Available online at http://bajas.edu.iq
https://doi.org/10.37077/25200860.2021.34.1.01
College of Agriculture, University of Basrah

Basrah Journal of Agricultural Sciences

ISSN 1814 – 5868 Basrah J. Agric. Sci., 34(1): 1-13, 2021 E-ISSN: 2520-0860

The Application of the CART and CHIAD Algorithms in Sugar Beet Yield Prediction

Nasim Monjezi

Department of Biosystems Engineering, Faculty of Agriculture, Shahid Chamran University of Ahvaz, Ahvaz, Iran

***Corresponding author-mail:** n.monjezi@scu.ac.ir
Received 25 September 2020; Accepted 18 December 2020; Available online 4 February 2021

Abstract: Yield prediction is a very important agricultural problem. Any farmer would like to know, as soon as possible, how much yield he can expect. The problem of predicting yield production can be solved by employing data mining techniques. This study evaluated the feasibility to predict the yield at Khuzestan Province in Iran using CART and CHAID algorithms. The analyses were performed using IBM SPSS Modeler 14.2. Three cropping seasons from 125 farms were selected between 2015 and 2018. The most important attributes were selected and the average yield was classified according to a decision tree. The data was partitioned into training (70%) and testing (30%) samples. The decision tree, including nine independent variables and 29 nodes, was produced through CART method. The decision tree, including nine independent variables and 39 nodes, was produced through the CHAID method. The CART and CHAID algorithms were evaluated using linear correlation and mean absolute error (MAE). Maximum precision of model in training part relevant to CART algorithm was equal to 95%, in testing part relevant to CART algorithm was equal to 93%. According to models′ precision, the results showed that CHAID and CART models were stable and suitable for prediction of sugar beet yield.

Keywords: Yield prediction, Decision tree, Classification and Regression Trees (CART), Chi-squared Automatic Interaction Detection (CHAID).

Introduction

Sugar beet is one of most important crops which are grown in tropical and sub-tropical areas of world (Abbas *et al*., 2016). Sugar beet agro-industrial companies work with large amount of data. Timely predicting sugar beet yield is a very important task. Accurate information about the nature of historical yield of crop is important. Modeling input, which is helpful to farmers and government organization for decision making process in establishing correct policies. The advances in computing and information storage have provided a vast amount of data. The challenge has been to extract knowledge from this raw data. This has led to new methods and techniques such as data mining that can bridge knowledge on crop yield estimation (Raorane & Kulkarni, 2012; Oliveira *et al.*, 2017; Thomas, 2017). Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories using pattern recognition technologies as well as

statistical and mathematical techniques (Khedr *et al*., 2015). Data mining is mainly categorized as descriptive and predictive data mining. In the agricultural area predictive data mining is mainly used (Raorane & Kulkarni, 2012). Data mining in agriculture provides many opportunities for exploring hidden patterns in these collections of data. This paper has advanced a new method of sugar beet yield prediction which is based on data mining .

Related works: Raorane & Kulkarini (2012) discussed the role of data mining as an effective tool for yield estimation in the agricultural sector. As crop production depends on geographical, biological, political and economic factors, data mining can solve the challenge of extracting knowledge from this raw data and estimate the amount of crops production. Accurate and reliable information about historical crop yield is important for decisions relating to agricultural risk management. An accurate estimate of crop size and risk helps in planning supply chain decision like production scheduling. Salame (2011) applied data mining techniques to evaluate applications for agricultural loans. The study used Logistic regression, neural network and decision tree to identify the financial and non-financial variables that signal the capacity of borrowers to pay back the loan, and determine the best model (s) to evaluate credit risk. Financial institutions that serve agriculture need to continuously evaluate their models and methods to assess the probability of default on loans, especially when assessing the probability of default of a new borrower by examining the performance of three different methods.

Machine learning algorithm was used to develop model to predict sugarcane yield (Hill *et al*., 2014). Machine learning algorithms are quite widely used in agricultural application, particularly for GIS, soil science, hydrology, precision agriculture, yield prediction and produce quality assurance (e.g. Robinson & Mort, 1997; Papageorgiou *et al*., 2011; Mollazade *et al*., 2012; Ahmadali *et al*., 2013; Pena *et al*., 2014;; Rodriguez-Galiano *et al*., 2014; Shainfar *et al*., 2014). The decision tree is one of the popular classification algorithms in current use in Data Mining and Machine Learning. Classification and regression techniques (decision tree) seem to be very popular. For example, Ekasingh *et al*. (2003) discuss the classification of farmers′ cropping choices using decision trees. Agriculture-related applications include Holmes (1998) for apple bruising, Cunningham & Holmes (1999) for mushroom grading and Michalski & Chilausky′s (1980) soybean disease diagnosis work, which is a classic benchmark problem in machine learning. Broadly, Classification and regression tree (CART) analysis has been used for detecting patterns in diverse areas, such as, epidemiology (Marshall, 2001), marine ecology (Dzeroski & Drumm, 2003), agricultural land use (Etter *et al*., 2006) and ecosystem classification (Dolan & Parker, 2005). In agriculture, the CART approach has been mainly used for detecting temporal and spatial variability in crop yields (Perez-Quezada *et al*., 2003). Ferraro *et al*. (2009) propose using CART to identify the dependence of sugarcane yield on the variation of both environmental and management factors. However, only few analyses have been carried out for detecting crop yield patters using farm-scale data (Roel *et al*., 2007). Other applications of data mining in agriculture can be seen in Folberth *et al*. (2012), Ureta *et al*. (2013), Xiao *et al*. (2014) and Meirelles & Zarate (2015).

The aim of this paper is to process the information, which was provided by a survey Shahid Beheshti agro-industrial from the

province of Khuzestan in Iran, using data mining techniques, with the aim of making the user the knowledge easier to handle. This work is organized as follows: In section 2 a review on the usage of data mining techniques in agriculture is presented. In Section 3, the methodology is applied and all stages taken to assembly the database are presented. In Section 4, the results are analysed. Finally, in Section 5, the conclusions of this work are presented.

Materials & Methods

Study area: The data for the study were collected from Shahid Beheshti agro-industrial Company. The data are obtained for the years from 2015 to 2018. The study area is located in Khuzestan Province which is major agricultural region in Iran.

