Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE
Paper ID #46960

BOARD #102: Work in Progress: Enhancing Transparency in Educational Decision-Making using XAI Technique

Eesha tur razia babar
Mr. Ahmed Ashraf Butt, University of Oklahoma

Dr. Ahmed Ashraf Butt is an Assistant Professor at the University of Oklahoma. He recently completed his Ph.D. in the School of Engineering Education at Purdue University and pursued post-doctoral training at the School of Computer Science, Carnegie Mellon University (CMU). He has cultivated a multidisciplinary research portfolio bridging learning sciences, Human-Computer Interaction (HCI), and engineering education. His primary research focuses on designing and developing educational technologies that facilitate various aspects of student learning, such as engagement. Additionally, he is interested in designing instructional interventions and exploring their relationship with first-year engineering (FYE) students' learning aspects, including motivation and learning strategies. Prior to his time at Purdue, Dr. Butt worked as a lecturer at the University of Lahore, Pakistan, and has been associated with the software industry in various capacities.

Work in Progress: Enhancing Transparency in Educational Decision-Making using XAI Technique

Introduction

In the evolving landscape of education, the integration of data-driven insights using Educational Data Mining (EDM) techniques has revolutionized how educators understand and enhance learning processes. EDM is an interdisciplinary field that uses techniques from data analysis, machine learning, and statistics to extract insights from educational data [1]. These advances have enabled significant progress in innovative educational practices, such as personalized learning, and predictive analytics[2]. However, a prominent constraint of most EDM techniques is the lack of transparency[3] in their decision-making process.

Most traditional data mining algorithms, such as classification, have the tendency to provide high prediction accuracy in various educational tasks, such as predicting student performance (e.g., [4]), detecting learning disabilities (e.g., [5]), and recommending learning content based on learning styles (e.g., [6]). However, these algorithms act as “black boxes”, often don’t provide reasoning behind their decisions. This lack of transparency can be a challenge for educators and researchers who need to understand the underlying factors that influence the outcomes of these decisions. For example, if an algorithm identifies a student at risk of failing in a classroom without explaining the factors contributing to this decision. It would be harder for the instructor to determine whether the prediction is based on attendance, grades, or other factors, making it harder to provide the needed support.

Without this transparency and understanding, the full potential of data mining techniques to extract data-driven insights in education remains untapped. This limitation highlights the necessity of techniques that can reveal how traditional data mining algorithms make decisions, a field known as Explainable AI (XAI). XAI addresses this challenge by improving the interpretability of data mining algorithms. Hence, the objective of this study is to understand how XAI can enhance the transparency of EDM algorithms, which can help educators and researchers make informed decisions to improve student learning outcomes.

The remainder of the paper first briefly discusses previously used EDM techniques and the need for XAI. After that, data collection, data pre-processing, implementation of EDM techniques on processed data, and results of implementing XAI are covered. In the end, we discuss our study's findings and limitations.

Literature Review

In recent years, educational institutions and researchers have increasingly used EDM techniques to understand various aspects of students' learning and enhance their learning experience [7][8][9][10][11]. For example, studies have used classification techniques to understand the different groups of learners [e.g.,[12][13]]. Also, ensemble learning techniques (e.g., Random Forest) have been used to understand students' performance on tasks [14][15]. Furthermore, the prediction of students' dropout rate is another critical area of application of EDM techniques. For example, Adejo et al. used a heterogeneous multi-modal ensemble approach to predict student academic performance and help identify students at risk of dropping out [16]. Another study used a Neuro-Fuzzy Algorithm for predictive analysis to understand student retention in an engineering program[17].

Another significant application of educational data mining is examining students' mental health, enabling early intervention. Several EDM techniques, such as Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Ensemble methods, and Linear Discriminant methods, have been employed by researchers to identify students' mental health problems[18][19][20]. As evident from the literature, significant progress has been made in extracting learning analytics using EDM techniques. However, a persistent challenge is the lack of transparency in how these models make decisions. Sujan et al. [21] argued that although data mining algorithms can provide high accuracy, their “black boxes” nature makes it difficult for educators to understand the underlying factors of their decisions. In this regard, XAI has emerged as a solution to this transparency issue. Recent efforts to integrate XAI into educational data mining have been limited but promising.

