Application of statistical data analysis algorithms and determination of the most significant diagnostic factors

Alena Rozhkova[1], *Vladislav* Kukartsev[2,3], *Mikhail* Kvesko[3], *Anna* Glinskaya[3*], and *Oksana* Kukartseva[3]

[1]Krasnoyarsk State Agrarian University, 90, Mira Avenue, 660049, Krasnoyarsk, Russia
[2]Bauman Moscow State Technical University, Bldg. 1, 5, 2nd Baumanskaya Str., 105005, Moscow, Russia
[3]Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy Av., 660037, Krasnoyarsk, Russia

Abstract. The article examines the application of statistical data analysis algorithms in diagnostics and the identification of significant factors influencing observed phenomena. The use of statistical methods, such as multiple regression, logistic regression, and neural networks, is discussed. The study focuses on determining the most impactful factors, establishing relationships between variables, and evaluating the effectiveness of technologies and methods. The dataset, provided by Universidad Cardenal Herrera, CEU, Madrid, Spain, includes factors related to patients' demographics, health conditions, and lifestyle. The analysis involves deductive analysis, correlation analysis, and Kohonen maps to identify relevant factors. Decision tree analysis is conducted using different combinations of factors, including average glucose levels and body mass index. The results show varying error rates for different methods and factor combinations. Overall, statistical data analysis demonstrates its potential for faster and more accurate identification of significant diagnostic factors.

1 Introduction

Statistical data analysis algorithms can be highly useful in diagnostics and identifying the most significant factors influencing the observed phenomena [1-3].

Using statistical methods can help determine which specific factors have the greatest impact [4, 5]. Multiple regression methods, logistic regression, or similar algorithms can be employed for this purpose [6]. Analyzing the results of such algorithms helps gain a better understanding of the influence of each variable [7-9]. Additionally, statistical data analysis can help determine the presence of relationships between different factors and describe their statistical significance [10, 11].

Statistical data analysis can also be used to assess the effectiveness of technologies and methods [12]. Comparative analysis allows for the comparison of results obtained using

* Corresponding author: anna_glinskaja@rambler.ru

different methods and helps in selecting the optimal method based on the observed outcomes [13, 14].

Statistical data analysis can handle large volumes of data, which is particularly beneficial when processing vast amounts of information gathered from various sources [15-17].

One of the methods of statistical data analysis is shallow networks [18]. It is worth noting that shallow networks are highly effective diagnostic tools when the amount of training data is relatively small, and the dynamics of the system under investigation are well-known [19, 20]. However, they require preprocessing of diagnostic signals to extract fault features and feed them as input to the network [21]. Therefore, the vast majority of applications of neural networks in diagnostics are related to the use of multilayer perceptron (MLP) [22, 23].

Kohonen Maps is a self-learning neural network that is designed to classify, organize, and visually represent large amounts of data [24]. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the characteristics of the data [25].

Correlation is a statistical relationship between random variables in which a change in one of the random variables results in a change in the expectation of the other. It is important that the features selected in the recognition model are highly correlated with the target variable in order to increase the accuracy of prediction, and it will also help to simplify the training of the recognition model and help to increase the efficiency of prediction [26, 27].

Therefore, the application of statistical data analysis algorithms holds tremendous potential [28, 29]. They enable faster and more accurate identification of the most significant factors in diagnostics [30].

2 Materials and methods

The data set was created by Universidad Cardenal Herrera, CEU, Madrid, Spain.

Patients are aged between 17 and 88 years.

Factors for which data were collected:

- Sex.
- Age.
- Hypertension.
- Heart disease.
- Ever married.
- Kind of activity.
- Place of residence.
- Average glucose level
- Body mass index.
- Status of smoking.

Data of 40,911 rows was loaded into deductor for analysis. The data were then correlated, resulting in the following significant factors from 10 factors:

- Age. (Fig. 1a)
- Hypertension. (Fig. 1b)
- Heart disease. (Fig. 1c)
- Ever married (Fig. 1d)
- Kind of activity. (Fig. 1e)
- Place of residence. (Fig. 1f)
- Average glucose level. (Fig. 1g)

- Body mass index. (Fig. 1h)
- Status of smoking. (Fig. 1i)

Figure 1 describe values of factors. Figure 2 presents the relevant factors derived from the correlations.

