Health and Technology (2023) 13:769–780
https://doi.org/10.1007/s12553-023-00781-z

ORIGINAL PAPER

Explainable and transparency machine learning approach to predict diabetes develop

Francesco Curia[1]

Received: 3 January 2023 / Accepted: 22 August 2023 / Published online: 27 September 2023

Abstract

Purpose This study aims to address the problem of type 1 diabetes by utilizing machine learning techniques and developing a decision support system based on Explainable Artificial Intelligence (XAI). The main research question is to predict the risk of developing type 1 diabetes in a population using different machine learning algorithms, while ensuring interpretability and transparency of the decision support system. The study builds upon a case-control study conducted by previous researchers, who approached the problem from a statistical-parametric perspective.

Method In this work, various machine learning algorithms, including Decision Trees (DT), Deep Neural Networks (DNN), XGBoost (XGB), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Classifier (SVC), are employed. The algorithms are evaluated based on their ability to predict the disease risk accurately and consistently on both the training and validation datasets. Additionally, Explainable AI techniques such as LIME (Local interpretable model-agnostic explanations) are employed to contextualize and interpret each prediction and assess the importance of various characteristics influencing the probability of developing the disease.

Results The results obtained from the application of machine learning algorithms show promising outcomes on both the training and validation datasets. However, the best-performing algorithms are not necessarily those with the highest accuracy, as they may suffer from overfitting. Instead, algorithms such as DNNs (97%) or KNNs (93%) exhibit similar behavior on both training and test datasets, making them more reliable, LR and SVC both around (98.3%). The adoption of Explainable AI techniques enables the measurement of each characteristic's importance and the analysis of factors influencing the disease's development probability. This allows the development of a clinical decision support system (CDSS) that is immediately understandable, transparent, and interpretable. By leveraging machine learning techniques and Explainable AI, this study addresses the challenge of type 1 diabetes prediction and decision support.

Conclusion The results indicate that algorithms like DNNs and KNNs offer reliable performance in predicting the risk of developing type 1 diabetes. The integration of Explainable AI techniques, specifically LIME, enhances the interpretability of predictions and provides insights into the factors influencing the disease. The developed CDSS based on XAI can potentially assist healthcare professionals in making informed clinical decisions, thereby improving patient care and management of type 1 diabetes.

Keywords Decision support system · Diabetes · Explainable AI · Interpretable machine learning · Extreme gradient boosting · Neural networks

1 Introduction

1.1 Motivations and purpose

Technological and scientific progress currently makes it possible to have effective tools in the fight against diseases. Thanks to expert systems (**SE**) that use artificial intelligence (**AI**) it is possible to obtain effective help and valid support in the early diagnosis and treatment of various pathologies. The exponential growth of the amount of data available, the so-called *Big Data*, and the possibility of connecting monitoring devices, or the *Internet of things* (IoT), make it possible to monitor affected patients in real time and almost in real time from pathologies. Diabetes is one of the diseases that can currently be treated through intelligent systems that make use of AI and its sub-branches such as machine learning (**ML**) and deep learning (**DL**). Nowadays it is possible to connect to

✉ Francesco Curia
francesco.curia1985@libero.it

[1] Rome, Italy

softwarte applications (Apps) installed on your devices such as tablets and smartphones to have a real-time monitoring of the parameters that affect people with this disease, such as blood glucose levels during the day. Predicting the evolution of a given phenomenon, such as the risk of developing this pathology, is a task that scientists and researchers are carrying out brilliantly; more and more AI/ML tools are used in order to predict the risk or evolution of the disease. Given a series of characteristics (features) as input to a given form of artificial intelligence, the machines learn independently and are able to provide a certain output, in the form of probability, label or point value, depending on whether the problem is one of classification or regression, and whether the problem is supervised or not. It arises spontaneously to ask ourselves, for example, once we have obtained a certain probability of onset of a certain disease, what were the factors that influenced its calculation, and how they are configured in the clinical picture. Another question that could be asked concerns the responsibility that a clinical practitioner has when making a decision and that decision has been achieved through intelligent systems. The clinical operator is the decision-maker on whom the responsibility for the decision weighs, be it a pharmacological treatment, be it a transplant or a diagnosis. These questions need precise answers and it is here that researchers are working in the direction of the known branch of AI Explainable AI (**XAI**), that is methods and models capable of interpreting and explaining an AI/ML algorithm called black-box. It could reasonably be assumed that a prediction or a regression, without interpretation of the components that define the algorithm, is an end in itself, moreover with the advent of European laws known as GDPR, and specifically the *right to explanation*, the service provider that makes use of AI is by law required to explain how the algorithm arrived at a certain decision. The purpose of this work is to analyze a complex problem such as that of type 1 diabetes, using machine learning algorithms and above all to provide a clinical decision support system (**CDSS**) that makes use of AI but at the same time is explainable and interpretable through XAI tools.

2 Clinical decision support systems

2.1 Diabetes

Several studies have also been conducted in the context of diabetes treatment using the CDSS. Specifically, [1] present a work based on the evaluation of the impact of an electronic system to support clinical decisions on diabetes, relating to medical records on the control of glycated hemoglobin A1c, blood pressure and cholesterol levels (LDL) in adults with diabetes. The study is relative to the period 2006-2007 on 2,556 diabetic patients. The CDSS was designed to improve care for those patients whose hemoglobin A1c, blood pressure or LDL levels were higher than the target through the application of general and generalized linear mixed models with repeated time measurements. In [2] Georga et al. it is present a clinical diabetes management system to support the follow-up and treatment processes of diabetic patients and also the authors propose a data mining of time models as a tool to predict and explain the long-term course of the disease. In the context of methods for multi-criteria decisions, Rung-Ching et al. [3] propose a TOPSIS based method to calculate the ranking of anti-diabetic drugs; the CDSS presents a utility of 87% through a recommendation system for outpatients. The authors also discuss the fact that in addition to helping the clinical diagnosis of doctors, the system can not only serve as a guide for specialist doctors, but it can also help non-specialist doctors and young doctors to prescribe medications. Diabetes is a chronic disease characterized by an excess of glucose in the blood, the International Diabetes Federation has estimated an alarming rise in the number of diabetics by the year 2030 [4, 5]. This disease is divided into two forms, type 1 diabetes and type 2 diabetes. Hyperglycemia can be caused by insufficient insulin production (i.e. the hormone that regulates the level of glucose in the blood) or by its inadequate action. Type 1 diabetes is characterized by the total absence of insulin secretion, while type 2 diabetes is determined by a reduced sensitivity of the organism to insulin and this disease can progressively worsen over time and is established on the basis of a pre-existing condition of insulin resistance. Type 2 diabetes is a disease with a high spread all over the world also due to the lifestyle of today, such as an unhealthy diet and/or little or no physical activity. In type 1 diabetes, affected people must necessarily take insulin by injection, in type 2 diabetes an appropriate drug therapy associated with a healthy lifestyle allows to contain the negative effects of the disease. Often the presence of hyperglycemia does not give any symptoms or signs, for this reason diabetes is considered a subtle disease. The associated symptomatology in acute cases is characterized by fatigue, increased thirst (polydipsia), increased diuresis (polyuria), unsolicited weight loss, sometimes even concomitant with increased appetite, malaise, abdominal pain, up to to arrive, in the most serious cases, to mental confusion and loss of consciousness. The major complications deriving from diabetes can cause the patient various damages, which are divided into: Ocular (retinopathy): caused by chronic hyperglycemia and hypertension leading to alteration of blood vessels with consequent worsening of vision up to blindness, Cardiocerebrovascular: myocardial infarction or ischemic heart disease, stroke, Renal (nephropathy): damage to the filtering structures of the kidney which can lead in extreme cases to dialysis, Neurological (neuropathy): anatomical

