DEVELOPING A PREDICTED MODEL FOR DIABETES TYPE 2 TREATMENT PLANS BY USING DATA MINING

TARIG MOHAMED AHMED
Assoc. Prof., Department of MIS, Prince Sattam Bin Abdalaziz University, KSA
Assoc. Prof., Department of C, omputer Sciences , University of Khartoum, Sudan
E-mail: Tarig_Harbi71@hotmail.com

ABSTRACT

Diabetes is a chronic disease in which blood sugar levels are too high. It can lead to very dangerous complication such as: heart disease and stroke, and even to complications requiring amputation of one of the parties. In this paper a new model was developed to classify diabetic type 2 treatment plans such as Insulin, medication and diet. These treatment plans could help the diabetic to control the blood glucose level. The model dataset was collected from JABER ABN ABU ALIZ clinic center for diabetes in Sudan. It was about 318 medical records. After conducting comprehensive experiments among data mining algorithms, J48 algorithm was selected to develop the proposed model based on accuracy results. By using WEKA application, the model was implemented. According to the results, the accuracy of the model was 70.8% and all others measures of validation were accepted. Also, the research mentioned that the duration of illness and the age factors were played very important role in specifying an appropriate treatment plans.

Keywords: *Diabetes type 2; data mining; classification; J48 algorithm*

1. INTRODUCTION

Diabetes is a chronic disease in which blood sugar levels are too high. Sugar from the foods we eat comes, and insulin is the hormone that helps the entry of sugar into the cells in order to give the energy[1]. In diabetes type I the body does not secrete insulin, and Type II diabetes, which is the type most prevalent, the body does not make insulin and cannot use them properly; it is an insufficient amount of insulin and sugar stays in the blood is present[2]. Over time can result in the presence of too much sugar in the blood to the emergence of serious problems as it can damage the eyes , kidneys and nerves .Diabetes can lead to heart disease and stroke, and even to complications requiring amputation of one of the parties[3].To avoid all these serious complications, the blood glucose level should be under control. One of the strong indictors that measures the diabetic, is the accumulative blood glucose level for previous 3 months (HPA1c). HPA1c should be less than 7 to consider the diabetic under control[6].

Data Mining is the process of integrating traditional methods for analyzing data with complex algorithms to extract accurate information, useful from among the huge amount of data is used, you may use those data later in the expectation an event in the future[4]. Classification in data mining is to organize data in certain categories. Also it is known as classified under the supervision and classification of the uses given row labels to arrange the objects in the collection of data. Classification use the training set, where all objects are already associated with known class labels. The classification algorithm is to learn from the training set and it builds a model. The model is used to classify new objects [5].

In this research a new classification model was developed. The new model aimed to classify diabetic patient type 2 treatment plans. The treatment plans had HPA1c less than 7. Three categories for treatment plans were selected such as Insulin , Medications and diet. The dataset of the model was collected from JABER ABN ABU ALIZ clinic center for diabetes in Sudan. The size was 318 medical records with 9 attributes. All medical records were related to under control diabetic type 2. In addition, a comprehensive experiments were conducted to select the classification techniques for the model. As result , the algorithm J48 was selected as best one with accuracy rate 70.8 % and ROC (Receiver operating characteristic) rate 0.624. These results were accepted to use the model.

As, application, WEKA was used to implement the model and It has a wide range of data maiming algorithms and this application was developed for research purpose.

This paper has been organized as sections which explain in details the phases of building the proposed model. In section No. 2, many previous researches have been reported as related works. Section No. 3 described in comprehensive way the proposed model, the finding and results discussion . Finally, section 4. concluded this work with specific recommendations as future works.

2. RELATED WORK

Many researches have been conducted in this area. In this section, many papers have been reviewed and summarized such as follows:

