*Advances in Economics, Business and Management Research, volume 138*
**2nd International Scientific and Practical Conference "Modern Management Trends and the Digital Economy: from Regional Development to Global Economic Growth" (MTDE 2020)**

# Development of a Model for Predicting Treatment of Cardiovascular Diseases Based on Machine Learning Methods

Bolodurina I.P.* Parfenov D.I. Zhigalov A.Yu. Zabrodina L.S.

*Department of Applied Mathematics, Orenburg State University, Orenburg, 460018, Russia*
**Corresponding author. Email: prmat@mail.osu.ru*

**ABSTRACT**
This study aims to build a model for predicting cardiovascular disease in patients based on the analysis of personalized patient data cards. The forecast for the treatment of the heart disease clinic was determined using the method of logistic regression, random trees for the algorithm for constructing ID3 decision trees and the ensemble training method - random forest. As part of an experimental study, the effectiveness of the application of the considered methods for forecasting was evaluated based on the analysis of the ROC curve and the AUC metric. Experiments on real datasets of patient visits to the clinic showed that for short-term forecasting, the ID3 algorithm for constructing decision trees showed better results, and with an increase in the period under consideration, the method of logistic regression turned out to be more effective.
***Keywords:*** *logistic regression, decision trees, random forest, heart disease, learning algorithm*

## 1. INTRODUCTION

Currently, Big Data technologies have become one of the dominant trends in the development of information technology. It is assumed that working with colossal volumes of unstructured data will have the greatest impact on production [1], government [2], trade [3] and medicine [4].
Thanks to the methods of data mining, it became possible to study the effectiveness of treatment by processing all available information about treatment practice [5]. Based on the analysis of all known medical histories and diagnostics, the practice of doctors includes the widespread use of decision support systems that allow access to the experience of thousands of colleagues across the country.
Methods of personal and preventive medicine based on remote monitoring of patients will lead to a significant reduction in costs and an increase in the quality of life [6]. The proliferation of various sensors of human body activity connected to gadgets reduces the need for laboratory tests, prevents unexpected complications, and an automatic reminder of the need for independent treatment and prophylactic manipulations will increase the quality of the prescribed treatment [7].
One of the effective methods of researching patient data and their medical history is machine learning methods. In the last decade, due to the inaccessibility of personal data about patients, methods based on intuition and heuristics have been used to solve forecasting problems - making diagnoses that do not always give adequate and timely results. Currently, each patient is listed in the clinic database and has a history of visits that allow you to calculate statistical characteristics, such as the average number of visits with certain types of diseases, the average length of stay in the clinic, diagnoses, and others. Due to the accumulated data array, doctors can move away from the heuristic definition of the disease and, using the experience of colleagues, faster, more timely, and most importantly more accurately make the correct diagnosis [8].
Machine learning-based data analysis approaches include several well-known problem-solving methods that allow you to analyze unbalanced data sets for classification and prediction. These include logistic regression, hidden Markov chains, decision trees, and random forests. Each of these methods has its advantages over similar methods for solving forecasting problems. Nevertheless, it is worth noting that today, in the practice of applying machine learning methods, there are examples of the effectiveness of various approaches, therefore, before making a forecast, it is necessary to compare the efficiency of the algorithms for the data set in question.

### *1.1. Related Work*

Research of medical data for forecasting, classification, and automation of internal processes is carried out all over the world.
The author's team from The national research University of the Higher School of Economics in the publication [9] considers the use of modern cloud technologies in the storage and processing of cardiac information. In particular, the paper [10] by researcher E. Yu. Zimina

