Predicting Threat Degree for Onset of Type 2 Diabetes Mellitus Based on Machine Learning Methods

Gyuzel Shakhmametova[1](✉), Nikita Vakkazov[1], and Sofya Klimets[2]

[1] Computer Science and Robotics Department, Ufa State Aviation Technical University, Ufa, Russia
g.shakhmametova@gmail.com

[2] Faculty of General Medicine, Bashkir State Medical University, Ufa, Russia

Abstract. This article discusses the use of machine learning methods to predict the degree of threat for onset of type 2 diabetes mellitus in patients aged 25 years. Type 2 diabetes mellitus is a disease that complicates the course of other concomitant diseases, and predicting the threat of its occurrence is an important element in forming the trajectory of diagnosis and treatment of the patient. A binary classification model based on the logistic regression method has been proposed and developed. The developed approach identifies the threat of diabetes mellitus onset using indicators such as age, sex, body mass index (BMI) and glycated hemoglobin (HbA1c). The paper describes how to build a model, and how to generate and prepare data for binary logistic regression. To implement the described approaches, the Python programming language, the Jupyter Lab development environment and scikit-learn, scipy, pandas, and numpy packages were used. Performance analysis showed accuracy of the proposed model as 0.98 on test data. The developed software can be used as a separate application and be built as a module into the clinical decision support system #COMESYSO1120.

Keywords: Machine learning · Predicting · Logistic regression · Type 2 diabetes mellitus

1 Introduction

The introduction of IT technologies in healthcare facilitates the ability to collect, store and process huge amounts of medical data which can be presented in the form of numbers, text, images, videos, or sound recordings. In the vast majority of cases, the data are heterogeneous and poorly structured [1]. Currently existing machine learning methods provide an effective set of tools for the analysis and processing of biomedical data. Utilizing machine learning methods allows solving a number of problems related to diagnosis of diseases, monitoring and prediction of patient's condition, research and development of new drugs, and other problems [2].

R. Silhavy et al. (Eds.): CoMeSySo 2022, LNNS 596, pp. 770–779, 2023.
https://doi.org/10.1007/978-3-031-21435-6_66

Subsequent paragraphs, however, are indented. Despite the rapid development of video and sound analysis, a significant part of medical data is stored, as a rule, in tabular form [3]. Data is provided as a set of rows and columns that contain numeric or text information. In this form, it is convenient to store various information about the patient, about the results of medical research, about the medical history of a particular patient, etc. Data stored in tabular form are used in endocrinology to record the results of studies of various patient indicators and further analyze these indicators to detect the dynamics of disease development [4]. Endocrinological monitoring is very important for controlling diabetes mellitus, as the disease can very quickly get out of control and lead to death. The WHO global report [5] notes that the number of people with diabetes has increased by a factor of 4 since 1980. And by 2030, diabetes could be the 7th leading cause of death worldwide. Analysis of endocrinological monitoring results makes it possible not only to control the course of diabetes mellitus, but also to detect the threat of its appearance. To detect the degree of threat for the onset of this disease, it is proposed to use machine learning and data analysis methods.

This article describes an algorithm for predicting the degree of threat for the type 2 diabetes mellitus based on logistic regression. For model training, a sample consisting of 71,833 records was used, containing information on all patients with diagnosed type 2 diabetes mellitus in a single region of the Russian Federation, and including a synthesized data sample containing corresponding indicators of healthy people. The synthesized sample was generated using a developed algorithm based on data analysis of patients with type 2 diabetes mellitus. The second part of the article presents existing solutions in the field of predicting the degree of threat of diabetes mellitus. In the third part, the algorithm for predicting the degree of threat of type 2 diabetes mellitus is discussed in detail. The fourth part presents the results of testing and analysis of the effectiveness of the developed algorithm.

