UDC 519.2
DOI: https://doi.org/10.17721/1812-5409.2024/1.23

Hanna LIVINSKA, PhD (Phys. & Math.), Assoc. Prof.
ORCID ID: 0000-0001-9676-7932
e-mail: hanna.livinska@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Daria SKRYPNYK, Student
e-mail: skrypnikdaria@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

José Luis GALÁN-GARCÍA, PhD (Math.), Assoc. Prof.
ORCID ID: 0000-0002-8773-6998
e-mail: jlgalan@uma.es
Universidad de Málaga, Málaga, Spain

COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR DIABETES DETECTION USING BINARY DATA

The prevalence of diabetes is constantly increasing, and timely diagnosis of this disease is extremely important in the health care system. Nowadays, in addition to advanced medical technologies, there is an opportunity to process large volumes of information quickly and efficiently, in particular medical information. Patients with diabetes, even in the early stages, have certain similar symptoms and patterns, which makes it possible to effectively diagnose the disease, based on easily obtained clinical data.

In this work, we compare the accuracy of different machine learning algorithms for diagnosing diabetes on binary data that do not include laboratory tests, and can be obtained by patient questionnaires. All the classification models give good results and show the feasibility of identifying individuals probably having undiagnosed diabetes, based on questionnaire data. The best results are provided by random forest classifier. In order to improve the accuracy of the classification, an analysis of the relationships between the variables of the dataset was carried out. This made it possible to reduce the dimensionality of the model by removing variables that do not carry useful information. Moreover, the data set is additionally balanced. The resulting models demonstrate higher efficiency than the classification models built on the base dataset.

K e y w o r d s : ***Diabetes prediction, random forest classifier, machine learning, classification accuracy.***

AMS 2020 classification: 62H30, 68T09.

Introduction

Diabetes Mellitus, commonly known as diabetes (Diabetes), is a chronic disease caused by a violation of glucose metabolism, which manifests itself in a decrease in the sensitivity of body tissues to insulin, a violation of glucose absorption by tissues, and damage to various organs and systems. There are several types of diabetes, but the most common are type 1, type 2, and gestational diabetes. The level of glucose in the blood of patients with diabetes increases, and with further progression of the disease, a high level of glucose in the blood leads to disturbances in the activity of the cardiovascular system, nervous and immune systems. Diabetes can have a serious impact on patients' health and quality of life. Uncontrolled diabetes causes permanent damage to the body and can cause other serious health problems such as cardiovascular disease, chronic kidney failure, vision problems, etc. People with diabetes may also experience reduced quality of life, emotional difficulties, and limitations in eating and physical activity.

The prevalence of diabetes is steadily increasing. Type 2 diabetes used to be called adult-onset diabetes because it occurs more often in older people. But the increase in the number of obese children has also led to an increase in the incidence of type 2 diabetes in young patients. Some studies have shown that environmental pollutants may also play a role in the development and progression of type 2 diabetes (Hectors et al., 2011). According to the World Health Organization, at the beginning of 2021, about 422 million people in the world were living with diabetes, which is approximately 8,5 % of the population. This number is projected to grow to 700 million by 2045. So, in the modern world, the problem of diabetes is one of the most important in the field of medicine, because the growing prevalence of this disease endangers the health and quality of life of millions of people.

Because of the long asymptomatic phase, early detection of diabetes is always critical to achieving clinically meaningful outcomes. About half of all people suffering from diabetes do not have a confirmed diagnosis due to the long symptom-free period of this disease. Symptoms of diabetes can be mild, and many patients do not notice them for years. Typical symptoms of diabetes include polyuria (frequent and abundant urination), polydipsia (constant feeling of thirst), signs of dehydration, which are mostly moderately pronounced (decreased skin elasticity, dryness of the skin), weakness caused by dehydration, apathy, weight loss, constant feeling of hunger, decreased vision, tingling or numbness in the hands or feet, ketoacidosis, susceptibility to purulent skin infections or infections of the genitourinary system. These symptoms can appear in various combinations. And during a physical examination in the early stages of diabetes, diabetes is often not diagnosed. However, if it is not treated, it leads to the development of chronic complications, especially cardiovascular complications, which are the main cause of death. In the later stages of diabetes, damage to certain organs is often detected as a consequence of the disease.

Nowadays, modern equipment and technologies make it possible to collect and process incredibly large data sets. This allows the discovery of hidden useful knowledge and patterns that may be important for understanding diseases, predicting treatment outcomes, and improving medical procedures. Accordingly, multivariate statistical analysis in medicine can help solve many urgent problems, in particular, problems of classification and diagnosis of diseases. The use of data analysis and machine learning methods can significantly improve the diagnosis of diabetes and help to take timely measures for its prevention and treatment. Due to the relevance of this direction, machine learning methods have been actively developed in recent years (Sadhu, & Jadli, 2021; Battineni et al., 2019; Wei, Zhao, & Miao, 2018; Fregoso-Aparicio et al., 2021; Kangra, & Singh, 2023; Aguilera-Venegas et al., 2023 etc.). Works in this area demonstrate high efficiency of the classification models for various data sets.

ISSN 1812-5409 (Print), ISSN 2218-2055 (Online)

Our work is devoted to the analysis of the dataset from (Islam et al., 2020), available in Kaggle, which contains mostly binary data (except for the patient's age), formed without additional laboratory examinations based on a questionnaire. Binary data have certain specifics in their processing and interpretation, not all methods of data analysis are effective for them, so the choice of technique should be careful. For classification tasks for binary data, the technique built on decision trees is recognized as the best one.

