Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 65 (2015) 633 – 642

Procedia Computer Science

International Conference on Communication, Management and Information Technology (ICCMIT 2015)

Proposed framework for implementing data mining techniques to enhance decisions in agriculture sector

Applied case on Food Security Information Center

Ministry of Agriculture, Egypt

Ayman E. Khedr[a], Mona Kadry[b], Ghada Walid[b]*

[a]*Faculty of Computers and Information, Department of Information Systems, Helwan University, Cairo, Egypt*
[b]*Arab Academy for Science and Technology and Maritime Transport,Department of Information Systems, Cairo, Egypt*

Abstract

Egypt is facing a problem of food insecurity combined with poverty especially in rural Upper Egypt. Food availability is one of the main pillars of food security including production and importation. Information about food availability is collected from different sources but prediction of the needed amounts of the main strategic crops for the upcoming years is not automatically calculated. This paper aims to predict the needed amount of crops to satisfy the Egyptian citizens' needs for the upcoming years by building process of Artificial Neural Networks (ANNs) via WEKA using Multilayer Perceptron (MLP) function as a data mining predictive technique. Results showed that using this data mining framework succeeded to predict the annual needed amounts of the main strategic crops (Wheat, Rice and Beans) up to the year 2020. The obtained results could help decision makers for achieving food security and the country's productivity for the upcoming years continuously.

Peer-review under responsibility of Universal Society for Applied Research

Keywords: Data mining; Food Security; Artificial Neural Network (ANNs); Multilayer perceptron (MLP); Knowledge Discovery (KDD); Food Security Information Center (FSIC).

* Corresponding author: Ghada Walid. Tel.: +20 0100 301 9704.
E-mail address: ghadawalid@aast.edu

Peer-review under responsibility of Universal Society for Applied Research
doi:10.1016/j.procs.2015.09.007

1. Introduction

Food security achieving is a significant and growing challenge and highly critical to reduce poverty rate. Planners, researchers, development agencies, farmers and citizens require information on the production of crops and nutrition requirements for further studies and evolving realistic strategies for improvement of the Egyptian production and to protect the citizens' health from malnutrition and different diseases. Data is also required for monitoring the movement of production prices of crops compared to imports.

Food Security Information Center (FSIC) is responsible for collecting data about the three pillars of food security, analyzing and reporting to decision makers to act with solutions to improve nutritional and health status of Egyptian population especially those in Upper Egypt. In the meantime a lot of data are collected from different resources while they are facing a real problem in analyzing the available data†. Therefore, the study problem could be stated as the Difficulty in analyzing the massive data concerning food security in order to enable decision makers to take appropriate actions for solving problems of food security.

Accordingly this study attempts to implement data mining techniques on Food Security Information Center (FSIC) massive data in order that the Egyptian Ministry of Agriculture can obtain their benefits and predict the required quantities of food and cost to cover the needs for the upcoming years for a healthy life of Egyptian citizens according to the population growth. The analysis of these data sets obtained from FSIC with applying data mining technique (Artificial Neural Network and Visualization) may yield useful outcomes to researchers in the Agriculture Research Center and decision makers. The study applies WEKA application to conduct qualitative analysis and to create a benchmark for the analysis of the dataset. The dataset was then analyzed by building artificial neural network using Multilayer Perceptron (MLP) within the data mining software.

The remaining of this paper is organized as follows; Section 2 introduces theoretical background of study, it is divided into two main issues; first is about Food Security, second is for DM concepts. Section 3 introduces DM in Agricultural Sector. Section 4 presents the proposed Data Mining System. Section 5 shows findings analysis and discussion. Finally, section 6 presents the conclusion and future work.

2. Theoretical Background

This section will discuss the main concept of Food security, then a brief theoretical background on data mining and previous applications of data mining in agriculture sector and food security.

