra Osorio [1], Jimmy Aurelio Rosales Huamani [2], José Luis Castillo Sequera [3]*

[1] Facultad de Ingenieria Industrial y de Sistemas, Universidad Nacional de Ingenieria, Lima, Perú

[2] Facultad de Ingenieria Geologica Minera y Metalurgica, Universidad Nacional de Ingenieria, Lima, Perú

[3] Departamento de Ciencias de la Computación, Universidad de Alcalá Madrid, España

* Correspondence: hparra@uni.edu.pe

(Article history: Received January 05, 2025; Received in revised form January 27, 2025; Accepted February 23, 2025; Available online March 15, 2025)

Abstract: The agri-food industry is currently migrating to Agri-Food 4.0. This trend involves incorporating the latest technological innovations with the goal of optimizing resources and making better decisions thanks to the information and data obtained through them. In this sense, this technology, known as Agroindustry 4.0, will allow us to develop crop yield estimates, which are vital for good agronomic management. This will allow for better land management in the face of increasing food demand due to constant population growth. This research proposes a crop yield forecasting model using the Cross Industry Standard Process for Data Mining (CRIP-DM) method with the Nonlinear Autoregressive Network with Exogenous Inputs (NARX) methodology using artificial neural networks (ANN). Applying this methodology will allow for better analysis of information on possible production scenarios for a given crop. This will help agricultural policymakers and stakeholders involved in food and nutrition security, project the real-time performance of a given crop´s production yield. This proposed model promises to yield results that provide significant contributions to addressing the problem of crop production, prediction and management. Ultimately, we can affirm that this study will allow us to achieve sustainable and responsible agriculture.

Keywords: Forecasting; Precision Agriculture; CRIP-DM; NARX-ANN

1. Introduction

The International Conference on Nutrition (ICN), convened by the Food and Agriculture Organization of the United Nations (FAO) and the World Health Organization (WHO) [1], noted with great concern the unacceptable fact that some 780 million people in developing countries (20 percent of the population) still do not have access to enough food to meet their

© 2025 ITMA. All rights reserved.

basic daily needs. In Yulian's work [2], I state that the incessant growth of the world's population directly pressures the food production sector, placing it in an inefficient reality with respect to current cultivation procedures. Therefore, it is of utmost importance to control the indiscriminate use of pesticides and agrochemicals in regard to responsible and sustainable cultivation.

Currently, agriculture is experiencing exponentially increasing use of information and communications technologies (ICTs). Autonomous robotic vehicles have been developed for agricultural purposes, such as mechanical weeding, fertilizer application, or fruit harvesting, as well as unmanned aerial vehicles with autonomous flight control. A key example is the development of lightweight, powerful hyperspectral snapshot cameras that can be used to calculate biomass growth and crop fertilization status.

Data mining will be used for data processing. Khumaidi [3] proposed a very appropriate method for conducting this research. In his work, the author indicated that the Cross-Industry Standard Process for Data Mining (CRISP-DM) is a method that provides a standard data mining process for solving most crop problems.

In Talavia [4] it was mentioned to have improved the efficiency of crop production in general and to have reduced the workload of farmers using Machine Learning, for which the author analyzed the resolution of this problem using fuzzy logic and artificial neural networks (ANN). From this it can be said that the accuracy of the forecast and the performance of the network of a NARX-ANN model, based on the analysis of time series, is significantly better than that of other models.

This paper consists of the following sections: Section II describes the existing literature on precision agriculture using various models. Section III develops the CRISP-DM and NARX-ANN proposal, fully applied to agriculture for use with a specific crop. At this stage, the basic requirements are outlined. Finally, Section IV describes the expected contributions of this model.

2. Styles and Formats

In recent years, with the adoption of Machine Learning (ML) in precision agriculture, several research works have already been carried out in various aspects of agriculture. To describe the methodologies used to implement forecasting systems in agriculture, we must first mention CRISP-DM, which is a proven method to guide works using data mining. In Altaleb [5], it was indicated that there is a potential expansion in the agricultural machinery industry and the use of Big Data is already being used in other industries such as e-commerce to improve decision-making processes.

This work uses the CRISP-DM method and is able to provide guidelines on how to manage and structure data in agricultural processes to achieve defined data mining objectives.

© 2025 ITMA. All rights reserved.