Our research involves analyzing available medical data, selecting, and configuring appropriate data analysis methods and machine learning models for predicting the presence of diabetes in patients, and evaluating the effectiveness of the developed method. First, we conduct pre-processing and correlation analysis of the available information, identify variables that have a significant impact or interfere with effective classification. Comparing different models of classifiers, we obtained the highest accuracy when building the Random Forest classifier based on the set of classification accuracy indicators. Additional balancing of the dataset made it possible to avoid underestimation of the classification error. Removing variables of insignificant influence on the classification result yields classification accuracy compared to the classification results of the base dataset.

This paper is organized as follows. In Section 2, we give a description of the used dataset. Some basic assumptions and ideas about the methods of analysis are given. Descriptive and correlation analysis are provided in Section 3. In Section 4, some classification accuracy evaluation metrics are described, and the obtained classification accuracy metrics are compared to select the optimal classification algorithm for our data. The classification results after some improving of the data are given and analyzed as well. The obtained accuracy of the improved model is shown. Finally, conclusions are given in Section 5.

1. Description of the data and selection of research methods. The "Early Diabetes Classification" dataset (Kaggle) is available on the Kaggle platform and is intended for conducting research in the field of medical diagnosis of diabetes. This dataset contains 520 observations of 17 traits collected through direct surveys and patient diagnoses at a diabetes hospital in Sylhet, Bangladesh. The data has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor.

The data includes a variety of biomedical indicators and characteristics that may be important for the diagnosis of this disease. This data set includes information on both patients with a confirmed diagnosis of diabetes and those without diabetes:

Age (16-90); Sex (1. Male, 2.Female); Polyuria (1.Yes, 2.No); Polydipsia (1.Yes, 2.No); Sudden weight loss (1.Yes, 2.No); Weakness (1.Yes, 2.No); Polyphagia (1.Yes, 2.No); Genital thrush (1.Yes, 2.No); Visual blurring (1.Yes, 2.No); Itching (1.Yes, 2.No); Irritability (1.Yes, 2.No); Delayed healing (1.Yes, 2.No); Partial paresis (1.Yes, 2.No); Muscle stiffness (1.Yes, 2.No); Alopecia (1.Yes, 2.No); Obesity (1.Yes, 2.No); Class (1.Positive, 2.Negative).

The "Class" variable here is the target variable that we predict. It takes the value 1 if diabetes is present and 0 if diabetes is absent.

To visualize the distribution of patients according to various characteristics, we use pie charts and bar charts.

When studying the relationship between variables, the correct choice of the correlation coefficient is important. Obviously, classical methods, such as Pearson's correlation, or Spearman's or Kendall's coefficients, which are used to analyze the dependence of categorical variables, are not effective for measuring the dependence between binary variables. Among the coefficients that are intended for working with binary variables, we chose Phi-coefficient.

The Phi correlation coefficient is a statistical index used to measure the degree of association between two binary variables (variables with two possible values, such as "yes" and "no", "1" and "0"). The advantages of the Phi correlation coefficient for binary data are its interpretability and convenience. It provides a numerical estimate of the degree of relationship between two categorical variables, which makes it easy to compare different pairs of variables and determine whether there is a statistically significant relationship between them, especially in the case of categorical and binary variables.

To detect the presence of dependence between the quantitative variable "age" and the variable "class", we used the Wilcoxon rank test comparing the mean values of two independent samples.

To build classification models, we used traditional algorithms: random forest classifier (rf), extra trees classifier (et), decision tree classifier (dt), gradient boosting classifier (gbc), light gradient boosting machine (lightgbm), quadratic discriminant analysis (qda), ada boost classifier (abc), logistic regression (lr), naive bayes (nb), ridge classifier (rc), linear discriminant analysis (lda), k neighbors classifier (knn), SVM-linear kernel (svm). Note that in cases where the information used for classification is also mostly expressed by binary variables, the most logical is to use decision trees. Decision trees are used to create models that partition a data set based on sequential decision making and are a very powerful classification algorithm.

Classification tree is a popular machine learning technique used to partition data into categories or classes based on features. It belongs to supervised learning category. The idea is to break the data into smaller groups, using the most informative features at each step. In this case, the partitioning continues until a given stopping criterion is reached, such as the maximum depth of the tree or the minimum number of examples in each leaf node.

Advantages of the classification tree method for binary data include:

– Simplicity and interpretability: Decision trees are easy to interpret because they can be visualized as graphical structures. This makes it easy to understand which features are important for classification and how exactly they affect the result.

– Good for Binary Data: Decision trees are effectively used for binary data because they allow you to do binary splits at each level. This allows the model to easily separate the data into two classes.

– Ability to work with categorical and numerical data: Decision trees can process both categorical and numerical data, without the need to predict values or use additional data transformations.

– Performance on large data sets: Decision trees can be quite fast for training and classification, even on large amounts of data.

– Ability to be used in ensembles: Decision trees can be used in ensembles, such as a random forest (as in our case), to improve classification quality.

