Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Educational Data Mining: Investigating Performance Assessment and the Interplay of Subjects from a Statistical Perspective

Prakriti Kapoor[1], Santam Kaushik[2], Dr. Balaji Gopalan[3], Shikhar Gupta[4], Anupamaa Bijlani[5], Onkar Bagaria[6]

[1]Centre of Research Impact and Outcome, Chitkara University, Rajpura- 140417, Punjab, India, Email Id- prakriti.kapoor.orp@chitkara.edu.in
[2]IT Head, Department of IT, Sanskriti University, Mathura, Uttar Pradesh, India, Email Id- it.office4@sanskriti.edu.in
[3]Assistant Professor, Department of Decision Science, JAIN (Deemed-to-be Univesity), Bangalore, Karnataka, India, Email Id- dr.balaji_gopalan@cms.ac.in
[4]Chitkara Centre for Research and Development, Chitkara University, Himachal Pradesh-174103 India, Email Id- shikhar.gupta.orp@chitkara.edu.in
[5]Faculty, Department of ISME, ATLAS SkillTech University, Mumbai, Maharashta, India, Email Id- anupamaa.bijlani@atlasuniversity.edu.in
[6]Department of Management Studies, Vivekananda Global University, Jaipur, India, Email Id- bagaria.omkar@vgu.ac.in

Introduction: Teaching professionals and other educational participants have a significant challenge in improving students' performance since it is primarily dependent on a variety of essential indicators and ancillary academic elements.
Objective: From the perspective of Educational Data Mining (EDM) procedures, the interaction between these factors presents itself in a mysterious fashion that is unknown. As a result, it grows essential to use variety of data mining approaches to analyze and find data generated from diverse educational sources to identify numerous unclear patterns.
Method: This study aimed to identify and distinguish elements from instructional data that affect student performance development. A unique effort has been made to uncover the impact of demographic factors and individual subjects on overall performance by using variety of mining approaches in relation to Analysis of variance (ANOVA) and Structural Equation Modeling (SEM).
Result: The results, evaluated as a whole, have unequivocally shown that proficiency in the English language is crucial for its overall effectiveness.
Conclusion: The majority of the findings unequivocally demonstrate that ability in the English language is critical to perform in the final product yet they urge academicians and other relevant stakeholders to prioritize this subject to improve the educational outcomes of students.

Keywords: F Educational Data Mining (EDM); Structural Equation Modeling (SEM); Analysis of variance (ANOVA); Students.

1. Introduction

The discovery of data is known as Data Mining (DM). This field deals with interpreting significant data to derive current beneficial ideas. (1) Data mining methods is used to forecast patterns and behaviors, which can used to inform organizational choices. (2)

To enhance the setting for learning and instruction, language proficiency is critical for understanding how students behave in educational environments as well as to distinguish successful and unsuccessful students. (3-4)

To investigate hidden trends, program collects and analyzes data. Numerous fields, including medical care, business, higher education, financial research and intrusion detection, can benefit from the use of data mining. (5) The factors or qualities of the educational environment, both internal and external, impact the academic success of students. (6)

Student performance assessment constitutes a few of the most significant applications in higher education. (7) Annual grades are typically used by a large number of higher education institutions to forecast student achievement (8).

To offered a balanced viewpoint on the variables impacting student achievement and their complex relationships across many academic fields. (9) In the field of EDM, a variety of data mining techniques are dedicated that is utilized to find educational information and classify parameters responsible for improved academic accomplishments. (10)

The paper (11) determined the learners engage with education monitors an accurate prediction of their achievements from a virtual classroom and how much interaction data utilized to predict and guide learners' academic success. The study (12) assessed the possibility of using artificial intelligence for education, as well as the methods used by academics over time and current developments in data mining for educational study. The study (13) focused on data mining tools for predicting college students' academic success, highlighting how crucial an ethical application process was choosing physically qualified applicants. The paper (14) explored the use of video learning in higher education institutes, highlighting the effectiveness of this method by examining the digital footprints left by these interactions.

The study (15) focused on an innovative approach that predicts the academic achievement of students with learning obstacles by analyzing their procrastinating behavior while submitting assignments. The study (16) enhanced the educational process through learning analysis, enabling educators to identify and intervene in student safety through behavior assessment in education statistics. The paper (17) provided a unique three-stage unbalanced big data mining structure aimed at enhancing the efficiency of optimization methods by removing the local maximum difficulty. The study (18) examined each learner's historical e-learning system usage through the utilization of techniques for data mining. The paper (19) presented a collaborative system that evaluates, adjusts and improves the education experience for students as well as educators by using educational information produced by online educational institutions.

