www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 17(2), 2026, 312-323. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2026.v17i2.1190

Web platform for diabetes prediction using weighted machine learning techniques based on personal and clinical indicators

Viridiana Barrera Soto, Juan Carlos Huerta Mendoza, José Lázaro Martínez
Unidad Académica Multidisciplinaria Reynosa Rodhe – Universidad Autónoma de Tamaulipas.
E-mails: a2213728001@alumnos.uat.edu.mx, jchuerta@docentes.uat.edu.mx, lazaro.martinez@uat.edu.mx

Abstract. Diabetes is a chronic metabolic disease characterized by elevated levels of glucose in the blood (or blood sugar), which over time leads to severe damage to the heart, blood vessels, eyes, kidneys, and nerves. The most common type is type 2 diabetes, usually in adults, which occurs when the body becomes resistant to insulin or does not produce enough insulin. By using artificial intelligence (AI) techniques in complex problems such as disease diagnosis, a degree of certainty in the results has been achieved to identify a specific type of disease. These applications have been advantageous because large amounts of patient data can be analyzed to find patterns. This work proposes a platform for the prediction of type 2 diabetes based on clinical or personal indicators. To do this, two supervised classification models were constructed using the PIMA Indian Diabetes dataset and the Centers for Disease Control and Prevention (CDC) dataset, integrating both into a web platform for prediction with new data to support the decisions of doctors and healthcare professionals. By integrating different algorithms into the final predictive model through voting weighting, the accuracy percentage in prediction has been increased.

Keywords: Artificial intelligence, predictive models, machine learning, diabetes.

Article Info
Received July 30, 2025
Accepted Dec 11, 2025

1 Introduction

Diabetes is a disease characterized by high blood glucose levels, either because the pancreas does not produce enough insulin (type 1 diabetes) or the body does not properly use the insulin it produces (type 2 diabetes). High blood glucose levels over a prolonged period of time can cause problems in various organs such as the heart, eyes, and kidneys (World Health Organization, 2019). The diagnosis of diabetes is based on the observation of fasting blood glucose above 1.26 g/l; this must be demonstrated after two observations, but a single figure above 2 g/l is enough to determine the diagnosis (Rigalleau, et al., 2019).

In the healthcare sector, large volumes of data are generated daily through the use of medical records and patient monitoring platforms. In the case of diabetes management, various platforms have been proposed to carry out continuous monitoring and facilitate communication between patient and doctor through the use of mobile technologies based on the Internet of Things (Shan et al., 2019). Artificial Intelligence (AI) techniques applied to disease diagnosis have been used in studies of complex problems, achieving a degree of certainty in the results obtained in the identification of a specific type of disease. These applications have been advantageous because these large amounts of patient data can be leveraged to analyze them and find patterns, aiding in prevention and diagnosis, thus providing better medical care in general (Ávila-Tomás et al., 2021).

In this way, it is possible to apply disease prediction techniques in hospitals and healthcare systems to facilitate diagnosis using various key indicators collected through surveys or clinical records. It is worth mentioning that blood sample analysis is required to obtain some of the indicators listed below (Barzallo and Barzallo, 2019):

- Plasma glucose concentration.
- Blood pressure (diastolic).
- Triceps skinfold thickness.
- 2-hour serum insulin.
- Diabetes pedigree function.

This study focuses on developing a predictive model for the diagnosis of type 2 diabetes based on lifestyle indicators such as diet, physical activity, body mass index, among others, without requiring specialized analysis or medical help of any kind. Prediction is also studied using clinical indicators from the CDC (Teboul, 2021) and PIMA (UCI MACHINE LEARNING, 2016) data sets such as blood glucose level, blood pressure, among others. With both prediction methods, a web platform is proposed integrating different machine learning techniques such as data mining for analysis and pre-processing, thus increasing the quality of the data, in addition to data balancing, that is, obtaining a uniform distribution between records of diabetic and non-diabetic people, in this way better results can be obtained by the different prediction algorithms integrated by means of a weighting of votes (Batista et al., 2004), all in architectures that can be of the ensemble type and available to the general population, which is useful as a first approach in the diagnosis of said disease.

2 Experimental procedures

Figure 1 shows the steps to follow for the construction of the diabetes predictive model.

