Abstract: Canola production increased significantly in Iran due to its high yield in recent years. Khuzestan province is the main centre of canola production in Iran. This paper presents a data mining study of samples of canola obtained from farms in Amirkabir Agro-Industry of Khuzestan province. Data were collected from 48 farms. The farms were chosen by random sampling method. The purpose of this study is to determine energy consumption of input and output used in canola production. And Output energy of canola farms is predicted using data mining and multi-layer perceptron neural network. This is an analytic research and its database consists of 432 records. Data required for this research was obtained during growing seasons in 2017-2018. Data analysis was done by IBM SPSS modeler 14.2. The results showed that the amount of energy consumed in canola production was 28927.43 MJ ha⁻¹. About 40% of this was generated by fertilizers and 37% from electricity and diesel fuel. Concerning the model used in the research, it was found that variables of chemical fertilizer, fuel, electricity energy and irrigation, respectively had the highest effect on output variable (productive energy). Amount of prediction precision in neural network algorithm meaning ratio of correctly predicted records to total records was 86.5%. Also, linear correlation between actual values and predicted values was 0.84 and 0.88 respectively for training data and testing data suggesting strong correlation. Results obtained in this research can be effective for canola farmers in Amirkabir Agro-Industry in direction of evaluation and optimization of energy consumption in process of canola production and reduction of consumption of energy inputs.

Keywords: Energy, Predict, Data mining, Artificial Neural Networks (ANN), Canola.

Introduction
Timely predicting canola output energy is a very important task which we face currently. The relation between agriculture and energy is very close. Agricultural itself is an energy user and energy supplier in the form of bio-energy (Namdari, 2011). Until now, considerable studies have been conducted in different countries on energy use in agriculture (Yilmaz et al., 2005; Ozkan et al., 2011; Sefeedpari et al., 2014; Almaliki et al., 2016; Almaliki, 2017). Efficient use of energy resources is one of the major assets of eco-efficient and sustainable production, in agriculture (Taheri-
Garavand et al., 2010). Therefore optimization and modeling of energy consumption are very important in production of agricultural products especially canola. On the other hand, collected data are various with complex relationship and it is difficult to analyze and manage them via empirical and statistical analyses and experiences. Data mining is a powerful technology in management and organization of a large size of information. Data mining means searching in one database for finding inter-data patterns. Several researches have been conducted on application of data mining techniques in agriculture. Medar & Rajpurohit (2014) present the various crop yield prediction methods using data mining techniques. Different data mining techniques such as K-Means, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN) and Support Vector Machines (SVM) for very recent applications of data mining techniques in agriculture field. In addition, the references included in parentheses confirmed the ability and necessity of using data mining technique in agriculture (Raorane & Kulkarni, 2013; Jeyenthil et al., 2014; Kalpana et al., 2014; Geetha, 2015; Raorane & Kulkarni, 2015).

The main objectives of this study are to predict output energy of canola cultivation in Khuzestan province of Iran by means of data mining technique (Artificial Neural Networks).

Materials & Methods

Selection of case study region and data collection

This study was conducted in Amirkabir Agro-Industry of Khuzestan province of Iran. The collected information belonged to the 2017-2018. Before collecting data, a pre-test survey was conducted by a group of randomly selected farms. The required sample size was determined using simple random sampling method. The equation is as below (Khoshnevisan et al., 2015):

\[ n = \frac{N \times s^2 \times t^2}{(N-1)d^2 + (s^2 \times t^2)} \]  

(1)

In the formula, the below signs and letters represent:

- \( n \) is the required sample size,
- \( N \) is the number of canola farms in target population,
- \( s \) is the standard deviation in the pre-tested data,
- \( t \) is the \( t \) value at 95% confidence limit (1.96), and
- \( d \) is the acceptable error. The permissible error in the sample size was defined to be 5% for 95% confidence. Thus the sample size was found to be 48 for collecting data appropriate questionnaire designed.

Energy Equivalents of Inputs and Output

In order to predict output energy, the data were converted into output and input energy levels using equivalent energy values for each commodity and input. Energy equivalents shown in table (1) were used for estimation.

Firstly, the amounts of inputs used in the production of canola were specified in order to calculate the energy equivalences in the study. Energy input includes human labor, diesel fuel, electricity, irrigation, machinery, fertilizers, chemical and seed amounts and output yield include canola. The units in Table 1 were used to find the input amounts. The amounts of input were calculated per hectare and then, these input data were multiplied with the coefficient of energy equivalent. The previous studies (cited in Table 1) were used to determine the energy equivalents’ coefficients.
Table (1): Energy equivalent of inputs and output in canola production.

