ISSN 2007-9737

Diabetes Diagnosis Using Supervised Learning Technique

Gerardo Martínez Guzmán, Carmen Cerón Garnica*, Yolanda Moyao Martínez,
Mariano Larios Gómez

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Mexico

{gerardo.guzmanm, carmen.ceron, yolanda.moyao, mariano.larios}@correo.buap.mx

Abstract. In this paper, an analysis of variables related with diabetes detection is done, using a decision tree which is a kind of algorithm of automatic supervised nonparametric learning, and that is used for classifying tasks from a set of classified objects, each described by an attribute vector and a class label. The classifying objects that have known class label are taken as example to train a mathematical model able to predict the class labels of new objects that have not been classified yet.

Keywords. Decision tree, diabetes, ID3 algorithm, supervised learning.

1 Introduction

Diabetes is a chronic metabolic disease characterized by elevated levels of blood glucose, which leads to serious damage to the heart, blood vessels, eyes, and kidneys. Type 2 diabetes is the most common form in adults and occurs when the body becomes resistant to insulin or does not produce enough of it. Over the past three decades, the prevalence of type 2 diabetes has risen dramatically in nearly every country, as seen in Figure 1, with a trend in the number of people between 20 and 79 years with diabetes around the world and projected for 2030 and 2045.

For people living with diabetes, access to affordable treatment, including insulin, is critical for their survival. To prevent diabetes, policies and practices are being implemented across entire populations and within specific settings: such as schools, homes, and workplaces, which contribute to good health for everyone.

This disease can be controlled through special treatments, only if detected on time. However, to detect people with the risk of developing this disease, can be a challenge, due to diverse variables that intervene in its development and also due to the complexity of the risk conditions.

For the population with diabetes, an early diagnosis is essential for living well: the longer a person lives with undiagnosed and untreated diabetes, the worse their health outcomes are likely to be. Therefore, easy access to services, such as blood glucose monitoring, should be available in primary healthcare. This should be combined with a healthy diet, physical activity, and, if necessary, medication to control blood pressure and lipids to reduce cardiovascular risk and other complications.

Diabetes is a leading cause of blindness, kidney failure, heart attacks, strokes, and lower limb amputation. Poorly controlled diabetes increases the likelihood of these complications and premature mortality.

For this reason, it is important to count with tools that support to tackle this problem of early detection of diabetes.

Moreover, these tools must be able to give an early diagnosis and thus take fast and well-informed decisions, that allow a better quality of life of the patients, through its appropriate tracking and treatment.

There are other investigations where they have employed statistical methods to discriminate this type of data [1-3, 7, 10]. Artificial intelligence techniques have been used such as Logistic Regression (LR), K Nearest Neighbors (KNN), Support Vector Classifier (SVC), Naive Byes (NB), Decision Tree (DT), Random Forest (RF), Stochastic gradient descent (SGD), [4-14].

doi: 10.13053/CyS-29-4-6124

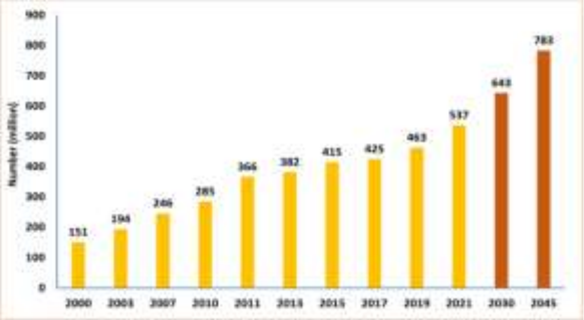


Fig. 1. People with diabetes around the world
Note: Data retrieved from The Diabetes Atlas (10th ed.) of International Diabetes Federation https://diabetesatlas.org.

This paper is organized as follows. In section 1, the review of previous works related with the topic of diabetes and its detection is presented, exploring the risk factors and the strategies that currently are used to solve this issue. In section 2, some conceptual and scientific fundaments needed are presented to understand the study that is presented. In section 3, the used methodology is described for the development of the predictive model, the used type of data is detailed, as well as the processing of these, and finally, the developed algorithm is presented. In section 4, the obtained results are presented from the tests realized with the application of the developed model. In section 5, the interpretation and implications of results are discussed, as well as some of the limitations of the model and research work that could be developed in the future.

2 State of the Art

Supervised learning has demonstrated to be an effective tool in classification problems and diagnosis of chronic diseases, particularly of diabetes mellitus type 2 (DM2), which currently has seen an increment and thus it represents one of the main factors of conditions and deaths globally [22].

Therefore, the detection on time of this disease is essential to reduce the appearance of critical health problems and also the use of traditional methods backed up by clinic results supported with innovative techniques of Machine Learning (ML) that help to develop effective and precise predictive models [19], [20].

There are key variables that impact the process of prediction, particularly biomedical parameters like glucose levels, body mass index, and arterial pressure, which play an important role in improving the efficiency of these models [20].

The incorporation of supervised learning tools has opened up new possibilities in the preventive health area, supplying strategies and early and individualized interventions with affordable prices and with the warranty of a major availability [18].

In this sense, different research works have been presented such as the Hunt algorithm, developed in the 60s to model human learning in psychology. It functions as base of many algorithms of decision trees like ID3, which is attributed to Ross Quinlan and this algorithm is based in the use of entropy and the gain of information, such as metrics to assess the candidates' divisions.

