Al-Nahrain Journal of Science

ANJS, Vol.25 (1), March, 2022, pp. 51-58

An Efficient Method of Classification the Gestational Diabetes Using ID3 Classifier

Safa A. Hameed

Department of Computer Science, College of Engineering and Science, Bayan University, Erbil, Kurdistan Region-Iraq

Article's Information

Received:
29.12.2021
Accepted:
28.02.2022
Published:
28.03.2022

Keywords:
Classification
Diabetes
Discretization
ID3
Machine Learning

Abstract

Artificial intelligence algorithms have an important and effective role in the medical field, especially in the field of diagnosing diseases. This research focuses on predicting the diagnosis of gestational diabetes by using the Iterative Dichotomiser3 (ID3) classifier algorithm, which is utilized to identify gestational diabetes; it was one of the most significant algorithms employed in this study. Training and testing are two critical phases of the research study. This study employed the Pima Indians Diabetes dataset, which comprised 768 women aged 21 and above with the eight reported traits. A feature selection stage, a discretization step, and using the classifier model for producing decision rules are all part of the Pima Indians diabetes data gathering process (Diabetes Dataset). In this study, the decision tree is employed to develop the classifier model, which is based on Diabetes training. The Iterative Dichotomiser3 (ID3) technique may be used to run the decision tree classification process. Diabetes is tested using decision rules, and the classifier implementation confusion matrix was retrieved from the testing portion. The system delivered high-quality results with a 94 percent accuracy rate.

DOI: 10.22401/ANJS.25.1.09

*Corresponding author: safa.hamid@bnu.edu.iq

1. Introduction

Building a system that is based on real-data training and can be used to forecast results and provide high-performance solutions is critical to solving issues in a variety of aspects in the life; it can be used in the health field. Healthcare data is exploding these days, needing an efficient, effective, and timely solution to reduce mortality rates. One of the most dangerous chronic health problems is diabetes. People nowadays are vulnerable to a wide range of illnesses that cause them to live shorter lives. The majority of people, especially women, are impacted by diabetes, which is a chronic condition. [1]. One of the most common disorders that affect pregnant women is gestational diabetes mellitus (GDM). This condition affects approximately 7% of pregnant women on average [2]. Gestational diabetes (GDM) is defined as the normal glucose metabolism before pregnancy and the development of diabetes during pregnancy. It can happen to women who have never had diabetes before [3,4]. To diagnose Diabetes, medical practitioners require a trustworthy prediction approach [5]. Considering what variables are affecting its adoption in various technologies. Hs-CRP and SHBG are two significant markers that regulate blood sugar levels in pregnant women. Compare blood sugar levels with uric acid, albumin, and creatinine values, if available [6]. Machine learning is a branch of science that studies how machines learn from their experiences [7]. The prediction of diabetics will be easier with these techniques, and mankind will gain a lot of benefits [8]. The use of data mining techniques to help individuals forecast diabetes has become a hot issue [9]. The decision tree technique is a key component of data classification mining, and the ID3 (Iterative Dichotomiser 3) approach is a well-known example that has produced impressive results in the field [10]. The ID3 method is a broad classification function that has several advantages, including easy-to-understand judgment rules and a simple model [11]. In this study, a smart approach was developed based on actual data to design a system that can identify gestational diabetes. This work involved numerous processes and two stages: training and testing, and it based on particular features as mentioned in the following paragraphs.

2. Related Work

Several works studied the learnability of decision trees. A decision tree is one of the most widely used classification methods in data mining. There are wide methods and applications in this field.