Data pre-processing of surveys

To analyses large data sets, data pre-processing operations are carried out first before applying data mining algorithm. Data pre-processing includes preparation of data in desired form to work, which is clean and free from any noise. It is also used for reduction of large data into summative workable data to avoid unnecessary processing of unwanted, meaningless data (Rathod & Garg, 2016). In this study, after creating the database, we pre-processed the data before the data mining step. This procedure consisted of data cleaning and null values substitution (i.e., values that are unknown or not present).

Data mining methods

The goal of data mining is to discover hidden knowledge in data sets which the human eye or conventional statistical analysis cannot uncover. There are a wide variety of techniques, called classification models, which are available to aid and perform predictive analysis. Classification models implement supervised learning: they use a set of labeled training data in order to compute a function capable of mapping a set of variables into a class label (Tan *et al.*, 2005; Maione *et al.*, 2016a).

Decision trees (Navada *et al.*, 2011) refers to one of the oldest classification models. This model is very popular due to its simplicity, low computational costs and quick generalization of new samples. Each variable of the dataset is individually questioned, and these questions and their answers can be arranged in a hierarchical structure called trees. Each node of the tree corresponds to a variable, and each edge originating from a node x represents a value, or a range of values, for the tested variable x. Leaf nodes store class labels, the final point of the classification process. When a new test sample must be classified, each of its variables is questioned, following an existent path in the tree until a leaf node is reached and the class label associated to this leaf node is set as the new sample's class label. The Hunt algorithm is one of the most popular algorithms for building decision trees. This algorithm arranges the nodes in the tree in order to maximize the information gain, i.e., the questioned variables will be capable of dividing the dataset in pure partitions, with a more frequent value of one class label (Maione *et al.*, 2016b). Another advantage of decision trees is that trees can be easily visualized and interpreted. By observing the decision rules exposed in the tree's paths, it is possible to generate hypotheses about the individual influence of each variable in the classification results and its relationships. In this study, we use decision trees based on CART and CHAID, which implements the Hunt algorithm. CART is called supervised learning algorithm, as the outcome variable of interest is previously know and supervises the process.

The models are obtained recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning is represented graphically as a decision tree. CART algorithm was popularized by Breiman *et al.* (1984). CART algorithm use Gini index measure as the splitting criteria. CHAID algorithm was developed by Kass (1980) from a method called automatic interaction detection. As it uses Chi-squared test for tree splitting strategy it is called Chi-squared Automatic Interaction Detection (CHAID). CHAID algorithm provides a splitting condition that is either dichotomist or multiple through a Chi-square test; each element of the tree generates two or more nodes (Luca *et al.*, 2016). The methodology used in this study is based on method of classification and prediction of data mining as well as method of supervision with algorithms of CART and CHAID using IBM SPSS Modeler 14.2.

Data

The data is taken in ten input variables. Table (1) shows the variables, seed, pesticide, fungicide, herbicide, chemical fertilizer (Nitrogen), chemical fertilizer (Phosphate), chemical fertilizer (Potassium), fuel and electricity considered for this work. Here the target is yield.

Table (1): Description of continuous sugar beet variables used for present study.

Variable name	Unit	Variable's Type	Usage (role)	description				
				minimum amount	maximum amount	Average	Standard deviation	number of valid records
chemical fertilizer (Nitrogen)	Kg ha^{-1}	Continuous	Input	50	350	234.65	32.60	125
chemical fertilizer (Phosphate)	Kg ha^{-1}	Continuous	Input	20	250	124.72	45.92	125
chemical fertilizer (Potassium)	Kg ha^{-1}	Continuous	Input	0	100	34.80	23.93	125
Electricity	kwh ha^{-1}	Continuous	Input	111.47	6688.46	1345.82	143.92	125
fuel	Lit ha^{-1}	Continuous	Input	35	200	180.53	32.01	125
Pesticide	Lit ha^{-1}	Continuous	Input	0	3	1.12	0.43	125
Herbicide	Lit ha^{-1}	Continuous	Input	1	8	3.31	1.23	125
Fungicide	Lit ha^{-1}	Continuous	Input	0	2	0.24	0.52	125
Seed	Kg ha^{-1}	Continuous	Input	3	24	5.58	4.67	125
yield	ton ha^{-1}	Continuous	target	30	80	70.56	7.29	125

The sample data was first partitioned into a training sample (70%) to build the models and a testing sample (30%) to validate the models. The training sample data is used to build the models, while the testing sample data is for validation of the models. Fig. (1) depicts the data modeling process using IBM SPSS Modeler 14.2

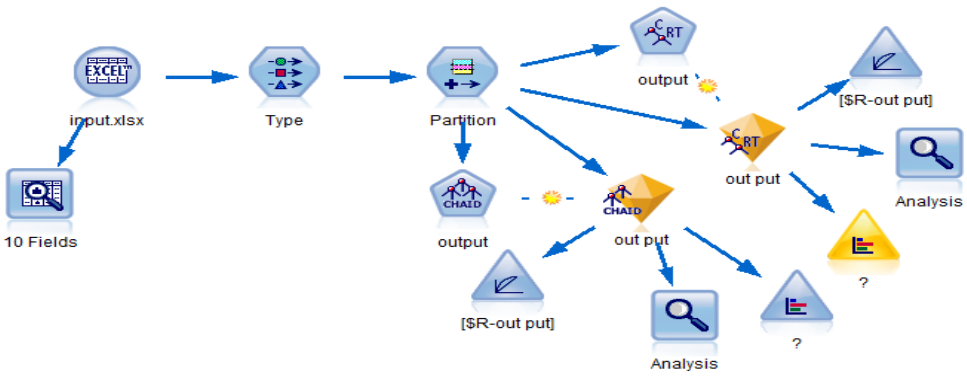

Fig. (1): Decision tree modeling in IBM SPSS Modeler 14.2.

The pentagon-shaped nodes show the construction of the models using CART and CHAID. The diamond-shaped nodes show the model outputs of the respective models. For decision trees, the CART and CHAID models were compared using the Analysis and Evaluation nodes. Then, the two predictive models which are CART and CHAID are connected to the Analysis node which provides the computation of accuracy rates, while the Evaluation node produces the Gain charts.

