Improving statistical methods of data processing in medical universities using machine learning

Ainur Orynbayeva†, Nurzhan Shyndaliyev† & Ainura Aripbayeva‡

L.N. Gumilyev Eurasian National University, Astana, Kazakhstan†
Astana Medical University, Astana, Kazakhstan‡

ABSTRACT: Recently, machine learning has been the most advanced big data processing model in various fields, including medicine, where it has been applied to medical data analysis, medical diagnostics and medical technology in general. This study was focused on improving the *Basics of Biostatistics* subject at Astana Medical University in Kazakhstan, and the content thereof was considered in view of practical approaches of using statistical methods with machine learning in the R programming environment. Successful improvement of this subject is necessary for planning, conducting research and data analysis to assess various situations and trends in health care, as well as for conducting scientific research in the field of medical biology, clinical and health care. The methods described in the article are proposed to be used in educational programmes in medical universities, as well as in interdisciplinary courses in other areas.

INTRODUCTION

Nowadays, machine learning techniques are used in different domains, from spam detection and fraud to image and music recognition [1]. The use of medical diagnostics is particularly important and promising. In particular, it is applied to forecast the condition of the patient, the differential diagnosis of the diseases, to test the effectiveness of the medications, etc.

The application of machine learning in medicine is related to the accumulation in medicine of vast amounts of heterogeneous data, the improvement of new medical technologies, including computer science. This requires the use of machine learning techniques to process and analyse data, as well as new knowledge.

In this article, the authors outline a review of training courses on machine learning methods at medical universities in Kazakhstan, and based on their own experience, they present the educational and methodological content for the analysis of medical data using statistical methods. At the same time, Astana Medical University effectively uses machine learning methods within the subject *Basics of Biostatistics* (60 hours), included in the educational programmes:

- 6B10107 - General Medicine;
- 6B10108 - Dentistry;
- 6B10118 - Paediatrics;
- 6B10119 - Public Health.

For practical implementation in the educational process, sharing the methods of application of statistical methods in the processing of medical data the use of the programming language R offers educational and methodological support in the subject delivery.

The purpose of the article is to describe the practical approaches of using statistical methods with machine learning in the R programming environment, and demonstrate the improvement of the *Basics of Biostatistics* subject in Astana Medical University with this specific tactic.

LITERATURE REVIEW

The analysis of data through machine learning in medical universities, including the use of statistical methods of medical data processing, is an innovative direction in the world practice and in Kazakhstan. However, there are already several universities that cover these advanced methods in the content of their educational programmes, some of them in Kazakhstan and several abroad.

In the framework of the master-class *Big data in the field of health* at Karaganda Medical University in Kazakhstan, the director of research XinShi analysed the importance of using clinical and non-clinical, real-time and high-volume data for early diagnosis and prevention of various diseases [2].

The possibilities of machine learning in medicine are considered by the researchers Cherikbayeva and Turkistan of *Al-Farabi* Kazakh National University Almaty, Kazakhstan, in their recent article, where they include a brief history of machine learning, and provide some background knowledge about the methods and current state of this technology in healthcare [3].

The *Information and Communication Technologies* (ICT) subject is an addition to the curriculum in the Department of Medical Biophysics and Information Technology at South Kazakhstan Medical University, Shymkent, Kazakhstan. The subject content is focused on the formation of the following skills: the role of ICT in the main sectors of society development, introduction to computer systems, software, human interaction with computers, database systems, data analysis, machine learning, big data, data management, Internet technologies, basics of cloud, mobile technologies [4].

In the scope of this subject, which is carried out in almost all Kazakh medical universities, only a general idea of the basics of machine learning in the analysis and processing of data is given. It should be noted that some researchers from Kazakh medical universities considered studies in this direction at master classes as separate scientific work. It follows that the introduction of machine-based methods into the medical field in the country requires a new direction in research.

Examining the overall practice of analysing statistical methods of data processing at medical universities using machine learning methods, the following achievements should be mentioned: The Australian Institute for Machine Learning in Adelaide, Australia, pursues globally competitive research and development in the fields of machine learning, artificial intelligence, computer-based viewing and deep learning. Currently, they use medical machine training in the following areas: cardiology, cancer, obstetrics and gynaecology, orthopaedics (arthroscopy, hip endoprosthesis), neurology (transient ischemic attack, stroke, vascular dementia), public health [5].

The researchers Kolachalama and Garg explain the importance of studying the methods of machine learning by medical professionals [6]. The theoretical case of medicine based on machine learning is discussed by Handelman at al who focus on general algorithms of machine learning used in medicine, as well as the future and importance of machine learning in medicine [7].

