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 & Kulkarni (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 patterns 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 assemble 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>
<thead>
<tr>
<th>Variable name</th>
<th>Unit</th>
<th>Variable’s Type</th>
<th>Usage (role)</th>
<th>minimum amount</th>
<th>maximum amount</th>
<th>Average</th>
<th>Standard deviation</th>
<th>number of valid records</th>
</tr>
</thead>
<tbody>
<tr>
<td>chemical fertilizer (Nitrogen)</td>
<td>Kg ha⁻¹</td>
<td>Continuous</td>
<td>Input</td>
<td>50</td>
<td>350</td>
<td>234.65</td>
<td>32.60</td>
<td>125</td>
</tr>
<tr>
<td>chemical fertilizer (Phosphate)</td>
<td>Kg ha⁻¹</td>
<td>Continuous</td>
<td>Input</td>
<td>20</td>
<td>250</td>
<td>124.72</td>
<td>45.92</td>
<td>125</td>
</tr>
<tr>
<td>chemical fertilizer (Potassium)</td>
<td>Kg ha⁻¹</td>
<td>Continuous</td>
<td>Input</td>
<td>0</td>
<td>100</td>
<td>34.80</td>
<td>23.93</td>
<td>125</td>
</tr>
<tr>
<td>Electricity</td>
<td>kwh ha⁻¹</td>
<td>Continuous</td>
<td>Input</td>
<td>111.47</td>
<td>6688.46</td>
<td>1345.82</td>
<td>143.92</td>
<td>125</td>
</tr>
<tr>
<td>fuel</td>
<td>Lit ha⁻¹</td>
<td>Continuous</td>
<td>Input</td>
<td>35</td>
<td>200</td>
<td>180.53</td>
<td>32.01</td>
<td>125</td>
</tr>
<tr>
<td>Pesticide</td>
<td>Lit ha⁻¹</td>
<td>Continuous</td>
<td>Input</td>
<td>0</td>
<td>3</td>
<td>1.12</td>
<td>0.43</td>
<td>125</td>
</tr>
<tr>
<td>Herbicide</td>
<td>Lit ha⁻¹</td>
<td>Continuous</td>
<td>Input</td>
<td>1</td>
<td>8</td>
<td>3.31</td>
<td>1.23</td>
<td>125</td>
</tr>
<tr>
<td>Fungicide</td>
<td>Lit ha⁻¹</td>
<td>Continuous</td>
<td>Input</td>
<td>0</td>
<td>2</td>
<td>0.24</td>
<td>0.52</td>
<td>125</td>
</tr>
<tr>
<td>Seed</td>
<td>Kg ha⁻¹</td>
<td>Continuous</td>
<td>Input</td>
<td>3</td>
<td>24</td>
<td>5.58</td>
<td>4.67</td>
<td>125</td>
</tr>
<tr>
<td>yield</td>
<td>ton ha⁻¹</td>
<td>Continuous</td>
<td>target</td>
<td>30</td>
<td>80</td>
<td>70.56</td>
<td>7.29</td>
<td>125</td>
</tr>
</tbody>
</table>

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.
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):

\[
\text{Gain} = \frac{\text{Number of hits in quantile}}{\text{total number of hits}} \times 100\% \quad (1)
\]

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 $12 < x \leq 12$, and fuel (Gini improvement measure = 551.22), which is split into $167.5 < x$ and $167.5 \geq x$, 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 ($\leq 167.21$); $(167.21-278.68)$, $(278.68-445.89)$, $(445.89-1114.74)$, $(1114.74-1672.11)$, $(1672.11-2229.48)$, $(2229.48-2786.86)$ and $(> 2786.86)$. The second classifier is herbicide, fuel and seed, and so on.

![Decision tree resulting from execution of CHAID algorithm.](image)

**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).

![Normalized importance of variable in CART algorithm.](image)
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).
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.
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


