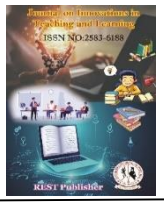
DOI: https://doi.org/10.46632/jitl/1/1/8

Journal on Innovations in Teaching and Learning
Vol: 1(1), 2022
REST Publisher; ISSN:2583 6188
Website: http: //restpublisher.com/journals/jitl/



Demographic performance in educational data mining for medical student academic success

***A. Neelamadheswari, M. Dhurgadevi, R. Kalaivani, P. Priyadharshini**
Mahendra Engineering College, Namakkal, Tamil Nadu, India.
*Corresponding author email: neelamadheswaria@mahendra.info

Abstract: Educational Data Mining (EDM) is a powerful approach for discovering hidden patterns over educational data as well as forecasting students' academic performance. This paper presents a novel approach with respect to Machine Learning (ML) methods in predicting the final test marks of undergraduate students's using midterm exam grades as primary data. Despite employing demographic variables for predicting students performance using EDM has been deemed extremely problematic with respect to fairness as well as is now being critically explored in the area, but are commonly employed in practice with no significant reasoning. The implementation and discussion are based primarily on the notion of how they contribute to excellent model performance. In this study, the processes that make them significant for prediction has been examined as well as the implications for ideas of equity with Logistic Regression (LR) method. These findings emphasize the importance of learning more regarding the causes and effects behind the data and thinking cautiously about demographic aspects in each unique situation, as well as conducting additional research to understand how demographic features impact educational achievement.

Keywords: Educational Data Mining, Logistic Regression, demographic features, machine learning, educational data, logistic regression

1. INTRODUCTION

EDM signifies the application of classical DM methods for addressing educational challenges and it is DM the application to educational data that includes student details, exam results, educational records, student attendance in class, and the question about students inquire frequently. At present, EDM play as a vital method to find out hidden trends in educational data, predicting academic progress, and the learning improvement as well as teaching environment. The usage of EDM has expanded the learning analytics scope [1].

Learning analytics encompasses the numerous components of student information collection, better knowledge about learning environment by conducting examination and analysis, as well as identifying the optimal student and teacher performance. One of the recent disputes in EDM is over whether demographic characteristics should be utilized in prediction models. Demographic aspects are specific characteristics of a population namely age, gender, education, religion, race, ethnicity and income. Researchers define demographic features in the EDM context in particular, declaring that these characteristics represent data providing context about students' experiences and backgrounds which it offer with them into school as well as other than academic achievement, demographic variables are incapable of being influenced [2].

Even if the demographic features need to be incorporated over predictive models have associated with the concept of justice via unawareness [3]. The fairness concept by unawareness is addressed about benefits in ML for a long time as well as possesses lately received attention in EDM [4], [5]. The underlying premise underlying this fairness concept is the fact that forecasting algorithms prioritize demographic variables has unfair since persons deemed comparable may receive distinct projections as an outcome [6]. As an instance, assume by employing a model that assigns students to a course-based on their predicted completion rate. When demographic variable influences about prediction and the student who are similarly compatible by all other criteria may have a different prognosis based purely on demographic factors. Evidently, most people would consider this unfair, and many urge for eliminating demographic features to

avoid it [7]. This viewpoint is not always shared by all researchers. However, the demographic information ought to be provided, and failure to accomplish so it could potentially interpreted as supporting a 'colorblind' attitude [8].

As a result, EDM is a growing multidisciplinary study field that combines education as well as technology, with the potential in extracting and converting enormous educational data towards usable data patterns and vital ideas. Several tasks like prediction, description, estimation, clustering, association and classification discovering, contribute towards the creation of many different approaches such as LR, naïve Bayes, Decision Tree (DT), neural networks (NN), Apriori, FP-Growth, K-means, etc. [9]. In practice, several EDM research employ demographic characteristics into their models for prediction [10]. From 2007 to 2018, researchers assessed the features most commonly utilized in EDM studies to predict student achievement. Six demographic factors namely age, gender, marital status, income, nationality, and job status constitute the ten most frequently utilized with gender being the most popular [11]. The underlying and occasionally apparent explanation behind both the presented reasoning and the frequent application of demographic variables is that these features are important in the education context. The key outcome was academic achievement, as measured by the cumulative grade point average (GPAX) of medical school graduates. The elements influencing the results obtained have been assessed in the literature and classified into four categories academic activities, demographics, learning style or psychology and environment,. According to student data, the requirements for inclusion are graduation from April 2010 to April 2023 for minimizing the recollection bias as well as assure voluntary achievement about all the survey questions.

