I.J. Modern Education and Computer Science, **2026, 5, 87-103**
Published Online on October 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2026.05.05

Specialized Software Tools for Educational Data Mining of Technical Students: Image Segmentation, Clustering, and Correlation-Regression Analysis

Oleksandr Derevyanchuk*
Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
E-mail: o.v.derevyanchuk@chnu.edu.ua
ORCID iD: https://orcid.org/0000-0002-3749-9998
*Corresponding author

Serhiy Balovsyak
Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
E-mail: s.balovsyak@chnu.edu.ua
ORCID iD: https://orcid.org/0000-0002-3253-9006

Shaocheng Qu
Department of Electronics and Information Engineering, Central China Normal University, Wuhan 430079, China
E-mail: qushaocheng@mail.ccnu.edu.cn
ORCID iD: https://orcid.org/0000-0002-4874-4186

Yurii Ushenko
Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine
E-mail: y.ushenko@chnu.edu.ua
ORCID iD: https://orcid.org/0000-0003-1767-1882

Nataliia Tytova
Dragomanov Ukrainian State University, Kyiv, 01601, Ukraine
E-mail: n.m.tytova@udu.edu.ua
ORCID iD: https://orcid.org/0000-0002-9415-4427

Hanna Kravchenko
Chernivtsi Applied College of Transport, Chernivtsi, 58000, Ukraine
E-mail: hannakravchenko81@gmail.com
ORCID iD: https://orcid.org/0009-0004-7609-0345

Received: March 25, 2026; Revised: May 10, 2026; Accepted: June 26, 2026; Published: October 8, 2026

Abstract: This article presents a set of specialized software tools for educational data mining and describes the integration of the developed tools into the educational process of students of technical specialties. The developed set consists of three programs that implement the main tasks of educational data mining. Program No. 1 provides segmentation, analysis, and visualization of images of educational materials; Program No. 2 performs data clustering; Program No. 3 performs correlation and regression analysis to identify relationships and predict learning outcomes. The developed programs can be used to implement a systematic analysis of education quality, during which automated processing of educational data and the formation of recommendations for improving the educational process are performed. The programs are implemented in Python. A scheme for integrating specialized programs into the educational process has been developed. The scheme reflects the relationships between the programs that process educational materials, analyze the results of the educational process, influence students' educational trajectories and the educational process.

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

Program No. 1, "SegmentFuzzy24," is designed for segmentation, analysis, and visualization of images of educational materials. Such processing makes it possible to highlight the investigated elements of technical devices in order to focus students' attention, as well as to determine the quantitative and geometric characteristics of multimedia presentation slides, including the number and size of symbols, the area of graphical objects, etc. Segmentation is performed using the region growing method. The determined characteristics are used for the classification of educational materials or are transferred to Programs No. 2 and No. 3 for further cluster, correlation, and regression analysis.

Program No. 2, "ClusterFuzzy23," is designed for clustering educational data, including students' learning outcomes and the characteristics of educational materials obtained from Program No. 1. Objects are divided into groups with similar characteristics using the K-means method. During the clustering process, the sum of the squared Euclidean distances between each object and the centroid of its cluster is minimized. Data clustering ensures their differentiated processing and contributes to the adjustment of educational trajectories. The clustering results can be transferred to Program No. 3 for further analysis. Fuzzy membership functions are used to determine the degree of membership of objects in overlapping clusters.

Program No. 3, "CorrelRegres25," is designed for correlation and regression analysis of educational data obtained directly or as a result of clustering in Program No. 2. Correlation analysis is used to establish relationships between the characteristics of the educational process, while regression analysis is applied to model their dependencies using polynomial functions and to predict learning outcomes. The obtained models are used to adjust educational trajectories, while objects that have significant deviations from regression dependencies are classified as outliers.

The developed programs have been demonstrated on sample data to process real educational data, namely for image segmentation and visualization, clustering of educational data, and their correlation and regression analysis.

Data clustering and correlation–regression analysis were performed using the grades of $N_0 = 76$ students in 12 subjects, assessed on a 100-point scale, for courses completed during the first and second years of study. The initial sample of N_0 students was randomly divided into a training set ($N = 60$, 80% of the data) and a validation set ($N_V = 16$, 20% of the data). The quality of the clustering results was evaluated using the Silhouette coefficient, Calinski–Harabasz index and Davies–Bouldin index. The analysis indicated that three or four clusters provide an appropriate representation of the data. The optimal polynomial degree $p_A = 3$ for the regression model was determined based on the minimum root mean square error RmseV = 12.37458, obtained on the validation set. The relatively high RmseV value can be attributed to the influence of a considerable number of factors affecting students' academic performance. The resulting regression model was used to predict students' grades in a subject for the subsequent academic period based on their grades in the same subject during the previous period. A promising direction for further improvement of the developed prediction system is the integration of artificial neural networks.

Index Terms: Educational Data Mining, Systematic Analysis of Education Quality, Prediction, Correlation, Regression, Clustering, Technical Students, Software.

1. Introduction

To ensure the quality of the modern educational process, large volumes of data are processed, which, in particular, describe students' learning conditions, their educational achievements, needs, and wishes [1]. Students' educational achievements are assessed based on the results of their current, intermediate, or final assessment. Significant volumes of data are obtained as a result of various types of testing conducted at all stages of assessment. To obtain information about students' needs and wishes, surveys are conducted during classes and through questionnaires. Ensuring the quality of the educational process is a necessary condition for a high level of professional training of technical students. A specific feature of the educational process for such students is the presentation of a significant amount of educational material not only in the form of text, but also in the form of diagrams, drawings, and photographs of devices. Therefore, in the training of technical students, there is a need to apply educational data mining (EDM) methods, which make it possible not only to automatically analyze the characteristics of the educational process and multimodal educational materials [2], but also to establish relationships between them and formulate recommendations for improving learning outcomes [3]. Important tasks of such analysis are data clustering, prediction of students' educational achievements, and identification of relationships between the characteristics of the educational process [4]. Another area of application of EDM methods is the automated preparation and analysis of educational materials [5]. Such preparation of educational materials consists, for example, in the segmentation of images of the objects under study in order to focus students' attention on them. The analysis of educational materials, for example, presentations, makes it possible to investigate the influence of their characteristics on students' perception of such materials.

However, when EDM methods are implemented in the educational process of technical students, there is a need to use appropriate software tools. Many developed software tools, such as Orange Data Mining [6] and MATLAB [7], implement data mining methods. The problem is that existing programs are generally quite universal and do not take

into account the specifics of processing educational data for technical students. In addition, in the software tools considered, data mining methods are implemented separately, while approaches to the integrated processing of data using different methods have not been sufficiently developed. For example, in MATLAB, clustering, regression, image segmentation and fuzzy logic are implemented as separate methods within the corresponding toolboxes.

Therefore, the topic of this study, which involves the development of specialized software tools for EDM methods and their integration into the educational process of technical students, is relevant. The developed software package, which contains three programs, implements the most common EDM tasks [8]. Program No. 1 implements image segmentation, analysis, and visualization. Program No. 2 is intended for clustering educational data. Program No. 3 performs regression and correlation analysis of data, which is used to predict educational achievements and establish relationships between the parameters of the educational process. The integrated use of the developed programs makes it possible to automatically analyze all available educational data and formulate recommendations based on them to improve the quality of the educational process.

1.1. Outline

A review of works devoted to educational data mining (EDM) methods and their implementation in the educational process is presented in Section 2, which also describes general issues of systematic analysis of education quality, the principles and practical application of EDM methods. The methodology of the study is presented in Section 3, which describes the scheme for integrating specialized software tools into the educational process and the principles of operation of the EDM methods used. Section 4 describes the software implementation of EDM methods. An example of processing educational data using the developed specialized software tools is presented in Section 5. Section 6 describes the conclusions of the study and directions for further research.

2. Literature Review

An important direction for optimizing the educational process is the systematic analysis of education quality, which provides a comprehensive approach to evaluating and studying the quality of the educational process [1]. The main purpose of systematic analysis is to identify relationships between the components of the educational process, determine the factors that influence learning outcomes, and develop measures to improve the quality of education. For this purpose, educational data mining (EDM) methods, in particular, are used.

Currently, there are several classifications of EDM methods. According to the well-known classification [4], the following groups of methods are distinguished:

1. Prediction.
2. Clustering.
3. Relationship mining.
4. Distillation of data for human judgment.
5. Discovery with models.

The group of EDM methods “Discovery with models” is the most complex and is based on the methods of the previous groups. Other classifications of EDM methods [8, 9] supplement the classical classification [4]. In particular, the study [9] considers such modern EDM methods as “Text Mining” and “Social Networks Analysis,” as well as machine learning methods (Decision Trees, Neural Networks, etc.). Each of the described groups of methods has a certain specialization and scope of application. Accordingly, the joint application of several groups of methods provides an in-depth analysis of data and a wide scope of application, which contributes to the successful solution of applied tasks of educational data processing. Therefore, at the present stage, complete software processing of educational data should use at least groups No. 1–No. 4 of EDM methods. Let us consider these groups of methods in more detail.

