Classification of Diabetes Disease Using Logistic Regression Method

Andrew Hrimov¹, Ievgen Meniailov¹, Dmytro Chumachenko¹(✉), Kseniia Bazilevych¹, and Tetyana Chumachenko²

¹ National Aerospace University “Kharkiv Aviation Institute”,
17 Chkalova Str., Kharkiv 61070, Ukraine
andrew.hrimov@gmail.com, evgenii.menyailov@gmail.com, dichumachenko@gmail.com, ksenia.bazilevich@gmail.com

² Kharkiv National Medical University, 4 Nauky Av., Kharkiv 61022, Ukraine
tatlchum@gmail.com

Abstract. At the moment, there are many methods of analysis and classification aimed at building the most accurate and effective mathematical models that are widely used in medicine as a decision-making tool. Existing methods make it possible to identify the relationships between input and output variables in the sample, build models reflecting these relationships, compare them in terms of accuracy, profitability and costs, and choose the most effective model. The increase in the incidence of diabetes not only in the world, but also in Ukraine, dictates the need to introduce a mathematical apparatus for automatic diagnosis of the disease. Within the framework of the study, the classification of patients with diabetes by the logistic regression method was implemented. Python is used for software implementation.

Keywords: Machine learning · Diagnostics · Classification · Diabetes · Logistic regression

1 Introduction

The global coronavirus pandemic has once again demonstrated the need to introduce digital tools into healthcare [1]. The digitalization of medicine is one of the most urgent tasks in the modern world, and already created products and solutions are used in various areas of public health: in the management of medical institutions [2–5], diagnostics of various diseases [6–8], modeling epidemic processes [9, 10] and predicting morbidity [11], surgical treatment [12, 13] and even in training of medical personnel [14–17].

The modern development of information technologies makes it possible to develop not just shells for automating the work of medical institutions [18], but also complex systems based on methods and means of artificial intelligence [19], multi-agent modeling [20], game theory [21], decision theory [22], machine learning [23, 24], computer vision [25, 26] and other modern methods and approaches.
The worldwide attention to the incidence of Covid-19 not only does not diminish the importance of global epidemics of other diseases, but often exacerbates them. One of these diseases is diabetes.

Diabetes is a chronic disease that develops when the pancreas does not produce enough insulin, or when the body cannot use the insulin it makes efficiently. Insulin is a hormone that regulates blood sugar levels. A common result of uncontrolled diabetes is hyperglycemia, or elevated blood sugar levels, which over time leads to severe damage to many systems in the body, especially nerves and blood vessels.

According to the latest official data from the Ministry of Health, there were 1.27 million people with diabetes in Ukraine. Among them, almost 200,000 patients require daily insulin intake. From 2010 to 2017, the total number of patients increased by 4%, and the rate per 100 thousand population – by 12%. The specific weight of diabetes mellitus cases among all diseases during this period increased by 0.3% (from 1.4% to 1.7%). According to the Public Health Center, in Ukraine, almost half of the patients with diabetes are not diagnosed.

One of the tools in solving the problem of diabetes diagnosis is the use of a machine learning apparatus to identify infected based on test data. Thus, the aim of the research is the automated classification of patients with suspected diabetes based on the logistic regression method.

2 Materials and Methods

At the moment, there are many methods of analysis and classification aimed at building the most accurate and effective mathematical models that are widely used in medicine as a decision-making tool [27]. Existing methods make it possible to identify the relationships between input and output variables in the sample, build models reflecting these relationships, compare them in terms of accuracy, profitability and costs, and choose the most effective model [28, 29]. In our case, this may be the presence or absence of diabetes.

Linear regression is used to model linear relationships between a continuous output variable and a set of input variables [30]. Under certain conditions, the linear regression equation serves as an irreplaceable and very high-quality tool for analysis and forecasting. The linear regression model is the most common and simplest equation for the relationship between input and output variables. In addition, the constructed linear regression equation can be the starting point for data analysis.