2. LITERATURE REVIEW

According to the literature study, it is necessary for enhancing educational quality by forecasting the performance of students' academic as well as aiding students who are at risk. the study review , the academic performance prediction was made using numerous and different factors, such as different digital traces departed by students on the internet through browsing, lesson time, percentage of participation as well as student's demographic characteristics namely gender, economic status, age, number of courses attended, internet access, etc. [12], skills for learning, study methodologies, learning strategies, study habits, social support perception, motivation, socio-demography, health status, academic performance features [13], assignments, projects, quizzes, etc. [14]. Bernacki et al. investigated if log records in the educational management system were adequate for predicting performance. The behaviour-based model for prediction accurately identified 75% of individuals who might be required to retake the training [15]. According to Alshanqiti, A., and Namoun, A., introduced this approach allows students that may struggle in coming semesters to be identified as well as supported. Predicting the student academic performance accuracy needs an in-depth knowledge of the factors along with features which affect student results as well as accomplishment [16].

The researchers used student academic data, student activity, as well as student video interactions in predicting whether a student would receive a passing or failing grade at the conclusion of the semester. Another investigation used quizzes, discussions, assignments, attendance, and lab work for predicting Semester Grade Point Average (SGPA). Pre-university traits as well as preceding academic achievement have been used for forecasting SGPA, which predicted total performance based on grades from the preceding four semesters of study (17). Modern techniques have been employed to forecast final performance on tests based on demographics, previous performance and student engagement. The Artificial Neural Networks (ANN) method obtained excellent precision using student involvement as well as historical performance, but demographic factors have been shown to be insignificant. In accordance with current e-learning methodologies, the behavior classification-based E-learning Performance (BCEP) along with the process behavior classification (PBC) models have been proposed [18]. The research studies have been conducted on the Open University Learning Analytics Dataset (OULAD) to predict e-learning performance, and the findings revealed that the proposed models outperformed established methods. In contrast to conventional medical education is based on passive memorizing of knowledge, current research is developed additional techniques that emphasize creativity and inventive engagement in learning [19]. Notably, educational evaluation may work with a wide range of dimensions along with information sources, including categorical data over attitude or learning style as well as GPA data with continuous category. As a result, it is a one-of-a-kind tool capable of collecting and mining various sorts of data for specific outputs.

3. RESEARCH METHODOLOGY

The retrospective cohort investigation was carried out utilizing an online standard questionnaire with 13 questions along with a test of learning style. The questionnaire has been distributed to all medical students who graduate from Phramongkutklao College of Medicine from April 2010 to April 2023. It took place between January 2022 and April 2023 which contained questions about various periods of the 6-years medical student career.

TABLE 1. Study factor and its description for various educational domains

Educational Domains	**Study factors**	**Description**
Demographics	Sex	Male / Female
	Family Address	Metropolis / Others
	Relationship status	Single / Married
	BMI Status	Underweight/Normal/Overweight/Obesity
	Family Income	Amount in Rs. Per Month
Academic Activities	Online Learning	Hybrid to entire Online / None
	Time for study	1/2/3/4
	Course session	3.5/3.5-3.75/≥3.75
	Attendance	Medicine/surgery/pediatrics/ Obstetrics & Gynecology
	Interest in specialties	<1/1-2/ >2h per day
	GPAX in pre-medical school	<3.25 (low grade) / ≥3.25 (High grade)
Environment	Learning Resource	Wi-Fi / Computer / iPad
	Transportation Time	30/30 – 60/>60 min
Psychology	Learning Style	Visual Score (1 to 16)
		Reading (1 to 16)
		Auditory (1 to 16)
		Kinetics (1 to 16)

