J. Of University Of Anbar For Pure Science : Vol.12 : No.3: 2018

ISSN: 1991-8941

Diabetes Classification Using ID3 and Naïve Bayes Algorithms

Khalid Shaker Jassim ، Hadeel M.Saleh
College of computer Science and IT, University of Anbar, Ramadi, Anbar, Iraq

Received:15/6/2017 / Accepted:21/11/2018

ABSTRACT: Diabetes can be defined as a chronic disease identified by high levels of blood glucose that result from issues in the way insulin is generated, the way insulin works, or both those reasons. The aim of this research is to propose a technique using the Decision Tree (ID3) and Naive Bayes to categorize diabetes and reduce classification errors by increasing the accuracy of the classification. The results of the proposed method were evaluated by comparing them with other results through the application of the proposed system to Pima India Diabetes data set, obtained from the UCI database site. The experimental results show that the ID3 recorded a precision ratio of 91% and the naive class corrected it to 94% for the same number of the test group.

1. INTRODUCTION

Medical data is an important factor to aid diagnosis and treatment; moreover, they might be advantageous in the area of education for health care researchers .Interpreting and analyzing medical data are significant and interesting components of the classification [1].

There is a great amount of medical data, which might be useful in healthcare[2].A wide range of data mining approaches are implemented by various scientists for diagnosing and treating various diseases like the diabetes, stroke, heart disease and cancer[3].

Data mining is procedure of deducing data from huge database. In medical system great amount of data are current. Data mining on medical records has drawn a great amount of attention, however ,it's still at an early phase of practice [4]. One of the most important approaches of data mining are classification and prediction of medicinal data which they play a great part in the extraction of knowledge from the already existing data-base. This type of data analyzing might be used for extracting models that describe significant data classes or for predicting upcoming trends of data [5].

The approaches of classifying that have been suggested during the past ten years include, ANN, genetic algorithms, Bayesian, decision tree and other approaches [6].

2. LITERATURE REVIEW

There are several methods of developing in different areas for various reasons. In the following, a set of proposals have been derived from the scientific literature:

In 2011, Selvakuberan K and Harini B [7],highlighted the efficiency of data mining approach in the area of Health Care applications. Diabetes is one of the basic reasons lead to early

illness and death all over the world. This study discussed the applications of data mining approaches in detecting diabetes in PIMA Indian Diabetes Dataset (PIDD). It has proposed a Feature Selection method that used a combination of Ranker Search approach. The precision rate of the classification has reached 81% and this method proved to be better compared to previous methods.

In 2013 Velu.C.M and Kash wan.K.R [8], a paper was introduced, wherethe researchers have used three approaches that are EM algorithm, H-means and clustering and Genetic Algorithm (GA), in order to classify the diabetes patients. The performance for H-means and showed better results than the rest when all the identical symptoms were collected into clusters with the use of those approaches.

In 2014 Saba Bashir, Usman Qamar, Farhan Hassan Khan [9],a work has been suggested; used several groups of classification approaches for data-sets of diabetes. Three kinds of decision trees ID3, C4.5 and CART are implemented in the form of base classifiers. The ensemble techniques utilized are Majority Voting, Bagging, Bayesian Boosting, Adaboost and Stacking. Two benchmark datasets have been used from UCI and BioStat repositories.

The results of experiments and the evaluation indicated that Bagging ensemble approach showed more efficient implementation when compared with single or other ensemble approaches.

In 2014 Aishwarya and Anto [10],proposed a model has been proposed, which was used to diagnose diabetes, the dataset that was used in that work is Pima India diabetes dataset of UCI machine learning repository. The approaches that have been utilized in this work are Genetic algorithm (GA) for the reasons of feature selection and Extreme Learning machine for classification purposes.

The implementation of the suggested system has undergone analysis according to several criteria, like the accuracy of classification, sensitivity and specificity using 10-fold cross-validation and matrix of confusion. The precision of the suggested system has reached the percentage of 89.54% for genetic algorithm.