and functional alteration of the central, peripheral and Voluntary nervous system, sensory, motor, visual, acoustic deficits. According to scientific studies, the individuals who are most likely to develop diabetes are: fasting blood glucose between 100 and 126 mg/dl, first degree family members for type 2 diabetes, Body Mass Index, i.e. weight ratio in kilos/height in m2, with a value > 25 kg/m2.

2.2 Machine learning for diabetes prediction

Much of the studies performed on diabetes using machine learning techniques use as given the well-known dataset Pima Indian Diabetes Database (PIDD) provided by the National Institute of Diabetes, Digestive and Kidney Diseases, several authors have proposed algorithms and methods to predict and classify diabetes. The data set consists of 768 patients (called *examples*) each with 9 numerical features and the data refer to women aged 21 to 81 years. The target variable under study is the class variable (diabetes = 1 (yes), diabetes = 0 (no)) [6]. Deepti and Dilip [7] use PIDD data for the classification of diabetes through three machine learning algorithms, the authors specifically apply Naive Bayes (NB), Decision tree and Support Vector Machine, obtaining respectively in terms of accuracy a value of **76.30**% for the NB, 73.82% for the DT and the lowest for the SVM equal to 65.10% and the maximum recall value is reached by the NB equal to 0.763. Han et al. [8] again on this PIDD dataset apply an algorithm based on two steps, in the first they apply an improved *k*-means and in the second step a Logistic regression. In this work, where other data and not the PIDD will be used, the main purpose is to show that it is possible to build a CDSS based on machine learning and deep learning and that this system is interpretable, and in every single component, as an intrinsic requirement of interpretability defined by [9] and both from a statistical point of view, in terms of the probability of disease development, and above all, which factors (or features) influence the final prediction. Through this method the authors reach an accuracy of 3% higher than the results present in other works (**95.42**%), such as that of Patil et al. [10]; the authors obtain an accuracy result equal to 92.38% through their method called Hybrid Prediction Model (HPM) which uses the Simple *k*-means clustering algorithm aimed at validating the chosen class label data (incorrectly classified instances are removed, the model is extracted from the original data) and then apply the classification algorithm to the resulting data set. The C4.5 algorithm is used to create the final classification model using the *k*-fold cross-validation method. The purpose of this case study is not to obtain a better classifier in terms of metrics like accuracy, recall and precision, but rather to provide a valid method in terms of explainability of these results that in the authors cited in literature examined have not provided. However, the studies cited employ supervised models only to find patterns in the data that can help predict disease development in advance, while they are not aimed at "explaining" or what characteristics define the possible development nor an interpretation of the models used. A limitation in the context of the CDSS is precisely the interpretability of algorithms and solutions. Dagliati et al. [11] within the EU-funded MOSAIC project developed a series of data mining algorithms and predictive models to predict complications of type 2 diabetes mellitus (T2DM). Basing the study on data collected through the electronic medical records of nearly a thousand patients. Missing data were imputed using the well-known Random Forest (RF) model. The predictive part was carried out through logistic regression to predict the onset of retinopathy, neuropathy or nephropathy, in different temporal scenarios. The variables considered by the authors in the study refer to: sex, age, time since diagnosis, body mass index (BMI), glycated hemoglobin (HbA1c), hypertension and smoking. The accuracy of the model was 83 %. Zou et al. the [12] use several machine learning and deep learning algorithms to predict diabetes mellitus. Specifically, the authors apply Decision Trees, Random Forest, and Neural Networks to a hospital physical exam dataset in Luzhou, China. The data consists of 14 features and the training and construction of the models was achieved through the use of the technique known as cross-validation. The authors randomly select 68994 data from healthy people and diabetic patients, respectively, as a training set. The sample was highly unbalanced. In order to reduce the data size the authors use Principal Component Analysis (PCA) and Maximum Relevance Minimum Redundancy (mRMR) to reduce dimensionality. The best performance was obtained from the Random Forest with an accuracy value equal to 80 %. Among the more recent works is that of Butt et al. [13], whose authors also use the PIDD dataset. The authors propose a machine learning-based approach to classifying, early-stage identification and prediction of diabetes. The interesting part that the authors propose is that relating to a possible diabetes monitoring system based on IoT sensors in order to monitor its blood glucose level. For the classification of diabetes, three different algorithms were used: Random Forest (RF), Multilayer Perceptron (MLP) and Logistic Regression (LR). The methodological approach is interesting: for the part relating to predictive analysis, they use the so-called Long Short Term Memory (LSTM), Moving Averages (MA) and Linear Regression (LR). The results are encouraging, use of MLP outperforms other classifiers with an accuracy rate of 86.08%, while use of LSTM improves prediction with a significant 87.26% accuracy of diabetes.

3 Data and modeling

3.1 Data

The data collected comes from a questionnaire administered to 306 people in Dhaka, Bangladesh [14]. The study includes 22 risk factors extracted from research on the prevention of type 1 diabetes disease. The sample is divided into two groups, case-control, with 152 people with the disease and 154 who do not have the disease. The authors of the study from which these data were collected extracted significant associations using data mining techniques and a statistical approach. The authors also use a probabilistic approach based on decision trees in order to show the efficiency and robustness of data, which can be used for future research purposes, such as machine applications and deep learning. The risk factors analyzed by the authors include characteristics such as age, sex, BMI, weight and height; whether or not the subject has hypoglycemia, autoantibodies, whether or not the individual is an insulin taker and in what modality; genetic and hereditary characteristics such as if there is a clinical history of diabetes in the family, both type 1 and type 2. Characteristics related more to the socio-geographical context on the mother's education, on the adequate nutrition, if there are already previous pathologies, duration of the disease and residence.