Figure 1 illustrate the four different process namely problem definition and formulation, data collection or gathering, Analysis and prediction. The study modified from IBM's 2015 foundation Approach for Data Science (FAD). The GPAX is classified as low grade or high grade level by 3.25 criterions. The analysis encompasses with supervised learning with LR method has employed for classifying outcome with respect to their likelihood of occurrence. In the present research, LR method was chosen as a suitable approach for examining the association among multiple predictor factors including a binary result in academic performance. Its ability to determine modified risk parameters or relationships when optimizing for variables and covariates is consistent with the research's objectives. The Adjusted Prevalence Ratio (APR) was employed to assess the risk related to the likelihood of a scenario happening. Moreover, the'margins' function was examined to determine projected probabilities.

Problem definition and formulation

According to the dataset considered in this research involves 145 medical students who have involved in the survey questionnaires with totally 13 variables which is segregated into four different educational domains of the students. This is done through the influence factors of the medical student success and even understands the education al system of the medical student progress.

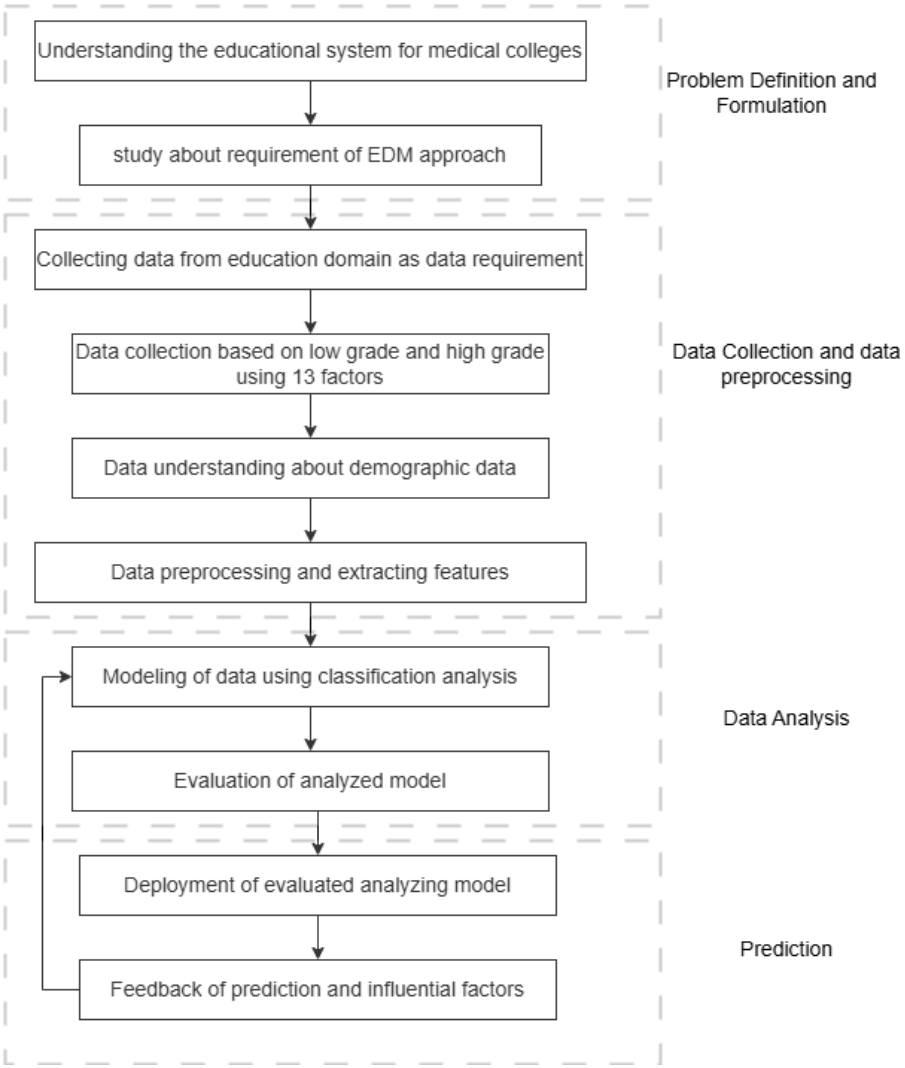


FIGURE 1. Flow chart for EDM for analyzing medical college data

Data collection and data preprocessing

This process involves collection response for the questionnaires from 145 medical students who have involved in the survey with totally 13 variables which is segregated into four different educational domains of the students. In data collection, the data requirement is collected and segregated through four domains namely demographic, academic activity, environment and life style. The student population is segregated through target value as low grade or high grade using 13 different factors. Moreover, demographic data is a descriptive type of dataset. Finally the process end with data preparation using data cleansing and eliminating features that are correlated to avoid confusion between correlation and causation. Features had been selected in accordance with p-value criterion of <0.20 or clinical significance.

Data Analysis

Backward step-wise optimization selection was used to discover which feature was most helpful attribute in classifying cases with the generation with no improvements is considered to be 1. Univariate and multivariate LR method were used to obtain frequency ratios as well as adjusted prevalence ratios. The p-value <0.05 represents as statistical significance. The two groups have significantly different baseline characteristics, including age, BMI, and baseline (pre-med) as GPAX. The high-grade group average age is projected gettable two years younger, implying a shorter time to graduation. However, there were no variations between educational course sessions. As a result, age and interval time was not deemed influential factors, as well as course sessions were employed for analysis