When analyzing data, there are often problems where the output variable is categorical, and then the use of linear regression is difficult. Therefore, when looking for relationships between a set of input variables and a categorical output variable, logistic regression has become widespread. Logistic regression is a binary classification method. It allows you to estimate the probability of realization (or non-realization) of an event depending on the values of some independent variables. The logistic regression line, unlike the linear one, is not straight.

ROC curve (Receiver Operator Characteristic) is a curve used to represent the results of binary classification in machine learning. Since there are two classes, one of them is called a class with positive outcomes, the other – with negative outcomes.
The ROC curve shows the dependence of the number of correctly classified positive examples on the number of incorrectly classified negative examples.

In the terminology of ROC analysis, the former are called true positive, and the latter, false negative. In this case, it is assumed that the classifier has a certain parameter, by varying which, we will obtain one or another division into two classes. This parameter is often called the threshold, or cut-off value. Depending on it, different values of type I and II errors will be obtained.

In logistic regression, the cut-off threshold ranges from 0 to 1 – this is the calculated value of the regression equation. Let's call it a rating.

To understand the essence of type I and II errors, consider a four-field confusion matrix (Table 1), which is built on the basis of the results of classification by the model and the actual (objective) belonging of the examples to classes.

| Model | Actually positive | Actually negative |
|-------|-------------------|-------------------|
| Positive | *True Positives* are correctly classified positive examples (the so-called true positive cases) | *False Positives* are negative examples classified as positive (type II error). This is a false detection because in the absence of an event, a decision is mistakenly made about its presence (false positive cases) |
| Negative | *False Negatives* are positive examples classified as negative (Type I error). This is the so-called “false pass” – when the event of interest to us is mistakenly not detected (false negative examples) | *True Negatives* are correctly classified negative examples (true negative cases) |

What is positive and what is negative depends on the specific task.

When analyzing, they often operate not with absolute indicators, but with relative shares (rates), expressed as a percentage:

- the proportion of True Positive Rate:

\[
TPR = \frac{TP}{TP + FN} \cdot 100\%;
\]

- the proportion of False Positive Rate:

\[
FPR = \frac{FP}{TN + FP} \cdot 100%.
\]

Let us introduce two more definitions: sensitivity and specificity of the model. They determine the objective value of any binary classifier.
Sensitivity is the proportion of truly positive cases:

\[ S_e = TPR = \frac{TP}{TP + FN} \cdot 100\% . \]

Specificity is the proportion of true negative cases that were correctly identified by the model:

\[ S_p = \frac{TN}{TN + FP} \cdot 100\% . \]

A high-sensitivity model often gives a true result if there is a positive outcome. Conversely, a model with high specificity is more likely to give a true result when there is a negative outcome (it detects negative examples). If we talk in terms of medicine - the problem of diagnosing a disease, where the model for classifying patients into sick and healthy is called a diagnostic test, then we get the following:

- a sensitive diagnostic test manifests itself in overdiagnosis – the maximum prevention of missing patients;
- a specific diagnostic test only diagnoses patients with certainty. This is important in the case when, for example, the treatment of a patient is associated with serious side effects and overdiagnosis of patients is not desirable.

## 3 Results

For implementation the classification method we have used open dataset of Diabetes patients: PIMA Indians Diabetes Database. Each instance represents individual patients and their various medical attributes along with diabetes classification. Database has 768 instances and 9 attributes (Table 2).

| Parameter name | Description | Data type |
|----------------|-------------|-----------|
| Pregnancies    | Number      | Decimal   |
| PG Concentration| Count       | Integer   |
| Diastolic BP   | Count       | Integer   |
| Tri Fold Thick | Count       | Integer   |
| Serum Ins      | Count       | Integer   |
| BMI            | Count       | Integer   |
| DP Function    | Count       | Integer   |
| Age            | Years       | Decimal   |
| Diabetes       | Present or not | 0/1         |
First step of data analysis is data preprocessing. For analysis and program realization we have used Python language. First of all, we need to import necessary modules and upload data from database (Fig. 1).

