ISSN 2007-9737

Selection of the Decision Variables for the Habanero Chili Peppers (Capsicum chinense Jacq.) Using Machine Learning

Blanca C. López-Ramírez[1], Francisco Chablé-Moreno[2], Francisco Cervantes-Ortiz[2]

[1] Tecnológico Nacional de Mexico/IT.Roque,
Department of Systems and Computing,
Mexico

[2] Tecnológico Nacional de Mexico/IT.Roque,
Department of Agricultural Sciences,
Mexico

{blanca.lr, francisco.cm, francisco.co}@roque.tecnm.mx

Abstract. Data science is an area that allows a gathering of data from several prospects, being at once, collaborative and multidisciplinary. It is an area so promising and open to research from different problems, including the challenges of agronomy science throughout the study of the exploitation of database knowledge. In this work, we will study if it is possible to identify some determined variable that could allow to response to the questions, ¿Is it possible to know the genotype from a habanero pepper plant, knowing some plant measure? also, ¿Is it possible to identify the yield through the plant height? The goal is to identify the proficiency of each one of the involved areas on the preparation, processing, and database, as the necessary methods and tools to gather relevant information to the expertise; derivable from techniques as Neural Networks Algorithms and Statistics. The outcome earned, prove even tho the statistics operations revealed results by a descriptive category besides a predictive one; The Neural Networks Algorithms find results within the prescriptive category, displayed on work and that represent a very interesting answer resulting from applying questions that were not obviously in basic analysis.

Keywords. Data analytics, rescriptive analysis, neural networks algorithms, post-decision, state variable.

1 Introduction

This work describes the study of the behavior of growing variables from 8 habanero chili genotype(Capsicum chinense Jacq.) through machine learning. The objective is to find the existence of a relationship between the phenomenological development variables and yield. It is interesting to know if the yield is directly related to the plant height or if the leaf growth is related to the regional weather.

Gathering this information, different genotypes had been cultivated under the same weather, and measuring and data collection too. The information that this research presents, not only implies raw data of variables but also presents a descriptive and prescriptive data analysis, applying data science techniques as statistics and neural networks.

Agricultural production has had 3 stages; the first one has been when the labor was rich and intense until 1920, later, the second stage when the industrial revolution provide heavy machine to the fieldwork and the seed study to improve farming, finally, the third one began since 2010 until today, which it is known as the Agriculture 3.0, this stage is considered by the use of innovative technology for different studies that allow a genetic and phenologic improvement of farms assisting to take decisions based on data analysis and data obtained from external sources.

The good decisions based on agricultural information give higher productivity, practicing

sustainability even helping to provide transparency to consumers who may want to know more about their foods [8]. Kumar [17] in 2017, mentions that one of the aspects which impacts the livelihoods and rural prosperity is agricultural management. The options of agricultural growth and connections with farm investment are the key element to agricultural development strategy.

In 2019, Mexico was in third place in the agricultural sector in America Latina and 11th place at the global level. Agriculture is the mainstay sector in the country's economy which wields a socio-economic impulse; with the culture of continuous improvement and the incorporation of technologies on the field, Mexico in the last year became within the 10 first agri-food products exporter [20]. The importance of habanero chili production has been used by different fields; in the pharmaceutic industry, according to Lopez-Puc [19] who studies the implementation of biotechnology in gathering varieties of habanero chili by its high capsaicinoides and capsaicina content, which are the raw material to the ointment production that relieves arthritis pain. In the agro-alimentary industry, for its proceeding in several foods that contain, habanero chili has elevated vitamin and minerals index [29].

It is used as an electrical system and irrigation coating to avoid rodents attacks. Lopez [19] also indicates in a significate way that, habanero chili is one of the least harmful chilies and is considered powerful healing, besides, it helps with gastritis and hemorrhoid problems. The 80% from chili habanero production is marked as dried fruit and 20% as sauce production like pasta and dehydrated [7]. Habanero chili it has been cultivated in the Yucatan's peninsular and is the leading production all over Latinoamerica in 2018 [30, 35]. Like farming, it has strong economic importance to vegetable producers in Yucatan state: remains in second place after the tomato farming concerning ground farming and, due to their demands to weather, the use of the controlled environment is higher [18].

There have been already published research documents terms as agronomics, climatic, genetic, chemical, among others as Santana et al. [31]. who investigated the formation of shoots of habanero chili plants with supplements in the shoots applying variables concentration of kinetina, benciladenina, and tidiazuron. On the other hand, as times go by and the social requirements for having data control, its management and generation of information in areas like medicine, astronomy, chemistry, biology, among others, has been supporting technological and specialization challenges cause of the data volume demands, the idea of gathering not only useful and in time information, but the knowledge acquired [24, 32].

However, with technological advance and social requirements, the statistic has been involved by its essence which is identified as science and classified depending on the purpose of data in statistic descriptive, demographic, probabilistic, or administrative in order to identify a phenomenon, concept, or incident [13, 3, 34]. In 1999, Witten et al. has taken the first steps to design learning methods through data exploration techniques, applying systems through data analysis with the WEKA tool, using the RJ48 algorithm, they isolate qualified attributes for a market and price of mushrooms labeling. Witten et al. claim to support a minimum work of programming to achieve the learning.