Prediction methods make it possible to predict students’ academic performance in certain subjects and other characteristics of the educational process [10, 11]. This is useful for adjusting students’ educational trajectories and identifying possible learning difficulties. Common prediction methods include regression analysis [12] and classification [13, 14]. In complex cases, prediction is performed using artificial neural networks (ANN) [15, 16] and machine learning methods [17, 18]. The effectiveness of using ANN to predict the results of the educational process is explained by the fact that such results are often influenced by a significant number of factors that are difficult to take into account using analytical methods. However, ANNs require training on specially selected datasets. Therefore, prediction of educational data is often performed using the classical regression method [12].

In study [19], student performance prediction was performed using machine learning methods and regression analysis. The results demonstrated that, under certain conditions, classical regression methods can be competitive with artificial intelligence-based methods. Study [20] focused on predicting students’ academic achievement using explainable artificial intelligence methods. The study highlighted the importance of assessing the significance of the factors that influence prediction outcomes.

Clustering methods are widely used to simplify and improve the efficiency of educational data analysis, as they make it possible to divide large arrays of source data into separate homogeneous groups (clusters) [21]. Clustering belongs to unsupervised machine learning methods that do not require the prior definition of classes of the investigated objects. The application of clustering contributes to the structuring of educational data, as well as to the identification of hidden patterns, similarities, and relationships. The results of clustering can have practical significance for analysis and decision-making in the educational process. In addition, decomposition of the overall data array into separate clusters makes it possible to carry out their further analysis step by step, which can reduce the complexity of such processing. Clustering contributes to the personalization of learning by identifying students with similar characteristics, interests, and educational needs, which makes it possible to adapt educational materials and approaches to individual groups [22]. Common clustering methods include, in particular, K-means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and hierarchical clustering [23]. The K-means method [24] is used quite widely, which is due to a number of its advantages: simplicity of software implementation, high performance even when processing large datasets, versatility, and the ability to process different types of data (educational data, image parameters, device characteristics, etc.). The obtained clusters may overlap to a significant extent, i.e., one object may partially belong to several clusters. Therefore, fuzzy logic tools are used to correctly describe such clustering results, and the membership of objects in clusters is described by fuzzy membership functions. The possibilities of fuzzy clustering are described, in particular, in study [25].

Relationship mining is intended to identify statistically significant and practically useful relationships between object characteristics [26]. In EDM, it makes it possible to determine the influence of various factors, including independent work and participation in group projects, on learning outcomes. Relationship mining is implemented, in particular, using correlation analysis and ANN. Correlation analysis makes it possible to determine the strength of the relationship between two variables, as well as to identify object characteristics that have the strongest relationship. In educational data, it can, in particular, be used to investigate the relationship between the duration of independent learning and students' academic results. Correlation analysis also makes it possible to identify pairs of academic subjects for which students' educational outcomes are significantly related.

Methods of distillation of data for human judgment are aimed at transforming complex data arrays into forms convenient for perception, interpretation, and decision-making [4, 26, 27]. For this purpose, visualization, infographics, image processing, summary tables, and dimensionality reduction methods are used. In education, they contribute to the effective analysis of large volumes of data in different formats, which is especially important for technical students. Data visualization involves presenting information using graphs, charts, images, diagrams, maps, and 3D models, which facilitates the identification of patterns, trends, and anomalies. Such processing may use image segmentation to automatically identify meaningful regions and focus students' attention on them. Visualization is important when presenting the results of modeling technical objects [28] and in the process of implementing STEM projects [29-31].

Discovery with models involves the use of created models to obtain new knowledge about learning processes, student behavior, and the effectiveness of learning strategies [4]. This direction includes modeling educational trajectories, developing adaptive systems, and predicting outcomes, in particular using clustering and prediction. Models that describe educational processes can be constructed using fuzzy cognitive maps and the Mental Modeler software package [32].

The review of current scientific works substantiates the need to integrate EDM methods into the educational process. It has been shown that prediction, correlation and regression analysis, clustering, and data visualization methods are appropriate for solving common tasks of educational data processing. The analysis of the conducted studies confirmed the need to develop specialized software tools for EDM and integrate them into the educational process.

3. Methodology

3.1. Scheme for Integrating Specialized Software Tools into the Educational Process

The integration of the developed specialized Programs No. 1–No. 3, which implement EDM methods, into the educational process is described by the following scheme (Fig. 1).

In Program No. 1, segmentation, analysis, and visualization of images obtained from educational materials are performed. The program makes it possible to process images of various devices studied by technical students. In this case, the investigated parts are identified as segments in the image of the device and are highlighted during visualization in order to focus attention on them. The obtained educational materials are used in the educational process. Program No. 1 also makes it possible to analyze images of educational materials, for example, slides of multimedia presentations. After segmentation of such images, their main characteristics are determined: the number of symbols, their size, the area of figures, etc. The obtained characteristics of images of educational materials can be analyzed directly in Program No. 1 (for example, for classifying images by the number of symbols) or transferred to Program No. 2 for cluster analysis and to Program No. 3 for correlation and regression analysis. Data exchange between the developed Programs No. 1–3 is implemented using a common data format based on text files. The first line of each file contains the parameter names, while the subsequent lines contain the corresponding parameter values; each column represents a

specific parameter.

Program No. 2 is designed for clustering educational data: the results of the educational process or the characteristics of educational materials obtained from Program No. 1. In particular, students' final grades in certain subjects are used as the results of the educational process. As a result of clustering, all investigated objects are divided into clusters, i.e., groups of objects with similar parameters. For example, students with high grades in certain subjects can be identified as a cluster. An important advantage of clustering is the possibility of targeted processing of the data of each cluster. For example, students who belong to a cluster with high grades are offered tasks of a high level of difficulty. Thus, based on the clustering results, students' educational trajectories are adjusted.

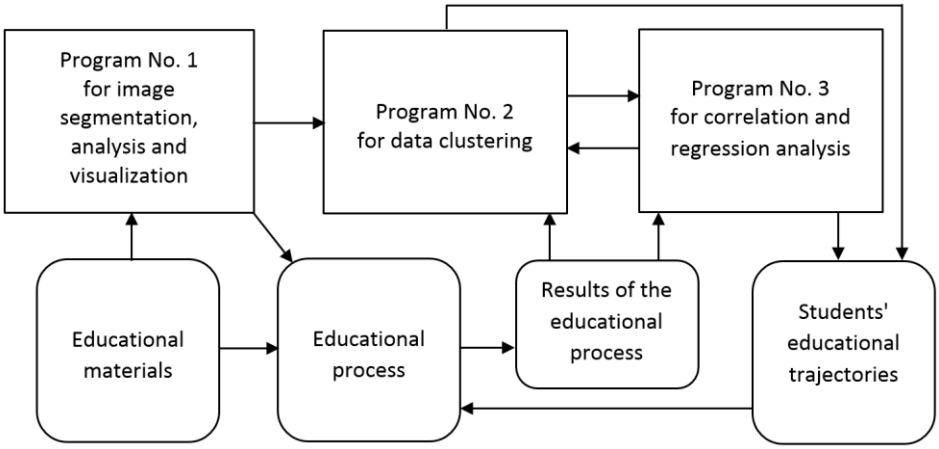


Fig. 1. Structural Scheme for Integrating Specialized Educational Data Mining Software Tools into the Educational Process.

The obtained clusters can also be transferred to Program No. 3. In the case of overlapping clusters, fuzzy membership functions are used to describe the degree of membership of objects in such clusters.

In Program No. 3, correlation and regression analysis of the results of the educational process is performed either directly or after their clustering in Program No. 2. Correlation analysis is used to establish relationships between the characteristics of the educational process; for example, pairs of subjects with the maximum correlation coefficients for students' grades are determined. The obtained pairs of parameters (subjects) can be transferred to Program No. 2 for cluster analysis. This is particularly useful when analyzing a large number of parameters, since the clustering method employed is designed to process only two parameters, represented by the x and y coordinates of an object. As a result of regression analysis of the data, dependencies of one characteristic of the educational process on another are determined. Thus, it is possible to establish the dependence of final grades in a subject in the 3rd semester on grades in another subject in the 1st semester. Based on the obtained regression dependencies, the characteristics of the educational process are predicted, which is used to adjust students' educational trajectories. Objects that deviate significantly from the obtained regression dependencies are processed as outliers.

Thus, the developed specialized programs provide the implementation of the main groups of EDM methods [4]:

1. From the group of methods "Prediction," the regression analysis method (Regression) is implemented in Program No. 3.
2. From the group of methods "Clustering," the "Clustering data" method is implemented in Program No. 2.
3. From the group of methods "Relationship mining," the "Correlation mining" method is implemented in Program No. 3.
4. From the group of methods "Distillation of data for human judgment," the data visualization method is implemented in Program No. 1.