<table>
<thead>
<tr>
<th>Item</th>
<th>Unit</th>
<th>Energy equivalent (MJ kg⁻¹)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Human labor</td>
<td>h</td>
<td>1.96</td>
<td>Taheri-Garavand et al., 2010</td>
</tr>
<tr>
<td>2. Machinery</td>
<td>Kg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Tractor</td>
<td></td>
<td>138</td>
<td>Taheri-Garavand et al., 2010</td>
</tr>
<tr>
<td>b. Plow</td>
<td></td>
<td>180</td>
<td>Taheri-Garavand et al., 2010</td>
</tr>
<tr>
<td>c. Disk harrow</td>
<td></td>
<td>149</td>
<td>Salami et al., 2010</td>
</tr>
<tr>
<td>d. Planter</td>
<td></td>
<td>133</td>
<td>Taheri-Garavand et al., 2010</td>
</tr>
<tr>
<td>e. Equipment of fertilizing</td>
<td></td>
<td>129</td>
<td>Taheri-Garavand et al., 2010</td>
</tr>
<tr>
<td>f. Sprayer</td>
<td></td>
<td>129</td>
<td>Taheri-Garavand et al., 2010</td>
</tr>
<tr>
<td>g. Combine</td>
<td></td>
<td>116</td>
<td>Taheri-Garavand et al., 2010</td>
</tr>
<tr>
<td>h. Other machinery</td>
<td></td>
<td>62.7</td>
<td>Mousavi-Avval et al, 2011</td>
</tr>
<tr>
<td>3. Diesel fuel</td>
<td>Lit</td>
<td>47.8</td>
<td>Banaeian et al., 2011</td>
</tr>
<tr>
<td>4. Chemicals</td>
<td>Kg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Herbicide</td>
<td></td>
<td>238</td>
<td>Erdal et al., 2007</td>
</tr>
<tr>
<td>b. Pesticide</td>
<td></td>
<td>101.2</td>
<td>Erdal et al., 2007</td>
</tr>
<tr>
<td>c. Fungicide</td>
<td></td>
<td>216</td>
<td>Erdal et al., 2007</td>
</tr>
<tr>
<td>5. Fertilizers</td>
<td>Kg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Nitrogen (N)</td>
<td></td>
<td>78.1</td>
<td>Pishgar Komleh et al., 2011</td>
</tr>
<tr>
<td>b. Phosphate (P₂O₅)</td>
<td></td>
<td>17.4</td>
<td>Pishgar Komleh et al., 2011</td>
</tr>
<tr>
<td>c. Potassium (K₂O)</td>
<td></td>
<td>13.7</td>
<td>Pishgar Komleh et al., 2011</td>
</tr>
<tr>
<td>6. Water for irrigation</td>
<td>m³</td>
<td>0.63</td>
<td>Pishgar Komleh et al., 2011</td>
</tr>
<tr>
<td>7. Electricity</td>
<td>kWh</td>
<td>3.60</td>
<td>Heidari &amp; Omid, 2011</td>
</tr>
<tr>
<td>8. Seed</td>
<td>Kg</td>
<td>25</td>
<td>Beheshti Tabar et al., 2010</td>
</tr>
</tbody>
</table>

The energy equivalences of unit inputs are given in Mega Joule (MJ) unit. The total input equivalent can be calculated by adding up the energy equivalences of all inputs in MJ.

Data mining technique

The goal of data mining is to discover hidden knowledge in datasets which the human eye or conventional statistical analysis can-not uncover. There are a wide variety of techniques, called Predictive models, which are available to aid and perform predictive analysis (Maione, 2016). In this study, we used a popular technique which has yielded good results in the recent data mining literature: Artificial neural networks. A brief description of this technique follows.