alternatively. Although the whole sample's about average BMI suggested underweight, the group of low-grade had an average weight distribution, if somewhat lower and close to the threshold limit.

Prediction

Univariate and multivariate LR method are used to obtain frequency ratios as well as adjusted prevalence ratios. The p-value <0.05 indicated as statistical significance. This survey is completed by 145 students, 72 of whom are male and 73 were female. The bulk of responders (63.4%) are from the high-grade group, with an average age of 27.4 ± 3.1 years. A large percentage of students (79.3%) lived in metropolitan regions and commuted to school in under thirty minutes (72.4%), relationship statuses with a balanced distribution. Ultimately, the starting point pre-med GPAX is decreased by 4% in the low-grade group, average about 3.62 compared with 3.78 over group of high-grade. There are no disparities among the two groups with respect to gender, family address, relationship status, or income. The majority of medical students' learning styles are kinesthetic (36.3%), which was investigated using four metrics like visual, mixed, reading and auditory. There are several discrepancies in traveling period and specialty of interest, which will be investigated subsequently for better prediction of medical student success.

4. RESULT AND DISCUSSION

The research conducted effectively used LR, FOR uncover factors influencing academic achievement between medical students with a special emphasis on visualization approaches, as depicted in the tree-based framework. Each method contributed new findings, expanding the comprehension of about academic, environmental, psychological and demographic factors influence student performance. LR method revealed that a pre-med GPAX of ≥ 3.75 as well as a keen interest over medical education are statistically significant educational achievement prediction (APR 1.72, p = 0.005; APR 1.51, p = 0.041) correspondingly. Margins analysis revealed a predicted mean GPAX is 3.41 (p < 0.005) to the students with a pre-med GPAX of ≥3.75 shown in table 2. The outcome is consistent with earlier research demonstrating GPAX is academic performance with good predictor. Although baseline GPAX has unalterable, continued support as well as tailored interventions might assist students with less pre-med GPAX, highlighting the importance of prior detection and assistance of academic. Furthermore, motivation for area of expertise selection has been shown to impact academic result with internal medicine demonstrating a substantial connection.