Another algorithm is the C4.5 that considers a posterior iteration of ID3 and also was developed by Quinlan. This is based on using the gain of information or the proportions of gains to assess the division points within the decision trees [16, 17, 21].

Thereafter, Quinlan presents the C4.5 algorithm, as an updated and improved version of ID3. In this version, more specialized techniques are used to perform the calculus of division point, and it also takes into account the pondered gain of information. This modification favored the management of more detailed datasets and gave as a result decision trees more consistent and precise.

Additionally, alternative versions and expansions have been developed, such as C5.0 along with other classification trees, that lift the level of performance and the prediction capacity.

There are other investigations where statistical methods have been employed to discriminate this type of data [1, 2, 3, 7, 10]. Artificial Intelligence techniques have been utilized like Logistic Regression (LR), which its main function is to model the probability of occurrence of an event of binary type; other outstanding method is K Nearest Neighbors (KNN), this generates predictions taking as a base the closeness in the context of properties that characterize the data points in the training set; and Bayesian classifiers such as Naïve Bayes (NB), which use probability foundations to realize the classification.

On the other hand, other more developed artificial intelligence techniques have been used, which include Support Vector Classifier. (SVC), based on increasing the margin between classes in order to improve generalization. There is also Random Forest (RF), which is based on the principle of integrating multiple decision trees to achieve an increase in consistency and accuracy. Finally, there are optimization methods such as Stochastic Gradient Descent (SGD), used for training more sophisticated models like neural networks [4, 5, 6, 8, 9, 10, 11, 12, 13, 14].

Different studies have applied these techniques with the goal of early diabetes prediction, achieving various levels of accuracy and specificity. For example, recent works such as the study presented by [15] where machine learning algorithms like Multiple Logistic Regression, Support Vector Machine, and KNN are employed to determine which of them offers the best accuracy in diabetes prediction. The study concludes that KNN is the best method, as it achieves better performance, with an accuracy of 0.8846 and a specificity of 0.9462.

In addition to algorithms, another important aspect to consider is the selection and analysis of the key variables or attributes that impact diabetes detection. Clinical, demographic, and epidemiological variables have been taken into account in the design of predictive models, with the objective of appropriately assessing the risk of diabetes in different study populations. Thus, the aim is to enhance both the model's performance and its usability in real-world clinical settings [2, 11, 10].

The main contribution of this research work consists in the analysis of variables related with the detection of diabetes, through a decision tree, which is defined as a supervised nonparametric automatic learning, whose function is to realize the process of classification. To that end, an initial set of classification objects, and each one of these sets are described through an attribute vector and a class label.

The literature shows that various studies have applied decision trees to predict type 1 (T1DM) and type 2 (T2DM) diabetes mellitus, achieving reliable performance in patient classification and management. For example, [23] in their study analyzed data from 50 patients with T1DM, applying the decision tree model as well as mathematical models to determine the ability to regulate glycemic levels. Their results show an overall classification error of 13.7%, and a lower error in T1DM classification, especially when specific attributes are integrated [23].

In another study, [23] utilized the C4.5 algorithm with the objective to identify attributes that impact glycemic control, using a sample of 2064 patients with T2DMm identifying key variables, such as diabetes education and lifestyle, that significantly influence changes in glycated hemoglobin levels, which is a key factor in the evolutive state of the disease.

Furthermore, it is reported that in these studies, the performance of decision trees is evaluated using metrics such as accuracy, sensitivity, and specificity. According to [23] in their study, accuracies of approximately 73% were achieved in predicting diabetes, based on different populations in comparative studies.

Other studies compare the efficiency of decision tree classifiers with that of artificial neural network classifiers, observing that although both models reach favorable prediction results, decision trees provide greater clinical understanding. This facilitates their implementation in medical settings for well-informed and personalized decision-making [24].

It can be concluded that the research reported in the literature trends towards integrating optimized decision tree algorithms with advances in identifying key variables to design robust predictive models with a high level of accuracy and clinical utility for the early diagnosis of T1DM and T2DM, thus enhancing prevention capabilities in the healthcare sector through individualized and consistent diagnoses.

Finally, the literature also highlights the relevance of the clinical interpretation of these models and the incorporation of these prediction tools into healthcare systems. The objective is to have sufficient information to increase the effectiveness of decision-making and the individualized treatment of patients in the early stages or at risk of diabetes.

3 Theoretical Framework

In this section, as series definitions of some measures of uncertainty are shown, which are used in the algorithm of decision tree developed in this research work. These measures allow to determine the purity or impurity of a dataset and to guide the process of node division in the tree.

3.1 Entropy

In the technical sense, the word information is a capacity and must be distinguished from the meaning, if the numerical value is calculated for the information, it must always be a capacity. Claude Shannon was the first in suggesting that this capacity can be quantified and develops a specific form of this measure called entropy.

