Machine Learning Applied to Predicting Diabetes Mellitus, Using Socioeconomic and Environmental Information from Users of the Health System*

Aprendizaje automático aplicado a la predicción de diabetes mellitus, utilizando información socioeconómica y ambiental de usuarios del sistema de salud

Aprendizagem automática aplicada à predição de diabetes mellitus, utilizando informação socioeconômica e ambiental de usuários do sistema de saúde

Jessner Alexander Mejía[1]; Mario Andrés Oviedo Benalcázar[2]; José Armando Ordoñez[3]; José Fernando Valencia Murillo[4]

[1] Engineer, Lumen Innovations, Colombia. jessner@lumeninnovations.org, ORCID: https://orcid.org/0000-0002-0729-4924

[2] Engineer, ASMET Salud EPS SAS, Colombia. marioandres7@hotmail.com, ORCID: https://orcid.org/0000-0002-4350-5346

[3] PhD, Universidad ICESI, Colombia. jaordonez@icesi.edu.co, ORCID: https://orcid.org/0000-0001-6544-0283

[4] PhD, Universidad de San Buenaventura, Colombia. jfvalenc@usbcali.edu.co, ORCID: https://orcid.org/0000-0003-2997-2121

Received: 07/10/2022. Approved: 15/02/2023. Published: 27/03/2023. Traduced: 28/02/2024

Mejía JA, Oviedo-Benálcazar MA, Ordoñez JA, Valencia JF. Machine Learning Applied to Predicting Diabetes Mellitus, Using Socioeconomic and Environmental Information from Users of the Health System. Rev. Fac. Nac. Salud Pública. 2023;41(2):e351168. DOI: https://doi.org/10.17533/udea.rfnsp.e351168

Abstract

Objective: The objective was to apply models based on machine learning techniques to support the early diagnosis of diabetes mellitus, using environmental, social, economic and health data variables, without dependence on clinical sample collection. **Methodology:** Data from 10,889 users affiliated with the subsidized health system in the southwestern area of Colombia, diagnosed with hypertension and grouped into users without (74.3%) and with (25.7%) diabetes mellitus, were used. Supervised models were trained using k-nearest neighbors, decision trees, and random forests, as well as ensemble-based models, applied to the database before and after balancing the number of cases in each diagnostic group. The performance of the algorithms was evaluated by dividing the database into training and test data (70/30, respectively),

[*] This text is product of the research: “Prediction of type-II diabetes mellitus from environmental, social, economic, and health variables of the population in the subsidized regime in Colombia”, Master’s Thesis in Information Technologies for Data Analytics, Universidad de San Buenaventura Cali (Valle del Cauca). Date begun: 13 October 2021; date of completion: 7 September 2022

and metrics of accuracy, sensitivity, specificity, and area under the curve were used. **Results:** Sensitivity values increased significantly when using balanced data, going from maximum values of 17.1% (unbalanced data) to values as high as 57.4% (balanced data). The highest value of area under the curve (0.61) was obtained with the ensemble models, by applying a balance in the amount of data for each group and by coding the categorical variables. The variables with the greatest weight were associated with hereditary aspects (24.65%) and with the ethnic group (5.59%), in addition to visual difficulty, low water consumption, a diet low in fruits and vegetables, and the consumption of salt and sugar. **Conclusions:** Although predictive models, using people's socioeconomic and environmental information, emerge as a tool for the early diagnosis of diabetes mellitus, their predictive capacity still needs to be improved.

---------*Keywords:* machine learning, diabetes mellitus, environmental factors, socioeconomic factors, predictive model

Resumen

Objetivo: Se propuso aplicar modelos basados en técnicas de aprendizaje automático como apoyo para el diagnóstico temprano de la diabetes mellitus, utilizando variables de datos ambientales, sociales, económicos y sanitarios, sin la dependencia de la toma de muestras clínicas. **Metodología:** Se utilizaron datos de 10 889 usuarios afiliados al régimen subsidiado de salud de la zona suroccidental en Colombia, diagnosticados con hipertensión y agrupados en usuarios sin (74,3 %) y con (25,7 %) diabetes mellitus. Se entrenaron modelos supervisados utilizando *k* vecinos más cercanos, árboles de decisión y bosques aleatorios, así como modelos basados en ensambles, aplicados a la base de datos antes y después de balancear el número de casos en cada grupo de diagnóstico. Se evalúo el rendimiento de los algoritmos mediante la división de la base de datos en datos de entreno y de prueba (70/30, respectivamente), y se utilizaron métricas de exactitud, sensibilidad, especificidad y área bajo la curva. **Resultados:** Los valores de sensibilidad aumentaron considerablemente al utilizar datos balanceados, pasando de valores máximos del 17,1 % (datos sin balancear) a valores de hasta 57,4 % (datos balanceados). El valor más alto de área bajo la curva (0,61) fue obtenido con los modelos de ensambles, al aplicar un balance en el número de datos por cada grupo y al codificar las variables categóricas. Las variables de mayor peso estuvieron asociadas con aspectos hereditarios (24,65 %) y con el grupo étnico (5.59 %), además de la dificultad visual, el bajo consumo de agua, una dieta baja en frutas y verduras, y el consumo de sal y azúcar. **Conclusiones:** Aunque los modelos predictivos, utilizando información socioeconómica y ambiental de las personas, surgen como una herramienta para el diagnóstico temprano de la diabetes mellitus, estos aún deben ser mejorados en su capacidad predictiva.

----------*Palabras clave:* aprendizaje automático, diabetes mellitus, factores ambientales, factores socioeconómicos, modelo predictivo.