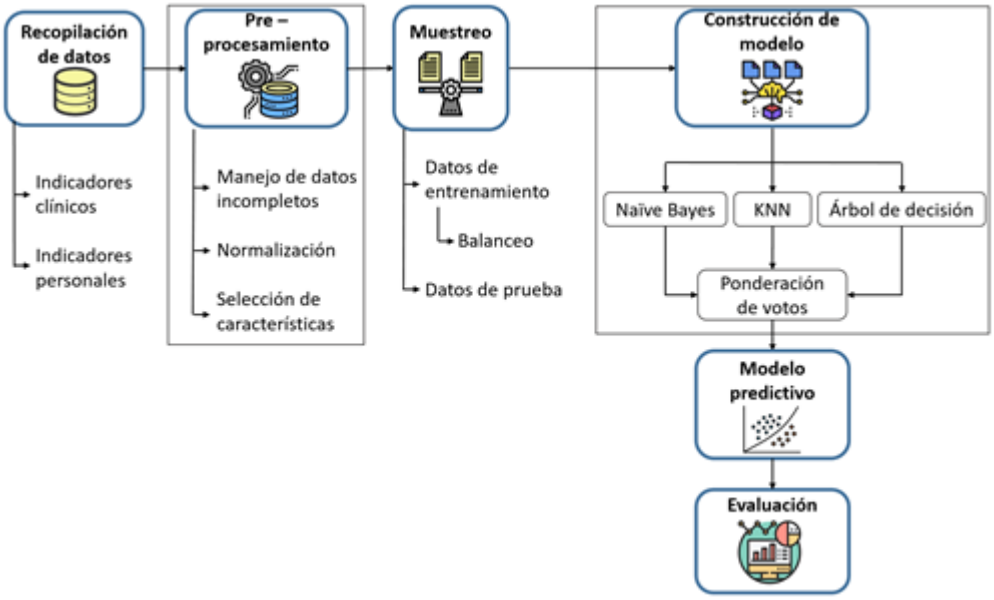

Fig. 1. Stages of building a diabetes predictive model.

2.1 Data collection

It is necessary to collect data on clinical and personal indicators. Two data entry options will be available to perform the prediction, achieving greater reach among the population. The data collected for this study are described below for both types of indicators.

Clinical indicators.

- Comes from the "Pima Indian Diabetes Dataset" compiled by the National Institute of Diabetes and Digestive and Kidney Diseases (UCI MACHINE LEARNING, 2016).
- Consists of 768 samples.
- Female patients at least 21 years old and of Pima Indian heritage.

Personal indicators.

- The dataset used was collected through a telephone survey by the U.S. Centers for Disease Control and Prevention (CDC) (Teboul, 2021).
- It consists of 253,680 samples of men and women aged 18 and older.

2.2 Preprocessing

The data is prepared and made suitable for training a prediction model. The procedures to be applied depend on the conditions of the data set. The tasks typically applied are the following:

- Incomplete values. These values must be completed or filled in using various techniques, assigning the mean or median of the feature, predicting these values using regression or classification techniques, or discarding the samples. The technique will depend on the number of missing values, the type of values, and the state of the data set.
- Outliers. Data are normalized by detecting outliers using data visualization techniques such as box plots and keeping outliers within the range of the 25th and 75th percentiles.
- Feature selection. Based on the correlation graph of the variables and the chi-square test, because this test analyzes categorical data and indicates the independence of the variables with the label, the best features that help predict the response label are selected from here, in this case, whether the sample corresponds to a diabetic or non-diabetic person.

At this stage, incomplete data or outliers that would affect the predictive capacity of the final model are handled. In addition, a feature selection is performed based on a correlation matrix and using the chi-square test to observe the relationship between the features. The preprocessing techniques performed on each data set are described below. It should be noted that these differ due to the state of the data.

Preprocessing of personal indicators: This data set does not contain incomplete values because it was previously cleaned and, being categorical characteristics, it does not have outliers. The preprocessing performed on this data set is described below.

- **Data type change.** The values in this data set are floats; however, it does not contain decimal values (with the exception of the "BMI" feature), so the data types are changed to integers, leaving the "BMI" feature in its original float data type.
- **Duplicate samples.** Duplicate samples bias the trained model because identical records may be present in both the training and test sets. Therefore, 23,899 samples were eliminated, leaving a total of 229,781 records.
- **Feature selection.** To select the features most closely related to the class label, the chi-square test is performed, which yields a score for each feature, as shown in Table 1.

