O. M. Kozachko, Cand. Sc. (Eng.), Assistant Professor; S. O. Zhykov, Cand. Sc. (Eng.), Assistant Professor; T. E. Vuzh, Cand. Sc. (Eng.), Assistant Professor; A. O. Lotoskyi

METHOD OF THE ANALYSIS OF FOREIGN LANGUAGE KNOWLEDGE LEVEL OF THE STUDENTS OF HIGHER EDUCATION ESTABLISHMENT ON THE BASE OF MACHINE LEARNING

Paper is devoted to the development of the method of the analysis of the level of foreign language knowledge of the students of higher education establishment on the base of machine learning methods. By means of machine learning methods the regularities and trends may be determined which enable to improve the level of knowledge of the students of higher education establishments. The tasks of the paper, according to aim, put forward is versatile analysis of the data: analysis of the subject area, intelligence analysis, model construction and determination of the subjects, influencing the level of foreign language study. For the search of the type of theclassifier mathematical model the decision trees are used in the given research.

For the development of the method of the analysis of thelevel of foreign language knowledge of the students the technologies of machine learning are used, by means of these technologies various models of decision trees are developed with further selection of the best one. Regularities determination is carried out as a result of decision trees construction on the samples of grades, obtained by the students of M. I. PirogovVinnytsiaNational Medical University in the 2nd, 4th and 6th semesters. Within the frame of this study the search of the classifier type was carried out on the base of gradient boosting and logistic regression. The experiments, carried out, showed that the rules, obtained by means of regression model more accurately forecast the level of foreign language knowledge. On the base of these studies adequate and rather accurate conclusions regarding the revealed regularities are mode.

The suggested method enables to reveal the regularities of determination of the level of foreign language knowledge of the students of higher education establishments, using the methods of machine learning and determine subjects, having greatest impact of the level of foreign language knowledge. For the determination of the level of foreign language knowledge of the students of higher education establishments the program module in the form of the web-system, using the basic web-technologies was developed, the module allows to solve the task, put forward, applying automatic facilities and provide recommendations, regarding the improvement of foreign language knowledge. Program module comprises the web-site with the connected data base. The results of the research can be efficiently used for the improvement of modern education process.

Key words: *decision tree, intelligence analysis, models construction, revealing of regularities, impact factor, Python.*

Introduction

Rapid development of Information Technologies leads to the growth of the volume of the stored information. The data can be used for carrying out information analysis for revealing general trends and features. By means of the determinedtrends and characteristics the work of the system can be optimised, improve the performance, reduce the amount of losses. The importance of studying English nowadays is obvious. English may open limitless possibilities in every-day and professional life, that is why the problem is urgent.

Taking into account the challenges and problems, facing modern education process, the level of usage of modern intelligent systems and algorithms for improvement education level and teaching in higher education establishments becomes higher. Numerous studies, aimed at the determination of regularities in this sphere are carried out. These studies can be efficiently applied for determining and revealing problems, existing in the sphere of education and for the determination of individual and collective peculiarities of the students by means of introduction of the classification process and regression analysis of the data set.

Some problems and approaches to their solution will be considered, they will help to determine the direction of the research and avoid the advent of the similar problems.

In the study [1] results, obtained due to the application of the algorithms of data analysis are described and demonstrated. Basic peculiarities, dependences and methods of allocation of the main features and factors from the data set are determined. Forecasting of the students' characteristics will help to divide them into various classes, that will help these students to develop their communication, leaders skills and self governing skills while their study at the university or other education establishments. The results show that assessment of the efficiency indices is an integral part of the improvement of modern education process.

The next example is the paper [2], where many examples and ideas from other sources are considered and systematized. Computational thinking is the terminology that comprises complex set of thinking processes, carried out for the problem setting and solution by means of computational tool.The ability to systematize problems and solve them by means of these tools is considered to be the skill which must be developed by the students along with the language, mathematics and other subjects.

Taking into consideration that informatics has numerous roots in the branch of mathematics, it is expedient to consider if it is possible to influence the study of mathematics, suggesting the students the measures, connected with computational thinking. In this sense the given paper presents the systematic survey of the literature, regarding the proofs concerning teaching mathematics in the activity, aimed at the development of skills of the computational thinking. Forty two papers, containing the solutions for the assessment of the results of teaching, published from 2006 till 2017 were analysed.

