DOI:10.2478/bile-2025-0008

Biometrical Letters
Vol. 62 (2025), No. 2, 113-140

Application of machine learning in diabetes diagnosis

Aleksandra Rosińska, Łukasz Smaga

Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland, e-mail: rosinskaola@poczta.fm, ls@amu.edu.pl

Summary

Diabetes is a serious global health problem that affects millions of people and leads to many complications if not diagnosed early. Early and accurate diagnosis is very important for improving patient outcomes and reducing healthcare costs. Machine learning can help to analyze medical data and predict diabetes more effectively. This study compares three machine learning models – logistic regression, random forests, and XGBoost – for predicting diabetes based on medical data. The models were tested in their basic forms and with different techniques for balancing the dataset, such as undersampling, oversampling, SMOTE, and an asymmetric approach. Additionally, variable reduction and probability averaging as a form of ensemble learning were applied. The experiments are based on the dataset available on the Kaggle platform, which contains 100,000 observations. The problem is interesting because diagnostic criteria based on glycated hemoglobin and blood glucose levels do not enable automatic and unambiguous diagnosis in this dataset. However, they will be important independent variables in the classification models considered. The results of the evaluation show the potential of machine learning in supporting specialists in diabetes diagnosis, and highlight the importance of proper data preprocessing for achieving better model performance.

Key words: classification, data balancing, diabetes, machine learning

1. Introduction

1.1. Diabetes

Diabetes is a chronic metabolic disease whose main feature is hyperglycemia, or elevated blood glucose levels caused by impaired insulin secretion or function. Insulin is a hormone that is essential for the proper metabolism of sugars, fats, and proteins. A lack of adequate insulin or impaired insulin function leads to the accumulation of glucose in the blood, which in the long

term can cause serious damage to many organs and systems of the body, including the kidneys, eyes, nerves, heart, and blood vessels (Włodarczyk, 2018-2020).

There are many types of diabetes, but the most common are type 1 and 2 diabetes (Fiłon, 2019). Type 1 diabetes is an autoimmune disease in which the immune system attacks and destroys the β cells of the pancreas responsible for producing insulin. This process leads to a complete deficiency of insulin, which results in a sudden increase in blood glucose levels. Although this type of diabetes accounts for about 10–15% of all cases, it is dominant among children and adolescents. Type 2 diabetes, also known as non-insulin-dependent diabetes, is the most common form of the disease, accounting for 80–90% of all cases. It develops gradually, often asymptomatically for many years, and therefore is usually diagnosed only when complications appear. The main cause of type 2 diabetes is insulin resistance, or a decrease in tissue sensitivity to insulin, combined with dysfunction of the β cells of the pancreas, which do not produce enough of this hormone. This disease most often appears in adults, usually after the age of 40, and its risk increases with age. However, in recent years, an increasing number of cases of type 2 diabetes have been observed in younger people, which is closely related to environmental factors such as lack of physical activity, obesity, or improper diet.

In Poland, around 3 million people suffer from diabetes; this is 8–9% of the country's population. It is predicted that by 2030, over 10% of the population will suffer from this disease (`https://www.gov.pl/web/gov/szukaj?scope=zdrowie&query=cukrzyca`). In 2023, the rate of deaths from diabetes in Poland was 25.7 per 100,000 inhabitants (`https://sdg.gov.pl/statistics_nat/3-1-e/`). These numbers show how important it is to detect diabetes early.

1.2. Machine learning in medical applications

Forecasting, based on the identification of risk factors such as overweight, insulin resistance, unhealthy diet, lack of physical activity, or genetic predisposition, enables earlier intervention before the disease develops. An increasingly important role in this area is played by algorithms using artificial intelligence and data analysis (Wojdan and Moniuszko, 2022), which, based on metabolic and clinical parameters, can estimate the probability of developing diabetes. In this way, it is possible to identify people at increased risk early and refer them for screening tests and to appropriate specialists.

Machine learning and artificial intelligence algorithms have already been used in the context of diagnosing and preventing many diseases. For example, McKinney et al. (2020) created an artificial intelligence algorithm that detects breast cancer cases based on mammography. They were prompted to do so by the fact that clinicians are unable to diagnose breast cancer in about 20% of cases, and 50% of women who regularly check themselves will receive a false positive result within 10 years. As a result, this algorithm turned out to be much more effective in detecting cases than the tested group of radiologists. Another example is given in the paper by Ogłoszka and Smaga (2022), who analyzed the use of classic machine learning methods on tabular data in the diagnosis of breast cancer. The best results were achieved using the XGBoost algorithm, which achieved a high accuracy of 99% and a sensitivity of 100%.

Ardila et al. (2019) created an algorithm that helps diagnose lung cancer from low-dose chest CT scans. The algorithm was found to be more effective than a group of six radiologists when previous CT scans were not available, and as effective as radiologists when previous scans were available. In Rajpurkar et al. (2018), the operation of the CheXNet algorithm, which performs diagnoses based on lung X-rays, was presented. This algorithm can diagnose 14 lung diseases, and in 11 out of 14 cases it performed better than doctors. In addition, it completed its analysis in about 2 minutes, while radiologists took about 3 hours to make a diagnosis.

Research on brain imaging data has been conducted, for example, by Reszke and Smaga (2023). In that study, convolutional neural networks were used to detect brain tumors from MRI images. The models achieved high efficiency, with an accuracy of 92.59% and an AUC of 0.95–0.96.

Disease detection can also be based on audio recordings. Laguarta et al. (2020) used cough recordings made with a mobile phone to detect asymptomatic COVID-19 infection. As a result, on a test sample containing 1,000 cough recordings, the AI algorithm achieved an efficiency of 98.5%, while these changes were not recognizable to the human ear.

To sum up, machine learning can provide valuable support in diagnosing various diseases, offering additional insights that may not be immediately apparent to clinicians. However, these methods should not be relied upon in isolation, but rather used as decision-support tools to enhance the work of medical specialists.

1.3. Objectives and structure

As we saw in the previous section, machine learning and artificial intelligence algorithms are finding more and more applications in the context of medicine, where they can be used to help with diagnosing various diseases and conditions. In most cases, they turn out to be much more effective than human beings, which indicates that they should be further developed and will likely be in high demand in this field. Therefore, in this work, we focus on the use of machine learning methods to predict diabetes.

For this purpose, we compare three machine learning models, namely logistic regression, random forests, and XGBoost. These models were selected due to their proven effectiveness in handling classification tasks in medical applications. For instance, logistic regression has been successfully used for predicting cardiovascular diseases based on patient risk factors (Ambrish et al., 2022), while random forests have shown strong performance in identifying patterns in cancer diagnosis (Dai et al., 2018). Other applications of these machine learning methods can be found in Stoltzfus (2011), Kasperczuk and Dardzińska (2017), Konukoglu and Glocker (2020), and Ranjbarzadeh et al. (2023). In this work, the models were evaluated in their basic configurations with the optimization of hyperparameters and after applying various data balancing techniques, including undersampling, oversampling, SMOTE, and an asymmetric approach, to address the issue of imbalanced datasets, commonly encountered in medical data. Furthermore, techniques such as variable reduction and probability averaging, as a form of ensemble learning, were employed to enhance model performance. The evaluation was performed using key classification measures: accuracy, sensitivity, specificity, precision, F1-score, AUC, and other measures specified for imbalanced classes (such as balanced accuracy and PR-AUC). By exploring these approaches, this study aims to demonstrate the potential of machine learning in improving the accuracy and reliability of diabetes diagnosis, contributing to better patient outcomes and more efficient healthcare systems.