TABLE 2. Multivariable LR analysis of factor correlate with High GPAX of medical students

Attributes	Total n (%)	High-Grade of n (%)	Multivariable	
			APR (95% CI)	P-Value
Gender				
Male	72(49.8%)	44(61.2%)	1	
Female	73(50.2%)	48(65.7%)	1.12(0.90 – 1.42)	0.278
Relationship Status				
Single	71(49%)	47(66.2%)	1	
Married	74(51%)	45(60.8%)	0.92(0.72 – 1.18)	0.501
Family Address				
Metropolis	115(79.3%)	16(53.3%)	1	
others	30(20.7%)	76(66.1%)	1.24(0.87 – 1.78)	0.242
BMI				
18.5 - 22.9	56(38.6%)	30(53.6%)		
< 18.5	74(51%)	54(73.0%)		
23 – 24.9	8(5.5%)	5(62.5%)		
≥ 25	7(4.8%)	3(42.9%)		
Course session				
Session 1 (< 3 years)	61(42.1%)	40(65.6%)	1	
Session 2 (4 – 5 years)	28(19.3%)	16(57.1%)	0.61(0.34 – 1.11)	0.108
Session 3 (6 – 8 years)	40(27.6%)	24(60.0%)	0.70(0.40 – 1.22)	0.206
Session 1 (> 8 years)	16(11%)	12(75.0%)	0.96(0.57 – 1.61)	0.865
GPAX (Pre-medical school)				
< 3.5	21(14.5%)	8(38.1%)	1	
3.5 – 3.75	31(21.4%)	20(64.5%)	1.69(0.98 – 2.91)	0.058
≥ 3.75	93(64.1%)	64(68.8%)	1.73 (1.02 – 2.91)	0.040
Family Income (Thai Baht per month)				
< 20,000	7(4.8%)	5(71.4%)		

20,000 – 1,000,000	88(60.7%)	58(65.9%)		
> 1,000,000	50(34.5%)	29(58.0%)		
Learning Resources				
Wi - Fi	3(2.1%)	3(100.0%)		
Desktop	11(7.6%)	6(54.5%)		
iPAD	131(90.3%)	83(63.4%)		
Online learning Attendance				
Hybrid to complete online	75(51.7%)	52(69.3%)		
No online learning	70(48.3%)	40(57.1%)		
Transportation time				
<30 min.	105(72.4%)	74(70.5%)	1	
30 to 60 min.	30(20.7%)	13(43.4%)	0.74(0.52 – 1.04)	0.080
>60 min.	10(6.9%)	5(50.0%)	0.59(0.30 – 1.14)	0.118
Specialties of Interest				
Minor	60(31.9%)	30(50.0%)	1	
Internal Medicine	35(30.4%)	32(91.4%)	1.52(1.14 – 2.03)	0.005
Obstetrics & Gynecology	20(15.4%)	9(45.0%)	0.54(0.20 – 1.49)	0.235
Surgery	17(12.8%)	11(64.7%)	0.84(0.34 – 2.08)	0.709
Pediatrics	13(9.5%)	10(76.9%)	0.98(0.41 – 2.33)	0.967
Time for study per day				
<1Hr	68(46.9%)	42(61.8%)		
1 to 2 Hr	66(45.9%)	42(63.6%)		
>2Hr	11(7.6%)	8(72.7%)		
Learning Style				
Visual	35(24.1%)	21(60.0%)	1	
Auditory	24(16.6%)	15(62.5%)	1.06(0.74 – 1.52)	0.743
Reading	13(9%)	8(61.5%)	1.06(0.68 – 1.67)	0.798
Kinetics	53(36.6%)	34(64.2%)	1.03(0.75 – 1.41)	0.853
Mixed	20(13.7%)	14(70.0%)	1.18(0.84 – 1.66)	0.348

The studies shown above demonstrate the importance of factors in "academic activities." However, it observed no statistical significance for environmental, demographic, or psychological variables. Hence, the discovery of the styles of learning as well as students with high-grade refers to highly achieving students, suggests that these students may learn in various fashions.

5. CONCLUSION

Researchers used three analytical methodologies to identify critical parameters influencing medical students' academic progress. LR has identified pre-med GPAX as well as interest in medical education as significant factors. These findings highlight the necessity of a comprehensive educational planning approach that takes into account learning inclination, personal wellbeing, and supporting settings. In conclusion, using a multi-domain framework over educational planning may lead to improved outcomes and tailored learning opportunities. Individualized learning styles, adaptable teaching methodologies, and encouraging environments may assist medical students accomplish their professional and educational objectives.