3.2 Data Analysis

The [14] authors who collected the data performed different types of analyzes; specifically they used association measures (given the qualitative nature of the variables) such as the Gini index, the info Gain, the Gain Ratio and the Chi-Square, in addition to showing a significance of the factors and subfactors the p -value is calculated for each of them. The results show that high significance factors are family history for type 1 and type 2 diabetes, gender is a highly statistically significant variable as the p -value shows as it is for age. The measure of association based on the Chi-Square shows how the variable concerning insulin intake and also the information on how it is taken, has a strong association value with the disease. Through the other association indicators such as Gain Ratio, Info Gain and Gini and Chi-Square index it emerges that factors such as HbA1c and hypoglycemia are significant of the disease.

From the plot 1 it is possible to deduce some characteristics, such as BMI, weight and height, in patients with type 1 diabetes and in healthy individuals. From the distributions it can be seen that weight and height are greater in the group of people affected by the disease, while the BMI seems to be almost the same for the two groups, slightly higher in patients with type 1 diabetes. From the scatterplot always in the Fig. 1 there is a positive trend between weight and

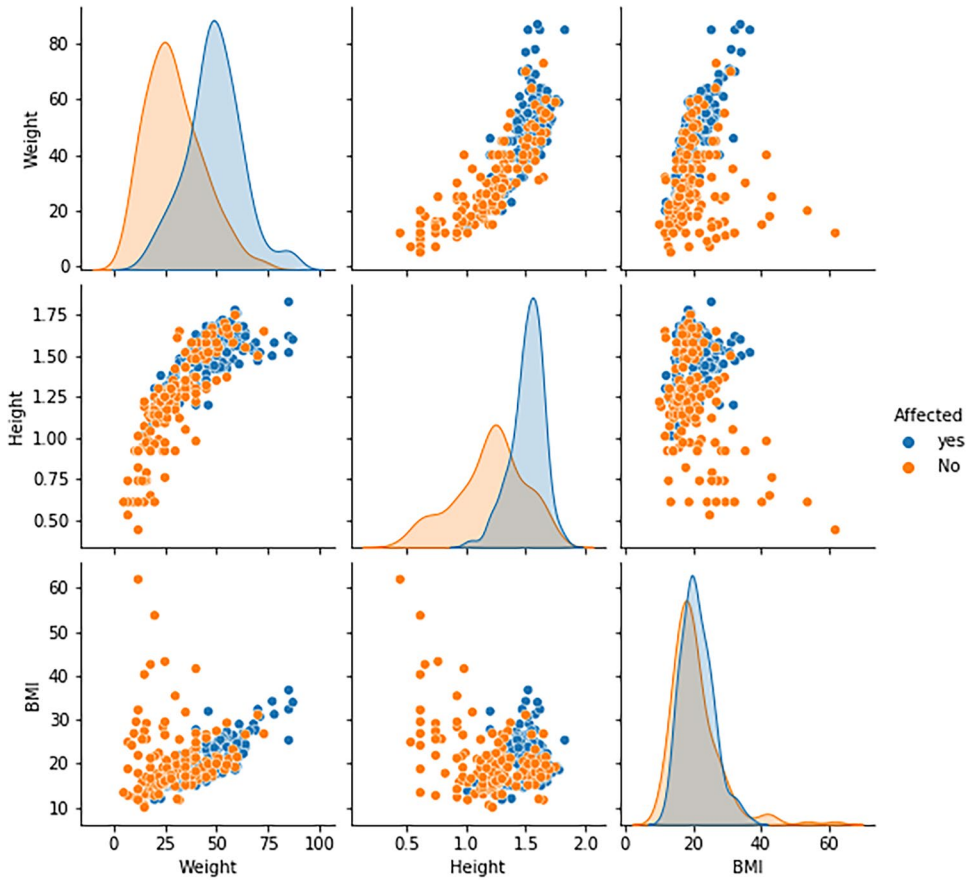

Fig. 1 Distributions of some features

height, with no linear functional relationship between these two variables with BMI. The Fig. 2 shows the distribution of BMI for individuals affected and not by type 1 diabetes, compared to the patient's family history for type 2 diabetes and the value of HbA1c (over or less than 7.5%, threshold value considered). It is therefore possible to show once again how in individuals without a pathological family condition (type 2 diabetes) with threshold values less 7.5% of HbA1c (therefore not at risk) the BMI value is high (plot below left) but slightly below that of individuals with possible inheritance due to family condition of type 2 diabetes (plot below right), with a value greater than 25.

The Fig. 3 instead compares the value of HbA1c with respect to the BMI for individuals with a pathological family history of type 1 diabetes; they are very low, for values over the 7.5% HbA1c threshold: the 7.5% threshold was chosen in relation to the fact that for values equal to or greater than 6.5% in two measurements conducted at different times, we are in the presence of a diagnosis of diabetes. This could lead to the belief that the pathological state of type 1 diabetes is unrelated to a high BMI value and this makes sense for this form of diabetes, while we know that weight and obesity are factors related to the onset of diabetes type 2. Among the other statistics it is interesting to note that 78 % of those observed did not have regressed pathologies, 2.6 % eye problems and 1.6 % stomach problems. For the group of people with type 2 diabetes, it is observed that 57 % have no previous pathologies but 5 % have eye problems and 3.2 % stomach problems. 100 % of the subjects in the control sample (not suffering from type 2 diabetes) do not have any previous diseases. As regards the duration of the disease (sample of 152 people) almost 43 % have a clinical history of the disease between 2 and 6 years, 3 % between 15 and 16 years.

Figure 4 shows the relationship between the number of years of disease (for those affected, target = 1), sex and age. It is possible to note that the course in years of the disease in affected people is concentrated in the female population aged over 15, with an average of 6 years and a maximum of 16 years, in men the average drops to 5.5; in individuals of both sexes, under the age of 15, the average drops to 2.