SUBJECT DESIGN

Examining the statistical methods of machine learning in the field of informatics in medical universities, it should be underlined that proper approaches can:

- contribute to the correct, rapid and unmistakable decision-making of students in the conduct of complex statistical calculations used in the field of medicine;
- foster the ability of students in the process of any statistical calculations to determine and make adjustments to the cause of the fault or malfunction;
- help to identify the cause of the increase in total costs and reduce waste;
- contribute to the efficiency of different processes when considering certain design works [8].

On the basis of the Bioinformatics and Statistics Department at Astana Medical University, Kazakhstan, educational and methodological support has been developed for the analysis of large volumes of medical data with the help of machine training in carrying out statistical calculations in the *Basics of Biostatistics* subject included in the educational programmes 6B10107 - General Medicine, 6B10108 - Dentistry, 6B10118 - Paediatrics, 6B10119 - Public Health. Also, research aimed at the formation of professional competencies currently required will continue in this institution [9].

The following section outlines the content of the pedagogical and methodological support included in the updated content of training.

RESEARCH METHODOLOGY AND RESULTS

On the basis of machine learning five basic methods of statistical analysis are considered:

- Mean - average mean is one of the most popular methods of statistical analysis. The average mean determines the overall data trend. The average mean is calculated by adding the numbers in the data set and then dividing the sum by the number of data points. Despite the simplicity of the calculation and its advantages, it is not recommended to use the average as the only statistical indicator, as this may lead to incorrect decision-making.
- Standard deviation - standard deviation is another widely used statistical tool or method. It analyses the deviations of different data points from the average of the whole data set. It defines how the data in the data set is distributed by average.

- Regression - regression is a statistical tool that helps define the causal relationship between variables. It defines the relationship between the dependent variable and the independent variable. It is usually used to predict future trends and events.
- Hypothesis test - hypothesis testing can be used to verify the validity or veracity of a conclusion or evidence compared to data sets. However, the prediction made at the beginning of the study may be correct or false depending on the results of the analysis.
- Determining the size of the sample or sampling data - this method is used when the population is very large [10].

In this research, the R programming environment was used to study statistical methods based on machine learning in medical data processing. Here, the analysis of data in the medical storage system in the programming environment R is the basis. The R language is the universal environment for statistics and programming. This language includes several convenient and effective machine learning packages [11][12].

The Law on Compulsory Social Health Insurance in Kazakhstan was adopted in 2015. Then, the Social Health Insurance Fund (SHIF) was established, contributions to which began to be received in July 2017. The actual SHIF system was launched in 2018. This means that, as of 1 January each year, every citizen must pay 1 percent of his or her salary to the Social Health Insurance Fund. The collected data can be analysed using statistical methods. In the first stage, medical costs were analysed using linear regression. Then, the data were imported from the earlier created Excel table and edited in the R environment.

```
install.packages("readxl")
install.packages("dplyr")
library("readxl")
library(dplyr)
```

Libraries were installed and launched.
To import data from a spreadsheet, one creates a variable:

```
Dinfo=read_excel("C:\\12.xlsx")
```

As a result, one can load data equivalent to the Dinfo variable (Table 1).

Table 1: Health insurance database.

No	Age	Sex	Contribution	Children	Region	Policlinic
1	19	Female	27900	0	Astana	10
2	25	Male	33770	1	Oskemen	1
3	36	Male	14300	3	Kyzylord	2
4	54	Female	12000	2	Shymken	3
5	56	Female	11000	3	Almaty	12
6	45	Female	9000	2	Almaty	10
7	70	Male	8700	3	Almaty	1
8	81	Female	15300	4	Astana	6
9	21	Male	14500	1	Astana	7
10	25	Male	25100	2	Astana	7
11	31	Male	26000	2	Karagan	2
12	34	Female	14000	2	Almaty	2
13	35	Female	21200	2	Atyrau	1
14	36	Male	36000	2	Aktau	1
15	19	Male	12400	0	Semei	1
16	24	Male	17800	1	Taraz	2
17	23	Female	16500	1	Taraz	2
18	25	Female	14200	1	Astana	10
19	36	Female	13200	2	Almaty	10
20	38	Male	14500	3	Almaty	1

With the help of algorithms and tools of machine learning, the information in Table 1 was obtained in order to effectively convey statistical methods of data processing to students. These data are presented in the form of medical tax payment indicators that is contributions of all people (men, women) aged 19 to 65 (before retirement age) in different regions of Kazakhstan. On the basis of these data, further work was carried out with statistical methods using the R programming language. A correlation matrix was created, and statistical analyses were conducted on the relationship between medical care payments and the contributors' age, as well as the effect of medical payments depending on or independent of the person's age

The following steps were undertaken in the R environment. When calculating the amount of medical expenses, statistical values were deducted (Figure 1):

summary(Dinfo$Contribution)

```
> summary(Dinfo$Contribution)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8700   13000   14500   17869   22175   36000
```

Figure 1: Medical expenses amount.