3.2. Fundamentals of Image Segmentation and Visualization Methods

Image segmentation consists in the spatial extraction of meaningful regions in images (for example, people, symbols, or device parts) that have similar (or identical) brightness, color, or texture [33]. Program No. 1 implements image segmentation using the classical region growing method. This segmentation method is one of the simplest, so the developed program can be used to study the principles of image segmentation. The region growing method is effective when the variability of pixel brightness or color within a segment is insignificant. This is characteristic of many images contained in educational materials (for example, images of text, diagrams, and drawings).

The principle of operation of the region growing method consists in selecting the initial pixels of segments and adding new pixels with close brightness and/or color values to them. The initial data for segmentation are digital color image $f_{RGBB}(i, k, c)$, where $i = 0, ..., M_B-1$; $k = 0, ..., N_B-1$; c = 0,..., Q_C-1; M_B is the image height (in pixels); N_B is the image width (in pixels); $Q_C = 3$ is the number of color channels (Red, Green, Blue).

The developed program makes it possible to perform image segmentation in the mode of selecting the investigated objects by color and in the mode of segmenting all objects. In the object selection by color mode, based on the initial color image f_{RGBB} a grayscale image f_P is calculated, whose pixel values indicate the degree of similarity of the corresponding pixels to the color reference f_{RGBE} with the mean values of the color channels $C_E(c)$, where $c = 0,..., Q_C-1$ (in the RGB color model). The color reference f_{RGBE} ($M_R \times N_R$ pixels) is selected as a rectangular image fragment of the minimum size that contains the investigated object.

The pixel values of the image f_P are calculated using the empirical formula:

$$f_P(i,k) = 1 - \sqrt{\frac{1}{3}\sum_{c=0}^{QC-1} \left| f_{RGBB}(i,k,c) - C_E(c) \right|} \quad (1)$$

where $i = 0, ..., M_B-1$ is the pixel row index; $k = 0, ..., N_B-1$ is the pixel column index; c is the color channel index.

The application of formula (1) makes it possible to separate pixels whose color is similar to the reference from other (background) pixels. Thus, in the image f_P, the maximum values correspond to the pixels whose color is most similar to the reference color.

In the mode of segmenting all objects in the image (which do not belong to the background), the image f_P is obtained from the color image by converting f_{RGBB} to grayscale (with equal weighting coefficients for all color channels).

Based on the image f_P, a binarized image f_B is calculated with a specified binarization threshold *LimB*, which is calculated using the Otsu method. After that, segmentation of the binarized image f_B is performed using the region growing method, as a result of which the segment image f_{S1} is calculated. The value of each pixel of the image f_{S1} is equal to the number of the segment to which it belongs. In the obtained image f_{S1}, segments are selected by height, width, and area, which makes it possible to identify only the investigated objects. The ranges of permissible values for segment dimensions and/or areas are specified by the user.

3.3. Fundamentals of the Data Clustering Method Using Fuzzy Logic

Data clustering consists in dividing the initial large set of objects into subsets (clusters). Such division makes it possible to structure the data and process the objects of individual clusters taking into account their specific features. In EDM, participants in the educational process are often considered as clustering objects.

Clustering of educational data is implemented in the "ClusterFuzzy23" program using the common K-means method (K-means clustering), which is characterized by simplicity of software implementation and low computational complexity [21, 22]. The K-means method consists in dividing a set of N objects into k clusters. For the clustering task, each object is described by two parameters x and y, which are considered as its coordinates. For example, a student can be characterized by his or her grades in two subjects. During the clustering process, the value of the metric J is minimized. This metric is defined as the sum of the squared Euclidean distances d (along the x and y coordinates) between each object and the centroid of its cluster.

If the clusters overlap significantly, the problem of separating objects at the cluster boundaries arises. The membership of such objects in certain clusters is described by fuzzy membership functions. The values of the fuzzy membership function $\mu_A(x)$ of an element x in a set A show the degree to which the element belongs to the set. A specific feature of the function $\mu_A(x)$ is that it can take any value from 0 to 1. For example, if $\mu_A(x) = 1$, then the element x clearly belongs to the set A; if $\mu_A(x) = 0$, then the element x clearly does not belong to A. When $0 < \mu_A(x) < 1$, the element belongs to the set fuzzily. The larger the value of the membership function, the greater the degree to which the element belongs to the set.

The fuzzy triangular membership function $\mu_d(d)$, which represents the degree of membership of objects in a particular cluster, is defined by the following equation:

$$\mu_d(d) = 1 - \frac{d}{R_c} \quad (2)$$

where d is the distance from the object to the center of cluster Cs; R_c is the cluster radius.

3.4. Fundamentals of Correlation and Regression Analysis

Correlation and regression analysis is implemented in Program No. 3, "CorrelRegres25." Correlation analysis of data consists in calculating the linear relationship between two characteristics of the educational process, which are mathematically described as parameters x and y, respectively [12]. For example, parameter x describes students' grades in a certain subject in the 1st semester, while parameter y describes their grades in another subject in the 4th semester.

From the values of parameters x and y, a training sample of N elements is formed using random sampling and written into one-dimensional arrays $X = (x_i)$ and $Y = (y_i)$, respectively, where $i = 0,..., N-1$. Similarly, a validation sample of N_V elements is formed using random sampling and written into one-dimensional arrays X_V and Y_V. The training sample is intended to establish the relationship between the parameters, while the validation sample is used to verify the

results of correlation and regression analysis.

The relationship between parameters *x* and *y* is quantitatively characterized by the Pearson correlation coefficient *Corr*. Its use makes it possible to assess the degree of relationship between learning outcomes in different subjects and to identify pairs of parameters for which the strongest correlation is observed. The value of the Pearson correlation coefficient *Corr* is determined from the training sample data according to the formula:

$$Corr = \frac{\sum_{i=0}^{N-1}(x_i - X_C)(y_i - Y_C)}{\sqrt{\sum_{i=0}^{N-1}(x_i - X_C)^2} \cdot \sqrt{\sum_{i=0}^{N-1}(y_i - Y_C)^2}} \quad (3)$$

where

x_i, y_i are the values with index *i* for parameters *x and y, respectively;*

N is the number of parameter values;

X_C, Y_C are the mean values of parameters *x* and *y*.

The values of the correlation coefficient *Corr* belong to the interval from −1 to +1. *Corr* values close to +1 characterize a strong positive correlation, in which an increase in one parameter is accompanied by an increase in the other. In contrast, values close to −1 indicate a strong negative correlation, when an increase in one parameter is accompanied by a decrease in the other. *Corr* values close to zero indicate the absence of a pronounced linear relationship between the investigated parameters. If more than two parameters of the investigated objects are analyzed (e.g., students' grades in several subjects), the obtained correlation coefficient values *Corr* are used to make an informed selection of parameter pairs with high *Corr* values.

The dependence of parameter *y* on *x* can be represented by the regression equation $y_A(x)$, which is approximated by a polynomial of degree *p* according to the formula:

$$y_A(x) = a_0 + \sum_{n=1}^{p} a_n \cdot x^n \quad (4)$$

The approximated values y_A are stored in the arrays Y_{AN}. The coefficients of the polynomial model $a_0,\ldots,a_p$ are determined by the least squares method by minimizing the difference between the original values y_i and the corresponding approximated values y_A. The approximation accuracy is evaluated using the Root Mean Square Error (Rmse), calculated for the training sample data, and the error RmseV, determined for the validation sample.

The constructed regression model can be used to diagnose students' educational achievements using the "outlier detection" method. In this case, an outlier is considered to be a result in which a student's actual grade in a certain academic discipline significantly deviates from the value predicted by the regression model. By default, the threshold value for identifying outliers is set to twice the standard deviation. Detection of such deviations makes it possible to identify students whose learning outcomes require additional analysis.

4. Software Implementation of the Methods

The computer programs "SegmentFuzzy24," "ClusterFuzzy23," and "CorrelRegres25," which implement the main groups of EDM methods, have been developed. The programs are implemented in Python (Python version 3.11.5), which has relatively low minimum hardware and software requirements for a computer (Windows 8.1, a processor with a frequency of at least 1 GHz, 4–8 GB RAM).

The "SegmentFuzzy24" program is designed for segmentation, analysis, and visualization of images used in the educational process. Image reading, visualization, and saving are performed using the methods of the "matplotlib" library. The initial color images f_{RGBB} are read from graphic files. In order to reduce the influence of noise and extraneous objects on the segmentation result, the images f_P are smoothed using a Gaussian filter implemented by the "gaussian_filter" method of the "ndimage" library. The standard deviation of the Gaussian filter kernel σ_G (e.g., $\sigma_G = 8$ pixels) was selected such that, after smoothing, the object under study exhibited a uniform color and was identified as a single segment. In the software implementation, segmentation of the binarized images f_B is performed using the "label" method from the "ndimage" library. By default, a 4-connected neighborhood is used for segmentation.

