UDC 004.891.3 **doi: 10.26906/SUNZ.2021.4.073**

Olexander Shmatko[1], Anna Goloskokova[1], Olha Korol[2], Irada Rahimova[3]

[1] National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine
[2] Azerbaijan Technical University, Baku, Azerbaijan
[3] Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine

# COMPARISON OF MACHINE LEARNING METHODS FOR A DIABETES PREDICTION INFORMATION SYSTEM

**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 algorithms. The paper discusses a number of machine learning methods such as random forest, AdaBoost algorithm, multilayer perceptron, neuro–like structure of Consecutive Geometric Transformations Models (CGTM), linear regression based on the stochastic gradient descent, generalized regression neural network and regression based on the support vector machine. The Pima Indian Diabetes dataset collected from the UCI machine learning repository was used in the research. The dataset contains information about 768 patients and their corresponding nine unique attributes: the number of times of pregnancy; plasma glucose concentration for two hours in an oral glucose tolerance test; diastolic blood pressure; the thickness of the folds of the skin of the triceps; the concentration of serum insulin for two hours; body mass index; a function of diabetes heredity; the age of a person; the result of a variable class (0 – no diabetes, 1 – a sick person). The research has been carried out to improve the prediction index based on the Recursive Feature Elimination method. It was found that the logistic regression model performed well in predicting diabetes. It has been 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 of the heredity functions "Diabetes Pedigree Function".

**Keywords**: machine learning, data mining, neural network, diabetes prediction information system, logistic regression, decision tree.

## 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 [1].

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.

**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. [1-5].

In [1], artificial neural networks were used to predict diabetes based on the PIDD dataset, which showed a prediction accuracy of 75.7%. The author of [3] showed that among the applied methods of machine learning, like support vector machine, naive Bayesian classifier, decision trees on PIDD and naive Bayesian classifier shows the best accuracy – 76.30%. In [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 and decision trees. 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 this paper, the Recursive Feature Elimination method was used 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 random forest, AdaBoost algorithm, multilayer perceptron, neuro–like structure of Consecutive Geometric Transformations Models (CGTM), linear regression based on the stochastic gradient descent, generalized regression neural network and regression based on the support vector machine.

## Main part

**System design.** The system architecture for the Diabetes Prediction System, shown in Fig. 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. Fig. 1 shows the components of the system architecture.

**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.

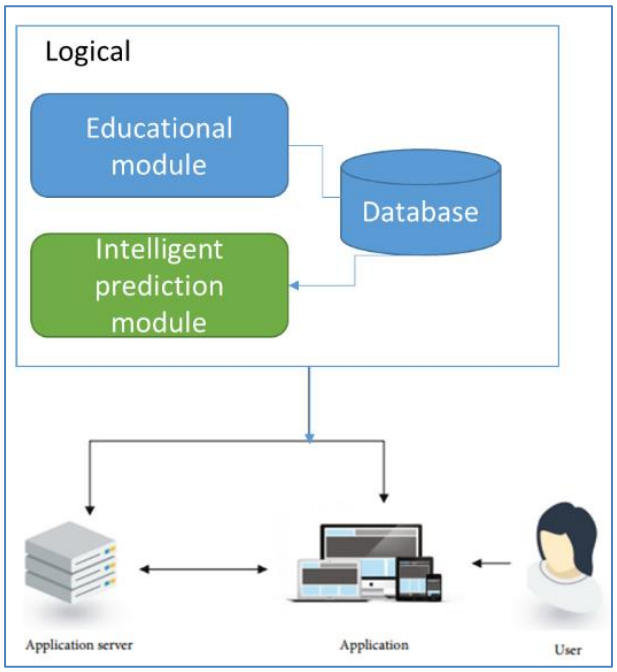


**Fig. 1.** System architecture

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.

**An example of solving the problem of classification using machine learning to predict the incidence of diabetes. 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" (Fig. 2).

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

| | 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 |

**Fig. 2.** Example data in the Pima Indians Diabetes Database sample

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

- "Age" – the age of a woman.

The available data show the following distribution: 500 people are healthy (i.e. 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 (Fig. 3).

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

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

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.