3.3 Processing

The data collected through the questionnaire are presented in unstructured form, therefore it was necessary to encode the variables using binary or multiclass form, where the number of classes was ≤ 4, for the other variables with a greater number of classes a transformation was necessary dummy type, with the exception of the only three continuous features such as weight, height and BMI. As the data comes from a case-control study, the sample is balanced and it was not necessary to employ balancing techniques (ie. oversampling, undersampling, ...). A feature was kept out of the analysis as there was a percentage ($\geq 50\%$) of missing observations, this is the "Duration of disease" feature, although it is an important information, this missing percentage could alter and inefficient subsequent inferences. However, this information can be subsequently correlated with the other information in a general analysis framework, possibly considering a subsample of people with the disease. There was no need to scale or normalize the variables or to impute missing values,

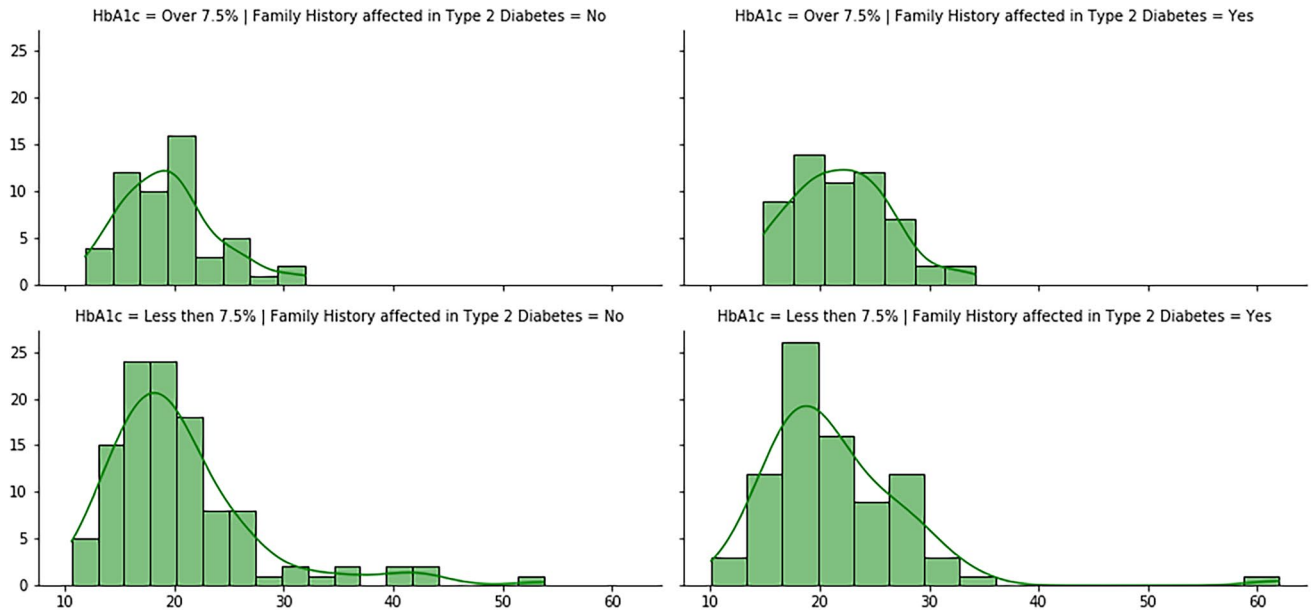

Fig. 2 BMI distributions for HbA1c, Family History (diabetes type 2) for affected

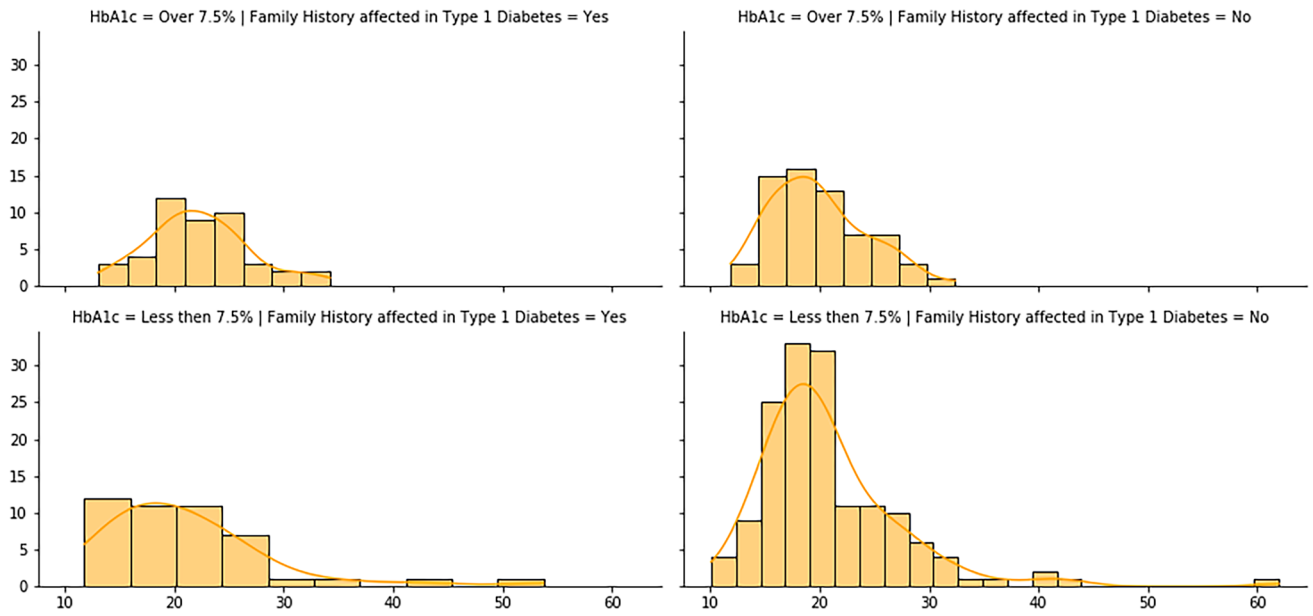

Fig. 3 BMI distributions for HbA1c, Family History (diabetes type 1) for affected

so the construction of the models was very fast. The initial sample, consisting of 306 observations and 22 features, was divided as usual, in a part of the training set, equal to 70% to build and train the model, a part equal to 20% for the validation of the models and a remainder 10% for the test phase on the predictive capacity of the model on so-called unknown data. The split of the dataset occurs randomly and this variability is controlled by using a "seed" at the time of subsampling in order to be able to repeat the tests under the same conditions in which this experiment was conducted.