Considering the technique of decision trees, this method and the ensemble of trees (random forest) seem to be optimal for identifying key binary factors contributing to the presence of diabetes and creating prognostic rules, which was confirmed by the results of calculating the classification accuracy indicators for different methods. In our models, decision trees are used to classify patients into two classes (0 and 1) based on several binary features, such as polyuria, polydipsia, weight loss, and

others. This makes it possible to quickly and efficiently classify patients based on simple signs and quickly and efficiently make decisions about the possible presence of diabetes.

The algorithms were implemented using 'pycaret', 'pandas', 'seaborn', 'phik', 'scipy.stats', 'matplotlib' libraries of Python programming language. The dataset was divided into training and test parts in a ratio of 70:30. Data classification models were built using the above methods. In the model, a 10-fold cross-validation set is used. When training each tree, a random subsample of the data was used to avoid overtraining. Accuracy, recall, precision, and F1-score values were used to assess accuracy and compare the effectiveness of classification methods.

2. Descriptive and correlational analysis. First, we conduct an exploratory analysis of the data. Conducting exploratory analysis before predicting diabetes helps to understand the data, identify dependencies and anomalies, identify important variables, and prepare the data for further modeling and prediction. Figure 1 gives the proportion of sick and healthy patients (healthy patients hear and later mean those who have not been diagnosed with diabetes). Note that the dataset contains more cases of diabetes (320 positive values) than cases of its absence (200 negative values). This means that the dataset is not well balanced, and we will balance it when building the model for avoiding underestimation of probabilities of error.

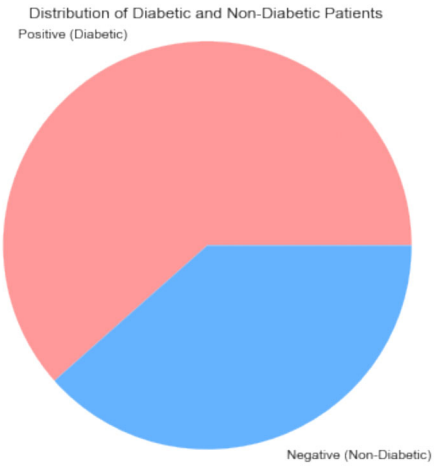

Fig. 1. Distribution of Diabetic and Non-Diabetic Patients

The boxplot in Figure 2 shows the distribution of patients according to the presence of the disease, age and sex, in Figure 3 shows histograms of the age distribution according to the presence of the disease.

Diagrams on Figures 4–7 show part of diabetic patients by different variables indicated in the dataset. Based on these graphs, it can be assumed that the presence of "polydipsia", a state of excessive thirst ("polyphagia"), "polyuria", and "sharp weight loss" are the most significant for the diagnosis of diabetes. Also, some influence of "age" and "gender" can be observed. On the other hand, the diagrams show that diabetes seems not dependent on "itching", "alopecia" and "delayed healing".

Next, we build a correlation map. We use the phi-coefficient, which is one of the best coefficients for measuring the degree of association between binary variables. It is calculated using the correspondingly normalized statistics chi-squared. The map acquires non-negative values. Values close to 1 indicate a stronger degree of connection between variables, values close to 0 indicate a weak connection or its absence.