In 2014, Ravi s, Smt.T [11],a study using data mining techniques for analyzing the data-bank of Diabetes disease and diagnosing that disease. This study also included implementing Fuzzy C-Means and Support Vector Machine and evaluating it on a group of medical data concerned with diabetes problem of diagnosis. The best result is by Fuzzy C-Means reached a precision of 94.3% and positive predictive value which is 88.57%. Support Vector Machine has a precision of 59.5% that is rather low.

In 2015 M.Dura raj [12],the highlighted a fact that NNs are one of the soft computing approaches that can be utilized for making predictions concerning medical data. NNs are known as the Universal predictors.

Diabetes mellitus or diabetes is an illness that occurs because of the increased amount of blood glucose .Different conventional approaches, depending on physical and chemical tests, are available to diagnose diabetes. ANNs-based systems are capable of being effectively implemented for the prediction of high blood pressure risk. This developed system splits the data set into one of the2subsets. The earlier detection with the use of soft computing approaches has helped the doctors in reducing the possibility of getting serious the disease. The data set selected to classify and test the simulation depends on Pima Indian Diabetic Set from (UCI) Repository of Machine Learning databases. In this study, a detailed survey was held on the application of various soft computing approaches for predicting diabetes. This study has aimed to recognize and suggest an efficient approach to earlier predict the disease.

The contributions of this research is suggested a new idea used to reducing the medical classification problems ,and increase the accuracy by using the preprocessing methods and classification techniques .

3. AN OVERVIEW OF THE PROPOSED SYSTEM

The structure of the Diabetes Classification System consist of two stages the first stage has specific function are read Diabetes Dataset is Pima India diabetes data set ,feature selection stage ,Discretization stage ,and the classifier model used to generate the decision Rules .to apply the classification process .the second stage is testing stage consist of two specific functions are read data set ,discretization ,decision rules and out put .Figure (1) describes the structures of the training and testing phases of diabetes classification system respectively.

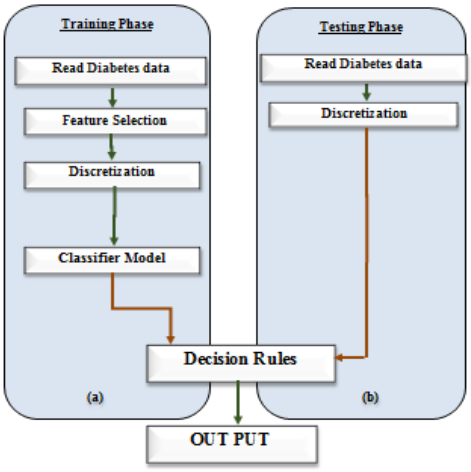

Figure (1):Phase structure of the phase Diabetes Classification System: a) the main block diagram of training phase of Diabetes Classification System. b) The main block diagram of testing phase of Diabetes Classification System.

3.1 DATA SET

The Pima Indian Diabetes Data set (PIDD) implemented in this research was taken from the UCI machine learning respiratory related to the Diabetes research. Every patient in this data base is a Pima Indian woman at least 21 years old and living near Arizona, the United States. Every sample of the dataset has eight properties (attributes). Those properties as depicted in table (1).

Table1: Number of Feature of PIDD.

Feature	Description	Range
1	Number of times pregnant	1-4
2	Plasma glucose concentration a 2 h in an oral glucose tolerance test.	120-140
3	Diastolic blood pressure (mm Hg).	80-90
4	Triceps skin fold thickness (mm).	12 mm(male)-23(female)
5	2-h serum insulin (lU/ml).	16-166 mlu/L
6	Body mass index (weight in kg/(height in m)^2)	18-24.9 kg/m2
7	Diabetes pedigree function Feature 8: 2-h serum insulin (lU/ml).	1-3
8	Age (years).	21 and above

It is made up of two classes and 768 samples. The class distribution is:

- Class 1: normal (500) samples.
- Class 2: abnormal (268) samples.

