Örgütsel Davranış Araştırmaları Dergisi
Journal Of Organizational Behavior Research
Cilt / Vol.: 7, Sayı / Is.: S, Yıl/Year: 2022, Kod/ID: 22S0-787
2528-9705

The design of a data mining algorithm to predict gestational diabetes in Iran

Paniz Arshadi Kalvanagh

Department of Computer and Information Technology, West Tehran Branch, Islamic Azad University, Tehran, Iran

*Corresponding Author
E-mail: panizarshadi2011@gmail.com

ABSTRACT

This study examines diabetes in pregnant women and the necessary parameters for timely diagnosis of gestational diabetes. Diabetes screening and diagnosis test is proposed for all pregnant women and for non-diabetic (normal and pre-diabetic) pregnant women. This research has used data mining methods to determine and apply the appropriate data mining algorithm and predict gestational diabetes in Iran. Available information can be accessed in the form of text files in the databases of public hospitals in the city of Ahvaz. The paper files of the patients of these hospitals are also used if needed. In these hospitals, the data of 8,882 pregnant women who visited midwives from 2013 to 2017 for prenatal tests are analyzed. This information is stored in the form of a database and completely coherent and separated in a similar system called Hospital information system (HIS)[1]. This study has used various data mining algorithms such as decision tree for data analysis and also Rapidminer tool and Rstudio[2] has been used for data analysis. According to the results, data mining software has correctly predicted 95.99% of the data, which indicates that it is a suitable method for predicting diabetes in pregnant women. According to the results obtained on the set of diabetic patients, among the classification methods, the decision tree obtains a better result, and K Nearest Neighbour[3] also has a higher accuracy than the classification method, which can be placed as a support method along with doctors' decisions to improve the accuracy of predicting diabetes in pregnant women.

Keywords: *diabetes, pregnant women, data mining, decision tree, nearest neighbor algorithm*

INTRODUCTION

Gestational diabetes is diabetes diagnosed for the first time during pregnancy (gestation). A number of worldwide studies have reported different prevalence rate of gestational diabetes prevalence varying from 4% to 17%. If this disease, one of the important issues in the field of health, is not treated, it will bring many problems and side effects for the mother and her child. In order to prevent unwanted complications that always plague them, it requires serious attention and timely warning in the early months of pregnancy. The probability of getting diabetes is much higher than in the past due to the change in lifestyle in industrial cities and

[1] A hospital information system is an element of health informatics that focuses mainly on the administrational needs of hospitals.
[2] RStudio is written in the C++ programming language and uses the Qt framework for its graphical user interface. Work on RStudio started at around December 2010, and the first public beta version (v0. 92) was officially announced in February 2011.
[3] K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classifies a data point based on how its neighbours are classified.

reduced mobility, increased obesity and the number of diabetics, less physical activity and increasing age of marriage and other effective environmental factors. In recent years, diagnostic systems have been developed for diseases such as diabetes, which can be used in different situations. Different traditional methods for diagnosing diabetes are based on physical and chemical tests. These methods can be accompanied by errors due to various uncertainties, for this reason, a number of data mining algorithms have been designed to overcome this uncertainty, such as Expectation Maximization (EM) algorithm, k-nearest neighbors (KNN) algorithm, K-means algorithm, KNN amalgam algorithm and adaptive network-based fuzzy inference system (ANFIS) algorithm.

It is possible to refer to the research conducted in [1] as one of the internal research conducted in the field of using data mining in the prediction and prevention of gestational diabetes. Studies have been done in [1] in an applied-survey research about data mining based on neural network and decision tree methods in early diagnosis of the risk of gestational diabetes, and two neural network and decision tree approaches have been used in data mining for experimental data analysis and prediction. The extracted data have been normalized, and analyzed after preparation in Matlab. This study seeks to find the answer to the question that "Are the two data mining methods of neural network and decision tree having the necessary accuracy in timely and correct diagnosis of the risk of gestational diabetes?" And can they be used for correct diagnosis?. According to the results of the research, data-oriented methods are effective in improving the accuracy of forecasting, they have a suitable performance in discovering implicit knowledge and identifying hidden relationships between data, and the decision-making error in both methods is acceptable and very close to each other. The authors have acknowledged in [2] that the purpose of this reference is to find solutions for disease diagnosis by analyzing the patterns in the data through classification analysis using simple Bayes and decision tree classification algorithms. The researchers have believed to provide a faster and more efficient method to diagnose the disease that leads to the timely treatment of the patients. Eyre et al [2] have acknowledged that automatic diagnosis of diabetes is an important medical problem in the real world. Diagnosing diabetes in the early stages is considered as a key to its treatment. Their experimental results prove the effectiveness of these models. The performance of techniques for diagnosing diabetes has been studied and these algorithms are known to be suitable [2].

It has been acknowledged in [3] that during pregnancy, when a woman has high blood glucose (glucose) and does not have pre-pregnancy diabetes, the diagnosis is gestational diabetes. According to the criteria of diabetes, it has recently been announced that about 18% of pregnant women have gestational diabetes. The purpose of this reference is to improve the diagnosis of gestational diabetes using data mining techniques. Also, this authority has investigated the performance of supervised learning algorithms, simple Bayes classifier, $C_{4\,5}$ and random tree in this field. According to the experimental results, the random tree is known as one of the best and most accurate algorithms in this field. Meri et al [4] have acknowledged that a very large number of researches have been conducted in the field of predicting diabetes, J48 and Cardiac resynchronisation therapy (CRT) are other examples of these methods. In [4], the authors have used algorithms in this field, including mathematical expectation-maximization (EM) algorithm, genetic algorithm and H-means clustering to classify diabetic patients. The performance of these algorithms is better than other algorithms when all the symptoms are similar to the clusters.