Definition. Let X be a random discrete variable of finte range $I = \{x_1, x_2, \dots x_n\}$ and let $p(x) = P(X = x)$ be its function of probability, where no values with zero probability exist. Then, we define entropy of the random variable X as

$$H(X) = \sum_{i=1}^{n} p(x_i)\, log \frac{1}{p(x_i)}\,. \tag{1}$$

Some important conditions are:

1) Given that this equality is satisfied:

$$log \frac{1}{p(x_i)} = -\, log\, p(x_i). \tag{2}$$

Entropy can be written in the following way:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\, log\, p(x_i). \tag{3}$$

2) For reasons of generality and given that the following limit is satisfied:

$$\lim_{p(x_i)\to 0} p(x_i)\, log\, p(x_i) = 0. \tag{4}$$

The definition of entropy can be extended to include the case of zero probability without changing the value of entropy, adopting convention that if $p(x_i) = 0$ then,

$$p(x_i)\, log\, p(x_i) = 0\,. \tag{5}$$

3) The numerical value of entropy depends on the base b of logarithm which is why it is also often expressed as $H_b(X)$, however, a change of base does not suppose more than a change of scale. For certain values of the base, the unities usually receive special denominations:

$Base\ 2$	the unities are called bits,
$Base\ e$	the unities are called nats,
$Base\ 10$	the unities are called hartleys.

The equivalence among different measure units are obtained in virtue of the known formula of change of base:

$$log_a\, x = \frac{log_b\, x}{log_b\, a}. \tag{6}$$

Using this formula, we can prove that:

$$H_2(X) = \frac{1}{Ln\ 2} H_e(X). \tag{7}$$

4) A new variable can be defined depending on the random variable X as:

$$I(X) = log\ \frac{1}{p(X)} = -log\ p(X). \tag{8}$$

Whose average value is

$$E[I(X)] = -\sum_{x} p(x)\, log\, p(X) = H(X). \tag{9}$$

Thus, if $I(x_i)$ is interpreted as the information that is obtained when x_i occurs, then $H(X)$ is the average information of the values of the random variable X.

3.2 Joint Entropy

Let X be a random discrete variable of finite range $I = \{x_1, x_2, \dots x_n\}$ with a probability function of $p_X(x) = P(X = x)$ and let Y be another random discrete variable of finite range $\varkappa = \{y_1, y_2, \dots y_m\}$ with probability function $p_Y(x) = P(Y = y)$. Consider the random bidimensional variable (X, Y) whose range is,

$$I \times \varkappa = \{(x_i, y_i)\ :\ x_i \in\ I\ , y_i \in\ \varkappa\ \}. \tag{10}$$

And whose joint density function is:

$$p(x, y) = P(X = x, Y = y). \tag{11}$$

Definition. The entropy of a random bidimensional variable (X, Y) will be given by

$$H(X, Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p\big(x_i, y_j\big)\, log\, p\big(x_i, y_j\big). \tag{12}$$

Certainly, the previous definition can be extended without any difficulty to any arbitrary number of random discrete variables.

In general, the following property is verified:

$$\mathrm{H}(\mathrm{X},\mathrm{Y}) \leq \mathrm{H}(\mathrm{X}) + \mathrm{H}(\mathrm{\ Y}). \tag{13}$$

The equality is verified if the random variables are independent.

3.3 Conditional Entropy

We represent by $p(y_j) = P(Y = y_j)$ the different values of a posteriori probability function of the random variable X and we represent by $H(Y = y_j)$ the entropy of probability function, that is to say,

$$H(X|Y = y_j) = -\sum_{i=1}^{n} p(x_i|y_j) \log p(x_i|y_j). \tag{14}$$

Thus, for each value of y_j the random variable Y, there is an entropy $H(Y = y_j)$ for which such entropy can be considered as a function of the random variable Y, being its average value:

$$\sum_{j=1}^{m} p(y_j)\, H(X|Y = y_j). \tag{15}$$

Said value receives the name of conditional entropy

Definition. The conditional entropy of the random variable X, given Y, is defined as:

$$H(X|Y) = \sum_{j=1}^{m} p(y_j) H(X|Y = y_j). \tag{16}$$

Which can also be written as:

$$\mathrm{H}(X|Y) = -\sum_{j=1}^{m}\sum_{i=1}^{n} p(x_i, y_j) \log p(x_i|y_j). \tag{17}$$

Indeed,

$$\begin{aligned} &H(X|Y) \\ &= \sum_{j=1}^{m} p(y_j) H(X|Y = y_j) \\ &= -\sum_{j=1}^{m} p(y_j) \sum_{i=1}^{n} p(x_i|y_j) \log p(x_i|y_j) \end{aligned} \tag{18}$$

$$\begin{aligned} &-\sum_{j=1}^{m}\sum_{i=1}^{n} p(y_j) p(x_i|y_j) \log p(x_i|y_j) \\ &= -\sum_{j=1}^{m}\sum_{i=1}^{n} p(x_i, y_j) \log p(x_i|y_j). \end{aligned} \tag{19}$$

An important property is the following:

$$H(X,Y) = H(X) + H(X) = H(Y) + H(Y). \tag{20}$$

An immediate consequence of the previous property is:

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z) \tag{21}$$

The gain of information I_G is defined as the difference between the entropy $H(X)$ and $H(Y)$, such that,

$$I_G = H(X) - H(X|Y) \tag{22}$$

The gain of information is always positive.

Also, the relative gain of information can be defined as:

$$RI_G = \frac{I_G}{H(X)}. \tag{23}$$

4 Methodology

The learning of decision tree employs the strategy of “divide and conquer”, through a search that has as objective to identify the optimal division points within the tree. This process of division repeats itself recursively from top to bottom until all or the majority of registers have been classified, with labels of specific classes.

The data that is available to realize the study integrate a sample formed by 519 registers and each one of them has 16 features that were retrieved through questionnaires directly from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh, by M. M. Faniqul Islam, Rahatara Ferdousi, Sadikur Rahman and Humayra Yasmin Bushra.