Resumo

Objetivo: Propôs-se aplicar modelos baseados em técnicas de aprendizagem automática como apoio para o diagnóstico precoce da diabetes mellitus, utilizando variáveis de dados ambientais, sociais, econômicos e sanitários, sem a dependência da coleta de amostras clínicas. **Metodologia:** Usaram-se dados de 10.889 usuários filiados ao regime subsidiado de saúde da zona sudoeste da Colômbia, diagnosticados com hipertensão e agrupados em usuários sem (74,3%) e com (25,7%) diabetes mellitus. Foram treinados modelos supervisionados utilizando k vizinhos mais próximos, árvores de decisão e florestas aleatórias, assim como modelos baseados em montagens, aplicados à base de dados antes de depois de equilibrar o número de casos em cada grupo de diagnóstico. Avaliou-se o desempenho dos algoritmos por meio da divisão da base de dados de treino e teste (70/30, respectivamente), e utilizaram-se métricas de exatidão, sensibilidade, especificidade e área sob a curva. **Resultados:** Os valores de sensibilidade aumentaram de maneira significativa ao utilizar dados equilibrados, passando de valores máximos de 17,1% (dados sem equilibrar) a valores de até 57,4% (dados equilibrados). O valor mais elevado de área sob a curva (0,61) foi obtido com os modelos de montagens, ao aplicar um balanço no número de dados por cada grupo e codificar as variáveis categóricas. As variáveis de maior peso estiveram associadas com fatores hereditários (24,65%) e com o grupo étnico (5,59%), além da dificuldade visual, o baixo consumo de água, um regime baixo em frutas e vegetais e o consumo de sal e açúcar. **Conclusões:** Embora os modelos preditivos, utilizando informação socioeconômica e ambiental das pessoas, surgem como uma ferramenta para o diagnóstico precoce da diabetes mellitus, ainda devem ser melhorados em sua capacidade preditiva.

---------*Palavras-chave:* aprendizagem automática, diabetes mellitus, fatores ambientais, fatores socioeconômicos, modelo preditivo

Introduction

Diabetes mellitus (DM) is one of the 10 most serious diseases globally and it is characterized by progressive complications that include cardiovascular diseases, retinopathy, cerebrovascular diseases, amputation of body parts, and sometimes death [1]. According with Bernardini [2], between 6% and 8% of the global population is affected by DM, with nearly 400-million people diagnosed and receiving treatment. This same study estimates that costs in caring for DM by 2030 will reach $490-billion US Dollars, equivalent to 12% of the medical expenses of all diseases [2]. For America, it is expected that 109-million people will be diagnosed with DM by 2040, with higher concentration in low- and middle-income countries [3].

In Colombia, according with the study by the High-Cost Account [4], regarding the situation of chronic kidney disease, arterial hypertension (AHT) and DM in the country for 2020, it is observed that, in the last six years, the prevalence of AHT and DM has increased, revealing that for that year there were 4,527,098 prevalent cases of AHT and 1,426,574 cases of DM. The same study indicates that for the period between July 2019 and June 2020, there were 31,316 deaths of people with DM diagnosis, which means an overall mortality rate of 62.78 cases for every 100 000 inhabitants [4]. Given that, according to the World Health Organization [5], nearly 45% of the population with DM ignores that they suffer from it; a crucial task for health systems is to have a timely diagnosis of patients that suffer it or are at greater risk of developing it to be able to promote treatments that allow having more efficient therapeutic management of patients.

Use of models based on machine learning (ML) techniques has demonstrated excellent results in different areas and specializations in medicine, supporting early and effective diagnosis of diseases, to start treatments in timely manner [6-9]. Particularly, efforts have been made to study and propose progress in DM prevention [10,11], including the creation of recommendation systems to promote healthy lifestyles [12,13], construction of body area networks for blood glucose monitoring [14], and creation of models that permit predicting DM [1,15-20]. In these studies, models based on ML used for their training entries derived from parameters taken from physiological variables, as well as from sociodemographic, environmental, and lifestyle variables registered in populations from India [1,19], Mexico [15], China [16], and the United Kingdom [18].

Although the best predictors of DM are associated with physiological variables extracted from clinical samples, such as blood samples for glucose analysis, these tests are costly and their study can take considerable time, for the handling of samples and the logistics required in transport to analysis laboratories. Accessing these data may be quite complex in developing countries, like Colombia, especially in rural zones. For these cases, it is much more feasible to have information from patients through sociodemographic, environmental, economic, and lifestyle variables. Thereby, from the analysis of environmental, social, economic, and health data, without depending on taking clinical samples, this study proposes the development of models based on ML techniques to support early diagnosis of DM or its prediction, to allow health professionals to establish prevention strategies or timely treatment of DM.

Methods

This section describes the database, techniques, and procedures used for the training, validating, and testing of the models proposed in this study for early diagnosis of DM.

Database

The data used in this study were taken from the company ASMET Salud, a health promoter entity (EPS) of the subsidized health regime in Colombia with broad coverage in rural zones and of difficult access, mainly in the southwest and northeast of the country.

Among its processes to identify health risks, ASMET Salud conducts a survey, where it obtains variables associated with lifestyle, eating habits, family background, life conditions, economic level, and environmental conditions for its affiliates, which, in turn, are related with the risk group to which each affiliate belongs (cancer, human immunodeficiency virus, AHT, diabetes, chronic kidney disease, hemophilia, etc.).

The database has 128,501 registries of users. Of these, initially 11,423 were selected, corresponding to users previously identified in the AHT risk group. Among them, 2,883 are users also diagnosed with DM.

Thereafter, and considering that these two pathologies (AHT and DM) occur mainly in the group of people over 40 years of age [21], the study excluded those who were under 40 years of age, leaving 10,889 registries with AHT, of which 2,08 suffers from DM (25.7%).

Data that could individualize the affiliate member was eliminated from each of the registries, like identification document, names, last names, and information that does not contribute to this study, like picture of the home façade, data of the person conducting the survey, telephone numbers, e-mail, etc.

Six registries were excluded in which all their variables were empty, leaving 10,883 registries, with 80% belonging to economic strata 1 and 2. Each registry has

69 variables, of which seven are continuous and 62 are categorical.

