International Journal of
Intelligent Systems and
Applications in Engineering
ISSN:2147-6799 www.atscience.org/IJISAE
Advanced Technology and Science
Original Research Paper

Knowledge Mining Approach For Healthy Monitoring From Pregnancy Data With Big Volumes

Yunus Santur[1], Sinem Güven Santur[2], Mehmet Karaköse*[3]

Accepted 3rd September 2016

Abstract: The process for obtaining information that will create value on a large-scale data stack is called data mining by its general name. Data mining is commonly used in sales and marketing departments, in determining strategies and making critical decisions for the future in many sectors. Similarly, data mining is used in the determination of health policies, more effective implementation of health services and in the management of resources and institutions in the health sector. In this study, it was aimed to create a software architecture of data mining that will help the personal monitoring of the pregnancy process in a more effective way in the health sector. Many different types of data such as age, gender, location, education, physical characteristics, lifestyle habits and medical history of the people that could be used for this purpose are stored online by health institutions. The machine learning algorithms have been created to determine classification, clustering and association rule on these data.

Keywords: *Data Mining, Knowledge Mining, Healthy Monitoring, Classification, Clustering, Machine Learning*

1. Introduction

The data generated by the computer systems do not make a sense by themselves, and their conventional analysis by human intervention is also impossible. Today, the diversity of data generated in fields such as stock market, trade, education, communication and e-commerce in addition to the fact that these data are large by volume and continue to increase are the main reasons of this.

In today's world, obtaining valuable information by interpreting these data by means of intelligent algorithms is called data mining. In other words, data mining is the discovery and interpretation of the significant relationships through computer programs that could be effective in providing future predictions from among large data stacks [1, 2].

Data mining has a great importance in many aspects from sales strategy of businesses to determination of Research and Development activities.

The analysis and interpretation of large-scale data and uncovering the relationships between data are difficult and long processes by human labour and talent. Today, performing knowledge retrieval using data mining is commonly used in many fields such as marketing, banking, stock market, communication systems, education, health and engineering.

In the marketing field, an attempt to bring out user patterns using data mining was made by Ünal et al. in order to determine the customer portfolio and to describe the user's shopping behaviors by using data of an internet retailer [3].

In the banking field, Tosun et al. tried to determine the customer losses and the reasons of the losses by constructing the profile of the 30,000 customers of Yapı Kredi Bank [4].

Between the years 2007-2013, Öcal et al. used data mining to make a qualitative classification according to the financial success or failure of the companies and the capacity situations of the companies by using the data of Borsa İstanbul [5].

A large telecommunication company's customers who showed a tendency to leave were determine by Gürsoy in 2010, and new campaigns for these customers and marketing strategies for the customer profile were developed by classification method [6].

In the research carried out by Gülen at al. on gifted children aged 7 years and older in Ankara science and art center, the education for the children's individual needs was differentiated through educational data mining and the course schedules were organized according to children's field of interest [7].

In the engineering field, Kaya et al. compared the open source data mining softwares Keel, Knime, Orange, R, RapidMiner (Yale) and Weka. In the study, which data mining method and which software would be more effective on which data set was investigated [8].

Tantuğ et al. examined the techniques used in data mining and especially investigated the clustering techniques. An efficient software architecture to reduce memory complexity was designed in the study. An efficient system with a lower complexity capable of clustering on large-scale data was developed by the software architecture designed [9].

The healthcare field is among the extensive areas of usage of data mining. Ertuğrul et al. constructed the patient profiles by processing and using the data in the patient information management system of Pamukkale University Hospital. Based on the patient profile information, healthcare professionals established a decision support system to provide taking most objective and effective decisions about the patient by reaching up-to-date information about the patient [10].

Irmak et al. processed the patients' information in the database of a hospital continuing to provide services via data mining

[1,3] *Computer Engineering Department, Engineering Faculty, Fırat University, 23100, Elazığ/Turkey*
[2] *Papatyasoft Software, Fırat Teknokent, 23100, Elazığ /Turkey*
* *Corresponding Author: Email: mkarakose@firat.edu.tr*
Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03, 2016.

techniques and developed an application for estimating the patient density of the hospital in the future [11].

By descending to particulars, in the applications in the field of health, Yıldız et al. developed an application that classifies the breast cancer types by using gene algorithm and data mining about breast cancer [12].

In this study, it was aimed to build software architecture of data mining for the effective personal follow-up of women during their pregnancy period.