Majumdar [20] in 2017 analyzed the data sets from different public databases through data mining techniques applied in the agronomic area, particularly, clustering techniques. The work was done with wheat farming. The groups were divided by districts assessing different variables like atmospheric humidity, pH, temperature, among others. The efficiency of the annual crop was obtained by linear regression. As well, Amato et al. [2] in 2013 make sure about the nonparametric techniques usually exceed the parametric. Amarato realized research with discriminating data analysis tools that adapt to images hiperespectrality to identify the use of farmed agriculture soil. The good results he acquired for the classification using the previous transformation of data and he remarked that the false positives could be prevented, also using a group of data little training and reaching robustness and capacity to identify the categories in its study. Conversely, Kanahal [16] in 2019 studied the needs of farmers basing on a questionnaire applied

to experts in consideration of different variables as revenue, farm size, and an agricultural occupation for modeling, and the adoption of a predictive direction system. In his work, he mentions that the farm size, revenue level, and agriculture occupation are important facts in the adoption modeling and the applying of the GPS system.

Issad [1] in 2019 submits a revision of the implementation and study of Data Mining Techniques in agriculture. In this work, he mentions the Padalulu's et al. study, his proposal is the estimate of fertilizer and irrigation, as well as Perea et al. [12], in 2019 they had worked on decision trees and genetic algorithms to predict irrigation events. The wheat production was reinforced by satellite climatic studies in Australia, creating empiric models to predict the efficiency, however, Cai et al. [5] in 2019 observed the benefit of not only gathering the climatic data via satellite but they compared climatic data already acquired. They had used the regression method known as LASSO and three learning methods to build prediction empiric models. They claim that the method based on automatic learning overcome the regression method. And due to the successful work achieved, their suggestion is to apply the same techniques to other crops.

Rajeswari [27] in 2017, create a model for data analysis bigdata through the cloud, he analyzed variables as fertilization, growth, market, as well as requirements for growth. Initially, his proposal was the extraction of information through the digital interconnection of routine objects with the internet (IoT Devices), storing it in one of the databases in the cloud, subsequently, pre-processing and labeling were made, getting a selection of attributes completing the application of a pattern algorithm of MapReduce prediction.

It is well known that problems in the real world are even complex, nonlinear, and a stage that could be charged with multivariable uncertainties, multimodal, discontinuous, or exponential. This is an immersion to get to the point; studying a data collection or getting a processing and analysis algorithm, does not ensure the information gathered [35, 6]. Although it is been a long time since the statistic was considered by the specialists as an individual science in the research study, like Stigler [32] in 1986 mentioned in his book the requirement of intervention in some disciplines in the employment of the data analysis management, consequently, this premise has been fundamental support to get to the source called the Data Science [4, 10]. Data Science is an interdisciplinary field for the research generation that is useful, relevant, and innovative, converted into knowledge, the multidisciplinary thought from fields as programming, communication, management, sociology, and mastery of topical, are intended to a reflection of Data Science [35, 6, 34, 33].

The researches already realized in the agronomic area apply theories after the data collection, either, using data in a computing system to produce information with a specific target, also there are works where technologies and smart algorithms have been applied for data analysis like [22, 36, 15]. However, studies are ensuring that the model system is important on data knowledge, this is, the recognition of involved disciplines in a solution to make a collaborative work and systematic that performs a succeed on the project [26].

The expansion of the system consists of several phases: Delay phase (phase "lag) is a short period of adaptation or an increasing of startup to the half; transition phase, accelerated growth, which leads to exponential growth (logarithm phase); negative acceleration phase, imbalance phase, which leads to a stationary phase: characterized by the net coefficient of increasing declares null. Based on studies realized by Hernandez et al. [14], in grafted cucumber plant had been observed that fruit weight by plant correlated positively and significantly with plant height r=0.63*, stem diameter r=0.59* and leaf number r=0.54*, revealing these characters are important for this production system.

While Estrada et al. [9], evaluating tomato genotypes, found that the leaf length correlated with the fruit diameter, the highest correlation match to the length and wide of leaf $r = 0.96$, also highlight, by the number of leaves by plant, the number of clusters by plant, number of fruits by cluster, the average weight of fruits, and the fruit diameter, which means that are correlated. Other studies, like Pinedo [25] realized, studying the

genetic improvement of Camu-camu, had found a correlation between the leaf length, petiole length, and the fruit weight, that is a specie with high vitamin C content; as equal as Nieto et al. [23] evaluating the chirimoya selections, they had found that the limbo area, the leaf perimeter, the petiole length, and the longitudinal axis, were variables that had higher correlations with a highly significative level $(P \leq 0.01)$, proving its dependence; in this studies that have been realized in different species, its been determined a high correlation besides submitting a logistic type of growing that match absolutely in the gathered results on the current investigation.