2 Problem Definition and Suggested Decision

2.1 State of Art

Presently, there is a large variety of software products for monitoring the condition of a patient with diabetes. This list includes:

- electronic diaries and journals;
- programs for monitoring measuring devices;
- calorie programs;
- electronic consultants.

Let us review some of them:

Diasend [6] is an online data management system to download data from various glucometers and insulin pumps. This is a web application, so its advantage is ease of access in case of working Internet. Its disadvantage is lack of intelligent analysis.

Glucose Buddy—Diabetes Helper 2.0 [7] is a mobile app that stores information about glucose levels, carbohydrates consumed, insulin, necessary medications and activity logs. There are also built-in reminders about taking medications and warnings. On the plus side, it allows adding product information using barcodes. Its main disadvantage is absence of Intelligent Data Processing System.

Health Tracker [8] is a program that helps the user track and display any health-related measurements. Its advantages are Windows and Mac platforms and simple user interface. Disadvantages include paid distribution and lack of data analysis functions.

Diabetes Pilot [9] makes it possible to print reports, save data for use in other programs or e-mails as well as edit and analyze data on your computer. Its advantages are a large list of parameters available for monitoring. Disadvantages include no version for Android.

GNU Gluco Control [10] is intended for type 1 diabetes. It supports storage of insulin, HA and food data. Its advantage is being an open source. However, this project rarely updated, which is its major flaw.

The main disadvantage of all ready-made solutions is the orientation towards the already existing disease of diabetes mellitus in the patient and the lack of functionality to predict the degree of threat of its occurrence, which today is an urgent and sought-after option.

2.2 Date Structure and Synthesis

The full sample of data for all patients with diabetes mellitus contains 154,970 records and has the following structure:

1. Patient Code.
2. Date of birth.
3. Sex.
4. Year of diagnosis.
5. Type of diabetes mellitus.
6. Class diagram.
7. Year of beginning of insulin therapy.
8. HbA1c, % (Last Visit).
9. Weight, kg (Last visit).
10. Height, cm (Last visit).
11. BMI, kg/m^2. (Last visit).
12. Others.

To construct a model for predicting the degree of threat of type 2 diabetes mellitus, records of patients over 25 years old and with type 2 diabetes mellitus disease were selected. This sub-sample included 36,833 records, each of which was a set of features for a unique patient. The following fields were selected as characteristics of experts:

1. Sex.
2. Age.
3. Body mass index (BMI).
4. Glycated hemoglobin (HbA1c), %.

This dataset contains 36,833 records of people with already developed diabetes mellitus. To teach the model, it is necessary to obtain similar data from healthy people. To do this, 35,000 records were synthesized with corresponding measures of healthy people.

Data generation is performed according to the following rules:

1. ***Sex*** $\in \{0, 1\}$. 50% share of one class.
2. ***Age*** $\in [25, 90]$ and has a normal distribution with $\mu = 60$, $\sigma = 30$. Noise generated from normal distribution is added to the data $\mu = 0$, $\sigma = 0.2$ (Fig. 1).