Then Khumaidi [3] mentioned that it is necessary to predict the yield of coffee trees, to estimate the planting, as well as forecast the most important variables to consider, to achieve the goal. Coffee productivity is influenced by internal and external factors, that is, the quality of the plant itself, the soil, altitude, and climate. The method used in this study is the CRISP-DM method and a multiple linear regression algorithm to predict the amount of coffee production and determine the relationship between the variables. Also, Mulatu [6] used the CRISP-DM methodology to predict rust disease in wheat crops using data mining with biotic factors such as rainfall, temperature, and relative humidity. Oliveira [7] used a CRISP-DM based platform with Machine Learning for soybean classification for quality issues.

Over the years, the replacement of traditional characterization of crop productivity and yield, with prediction using artificial intelligence methods and techniques has been proposed. Thus, there are many researches works that address these topics, such as Pascual [8], who applied neural networks for forecasting to improve pasture quality. Similarly, De la Piedra [9] presented a design that uses a feedforward neural network with a backpropagation algorithm, adding a Takagi-Sugeno type neurofuzzy system, simulated in MATLAB and implemented on an Arduino board. The same is true for Saldaña [10], who developed an artificial neural network prediction model with a feedforward backpropagation structure to forecast yield, but this time using a historical database for said prediction.

The research carried out by Mendoza-Haro [11] showed how important artificial neural networks are within some industries that want to be on par with the trends of agroindustry 4.0. In this case we put the sugar industry as an example, in which the historical climatological database was analyzed using MATLAB. Cao [12] used recurrent neural networks (RNN) for more complex short-term forecasting issues.

NARX models have many applications. Cárdenas [13] used them to predict wind speeds. Zhang [14] used them to solve classification problems using support vector machines (SVMs). Nedjah [15] used them for pattern recognition. Wang [16] used CA-NARX to predict water sources.

In recent years, neural networks have come in a wide variety of architectures and topologies. The most prominent type of neural networks are recurrent neural networks (RNNs), which help efficiently process directional sequences of data. Studies by Skolkovo [17] have shown how RNNs can be used in combination with computer vision algorithms to handle the task of crop growth prediction; this system was implemented on a Raspberry Pi computer.

Likewise, there are many studies referring to obtaining a better forecast of a crop and they are usually oriented towards the use of a group of topologies and algorithms. Arteaga [18] addresses the prediction of crop yield. In it, topologies such as dense neural network (DNN),

© 2025 ITMA. All rights reserved.

Random Forest (RF), decision trees (DT), and finally support vector machines (SVM) are compared. Also, Thilakarathne [19] developed a recommendation platform with a cloud-based service with the intention of offering precision agriculture solutions that are free and open source. In it, he evaluates five predictive machine learning algorithms; K-nearest neighbors (KNN), decision tree (DT), random forest (RF), extreme gradient boosting (XGBoost) and Support Vector Machine (SVM).

3. Methodology

To achieve the study objectives, a two-stage methodology was adopted. The first stage includes CRISP-DM, and the second stage applies NARX-ANN. The proposed system is shown in Figure 1. These methods are explained in detail below, and their significance and stages are analyzed in each step.

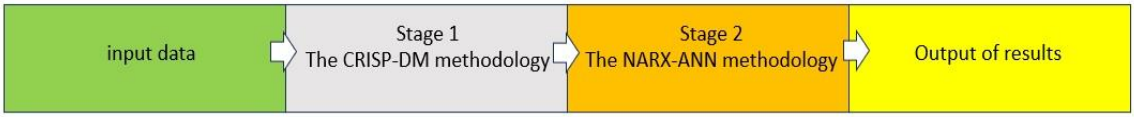


Figure 1. Proposed stages for the diagnosis of agricultural production

This work proposes two models for crop forecasting, a first CRISP-DM model where the data processing will be carried out and a second model where the forecast will be made with the NARX-ANN methodology. Both models are detailed below.

3. 1 Methodology for CRISP-DM model

Wirth [20] mentions that the CRISP-DM procedure is a comprehensive process model used to carry out data mining projects. This process model is independent of both the industry sector and the technology used. The CRISP-DM model provides an excellent basis for developing a specialized process model that prescribes the steps to be followed in detail. The CRISP-DM reference model for data mining provides an overview of the project life cycle and is divided into six phases shown in Figure 2. Below, we briefly describe each phase:

3. 1. 1 Business Understanding

This stage involves gathering the most relevant sources of information for business understanding. The following is a list of the parameters to be considered: the study objectives, the problem to be solved, resource requirement inventories, an assessment of the current external and internal environment, pest risks and contingencies, a glossary of terms, costs, and benefits, a crop study plan, which must accurately contain the plant's taxonomy and determine its classification in terms of species, genus, order, and class, an initial evaluation of tools and techniques, a description of the crop's characteristics, and the data mining success criteria.