Table 1. Chi-square test of personal indicators

Feature	Score	Feature	Score
PhysHlth	103705.813420	HvyAlcoholConsump	973.676965
BMI	16667.006176	PhysActivity	656.187604
MentHlth	13499.056578	Education	537.068918
Age	9400.408945	Smoker	268.924145
HighBP	8633.562752	NoDocbcCost	163.138397
DiffWalk	8310.280494	Sex	136.998866
GenHlth	8142.602678	Veggies	89.554915
HeartDiseaseorAttack	6015.661944	Fruits	57.666728
HighChol	5381.985473	CholCheck	54.155676
Income	3748.254712	AnyHealthcare	7.883473
Stroke	2212.060848		

According to the results of the chi-square test, the four features with significantly lower scores than the rest (Veggies, Fruits, CholCheck, AnyHealthcare) are discarded. Additionally, three features (Income, Education, NoDocbcCost) are discarded because they are closely related to the sample collection area. The DiffWalk feature is also discarded because it is considered redundant with the PhysHlth property, thus obtaining 13 final features.

Pre-processing of clinical indicators: For this dataset, additional cleaning techniques are required to handle null or incomplete values. The preprocessing is described below.

- **Null values.** To detect these values, the “0” is replaced by “NaN” in the features “Glucose”, “BloodPressure”, “SkinThickness”, “Insulin” and “BMI”, because they represent measurements resulting from blood tests that by their nature cannot result in “0”. In the Table 3 shows the number of null values for each feature in the data set.

Table 3. Null values of clinical indicators.

Feature	Null quantity
Age	0
Pregnancies	0
Body mass index	11
Diastolic blood pressure	35
Plasma glucose concentration	5
2-hour serum insulin	374
Diabetes Pedigree Function	0
Triceps skinfold thickness	227

It's worth mentioning that most of the samples with missing insulin also have null values for the other characteristics. After removing all the samples with null values, we have a total of 394 samples.

- **Feature selection.** In the same way, the chi-square test is performed to evaluate the relationship of the characteristics with the class label, the scores obtained are shown in Table 4.

Table 4. Chi-square test of clinical indicators.

Feature	Score
2-hour serum insulin	3151.803403
Plasma glucose concentration	836.037745
Age	169.002043
Triceps skinfold thickness	99.320379
Pregnancies	82.598850
Body mass index	42.847984
Diastolic blood pressure	31.834948
Diabetes Pedigree Function	3.971641

With the results obtained, the diabetes pedigree function is discarded due to its low correlation with the class label. The number of pregnancies is also discarded since this prediction is intended to be applied to the entire adult population. Finally, the triceps skinfold thickness is discarded because this characteristic is used to indicate obesity. However, the body mass index is already available, which is easier to measure with height and weight. In this way, 5 final characteristics are obtained.

2.3 Sampling

After preparing the data, the dataset is split into two parts, for training and testing, in an 80/20 ratio, respectively. This split involves including the "random_state" parameter with any integer value. This results in identical datasets between experiments, allowing for comparison of the results of algorithms trained with the same samples. This is done in the same way for both datasets.

Additionally, an “Oversampling” type resampling is applied to the training set, in order to obtain the same number of negative and positive samples by increasing the number of samples from the minority class. To do this, the “SMOTEENN” technique from the “imblearn” library is used.

It is worth mentioning that, before starting to build the model, the training data must be balanced, that is, a process must be performed to obtain the same number of samples from diabetic and non-diabetic people. This is because the number of samples usually from the "non-diabetic" class tends to predominate and thanks to this process it helps prevent the model from leaning towards said class (it avoids biases). However, if too many synthetic samples are made, it can also affect the model's performance since they are created from existing samples.

2.4 Model construction

This stage begins with the training and evaluation of various prediction models (Extra Trees, Gradient Boosting, AdaBoost, KNN, Random Forest, XGB, Decision Tree, Naive Bayes and Neural Networks), in such a way that by adjusting parameters through meshes the best configuration and the highest percentage of accuracy are obtained for each algorithm, in order to select the algorithms with the best performance for each data set. Then, the selected and trained models are integrated through a weighted ensemble of votes, that is, the probability of belonging to the positive and negative class returned by the prediction of each individual algorithm is averaged, this strategy is observed in Figure 2. With this integration, the aim is to increase the percentage of accuracy compared to the individual models, it is worth mentioning that the number of models to be integrated is according to each data set since a greater number of models does not mean better performance, for this purpose the ensemble is evaluated with the two, three or four best models.