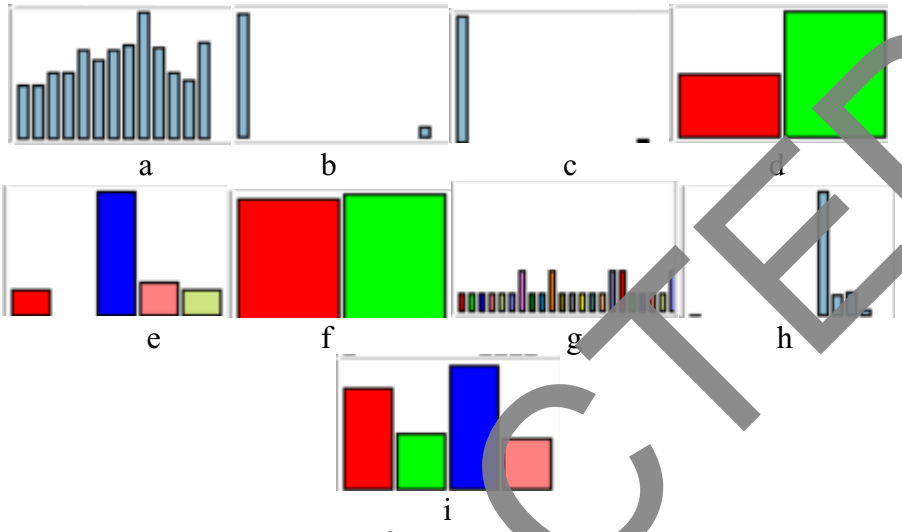


Fig. 1. Value of indicators.

Nº	Input field Field	Correlation with output fields Total indicator
1	Indicator 2	0,059
2	Indicator 3	0,257
3	Indicator 4	0,224
4	Indicator 5	0,182
5	Indicator 6	0,026
6	Indicator 7	0,012
7	Indicator 8	0,265
8	Indicator 9	0,018
9	Indicator 10	0,068

Fig. 2. Relevant factors resulting from correlation

Table 1. Describes the correlation factor:

Table 1. Description of data factors.

Value	Sex	Hypertension	Heart disease	Ever married	Kind of activity	Place of residence	Status of smoking	Stroke
0	M	No	No	No	Unemployed	City	Did not smoke/quit > 5 years ago	No
1	F	There is	There is	Yes	On maternity leave	Rural area	Smoke	There is
2					Budget			
3					Businessman			
4					Private			

Kohonen maps were created (Fig. 3). The results of errors are presented in table 2.

Table 2. Decision tree research results.

Research error, 0%			
	Classified		
Actually	0	1	Total
0	20450		20450
1		20460	20460
Total	20450	20460	40910
Research error, 1.6%			
0	19781	669	20450
1		20460	20460
Total	19781	21129	40910
Research error, 29.27%			
0	15956	4484	20450
1	7482	12978	20460
Total	23438	17472	40910
Research error, 29.5%			
1	14240	6210	20450
2	5856	1604	20460
Total	20096	20814	40910
Research error, 31%			
0	13771	6679	20450
1	6310	14150	20460
Total	20081	20829	40910
Research error, 40.9%			
0	7150	13300	20450
1	3440	17020	20460
Total	10590	30320	40919

The study error of 29.27% is shown in table 2 (Experiment 1).