Mirsharif et al [5] have designed a fuzzy expert system[4] and a fuzzy model approach. Fuzzy inference expert system has been used to diagnose gestational diabetes in conditions of lack of access to a doctor. Fuzzy expert system is a method based on a model that needs to collect the knowledge of experts. In the fuzzy expert system model, rules are extracted and collected in the knowledge base under the supervision of experts, experts and medical reference books, which can be used as decision support. It has been acknowledged that the mean square error level in the fuzzy inference system is 0.2%. The reference [6] has used two old classification methods (logistic regression and Fisher's linear analysis and machine learning classifiers (neural networks, support vector machines, neural fuzzy clustering (c-mean) and random forests) to classify diabetic and non-diabetic people. The database used in this study included 6,500 cases. Sampling was done based on a cluster sampling of the population of Iran that was done in 2005 and 2007 in order to evaluate the prevalence of major risk factors of non-communicable diseases in hospitals of Tehran. Ten risk factors that are usually associated with diabetes were selected to compare the performance of six cases in terms of sensitivity, specificity, complete accuracy and area under the receiver factor curve criterion. In [6], a comparison (based on real information) has been made in the field of data mining methods in predicting diabetes in Iran. In this reference, it has been acknowledged that the discriminating performance of support vector machine models has been better than other methods. Therefore, it can be done with simple clinical measures to successfully diagnose diabetes. It has been acknowledged in [7] that those who have the potential to develop diabetes can be identified by examining risk factors such as body mass index and family history of diabetes such as logistic regression. The authors believe that it is important to predict the predictive power of diabetes prediction models through learning theory and data mining technologies for classification that does not require distributional assumptions. Classical techniques such as logistic regression and Fisher's linear discriminant analysis have been widely used to classify different problems, especially in cases where the dependent variable is bipolar. It seems that there is still a need for more investigation regarding the positive performance of data mining methods with dividers such as neural networks, support vector machines, fuzzy C-Mean and random forests. Reference [8] has also investigated the use of data mining in predicting diabetes and the methods of Gaussian Mixture Model (GMM), support vector machine (SVM), logistic regression, neural network, etc. The authors have acknowledged that the use of artificial network has brought more suitable results in the prediction of diabetes. It is mentioned in [9] that one of the popular and common classifications for classification and prediction is the decision tree, which it is possible to implement it easily and simply and interpret its results. Decision trees are able to produce human-understandable descriptions of the relationships in a data set and provide their predictions in the form of rules that are suitable in terms of statistical parameters. This learning method is used for discrete functions and erroneous data and helps to discover knowledge. According to the results of various researches, each of these researches have tended to select and examine only a few data mining algorithms, and finally, with the results obtained, they have given their suggestion about the best algorithm. In addition to this, few researches have been

[4] Fuzzy expert systems, which work based on the fuzzy-logic approach, can model the rules obtained from fuzzy preferences of experts and can provide outputs by using these rules. The main elements of a fuzzy expert system are fuzzy logic, fuzzy base rule, fuzzy inference, and learning method (Siler and Buckley, 2005).

done in our country in the field of gestational diabetes, and there is still a lot of discussion about the appropriate algorithm for the Iranian sample. In addition, past researches usually have a limited statistical population, and it can be very useful to use a suitable statistical population. Much attention has been paid to decision trees and neural networks [1]. However, more studies are still needed in the field of selecting the appropriate algorithm type. Meanwhile, the type of algorithms and their verification are the most important factors to consider. Algorithms accuracy may also depend on regions, for this reason, input data and verification should be based on information obtained from domestic hospitals. It is also necessary to check the results of the algorithm with data samples from abroad. In this research, the main issue is to check the accuracy of various data mining algorithms for predicting gestational diabetes in Iran.

Proposed method for diagnosing and predicting diabetes

This study is considered as an applied research in terms of its purpose, and its results are specific and objective and have a practical aspect. The descriptive-predictive method has been used in the present study and the statistical population includes information and parameters of all pregnant women who referred to Ahvaz Amir Kabir Hospital, Ahvaz Amir_Ol_Momenin Hospital, Shahid Rajaei Polyclinic Ahvaz, Ahvaz Isar Clinic and Ahvaz Imam Hossein (AS) Clinic and were visited and their information has been stored in the Hospital Information System (HIS) database. This information was analyzed by data mining software. To investigate and diagnose gestational diabetes in pregnant women, the information available in the databases of Ahvaz public hospitals can be accessed in the form of an Excel text file. The paper files of the patients of these hospitals and the HIS of the hospitals were used to extract information from these files. This study has dealt with part of the methods of diagnosis of diabetes during pregnancy based on data mining, which include decision tree, simple Bayes classifier and SVM. Then, a part of data analysis and modeling was implemented by Rstudio. Then the continuation of data analysis was done in Raidminer, the main data model was executed based on Raidminer.