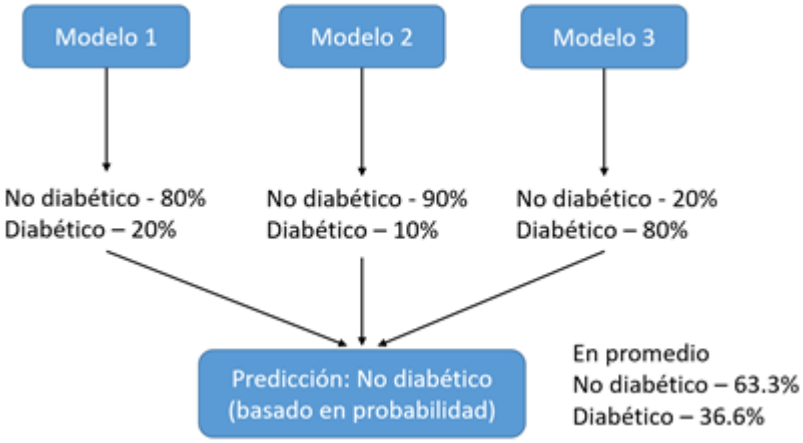

Fig. 2. Integration of algorithms.

Predictive model: To implement this model integration, the "VotingClassifier" function from the "scikit-learn" library is used, with which weights are assigned as parameters with the purpose of giving greater importance to the model with the best performance. Figure 3 shows its coding.

```
votingC = VotingClassifier(estimators=[('gb', gb),
('adab', adab),
('extraT', extraT)], voting='hard')
votingC = votingC.fit(X_train, Y_train)
y_predE=votingC.predict(X_test)
matrixensem = classification_report(Y_test,y_predE)
print("Voting")
print(matrixensem)
scores = cross_val_score(votingC, X_train, Y_train,
cv=5, scoring='accuracy')
print(scores)
print(scores.mean())
```

Fig. 3. VotingClassifier Encoding.

2.5 Web platform

One of the main features of the solution proposed in this research is an easy-to-use and interactive interface that maintains correct visualization on different devices. This interface allows users to enter the required data, depending on clinical or personal indicators, and obtain a prediction. Figure 4 shows the data flow between the server and the client.

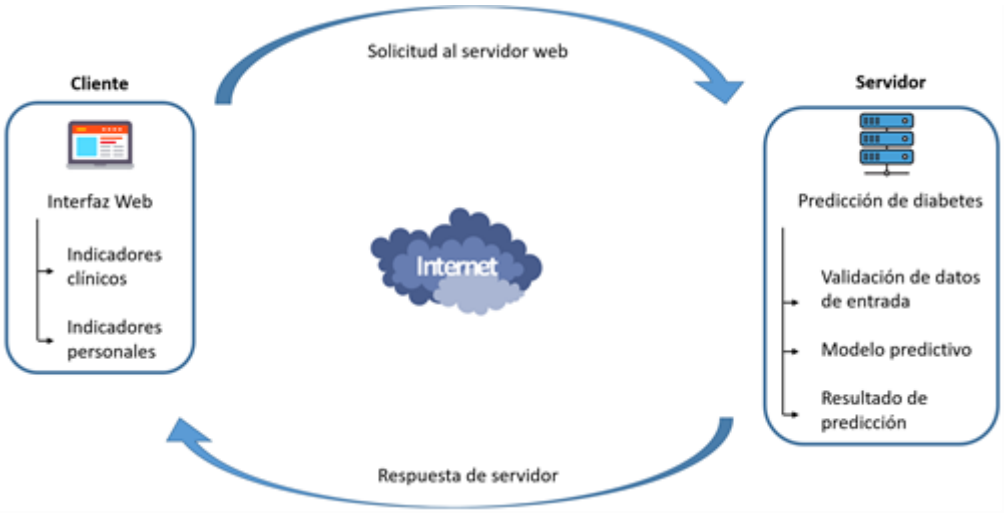

Fig. 4. Information flow between client and server.

Welcome interface: The interface contains information about diabetes and its types, also an introduction to the prediction platform, as shown in Figure 5.

Fig. 5. Welcome interface.

Personal Indicators Interface: The web platform has a specific interface for predicting diabetes using personal indicators. Information is requested from the user via drop-down menus, resulting in an intuitive questionnaire and numeric fields for age, weight, and height, as shown in Figure 6.