The remainder of the paper consists of the following sections. Section 2 is devoted to the description of the dataset on which the research is based. In section 3 we focus on the description of the methods used in the experiments, the evaluation measures, and the problem of imbalanced classes. Section 4 presents detailed results from a comparison of the methods. Finally, section 5 contains a summary and conclusions from the research.

This paper is based on a degree thesis by Rosińska (2025).

2. Diabetes dataset

The analyzed dataset comes from the Kaggle service `https://www.kaggle.com/`, and is called the "Diabetes prediction dataset." It is available at `https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset`. The dataset contains data from 100,000 patients examined for their risk of developing diabetes. The dataset includes the following nine variables. We mainly focus on comparing the results for individual variables with respect to diabetes incidence. The description of each variable can be found on the Kaggle page linked to above.

Diabetes This is the target variable. Healthy patients (91.5%) are marked as 0, and diabetic patients (8.5%) are marked as 1. Thus, the numbers of healthy and sick patients are very uneven. For this reason it will be useful to apply data balancing methods in subsequent steps.

Gender Of all the patients studied, women with diabetes constitute 4.46%, and healthy women 54.09%. In turn, men with diabetes constitute 4.04% of the studied patients, and healthy men 37.39%. (For some people, a gender was not indicated.) The proportion of diabetes in women is 7.62%, and the proportion of diabetes in men is 9.75%. To analyze the significance of gender for illness, the proportion test for two samples was applied. This implies that men are more likely to suffer from diabetes than women.

Age The median age in the non-diabetic group was 40 years, and in the diabetic group it was 62 years. A more detailed age distribution of patients using a box-and-whisker plot is presented in Figure 1. The Wilcoxon test was used to examine significant differences in the ages of patients with and without diabetes. The test showed that older people were more likely to develop diabetes.

Hypertension The proportion of people with diabetes in the group with hypertension is 27.90%, and the proportion of people with diabetes in the group without hypertension is 6.93%. The test for two proportions showed significant differences, which implies that people with hypertension are around four times more likely to develop diabetes.

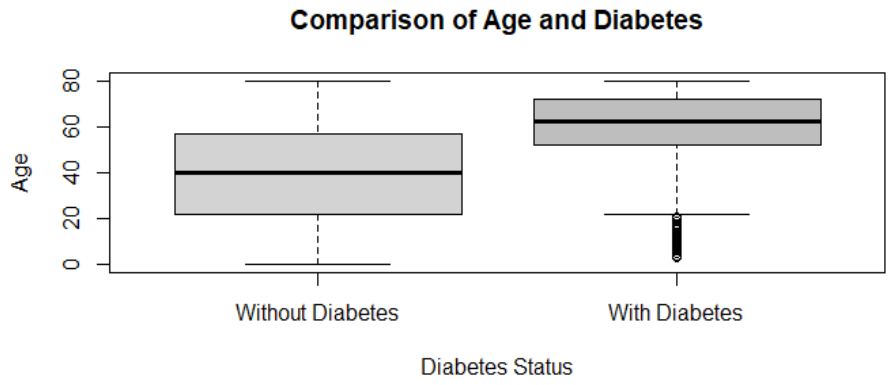

Figure 1. Box-and-whisker plot illustrating the relationship between age and diabetes

Heart disease The proportion of people with diabetes in the heart disease group is 32.14%, and in the group without heart disease it is 7.53%. The same test as before also shows that people with heart disease are around four times more likely to have diabetes than those without.

Smoking history The variable storing smoking history is more diverse than the other variables, containing values such as *current*, *ever*, *former*, *never*, *not current*, and *No Info*. The value *No Info* will not be considered in further analyses, because it does not provide any significant information. The largest percentage of people with diabetes is in the *former* group, at 17%. The second largest group with diabetes is the *ever* group, at 11.8%. The smallest percentage of diabetics is in the *never* group, at 9.53%. It can be seen that former smokers and people who currently smoke seem to have a higher percentage of diabetes than people who have never smoked. The χ^2 test of independence is used to analyze these results. It confirms that there is a significant relationship between smoking history and the occurrence of diabetes. It can therefore be assumed that people who are current smokers or who have smoked in the past have a greater risk of developing diabetes than people who have never smoked.

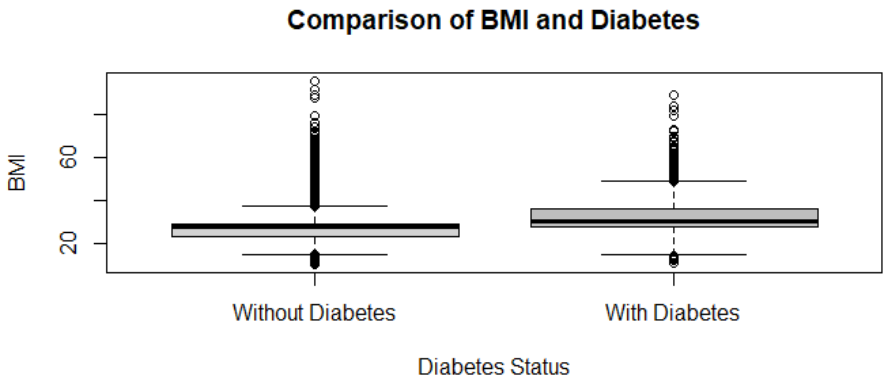

Figure 2. Box-and-whisker plot illustrating the relationship between BMI and diabetes

BMI A healthy adult should have a BMI in the range 18.5–24.99, those who are overweight lie in the range 25–29.99, and a higher BMI indicates obesity. The median BMI value in the group of people without diabetes is 27.32, and in the group of people with diabetes it is 29.97. The distribution of BMI in both groups is illustrated in Figure 2. The Wilcoxon test confirms that people in the group with diabetes have a statistically higher BMI than those without it. This means that BMI is an influential factor in the incidence of diabetes.

HbA1c level The level of glycated hemoglobin, i.e. HbA1c, is a direct indicator of diabetes risk. In nonpregnant individuals, the normal value for a healthy person is considered to be below 5.7%, and a result in the range 5.7–6.4% indicates prediabetes, while a value equal to or greater than 6.5% suggests diabetes (American Diabetes Association Professional Practice Committee, 2025). In our dataset, the median in the group of healthy people is 5.8%, and in the group of people with diabetes it is 6.6%. Figure 3 illustrates the distribution of HbA1c levels in both groups. The Wilcoxon test showed a statistically significant difference in HbA1c levels between the groups. The analysis suggests that a higher HbA1c level is significantly associated with the occurrence of diabetes, because it reflects diabetics' difficulty in maintaining blood glucose levels within the norm.