3.4 Machine learning modeling

In order to build a robust binary classifier capable of classifying the target variable under study (1 = affected, 0 = no) with a certain predictive capacity, some of the best

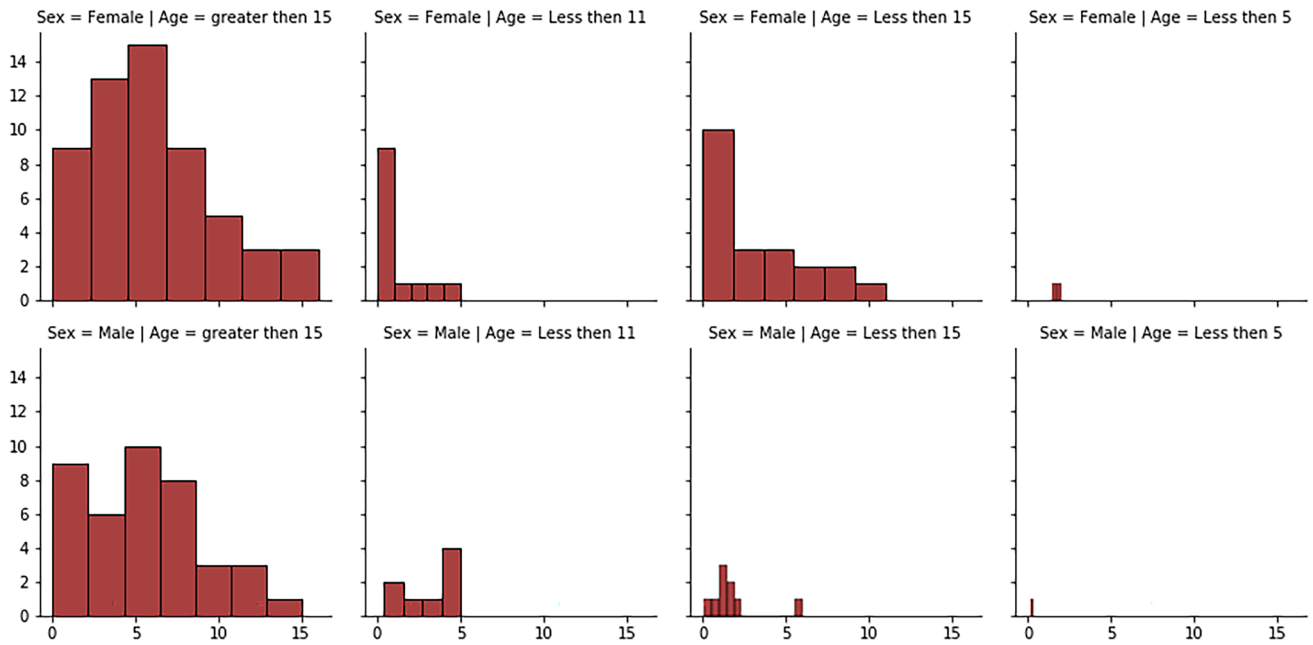

Fig. 4 Distributions of disease years, Age and Sex

known supervised methods were applied to the available data. Specifically, models such as Support Vector Machines, XGBoost, Logistic Regression, KNN, Decision Trees and Deep Neural Networks were trained Table 1, validated and tested: these algorithms were compared with each other in terms of performances, using metrics such as accuracy, sensitivity and specificity. The sample examined made up of 306 observations, deriving from a case-control study conducted in Bangladesh, is balanced, precisely due to the nature with which the clinical studies are carried out, therefore there was no need to adopt balancing techniques. By virtue of this balance of cases with diabetes and unaffected cases, the estimates made are consistent. The predictive capacity in the models used is highly precise, the number of features examined equal to 21 and the limited sample size, also typical of the nature of clinical studies, allowed to train each model very effectively from a computational point of view.

The use of the resulting models in the Table 1 is justified by providing a general overview to the reader of the types that can be used in a CDSS. Models such as Logistic Regression provide easier interpretation of parameters and predictions, as output can be explained through linearity generating a directly observable cause-effect relationship. Contrary to this ease of interpretation, since this property should be intrinsic in a CDSS, a Deep Neural Network is employed; whose parameters during the training phase increase exponentially with the increase of the layers used. The SVC model is an algorithm that is strongly affected by the sample size but widely usable in a clinical problem, where the datasets are not so large, preserving the simplicity of explanation. The KNN, on the other hand, is one of the simplest models and one that suffers a lot from the problem of overfitting: it was used to show how its simplicity of interpretation generates the trade-off between complexity and accuracy typical of intelligent systems that use machines. and deep learning. Among the simplest models to explain is the decision tree which at the base has a very simple logic of partitioning of the instances based on successive splits and homogeneity in the nodes. Being an ensemble method based on trees, XGBoost has the simplicity of interpretation of Decision Trees but the accuracy of a more complex method, since the objective function is solved with an approximation method such as the gradient descent and it is therefore a locally optimal solution, thus not ensuring the transparency of the AI algorithms.

3.5 Statistical analysis and performances

This section analyzes the performances of the algorithms used; specifically, there are some premises that must be made in order to understand the purpose of the work. With current methodological tools, building clinical decision support systems using AI and ML tools is getting easier. Algorithms are increasingly performing and the results in terms of accuracy and predictive capacity are highly efficient. The point is that obtaining models that predict a certain characteristic or a certain risk, or that classify a phenomenon, is an end in itself if the model and the solution found (and the answer to the problem) cannot be explained. With the data we have available in this work, it is easy to build a model that classifies an instance as possible for the development of diabetes, but the crucial purpose is to understand which factors to keep an eye on over time in order to understand its possible evolution. Some data considerations are a must:

(a) the data used in this study by questionnaire bring to noisy and approximate biased data
(b) the sample size is small: there is insufficient data in this study for only some of the machine learning models chosen in this work
(c) the data available as datasets have already been preprocessed by the authors in the original work

Given these premises, it is clear that it is possible to build a model that is highly capable of recognizing a behavior, but also we must instead investigate whether we have methods available to understand which characteristics in the model

Table 1 main parameters for each model

Logistic regression	XGBoost	Deep Neural Network
tollerance: 0.001	n-estimators: 500	layer: (10,10,10,10,10)
iterations: 250	learning rate: 0.1	activation: ReLu
solver: liblinear	max depth: 5	learning rate: 0.1
C: 0.75	objective: binary logistic	optimizer: adam
Support Vector Classifier	KNN	Decision Trees
C: 0.6	neighbors: 5	criterion: entropy
kernel: linear	weights: uniform	min sample leaf: 5
degree: 1		max depth: 2
tollerance: 0.01		min impurity: 0.01

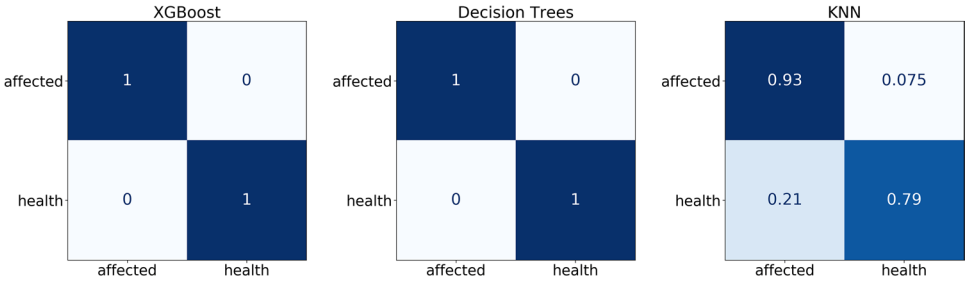

Fig. 5 Confusion matrix: Xgboost, Decision Trees and KNN

can provide information in advance, especially on how the probability or class of belonging is been calculated. Having said that we can present the results of the six classifiers used: the **XGBoost** and the **Decision Tree** showed the highest predictive capacity in terms of accuracy, both reaching **100%** on the test data, while the **Logistic Regression** and the **SVC** around **98.3%**. The **Neural Network** obtained an accuracy of **97%** and finally the **KNN** a score of **93%**. It is possible to visually compare the results in the Figs. 5 and 6.