Many different types of data such as the age, gender, location, education, physical characteristics, lifestyle habits and medical history of the pregnant could be collected and processed with data mining techniques.

Thus, the control frequency could be determined according to the personalized care, nutrition content, situations that require education and risk status during pregnancy by the expectant mothers' data interpreted. Consequently, an approach that uses data mining in following the pregnancy process was proposed in the study.

2. Data Mining

As a general definition, data mining is the process of mining information by reaching the information from among large-scale data. With another definition, it is to search for the connections that could allow us to make predictions for the future from among the large-scale data stacks using computer programs [13].

In almost every stages of our life, electronic recording systems such as mobile phones, sensors and computers provide convenience for recording and storing information and reaching them where necessary along with the technological developments. To be able to make sense out of these data which are rapidly increasing by volume, to establish relationships by them and to finalize data by a decision or conclusion have led researchers to carry out much more investigations in the field of data mining.

Several processes and some techniques are used while performing data mining studies independently of the problem [13, 14]. These processes involve the use of data cleaning, data integration, data reduction, data conversion and data mining algorithms and ultimately the interpretation of the obtained results and the steps of verifying these results, respectively [15,20].

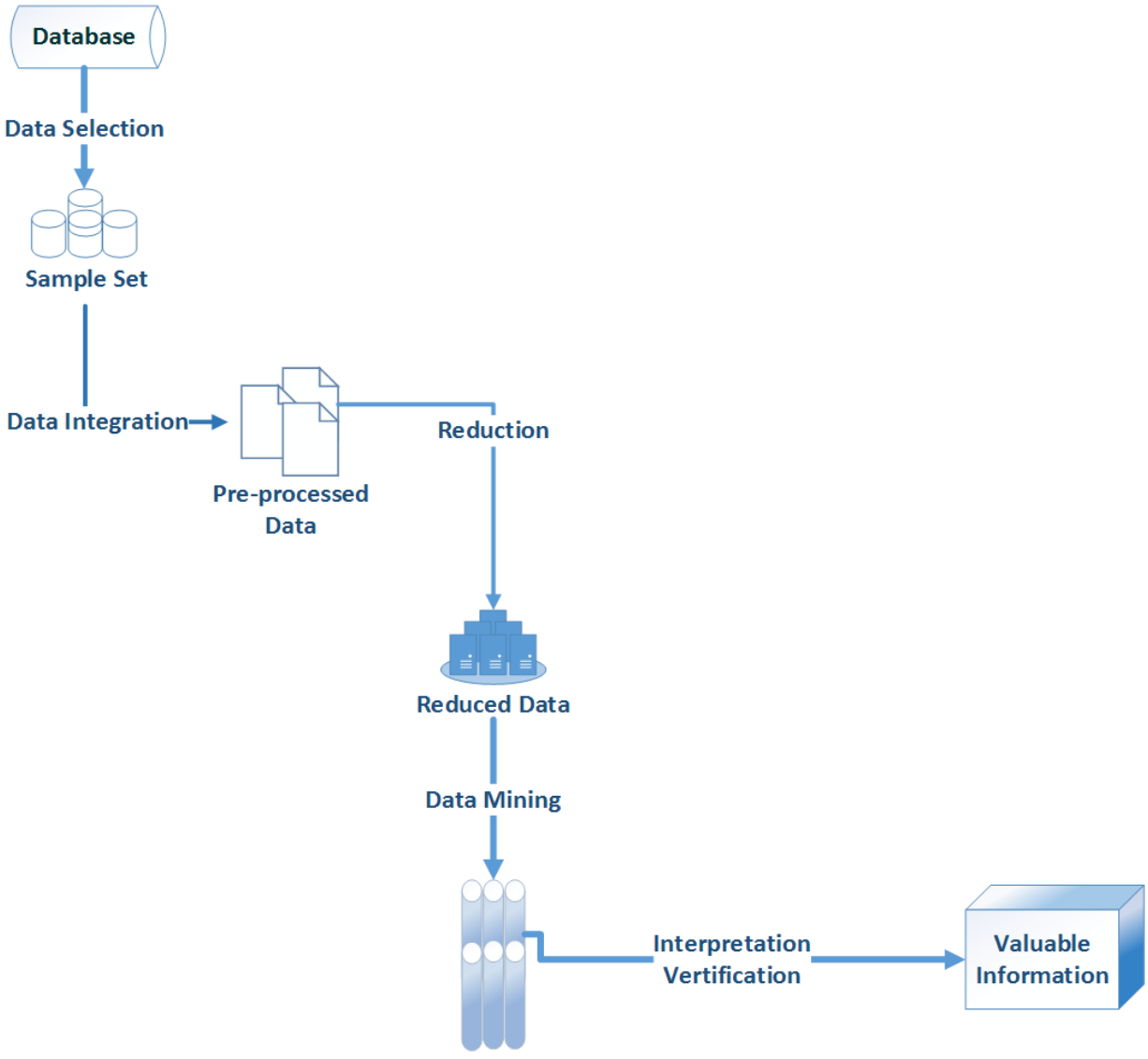

Figure 1. Data mining proces

Table 1. Data mining process

Stage	Explanation
Data Cleaning	The bad and inconsistent data found in the database are called noise. These data can exit the process or the processing can be performed by setting a fixed value instead of them. This is called data cleaning.
Data Integration	It is the process of combining different types of data taken from different databases by converting into a single data type.
Data Reduction	The number of data or variable can be reduced by numerous data reduction methods such as data compression, merging, sampling and generalization.
Data Conversion	It is the process of changing the format of the data by protecting its content according to the algorithm model to be used.
