Predicting Students' Academic Performance Based on the Cluster Analysis Method

Olha Pronina[1] and Olena Piatykop[1]

[1] *State Higher Educational Institution «Pryazovskyi State Technical University», University str., 7, Mariupol, 87500, Ukraine*

Abstract

A mathematical model for assessing student performance based on cluster analysis has been developed. Assessment is based on two stages: assessment of applicants to identify patterns between the results of external independent testing and assessment based on the results of current assessments and exams of students.

The developed model gives an idea of how well the student will study in the areas of disciplines that are related to external independent testing. And also, about how well the student will study, based on the results of intermediate certifications and sessions. This, in turn, makes it possible to draw attention to specific students who may have problems in a number of disciplines and topics. And also, on the contrary, to identify students that have a clear inclination to discipline and it may be worth deepening this knowledge. Thus, allowing the introduction of an individual training vector.

Experimental studies have been carried out to confirm the effectiveness of the developed model for the first and second stages of assessment. The developed model based on cluster analysis makes it possible to use it in the future to analyze and predict the progress of students of any higher educational institution in Ukraine.

Keywords

Predicting student learning progress, students' academic performance, educational data mining, cluster analysis, Kohonen map, nearest neighbor distance.

1. Introduction

Education is one of the most important components of the life of society. In the modern world, specialists with a high level of knowledge and skills play an important role. In order to obtain such specialists, an integrated approach is needed, one of the stages of which is obtaining a higher education.

The main task of universities is to educate students in their specialty. To accomplish this task, high quality education is required. The quality of education is a social category that determines the state and effectiveness of the education process in society, its compliance with the needs and expectations of society in the development and formation of civil, domestic and professional competencies of the individual.

One of the most important components of the quality of education is the success of education. The success of training is understood as a complex indicator that characterizes the quality of the various knowledge and competencies of the graduate formed over the years of training. The following stages of monitoring the success of training can be distinguished: monitoring the flow of applicants; monitoring progress and other activities; monitoring the success of future professional activities. Today, more than ever, it is important to have the most accurate assessment of the quality of education in order to assess the development trends of our future. One of the most common criteria of which is the assessment of a student's knowledge by experts (teachers or employers).

WS MROL: ICTERI-2021, Vol II: Workshops, September 28 – October 2, 2021, Kherson, Ukraine
EMAIL: pronina.lelka@gmail.com (O. Pronina); piatykop_o_ye@pstu.edu (O. Piatykop)
ORCID: 0000-0001-7085-8027 (O. Pronina); 0000-0002-7731-3051 (O. Piatykop)

In this work, the object is the first and second stages of monitoring and its indicators - the results of entrance examinations and student performance in the learning process. At the stage of study in a higher education institution, the most important stage of knowledge testing is the assessment by teachers. Grades of students' progress are assigned in accordance with their knowledge in learning, assessed: tests, exams, tests, laboratory work, practical work, independent work, abstracts, etc. The scores obtained are a natural and universal indicator that allows you to identify the degree of mastery of the material; also, on the basis of assessments it is possible to determine the trends affecting the success [1].

Currently, in almost all areas of human activity, there is a general need for the study of statistical data describing the behavior of observed objects, events, processes or phenomena. One of the most urgent and practically demanded tasks of data analysis is the problem of dividing objects into comparatively homogeneous groups (subsets), which are called clusters.

At the moment, the number of methods for dividing groups of objects into clusters is quite large - several dozen algorithms and even more of their modifications. Cluster analysis should take into account the temporal dynamics of academic performance, which will make it possible to more fully reveal the picture of the quality of education. An in-depth analysis of the data reveals key trends in the success of education. Subsequently, it makes it possible to influence the quality of the educational process by making optimal management decisions. Therefore, cluster analysis was chosen for this work.

2. Literature review

Significant experience in studying the relationship between pre-university testing results and further success at the university has been accumulated in the United States, where many universities take into account the results of standardized tests, such as SAT (Scholastic Aptitude Test, Scholastic Assessment Test) or ACT (American College Testing), conducted by private organizations for decades.

Research [2] is devoted to determining whether there is a relationship between the SAT or ACT, HSGPA and the final grade of students in the freshman English course. The authors of [3] investigated the high school general education score (HSGPA) and ACT scores as predictors of college graduation. The relationship of HSGPA to college graduation results has been found to be strong and consistent and greater than the relationship to ACT score ratio.

In Ukraine, at present, as a criterion by which the ranking of applicants is carried out, the summation of points obtained in external independent testing in several subjects is used. The list of subjects from three or four entrance examinations for each specialty is determined by the rules for admission to a higher educational institution.

Obviously, the choice of the disciplines of entrance examinations and weighting factors for various subjects when summing up points plays an important role in the selection procedure for applicants. Weighting factors determine the contribution of individual disciplines to the final criterion. For example, in article [4], the possibility and admissibility of using the results of external independent testing to identify applicants who can successfully continue their education in higher education are considered. Mathematically sound recommendations for improving the quality of competitive selection are proposed to identify applicants and their potential for study. The recommendations are based on a statistical analysis of data on the performance of applicants for higher education of the first year of study in the specialty 226 "Pharmacology" of the branch of knowledge "Health", who joined the National University of Pharmacy in 2015, in comparison with the competitive score. It has been proved that by purposefully changing the weights used in calculating the competitive score, a higher educational institution can positively influence the qualitative composition of applicants for higher education.