 

considers ways to solve the problem of diagnosing the patient's heart health using data Mining classification methods for processing cardiac data. The cluster analysis was based on the search for similar forms of Fourier spectra obtained by modeling the work of the heart using the Fermi-Layer-Ulan decomposition.
The authors of the study [11] note the prospects of the Big Data analysis method for evaluating qualitative and quantitative indicators of pharmacotherapy in patients with arterial hypertension. In the framework of the publication [12], a review of methods and systems for the intellectual analysis of medical data is performed, and an architecture and software platform for analyzing heterogeneous sources of structured and unstructured data is proposed.
In the I.V. Stepanyan's dissertation research is devoted to the development of theoretical and methodological aspects of risk management using bioinformatics technologies to predict employee health disorders [13]. The bionic self-organizing Kohonen network was used for cluster analysis.
The authors of the study [14] ask the question of using machine learning methods to improve the prediction of the risk of cardiovascular diseases, based on the processing of clinical Practice Research Datalink (CPRD) arrays of clinical data. Experimental results show the improvement of prediction accuracy. Scientist Shankar M. Krishnan from the Wentworth Institute of Technology (USA) notes in [15] that the use of Analytics in the health sector together with effective organization, optimization, and analysis of big data provides fast and accurate diagnosis, as well as reducing the number of preventable errors.
The publication [16] assesses the global risk management of cardiovascular diseases in clinical practice among physicians divided into groups according to the use of conventional or electronic support for the collection and registration of clinical data.
Thus, the review of studies has shown that the use of machine learning technologies in the processing of cardiac data to solve the problem of diagnosis is one of the most pressing issues at the moment.

### *1.2. Our Contribution*

This study is aimed at making a forecast for the patient to determine whether they will go to the polyclinic with cardiovascular diseases (CVD) in the next month or 3 months based on the analysis of personalized cards.
The forecast for treatment in a polyclinic with heart diseases was determined using the logistic regression method, the ID3 decision tree construction algorithm, and the ensemble training method - random forest. As part of the experimental study, the effectiveness of the considered methods for forecasting based on the analysis of the ROC curve and the AUC metric was evaluated.

### *1.3. Paper Structure*

The article is organized as follows. Section 2 presents a mathematical formalization of the problem. Section 3 briefly describes how to solve the problem of predicting cases of cardiovascular diseases using machine learning methods such as logistic regression, the ID3 decision tree algorithm, and random forests. Section 4 is devoted to experiments with real data from the history of patient visits to the polyclinic. Section 5 concludes the paper and presents direction for future research.

## 2. THE MODEL FOR PREDICTING TREATMENT OF CARDIOVASCULAR DISEASES BY USING MACHINE LEARNING METHODS

### *2.1. Problem statement*

Let's look at the clinic's database, which consists of patients $U = \{u_1, u_2, ..., u_k\}$ and records that characterize individual visits $C = \{c_1, c_2, ..., c_p\}$. The source data set is represented by a table and each column corresponds to one of the following characteristics of the visit record:

1. ID – ID number of the patient in clinic's;
2. MO – code of the medical institution;
3. CODE_MP – medical care code;
4. DATE_IN – date of arrival at the medical institution;
5. DATE_OUT – date of exit from the medical institution;
6. DLITELN – duration of treatment;
7. ICD – disease code for ICD;
8. POL – gender code (1-male, 2-female);
9. VOZRAST – the current age of the patient;
10. ATER – whether atherosclerosis was detected when the patient was admitted;
11. ISHEM - whether ischemia was detected when the patient was admitted;
12. GIPER - whether hypertension is detected when receiving a patient;
13. STENOK - whether stenocardia was detected when the patient was admitted;
14. INF_MIOK - whether a myocardial infarction was detected when the patient was admitted.

Based on the presented initial data set, it is quite difficult to estimate the forecast of a patient's arrival after a fixed time with cardiovascular diseases, so it is necessary to prepare the data.
It is difficult to take into account an important criterion for forecasting – time. To account for it, you can artificially create features that aggregate some patient indicators over a certain time. For example, this can be the number of visits over the past six months. The signs that identify the visit and patient information will remain unchanged. This

study collects statistics on the number of visits to each disease and disease class (ICD) for the last 3 and 6 months.

Note that for a more accurate forecast, you need to throw out records for the first months and do not take them into account in the forecast (exclude patients with a missing medical history), this also applies to the last records (exclude patients for whom it is impossible to check the forecast).