The authors of [1] test the diagnostic usefulness of glycated hemoglobin A1c (HbA1c) in identifying gestational diabetes mellitus, researchers used a technique (GDM). This prospective research included pregnant women who were 24 weeks or older and had an abnormal GDM screening test. The HbA1c level was measured at the same time as the 3-hour 100-gram oral glucose tolerance test (OGTT). A 3-h 100 g OGTT was used to diagnose GDM, according to the National Diabetes Data Group. The pregnancy outcome was tracked in terms of gestational age at delivery, preeclampsia, primary cesarean section rate, birth weight, fetal macrosomia, and neonatal intensive care unit hospitalization. A total of 141 ladies were registered. NDDG criteria identified 35 women with GDM, and there was a 74.6 percent accuracy rate. The work in [4] was to enhance diagnostic accuracy by improving data quality and selecting appropriate classifiers. The Random decision forests (RF) method was used with estimator selection for classification, and the model was utilized to predict GDM using patient health data. When compared to other classifiers, the suggested model has a high accuracy of 91 percent. The suggested technique in [5] uses multiple classification algorithms to detect diabetes using a given dataset, including ID3, C4.5, LDA, and Nave Bayes. With 91 percent accuracy and a 0.938 error rate, C4.5 was determined to be the top algorithm.

In [6] a total of 400 pregnant women took part in the prospective observational research. Before 15 weeks of pregnancy, the maternal blood sex hormone-binding globulin (SHBG), high-sensitive C-reactive protein (hs-CRP), uric acid, creatinine, and albumin were tested, with an overall accuracy of 75.46 percent. [8] employed a sample of Pima Indian diabetes dataset and used the R tool to do statistical analysis for producing the graph and computing the Gini index, as well as developing a prediction model using the K-Nearest Neighbor (KNN) Algorithm, which had a79 percent efficiency.

The dataset utilized in another research article given by [9] was the Pima Indian diabetes dataset. To increase the quality of the data, preprocessing was performed. To create the Nave Bayes model, a classifier was applied to the changed dataset. Finally, weka was used to simulate the results, and the model's accuracy was 72.3 percent. Where as [14] in this research work used the publicly available Pima Indian diabetic database (PIDD), they tested data mining algorithms using ROSETTA software to predict their accuracy in predicting diabetic status from the 8 variables given. The accuracy of predicting diabetic status on the PIDD was 82.6% on the initial random sample, which exceeds the previously used machine learning algorithms that ranged from 66-81%. The average accuracy of a group of ten random samples was 73.2 percent.

The implementation in [15] was done in Tanagra using the C4.5 Algorithm, which has a 0.11 error rate. Using the data mining tool Tanagra, this article proved the efficacy of categorization on a dataset including records of both diabetes and non-diabetic patients. The dataset was hand-crafted using reports from the Jyoti Diagnostic & Research Centre hospital. In this research, a system based on real data was used which is Pima India diabetes data set by uses the ID3 algorithm to diagnose gestational diabetes and predicts the results. The accuracy of the system has reached 96%.

3. Method

One of the most essential approaches in the health industry is the use of smart technologies to diagnose illnesses [12] [13]. When it comes to identifying illnesses, artificial intelligence algorithms play a critical and accurate role [14] [15].

The proposed system in this paper is the Diabetes Classification System which consists of two phases are the training and testing phase as shown in Figure 1, in the first phase: read a function that has been specified Diabetes Information Set, including the Pima India diabetes data set as the dataset, as well as the following steps: feature selection, the Discretization, and the classifier approach employed in the decision-making Rules to use the classification system. The second phase is the testing phase: in this phase, we use two distinct functions to read the data collection, discretize it, apply the decision criteria, and print it out. Pima Indians diabetes data collection involves a feature selection step, a discretization step, and applying the classifier model for generating decision rules (Diabetes Dataset). The second phase is the research stage, which has two unique functions: read data and write data.

3.1 Data collection:

The data used in the proposed system is taken from UCI Machine Learning Repository. The data set included 768 women aged 21 and up:

- 500 negative cases and
- 268 positive cases

With the following eight documented features:

1. The number of pregnancies you've had before.
2. In an oral glucose tolerance test, plasma glucose levels at 2-hours.
3. Blood pressure in the diastole (mm Hg).
4. Skinfold thickness of the triceps (mm).
5. Insulin serum (mu U/ml) after 2 hours.
6. Body mass index BMI (weight in kilograms divided by height in meters).
7. Pedigree function in diabetes.
8. Age (years).