Karahoca, Adem, and M. Alper Tunga used multivariate data partitioning method which is called Indexing HDMR to manage drug dosage. As dataset of this study, 142 diabetic of type 2 medical records were used to implement the model. As the result of this study a polynomial structures was obtained for the dosage planning by using Indexing HDMR method with high performance[7]. Liu, Haifeng, et al. proposed a model by using data mining techniques. The model was used to help physicians to control the glucose level in diabetes type 2 only. In the first, all factors affect the treatment plan were identified. The model performance was validated by using HPA1c value[8]. Habibi et al. proposed a model to diagnosis diabetes type 2 by using decision tree (J48). To generate the model, 22,398 medical records were used. The precision of the model was 0.717. The age factor was found as very important factor in the classification tree. The ROC curve was indicated that the model has high quality [9]. Toussi, Massoud, et al. analyzed the type 2 management for French national guidelines. By using C5.0 algorithm, the physicians' prescriptions were obtained. About 463 medical records were used to developed the model. The model was consisted of 72 rules and 12 of them are related to the treatment types. As result, the model was useful a tool for the physicians in taking their decision in treatment plans [10]. Ramezankhani, Azra, et al. developed a prediction model to identify low factors for the type 2 diabetes incidence. About 6647 records and decision tree method were used to generate the model. The result mentioned that the accuracy was 90.5% and 97.9% specificity[11]. ALjumah et al proposed a classification model by using support vector machine algorithm in the Oracle Data Miner (ODM). The model was aimed to treating diabetes. From World Health Organization (WHO), the datasets were collected to generate the model. All datasets were related to Saudi diabetic patients. The model generated several treatment types for two age groups (Old and young) [12]. Patil et al. developed a hybrid classification model for identifying type 2 diabetic patients. The model was used several data analysis methods. Also, it investigated the characteristics and measures that related to diabetes. The class labels were selected by Simple K-means clustering and the classifier was built by C4.5 . the model dataset was collected from Pima Indians diabetes - university of California .After testing the model, 92.38% accuracy was obtained[13]. Sanakal and Jayakumari implemented a model for diagnosing diabetes by using 9 attributes for diagnosing. The model was developed by using Fuzzy C-means clustering (FCM) algorism. about 768 cases were used to build the model. After testing the model, FCM obtained 94.3% accuracy [14]. Vasudevan implemented a statistical analysis on real dataset from National Institute of Diabetes and Digestive and Kidney Diseases. By using logistic regression method , the analysis was performed. SPSS 7.5 tool was used. The analysis obtained significant factors. Iterative Dichotomiser-3 algorithm was used to build the prediction model. The model detected the diabetes disorder risks as result[15]. Liu, Haifeng, et al. recognized the factors that to control blood glucose level. They implemented experiments were based on HbA1C value. The results were validated and compared with a clinical guideline. The model of the study was used for new diabetic patients[16].

3. MATERIAL AND METHOD

3.1 Data Collection

The dataset of this research was collected from JABER ABN ABU ALIZ clinic center for diabetes in Sudan. It consists of 11694 records and 115 attributes were used to describe comprehensive information about diabetes type 2. In addition, the dataset represents the demographic information, the complication diseases, basic control and others.

3.2 Data Preprocessing

This research aimed to classify the diabetic patients type 2 into controlled treatment plan. So, some preprocessing tasks were implemented to meet the research goal. For example, the records related to diabetes type 2 with value of HPA1C less than 7 were selected. Also, all non-relevant attributes were eliminated. After completing the preprocessing tasks, the new dataset consists of 318 records and 9 attributes. Table 1 describes the final attributes of dataset.

Table 1. List of Attributes Descriptions

Feature name	Type
Gender	Nominal
Age	Nominal
Smoking	Nominal
History of hypertension	Nominal
Renal problem	Nominal
Cardiac problem	Nominal
Eye problem	Nominal
Duration of Diabetes	Nominal
Basic control (Class Labels Attribute)	Nominal

The following tables (Table 2 to Table 5) present the frequency of some of important attributes which had effected the model.

TABLE 2. AGE INFORMATION

Items	Frequency	Percent
20 - 30 years	2	.6
31 - 40 years	13	4.1
41 - 50 years	45	14.2
51 - 60 years	97	30.5
61 - 70 years	104	32.7
71 - 80 years	47	14.8
> 80 years	10	3.1
Total	318	100.0

Table 3: Gender Information

Items	Frequency	Percent
Male	221	69.5
Female	97	30.5
Total	318	100.0

Table 4: Duration Information

Items	Frequency	Percent
Newly discovered	19	6.0
< 5 years	55	17.3
5 - 10 years	91	28.6
11 - 15 years	49	15.4
16 - 20 years	46	14.5
21 - 25 years	28	8.8
26 - 30 years	18	5.7
> 30 years	12	3.8
Total	318	100.0

Table 5: Basic Control Information

Items	Frequency	Percent
Oral Hypoglycemic	188	59.1
Insulin	113	35.5
Diet	17	5.3
Total	318	100.0

3.3 Tools and techniques

After comprehensive survey among data mining techniques, one of them was selected to develop the model. The final model was selected based on best evaluation criteria. The following sections present three experiments that were used to find out the best model for classifying good treatment plan for diabetic type 2 based on control level.