Fold creation To apply cross validation technique, the entire labeled dataset is divided in to mutually exclusive folds. The 768 instances in this thesis were selected from the PIDD database, which means that:

- 499 are utilized for the training-set, 325 Instances are normal, 174 are abnormal.(i.e.65%)
- 269 are utilized for testing, 175 Instances are normal, 94 are abnormal. (i.e.35%)
- As depicted in table (2)

Table (2) Number of cases tested and trained

	No. Of cases	Normal	Abnormal
Training	499	325	174
Testing	269	175	94

3.2 FEATURE SELECTION STAGE:

Feature Selection is the most significant step in the conception of a Diabetes Classification System due to the fact that the effectiveness relies on the correctness of the selection.

Every sample of the PIDD data-set has eight properties. Sub-set property chosen from eight properties according to the entropy of property with class. In case of choosing more than five properties so more time consuming. In the opposite case, when selecting less than five properties makes it less time consuming but less accurate Diabetes Classification System. The optimal five properties (1,3,4,5,7) according to the higher entropy are chosen for representing the Diabetes and will be utilized subsequently as an input to the part of classification.

3.3 DISCRETIZATION

Diabetes is not easily explained, and it's very important improving PIDD for the sake of making the classification stage easier and more efficient. In the Diabetes Classification System an important phase, prior to the classification procedure, has to be performed. The numerical values of eight properties it has to be converted to categorical values. This is performed via dividing the range of values of eight properties into k equal sized bins, or in other words equal intervals of width, where k is a parameter chosen by a user according to length of data. Algorithm (1) depicts Equal Width Interval Discretization (EWID).

Algorithm (1) : EWID

Input: numerical values of eight features.

Output: categorical values of eight features.

```
Begin
min1 = f1(0) : max1 = f1(0) : min2 = f2(0) : max2 = f2(0) : min3
= f3(0) : max3 = f3(0)
min4 = f4(0) : max4 = f4(0) : min5 = f5(0) : max5 = f5(0) : min6
= f6(0) : max6 =f6 (0)
min7 = f7(0) : max7 = f7(0) : min8 = f8(0) : max8 = f8(0)
For i = 1 To length of data set {
If f1(i) > max1 Then max1 = f1(i) : If f1(i) < min1 Then min1 =
f1(i)
If f2(i) > max2 Then max2 = f2(i) : If f2(i) < min2 Then min2 =
f2(i)
If f3(i) > max3 Then max3 = f3(i) : If f3(i) < min3 Then min3 =
f3(i)
If f4(i) > max4 Then max4 = f4(i) : If f4(i) < min4 Then min4 =
f4(i)
If f5(i) > max5 Then max5 = f5(i) : If f5(i) < min5 Then min5 =
f5(i)
If f6(i) > max6 Then max6 = f6(i) : If f6(i) < min6 Then min6 =
f6(i)
If f7(i) > max7 Then max7 = f7(i) : If f7(i) < min7 Then min7 =
f7(i)
If f8(i) > max8 Then max8 = f8(i) : If f8(i) < min8 Then min8 =
f8(i) }
K=3  //number of possibilities
w1 = Round(((max1 - min1) / k), 3) : w2 = Round(((max2 - min2)
/ k), 3)
w3 = Round(((max3 - min3) / k), 3) : w4 = Round(((max4 - min4)
/ k), 3)
w5 = Round(((max5 - min5) / k), 3) : w6 = Round(((max6 - min6)
/ k), 3)
w7 = Round(((max7 - min7) / k), 3) : w8 = Round(((max8 - min8)
/ k), 3)
// determine k ranges for 8 features
Lowf1 = (min1 + w1) : mediumf1 = (min1 + (2 * w1)) : highf1 =
(min1 + (3 * w1))
Lowf2 = (min2 + w2) : mediumf2 = (min2 + (2 * w2)) : highf2 =
(min2 + (3 * w2))
Lowf3 = (min3 + w3) : mediumf3 = (min3 + (2 * w3)) : highf3 =
(min3 + (3 * w3))
Lowf4 = (min4 + w4) : mediumf4 = (min4 + (2 * w4)) : highf4 =
(min4 + (3 * w4))
Lowf5 = (min5 + w5) : mediumf5 = (min5 + (2 * w5)) : highf5 =
(min5 + (3 * w5))
Lowf6 = (min6 + w6) : mediumf6 = (min6 + (2 * w6)) : highf6 =
(min6 + (3 * w6))
Lowf7 = (min7 + w7) : mediumf7 = (min7 + (2 * w7)) : highf7 =
(min7 + (3 * w7))
Lowf8 = (min8 + w8) : mediumf8 = (min8 + (2 * w8)) : highf8 =
(min8 + (3 * w8))
End
```