SHAP (SHapley Additive exPlanations)[22] is an XAI technique widely used to explain model predictions in various domains, including education. This technique quantifies the contribution of each factor (or feature) that influences an AI model's decision, increasing its transparency and interoperability. Yeonju Jang et al. [23] used XAI to enhance understanding of at-risk students. In this study, SHAP global explanations identified specific homework evaluation scores as the key predictor, while local explanations highlighted specific assignments and peer enrollment as the most important features to find at risk students. Researchers concluded this insight could help educators provide targeted support. Asma Ul Hussna et al. [24] applied SHAP to assess the impacts of COVID-19 on students' social lives, mental health, and education using a range of predictive models. These studies show the growing importance of SHAP in making complex models more interpretable, thereby enhancing the transparency of educational data mining processes.

Despite significant advancements in EDM algorithms, the integration of XAI techniques, such as SHAP, remains limited, with most studies focusing on a single application. This paper addresses this gap by demonstrating the applicability of XAI across three distinct EDM scenarios: predicting students' self-reported physical health status, academic success and dropout risk, and overall academic performance[25][26][27]. Each scenario is analyzed using a separate dataset, focusing on classification tasks while incorporating SHAP to enhance interpretability and transparency.

Methods

The first dataset includes data on 886 Swiss medical students, and 20 characteristics such as students' demographics (age, sex, etc.), study details (weekly study hours), personal status (partner, job), health metrics (self-reported health, depression, anxiety) and empathy/burnout scores. The second dataset has the data of 4,423 students, with 37 features that include students' marital status, previous qualification, nationality, parents' qualifications, etc. The third dataset has data of 80 students with 16 features, such as gender, nationality, place of birth, and classroom participation (raised hands, visited resources, announcements viewed, discussion participation), etc. A detailed description of each dataset is given in Appendix A.

Firstly, the necessary pre-processing procedures were conducted, such as applying various undersampling/oversampling techniques, e.g., SMOTE (synthetic minority oversampling

Table 1: Binary Classification Results of Medical Students' mental health dataset

Model	Accuracy	F1 Score	Recall	Precision
Decision Tree	0.73	0.82	0.76	0.89
Random Forest	0.74	0.82	0.87	0.78
Logistic Regression	0.67	0.75	0.81	0.69
Multilayer Perceptron	0.71	0.79	0.79	0.80
XG Boost	0.67	0.77	0.76	0.78
Ada Boost	0.71	0.79	0.76	0.81
Naive Bayes	0.71	0.78	0.73	0.84

technique) [28] to deal with class imbalance problem of datasets. After that statistical analysis and data visualization tasks were performed, and subsequently various EDM algorithms were applied. Finally, SHAP (i.e., XAI technique) was applied to provide global explanations for the decision-making process of EDM algorithms [29].

Experiments

Medical Students' mental health dataset

The self-reported physical health status of participant(1-5) was used as the dependent variable for this dataset with 1 means very dissatisfied, and 5 means very satisfied. Scores of 1-3 were considered as class 0 (low health index), and 4-5 as class 1 (high health index) for binary classification. The results of applying the EDM algorithms to this dataset are shown in Table 1. We found that RF achieved the highest accuracy (0.74) therefore SHAP explanations are generated for RF model using the original dataset (without SMOTE), and are discussed in the Results and Discussion sections.

Students' dropout and academic success dataset

In this dataset, there are three main classes in the dependent variable, showing whether a student “graduated” (class 0), stayed “enrolled” (class 1) or “dropped out” (class 2) at the end of the normal duration of an undergraduate degree. Therefore, the task in this data set is multi-class classification. Table 2 shows the result of applying different EDM techniques for this task. We found that XG Boost achieved the highest accuracy (0.79) therefore SHAP explanations are generated for XG Boost model using the original dataset (without SMOTE), and are discussed in the Results and Discussion sections.