The purpose of this study was to evaluate instructional data in order to determine which components influence the development of student performance. An inventive effort has been undertaken to determine the influence of individual topics and demographic characteristics on overall performance through the application of a range of mining techniques.

2. Methodology

Data Collection

The educational institution of Kashmir provided the data collection for the study, which included each of the educational institutions in Kashmir, involving the institutions in North, South and Central Kashmir. Each of the 24 universities included in this investigation was the subject of the study. During the pre-processing stage, identification codes were used to extract critical data, including statistics and college codes. The dataset initially had 2899 1 entries of Bachelor of Science (BSC) students along with nine characteristics.(20) Following the extraction of data for the stream mentioned above, 1793 counts were found in total. The collection of data included data regarding every student, such as their registration number, name, course code, parentage, capacity, topics, all grades received, total scores received and their overall rating.

Data pre-processing

The techniques that follow were considered as well as employed to analyze and obtain the data prepared for evaluation:

The data mining process is considered to be crucial. Factual data is illogical, loud and lacking. As a result, data must be chosen as well as changed into an additional reliable and consistent condition. Data extraction, data transformation, data cleaning and other operations were carried out as a component of the pre-processing stage of the data. The following characteristics were eliminated during the data pre-processing stage, after which various categories were eliminated out of the dataset. SQL 2008 was used to extract fields like English, Biology, Chemistry and Zoology from attribute topics. An attribute registration number was utilized to calculate the demographics. In addition, the categorization of the demographic information was performed in accordance with data processing requirements.

Hypothesis Development

The following hypotheses were examined and accurate overall results were obtained for each hypothesis.

Hypothesis 1: The total marks obtained are inversely correlated with chemistry.

Hypothesis 2: The total amount of marks earned and zoology are strongly correlated.

Hypothesis 3: The total number of marks earned has a negative correlation with biology.

Hypothesis 4: The total amount of scores received has a negative correlation with general English.

3. Analysis of Results

The study used a variety of methods, including multiple regression and linear regression, using a real dataset that was obtained from the academic of South India. The following sections provide illustrations of the outcomes obtained through the use of the methodologies above, accompanied by appropriate interpretations.

In our study, the structural framework requires four input variables in addition to one output component. The framework demonstrates the effect of such predicted elements on the output factors. ANOVA, which should be covered in the following section, is utilized to assess the outcome variable in the categories of rural and urban to determine if a rural region performs better than an urban area. Figure 1 illustrates how the following four different input factors GE_TOT, BO_TOT, CH_TOT and ZO_TOT determine the resultant factor, OBT_Scores.