2.1. Food Security

Food security exists when all people, at all times, have physical and economic access to sufficient safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life."[2]

The main target of Food Security Information Center is to help decision makers in agriculture sector with acceptable information about the main pillars of food security which are[3]:

- Food availability: sufficient quantities of food available on a consistent basis.
- Food access: having sufficient resources to obtain appropriate foods for a nutritious diet.
- Food use: appropriate use based on knowledge of basic nutrition and care, as well as adequate water and sanitation.

---

[1] Interviewing with Prof. Dr. Akila Saleh Hamza the founder and coordinator of Food Security Information Center[1]

2.2. Data Mining techniques

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 [4, 5].

The basic DM process shown in **Fig. 1** is as follows:

1. Understanding the business problem and define the main objectives.
2. Collect, describe and explore the metadata. Selected data identifies addressed attributes and variables. It often consists of selecting a time span, geography or product set.
3. Prepare, Clean and transform data into the appropriate format, because some of the mining functions accept data only in a certain format.
4. Apply various data mining algorithms depends on the mining objective, type of application and the computational power to achieve the desired results.
5. Visualize the discovered results to help the analyst to understand and communicate its remarks.
6. Represent the analysis result to decision makers in order to get the most benefit and take decision.

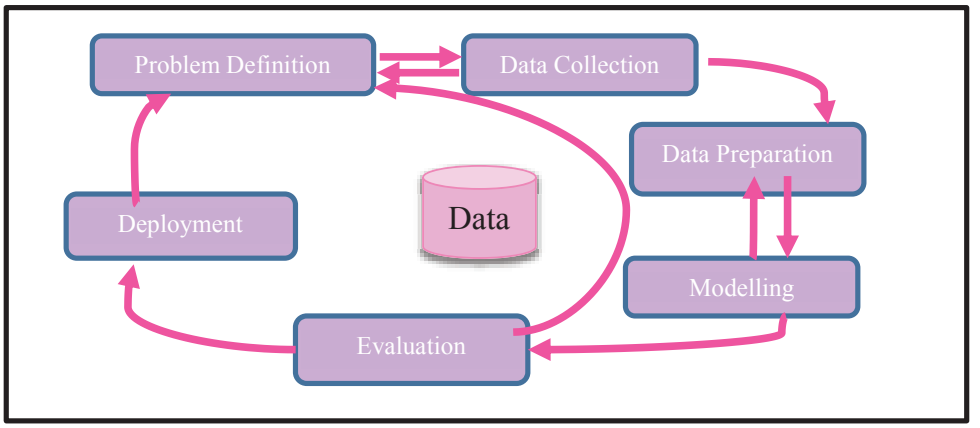

Fig. 1. DM process model[7]

2.2.1 Predictive and Descriptive DM techniques:

Data mining is mainly categorized as descriptive and predictive techniques as shown in **Fig. 2** to accomplish its different tasks to fit a model to the data. A predictive model makes prediction about future values using different and historical data. The predictive models include classification, regression, time series analysis and prediction.

A descriptive model identifies hidden patterns or relationships in data. It explores the properties of the data that has been examined. The descriptive models include clustering, summarization, association rules and sequence discovery. In the agricultural area predictive data mining is mainly used [6].

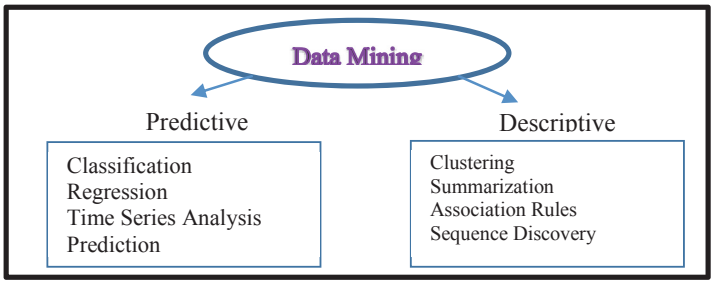

Fig. 2. DM Techniques

3. Related work

DM has been used in Agricultural sector as follows:

Rani[8] in his research about the efficient management of feed resources using data mining techniques mentioned that clustering of feed resource into different groups based on the composition can help in better feed management. The major findings of the study are for clustering techniques that could be used for classifying the feed resources into different groups with reasonably good accuracy. Further, the K-means found to be the best technique for classification. He found that the k means clustering was not affected by sample size and the distribution of different feed resources in different data sets as the outputs were almost consistent across groups.