Model Assessment

We can visualize the model performance gains charts. Gain is equal to percent of recodes in target class compared to total database included in this node. Gain is defined by relation 1 as follows (Alizadeh & Malekmohamadi, 2014):

$$\text{Gain}= \frac{\text{Number of hits in quantile}}{\text{total number of hits}} * 100\% \qquad (1)$$

Results & Discussion

Data from 125 farm was taken for this study and data has been entered and saved in MS excel. Then data was processed in IBM SPSS Modeler 14.2 and the results were obtained. In this paper the estimation of the crop yield is analyzed with respect to ten factors namely seed, pesticide, fungicide, herbicide, chemical fertilizer (Nitrogen), chemical fertilizer (Phosphate), chemical fertilizer (Potassium), fuel, electricity and yield. Greenland (2005) found relation between climate variables and annual sugarcane yield in Louisiana and it was possible to simulate the annual yield based on climate variables. In sugarcane cropping systems, yield variability has been mainly attributed to harvest time and crop cultivar (Lisson *et al.*, 2005), crop class (Evenson *et al.*, 1987) and soil properties (Nelson & Ham, 2000). However, few attempts have been made to characterize and quantify the factors that contribute to the variation in sugarcane crop yield (Lawes *et al.*, 2002b). Usually, the joint effects of different factors on crop yield are described in crop simulation models (Lark, 1997; Lisson *et al.*, 2005).

Decision tree built from the yield dataset To generate the decision tree, was entered this data base to the IBM SPSS Modeler 14.2 and obtained the decision trees shown in figs. (3 and 4).

CART algorithm

In CART algorithm about 70% of data is used as training data sets and about 30% of the data are employed as test data (Fig. 2). The Gini coefficient has been used. Also the tree has pruned from the fifth level.

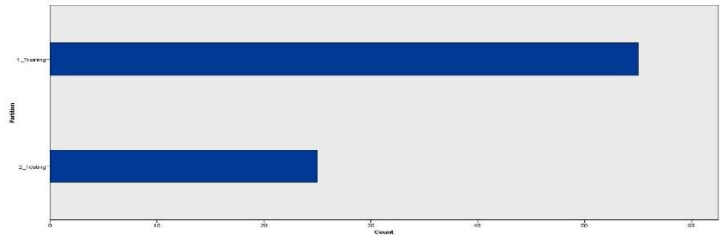

Fig. (2): Ratio of training data to testing data.

Fig. (3) shows the architecture of CART model. The CART classification tree obtained (Fig. 3) shows a tree with 29 nodes; fifteen of them are terminal nodes. The Gini index was selected as a splitting criterion. The first variable selected for splitting is fuel (Gini improvement measure = 14680.84). The next discriminators are seed (Gini improvement measure =5266.41), which is split into $12 <$ & $12 \geq$ and fuel (Gini improvement measure = 551.22), which is split into $167.5 <$ and$167.5 \geq$. , and so on. Percentages in each category and in each joint category are shown in fig. (3). The improvement (Fig. 3) measures the increase of the effect of child node on the dependent variable; it is determined by the largest difference in the proportions of the dependent variable in the child nodes (Lemon, 2003). Thus, improvements of 14680.84 means that fuel contribute 14680.84 in the discrimination between yield farms; seed makes an additional 5266.41 improvement, and so on.

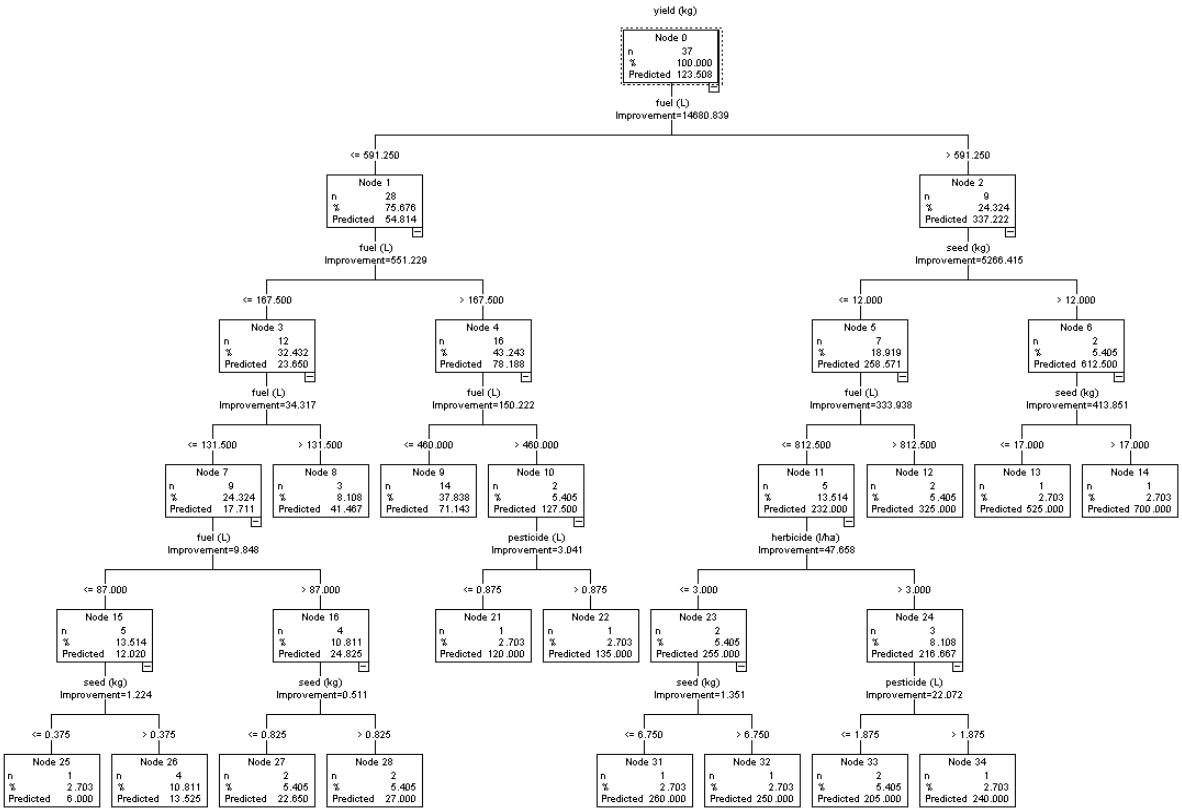

Fig. (3): The tree resulted from execution of CART algorithm.