A first selection of variables was carried out taking into account the medical criteria according to the variable's contribution to the study objective. Thus, only 18 variables were selected per registry: one continuous variable (age), three nominal categorical variables (blood type, Rh, and ethnic group), and 14 ordinal categorical variables (economic level, educational level, disability, physical activity, fruit and vegetable diet, water consumption, salt in foods, sugar in foods, visual difficulty, skin lesions, weakness in body, loss of strength, relative with diabetes, and diabetes). For this study, the target variable to predict was diabetes, categorized into two groups: patients with DM and without DM. The categorization was based on the medical diagnosis of diabetes registered in the ASMET Salud database.

Null values were identified for each variable, finding that the variables "economic level" and "visual difficulty" had the highest number of registries with lack of data. Null values were replaced by the mode, given that all the variables with null values were categorical. According with the categories registered in the database, the ordinal variables were encoded by scales. For example, the variables "fruit and vegetable diet" and "water consumption" used the following scale: 0-Never; 0,5-Less than once per week; 1-Once per week; 2-Between two and three times per week; 4-Between four and six times per week; 5-Every day.

The variable "relative with diabetes" used: 0-None; 2-Yes: Other relatives; 3-Yes: Grandparents, uncles/aunts, cousins; 5-Yes: Parents, siblings or offspring.

Dichotomous variables (disability, salt in foods, sugar in foods, visual difficulty, skin lesions, weakness in body, loss of strength, diabetes) were encoded with "0" and "1", where "1" indicates the presence of disease or exposure to some factor, and "0", its absence.

Moreover, encoding of nominal variables used the one-hot encoding method, which consists in creating a vector of *N* columns to encode the *N* classes of the nominal variable, and, for each registry, mark the column to which said registry belongs with a 1 and leave the rest with 0.

Prediction models

The models used were selected bearing in mind the following: 1) this is a retrospective diagnostic study applied to a two-class classification problem, one class with individuals with AHT, but without DM (No-DM), and another class that includes people with AHT and who also manifest DM (Yes-DM); 2) methods based on supervised learning can be applied, considering that the database contains the target variable to predict (diabetes); 3) the classes to classify belong to an unbalanced data set, that is, the percentage of No-DM users is much higher than the percentage of Yes-DM users (3 to 1 ratio); and 4) they are the most common models used or evaluated in the solutions found in the state of the art [11-16], both individually as in compound algorithms.

To preprocess the database, training, validation and evaluation of the models proposed, Python® 3.8 was used, with the Scikit-Learn, Pandas, NumPy and Plotly libraries, and Google® BigQuery was used as data centralization services.

The following describe each of the models used.

K nearest neighbors

It is an algorithm that assigns to an observation the most common class among the closest classes (*K nearest neighbors*, KNN) to said observation.

To evaluate closeness, the distance between the test point and each of the training observations is determined, using metrics, like the Euclidean distance, from Manhattan, from Minkowski and Chebyshev.

Upon obtaining the distances, the KNN are taken and the category to which each neighbor belongs is identified. The category with the greatest nearest neighbors will be that assigned to the observation being classified.

Normally, an odd *K* value is taken, to facilitate the tiebreak between which class is closest to the test point, and a small *K* value to reduce computing time, this due to the number of comparisons the algorithm must perform.

This algorithm does not generate a model explicitly and, on the contrary, must compare each instance or test observation with all the training observations [22].

Decision tree

These are versatile machine learning algorithms that can perform both classification and regression tasks, and even multiple output tasks. This algorithm is recognized for being easy to read by the human eye, given that it defines its paths by answering questions, which allows deciding or creating paths to reach a final decision.

Decision trees (DT) are adjusted are adjusted through numerous iterations and through the creation of decision thresholds, from the values taken by the characteristics of a data set [22].

Questions produce an input, question and output schema, which technically look like nodes and branches.

Different algorithms are proposed in the literature [23]. Among the most popular, two are highlighted: C4.5 and Classification and Regression Trees (CART). This work used the latter, provided by the Scikit-Learn library [22].

For the adjustment process, various hyperparameters are used, such as the depth of the tree (maximum number of internal node levels that must be created) and the impurity of a node, determined by entropy or the Gini index.

For the Gini index, which is a measure used by default for the training process in the Scikit-Learn library, a node is said to be pure if Gini = 0, while the closer the Gini value is to 1, the more likely it is that the algorithm will make a mistake in its prediction (impure node).

In the DT and other algorithms derived from them, it is possible to determine the level of importance of the input variables to the model, determined by calculating the mean and standard deviation of the accumulation of the decrease of impurities within each tree.

Random forest

It is a robust ensemble algorithm, which is composed of various versions of the same DT algorithm and where each version is trained with random subsets from the same database (*bagging*), which is divided into sections distributed among the DT that make up the random forest (RF).

Assignment of the samples to each DT is done randomly both in observations as in characteristics (*bootstrapping*), so that each DT is trained with slightly different data [22].

The RF permit controlling the hyperparameters characteristic of a DT, like the number of leaves, and add other hyperparameters, such as number of trees or number of central processing units used for training.

Hyperparameter optimization

The hyperparameter optimization used the grid search technique, a controlled method that allows iterating over a finite number of previously defined values [22].

This optimization is based on selecting a set of *N* values for each parameter from a total of *M* parameters, evaluating each possible combination of hyperparameters.

This work studied the following combinations: 1) for KNN, n_neighbors {3,5,7,9,11,13,15} and weights {'uniform','distance'}; 2) for DT, max_depth {5,10,15,20,30} and criterion {'gini','entropy'}; 3) for RF, n_estimators {20,30,60,100,150,180,200,300}, criterion {'gini','entropy'} and max_depth {5,10}.

Algorithm assembly

Consists in creating an algorithm from the union of multiple models, to improve the generalization of the predictions. The assemblies seek to minimize the model's prediction bias (average of the difference between the predicted value and the real value) and minimize the variance (capacity to respond to new data) [22].

Among the techniques to create assemblies, there are:

Assembly by vote: where diverse algorithms are joined and trained with data subset. Once trained, a new value is predicted of each of the algorithms and the mode is selected.