The "ClusterFuzzy23" program is used for clustering educational data using fuzzy logic. Programmatically, clustering is performed using the "kmeans" method of the "sklearn" library. As a result of clustering, the following arrays are calculated:

- *Nc* – cluster numbers for the investigated objects;
- *Rc* – cluster radii;
- *Cs* – coordinates of cluster centers (calculated as the coordinates of their centers of gravity);
- *Cq* – the number of points in the clusters.

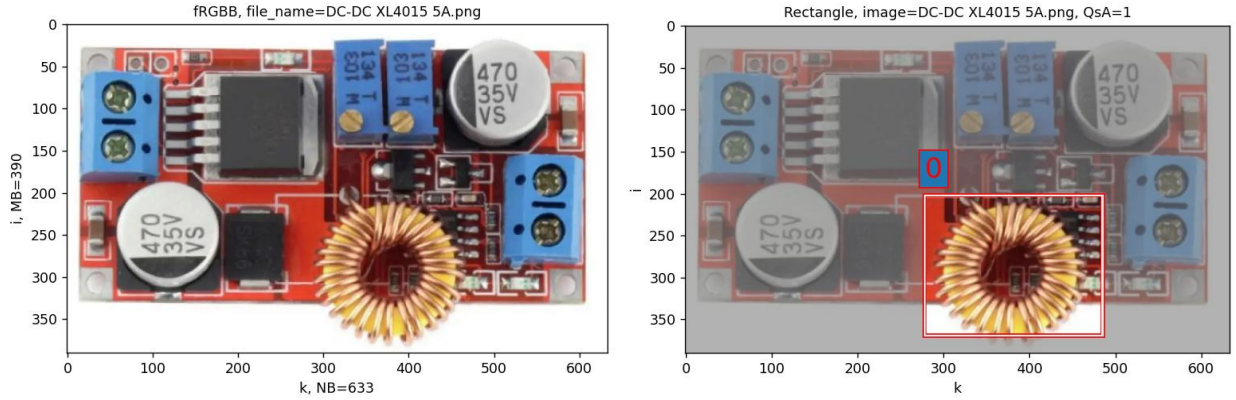
The clustering quality metrics, including the Silhouette coefficient, Calinski–Harabasz index and Davies–Bouldin index, are computed using the methods provided by the "sklearn" library.

The "CorrelRegres25" program is designed for correlation and regression analysis of educational data. The calculation of the correlation coefficient *Corr* between the data stored in the arrays *X* and *Y* is performed using the corrcoef(X, Y) function of the "numpy" library. The calculation of the coefficients of a polynomial of degree *p* is performed using the polyfit(X,Y,p) function of the "numpy" library.

5. Results and discussion

5.1. Results of Image Segmentation and Visualization

Using Program No. 1, "SegmentFuzzy24," test images were segmented by the region growing method. The initial images can be read both from graphic files and from video cameras [34]. In the considered example, the segmentation task was to identify the investigated element (the power inductor) in a series of device images (Fig. 2a). The color reference was obtained as a fragment of the image of the power inductor in one of the device images. Therefore, regions that are most similar in color to the reference are highlighted as segments in the image. For further processing, only segments whose sizes are within the allowable range are used. If the image contains different elements with a similar color, then, if necessary, segment selection by their numbers is performed. During visualization, the obtained segments are bounded by frames with corresponding labels (Fig. 2b).



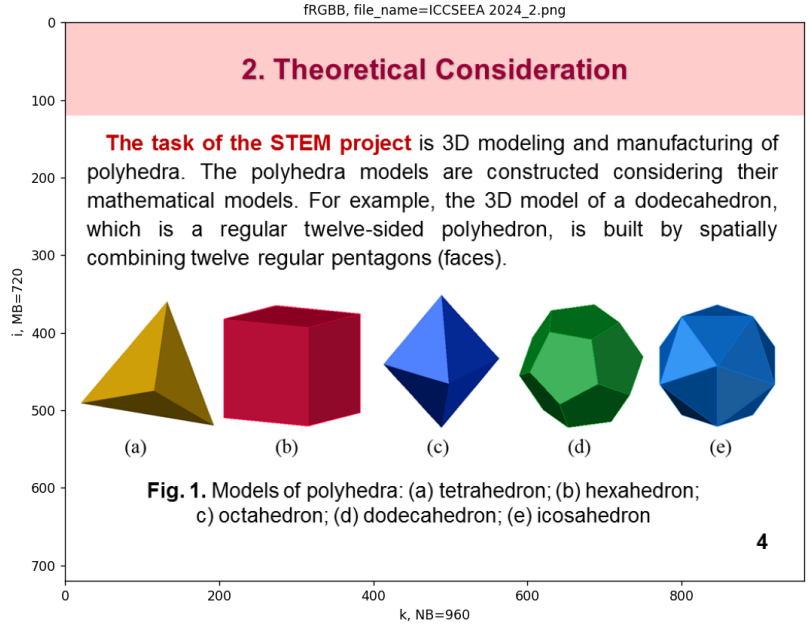
(a) (b)

Fig. 2. Initial image of the XL4015 DC-DC 5A step-down voltage converter (a) and the result of selecting the investigated element (the power inductor) (b); total number of segments QsA; segment numbering starts from 0.

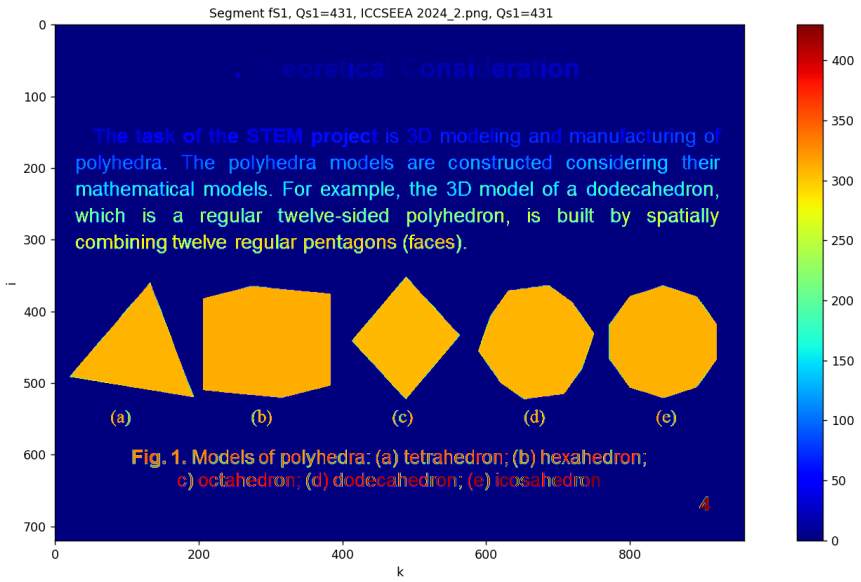
The image contrast outside the frames of the investigated objects is also reduced, while the contrast of the selected regions is increased [35]. Contrast changes are controlled visually or on the basis of image profiles [36]. As a result, this makes it possible to focus students' attention specifically on the required objects.

The "SegmentFuzzy24" program makes it possible, if necessary, to describe the membership of an image segment in the investigated object using fuzzy membership functions. The use of fuzzy logic is appropriate if the image contains several regions similar to the investigated object.

The determination of the characteristics of images of educational materials will be considered using the example of processing an image of a presentation slide (Fig. 3) with the developed program by segmenting all objects. After segmentation of such images, their main characteristics are determined: the number of symbols, their size, the area of figures, etc. For example, segments whose height and width are within certain ranges are identified as symbols (unlike figures). The obtained characteristics of images of educational materials can be analyzed directly in Program No. 1 (for example, for classifying images by the number of symbols) or transferred for processing by Programs No. 2 and No. 3.



(a)



(b)

Fig. 3. Initial image of a presentation slide (a) and the result of its segmentation (b); Qs1 is the number of segments; each segment has a color corresponding to its number in the color palette.

5.2. Data Clustering Results

Let us consider an example of data clustering using Program No. 2, "ClusterFuzzy23" by the K-means method. The initial educational data are read from a text file, in which each row contains the data of a certain student, and each column represents his or her grades in the corresponding subject (on a 100-point grading scale). Clustering is performed in the feature space *x*, *y*. In the considered case, students' grades in the subject "Computer_Graphics" are used as parameter *x*, while their grades in the subject "Web_Technologies_and_Web_Design" are used as parameter *y*. The number of clusters is set by the user as the value of the variable *k* (Fig. 4). The value of the student number i_v, whose data will be analyzed in detail, is also set. After clustering, the cluster boundaries are highlighted by ellipses (if necessary, the cluster boundaries can be highlighted by arcs for each quadrant separately, which is appropriate for clusters with an asymmetric shape). A quantitative assessment of clustering quality was performed using the following metrics: the sum of squared distances between objects and the centers of their respective clusters (*J*), the Silhouette coefficient, the Calinski–Harabasz index and the Davies–Bouldin index (Table 1). For each metric, the mean value and standard deviation (SD) were calculated based on the clustering results. To obtain reliable metric estimates, the clustering procedure was repeated $Q_l = 30$ times.