.

3.4 WEKA Application

WEKA application tool is a data maiming application tool which is built based on Java Language. There is more than a dozen years since the first public version of WEKA. At that time, the program was rewritten from scratch, and has evolved to a great extent, and now accompany text on the data extraction . These days, WEKA and enjoys wide acceptance in both the academic world and the business sector. It has an active community. Many algorithms have been implemented in WEKA [17].

3.5 Data Mining Techniques

To select an appropriate data mining algorithm, an intensive experiments have been done. Based on best results, the J48 algorithm was selected to be use in the proposed model. The J48 algorithm used the basic control attribute to classify the diabetic type 2 to one of the controlled treatments which were Insulin, Medication and Diet.

J48 algorithm is classification implementation of C4.5 algorithm which was developed by Ross Quinlan. Many researchers in health sectors used C4.5 algorithm. C4,5 is built based on a decision tree in two phases, phase one, building the classifier by using a training dataset. Phase two, test the model by using a testing dataset to measure the accuracy[18].

3.6 Model Validation

To validate the model, there are many factors could be used in confusion matrix. the confusion matrix, which demonstrate the correctly and incorrectly classification classes, the correct classifications denoted by (TP) True Positive and True Negative (TN). The incorrect classified occur when a false positive (FP), that is when an outcome is predicted as positive while it is actually negative, moreover, the same happened when a false negative (FN) if an outcome is classified as negative but it is actually positive [19].

Accuracy = $TP/(TP + FN)$

It is significant to use F-Measure, it produces a high result when Precision and Recall are both balanced, although Precision and Recall are valid measures, but the F-Measure was used because, Precision and Recall can be optimized at the overhead of the other.

Precision = $TP/(FN + TP)$

Recall = $TP/(FP + TP)$

F-Measure =

$$2 * \text{Recall} \times \text{Precision} /(\text{Recall} + \text{Precision})$$

It measures how good the classifier in recognizing instances of different classes is, it is a percentage of correctly classified instances in a test dataset [20].

3.7 Proposed Model

After using u the dataset and J48 algorithm, the model was generated . figure 1 describes the components of the model and figure 2 shows the model results after implementing by using WEKA application.

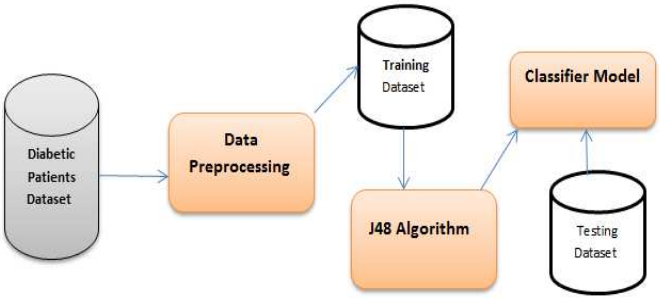
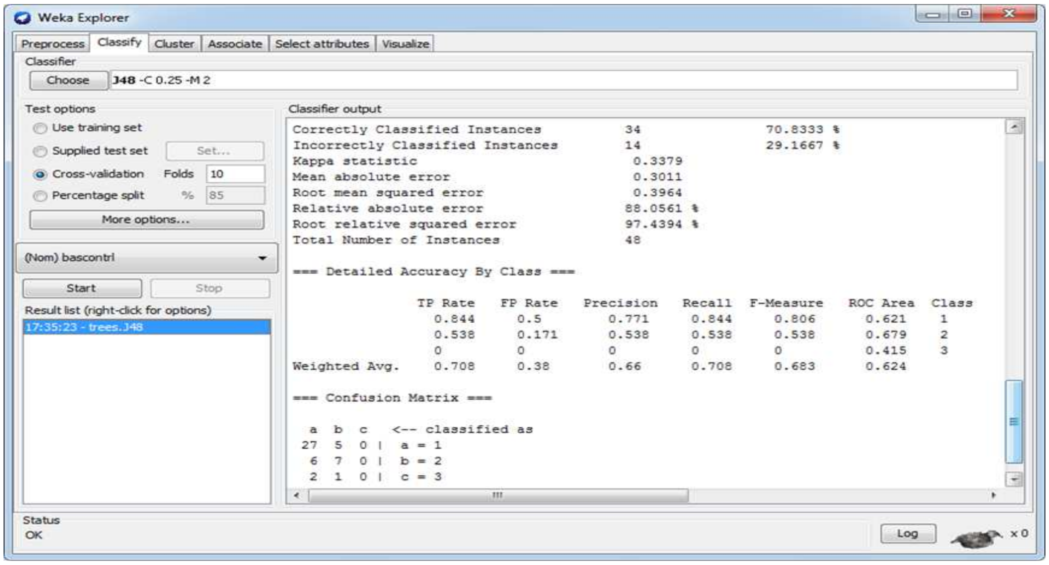