The values here are the minimum, 1st quarter average, total mean, 3rd quarter average, maximum values. Each indicator in the table allows to automatically extract indicators, such as minimum, maximum and average values from thousands of data. These statistical analyses in the R programming environment are more efficient than in the Excel spreadsheet. This, in turn, saves time and prevents any errors. These values represent the validity of health insurance values and are visualised as a diagram (Figure 2):

hist(Dinfo$Contribution)

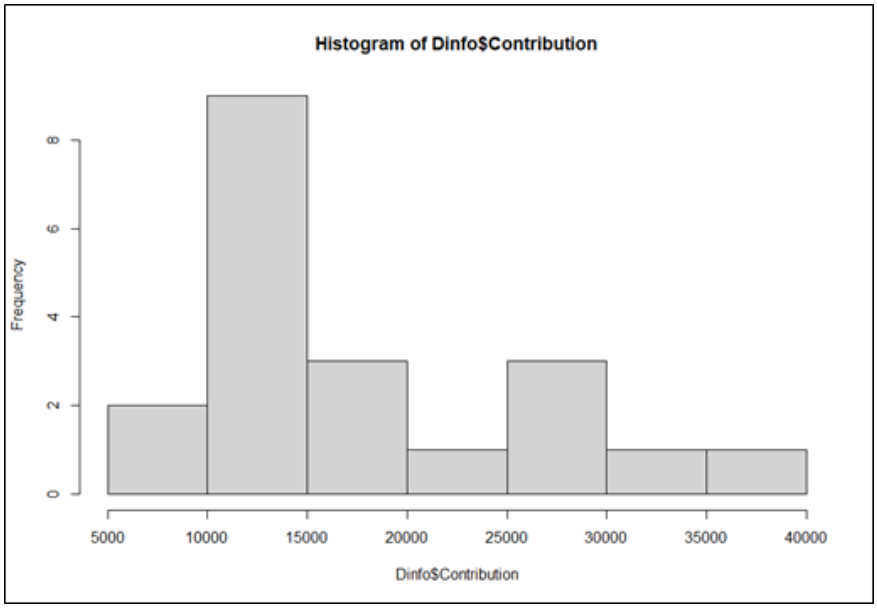


Figure 2: Diagram of medical expenses amount.

In order to determine the purpose of the data, data in the form of a table can be used (Figure 3). In order to consider the algorithms of statistical analysis, the data of medical insurance payments for each region were extracted. Based on these data, one can determine which regions most often paid for medical insurance.

table(Dinfo$Region)

```
> table(Dinfo$Region)

    Aktau    Almaty    Astana    Atyrau Karagandy Kyzylorda   Oskemen     Semei
        1         6         5         1         1         1         1         1
 Shymkent     Taraz
        1         2
```

Figure 3: Data assignment.

In the next stage, a correlation matrix was used to determine the relationship between variables.

cor(Dinfo[c(“Age”, “Contribution”)])

Here, the main values were the age and payment of the health insurance client. It can be observed that the perfect correlation between the age of the client and the payment is achieved (Figure 4):

```
> cor(Dinfo[c("Age", "Contribution")])
                    Age Contribution
Age          1.0000000   -0.3990065
Contribution -0.3990065   1.0000000
```

Figure 4: The result of the correlation matrix.

As a result of the correlation matrix, it can be seen that the implementation of medical insurance payments is not related to the age of a person. A diagram can be used to look at the ideal correlation coefficient. To determine the relationships between the data, the matrix diagram was used with the pairs function (Figure 5):

pairs(Dinfo[c("Age", "Contribution")])

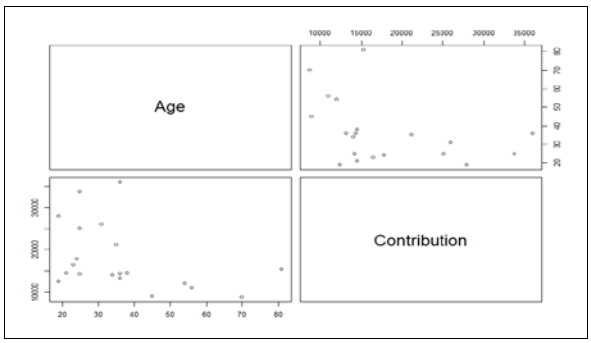


Figure 5: Scattering matrix diagram.