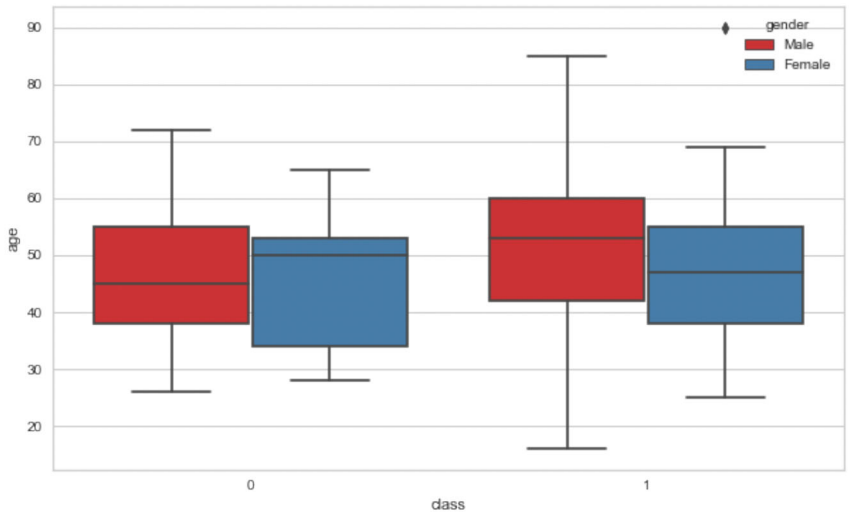

Fig. 2. Distribution of Diabetic and Non-Diabetic Patients by Age and Gender

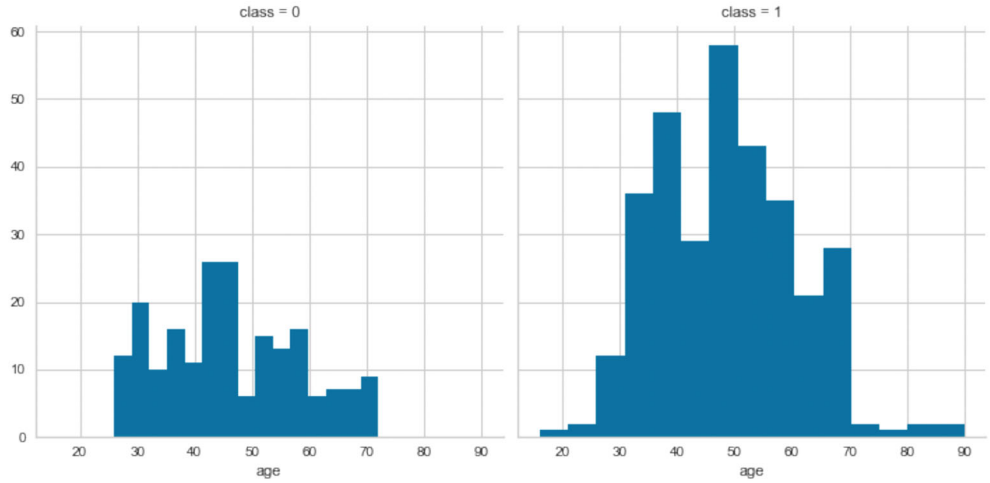

Fig. 3. Distribution of Patients Age by Class

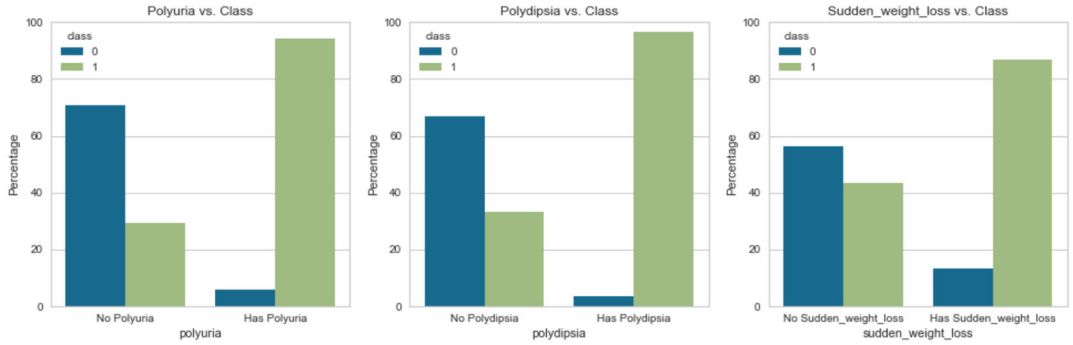

Fig. 4. Distribution of Patients with Polyuria, Polydipsia and Sudden Weight Loss by Class

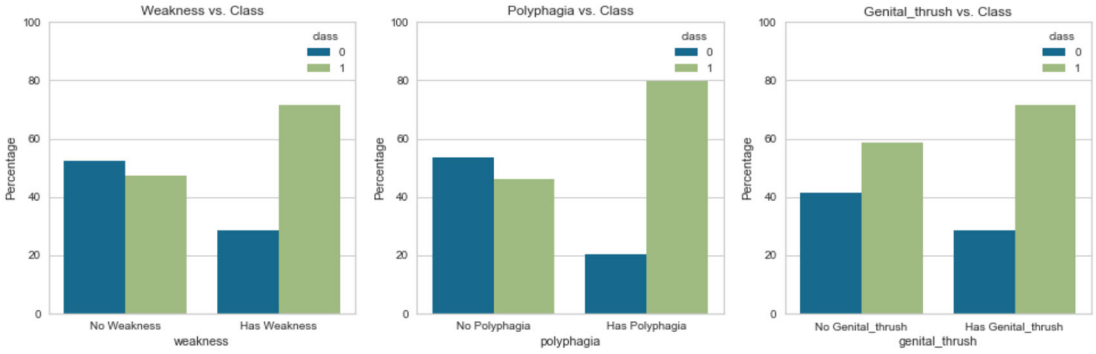

Fig. 5. Distribution of Patients with Weakness, Polyphagia and Genital Thrush by Class

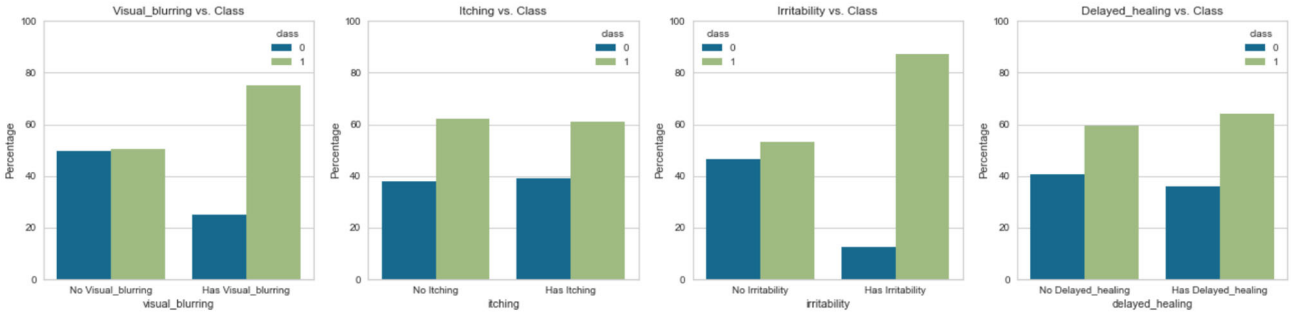

Fig. 6. Distribution of Patients with Visual Blurring, Itching, Irritability and Delayed Healing by Class