Paper [3] considers the problem what is common between the development of computer thinking and study of the English language, what methods are most efficient and simple. Efficient teaching of computational thinking for the pupils, studying English correlates with other forms of the content education. Analysis of the computer code can be used for the construction of meta-understanding of the computational semiotics and visual nature of certain programming languages, such as Scratch, can promote the development of literacy.

Most important is that the projects, connected with the computational thinking whether for stories creation or for the development of electronic text projects, give students wide possibilities for self-expression and development of their identity. Now the sphere of computational thinking in education is formed and it is important that our teaching computational thinking meets the requirements of different students.

Actuality of the work is that by means of the method of machine teaching the regularities and trends can be revealed, this enables to improve the students' knowledge level. That is why the studies in this domain are of great importance.

Object of the research is the analysis of foreign language knowledge of the students of M. I. Pirogov Vinnytsia National Medical University.

Subject of the research are the methods of the machine learning of the students of higher education establishments.

Aim of the study is determination of the subjects, influencing most the level of foreign language knowledge.

Task of the research according to the aim is versatile analysis of data: analysis of the subject area, intelligence analysis, construction of the object model and determination of the subjects, influencing the level of foreign language study. On the base of these studies adequate and sufficiently accurate conclusions, regarding the revealed regularities can be made.

Survey of the forecasting methods and problem set up

For the development of the method of analysis of the level of foreign language knowledge of the students of higher education establishment technologies of machine learning are used in the given research, by means of these technologies various models of the decision trees with further selection of the best one are developed. Initial data are grades on different subjects, presented in the form of200-points scale. Assume thatY-is the student grade in foreign language and $X_1, X_2, ...,X_n$ – are grades in different subjects, obtained by the student during the semester. In this case the regularity between the grade in foreign language and grades in other disciplines will be found in the form of the following relation:

$$Y = f(X_1, X_2, ..., X_n), \tag{1}$$

where n– is the number of disciplines

Grade in foreign language will be divided into four levels: “poor”, “satisfactory”, “good”, “excellent”. Then the relation (1) can be considered as the classifier which establishes the correspondenceto the level of foreign language knowledge for the input vector of grades X.

For searching the type of mathematical model of the classifier (1) decision trees, built according to such methods were used: boosting, bagging, stacking, etc. Boosting is an assembler meta-algorithm of machine learning for shift decrease and dispersion in training with a teacher and the family of machine learning algorithms, which convert weak students in strong ones. [4]. Bagging is meta-algorithm of compositional learning intended for the improvement of the stability and accuracy of machine learning algorithms, used in statistical classification and regression[5]. Stacking is one of the most popular methods of algorithms assembling, i.e., usage of several algorithms for the solution of one of the problems of machine learning [6].

For the automation of the above-mentioned methods there exist three most powerful libraries: Xgboost, Catboost andLightGBM. Taking into consideration large dimension of the dataset, great number of features and limited computational possibilities library LightGBM is suggested for usage. [7].

Analysis of input data of the models

Input data for determining the regularities of the level of foreign language knowledge of the students of higher education establishments is the set of grades, obtained by the students of Vinnytsia National Medical University for the disciplines in the 2nd, 4th, 6th semesters and results of the first stage of the Single National Qualification Exam (SNQE). Information regarding the disciplines, on the base of which the regularities of the impact on the level of foreign level knowledge, are presented in Tab. 1.

Table 1

Information about the subjects, influencing the level of foreign language knowledge

№	Designation	Name of the Subject	Minimal value	Maximum value	Average value
1	2	3	4	5	6
1	X_1	Latin language and medical terminology	122	200	167.6
2	X_2	Medical biology	124	200	158.3
3	X_3	Medical and biological physics	122	200	166.4
4	X_4	Fundamentals of economic theories	153	200	179.6
5	X_5	Fundamentals of psychology and pedagogics	120	197	158.7
6	X_6	Life safety and labour protection	132	200	173.0
7	X_7	Biological and bioorganic chemistry	123	198	151.8
8	X_8	Patient care	142	200	177.4
9	X_9	Logic, formal logic	128	196	163.1
10	X_{10}	Medical informatics	131	195	165.4
11	X_{11}	Physical education	135	200	171.1
12	X_{12}	Foreign language	122	200	157.6
13	X_{13}	Civil defence	122	195	154.5
14	X_{14}	Military hygiene	122	200	164.2
15	X_{15}	General surgery	131	200	170.7
16	X_{16}	Pathomorphology	99	200	156.0
17	X_{17}	Pathophysiology	92	195	152.8