For developing the classification model, the total dataset was divided in a training set, commonly two thirds of the total de data. These data are labeled, in this case the label id the feature “class”, whose values can be positive or negative, and in a test set which is used to assess the model,

Table 1. Sample registers of database

Dataset	Age	Gender	Polyuria	Polydipsia	Class	Obesity	Alopecia	Weakness	Polyphagia	Partial paresis	Visual blurring	Irritability	Genital thrush	Delayed healing	Muscle stiffness	Sudden weight loss	Itching
REG_1	40	Male	No	Yes	Positive	Yes	Yes	Yes	No	No	No	No	No	Yes	Yes	No	Yes
REG_2	58	Male	No	No	Positive	No	Yes	Yes	No	Yes	Yes	No	No	No	No	No	No
REG_3	41	Male	Yes	No	Positive	No	Yes	Yes	Yes	No	No	No	No	Yes	Yes	No	Yes
REG_4	45	Male	No	No	Positive	No	No	Yes	Yes	No	No	No	Yes	Yes	No	Yes	Yes
REG_5	60	Male	Yes	Yes	Positive	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes

this is made up by a third of the total data, which is the complement of the training set. The model developed from a training dataset is used to classify the test dataset.

Once the model is developed from the training set, we proceed to classify the test dataset, subsequently the test data labels are compared with the generated predictions by the model and to do so, the percentage of classification is calculated. If the precision of this percentage is right, then it is possible to use the model to predict the class of cases.

The data of the study is represented through 16 dichotomous variables, related to diabetes. Since the variables are dichotomous, decision trees are considered which are uses in these cases for task classification. With these hypotheses, the algorithm I3 is applied, with the objective to determine the degree of accuracy with which the algorithm can discriminate the data into positives and negatives, such that the algorithm indicates with certain security, if with the obtained answers, can be determined if any person suffers or not from diabetes. In case that the algorithm does not deliver good results or leave data without classification, a technique of pruning the branches of the tree.

For this study, the training dataset of 346 registers randomly selected represents two third part of the total of registers, on the other hand, the test dataset conformed by the 173 remaining registers, represent a third part of the total of registers.

To simplify the analysis and the interpretation of the age variable, the values were grouped into two groups, the first group includes the ages between 25 and 50 years and the second group includes ages of more than 50 years, with this the variable age is dichotomous. Table 1 shows the first five registers of the database.

4.1 Model Development

Initially, the entropy of classes is calculated. In this case, as shown in Table 2, only two classes are considered: negatives (C1) and positives (C2). The formula for entropy is the following:

$$H(X) = -\sum_{i=1}^{2} p(x_i) log\, p(x_i). \qquad (24)$$

It is obtained the following values.

A calculus of conditional entropy is realized for the first variable. Applying the formula,

$$\mathrm{H(Y)} = -\sum_{\mathrm{j}=0}^{1} \mathrm{p}(\mathrm{y_j}) \sum_{\mathrm{i}=1}^{2} \mathrm{p}(\mathrm{y_j}) \log \mathrm{p}(\mathrm{y_j}). \qquad (25)$$

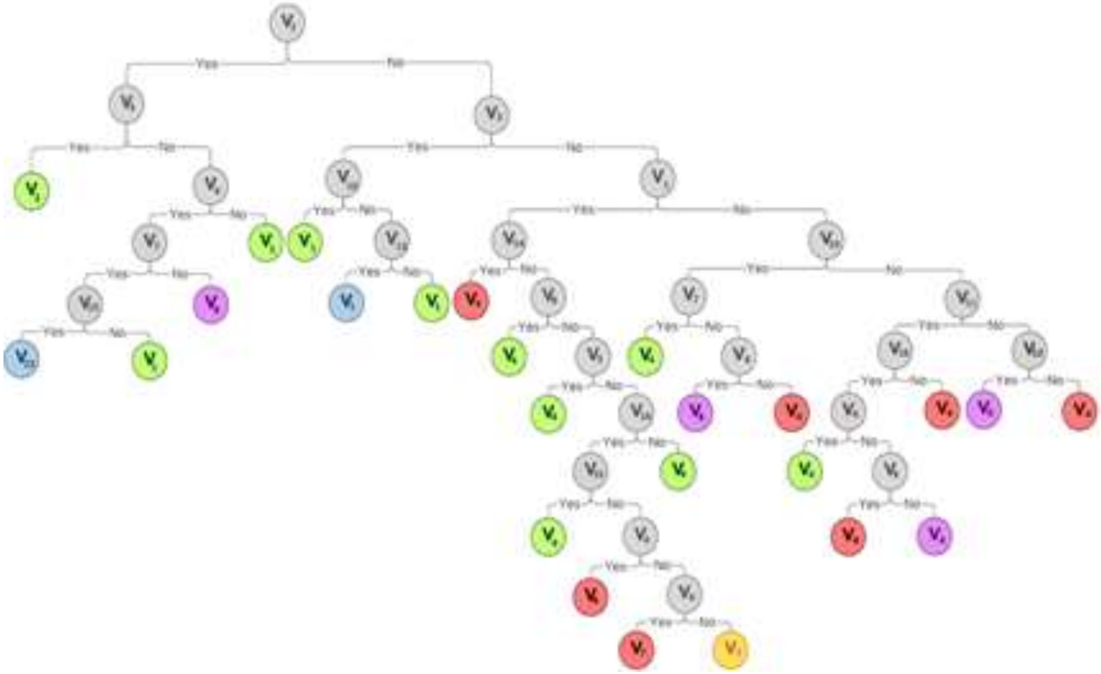
And making "$No = 1$", "$Yes = 0$" we have,