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix

df1 = pd.read_excel('Diabetes.xls')
df1.head()
```

![Fig. 1. Data import.](image1)

For correct analysis we have to change format of “infected” and “healthy” values to “0” and “1”. Let’s form data frames of characteristics and Boolean values of the disease. Next, we will set the data for training and validation and build a logistic regression model. Next, we will check it on test data and find the accuracy of its classification.

Figure 2 shows the error matrix for the constructed model. Here 39 is the number of correctly predicted healthy people, 35 are incorrectly predicted healthy people, 16 are incorrectly predicted patients and 141 people were correctly identified as sick, in other words, the model correctly predicted 39 + 141 = 180 people, and was mistaken in the case of 35 + 16 = 52 persons.

```
confusion_matrix = confusion_matrix(y_test, y_pred)
print(confusion_matrix)
```

```
[[39 35]
 [16 141]]
```

![Fig. 2. Error matrix.](image2)

To improve the accuracy of the model, it is necessary to analyze the characteristics that were used for the classification (Fig. 3). Here, you can see the differences in mean values for sick and healthy patients, in order to better understand the influence of each characteristic, we visualize their values (Fig. 4).

Next, let’s build several models, taking into accounts the factors Age, Pregnancies, Serum Ins, DP Function, and the second – PG Concentration, Diastolic BP, Tri Fold Thick and BMI, similarly creating data frames and setting test and data for training the model.
Fig. 3. Characteristics used for classification.

| Diabetes | Pregnancies | PG Concentration | Diastolic BP | Tri Fold Thick |
|----------|-------------|------------------|--------------|---------------|
| 0        | 4.865672    | 141.257463       | 70.024627    | 22.164179     |
| 1        | 3.290000    | 105.900000       | 68.104000    | 19.664000     |

| Diabetes | Serum Ins | BMI | DP Function | Age |
|----------|-----------|-----|-------------|-----|
| 0        | 100.335821 | 35.142537 | 0.550500 | 37.867164 |
| 1        | 68.792000  | 36.384200 | 0.429734 | 31.190000 |

Fig. 4. Data visualization.
The first model classifies with an accuracy of 0.6926. Second model with an accuracy of 0.7706. ROC-curves of the general model, the model of factors 1 and factors 2 are presented in Fig. 5, 6 and 7, respectively.

Analysis shows that the characteristics of the second group have a greater impact on the classification accuracy. The dashed line represents the ROC curve of a completely random classifier. A good classifier remains as far from it as possible (towards the upper left corner). In this case, the optimal classifiers can be called those presented in Fig. 5 and 7.

![Total Logistic Regression Model](image1)

**Fig. 5.** Model of all factors.

![First Group of Features](image2)

**Fig. 6.** Model of first group of factors (Age, Pregnancies, Serum Ins, DP Function).
4 Conclusions

The article presents a logistic regression method as a tool for developing a mathematical-statistical model for predicting the probability of an event of interest to the researcher in the presence of two possible outcomes. The ROC analysis method was selected and described in detail as a tool for assessing the quality of the model. The capabilities of these methods are demonstrated by a real example of creating and evaluating the effectiveness (sensitivity and specificity) of a model for predicting the likelihood of diabetes incidence.

The analysis showed that the factors PG Concentration, Diastolic BP, Tri Fold Thick and BMI most affect the accuracy of the disease detection.

The accuracy of the model was 77%. At first glance, the accuracy of the model is relatively low for the problem of diagnosing morbidity. However, in our case, the option with the maximum sensitivity and specificity of the tests was chosen, which indicates overdiagnosis of patients. In the task of diagnosing diabetes, this is the best option, because a false-positive result can threaten, for example, only an additional visit to the doctor, and a false-negative result can not reveal a dangerous, but curable disease.