CHAID algorithm

In CHAID, the tree was pruned at the three level. We generate decision trees based on the collected data (Fig. 4) CHAID algorithm is used to split (farms) into groups based on value of an independent variables. The tree diagram (Figure 4) shows tree construction based on the subsample of 88 farms. There are totally 39 nodes that consist of 28 terminal nodes; the first node placed in the tree is root node. The first discriminator (electricity) splits the root node into eight child nodes (≤ 167.21); (167.21-278.68), (278.68-445.89), (445.89-1114.74), (1114.74-1672.11), (1672.11-2229.48), (2229.48-2786.86) and ($>$ 2786.86). The second classifier is herbicide, fuel and seed, and so on.

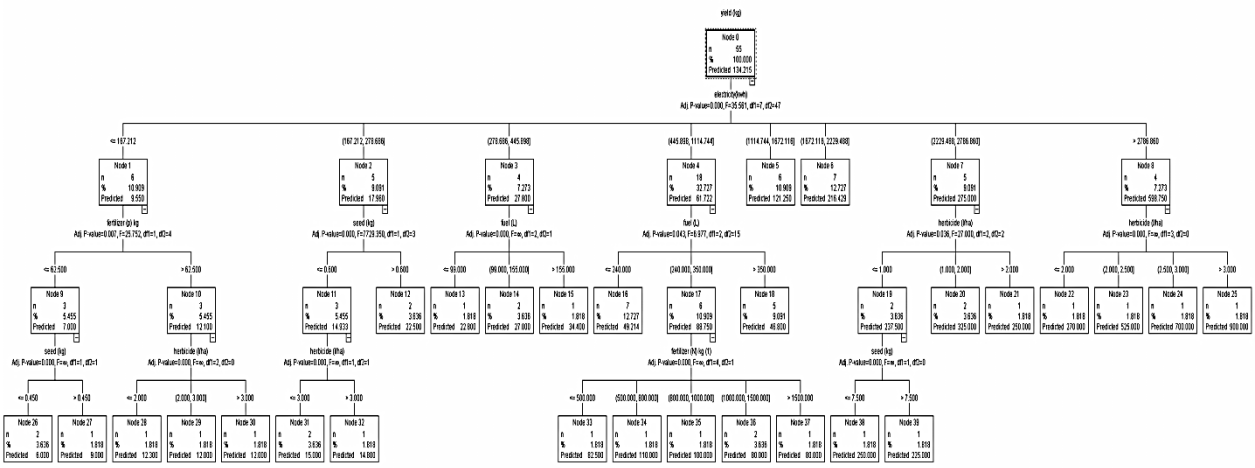

Fig. (4): Decision tree resulting from execution of CHAID algorithm.

Indicator importance: fuel, seed, fungicide, pesticide, chemical fertilizer (Nitrogen), chemical fertilizer (potassium), chemical fertilizer (Phosphate), electricity, herbicide are the most important fields (variables) for CART model (Fig. 5). In CHAID model, the predictor importance are electricity, seed, chemical fertilizer (Phosphate), chemical fertilizer (Nitrogen), herbicide and fuel (Fig. 6).

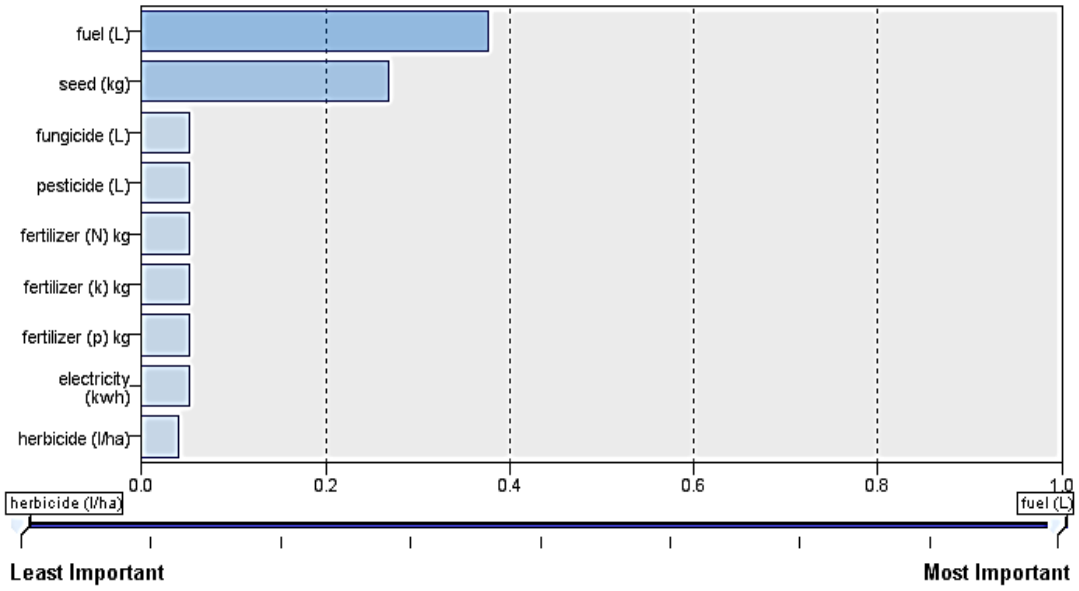

Fig. (5): Normalized importance of variable in CART algorithm.