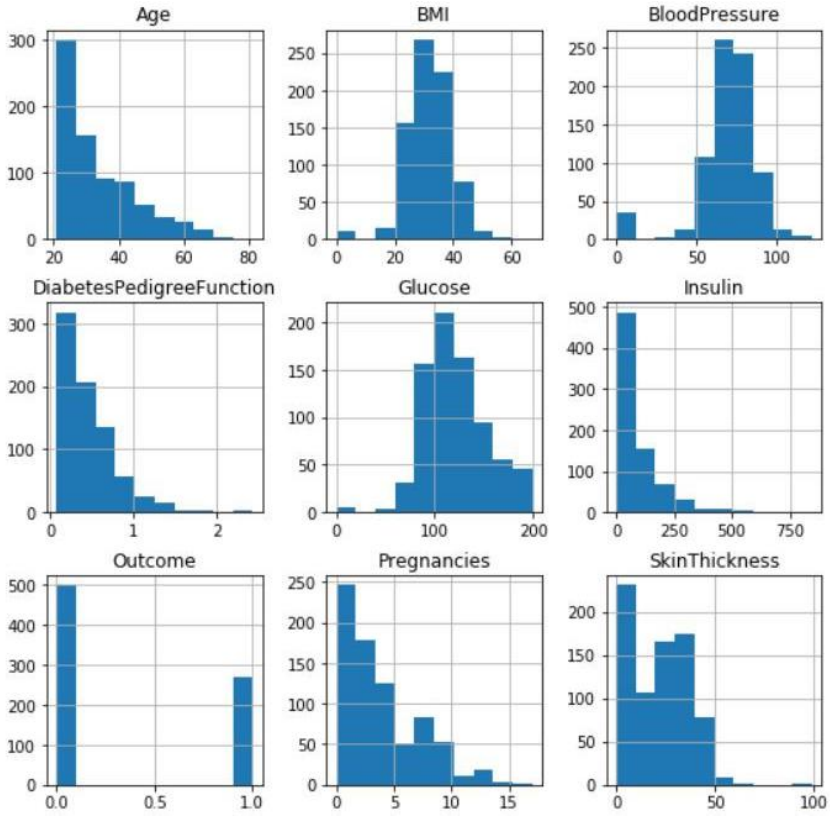


**Fig. 3.** Graphic representation of data distribution

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.

**Choice of classification method.** The methods of machine learning, which were studied in the article, are presented in table. 1

The first group of methods is ensemble. The paper covers the study of two different classes of these methods: bagging, which includes the Random forest algorithm, and boosting, which is represented by the AdaBoost algorithm. The peculiarity of Random forest is that it provides a stable and effective solution while minimizing problems with retraining [6]. In addition, it is resistant to emissions and scaling and is able to process large data sets with many features of each input vector [7]. Among the disadvantages of this method should be noted the lack of extrapolative properties and difficult interpretation of the results [6].

The composition of the AdaBoost algorithm involves an iterative process of building private models, where each subsequent algorithm is learned using information about the errors made in the previous stage [8]. Simplicity of implementation and high ability to generalize are the main advantages of the AdaBoost algorithm. Among the limitations of the algorithm should be noted the need for no noise in the data, which can lead to retraining [9]. In addition, an important role in the efficiency of this algorithm is played by the dimension of the sample for training [8]. The next group of methods for solving the regression problem under study is neural network. The possibility of using a multilayer perceptron, a generalized regression neural network and a neuro–like structure of a model of successive geometric transformations to predict the amount of costs of retail store

consumers is considered. Despite the possibility of a fairly accurate approximation, the multilayer perceptron is characterized by the durability of the training procedure for its iterative process. The generalized regression neural network is fast, but characteristics such as the size and structure of the data sample, the quality of the algorithm and software solution based on it, the use of parallelism, etc., in some cases make the network very large and slow. In addition, in the Random forest algorithm, it is not capable of extrapolating data [10]. The application of neuro–like structures of the model of successive geometric transformations to the solution of regression problems is characterized by a high speed of implementation of learning procedures and sufficient accuracy of the forecast [11]. However, large amounts of data may limit the use of this computing tool.

*Table 1* – **Parameters of the studied machine learning algorithms**

| Number of the method | The name of the machine learning method | Symbol of the method | Method parameters |
|---|---|---|---|
| 1 | Random Forest | Method 1 | maximum depth of the tree = 5 |
| 2 | AdaBoost algorithm | Method 2 | basic algorithm — decision tree (maximum depth = 4), number of weak (base) trees = 300 |
| 3 | Multilayer perceptron | Method 3 | 23 inputs, 23 neurons in the hidden layer, 1 output |
| 4 | Neuro–like structure of Consecutive Geometric Transformations Models (CGTM) | Method 4 | 23 inputs, 23 neurons in the hidden layer, 1 output |
| 5 | Linear regression based on the stochastic gradient descent | Method 5 | waste function = 'squared_loss', α = 0.0001 |
| 6 | Generalized regression neural network | Method 6 | σ = 0.4 (σ ϵ [0.1,1.5] |
| 7 | Regressor based on the support vector machine | Method 7 | kernel = rbf, epsilon = 0.001, maximum number of iterations = 200 |

Linear regression based on stochastic gradient descent, like the support vector machine, is characterized by a high speed [12], but not always by satisfactory prediction results.