Table 1. Clustering Metrics.

#	Metrics	k=3		k=4		k=5	
		Mean	SD	Mean	SD	Mean	SD
1	J	8868.64	1.13	6779.28	14.53	5180.73	27.20
2	Silhouette	0.4443	0.0008	0.4404	0.0017	0.4063	0.0125
3	Calinski–Harabasz	128.92	0.02	118.30	0.30	119.97	0.71
4	Davies-Bouldin	0.8712	0.0031	0.8001	0.0103	0.8499	0.0064

The decrease in the J metric with an increasing number of clusters k is expected, as the objects become closer to the centers of their respective clusters. For $k = 3$ the SD of the J metric is negligible (0.013% relative to the mean value), whereas for $k = 4$ it increases by an order of magnitude. The SD values are also negligible for all the metrics considered relative to their corresponding mean values, indicating high stability of the clustering results across different clustering runs. The Silhouette coefficient ranges from −1 to 1, with higher values indicating better separation of the data. The obtained Silhouette coefficient values indicate a distinct separation of the data into clusters, although some overlap between the clusters cannot be excluded. The Calinski–Harabasz index is defined as the ratio of between-cluster dispersion to within-cluster dispersion; therefore, a higher index value indicates better clustering quality. For the Davies–Bouldin index, lower values indicate better separation of the data into clusters. Thus, according to the Silhouette coefficient and Calinski–Harabasz index, the best clustering result is obtained for $k = 3$, whereas according to the Davies–Bouldin index the best clustering result is obtained for $k = 4$. Therefore, clustering with $k = 3$ or $k = 4$ clusters is considered appropriate.

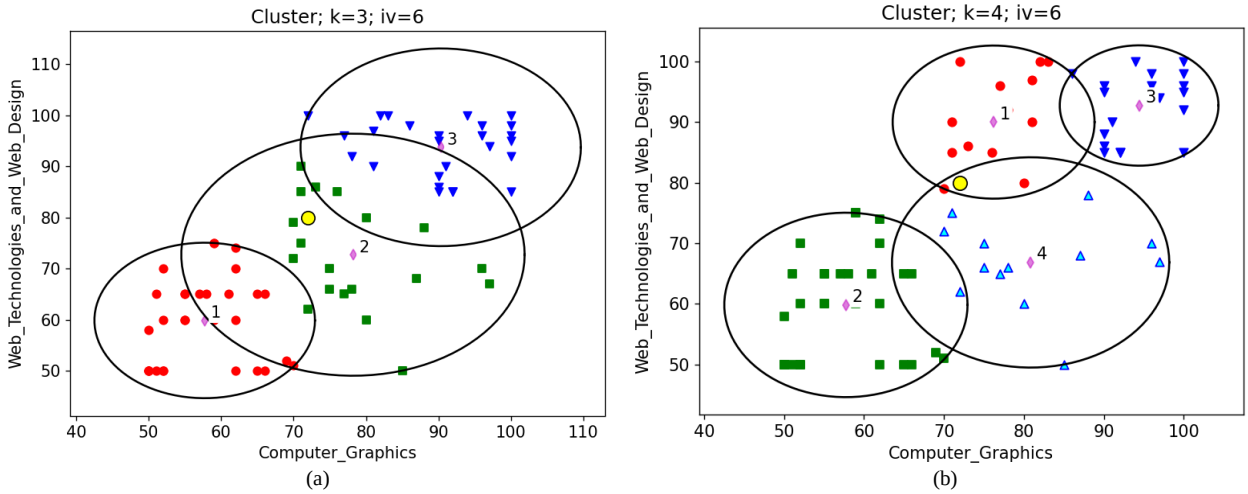


Fig. 4. Clustering of data in the space of the parameters (grades) “Computer_Graphics” (x-axis) and “Web_Technologies_and_Web_Design” (y-axis) with the number of clusters $k = 3$ (a) and $k = 4$ (b); the student with number i_v is marked with a marker (yellow circle).

Fuzzy clustering of the data of the student with number i_v is performed as follows. For each cluster with number n_k, a fuzzy triangular membership function $\mu_d(d)$ is calculated, which indicates the degree of membership of objects in this cluster depending on their distance d from the cluster center with radius Rc (Fig. 5a). The distance of the student with number i_v from the cluster center is equal to d_0. Accordingly, the student belongs to a certain cluster with the membership degree $\mu_d(d_0)$. In the considered case (Fig. 4b, Fig. 5), the student with number $i_v = 6$ belongs more strongly to cluster No. 1 and partially belongs to cluster No. 4.

Students belonging to each cluster have certain common characteristics. For example, cluster No. 1 (Fig. 4b) includes students with high grades in “Web_Technologies_and_Web_Design” and average grades in “Computer_Graphics”. The clustering results can be applied in the educational process, in particular, to determine the level of task difficulty for different students and to divide students into teams during the implementation of STEM projects. The obtained clusters can also be transferred for further processing in Program No. 3.

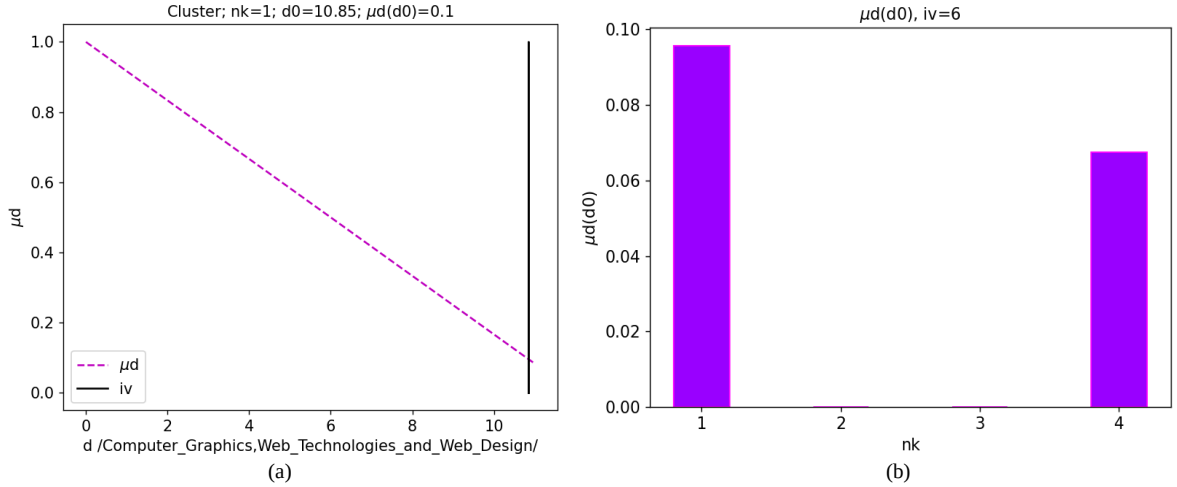


Fig. 5. Fuzzy membership functions for the clustering results (Fig. 4b): a) fuzzy triangular membership function $\mu_d(d)$ for cluster No. *nk = 1 with a radius Rc = 12.056*; b) values of the fuzzy membership $\mu_d(d_0)$ of the student with number $i_v = 6$ in clusters with numbers n_k

5.3. Results of Correlation and Regression Analysis

Let us consider an example of correlation and regression analysis of data using Program No. 3, "CorrelRegres25." The same educational data as in Program No. 2 were analyzed, namely, students' grades in 12 subjects. Subjects No. 1-6 were taught during Semesters No. 1-2 (first year of study), while Subjects No. 7-12 were taught during Semesters No. 3-4 (second year of study). In the considered example, the data of all $N_0 = 76$ students were analyzed; however, if necessary, it is possible to analyze only the data of students belonging to a certain cluster (Fig. 4).

In order to verify the results of the regression analysis, the investigated objects were randomly divided into a training set (number of values $N = 60$, 80% of the data) and a validation set (number of values $N_V = 16$, 20% of the data). By comparing the grades in subjects taught during Semesters No. 1-2 and Semesters No. 3-4, their correlation coefficients *Corr* were determined (Fig. 6). A high *Corr* value means that high grades in one subject correspond to high grades in another subject, and vice versa. Thus, correlation analysis makes it possible to establish relationships between subjects. The obtained relationships were used for the justified selection of pairs of subjects for which regression analysis of grades is appropriate, namely, pairs of subjects with high *Corr* values. Therefore, the pair of subjects "Computer_Graphics" (Semester No. 1) and "Web_Technologies_and_Web_Design" (Semester No. 3) was selected for further regression analysis, since for the subject "Computer_Graphics" the maximum value of the correlation coefficient Corr is obtained with the subject "Web_Technologies_and_Web_Design". Accordingly, the grades in the subject "Computer_Graphics" for the previous period were processed as parameter (x), while the grades in the subject "Web_Technologies_and_Web_Design" for the subsequent period were processed as parameter (y).