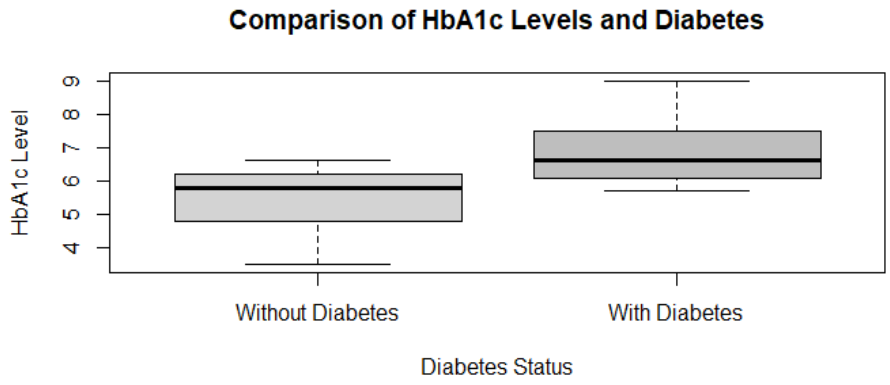

Figure 3. Box-and-whisker plot illustrating the relationship between HbA1c level and diabetes

Blood glucose level Glucose level is a key factor in detecting the disease, because it is the first indication that diabetes is developing in the patient's body. In nonpregnant individuals, the glucose level in a healthy person is from 70 to 99 mg/dL, and after a meal it should be less than 140 mg/dL (American Diabetes Association Professional Practice Committee, 2025). Values in the ranges 100–125 mg/dL and 140–199 mg/dL respectively indicate prediabetes, while values equal to or greater than 126 mg/dL and 200 mg/dL respectively suggest diabetes. In our dataset, the median blood glucose level in healthy patients is 140 mg/dL, and in people with the disease it is 160 mg/dL. The Wilcoxon test showed a significant difference between the medians in the two study groups, which is also visible in Figure 4. This analysis therefore indicates that higher blood glucose levels are strongly associated with the occurrence of diabetes.

Summarizing all of the above analyses, significant relationships can be observed between health variables and diabetes. Higher BMI, HbA1c levels, and blood glucose levels are strongly associated with the occurrence of diabetes, which is confirmed by the graphs and the Wilcoxon test results. Hypertension, heart disease, and smoking history also show a relationship with diabetes, indicating a higher risk of diabetes.

Remark 2.1. *Let us note one important feature of the dataset being considered. As mentioned above, both glucose level and glycated hemoglobin*

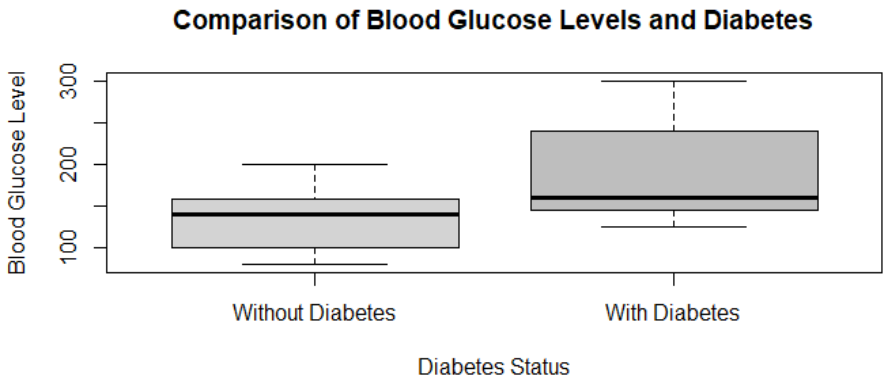

Figure 4. Box-and-whisker plot illustrating the relationship between glucose levels and diabetes

*(HbA1c) are official diagnostic criteria for diabetes and prediabetes. However, the diagnosis presented in the dataset was not based solely on these criteria; there are positively diagnosed individuals with these markers lower than the reference level, and healthy individuals with higher levels. These discrepancies may occur for various reasons, such as treatment, data errors, different diagnosis sources, measurements from different dates, or diagnoses based on history and symptoms. However, this cannot be determined at this point. Furthermore, the dataset's author mentions that he was unable to determine how glucose levels were measured or what type of diabetes was identified (*https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/discussion/405636, https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/discussion/403663*). Therefore, since the reference levels for glucose and glycated hemoglobin (HbA1) do not determine morbidity in this dataset, it makes sense to use them as independent variables in our classification models, along with the other variables described in this section. We expect these two variables to be of great importance in the classification of diabetes using machine learning methods, which we will also show in section 4.*

3. Methods

This section provides a brief description of the methods used in the study. These are mainly classification models, since we wish to assign patients as diabetics or non-diabetics. Additionally, measures evaluating the performance of the classifiers and methods for solving the problem of imbalanced data are described.

3.1. Classification methods

The study tested three popular machine learning algorithms used in the context of binary classification: logistic regression, random forests, and XGBoost. Each of these models represents a different way of learning, which allows them to be compared in terms of their effectiveness in the problem of diabetes prediction.

3.1.1. Logistic regression

Logistic regression models analyze the relationships between a binary dependent variable and independent variables. For the i-th observation, the probability p_i of a specified value (denoted by 1) of the dependent variable is modeled as:

$$p_i = P(Y_i = 1|\mathbf{X}_i) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_p X_{ip})}},$$

where $i = 1, 2, \ldots, n$. In this formula, $P(Y_i = 1|\mathbf{X}_i)$ denotes the conditional probability of the i-th observation belonging to class 1. The training data are defined as $(\mathbf{X}_i, Y_i)$ or more precisely $(X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i)$, and $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the unknown parameters. As a result of the logistic function, values between 0 and 1 are returned. By default, if $P(Y_i = 1|\mathbf{X}_i)$ is greater than 0.5, class 1 is predicted. If the probability value is less than 0.5, class 0 is predicted. Logistic regression is therefore a classification technique. First, the probability values of the observation belonging to each of the target classes are estimated. Next, the observation is assigned to the class for which this probability is the highest.

3.1.2. Random forests

Random forests are a special type of machine learning method based on bagging. Bagging is a technique where multiple classifiers are trained on

random samples of data and their results are combined using majority voting. Random forests use a special version of decision trees where a subset of features is selected and used to construct the tree (Fenner, 2019). The goal is to force the trees to be diverse and independent, which reduces the problem of overfitting. Furthermore, each tree is trained on a bootstrap sample (i.e. a random subsample with replicates) taken from the training set.

Each tree in the random forest is trained without pruning, i.e. it grows until it reaches a maximum depth at which each observation is in a separate leaf or other stopping criteria are met. Each partition s in the classification tree is conditioned by observations from the training set that belong to a given node t. For each node t we can define a certain measure of heterogeneity of elements in that node. In the case of classification, the Gini index is used as a measure of heterogeneity:

$$G(p_1, p_2, \ldots, p_k) = 1 - \sum_{i=1}^{k} p_i^2,$$

where k is the number of classes in the dataset, and p_i is the proportion of observations belonging to i-th class in a given node (Fenner, 2019).

In the case of a random forest, many decision trees are trained in such a way that: (1) many bootstrap samples are drawn, (2) for each bootstrap sample m features are drawn, which will be used to construct the decision tree. The hyperparameter m should be much smaller than the data dimension p; therefore for classification its value is most often assumed to be equal to $\sqrt{p}$ (Fenner, 2019).