This work suggests a study of determinate variables of a data collection from habanero pepper plants settled at the beginning of Data Science. It is analyzed the data representation complex, the vision of behavior over and as time goes by, preserves the mastery of knowledge of the data topic and its generalities on the experimental study.

As a result of our work, we have found that is possible to identify the relation between determinant variables. Even better, the Neural Network algorithm proved to be a great support widget in the research of experimental variables, emphasizing the relevance of the variable which has the reply to particular incidents, in other words, a prescriptive analysis of data has been developed with great success. The relevance of this study lies that no work has studied the variable knowledge relation of habanero pepper through computational algorithm techniques.

2 Related Work

The applied methodology was through the Data Science philosophy, where its discipline simulates a coordinate and systematic mechanism. In this section, the first steps were described, as well as the experimental design of the plants and their treatment, which is critical for the experimental replication as well as the procurement process and the data management of the experiment at the 3 harvest events conclusion. In addition, the Artificial Neural Network used for knowledge analysis has been presented.

Table 1. Genotypes of C. chinense in greenhouse experiments

Gen	Site	Name
G2	PGHBN2-130217-C1	Gliese 204 Chato1
G3	PGHBN3-130217-C1	Gliese 204 Chato1
G4	PGHBN4-130217-C1	Betelgeuse 2
G5	PGHBN5-130217-C1	Betelgeuse 2
G6	PGHBN6-130217-C1	Rigel 1
G7	PGHBN7-130217-C1	Gliese 204 Chato1
G8	PGHBN8-130217-C1	Rigel 1
G9	PGHBN9-130217-C1	Gliese 204 Chato1

Fig. 1. Growing habanero pepper plants

2.1 Experimental Crop Design

The experimental design was completely random with 8 treatments and 23 replications, getting 134 experimental units. The assessed variables in the habanero pepper plants were: plant height, leaf width, leaf length, bloom time al first and second shoot, and the second shoot, bloom time, number of fruits, and the greenhouse weather.

The selection of the seed was the beginning, until obtaining data from three harvests. In Table 1 the genotypes under study for their adaptation are presented. The conditions used in the greenhouse and the treatment of the plants used for the study variables are shown.

The measurement of the variables was performed in centimeters below describe each of them:

— Plant height (API). Seized from the bottom of the stem to the highest apex of the plant in centimeters.

— Number of primary stems (NTP). The number of secondary limbs of the leaf from the core stem was quantified.

— Number of secondary stems (NTS). The number of tertiary stems was quantified based on their secondary stem.

— Leaf length (LHj). The leaf development was quantified in the core rib at different sampling dates, using as fundamental the start of leaf limb.

— Leaf width (Anh). The leaf plant evolution was evaluated from its middle segment in terms of different sampling dates meanwhile its growth.

— Fruit length (LF) The fruit length will be measured.

— Fruit width (AF). By Measuring the widest fruit section, the width will be determined.

— Environmental temperature, in the greenhouse.

— Closed bloom (FC). To evaluate this variable the flower bud has to emerge on the three emergency days, closed bloom (days).

— Semi-open bloom (FSA). It had been qualified when the bloom is 30% open, but without pollination, the views as semi-closed bloom were selected.

— Open bloom (FA). It will be observed if the bloom is completely open and the anthers are dehiscences (days).

— Pollination (Pol). The bloom condition will be monitored if it bears fruit, the result will be recognized as pollination.

— Fruit Color (CF). The color might be considered based on the pigmentation acquired over the maturation process.

— green-dark fruit (CVOF). The fruit will reach this color as of the 20 days of pollination.

Fig. 2. Identification of habanero pepper plants

— Light-green color (CVCIF). After 3 or 4 weeks later of pollination change this color (visual).

— Light-green with red or orange pigmentation according to its genotype (CPRN). In a period of 4.5 to 5.5 weeks of pollination, this is the pigmentation produced in the fruit (Visual).

2.2 Design and Analysis of Data

With globalization advances and global warming, the agronomic field is concerned by different purposes of the farming process such as: cost reduction, resource optimization, and/or the genetic improvement of the product. The habanero pepper experts who have been contributing to this project, not only are interested in weather adaptation in protected crop to improve efficiency and production, but also the phenomenological condition from variables of the habanero pepper plant.

In the first analysis phase, the registered data were achieved and reviewed in numerical value with a spreadsheet. The data were evaluated and discerned through a set of continuous and discrete data values for a first descriptive diagnostic that were determined by the information gathered.

For the data analysis, a statistical function was investigated to reach the descriptive analysis as Pearson's correlation test between the variable of height-leaf, height-stem, and finally, fruit and plant of the three crop periods.

2.2.1 Data Modeling

A set of data values was used carefully evaluated with the intention of noise-canceling which might harm the analysis process, irrelevant variables were isolated, in other words, those that don not have a coincidence relation between them. Once the database has been already reviewed, an entraining function was applied to a set of data preparing it and using it in a Neural Network Algorithm. The tests were made once the model has been trained over the set of data to reach the accurate model.