The regression dependencies were described by polynomials of degree p (Fig. 7). The optimal polynomial degree pA is determined as the value of p that provides the minimum approximation error RmseV on the validation sample (Fig. 7b). If necessary, additional specified constraints may be taken into account when selecting the optimal degree, in particular, the requirement for monotonicity of the polynomial function within the investigated range. To determine pA, different values of the polynomial degree (p) within the specified range (for example, from 1 to 6) are considered sequentially, after which the value p_A corresponding to the minimum RmseV is selected.

In the case under consideration, the 95% confidence interval for the regression error RmseV = 12.48886 at a polynomial degree $p = 1$ is [7.43, 17.55], whereas that for the minimum regression error RmseV = 12.37458 at a polynomial degree $p = 3$ is [6.87, 17.88]. Thus, the reduction in the regression error RmseV when increasing the polynomial degree from $p = 1$ to $p = 3$ is insignificant, confirming that the relationships represented by the constructed regression equations are close to linear. The relatively high regression errors RmseV are attributable not only to grades obtained in previous periods, but also to a number of additional factors affecting students' grades.

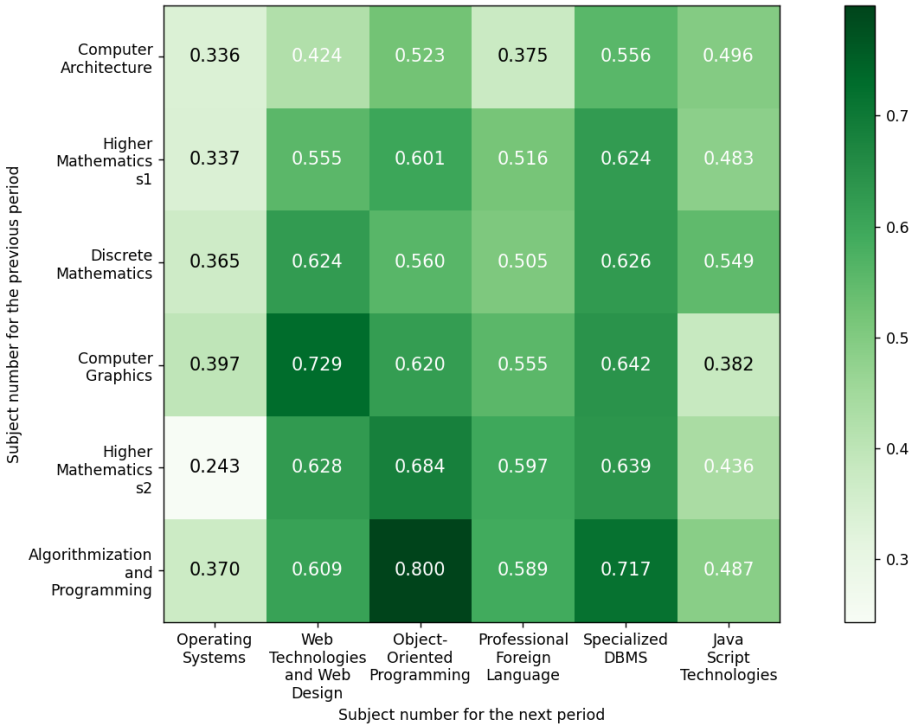


Fig. 6. Values of the correlation coefficients *Corr* between grades in subjects for the previous period (Semesters No. 1-2) and the next period (Semesters No. 3-4)

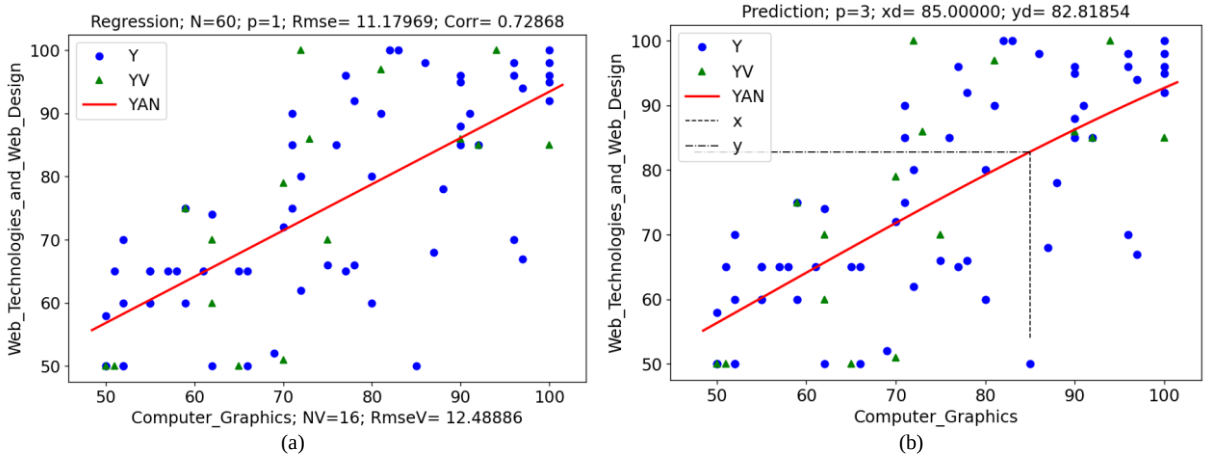


Fig. 7. Regression analysis of grades in subjects using a polynomial of degree: a) p = 1 (RmseV = 12.48886); b) p = 3 (RmseV = 12.37458); $p_A = 3$

Using the regression equation, the grade in the subject "Web_Technologies_and_Web_Design" for the subsequent period y_d was predicted based on the grade in the subject "Computer_Graphics" for the previous period x_d (Fig. 7b). Such prediction can be applied in the educational process, for example, to predict the academic performance of second-year students based on their grades during the first year of study. The predicted academic performance can also be used to recommend certain elective subjects (with the highest predicted academic performance).

If necessary, correlation and regression analyses are performed not for the entire dataset, but differentially for specific clusters identified using Program No. 2. For example, regression analysis of clusters No. 2 and No. 3 (Fig. 4a) makes it possible to obtain more accurate predictions of students' grades in subjects for the subsequent period y_d, based on their grades for the previous period x_d, specifically for students belonging to the respective clusters (Fig. 8). Correlation analysis of individual clusters also makes it possible to identify hidden and non-obvious relationships. In particular, for the entire dataset (Fig. 7a) a high positive correlation coefficient is observed (Corr = 0.72868); for cluster No. 2, the correlation coefficient is negative (*Corr* = -0.34960, Fig. 8a), whereas for cluster No. 3 it is close to zero (*Corr* = 0.07068, Fig. 8b).

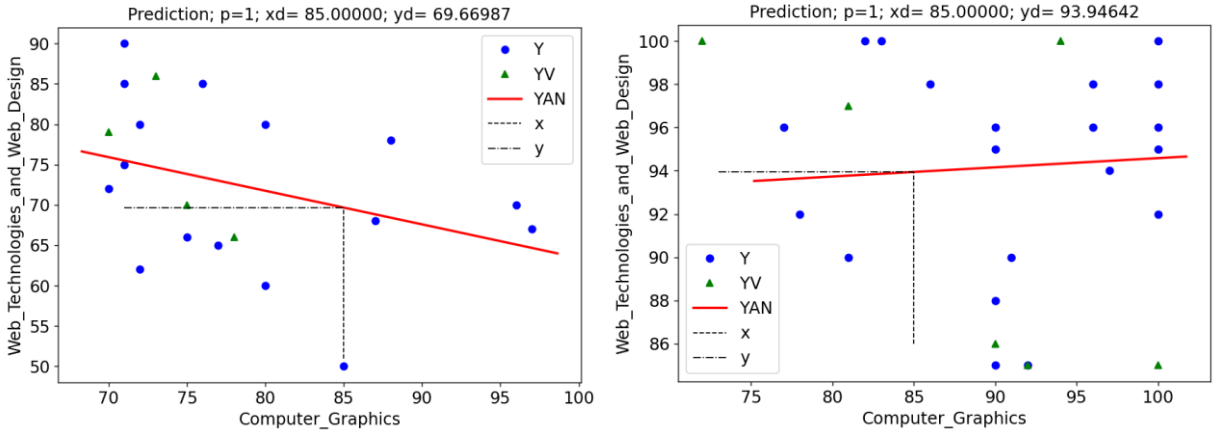


Fig. 8. Regression analysis of grades in subjects belonging to the clusters (Fig. 4a): a) No. 2 (Corr = -0.34960); b) No. 3 (Corr = 0.07068)

6. Conclusion and Future Research Directions

A set of specialized software tools for educational data mining has been developed, and the developed tools have been integrated into the educational process of technical students. The set consists of three programs that provide the implementation of the main data mining tasks. Program No. 1 is designed for image segmentation, analysis, and visualization. Program No. 2 implements clustering of educational data. Program No. 3 performs regression and correlation analyses, which are used to predict educational outcomes and identify relationships between the parameters of the educational process. The developed programs are intended for systematic analysis of education quality. Such analysis provides automated processing of available educational data and the formation of recommendations based on them aimed at improving the quality of the educational process.