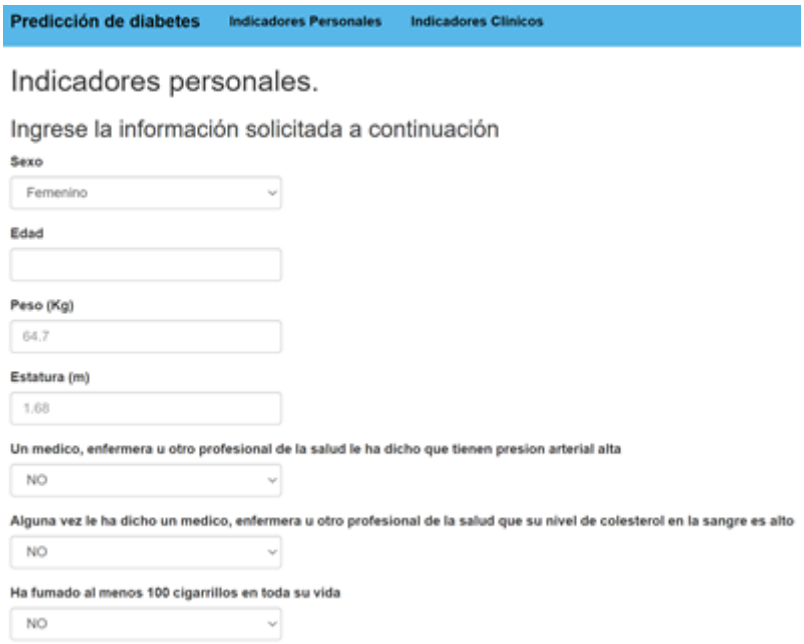

Fig. 6. Personal Indicators Interface.

Clinical Indicators Interface: In the same way, there is an interface to make the prediction with clinical indicators where all fields are numeric with a provisional text or "placeholder" to intuitively indicate to the user what type of information is required, as shown in Figure 7.

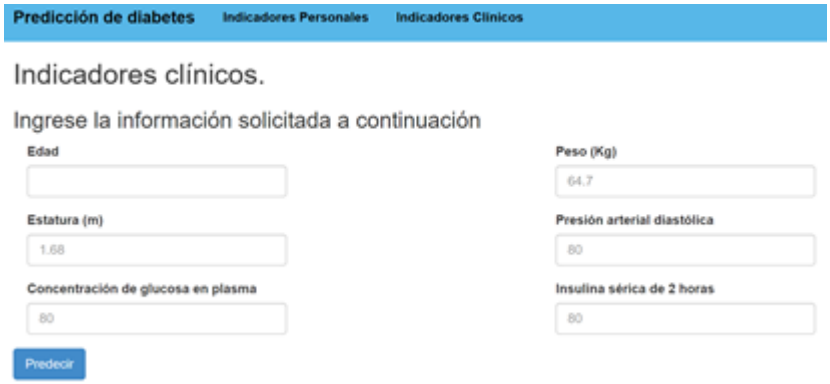

Fig. 7. Clinical Indicators Interface.

Prediction result interface: After entering the requested information for each type of indicators and pressing the "predict" button, the user is directed to a results interface where the highest probability (positive or negative) regarding the diagnosis of type 2 diabetes is presented, as shown in Figure 8.

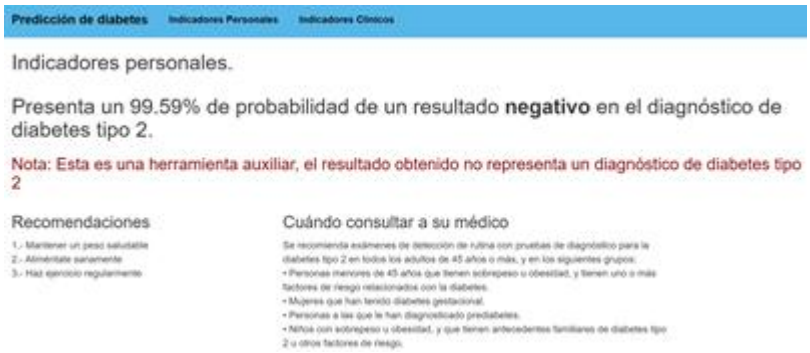

Fig. 8. Prediction result interface.

2.6 Testing with individual algorithms

After preparing the data sets, a selection of algorithms is made to be integrated into a final ensemble model. This model is constructed by integrating the selected algorithms through weighted voting. Nine different algorithms are tested to identify the best performing algorithms for each data set and use them in the final ensemble model.

Algorithm selection: This stage is of utmost importance since the algorithms with the best performance for each data set must be used. Therefore, tests are carried out with nine different algorithms (Extra Trees, Gradient Boosting, AdaBoost, KNN, Random Forest, XGB, Decision Tree, Naive Bayes and Neural Networks). For this purpose, a grid or list of configurable parameters is prepared according to each algorithm. The "GridSearchCV" function of the scikit-learn library is responsible for executing and selecting the best combination of parameters for each algorithm based on the accuracy metric by applying cross-validation. Once the best configuration is obtained, the area under the curve is measured, which indicates how well the model can distinguish between a positive and negative sample. With these results, the three best algorithms are selected based on accuracy and AUC metrics.