Table 1

1	2	3	4	5	6
18	X_{18}	Propedeutics of the internal medicine	122	200	159.0
19	X_{19}	Propedeutics of paediatrics	122	200	162.2
20	X_{20}	Nursing practice	150	200	179.5
21	X_{21}	Pharmacology	125	200	158.5

For performing intelligence analysis of the grades each semester was considered separately. Fig. 1 shows the correlation graph for the first dataset(thermal correlation chart).

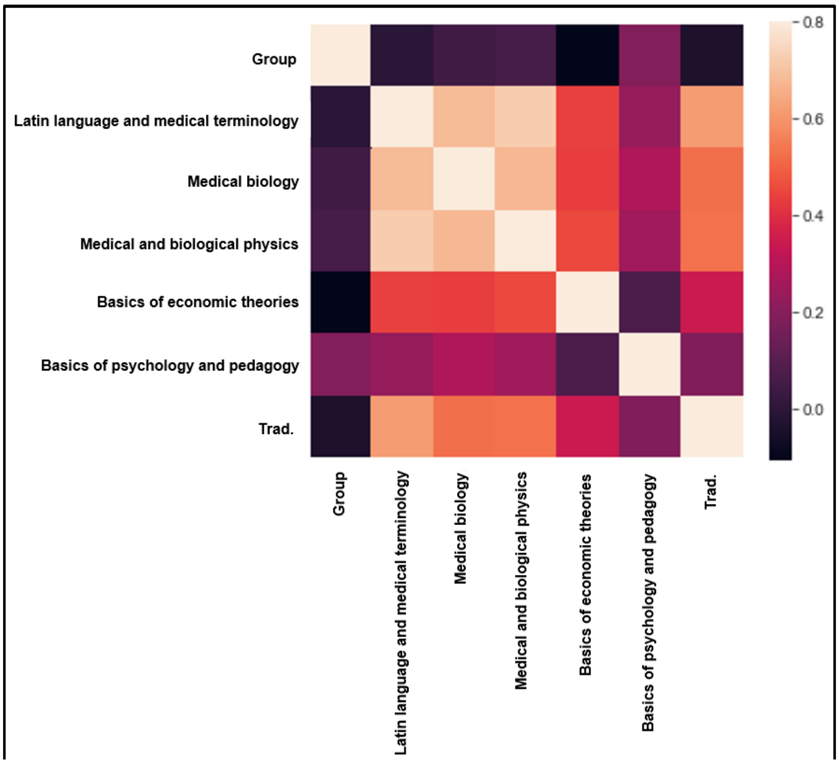
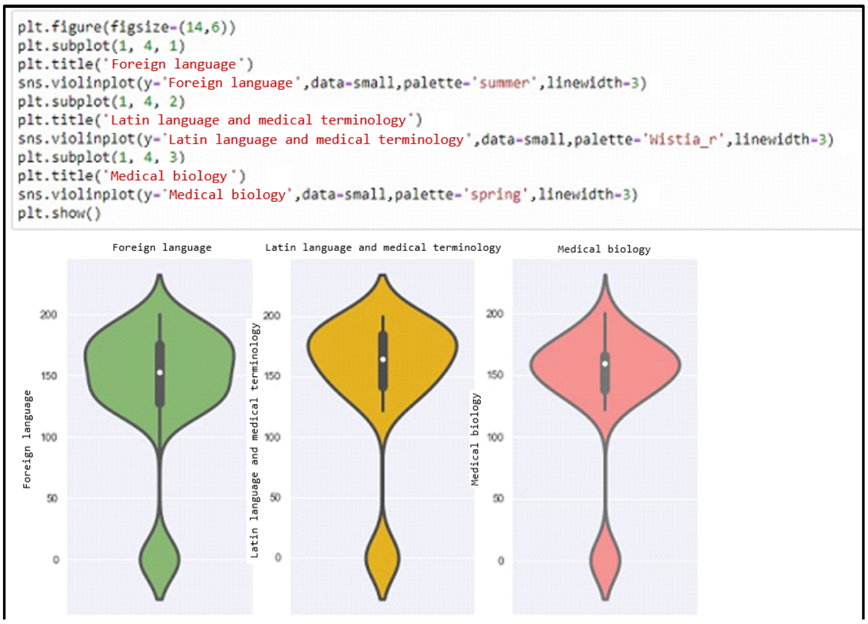