$$\begin{aligned} \mathrm{H(A_1)} = -\mathrm{p(A_1} = 1)[\mathrm{p(A_1} = 1) + \mathrm{p(A_1} = 1)] \\ - \mathrm{p(A_1} \\ = 0)[\mathrm{p(A_1} = 0) \\ + \mathrm{p(A_1} = 0)] \\ = -\frac{218}{346}\left[\frac{122}{218}\log\left(\frac{122}{218}\right) + \frac{96}{218}\log\left(\frac{96}{218}\right)\right] \\ - \frac{128}{346}\left[\frac{16}{128}\log\left(\frac{16}{128}\right)\right. \\ \left. + \frac{112}{128}\log\left(\frac{112}{128}\right)\right] = 0.82. \end{aligned} \qquad (26)$$

Calculus of the conditional entropy for the second variable.

Table 2. Classes considered for the model

Class	Elements	$p(x_i)$	- $p(x_i)$*(log $p(x_i)$)
Negative (C1)	138	0.399	0.529
Positive (C2)	208	0.601	0.441
TOTAL	346	1	0.970



$$\begin{aligned}\mathrm{H}(\mathrm{A}_2) = -\mathrm{p}(\mathrm{A}_2 = 1)[\mathrm{p}(\mathrm{A}_2 = 1) \\ + \mathrm{p}(\mathrm{A}_2 = 1)] \\ - \mathrm{p}(\mathrm{A}_2 \\ = 0)[\mathrm{p}(\mathrm{A}_2 = 0) \\ + \mathrm{p}(\mathrm{A}_2 = 0)],\end{aligned} \tag{27}$$

$$\begin{aligned}= -\frac{176}{346}\left[\frac{128}{176}\log\left(\frac{128}{176}\right)\right. \\ \left. + \frac{48}{176}\log\left(\frac{48}{176}\right)\right] \\ - \frac{170}{346}\left[\frac{10}{170}\log\left(\frac{10}{170}\right)\right. \\ \left. + \frac{160}{170}\log\left(\frac{160}{170}\right)\right] \\ = 0.59.\end{aligned}$$

Calculus of the conditional entropy for the third variable.

$$\begin{aligned}H(A_3) = -p(A_3 = 1)[p(A_3 = 1) \\ + p(A_3 = 1)] \\ - p(A_3 \\ = 0)[p(A_3 = 0) \\ + p(A_3 = 0)] \\ = -\frac{198}{346}\left[\frac{132}{198}log\left(\frac{132}{198}\right)\right. \\ \left. + \frac{66}{198}log\left(\frac{66}{198}\right)\right] \\ - \frac{148}{346}\left[\frac{6}{148}log\left(\frac{6}{148}\right)\right. \\ \left. + \frac{142}{148}log\left(\frac{142}{148}\right)\right] \\ = 0.63.\end{aligned} \tag{28}$$

Realizing the calculus for each variable, we arrive to the calculus of the conditional entropy for the last variable.

$$\mathrm{H}(\mathrm{A}_{15}) = -\mathrm{p}(\mathrm{A}_{15} = 1)[\mathrm{p}(\mathrm{A}_{15} = 1) + \mathrm{p}(\mathrm{A}_{15} = 1)] - \mathrm{p}(\mathrm{A}_{15} = 0)[\mathrm{p}(\mathrm{A}_{15} = 0) + \mathrm{p}(\mathrm{A}_{15} = 0)]$$

$$= -\frac{289}{346}\left[\frac{121}{289}\log\left(\frac{121}{289}\right) + \frac{168}{289}\log\left(\frac{168}{289}\right)\right] - \frac{57}{346}\left[\frac{17}{57}\log\left(\frac{17}{57}\right) + \frac{40}{57}\log\left(\frac{40}{57}\right)\right] = 0.96. \quad (29)$$

The gain of information I_G for each one of the variables $A_1, A_2, \dots, A_{15}$ is the following:

$$\begin{aligned} \mathrm{I}_{\mathrm{GA}_1} &= \mathrm{H(C)} - \mathrm{H(A1)} = 0.970 - 0.82 = 0.15, \\ \mathrm{I}_{\mathrm{GA}_2} &= \mathrm{H(C)} - \mathrm{H(A2)} = 0.970 - 0.59 = 0.38, \\ \mathrm{I}_{\mathrm{GA}_3} &= \mathrm{H(C)} - \mathrm{H(A3)} = 0.970 - 0.63 = 0.34, \\ &\dots\dots\dots. \\ \mathrm{I}_{\mathrm{GA}_{15}} &= \mathrm{H(C)} - \mathrm{H(A4)} = 0.970 - 0.96 = 0.01. \end{aligned} \quad (30)$$

The relative gain of information for each one of the variables $A_1, A_2, \dots, A_{15}$ is:

$$\begin{aligned} \mathrm{RI}_{\mathrm{GA}_1} &= \frac{\mathrm{I_G}}{\mathrm{H(X)}} = \frac{0.15}{0.97} = 0.15, \\ \mathrm{RI}_{\mathrm{GA}_2} &= \frac{\mathrm{I_G}}{\mathrm{H(X)}} = \frac{0.38}{0.97} = 0.39, \\ \mathrm{RI}_{\mathrm{GA}_3} &= \frac{\mathrm{I_G}}{\mathrm{H(X)}} = \frac{0.34}{0.97} = 0.35, \\ &\dots\dots \\ \mathrm{RI}_{\mathrm{GA}_{15}} &= \frac{\mathrm{I_G}}{\mathrm{H(X)}} = \frac{0.01}{0.97} = 0.01. \end{aligned} \quad (31)$$

From results obtained, we can deduce that the variable A_2 is the one with greater relative gain of information, which indicates that we must choose this variable to realize a new division, that is to say, the base is divided in two parts; the first part considers answers with value “yes” and second part considers answers with value “no”, regarding the variable A_2. This process continues until finishing with all the possible divisions and by obtaining a decision tree.