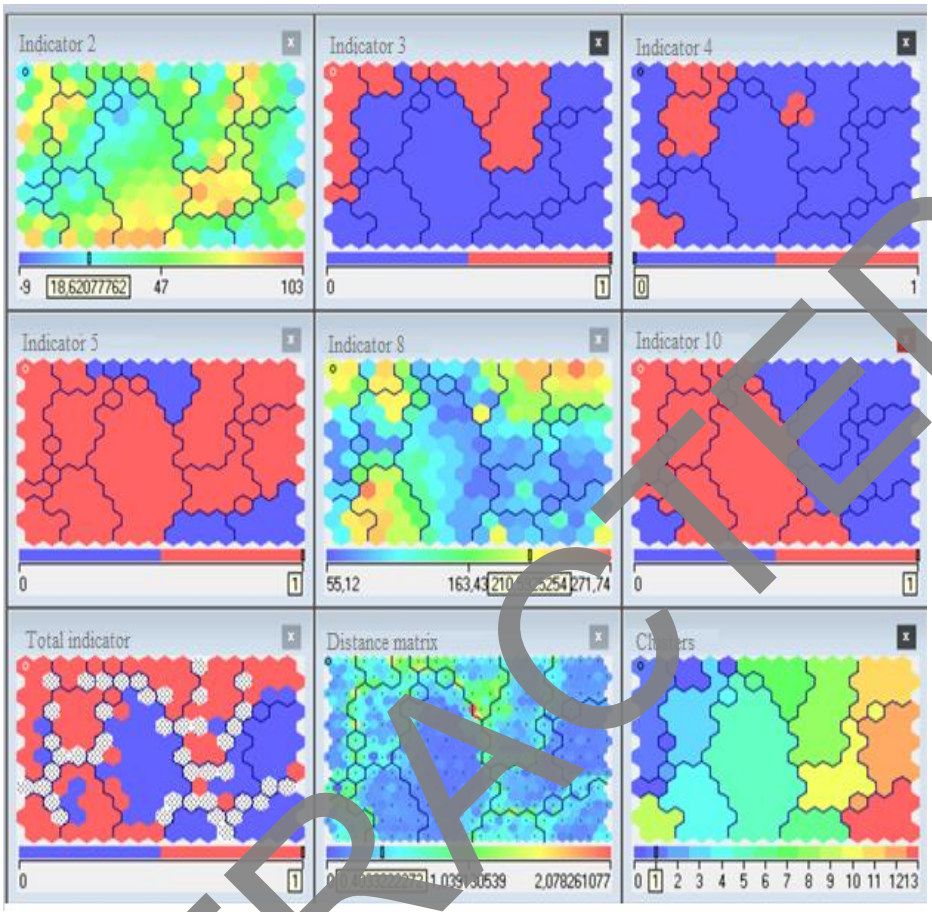


Fig. 3. Kohonen maps.

3 Results and discussions

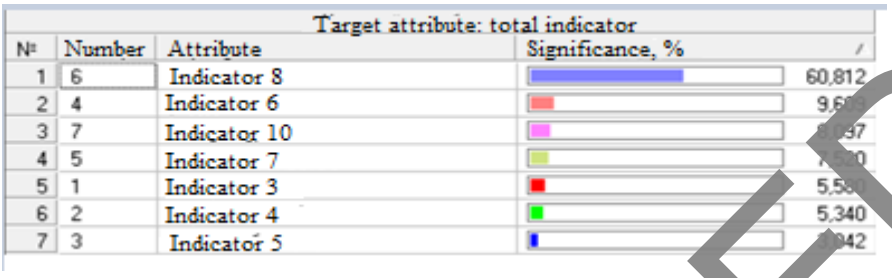
The research continued using the decision tree. When using all factors in the method, the most significant was indicator 8. The significance of all factors is shown in Figure 4. The result of the study error of 0% is shown in Table 2 (Experiment 2).

Target attribute: total indicator

№	Number	Attribut	Significance, %
1	7	Indicator 8	76,461
2	8	Indicator 9	15,275
3	2	Indicator 3	3,087
4	5	Indicator 6	1,577
5	6	Indicator 7	1,526
6	4	Indicator 5	1,260
7	9	Indicator 10	0,482
8	3	Indicator 4	0,333
9	1	Indicator 2	0,000

Fig. 4. Importance of factors.