© 2025 ITMA. All rights reserved.

3. 1. 2 Data Comprehension

This phase begins with the initial data collection. It must specify details of the location where the crop was grown. Therefore, the following information must be available: The geographic location, the growing area, the start and end dates of the crop, the meteorological data for the area such as relative humidity, maximum and minimum ambient temperatures, rainfall, flow, and wind speed. It is also essential to perform a soil and water analysis in certified laboratories. The soil analysis must include electrical conductivity (EC), texture, structure, consistency, and color of the soil, pH values, density, organic matter, concentrations of nitrogen (N2), phosphorus (P2O3), potassium (K2O), sodium, calcium, aluminum, and boron, and micronutrients such as iron, copper, zinc, and manganese. Regarding water analysis, this should include parameters such as electrical conductivity (EC), pH, and the concentration of calcium, sodium, magnesium, and potassium cations. Regarding anions, information on chlorides, nitrates, sulfates, carbonates, and bicarbonates should be included. Additional information on toxic ions such as cadmium, chloride, arsenic, lead, and mercury should also be included, if necessary. This data should include the use of fertilization programs that provide adequate amounts of nutrients throughout the growing season.

3. 1. 3 Data Preparation

Data Preprocessing

It is quite possible to have erroneous data due to random fluctuations at the time the variables are captured by the sensors. To correct this, techniques that guarantee their detection will be used. The Mahalanobis Distance (MD) algorithm is recommended, as it allows working with the inverse variance and covariance matrix of the sample set. When collecting data in real time, it is quite common for null or empty data to exist. This phenomenon is usually caused by the desynchronization of the microcontroller program with the peripherals (sensors/actuators), which must be eliminated and corrected. It is recommended that the series be completed correctly, and therefore each point should be analyzed individually to establish the interpolation of the missing data. To replace discarded values in the series, it is recommended to use the linear trend technique at one point. This technique involves regressing the series on a scalable index of the variable, from the value 1 to the total number of data points in the sample; the remaining values are then replaced by their predicted values.

Data Normalization

Data must be characterized according to their different dimensions as appropriate, therefore, the data sample set must be normalized using the formula: $y = (x-xmin) / (xmax-xmin)$, where $x, y \in Rn$; $xmin=min(x)$; $xmax=max(x)$, x are the data that were normalized,

© 2025 ITMA. All rights reserved.

and the results after being normalized using the formula. The results of the original data were normalized to the range of [0, 1].

3. 1. 4 Modeling

It is very important and necessary to determine whether all the crop parameters analyzed in this study should be incorporated into the model. To determine this, we used the Granger cointegration methodology. The use of this econometric technique applied to time series was proposed by Hocaoglu [21].

To do so, three well-known procedures must be developed: first, the graphical representation of the time series; second, the generation of a correlogram for each of the variables; and third, the Dickey-Fuller unit root statistical test.

The modeling will then be developed using OLAP (On-Line Analytical Processing) tools, the objective of which is to streamline the query of large amounts of data. To achieve this, it uses diverse, typically multidimensional, data structures (or OLAP Cubes). A network-like structured model is proposed, containing two types of tables: the measurement table, which should contain the numerical values of the parameters, and another dimension table, where the measurements are described, such as crop types, plant taxonomy, crop phenology, season and expected time of cultivation, soil and water characteristics, etc.

3. 1. 5 Evaluation

To evaluate the modeling process, data must be tested to observe distortions, perform comparisons, and validate information. Simulations with Bayesian templates must be performed to evaluate the results and submit a results report.

3. 1. 6 Implementation

In this final stage, the model will be implemented. This requires the operational requirements for each specific need, such as implementation and maintenance plans, crop parameters used and measured by sensors, enabled or operational devices, type of module used (microprocessor), software and operating system versions, availability of IoT and service providers, type of power source (battery, solar, etc.), characteristics, and availability of the land under study.

All of this must be indicated in the Checklist forms for the final report, which include observations and difficulties (conclusions and recommendations).

Finally, Figure 2 shows a summary of the stages for implementing the model.

© 2025 ITMA. All rights reserved.