The tests that were made in this stage had the intention to create a machine learning model to explain if the plant height might have an impact on the production, or any determinant variable. Through the multiple linear regression function with function 1, a diagnostic and descriptive model has been presented describing the issue and what can be achieved through an set of exposed data:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon, \tag{1}$$

whereβ_0 is the independent term and Y is the desired value.β_1, β_2,...,β_p are the biased coefficients of regression. ϵ is the perceived error due to no controlled variables. Then, the model of multiple linear regression is interpreted with function 2:

$$\begin{aligned} R' = \beta_0 + \beta_1 Higth + \beta_2 Leaf1L + \\ \beta_3 Leaf1 + \beta_4 Leaf2L + \beta_5 Leaf2w \\ \beta_6 Temperature + \beta_7 Plants + \epsilon. \end{aligned} \tag{2}$$

Statistics models help us to qualified the validation and reliability of the gathered data. The study of correlation has been used to know the relation level during the variables growing in the study, another one, is the regression to identify the dependent and independent variable of the leaf and fruit growing, to evaluate and validate the relationship between both.

Among the logarithmic and exponential functional models, a variant of the regression of each genotype was analyzed by genotype and to know the data prove. Then, an analysis of lineal regression had been performed providing the dependency between the expressed data, in other words, the dependent and independent variables.

Once the behavior of the dependent variables has been recognized, machine learning through supervised learning has been suggested. The data already shown have been analyzed as the dependence of the variable to the response variable. In order to know the effect of temperature on plant growth or a variable that directly affects, or even, to identify the growth-production based on the genotype, the paradigm of the Artificial Neural Network was presented which supports the optimization and prediction of the reply variable.

2.2.2 Machine Learning: Artificial Neural Network

Since 1888, Ramón and Cajal [28] prove that the nervous system is formed by interconnected neurons that learn the influence of external information. There are different ways to learn as they are: through new connections, through connection breaks, links between neurons, or the neurons reproduction. Artificial paradigm of Neuronal Networks simulate the biological Neuronal Network, the standard model arises in 1986 by Rumelhart and McClelland as defined in function 3, we become aware of Artificial Neural Network in [21]:

$$f_i \left(\sum_{i=1}^{n} w_{ij} x_j - \theta_i \right), \tag{3}$$

where x_j represents an entry and w_{ij} the synaptic weights, the propagation rule used is to mix linearly the entries and the synaptic weights depending on the function 4:

$$h_i(x_1, \ldots, x_n, w_{i_1}, \ldots, w_{i_n}) = \sum_{i=1}^{n} w_{ij} x_j. \tag{4}$$

The Adaline Artificial Neural Network (AANN) was used in Adalina, introduced by Winrow and Hoff in 1960, and the name comes from ADAptative LInear Neuron, operational input of continuous values intended to classify data. Different from other neuronal networks, one additional parameter

ISSN 2007-9737

Algorithm 1 ANN training

```
Train_ANN (A,D) [ANN: Network Neural, D:
Data set, IL: Input Layer,
OL:Output Layer, e:error]
w randomly chosen weights --> ANN  [-1,1]
Repeat
    For (x,c) de D
        Insert x en IL
        Propagate the values of
             neurons from the IL (forward)
        Read the output y(x) = OL
        Calculate the(x) = |c-y(x)|
        Use error e(x) to adjust w
             and minimize the x error
Return w
```

has been incorporated called bias giving a free degree to the model (see function 3):

$$E(W) = \frac{1}{2}\sum_{r=1}^{N}\sum_{i=1}^{m}(c_i^r - y_i^r)^2. \tag{5}$$

The notable thing about this model is the learning rule, identified as LMS (Least Mean Square), learning is regulated by the weights selected to the error made by the neuron. With the cost function, network optimization has been acquired (see function 5). And the Artificial Neural Network training steps are in the algorithm 1.

3 Results and Discussions

The study presents a close relationship between the fruit and leaf variables, as we can see in Figure 3, where the value of the determining variables has a logistic behavior. During the first data gathering, the fruit and leaves variables show fast germination and tend to converge towards slower germination.Coefficient of determination is $R^2 = 0.94$ of the LHj and AnH variables and the length growth of leaf which is already explained in the width leaf development.

Difference between leaf 1 and leaf 2 growing by its position in the plant could be considered, due to

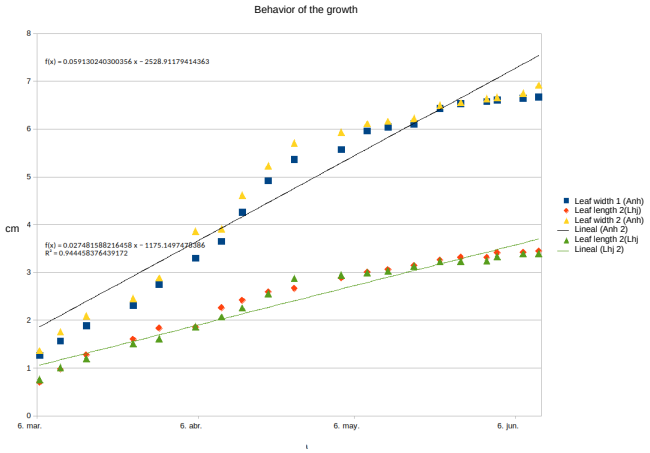

Fig. 3. Behavior of the growth rate of the variables leaf width and fruit width of the Rigel genotype and their relationship as a function of temperature

the apical dominance that is in the top section, its development is faster than the bottom of the plant.