Further, the scattering function of the pair matrix diagram was used to determine the data relationship (Figure 6):

pairs.panels(Dinfo[c("Age", "Contribution")])

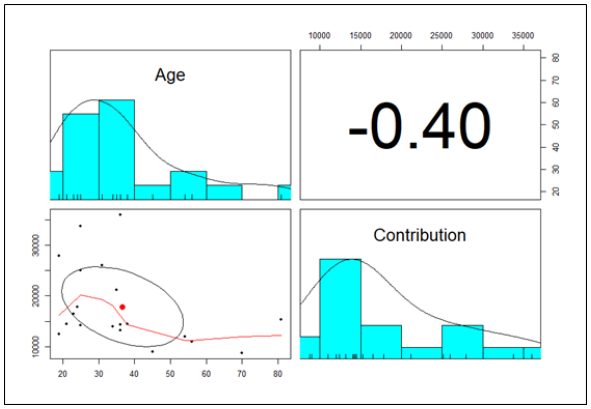


Figure 6: The result of the pairs function.

The graph defining the relationship between the data column and the path of each variable has a point. The authors needed to construct its model to add more information to the bitmap. In this case, the lm function was used based on the linear regression model. This was necessary for data learning. Therefore, one can enter the name of the required columns using the lm function.

```
ins_model<-lm(Contribution ~ Age+Sex+Region,data=Dinfo)
ins_model<-lm(Contribution~.,data=Dinfo)
ins_model
```

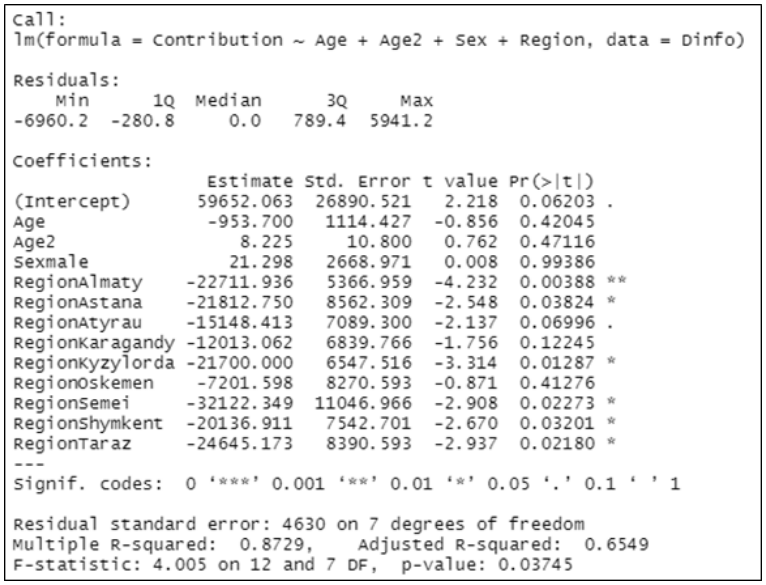
```
Call:
lm(formula = Contribution ~ ., data = Dinfo)

Coefficients:
    (Intercept)             `№`             Age         Sexmale        Children
        48874.8          -437.6          -320.0         -1896.0          3489.5
   RegionAlmaty    RegionAstana    RegionAtyrau RegionKaragandy RegionKyzylorda
       -23455.4         -16843.0        -17453.6        -12602.1        -29692.3
 RegionOskemen     RegionSemei  RegionShymkent      RegionTaraz      Policlinic
        -7511.8         -21623.8        -23890.1        -18904.0          -310.8
```

Figure 7: Data model.