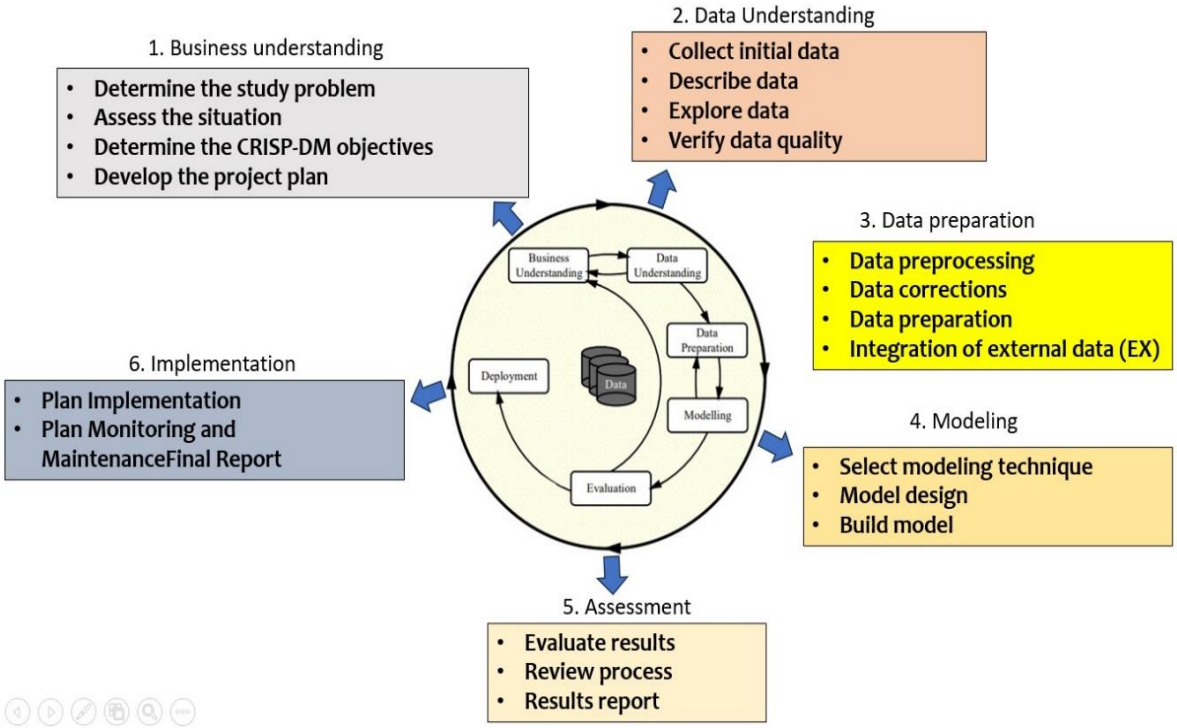


Figure 2. Steps of the CRISP-DM based methodology for agriculture 4.0 (Checklist)

3. 2 Forecasting methodology with the NARX-ANN model

Buitrago [22] describes the dynamic behavior of the NARX model using ANN. The construction starts by initially proposing a feedforward perceptron neural network, in which a nonlinear regression function behavior of the model is considered.

For the design and simulation of the NARX-ANN model, the development of the Neural Net Time Series Toolbox in MATLAB is proposed. These dynamic neural networks are suitable for time series prediction. There are multiple neural networks in the MATLAB libraries, as can be seen in the App and verified by the Machine Learning option. MATLAB classifies these networks into three categories: NARX networks applied in open-loop mode, closed-loop mode, and open/closed-loop multi-step prediction.

This tool allows you to solve three types of nonlinear time series problems, as detailed below:

- Nonlinear autoregressive systems with external inputs (NARX).
- Nonlinear autoregressive systems (NAR).
- Nonlinear input-output systems.

The development stages of a NARX-ANN model are detailed below (Figure 3).

© 2025 ITMA. All rights reserved.

To perform the respective simulations in the NARX type system, the Nonlinear Autoregressive with External (Exogenous) Input series of MATLAB was used. To start this library, the Dataset is loaded with an output (Targets) made up of a matrix where the information records are based on time. The recommended inputs are the phenological stages, the date of shipment, soil temperature and humidity (%), temperature (°C) and humidity (%) of the environment, nitrogen, phosphorus and potassium concentrations in the soil (mg / kg), electrical conductivity (mhos / cm), pH, heliophane (hr), UV radiation, seed quality, planting density, prolificacy. The output is the crop yield, which can be associated with the weight of the plant or grain, plant height, development of the number of leaves, length and diameter of a fruit or leaves. The training, validation and testing phase of the Dataset then proceeds. It is suggested to use 70% for training the neural network, 15% for validation and 15% for testing.