Figure (1) EWID algorithm pseudo code

3.4 CLASSIFIER MODEL

Constructing classifier model is the most widely known task classification. This structure utilized to predict the Diabetes class, the class might be classified as a normal or abnormal .The

database of Diabetes (Training-Set) is made up of attribute value representation with five categorical attributes (1, 3, 4, 5and 7) and class attribute for a large number of patients. Those attributes are the input to the classifier model for learning. Prediction of the new case is based on classifier model. The classifier model described in figure (2)

The classifier model in this study is constructed with the use of decision tree according to training Diabetes.

Figure (2) Classifier Model

3.5 DECISIONTREE AND DECISION RULES

The decision tree classification algorithm might be performed by the use of the algorithm of Iterative Dichotomiser 3(ID3)this algorithm as shown in figure(6) . This algorithm is capable of generating decision rules via a decision tree according to the attributes that take an important part in classifications that are based on entropy and information gain.Due to computing ID3 algorithm on the training set, creating the decision rules with the use of if- then format that is depicted subsequently. For the sake of simplifying a decision tree transforming it to decision rules that are simpler to be comprehended and implemented on a computer. In the stage of testing, the classifier model performs a classification of the tested data as normal, or abnormal according to decision rules that are constructed by ID3 algorithm. Each step in the testing stage is as in the training stage except the fact that the Diabetes is tested according to decision rules.

Algorithm (2) ID3
Input: Training set consist of categorical attributes
Output: Decision tree and Decision rules

Begin:
Create a Root node for the tree based on high info gain.
If all have the same value of the target attribute, Return the single node tree Root with label = the value of target attribute.
If Attributes is empty, Return the single node tree Root with label = most common value of Target attribute
Else
Let A = the attribute from Attributes that best classifies.
Let the decision attribute for Root = A.
ForEach possible value, vi, of A
Add a new tree branch below Root corresponding to the test A = vi.
Let Examples vi be the subset of Examples that have value vi for A.
If Examples vi is empty Then
Add a leaf node with label = most common value of Target attribute to new branch.
Else
Add a sub tree ID3(Examples vi , Target_ attribute ,Attributes– {A}) below this new branch.
End For
End
Return Root
End

Figure (3) ID3 algorithm

3.6 NAÏVE CLASSIFIER

The Naïve Classifier algorithm can be implemented as shown in Algorithm (3). Naïve Classifier has the ability to predict class membership probabilities of Diabetes as normal or abnormal such as the probability that a given sample belongs to a particular class. Through