The methodology of the study is described, according to which a scheme for integrating specialized software tools into the educational process was developed. The scheme shows the relationships between the developed Programs No. 1–No. 3, which process educational materials, analyze the results of the educational process, influence students' educational trajectories and the educational process. The programs are implemented in Python.

Program No. 1, "SegmentFuzzy24," provides segmentation, analysis, and visualization of images of educational materials. This makes it possible to highlight the investigated parts in images of technical devices in order to focus students' attention, as well as to determine the characteristics of multimedia presentation slides, including the number and size of symbols, the area of figures, etc. Image segmentation is performed using the classical region growing method. The obtained parameters are used for the classification of materials or are transferred to Programs No. 2 and No. 3 for cluster, correlation, and regression analysis.

Program No. 2, "ClusterFuzzy23," is designed for clustering educational data, including students' learning outcomes and the characteristics of educational materials obtained from Program No. 1. Clustering is performed using the K-means method and provides the division of objects into groups with similar parameters, which makes it possible to perform their differentiated processing and adjust educational trajectories. The formed clusters can be transferred to Program No. 3 for further analysis. The degree of membership of objects in overlapping clusters is described using fuzzy membership functions.

Program No. 3, "CorrelRegres25," implements correlation and regression analysis of educational data directly or after their clustering in Program No. 2. Correlation analysis identifies relationships between the characteristics of the educational process, while regression analysis determines dependencies in the form of polynomials and provides prediction of learning outcomes. The obtained dependencies are used to adjust educational trajectories, while objects that significantly deviate from the regression models are identified as outliers.

Testing of the developed programs showed their correct functioning in image segmentation and visualization, clustering of educational data, and their correlation and regression analysis.

The developed specialized software tools can be improved through the use of artificial intelligence tools and technologies. In particular, for accurate segmentation of complex images with significant variability in pixel brightness and color, the application of convolutional neural network models with the YOLO architecture is promising. For the recognition of symbol images, the Tesseract Optical Character Recognition (OCR) system can be used, which employs Long Short-Term Memory (LSTM) neural networks [37].

All the Declarations and Statements

Author Contributions Statement

Oleksandr Derevyanchuk – Conceptualization, Methodology, and Supervision: Proposed research ideas, Constructed the overall framework, and supervised project execution.
Serhiy Balovsyak – Data Curation and Software Implementation, and Supervision: Proposed research ideas, Handled data acquisition, dataset preprocessing, and implementation of the research model.
Shaocheng Qu – Writing – Review and Editing, and Project Management: Reviewed and edited the manuscript, and supervised project execution.
Yuriy Ushenko – Formal Analysis, Visualization, and Statistical Analysis: Performed in-depth analysis of experimental results; and helped coordinate project milestones and deadlines.
Nataliia Tytova – Writing – Review and Editing, and Project Management: Reviewed and edited the manuscript, ensured clarity and coherence.
Hanna Kravchenko – Writing – Drafted the initial manuscript, contributed to the literature survey, and documented the technical background of the study.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

This research was not supported by a grant.

Data Availability Statement

This study did not use publicly available datasets.

Ethical Declarations

Statements of ethical approval for studies involving human subjects and/or animals.

Acknowledgments

We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

Declaration of Generative AI in Scholarly Writing

https://www.mecs-press.org/ijcnis/author.html#DeclarationofgenerativeAIinscholarlywriting
AI and AI-assisted technologies were used to translate part of the text.

Abbreviations

The following abbreviations are used in this manuscript:

ANN – Artificial Neural Networks
EDM – educational data mining
LSTM – Long Short-Term Memory
OCR – Optical Character Recognition
RMSE – Root Mean Square Error
STEM – Science, Technology, Engineering, Mathematics

References

[1] A. Biloshchytskyi, S. Omirbayev, A. Mukhatayev, O. Kuchanskyi, M. Hlebena, Y. Andrashko, N. Mussabayev, and A. Faizullin, “Structural models of forming an integrated information and educational system “quality management of higher and postgraduate education,” *Frontiers in Education*, vol. 9 (1291831), pp. 1-20, 2024. doi: 10.3389/feduc.2024.1291831.
[2] M. Trifonova, and G. Kiryakova, “Multimodal Learning Resources: A Way to Engage Students’ Senses,” *Engineering Proceedings*, vol. 103, no. 1:13, article no. 13, 2025. doi: 10.3390/engproc2025103013.
[3] M. Fazlikhani, and I. Bagheri, “Educational Computer Science Development Based on Data Mining,” In *5th International Conference on Science, Engineering, and role of Technology in new Businesses*, Copenhagen, Denmark, pp. 1-8, December 16, 2023.

[4] R. Baker, and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3-17, 2009. doi: 10.5281/zenodo.3554658.
[5] S. Yi, J. Matsugami, H. Yumoto, and T. Yamasaki, "An Online Presentation Slide Assessment System Using Visual and Semantic Segmentation Features," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 13, pp.16494-16496, 2023. doi: 10.1609/aaai.v37i13.27090.
[6] *Orange Data Mining*. 2026. URL: https://orangedatamining.com.
[7] *MATLAB*. 2026. URL: https://www.mathworks.com/products/matlab.html.
[8] C. Romero, and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery*, vol. 3, no. 1, pp. 12-27, 2013. doi: 10.1002/widm.1075.
[9] E. Kalita, S. S. Oyelere, S. Gaftandzhieva, et al., "Educational data mining: a 10-year review," *Discover Computing*, vol. 28, article no. 81, 2025. doi: 10.1007/s10791-025-09589-z.
[10] S. Batool, J. Rashid, M.W. Nisar, et al., "Educational data mining to predict students' academic performance: A survey study," *Educ. Inf. Technol*, vol. 28, pp. 905-971, 2023. doi: 10.1007/s10639-022-11152-y.
[11] Zareen Alamgir, Habiba Akram, Saira Karim, and Aamir Wali, "Enhancing Student Performance Prediction via Educational Data Mining on Academic data," *Informatics in Education*, vol. 23, no. 1, pp. 1-24, 2024. doi: 10.15388/infedu.2024.04.
[12] Rolly T. Dagdagui, "Predicting Students' Academic Performance Using Regression Analysis," *American Journal of Educational Research*, vol. 10, no. 11, pp. 640-646, 2022. doi: 10.12691/education-10-11-2.
[13] D. Arie, R. Kvyetnyy, O. Bisikalo, O. Sofina, and Y. Bunyak, "Students classification basing the numerical model to compile playful teams on the Unispher™ platform," *Lecture Notes in Networks and Systems*, V. Kazymyr et al. (Eds.), Springer, Cham, vol. 1091, 2024. doi: 10.1007/978-3-031-67348-1_16.
[14] M. Talakh, V. Dvorzhak, M. Maksymovych, and Y. Ushenko, "Evaluating trade-offs between data compression and information preservation in classification tasks," In *Artificial Intelligence, Medical Engineering and Education: Proceedings of the 8th International Conference (AIMEE 2024)*, Huangshi, China, pp. 244-251, October 26-27, 2024. IOS Press, 2025. doi: 10.3233/ATDE250132.
[15] N, Ahmad, N. Hassan, H. Jaafar, and N.I.M. Enzai, "Students' Performance Prediction using Artificial Neural Network," *IOP Conf. Ser.: Mater. Sci. Eng*., vol. 1176, pp.1-8, 2021. doi: 10.1088/1757-899X/1176/1/012020.
[16] C. F. Rodriuez-Hernandez, M. Musso, E. Kyndt, and E. Cascallar, "Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation," *Computers and Education: Artificial Intelligence*, vol. 2, pp. 1-14, 2021. doi: 10.1016/j.caeai.2021.100018.
[17] S. Sarsa, J. Leinonen, and A. Hellas, "Empirical Evaluation of Deep Learning Models for Knowledge Tracing: Of Hyperparameters and Metrics on Performance and Replicability," *Journal of Educational Data Mining*, vol. 14, no. 2, 2022. doi: 10.5281/zenodo.7086179.
[18] P. Nayak, S. Vaheed, S. Gupta, et al., "Predicting students' academic performance by mining the educational data through machine learning-based classification model," *Educ. Inf. Technol*, vol. 28, pp. 14611-14637, 2023. doi: 10.1007/s10639-023-11706-8.
[19] B. Sonnleitner, T. Madou, M. Deceuninck, F. Theodosiou, and Y.R. Sagaert, "Evaluation of early student performance prediction given concept drift," *Computers and Education: Artificial Intelligence*, vol. 8, article no. 100369, 2025. doi: 10.1016/j.caeai.2025.100369.
[20] E. Tiukhova, P. Vemuri, N.L. Flores, A.S. Islind, M. Oskarsdottir, S. Poelmans, B. Baesens, and M. Snoeck, "Explainable Learning Analytics: Assessing the stability of student success prediction models by means of explainable AI," *Decision Support Systems*, vol. 182, article no. 114229, 2024. doi: 10.1016/j.dss.2024.114229.
[21] T. Omar, A. Alzahrani, M. Zohdy, "Clustering approach for analyzing the student's efficiency and performance based on data," *Journal of Data Analysis and Information Processing*, vol. 8, no. 03, pp. 171-182, 2020. doi: 10.4236/jdaip.2020.83010.
[22] M. Nafuri, A. Fikri, N.S. Sani, N.F.A. Zainudin, A.H.A. Rahman, and M. Aliff, "Clustering Analysis for Classifying Student Academic Performance in Higher Education," *Applied Sciences*, vol. 12, no. 19, pp. 9467, 2022. doi: 10.3390/app12199467.
[23] M. Ali, "Clustering in machine learning: 5 essential clustering algorithms," *DataCamp*. URL: https://www.datacamp.com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms.
[24] M. Wang, and X. Wei, "Research on Logistics Center Location-Allocation Problem Based on Two-Stage K-Means Algorithms," *Advances in Computer Science for Engineering and Education*, Hu, Z., Petoukhov, S., Dychka, I., He, M. (Eds.), Springer International Publishing, vol. 1247, pp. 52-62, 2021. doi: 10.1007/978-3-030-55506-1_5.
[25] A. Heni, I. Jdey, and H. Ltifi, "K-means and fuzzy c-means fusion for object clustering," In *8th International Conference on Control, Decision and Information Technologies (CoDIT)*, Istanbul, Turkey, pp. 177-182, 2022. doi: 10.1109/CoDIT55151.2022.9804078.
[26] A. Algarni, "Data Mining in Education," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, no. 6, pp. 456-461, 2016.
[27] O. Chaban, E. Manziuk, and P. Radiuk, "Method of adaptive knowledge distillation from multi-teacher to student deep learning models," *Journal of Edge Computing*, vol. 4, no. 2, pp. 159-178, 2025. doi: 10.55056/jec.978.
[28] V. Ravluk, Y. Derevianchuk, O. Derevyanchuk, A. Krychun, and H. Kravchenko, "Investigation of the statistical data on the technical condition of brake equipment components of passenger carriages in operation," *Edelweiss Applied Science and Technology*, vol. 8, no. 6, pp. 5957-5970, 2024. doi: 10.55214/25768484.v8i6.3292.
[29] O. Derevyanchuk, V. Kovalchuk, S. Balovsyak, M. Dominikov, H. Kravchenko, O. Pshenychnyi, and A. Kovalchuk, "Implementation of the STEM Project "Modeling of Spatial Images of Polyhedra" in the Professional Training of Future Specialists in Engineering and Pedagogical Specialties," in: Hu Z., Yanovsky F., Dychka I., He M. (eds), *Advances in Computer Science for Engineering and Education VII. ICCSEEA 2024. Lecture Notes on Data Engineering and Communications Technologies*. Springer, Cham, vol. 242, pp. 683-692, 2025. doi: 10.1007/978-3-031-84228-3_59.