The influence of the results of general independent assessment in mathematics of external independent testing, together with other indicators, is also taken into account in [5] to predict the potential of freshmen and apply the differential approach by a teacher. Correlations between the current performance of applicants for higher education for the first year and the components of the competitive score are also carried out by the authors of work [6] for specialty 151 "Automation and computer-integrated technologies". The relevance of methods for automated data analysis in educational systems is confirmed by a number of studies of learning processes [7 – 12]: dynamic systems when modeling

educational trajectories; research on nonlinear methods and statistical tools to describe changes in student performance; the use of analytical tools for decision-making processes in the context of the growth of educational data, requiring the extraction of knowledge from data sets; management of the quality and content of training programs that meet the requirements of the labor market; the use of an agent-based approach to build the architecture of a learning management system, taking into account various types of educational activities of students; creation of stereotyped models of groups of students; the application of tools and standards for the analysis and structuring of managed learning content.

The main approaches of the methods of data mining with educational data (Educational Data Mining, EDM) are used for forecasting, clustering, analysis of relationships, identifying and extracting data for the purpose of human decision-making [12-13]. Education uses many data mining techniques to extract hidden knowledge from student performance data [12-20].

The paper [14] also explores the application of data mining to educational data. To assess the progress of students, assessments, tests, intermediate and final exams, assignments, laboratory work, etc. were used. As a result, the author determined that a slight modification of the K-means method is required. Then this clustering method may give better results in predicting student performance. In the article [15], using data mining methods, the influence of other factors on student performance is assessed. The author investigated the impact of student behavior and classroom absenteeism on student performance using methods: Decision Tree, K-Nearest Neighbor algorithms, Artificial Neural Network.

The work [16] is also devoted to predicting student performance. The authors applied a hybrid algorithm combining clustering and classification to the academic, behavioral, and complementary characteristics of the student dataset. The proposed model can help teachers recognize weak learners in order to change the learning process and reduce student failure rates. The authors used different methods in their work [17]. The main goal of this study was to examine the effect of grades and online activity data on student performance. To predict based on classification, the authors applied five classification algorithms: decision tree, random forest, sequential minimal optimization, multilayer perceptron, and logistic regression.

In the work [18], methods of data analysis are proposed in order to extract hidden knowledge from data by performing the tasks of pattern recognition and predictive modeling. The author applies cluster analysis and decision trees to educational data of a higher education institution in Croatia. The author proposes to use the obtained results to modify the course, which should be given more attention. Thus, change the methods of motivation to prepare for the exam.

Researchers in their work [19] also use EPM (Educational process mining) methods to analyze learning processes. Educational process mining is an evolving discipline that provides evidence-based understanding and support for processes and provides new tools to improve educational processes. There are significant advantages to combining educational data with data mining techniques to understand educational processes.

Thus, a large number of researchers are engaged in the issue of predicting student performance. The authors analyze methods that can be used to make predictions or look for methods that can improve predictions [2-20]. At the same time, the authors use different attributes for forecasting: classroom grades, examinations, final exams, behavior, attendance, and more [2-20]. Also, the authors carry out grouping and classification according to different parameters. Therefore, it is difficult to compare the studies of different authors since each study has its own purpose. Nevertheless, all research in the area of predicting student performance is relevant and important. As of today, there is still no unified approach, so research needs to be continued.

The aim of this work is to predict student performance based on the method of cluster analysis. For this, a number of tasks are considered. It is necessary to form a mathematical model that will allow predicting the level of student performance based on the data of external independent testing and control activities. It is also necessary to conduct modeling, training, testing, and evaluate the results. This article is dedicated to describing this.

3. Mathematical model

The developed model for assessing student performance based on a variety of input factors is customizable. To solve the problem, it must be divided into two stages. At the first stage, it is necessary

to predict the numerical values of assessments of student learning outcomes. For example, predicting the average student learning scores for a semester or for the entire period of study, depending on the values of the input variables (for example, scores of external independent testing, academic performance in the semester, academic performance in the first part of the discipline, that is, the first certification, etc.). This requires the structure and values of the parameters of the data analysis model, input variables and factors. At the second stage, adjust the finished model according to the parameters of the independent variables and the dependent variable. This will ensure the maximum quality of the model under the feasible conditions specified in the form of constraints.

Determination of independent variables corresponding to a specific subject area and forecasting goal is a separate task. In addition to the scores of students' progress, it was found that competencies were proposed as independent variables of the model, since the study of any discipline involves the acquisition and improvement of a number of competencies.

The formation of a data analysis model (second stage) is carried out using the following procedures: creation of training and test samples of data from external independent assessment and assessments of the results of educational activities of students on the basis of a priori data; preparation of data characterizing the parameters of competencies for each discipline for students; building cluster models using various structures, significant factors input variables and methods of cluster analysis; evaluating the quality of models on test samples.