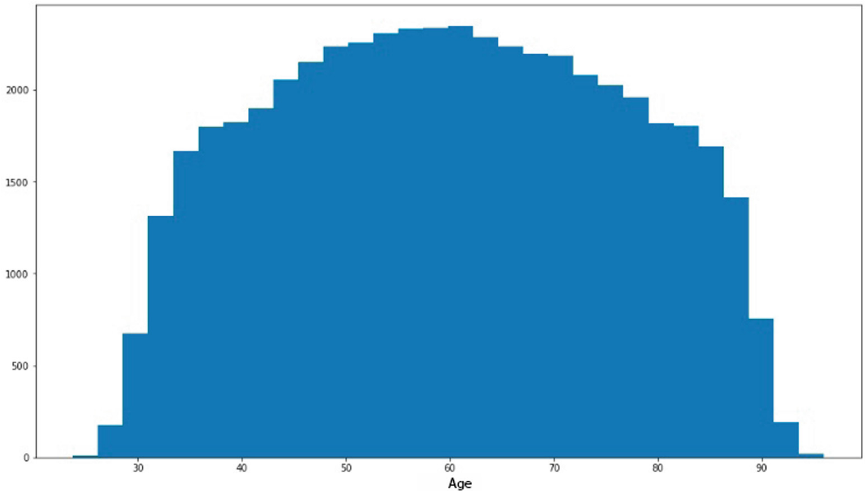


Fig. 1. Characteristic distribution ***Age***

3. ***Body mass index (BMI)*** $\in [19.5, 25.9]$ and has a normal distribution with $\mu = 21$, $\sigma = 2.5$, Noise generated from normal distribution is added to the data $\mu = 0$, $\sigma = 0.5$ (Fig. 2).

Glycated hemoglobin (HbA1c), %, $\in [3.6, 5.6]$ and has a normal distribution with $\mu = 4.6$, $\sigma = 0.5$. Noise generated from normal distribution is added to the data $\mu = 0$, $\sigma = 0.2$ (Fig. 3).

Data analysis and generation were performed in the Jupyter Lab environment [11] using the Python 3.9 programming language and Scipy [12], Matplotlib [13] libraries.

```
lower = 3.6
upper = 5.6
mu = 4.6
sigma = 0.5
```