Students' academic performance dataset

For this data set, students' end-semester exam marks, were used as the dependent variable. The student's marks in the end-semester exams are divided into three classes: low (class 0), medium (class 1), high (class 2), making them suitable for a multi-class classification problem. The results of applying the EDM algorithm to this data set are shown in Table 3. We found that XG Boost achieved the highest accuracy (0.85) therefore SHAP explanations are generated for XG Boost

Table 2: Multi-Class Classification Results of Students' dropout and academic success dataset

Model	Accuracy	F1 Score	Recall	Precision
Decision Tree	0.71	0.65	0.65	0.65
Random Forest	0.76	0.70	0.71	0.70
Logistic Regression	0.74	0.70	0.70	0.70
Multilayer Perceptron	0.68	0.63	0.63	0.63
XG Boost	0.79	0.73	0.74	0.72
Ada Boost	0.74	0.69	0.69	0.70
Naive Bayes	0.63	0.61	0.62	0.63

Table 3: Multi-Class Classification Results of Students' academic performance dataset

Model	Accuracy	F1 Score	Recall	Precision
Decision Tree	0.79	0.78	0.78	0.78
Random Forest	0.83	0.82	0.83	0.82
Logistic Regression	0.81	0.81	0.83	0.79
Multilayer Perceptron	0.76	0.75	0.77	0.74
XG Boost	0.85	0.85	0.86	0.84
Ada Boost	0.80	0.79	0.79	0.79
Naive Bayes	0.77	0.77	0.79	0.76

model using the original dataset (without SMOTE), and are discussed in the Results and Discussion sections.

Results and Discussions

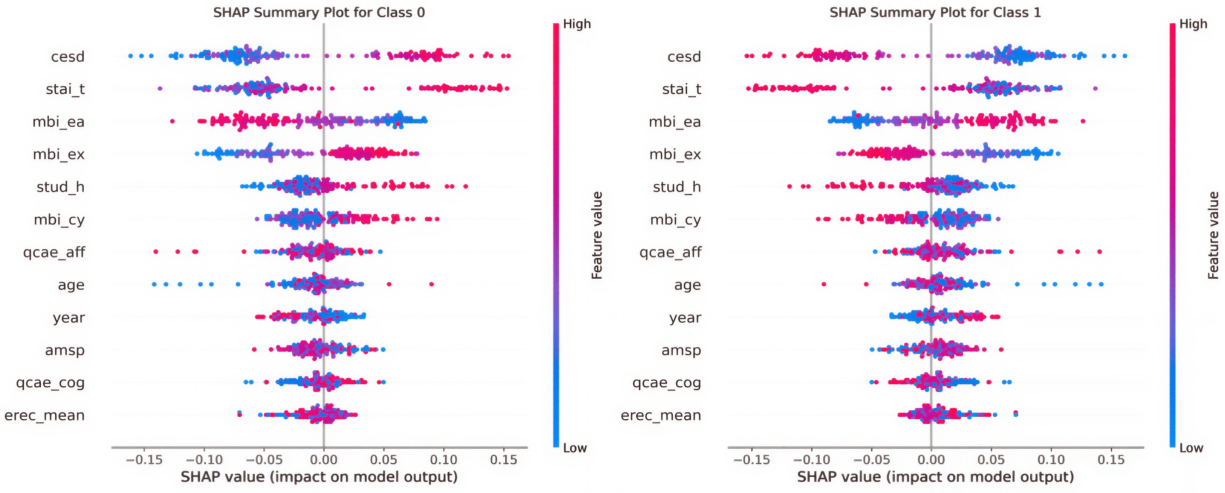

Figure 1: SHAP summary plots highlighting key features and their impact on RF's decisions for each class in the medical students' mental health dataset