ISSN 1728-3817

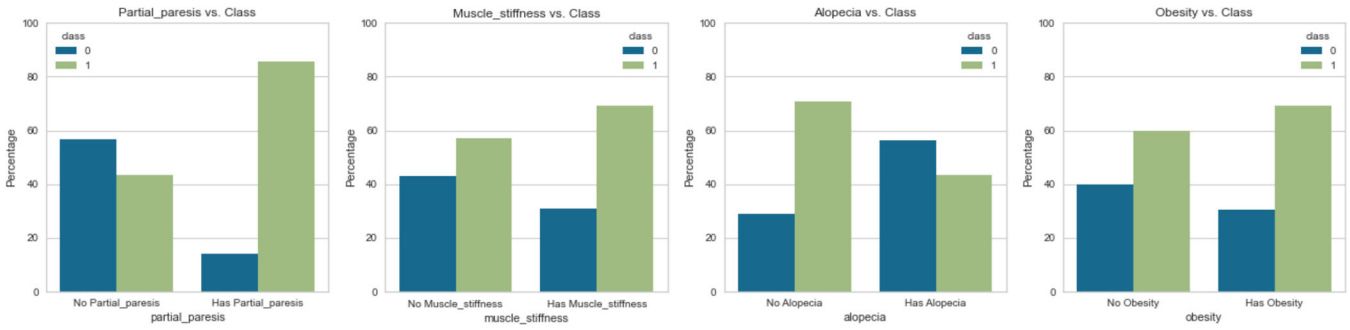

Fig. 7. Distribution of Patients with Partial Paresis, Muscle Stiffness, Alopecia and Obesity by Class

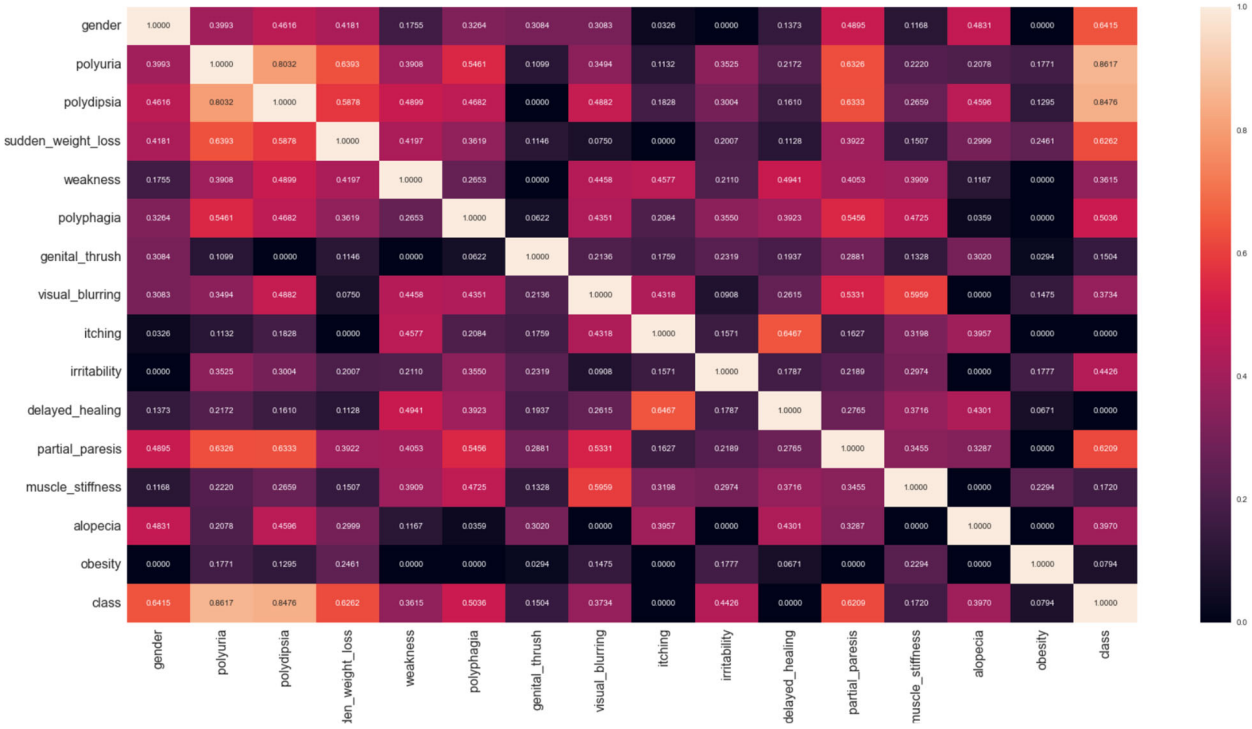

Fig. 8. Correlation Map with Phi-Coefficients

Figure 8 shows that our hypothesis about the strong influence of "polydipsia" and "polyuria" and the significant influence of "gender", "sudden weight loss", "partial paresis" on the presence of diabetes is true. Also, we see a significant effect of variables "weakness", "visual blurring", "irritability". However, there are doubts about the influence of "obesity", "itching", "delayed healing". Therefore, we will exclude these three variables from our basic model. Checking the equality of the average values of age in the groups of sick and healthy patients showed a statistically significant difference between these groups.