The outputs of Rani[8] study demonstrate that clustering could be effectively used for grouping of different feed resources without the aid of experts to an extent of 70% and thus can form a sound basis for efficient feed management. The application of the study was that it can be used for clustering new feed resource into a particular category which will be useful in determining the extent of usage of the new feed. Clustering the feeds into different groups provides multiple options to the end user (farmer or feed industry) to choose from a wide range of feed resources that is best suited in terms of local availability or price advantage.

Megala and Hemalatha[9] used data mining approach to determine the vanished agricultural land in Tamilnadu, India. His research aimed to establish if data mining techniques can be used to assist in the clustering methods by determining whether meaningful patterns exist across various land profiles at various research sites across Tamil Nadu in India. The overall aim of the research was to determine the land utilization for agriculture and non-agriculture areas for the past ten year. The research used Software to conduct qualitative analyses and to create a benchmark for the analysis of the dataset. The dataset is analyzed using a clustering process within the data mining software.

Salame[10] applied Data mining techniques to evaluate applications for agricultural loans. The study used Logistics 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.

Raorane and Kulkarini[11] discussed the role of data mining as an effective tool for yield estimation in the agricultural sector. As crop production depends on climatic, 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.

Wang et al[12] mentioned in their paper that evaluation and early warning prediction that Food safety risk early warning is one of the most important contents in food security management. An effective early warning system could improve the management of food security status. A great number of researchers are engaged in the research of food security early warning and they have got a lot of achievements, Data were collected via anonymous questionnaires, surveys, etc. Liu et al.[13] proposed a data-mining technique to predict food quality using a back-propagation neural network. To reduce errors, data near the threshold values are selected to train their proposed system.

Table 1. shows the different applications of DM techniques in agriculture sector and Food Security sector in prior research studies from different perspectives.

Table 1. Previous studies and findings

Year	Case	Techniques	Findings
Rani8	Efficient management of feed resources using data mining techniques.	Clustering	Clustering can be effectively used for grouping of different feed resources without the aid of experts to an extent of 70% and thus can form a sound basis for efficient feed management.
Megala and Hemalatha9	Data mining approach to determine the vanished agricultural land in Tamilnadu, India.	Classification/ Clustering/ Visualization	Reduced the time taken to undertake data analysis. Increased automation of the process.

			The integrity of the data is critical to ensure that results are not affected by outliers and null values in the data set, or other adverse factors. The establishment of clusters in the data required a large amount effort by the researchers when using current methods. Maps classification have been shown in order to understand the land lost with in the past seven years. Which helps in understanding of the biophysical and environmental management.
Raorane and Kulkarini11	Data mining as an effective tool for yield estimation in the agricultural sector.	Classification, Clustering	Artificial Neural Network is a new technique used in flood forecast. Decision tree is one of the classification algorithms. Support Vector Machine is able to classify data samples in two disjoint clusters. SVM are a set of related supervised learning method used for classification and regression. K means method is used to forward the pollution in atmosphere. Different changes of weather are analyzed using SVM. K means approach is used to classify the soil and plants.
Wang et al12	Data mining to make an evaluation and early warning prediction that Food safety risk early warning.		

4. The proposed framework

In this study, a framework has been developed that attempt to explain the key processes that occur in data mining. Through an understanding of these phenomena and processes and by relating these processes to the outcomes of data mining, so as to build a better understanding of how an attempt to improve data mining and their outcomes in making decisions is made.