Results

Modeling in Rapidminer

This view is at the center of the design perspective, and you can place them in this view by choosing the appropriate operators for your work. In fact, a process arises from the connection of several operators in this view. Operators are the constituent components of a process, which you can see in the figure below, the constituent components of an operator:

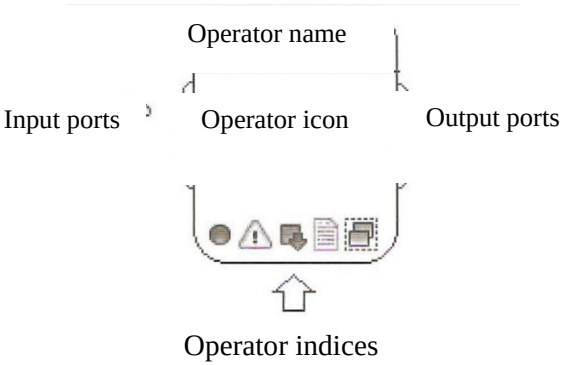


Figure 1- Overview of operators

After executing the operators, it will show us the data matrix in the Result section:

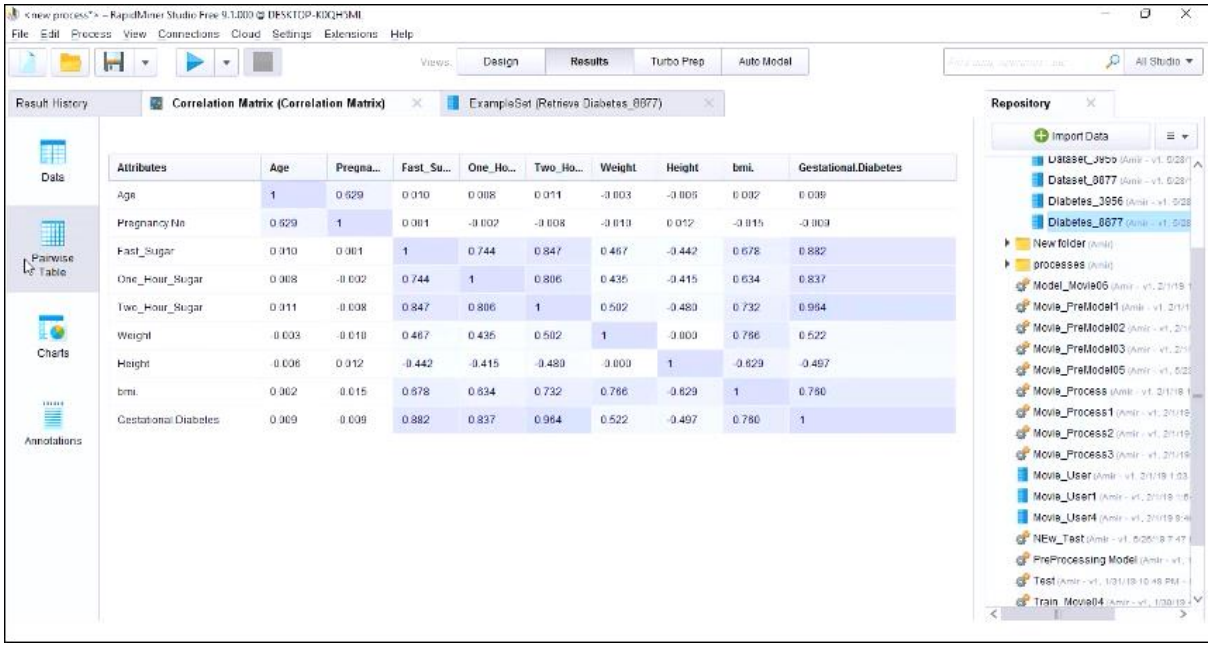


Figure 2: The result of the data matrix

Then, the data chart can be executed from the chart section as shown below:

Figure 3: Data chart diagram

Then, the correlation of each of the attributes with the target attribute is determined as shown below by adding the Set Role operator, with the help of this operator you can determine a role for the specific attributes.