**Development of the problem solving concept based on the machine learning algorithms.** MN or an artificial intelligence (AI), is a research discipline that deals with the methods and algorithms of experiential learning. For some researchers, the phrase MN is part of AI, provided that the ability to learn is a crude attribute of an intellectual individual.

The purpose of machine learning is to develop computer systems that can learn and respond based on their previous observations. The purpose of artificial intelligence is to develop an intelligent agent or assistant that uses a variety of machine–based learning methods based on the solution.

Database knowledge research (DBKR) is a discipline that includes hypotheses, approaches, and strategies

$$RMSE = \sqrt{\sum_{i=1}^{n}\left(y_i^{pred} - y_i^{true}\right)^2}, \qquad (1)$$

that seek to understand data and extract valuable facts from it. It is known that this is a multistage method (selection, pre–processing, transformation, MN / AI, understanding / evaluation), defined in Fig. 4.

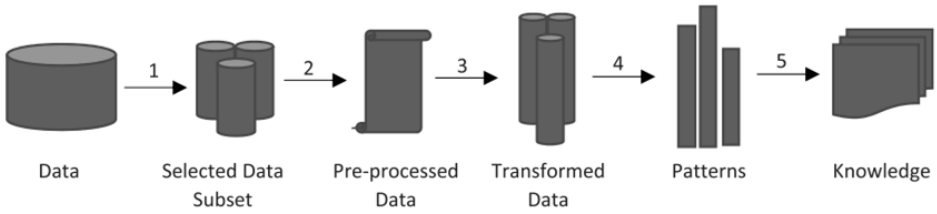


**Fig. 4.** Multistage method of knowledge extraction

The most critical phase in the whole DBKR method is MN / AI, which illustrates the use of MN and AI algorithms in data processing. MN processes are usually divided into three specific groups. These include:

- supervised learning (SL), where the scheme indicates the functionality of the marked learning data;
- unsupervised learning (UL), where the system tries to deduce the nature of unidentified data;
- reinforced learning (RL), where the machine interacts with a dynamic context. Artificial intelligence is used to develop intellectual assistants who will help in self–management and personalization of disease therapy.

**Evaluation of research results.** The simulation of the available ML methods (table 1) was performed on a real data set "Pima Indians Diabetes Database".

The simulation was performed on a data set from which all observations with spaces were removed. The data sample was randomly divided into training and test. The percentage of such division was 70% to 30%, respectively.

All parameters of the studied methods are presented in table 1. Evaluation of the effectiveness of the forecasting task was based on two indicators:

- The root–mean–square error (RMSE):
- Duration of the training procedure.

In Fig. 5 presents a comparison of the values of the root mean square error of all studied methods.

The values of the error are plotted on the x–axis, and the studied methods – on the y–axis. The black (dark) columns of the histogram indicate the error obtained in

the training mode (rmse training), grey (light), respectively, the error of the application mode (rmse test).

The worst result on accuracy of the decision of the set task is shown by the method constructed on the basis of the support vector machine (fig. 5). The best result is obtained using the AdaBoost algorithm. The difference in accuracy based on (fig. 5) between the two methods is more than 28%. The Random forest algorithm in comparison with AdaBoost shows slightly worse results in the application mode, but much better in the training mode. It should be noted that in Fig. 5 the error in the training mode for method 6 does not show, because it does not require training as such. However, as can be seen from Fig. 5, the accuracy of its operation is unsatisfactory.

In addition to the accuracy of work, no less important feature of the information system is the ability to work online. That is why the duration of the training procedures of the methods underlying such systems is critical. Given this, the study conducted an experimental study to assess this indicator.

Figure 6 shows the duration of training procedures for all studied methods (in seconds).

As can be seen from Fig. 6, the training procedure of method 5 is very fast, and the duration of the learning algorithms underlying methods 1 and 2, respectively, is 6 and 12 times slower. The usage of a multilayer perceptron to solve the problem leads to significant time delays. In particular, it is slower than method 1 by more than 549 times. However, the worst results, given the operating time, were obtained using a neural network of generalized regression. Its usage for the problem solution lasts more than 130 seconds.