3.1.3. XGBoost

Extreme gradient boosting (XGBoost) is a very popular gradient boosting algorithm (Fenner, 2019) in machine learning. Gradient boosting is a technique that iteratively builds an ensemble model by adding successive decision trees. Each new tree is built in such a way as to minimize the prediction error of the previous trees by fitting to the residuals, which are the differences between the actual values and the predicted values. For this purpose, an optimization strategy is used that generates better model weights (Starmer, 2022).

3.2. Classifier performance measures

It is necessary to evaluate the performance of the classification methods. The appropriate measures are based on the confusion matrix, which is presented in Table 1.

Table 1. Confusion matrix

	Reference Negative	Reference Positive
Prediction Negative	TN	FN
Prediction Positive	FP	TP

The rows in Table 1 correspond to the model predictions, and the columns to the actual classes in the test set. The table cells contain the following values:

1. TN – True Negatives – cases correctly classified as negative,
2. FN – False Negatives – cases incorrectly classified as negative,
3. FP – False Positives – cases incorrectly classified as positive,
4. TP – True Positives – cases correctly classified as positive.

The following popular classification quality measures were used in the research:

1. Accuracy – the proportion of well classified observations:

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

2. Sensitivity (recall) – the ratio of cases correctly classified as positive to all cases that are actually positive:

$$\text{Sensitivity} = \frac{TP}{TP + FN},$$

3. Specificity – the ratio of cases correctly classified as negative to all cases that are actually negative:

$$\text{Specificity} = \frac{TN}{TN + FP},$$

4. Precision – the ratio of cases correctly classified as positive to cases classified as positive:

$$\text{Precision} = \frac{TP}{TP + FP},$$

5. F1-score – the harmonic mean of precision and sensitivity:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}},$$

The F1-score represents a compromise between precision and sensitivity.

6. Balanced Accuracy – the arithmetic mean of specificity and sensitivity:

$$\text{Balanced Accuracy} = \frac{\text{Specificity} + \text{Sensitivity}}{2},$$

7. NPV – Negative Predictive Value – the ratio of cases correctly classified as negative to all those classified as negative:

$$\text{NPV} = \frac{TN}{TN + FN},$$

8. MCC – Matthews Correlation Coefficient, a measure that takes into account both correct and incorrect classifications for both classes:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

If the denominator is 0, the MCC value is 0.

The higher the value of these measures, the better the model.

In medical research, precision is also called positive predictive value (PPV). It is the probability that a person with a positive test result actually has the disease. PPV answers the question, "If my test result is positive, what is the probability that I actually have the disease?" NPV, on the other hand, is the probability that a person with a negative test result is actually healthy. NPV answers the question, "If my test result is negative, what is the probability that I actually am healthy?" PPV and NPV are

most useful when assessed in the context of the specific clinical application and patient population in which the test will be used. They are strongly dependent on the disease's prevalence. For example, if the prevalence is low (the disease is rare), the test's PPV will be low and the NPV high; then even a very good test in a low-prevalence population can generate many false positives.

Balanced accuracy is a metric for evaluating binary classification models that is particularly useful when classes are imbalanced. In the case of imbalanced classes, a model can achieve high accuracy by simply classifying most samples as the majority class. Balanced accuracy addresses this issue by treating both classes (positive and negative) with equal weight, regardless of their size. By calculating the average of sensitivity and specificity, it provides a fairer picture of the model's performance.

The MCC, also known as the phi coefficient, is a measure of the quality of binary classifications. Unlike other popular metrics, such as F1-score, MCC considers all four elements of the confusion matrix. MCC is considered one of the most reliable metrics because it provides a balanced score regardless of the class distribution. The MCC value ranges from -1 to 1. A value of 1 represents perfect classification (all predictions are correct), a value of 0 indicates that the model performs no better than a random guess, and a value of -1 indicates a complete discrepancy between predictions and actual labels (the model confuses everything). The main advantage of MCC is that it is a balanced measure and is robust to imbalanced data. For example, in a dataset with 99% negative and 1% positive cases, a model that always predicts the negative class will achieve 99% accuracy. This result is misleading because the model effectively fails to correctly identify the rare positive class. In such a situation, the MCC will yield a score close to 0, which much better reflects the model's lack of predictive value.

In addition to the above standard measures, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are widely used measures for evaluating the performance of binary classification models. The ROC curve plots the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) or equivalently specificity (see Figure 5) across various classification thresholds, providing a visual representation of the trade-off between sensitivity and specificity. The AUC quantifies the overall performance of the classifier, with a value of 1 indicating perfect discrimination and 0.5 suggesting performance no better than random guessing. When the classes in a dataset are highly imbalanced

and the positive class is rare, a better measure is the PR-AUC (Area Under the Precision–Recall Curve). The PR curve is a graph that shows the tradeoff between a model's precision and sensitivity for different decision thresholds (Saito and Rehmsmeier, 2015). Typically, increasing sensitivity (e.g., by lowering the decision threshold) often reduces precision because the model predicts more samples as positive, leading to more false positives. The PR curve visualizes this tradeoff (see Figure 6). The PR-AUC is the area under this curve.

3.3. Imbalanced data

As we noted in section 2, our dataset is quite imbalanced. In the training set (see section 4), the sample of people with diabetes has 6,022 observations, and the sample of healthy people has 63,978 observations. This imbalance caused the models to tend to predict the dominant class while ignoring the minority one, which resulted in sensitivity in the range 60–70%. This problem is called the class imbalance problem. To deal with it, various methods of data resampling are used, involving changing the number of observations in the imbalanced classes; see e.g. Małowiecki (2023) and Ogłoszka and Smaga (2022). We also considered a method of combining classifiers. More specifically, the following data balancing results were used:

1. Undersampling is a resampling method that works by removing part of the data from the majority class so that both classes have an equal number of observations.
2. Oversampling works in a way opposite to undersampling: here the size of the smaller class is increased to the size of the larger class by randomly selecting records of the smaller class.
3. Asymmetric approach – this consists of partial balancing of classes by controlled undersampling of the majority class. In this variant, the number of observations of class 0 is limited to twice the number of observations in class 1, while maintaining the original size of the minority class. This solution allows us to mitigate the problem of imbalanced data with less restrictive interference in the structure of the training set than in the case of undersampling and oversampling.
4. SMOTE (*Synthetic Minority Over-Sampling Technique*) works similarly to oversampling, in that it increases the size of the minority class. However, instead of duplicating existing minority class samples,

SMOTE creates new, synthetic samples from existing data by interpolating them in feature space. The algorithm randomly selects a sample from the minority class and its nearest neighbors and then generates new samples as linear combinations of these points, which increases the diversity of the data and helps balance the classes in the dataset.

5. The average of probabilities is one of the simple methods of ensemble learning, which combines the predictions of several classifiers by averaging their posterior probabilities, i.e. the probabilities of belonging to the positive class, in this case to the class of patients with diabetes. For our models, we can express this average with the formula:

$$P_{av}(y=1) = \frac{1}{L}\sum_{l=1}^{L} P_l(y=1),$$

where $P_{av}(y=1)$ is the average of the posterior probabilities, $P_l(y=1)$ is the probability of belonging to class 1 returned by the l-th classifier, and L is the number of classifiers included in the model.

