Comparison of Machine Learning Methods for a Diabetes Prediction Information System

Olexandr Shmatko [1], Olha Korol [2], Andrey Tkachov[3], Vasyl Otenko[4]

[1] *National Technical University "Kharkiv Polytechnic Institute" st. Kirpichova, 2, Kharkiv, 61000, Ukraine*
[2,3,4] *Simon Kuznets Kharkiv National University of Economics, ave. Nauki, 9-A, Kharkiv, 61166 Ukraine*

Abstract

Diabetes is a disease for which there is no permanent cure; therefore, methods and information systems are required for its early detection. This paper proposes an information system for predicting diabetes based on the use of data mining methods and machine learning (ML) algorithms. The paper discusses a number of machine learning methods such as decision trees (DT), logistic regression (LR), k-Nearest Neighbors (k-NN). For our research, we used the Pima Indian Diabetes (PID) dataset collected from the UCI machine learning repository. The dataset contains information about 768 patients and their corresponding nine unique attributes. Research has been carried out to improve the prediction index based on the Recursive Feature Elimination method. We found that the logistic regression (LR) model performed well in predicting diabetes. We have shown that in order to use the created model to predict the likelihood of diabetes mellitus with an accuracy of 78%, it is necessary and sufficient to use such indicators of the patient's health status as the number of times of pregnancy, the concentration of glucose in the blood plasma during the oral glucose tolerance test, the BMI index and the result of the calculation. heredity functions "DiabetesPedigreeFunction".

Keywords

Machine learning, Data Mining, Neural Network, Diabetes Prediction Information System, KNN, Logistic regression, Decision tree.

1. Introduction

Diabetes mellitus is an "epidemic of the XXI century", an incurable disease of the pancreas, which develops due to absolute or relative insufficiency of the hormone insulin. It is characterized by a steady rise in blood glucose levels, which in turn can lead to complications.

To achieve compensation for diabetes, constant monitoring is required. In addition to taking oral medications and insulin, following a strict diet, exercise, daily routine, checking your blood glucose regularly, and keeping a special diary, your diabetic should see an endocrinologist regularly for advice and appropriate measures to improve or maintain the condition.

Normally, the level of glucose in the blood varies within fairly narrow limits: from 3.3 to 5.5 mmol / liter. This is due to the fact that in a healthy person, the pancreas produces or stops the release of insulin depending on the actual level of glucose in the blood. In case of insufficiency or complete absence of insulin (type 1 diabetes mellitus) or in case of impaired interaction of insulin with cells (type 2 diabetes mellitus), glucose accumulates in the blood in large quantities, and the body's cells are unable to absorb it. As of 2019, in addition to the already mentioned type 1 and type 2 diabetes, there are gestational diabetes (gestational diabetes), MODY-diabetes and LADA diabetes [2].

ISIT 2021: II International Scientific and Practical Conference «Intellectual Systems and Information Technologies», September 13–19, 2021, Odesa, Ukraine
EMAIL: oleksandr.shmatko@khpi.edu.ua (A. 1); korol.olha2016@gmail.com (A. 2); andrew.tkachov@hneu.net (A. 3); ovi@hneu.edu.ua (A. 4)
ORCID: 0000-0002-2426-900X (A. 1); 0000-0002-8733-9984 (A. 2); 0000-0003-1428-0173 (A. 3); 0000-0002-5979-1084 (A. 4)

Depending on the specifics of the diagnosis, treatment of patients with diabetes involves the use of oral agents to improve insulin permeability to body tissues and / or replacement therapy with subcutaneous insulin injections of varying duration to mimic the normal functioning of the pancreas. With mild diabetes, you can do without medication, but a strict diet with a clear understanding of the amount of nutrients consumed, moderate exercise, daily routine, blood glucose control and diary of self-monitoring are mandatory for all patients with this diagnosis.