Then, all factors were investigated without taking into account indicator 8. In this case, the most significant factor was indicator 9. The remaining factors are shown in Figure 5. The study error of 1.6% is shown in Table 2 *(Experiment 3).*



Target attribute: total indicator

№	Number	Attribute	Significance, %
1	6	Indicator 8	60,812
2	4	Indicator 6	9,6
3	7	Indicator 10	8,097
4	5	Indicator 7	7,530
5	1	Indicator 3	5,580
6	2	Indicator 4	5,340
7	3	Indicator 5	3,042

Fig. 5. Importance of factors.

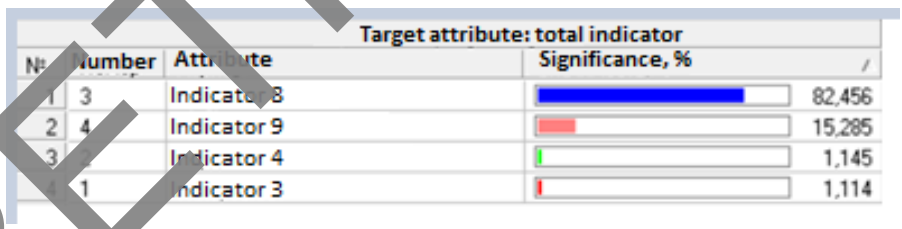
The data were then examined without taking into account to indicator 9 (Figure 6). The result of the 31% error of the third study is shown in Table 2 (Experiment 4).

Target attribute: total indicator

№	Number	Attribute	Significance, %
1	1	Indicator 3	31,006
2	2	Indicator 4	21,028
3	3	Indicator 5	20,453
4	4	Indicator 6	11,034
5	6	Indicator 10	8,326
6	5	Indicator 7	8,155

Fig. 6. Importance of factors.

After that, the factors data were examined in a laboratory and by medical examination and medical report. The significance of the factors is shown in Figure 7. The result of the study error of 0% is shown in Table 2. (Experiment 5).



Target attribute: total indicator

№	Number	Attribute	Significance, %
1	3	Indicator 8	82,456
2	4	Indicator 9	15,285
3	2	Indicator 4	1,145
4	1	Indicator 3	1,114

Fig. 7. Importance of factors.

The data set was then examined on the factors collected in a laboratory and by medical examination and doctor's report, including the patient's indicator 10 (figure 8). The study error of 0% is shown in Table 2 (Experiment 6).

Target attribute: total indicator			
№	Number	Attribute	Significance, %
1	3	Indificator 8	78,310
2	4	Indificator 9	17,551
3	1	Indificator 3	2,991
4	5	Indificator 10	0,605
5	2	Indificator 4	0,543

Fig. 8. Importance of factors.

The dataset was then investigated by factors collected by interviewing patients, not including the indicator 10 (Figure 9). The study error of 40.9% is shown in Table 2 (Experiment 7).

Target attribute: total indicator			
№	Number	Attribute	Significance, %
1	1	Indificator 5	45,763
2	2	Indificator 6	35,712
3	3	Indificator 7	18,524

Fig. 9. Importance of factors.

The data were then processed by factors collected through patient interviews, including indicator 10 (Figure 10).

Target attribute: total indicator			
№	Number	Attribute	Significance, %
1	1	Indicator 5	40,169
2	2	Indicator 6	31,636
3	3	Indicator 7	18,012
4	4	Indicator 10	10,184

Fig. 10. Importance of factors.

The data set was then investigated through the creation and training of a neural network (figure 11).