Acknowledgment. The study was funded by the Ministry of Health of Ukraine for the state budget in the framework of the research work on the theme “To develop a scientifically substantiated strategy of prevention antibiotic resistance of the bacteria causing of healthcare-associated infections in healthcare facilities” (State registration number 0118U000944).
References

1. Ohannessian, R., Duong, T.A., Odone, A.: Global telemedicine implementation and integration within health systems to fight the COVID-19 pandemic: a call to action. JMIR Public Health Surveill. 6(2), e18810 (2020). https://doi.org/10.2196/18810

2. Dotsenko, N., Chumachenko, D., Chumachenko, I.: Modeling of the process of critical competencies management in the multi-project environment. In: 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT), vol. 3, pp. 89–93. IEEE, Lviv (2019). https://doi.org/10.1109/STC-CSIT.2019.8929765

3. Dotsenko, N., Chumachenko, D., Chumachenko, I.: Management of critical competencies in a multi-project environment. In: CEUR Workshop Proceedings, vol. 2387, pp. 495–500 (2019)

4. Dotsenko, N., Chumachenko, D., Chumachenko, I.: Project-oriented management of adaptive teams’ formation resources in multi-project environment. In: CEUR Workshop Proceedings, vol. 2353, pp. 911–920 (2019)

5. Dotsenko, N., Chumachenko, D., Chumachenko, I.: Modeling of the processes of stakeholder involvement in command management in a multi-project environment. In: 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), vol. 1, pp. 29–32. IEEE, Lviv (2018). https://doi.org/10.1109/STC-CSIT.2018.8526613

6. Bazilevych, K., Meniailov, I., Fedulov, K., et al.: Determining the probability of heart disease using data mining methods. In: CEUR Workshop Proceedings, vol. 2488, pp. 383–394 (2019)

7. Strilets, V., Bakumenko, N., Chernysh, S., et al.: Application of artificial neural networks in the problems of the patient’s condition diagnosis in medical monitoring systems. In: Nechyporuk, M., et al. (eds.) Integrated Computer Technologies in Mechanical Engineering. AISC, vol. 1113, pp. 173–185. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37618-5_16

8. Bakumenko, N., Strilets, V., Ugryumov, M.: Application of the C-means fuzzy clustering method for the patient’s state recognition problems in the medical monitoring systems. In: CEUR Workshop Proceedings, vol. 2362, pp. 218–227 (2019)

9. Chumachenko, D., Meniailov, I., Bazilevych, K., et al.: Development of an intelligent agent-based model of the epidemic process of syphilis. In: 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT), vol. 1, pp. 42–45. IEEE, Lviv (2019). https://doi.org/10.1109/STC-CSIT.2019.8929749

10. Chumachenko, D., Chumachenko, T.: Intelligent agent-based simulation of HIV epidemic process. In: Lytvynenko, V., et al. (eds.) Lecture Notes in Computational Intelligence and Decision Making. ISDMCI 2019. AISC, vol. 1020, pp. 175–188. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-26474-1_13

11. Polyvianna, Y., Chumachenko, D., Chumachenko, T.: Computer aided system of time series analysis methods for forecasting the epidemics outbreaks. In: 2019 IEEE 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), pp. 1–4. IEEE, Polyana (2019). https://doi.org/10.1109/CADSM.2019.8779344

12. Nechyporenko, A.S., Krivenko, S.S., Alekseeva, V., et al.: Uncertainty of measurement results for anatomical structures of paranasal sinuses. In: 2019 8th Mediterranean Conference on Embedded Computing (MECO), pp. 1–4. IEEE, Budva (2019). https://doi.org/10.1109/MECO.2019.8760032
13. Alekseeva, V., Lupyr, A., Urevich, N., et al.: Significance of anatomical variations of maxillary sinus and ostiomeatal components complex in surgical treatment of sinusitis. Novosti Khirurgii 27(2), 168–176 (2019). https://doi.org/10.18484/2305-0047.2019.2.168

14. Herasymova, A., Chumachenko, D., Padalko, H.: Development of intelligent information technology of computer processing of pedagogical tests open tasks based on machine learning approach. In: CEUR Workshop Proceedings, vol. 2631, pp. 121–130 (2020)