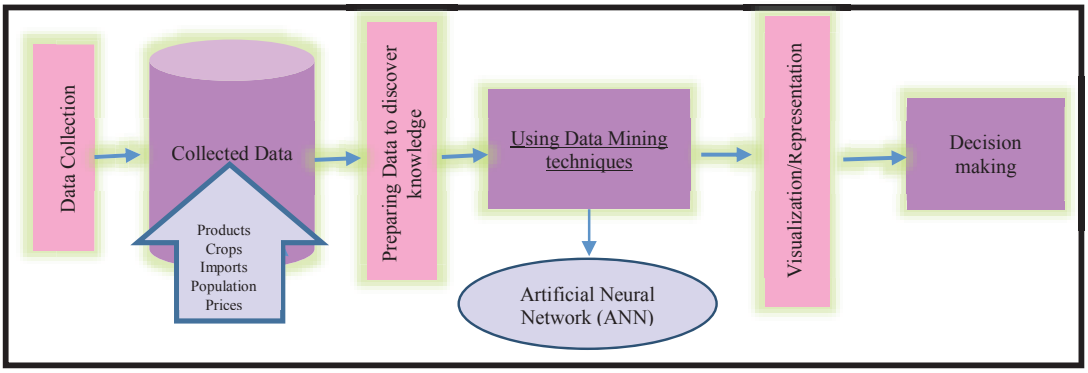

Fig. 3. Proposed Framework

The proposed framework as depicted in **Fig. 3** aims to build a data mining system for predicting the population needs in agriculture sector to overcome food insecurity in Egypt. After defining the goals and objectives of the study; KDD will be started through data selection, pre-processing, processing and applying data mining techniques and visualizing the results to make the appropriate decision.

4.1. Data collection

The data gathered from Food Security Information Center which have been gathered from FAO, WHO, CAMPAS, and real studies about the status of food security in the Egyptian governorates about the rate of population growth, the amount of food needed to satisfy the needs of all citizens, the productivity of strategic crops, the rate of imports of each crop with its price.

4.2. The Choice of Data Mining Tool

WEKA stands for Waikato Environment for Knowledge Learning. It is a popular suite of machine learning software written in Java. It was developed by The University of Waikato, New Zealand.

WEKA is an open source software which consists of a collection of machine learning algorithms for data mining tasks. WEKA is freely available and it is also platform-independent.

4.3. Preparing data to discover knowledge

The data preparation is an important function to data mining which reprocess to the prediction system needs, check the integrity and consistency of data, remove the inaccurate and Worthless data and prepare the mathematical model.

4.4. Artificial Neural Networks (ANN)

Artificial Neural Network is used in modeling the needs of the required amounts of food according to the population growth rate. 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. It consists of an interconnected group of artificial neurons and it processes information using a connectionist approach to computation. Most of times ANN act as an adaptive system that changes its structure based on internal or external or information that flows through the network during the learning phase[15].

Neural networks are typically organized in layers. Layers are made up of number of interconnected ‘nodes’ which contain an activation function as shown in **Fig. 4**. 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[15].

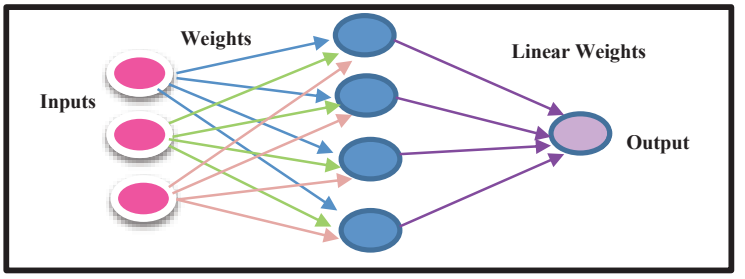

Fig. 4. Neural Network[16]

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 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 hyper-plane[14]. Multilayer perceptron using a back-propagation algorithm are the standard algorithm for any supervised-learning pattern recognition process.