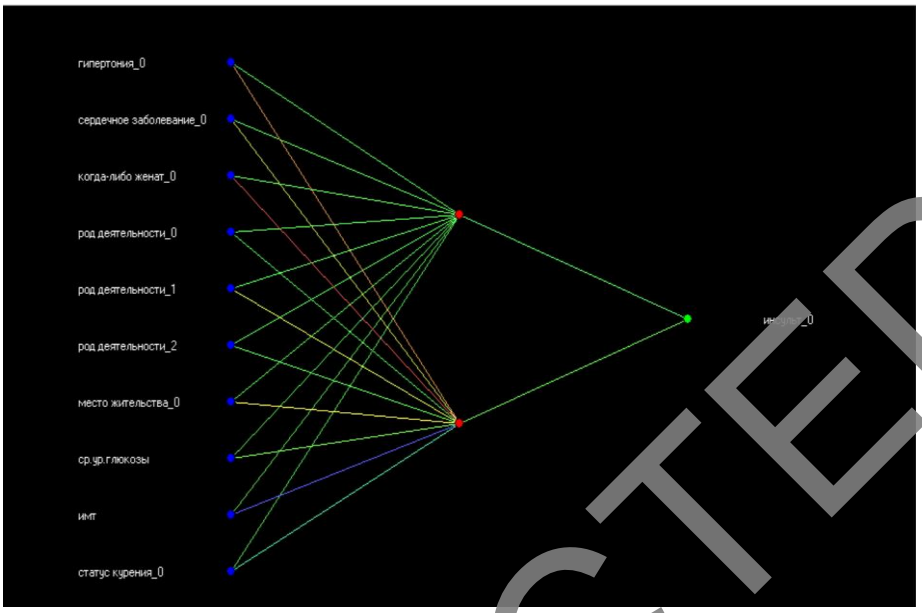


Fig. 11. Neural network.

The study error of 29.5% is shown in Table 2

The error value of all methods used is presented in comparative table 3.

It is evident in Figure 12 and Figure 13 in more detail, the number of errors and the error percentage:

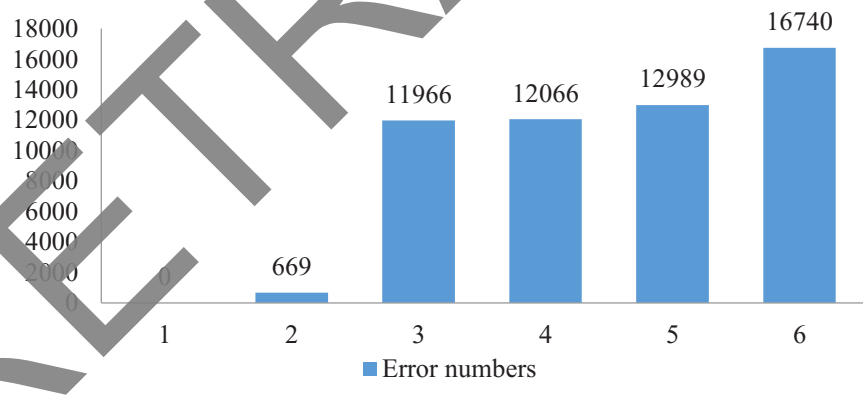


Fig. 12. Number of errors in the experiments.

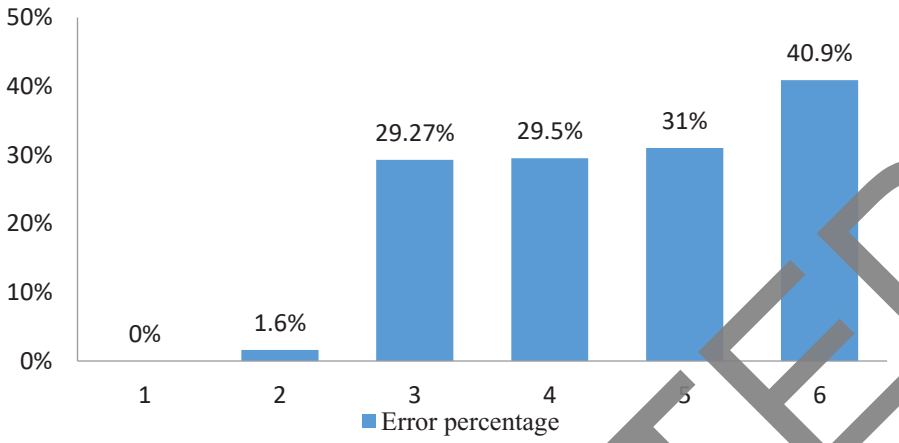


Fig. 13. Percentage of errors in the experiments.

Table 3. Comparison of error values.

