Correlation Analysis of Data in the Education Management System

Gevorg Margarov
NPUA
Yerevan, Armenia
e-mail: mgi@polytechnic.am

Kristine Hambardzumyan
NPUA
Yerevan, Armenia
e-mail: hambardzumyan.k@polytechnic.am

Marine Usepyan
NPUA
Yerevan, Armenia
e-mail: musepyan@ polytechnic.am

***Abstract*—The task of raising the level of education is to identify all the factors that affect the quality of education. Research has shown that there are many such factors. An important factor influencing the quality of education is also a properly constructed curriculum, which should contribute to the development of the necessary competencies of students and ensure the competitiveness of graduates in the labor market. It follows from the foregoing that in order to improve the quality of education, it is necessary to collect all the factors and, after processing them, identify important factors that affect the quality of education. To solve this problem, this article proposes using modern information technologies to collect all the necessary data that affect the quality of education, prepare them for processing, and conduct a correlation analysis of the data. As a result of this analysis, interconnected data are identified, and data can be identified that have the greatest impact on the quality of education. The article also proposes the use of correlation analysis to determine the correlation dependence of disciplines and identify the order of study of various topics and disciplines, which contributes to the correct assimilation of the material and the formation of the necessary competencies. The results obtained will help improve the effectiveness of courses, optimize curricula, and as a result, improve the educational system.**

***Keywords*—Correlation analysis, correlation coefficient, Pearson correlation coefficient, collected education data, learning outcomes.**

I. INTRODUCTION

Data analytics techniques have been gaining more space in the scientific environment with applications in various areas of knowledge, including education [1].

Educational data analytics is used to study the data available in the educational field and bring out the hidden knowledge from it. Analytics is a process of discovering, analyzing, and interpreting meaningful patterns from large amounts of data. Data analytics relies on the techniques of data mining such as classification, association, correlation, categorization, prediction, estimation, clustering, trend analysis and visualization [2].

Attribute /Feature selection methods are used to reduce the dimensionality of the data by removing the redundant and irrelevant attributes in a data set.

The feature selection process has many benefits: it allows visualization and understanding of data more easily, reduces the time and storage required for the mining process, and improves the performance of the algorithms by avoiding the curse of dimensionality [4].

We will use correlation analysis to identify relationships between data, presenting strong and weak relationships.

Which, in turn, will serve as the basis for extracting data that has the greatest impact in this system.

II. CORRELATION ANALYSIS

Correlation is the relationship or the dependence between people, things, ideas, etc. As an analogy, in the statistics, we can say: interdependence between two or more variables [1].

Correlation is used to test relationships between quantitative variables or categorical variables.

A high correlation points to a strong relationship between the two variables, while a low correlation means that the variables are weakly related.

Researchers use correlation analysis to analyze quantitative data collected through research methods like surveys and live polls. They try to identify the relationship, patterns, significant connections, and trends between two variables or datasets.

There is a positive correlation between two variables when an increase in one variable leads to the increase in the other. On the other hand, a negative correlation means that when one variable increases, the other decreases and vice-versa.

Correlations are useful because if you can find out what relationship variables have, you can make predictions about future behavior.

An intelligent correlation analysis can lead to a greater understanding of your data.

III. CORRELATION COEFFICIENTS

It is often necessary to evaluate the degree of relationship between two or more variables. We can discover accurately how much one variable interferes with the outcome of another.

However, in some situations, the relationship between two variables is not linear, or one of them is not continuous, or the observations are not randomly selected.

The correlation coefficient is the unit of measurement used to calculate the intensity in the linear relationship between the variables involved in a correlation analysis.

Correlation is a bivariate analysis that measures the strength of the association between two variables and the direction of

the relationship between the measurements obtained. In terms of the strength of the relationship, the value of the correlation coefficient (r) ranges between + 1 and -1. A value indicates a perfect degree of association between two variables, where the sign indicates the direction of this association; a + sign indicates a positive relationship and a - sign indicates a negative relationship [1].

IV. Pearson Correlation Coefficient

Pearson correlation coefficient or Pearson's correlation coefficient or Pearson's r is defined in statistics as a measurement of the strength of the relationship between two variables and their association with each other.

In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

This approach is based on covariance and, thus, is the best method to measure the relationship between two variables.

- **Positive correlation:** A positive correlation between two variables means both the variables move in the same direction. An increase in one variable leads to an increase in the other variable and vice versa.
- **Negative correlation:** A negative correlation between two variables means that the variables move in opposite directions. An increase in one variable leads to a decrease in the other variable and vice versa.
- **Weak/Zero correlation:** No correlation exists when one variable does not affect the other.

The Pearson correlation coefficient is presented in Fig. 1.