Create Likelihood by finding the probabilities based on the Bayes theorem.

```
Algorithm( 3 ) naïve classifier
Input: Training set consist of categorical attributes
Output: decision
Begin
For i = 0 To 1  // two class
C = 0
For j = 0 To length of Training set
If fcn(i) = fc(j) Then C = C + 1
End for
pfc(i) = Math.Round((C / count_sw), 3)
end for
For m = 1 To 8
If m = 1 Then z = ns1 : If m = 2 Then z = ns2 : If m = 3
Then z = ns3
If m = 4 Then z = ns4 : If m = 5 Then z = ns5 : If m = 6
Then z = ns6
If m = 7 Then z = ns7 : If m = 8 Then z = ns8
For i = 0 To z
For k = 0 To 1
C = 0 :u = 0
For j = 0 To length of Training set
If m = 1 Then If fen1(i) = fe1(j) And fcn(k) = fc(j) Then C =
C + 1
If m = 2 Then If fen2(i) = fe2(j) And fcn(k) = fc(j) Then C =
C + 1
If m = 3 Then If fen3(i) = fe3(j) And fcn(k) = fc(j) Then C =
C + 1

If m = 4 Then If fen4(i) = fe4(j) And fcn(k) = fc(j) Then C =
C + 1
If m = 5 Then If fen5(i) = fe5(j) And fcn(k) = fc(j) Then C =
C + 1
If m = 6 Then If fen6(i) = fe6(j) And fcn(k) = fc(j) Then C =
C + 1
If m = 7 Then If fen7(i) = fe7(j) And fcn(k) = fc(j) Then C =
C + 1
If m = 8 Then If fen8(i) = fe8(j) And fcn(k) = fc(j) Then C =
C + 1
If m = 9 Then If fen9(i) = fe9(j) And fcn(k) = fc(j) Then C =
C + 1
If m = 10 Then If fen10(i) = fe10(j) And fcn(k) = fc(j) Then
C = C + 1
Next
// likelihood tables
If m = 1 Thenpl1(i, k) = Math.Round((C / (pfc(k) *
count_sw)), 3)
If m = 2 Thenpl2(i, k) = Math.Round((C / (pfc(k) *
count_sw)), 3)
If m = 3 Then pl3(i, k) = Math.Round((C / (pfc(k) *
count_sw)), 3)
If m = 4 Then pl4(i, k) = Math.Round((C / (pfc(k) *
count_sw)), 3)
If m = 5 Then pl5(i, k) = Math.Round((C / (pfc(k) *
count_sw)), 3)
If m = 6 Then pl6(i, k) = Math.Round((C / (pfc(k) *
count_sw)), 3)
If m = 7 Then pl7(i, k) = Math.Round((C / (pfc(k) *
count_sw)), 3)
If m = 8 Then pl8(i, k) = Math.Round((C / (pfc(k) *
count_sw)), 3)
End for
End for
End for
End
```

Figure (4) naïve Bayes algorithm

4. EXPERIMENTAL RESULTS

As it was depicted before, the input data of the Diabetes Classification System was Pima Indian diabetes disease measurements.

There is a significant stage prior to classification, the numerical values that are extracted from properties must be transformed to categorical values, which will be utilized to train the classifier with the use of the EWID algorithm which has been depicted in the third chapter. The Diabetes Classification System categorical values of the five properties are depicted in Table (2). This table is made up of three fields as Attributes, Attribute-value and Range of Values. The first field corresponds to the five properties (How many times pregnant, Diastolic blood pressure, Triceps skin fold thickness, serum).

Insulin and Diabetes pedigree function Feature. The second field corresponds to the categorical values of five properties. The third field corresponds to the range of values obtained by EWID algorithm.