Fig. 1. Thermal correlation chart

On the base of this chart the conclusion can be made that there exists certain dependence between first three disciplines. But we would like to know what influences the results in English language. The greatest correlation of all available disciplines has “Latin language and medical terminology”, that is quite logical - ability for learning one language influences the learning of other languages. Fig. 2 presents the graphs of values distribution for all the subjects.

On the base of the distribution graphs the conclusion can be made that the most similar to the distribution of the subject “Foreign language” is the value of grades “Latin language and medical terminology”. In the same way other datasets were analysed.



```
plt.figure(figsize=(14,6))
plt.subplot(1, 4, 1)
plt.title('Foreign language')
sns.violinplot(y='Foreign language',data=small,palette='summer',linewidth=3)
plt.subplot(1, 4, 2)
plt.title('Latin language and medical terminology')
sns.violinplot(y='Latin language and medical terminology',data=small,palette='Wistia_r',linewidth=3)
plt.subplot(1, 4, 3)
plt.title('Medical biology')
sns.violinplot(y='Medical biology',data=small,palette='spring',linewidth=3)
plt.show()
```

Foreign language
Latin language and medical terminology
Medical biology
Foreign language
Latin language and medical terminology
Medical biology
200
150
100
50
0

Fig. 2. Graphs of the distributionof the values of the first three subjects

Development of object model

Within the frame of the given research the form of the classifier (1) is carried out on the base of the gradient boosting [3] and logistic regression [4].

The problem of the classifier construction will be formulated in the following way. Set of the objects grades *X* and their classes *Y*={2,3,4,5} is given as well asthe objective function *y**, value of which $y_i = y^*(x)$ are known at certain subset of classes $\{x_i\}$ of the set *X*, where x_i – is dimensionality vector *n*. It is necessary to construct the function, which by the known data (x_i, y_i) approches the objective function *y* on the whole data sample, $i = 1,..m$.

For the construction of the classifier the gradient boosting will be considered. The idea of gradient boosting lies in the construction of the set of decision trees so that each next tree tried to improve the quality of the whole combination of trees.

Classification is carried out by the following formula [8]:

$$obj\_boost(X) = \underset{y \in Y}{argmax} \sum_{k: b_k(x)=y}^{n} a_k, \tag{2}$$

where $b_k(x)$ – is the answer of *k*th tree to the object *x*; a_k – is the contribution of *k*th tree in the composition.

In the process of teaching *K* decision trees are constructed at all *m* objects and *s* randomly chosen features from the total amount of features *n*. After teaching of the tree, weights of the incorrectly classified objects increase, thus the following tree performs focusing mainly on them.

Another classifier, considered in the paper is logistic regression, that presents statistical linear model of the classification, enabling to forecast aposterior probabilities of classes by means of the logistic curve. Object is referred to the class with the greatest probability, determined by the following formula [8]:

$$obj\_reg(X) \;=\; \underset{y \in Y}{argmax}\, P(y^*(x) = y), \tag{3}$$

$$P(y^*(x) = y) = \frac{e^{<x,a_y>}}{\sum_{k=1}^{K} e^{<x,a_k>}}, \tag{4}$$

that is, object x is assigned the class with the greatest probability, which is calculated according to softmax-function a_k – is the vector of the regression coefficients, connected with class k, and $x = (x_1, \dots x_n)$ – is the vector of features.

Studies carried out, showed that logistic regression is the best model of classification. Figs. 3-5 show decision trees, obtained by means of logistic regression (3) on the dataset of the results of study in the 2nd, 4th and 6th semesters, correspondingly. It is seen from the constructed decision trees that main disciplines, influencing the level of foreign language knowledge are the following: “Latin language and medical terminology”, “Biological and bioorganic chemistry”, “General surgery”. In the same way the dataset of the data of the results of Single National Qualification Exam(SNQE) is modelled.