Application of Data Mining Algorithms	It includes the processes of applying machine learning algorithms [21-23] on data which were made ready through the above processes for classification, clustering or determining the association rule..
Interpretation and Verification	The data on which algorithm was applied are verified by bringing out certain significance relationships.

There are three main methods of data mining to make interpretation and verification [15, 16]. These methods are called as classification, clustering and determining the association rule,

1) Classification: It is one of the fields where data mining is commonly used. People often adopt this method because they always classify, categorize or grade the data. Rules are determined by keeping some of the data found in the database separate for education, and how to decide when a new situation arises is determined by this rule.

2) Clustering: It is the process of grouping data by taking into account the similarities between them. By this method, clusters and sub-clusters are generated based on the similarities and differences between variabilities.

3) Determining the Association Rule: It is a data mining method which is applied on the possibility of which situations can occur simultaneously by identifying the relationships of the data included in the database with each other.

3. Proposed Methodology

In this study, expectant mothers' information shown in Fig. 2 were used and the process of knowledge retrieval was performed using data mining on these data.

The software architecture of the proposed method, the block diagram of which is given in Fig. 2, works in four stages. In the first step, a no-sql based database was created for storing data. In the second step, preprocessing was performed to extract information on data set. In the third step, classification, clustering and association rule were identified by running three separate data extraction algorithms. In the fourth and final step, estimation on new data, identifying the cluster to which it belongs and the process of finding the other data to which it is related were performed.

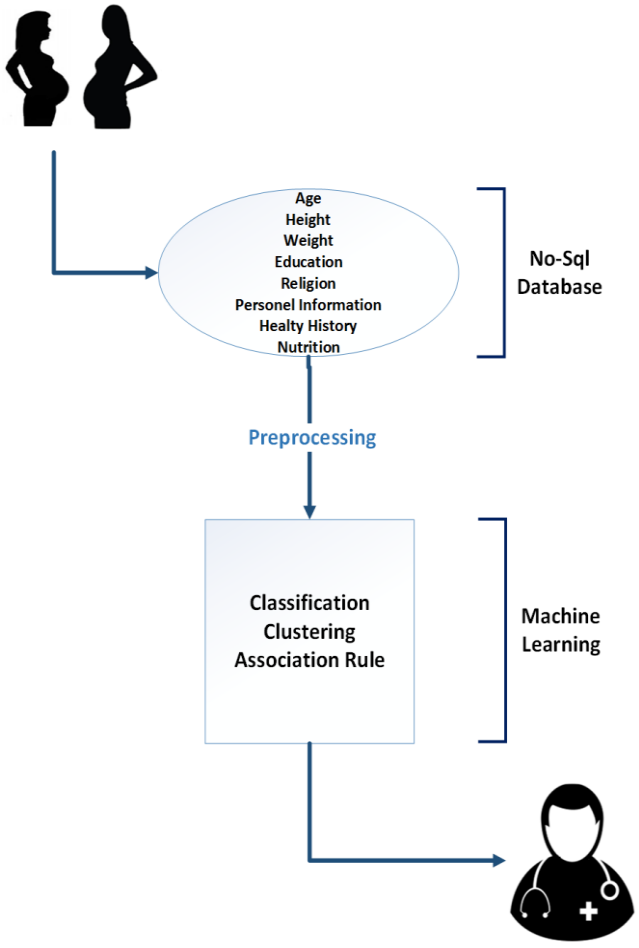

Figure 2. Block diagram of the proposed system

In the study, expectant mothers' information shown in Fig. 2 were used and the process of knowledge retrieval was performed using data mining on these data. The sample training set, main data types and the normalization of the data type into a certain range if necessary are given in Table II.