Table (2) Categorical values of features

Attributes	Attribute-value	Range of Values
Number of times pregnant	low	[0 To 2]
	Medium	[3 To 5]
	high	[6 To 17]
Diastolic blood pressure	low	[0 To 80]
	medium	[80 To 100]
	high	[100 To 122]
Triceps Skin	low	[0 To 20]

fold thickness	medium	[20 To 60]
	high	[60 To 99]
serum insulin	Normal	[0 To 280]
	Abnormal	[280 To 860]
Diabetes pedigree function Feature	low	[0.084 To 1.251]
	high	[1.251 To 2.42]

5. PERFORMANCE MEASURES OF THE DIABETES CLASSIFICATION SYSTEM:

The implementation of the DCS is evaluated by the use of confusion matrix, running time and the precision of the classification. The main method to evaluate the precision of classification is k-fold cross-validation that splits the data into k (k = 3 in this work) sub-sets, typically of identical sizes, via random sampling and probably with classification. Then every one of the subsets in turn is utilized to test and the rest is for the training. The estimates of precision are averaged for yielding a general estimation of precision. The confusion matrix of DCS implementation has been extracted from the testing part with the use of ID3 and the naive Bayesian might be depicted in Tables (3) and (4). The matrices of confusion must be read as follows: rows mean the object to be recognized and columns mean the label the classifiers associates at this object. The running time and precision of classification of DCS implementation can be listed in Table (5) the running time measured in seconds, which is computed by the difference between the time of the beginning of the implementation of the DCS and the time of its end. The running time is computed 5 times for the sake of making sure that the outputs are clear, due to the fact that the CPU could be busy with other processes.

Table (3) confusion matrix using ID3 classifier

Predicate Class	**Actual Class**

	Normal	**Abnormal**
Normal	159	8
Abnormal	20	86

Table (4): Confusion matrix using naïve classifier.

Predicate Class	**Actual Class**	
	Normal	**Abnormal**
Normal	165	8
Abnormal	8	88

Table (5): Performance measures of the DCS.

Class Type	**ID3**		**naïve classifier**		**k-fold**
	Running time	Accuracy	Running time	Accuracy	3
Average	35 Sec	91%	27 Sec	94%	

6. PROPOSED SYSTEM VS. RELATED WORK SYSTEM:

The performance of the suggested system was depicted by 2kinds of properties and compared with other related system as listed in table (6).

Table (6) comparison proposed system with different works

Ref.	**Authors**	**year**	**Technique used**	**Accuracy**
[17]	1Saba Bashir, 2Usman Qamar, 3Farhan Hassan Khan, 4M.Younus Javed	2014	C4.5 and ID3 and CART	91%
[18]	Aisarya S^1 and Anto S^2	2014	GA and Extreme Learning	89.54%
[21]	Ravi s, Smt.T	2014	Fcm+Svm	94.3%
[22]	A. Khan & K. Revett	2004	Genetic algorithm based on rough set	94.3%

[23]	Selvakuberan K& Harini B	2011	Combination ranker Search approach	81%
[24]	M. NirmalaDevi	2008	k-mean +k-nearest neighbor(KNN)	97.4%
[25]	C. M. Velu and K. R. Kashwan	2013	EM algorithm +H-mean +clustering and Genetic algorithm	79%
[26]	M. Dura raj, G. Kalaiselvi	2015	ANN +SVM+C4.5+K-NN	89%
	Proposed system		ID3 Naïve bayes	91 % 94%

7. CONCLUSIONS AND FUTURE WORK

Classifying the Diabetes situations as normal or abnormal will offer a second opinion to the physician concerning the treatment of patients and therefore, the suggested system will be assisting the doctors in improving the disease diagnoses.

Minimized time of execution of tree constructing and naïve possibility via getting rid of irrelevant and redundant properties. The estimated time needed for the suggested system (measured by seconds) is minimized and estimated of about 35 seconds for ID3 classifier compared to 27 s for the naïve classifier, each of the 5 properties that is chosen according to entropy was used in this work are necessary to build a good classifier. The experimental outputs indicate that the ID3 classifier scored a percentage of precision of up to 91% and an accuracy percentage of up to 94% with the use of the naïve Classifier for the same number of the test group. For the future works the DCS could be updated according to some method of enhancing the system's quality are developing the chosen property by using another criterion such as Time Frequency, Term Frequency Inverse Frequency, and so on, implementing a different classification algorithm for constructing the classifier like the Artifial Neural Networks or Support Vector Machine.