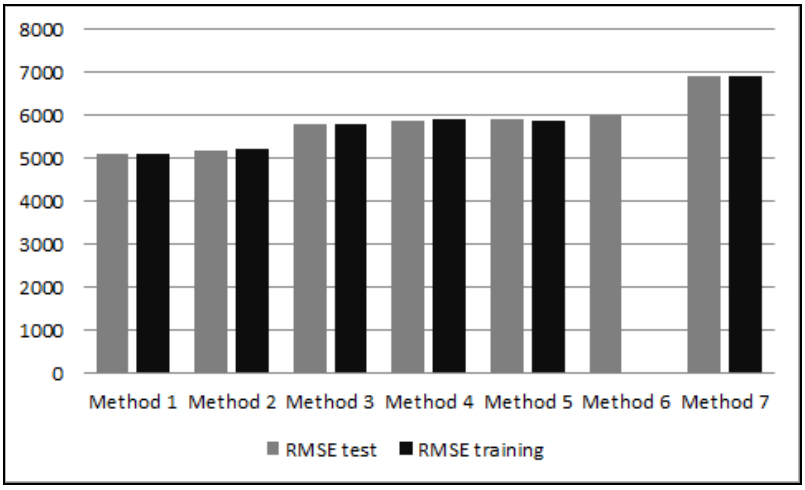


**Fig. 5.** The root mean square error value of the RMSE modes of training and usage

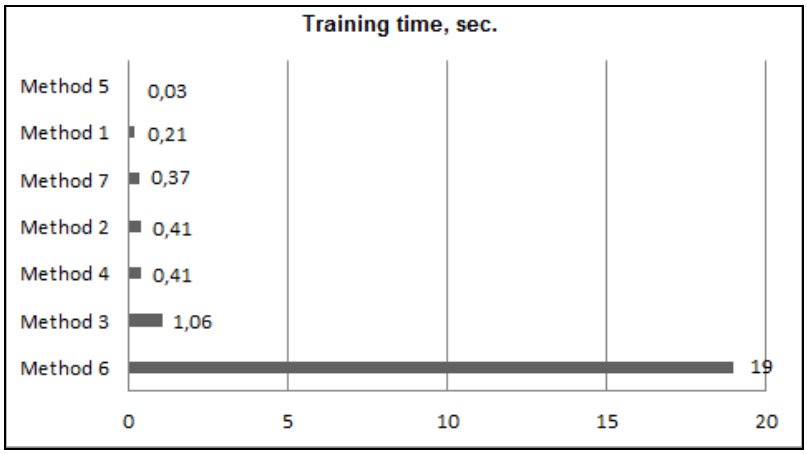


**Fig. 6.** Duration of training procedures of the studied methods

Based on the analysis of both the accuracy of ML methods (Fig. 5) and the duration of their training procedures (Fig. 6), it can be argued that the most effective solution to the problem is provided by ensemble methods according to the schemes of improved association (bagging), and improved cross–section (boosting).

That is, methods 1 and 2, respectively. Artificial neural networks (methods 3 and 4) do not provide sufficient accuracy.

## Висновки

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 abandoned four parameters.

We used four input parameters ("Pregnancy", "Glucose", "BMI", "Pedigree function of diabetes") and one

output parameter ("Result") in the PIMA data set. Seven different machine learning algorithms were used, including random forest, AdaBoost algorithm, multilayer perceptron, neuro–like structure of Consecutive Geometric Transformations Models (CGTM), linear regression based on the stochastic gradient descent, generalized regression neural network and regression based on the support vector machine for providing the diabetes prediction. The performance of these models was evaluated by various parameters.

Thus, the accuracy of the models was assessed both in the training mode and in the test mode. The best performance on these parameters was determined in models developed by the methods of AdaBoost and Random Forest. The worst result in terms of the accuracy of solving the problem was showed by a method based on the machine of reference vectors. Another indicator that has been studied was the assessment of the training duration. The least time was spent on learning the method of linear regression based on stochastic gradient descent, while the training of models based on the building and training of neural networks is the slowest. Taking into account the analysis of methods for both indicators, it is established that the most effective solution of the problem is provided by ensemble methods according to the schemes of both improved bugging and improved boosting, i.e. Random Forest and AdaBoost methods, respectively.

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.