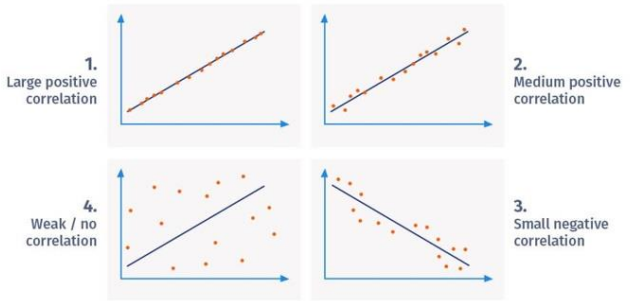


Fig. 1. Pearson correlation coefficient

The Pearson coefficient correlation has a high statistical significance. It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. The relationship of variables is measured with the help of the Pearson correlation coefficient calculator. This linear relationship can be positive or negative.

V. Pearson Correlation Coefficient Formula

The correlation coefficient formula finds out the relation between the variables. It returns the values between -1 and 1.

$$r = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}} \quad (1)$$

where:

N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

VI. Correlation Analysis of Collected Data in the Education Management System in Real Time

Considering the offered possibilities of current information technologies, nowadays we are able to collect a large amount of data in the educational system and process them. This study investigates the significance of using multi-source data, such as student application data, university platform records, and survey findings, to undertake a thorough analysis of student admission and academic performance in the university setting. The study analyzes crucial elements impacting students' achievement by combining data from numerous sources, such as entrance standards, university assistance, and non-cognitive qualities. The findings emphasize the need of personalized support systems and data-driven decision-making in improving admission procedures and overall student success. This study makes an important addition to the academic community by demonstrating the possibility of using varied data sets for in-depth analysis of educational phenomena and guiding future research in related fields. All possible data of the student are collected, namely: age, gender, address, standard school, spatial school, college, mathematic score, physics score, English score, information score, mathematic Olympiad, physics Olympiad, other interests, attendance, social status, health condition, religion, work, spatial work, MOG.

It is necessary to find a correlation between the data in order to ascertain which features have an impact on MOG.

The collected data are presented in Fig. 2.

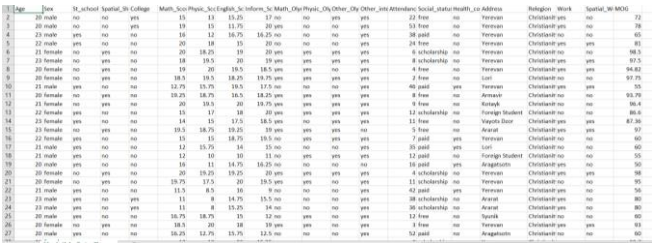


Fig. 2. Collected data

Data preprocessing is one of the most complex data mining tasks, which includes preparation and transformation of data into a form suitable for mining procedure. Data preprocessing aims to reduce the data size, find the relations between data, normalize data, remove outliers, and extract features for data. It includes several techniques like data cleaning, integration, transformation, and reduction [3].

First, data cleaning is performed in the collected data, during which the missing data are filled with relevant data, then data transformation is performed. Data transformation implies making all text characters discrete (digital), which will ensure further work with them. The transformed data are presented in Fig. 3.

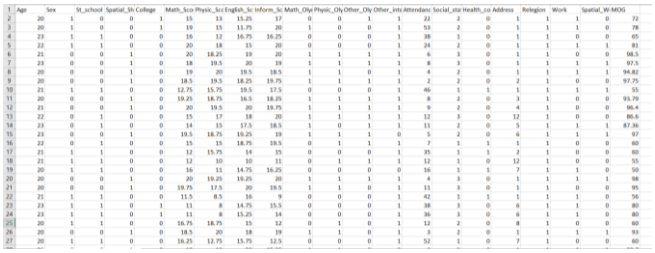


Fig. 3. Transformed data

We apply the correlation to the modified twills, the result is shown in Fig. 4.

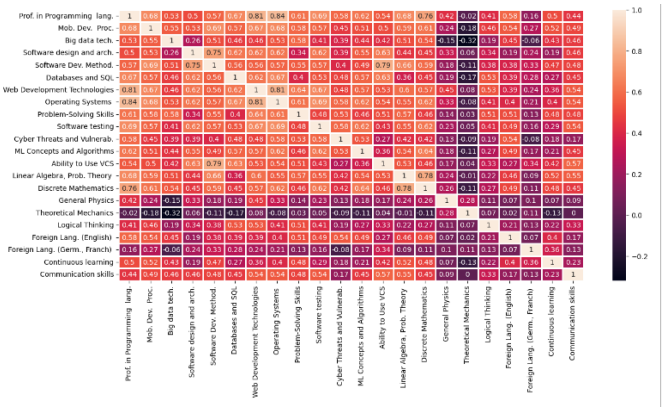


Fig. 4. Correlation analysis

The obtained results allow us to identify weak and strong connections between features. It is important to identify that features have an effect on GPA. From the obtained results we can single out the data that show a strong connection, these are the cases when the correlation value is in the range from 0.5 to 1 or from -0.5 to -1.