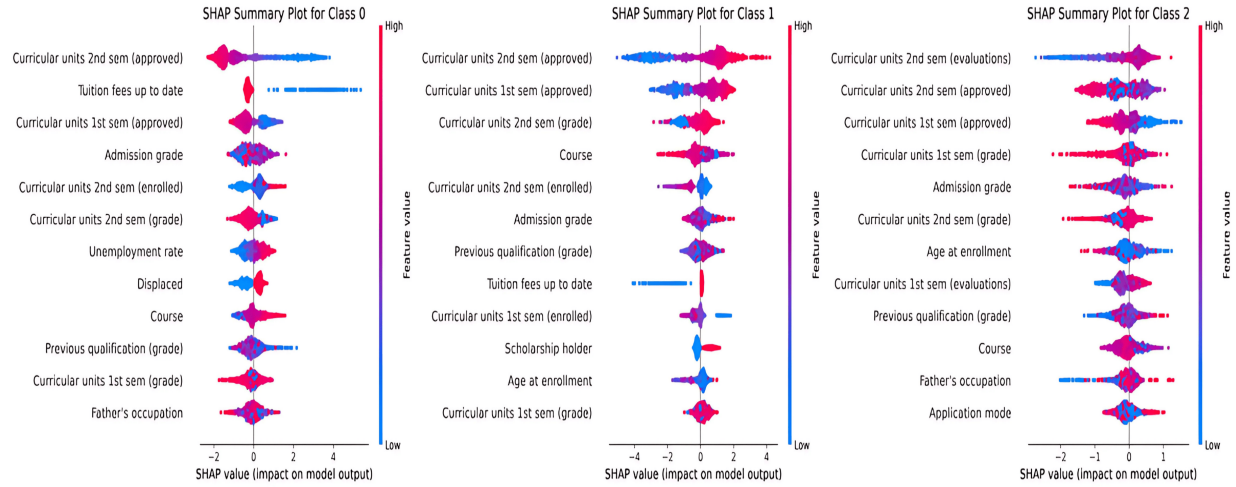

Figure 2: SHAP summary plots highlighting key features and their impact on XGBoost's decisions for each class in the students' dropout and academic success dataset

For each dataset, we used SHAP-based methods for feature selection by training models on the entire data set to generate SHAP values. In this method, SHAP values are computed for each data point, which are then aggregated to find the average absolute SHAP value for each feature. This value indicates the feature's overall impact on model predictions. By ranking features in descending order based on their average absolute SHAP values, we identified the most important features for each data set[30]

For medical student data set, average absolute SHAP values showed that cesd (Center for Epidemiologic Studies Depression scale of the participant (self-reported depression level)) and stat_i (State-Trait Anxiety Inventory scale of the participant (self-reported anxiety level)) are the top two most important features impacting the perceived health status of students.

After finding the overall most important features, we generated SHAP summary plots for this data set to highlight key features for each class and their relationships with that class. In SHAP summary plot, y-axis ranks features by importance, while the x-axis represents SHAP values for these features. Each point corresponds to a data row. Fig. 1 shows that high cesd values (red points) have positive SHAP values for class 0, meaning high cesd favors class 0, while low cesd (blue points) favors class 1. Similarly, high stai_t values favor class 0, and low values favor class 1. This indicates students with higher cesd and stai_t scores are more likely to belong to class 0 and vice versa. These findings align with prior research [31], showing that mental health issues often lead to physical health problems, increasing our understanding of mental and physical health issues faced by college students.

For the academic success and dropout dataset, SHAP values identified curricular units approved in the second semester as the most important feature. Fig. 2 shows higher values of it negatively correlate with class 0, moderately positive with class 1, and distributed between moderate positive to moderate negative with class 2. This suggests students with fewer approved units in the second semester are more likely to graduate on time, aligning with prior studies and enhancing our