3. Classification models. We can now compare different machine learning models and determine which one is best suited to the task of predicting the presence of diabetes in our dataset. To evaluate the effectiveness of the model, we will use accuracy, recall, precision, and F1-score. Accuracy is computed based on the predicted values and true values. Recall is defined as how well the predictions are out of all the positive classes. Precision is defined as the number of classes that are positive from correctly predicted positive classes. F1-score is used to estimate precision and recall simultaneously. It gives possibility to compare models with low precision and high recall and vice-versa. Given the confusion matrix, the evaluation metrics are calculated as follows:

$$Recall = \frac{TP}{TP + FN}, \tag{1}$$

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

$$F1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}, \tag{4}$$

where TP (True Positive) is the number of predicted values are positive and expected values are true; TN (True Negative) is the number of predicted values are negative and the actual values are also true; FP (False Positive) is the number of predicted

values are positive when the actual values are false, it is also called Type 1 Error; FN (False Negative) is the predicted values are negative but the actual values are false, it is also called Type 2 Error. All the accuracy characteristics should be as high as possible. Figures 9 and 10 show accuracy scores and recall scores for different classification algorithms we use. Note that the most effective model is the random forest classifier. The obtained accuracy score for it is equal to 0.9696, recall score – 0.9686, F1-score – 0.9749, precision score – 0.9822. The optimal model consists of 100 decision trees.

Fig. 9. Accuracy Scores for Different Classification Algorithms

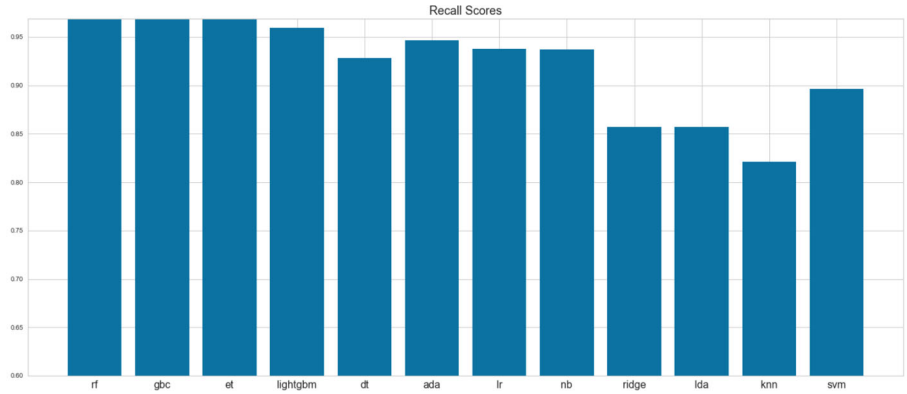

Fig. 10. Recall Scores for Different Classification Algorithms

The most significant signs for the classification of diabetes are shown in Figure 11. These are primarily "polyuria" and "polydipsia".

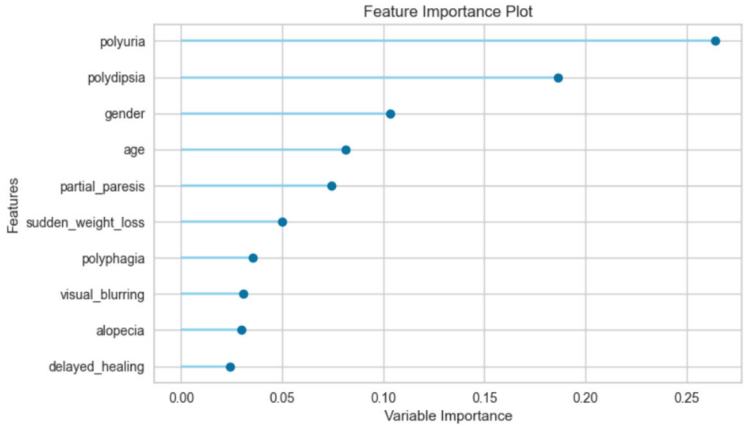

Fig. 11. Variable Importance in the Basic RF-Model

To reduce the dimension of the model, we remove non-influential signs for which the value of phi-coefficients is very small: "delayed healing", "itching" and "obesity". New models are built, and the accuracy of classification models are compared again.

Figures 12 and 13 show that random forest gives again best results, and now we have better accuracy. For random forest algorithm the accuracy score get value 0.9696, recall score – 0.9971, F1-score – 0.975, precision score – 0.9822. Actually, this model has less dimension but even better accuracy scores than the basic one.

Now let us balance the classes and repeat the algorithm. Figures 14 and 15 show accuracy results for the balanced model. We can see, that on the balanced data, the random forest classifier shows the best results in accuracy score. For random forest algorithm the accuracy score got value 0.9777, recall score – 0.9644, F1-score – 0.9774, precision score – 0.9913. The main values for the classification are shown in Figure16.