The existence of missing data is one of the most difficult PID data collections. There are a variety of causes for the lack of comprehensive data, including patient death, equipment failures, and respondents' reluctance to answer some questions. In more than half of the cases, one or more characteristics have missing data.

The Pima Indian Diabetes Data set (PIDD) used in this study was derived from UCI's machine learning respiratory research on diabetes. Every patient in this database is a Pima Indian ladies with at least 21 years who live in or near Arizona. There are eight properties in each dataset sample (attributes), as shown in Table 1. There are 768 samples in all, divided into two groups.

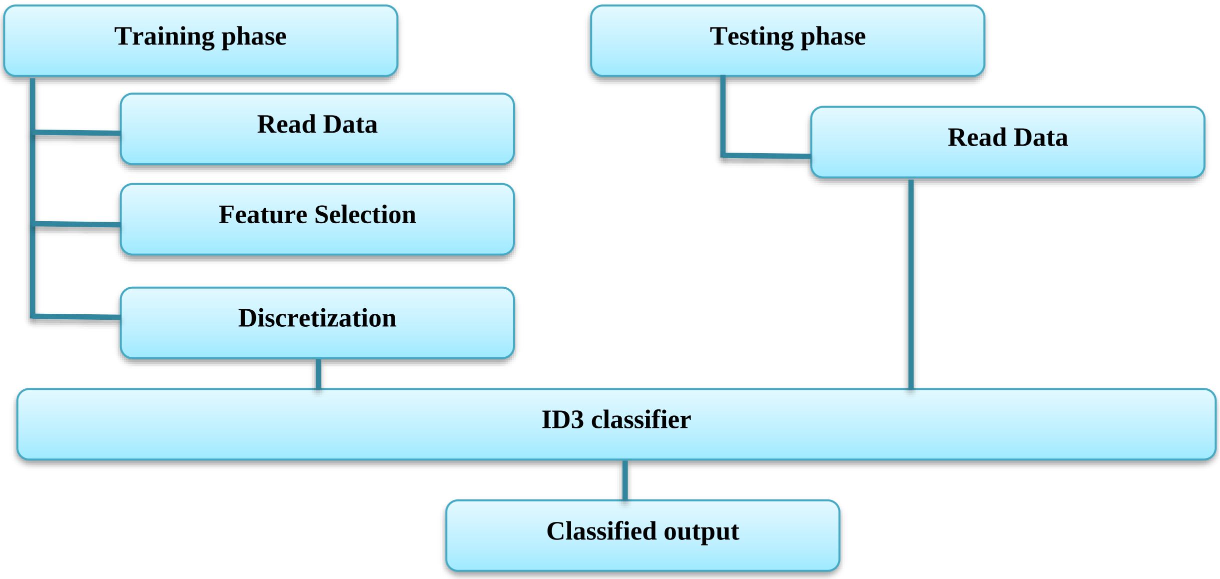

Figure 1. The main block diagram of the training and testing phases of the Diabetes Classification System.

Table 1. Pima Indian diabetes features.

Feature number	Feature Description	Domain
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

Formation of folds: The complete labeled dataset is separated into mutually exclusive folds to perform the cross-validation procedure. The following is the class distribution:

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

The 768 examples in this paper were chosen from the PIDD database, which means:

- The training set contains 499 instances, 325 of which are normal and 174 of which are abnormal (i.e.; 65% of the total).
- 269 instances are used for testing, 175 of which are normal and 94 which are abnormal (i.e.; 35% of the total) as seen in Table 2.

3.2 Stage of feature selection:

The stage of feature selection is the most important in the development of a Diabetes Classification System since the efficacy is dependent on the accuracy of the selection.