The NARX model's performance is evaluated through developed tests that examine its performance in terms of correlation coefficient (R2) and MSE values. The R2 identifies the relationship between measured and predicted values, ranging from 0 (zero) to 1 (one). An R2 value closer to 1 (one) indicates a better fit of the network model. The MSE value, on the other hand, shows the distinction between true and target outputs based on the squared average; Its lower value indicates better prediction accuracy in experiments. Low RMSE values indicate that the model promises very good prediction results with minimal error.

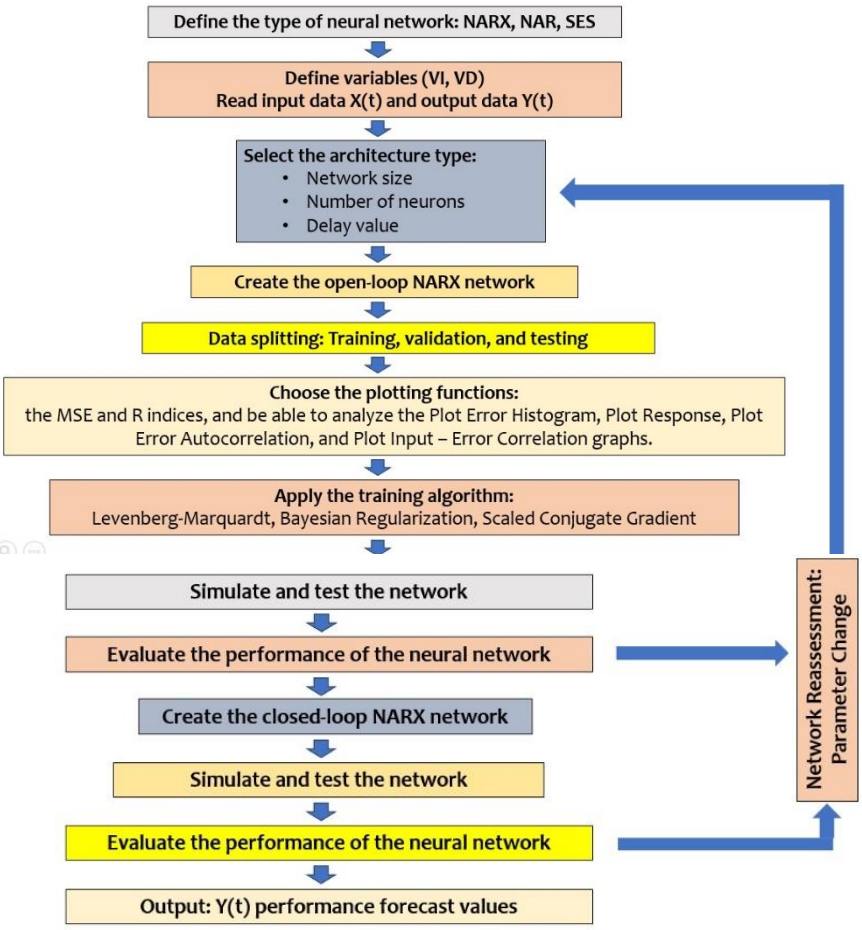


Figure 3. Block diagram of the NARX-ANN model

© 2025 ITMA. All rights reserved.

The graphs used by the NARX model allow us to efficiently validate network performance. The behavior of the error autocorrelation function (ACF) tells us how prediction errors relate over time to the network performance curve.

The ACFP graph evaluates the input error given by the cross-correlation and describes the correlation of errors with the input pattern; if all correlations are equal to zero, then the prediction model is perfect. To complement the analysis, the training statistics, Error Histogram, Regression, Time-Series Response, Error Autocorrelation, and InputError Cross-Correlation can be used.

For crop productivity performance forecasting, the factors to be taken into account are: Crop area and environment, production quantity, climatic factors, soil characteristics, as well as the genotype (G) and phenotype (F) of the plant. Therefore, in this context, this work proposes the application of a CRISP-DM model using NARX-ANN with IoT.