[30] O.V. Derevyanchuk, V.I. Kovalchuk, V.M. Kramar, H.O. Kravchenko, D.V. Kondryuk, A.V. Kovalchuk, and B.V.Onufriichuk, "Implementation of STEM education in the process of training of future specialists of engineering and pedagogical specialties," *Proceedings of SPIE*, vol. 12938, pp. 214-217, 2024. doi: 10.1117/12.3012996.
[31] V. Kovalchuk, A. Androsenko, O. Derevyanchuk, N. Volkova, and Y. Piven, "Development of pedagogical skills of students of technology and pedagogical specialties using STEM technologies," *Edelweiss Applied Science and Technology*, vol. 8, no. 4, pp. 498-506, 2024. https://doi.org/10.55214/25768484.v8i4.1125.
[32] *Mentalmodeler*. URL: https://www.mentalmodeler.com.
[33] O. Berezsky, M. Zarichnyi, and O. Pitsun, "Development of a metric and the methods for quantitative estimation of the segmentation of biomedical images," *Eastern-European Journal of Enterprise Technologies*, vol. 6, no. 4(90), pp. 4-11, 2017. doi: 10.15587/1729-4061.2017.119493.
[34] S. Balovsyak, Kh. Odaiska, O. Yakovenko, and I. Iakovlieva, "Adjusting the Brightness and Contrast parameters of digital video cameras using artificial neural networks," *Proceedings of SPIE*, vol. 12938. pp. 129380I-1 - 129380I-4, 2024. doi: 10.1117/12.3009429.
[35] S.V. Balovsyak, O.V. Derevyanchuk, H.O. Kravchenko, O.P. Kroitor, and V.V. Tomash, "Computer system for increasing the local contrast of railway transport images," *Proc. SPIE, Fifteenth International Conference on Correlation Optics*, vol. 12126, pp. 121261E1-7, 2021. doi: 10.1117/12.2615761.
[36] S.V. Balovsyak, O.V. Derevyanchuk, and I.M. Fodchuk, "Method of calculation of averaged digital image profiles by envelopes as the conic sections," *Advances in Intelligent Systems and Computing*, Hu Z., Petoukhov S., Dychka I., He M. (Eds.), Springer International Publishing, vol. 754, pp. 204-212, 2019. doi: 10.1007/978-3-319-91008-6_21.
[37] *Tesseract_OCR*. URL: https://tesseractocr.org.

Authors' Profiles



Oleksandr Derevyanchuk: Received the master of engineering degree (1999) and Ph.D. of Physics and Mathematics (2014) at the Yuriy Fedkovych Chernivtsi National University. He is a Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Professional and Technological Education and General Physics, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine.
Research Interests: Education Technology, Educational Data Mining, Digital Processing of Signals and Images, Programming, Pattern Recognition, Artificial Neural Networks.



Serhiy Balovsyak: Graduated from Chernivtsi State University (1995). In 2018, he defended his doctoral dissertation in the specialty "Computer systems and components".
Currently position – associate professor at the Department of Computer Systems and Networks of Yuriy Fedkovych Chernivtsi National University, Ukraine.
Research Interests: Education Technology, Digital Processing of Signals and Images, Programming, Pattern Recognition, Artificial Neural Networks.



Shaocheng Qu received the B.Sc. degree in process control and instrument from the Wuhan University of Chemical and Technology, China, in 1994, the M.Sc. degree in control engineering from the Naval University of Engineering in 2000, and the Ph.D. degree in control theory and control engineering from the Huazhong University of Science and Technology, China, in 2004. Since 2004, he joined Department of Electronics and Information Engineering, College of Physical Science and Technology, Central China Normal University, China, where he is currently a Professor and the Director of the Research Center of Artificial Intelligence and Internet of Things. He was a Visiting Professor with Bristol Institute Technology, University of the West of England, UK, from 2010-2011. He was a Senior Research Scientist with Department of Electrical, Electronic & Computer Engineering, University of Pretoria, SA, from 2012-2013. He has published more than 70 papers on these topics of Control theory and applications, embedded systems, and the IoT system applications, and provided consultancy to various industries.



Yuriy Ushenko: M.Sc. in Telecommunications (2003).
PhD in Optics and Laser Physics (2006). D.Sc. in Optics and Laser Physics, Taras Shevchenko National University of Kyiv (2015).
Current position – Professor, Head of Computer Science Department, Yuriy Fedkovych Chernivtsi National University, Ukraine.
Research Interests: Data Mining and Analysis, Computer Vision and Pattern Recognition, Optics & Photonics, Biophysics.



Nataliia Tytova: Doctor of Pedagogical Sciences (2019), Professor (2021), Head of Department of Department of Professional Training, Document Studies and Public Administration Dragomanov Ukrainian State University.
Research Interests: psychological and pedagogical training of vocational education teachers, professional training of specialists in the field of document studies, public management and administration.



Hanna Kravchenko: Teacher of the Chernivtsi Applied College of Transport, Chernivtsi, Ukraine.
Research Interests: Education Technology, Educational Data Mining, Digital Processing of Signals and Images, Programming, Pattern Recognition, Artificial Neural Networks.

How to cite this paper

Oleksandr Derevyanchuk, Serhiy Balovsyak, Shaocheng Qu, Yurii Ushenko, Nataliia Tytova, Hanna Kravchenko, "Specialized Software Tools for Educational Data Mining of Technical Students: Image Segmentation, Clustering, and Correlation-Regression Analysis", International Journal of Modern Education and Computer Science(IJMECS), Vol.18, No.5, pp. 87-103, 2026. DOI:10.5815/ijmecs.2026.05.05