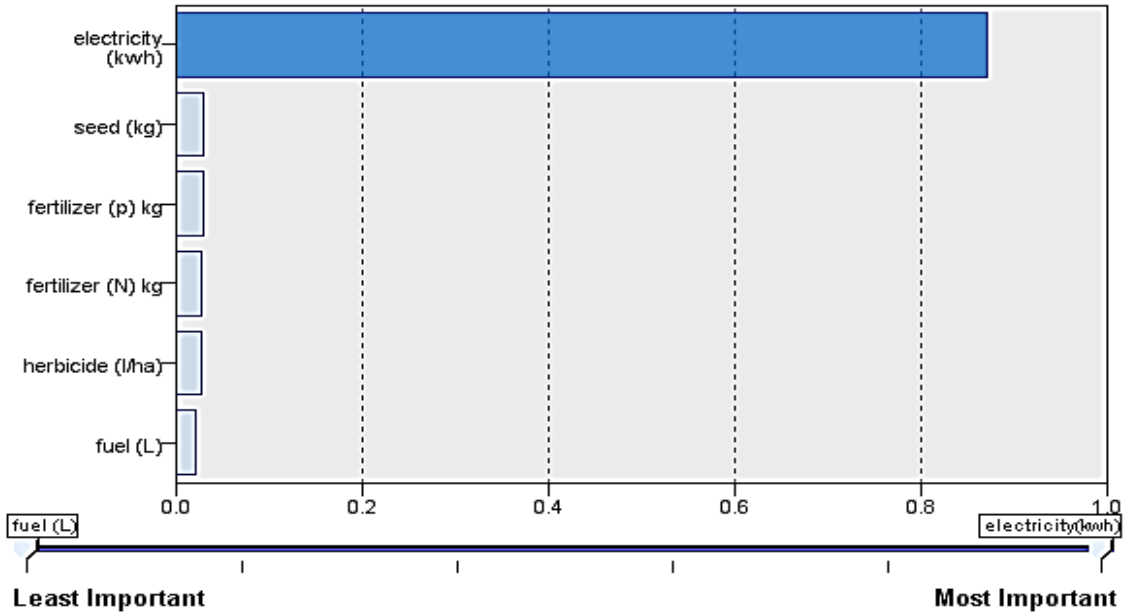

Fig. (6) Normalized importance of variable in CHAID algorithm.

Finding from evaluating models′ precision

Gain chart: The gain charts (Figs. 7 and 8) provide a visual summary of the usefulness of the information provided by one or more statistical models for predicting categorical dependent variable. Specifically, the chart summarizes the utility that one can expect by using the respective predictive models, as compared to using baseline information only. The baseline indicates the expected result if no model were used at all. The statistical analyzers expect that the models having high precision be close to the best line presented in the figures and, in fact, be bowl-shaped. In these figures, the horizontal axis is disjunction points and the vertical axis is cumulative percentage of samples, which are located under these disjunction points. On the other hand, the linear model closer to the best line, namely "BEST", is a better model. Corresponding value of Gains can be computed for each percentile of the population to determine the percentile of cases that should be targeted to achieve a certain percentage of predictive accuracy. You can see from the graph (Figure 7) that the Gains values for different percentiles can be connected by a line and it will typically ascend slowly and merge with the baseline if all farms (100%) were selected. In testing, the CHAID was more effective (Fig. 8). **Modeling accuracy**

After training the model we are able to make the prediction of revenue increase. We can test the results of the learning process with changing the input data and executing the Analysis node. From the linear correlation between the predicted increase and the correct answers, we are able to find that the trained system predicts with a high degree of success. The precision of CART model in training and testing sections is 95% and 93%, respectively (Table 2).

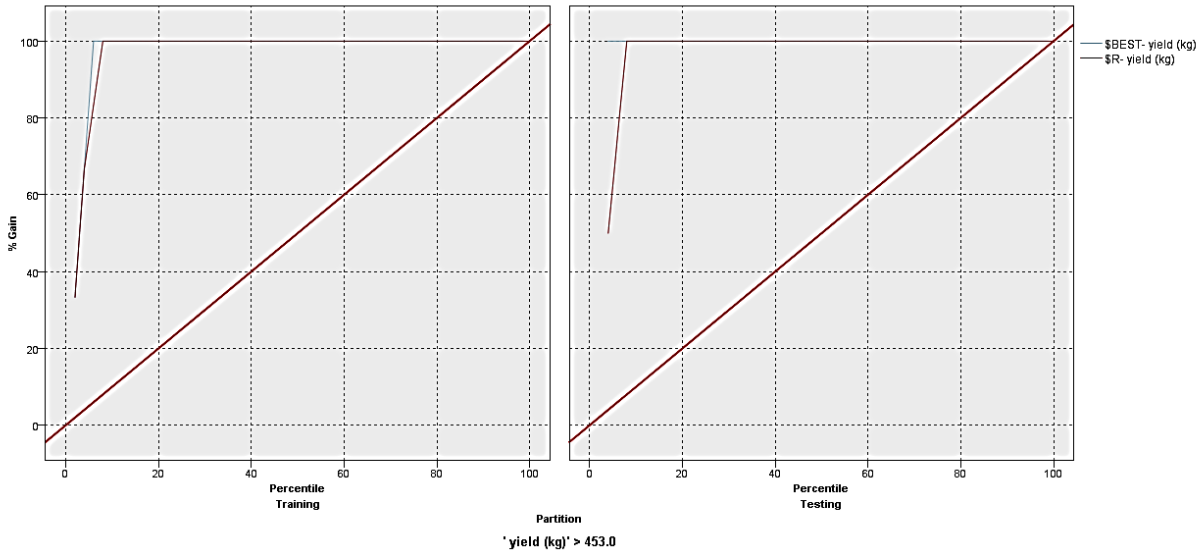

Fig. (7): Gain chart for CART algorithm.