The germination average of the habanero pepper leaf of genotype Rigel is 4.71 cm, considering that can reach the beginning of the growth of the leaf development, this variable has a variant value of 3.71, this reply is because the data of the variables was scattered and with more variability. The behavior of the width leaves in the same genotype was an average of 2.53 cm, which was found in the same days of the length germination, the variance is 0.74 lower due to the data is closer in relation to the average value of the width leaf (see Table 2).

A factor that influenced the growth rate and the production of the habanero pepper crop was the temperature. The range of temperature values in the greenhouse is between 48.75ºF and 91.85 ºF.

In Figure 4 the increase of the leaf growth and the fruit in temperature function have been observed. On the axis of the independent variables, the values of the average temperature have been considered and it is relational with the growing rate of the leaf length at that moment, and the fruit length is presented at the same time.

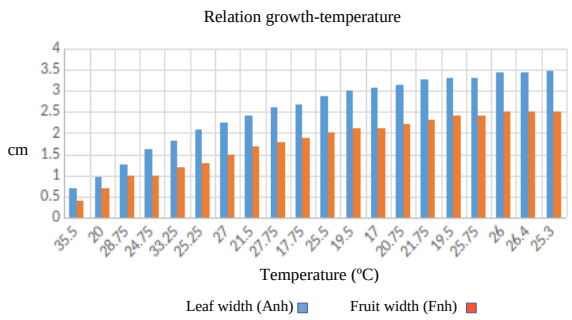

Fig. 4. Behavior of the growth rate for the variables width of the plant leaf and the width of the fruit of the Rigel genotype and their relationship as a function of temperature

Table 2. Determination coefficients

Groups	Rep	SC	$\overline{x}$	S^2
Large(LH)	20	94.24	4.71	3.71
Width(AnH)	20	50.67	2.53	0.74
Fruit	20	50.67	2.53	0.74

Table 3. Coefficients of determination and standard deviation to variables fruit width and plant leaf from chilli Gliese genotype and Riegel

Study factor	Betelgeuse	Riegel	Gliese
Coef. of det. R^2	0.91	0.95	0.95
Adjusted R-squared	0.91	0.95	0.95
Error	0.2 1	0.18	0.17

The increase in each variable is distinguished, even though the fruit length is slower.

In the linear regression analysis between the variables length leaf width (AnH) and fruit width (AF) a highly significant correlation was obtained, where the sum of squares was 9.88 and the critical value of F was $9.1898E-12$, regarding the length of the fruit and the leaf, the sum of squares is 12.18 and the critical value of F is $2.54E-14$. Factors study genotypes can be seen in Table 3.

The results of the multiple linear regression technique showed a significant association in those

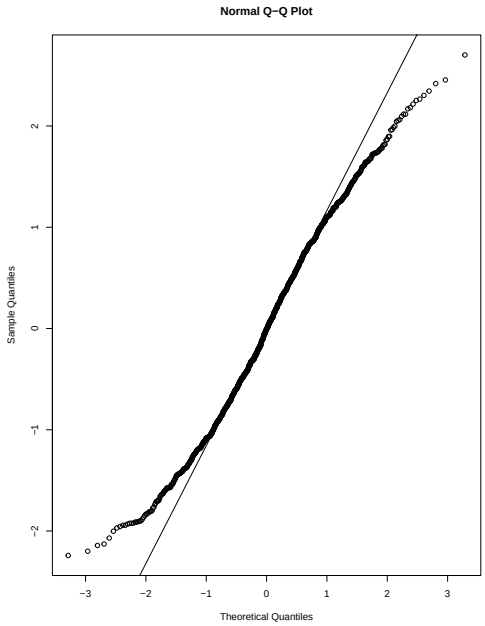

Fig. 5. Residual diagnostics in the data set

predictive variables such as height, temperature and type of pepper genotype, Figure 6.

On the other hand, the estimated line obtained from the regression diagnosis can be observed in the equation 6, the $Leaf2L$ variable was eliminated because it does not present significant values in the response variable.

$$R' = -1.04 - 0.048 * Higth - 0.18925 * Leaf1L + 0.493 * Leaf1w - 0.447 * Leaf2w - 0.03805 * Temperature + 0.212 * Plants + \epsilon. \quad (6)$$

Furthermore, these results showed a growth trend with respect to yield for each treated genotype (see Figure 7). The normality of the studied data set can be observed in Figure 5, this represents a long-tail behavior in the residuals, which means that the proposed model must be studied.