Bagging method: the same algorithm is trained with different data subsets produced from the training set. The final prediction will be the mode of the predictions obtained for each of the subsets.

Boosting method: is the union in sequence of simple models that transfer their learning rate to each other; this means that given a model *M*, the following algorithm *M1* will learn from the training errors of *M*.

Gradient boosting refers to the gradient descent optimization algorithm used to adjust the loss function when a model is trained. The *eXtreme Gradient Boosting* (XGB) is an efficient open source implementation of gradient boosting algorithm (Python library), designed to be computationally efficient.

Evaluation of algorithms

Given that it is an unbalanced database, with 3:1 ratio in the groups of users diagnosed without diabetes (No-DM) with respect to those diagnosed with diabetes (Yes-DM), balancing was conducted in the number of registries per class. For such, a subsampling technique was applied with which registries in the group with the most data (No-DM) were randomly deleted, until reaching a size similar to the group with less data (Yes-DM).

To evaluate the algorithm performance, the database was divided into two subsets: *training data* (*Train*) and *test data* (*Test*), in 70/30 proportion, respectively.

The *performance metrics* for each model were determined from the confusion matrix, identifying true negatives (TN), false negatives (FN), true positives (TP), and false positives (FP):

1. *Accuracy (Acc):* permits knowing the proportion of elements classified correctly. For this, the relationship between the correct predictions (TP + TN) is considered with respect to the total of observations made (TP + TN + FP + FN).
2. *Sensitivity (Sen):* provides information about the number of positive cases that the model can predict correctly. This metric is of utmost importance in this work, given that the prediction model will be used by ASMET Salud to determine which patients have greater probability of developing DM and, thus, promote early contact with these patients.
3. *Specificity (Spe):* delivers information about the number of negative cases that the model can predict correctly.
4. *Area under the curve (AUC-ROC):* permits knowing the capacity a prediction model has to classify correctly instances that arise. It can also be seen as a separation measure among the classes of a model. It is calculated from the receiver operating characteristic (ROC) curve, which reflects the ratio between the TP rate and the FP rate for different threshold levels.

5. *F1 Score:* combines accuracy and sensitivity in a single value, which permits observing the behavior of these two values in a single measurement.

The following present the equations used to calculate metrics Acc, Sen, Spe, and F1 Score:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Ecuación 1}$$

$$Sen = \frac{TP}{TP + FN} \qquad \text{Ecuación 2}$$

$$Esp = \frac{TN}{TN + FP} \qquad \text{Ecuación 3}$$

$$F1\,Score = 2 * \frac{TP/(TP + FP) * Sen}{TP/(TP + FP) + Sen} \qquad \text{Ecuación 4}$$

Results

The following present the results obtained during training, validation, and test of the models proposed in this study for early diagnosis of DM, applied to the ASMET Salud database.

Distribution of affiliates by age range

Figure 1 shows the population pyramid of the affiliates, from the ASMET Salud database, with AHT. It is noted that the majority (95%) of the affiliates who suffer AHT are older or equal to 40 years of age. Likewise, it was found that 93% of the patients with DM are older or equal to 40 years of age. In light of this situation, this study decided to only consider registries from affiliates who are older or equal to 40 years of age.

To know if the population pyramid obtained with the ASMET Salud database follows the same distribution reported for the Colombian population, Table 1 compares the distribution of affiliates by age and by pathology (AHT and DM) in this study, with respect to those presented by the High-Cost Account [4].

Although the ASMET Salud database contains users mainly from rural zones form the southwest and northeast of the country, Table 1 shows that this population follows a distribution similar to that reported throughout Colombia by the High-Cost Account [4]. The similarity is more evident when comparing the group of affiliates diagnosed with AHT, for which the differences between both populations are < 0.5% for age ranges from 50 to 74 years (62.58% vs. 61.24%) and from 60 to 64 years (14.16% vs. 14.26%). For the group of affiliates diagnosed with DM, the differences are also < 0.5% for age ranges from 60 to 64 years (15.27% vs. 15.41%), but for ages between 50 and 74 years, the difference is of approximately 10% (65.73% vs. 55.80%).

Classification de los risk groups

The results, obtained in the classification of subject according to risk group, are presented grouped for cases that use data without and with encoding, and applying or not balancing techniques in the number of registries per class.

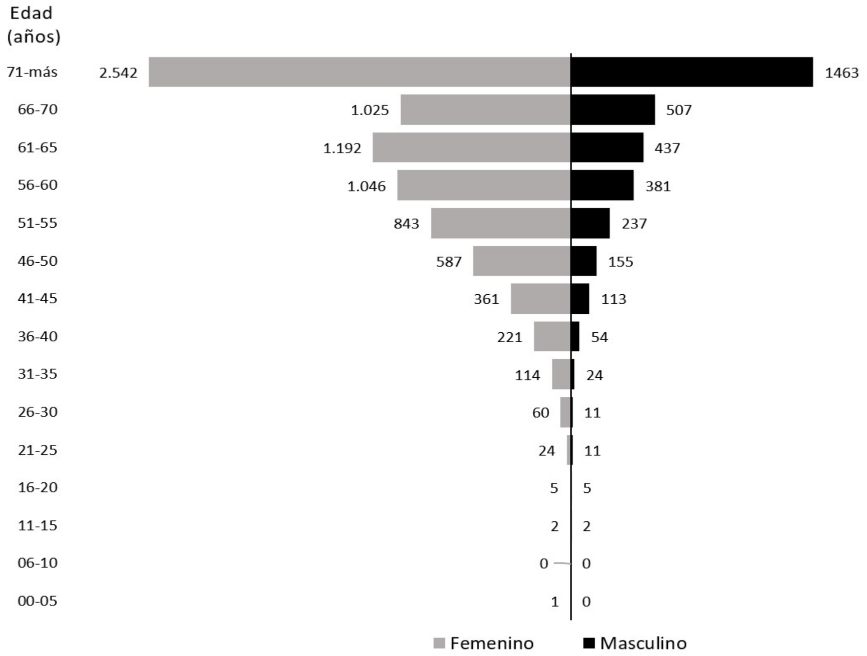

Figure 1. Population pyramid: affiliates diagnosed with hypertension (AHT).