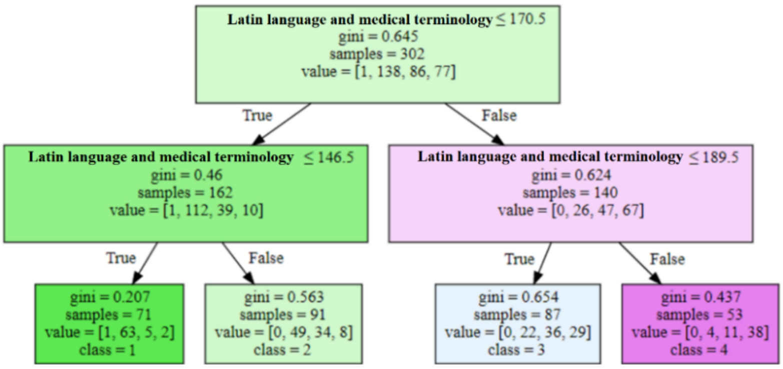


Fig. 3. Decision tree of the first dataset by the results of the second semester

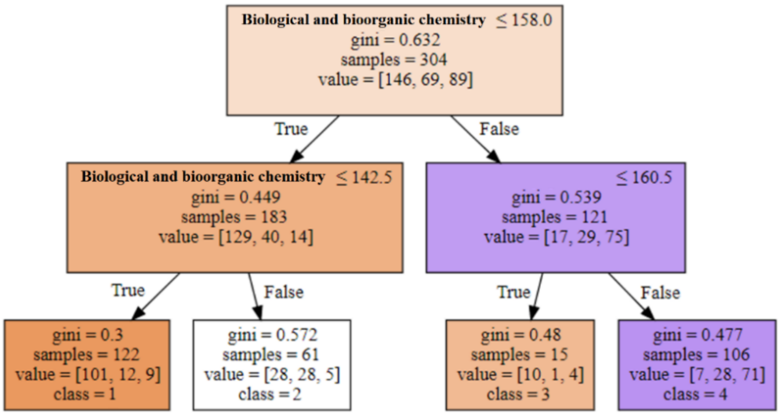


Fig. 4. Decision tree of the second dataset by the results of the fourth semester

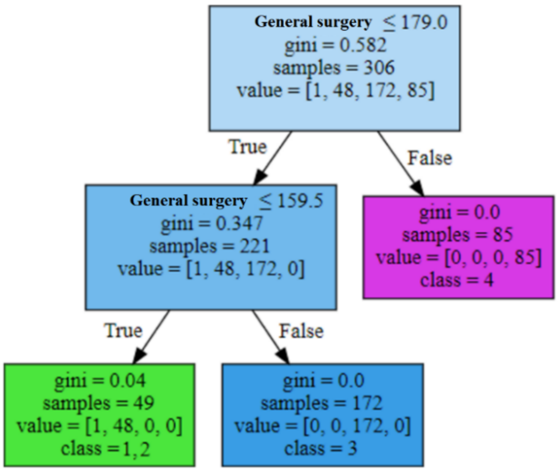


Fig. 5. Decision tree of the third dataset by the results of the sixth semester

From Fig. 3 the regularity is seen, that can be interpreted in the form of the rule "If-Then":

– if student obtained the grade in the subject "Latin language and medical terminology" less than 146.5 points, then the grade of foreign language will be "poor";

– if the grade in the subject "Latin language and medical terminology" is within the range from 146.5 to 170.5 points, then the grade in foreign language will be "satisfactory";

– if the grade in the subject "Latin language and medical terminology" is within the range from 170.5 to 189.5 points, then the grade in foreign language will be "good";

– if the gradein the subject "Latin language and medical terminology" is greater than 170.5, then the grade in foreign language will be "excellent".

In the same way decision trees, shown in Figs.4,5 are interpreted.

Accuracy of the classification of the machine learning models is presented in Table 2.