REFERENCES

1. Alam, Talha Mahboob, et al. A model for early prediction of diabetes. Informatics in Medicine Unlocked. 2019. No. 16. P. 100–204.
2. Sisodia, Deepti, and Dilip Singh Sisodia. Prediction of diabetes using classification algorithms. Procedia computer science. 2008. No. 132. P. 1578–1585.
3. Tigga, Neha Prerna, and Shruti Garg. Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*. 2020. No. 167. P. 706–716.
4. Diwani, Salim Amour, and Anael Sam. Diabetes forecasting using supervised learning techniques. *Adv Comput Sci an Int J*. 2014. No. 3. P. 10–18.
5. Zou, Quan, et al. Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*. 2018. No. 9. P. 515–523.
6. Joshi R., Gupte R., Saravanan P. A Random Forest Approach for Predicting Online Buying Behavior of Indian Customers. *Theoretical Economics Letters*. 2018. No. 08. P. 448–456.
7. Wu X., Meng S. E–commerce customer churn prediction based on improved SMOTE and AdaBoost. *13th International Conference on Service Systems and Service Management (ICSSSM)*. Kunming. 2016. P. 1–5.
8. Cao Y., Miao Q–C., Liu J–C., Gao L. Advance and Prospects of AdaBoost Algorithm. *Acta Automatica Sinica*. 2013. Vol. 39. No. 6. P. 745–758.
9. Alomair O. A., Garrouch A. A. A general regression neural network model offers reliable prediction of CO2 minimum miscibility pressure. *Journal Petrol Explor Prod Technol*. 2016. No. 6. P. 351–365.
10 Tkachenko R., Izonin I. Model and Principles for the Implementation of Neural–Like Structures Based on Geometric Data Transformations. *Advances in Computer Science for Engineering and Education. Springer International Publishing*, Cham. 2019. P. 578–587.
11. Izonin I., Trostianchyn A. et al. The Combined Use of the Wiener Polynomial and SVM for Material Classification Task in Medical Implants Production. *International Journal of Intelligent Systems and Applications*. 2018. No. 10. P. 40–47.
12. Tepla T. L., Izonin I. V., Duriagina Z. A. et al. Alloys selection based of the supervised learning technique for design of biocompatible medical materials. *Archives of Materials Science and Engineering*. 2018. No. 1. P. 32–40.

Received (Надійшла) 28.10.2021
Accepted for publication (Прийнята до друку) 24.11.2021

**Порівняння методів машинного навчання для інформаційної системи прогнозування діабету**

О. В. Шматко, А. О. Голоскокова, О. Г. Король, І. Р. Рагімова

**Аннотація.** Діабет – це хвороба, від якої немає постійного лікування; тому для його раннього виявлення потрібні методи та інформаційні системи. У цій статті пропонується інформаційна система для прогнозування діабету, яка ґрунтується на використанні методів інтелектуального аналізу даних та алгоритмів машинного навчання. У статті обговорюється ряд методів машинного навчання, таких як випадковий ліс, алгоритм AdaBoost, багатошаровий персептрон, нейроподібна структура моделей послідовних геометричних перетворень, лінійна регресія на основі градієнтного стохастичного спуску, узагальнена регресійна нейронна мережа і регресія на основі машини опорних векторів. Для даного дослідження був використаний набір даних Pima Indian Diabetes, зібраний із репозиторію машинного навчання UCI. Набір даних містить інформацію про 768 пацієнтів та їх відповідні дев'ять унікальних атрибутів: кількість вагітностей; концентрація глюкози у плазмі протягом двох годин при пероральному тесті на толерантність до глюкози; діастолічний артеріальний тиск; товщина складок шкіри трицепса; концентрація інсуліну сироватки за дві години; індекс маси тіла; функція спадковості при діабеті; вік людини; результат змінної класу (0 – немає діабету, 1 – хворий). Були проведені дослідження щодо покращення індексу прогнозування на основі методу виключення рекурсивних ознак. Виявлено, що модель логістичної регресії добре підходить для прогнозування діабету. Показано, що для використання створеної моделі для прогнозування ймовірності цукрового діабету з точністю до 78% необхідно і достатньо використовувати такі показники стану здоров'я пацієнтки, як кількість вагітностей, концентрація глюкози в плазмі при пероральному тесті на толерантність до глюкози, індекс ІМТ та результат розрахунку функції спадковості «Родовідна функція діабету».

**Ключові слова:** машинне навчання, інтелектуальний аналіз даних, нейронна мережа, інформаційна система прогнозування діабету, логістична регресія, дерево рішень.