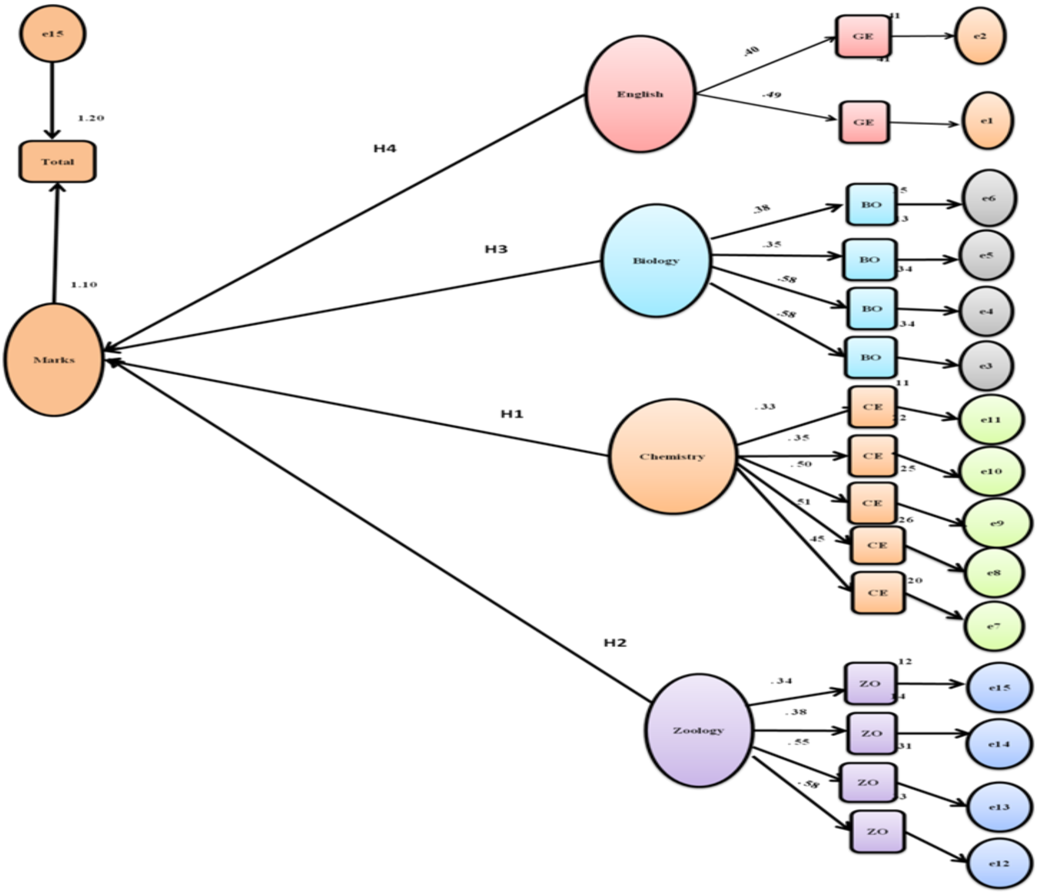

Figure 1. Structural Model [Source: Author]

Following the computational modeling's execution, a number of modeling predictions have been shown to be significant, including Chi-square = 2 482,627, p <,000. Furthermore, the results confirm the idea that fields including Chemistry, Zoology, Biology and English possess a significant influence on final grades.

Additionally, Table 1 demonstrates that the route assessment connecting English and scores is essential when the value is below 0,05. The equivalent occur involving approach calculations that involve chemistry and scores, biology and scores, along with zoology and scores. According to the architectural approach, English and the total score possess the highest average of 0,95, indicating that language arts as an area of study has a significant influence on a student's overall assessment.

Table 1. Sodium Level Arbitrary Box Chart [Source: Author]

Variables	Standard Error	Evaluate	P value	Critical Region	Identification
Marks-English	0,176	0,822	***	4,70	significant
Marks-Chemistry	0,097	0,352	***	3,67	significant
Marks-Zoology	0,087	0,358	***	4,16	significant
Marks-Biology	0,242	-0,622	***	-2,59	significant

The current study's hypotheses were assessed according to the evaluation of Critical Region (CR), while p values and Table 2 provide an overall expansive collection of findings related to the hypothesis investigated. Furthermore, the three ratings and p-value of 001 which is below 05 in scores-English in Table 2, that indicate the association is relevant. The findings additionally demonstrate the significant relationship between each of the additional interactions, such as Marks-Biology, Marks-Chemistry and Marks-Zoology, since the p-value for each of those relationships is below 05 as well as is 001 as shown by the triple stars. Similar to the previous example, the findings demonstrate that there is indeed a significant connection between the Marks-Biology, Marks-Chemistry and Marks-Zoology relationships. The p value for those relationships is 001, which is below the threshold of 05 and it is denoted by triple stars.

Table 2. Findings related to the assessment hypothesis [Source: Author]

Hypothesis	Standard Error	Evaluate	P	Critical Region	Outcome
H1	0,092	-0,671	***	-7,29	Validated
H2	0,086	0,357	***	4,15	Validated
H3	0,241	-0,621	***	-2,59	Validated
H4	0,198	-0,818	***	-4,4	Validated

To obtain outcomes for students with diverse demographic regions, an ANOVA was utilized for factors including English, Biology, Chemistry, Zoology and Total Scores. This allowed for additional progress toward solving the present challenge. To ascertain whether the student demographic category makes a more significant contribution to overall efficiency, extensive research was done to learn more about students who are located in rural or urban areas. According to the data presented in the tables below, students living in urban regions tend to have better English scores on the GE_TOT (English) assessment than those living in rural areas. According to our results, urban pupils have a more robust command of the English language compared to their rural counterparts. In addition, pupils from rural areas outperform their urban counterparts in other topics, which include