15. Zabolotnyi, O.V., Zabolotnyi, V.A., Koshevoi, M.D.: Conditionality examination of the new testing algorithms for coal-water slurries moisture measurement. Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu 1, 51–59 (2018)

16. Chumachenko, D., Balitskii, V., Chumachenko, T., et al.: Intelligent expert system of knowledge examination of medical staff regarding infections associated with the provision of medical care. In: CEUR Workshop Proceedings, vol. 2386, 321–330 (2019)

17. Zabolotnyj, A.V., Koshevoj, N.D.: Development, examination and optimization of the device for quality control of dielectric materials. Pribory i Sistemy Upravleniya 1, 39–42 (2004)

18. Mazorchuck, M., Dobriak, V., Chumachenko, D.: Web-application development for tasks of prediction in medical domain. In: 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), vol. 1, pp. 5–8. IEEE, Lviv (2018). https://doi.org/10.1109/STC-CSIT.2018.8526684

19. Mashtalir, V.P., Shylakho, V.V., Yakovlev, S.V.: Group structures on quotient sets in classification problems. Cybern. Syst. Anal. 50(4), 507–518 (2014). https://doi.org/10.1007/s10559-014-9639-z

20. Chumachenko, D., Chumachenko, K., Yakovlev, S.: Intelligent simulation of network worm propagation using the code red as an example. Telecommun. Radio Eng. 78(5), 443–464 (2019). https://doi.org/10.1615/TelecomRadEng.v78.i5.60

21. Chumachenko, D., Meniaiof, I., Bazilevych, K., Chumachenko, T.: On intelligent decision making in multiagent systems in conditions of uncertainty. In: 2019 Xth International Scientific and Practical Conference on Electronics and Information Technologies (ELIT), pp. 150–153. IEEE, Lviv (2019). https://doi.org/10.1109/ELIT.2019.8892307

22. Mashtalir, V.P., Yakovlev, S.V.: Set methods of clusterization of standard information. Cybern. Syst. Anal. 37(3), 295–307 (2001). https://doi.org/10.1023/A:1011985908177

23. Karatekin, T., Sancak, S., Celik, G., et al.: Interpretable machine learning in healthcare through generalized additive model with pairwise interactions (GA2M): predicting severe retinopathy of prematurity. In: 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), pp. 61–66. IEEE, Istanbul (2019). https://doi.org/10.1109/Deep-ML.2019.00020

24. Piletskiy, P., Chumachenko, D., Meniaiof, I.: Development and analysis of intelligent recommendation system using machine learning approach. In: Nechyporuk, M., et al. (eds.) Integrated Computer Technologies in Mechanical Engineering. AISC, vol. 1113, pp. 186–197. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37618-5_17

25. Gargin, V., Radutny, R., Titova, G., et al.: Application of the computer vision system for evaluation of pathomorphological images. In: 2020 IEEE 40th International Conference on Electronics and Nanotechnology (ELNANO), pp. 469–473. IEEE, Kyiv (2020). https://doi.org/10.1109/ELNANO50318.2020.9088898

26. Nechyporenko, A., Reshetnik, V., Alekseeva, V., et al.: Assessment of measurement uncertainty of the uncinated process and middle nasal concha in spiral computed tomography data. In: 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), pp. 585–588. IEEE, Kyiv (2019). https://doi.org/10.1109/PICST47496.2019.9061557
27. Bondarenko, A.V., Pokhil, S.I., Lytvynenko, M.V., et al.: Anaplasmosis: experimental immunodeficient state model. Wiadomosci Lekarskie 72(9), 1761–1764 (2019)

28. Kozko, V.M., Bondarenko, A.V., Gavrylov, A.V., et al.: Pathomorphological peculiarities of tuberculous meningoencephalitis associated with HIV infection. Intervent. Med. Appl. Sci. 9 (3), 144–149 (2017). https://doi.org/10.1556/1646.9.2017.31

29. Lytvynenko, M., Shkolnikov, V., Bocharova, T., et al.: Peculiarities of proliferative activity of cervical squamous cancer in HIV infection. Georgian Med. News 270, 10–15 (2017)

30. Kim, K., Seol, J.A., Kim, D.B., et al.: Health management based on history of personalized physiological data using linear regression analysis. In: 2018 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE), p. 1. IEEE, Porto (2018). https://doi.org/10.1109/GMEPE-PAHCE.2018.8400736