Prediction of learning outcomes for new groups of students based on the developed models (first stage) is carried out using the following procedures: collection of data from external independent assessment and results (assessments, etc.) obtained at a certain stage of forecasting; diagnostics of the parameters of the competency model; diagnostics of the initial forecast of the learning outcomes of 1st year students until the end of the first year of study and until the end of the bachelor's degree; consistent refinement of the results of individual predictions for each student, taking into account the marks obtained in the learning process and competencies.

Assessment of students in a point-based system is a set, which is a combination of the results of completing individual blocks of tasks, which is assessed by the teacher:

$$R=\{C_1 \cup C_2 \cup \ldots \cup C_n\} \quad (1)$$

where n – the number of controls, R – the rating; C – the control.

Moreover, each type of control can include several tasks. So, for example, the assessment of the subject consists of laboratory works, tests, independent robots and the final assessment on the exam. In turn, the number of laboratory work varies from eight, independent and control from two. Thus, each type of control event can be represented as a set:

$$C_n=\{c_1;c_2;\ldots;c_m\} \quad (2)$$

where C_n are the type of control; m is the number of control measures of one type of control.

Thus, the grade for the subject can be presented in the following form:

$$R=\{(c_1^1;c_1^1;\ldots c_m^1) \cup (c_1^2;c_1^2;\ldots c_m^2) \cup \ldots \cup (c_1^n;c_1^n;\ldots c_m^n)\} \quad (3)$$

Since each type of control is assessed with a different number of points, then each $c_m^n=g_i$. Then:

$$C_n=S_i(g_i) \quad (4)$$

where i is the number of the set of estimates; S is the dependency function.

In this case, the function S can take any form, both percentage and summation, or it can be in the form of any complex function that characterizes the dependence of the assessment on the type of control. Thus, the final expression of the assessment for the subject can be represented as the sum of all types of control:

$$R=\sum S_i(g_i) \quad (5)$$

A similar methodology is used to form competencies in each discipline. Taking into account the fact that several competencies can be formed in one discipline, the student's rating can be presented in the following form:

$$R=\{K_1 \cup K_2 \cup \ldots \cup K_t\} \quad (6)$$

where t is the number of competencies in this discipline.

The formation of competencies is also associated with blocks of assignments for each discipline.

$$K_t=\{C_1 \cup C_2 \cup \ldots \cup C_n\} \quad (7)$$

It is worth noting that the verification of the formation of competencies can be carried out not at all control events. That is, the number of competencies may be less and checked less often.

To receive one hundred points for a discipline, it is necessary to score both the maximum in all control events, and the number of competencies should also be maximum.

$$\sum K_t = R_{max} \tag{8}$$

When checking these two parameters, namely the number of points and the quality of mastering the material in the form of the number of competencies, we can talk about the complete study of the discipline provided. The resulting mathematical model can be used to calculate the rating of each student. As a further action, students can be grouped according to their results into clusters. The general performance of students can be conditionally divided into three clusters: unsatisfactory, which corresponds to an insufficient level of knowledge; good and enough, which corresponds to a sufficient level of knowledge; excellent, which corresponds to a high level of knowledge. On the basis of which the forecasting of students' success takes place, with the possibility of continuous recalculation of ratings to check the success of the forecast.

4. Modeling

To group data into clusters, we need to define their structure. The structure is understood as the distribution of students according to the calculation of their rating, which includes the amount of points and the number of competencies. In addition to defining the structure, it is necessary to determine the methodology for getting future data into organized clusters. For this, it is customary to calculate the distance between each cluster center and the new data position from the sample.

There are many options for calculating the distance between cluster objects. In the work, a method for combining into clusters was chosen - the "nearest neighbor rule" (single link method) to determine the distance between clusters. In order to determine the size of the clusters and their homogeneity, after calculating all the distances between the centers of the clusters, it is necessary to set the maximum allowable distance. By estimating the distance and changing it, it is possible to determine the size of the cluster and how homogeneous it is. During the training process, the model allows you to determine the ownership of data in existing clusters or identify new clusters.

During the tuning of the neural network, it was found that the maximum error should be less than 0.05. For the training sample, the training interval was adjusted, namely, the recalculation of the training error every 20 epochs, and after that the data was mixed. The training radius was set to 4 at the beginning of training and 0.1 at the end of training. The neighborhood function is stepwise.

The learning process lasted 223 epochs, 100% of the data was recognized and processed. At the time of training, the maximum error reached 0.629331, at the time of testing, the maximum error reached -0.392486.

Since the data in the sample are evaluated on different scales, for example, a semester is a ten-point scale, an assessment is a hundred-point scale, it is necessary to normalize the data. Normalization is understood as the transformation of the original data, which translates them into dimensionless quantities. The transition from traditional units of measurement to normalized and vice versa using the linear normalization method is carried out using the following calculated ratios [11]:

When normalizing and denormalizing within [0; 1], we use formula (9).

$$\tilde{x}_{ik} = \frac{x_{ik} - x_{min_i}}{x_{max_i} - x_{min_i}}, \qquad y_{jk} = y_{min_j} + \tilde{y}_{jk}(y_{max_j} - y_{min_j}) \tag{9}$$

where x_{ik} and y_{jk} are the i-th input and j-th initial value of the k-th example of the original sample in traditional units of measurement adopted in the problem being solved; $\tilde{x}_{ik}$, $\tilde{y}_{jk}$ are ascertained and normalized input and output values; N is the number of examples of the training sample.