Under conditions of poor or insufficient treatment (decompensation or subcompensation of diabetes), blood glucose levels in the human body are consistently high. Against this background, complications of diabetes develop, which not only worsen the patient's standard of living, but can also be fatal. These complications include:

ketoacidosis (accumulation of a dangerously large number of ketone bodies in the blood), hypoglycemia (decrease in blood glucose below 3.3 mmol / l), hyperosmolar and lactic acidotic coma, polyneuropathy (peripheral nerve damage), nephropathy (kidney damage), retina retinal vessels), angiopathy (impaired vascular permeability), diabetic foot syndrome, etc.

To achieve compensation for diabetes - a condition in which the patient has achieved stable normal blood glucose levels during treatment and the risk of complications is reduced - constant monitoring is required. In addition to the above measures, this control also includes regular visits to the endocrinologist for advice and appropriate measures to improve or maintain the patient's health.

2. Literature review

There are a number of studies on predicting diabetes based on machine learning (ML) methods for the Pima Indian Diabetes Dataset (PIDD). Pima Indian Diabetes Dataset (PIDD) containing: 9 attributes, 768 records describing female patients. [1], [2], [3], [4], [5].

In [2], artificial neural networks were used to predict diabetes based on the PIDD dataset, which showed a prediction accuracy of 75.7%. Sajida The authors of [3] showed that among the applied machine learning methods SVM, NB and DT on PIDD, the NB classifier shows the best accuracy - 76.30%. [4] applied logistic regression to PIDD to predict diabetes. The model proposed in this paper showed a fairly good forecast with an accuracy of 75.32%. In the study [5], all patient data were used to train and test a classifier based on Naive Bayes (NB) and decision trees (DT). The research results showed that the best algorithm is the naive Bayesian algorithm with an accuracy of 76.3021%.

The most important problem in a machine learning method is the choice of training parameters and the corresponding classifier. In our work, we used the Recursive Feature Elimination method to improve the prediction rate. Our research work is to select the best classifier for the diabetes prediction information system. In this work, various machine learning classification algorithms are used to predict diabetes in a patient, such as Linear Regression (LR), K-Nearest Neighbor (KNN), Decision Tree (DT).

3. System design

The system architecture for the Diabetes Prediction System, shown in Figure 1 below, is a conceptual model that defines the structure, behavioral interactions, and several systemic representations that underlie the system. The figure shows a formal description of the system, submodules of the system, as well as data flows between them.

Figure 1 shows the components of the system architecture.

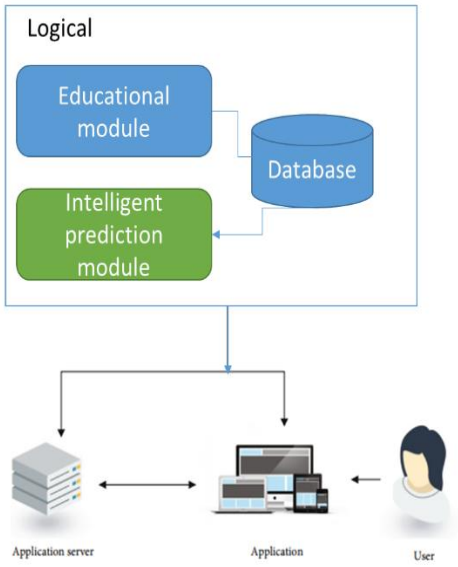

Figure 1: System architecture.

4. Methods

Based on the comparison and analysis of the functional properties of leading software solutions in the field of medicine, it was determined that the option "Obtaining prediction of the probability of the patient's disease" is not implemented in modern diabetes management information systems. However, due to statistics on the fate of patients with misdiagnosis, it becomes impossible to deny the need to implement this useful function.

The problem of predicting the incidence of diabetes can be solved using the methods of classification of machine learning.

In the tasks of medical diagnostics, patients act as objects. The characteristic description of the patient is, in fact, a formalized medical history. Having accumulated a sufficient number of precedents in electronic form, you can use the methods of classification of machine learning and predict the likelihood of the patient's disease.