As a result of the correlation analysis for the presented data, it turns out that the average qualitative grade has a strong positive relationship with the following characteristics: special school graduate, mathematics, physics, informatics, English grade, mathematics, physics Olympiad participant, social status, and has a strong negative relationship with the student's gender, regular school, attendance, health status, indicating that the increase of the given characteristic leads to a decrease in the average qualitative grade (opposite effect), and the following characteristics have a weak relationship with other interests, address, college and work. There are certain characteristics that have borderline values, which are participation in other Olympiads and professional work. The obtained results allow to understand, in the educational process, which features are necessary and have an impact on the average quality assessment, and which features are not necessary and can be ignored (reduced). The same principle can be applied to any amount of data to obtain the necessary features. On the same principle, the analysis can also be carried out for the assessment of each subject, choosing those features that have an impact on the improvement of the quality of the course.

VII. Correlation Analysis to Reveal Strong and Weak Relationships Between Learning Outcomes

Modern approaches to the development of university curricula are based on considering the requirements of the labor market and the needs of students. This means that the curriculum should be focused on specific professional competences, that will be in demand in the future professional activity of the student [5].

Since the requirements of the labor market are permanently changing, it becomes necessary to regularly review academic disciplines and curricula. For these purposes, it is necessary to conduct a survey of all stakeholders, these are: representatives of the labor market, students, graduates and university professors. The questionnaire includes main learning outcomes that are necessary for a specialist in the chosen field. The research was carried out for the specialty "Software Engineering". The table presents the results of the stakeholder survey (Fig. 5).

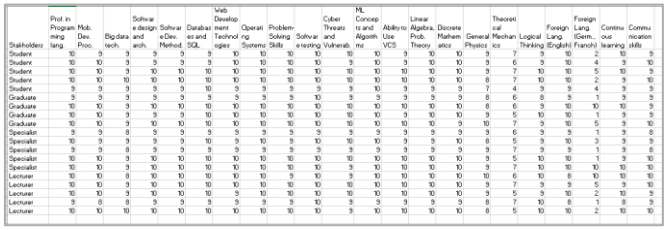


Fig. 5. Survey results for all stakeholders

The results of the correlation analysis of cooperation performance are presented below (Fig. 6) for all the questioned administrative departments indicated.

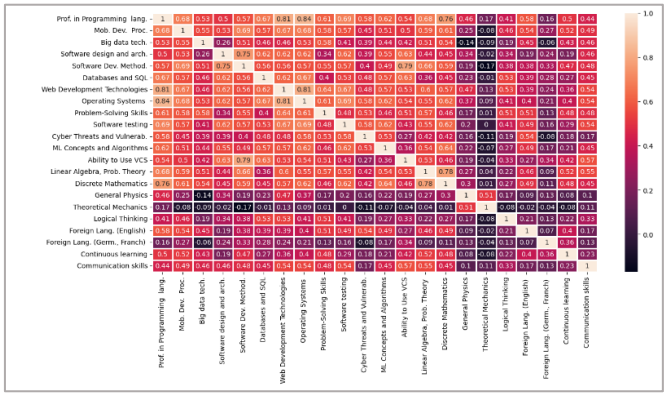
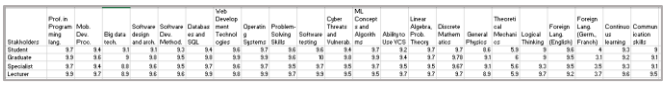


Fig. 6. Correlation analysis for all data

The table presents the dependencies between the final results, differentiated by colors. Results corresponding to the purple color indicate outcomes influenced by low performance. For instance, Theoretical Mechanics and Foreign Language (German, French) show low achievers in the remaining final results. On the other hand, the correlation of the Linear Algebra and Discrete Mathematics learning outcomes with the performance of the remaining final results is considerably higher.

For comparison purposes, the average number calculation was performed for four administrative departments based on the results presented in the table (Fig. 7).