Although a statistical model was found and this represents a proposed solution, the problem is complex as can be seen in Figure 7, where it only shows the correlation of two variables.

It is necessary and advisable to continue with the study to find the information of answers to questions as; If there is a relationship between the determining variables type of plant and yield, or also pre-write which variable could identify the type of plant in question, before having the fruit, that is, without having to invest in time.

One of the first Artificial Neural Network experiments was to identify if the value of the length of the leaf can help predict the height of the plant. Among the parameters that he included in the algorithm were the variable $h1$ and the response variable $height$, the error function was also used the equation 7 and the activation function applied was 8. The results showed a correspondence between both variables:

$$err(x, y) = \frac{1}{2} * (y - x)^2, \tag{7}$$

$$f(x) = \frac{1}{1 + exp(-x)}. \tag{8}$$

Another interesting piece of data is that the Garson [11] method of the Artificial Neural Network paradigm was used in the search for response variables among all the variables that were included in the study. The method made a significant connection between all nodes, giving a weighting value to each connection and, later, the technique considered a disconnection between the evaluation of weights on variables that do not have a significant relationship that is evaluated by the paradigm, since the connections were scaled and disarticulated, those variables that were not relevant to find the response variable were eliminated.

Figure 8 shows the variables of the data set that obtained a significant level of importance (x's axis) in relation to the plant genotype (y's axis), this predictive analysis will allow the expert to decipher the type of habanero chili without waiting for the plant to expose the fruit. In the same way, the study was considered very significant because the power of a neural network was verified in a prescriptive-level data analysis. Another example that showed its importance was the identification of an explanatory variable for performance, the result of the Artificial Neural Network can be seen in Figure 9.

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.1585 -1.7914 -0.0087  1.8129  6.2121