This proposal will allow associating the most important variables involved in plant growth. The main objective of this study is to improve the production yield forecasting of any crop. Finally, the proposed agricultural system based on CRISP-DM and NARX-ANN is shown in Figure 4.

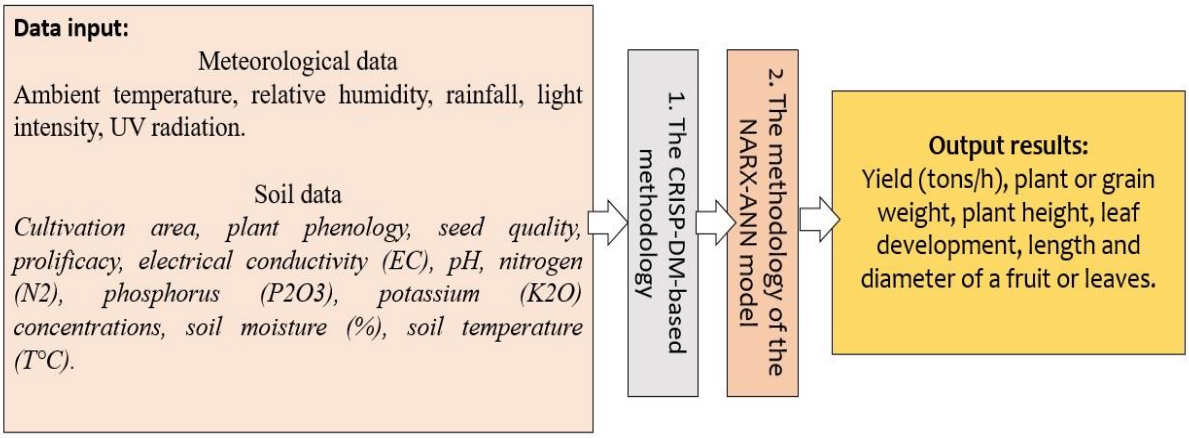


Figure 4. Final proposal of the CRISP-DM and MARX-ANN model

4. Conclusions and future work

This work aims to propose a CRISP-DM model using NARX-ANN to contribute to the use of modern technology in agriculture. This research incorporates various methodologies aimed at efficiently meeting crop forecasts for society's constant demand for food. These tools are extremely important today and will allow us to solve the prediction problem and thus help improve the yield of any crop.

© 2025 ITMA. All rights reserved.

Population growth over the years is a very important factor to consider, as the immediate effect would be that the relationship between the number of people and the amount of food available impacts people's nutrition, resulting in reduced food availability and, consequently, malnutrition. This work seeks to solve the problem of increasing the use of harvested areas for crops, thus preventing soil depletion worldwide.

It should be noted that a prototype smart module has already been designed using IoT, which uses mineral crop parameters as inputs to predict the yield of hybrid yellow corn in the city of Chilca, Lima, Peru. Once the training is complete, the module will be applied to different types of crops in the study area (cotton, mandarin and apple).

Conflicts of Interest: The author declares no conflict of interest.

References

[1] Informedel Senador Fidel Demédicis Hidalgo de la Segunda Conferencia Internacional sobre Nutrición y la Reunión de la Unión Interparlamentaria sobre Nutrición [Report of Senator Fidel Demédicis Hidalgo of the Second International Conference on Nutrition and the Meeting of the Interparliamentary Union on Nutrition] Roma, Italia, Roma, Italia, 2014. https://infosen.senado.gob.mx/sgsp/gaceta/62/3/2014-1215-1/assets/documentos/Informe_Roma.pdf

[2] Yulian Adalberto Sepulveda Casadiego, (2020) Inteligencia artificial y sus aplicaciones en la agricultura moderna [Artificial intelligence and its applications in modern agriculture], Conference – Paper, September

[3] Khumaidi, A. (2020). Data mining for predicting the amount of coffee production using CRISP-DM method. Jurnal Techno Nusa Mandiri, 17(1), 1-8.

[4] Talaviya, T., Shah, D., Patel, N., Yagnik, H., & Shah, M. (2020). Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. Artificial Intelligence in Agriculture, 4, 58-73.

[5] Altaleb, M., Deeken, H., & Hertzberg, J. (2022). A data mining process for building recommendation systems for agricultural machines based on big data.

[6] Mulatu, W. B., Robe, B., Robe, B., & Robe, B. (2020). Prediction of wheat rust diseases using data mining application. Open Access Library Journal, 7(09), 1.