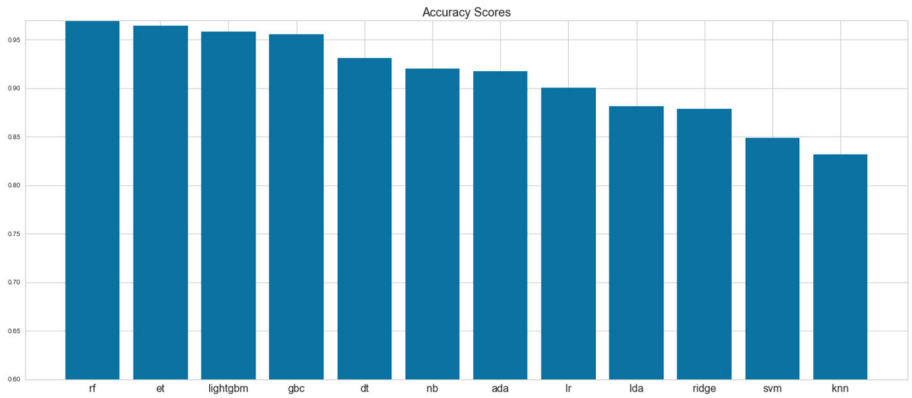

Fig. 12. Accuracy Scores for Model with Reduced Dimension

Fig. 13. Recall Scores for Model with Reduced Dimension

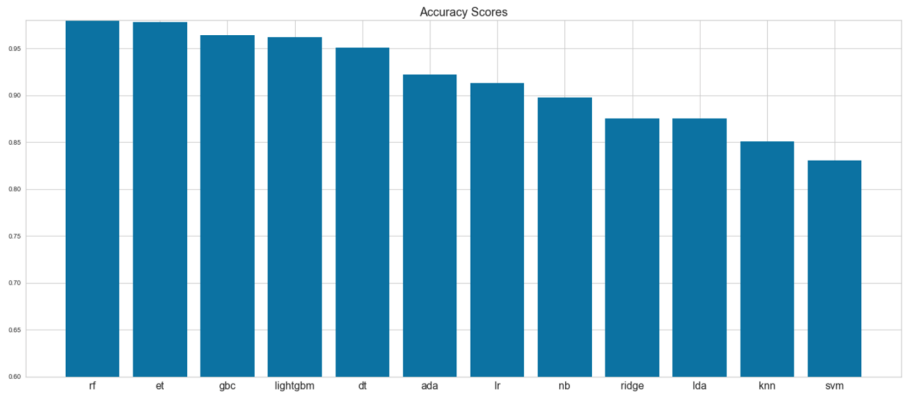

Fig. 14. Accuracy Scores for Balanced Model with Reduced Dimension

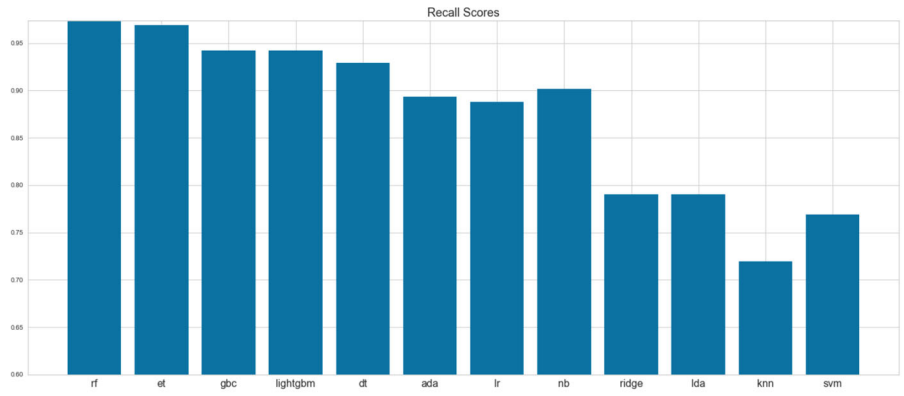

Fig. 15. Recall Scores for Balanced Model with Reduced Dimension

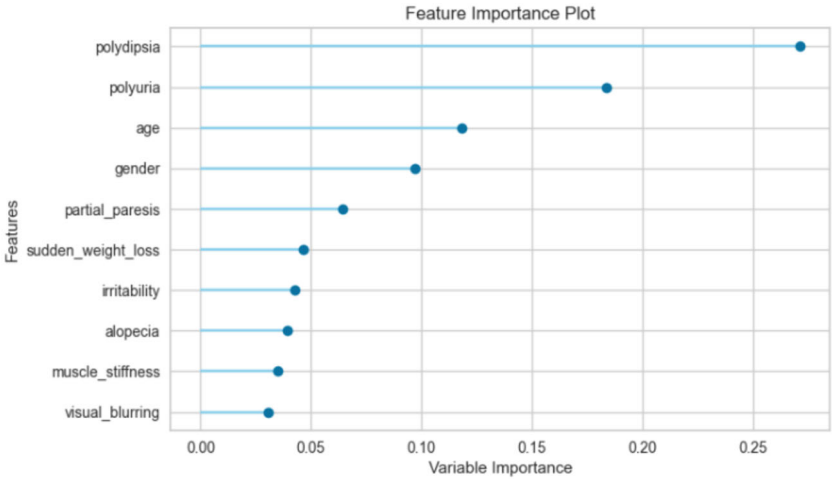

Fig. 16. Variable Importance in the Balanced Model with Reduced Dimension

Discussion and conclusions

Many cases of diabetes develop asymptomatically; therefore, it is recommended to focus on early diagnosis of this disease. Machine learning algorithms can greatly help doctors to establish a diagnosis even based on easily obtained medical questionnaire data. If diabetes is suspected, additional screening tests should be prescribed, especially in high-risk groups for the development of diabetes. Determining the type of diabetes and differential diagnosis are also very important.