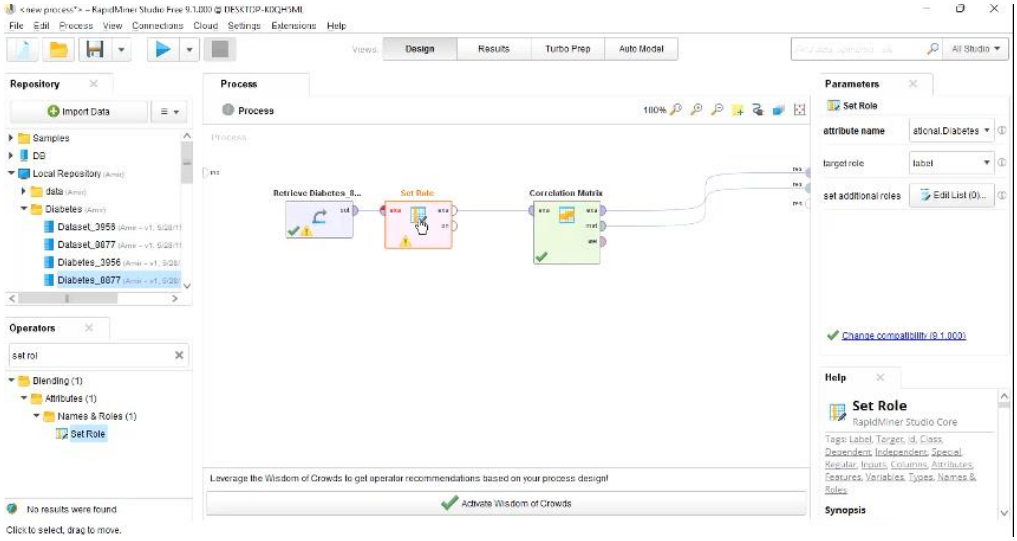


Figure 4: Adding the Set Role operator

Then, we save it in the Diabetes section called Correlation processes. Now we want to draw the Information Gay model in the perspective view using appropriate operators. As shown below, we enter Diabetes data as an operator in the screen:

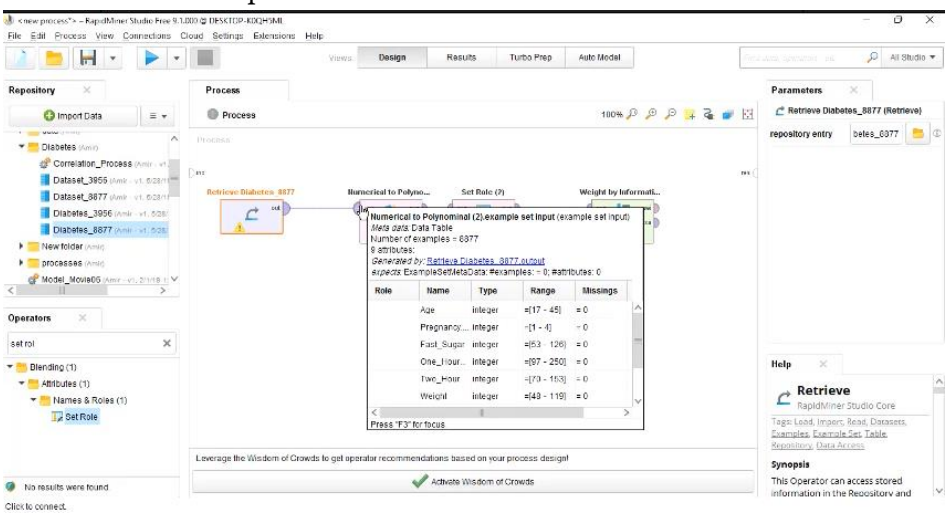


Figure 5- Adding Diabetes data

According to the figure above, our data is numerical, so we forcibly use a Set Role operator, which is known as Lable. Then we get the output as shown in the figure below, and according to the output, it is determined that the same as the previous outputs of the first column and Second, the data is not suitable, it should be deleted and the data should be separated, also the last three columns are balanced.

Figure 6 - Data output

And finally, we use the diabetes data and the Numeric to polynominal operator to convert numerical data to nominal data by adding appropriate operators such as the Weight operator, and draw the model, and run it to Manual:

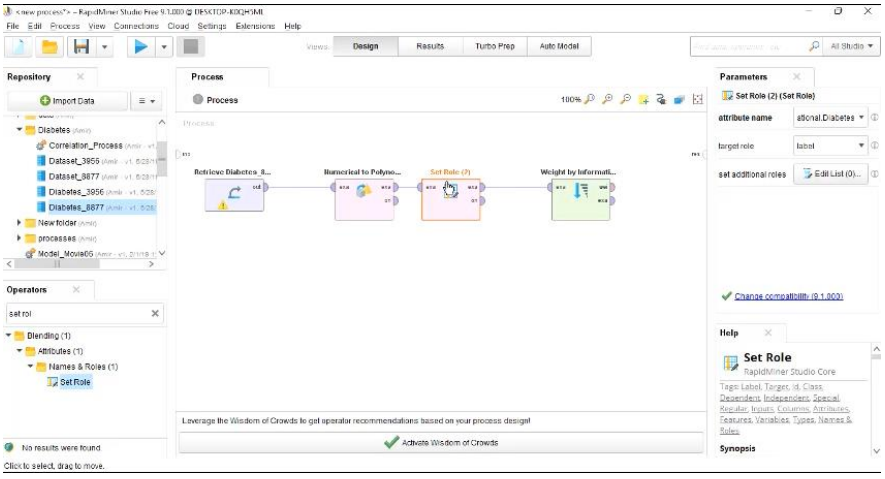


Figure 7 - Adding the Weight operator from diabetes data and the Numeric to polynominal operator

After implementing the Correlation and Information Gay model obtained in the R software, we are now running our main model according to the following figure. First, we enter the diabetes

data into the perspective, then we convert our data into nominal data using the Numeric to polynominal operator. Now, we want to predict the target column. We select the Set Role operator, which determines this column is the target and Lable column for us. Then, we use the Cross validation operator to build our prediction model, and manually connect the operators as shown below, and save the model:

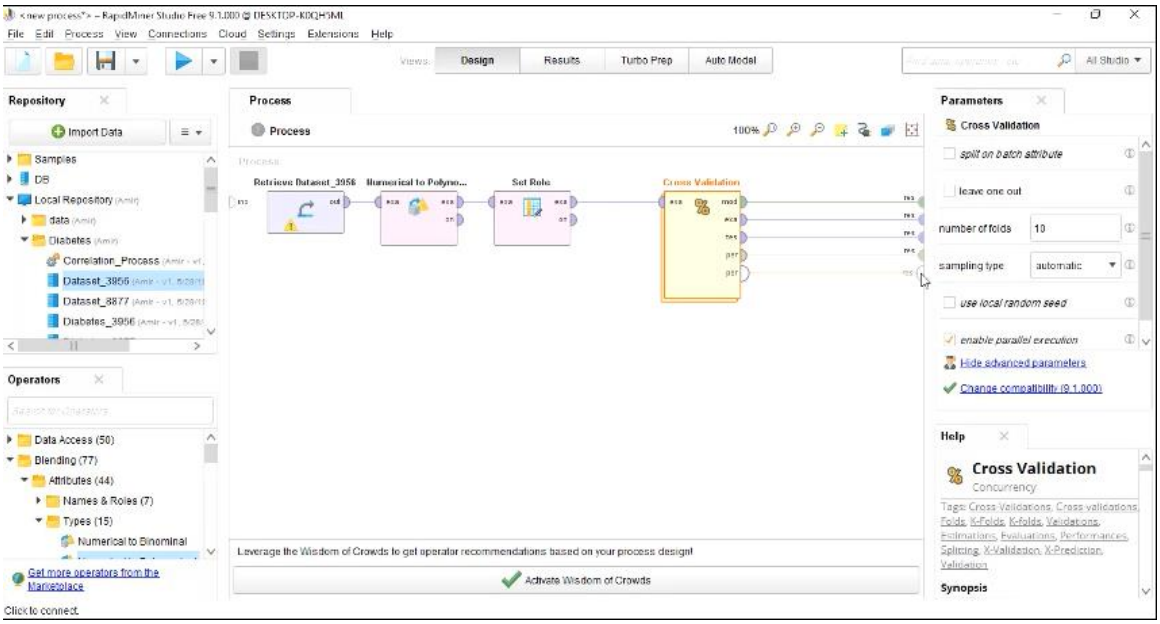


Figure 8- Adding the Cross validation operator

The following page opens by right-clicking on the Cross validation operator. In the following page, it is shown that ID3 is the same as J48, both of which work with Gay Information, but the ID3 model is used because we wrote J48 in Java, and it does not available in Rapidminer. According to the model below which we implemented, then we tested it, and according to the Performance operator, the two available values and the predicted values indicate the percentage of accuracy:

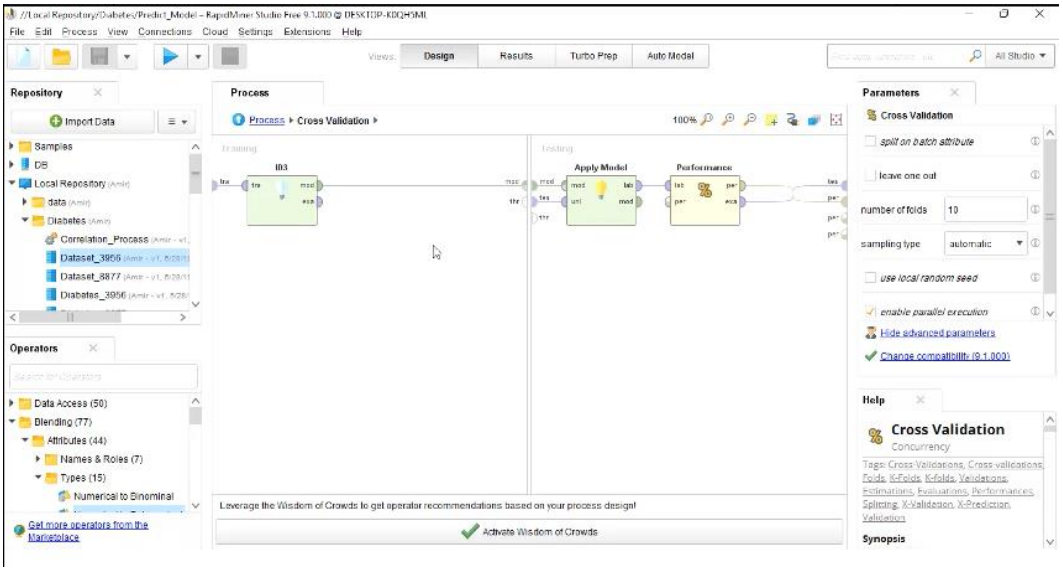


Figure 9- The model predicted of the accuracy percentage

Then, we ran the model and the following outputs were obtained. The first output of the tree model is:

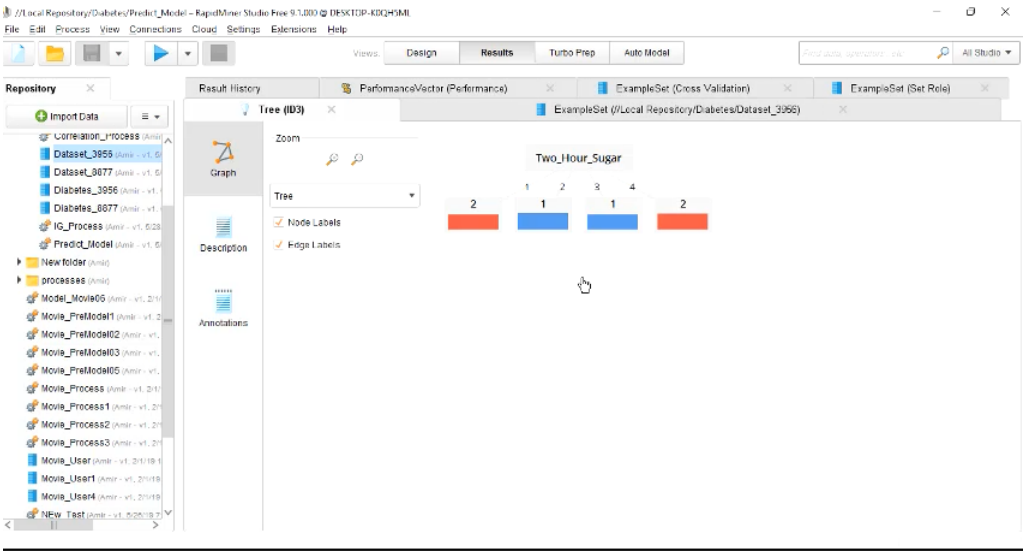


Figure 10: The tree model output: accuracy %

Also, according to the performance result shown in the figure below, 100% accuracy has not been achieved and a wrong number has been predicted, and in this way, the number of correct and incorrect predictions can be checked:

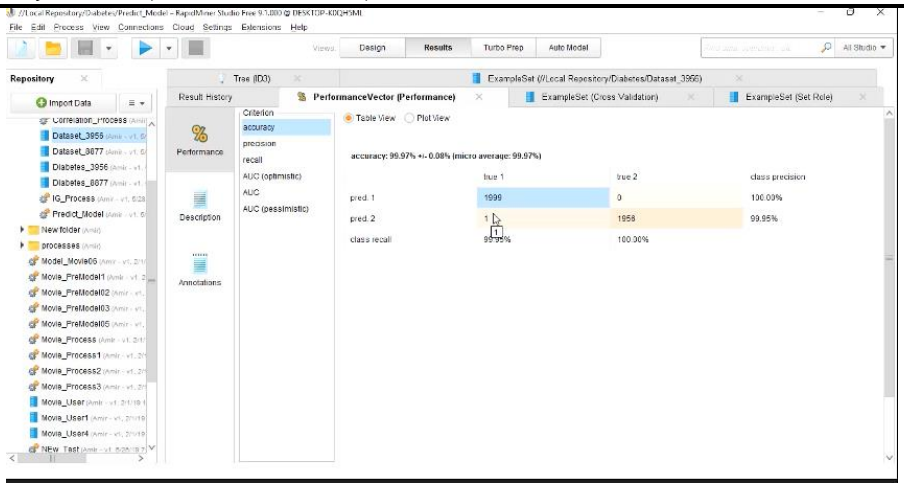
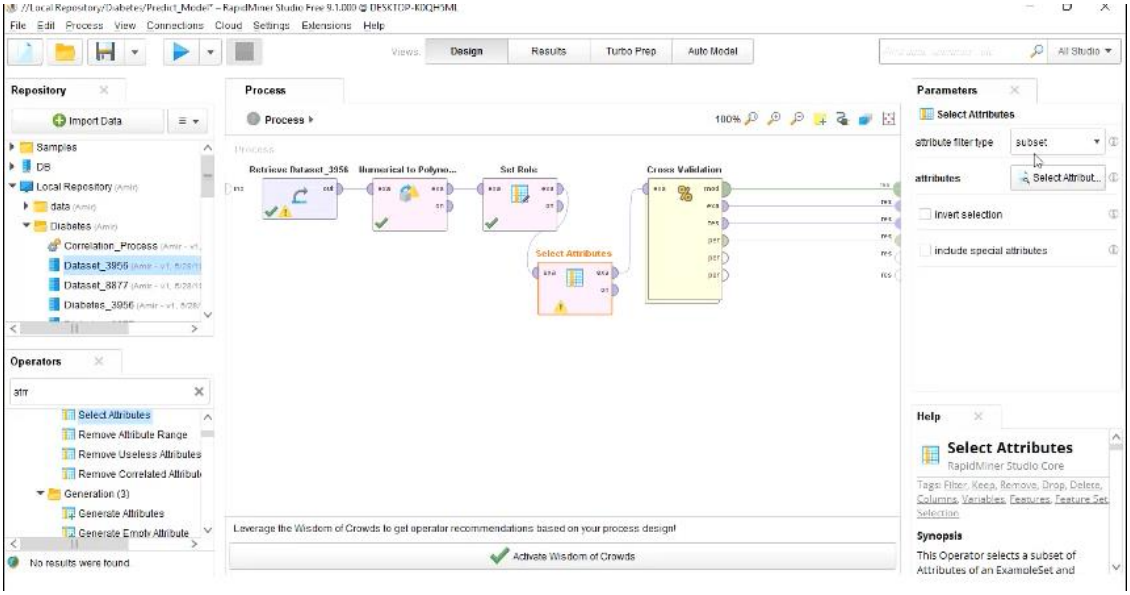


Figure 11: Performance result

Then we return to Prosses, and as shown below, we delete the two-hour glucose column after the Set Role operator, and add the Select Attributes operator to the model:



Then; according to the following figure, the output of the Cross validation section shows the real values and the predicted values:

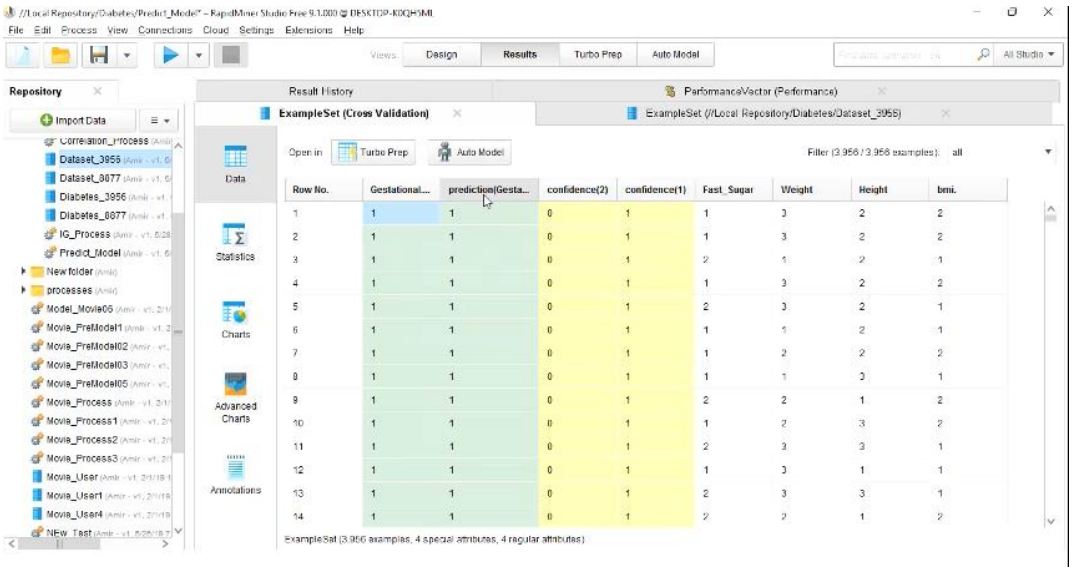


Figure 13: Actual values and predicted values

According to the figure below, you can use the Naive Bayes operator instead of the ID3 operator to run in this software the models that we ran in the R software:

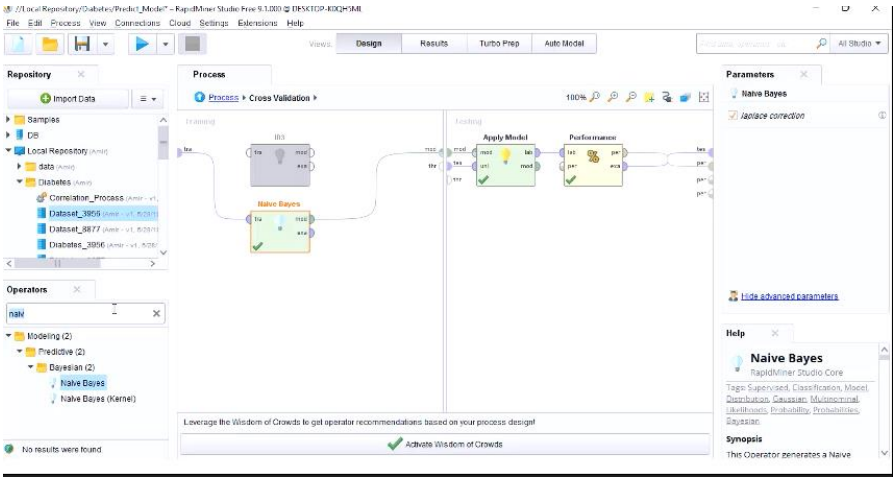


Figure 14: Using the Naive Bayes operator and building a new predicted accuracy percentage model

Finally, the general data model is executed and implemented according to the figure below and shows the percentage of accuracy according to the Performance operator and the values that exist and the values that are predicted and correctly displays the values that are correctly predicted. And the values that he predicted wrongly:

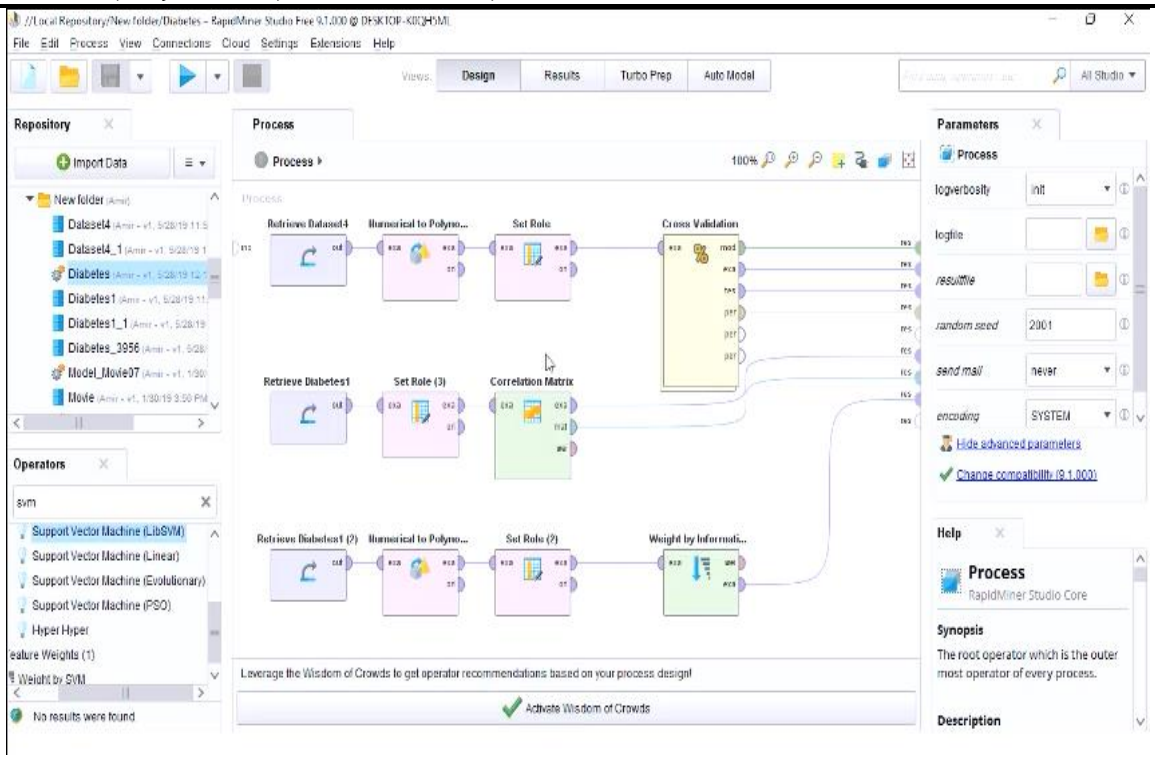


Figure 15: General data model of pregnant women's diabetes

Conclusion

In this study, the application of data mining in pregnant women's diabetes was investigated due to the importance and sensitivity of data mining in medicine, as well as the urgent need of this industry to move from traditional medicine to evidence-based medicine. The content presented is not a guarantee that the simple Bayesian network is completely useful in predicting diseases, whether an algorithm is useful or not depends on the motivation of health care providers in using such algorithms. On the other hand, proper training, motivation, and ease of use are very effective in making these algorithms useful for the user. On the other hand, users should be taught that the use of these algorithms is only to confirm or reject the decision made by them and has a supporting role.

Also, this study has investigated whether different data mining algorithms are suitable for predicting gestational diabetes and investigating the limitations and obstacles of using data analysis methods in the field of health using the aforementioned algorithms and data mining tools. The results of the analysis of the available parameters in the analysis of the data of pregnant women showed that the column of age and the number of pregnancies are very close to 0, that is, they are not suitable parameters for checking diabetes in pregnant women and should be

removed. Also, the best column of the data table is fasting blood sugar level[5], 1-h glucose and 2-h blood glucose, which diagnosed diabetes in pregnant women based on their values.
The results of the investigations show that the data mining software has correctly predicted 95.99% of the data, which indicates that it is a suitable method for predicting diabetes in pregnant women. According to the results obtained on the set of diabetic patients, among the tree classification methods, the decision tree obtains a better result, and, nearest neighbor algorithm also has a higher accuracy than the classification method, which can be used as a support method in addition to Doctors' decisions should be made to improve the accuracy of predicting diabetes in pregnant women. But the results of other articles in the field of diabetes diagnosis show that an algorithm can never be introduced as an optimal algorithm. As a result, an algorithm can be introduced as the optimal algorithm for each application according to the data set used.

Acknowledgment: None

Conflict of Interest: None

Funding: None

Ethical statements None

References

1. Mirsharif and Rouhani, data mining based on neural network and decision tree methods in early diagnosis of gestational diabetes risk, Journal of Health and Biomedical Informatics Medical Informatics Research Center, 139
2. Iyer et al, diagnosis of diabetes using classification mining techniques, International Journal of Data Mining
3. Nagarajan et al, Data Mining Techniques for Performance Evaluation of Diagnosis in Gestational Diabetes, ijsr 2014
4. C. M. Velu and K. R. Kashwan, “Visual Data Mining Techniques for Classification of Diabetic Patients”, 3rd IEEE International Advance Computing Conference (IACC), 2013.
5. Mirsharif M, Alborzi M. A fuzzy expert system & neuro-fuzzy system using soft computing for gestational diabetes mellitus diagnosis. International Journal of Information, Security and Systems Management 2014;3(1):249-52.
6. Tapak et al, Real-Data Comparison of Data Mining Methods in Prediction of Diabetes in Iran, HIR, 2013.
7. Poorolajal J, Zamani R, Mir-Moeini R, Amiri B, Majzoobi M, Erfani H, et al. Five-year evaluation of chronic diseases in Hamadan, Iran: 2005-2009. Iran J Public Health 2012;41(3):71-81

[5] This measures your blood sugar after an overnight fast (not eating). A fasting blood sugar level of 99 mg/dL or lower is normal, 100 to 125 mg/dL indicates you have prediabetes, and 126 mg/dL or higher indicates you have diabetes.

8. Kim S, Kim W, Park RW. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthc Inform Res 2011;17(4):232-43.

9. Sapra RL, Mehrotra S, Nundy S. Artificial neural networks: prediction of mortality/survival in gastroenterology. Current Medicine Research and Practice 2015;5(3):119-29.