4. Experiments

This section describes the experiments and summarizes the results of diabetes classification on the considered dataset. The experiments were run in the R programming language (R Core Team, 2025). The code for all experiments is given in the GitHub repository available at `https://github.com/ls-git-17/ml_in_diabetes_diagnosis`.

4.1. Set-up

In order to evaluate the effectiveness of selected classification models, the dataset was divided randomly into a training set, which contains 70% of the data, and a test set, which contains 30% of the data. This division will enable an objective evaluation of the models on the data.

During the experiment for logistic regression, the default settings of the hyperparameters of the `glm()` function available in the R language were adopted, because logistic regression has relatively few of them, and the default values are often sufficient to obtain good results. However, we also studied the effect of optimizing hyperparameters (see section 4.2.3). In the

implementation, the logistic regression uses treatment contrasts with a reference group encoding of categorical variables.

In the basic application of random forests, the number of trees was set to 500, which enables stable and repeatable prediction results. The number of variables selected for each division was set to 3, since we have 8 independent variables in the dataset. This increases the diversity of trees and reduces the problem of overfitting. In section 4.2.3, we investigate the optimization of hyperparameters for the random forests.

For the basic XGBoost model, the learning rate η was set to 0.1, which allows gradual model training and reduces the risk of overfitting, and the maximum tree depth was set to 6 to avoid overfitting the model to the training data. However, we also consider hyperparameter optimization for this model (see section 4.2.3).

This selection of settings allows a fair comparison of the performance of individual models, while also taking into account potential problems of overfitting and prediction stability.

4.2. Results

In this section, we present and discuss the results of the evaluation process. First, we investigate the basic models, which are evaluated on the test set (section 4.2.1). Next, we check the performance for the reduced models and class balancing methods (section 4.2.2). Finally, we evaluate the classification models using stratified cross-validation and investigate the optimization of hyperparameters (section 4.2.3).

4.2.1. Basic models evaluated on the test set

First, let us take a look at how the results of the basic models compare, without applying the methods described in section 3.3. We evaluate the models on the test set in a standard way. The confusion matrices for the tested models are given in Tables 2–4.

Table 2. Confusion matrix for logistic regression

	Reference 0	Reference 1
Prediction 0	27251	906
Prediction 1	271	1572

Table 3. Confusion matrix for random forests

	Reference 0	Reference 1
Prediction 0	27447	744
Prediction 1	75	1734

Table 4. Confusion matrix for XGBoost

	Reference 0	Reference 1
Prediction 0	27483	749
Prediction 1	39	1729

The results are presented in Table 5. Overall, XGBoost is the best model for predicting diabetes, as it achieves the best results for 7 out of 10 evaluation measures. It has better performance than logistic regression and random forests, in terms of both accuracy and other measures such as specificity, precision, F1-score, MCC, ROC-AUC and PR-AUC. However, the XGBoost model is slightly worse than random forests in terms of balanced accuracy and NPV. Random forests achieve very good results, especially in specificity, NPV, and ROC-AUC, which makes them a good choice, especially when we are interested in detecting healthy cases. Logistic regression, while still effective, has some limitations compared with more complex models such as XGBoost and random forests, especially in terms of sensitivity, precision, F1-score, MCC, and PR-AUC. Unfortunately, we see lower sensitivity values relative to other measures, meaning that the models are less good at predicting diabetes. This is an important aspect to which we will return in section 4.2.2. Let us note that the balanced accuracy is much smaller than the accuracy, which is natural for imbalanced cases. However, the balanced accuracy is quite high for all basic models. The same is true for ROC-AUC and PR-AUC. We discuss these measures in more detail below.

Table 5. Evaluation measures (as percentages) of the classification methods for the test set

Criterion	Logistic regression	Random forests	XGBoost
Accuracy	96.08	97.26	97.37
Balanced accuracy	81.23	84.84	84.82
Sensitivity	63.44	69.98	69.77
Specificity	99.02	99.71	99.86
Precision	85.30	95.64	97.79
NPV	96.78	97.36	97.35
F1-score	72.76	80.82	81.44
MCC	71.60	80.51	81.40
ROC-AUC	96.23	97.12	98.00
PR-AUC	81.43	87.40	88.94

For each of the basic models, ROC and PR curves were also created, as summarized in Figures 5 and 6. The ROC curves are very close to each other, which indicates comparable performance of the models in terms of this criterion. The most efficient model is the one whose curve is closest to the upper left corner of the graph. All models achieved very good results, but a slight advantage is visible for the XGBoost model, whose curve is slightly above the others. We obtain greater distinction in the models' properties from the PR curves, due to class imbalance (Figure 6). The logistic regression model is clearly worse, and the XGBoost model is slightly better, than the random forest model.

The models were also subjected to an assessment concerning the importance of variables and a possible reduction of variables in the basic model. In the case of logistic regression, stepwise regression was used with AIC, which did not remove any of the variables, and so it does not simplify the model. To illustrate the reduction of the logistic regression model, the smoking history variable will be removed due to the large incidence of the value *No Info* (see section 4.2.2), as this is effectively missing data.

In the case of random forests, the importance of variables is shown in Figure 7. The most important variables in this model are HbA1c level and blood glucose level. Of the remaining variables, the most important are age and BMI. Based on these results, in the reduced random forest model, only these four independent variables were taken into account.

The variable importance plot in the XGBoost model is shown in Figure 8. It can be seen that the most important variables in the XGBoost

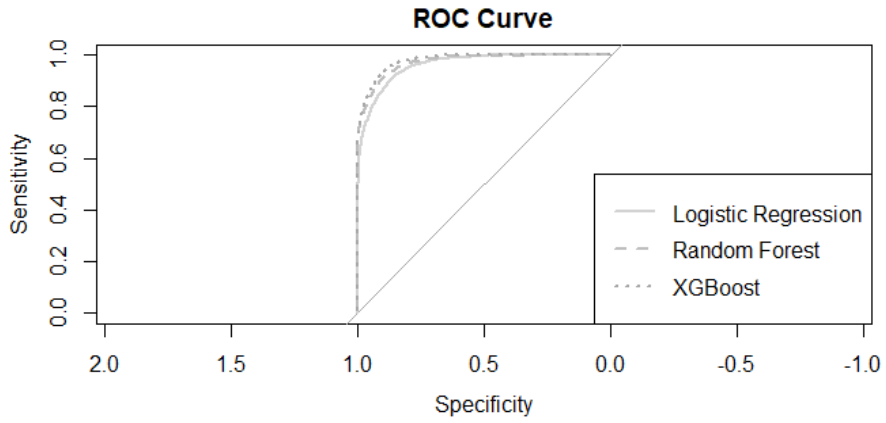

Figure 5. ROC curves for logistic regression, random forests, and XGBoost in the test set