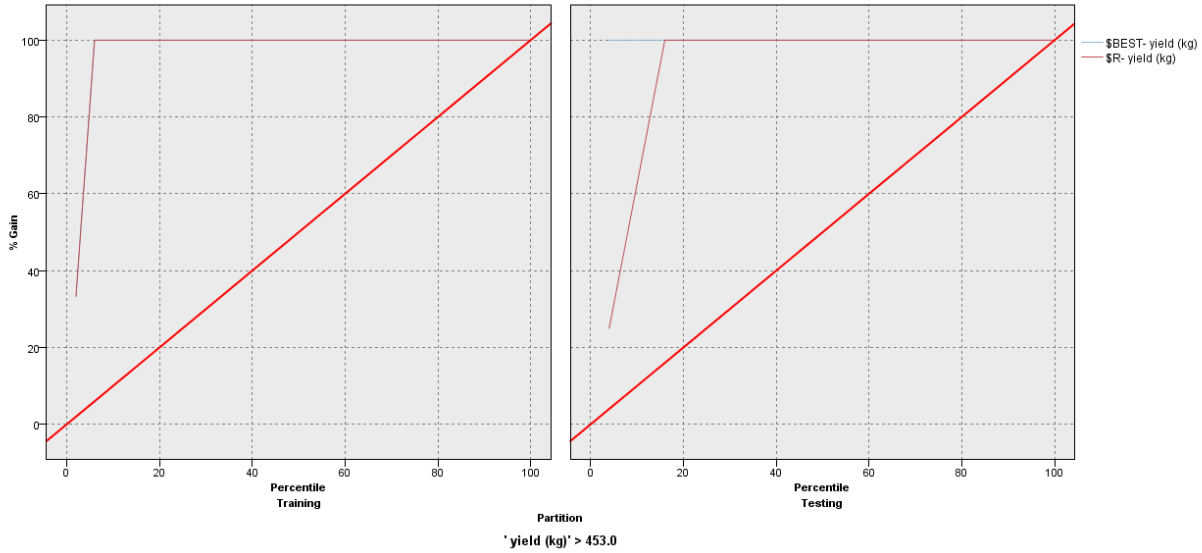

Fig. (8): Gain chart for CHAID algorithm.

The CHAID model's is correspondingly (Table 3). The model′s precision in training and testing sections is 90% and 85%, correspondingly. As can be easily seen on average values, CART is the leading algorithm with most accurate prediction results. CHAID is only slightly inferior. Generally, the prediction accuracy of the models in this study, from the range of 85% to 95% is satisfactory as it is supported by Bozkir & Sezer (2011).

Nonetheless, there is one issue that must be addressed for decision tree analyses made in this study. Although CART performs the best accuracy on average values, it does not mean that CART is the best. Eventually, the results of this paper show that the decision tree models from CART and CHAID data mining approaches are quite effective tools in this process and the decision trees produced in this study could be useful to analysts of yield prediction.

Table (2): Results for output field yield (CART algorithm).

partition	training	testing
Minimum Error (ton ha^{-1})	-275.0	-125.0
Maximum Error (ton ha^{-1})	200.0	60.0
Mean Error (ton ha^{-1})	-6.75	-8.56
Mean Absolute Error (ton ha^{-1})	22.69	31.38
Standard Deviation	52.19	44.29
Linear correlation	0.952	0.933
Occurrences	88	37

Table (3): Results for output field yield (CHAID algorithm).

partition	training	testing
Minimum Error (ton ha^{-1})	-56.42	-600.0
Maximum Error (ton ha^{-1})	43.75	100.0
Mean Error (ton ha^{-1})	0.0	-74.55
Mean Absolute Error (ton ha^{-1})	9.50	122.50
Standard Deviation	17.58	206.71
Linear correlation	0.90	0.85
Occurrences	88	37

The study should probably be repeated with several data mining tools and the same datasets to provide a comparative analysis and see if the results vary across tools.

Conclusion

Agricultural organizations and their management try every day to find information (knowledge) in large databases for business decision making. Data mining, through better management and data analysis, can assist agricultural organizations to achieve greater profit. Data mining technology provides user oriented access to new and hidden patterns in data, from which knowledge is generated which can help with decision making in agricultural organizations. Yield prediction is a very important agricultural problem that remains to be solved based on the available data. The yield prediction problem can be solved by employing Data Mining techniques such as decision trees. This paper aims to predict sugar beet yield by building process of CART and CHAID via IBM SPSS Modeler 14.2 as a data mining predictive technique. The CART and CHAID algorithms were evaluated using linear correlation and mean absolute error (MAE). As a consequence, decision tree models and decision support system developed with decision trees have significant potential for decision makers in resource optimization. The performance of predictive models depends on the data structure, data quality and variable selection. With the availability of data mining software, data mining models are easy to construct and apply in agriculture.

Acknowledgement

This paper was supported by Shahid Chamran university of Ahvaz, Iran.

Conflict of interest: The authors declare that they have no conflict of interest.

ORCID:

N. Monjezi: 0000000182297706

References

Abbas, H.T.; Sahi, S.T.; Habib, A., & Ahmed, S. (2016). Laboratory evaluation of fungicides and plant extracts against strains of *Colletrichum falcatum* the cause of red rot of sugarcane. *Pakistan Journal of Agricultural Sciences, 53,* 181-186. https://doi.org/10.21162/PAKJAS/16.4655

Alizadeh, S., & Malekmohamadi, S. (2014). *Data mining and knowledge discovery step by step with Clementine*. Khajeh Nasir University. K. N. Toosi Univ. Technology Press. Tehran: 367pp. https://doi.org/10.5772/6438

Bozkir, A. S., &. Sezer, E. A. (2011). Predicting food demand in food courts by decision tree approaches. *Procedia Computer Science, 3,* 759-763. https://doi.org/10.1016/j.procs.2010.12.125

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC. New York: 368pp. https://www.routledge.com/Classification-and-Regression-Trees/Breiman-Friedman-Stone-Olshen/p/book/9780412048418

Cunningham, S. J., & Holmes, G. (1999). Developing innovative applications in agriculture using data mining. In: Proceeding Southeast Asia Regional Computer Confederation Conference. https://www.cs.waikato.ac.nz/~ml/publications/1999/99SJC-GH-Innovative-apps.pdf

Dolan, B. J., & Parker, G. R. (2005). Ecosystem classification in a flat, highly fragmented region of Indiana, U.S.A. *Forest Ecology and Management, 219,* 109-131. http://dx.doi.org/10.1016%2Fj.foreco.2005.08.045

Dzeroski, S., & Drumm, D. (2003). Using regression trees to identify the habitat preference of the sea cucumber (*Holothuria leucospilota*) on Rarotonga, Cook Islands. *Ecological Modelling, 170,* 219-226. https://doi.org/10.1016/S0304-3800(03)00229-1