Table 2

Classification accuracy of the machine learning models

	2 semester	4 semester	6 semester	SNQE
XGBoost	71%	65%	77%	60%
Logistic regression	84%	73%	99%	78%

Accuracy of the classification is determined by the error rate, according to the formulas:

$$F(X) = \sum_{j=1}^{m} \frac{\Delta_j}{m}, \tag{5}$$

$$F(X) = \begin{cases} 1, \ if\ obj_j(X) = Y_j, \\ 0, \ if\ obj_j(X) \neq Y_j. \end{cases} \tag{6}$$

For the determination of the level of foreign language knowledge of the students of higher education establishment the program module in the form of the web-system, applying main web-technologies was developed: hyper textmarkup languages (HTML),CSS and templates of Bootstrap library for the construction of the layout of the site, local serverDenwer, web-interface phpMyAdmin for the work with the database, written by the query language SQL. Also PHP was used for the connection of the site layout with the database. Program module comprises the website with the connected database.

Conclusions

Method of revealing the regularities of the level of foreign language knowledge of the students of higher education establishments is suggested in the research, unlike the available technologies it uses the machine learning technologies and provides the identification of the subjects , study of which influences the level of foreign language knowledge. Revealing of the regularities is carried out as a result of decision trees construction on the samples of the grades, obtained by the students of M. I. Pirogov Vinnytsia National Medical University in the 2nd, 4th and 6th academic semesters. Decision trees were constructed by means of boosting methods and logistic regression. Experiments, carried out, showed that the rules, obtained by the regression model forecast the level of foreign language level more accurately. Besides, the subjects, influencing the improvement of foreign language knowledge were identified, namely: "Latin language and medical terminology" is the most influentialsubjects in the 2nd semester, "Biological and bioorganic chemistry" in the 4th semester and "General surgery" in the 6th semester.

REFERENCES

1. Predicting Pupil's Successfulness Factors Using Machine Learning Algorithms and Mathematical Modelling Methods [Electronic resource]. Access mode: https://link.springer.com/chapter/10.1007/978-3-030-16621-2_58.

2. Mathematics Learning through Computational Thinking Activities: A Systematic Literature Review [Electronic

resource]. Access mode: http://jucs.org/jucs_24_7/mathematics_learning_through_computational/jucs_24_07_0815_0845_barcelos.pdf.

3. Teaching computational thinking to english learners [Electronic resource]. Access mode: https://par.nsf.gov/servlets/purl/10073683.

4. Zhi-Hua Z. Ensemble Methods: Foundations and Algorithms / Zhou Zhi-Hua. – New York : Chapman and Hall/CRC, 2012. – 236 c.

5. Shinde A. Preimages for Variation Patterns from Kernel PCA and Bagging / A.Shinde, A. Sahu, D. Apley, G. Runger. // IIE Transactions. – 2014. – Vol. 46. – P. 429 – 456. DOI: 10.1080/0740817X.2013.849836.

6. A Kaggler's Guide to Model Stacking in Practice [Electronic resource]. Access mode: http://blog.kaggle.com/2016/12/27/a-kagglers-guide-to- model-stacking-in-practice/.

7. Random classification noise defeats all convex potential boosters [Electronic resource] / Philip M. Long, Rocco A. Servedio // Machine Learning. – 2010. – 78. – P. 287 – 304. DOI: https://doi.org/10.1007/s10994-009-5165-z.

8. Module pandas_profiling. [Electronic resource]. Access mode: https://pandas-profiling.ydata.ai/docs/master/.

9. Order of the Ministry of Healthcare of Ukraine of 22.01.2021 №106 [Electronic resource]. Access mode: https://zakon.rada.gov.ua/laws/show/z0269-21#n7. (Ukr).

10. Some Studies in Machine Learning Using the Game of Checkers [Electronic resource]. Access mode: https://ieeexplore.ieee.org/document/5392560.

Editorial office received the paper 20.12.2021.
Thepaperwasreviewed 25.12.2021.

Kozachko Oleksii – Cand. Sc. (Eng.), Assistant Professor with the Department of System Analysis and Information Technologies.

Zhukov Serghiy – Cand. Sc. (Eng.), Assistant Professor with the Department of System Analysis and Information Technologies.

Vinnytsia National Technical University.

Vuzh Tetiana – Cand. Sc. (Eng.), Assistant Professor with the Department of Biological Physics, Medical Equipment and Informatics.

M. I. Pirogov Vinnytsia National Medical University.

Lototskyi Andriy – Master Student with the Department of System Analysis and Information Technologies.

Vinnytsia National Technical University.