Method	Error value
Kohonen maps	29.27%
Decision tree for all factors	0%
Decision tree without average glucose	1.6%
Decision tree without average glucose level and body mass index	31%
Decision tree for laboratory collected factors	0%
Decision tree for laboratory collected factors with smoking status	0%
Decision tree from a survey without smoking status	40.9%
Decision tree of the survey based on smoking status	39%
Neural network	29.5%

4 Conclusion

The study identified the main risk factors, built maps of Kohonen taking into account errors in the analysis [31]. Several tests were also conducted on various criteria, indicating the percentage of error results shown in table 2. For ease of reference, a comparison table 3 of error probabilities for Kohonen maps and decision trees was selected.

References

1. N. V. Martyushev et al., Energies **16(2)**, 729 (2023)
2. A. Shutaleva et al., Sustainability **15(4)**, 3011 (2023)
3. V. A. Rezanov et al., Metals **12(12)**, 2135 (2022)
4. N. V. Martyushev et al., Materials **16(9)**, 3490 (2023)
5. V. A. Kukartsev et al., Metals **13(2)**, 337 (2023)
6. O. A. Kolenchukov et al., SOCAR Proceedings **1**, 29-34
7. V. V. Bukhtoyarov et al., SOCAR Proceedings **1**, 12-20 (2022)
8. O. A. Kolenchukov et al., Energies **15(22)**, 8346 (2022)
9. K. A. Bashmur, Sustainability **14(20)**, 13083 (2022)

10. K. Degtyareva et al., *Use of computer simulation tools to simulate processes at the foundry*, in Proceedings of the 23rd International Symposium INFOTEH-JAHORINA (INFOTEH), pp. 1-5, IEEE (2024)
11. K. Degtyareva et al., *Automated system for accounting of customers and orders*, in Proceedings of the 23rd International Symposium INFOTEH-JAHORINA (INFOTEH), pp. 1-4, IEEE (2024)
12. V. I. Golik et al., MIAB. Mining Inf. Anal. Bull. **11(1)**, 175-189 (2023)
13. T. A. Panfilova et al., MIAB. Mining Inf. Anal. Bull. **11(1)**, 239-251 (2023)
14. E. Suprun et al., BIO Web of Conferences **84**, 01008 (2024)
15. V. Orlov et al., E3S Web of Conferences **460**, 07002 (2023)
16. K. Kravtsov et al., E3S Web of Conferences **458**, 09022 (2023)
17. V. S. Tynchenko et al., E3S Web of Conferences **458**, 01011 (2023)
18. E. Semenova et al., *Using UML to describe the development of software products using an object approach*, in Proceedings of the 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), pp. 1-4, IEEE (2022)
19. V. S. Tynchenko et al., AIP Conference Proceedings **2700(1)**
20. N. Chernykh et al., *Comparative analysis of existing measures to reduce road accidents in Western Europe*, in Proceedings of the 22nd International Symposium INFOTEH-JAHORINA (INFOTEH), pp. 1-6, IEEE (2023)
21. E. Volneikina et al., *Simulation-dynamic modeling of supply chains based on big data*, in Proceedings of the 22nd International Symposium INFOTEH-JAHORINA (INFOTEH), pp. 1-6, IEEE (2023)
22. O. A. Filina et al., Energies **17(1)**, 17 (2023)
23. I. P. Boychuk, A. V. Grinek, N. V. Martyushev, et al., Energies **16(24)**, 8101 (2023)
24. V. I. Golik et al., Materials **16(21)**, 7004 (2023)
25. B. V. Malozyomov et al., Energies **16(13)**, 5046 (2023)
26. I. P. Malashin et al., Polymers **16(1)**, 115 (2023)
27. B. V. Malozyomov et al., Energies **16(13)**, 4907 (2023)
28. V. S. Tynchenko et al., Journal of Physics: Conference Series **2373(6)**, 062015
29. V. A. Nelyub et al., Correlation Analysis and Predictive Factors for Building a Mathematical Model. In Proceedings of the Computational Methods in Systems and Software, pp. 14-25, Cham, Springer International Publishing (2023)
30. K. V. Degtyareva et al., E3S Web of Conferences **458**, 02002 (2023)
31. A. Gantimurov et al., E3S Web of Conferences **431**, 03005 (2023)