The PIDD data set has eight attributes in each sample. According to the entropy of the property with class, Assume S (training data) is a collection of C outputs. Let P (I) be the fraction of S that belongs to class I, where I is different from one to C in the classification problem with C classes.

P is the basic diversity metric (I) (1)

Entropy is an information–theoretical way of measuring the quality of a split. It determines the amount of data in a specific attribute.

$$\text{Entropy}(S) = \sum\nolimits_{I=1}(-P(I)P(I)) \quad (2)$$

Gain is the information gain of the example set S on the attribute A. (S, A).

$$\text{Gain}(S,A) = \text{Entropy}(S) - \Sigma\left(\left(\frac{|SV|}{|S|}\right) \times \text{Entropy}(SV)\right) \quad (3)$$

where SV = subset of S in which feature A has value V, |SV| = amount of data in SV, and |S| = amount of elements in S, and is over every value V of every conceivable value of the attribute A.

A subset of eight properties is chosen. When there are more than five homes to choose from, the process takes longer. Selecting less than five attributes, on the other hand, takes less work but results in a less accurate Diabetes Classification System. The ideal five attributes (1, 3, 4, 5, and 7) for depicting Diabetes are chosen based on increased entropy and will be used as an input to the classification section later.

3.3 Discretization

Diabetes is difficult to explain, thus it’s critical to improve PIDD to make the categorization step easier and more efficient.

Before the categorization procedure in the Diabetes Classification System, a crucial phase must be completed. It is necessary to transform the numerical values of eight attributes into category values. This is accomplished by splitting the range of values for eight characteristics into k equal-sized bins, or in other words, equal-width intervals, where k is a user-selected quantity based on data length. Equal Width Interval Discretization [16,17] is depicted in Algorithm 1: (EWID) in Figure 2.

Table 2. Samples of categorical features values.

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

Algorithm 1. Equal Width Interval Discretization (EWID)

```
Input: Eight attributes have numerical values.
Output: Eight attributes have categorical values.
Begin
minimum1 = Op1(0) : maximum1 = Op 1(0) : minimum 2 = Op 2(0) : maximum 2 = Op 2(0) : minimum 3 = Op 3(0) : maximum 3 = Op 3(0)
minimum4 = Op 4(0) : max4 = Op 4(0) : minimum 5 = Op 5(0) : maximum 5 = Op 5(0) : minimum 6 = Op 6(0) : maximum 6 = Op 6 (0)
minimum 7 = Op 7(0) : maximum 7 = Op 7(0) : minimum 8 = Op 8(0) : maximum 8 = Op 8(0)
For x = 1 To Len of data set {
If Op1(x) > maximum1 Then maximum1 = Op1(x) : If Op1(x) < minimum 1 Then minimum 1 = Op1(x)
If Op2(x) > maximum2 Then maximum2 = Op2(x) : If Op2(x) < minimum 2 Then minimum 2 = Op2(x)
If Op3(x) > maximum3 Then maximum3 = Op3(x) : If Op3(x) < minimum 3 Then minimum 3 = Op 3(x)
If Op4(x) > maximum4 Then maximum4 = Op4(x) : If Op4(x) < minimum 4 Then minimum 4 = Op4(x)
If Op5(x) > maximum5 Then maximum5 = Op5(x) : If Op5(x) < minimum 5 Then minimum 5 = Op5(x)
If Op6(x) > maximum6 Then maximum6 = Op6(x) : If Op6(x) < minimum 6 Then minimum 6 = Op6(x)
If Op7(x) > maximum7 Then maximum7 = Op7(x) : If Op7(x) < minimum7 Then minimum 7 = Op7(x)
If Op8(x) > maximum8 Then maximum8 = Op8(x) : If Op8(x) < minimum8 Then minimum8 = Op8(x) }
K=3 // possibilities number
G1 = Round(((maximum1 - minimum1) / k), 3) : G2 = Round(((maximum2 - minimum2)/ k), 3)
G3 = Round(((maximum3 - minimum3) / k), 3) : G4 = Round(((maximum4 - minimum4)/ k), 3)
G5 = Round(((maximum5 - minimum5) / k), 3) : G6 = Round(((maximum6 - minimum6)/ k), 3)
G7 = Round(((maximum7 - minimum7) / k), 3) : G8 = Round(((maximum8 - minimum8)/ k), 3)
// Identify k ranges for each of the eight attributes.
LowOp1 = (minimum1 + G1) : medOp1 = (minimum 1 + (2 * G1)) : highOp1 = (minimum1 + (3 * G1))
LowOp2 = (minimum2 + G2) : medOp2 = (minimum2 + (2 * G2)) : highOp2 = (minimum2 + (3 * G2))
        = (minimum3 + G3) : medOp3 = (minimum3 + (2 * G3)) : highOp3 = (minimum3 + (3 * G3))
LowOp4 = (minimum4 + G4) : medOp4 = (minimum4 + (2 * G4)) : highOp4 = (minimum4 + (3 * G4))
LowOp5 = (minimum5 + G5) : medOp5 = (minimum5 + (2 * G5)) : highOp5 = (minimum5 + (3 * G5))
LowOp6 = (minimum6 + G6) : medOp6 = (minimum6 + (2 * G6)) : highOp6 = (minimum6 + (3 * G6))
LowOp7 = (minimum7 + G7) : medOp7 = (minimum7 + (2 * G7)) : highOp7 = (minimum7 + (3 * G7))
LowOp8 = (minimum8 + G8) : medOp8 = (minimum8 + (2 * G8)) : highOp8 = (minimum8 + (3 * G8))
End
```