Thus, as a result of statistical data processing, the following additional features were identified:

1. ICD – ICD disease code not detailed;
2. COUNT_CODE_MP – number of visits for each type of institution in the last six months;
3. BOOL_CODE_MP – whether there was a visit for each type of institution in the last 3 months;
4. ATER_6M – how many times has atherosclerosis been detected when receiving a patient in the last 6 months;
5. ATER_3M – whether atherosclerosis was detected when receiving a patient in the last 3 months;
6. ISHEM_6M – how many times was ischemia detected when receiving a patient in the last 6 months;
7. ISHEM_3M – whether ischemia was detected when receiving a patient in the last 3 months;
8. GIPER_6M – how many times hypertension was detected when receiving a patient in the last 6 months;
9. GIPER_3M – whether hypertension was detected when receiving a patient in the last 3 months;
10. STENOK_6M – how many times was stenocardia detected when receiving a patient in the last 6 months;
11. STENOK_3M – whether stenocardia was detected when receiving a patient in the last 3 months;
12. INF_MIOK_6M – how many times has a myocardial infarction been detected when receiving a patient in the last 6 months;
13. INF_MIOK_3M – whether a myocardial infarction was detected when receiving a patient in the last 3 months;
14. ICD_CLASS_COUNT_6M – the sum of the number of cases with different classes of diseases by ICD for the last 6 months;
15. IS_AGAIN_1M – whether the patient will arrive in the next month after DATE_OUT or not;
16. ICD_CLASS_BOOL_3M – was there any treatment of different classes of ICD diseases in the last 3 months;
17. ICD_CLASS_BOOL_1M – whether the handling of different classes of diseases according to ICD for the last month.

The ICD_CLASS_BOOL_NM field is responsible for re-visiting the patient, if 0 - there is no visit, if 1 - there is a visit.

These fields were obtained by simple aggregation with one-hot and bug-of-words encoding.

This study is aimed at making a forecast for the patient to determine whether they will go to the polyclinic with cardiovascular diseases in the next month or 3 months based on the analysis of personalized cards. In this regard, the output field for training the classifier is both ICD_ HEART_BOOL_1M and ICD_ HEART_BOOL_3M.

Let function $f(C_i)$ describe a specific classifier that gets a vector of characteristics of visits $C_i$ to patients $u_i$.

The function determines a certain value $Y \in \{0;1\}$, whether he will go to the polyclinic with cardiovascular diseases in the next 3 months.

Under certain conditions, the value $Y$ can be converted to the probability that the patient will go to the clinic. In this case, the condition of monotony is valid, meaning that the higher the value $Y$, the higher the probability of its arrival. It is necessary to find such parameters at which the classifier will give the best probabilistic estimates in terms of the selected metrics.

As a result, we obtain the forecasting problem, which we will solve using the logistic regression method, the ID3 decision tree construction algorithm, and the ensemble training method — random forest

## *2.2. Logistic Regression Method*

**Logistic regression** is a type of generalized linear model (GLM) that uses a logistic function to predict a binary characteristic based on any kind of independent input parameter.

The coefficients of the logistic regression algorithm should be estimated based on the training sample using the maximum likelihood estimation method, which is the most common learning algorithm used by various machine learning algorithms.

The main idea of the maximum likelihood method for logistic regression is that the algorithm looks for values for the coefficients of the logistic function that minimize error in the probabilities predicted by the model from the values in the data.

## *2.3. Algorithm for building the ID3 decision tree*

The ID3 algorithm builds a top-down decision tree. The ID3 algorithm implements a kind of "greedy" search in the space of all possible trees: it adds a subtree to the current tree and continues the search without making any returns. By this approach, the algorithm becomes very effective, however, it is highly dependent on the procedure for selecting the next property for testing.

We can assume that each property of an object contributes a certain amount of new information to the solution of the classification problem and reduces uncertainty.

In General, in information theory, entropy is calculated:

$$I(P) = -\sum_{i=1}^{n} p_i \log p_i \qquad (1)$$

The ID3 algorithm selects a specific property for the role of the root of the current subtree based on the amount of information received as a result of its verification: the root of the subtree selects the property that gives the most

information when checking (it reduces uncertainty most of all).