Table 3. Sample registers of database

V_1 = Gender	V_2 = Polyuria	V_3 = Polydipsia	V_4 = Sudden weight loss
V_5 = Weakness	V_6 = Polyphagia	V_7 = Genital thrush	V_8 = Visual blurring
V_9 = Itching	V_{10} = Irritability	V_{11} = Delayed healing	V_{12} = Partial paresis
V_{13} = Muscle stiffness	V_{14} = Alopecia	V_{15} = Obesity	V_{16} = Age
V_1 = Gender	V_2 = Polyuria	V_3 = Polydipsia	V_4 = Sudden weight loss

Subsequently, a script made in Python was implemented for the creation of the decision tree, said tree achieved to classify all the individuals, except one register. For this reason, it is not necessary to apply some method of pruning to avoid overfitting and improve the generalization of the model.

Table 3 shows the nomenclature used to name the variables, and the same nomenclature was used in the decision tree shown in Figure 2.

Once built the model from the training dataset, this is used to realize the classification of the test dataset, which consists of the 173 remaining registers. The real labels of test data are compared with the labels predicted by the model, with the purpose of calculating the percentage of coincidences. If the precision of the percentage of classification is high, then it is concluded that the model can be used to classify new data.

The classification obtained to classify new data is the following:

The green nodes represent the positive class, independently from the value observed by the variable. The blue nodes represent the positive class, as long as the value of the variable is positive.

The red nodes represent the negative class regardless the value of the variable. The purple nodes represent the class that is opposite to the value of the variable.

The unique case for which the system cannot give a classification is the orange node.

Table 4. Classification results

Register	Age	Gender	Polyuria	Polydips	Delayed healing	Muscle stiffness	Partial paresis	Genital thrush	Visual blurring	Weakness	Polyphagia	Itching	Irritability	Alopecia	Sudden weight loss	Obesity	Class	Result
1	40	Male	No	Yes	Yes	Yes	No	No	No	Yes	No	Yes	No	Yes	No	Yes	Positive	Negative
2	58	Male	No	No	No	No	Yes	No	Yes	Yes	No	No	No	Yes	No	No	Positive	Positive
3	41	Male	Yes	No	Yes	Yes	No	No	No	Yes	Yes	Yes	No	Yes	No	No	Positive	Negative
5	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive	Positive
10	70	Male	No	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Positive	Positive
15	60	Male	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	No	No	No	No	Positive	Positive
16	58	Male	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	No	No	No	No	Positive	Positive
17	54	Male	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Yes	No	Positive	Positive
21	62	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No	No	No	Positive	Positive
22	54	Male	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Positive	Positive

Table 5. Confusion matrix

	Positive	Negative
Positive	TP: 104	FP: 8
Negative	FN: 7	TN: 54

5 Results and Discussion

With the application of the developed model, the following results were obtained. From the 173 cases of the test set, the system coincided with the expected result in 158 cases, with a difference in only 15 cases. This represents a percentage of 91.3%. Given the high level of precision, the developed model can be considered appropriate to classify new cases with a reliability of 90%. Some of these results are shown in Table 4.

Now, specificity is assessed, that is to say, the capacity of the model to correctly identify the real negatives, for this case, the people without diabetes. Namely, the proportion of healthy people that the model correctly classifies as negatives. For this, the data is taken from Table 5. According to the results, an inverse relationship between true negatives and false positives can be seen.

Now, some evaluation metrics are applied, such as accuracy, sensitivity, specificity, and precision to determine the model's ability to detect diabetes.

- True Positive (TP): The number of positive cases that were correctly identified as positive.
- True Negative (TN): The number of negative cases that were correctly identified as negative.
- False Positive (FP): The number of negative cases that were incorrectly identified as positive.
- False Negative (FN): The number of positive cases that were incorrectly identified as negative.

Accuracy is used to evaluate the performance of any classifier based on correctly predicted instances versus the total number of instances and is calculated by the formula:

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} = 0.91. \qquad (32)$$

Precision calculates the ratio of true positives to all positive predictions (both true and false positives):

$$\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) = 0.93. \qquad (33)$$

Sensitivity, also known as Recall, calculates the ratio of true positives to all actual positive cases:

$$\mathrm{Recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) = 0.94. \tag{34}$$

Specificity calculates the ratio of true negatives to all actual negative cases. It measures the proportion of people without the disease who correctly receive a negative test result:

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = \frac{54}{62} = 0.87. \tag{35}$$

The metric of specificity gives a result of 87%, which shows that is a study in which many true negatives and very few false positives exist. Therefore, this proofs that the developed algorithm for this study is an efficient method of prediction of this disease.