Table 1. Comparison: ASMET Salud database vs. High-Cost Account [4]

Disease	Age range	High-Cost Account [4]	Database (ASMET Salud)
	(years)	(%)	(%)
Diabetes mellitus (DM)	Between 50-74	65.73	55.80
	Between 60-64	15.27	15.41
	Under 35	3.46	3.32
Hypertension (AHT)	Between 50-74	62.58	61.24
	Between 60-64	14.16	14.26
	Under 35	3.89	2.66

Classification using data without encoding

Table 2 contains the performance results of the classification models, applying ML techniques, to classify AHT patients with or without DM. These results correspond to those obtained when using the dataset in pure state, that is, without making any transformation to the categorical and continuous variables. Additionally, the first part of the table shows the results obtained without balancing the number of registries per class (AHT affiliates with and without DM). The second part shows the results after balancing the number of registries per class, applying the subsampling technique.

From Table 2, for unbalanced data, it is noted that: 1) RF obtained the highest Acc value in the *Train* group, but – in turn – one of the lowest values in the *Test* group, suggesting possible overtraining of the model and poor adjustment for generalization; 2) Sen values were quite low (between 1.43% and 15.3%), indicating that the capacity to detect AHT affiliates with DM is very low; 3) Spe values were quite high (between 88.37% and 99.44%), indicating that models mostly learn to detect healthy samples (AHT affiliates without DM); 4) Sen and F1 Score values were low, being the worst results those obtained with the DT model; and 5) the AUC-ROC parameter is between 0.5043 and 0.5207, meaning that the models cannot discriminate correctly instances in each possible class.

Likewise, from Table 2, for balanced data, note that compared with the results achieved for unbalanced data: 1) Acc values diminished in the *Train* and *Test* groups; 2) the Sen value increased, reaching values between 47.01% and 54.15%, improving the capacity to detect AHT affiliates with DM; 3) the Spe parameter decreased to values between 51.38% and 66.06%, indicating reduction in detecting healthy samples (AHT affiliates without DM); 4) the Sen and F1 Score parameters increased in value, with all models obtaining similar values; and 5) the AUC-ROC parameter value increased slightly, between 0.5147 and 0.5656, again indicating that the models cannot discriminate correctly instances in each possible class.

Table 2. Performance of the classification models using data without encoding

Model	Train			Test		
	Acc(%)	Acc(%)	Sen(%)	Spe(%)	F1 Score	AUC-ROC
Unbalanced data						
KNN-5	78.72	69.96	15.1	88.99	0.2063	0.5207
DT	74.29	74.18	1.43	99.44	0.0277	0.5043
RF	98.56	69.55	15.3	88.37	0.2060	0.5185
Balanced data						
KNN-5	70.27	51.47	51.56	51.38	0.5223	0.5147
DT	60.63	56.28	47.06	66.06	0.5256	0.5656
RF	98.86	54.05	54.15	53.94	0.5482	0.5405

Train: training data; Test: test data; KNN-5: 5 nearest neighbors; DT: decision tree; RF: random forest; Acc: accuracy; Sen: sensitivity; Spe: Specificity; AUC-ROC: area under the curve.

Classification using encoded data

The following step, again evaluated the classification algorithms, but this time encodes the categorical variables. The first part of Table 3 presents the results by applying encoding, but without balancing the number of registries per class (AHT affiliates with and without DM). The second part displays the results obtained after balancing the number of registries per class, through the subsampling technique.

From Table 3, for unbalanced data, note that, overall, the values of the Acc, Sen, Spe, F1 Score, and AUC-ROC metrics were very similar to those collected with data without encoding (Table 2), although a slight increase in these metrics occurs when encoded data is used, especially for the DT model, which reached an important increase in sensitivity value, going from 1.43% to 10.16%. However, the general behavior of the three algorithms is deficient, ranging from 10.16% to 17.11% in sensitivity. In this sense, it may be inferred that the models: 1) had very low capacity to detect AHT affiliates with DM (low Sen); and 2) learn to mostly detect AHT affiliates without DM (high Spe).

Likewise, by analyzing the results obtained with balanced data, similar behavior is observed to that reached using data without encoding and balanced (Table 2), again showing a slight increase of said metrics when using encoded data, especially for the DT model. That is: 1) the Sen value increased, reaching values between 46.19% and 57.44%, improving the capacity to detect AHT affiliates with DM; 2) the Spe metric diminished to values between 53.76% and 74.31%, which indicates that detection is reduced of healthy samples (AHT affiliates without DM); and 3) the AUC-ROC parameter value increased slightly, between 0.5304 and 0.6025, with the DT model having the best performance.

Registries with encoded and balanced data (applying subsampling) were used for the following analyses.

Table 3. Performance of the classification models using encoded data

Model	Train			Test		
	Acc	Acc	Sen	Spe	F1 Score	AUC-ROC
	(%)	(%)	(%)	(%)		
Unbalanced data						
KNN-5	78.26	70.60	17.11	89.17	0.2308	0.5314
DT	78.07	72.67	10.16	94.37	0.1608	0.5226
RF	98.40	70.28	15.69	89.23	0.2139	0.5246
Balanced data						
KNN-5	70.07	52.98	51.21	54.86	0.5286	0.5304
dt	61.03	59.84	46.19	74.31	0.5421	0.6025
RF	98.80	55.65	57.44	53.76	0.5714	0.5560

Train: training data; Test: test data; KNN-5: 5 nearest neighbors; DT: decision tree; RF: random forest; Acc: accuracy; Sen: sensitivity; Spe: Specificity; AUC-ROC: area under the curve.

Hyperparameter adjustment

Hyperparameter optimization yielded the following results: 1) KNN: KNeighborsClassifier(n_neighbors = 15); 2) Decision Tree: DecisionTreeClassifier(criterion = 'entropy', max_depth = 5); and 3) Random Forest: RandomForestClassifier(max_depth = 10, n_estimators = 100).