Figure 1: Model Components

3.8 Result and Discussion

As mentioned in figure 2 , the accuracy of the model was 70.8% . So, the model can classify the diabetic patient type 2 treatment plans that can help them to control the glucose blood level within 70.8% accuracy . Also , the Precision = 0.66, recall =0 .708 and F-measure = 0.683 results were good and acceptable. In addition , the ROC = 0.624 and that means the model could be used to classify new patients treatment plans. Moreover, the model revealed that the attributes Duration and Age were played very important role to specify the treatment plans for diabetic patient type 2.



Classifier
J48 -C 0.25 -M 2

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 85

(Nom) bascontrl

Result list (right-click for options)
17:35:23 - trees.J48

Classifier output

```
Correctly Classified Instances          34               70.8333 %
Incorrectly Classified Instances        14               29.1667 %
Kappa statistic                          0.3379
Mean absolute error                      0.3011
Root mean squared error                  0.3964
Relative absolute error                 88.0561 %
Root relative squared error             97.4394 %
Total Number of Instances               48

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.844     0.5        0.771     0.844     0.806      0.621    1
                 0.538     0.171      0.538     0.538     0.538      0.679    2
                 0         0          0         0         0          0.415    3
Weighted Avg.    0.708     0.38       0.66      0.708     0.683      0.624

=== Confusion Matrix ===

  a  b  c   <-- classified as
 27  5  0 |  a = 1
  6  7  0 |  b = 2
  2  1  0 |  c = 3
```

Status
OK

Figure 2: Model implementation

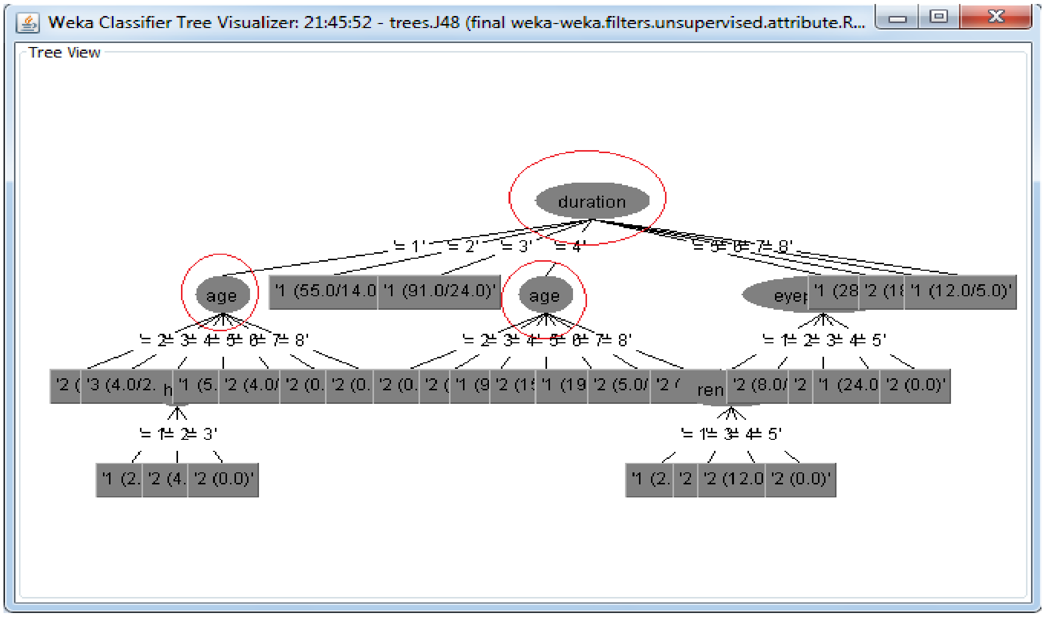


Figure 3: Model Visualization Tree

Figure 3 mentions that the duration of the illness and the age of the patient are very important factors to determine a good treatment plan.