Tests with algorithm integration: Once the algorithms have been selected, the next step is integration into a final predictive model by means of a weighting of results. For personal indicators, the best algorithms were Gradient Boosting, XGB and Neural Network with accuracy results of 0.8541, 0.8542 and 0.8524 respectively, regarding clinical indicators, the best algorithms were Naive Bayes, Gradient Boosting and Random Forest with results of 0.8607, 0.8481 and 0.8101 respectively. To perform this integration, the "VotingClassifier" function of the "scikit-learn" library is used, weights are assigned in order to give greater importance to the model with the best performance, The Table 5 shows the weights assigned to each algorithm.

Table 5. Weights assigned in the integration of algorithms.

Feature	Score	
Personal indicators	Gradient Boosting	3
	XGB	2
	Red neuronal	1
Clinical indicators	Naive Bayes	3
	Gradient Boosting	2
	Random forest	1

Two final predictive models are obtained, one for personal indicators and one for clinical indicators. These models are capable of predicting a positive or negative outcome with a sample containing 13 characteristics for personal indicators and 5 values for clinical indicators.

3 Results

3.1 Performance of algorithms on personal data sets

The Table 6 presents the training and testing results for the 9 classification algorithms considered in the experiments. The training results show an accuracy rate of over 80% in all cases, with the model trained with the Adaboost algorithm standing out at 95%. In the testing stage, the performance in a real environment is observed since the data has never been seen by the predictive model. Based on the accuracy and area under the curve results, the algorithms that make up the final weighting model are selected. It can be observed that the GradientBoosting, XGB, and neural network algorithms stand out with an accuracy rate of 0.8541, 0.8542, and 0.8524, respectively.

Table 6. Training and testing results with personal indicators.

	Training		Test	
Algorithm	Accuracy	AUC	Accuracy	AUC
Gradient Boosting	0.8537	0.8228	0.8541	0.8196
XGB	0.8533	0.8240	0.8542	0.8189
Neural Network	0.8517	0.8185	0.8524	0.8188
Random Forest	0.8596	0.8534	0.8531	0.8152
Extra Trees	0.8699	0.8825	0.8509	0.8063
Decision Tree	0.8507	0.8093	0.8503	0.8057
KNN	0.8518	0.8283	0.8480	0.7864
Naive Bayes	0.8429	0.7113	0.8435	0.7122
AdaBoost	0.9561	0.9887	0.8240	0.6457
Decision Tree	0.8507	0.8093	0.8503	0.8057

3.2 Performance of algorithms on clinical datasets

Similarly, experiments were performed on the clinical dataset with the aim of selecting the algorithms with the best performance in terms of accuracy percentage and area under the curve. Table 7 shows the training results, where it can be observed that the majority of the algorithms reach 100% accuracy, which may indicate overtraining due to the limitation in the number of samples. In the results with the test dataset, a significant decrease in performance is observed in all cases; however, the selected algorithms have an accuracy above 80%, these being Naive Bayes, Gradient Boosting and Random Forest with an accuracy of 0.8607, 0.8481 and 0.8101 respectively.

Table 7. Training and testing results with clinical indicators.

	Training		Test	
Algorithm	Accuracy	AUC	Accuracy	AUC
Naive Bayes	0.7393	0.8325	0.8607	0.8703
Gradient Boosting	0.8625	0.9228	0.8481	0.8618
Random Forest	1.0000	1.0000	0.8101	0.8550
Decision Tree	0.9478	0.9928	0.7974	0.7881
Extra Trees	1.0000	1.0000	0.7848	0.8436
XGB	1.0000	1.0000	0.7848	0.8518
AdaBoost	1.0000	1.0000	0.7721	0.7289
KNN	1.0000	1.0000	0.6835	0.6260
Neural Network	0.6682	0.7367	0.6455	0.7314

3.3 Voting model performance

In the Figure 9 shows the results with the test data set obtained with the final voting model for each type of indicator. In the case of clinical indicators, an accuracy increase of 3% is obtained with respect to the individual algorithm with the best performance that makes up the ensemble type integration.