### *2.4. Random Forest*

**Random forest** is a set of decision trees. In the regression problem, their responses are averaged, and in the classification problem, a majority vote is taken. All trees are built independently according to the following scheme:
• A sub-sample of the training sample of the sample size is selected – a tree is built on it (each tree has its sub-sample).
• To build each split in the tree, view the max_features of random features (for each new split, we have our random features).
• Select the best feature and split it (according to pre-set criteria). The tree is usually built until the selection is exhausted (until only one class remains in the leaves), but in modern implementations, some parameters limit the height of the tree, the number of objects in the leaves, and the number of objects in the sub-sample at which splitting is performed.
It is clear that such a construction scheme corresponds to the main principle of ensembling (building a machine learning algorithm based on several, in this case, solving trees): the basic algorithms must be diverse (so each tree is built on its training sample and there is an element of randomness when choosing splits).

## 3. EXPERIMENTAL RESULTS

All experiments performed in this work were carried out on real data from the history of visits to patients of one clinic.
The data set contains information on patient visits, supplemented by the statistical characteristics defined above with a mark on whether he will come in one or three months with cardiovascular diseases.

### *3.1. Decision Tree Analysis*

It is worth noting that the advantage of the algorithm of building the ID3 tree is also the simplicity of presentation and the interpretation of the results. In this regard, to confirm that the built forecast can correspond to real history, we will analyze the built decision tree, highlighting the basic rules.

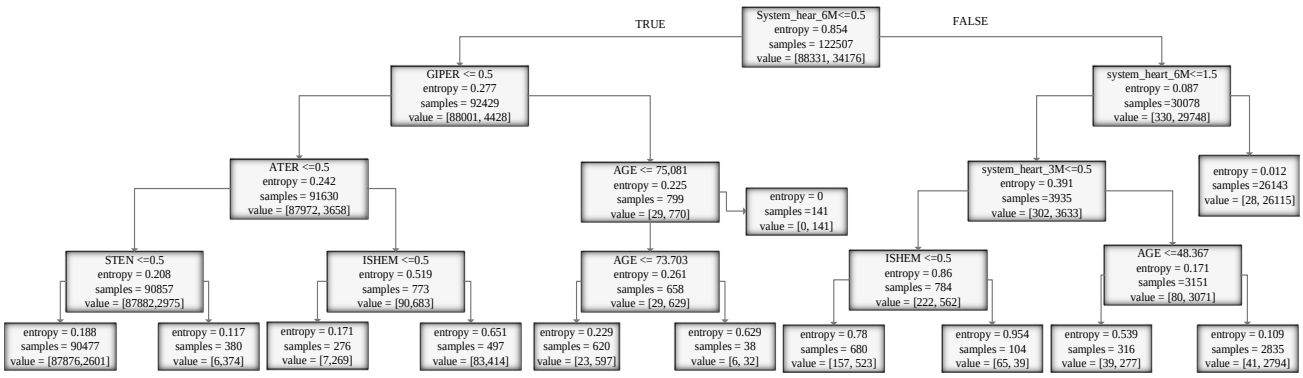


**Figure 1** Decision tree for ICD_HEART_BOOL_3M forecast on the training set.

The training set with the ICD_CLASS_BOOL_3M field of patients (whether patients will 3 to the polyclinic with cardiovascular diseases in the next three months) has only 122507 unique records. Their: 88331 - will not appeal to the clinic with GCC, 34176 - will appeal. The decision tree constructed by the ID3 algorithm is presented above and has a depth of 4. The accuracy of the built decision tree on the test set is 0.8546.
According to the Solutions Tree we can highlight the most important generalizing rules:

**IF** Patient did not arrive for the last 6 months with CVD **THEN**

**IF** Patient has no hypertension, atherosclerosis and angina **THEN** patient will not come next 3 months with CVD

**ELSE**

**IF** the patient has hypertension / atherosclerosis / angina pectoris **THEN** will come in the next 3 months with CVD.