4.1 An example of solving the problem of classification using machine learning to predict the incidence of diabetes

4.1.1. Description of the source data

To implement the considered methods of classification of machine learning, we will use the popular service "UCI Machine Learning Repository", which provides a large number of sets of real data, and consider the initial data presented in the sample "Pima Indians Diabetes Database" (figure 2)

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 2: Example data in the Pima Indians Diabetes Database sample

There are a total of 768 records in the sample, each of which is characterized by the following nine parameters.

1. "Pregnancies" - the number of times of pregnancy
2. "Glucose" - plasma glucose concentration (in mg / dl) for two hours in an oral glucose tolerance test
3. "BlooodPressure" - diastolic blood pressure (in mm Hg)
4. "SkinThickness" - the thickness of the folds of the skin of the triceps (in mm)
5. "Insulin" - the concentration of serum insulin for two hours (in μg / ml)
6. "BMI" - body mass index, calculated by the formula: weight in kg / (height in m) 2
7. "DiabetesPedigreeFunction" - a function of diabetes heredity
8. "Age" - the age of man
9. "Outcome" - the result of a variable class (0 - no diabetes, 1 - a sick person)

The available data show the following distribution: 500 people are healthy (ie their "Outcome" parameter is zero) and 268 have diabetes (their "Outcome" parameter is equal to one).

In graphical form, the data "Pima Indians Diabetes Database" can be represented as follows (figure 3).

As can be seen from Figure 3, inaccurate data are found in the sample. For example, these are:

1. blood pressure equal to zero (35 cases);
2. zero blood glucose concentration (5 cases);
3. skin fold thickness less than 10 mm (227 cases);
4. BMI approaching zero (11 cases);
5. zero level of insulin concentration in the blood (374 cases).

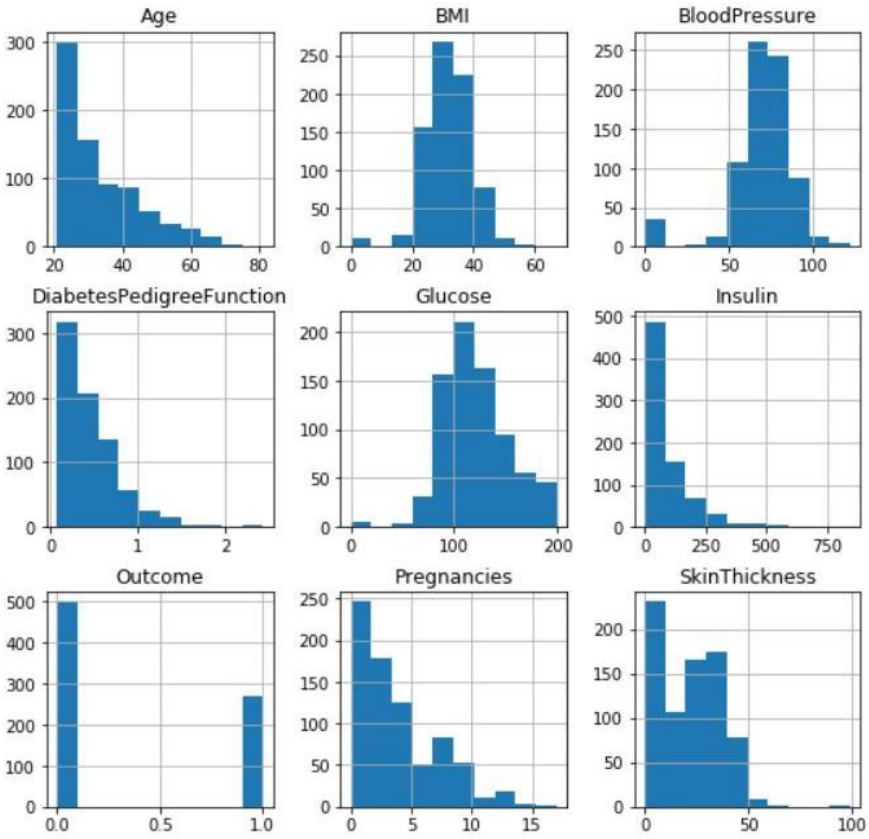

Figure 3: Graphic representation of data distribution

To eliminate the above problems, the following options are proposed:

- Delete or ignore records. An undesirable option, because it means the loss of valuable information. The sample contains too many records with zero skin thickness and blood insulin concentration, but this tool can be applied to the fields "BMI", "Glucose", "Blood pressure".
- Using averages. This option may be the case for some samples, but using a mean value, such as blood pressure, will be the wrong signal for the model.
- Avoid the use of problematic characteristics. This option could work for the thickness of the skin, but at this stage it is difficult to predict the result.

After analyzing possible ways to solve the problem of incomplete data, we decide to remove from the sample rows in which the attributes "BMI", "Glucose" and "Blood pressure" are zero. As a result, 724 records remain in the database.

4.1.2. Choice of classification method

In order to choose the method of classification of machine learning, which is better suited for the task of predicting the incidence of diabetes, it is advisable to choose the method whose accuracy in the selected sample will be the highest.

Avoidance of training and testing on the same data is a common practice, which is explained by the fact that the purpose of the model is to provide data other than the sample. In addition, the model can be overly complex, leading to retraining. To avoid the above problems, there are two precautionary methods:

• retention method - part of the training set can be postponed and used as an affirmative (test) set;

• cross-checking - repeating the method of retention several times, ie repeating the division of the sample into training and approval sets.

Calculations of the accuracy values of the selected classification methods will be performed using Python programming language, namely using the methods of the library "scikit-learn" [39]. As input parameters "x" we will give models data from columns "Pregnancies", "Glucose", "Blood Pressure", "Skin Thickness", "Insulin", "BMI", "Diabetes Pedigree Function" and "Age". As the expected result "y" - data from "Outcome".

The results of the calculations presented in the table 1.

Table 1

The results of calculating the accuracy of the classification methods by the method of retention and the method of cross-checking

Method	retention method	cross-checking
KNN	0.711521	0.711521
LR	0.776440	0.776440
DT	0.681327	0.685494

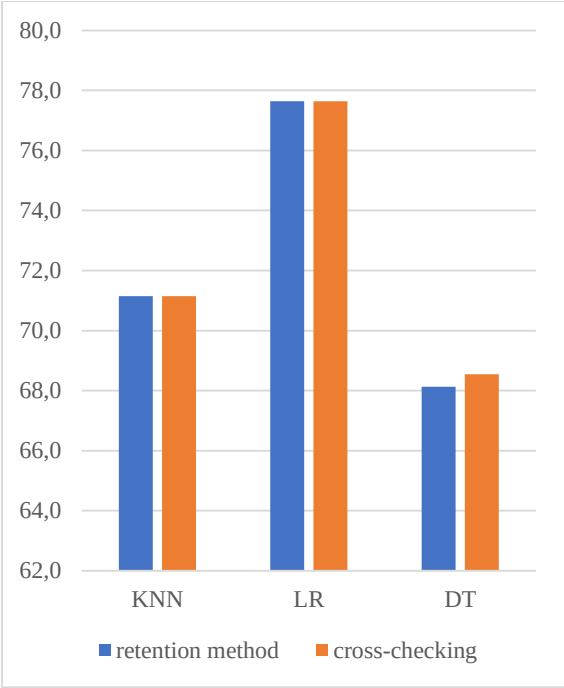

Figure 4: Comparison of the obtained results of accuracy of classification methods

Based on the obtained data, it can be stated that the method of logistic regression has a higher accuracy of the sample "Pima Indians Diabetes Database" than the method of kNN-classification and the method of the decision tree (figure 4), and therefore it can be used to implement the function "Diabetes prediction" in the information system to automate the process of admission of patients with diabetes in endocrinologists.