Coefficients:
             Estimate   Std. Error  t value   Pr(>|t|)
(Intercept) -1.03911    1.05112    -0.989    0.323115
No_data      1.34144    0.02048    65.497    < 2e-16 ***
lants        0.21267    0.03386    6.281     5.05e-10 ***
Leaf1L      -0.18925    0.08247    -2.295    0.021957 *
Leaf1w       0.49331    0.12257    4.025     6.14e-05 ***
Leaf2L      -0.05921    0.07744    -0.765    0.444726
Leaf2w      -0.44678    0.12866    -3.473    0.000538 ***
Temp        -0.03805    0.01313    -2.897    0.003854 **
Height      -0.04789    0.01519    -3.153    0.001667 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.307 on 979 degrees of freedom
Multiple R-squared:  0.9169,  Adjusted R-squared:  0.9162
F-statistic:  1350 on 8 and 979 DF,  p-value: < 2.2e-16
```

Fig. 6. Results of data analysis

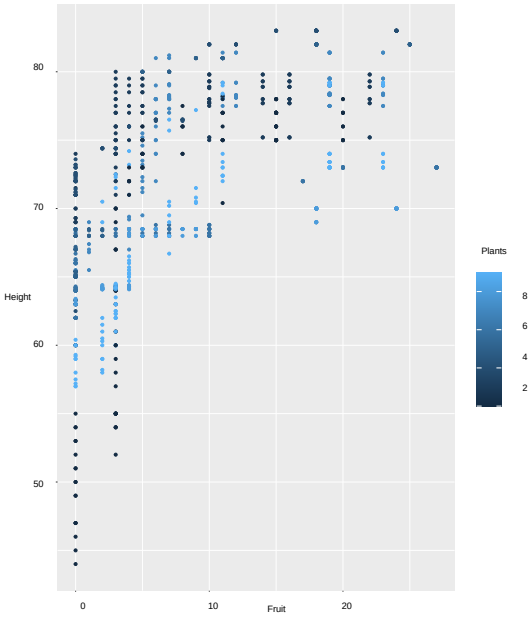

Fig. 7. Graphical representation of the dispersion between variables

ISSN 2007-9737

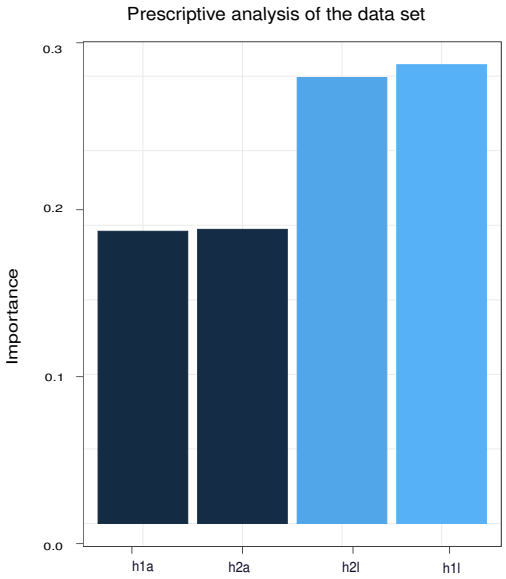

Fig. 8. ANN paradigm in the data set

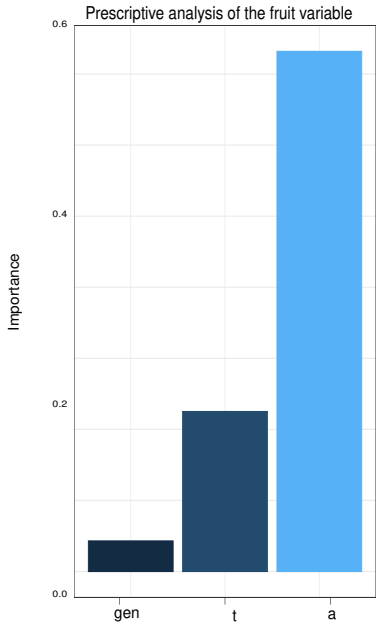

Fig. 9. Importance of the explanatory variable for the response variable Fruits

4 Conclusion and Future Work

The study of experimental data in the habanero pepper crop was performed by a multi-disciplinary team, where team members whose area of expertise is computer science, interacted with experts in agricultural science. In this research, the experimental data obtained was analyzed by statistical and data mining techniques with the main purpose of finding interesting information beyond the prediction that a multiple regression analysis could produce.

In fact, the linear regression model appears to be insufficient for finding an explanatory variable that fully determines the response variable. We found that the Artificial Neural Network approach was quite valuable for carrying out our analysis. This work is a process that we find attractive for the analysis of multivariable linear and not linear data, which are required for obtaining a diagnostic that is both, predictive and prescriptive.

The studies carried out show that determining variables such as plant height, leaf length, leaf width and fruit have a positive and significant correlation, as did Hernandez et al. [14] with they study of the cucumber plant.

On the other hand, Estrada [9] in his study of tomato genotype found a correspondence relationship between the values of the fruit and the plant with respect to the data of the clusters, we found a strong relationship between the fruit and the values of the plant's leaf $r^2 = 0.94$.

The study of the data of a habanero pepper crop was carried out in a multidisciplinary study, the data obtained was analyzed by statistical and data mining techniques in order to find interesting information beyond a multiple regression study.

The regression model does not seem sufficient to find an explanatory variable that determines the response variable, being the ANN technique very helpful for this objective.

Acknowledgments

We would like to thank to TecNM/IT Roque for the support to carry out the project in the greenhouse and the facilities. Almost, thanks to the MC.

Chablé for his advice and support in conducting the machine learning study. On the other hand, to Román García for the support in the translation. This work is funded by agronomy department of TecNM/IT Roque.

References

1. **Ait, H., Aoudjit, R., Rodrigues, J. (2019).** A comprehensive review of data mining techniques in smart agriculture. Engineering in Agriculture, Environment and Food, Vol. 12.
2. **Antoniadis, A., Carfora, M. F., Colandrea, P., Cuomo, V., Franzese, M., Pignatti, S., Serio, C., Amato, U. (2013).** Statistical classification for assessing prisma hyperspectral potential for agricultural land use. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 6.
3. **Badii, M., Araiza, L. A., Guillén, A. (2010).** Esenciales de la estadística: Un acercamiento descriptivo (essentials of statistics: A descriptive approach). df, pp. 4.
4. **Ben, D. (2017).** Big data and data science: A critical review of issues for educational research. British Journal of Educational Technology.
5. **Cai, Y., Guan, K., Lobell, D., Potgieter, A., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., Peng, B. (2019).** Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agricultural and Forest Meteorology, Vol. 274, pp. 144–159.
6. **Cao, L. (2017).** Data science: A comprehensive overview. ACM Comput. Surv., Vol. 50, No. 3.
7. **CastellónMartínez, E., ChávezServia, J., Carrillo-Rodríguez, J., Vera-Guzmán, A. (2017).** Preferencias de consumo de chiles (capsicum annuum l.) nativos en los valles centrales de Oaxaca. Revista fitotecnia, mexicana, Vol. 35, No. 5, pp. 27–35.
8. **Creutzberg, G. (2015).** Agriculture 3.0: A New Paradigm for Agriculture. Nufflier Canada.
9. **Estrada, Y., Lescay, Y., Vázquez, F., Celeiro, F. (2012).** Variabilidad genética y correlaciones fenotípicas en germoplasma de tomate (solanum lycopersicum l). Granma Ciencia, Vol. 16.