**ELSE**

**IF** Patient came last 6 months more than 1 times with CVD **THEN**

**IF** the patient did not come the last 3 months with CVD **THEN**

**IF** the patient has ischemia **THEN** will come in the next 3 months with CVD.

**ELSE** will not come in the next 3 months with CVD.

**ELSE**

**IF** age > 48 **THEN** will come in the next 3 months with CVD.

**ELSE** patient will not come next 3 months with CVD.

**ELSE** the patient will come the next 3 months with CVD.

Because the built decision tree performs classification of the training set by output characteristic with sufficient accuracy, it is possible to speak about adequacy of the built model.

### *3.2. Comparative analysis of algorithm efficiency*

This study evaluated the effectiveness of the methods considered for prediction based on the ROC curve analysis and the patient 's ACC metric.
The forecast was built separately for applications in the next month (Figure 2) and 3 months (Figure 3) to the polyclinic with cardiovascular diseases on the basis of analysis of personalized maps.

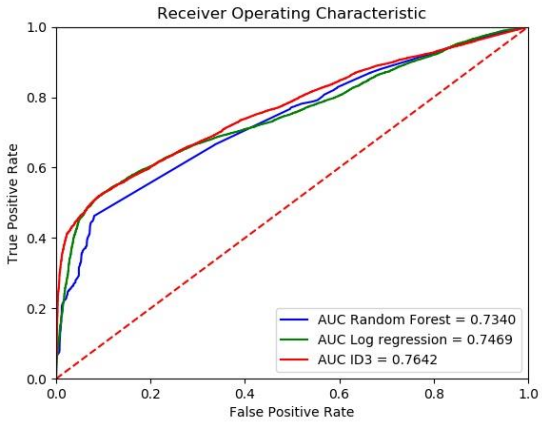


**Figure 2** ROC curve for ICD_HEART_BOOL_1M

It is worth noting that, according to the ACC metric, the ID3 algorithm showed the best prediction results (ACC ID3 = 0.7642) on the test set for analysis of cases in the coming month. However, to predict for a longer time (3 months), a higher accuracy was shown by the logistics regression construction method (ACC Log. Regression = 0.8464).

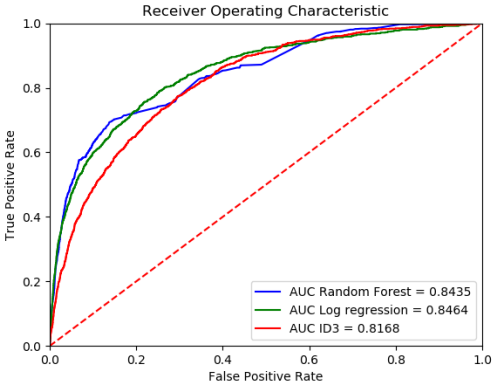


**Figure 3** ROC curve for ICD_HEART_BOOL_3M

Besides, we note that the accuracy of the forecast of visits for the next month (0.75) is lower than that for the next 3 months (0.84). This is because the exacerbation of cardiovascular diseases is quite rare, with regular examinations of patients, and when the period under review increases for the forecast, the accuracy should increase.

## 4. CONCLUSION

Within the framework of this study, a forecast was built, the patient's appeals in the next month or 3 months to the polyclinic with cardiovascular diseases based on analysis of personalized maps. The forecast was determined by logistic regression, the algorithm for constructing ID3 decision trees and the ensemble training method — random forests.
The constructed models showed a good result since they had high generalizing ability and accuracy. As part of an experimental study, the effectiveness of the application of the considered methods for predicting real clinic data based on the analysis of the ROC curve and the AUC metric was evaluated.
Each of the methods considered has its advantages over similar methods of solving forecasting problems. However, it is worth noting that for a short time of prediction (1 month), the algorithm of ID3 construction of decisive trees showed higher results, and when the period under consideration was increased to 3 months, the best results were shown by the method of logistics regression.

## ACKNOWLEDGMENT

The study was carried out with the financial support of the Russian Federal Property Fund in the framework of scientific projects No. 18-07-01446, No. 20-07-01065, as well as a grant from the President of the Russian Federation for state support of leading scientific schools of the Russian Federation (NSh-2502.2020.9).