[7] Oliveira, D. C. D., Barbosa, U. C., Bergland, A. C., Resende, O., & de Oliveira, D. E. (2022). G-Soja-website with prediction on soybean clasification using machine learning. Engenharia Agrícola, 42, e20210140.

[8] Pascual, I. D. L. A., Ramírez, J. L., & Ortiz, A. (2016). Métodos de Inteligencia Artificial para la predicción del rendimiento y calidad de gramíneas [Artificial Intelligence methods for predicting the performance and quality of grasses]. REDVET. Revista Electrónica de

© 2025 ITMA. All rights reserved.

Veterinaria, 17(12), 1-9.

[9] De la Piedra, M. C., Martínez, A. G. C., de Tuxtla Gutiérrez, I. T., Molina, J. A. L., Navarro, N. A. M., & Sánchez, J. O. G. (2016). Modelo neurodifuso para el control de humedad del suelo en cultivo hidropónico para la planta de tomate [Neurofuzzy model for soil moisture control in hydroponic cultivation for tomato plants]. Revista Tecnología Digital Vol, 6(1), 43-56.

[10] Saldaña-Robles, A., Jöns, S., Saldaña-Robles, N., & Damián-Ascencio, C. E. (2010). Pronósticos para el Rendimiento Del Trigo Usando Redes Neuronales Artificiales [Wheat Yield Forecasting Using Artificial Neural Networks]. Conciencia Tecnológica, (40), 25-28.

[11] Mendoza-Haro, I., & Marquetti-Nodarse, H. (2019). Redes Neuronales Artificiales: factores que determinan la cosecha de caña en la industria azucarera [Artificial Neural Networks: factors that determine the cane harvest in the sugar industry]. Revista Ciencia UNEMI, 12(29), 36-50.

[12] Cao, Q., Ewing, B. T., & Thompson, M. A. (2012). Forecasting wind speed with recurrent neural networks. European Journal of Operational Research, 221(1), 148-154.

[13] Cadenas, E., Rivera, W., Campos-Amezcua, R., & Heard, C. (2016). Wind speed prediction using a univariate ARIMA model and a multivariate NARX model. Energies, 9(2), 109.

[14] Zhang, J., Yin, Z., & Wang, R. (2017). Nonlinear dynamic classification of momentary mental workload using physiological features and NARX- model-based least-squares support vector machines. IEEE Transactions on Human-Machine Systems, 47(4), 536-549.

[15] Nedjah, N., & de Macedo Mourelle, L. (2016, November). Online pattern recognition for Portuguese phonemes using multi-layer perceptron combined with recurrent non-linear autoregressive neural networks with exogenous inputs. In 2016 IEEE Latin American conference on Computational Intelligence (LA-CCI) (pp. 1-6). IEEE.

[16] Wang, J., Geng, Y., Zhao, Q., Zhang, Y., Miao, Y., Yuan, X., & Zhang, W. (2021). Water quality prediction of water sources based on meteorological factors using the CA-NARX approach. Environmental Modeling & Assessment, 26, 529-541.

[17] Skolkovo Institute of Science and Technology (2019) Agricultura del futuro: Redes neuronales pueden predecir el crecimiento de las plantas, https://www.engormix.com/MA-agricultura/noticias/agricultura-futuro- redes-neuronales-t25801/p0.htm

[18] Arteaga, J. J. G., Zambrano, J. J. Z., Cevallos, R. A., & Romero, W. D. Z. (2020). Predicción del rendimiento de cultivos agrícolas usando aprendizaje automático [Predicting Agricultural Crop Yield Using Machine Learning]. Revista Arbitrada Interdisciplinaria Koinonía, 5(2), 144-160.

[19] Thilakarathne, N. N., Bakar, M. S. A., Abas, P. E., & Yassin, H. (2022). A cloud enabled crop recommendation platform for machine learning- driven precision farming. Sensors, 22(16),6299

[20] Wirth R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of

© 2025 ITMA. All rights reserved.

knowledge discovery and data mining (Vol. 1, pp. 29-39).

[21] Hocaoglu F. O., & Karanfil, F. (2013). A time series-based approach for renewable energy modeling. Renewable and Sustainable Energy Reviews, 28, 204-214.

[22] Buitrago, J., & Asfour, S. (2017). Short-term forecasting of electric loads using nonlinear autoregressive artificial neural networks with exogenous vector inputs. Energies, 10(1), 40.

© 2025 ITMA. All rights reserved.