For the task at hand, normalization was carried out for a number of indicators by formula (9). In this work, cluster analysis was used, the entire collected data set was divided into two samples: training and test. After training and testing the neural network, an experiment was carried out on a small amount of data. The experiment used data that had not previously passed through the mathematical model.

The data set, which was launched for testing, contains the results of the first certification, the second certification and the results of the session of students of the Faculty of Information Technologies, the Department of Informatics of the SHEI "Pryazovskyi State Technical University". The model provides for an increase in the criteria for evaluating additional indicators, namely, grades for laboratory work, practical work, independent work, tests, abstracts, and the like.

In the future, this method can be used for big data, namely, data on all faculties of the SHEI "Pryazovskyi State Technical University", or when changing the data structure for another university. To construct a cluster analysis, significant indicators were identified for each of the separation options. The division into clusters was carried out in the Deductor Academic environment [15]. At the next stage, the similarities between the elements of the cluster were determined and the Kohonen map was constructed [16]; for this, errors were found for the training set.

For the qualitative construction of clusters, it is necessary to determine not only the initial data, but also indicate the initial data and calculated data. The next step was to build communication between the clusters. For clarity of the results, a self-organizing Kohonen map was built, for each significant characteristic the interpretation of the results is shown in Figure 1. To construct a cluster analysis, significant indicators were identified: semester, group, student, subject, type of final control, assessment. For each indicator it is necessary to carry out a splitting variant. An example of splitting for the KM-19-M group, GVUZ "PSTU" is shown in **Figure 1**.

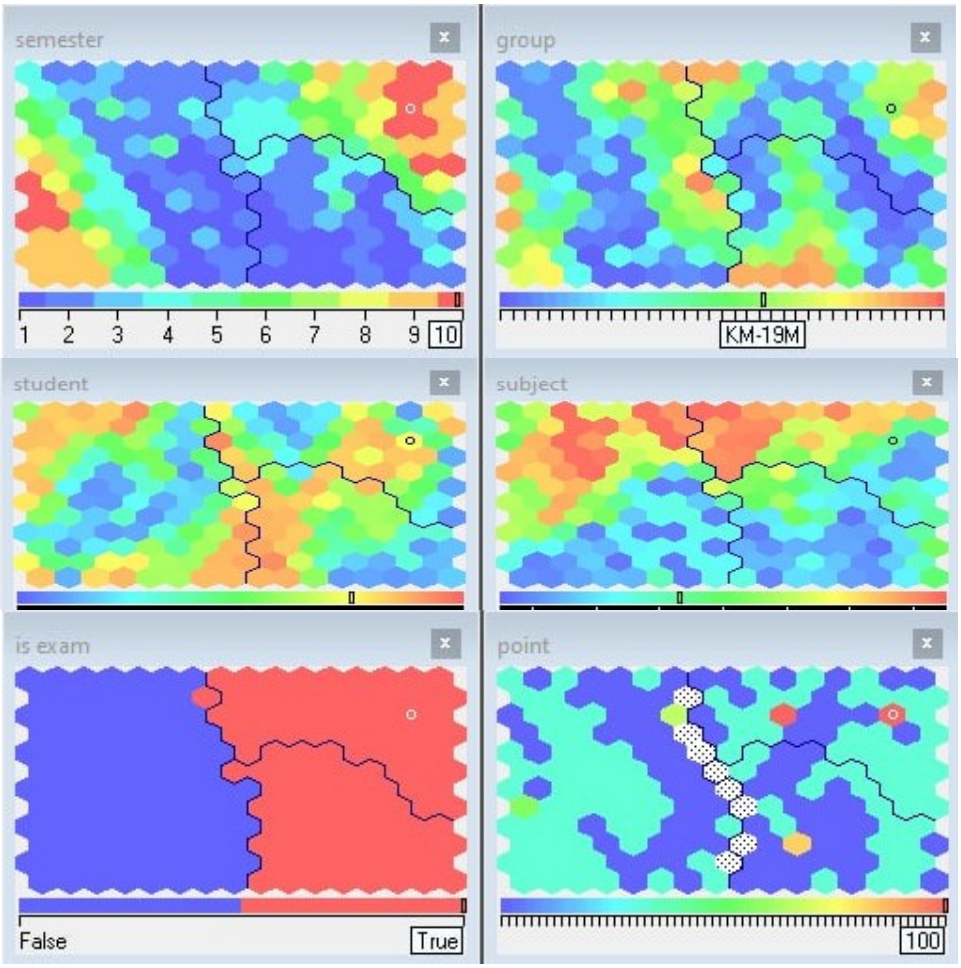

Figure 1: Constructed clusters to predict student learning success

Analyzing the results, which are presented in **Figure 1**, you can see that the entire partitioning occurs in three established clusters (red, blue and green). The color scale shows the proximity of the data to one of the clusters. So the orange dots refer to the red cluster, but the distance to the center of the red

cluster is already maximum. Shades of cyan are deleted data, but still belong to the blue cluster. Yellow and light turquoise are the distant points of the green cluster.