Artificial neural networks

The simplest definition of an artificial neural network (ANN), is a simulation of biological neural system, it is a mathematical model or computational model based on biological neural networks (Arockiaraj, 2013). Neural networks are typically organized in layers. Layers are made up of number of interconnected ‘nodes’ which contain an activation function. Patterns are presented to the network via the ‘input layer’, which communicates to one or more ‘hidden layers’ where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to an 'output layer'. Most ANNs contain some form of 'learning rule' which modifies the weights of the connections.
according to the input patterns that it is presented with. Therefore ANN is like human biological brain learns and recognize by examples and testing. The learning phase based on providing samples of real cases as inputs and use training methods to extract the relations between them to get another output from the given inputs, this is done by calculating the connective weights between layers that work as data processing on input to obtain better output. In learning phase large sample of dataset must cover most of the possible inputs that is considered as an expected rate of error which needs to be minimized as much as possible helping the proposed model to generalize on all possible data. Testing phase: based on providing a sample data considered as other inputs that ANN trained on and extract output to be compared with results to measure the system’s ability to learn, and its reliability to predict and forecast the needed results to be adapted in the real life to be able to solve problems. A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate output. It is a modification of the standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions, and is more powerful than the perceptron in that it can distinguish data that is not linearly separable, or separable by a hyperplane. Multilayer perceptron using a back-propagation algorithm are the standard algorithm for any supervised-learning pattern recognition process (Ayman et al., 2015).

Data analysis

Data mining software of IBM SPSS Modeler 14.2 has been used in the research for modeling neural network and result validation. Data were divided into training and testing data. 70% of data were training and 30% of them were testing data.

Model assessment

In this stage, output model of training data was used for testing data. In other words, testing data (30% of data in the research) are predicted by derived model and results of prediction are compared with reality using this model.

Results & Discussion

Analysis of energy input and output in canola production

The input and output energy values used in canola production are illustrated in table (2). It is evident that, the total energy input for canola production was about 28927.43 MJ ha⁻¹, from which, the highest share was consumed by fertilizers, followed by diesel fuel inputs and electrical energy. Several researchers reported that chemical fertilizers spatially nitrogen were the most energy consuming inputs followed by diesel energy (Ozkan et al., 2004; Mohammadi et al., 2008). On the other hand, the output energy was equal to 58638.75 MJ ha⁻¹ in the farms that were studied. In the other columns of table (2) the standard deviation values, maximum values and minimum values for the inputs and output used in ANN model are presented.

Data mining

In present research, information collected from canola farms of Amirkabir Agro-industry was used as database. The stages of the second research phase are divided into six parts such as data identification and selection, data description, data pre-processing and preparation, data modeling and analysis, data evaluation and validation and model development.
Table (2). Amounts of energy inputs and output for canola production.

<table>
<thead>
<tr>
<th>Item</th>
<th>Total energy Equivalent (unit ha⁻¹)</th>
<th>Percentage (%)</th>
<th>standard deviation (SD) (unit ha⁻¹)</th>
<th>Max (unit ha⁻¹)</th>
<th>Min (unit ha⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inputs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human labor (MJ)</td>
<td>183.05</td>
<td>0.63</td>
<td>35.71</td>
<td>238.18</td>
<td>120.22</td>
</tr>
<tr>
<td>Machinery (MJ)</td>
<td>2398.49</td>
<td>8.29</td>
<td>170.55</td>
<td>2737.98</td>
<td>1940.06</td>
</tr>
<tr>
<td>Diesel fuel (MJ)</td>
<td>5980.22</td>
<td>20.67</td>
<td>805.21</td>
<td>7543.29</td>
<td>4212.77</td>
</tr>
<tr>
<td>Chemicals (MJ)</td>
<td>919.45</td>
<td>3.18</td>
<td>142.87</td>
<td>973.81</td>
<td>420.33</td>
</tr>
<tr>
<td>Fertilizers (MJ)</td>
<td>11450.45</td>
<td>39.58</td>
<td>125.67</td>
<td>12980.30</td>
<td>9073.67</td>
</tr>
<tr>
<td>Electricity (MJ)</td>
<td>4740.15</td>
<td>16.39</td>
<td>297.78</td>
<td>6065.79</td>
<td>4000.55</td>
</tr>
<tr>
<td>Irrigation (MJ)</td>
<td>3025.62</td>
<td>10.45</td>
<td>205.61</td>
<td>2760.92</td>
<td>3500.73</td>
</tr>
<tr>
<td>Seed (MJ)</td>
<td>230.00</td>
<td>0.79</td>
<td>5.04</td>
<td>237.50</td>
<td>200.00</td>
</tr>
<tr>
<td>Total energy input (MJ)</td>
<td>28927.43</td>
<td>100</td>
<td>4311.92</td>
<td>17579.41</td>
<td>29422.83</td>
</tr>
<tr>
<td><strong>Output</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>canola yield (kg)</td>
<td>2345.55</td>
<td>-</td>
<td>307.65</td>
<td>1990.50</td>
<td>2700</td>
</tr>
<tr>
<td>Total energy output (MJ)</td>
<td>58638.75</td>
<td>-</td>
<td>1209.80</td>
<td>48886.51</td>
<td>65708.52</td>
</tr>
</tbody>
</table>

Stage of data identification and selection
In the first stage, information available about consumed inputs in process of canola production has been studied. Concerning collected information, nine effective variables required for modeling energy consumption in process of canola production were defined and stored in informational bank of farms as separate fields. These eight variables are independent variables of the research and they are as follows: human labor, diesel fuel, electricity, irrigation, machinery, fertilizers, chemical and seed. Variable of output energy of the farm is dependent variable of the research.