For the specific case of this study, no notable difference was observed in the performance of the entropy-optimized algorithm or with Gini-based optimization. Table 4 shows the results obtained in each of the classification models with optimized hyperparameters (registries with encoded and balanced data)

Table 4 demonstrates that parameter optimization generated models with better generalization, which follows from the very similar values obtained in Acc during training and validation. The models with the highest value in the AUC-ROC metric were DT-adj (60.45%) and RF-adj (60.37%). Particularly, the DT-adj had a Sen of 45.67% and capacity to detect healthy patients of 75.23%.

Algorithm assembly

Results of classifiers, using assembly-based models, are shown in Table 5. In general, the three assembly methods obtained very similar values in the metrics used to measure the performance of the models, with AUC-

Table 4. Performance of classification models with optimized hyperparameters (registries with encoded and balanced data)

Model	Train	Test				
	Acc(%)	Acc(%)	Sen(%)	Spe(%)	F1 Score	AUC-ROC
KNN-adj	53.87	53.87	52.77	55.05	0.5408	0.5391
DT-adj	60.02	60.02	45.67	75.23	0.5404	0.6045
RF-adj	59.93	59.93	45.33	75.41	0.5380	0.6037

KNN-adj (nearest neighbors), DT-adj (decision tree) and RF-adj (random forest) are the classification models adjusted in their hyperparameters. Train: training data; Test: test data; Acc: accuracy; Sen: sensitivity; Spe: Specificity; AUC-ROC: area under the curve.

Table 5. Performance of classification models using assemblies

Model	Train	Test				
	Acc(%)	Acc(%)	Sen(%)	Spe(%)	F1 Score	AUC-ROC
EVM	63.95	59.84	45.67	74.86	0.5393	0.6027
GBM	60.11	60.11	48.96	71.93	0.5582	0.6044
XGB	60.55	60.55	48.44	73.39	0.5583	0.6092

EVM: assembly by vote; GBM: gradient boosting; XGB: *eXtreme Gradient Boosting*; Train: training data; Test: test data; Acc: accuracy; Sen: sensitivity; Spe: Specificity; AUC-ROC: area under the curve.

ROC values of approximately 0.60, with the eXtreme Gradient Boosting (XGB) method obtaining the slightly highest value (0.6092). This assembly had a sensitivity of 48.44% and Specificity of 73.39%.

Variables of greatest importance in the prediction models

Figure 2 shows the variables that obtained a level of importance > 2% for the assembly-based classifier with the XGB method. The degree of consanguinity is the most relevant variable as decisive factor (22.6%, considered accumulated value between relative_diabetes_x), followed by the mestizo ethnic group (5.6%). Other variables are related with visual difficulty, low consumption of water, a diet low in fruits and vegetables, and consumption of salt and sugar.

Discussion

This work applied models based on ML techniques to predict the groups of affiliates who have AHT with and without DM, from data registered in the database of the healthcare provider entity ASMET Salud.

The variables selected and included in the database analyzed center on sociodemographic, anthropometric, and lifestyle variables, excluding metabolic variables, to focus on characteristics that describe the patient's lifestyle and context, which are variables that can even be taken through remote measuring instruments.

From this population of subjects, it is observed that the percentages of affiliates distributed by age range (population pyramid in Figure 1) correspond mostly to that reported in the study of the High-Cost Account [4], where – perhaps – the biggest difference is found in the range of users between 50 and 74 years of age who have diabetes, with 65.73%, which is higher than the 55.80% obtained with the data from ASMET Salud. This difference may be related with the type of population included in the ASMET Salud database, which corresponds principally to users from rural zones in the country's southwest and northeast zones, while the High-Cost Account includes information of users from health entities throughout Colombia.

From the results herein and reported in Table 1, it may be stated that AHT and DM are diseases that, for the study population, have higher prevalence in people over 40 years of age, without meaning to say that the population under 40 years of age is exempt from suffering these diseases.

The performance of the classifiers used in this work (KNN, DT, and RF) is largely affected by the imbalance in the distribution of the two classes modelled: affiliates with AHT, but without DM (No-DM), and affiliates with AHT and who also have DM (Yes-DM). This is reflected mainly in the sensitivity (Sen) metric, which provides information about the percentage of positive patients (Yes-DM group) that the model can predict. Particularly, it goes from values that are around 15% in the unbalanced database, to values of up to 54% in the balanced database.

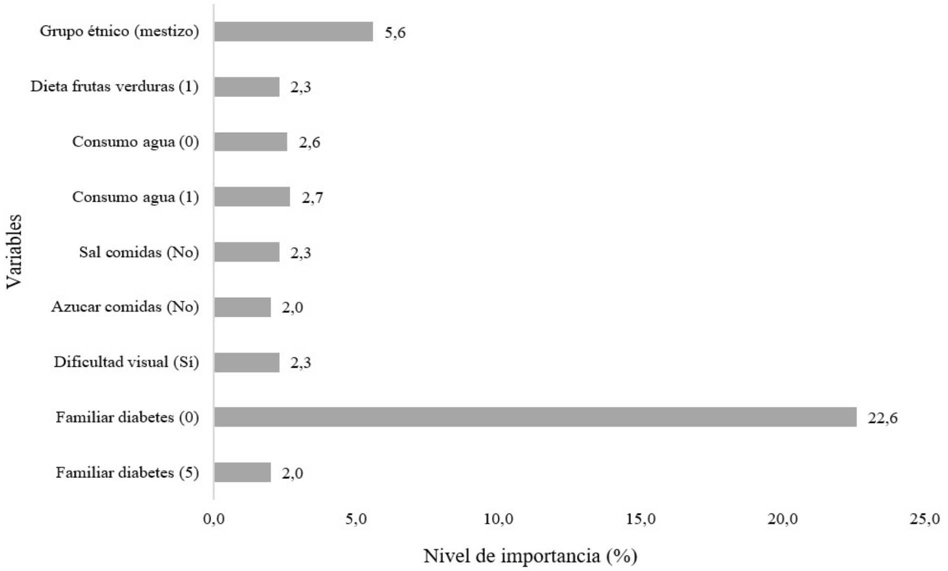

Figure 2. Variables of greater importance for the assembly-based classifier with the eXtreme Gradient Boosting (XGB) method. Fruit and vegetable diet (1): Once per week; water consumption (0): Never; water consumption (1): Once per week; relative with diabetes (0): None; relative with diabetes (5): Yes - Parents, siblings or offspring.