As a statistical measure of interest we can consider the **F1-Score**, defined as the harmonic mean of *precision* and *recall*. This metric is clearly in line with good accuracy, as it takes into account, as a test, the predictive ability of the model to recognize true positives (target = 1) with respect to the total of true positives and false positives (target = 0); this measure is called *precision*, and the model's ability to recognize true positives and false negatives. This last quantity is called *recall*. The F1-Score has a maximum value equal to 1, and in this work the **XGBoost** and the **Decision Tree** obtain this value, while 0.98 for the **Logistic regression** and the **SVC**; 0.97 for the **Neural Network** and 0.94 for the **KNN**. In Fig. 7 it is possible to compare the results of the models on both train and test data.

3.6 Importance and explainable features

The part of explainability and importance of features in a CDSS system is fundamental. Being able to analyze and understand the influence of every single component of the decision-making process in the final predictions is impractical. In this regard, this section presents a features analysis framework, based on the importance of the characteristics for each model used, with the exception of the KNN, which by its nature is a very simple algorithm in its functionality. The concept of feature importance is calculated using a decision tree approach, using XGB and considering the quality of the prediction with respect to each feature inserted and removed during each iteration. For each of the features that make up the training dataset, on which the models have been trained, a score is provided, which represents the importance that variable had for predicting the target variable. The framework is based on the following steps:

(a) features importance is evaluated for each algorithm
(b) the score of each variable for each algorithm used is weighted and averaged
(c) a general final score is obtained for each feature
(d) the confidence interval for each feature is calculated with a confidence level of 95 % assuming a Gaussian distribution for the mean parameter

In the Table 2 it is possible to observe the importance score for each feature, with respect to the method used depending on the type of classifier used. The mean and standard deviation of the distribution of the values associated with the feature score were calculated and the 95 % confidence interval of the "importance" parameter was calculated. The first ten features show a significant importance in statistical terms: factors such as insulin intake, duration of the disease and HbA1c values it is noted that they obtained very high scores, thus indicating an importance (as it is reasonable to expect) of these characteristics. in order to build a CDSS useful both to predict and to explain the pathology. BMI, mother's education, weight and height, as well as area

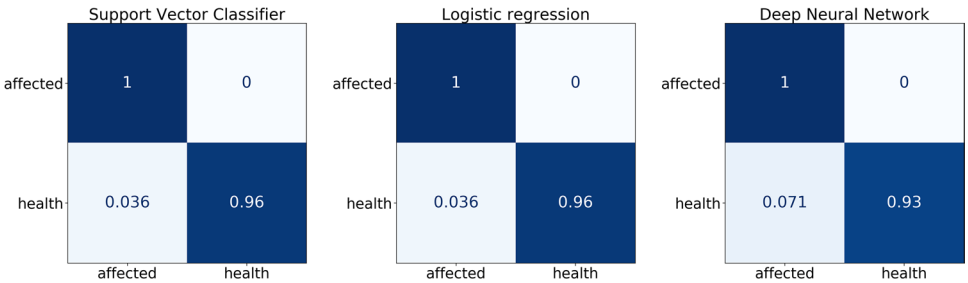

Fig. 6 Confusion matrix: SVC, Logistic regression and DNN

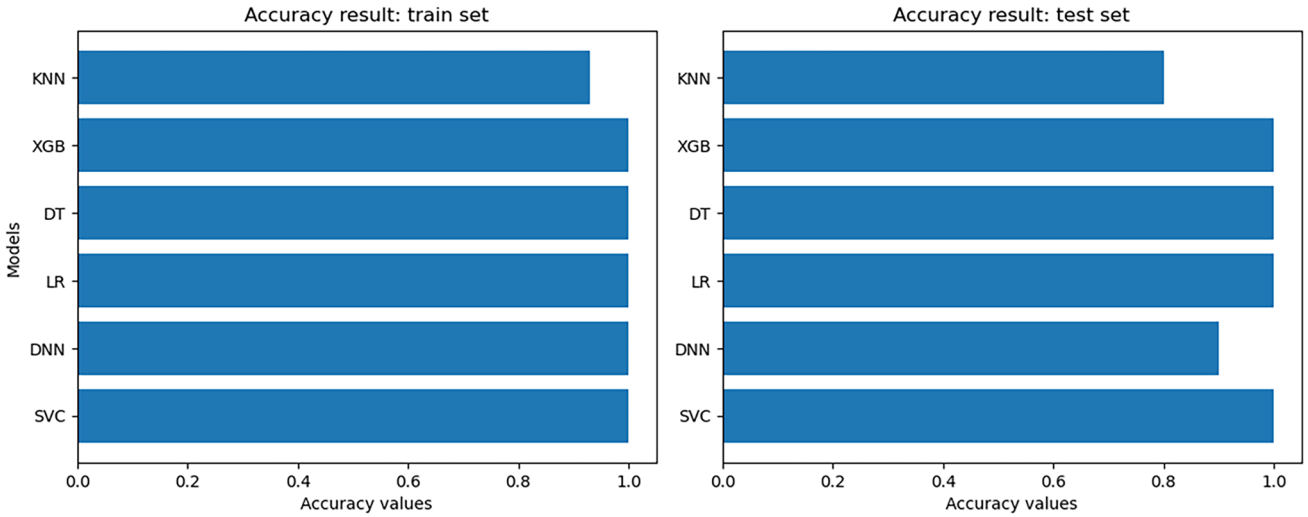

Fig. 7 Accuracy values on train (90%) and test (10%) set

of residence and adequate nutrition do not appear to make a significant contribution to CDSS, with scores close to zero or even negative. In the barplot of Fig. 8 it is possible to observe the contribution of importance for those features with a score greater than 10 %.