The calculated distance between the constructed clusters and the quantization error matrix are shown in the **Figure 2**. In the figure, you can see the color gradation of the distance between the clusters.

Distance matrix is used to visualize the structure of clusters obtained as a result of map training. Great value indicates that this neuron is very different from those around it and belongs to a different class. The quantization error matrix displays the average distance from the location of the samples to the center of the cell. The quantization error matrix shows how well the Kohonen network is trained. The smaller the average distance to the center of the cell, the closer the examples are to it and the better the model (**Figure 2**).

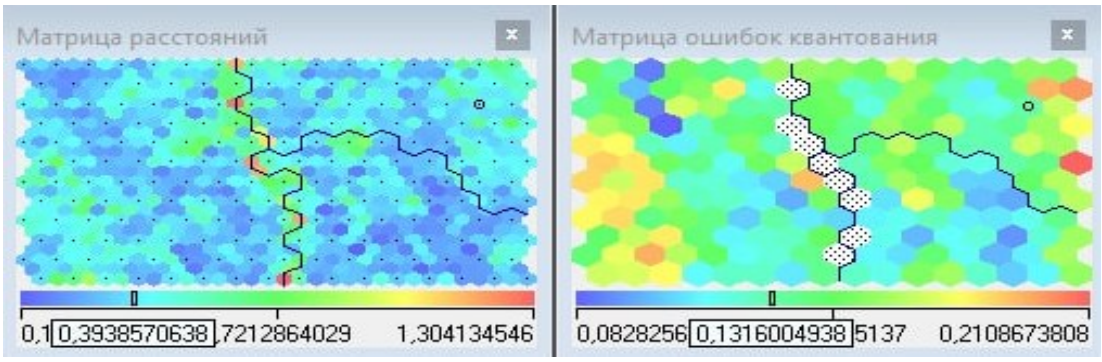

Figure 2: Cluster distance matrix and quantization error matrix

A general Kohonen map was built [16], in it there are three main clusters of learning forecast: high level, sufficient level, insufficient level, presented in **Figure 3**.

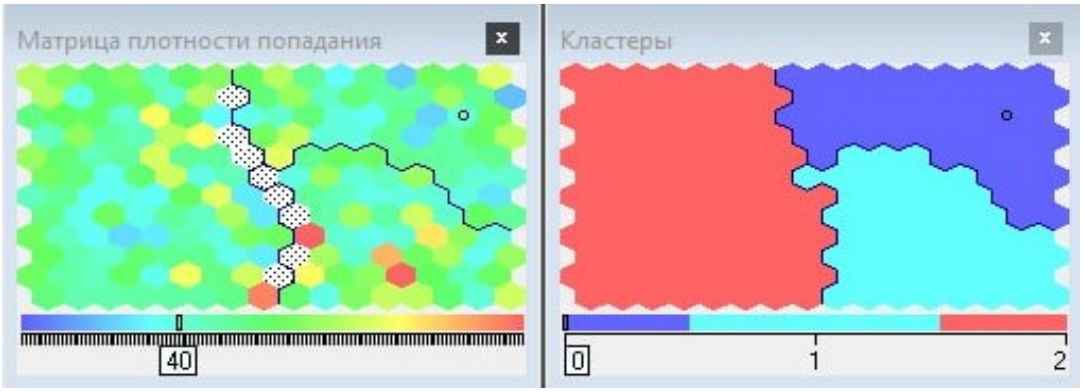

Figure 3: Results of fuzzy modeling output variable specialty, indicating the term of a specific specialty

The color demonstrates the created clusters and the hit density matrix. The hit density matrix displays the number of objects that hit the cell. Clusters - cells of the Kohonen map, combined into clusters by the k-means algorithm.

5. Experimental research

To test the developed model, a two-stage study was carried out. At the first stage, the connection between the results obtained during external independent testing and further distribution by score was checked. At the second stage, a comparison was made between the predicted grades of students and their real grades obtained during training.

The first stage of model validation. For the analysis, the data of students enrolled in the specialty "Computer Science" of the educational program "Computer Science" in 2018-2020 were selected.

According to the rules of admission, the components of the competitive score established grades in three subjects of external independent testing and an average score in the certificate. Obligatory subjects of external independent testing were "Ukrainian language and literature" and "Mathematics". The third subject in 2018 could be chosen "Foreign language" or "Physics", since 2019 it was also possible to present the results of external independent testing on the "History of Ukraine". Data on the results of external independent testing were selected on the website "Applicants search service" [17].

The data on the distribution in the third subject of external independent testing are shown in Table 1. Each of the disciplines of the UPE is evaluated from 100 to 200 points. Descriptive statistical data of the results of external independent testing in the subjects of 85 students who entered the Department of Computer Science in 2018-2020 are shown in Table 2

One-dimensional distributions and characteristics do not give an idea of how the results of UPE in various disciplines are interrelated.