4.1.3. Improving the accuracy of prediction

Usually, not all source data improve the performance of the model. In order to correctly identify which of the available attributes have a greater impact on the resulting model, we use the method of recursive Feature Elimination (RFE).

The essence of the method is that it recursively removes attributes and builds models based on those attributes that remain. RFE uses model accuracy to determine which attributes or combinations of them contribute most to target prediction.

Using the library "scikit-learn" we build a graph of the accuracy of the prediction of diabetes from the number of initial parameters (Fig. 5).

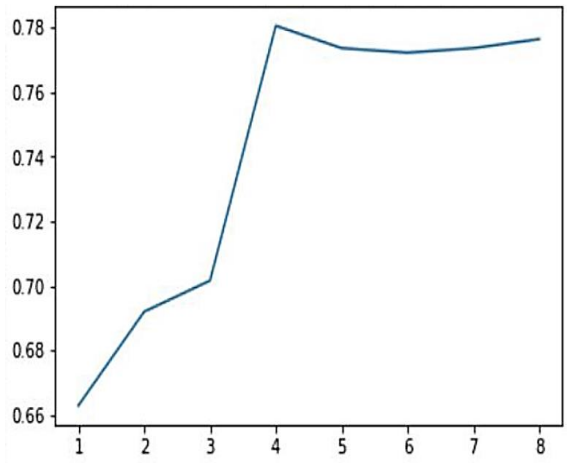

Figure 5: Dependence of accuracy of prediction of diabetes mellitus by logistic regression method on the number of initial parameters

As you can see, the best accuracy of the model is achieved by using only four attributes: "Pregnancies", "Glucose", "BMI", "DiabetesPedigree Function". A comparison of the results of the accuracy calculation is given in the following table 2/

Table 2

The results of calculating the accuracy of the method of logistic regression after reducing the number of output attributes

Number of output parameters	8	4
Accuracy of logistic regression method	0.776440	0.780588

Thus, to further use the created model to predict the probability of diabetes with an

accuracy of 78%, it is necessary and sufficient to use such indicators of the patient's health as the number of pregnancies, plasma glucose concentration in the oral glucose tolerance test, BMI and the result of calculating the heredity function "DiabetesPedigreeFunction".

5. Conclusions

Early detection of diabetes is one of the major health problems. This paper proposes a system architecture and classifier for an information system that can predict diabetes with high accuracy. We have pre-processed the data. Using the method of reducing the number of functions, we have abandoned four parameters. We used four input parameters ("Pregnancies", "Glucose", "BMI", "DiabetesPedigree Function") and one output parameter (result) in the PIMA dataset. We used three different machine learning algorithms, including DT, KNN, LR on PIDD, to predict diabetes and evaluated the performance on various parameters. All models show good results in some parameters. All models provided over 70% accuracy. LR provided an accuracy of approximately 77–78%. The use of improving the prediction index based on the Recursive Feature Elimination method allowed us to reduce the number of parameters from 8 to 4. Among all the proposed models, the forecasting accuracy for logistic regression (78.05%) was better than the accuracy in [1] (LR 75.32%), [2] (NB - 76.30%), [3] (NB - 76.3021%), [4] (RF - 77.21%) and [5] (ANN 75.7%).

6. References

[1] Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." Procedia computer science 132 (2018): 1578-1585.

[2] Alam, Talha Mahboob, et al. "A model for early prediction of diabetes." Informatics in Medicine Unlocked 16 (2019): 100204.

[3] Tigga, Neha Prerna, and Shruti Garg. "Prediction of type 2 diabetes using machine learning classification methods." Procedia Computer Science 167 (2020): 706-716.

[4] Diwani, Salim Amour, and Anael Sam. "Diabetes forecasting using supervised learning techniques." Adv Comput Sci an Int J 3 (2014): 10-18.

[5] Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.