The construction of the data model by introducing a nonlinear link was programmed as follows:

```
Dinfo$Age2<-Dinfo$Age^2
Dinfo$Contribution18<-ifelse(Dinfo$Contribution>=19,1,0)
ins_model2<-lm(Contribution ~ Age+Age2+Sex+Region,data=Dinfo)
summary(ins_model2)
```



```
Call:
lm(formula = Contribution ~ Age + Age2 + Sex + Region, data = Dinfo)

Residuals:
    Min      1Q  Median      3Q     Max
-6960.2  -280.8     0.0   789.4  5941.2

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      59652.063  26890.521   2.218  0.06203 .
Age               -953.700   1114.427  -0.856  0.42045
Age2                 8.225     10.800   0.762  0.47116
Sexmale             21.298   2668.971   0.008  0.99386
RegionAlmaty    -22711.936   5366.959  -4.232  0.00388 **
RegionAstana    -21812.750   8562.309  -2.548  0.03824 *
RegionAtyrau    -15148.413   7089.300  -2.137  0.06996 .
RegionKaragandy -12013.062   6839.766  -1.756  0.12245
RegionKyzylorda -21700.000   6547.516  -3.314  0.01287 *
RegionOskemen    -7201.598   8270.593  -0.871  0.41276
RegionSemei     -32122.349  11046.966  -2.908  0.02273 *
RegionShymkent  -20136.911   7542.701  -2.670  0.03201 *
RegionTaraz     -24645.173   8390.593  -2.937  0.02180 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4630 on 7 degrees of freedom
Multiple R-squared:  0.8729,    Adjusted R-squared:  0.6549
F-statistic: 4.005 on 12 and 7 DF,  p-value: 0.03745
```

Figure 8: Nonlinear analysis model.

Here, one can perform statistical analyses of data on the impact of a person’s age on medical bills or whether it is dependent or independent of the location.

The implementation of the described statistical methods from the point of view of machine learning allows to make predictions on the payments received for medical insurance, that is, to analyse the funds coming into the budget by regions, to see the operations carried out in several steps in the form of a single graph from a centralised database. Presentation of these methods to students of medical educational institutions allows to improve their IT competence. It is crucial to establish, evaluate and assimilate evidence based on scientific research related to public health issues in order to improve the quality of health care services.

CONCLUSIONS

It can be noted that the statistical machine learning methods used in the study are important for data analysis in the field of medicine. The statistical methods considered are universal for data analysis and the software environment for their implementation is an accessible, open resource.

The improved subject contributes to the acquisition of new knowledge in the field of natural sciences, biomedical and clinical sciences in integration with IT. The application of constantly evolving IT in various types of professional activities of medical doctors is necessary for their successful practice and should be included in continuing education.

REFERENCES

1. Serik, M., Nurgaliyeva, S. and Balgozhina, G., Introducing robotics with computer neural network technologies to increase the interest and inventiveness of students. *World Trans. on Engng. and Technol. Educ*., 20, **1**, 33-38 (2022).
2. Bigdata: Advancing Healthcare Research (2019), 21 October 2022, https://qmu.edu.kz/ru/news/view/3381.
3. Cherikbayeva, L.S. and Turkistan, B.Y., Application and research of effective machine learning algorithms in medical data processing. *Bulletin Abai KazNPU. Series of Phys. & Math. Sciences*, 78, **2**, 179-187 (2022).
4. Ivanova, M.B., Berdiyeva, M., Abdrimova, Z., Khalmenov, Z. and Maulenova, A., Syllabus on the Subject *Information and Communication Technology*. South Kazakhstan Medical Academy, 2-5 (2022).
5. The Australian Institute for Machine Learning (AIML) Conducts Globally Competitive Research and Development in Machine Learning, Artificial Intelligence, Computer Vision and Deep Learning, 28 October 2022, https://www.adelaide.edu.au/aiml/about-us
6. Kolachalama, V.B. and Garg, P.S., Machine learning and medical education. *NPJ Digital Med*., 1, **54,** 1-3 (2018).
7. Handelman, G.S., Kok, H.K., Chandra, R.V., Razavi, A.H., Lee, M.J. and Asadi, H., eD octor: machine learning and the future of medicine. *J. of Internal Medicine,* 284, **6**, 603-619 (2018).
8. Serik, M., Akhmetova, B., Shyndaliyev, N., and Mukhambetova, M., Supervising and managing STEM projects for school students by the school-university model. *World Trans. on Engng. and Techn. Educ*, 20, **2**, 95-100 (2022).
9. Orynbayeva, A. Syllabus on the discipline *Basics of Biostatistics,* Astana Medical University, 1-17 (2022).
10. Dixon, W.J. and Massey, F.J. Jr., *Introduction to Statistical Analysis:* McGraw-Hill, 55-58 (1951).
11. Sidey-Gibbons, J.A. and Sidey-Gibbons, C.J., Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19, **1**, 1-18 (2019).
12. Serik, M., Nurbekova, G., Mukhambetova, M. and Zulpykhar, Z., The educational content and methods for big data courses including big data cluster analysis. *World Trans. on Engng. and Technol. Educ*., 20, **3**, 203-208 (2022).