Table 1

Distribution according to the selected third subject of external independent testing

Year	Subject of external independent testing		
	Foreign language	Physics	History of Ukraine
2018	58%	42%	-
2019	75%	15%	10%
2020	62%	31%	7%

Table 2

Descriptive Statistics of External Independent Testing Scores by Subject

Year	Value	Subject of external independent testing				
		Ukrainian language and literature	Mathematics	Foreign language	Physics	History of Ukraine
2018	the average	160,35	157,55	172,73	155,89	-
	the minimal	104	104	148	104	-
	the maximum	192	196	195	192	-
2019	the average	156,00	151,90	150,80	150,3	127,0
	the minimal	115	100	118	136	102
	the maximum	191	183	182	171	152
2020	the average	147,90	145,13	150,42	141,08	114,67
	the minimal	104	103	100	109	103
	the maximum	192	194	195	180	124

In the first year, students study the following disciplines: social and humanitarian training ("History and culture of Ukraine", "Foreign language", "Business Ukrainian language"); fundamental and natural science training ("Higher Mathematics", "Discrete Mathematics", "Physics"); professional training ("Algorithmization and Programming", "Introduction to Computer Science", "Computer Circuitry and Computer Architecture", "Professional and Personal Development of a Student", "Computer Design").

The indicator that characterizes the success of training is the session. At the session, a number of disciplines end with an exam, and others - with a credit. Grades per session are set on a 100-point scale, therefore, for further comparison, the results of external independent testing and the average score of the certificate are also normalized to this scale.

External independent testing in mathematics and the Ukrainian language is also required in the first year of study "Higher mathematics", "Discrete mathematics" and "Business Ukrainian language". Therefore, it is possible to assess the correlation and identify the relationship between these indicators. Table 3 shows the coefficient of correlation between the marks of external independent testing and mathematical disciplines in the first year.

Table 3

Correlation coefficient for the subject of external independent testing "Mathematics"

Year	Subject of external independent testing	Discipline for 1 course	
		"Higher mathematics"	"Discrete mathematics"
2018	Mathematics	0,76	0,65
2019	Mathematics	0,76	0,69
2020	Mathematics	0,44	0,55

It is worth noting that students with results of external independent testing less than 120 points also had a low level of results in mathematical disciplines. So, it coincides in 2018 by 100%, in 2019 by 80%. It should be noted that "Discrete Mathematics" ends with a test, so it was enough for students to score at least 60 points. This also influenced the result. A similar situation is with the disciplines "Foreign language" and "Business Ukrainian language", which also end with a test and do not require students to show all their abilities. Therefore, the correlation between these external peer test results and related disciplines is also low. But you can analyze the low performance in these disciplines and check the connection with the result of external independent testing in the relevant subjects. The analysis showed that students who either did not pass external independent testing in the subject, or had low scores (<125), had low progress in the "Foreign language" discipline. In the discipline "Business Ukrainian language" has no low academic performance. This also applies to the discipline "History and Culture of Ukraine". So, the average score in this discipline was: in 2018 - 72 points, in 2019 - 78 points, in 2020 - 77 points. Several low grades are still present in each year, but such students have problems not only with one discipline.

Therefore, it was necessary to analyze the dependence on the whole by the level of training, and not by individual disciplines. Table 4 shows the correlation coefficient between the results of the first year, the average grade of the certificate and the average score of external independent testing for each student. The correlation according to the average score is higher than with the average score of the certificate. This is due to the fact that the grades of the certificate are given in educational institutions of different levels, by different teachers and in classes of different profiles. And the general independent assessment exams are the same for everyone. Also, for the 2018 students of the introduction, the correlation of success was assessed not only for 1 course, but for success in two years of study. The results showed that the correlation of academic performance for two years with the average grade of the certificate is 0.835, and with the average score of external independent testing - 0.73. This shows that students who did well in school also do well at university, and students who did not develop their personality and diligence at school also perform poorly at university.

Table 4

Correlation matrix for first-year performance

Year	Student	Average mark of the certificate	Average mark of external independent testing
2018	Grade point average in the 1st year	0,75	0,84
2019	Grade point average in the 1st year	0,43	0,79
2020	Grade point average in the 1st year	0,66	0,77

Thus, it is possible to predict the success of first-year students based on the results of their preliminary education.

Second stage of model validation. The data for the second part were the results of the first certification, the second certification and the results of the session of students of the Faculty of Information Technologies, the Department of Computer Science of the State Higher Educational Institution "Pryazovskyi State Technical University". The model provides for an increase in the evaluation criteria for additional indicators, namely, grades for laboratory work, practical work, independent work, control work, abstracts, and the like.

For the experiment, 5 groups were taken with a total of 126 students. The initial data were the grades for the first certification, that is, laboratory, independent and control activities in the first half of the semester. The grades for the second attestation were the forecast for each discipline. At the end of the study of the subjects of the semester, the results of the second certification were compared with the forecast. The analysis of the forecast for the entire semester in the chosen discipline and the real grade in the semester was also carried out. Comparison of the results on a random sample for the subject "Introduction to Computer Science" for an example is shown in **Figure 4**. A similar prediction was made for each subject that was studied in each group.