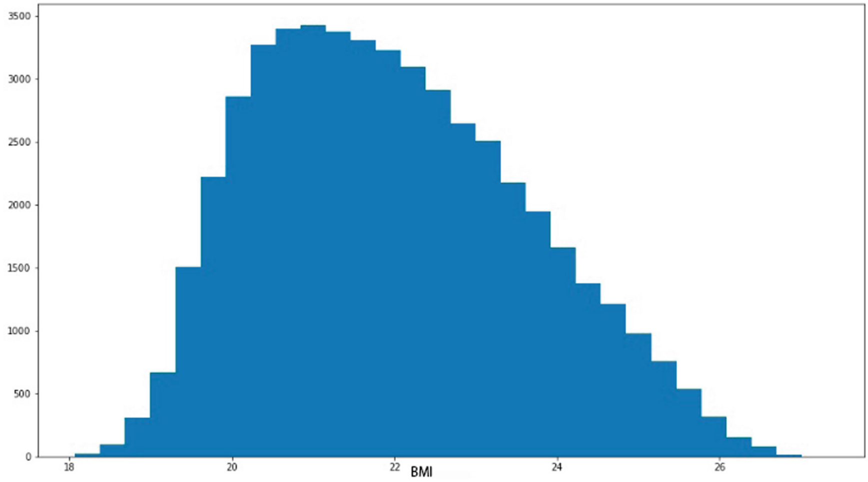


Fig. 2. ***BMI*** characteristic distribution

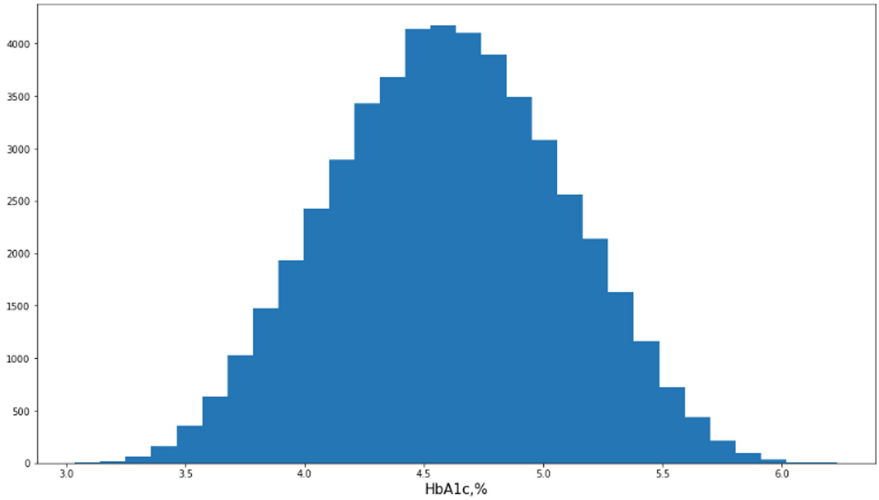


Fig. 3. Characteristic distribution HbA1c

```
hba1c_norm = scipy.status.truncnorm.rvc((lower-mu)/sigma,
                        (upper-mu)/sigma, loc=mu,
                        scale=sigma,
                        size=N)
hba1c = hba1c_norm + hba1c_noise
```

[Computer program code. Example of characteristic generation HbA1c]

The RVS method from the Scipy library stats.truncnorm module was used to generate normally distributed data. To generate noise, the normal method from the random library module Numpy was used. At the end, the sample and noise were added up to obtain the final generation result.

2.3 Construction of Logistic Regression Model

The task of predicting the degree of threat of type 2 diabetes mellitus can be reduced to the task of binary classification with the determination of the probability of belonging to each class. To solve this problem, a mathematical model based on the logistic regression method was chosen.

Logistic regression [14] is a method of constructing a linear classifier that allows evaluating the posterior probabilities of belonging to classes of objects.

Logistic regression equation:

$$\sigma(Z) = \frac{1}{1 + e^{-\hat{y}}} \tag{1}$$

$$\hat{y} = \sum_{i=0}^{n} W_i x_i \tag{2}$$

Error functionality:

$$L\big(\hat{y},\, y\big) = -\big(y\ \log\ \hat{y} + (1 - y)\log\big(1 - \hat{y}\big)\big) \tag{3}$$

The work uses binary logistic regression. The result of this algorithm is the probability of belonging to each class (class 0 - there is no diabetes, class 1 - there is diabetes). Classification threshold is 0.5.

To build the model, the Python programming language and the Sklearn [15] linear_mode module with the implementation of logistic regression were used. The random state parameter is set for the reproducibility of the experiment while multi class = ovr indicates that the model will be binary.

```
clf = LogisticRegression(random_state=0, multi_class=’ovr’)
                .fit(X_train, y_train)
```

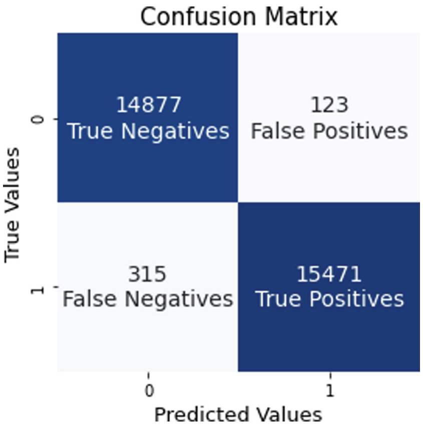
[Computer program code. Model object creation example and training run]

To assess the degree of threat, it is necessary to obtain the probability of the patient belonging to one of two classes. To do this, use the predict_proba method, which returns two numbers between 0 and 1. The first is the probability of belonging to class 0 (the patient does not have type 2 diabetes), and the second is the probability of belonging to class 1 (the patient has type 2 diabetes).

3 Results

Patient data from both the main sample and the synthesized set were used to test the model. The Sklearn.model_selection module was used for correct data splitting. Testing was carried out using the Sklearn.linear_model module. Metrics were calculated using the Sklearn.metrics module. Jupyter Lab was used.

Testing was performed on a sample of 30,786 records. Of these, 15,000 records belonged to class 0 (no diabetes of the 2nd degree) and 15,786 records belonged to class 1 (that is, diabetes of the 2nd degree). Figure 4 shows the error matrix.