Stakeholders	Prof. in Programming lang.	Mob. Dev. Proc.	Big data tech.	Software design and arch.	Software Dev. Method.	Databases and SQL	Web Development Technologies	Operating Systems	Problem-Solving Skills	Software testing	Cyber Threats and Vulnerab.	ML Concepts and Algorithms	Ability to Use VCS	Linear Algebra, Prob. Theory	Discrete Mathematics	General Physics	Theoretical Mechanics	Logical Thinking	Foreign Lang. (English)	Foreign Lang. (Germ., Franch)	Continuous learning	Communication skills
Student	9.7	9.4	9.1	9.1	9.3	9.4	9.6	9.7	9.6	9.6	9.6	9.4	9.7	9.2	9.7	9.7	8.6	5.9	9	9.6	4	9.3	9
Graduate	9.9	9.6	9	9.8	9.5	9.8	9.9	9.9	9.6	10	9.8	9.9	9.4	9.7	9.5	9.70	9.1	6	9	9.5	3.1	9.2	9.1
Specialist	9.7	9.4	8.8	9.6	9.5	9.7	9.6	9.7	9.5	9.7	9.5	9.7	9.5	9.5	9.7	9.67	9.1	5.6	9.2	9.5	3.5	9.2	9.1
Lecturer	9.9	9.7	8.9	9.6	9.6	9.9	9.8	9.9	9.7	9.9	9.5	9.5	9.7	9.7	9.7	9.7	8.9	5.9	9.7	9.2	3.7	9.6	9.5

Fig. 7. Average data for each group of stakeholders

By applying correlation analysis, the following correlation values are obtained on the average of the normalized values of the results for the four administrative departments. The collected analysis for average data is presented in Fig. 8.

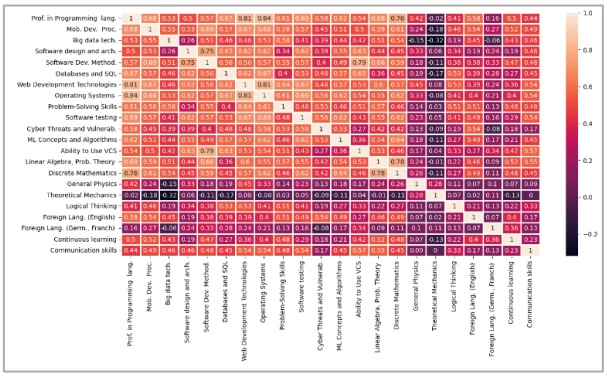


Fig. 8. Correlation analysis for average data

Thus, the obtained two correlation coefficient values do not differ significantly. Analyzing the acquired results allows us to conclude that the selected disciplines with a correlation coefficient greater than or equal to 0.5 are significantly connected to the final outcomes, while those disciplines with a correlation coefficient close to 0 can be disregarded.

Based on this analysis, an educational plan can be developed that incorporates the courses with strong correlations and, consequently, more class hours will be allocated to these courses compared to those with weak connections to the remaining subjects. Additionally, some subjects with no significant correlations can be omitted from the educational plan altogether, which will enable the students to have more focused study plans and elevate the overall educational quality, considering the dynamic demands of the job market.

VIII. Conclusion

Correlational analysis based on the foundation of exploratory research reveal that they contribute to the improvement of academic performance, which assumes acquiring essential indicators from a vast amount of data and minimizing insignificant data, as well as arranging positive educational plans for effective academic progress.

The correlational analysis was conducted at the foundation of the research carried out by Pearson, which, with the assistance of Parsonean collaboration, defined strong relationships between structured data and permitted correlations.

For clarity, the correlational analysis was performed on the average quality assessments of students throughout their years of study, allowing the identified factors to potentially have a correlation with the overall academic performance of students during the course of their academic journey.

Since the students' academic journey data are dependent on positive academic planning, the correlational analysis of final outcomes allows for the improvement of educational plans based on labor market demands.

Acknowledgment

The authors are grateful to students and stakeholders for participating in the survey and supporting the research.

References

[1] P.A.N. Prestes, T.E.V. Silva, G.C. Barroso, “Correlation analysis using teaching and learning analytics”, *Heliyon*, Issue 11, vol. 7, Published online: November 19, 2021

[2] Bharati Kawade, Dr. K S Wagh, “Data Analytics in Educational Management System”, *International Journal of Computer Applications (0975 – 8887), National Conference on Advances in Computer, Communication and Networking* 2016, Pune, Maharashtra, February, 2016

[3] Kristine Hambardzumyan, “Data Preprocessing in Real-time Education Management System”, *CSIT Conference*, Yerevan, Armenia, September 27 - October 1, 2021

[4] Isabelle Guyon, Andre Elisseeff, “An Introduction to Variable and Feature Selection”, *J. of Machine Learn. Res.*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[5] Manuela Epure, “University-business cooperation: adapting the curriculum and educational package to labor market requirements”, *Proceedings of the International Conference on Business Excellence*, Issue 1, vol. 11, pp. 339-349, July, 2017