All values shown in Table II were converted to integer numbers by normalization, if needed, depending on the primary data types. The values of the sets obtained after normalization constituted the input data of the training algorithm. According to what some values are normalized is explained as follows.

Body mass index: It is a value calculated based on height and weight, and it is divided into 3 groups in practice [17].

Age: It is normalized to 3 different groups by considering the fertility criteria [17].

Disease: Whether the person has a disease is discussed as a Boolean value.

Pregnancy History: It includes the statistics of the person's pregnancy history, if available. They are numerical data such as the number of children if she had a baby, whether the previous pregnancies were concluded with healthy results, cesarean and the number of normal delivery.

Other: The personal data such as the person's income level and nutritional habit apart from these were used by being converted into integer values by grouping once again.

Table 2. Training set data types and their normalization

Data	*Data Type*	*Normalized Data*	*Sample Data* 1	2	…	N
Age [17]	Integer [0-2]	0-19: Adolescent 19-35: Adult >35: Menopause	0	1		2
Body mass index [17]	Float [0-3]	<18.4: Underweight 18.5-24.9: Normal 25-29.9: Overweight >30-34.9: Obesity	1	1		2
Number of babies	Integer [0,n]		2	1		2
Number of abortions	Integer [0,n]		1	0		1
Number of natural birth	Integer [0,n]		0	0	…	2
Number of cesarean	Integer [0,n]		2	1		0
Prosperity Index	Integer [0,100]		50	25		60
Disease	Boolean {0,1}		0	1		0
		…				
Dietary	String [0,n]		2	1		3

Matlab and python (scikit library) are mostly used for data mining studies in computer environment. In addition to these, there are also various tools which are free and open source and also include sample data-sets [8, 19]. However, a data-set was generated in a synthetic dataset matlab environment in accordance with Table II because there was not any dataset specific to this study, and it was saved in json format as no-sql based.

In the sample application scenario, random forest for classification and k-means for clustering were used [18]. The target class labels chosen for classification in the training algorithm were identified as “healthy pregnancy” and “risky pregnancy”. In the test phase, the results were analyzed by achieving classification accuracy according to "1" based on the confusion matrix given in Table 2.

Table 3. Confusion matrix

		Actual Class	
Prediction Class		**P**	**N**
	P	TP	FP
	N	FN	TN

$$\textbf{Accuary} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{N}} \quad (1)$$

4. Experimental Results

In the study, classification was performed on the synthetic dataset generated in matlab environment by random forest method. 75% of the dataset was used in training. The remaining 25% was used for testing purposes, and the results were compared on confusion matrix by comparing with actual values. The classification accuracy performance was above 80% in the scenario chosen for the sample classification application.

Table 4. Confusion matrix of experimental result

	Actual		
Predicted		P	N
	P	210	55
	N	45	190

5. Conclusion and Future Work

Today, data mining is commonly used with the purpose of obtaining valuable information in fields such as banking, education, sales, stock market and health. The information obtained as a result of data mining are effective in making strategic decisions.

In this study, it was aimed to perform pregnancy follow-up process using data mining in the follow-up process of pregnancy and retrieve information related to this process. Thanks to the proposed method, it was aimed to use these information in the pregnancy periods of new expectant mothers. The classification application was performed by creating a sample scenario for the study.

In addition to the study, it is aimed to create other similar scenarios in the future and make comparison with other open source tools used for performing data mining.

Acknowledgements

This work was supported by the TUBITAK 1512 Programme (The Scientific and Technological Research Council of Turkey) under Grant No: 2

References

[1] Savaş S., Topaloğlu N and M. Yılmaz M. Veri madenciliği ve Türkiye'deki uygulama örnekleri, 2012.

[2] Wu X., Zhu X., Wu G. Q., Ding W. Data mining with big data, IEEE transactions on knowledge and data .eng., 26(1), 97-107, 2014.