Figure 2. The structure for the EWID algorithm code

3.4 Model of classifier:

The most well-known task classification is the constructing classifier model. This structure was used to predict the Diabetes class, which might be either normal or pathological, for a large number of patients,

The Diabetes (Training-Set) database is built up of attribute value representation with five categorical attributes (1, 3, 4, 5 and 7) and class attribute. Those characteristics are fed into the learning classifier model. A classifier model is used to predict the new case.

4. Result

The input data in Diabetes Classification System included Pima Indian diabetes illness measures, as previously stated. Before classification, the numerical values are taken from characteristics and must be converted to categorical values, which will then be used to train the classifier using the EWID method. Table 3 shows the categorical values of the five attributes according to the Diabetes Classification System. Attributes, Attribute-value, and the domain of values: are the three fields that make up this table.

Table 3. Feature categorical values.

Attributes	Attribute-value	Domain
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 Skinfold thickness	low	[0 To 20]
	Medium	[20 To 60]
	High	[60 To 99]
Serum insulin	Normal	[0 To 280]
	Abnormal	[280 To 860]
Diabetes pedigree	Low	[0.084 To 1.251]
	High	[1.251 To 2.42]

The five attributes are shown in the first field: How many times pregnant, Diastolic blood pressure, Triceps skinfold thickness, serum insulin, and diabetes pedigree. The categorical values of five attributes are shown in the second field. The domain of values obtained by the EWID method is represented in the third field. Table 4 shows the categorical values samples that were obtained by changing numerical property values of the Diabetes circumstance with the class attribute according to the Table 1 the range of value field. The findings of the top five attributes chosen for usage in the Classifier model are shown in Table 5. The classifier model was trained and tested on the following properties: Preg denotes the number of times a woman has been pregnant, Pres denotes the Diastolic blood pressure, and Skin denotes the thickness of the Triceps skin folds, Insu denotes serum insulin, and Pedi denotes the Diabetes pedigree function characteristic.

The entropy Eq. (1) is used to calculate the entropy for each feature.

Table 4. Categorical features values samples.