The foregoing is explained by the fact that to train the model, the Acc was considered as metric for parameter optimization, which permits knowing the global relation of users classified correctly, finding users classified in the Yes-DM group and in the No-DM group. This is why the value of the Specificity metric is so high when using the unbalanced database (values between 88% and 99%), compared with the value obtained using the balanced database (values between 51% and 66%). By having balanced data, there was better equilibrium between the specificity and sensitivity measurements from the three models proposed.

To evaluate the models proposed, metrics like the Acc, Sen, Spe, F1 Score, and AUC-ROC have been suggested. This study prioritized the Sen metric, bearing in mind that the best model will be used by collaborators at the ASMET Salud EPS, so it is of utmost importance to know the patients who may develop DM (true positives) to promote early contact of these patients, which will permit undertaking actions on prevention and promotion schemes that could avoid or slow down DM development.

The aforementioned, besides being reflected as a benefit in the quality of life of people are treated appropriately and on time, can become an economic benefit for the EPS, by allowing improvement of indicators measured by the High-Cost Account [4] to determine the percentage of compensation that the EPS must make in managing patients with DM.

Although studies, like those reported by Abbas *et al.,* [15] have achieved sensitivity values of 81.1%, it should be mentioned that these values emerged by including four variables in the models: two physiological variables, measured directly or derived from an oral glucose tolerance test, and wo sociodemographic variables (age and ethnicity). The same study [15] reports that using only the two physiological variables (area under the curve glucose at 2 h and plasma glucose level after 120 min) approximately 72% sensitivity is reached, highlighting the prediction power of these types of physiological variables. However, said study does not report sensitivity values, and includes only sociodemographic variables.

Given the nature of the models used (tree-based models) it is feasible to analyze the importance of each variable included in the model's construction process. This, aligned with a business desire, which is to detect the importance of the variables that affect the development of DM, will permit initiating detailed research of a measurement instrument that permits defining guidelines or internal programs to prevent the disease and its early detection.

Overall, the results demonstrate that consanguinity is the most significant variable with > 20% importance, followed by age and ethnic group, depending on the model. In turn, the frequency of consumption of fruits and water are decisive factors in over 2%, with this being a considerable value. That is, the aforementioned can be joined into four large groups of importance: 1) level of consanguinity of relatives with diabetes; 2) age;

3) healthy diet; and 4) ethnic group. The absence or presence of each of these factors determine > 2% the development of DM among the study population (users of the subsidized health regime located in Colombia's southwestern zone), which confirms the importance of genetic inheritance in the development of diabetes, enhanced by eating style, age, and race. Nevertheless, for ethnic group, a study is necessary on the distribution of ethnic groups in Colombia and of the affiliates to the EPS, which seeks to avoid bias about the prediction and which confirms that the affiliate distribution does not affect the prediction.

Although the highest AUC-ROC value reached in this work was 0.61, it must be considered that said value was achieved by using only sociodemographic, anthropometric, and lifestyle variables, excluding metabolic variables, which, although it is the aim of this study, may also be seen as a limitation.

Other studies in the literature [15-19] have reached AUC-ROC values > 80%, which has been possible specially because the models have introduced as input variables parameters taken from physiological variables, including analytics from blood samples. Therefore, future work should perform a more detailed analysis of the variables registered in the database, applying transformations in said variables or including new variables, such as the anthropometric measurements of users or others that are easy to acquire, and that may be associated with the conditions of AHT and DM.

Conclusion

Models, based on ML techniques, were developed to support early diagnosis of DM or its prediction to allow health professionals establish prevention strategies or timely treatment of DM. The models use variable inputs, derived from environmental, social, economic, and health data, without depending on collecting clinical samples, registered in users of ASMET Salud, an institution of the subsidized health regime in Colombia covering primarily the country's southwestern zone.

The best AUC-ROC values were obtained with assembly-based models, which integrate supervised models using KNN, DT, and RF. The assembly using the XGB technique obtained the highest AUC-ROC value (0.61), identifying as the most important variables those associated with hereditary aspects (24.65%) and with ethnic group (5.59%), in addition to visual difficulty, low water consumption, a diet low in fruits and vegetables, and consumption of salt and sugar.

Acknowledgments

To the directives at the entity ASMET Salud EPS SAS for facilitating the database and advisory in interpreting the information registered in it.

Declaration of funding

No funding source is declared.

Declaration of conflicts of interest

The authors declare having no type of conflict of interest.

Declaration of responsibility.

The authors declare that everything written and the different points of view is the responsibility of all the authors, who reviewed and approved the final manuscript. The institutions of affiliation are not responsible for that described in this article.

Declaration of contributions by authors

Jessner Alexander Mejía, Mario Andrés Oviedo Benalcázar, José Armando Ordoñez and *José Fernando Valencia Murillo* contributed substantially to the research design, analysis and interpretation of the data, critical review of its intellectual content, approval of the final version of the manuscript submitted, and are in capacity of responding for issues related with the accuracy or integrity of any part of the work.