	0	1
0	14877 True Negatives	123 False Positives
1	315 False Negatives	15471 True Positives

Fig. 4. Error matrix on test sample

Analysis of the error matrix shows that 30,348 records were classified correctly (of which 14,877 had no diabetes and 15,471 had diabetes), and 438 records were classified incorrectly (of which 123 records, the model classified the presence of diabetes and 315 labeled the absence of diabetes).

Calculation of metrics:

$$P = \frac{15471}{15471 + 123} = 0.99 \tag{4}$$

$$R = \frac{15471}{15471 + 315} = 0.98 \tag{5}$$

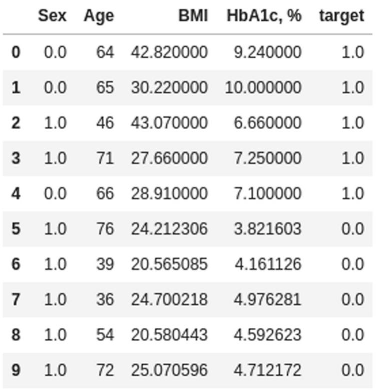
$$F = 2 \cdot \frac{0.99 \cdot 0.98}{0.99 + 0.98} = 0.99 \tag{6}$$

The metrics turned out to be almost the same. For the result, take the value of F measure.

$$\text{Error of the 1st kind} = \frac{123}{30,786} = 0.004 \tag{7}$$

$$\text{Error of the 2nd kind} = \frac{315}{30,786} = 0.01 \tag{8}$$

The efficiency of the algorithm was tested on 257 real new records. Figure 5 illustrates a fragment of 10 records (patients).



	Sex	Age	BMI	HbA1c, %	target
0	0.0	64	42.820000	9.240000	1.0
1	0.0	65	30.220000	10.000000	1.0
2	1.0	46	43.070000	6.660000	1.0
3	1.0	71	27.660000	7.250000	1.0
4	0.0	66	28.910000	7.100000	1.0
5	1.0	76	24.212306	3.821603	0.0
6	1.0	39	20.565085	4.161126	0.0
7	1.0	36	24.700218	4.976281	0.0
8	1.0	54	20.580443	4.592623	0.0
9	1.0	72	25.070596	4.712172	0.0

Fig. 5. Fragment of patient data and healthy people

4 Discussions

Each record is represented by the characteristics described earlier and the target field, which contains a class label (0—healthy, 1—sick). Figure 6 shows the classification errors.

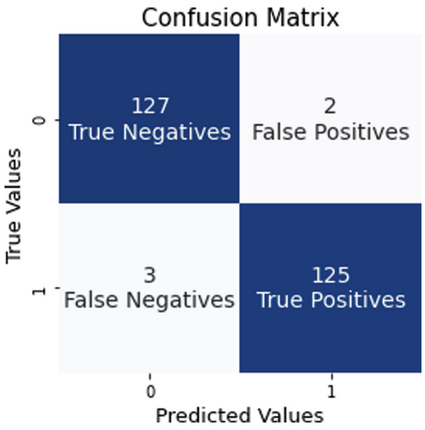


Fig. 6. Error matrix on a sample of 257 entries

Analysis of the error matrix shows that 252 records were classified correctly (of which 127 did not have diabetes and 125 had diabetes), and 5 real new records were classified incorrectly (of which 2 records, the model classified the presence of diabetes, and 3 marked the absence of diabetes).

Calculation of metrics:

$$P = \frac{125}{125 + 2} = 0.98 \tag{9}$$

$$\mathrm{R} = \frac{125}{125 + 3} = 0,98 \tag{10}$$

$$F = 2 \cdot \frac{0.98 \cdot 0.98}{0.98 + 0.98} = 0.98 \tag{11}$$

Next, we calculate errors of the 1st and 2nd kind:

$$\text{Error of the 1st kind} = \frac{2}{257} = 0.008 \tag{12}$$

$$\text{Error of the 2nd kind} = \frac{3}{257} = 0.012 \tag{13}$$

Thus, the proposed model with 98% accuracy classified all the examples provided.