BO_TOT (Biology), CH_TOT (Chemistry) and ZO_TOT (Zoology). Table 3 shows that when English was the topic, metropolitan areas had higher values for measures, including standard deviation and mean. The percentages are significant among the rural classification in chemistry, biology and zoology. Table 4 indicates that the findings possess the statistical significance shown by Sig. Each of the variables that is dependent, including GE_TOT, BO_TOT, CH_TOT and ZO_TOT, offers an amount lower than 0,051. F connected using the evaluation statistic is 19,853, a number larger than the calculated or essential amount at 1 and 1793 directions of independence. Meanwhile, ANOVA was done between demographic information categorized factor - rural areas and GE_TOT comprising constant factor. Moreover, the importance of the findings is shown by p = 00, a coefficient smaller than 05.

Table 3. Mean and standard deviation are high in the case of urban students taking English [Source: Author]

Variables		Mean	N	Std. Deviation
Biology	1	91,10	832	14,292
	2	95,63	965	13,356
	Overall	93,53	1 796	13,977
English	1	77,92	832	13,487
	2	75,35	965	11,651
	Overall	76,55	1 796	12,596
Zoology	1	88,37	832	13,983
	2	91,60	965	11,073
	Overall	90,11	1 796	12,604
Chemistry	1	89,83	832	12,358
	2	93,49	965	11,891
	Overall	91,79	1 796	12,243

Table 4.Statistical Significance (Sig) for the data [Source: Author]

ANOVA					
Total values	Sum of squares	Data Frame (df)	Mean Square	F	Significant
Biology					
Within Groups	3 41 268,0 451	1 794	190,335	-	-
Between Groups	9 137,0 559	2	9 137,559	48 ,009	0,001
Overall	3 50 406,009	1 795	-	-	-
English					
Within Groups	2 81 640,0 186	1 794	157,079	-	-
Between Groups	2 961,0 201	1	2 961,201	18 ,853	0,001
Overall	2 84 601,0 387	1795	-	-	-
Zoology					

Within Groups	2 80 319,0 754	1 794	156,342	-	-
Between Groups	4 640,0 958	1	4 640,958	29 ,686	0,001
Overall	2 84 960,0 711	1 795	-	-	-
Chemistry					
Within Groups	2 62 880,149	1 794	146,616	-	-
Between Groups	5 981,0 417	1	5 981 ,417	40 ,798	0,001
Overall	2 68 861,564	1 795	4 640 ,958	-	-

4. Discussion

The primary goal of this research was to identify and classify the elements that have the potential to enhance educational abilities and performance dramatically. According to the practical findings, the academic area of English is a crucial variable that influences the dependent factor, which is the total outcome, in comparison with different subjects like Biology, Chemistry and Zoology. Additionally, the results have established a solid foundation for the relationship among the predictive and criteria factors while limiting the impact of socioeconomic variables. Educational practice is going too impacted directly by the study's findings. Due to the focus on English competence, there is a demand for programs of focused language education that use cutting-edge teaching. Considering the impact of demographics necessitates inclusive teaching methods that take into account students' varied histories.

5. Conclusion

Students from urban locations could perform better in the English language compared to students from different areas for the reasons mentioned in the study. However, compared to the urban population, pupils from rural areas have shown much higher scores in zoology, biology and chemistry. Phenomenological limitations have hindered the field of EDM and how it is associated with various factors and an individual's general academic achievement, leading to limited efforts in the area. The results of this investigation are significant and might help several parties advanced in the growth of students' educational accomplishment. Considering every effort to ensure the study's validity and reliability, some limitations are typical of this type of study that can't be entirely eliminated. Researchers interested in the field of EDM can find valuable insights from this study, which will assist them in their future endeavors.