REFERENCES:

[1] Smitha, P., Shaji, L., & Mini, M. G. (2011). A review of medical image classification techniques. In International conference on VLSI, Communication & Intrumrnataiom (pp. 34-38).

[2] Canlas, R. D. (2009). Data mining in healthcare: Current applications and issues. School of Information Systems & Management, Carnegie Mellon University, Australia.

[3] Porter, T., & Green, B. (2009). Identifying diabetic patients: a data mining approach. *AMCIS 2009 Proceedings*, 500.

[4]Naik, J., & Patel, S. (2014). Tumor detection and classification using decision tree in brain MRI. International Journal of Computer Science and Network Security (IJCSNS), 14(6), 87.

[5] Mao, Y., Chen, Y., Hackmann, G., Chen, M., Lu, C., Kollef, M., & Bailey, T. C. (2011, December). Medical data mining for early deterioration warning in general hospital wards. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on* (pp. 1042-1049). IEEE.

[6] Esmaeli, M. (2011). A Scalable Parallel Algorithm for Decision Support from Multidimensional Sequence Data.

[7]Selvakuberan, K., Kayathiri, D., Harini, B., & Devi, M. I. (2011, April). An efficient feature selection method for classification in health care systems using machine learning techniques. In *Electronics Computer Technology (ICECT), 2011 3rd International Conference on* (Vol. 4, pp. 223-226). IEEE.

[8]Velu, C. M., & Kashwan, K. R. (2013, February). Visual data mining techniques for classification of diabetic patients. In *Advance*

Computing Conference (IACC), 2013 IEEE 3rd International (pp. 1070-1075). IEEE.

[9]Bashir, S., Qamar, U., Khan, F. H., & Javed, M. Y. (2014, December). An Efficient Rule-Based Classification of Diabetes Using ID3, C4. 5, & CART Ensembles. In *Frontiers of Information Technology (FIT), 2014 12th International Conference on* (pp. 226-231). IEEE.

[10] Aishwarya, S., & Anto, S. (2014). A Medical Expert System based on Genetic Algorithm and Extreme Learning Machine for Diabetes Disease Diagnosis. *International Journal of Science, Engineering and Technology Research (IJSETR)*, *3*(5).

[11]Sanakal, R., & Jayakumari, T. (2014). Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine. *International Journal of Computer Trends and Technology*, *11*(2), 94-8.

[12] Durairaj, M., & Kalaiselvi, G. (2015). Prediction of diabetes using soft computing techniques-A survey. *International journal of scientific and technology research*, *4*(3), 190-192.

تصنيف مرض السكري باستخدام الخوارزمية التكرارية وخوارزمية المصنف الساذج

خالد شاكر جاسم ، هديل محمد صالح

جامعة الانبار– كلية الحاسوب وتكنولوجيا المعلومات

Khalidalhity@gmail.com

الخلاصة: يمكن تعريف مرض السكري بأنه مرض مزمن سببه ارتفاع مستويات الجلوكوز في الدم تنتج عنه مشاكل في الطريقة التي يتم بها توليد الأنسولين، وطريقة عمل الأنسولين، أوكليهما. والهدف من هذا البحث هو اقتراح تقنية تعتمد على شجرة القرار (ID3) ونايف بايز لتصنيف مرض السكري والحد من أخطاء التصنيف عن طريق زيادة دقة التصنيف، وتم تقييم نتائج الطريقة المقترحة عن طريق مقارنتها مع نتائج أخرى من خلال تطبيق النظام المقترح على بيانات بيما الهندية لمرض السكري والتي تم الحصول عليها من موقع الـ UCI لقواعد البيانات. وتشير النتائج التجريبية إلى أن المصنف ID3 سجل نسبة من الدقة تصل إلى 91٪ ووصلت دقه المصنف الساذج إلى 94٪ لنفس العدد من مجموعة الاختبار.