In addition to classification, the degree of threat for onset of diabetes may also be predicted. Figure 7 shows the probability distribution of such forecasts.

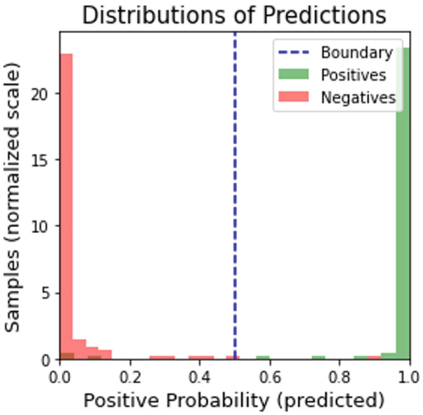


Fig. 7. Probability distribution of the forecast

Analysis of the probability distribution for predictions shows that most of them are close to 0 or 1. This suggests that in this case, the algorithm assigns a particular class label with high confidence.

5 Conclusions

The result of the study is the development of a model to assess the degree of threat for onset of type 2 diabetes mellitus in patients over 25 years of age based on machine learning methods. Model development involves 2 stages, i.e., preparation of real and synthesized data for further model training, and training in binary logistic regression. Statistical and data processing methods are used to implement the data preparation phase. The software is implemented in the Python 3.9 programming language; Pandas 1.1.3, Numpy 1.19.5, Scipy 1.6.2 packages were used for data processing, and Scikit-learn 0.23.2 packages were used for machine learning algorithms. The test results showed a high classification accuracy of 99%. The developed software is cross-platform and can function both as a separate application and as an embedded module in the clinical decision support system for endocrinologists.

At the moment, the developed software functions as a separate application, but in the future it will be built as a module into the clinical decision support system for the prevention and treatment of bronchopulmonary diseases.

Acknowledgements. The reported study was funded by RSF, project number 22-19-00471.

References

1. Abiteboul, S.: Querying semi-structured data. In: Afrati, F., Kolaitis, P. (eds.) Database Theory—ICDT '97. Lecture Notes in Computer Science, vol. 1186, pp. 1–18. Springer, Heidelberg (1997). https://doi.org/10.1007/3-540-62222-5_33
2. Muthalaly, R.G., Evans, R.M.: Applications of machine learning in cardiac electrophysiology. Arrhythm Electrophysiol. Rev. **9**, 71–77 (2020)
3. Mostafa, F., Hasan, E., Williamson, M., Khan, H.: Statistical machine learning approaches to liver disease prediction. Livers **1**, 23 (2021)
4. Bland, M., Peacock, J.: Statistical questions in evidence-based medicine (2000)
5. World Health Organization. https://www.who.int/ (date of the request: 15.08.22)
6. Diasend. https://diasend.com//us. Last Accessed 15 Jul 2022
7. Glucose Buddy. https://www.glucosebuddy.com (date of the request: 15.08.22)
8. Health tracker. http://www.blackcatsystems.com/software/healthtracker.html. Last Accessed 15 Jul 2022
9. Diabetes Pilot. http://www.diabetespilot.com/desktop. Last Accessed 15 Jul 2022
10. GNU Gluco control. http://ggc.sourceforge.net/. Last Accessed 15 Jul 2022
11. Jupyter documentation. https://jupyter.org/. Last Accessed 15 Jul 2022
12. Scipy documentation. https://docs.scipy.org/doc/scipy/. Last Accessed 15 Jul 2022
13. Matplotlib documentation. https://matplotlib.org/stable/index.html. Last Accessed 15 Jul 2022
14. Hosmer, D.W. Jr., Lemeshow, S., Sturdivant R.X.: Applied Logistic Regression, 3rd edn
15. Scikit-learn documentation. https://scikit-learn.org/stable/index.html. Last Accessed 15 Jul 2022