Id	Preg	Press	Skin	Insu	Pedi	Class
1	Low	Low	Low	Normal	Low	Normal
2	Low	Medium	Medium	Normal	High	Normal
3	High	Low	Medium	Normal	Low	Normal
4	High	Medium	Medium	Abnormal	High	Abnormal
5	Low	High	Medium	Normal	High	Abnormal
6	Medium	High	High	Abnormal	High	Abnormal

Table 5. Entropy of categorical features values.

Entropy	Preg	Plas	Skin	Skin	Ins	Mass	Pedi	Age
	0.71	0.13	0.49	0.88	0.55	0.02	0.64	0.08

4.1 ID3 classifier:

The decision tree is used to build the classifier model in this work, which is based on training Diabetes. The Iterative Dichotomiser 3 (ID3) algorithm may be used to execute the decision tree classification procedure, as illustrated in Figure 1. This technique can generate decision rules using a decision tree based on qualities that play a key role in classifications based on entropy and information gain [18]. Because the ID3 method was computed on the training set, the decision rules were created using the if-then structure.

The classifier model conducts a categorization of the tested data as normal or abnormal using decision rules created by the ID3 algorithm during the testing step.

These rules are according to the data that was utilized on during the discretization stage as shown:

1. If pres=”normal,” preg=”low,” skin=”abnormal high,” ins=”normal,” and Pedi=”G2,” output=”tested normal” is returned.
2. If pres equals “C4”, preg equals “a4”, and skin equals “d1,” output equals “tested positive.”
3. If pres is “C4”, preg is “a4”, and skin is “d3,” output is “tested negative.”
4. If pres is “C4”, preg is “a4”, and skin is “d2,” output is “tested negative.”
5. If pres equals “C4” and preg equals “a5”, output equals “tested positive.”
6. If pres=”C4” and preg=”a1,” output=”tested negative” is returned.
7. If pres=”C4” and preg=”a3,” output=”tested negative” is returned.

For the purpose of converting a decision tree into decision rules, that are easier to understand and apply on a computer.

Except for the fact that the Diabetes is tested according to decision rules, every step in the testing stage is the same as in the training stage. The ID3 classifier was trained using the Diabetes cases as illustrated in Figure 3.

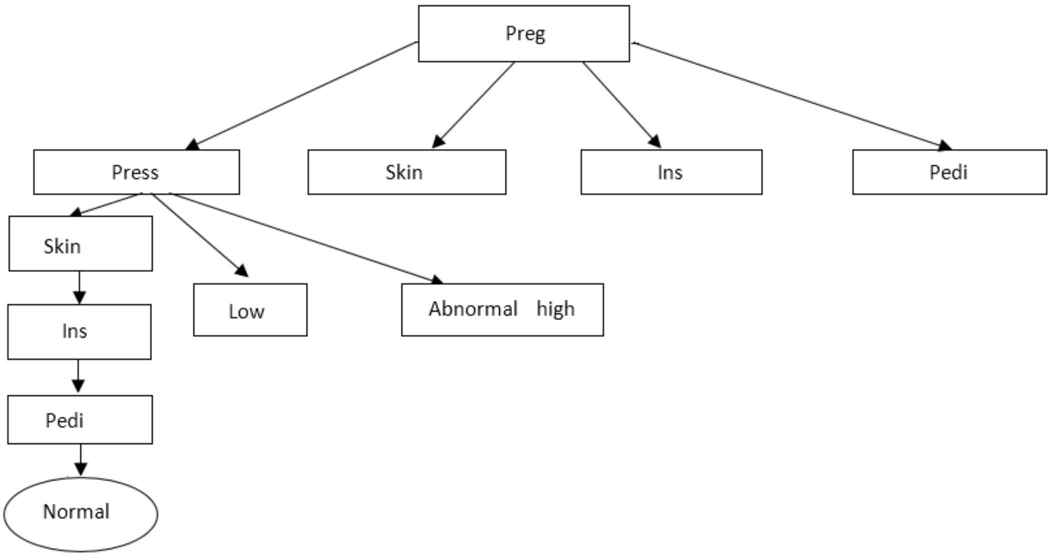

Figure 3. A decision tree in action.