References

1. Moscoso-Zea O, Saa P, Luján-Mora S. Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining. Australasian Journal of

Engineering Education. 2019 Jan 2;24(1):4-13. DOI: 10.1080/22054952.2019.1601063

2. Mimis M, El Hajji M, Es-saady Y, OueldGuejdi A, Douzi H, Mammass D. A framework for intelligent academic guidance using educational data mining. Education and Information Technologies. 2019 Mar 16;24:1379-93. DOI: 10.1007/s10639-018-9838-8
3. Injadat M, Moubayed A, Nassif AB, Shami A. Systematic ensemble model selection approach for educational data mining. Knowledge-Based Systems. 2020 Jul 20;200:105992. DOI: 10.1016/j.knosys.2020.105992
4. Sarra A, Fontanella L, Di Zio S. Identifying students at risk of academic failure within the educational data mining framework. Social Indicators Research. 2019 Nov;146:41-60. DOI: 10.1007/s11205-018-1901-8
5. Dabhade P, Agarwal R, Alameen KP, Fathima AT, Sridharan R, Gopakumar G. Educational data mining for predicting students' academic performance using machine learning algorithms. Materials Today: Proceedings. 2021 Jan 1;47:5260-7. DOI: 10.1016/j.matpr.2021.05.646
6. Lemay DJ, Baek C, Doleck T. Comparison of learning analytics and educational data mining: A topic modeling approach. Computers and Education: Artificial Intelligence. 2021 Jan 1;2:100016. DOI: 10.1016/j.caeai.2021.100016
7. Huang AY, Lu OH, Huang JC, Yin CJ, Yang SJ. Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. Interactive Learning Environments. 2020 Feb 17;28(2):206-30. DOI: 10.1080/10494820.2019.1636086
8. Dabhade P, Agarwal R, Alameen KP, Fathima AT, Sridharan R, Gopakumar G. Educational data mining for predicting students' academic performance using machine learning algorithms. Materials Today: Proceedings. 2021 Jan 1;47:5260-7. DOI: 10.1016/j.matpr.2021.05.646
9. Mahajan G, Saini B. Educational Data Mining: A state-of-the-art survey on tools and techniques used in EDM. International Journal of Computer Applications & Information Technology. 2020;12(1):310-6.
10. Hasan R, Palaniappan S, Mahmood S, Sarker KU, Abbas A. Modelling and predicting student's academic performance using classification data mining techniques. International Journal of Business Information Systems. 2020;34(3):403-22. DOI: 10.1504/IJBIS.2020.108649
11. Kokoç M, Altun A. Effects of learner interaction with learning dashboards on academic performance in an e-learning environment. Behavior & Information Technology. 2021 Jan 25;40(2):161-75. DOI: 10.1080/0144929X.2019.1680731
12. Salloum SA, Alshurideh M, Elnagar A, Shaalan K. Mining in educational data: review and future directions. InProceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020) 2020 (pp. 92-102). Springer International Publishing. DOI: 10.1007/978-3-030-44289-7_9
13. Mengash HA. Using data mining techniques to predict student performance to support decision making in university admission systems. Ieee Access. 2020 Mar 19;8:55462-70. DOI: 10.1109/ACCESS.2020.2981905
14. Hasan R, Palaniappan S, Mahmood S, Abbas A, Sarker KU, Sattar MU. Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. Applied Sciences. 2020 Jun 4;10(11):3894. DOI: 10.3390/app10113894
15. Hooshyar D, Pedaste M, Yang Y. Mining educational data to predict students' performance through procrastination behavior. Entropy. 2019 Dec 20;22(1):12. https://doi.org/10.3390/e22010012
16. Huang AY, Lu OH, Huang JC, Yin CJ, Yang SJ. Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and

learning logs. Interactive Learning Environments. 2020 Feb 17;28(2):206-30. DOI: 10.1080/10494820.2019.1636086

17. Hassib EM, El-Desouky AI, El-Kenawy ES, El-Ghamrawy SM. An imbalanced big data mining framework for improving optimization algorithms performance. IEEE Access. 2019 Nov 26;7:170774-95. DOI: 10.1109/ACCESS.2019.2955983
18. Daghestani LF, Ibrahim LF, Al-Towirgi RS, Salman HA. Adapting gamified learning systems using educational data mining techniques. Computer Applications in Engineering Education. 2020 May;28(3):568-89. DOI: 10.1002/cae.22227
19. Chytas K, Tsolakidis A, Triperina E, Skourlas C. Educational data mining in the academic setting: employing the data produced by blended learning to ameliorate the learning process. Data Technologies and Applications. 2023 Jun 14;57(3):366-84. DOI: 10.1108/DTA-06-2022-0252
20. Ashraf M, Zaman M, Ahmed M. Using ensemble StackingC method and base classifiers to ameliorate prediction accuracy of pedagogical data. Procedia computer science. 2018 Jan 1;132:1021-40. DOI: 10.1016/j.procs.2018.05.018