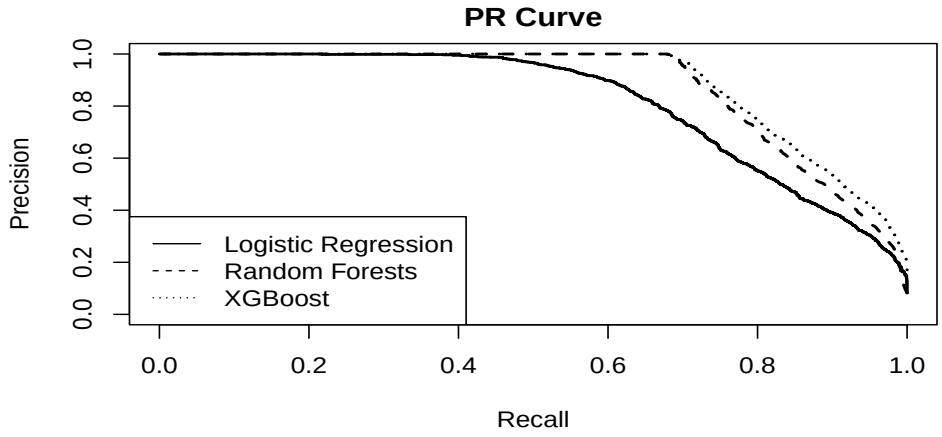

Figure 6. PR curves for logistic regression, random forests, and XGBoost in the test set

model are the same as in the case of random forests. Therefore, in the variable reduction model, the variables HbA1c level, blood glucose level, BMI, and age again participate. The results for the reduced models are given in section 4.2.2.

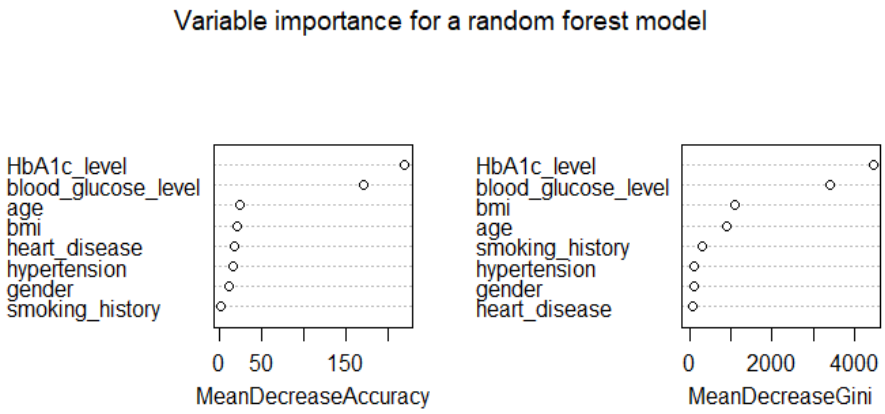

Figure 7. Importance of variables in the random forest model

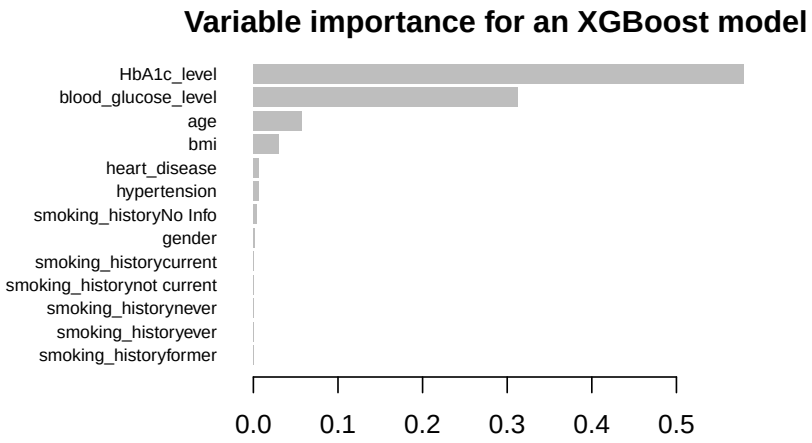

Figure 8. Importance of variables in the XGBoost model

4.2.2. Reduced models and class balancing evaluated on the test set

Let us move on to the analysis of the results for individual models subjected to data balancing (due to the large imbalance of classes) and variable reduction. These modifications aimed to increase the effectiveness of detecting

diabetes cases, which in practice means striving to improve the sensitivity results.

We will first look at logistic regression models, for which all collected results are presented in Table 6. Reduction of variables caused a small decrease in most measures, but slightly improved precision. This method produces a similar F1-score as the original model. The highest sensitivity was obtained for undersampling (88.94%) and oversampling (88.66%), which confirms the effectiveness of these methods in the context of identifying minority class cases. The SMOTE method also achieved very good sensitivity (87.93%), while maintaining a reasonable value of specificity (88.8%). Moreover, for each of the approaches described in section 3.3, we observe a significant increase in balanced accuracy and stable results for ROC-AUC and PR-AUC. Unfortunately, this was at the cost of accuracy, precision, F1-score, and MCC. In these approaches, the precision is more than two times lower than for the original model. In this example, we clearly see that the undersampling and oversampling deliberately bias the model toward the minority class, making it less conservative in its predictions. This increased willingness to correctly identify true positives (high sensitivity) simultaneously leads to more false alarms (false positives), which consequently reduces precision. We can distinguish the asymmetric method, which led to a sensitivity of 81.44%, and greater precision (53.09%) and F1-score (64.28%) than other resampling techniques. The use of the average probability allowed further improvement of the overall classification stability, achieving a sensitivity of 75.67% and a precision of 62.48%, which is a good balance without the need for drastic data modification.

Now let us consider the summary of results for the random forest models, presented in Table 6. The use of variable reduction did not significantly affect the quality of classification, except in the case of precision. The largest increases in sensitivity and balanced accuracy (as much as 92.21% and 91.06% respectively) were obtained thanks to the use of undersampling, which makes this variant the most sensitive for detecting diabetes in the entire set of models, but at a high cost in other measures, especially precision. The asymmetric approach is a compromise here – high balanced accuracy and sensitivity were obtained (90.36% and 85.76% respectively) with moderate precision (60.54%) and good overall accuracy (94.21%). The SMOTE method allowed us to obtain a sensitivity of 71.51% while maintaining a very high precision (89.86%) and specificity (99.27%),making it one of the most balanced options. Also, averaging the probabilities improved