Ekasingh, B., Ngamsomsuke, K., Letcher, R., & Spate, J. (2003). *A data mining approach to simulating land use decisions: Modelling farmer's crop choice from farm level data for integrated water resource management*. In: Singh, V. & Yadava, R. (Eds.). *Advances in Hydrology:* Proceedings of the International Conference Water Environment Research, 175-188. Corpus ID: 202602091

Etter, A., McAlpine, C., Wilson, K., Phinn, S., & Possingham, H. (2006). Regional patterns of agricultural land use and deforestation in Colombia. Agric. *Agriculture, Ecosystems & Environment, 114,* 369-386. https://doi.org/10.1016/j.agee.2005.11.013

Evenson, C. I., Muchow, R. C., El-Swaify, S. A., & Osgood, R. V. (1987). Yield accumulation in irrigated sugarcane. I. Effect of crop age and cultivar. *Agronomy Journal, 89,* 638-646. https://doi.org/10.2134/agronj1997.00021962008900040016x

Ferraro, D. O., Rivero, D. E., & Ghersa, C. M. (2009). An analysis of the factors that influence sugarcane yield in Northern Argentina using classification and regression trees. *Field Crops Research, 112,* 149-157. https://doi.org/10.1016/j.fcr.2009.02.014

Folberth, C., Taiser, T., Abbaspour, K. C., Schulin, R., & Yang, H. (2012). Regionalization of a large-scale crop growth model for sub-Saharan Africa: Model setup, evaluation, and estimation of maize yields. *Agriculture, Ecosystems and Environment, 151,* 21-33.https://doi.org/10.1016/j.agee.2012.01.026

Greenland, D. (2005). Climate variability and sugarcane yield in Louisiana. *The Journal of Applied Meteorology and Climatology, 44,* 1655-1666. https://doi.org/10.1175/JAM2299.1

Hill, M. G.; Connolly, P. G.; Reutemann, P., & Fletcher, D. (2014). The use of data mining to assist crop protection decisions on kiwifruit in New Zealand. *Computers and Electronics in Agriculture, 108,* 250-257. http://dx.doi.org/10.1016/j.compag.2014.08.011

Holmes, G., Cunningham, S., Dela Rue, B., & Bollen, A. (1998). Predicting apple bruising using machine learning. In: Proceedings of the Model-IT Conference. *Journal Acta Horticulture, 476,* 289-296. https://doi.org/10.17660/ActaHortic.1998.476.33

Kass, G. V. (1980). An Exploratory technique for investigating large quantities of categorical data. *Journal of Applied Statistics, 29,* 119-127. https://doi.org/10.2307/2986296

Khedr, A. E., Kadry, M., & Walid, G. (2015). Proposed framework for implementing data mining techniques to enhance decisions in agriculture sector. *Procedia Computer Science, 65,* 633-642. https://doi.org/10.1016/j.procs.2015.09.007

Lark, R. M. (1997). An empirical method for describing the joint effects of environmental and other variables on crop yield. *Annals of Applied Biology*, *131*, 141–159. https://doi.org/10.1111/j.1744-7348.1997.tb05402.x

Lawes, R. A., Lawn, R. J., Wegener, M. K., & Basford, K. E. (2002a). Understanding and managing the late time of ratooning effect on cane yield. *Proceedings of the Australian Society of Sugar Cane Technology*, *24*, 160-165. https://espace.library.uq.edu.au/view/UQ:98018

Lawes, R. A., McDonald, L. M., Wegener, M. K., Basford, K. E., & Lawn, R. J. (2002b). Factors affecting cane yield and commercial cane sugar in the Tully district. *Australian Journal of Experimental Agriculture*, *42*, 473-480. https://doi.org/10.1071/EA01020

Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. *Annals of Behavioral Medicine*, *26*, 172-181. https://doi.org/10.1207/S15324796ABM2603_02

Lisson, S. N., Inman-Bamber, N. G., Robertson, M. J., & Keating, B. A. (2005). The historical and future contribution of crop physiology and modelling research to sugarcane production systems. *Field Crops Research*, *92*, 321-335. https://doi.org/10.1016/j.fcr.2005.01.010

Luca, M. D., Abbondati, F., Pirozzi, M., & Zilioniene, D. (2016). Preliminary study on runway pavement friction decay using data mining. *Transportation Research Procedia*, *14*, 3751- 3760. https://doi.org/10.3390/su12093516

Maione, C., Batista, B.L., Campiglia A. D., & Barbosa Jr., F. (2016a). Rommel Melgaço Barbosa Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. *Computers and Electronics in Agriculture*, *121*, 101-107 https://doi.org/10.1016/j.compag.2015.11.009

Maione, C., Paula, E. S., Gallimberti, M., Batista, B. L., Campiglia, A. D., Barbosa Jr. F., & Barbosa, R. M. (2016b). Comparative study of data mining techniques for the authentication of organic grape juice based on ICP-MS analysis. *Expert Systems with Applications*, 49, 60-73. https://doi.org/10.1016/j.eswa.2015.11.024

Marshall, R. J. (2001). The use of classification and regression trees in clinical epidemiology. *Journal of Clinical Epidemiology*, *54*, 603-609. https://doi.org/10.1016/s0895-4356(00)00344-9

Meirelles, W. C. L., & Zarate, L. E. (2015). Data mining in the reduction of the number of places of experiments for plant cultivates. *Computers and Electronics in Agriculture*, *113*, 136-147. https://doi.org/10.1016/j.compag.2015.02.006

Michalski, R., & Chilausky, R. (1980). Learning by being told and learning by examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *Information Journal Policy analysis Information Systems*, *4*, 125-161. https://link.springer.com/article/10.1007/BF00130711

Mollazade, K., Omid, M., & Arefi, A. (2012). Comparing data mining classifiers for grading raisins based on visual features. *Computers and Electronics in Agriculture*, *84*, 124-131. https://doi.org/10.1016/j.compag.2012.03.004

Navada, A.; Ansari, A.; Patil, S. & Sonkamble, B. (2011). Overview of use of decision tree algorithm sin machine learning. 2011 IEEE. *Control and System Graduate Research Colloquium (ICSGRC)*, 37-42. https://doi.org/10.1109/ICSGRC.2011.5991826.