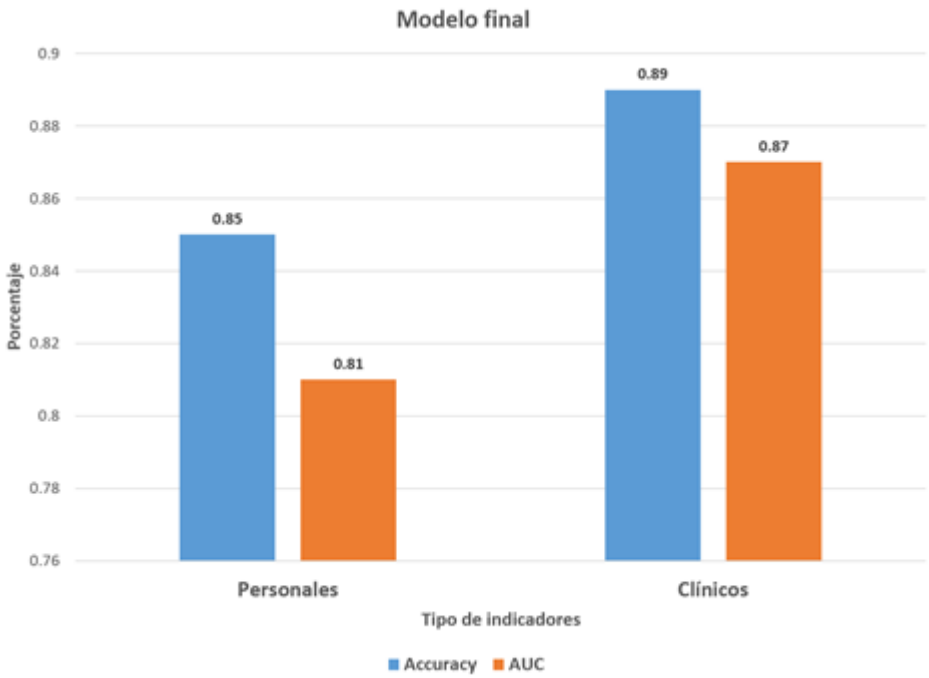

Fig. 9. Test results with final model.

The Figure 10 shows the difference in area under the curve of the individual algorithms and the final voting model for the personal indicators, resulting in an increase of 0.0010 for the vote ensemble.

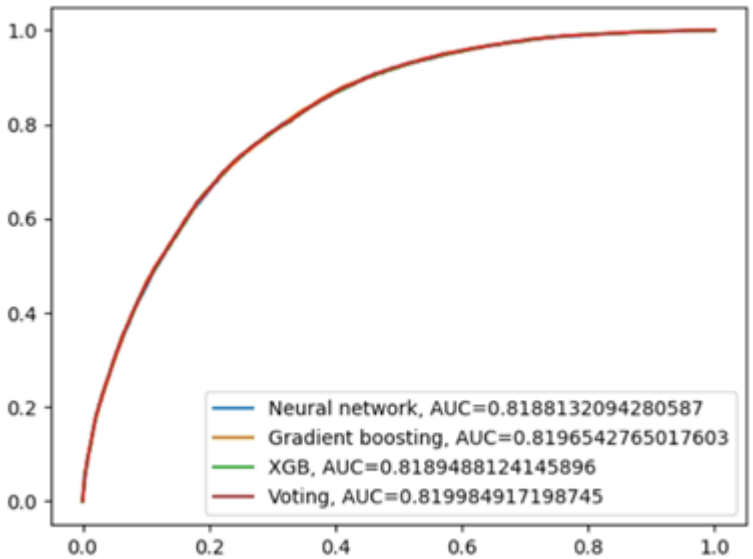

Fig. 10. Difference in area under the curve of algorithms for personal indicators.

The Figure 11 shows the difference in the area under the curve of the individual algorithms and the final voting model for the clinical indicators. In this case, no difference is shown with respect to the best performing algorithm (Naive Bayes)..

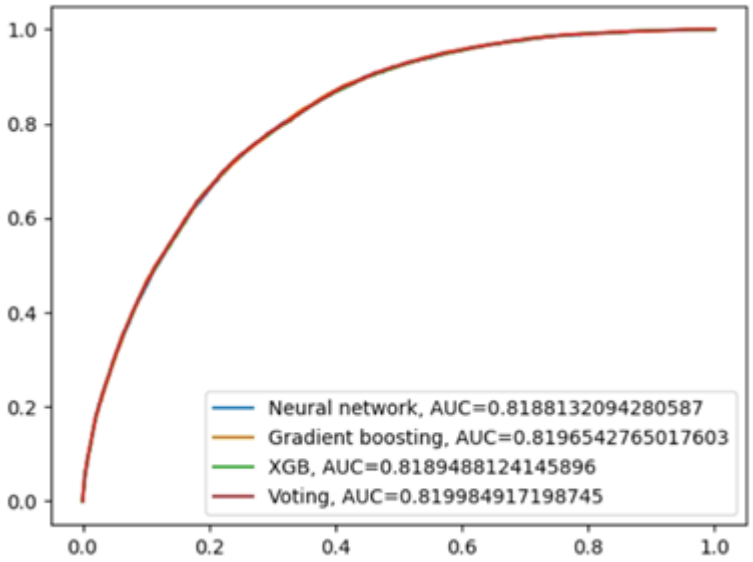

Fig. 11. Difference in area under the curve of algorithms for clinical indicators.