4. CONCLUSION AND FUTURE WORK

Diabetes is a chronic disease in which blood sugar levels are too high. Sugar from the foods we eat comes, and insulin is the hormone that helps the entry of sugar into the cells in order to give the energy. In this research a new classification model was proposed. The model aimed to classify diabetic patient type 2 treatment plans. The treatment plans had HPA1c less than 7 so, they were considered under control plans. Three categories for treatment plans such as Insulin , Medicine and diet had been selected as lassification labels. The dataset of the model was collected from JABER ABN ABU ALIZ clinic center for diabetes in Sudan. The dataset size was 318 medical records with attributes. The model was implemented using J48 algorithm. The result mentioned that the accuracy was 70.8% and all measurement results were accepted to use the model. Also, the research mentioned that the duration of illness and the age factors were played very important role in specifying an appropriate treatment plans.

As future work, the model could be extendable to include diabetic patients type 1 treatment plans by adding more attributes. Also, to increase the accuracy of the model some additional features need to be added such as nutrition system and exercise.

REFRENCES:

[1] Willett, Walter. Eat, drink, and be healthy: the Harvard Medical School guide to healthy eating. Simon and Schuster, 2011.

[2] Musselman, Dominique L., et al. "Relationship of depression to diabetes types 1 and 2: epidemiology, biology, and treatment." Biological psychiatry 54.3 (2003): 317-329.

[3] Gabbay, Robert A., et al. "Nurse case management improves blood pressure, emotional distress and diabetes complication screening." Diabetes research and clinical practice 71.1 (2006): 28-35.

[4] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17.3 (1996): 37.

[5] Li, Tsai-Chung, et al. "Glycemic control paradox: Poor glycemic control associated with higher one-year and eight-year risks of all-cause hospitalization but lower one-year risk of hypoglycemia in patients with type 2 diabetes ."Metabolism 64.9 (2015): 1013-1021.

[6] Kaur, Harleen, and Siri Krishan Wasan. "Empirical study on applications of data mining techniques in healthcare." Journal of Computer Science 2.2 (2006): 194-200.

[7] Karahoca, Adem, and M. Alper Tunga. "Dosage planning for type 2 diabetes mellitus patients using Indexing HDMR." Expert Systems with Applications39.8 (2012): 7207-7215.

[8] Liu, Haifeng, et al. "An efficacy driven approach for medication recommendation in type 2 diabetes treatment using data mining techniques. "Studies in health technology and informatics 192 (2012): 1071-1071.

[9] Habibi, Shafi, Maryam Ahmadi, and Somayeh Alizadeh. "Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining." Global journal of health science 7.5 (2015): 304.

[10] Toussi, Massoud, et al. "Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes." BMC medical informatics and decision making 9.1 (2009): 1.

[11] Ramezankhani, Azra, et al. "Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study." Diabetes research and clinical practice 105.3 (2014): 391-398.

[12] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, Application of data mining: Diabetes health care in young and old patients, Journal of King Saud University - Computer and Information Sciences, Volume 25, Issue 2, July 2013, Pages 127-136

[13] Patil, Bankat M., Ramesh Chandra Joshi, and Durga Toshniwal. "Hybrid prediction model for Type-2 diabetic patients." Expert systems with applications 37.12 (2010): 8102-8108.

[14] Sanakal, Ravi, and Smt T. Jayakumari. "Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine."International Journal of Computer Trends and Technology 11.2 (2014): 94-8.

[15] Vasudevan, P. "ITERATIVE DICHOTOMISER-3 ALGORITHM IN DATA MINING APPLIED TO DIABETES DATABASE." Journal of Computer Science10.7 (2014): 1151.

[16] Liu, Haifeng, et al. "An efficacy driven approach for medication recommendation in type 2 diabetes treatment using data mining techniques."Studies in health technology and informatics 192 (2012): 1071-1071.

[17] Hall, Mark, et al. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11.1 (2009): 10-18.

[18] Bruno Fernandes Chimieski, Rubem Dutra Ribeiro Fagundes, Association and Classification Data Mining Algorithms Comparison over Medical Datasets, 2013

[19] Duarte de Araujo, Flavio Henrique, Andre Macedo Santana, and Pedro de Alcantara dos Santos Neto. "Evaluation of Classifiers Based on Decision Tree for Learning Medical Claim Process." Latin America Transactions, IEEE (Revista IEEE America Latina) 13.1 (2015): 299-306.

[20] Vieira JAP. Algorithm development for physiological signals analysis and cardiovascular disease diagnosis - a data mining approach [thesis]. Coimbra: University of Coimbra; 2011.