5.1 Calculating the Standard Error Using the Bootstrap Method

The traditional approach to statistical inference is based on idealized models and assumptions. Often, expressions for measures of precision, such as the standard error, are based on asymptotic theory. A modern alternative to the traditional approach is the bootstrap method, introduced by Efron (1979), see [25] and [26]. The bootstrap algorithm for estimating the standard error of an estimator $\hat{\theta} = s(X)$ of parameter θ can be carried out through the following steps:

From the initial sample $X_1, X_2, \dots, X_n$, B independent bootstrap samples are generated,

$$\begin{aligned} \mathrm{X}^{*(1)} &\sim \mathrm{X}_1^{*(1)}, \dots, \mathrm{X}_n^{*(1)}, \\ \mathrm{X}^{*(2)} &\sim \mathrm{X}_1^{*(2)}, \dots, \mathrm{X}_n^{*(2)}, \\ &\dots \\ \mathrm{X}^{*(\mathrm{B})} &\sim \mathrm{X}_1^{*(\mathrm{B})}, \dots, \mathrm{X}_n^{*(\mathrm{B})}. \end{aligned} \tag{36}$$

Evaluate

$$\hat{\theta}^{*(\mathrm{b})} = \mathrm{s}\left(\mathrm{X}^{*(\mathrm{b})}\right); \quad \mathrm{b} = 1, \dots, \mathrm{B}. \tag{37}$$

The standard error $se(\hat{\theta})$ is estimated by the standard deviation of the B replicates.

$$\widehat{se}_{boot}(\hat{\theta}) = \left[\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{*(b)} - \hat{\theta}^{*(\cdot)}\right)^2\right]^{1/2} \tag{38}$$

Where

Metric	Result
Mean Accuracy	0.9390
Standard Deviation	0.0189
Maximum Accuracy	0.9936
Minimum Accuracy	0.8782.

6 Conclusions and Further Research

In this work, a system was developed with the main objective of creating a machine learning-based model to detect whether a person has diabetes. In the testing phase, the system achieved a high prediction accuracy of 92%, which allows us to use it for predicting new cases or at least as a tool to warn if confirmation with a specialist is necessary. Furthermore, the bootstrap algorithm gives us a slightly higher average and a very small standard deviation of 0.01, assuring us that the 92% prediction has very little bias, making it highly reliable.

The work achieves very positive results that encourage the use of artificial intelligence techniques for disease detection, as in this case with diabetes, thus providing a wider variety of available alternatives. It is worth mentioning that the problem can be addressed using other supervised machine learning techniques, such as neural networks, although we believe a similar accuracy to that found in this work can be achieved.

For future work, an exhaustive analysis of the database can be performed. After running the system several times, we noticed there are records the algorithm could not classify correctly. For example, some records have negative answers to all questions; however, some of these individuals have diabetes while others with the same answers do not. Therefore, we are confident that by eliminating these types of ambiguous records, it is possible to increase the system's prediction percentage, and the system could be designed to abstain from providing a prediction for those who answer all questions negatively. With this increased accuracy, we believe it would be difficult to find another technique that surpasses this prediction percentage. We will verify these claims in future work by comparing these results with a neural network, for example.

Finally, this work achieves very positive results that encourage the use of artificial intelligence techniques for detecting diseases like diabetes, thus providing a wider variety of available alternatives for prevention and detection. We reiterate that the problem can be addressed using other machine learning techniques [27], [28], [29], although we believe a similar accuracy to that found in this work can be achieved, but it would be difficult to surpass it.

References

1. **Afroz, A., Alam, K., Ali, L. et al., (2019).** Type 2 diabetes mellitus in Bangladesh: a prevalence-based cost-of-illness study. BMC Health Services Research, Vol. 19, pp. 601. doi: doi.org/10.1186/s12913-019-4440-3
2. **Yadav, D. C., Pal, S. (2021).** An Experimental Study of Diversity of Diabetes Disease Features by Bagging and Boosting Ensemble Method with Rule Based Machine Learning Classifier Algorithms. SN Computer Science, Vol. 2, No. 50. doi: doi.org/10.1007/s42979-020-00446-y
3. **Cho, G., Yim, J., Choi, Y., Ko, J., Lee, S.-H. (2019).** Review of Machine Learning Algorithms for Diagnosing Mental Illness. Psychiatry Investigation, Vol. 16, No. 4, pp. 262–269. doi: doi.org/10.30773/pi.2018.12.21.2
4. **Gogebakan, K., Sah, M. (2021).** A Review of Recent Advances for Preventing, Diagnosis and Treatment of Diabetes Mellitus using Semantic Web. 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications pp. 1–6. doi: doi.org/10.1109/hora52670.2021.9461282
5. **Baier, L. J., Hanson, R. L. (2004).** Genetic Studies of the Etiology of Type 2 Diabetes in Pima Indians: Hunting for Pieces to a Complicated Puzzle. Diabetes, Vol. 53, No. 5, pp. 1181–1186. doi: doi.org/10.2337/diabetes.53.5.1181
6. **Islam, M. M. F., Ferdousi, R., Rahman, S., Bushra, H. Y. (2019).** Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. Computer Vision and Machine Intelligence in Medical Image Analysis, Vol. 992, pp. 113–125. doi: doi.org/10.1007/978-981-13-8798-2_12
7. **Sharma, P., Choudhary, K., Gupta, K., Chawla, R., Gupta, D. Sharma, A. (2020).** Artificial plant optimization algorithm to detect heart rate & presence of heart disease using mahine learning. Artificial Intelligence in Medicine, Vol. 102, 101752, doi: doi.org/10.1016/j.artmed.2019.101752
8. **Williams, R., Karurangab, S., Malandab, B., Saeedib, P., Basitc, A., Besançon, S. et al. (2020).** Global and regional estimates and projections of diabetes-related health expenditure: results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes Research and Clinical Practice, Vol. 162, pp. 108072. doi: doi.org/10.1016/j.diabres.2020.108072
9. **Ghosh, S., Collier. A. (2012).** Diabetes (2nd ed). Churchill Livingstone/Elsevier.
10. **Le, T. M. , Vo, T. M., Pham, T. N. Dao, S. V. T. (2021).** A Novel Wrapper–Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic. IEEE Access, Vol. 9, pp. 7869–7884. doi: doi.org/10.1109/access.2020.3047942
11. **Kuo, K.-M., Talley, P., Kao, Y. Huang, C. H. (2020).** A multi-class classification model for supporting the diagnosis of type II diabetes mellitus. PeerJ, 8, e9920. doi: doi.org/10.7717/peerj.9920
12. **Abbas, H. T., Alic, L., Erraguntla, M., Ji, J.X., Abdul-Ghani, M., et al. (2019).** Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test. PLOS ONE, Vol. 14, No. 12, e0219636. doi: doi.org/10.1371/journal.pone.0219636
13. **Yu, W., Liu, T., Valdez, R., Gwinn, M. Khoury, M. J. (2010).** Application of support vector machine modeling for prediction of