Nelson, P. N., & Ham, G. J. (2000). Exploring the response of sugar cane to sodic and saline conditions through natural variation in the field. *Field Crops Research*, *66*, 245-255. https://doi.org/10.1016/S0378-4290(00)00077-0

Oliveira M. P. G., Bocca, F. F., & Rodrigues, L. H. A. (2017). From spreadsheets to sugar content modeling: A data mining approach. *Computers and Electronics in Agriculture*, 132, 14-20. https://doi.org/10.1016/j.compag.2016.11.012

Papageorgiou, E. I., Markinos, A. T., & Gemtos, T. A. (2011). Fuzzy cognitive map based approach for predicting yield in cotton crop production as a basis for decision support system in precision agriculture application. *Applied Soft Computing*, *11*, 3643-3657. https://doi.org/10.1016/j.asoc.2011.01.036

Pena, J. M., Gutierrez, P. A., Hervas-Martinez, C., Six, J., Plant, R. E., & Lopez-Granados, F. (2014). Object-based image classification of summer crops

with machine learning methods. *Remote Sensing, 6,* 5019-5041. https://doi.org/10.3390/rs6065019

Perez-Quezada, J. F., Pettygrove, G. S., & Plant, E. R. (2003). Spatial-temporal analysis of yield and soil factors in two four-crop-rotation fields in the Sacramento Valley. California. *Agronomy Journal, 95,* 676-687. https://doi.org/10.2134/agronj2003.0676

Raorane, A. A., & Kulkarni, R. V. (2012). Data mining: an effective tool for yield estimation in the agricultural sector. *International Journal of Emerging Trends & Technology in Computer Science, 1,* 75-79.

Rathod, R. R., & Garg, R. D. (2016). Regional electricity consumption analysis for consumers using data mining techniques and consumer meter reading data. *International Journal of Electrical Power and Energy Systems, 78,* 368-374. https://doi.org/10.1016/j.ijepes.2015.11.110

Robinson, C., & Mort, N. (1997). A neural network system for the protection of citrus crops from frost damage. *Computers and Electronics in Agriculture, 16,* 177-187. https://doi.org/10.1016/S0168-1699(96)00037-3

Rodriguez-Galiano, V., Mendes, M. P., Jose Garcia-Soldado, M., Chica-Olmo, M., & Ribeiro, L. (2014). Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). *Science of the Total Environment, 476,* 189-206. https://doi.org/10.1016/j.scitotenv.2014.01.001

Roel, A., Firpo, H., & Plant, R. E. (2007). Why do some farmers get higher yields? Multivariate analysis of a group of Uruguayan rice farmers. *Computers and Electronics in Agriculture, 58,* 78-92. https://doi.org/10.1016/j.compag.2006.10.001

Salame, E. J. (2011*). Applying data mining techniques to evaluate applications for agricultural loans*. Ph. D. Thesis, University of Nebraska, 162pp. https://digitalcommons.unl.edu/agecondiss/10/

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining.* Addison-Wesley Longman Publishing Co., Inc., Boston, M.A., 792pp. https://dl.acm.org/doi/book/10.5555/1095618

Thomas, E. (2017). An artificial neural network for real-time hardwood lumber grading. *Computers and Electronics in Agriculture, 132,* 71-75. https://doi.org/10.1016/j.compag.2016.11.018

Ureta, C., González-Salazar, C., Gonzalez, E. J., Alvarez-Buylla, E. R., & Martínez-Meyer, E. (2013). Environmental and social factors account for Mexican maize richness and distribution: a data mining approach. *Agriculture, Ecosystems and Environment, 179,* 25-34. DOI: 10.1016/j.agee.2013.06.017

Xiao, Y., Mignolet, C., Mari, J. F., & Benoit, M. (2014). Modeling the spatial distribution of crop sequences at a large regional scale using land-cover survey data: A case from France. *Computers and Electronics in Agriculture, 102,* 51-63. https://doi.org/10.1016/j.compag.2014.01.010

تطبيق خوارزميات CART و CHIAD في تنبئو انتاج البنجر السكر

نسيم منجزي

قسم هندسة النظم الحيوية ، كلية الزراعة ، جامعة شهيد جمران ، الأهواز ، إيران

المستخلص: يعتبرالتنبؤ بالناتج الزراعي مشكلة في غاية الاهمية. يود أي مزارع أن يعرف وفي أقرب وقت ممكن ، مقدار الناتج الذي يمكن أن يتوقعه. يمكن حل مشكلة توقع إنتاج الغلة باستخدام تقنيات التنقيب عن البيانات. قيمت هذه الدراسة جدوى توقع العائد في مقاطعة خوزستان في إيران باستخدام خوارزميات CART و CHAID. تم إجراء التحليلات باستخدام IBM SPSS Modeler 14.2. تم اختيار ثلاثة مواسم محصولية من 125 مزرعة بين عامي 2015 و 2018. تم اختيار أهم الصفات وتم تصنيف متوسط المحصول وفقًا لشجرة القرار. تم تقسيم البيانات إلى عينات تدريب (70٪) واختبار (30٪). تم إنتاج شجرة القرار ، بما في ذلك تسعة متغيرات مستقلة و 29 عقدة ، من خلال طريقة CART. تم إنتاج شجرة القرار بما في ذلك تسعة متغيرات مستقلة و 39 عقدة من خلال طريقة CHAID. تم تقييم خوارزميات CART و CHAID باستخدام الارتباط الخطي والذي يمثل الخطأ المطلق (MAE). كانت الدقة القصوى للنموذج في جزء التدريب ذي الصلة بخوارزمية CART تساوي 95 ٪ .اما في جزء الاختبار ذي الصلة بخوارزمية CART كانت تساوي 93 ٪. وفقًا لنماذج ' الدقة، أظهرت النتائج أن نماذج CHAID و CART كانت عالية الدقة ومناسبة للتنبؤ بمحصول البنجر السكر.

الكلمات المفتاحية: تنبؤ الناتج، شجرة القرار، التصنيف وشجرة الانحدار (CART) ، اكتشاف التفاعل التلقائي لمربع كاي (CHAID)