10. **Gibson, D., Ifenthaler, D. (2017).** Preparing the Next Generation of Education Researchers for Big Data in Higher Education, chapter 3. Springer, pp. 29–42.
11. **Goh, A. (1995).** Back-propagation neural networks for modeling complex systems. Artificial Intelligence in Engineering, Vol. 9, No. 3, pp. 43–151.
12. **González, R., Camacho, E., Montesinos, P., Rodríguez, J. (2019).** Prediction of irrigation event occurrence at farm level using optimal decision trees. Computers and Electronics in Agriculture, Vol. 157, pp. 180.
13. **Gutierrez, C. (1994).** Filosofía de la estadística. Universitat de Valencia Servei de Publicacions, EU, 1 edition.
14. **Hernandez, Z., Sahagun-Castellanos, J., Espinosa-Robles, P., Colinas-León, M. T., Rodriguez Perez, J. E. (2014).** Efecto del patrón en el rendimiento y tamaño de fruto en pepino injertado. Revista fitotecnia mexicana, Vol. 37, pp. 41–47.
15. **Jones, J. W., Antle, J. M., Basso, B., Boote, K., Conant, R. T., Foster, I., Charles, G., Herrero, M., Howitt, R., Janssen, S., Keating, B., Munoz-Carpena, R., Porter, C., Rosenzweig, C., Wheeler, T. (2017).** Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science. Agricultural Systems, Vol. 155, pp. 269–288.
16. **Khanal, R., M, A., Lambert, D., Paudel, K. (2019).** Modeling post adoption decision in precision agriculture: A Bayesian approach. Computers and Electronics in Agriculture, Vol. 162, pp. 466–474.
17. **Kumar, M., Nagar, M. (2017).** Big data analytics in agriculture and distribution channel. pp. 384–387.
18. **Lecona-Guzmán, C. (2017).** Mejoramiento genético de chile habanero: selección y registro de variedades mejoradas. Revista fitotecnia, mexicana, Vol. 35, No. 5, pp. 27–35.
19. **López-Puc, G., Canto-Flick, A. (2009).** El reto biotecnológico del chile habanero. Ciencia, Vol. 60, No. 3, pp. 30–35.
20. **Majumdar, J., Naraseeyappa, S., Ankalaki, S. (2017).** Analysis of agriculture data using data mining techniques: application of big data. Journal of Big Data, Vol. 4, No. 20.

21. **Minsky, M., Papert, S. (1969).** Perceptrons: An Introduction to Computational Geometry. MIT Press, Cambridge, MA, USA.

22. **Muthurasu, N., Sahithyan, S., Aravind, M. T., RamanagiriBharathan, A. (2018).** A prediction system for farmers to enhance the agriculture yield using cognitive data science. Journal of Advanced Research in Computer Science.

23. **Nieto, R., Andre, J., Barrientos-Priego, A., Martínez-Damián, M., González-Andrés, F., Segura, S., Gallegos-Vázquez, C. (2003).** Variación morfológico de la hoja del chirimoyo. Revista Chapingo, Serie Ciencias Forestales y del Ambiente, Vol. 10, pp. 103–110.

24. **Perero, M. (1995).** Historia e historias de matemáticas. Grupo Editorial Iberoamericano, México, 1 edition.

25. **Pinedo, M. (2012).** Correlation and heritability analysis in the genetic improvement of camu-camu. Scientia agropecuaria, Vol. 3, pp. 23–28.

26. **Pivoto, D., Waquil, P., Talamini, E., Finocchio, C. P. S., Dalla Corte, V. F., de Vargas Mores, G. (2018).** Scientific development of smart farming technologies and their application in brazil. Information Processing in Agriculture, Vol. 5, No. 1, pp. 21–32.

27. **Rajeswari, S., Suthendran, K., Rajakumar, k. (2018).** A smart agricultural model by integrating iot, mobile and cloud-based big data analytics. International Journal of Pure and Applied Mathematics, Vol. 118, pp. 365–369.

28. **Ramon, C. (1899).** Textura del sistema nervioso del hombre y de los vertebrados. Moya, Madrid, 1 edition.

29. **Rozete, M. (2019).** Caracterización fitoquímica y sensorial de variedades de chile habanero (Capsicum chinense Jacq.) Yucatán. Ph.D. thesis, Centro de Investigación Científica de Yucatán, A.C. Posgrado en Ciencias Biológica, Yucatán.

30. **Sagarpa (2020).** Planeación agrícola nacional 20172030 chiles y pigmentos mexicanos.

31. **Santana-Buzzy, N., Canto-Flick, A., Barahona-Pérez, F., Montalvo-Peniche, M., Zapata-Casillo, P., Gutierrez-Alonso, O. (2005).** Regeneration of habanero pepper (capsicum chinense jacq.) via organogenesis. HortScience: a publication of the American Society for Horticultural Science, Vol. 40, pp. 1829–1831.

32. **Stigler, S. (1986).** The History of Statistics: The Measurement of Uncertainty before 1900. The Belknap Press of Harvard University Press, EU, 1 edition.

33. **Tukey, J. W. (1962).** The future of data analysis. Ann. Math. Statist., Vol. 33, No. 1, pp. 1–67.

34. **Tukey, J. W. (1977).** Exploratory Data Analysis. Behavioral Science: Quantitative Methods. Addison-Wesley, Reading, Mass.

35. **Vasquez, B. A. (1999).** Ciencia de datos para gente sociable.

36. **Wolfert, S., Ge, L., Verdouw, C., J, B. M. (2017).** Big data in smart farming – a review. Agricultural Systems, Vol. 153, pp. 69–80.

Article received on 11/12/2021; accepted on 12/04/2022. Corresponding author is Blanca C. López-Ramírez.