common diseases: the case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making, Vol. 10, No. 16. doi: doi.org/10.1186/1472-6947-10-16

14. **Hasan, M. K., Alam, M. A., Das, D., Hossain, E. Hasan, M. (2020).** Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access, Vol. 8, pp. 76516–76531. doi: doi.org/10.1109/access. 2020.2989857
15. **Gil-Vera, V. D., Quintero-López, C. (2021).** Predicción del rendimiento académico estudiantil con redes neuronales artificiales. Información tecnológica, 32(6), 221–228. https://doi.org/10.4067/s0718-07642021000 600221
16. **Quinlan, J. R. (1986)**. Induction of decision trees. Machine Learning, 1(1), 81-106. http://dx.doi.org/10.1007/BF00116251
17. **Quinlan, J. R. (1993)**. C4.5: Programs for Machine Learning. Morgan Kaufmann. https://books.google.com/books/about/C4_5.html?
18. **Aracena, C., López, M., Rojas, P. (2022).** Aplicaciones de aprendizaje automático en salud. Revista Médica Clínica Las Condes, 33(2), 100–110. https://doi.org/10.1016/j.rmclc.2022.10.001
19. **Zhang, J. (2023)**. Early detection of type 2 diabetes risk: limitations of current diagnostic criteria and new approaches. Journal of Diabetes Research, 2023, Article ID 10665905. https://doi.org/10.1155/2023/10665905
20. **Moraes Morelli, D., Rubinstein, F. A., Santero, M., Gibbons, L., Moyano, D. L., Nejamis, A., Beratarrechea, A. G. (2023)**. Effectiveness of a diabetes program based on digital health on capacity building and quality of care in type 2 diabetes: a pragmatic quasi-experimental study. BMC Health Services Research, 23(1), Article 123. doi: doi.org/10.1186/s12913-023-09082-7
21. **Solé, R. S. I. (2025)**. Clasificación: árboles de decisión. Universitat Oberta de Catalunya.
22. **Aparicio-Montenegro, P. R. (2025)**. El abordaje de la diabetes mediante la Inteligencia Artificial. Open Journal. doi: https://doi.org/10.5281/zenodo.15565315
23. **Griva, L., Basualdo, M. (2018)**. Clasificación de pacientes con diabetes mellitus tipo 1 mediante técnicas de árbol de decisión. IX Congreso Argentino de Informática y Salud (CAIS), Jornadas Argentinas de Informática e Investigación Operativa (JAIIO), pp. 49–62.
24. **Cantú, A. G. (2019).** Árboles de decisión y su aplicación en el síndrome metabólico. Tesis de maestría, Centro de Investigación en Matemáticas, CIMAT. Repositorio CIMAT. https://cimat.repositorioinstitucional.mx/jspui/bitstream/1008/1013/1/MTY%20TE%203.pdf
25. **Efron, B. (1979).** Bootstrap methods: another look at the jackknife. Annals of Statistics 7, 1–26.
26. **Efron, B., Tibshirani, R. J. (1993).** An Introduction to the Bootstrap. Chapman & Hall.
27. **Jagannathan, R., Neves, J. S., Dorcely, B., Chung, S. T., Tamura, K., Rhee, M., Bergman, M. (2020).** The Oral Glucose Tolerance Test: 100 Years Later. Dove Medical Press, 13, 3787–3805.
28. **Sahoo, A.K., Pradhan, C., Das, H. (2020).** Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making. Nature inspired computing for data science, 871, 201–212. Cham: Springer International Publishing.
29. **Singh, N., Singh, P. A. (2020).** Stacked generalization approach for diagnosis and prediction of type 2 diabetes mellitus. Advances in intelligent systems and computing, pp. 559–570, Springer.

Article received on 10/06/2025; accepted on 02/10/2025.
**Corresponding author is Carmen Cerón Garnica.*