With the use of ID3, the confusion matrix of classifier implementation was taken from the testing section and displayed in Tables 6. This table includes both predicate and actual classes.

The accuracy and error rate for diagnosed cases are calculated using Table 6. The values used in calculating the accuracy of the ID3 classifier by using the accuracy Eq. (4), where the:

TP: true positive

TN: true negative

TP and TN are added together, then divided by the sum of all with FP (false positive) and FN (false negative), and the error is computed using the Eq. (5) [19].

$$\text{Accuracy} = (TP+TN)\backslash(TP+TN+FP+FN) \tag{4}$$

$$\text{Accuracy} = (165+88)/(165+8+8+88)$$
$$= 0.94$$

$$\text{Error} = (FP+TN)/(TP+TN+FP+FN) \tag{5}$$
$$= (8+8)/(165+8+8+88)$$
$$= 0.06$$

Table 6. The confusion matrix using ID3 classifier.

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

The suggested work was utilized to boost accuracy by employing the ID3 classifier approach, where the mechanism was conducted to avoid the mistakes of earlier works and in a higher performance way to achieve correct diagnosis and classification, this work has been compared to previous works as shown in the Table 7.

Table 7. Comparing the accuracy of the results between previous work.

Authors	The year	Dataset	The method used	Accuracy
Sumathi et al, [4]	2021	The Pima Indian Diabetes Dataset	Random decision forests (RF) method	91.447
Patcharaporn et al,[1]	2021	National Diabetes Data Group (NDDG)	A HbA1c test was carried out.	74.6
Divya Jain et al, [15]	2014	Reports obtained from the Jyoti Diagnostic & Research Centre hospital	C4.5 Algorithm in Tanagra.	89%
Ahmed et a, 2014	2014	A total of 400 pregnant women took part in this prospective observational research.	Hs-CRP and SHBG	75.46%.
Rajesh et al, [5]	2012	The Pima Indian Diabetes Dataset	C4.5 algorithm	91%
Joseph et al, [14]	2001	The Pima Indian Diabetes Dataset	data mining algorithms using ROSETTA software	73.2%
Anurag et al, [8]	2016	The Pima Indian Diabetes Dataset	K-Nearest Neighbor (KNN) Algorithm	79%.
Yang et al, [9]	2012	The Pima Indian Diabetes Dataset	The Pima Indian Diabetes Dataset	72.3%

5. Conclusion and Future Work

The proposed system assists doctors in improving disease diagnoses by classifying Diabetes situations as normal or abnormal, which will provide a second opinion to the physician regarding patient treatment. By removing unnecessary and repetitive features, we were able to reduce the time it took to form a tree. The estimated time needed for the recommended system (measured in seconds) is minimized, with an estimated duration of roughly 35 seconds for the ID3 classifier compared to 27s. Each of the 5 attributes chosen based on entropy in this study is required to develop a decent classifier. According to the results of the experiments, the ID3 classifier achieved a precision of up to 94 percent. It could be updated in the future based on some method of improving the system's quality, such as developing the chosen property using another criterion such as Time Frequency, Term Frequency, Inverse Frequency, and so on, or implementing a different classification algorithm for constructing the classifier, such as the Artificial Neural Network or Support Vector Machine.

References

[1] Patcharaporn S. and Vorapong P.; "Diagnostic accuracy of HbA1c in detecting gestational diabetes mellitus", 2021, The Journal of Maternal-Fetal & Neonatal Medicine, 33(20), 2020.

[2] Filho E.G.; Pinheiro P.R.; Pinheiro M.C.D.; Nunes L.C.; Gomes L.B.G. and Farias P.P.M.; “Support to Early Diagnosis of Gestational Diabetes Aided by Bayesian Networks. In: Silhavy R. (eds) Artificial Intelligence Methods in Intelligent Algorithms", CSOC 2019. Advances in Intelligent Systems and Computing, 985, Springer, Cham.; 2019 https://doi.org/10.1007/978-3-030-19810-7_36.