All machine learning algorithms in the classification of diabetes give a good result, but still give different accuracy depending on the structure of the patient data on which the diagnostic procedure is based. When processing data and building the most effective algorithm, pre-processing of data, identification of factors of significant influence, and therefore the correct selection of variables for classification, the correct choice of methods is crucial. For the classification of binary variables, which can be easily obtained by survey, the most effective classification methods are related to the construction of random trees and their combinations.

Undoubtedly, technology cannot be fully relied upon for diagnosis. Medical expertise should be involved in the diagnosis of diabetes as there are a lot of factors to be taken into consideration. However, machine learning classification systems can significantly help detect such a dangerous disease as diabetes in time.

Authors' contribution: José Luis Galán-García, Hanna Livinska – conceptualization, methodology; Daria Skrypnyk, Hanna Livinska – software, formal analysis; Daria Skrypnyk – data validation, spelling (original draft); Daria Skrypnyk, Hanna Livinska – writing (viewing and editing).

References

Aguilera-Venegas, G., López-Molina, A., Rojo-Martínez, G., & Galán-García, J. L. (2023). Comparing and tuning machine learning algorithms to predict type 2 diabetes mellitus. *Journal of Computational and Applied Mathematics*, 427, 115115.

Battineni, G., Sagaro, G., Nalini, C., Amenta, F., & Tayebati, S. K. (2019). Comparative Machine-Learning Approach: A Follow- Up Study on Type 2 Diabetes Predictions by Cross- Validation Methods, *Machines*, 7, 74.

Frank, E., Hall, M. A., & Witten, I. H. (2016). *The WEKA Workbench Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, Fourth Edition, doi:10.1016/B978-0-12-804291- 5.00024-6.

Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology & Metabolic Syndrome,* 13, 148. https://doi.org/10.1186/s13098-021-00767-9.

Hectors, T. L., Vanparys, C., van der Ven, K., Martens, G. A., Jorens, P. G., Van Gaal, L. F., Covaci, A., De Coen, W., & Blust, R. (2011). Environmental pollutants and type 2 diabetes: a review of mechanisms that can disrupt beta cell function. *Diabetologia,* 54(6), 1273-90.

Islam, M.M.F., Ferdousi, R., Rahman, S., & Bushra, H.Y. (2020). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. In M. Gupta, D. Konar, S. Bhattacharyya, S. Biswas (Eds.). *Computer Vision and Machine Intelligence in Medical Image Analysis. Advances in Intelligent Systems and Computing* (vol. 992). Springer, Singapore. https://doi.org/10.1007/978-981-13-8798-2_12

Kaggle. *Prediction of diabetes at early stage*. https://www.kaggle.com/datasets/andrewmvd/early-diabetes-classification

Kangra, K., & Singh, J. (2023) Comparative analysis of predictive machine learning algorithms for diabetes mellitus. *Bulletin of Electrical Engineering and Informatics*, 12(3), 1728–1737.

Sadhu, A., & Jadli, A. (2021) Early-Stage Diabetes Risk Prediction: A Comparative Analysis of Classification Algorithms. *International Advanced Research Journal in Science, Engineering and Technology*, 8(2), 193–201.

Wei, S., Zhao, X., & Miao, C. (2018). A comprehensive exploration to the machine learning techniques for diabetes identification. In *IEEE World Forum on Internet of Things, WF-IoT 2018.* Proceedings.

Отримано редакцією журналу / Received: 18.02.24
Прорецензовано / Revised: 12.04.24
Схвалено до друку / Accepted: 27.05.24

Ганна ЛІВІНСЬКА, канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0001-9676-7932
e-mail: hanna.livinska@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

Дарья СКРИПНИК, студ.
e-mail: skrypnikdaria@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

Хосе-Луіс ГАЛАН-ГАРСІЯ, PhD (мат.), доц.
ORCID ID: 0000-0002-8773-6998
e-mail: jlgalan@uma.es
Університет Малаги, Малага, Іспанія

ПОРІВНЯЛЬНИЙ АНАЛІЗ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ ДІАБЕТУ ЗА БІНАРНИМИ ДАНИМИ

Поширеність цукрового діабету постійно зростає, і своєчасна діагностика цього захворювання є надзвичайно важливою в системі охорони здоров'я. Нині, окрім передових медичних технологій, є можливість швидко та якісно обробляти великі обсяги інформації, зокрема й медичної. Хворі на цукровий діабет навіть на ранніх стадіях мають певну схожість симптомів і закономірностей, що дає можливість ефективно діагностувати захворювання на основі легко отриманих клінічних даних.

У пропонованій роботі ми порівнюємо точність різних алгоритмів машинного навчання для діагностики діабету на двійкових даних, які не охоплюють лабораторні тести і можуть бути отримані шляхом анкетування пацієнта. Усі моделі класифікації даних базового датасету мають хороші результати та надають можливість ідентифікації осіб, які, ймовірно, мають недіагностований діабет, на основі цих даних. Найкращі результати дає класифікатор випадкового лісу. Для підвищення точності класифікації проведено аналіз зв'язків між змінними датасету. Це дало змогу зменшити розмірність моделі, видаливши змінні, що не містять корисної інформації. Окрім того, додатково збалансувано множину даних. Отримані моделі демонструють вищу ефективність, ніж моделі класифікації, побудовані за базовим датасетом.

К л ю ч о в і с л о в а : ***діагностика цукрового діабету, класифікатор випадкового лісу, машинне навчання, точність класифікації.***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.