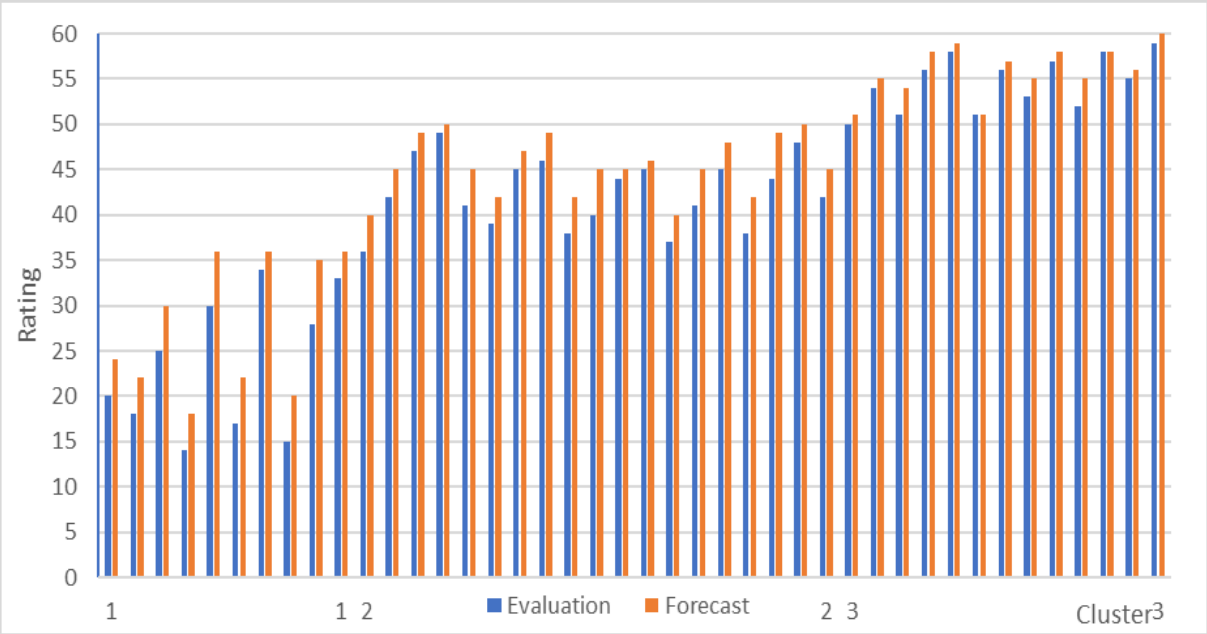

Figure 4: Comparative diagram of the results of the second certification and its forecast

Analyzing the results obtained, it can be noted that the forecast differs from the real assessment for the second certification insignificantly. Only in three cases the estimate from the forecast belongs to another cluster. All these cases are for those students who could have received a higher grade or who passed the work at the last moment, thereby increasing their final grade. In this case, the real estimate is at the junction of clusters. A similar study was carried out for all the results of the winter session of the 2020-2021 academic year.

To assess the quality of the splitting for the future forecast, the existing data was divided into additional subsets (group, subject), which made it possible to make forecasting modeling simpler. A comparison was made between the forecast results and real estimates. The quality of the partition was assessed using the mean square error value. The results showed that the error for such a volume of data is insignificant, and varies within the range (0.015; 0.03), which is a high indicator of accuracy.

The developed self-organizing Kohonen map demonstrates the results on the distribution of students according to academic performance for the future forecasting of the success of students in the entire university. In addition to predicting academic performance, a forecast of mastering the subject and assessing competencies in each subject was also carried out. For this, the forecasting of the results of the final mark obtained in the exam was carried out.

The results obtained confirm the relationship between the characteristics that influence the results of "student achievement". In addition, it is possible to determine how what result the student will demonstrate at the time of the session passing based on the certification. You can also simulate two-semester disciplines and predict learning outcomes for each course as a whole. It is also possible to carry out modeling and predict the study of disciplines that are related to each other or are a continuation of each other.

6. Conclusions

With the development of information technologies, modern computing systems are also developing, which allow not only accumulating data, but also further analyzing them, and making future forecasts. One of the tools for data analysis and forecasting is cluster analysis.

The work used cluster analysis to predict the progress of students during their studies at a higher educational institution. According to the obtained results of cluster analysis, for further predicting the success of student learning, it can be noted that the success data indicate that marks from external independent assessment can explain the variation in success in the first years and in a number of compulsory subjects. In addition, with an increase in the rate, the influence of assessments of external independent testing significantly decreases. The factors affecting student performance in information technology and math disciplines are similar, in contrast to math disciplines or information technology and humanities subjects.

The value of the root-mean-square error demonstrates the accuracy of the future forecast on the test sample, due to insignificant deviations from the real data. Analyzing the results, a regularity was revealed that the general level of passing the summer exams is higher than the winter ones. It should also be noted that the average grades in summer and winter exams are not uniform.