REFERENCES

[1]. Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*(October 2019), 106189. https:// doi. org/ 10. 1016/j. chb. 2019. 106189

[2]. BAKER, R. S., ESBENSHADE, L., VITALE, J., KARUMBAIAH, S., ET AL. 2023. Using demographic data as predictor variables: a questionable choice. Journal of Educational Data Mining 15, 2, 22–52.

[3]. CASTELNOVO, A., CRUPI, R., GRECO, G., REGOLI, D., PENCO, I. G., AND COSENTINI, A. C. 2022. A clarification of the nuances in the fairness metrics landscape. Scientific Reports 12, 1, 4209.

[4]. MEHRABI, N., MORSTATTER, F., SAXENA, N., LERMAN, K., AND GALSTYAN, A. 2021. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) 54, 6, 1–35.

[5]. CATON, S. AND HAAS, C. 2024. Fairness in machine learning: A survey. ACM Computing Surveys 56, 7, 1–38.
[6]. KIZILCEC, R. F. AND LEE, H. 2022. Algorithmic fairness in education. In The ethics of artificial intelligence in education. Routledge, New York, USA, 174–202.
[7]. HU, Q. AND RANGWALA, H. 2020. Towards fair educational data mining: A case study on detecting at-risk students. In 13th International Conference on Educational Data Mining. International Educational Data Mining Society, Ifrane, Morocco.
[8]. YU, R., LEE, H., AND KIZILCEC, R. F. 2021. Should college dropout prediction models include protected attributes? In Proceedings of the eighth ACM conference on learning@ Scale. Association for Computing Machinery, New York, USA, 91–100.
[9]. Aleem, A, and Gore, MM, editors. Educational data mining methods: A survey. 2020 IEEE 9th international conference on communication systems and network technologies (CSNT. 10–12. (2020).
[10]. PAQUETTE, L., OCUMPAUGH, J., LI, Z., ANDRES, A., AND BAKER, R. 2020. Who is learning? using demographics in edm research. Journal of Educational Data Mining 12, 3, 1–30.
[11]. ALTURKI, S., HULPUS, , I., AND STUCKENSCHMIDT, H. 2022. Predicting academic outcomes: A survey from 2007 till 2018. Technology, Knowledge and Learning 27, 1, 275–307.
[12]. Rebai, S., Ben Yahia, F., & Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. Socio-Economic Planning Sciences, 70(August 2018), 100724. https:// doi. org/ 10. 1016/j. seps. 2019. 06. 009
[13]. Costa-Mendes, R., Oliveira, T., Castelli, M., & Cruz-Jesus, F. (2020). A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach. Education and Information Technologies, 26, 1527–1547. https:// doi. org/ 10. 1007/ s10639- 020- 10316-y
[14]. Kardaş, K., & Guvenir, A. (2020). Analysis of the effects of Quizzes, homeworks and projects on final exam with different machine learning techniques. EMO Journal of Scientific, 10(1), 22–29.
[15]. Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. Computers & Education, 158(August), 103999. https:// doi. org/ 10. 1016/j. compe du. 2020. 103999
[16]. Alshanqiti, A., & Namoun, A. (2020). Predicting student performance and its influential factors using hybrid regression and multi-label classification. IEEE Access, 8, 203827–203844. https:// doi. org/ 10. 1109/ access. 2020. 30365 72
[17]. Tomasevic, N., Gvozdenovic, N. & Vranes, S. Computers & Education An overview and comparison of supervised data mining techniques for student exam performance prediction. Comput. Educ. 143, 103676 (2020).
[18]. Qiu, F. et al. Predicting students performance in e - learning using learning process and behaviour data. Sci. Rep. https:// doi. org/ 10. 1038/ s41598- 021- 03867-8 (2022).
[19]. Hindin, D. I., Mazzei, M., Chandragiri, S., DuBose, L., Threeton, D., Lassa, J., et al. (2023). A National Study on training innovation in US medical education. Cureus. 15:e46433. doi: 10.7759/cureus.46433.