4.4.1 Back-Propagation

Back propagation nets are the most common kind of ANN. The basic topology is that layers of neurons are connected to each other. Patterns cause information to flow in one direction, then the errors "back-propagate" in the other direction, changing the strength of the interconnections between layers. A very simple example of Neural Networks using back propagation this program is a simple example of Neural Networks using back propagation[16]. Back Propagation network learns by example to give weights that used to get output from provided input. Learning steps:

1. Initialized by setting up all its weights to be small random numbers.
2. Apply input and calculate output that different form target one.
3. Calculate error rate as target – Actual.
4. Used error rate to modify weights to get small error rate.
5. Apply step 2, 3, 4 more and more to get final trusted model.

There are different methods to calculate the error rate based on how training process work like:

1. If we used threshold in training the error function = target – actual.
2. No threshold the error function express as outα (1 - outα) (Targetα - outα).

Whatever how error calculate the new weight value will be calculated as:

W (new) = w (old) + n * error rate * output.

Based on this equation we calculate the new weight for each neuron, test error function and test how new weight reduce the different between the output and the target and work in same manner until reach the model's goal.

4.5. Visualization and representation of results to Decision makers

The usage of visualization technology and choice of suitable visualization tools let decision makers confirm the reliability of discovered knowledge and summarizing the data mining results which are likely to be varied on the manifestation, such as chart or the rule representation, provide a strong decision support for the decision maker. The decision makers can understand the results by visualization interface and do a decision.

5. Findings analysis and discussion:

Data obtained from the adapted framework by building process of Artificial Neural Networks (ANNs) via WEKA using Multilayer Perceptron (MLP) function as a data mining predictive technique shows the difference between the available amount of crops and the needed amount to achieve food security according to the population growth up to year 2020, data set trained from the years (2001- 2011) and predict data from the years (2012- 2020) and visualized as follow:

5.1. Wheat

Results shown in **Fig. 5** could explain the amounts of production quantities, imports, losses and other uses, the amount of food available for humans and the amounts required to achieve food security according to the calculated food basket proposed by the National Nutrition Institute (NNI) and FSIC to show the difference between the available

amounts and the needs to cover the requirements of wheat. The forecasting results show that Egypt will have a shortage to cover the needs of the Egyptian citizens from production and imports of wheat, which is considered a major obstacle since wheat is one of the key elements of bread production, which is indispensable in the Egyptian houses.

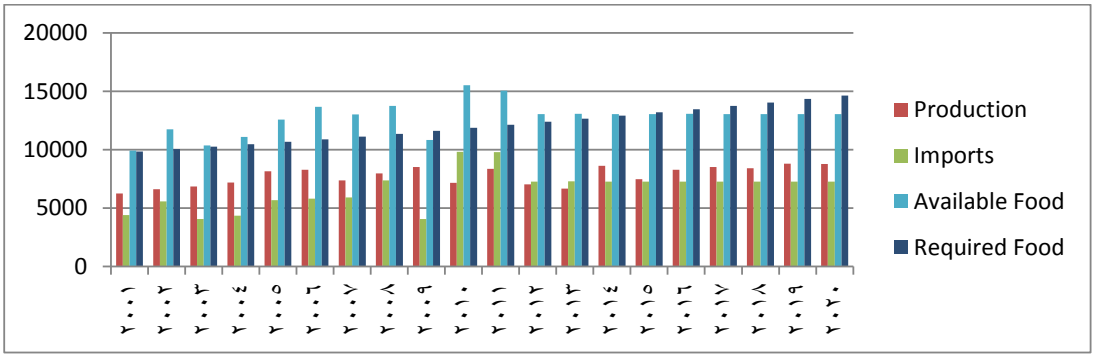

Fig. 5. Available data of Wheat from 2001 to 2011 and the predicted data form 2012 to 2020 in bar chart

5.2. Rice

Results shown in **Fig. 6** could explain the amounts of production quantities, imports, losses and other uses, the amount of food available for humans and the amounts required to achieve food security according to the calculated food basket proposed by the National Nutrition Institute (NNI) and FSIC to show the difference between the available amounts and the needs to cover the requirements of rice. The forecasting results show that Egypt may have a shortage to cover the needs of rice but it could be covered as it is not a large gap between available and required amounts needed to achieve food security to Egyptian citizens.