Referencias

1. Howlader KC, Satu MS, Awal MA, et al. Machine learning models for classification and identification of significant attributes to detect type 2 diabetes. Health Inf Sci Syst 2022;10(2). DOI: https://doi.org/10.1007/s13755-021-00168-2
2. Bernardini D. Sobre los aspectos económicos de la diabetes mellitus. Rev Cubana Aliment Nutr. [internet]. 2022 [citado 2022 ago. 26]; 30(Supl. 2):255-61. Disponible en: http://revalnutricion.sld.cu/index.php/rcan/article/view/1226/1701
3. Organización Mundial de la Salud. Informe mundial sobre la diabetes. Geneva, Switzerland: WHO [internet]; 2016 [citado 2022 ago. 26]. Disponible en: https://apps.who.int/iris/bitstream/handle/10665/254649/9789243565255-spa.pdf
4. Cuenta de Alto Costo, Fondo Colombiano de Enfermedades de Alto Costo. Situación de la enfermedad renal crónica, la hipertensión arterial y la diabetes mellitus en Colombia 2020. Bogo-

tá [internet]; 2021 [citado 2022 ago. 26]. Disponible en: https://cuentadealtocosto.org/site/publicaciones/situacion-de-la-enfermedad-renal-cronica-la-hipertension-arterial-y-diabetes-mellitus-en-colombia-2020/

5. Colombia, Ministerio de Salud y Protección Social. Prevenir la diabetes, clave desde los hábitos saludables. [internet]; 2021 [citado 2022 ago. 26]. Disponible en: https://www.minsalud.gov.co/Paginas/Prevenir-la-diabetes-clave-desde-los-habitos-saludables.aspx
6. Kruczkowski M, Drabik-Kruczkowska A, Marciniak A, et al. Predictions of cervical cancer identification by photonic method combined with machine learning. Sci Rep. 2022;12(1):3762. DOI: https://doi.org/10.1038/s41598-022-07723-1
7. Hameed Z, Zahia S, Garcia-Zapirain B, et al. Breast cancer histopathology image classification using an ensemble of deep learning models. Sensors. 2020;20(16):4373. DOI: https://doi.org/10.3390/s20164373
8. Konnaris MA, Brendel M, Fontana MA, et al. Computational pathology for musculoskeletal conditions using machine learning: Advances, trends, and challenges. Arthritis Res Ther. 2022;24(1):68. DOI: https://doi.org/10.1186/s13075-021-02716-3
9. Lee LS, Chan PK, Wen C, et al. Artificial intelligence in diagnosis of knee osteoarthritis and prediction of arthroplasty outcomes: A review. Arthroplasty. 2022;4(1):16. DOI: https://doi.org/10.1186/s42836-022-00118-7
10. Lazzarini PA, Raspovic A, Prentice J, et al. Guidelines development protocol and findings: Part of the 2021 Australian evidence-based guidelines for diabetes-related foot disease. J Foot Ankle Res. 2022;28:15. DOI: https://doi.org/10.1186/s13047-022-00533-8
11. Patel D, Msosa YJ, Wang T, et al. An implementation framework and a feasibility evaluation of a clinical decision support system for diabetes management in secondary mental healthcare using CogStack. BMC Med Inform Decis Mak. 2022;100(1):22. DOI: https://doi.org/10.1186/s12911-022-01842-5
12. Cerón-Rios GM, Lopez-Gutierrez DM, et al. Recommendation System based on CBR algorithm for the Promotion of Healthier Habits. Sanchez-Ruiz AA, Kofod-Petersen A, editors. Proceedings of ICCBR 2017 Workshops (CAW, CBRDL, PO-CBR), Doctoral Consortium, and Competitions co-located with the 25th International Conference on Case-Based Reasoning (ICCBR 2017). Trondheim, Norway, June 26-28, 2017. CEUR Workshop Proceedings [internet]; 2017. pp. 167-76 [citado 2022 ago. 26]. Disponible en: https://ceur-ws.org/Vol-2028/paper16.pdf
13. Li J, Huang J, et al. Application of artificial intelligence in diabetes education and management: Present status and promising prospect. Front Public Health. 2020;8:173. DOI: https://doi.org/10.3389/fpubh.2020.00173
14. Rohokale V, Rashmi Neeli, Prassad Ramjee. A cooperative internet of things (IoT) for rural healthcare monitoring and control. 2011 2nd International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE). 2011; 1-6. DOI: https://doi.org/10.1109/WIRELESSVITAE.2011.5940920
15. Abbas H, Alic L, Rios M, et al. Predicting diabetes in healthy population through machine learning. In: Proceedings - IEEE Symposium on Computer-Based Medical Systems. Institute of Electrical and Electronics Engineers Inc. [internet]; 2019. pp. 567-70 [citado 2022 ago. 26]. Disponible en: https://ieeexplore.ieee.org/document/8787404
16. Zhang L, Wang Y, Niu M, et al. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study. Sci Rep. 2020;4406(1):10. DOI: https://doi.org/10.1038/s41598-020-61123-x
17. Dinh A, Miertschin S, et al. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak. 2019; 211(1):19. DOI: https://doi.org/10.1186/s12911-019-0918-5
18. Fazakis N, Kocsis O, Dritsas E, et al. Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access. 2021;9:103737-57. DOI: https://doi.org/10.1109/ACCESS.2021.3098691
19. Shetty G, Katkar V. Type-II diabetes detection using decision-tree based ensemble of classifiers. In: 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA); 2019. pp. 1-5. DOI: https://doi.org/10.1109/ICCUBEA47591.2019.9129348

20 Haq AU, Li JP, Khan J, et al. Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data. Sensors. 2020;20(9):2649. DOI: https://doi.org/10.3390/s20092649

21. Leiva AM, Martínez MA, Petermann F, et al. Factores asociados al desarrollo de diabetes mellitus tipo 2 en Chile. Nutr Hosp. 2018;35(2):400-7. DOI: https://doi.org/10.20960/nh.1434
22. Géron A. Hands-on machine learning with Scikit-Learn and TensorFlow. CA: O'Reilly Media; 2017. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/
23. Priyam A, Abhijeeta, Gupta R, et al. Comparative analysis of decision tree classification algorithms. Int. J. Curr. Eng. Technol. 2013;3(2):334-7. https://inpressco.com/comparative-analysis-of-decision-tree-classification-algorithms/