Data description and definition of Variables used in the model
Data used in this research is entered in IBM SPSS modeler 14.2 in form of an Excel file with nine columns including eight independent variables (input) and one dependent variable (target).

Stage of data pre-processing and Preparation
Before modeling, raw data collected from canola farms were prepared and pre-processed for improvement of data quality. Data pre-processing and preparation include standardization of number of farms, standardization of unique values in different data sources, correction or deletion of inconsistent and irrational values and addition of different sources based on the number of standardized farm. Also, presence of unconventional records in database increased error factor in results of data mining. Therefore, identification and management of such unconventional data is very important in data preparation stage. Concerning statistical data description, there is no missing data among records in this stage.

Stage of data analysis and modeling
Data are modeled in this stage and necessary analyses are presented regarding model validation. In order to predict output energy of canola farms and to enter the software, final
The database had 432 records and 9 fields of input and output variables. In order to model data, field of output energy variable was introduced to the software as output and other variables (fields) were introduced to the software as input variables. The main algorithm used in this research for energy of canola farms is artificial neural network. The network used here is multilayer perceptron which is one of the most practical artificial neural networks. Number of hidden layers of the network is one layer with three neurons (Fig. 1). The condition for stopping modeling is that if error is not reduced, optimization will be stopped. In present model, data prediction precision was 86.5%.

Concerning description of input and output variables, it is noteworthy that some variables are more effective on prediction of model. In the graph of neural network in Fig. (2), descending trend of effect of variables has been determined on target variable.

**Fig. (1): Architecture of a multilayer perceptron.**

**Fig. (2): Predictor importance (Target: output energy).**
The first variable with the highest effect is fertilizers. Fuel and electricity energy are other factors affecting output variable (productive energy). Other variables are effective as well but their effectiveness is less than the three variables.

**Model evaluation and validation**

After preparation of model, it should be evaluated. Evaluation results cause improvement of model and they make it practical. Since the method presented in every research should be measured based on validity and concerning that the research method is data-oriented, validation method is as follows: data are divided into two training and testing data collections. Training data makes the model and testing data evaluates the model. In other words, testing data were predicted by extracted model and prediction results are compared with reality by this model.

Data were divided randomly by software. Regarding number of data in each set, number of training data is always more than testing data. In present research, number of training sets is 70% (34 farms) and the rest (30%) has been considered as testing data (14 farms). Validity value is tested by results of new data and training data are entered the algorithm as observer and evaluate results of its accuracy. Amount of prediction precision in neural network algorithm meaning the ratio of correct predicted records to total records was 86.5%. Therefore, the model made by neural network has high estimative precision. Another argument is how precision of prediction methods can be measured. In order to evaluate precision of prediction models (the difference between the real value and predicted value of dependent variable), index of mean absolute value of prediction error is used. Mean absolute error value (MAE) of training data and testing data is 1531.20 and 2077.38 respectively which is an optimal value (Table 3). It is clear that in order to increase precision of a predictive method, the values of abovementioned indices should be small. Also, linear correlation between real and predicted values is 0.88 and 0.84% for training data and testing data respectively and it suggests a strong correlation.

**Model development**

In this stage, the model and its results can be presented to Amirkabir Agro-Industry manager to be used in future for prediction of energy production during the process of canola production. Therefore, after necessary reports; it was explained the most effective variables on prediction of output energy of canola farms based on this model. Concerning that variables of chemical fertilizers, fuel and electricity and

<table>
<thead>
<tr>
<th>Partition</th>
<th>1_Training</th>
<th>2_Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Error</td>
<td>-1713.71</td>
<td>-239.03</td>
</tr>
<tr>
<td>Maximum Error</td>
<td>1278.11</td>
<td>1617.44</td>
</tr>
<tr>
<td>Mean Error</td>
<td>-532.70</td>
<td>-812.28</td>
</tr>
<tr>
<td>Mean Absolute Error</td>
<td>1531.20</td>
<td>2077.38</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>4990.18</td>
<td>5673.05</td>
</tr>
<tr>
<td>Linear Correlation</td>
<td>0.88</td>
<td>0.84</td>
</tr>
<tr>
<td>Occurrences</td>
<td>34</td>
<td>14</td>
</tr>
</tbody>
</table>

Table (3): Comparing predicted output energy with actual output energy.
irrigation are the most effective variables on output energy of canola farms, energy efficiency will be increased in farms if input variables are optimized.