As regards the methods known as Explainable AI or Explainable ML, the **LIME** (Local Interpretable Model-Agnostic Explanations) method was applied in order to deduce how the probability predicted by the classifier is influenced by certain features and how these components of the model black-boxes explain the result of the prediction. The Table 3 compares the contribution of each feature given a given condition, for a patient with diabetes, belonging to the group of cases and a healthy person without

Table 2 features importance and confidence intervals

Feature	XGBoost	DT	SVC	LR	DNN	Mean	CI 95%
Insulin taken	1	0.781	0.904	1	0.701	0.877	[0.866,0.888]
Duration disease	0.961	0.125	1	0.768	0.689	0.709	[0.679,0.738]
HbA1c	0.433	0.425	0.565	0.648	0.475	0.509	[0.501,0.517]
Hypoglycemis	0.452	0.423	0.417	0.459	0.317	0.414	[0.409,0.418]
Pancreatic disease in child	0.140	0.142	0.432	0.410	0.283	0.281	[0.270,0.293]
How Taken	0.518	1	-0.904	-0.909	1	0.140	[0.059,0.222]
Family History Type 2 Diabetes	0.017	0.004	0.215	0.228	0.194	0.132	[0.122,0.141]
Age	0.230	0.048	0.089	0.018	0.262	0.130	[0.120,0.139]
Height	0.222	0.031	0.001	0.114	0.206	0.115	[0.107,0.123]
Family History Type 1 Diabetes	0.015	0	0.141	0.103	0.111	0.074	[0.069,0.079]
Other diease	0.086	0.021	-0.140	-0.040	0.423	0.070	[0.052,0.088]
Weight	0.149	0.033	0.033	0.036	0.068	0.064	[0.060,0.068]
BMI	0.022	0	-0.079	-0.037	0.316	0.044	[0.031,0.057]
Standardized birth weight	0.013	0.003	-0.035	0.050	0.154	0.037	[0.031,0.043]
Education of Mother	0.145	0.028	-0.228	-0.122	0.348	0.034	[0.015,0.053]
Impaired glucose metabolism	0	0.001	-0.01	0.021	0.055	0.013	[0.011,0.015]
Growth-rate in infancy	0.029	0	-0.139	0.1	0.041	0.007	[0.001,0.015]
Sex	0.067	0	-0.202	-0.237	0.315	-0.010	[-0.028,0.008]
Autoantibodies	0.142	0.101	-0.233	-0.135	0.040	-0.016	[-0.030,-0.003]
Adequate Nutrition	0.097	0.096	-0.276	-0.215	0.192	-0.021	[-0.038,-0.003]
Area of Residence	0.126	0.060	-0.567	-0.508	0.777	-0.022	[-0.068,0.023]

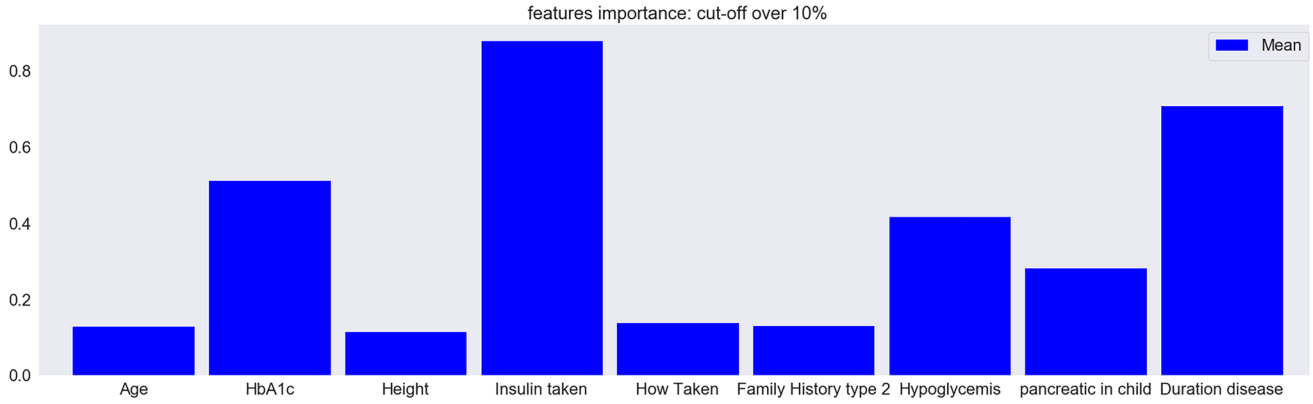

Fig. 8 Most important features selected

diabetic disease belonging to the control group; ID20 and ID15 respectively. The first is a male patient under the age of 15 without regressed pathologies, suffering from type 1 diabetes, with no hereditary history, with an average BMI and with a full-blown disease for 4 years. The second is a male individual under the age of 11, residing in the urban area of the country, in a state of obesity with a very high BMI, who does not have diabetic pathology, as can be seen from the Tables 4 and 5. The tables cited are closely related to the results obtained and published in the Table 3, from which it is possible to deduce how, for a given range of values assumed by the reference feature, the probability of developing the disease is increases or vice versa decreases. Patient ID20, as estimated by the model, has a **95%** probability of developing the disease while patient ID15 has a **2%** probability; this is the test carried out on the validation data of the model, in order to show the ability of the model to distinguish a case of diabetes from a non-diabetic one. It is therefore possible to observe how the model is highly efficient and the features are consistent with the pathological state of the subject analyzed. In the case of the ID20 patient we can observe that the duration of the disease, the insulin intake and the modality, not randomly, are factors with a certain weight: specularly you do not take insulin and the modality of assumption are factors that decrease the possibility of risk of developing the disease, as can be seen in patient ID15. Therefore, in order to build a transparent and explainable CDSS, classifiers such as **XGBoost**, used in the prediction and use of the LIME method, while being quite complex methods compared to linear but highly accurate methods, allow to be locally explained and analyzed in every their component. Clearly, for each classifier used it is possible to obtain a probability and the respective explanation, both of the outcome and of the features, locally and globally and this can be done for the entire population under consideration. In order to show the validity of the method, only two patients randomly extracted from the validation data population were analyzed.