[3] Fan D.; Weiyang Z.; Wei W.; Danhong P.; Tian X.; Jun W.; Gongdao W. and Fengzhen H.; "Prediction of pregnancy diabetes based on machine learning", The Third International Conference on Biological Information and Biomedical Engineering, BIBE, 2019.

[4] Sumathi A.; Meganathan S. and Vijayakumar V.; "Prognosis Model for Gestational Diabetes Using Machine Learning Techniques", Sensors and Materials, 33 (9): 3011-3025, MYU Tokyo, 2021.

[5] Rajesh K. and Sangeetha V.; "Application of data mining methods and techniques for diabetes diagnosis", International Journal of Engineering and Innovative Technology (IJEIT), 2(3): 224-9, 2012.

[6] Ahmed M. M.; Ghada A. F. M.; Walaa A. M. and Dalia A. H.; "Comparative study between different biomarkers for early prediction of gestational diabetes mellitus", The Journal of Maternal-Fetal & Neonatal Medicine, 27 (11), 2014.

[7] Ioannis K.; Olga T.; Athanasios S.; Nicos M.; Ioannis V. and Ioanna C.; "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal, 2017.

[8] Anurag K. S.; Chandankumar and Neha M.; "Analysis of diabetic dataset and developing prediction model by using HIVE and R", Indian Journal of Science and Technology, 9 (47): 1-5, 2016.

[9] Yang G.; Guohua B. and Yan H.; "Using bayes network for prediction of type-2 diabetes", International Conference for Internet Technology and Secured Transactions: 471-472, 2012.

[10] Yingying W.; Yibin L.; Yong S.; Xuewen R. and Shuaishuai Z.; "Improvement of ID3 algorithm based on simplified information entropy and coordination degree", Algorithms MDPI, 10 (4), 124, 2017. https://doi.org/10.3390/ a10040124

[11] Fayyad U.M. and Irani K.B.; "The attribute selection problem in decision tree generation", In Proceedings of the National Conference on Artificial Intelligence, San Jose, CA, USA, 12-16 July: 104-110, 1992.

[12] Suresh P. K. and Pranavi S.; "Performance analysis of machine learning algorithms on diabetes dataset using big data analytics", International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), Publisher: IEEE, 2017.

[13] Umamaheswari S. and Shobana A.; "A comprehensive system focusing on analyzing severity of heredity diabetes", Journal of Computational and Theoretical Nanoscience, 15 (6-7), 2018.

[14] Joseph L. B.; “Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?", Computing Science and Statistics Publisher, 2001. http://www.galaxy.gmu.edu/interface/I01/I2001Proceedings/JBreault/JBreault-Paper.pdf

[15] Divya J. and Sumanlata G.; "Predicting the Effect of Diabetes on Kidney using Classification in Tanagra", International Journal of Computer Science and Mobile Computing, 3 (4), April, 2014.

[16] Saleha A. and Nasari F.; “Implementation of equal-width interval discretization in naive bayes method for increasing accuracy of students' majors prediction”, Lontar Komputer, 9 (2), August, 2018. doi: 10.24843/LKJITI.2018.v09.i02.p05 Accredited B by RISTEKDIKTI Decree No. 51/E/KPT/2017.

[17] Dash R.; Paramguru R. L. and Dash R.; “Comparative analysis of supervised and unsupervised discretization techniques", International Journal of Advances in Science and Technology 2, 3: 29-37, 2011.

[18] Surya L. P. and Kiran R. K.; "ID3 and its applications in generation of decision trees across various domains-survey", (IJCSIT) International Journal of Computer Science and Information Technologies, 2015.

[19] Han J. and Kambar M.; "Data mining: Concepts and techniques”, 2nd ed., Morgan Kaufmann Publisher, 2006.