Table 6. Evaluation measures (as percentages) of the models for the test set

Model	Acc	BalAcc	Sens	Spec	Prec	NPV	F1	MCC	ROC-AUC	PR-AUC
Logistic regression										
Original	96.08	81.27	63.44	99.02	85.30	96.78	72.76	71.60	96.23	81.43
Reduced	96.07	81.15	63.28	99.02	85.36	96.77	72.68	71.53	96.13	81.19
Under.	88.82	88.88	88.94	88.81	41.71	98.89	56.79	56.19	96.24	81.29
Over.	88.82	88.75	88.66	88.84	41.70	98.86	56.72	56.07	96.25	81.27
As. appr.	92.52	87.48	81.44	93.52	53.09	98.24	64.28	62.03	96.24	81.34
SMOTE	88.70	88.35	87.93	88.80	41.35	98.79	56.25	55.49	96.16	81.12
Av. prob.	94.24	85.79	75.67	95.91	62.48	97.77	68.44	65.67	96.25	81.44
Random forests										
Original	97.26	84.84	69.98	99.71	95.64	97.36	80.82	80.51	97.12	87.40
Reduced	96.90	84.98	70.70	99.26	89.62	97.41	79.04	78.03	96.55	85.85
Under.	90.10	91.06	92.21	89.91	45.15	99.23	60.62	60.37	97.75	88.05
Over.	96.31	86.74	75.26	98.21	79.09	97.78	77.13	75.15	97.47	87.15
As. appr.	94.21	90.36	85.76	94.97	60.54	98.67	70.98	69.13	97.70	88.17
SMOTE	96.98	85.39	71.51	99.27	89.86	97.48	79.64	78.63	97.16	86.94
Av. prob.	95.90	88.25	79.10	97.41	73.33	98.10	76.10	73.92	97.75	88.17
XGBoost										
Original	97.35	84.64	69.41	99.86	97.84	97.32	81.21	81.19	98.02	88.96
Reduced	97.36	84.24	68.52	99.96	99.30	97.24	81.09	81.30	97.82	88.25
Under.	90.11	91.60	93.38	89.82	45.22	99.34	60.94	60.89	97.98	88.68
Over.	90.80	91.73	92.90	90.61	47.09	99.30	62.49	62.22	98.01	88.93
As. appr.	94.53	90.56	85.80	95.32	62.27	98.68	72.17	70.31	97.99	88.87
SMOTE	97.32	84.86	69.94	99.79	96.71	97.36	81.17	80.98	97.86	88.46
Av. prob.	96.13	88.35	79.02	97.68	75.37	98.10	77.15	75.06	98.01	88.90

the overall results.

Finally, let us consider the models based on XGBoost (Table 6). The use of variable reduction had practically no effect on the quality of classification. All measure values remained at a very high level, with slightly better precision (99.30%) and specificity (99.96%). This means that reducing the number of predictors does not worsen the model's performance, and can be beneficial in the context of simplifying the model's structure without impairing its quality. Similarly as with the previous models, we obtained a significant increase in sensitivity using data balancing methods. Undersampling allowed us to achieve the highest sensitivity (as much as 93.38%) among all variants of the XGBoost model, (its balanced accuracy is also almost the highest), though with a significant drop in precision to 45.22% and a slight drop in accuracy. Oversampling gave very similar results: the model's balanced accuracy and sensitivity were then 91.73% and 92.90% re-

spectively, and precision 47.09%. The asymmetric approach offered the best compromise – high sensitivity at a level of 85.8%, and significantly higher precision (62.27%) than other data balancing methods. The XGBoost model using SMOTE turned out to be almost as effective as the original version – all measures, except sensitivity, remained at a very high level. It is also worth noting the use of the average of probabilities, which made it possible to increase the model's sensitivity to 79.02% while retaining quite good results for the other measures.

In summary, for each of the three main models, the best sensitivity and balanced accuracy were achieved by the undersampling and oversampling models, making them the best choice when the most important aim is to detect as many cases of diabetes as possible. However, the precision was mainly quite low. The most balanced results were achieved by the asymmetric approach and the average of probabilities, both approaches offering good sensitivity and significantly higher precision than the data balancing methods.

Table 7. Evaluation measures (as percentages) of all models obtained by stratified nested cross-validation on the train set

Model	Acc	BalAcc	Sens	Spec	Prec	NPV	F1	MCC	ROC-AUC	PR-AUC
Logistic regression										
Original	96.0	80.9	62.6	99.1	87.4	96.6	72.9	72.0	96.2	81.9
Under.	88.6	88.5	88.3	88.6	42.2	98.8	57.1	56.2	96.2	81.8
Over.	88.7	88.5	88.2	88.7	42.4	98.8	57.3	56.3	96.2	81.8
As. appr.	92.6	87.2	80.7	93.8	54.9	98.1	65.4	62.8	96.2	81.8
SMOTE	89.0	88.3	87.5	89.2	42.2	98.7	57.8	56.7	96.2	81.8
Av. prob.	94.4	85.8	75.3	96.2	65.2	97.6	69.9	67.1	96.2	82.0
Random forests										
Original	97.1	83.7	67.4	99.9	99.0	97.0	80.2	80.4	97.4	87.5
Under.	89.9	90.7	91.5	89.8	45.8	99.1	61.0	60.4	97.6	87.8
Over.	93.7	89.7	85.0	94.5	59.3	98.5	69.9	67.8	97.5	87.5
As. appr.	94.3	89.4	83.5	95.3	62.8	98.4	71.6	69.4	97.5	87.8
SMOTE	97.0	84.3	69.0	99.6	94.2	97.2	79.6	79.2	97.3	87.0
Av. prob.	96.1	87.6	77.2	97.9	77.6	97.9	77.4	75.3	97.6	87.9
XGBoost										
Original	97.6	84.2	68.5	99.9	97.8	97.1	80.6	80.5	97.9	88.6
Under.	90.4	91.1	92.0	90.2	47.0	99.2	62.2	61.6	97.8	88.4
Over.	90.6	91.2	91.9	90.5	47.7	99.2	62.8	62.1	97.9	88.5
As. appr.	94.2	90.0	85.0	95.1	61.9	98.5	71.6	69.5	97.9	88.4
SMOTE	97.1	84.3	68.8	99.8	96.3	97.1	80.3	80.1	97.8	88.1
Av. prob.	96.0	88.1	78.5	97.7	76.0	98.0	77.2	75.1	97.9	88.5

4.2.3. Evaluation by stratified cross-validation

Additionally, model evaluation studies using stratified cross-validation (CV) (Varma and Simon, 2006) were conducted on the basic models and on the models using data balancing methods, to further verify the results obtained in the previous sections and potentially further improve the performance of the models. For verification, we used stratified nested cross-validation with 5-folds and 3-nested-folds on the train set. For attempting to improve the models by optimizing hyperparameters, we used: (1) 5-fold stratified cross-validation on the train set to optimize hyperparameters, and (2) evaluation of the best model on the test set. The ranges of hyperparameters for the three main models are given in the code available in the GitHub repository mentioned above.

The results are presented in Tables 7 and 8. We can easily observe that they are mostly very similar to the results presented in sections 4.2.1 and 4.2.2. The optimization of hyperparameters can improve the properties of the models, but in general this is not true for our models and data set. One exception is the random forest model with oversampling, for which we have differences between the measures for the test set (sections 4.2.1 and 4.2.2) and both CV evaluations. This may be because we use 1000 trees for CV evaluation and the optimization of hyperparameters. Therefore, in this case, the optimization of hyperparameters changes the results.

5. Conclusions

In the tests conducted, we focused on comparing the results of different classification models and the impact of data balancing methods on their effectiveness in diagnosing diabetes. The logistic regression, random forests, and XGBoost models were evaluated, both in their basic versions and after applying various techniques aimed at improving the measures, such as variable reduction and hyperparameter changes, as well as undersampling, oversampling, an asymmetric approach, and SMOTE. Additionally, the ensemble learning approach in the form of average posterior probabilities was tested.

In general, the XGBoost model with undersampling has proved to be the best in the context of diagnosing diabetes, when the aim is to achieve the best diagnosis of sick people. In the case when we wish to choose a model with the best measures and a balance between them, the XGBoost model with the average of probabilities will be the best.