[3] Ünal O.O. Internet Kullanım Analizi Ve Kullanıcı Betimleme Konularında Veri Madenciliği Uygulamaları, Doctoral dissertation, Fen Bilimleri Enstitüsü, 2015.

[4] Tosun T. Veri Madenciliği Teknikleriyle Kredi Kartlarında Müşteri Kaybetme Analizi, Doctoral dissertation, Fen Bilimleri Enstitüsü, 2015.

[5] Ocal N., Ercan M. K and E. Kadioglu. Predicting Financial Failure Using Decision Tree Algorithms: An Empirical Test on the Manufacturing Industry at Borsa Istanbul. International Journal of Economics and Finance,7(7), 189, 2015.

[6] Gürsoy U.T.Ş. Customer Churn Analysis in Telecommunication Sector", İstanbul University Journal of the School of Business Administration, V.39-1, 35-49, 2010.

[7] GÜLEN Ö and Özdemir S. Veri Madenciliği Teknikleri İle Üstün Yetenekli Öğrencilerin İlgi Alanlarının Analizi, Üstün Yetenekliler Eğitimi ve Araştırmaları Dergisi (UYAD), 1(3), 2013.

[8] Kaya M. and Özel S.A. Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması, Akademik Bilişim, 2014.

[9] Tantuğ A.C. Veri Madenciliğin Ve Demetleme, Doctoral dissertation, Fen Bilimleri Enstitüsü.

[10] Ertuğrul İ., Organ A. and Şavlı A. The Determination of Patient Profile at Pamukkale Unv. as Relater to the Application of Data Mining, 2013.

[11] Irmak S., Köksal C. D. and Asilkan Ö. Hastanelerin Gelecekteki Hasta Yoğunluklarının Veri Madenciliği Yöntemleriile Tahmin Edilmesi". Uluslararası Alanya İşletme Fakültesi Dergisi, 4(1), 2012.

[12] Yıldız O., Tez M., Bilge H. Ş., Akcayol M. A. and Güler İ. Meme kanseri Sınıflandırması İçin Veri Füzyonu Ve Genetik Algoritma Tabanlı Gen Seçimi, Journal of the Faculty of Engineering & Architecture of Gazi University, 27(3),2012.

[13] Sharma S., Osei-Bryson K. M. and Kasper G. M. Evaluation of an integrated Knowledge Discovery and Data Mining process model, Expert Systems with Applications, 39(13), 11335-11348, 2012.

[14] Braha D. Data mining for design and manufacturing: methods and applications, (Vol. 3). Springer Science & Business Media, 2013.

[15] Cios K.J., Pedrycz W., Swiniarski R. W. Data mining methods for knowledge discovery, 458, Springer Science & Business Media, 2012.

[16] Larose D.T. Discovering knowledge in data: an introduction to data mining, John Wiley & Sons, 2014.

[17] Schieve L.A., Cogswell M. E., Scanlon K. S., Perry G., Ferre C., Blackmore-Prince C. Prepregnancy body mass index and pregnancy weight gain: associations with preterm delivery, Obstetrics & Gynecology, 96(2), pp.194-200, NMIHS Coll. Working Group, 2010.

[18] Witten I.H., Frank E. Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, 2005.

[19] The New Stack (2016). Available: http://thenewstack.io/six-of-the-best-open-source-data-mining-tools.

[20] Santur S.G., Santur Y. Knowledge Mining Approach For Healthy Monitoring From Pregnancy Data With Big Volumes, International Conference on Advanced Technology & Sciences (ICAT'16) pp. 145-148, 2016.

[21] Santur Y., Karaköse M., Aydın İ., Akın E. IMU based adaptive blur removal approach using image processing for railway inspection", In 2016 International Conference on Systems, Signals and Image Processing (IWSSIP'16), pp.1-4, IEEE, 2016.

[22] Santur Y., Karakose M., Akın E. Random Forest Based Diagnosis Approach for Rail Fault Inspection in Railways", International Conference on Electrical and Electronics Engineering (Eleco'15), 9.th, pp.714-719, 2015.

[23] Santur Y., Karaköse M., Akın E. Condition Monitoring Approach Using 3d Modelling Of Railway Tracks With Laser Cameras, International Conference on Advanced Technology & Sciences (ICAT'16) pp. 132-135, 2016.