4 Conclusions

In this work, a web-based platform for predicting type 2 diabetes using clinical or personal indicators is proposed, which seeks to provide a first approximation for the general population in the diagnosis of this disease. To this end, two prediction models were developed for each type of indicator using the proposed methodology. The aim is to increase the accuracy percentage by integrating the best algorithms in a weighted voting system. This involves selecting the algorithms based on the results obtained during the experimentation stage. Based on the results obtained, the accuracy result was increased by 3% through the ensemble-type integration of algorithms for clinical indicators compared to the individual models used in said integration. However, for personal indicators, the metric of the algorithm with the best performance was not surpassed following the same methodology. Based on the experimentation and results, the integration of algorithms with similar prediction results is analyzed, that is, they share the same strengths and weaknesses, which may result in a lack of diversity in the integrated models and therefore does not significantly improve the ensemble-type model. To obtain new results, it is desired to integrate a prediction of type 2 diabetes using combined indicators, that is, to make the prediction by entering clinical and personal data, in addition to increasing the number of samples by collecting data from health centers in order to obtain even more precise models and reduce overtraining, as well as considering more metrics in the evaluation and selection of models, in turn exploring the development of predictive models to fill in missing values, thus avoiding discarding them.

References

Ahmed, N., Ahammed, R., Islam, M. M., Uddin, M. A., Akhter, A., Talukder, M. A., & Paul, B. K. (2020). Machine learning based diabetes prediction and development of smart web application. *International Journal of Cognitive Computing in Engineering, 2*, 229–241. https://doi.org/10.1016/j.ijcce.2020.09.001

Anand, A., & Shakti, D. (2015). Prediction of diabetes based on personal lifestyle indicators. In *2015 1st International Conference on Next Generation Computing Technologies (NGCT)* (pp. 253–257). IEEE. https://doi.org/10.1109/NGCT.2015.7375139

Ávila-Tomás, J. F., Mayer-Pujadas, M. A., & Quesada-Varela, V. J. (2021). La inteligencia artificial y sus aplicaciones en medicina II: Importancia actual y aplicaciones prácticas. *Atención Primaria, 53*(1), 81–88. https://doi.org/10.1016/j.aprim.2020.04.003

Barzallo, S., & Barzallo, P. (2019). La inteligencia artificial en medicina. *Ateneo, 21*(2), 81–94.

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems* (Lecture Notes in Computer Science, Vol. 1857, pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference* (pp. 372–378). IEEE. https://doi.org/10.1109/SAI.2014.6918213

Rigalleau, V., Monlun, M., Foussard, N., & Mohammedi, K. (2021). Diagnóstico de diabetes. *EMC – Tratado de Medicina, 25*(2), 1–7. https://doi.org/10.1016/S1636-5410(21)46036-4

Shan, R., Sarkar, S., & Martin, S. S. (2019). Digital health technology and mobile devices for the management of diabetes mellitus: State of the art. *Diabetologia, 62*(6), 877–887. https://doi.org/10.1007/s00125-019-4864-7

Srivastava, A. K., Kumar, Y., & Singh, P. K. (2020). A rule-based monitoring system for accurate prediction of diabetes. *International Journal of E-Health and Medical Communications, 11*(3), 32–53. https://doi.org/10.4018/IJEHMC.2020070103

Teboul, A. (2021). *Diabetes health indicators dataset*. Kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

UCI Machine Learning Repository. (2016). *Pima Indians diabetes database*. Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Verma, A. K., Pal, S., & Kumar, S. (2019). Classification of skin disease using ensemble data mining techniques. *Asian Pacific Journal of Cancer Prevention, 20*(6), 1887–1894. https://doi.org/10.31557/APJCP.2019.20.6.1887

World Health Organization. (2019). *Classification of diabetes mellitus*. WHO. https://www.who.int/publications/i/item/classification-of-diabetes-mellitus