## REFERENCES

[1] S. Windmann, A. Maier, Big Data Analysis of Manufacturing Processes. in: O. Niggemann, C. Frey (Eds.), Journal of Physics: Conference Series. 2015. DOI: https://doi.org/10.1088/1742-6596/659/1/012055.

[2] G. Kim, J. Chung Big Data Applications in the Government Sector: A Comparative Analysis among Leading Countries. Communications of the ACM, 57. (2014) 78-85. DOI: https://doi.org/10.1145/2500873.

[3] M. Alshura, A. Zabadi, M. Abughazaleh,. Big Data in Marketing Arena. Big Opportunity, Big Challenge, and Research Trends: An Integrated View. MANAGEMENT AND ECONOMICS REVIEW, vol. 3. (2018) DOI: https://doi.org/10.24818/mer/2018.06-06.

[4] A. Beam, I. Kohane. Big Data and Machine Learning in Health Care. JAMA. vol. 319. (2018) DOI: https://doi.org/10.1001/jama.2017.18391.

[5] T. Geldof, N. Damme. Patient-Level Effectiveness Prediction Modeling for Glioblastoma Using Classification Trees. in: I. Huys, W. Dyck (Eds.), Frontiers in Pharmacology, vol. 10. (2020) DOI: https://doi.org/10.3389/fphar.2019.01665.

[6] A. Blasiak, J. Khong, T. Kee. CURATE.AI: Optimizing Personalized Medicine with Artificial Intelligence. SLAS TECHNOLOGY: Translating Life Sciences Innovation 247263031989031. (2019). DOI: https://doi.org/10.1177/2472630319890316.

[7] R. Lin, Z. Ye. Chronic Diseases and Health Monitoring Big Data: A Survey. in: H. Wang (Eds.), IEEE Reviews in Biomedical Engineering. (2018). DOI:https://doi.org/10.1109/RBME.2018.2829704.

[8] K. Leung, A. Stevenson. (2018). Application of Big Data in Decision Making for Emergency Healthcare Management. International Journal of Research and Engineering, vol. 5. pp. 311-314. DOI: https://doi.org/10.21276/ijre.2018.5.2.2.

[9] E. Zimina, M. Novopashin, A. Shmid.. technologies in the problems of mathematical analysis of cardiological information, 2018, pp. 112-118. DOI: https://doi.org/10.18287/1613-0073-2018-2212-112-118.

[10] E. Zimina. Cluster analysis of cardiac data. Statistics and Economics, 2018, vol. 15, pp. 30-37. DOI:https://doi.org/10.21686/2500-3925-2018-2-30-37.

[11] I. Burykin, G. Aleyeva, R. Hafizyanova. Prospective Value of Big Data Analysis Method for Assessment of Pharmacotherapy Quality and Efficacy in Patients with Arterial Hypertension. Modern technologies in medicine, 2017, vol. 9, pp. 194. DOI: https://doi.org/10.17691/stm2017.9.4.24.

[12] A. Baranov, L. Baranova, I. Smirnov, D. Deviatkin, A. Shelmanov, E. Vishneva, E. Antonova. Technologies for Complex Intelligent Clinical Data Analysis. RAMS, 2016, vol. 71, pp. 160-171.

[13] I. V. Stepanyan. Scientific and methodological bases and bioinformatic technologies of occupational risk management in occupational medicine, dissertation of doctor of biological Sciences. Moscow, 2012.

[14] S. Weng, J. Reps, J. Kai. Can Machine-learning improve cardiovascular risk prediction using routine clinical data. PLoS ONE, 2017, vol. 12. DOI: https://doi.org/10.1371/journal.pone.0174944.

[15] S. Krishnan. Application of Analytics to Big Data in Healthcare, 2016, pp. 156-157. DOI: https://doi.org/10.1109/SBEC.2016.88.

[16] G. Tocci, A. Ferrucci, P. Guida. Use of Electronic Support for Implementing Global Cardiovascular Risk Management. High HEART Press Cardiovasc Prev, 2010, vol. 17, pp. 37–47. DOI: https://doi.org/10.2165/11311750-000000000-00000.