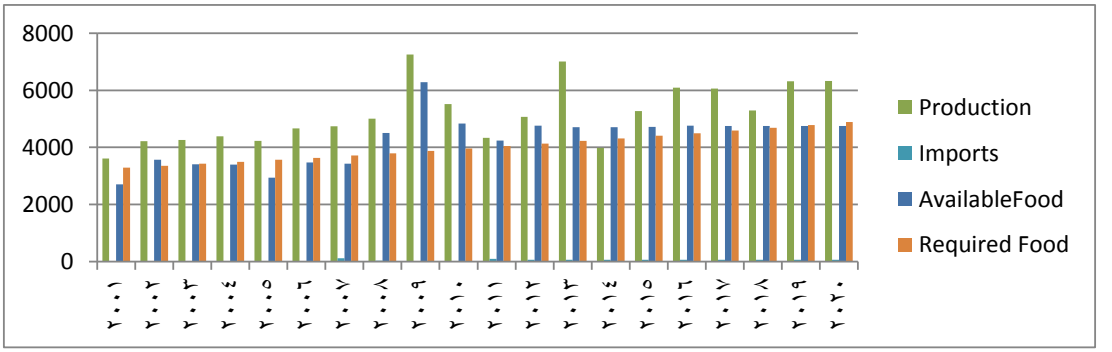

Fig. 6. Available data of Rice from 2001 to 2011 and the predicted data form 2012 to 2020 in bar chart

5.3. Beans

Results shown in **Fig. 7** could explain the amounts of production quantities, imports, losses and other uses, the amount of food available for humans and the amounts required to achieve food security according to the calculated food basket proposed by the National Nutrition Institute (NNI) and FSIC to show the difference between the available amounts and the needs to cover the requirements of beans. The forecasting results show that Egypt will have a shortage to cover the needs of the Egyptian citizens from production and imports of beans, which is considered a major obstacle since beans considered one of the main elements of the Egyptians daily dietary meal.

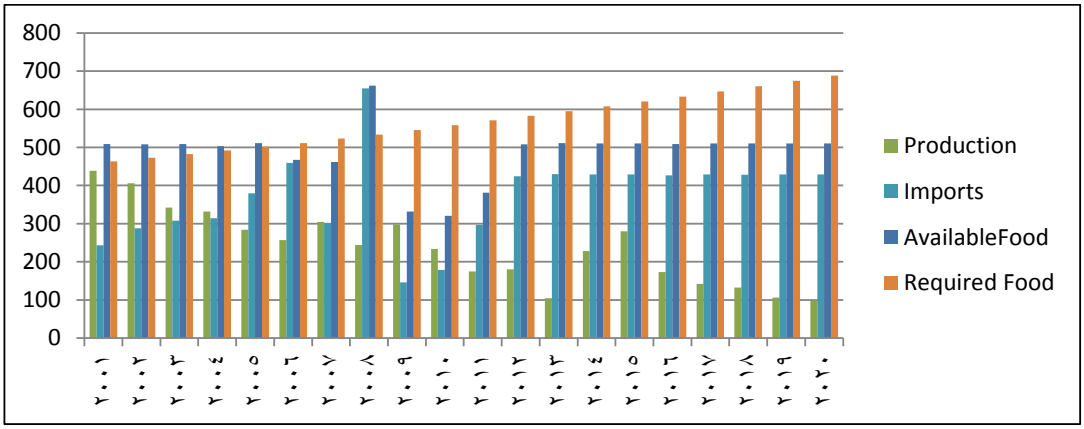

Fig. 7. Available data of beans from 2001 to 2011 and the predicted data form 2012 to 2020 in bar chart

6. Conclusion and future work

This paper attempt to choose the appropriate method in the field of food security for predicting the needed quantity of production and imports to cover the requirements of the strategic crops for Egyptian citizens related to population growth. It presents the difference between the available and the required amounts of food and shows the excess and shortage that should be covered to overcome the status of food insecurity. The results showed that Egypt will face a great problem in the availability of wheat and beans as production and imports will not cover the needs since they are considered the staple food in Egyptian daily diet, but Egypt could cover the needs of rice through decreasing the amount of exports as the production of rice exceeds the Egyptians needs. The researchers and consultants in FSIC agreed and admitted that the study's findings are going to be helpful, more accurate and reliable than their traditional statistical ways for their decisions making.