Table 8. Evaluation measures (as percentages) of the all models obtained on the test set for the best hyperparameters optimized by stratified cross-validation on the train set

Model	Acc	BalAcc	Sens	Spec	Prec	NPV	F1	MCC	ROC-AUC	PR-AUC
Logistic regression										
Original	96.1	81.0	62.9	99.1	86.2	96.7	72.7	71.7	96.2	81.4
Under.	89.0	88.6	88.1	89.1	42.0	98.8	56.9	56.1	96.3	81.3
Over.	88.8	88.7	88.6	88.9	41.7	98.9	56.8	56.1	96.3	81.3
As. appr.	92.6	87.5	81.5	93.6	53.4	98.3	64.4	62.2	96.3	81.3
SMOTE	89.0	88.4	87.8	89.1	42.0	98.8	56.8	56.0	96.2	81.3
Av. prob.	94.3	85.9	75.9	96.0	62.8	97.8	68.7	66.0	96.3	81.5
Random forests										
Original	97.4	84.4	69.0	99.9	98.6	97.3	81.1	81.2	97.6	88.2
Under.	90.5	91.3	92.3	90.3	46.2	99.2	61.6	62.3	97.8	88.3
Over.	93.5	90.2	86.3	94.2	57.1	98.7	68.7	67.0	97.7	88.2
As. appr.	94.6	89.7	83.8	95.6	63.0	98.5	71.9	69.8	97.8	88.3
SMOTE	97.2	85.1	70.7	99.5	93.1	97.4	80.4	79.7	97.4	87.3
Av. prob.	96.2	88.0	78.2	97.8	76.1	98.0	77.2	75.1	97.8	88.4
XGBoost										
Original	97.3	85.0	70.1	99.8	96.8	97.4	81.3	81.1	98.1	89.1
Under.	90.8	91.8	93.0	90.6	47.0	99.3	62.4	62.2	98.0	88.9
Over.	90.6	91.5	92.5	90.5	46.6	99.3	62.0	61.7	98.1	89.2
As. appr.	94.5	90.9	86.5	95.2	62.0	98.7	72.2	70.5	98.1	89.0
SMOTE	97.3	85.1	70.5	99.8	96.2	97.4	81.4	81.1	97.9	88.6
Av. prob.	96.2	88.6	79.5	97.7	75.5	98.2	77.5	75.4	98.1	89.1

Naturally, further investigation of classification models in diabetes research is necessary, exploring both their variations and diverse medical datasets. At the same time, attention must be paid to their limitations, to ensure that such models are used as supportive tools rather than standalone diagnostic systems. Nevertheless, we hope that our methods can be successfully generalized to other diabetes datasets.

Acknowledgements

We would like to thank the anonymous reviewers for their comments, which helped us to improve the manuscript.

References

Ambrish G., Ganesh B., Ganesh A., Srinivas C., Dhanraj Mensinkal K. (2022): Logistic regression technique for prediction of cardiovascular disease. Global Transitions Proceedings 3, 127-130.

American Diabetes Association Professional Practice Committee (2025): 2. Di-

agnosis and Classification of Diabetes: Standards of Care in Diabetes – 2025. Diabetes Care 48, S27-S49.

Ardila D., Kiraly A.P., Bharadwaj S., Choi B., Reicher J.J., Peng L., Tse D., Etemadi M., Ye W., Corrado G., Naidich D.P., Shetty S. (2019): End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine 25, 954-961.

Dai B., Chen R.C., Zhu S.Z., Zhang W.W. (2018): Using Random Forest Algorithm for Breast Cancer Diagnosis. International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 2018, pp. 449-452.

Fenner M.E. (2019): Machine Learning with Python for Everyone. Addison Wesley.

Fiłon J. (2019): Diabetes – a public health challenge in the 21st century. Medical University of Białystok. https://pbc.biaman.pl/Content/60948/Cukrzyca_wyzwanie_zdrowia_publicznego_w_XXI_w.pdf [accessed: June 30, 2025]. (in Polish)

Kasperczuk A., Dardzińska A. (2017): Logistic Regression Methods in Selected Medical Information Systems. T.K. Dang et al. (Eds.): Future Data and Security Engineering 2017, LNCS 10646, pp. 168-177.

Konukoglu E., Glocker B. (2020): Random forests in medical image computing. Handbook of Medical Image Computing and Computer Assisted Intervention, The Elsevier and MICCAI Society Book Series, pp. 457-480.

Laguarta J., Hueto F., Subirana B. (2020): COVID-19 artificial intelligence diagnosis using only cough recordings. IEEE Open Journal of Engineering in Medicine and Biology 1, 275-281.

Małowiecki A. (2023): Data Resampling Methods to Solve Data Imbalance Problem in Credit Card Fraud Detection. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław (in Polish). https://dbc.wroc.pl/Content/125401/Malowiecki_Metody_resamplingu_danych_w_rozwiazaniu.pdf

McKinney S.M., Sieniek M., Godbole V., Godwin J., Antropova N., Ashrafian H., Back T., Chesus M., Corrado G.S., Darzi A., Etemadi M., Garcia-Vicente F., Gilbert F.J., Halling-Brown M., Hassabis D., Jansen S., Karthikesalingam A., Kelly C.J., King D., Ledsam J.R., Melnick D., Mostofi H., Peng L., Reicher J.J., Romera-Paredes B., Sidebottom R., Suleyman M., Tse D., Young K.C., Fauw J.D., Shetty S. (2020): International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94.

Ogłoszka A., Smaga Ł. (2022): Classification methods in the diagnosis of breast cancer. Biometrical Letters 59, 99-126.

Rajpurkar P., Irvin J., Ball R.L., Zhu K., Yang B., Mehta H., Duan T., Ding D., Bagul A., Langlotz C.P., Patel B.N., Yeom K.W., Shpanskaya K., Blankenberg F.G., Seekins J., Amrhein T.J., Mong D.A., Halabi S.S., Zucker E.J., Ng A.Y., Lungren M.P. (2018): Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS medicine 15, e1002686.

Ranjbarzadeh R., Dorosti S., Ghoushchi S., Caputo A., Tirkolaee E., Ali S., Arshadi Z., Bendechache M. (2023): Breast tumor localization and segmentation using machine learning techniques: Overview of datasets, findings, and methods. Computers in Biology and Medicine 152, 106443.

R Core Team (2025): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R version 4.5.1 (2025-06-13). https://www.R-project.org/.

Reszke M., Smaga Ł. (2023): Machine learning methods in the detection of brain tumors. Biometrical Letters 60, 125-148.

Rosińska A. (2025): Statistical methods and machine learning algorithms in the diabetes prediction problem. Master's thesis in Data Science. Adam Mickiewicz University, Poznań (in Polish).

Saito T., Rehmsmeier M. (2015): The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE 10(3), e0118432.

Starmer J. (2022): The StatQuest Illustrated Guide To Machine Learning. StatQuest Publications.

Stoltzfus J. (2011): Logistic regression: A brief primer. Academic Emergency Medicine 18, 1099-1104.

Włodarczyk E., Diabetes prevention and early detection program. Łódź 2018-2020. https://rpo.lodzkie.pl/images/2017/808-nabor-10.3.2/zal14.pdf [accessed: June 30, 2025]. (in Polish)

Wojdan K., Moniuszko M. (2022): Artificial intelligence in medicine – current status and implementation challenges. NAUKA 3, 41-52. (in Polish)

Varma S., Simon R. (2006): Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7, 91.