The results obtained confirm the adequacy of the developed model and can be used in the future to analyze and predict the progress of students of any higher educational institution in Ukraine.

References

[1] A. Nagesh, Ch. Satyamurty, Application of clustering algorithm for analysis of Student Academic Performance. International Journal of Computer Sciences and Engineering. (2018), vol. 6, pp. 381-384. doi: 10.26438/ijcse/v6i1.381384.

[2] G. Cheryl, ACT Scores, SAT Scores, and High School GPA as Predictors of Success in Online College Freshman English, (2016). Doctoral Dissertations and Projects. 1558.

[3] EM. Allensworth, K. Clark, High School GPAs and ACT Scores as Predictors of College Completion: Examining Assumptions About Consistency Across High Schools. Educational Researcher. (2020), vol. 49(3), pp. 198-211. doi:10.3102/0013189X20902110.

[4] A. A. Kotvitskaya, N. V. Zhivora, S. V. Pogorelov, I. V. Krasovsky, O. A. Vislous Study of the influence of employed coefficients on prognostic duration of the competitive balance of health development participants. Pharmaceutical Review, (2017), vol. 4, pp.129-135 doi 10.11603/2312-0967.2017.4.8341 (Ukr)

[5] V.O. Tkach, O.A. Voytovich, Prognostication of progress of students studying fundamental disciplines, depending on their intellectual capabilities, Visnyk of Kherson National Technical University, (2018), vol 2, pp. 201-205. (Ukr)

[6] O. Podolian The quality analysis of the competitive selection of abiturients in accordance with the engineering specialty in higher education, Bulletin of the cherkasy Bohdan Khmelnytsky national university. Series "Pedagogical sciences". (2018), vol. 16 pp. 23-30. doi: 10.31651/2524-2660-2018-16-23-30. (Ukr)

[7] D. Stamovlasis Bifurcation and hysteresis effects in student performance: the signature of complexity and chaos in educational research. Complicity: An Intern. J. of Complexity and Education. (2014), vol. 11, issue 2, pp. 51–64.

[8] H. Steenbeek, P. van Geert The emergence of learning-teaching trajectories in education: A complex dynamic systems approach. Nonlinear dynamics, psychology, and life sciences. (2013), vol. 17, issue 2. pp. 233–267.

[9] P. Van Geert Dynamic modeling for development and education: from concepts to numbers. Mind, Brain, and Education. (2014), vol. 8, issue 2, pp. 57–73.

[10] B. Daniel Big data and analytics in higher education: Opportunities and challenges. British j. of educational technology. (2015), vol. 46, issue 5, pp. 904–920.

[11] V. Shevchenko Prognostication of students progress on the basis of cluster analysis methods, Bulletin of Kharkiv National Automobile And Highway University, (2015), vol. 68, pp. 15 – 18. (Ukr)

[12] E. Alyahyan, D. Düştegör, Predicting academic success in higher education: literature review and best practices. International Journal of Educational Technology in Higher Education, (2020), vol. 17, article number: 3. https://doi.org/10.1186/s41239-020-0177-7.
[13] R. Asif, A. Merceron. Analyzing undergraduate students' performance using educational data mining/ Computers & Education. (2017), vol. 113, pp.177–194.
[14] T. S. Sayana, Prediction of Students Academic Performance using Data Mining: Analysis, International journal of engineering research & technology (IJERT) RTPPTDM-2015, (2015), vol. 3, iss. 30.
[15] M. H. Rahman and M. R. Islam, Predict Student's Academic Performance and Evaluate the Impact of Different Attributes on the Performance Using Data Mining Techniques, 2nd International Conference on Electrical & Electronic Engineering (ICEEE), (2017), pp. 1-4, doi: 10.1109/CEEE.2017.8412892.
[16] B. K. Francis, S. S. Babu, Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. Journal of Medical Systems, (2019), 43, 162 doi: 10.1007/s10916-019-1295-4
[17] A. Alhassan, B. Zafar A. Mueen Predict Students' Academic Performance based on their Assessment Grades and Online Activity Data International Journal of Advanced Computer Science and Applications, (2020), vol. 11, no. 4, doi :10.14569/IJACSA.2020.0110425.
[18] Križanić, Snježana Educational data mining using cluster analysis and decision tree technique: A case study, International Journal of Engineering Business Management, (2020), vol. 12, pp. 1-9 doi:10.1177/1847979020908675.
[19] A. Bogarín, R. Cerezo, C. Romero A survey on educational process mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. (2018), vol. 8, issue 1, pp. 1–17.
[20] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, S. Nam Liao. Predicting academic performance: a systematic literature review. In Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018 Companion). Association for Computing Machinery, New York, NY, USA, (2018), pp. 175–199. DOI:https://doi.org/10.1145/3293881.3295783.
[21] Loginom, URL: https://basegroup.ru/deductor/download
[22] G. Detorakis, A. Chaillet, N.P. Rougier, Stability analysis of a neural field self-organizing map. Journal of Mathematical Neuroscience, BioMed Central, (2020), vol. 10 (20), doi:10.1186/s13408-020-00097-6.
[23] Service for the search for applicants URL: https://abit-poisk.org.ua/