**Conclusion**

Data of canola farms was used in present paper. In the first research phase, input and output energies of canola farms were analyzed. Total energy consumption in canola production was 28927.43 MJ ha\(^{-1}\). The energy input of chemical fertilizer has the biggest share within the total energy inputs followed by fuel. Energy output was calculated as 58638.75 MJ ha\(^{-1}\). In the second research phase (data mining), database was divided into training and testing parts. Then, a model was created based on artificial neural network technique of data mining in order to estimate energy in process of canola production using IBM modeler 14.2 and training dataset. As a result, the model was assessed using testing datasets and could reach 86.5% precision for estimation of output energy. It is clearly evident that this model has a good accuracy for estimation of amounts of output energy. Also, results of the study indicate that prediction methods provide proper source allocation and increase efficiency of inputs by presenting more accurate picture of energy status in canola farms.

**Acknowledgements**

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**


تقيیم استهلاک الطاقة لنظام انتاج الكانولا في ايران (حالة الدراسة: منطقة امير كیبر الصناعیة الزراعیة)

نسم منجی؛ و اسماعیل حسین زاده

1 قسم هندسه النظم الحیویة، کلیة الزراعت، جامعة شهید جماران، الأهواز، ایران
2 ماجستیر مکنیة زراعیة، امير کیبر، الصناعیة الزراعیة، ایران

المستخلص: زاد انتاج الكانولا بشكل كبير في ایران بسبب ارتفاع إنتاجه في السنوات الأخيرة. تعد محافظة خوزستان هي المركز الرئیسي لانتاج الكانولا في ایران. هدف هذا البحث هو دراسة التنبؤ عن البيانات لنماذج من بذور الكانولا التي تم الحصول عليها من مزارع في منطقة امير کیبر الصناعیة الزراعیة في محافظة خوزستان. تم جمع البيانات من 48 مزرعة. تم اختيار المزارع بطريقة العينة العشویة. الغرض من هذه الدراسة هو تحديد استهلاک الطاقة من المدخلات والمخرجات المستخدمة في إنتاج الكانولا. وتم توقع الطاقة الناتجة من مزارع الكانولا باستخدام التنبؤ عن البيانات والشبكة العصبية متعددة الطبقات. وبدعم بحث تصمیم وقاعدة بيانات تتألف من 432 سجلًا. تم الحصول على البيانات المطلوبة لهذا البحث خلال مواسم النمو من 2017-2018. تم اجراء تحلیل البيانات بواسطة مصنف نماذج IBM SPSS 14.2. أظهرت النتائج أن كمية الطاقة المستهلكة في إنتاج الكانولا كانت 28927.43 ميجا جول هكتار -1. وكان القدرات الأكبر منها حوالي 40% ناتج عن الأسمادة و 37% من الكهرباء وقود الوقود. أما النموذج المستخدم في البحث فقد وجد أن متغيرات الأسمدة الكیمایة والوقود والطاقة الكهربائيه والري على التوالي كان لها التأثير الأکبر على متغير المخرجات (الطاقة الإنتاجیة). مدار دقة التنبؤ في خوارزمیة الشبکة العصیمیة والتي تعنی أن نسبة البيانات المتوقعة بشكل صحيح إلى إجمالي البيانات كانت 86.5%. كان الارتباط الخطی بين القيم الفعلیة والقيم المتوقعة 0.84 و 0.88 على التوالي لبيانات التدريب وبيانات الاختبار والتي تشير إلى ارتباط قوي. يمكن أن تكون النتائج التي تم الحصول عليها في هذا البحث فعالة لمزارع الكانولا في منطقة امير کیبر الصناعیة الزراعیة في اتجاه تبیین وتحسين استهلاک الطاقة في عملية إنتاج الكانولا.

الكلمات المفتاحیة: الطاقة، التنبؤ، التنبؤ في البيانات، الشبکات العصیمیة الاصطناعیة (ANN)، الكانولا.