As for future work, it is recommended to apply and generalize the proposed framework to other crops and also those of animal origin (Milk, Eggs and Meat). Reports based on this model could be great helpful to decision makers involved in the other food security objectives and goals.

Acknowledgements

I would like to express my special appreciation and thanks to Food Security Information Center (FSIC) that provided me with all the needed data.

References

1. Food Security Information Center (FSIC), 2014.
2. Ecker O, and Breisinger C. The Food Security System: A New Conceptual Framework, International Food Policy Research Institute, Paper 01166, 2012, pp. 2-3.
3. World Food Program (WHO), www.who.int/trade/glossary/story028/en/, [Access date: 4 January 2014].
4. Sharma M. Data mining: A literature survey, International Journal of Emerging Research in Management and Technology, Volume-3, Issue 2, February 2014, pp.1.
5. Khedr A. Adoption of new technologies in a highly uncertain environment: The case of Knowledge Discovery in Databases for Customer Relationship Management in Egyptian public banks, Ph.D dissertation, ISBN: 978 90 5291 104 5, Universiteit Leiden, Netherland, 2008, pp. 54.
6. Khedr A., El Seddawy A. , and Idrees A. Performance Tuning of K-Mean Clustering Algorithm a Step towards Efficient DSS, International Journal of Innovative Research in Computer Science & Technology (IJIRCST), ISSN: 2347-5552, Vol. 2, Issue 6, November 2014, pp. 111 - 118.
7. IBM, [Online document], available at: (www.publib.boulder.ibm.com), [Access date: 9 May 2014].
8. Rani V. Efficient management of feed resources using data mining techniques, Master of Philosophy, Department of Computer Science, Christ University, Bangalore, 2010, pp. 5, 87, 88.

9. Megala S., Hemalatha M. A Novel Datamining Approach to Determine the Vanished Agricultural Land in Tamilnadu, International Journal of Computer Applications (0975 – 8887) Volume 23– No.3, June 2011, pp. 23, 24, 26.
10. Salame F. Applying data mining techniques to evaluate applications for agricultural loans, University of Nebraska-Lincoln, 2011, pp. 164,165,166.
11. Raorane A. and Kulkarini F. Data Mining: An effective tool for yield estimation in the agricultural sector, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 1, Issue 2, July – August 2012, pp.1, 2, 4.
12. Wang Y., Tang J. and Cao W. Grey prediction model-based food security early warning prediction, School of Business, Jiangnan University, Wuxi, China, Grey Systems: Theory and Application Vol. 2 No. 1, 2012, Emerald Group Publishing Limited, 2012, pp. 2.
13. Liu Z., Meng L., Zhao W. and Yu F. Application of ANN in food safety early warning, The 2nd International Conference on Future Computer and Communication, Wuhan, Vol. 3, 2010, pp. 677-80.
14. Catur APP R., Rakhman A., Yong Sun G. Building Artificial Neural Network Using WEKA Software, Information System Department, Faculty of Information Technology, INSTITUT TEKNOLOGI SEPULUH NOPEMBER, 2009, pp. 3.
15. Arockiaraj M. Applications of Neural Networks in Data Mining, International Journal of Engineering and Science Vol.3, Issue 1 (May 2013), pp. 8-11.
16. Patel N., Patel R., Gupta A. Advanced Neural Network Applied In Engineering Science, International Journal of Engineering Sciences and Research Technology, Issue (November, 2014),pp. 737- 743.