Table 3 Explanations comparison with LIME

Condition (Case)	Contribution: ID20	Condition (Control)	Contribution: ID15
$0 < \text{Insulin taken} \leq 1$	0.28	$\text{Insulin taken} \leq 0$	-0.274
$\text{How Taken} \leq 0$	0.227	$\text{Duration disease} \leq 0$	-0.251
$0 < \text{Duration disease} \leq 4$	0.168	$0 < \text{How Taken} \leq 1$	-0.22
$\text{Hypoglycemis} < 0$	-0.141	$\text{Hypoglycemis} \leq 0$	-0.144
$0 < \text{HbA1c} \leq 1$	0.109	$\text{HbA1c} \leq 0$	-0.116
$\text{Height} > 1.53$	0.064	$\text{Height} \leq 1.22$	-0.067
$\text{Weight} > 51$	0.048	$\text{Weight} \leq 22$	-0.05
$0.00 < \text{Sex} \leq 1$	-0.041	$\text{Other disease} \leq 36$	0.039
$\text{pancreatic in child} \leq 0$	-0.039	$0 < \text{Sex} \leq 1$	-0.037
$\text{Other disease} \leq 36$	0.034	$\text{pancreatic in child} \leq 0$	-0.036

Table 4 Explanations values with LIME: Case ID20

Feature	Value	Feature	Value
Age	greater then 15	Standardized growth-rate in infancy	Highest quartiles
Sex	Male	Standardized birth weight	Middle quartiles
Area of Residence	Rural	Autoantibodies	Yes
HbA1c	Over 7.5%	Impaired glucose metabolism	No
Height	1.71	Insulin taken	Yes
Weight	57	How Taken	Injection
BMI	19.49	Family History type 1	No
Other disease	none	Family History type 2	No
Adequate Nutrition	Yes	Hypoglycemis	No
Education of Mother	No	Duration disease	4 years

Table 5 Explanations values with LIME: Control ID15

Feature	Value	Feature	Value
Age	Less then 11	Standardized growth-rate in infancy	Middle quartiles
Sex	Male	Standardized birth weight	Middle quartiles
Area of Residence	Urban	Autoantibodies	Yes
HbA1c	Less then 7.5%	Impaired glucose metabolism	No
Height	0.61	Insulin taken	Yes
Weight	15	How Taken	None
BMI	40.31	Family History type 1	No
Other disease	none	Family History type 2	No
Adequate Nutrition	Yes	Hypoglycemis	No
Education of Mother	Yes	Duration disease	0 years

4 Conclusions

Different types of supervised algorithms have been put in place in order to draw an exhaustive picture of a clinical decision-making process based on artificial intelligence methods. Models such as Logistic regression certainly offer interesting and easily interpretable predictive capabilities, but which suffer from overfitting and not very high accuracy. Same description for classifiers such as KNN, very simple in its logic and on the learning part based on a very simple cost function to be minimized; it has been observed that this method is very simple to explain, but that it suffers from poor accuracy on test data (unseen). Decision trees is a well-known method, its logic is perhaps the simplest to explain, and this justifies its wide use, as well as the possibility of using it for the part relating to features importance. XGBoost is certainly the most widely used algorithm in supervised machine learning problems and the results obtained confirm it, also based on so-called adjusted decision trees. Both the predictive ability and the ability to explain the features and results obtained. Models such as the SVM are strongly affected by the sample size and above all by the number of features involved (since we remember in the optimization part the function is quadratic and the method used in the resolution is known as *dual* form, and this method is certainly onerous from the computational point of view), although for small datasets such as the one treated, the performances have been remarkable. The Deep neural network has a high accuracy and a very high number of parameters involved and thanks to explanability methods such as the one used, LIME, it is possible to give an exhaustive explanation of the results obtained. Therefore it is possible to employ all these methods and obtain satisfactory results; it is also possible to insert in the pipeline of a CDSS also methods such as LIME in order to complete the picture and have all the elements to be able to make a decision on a meaningful basis. From the statistical point of view, the estimates obtained in Table 2 and the contributions to the probabilities assigned as a function of each features with respect to the classifier used in Table 3, are robust and consistent. The results obtained on two observations, one of the control group and one of the patients, were compared. From this comparison it is possible to deduce significant differences and to infer particular characteristics from which conclusions can be drawn that a classifier in itself cannot extract, since as already mentioned the result obtained in probabilistic terms if not supported by an adequate explanation, remains an end to itself and of little use.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12553-023-00781-z.

Declarations

Conflicts of interest The author declares under his own responsibility that there are no conflicts of interest in the realization of this work.

References

1. O'Connor PJ, Sperl-Hillen JM, Rush WA, Johnson PE, Amundson GH, Asche SE, Ekstrom HL, Gilmer TP. Impact of electronic health record clinical decision support on diabetes care: a randomized trial. Ann Fam Med. 2011.
2. Georga E, Protopappas V, Arvaniti E, Fotiadis D. The Diabino System: Temporal Pattern Mining from Diabetes Healthcare and Daily Self-monitoring Data; 2019.
3. Rung-Ching C, Hui Qin J, Chung-Yi H, Cho-Tsan B. Clinical Decision Support System for Diabetes Based on Ontology Reasoning and TOPSIS Analysis. Artif Intell Med Appl. 2017.
4. Amatul Z, Asmawaty AK, Aznan MAM. A comparative study on the pre-processing and mining of Pima Indian diabetes dataset. 2013.
5. International Diabetes Federation. 2021. https://idf.org/news/diabetes-now-affects-one-in-10-adults-worldwide/.
6. Pima Indian Diabetes Database, Schulz LO, Bennett PH, Ravussin E, Kidd JR, Kidd KK, Esparza J, Valencia ME. Effects of traditional and western environments on prevalence of type 2 diabetes in Pima Indians in Mexico and the US. Diabetes Care. 2006;29(8):1866–71. https://doi.org/10.2337/dc06-0138.
7. Deepti S, Dilip SS. Prediction of Diabetes using Classification Algorithms. Procedia Comput Sci. 2018;132:1578–85.
8. Han W, Shengqi Y, Zhangqin H, Jian H, Xiaoyi W. Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlocked. 2018;10:100–7.
9. Arrieta B, Rodriguez ADN, Del Ser J, Bennetot A, Tabik SB, González A, García S, Gil-López S, Molina D, Benjamins VR, Chatila RH, Francisco. Explainable Artificial Intelligence (XAI): Concepts. Opportunities and Challenges toward Responsible AI: Taxonomies; 2019.
10. Patil BM, Joshi RC, Durga T. Hybrid prediction model for Type-2 diabetic patients. Expert Syst Appl. 2010;37(12):8102–8.
11. Dagliati A, Marini S, Sacchi L, et al. Machine Learning Methods to Predict Diabetes Complications. J Diabetes Sci Technol. 2018;12(2):295–302. https://doi.org/10.1177/1932296817706375.
12. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus With Machine Learning Techniques. Front Genet. 2018;9:515. https://www.frontiersin.org/article/10.3389/fgene.2018.00515.
13. Butt UM, Letchmunan S, Ali M, et al. Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. J Healthc Eng Hindawi. 2021.
14. Asaduzzaman S, Al Masud F, Bhuiyan T, Ahmed K, Paul BK, Matiur Rahman SAM. Dataset on significant risk factors for Type 1 Diabetes: a Bangladeshi perspective. Data Brief. 2018;21:700–8 ISSN 2352-3409.