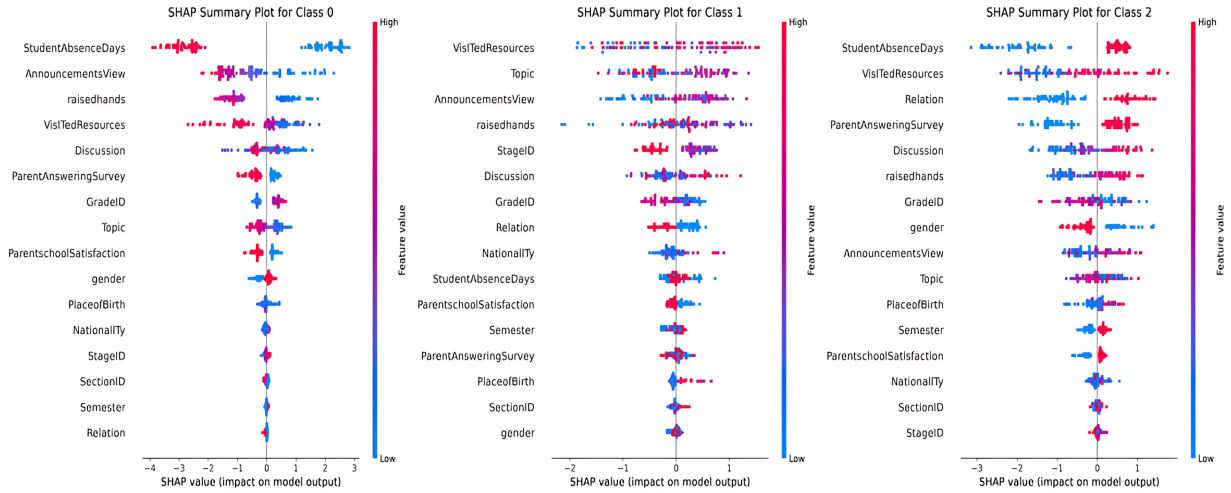

Figure 3: SHAP summary plots highlighting key features and their impact on XGBoost's decisions for each class in the students' academic performance dataset

understanding of factors impacting students academic success[32].

For the student academic performance dataset, SHAP values identified VisitedResources as the most important feature. Fig. 3 shows its relationship with each class: higher values have a strong negative correlation with class 0, a moderate positive correlation with class 1, and a positive correlation with class 2. This suggests that students who access resources more frequently tend to perform better. These findings align with previous studies, enhancing our understanding of factors influencing student performance [33].

In summary, SHAP values identified key factors influencing student mental health, academic success and dropout rates, and academic performance, aligning with prior studies. Previous studies mostly showed the important features; however, this study also showed their specific relations with each class. Hence, the specific contributions of this work are providing guidelines for future researchers to employ the XAI technique to enhance their finding's transparency and interpretability.

Conclusion and Future Directions

RF achieved the highest accuracy for the first dataset, highlighting cesd and stati as key features. For the second and third datasets, XGBoost performed best, with SHAP identifying curricular units approved in the second semester as critical for academic success and dropout prediction and visited resources as the top factor for academic performance. These findings show how XAI enhances transparency in EDM algorithms, helping educators understand student outcomes. Since it is an exploratory study, a limitation is the reliance on a single dataset per application, which may limit generalizability, and transparency may be skewed if the dataset is biased. Other limitations include focusing only on classification problems and studying only the top 1-2 features. Additionally, this work only includes global interpretability. Future work will expand to regression problems and incorporate local interpretability techniques like LIME and Eli5.

References

[1] O. Scheuer and B. M. McLaren, "Educational data mining," in Encyclopedia of the Sciences of Learning, Boston, MA: Springer US, 2012, pp. 1075–1079.

[2] F. Alshareef, H. Alhakami, T. Alsubait, and A. Baz, "Educational Data Mining Applications and Techniques," International Journal of Advanced Computer Science and Applications, vol. 11, 2020.

[3] T. Zarsky, "Transparency in data mining: From theory to practice," in Studies in Applied Philosophy, Epistemology and Rational Ethics, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 301–324.

[4] S. Roy and A. Garg, "Predicting academic performance of student using classification techniques," in 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), 2017, pp. 568–572.

[5] W. F. Horn and J. P. O'Donnell, "Early identification of learning disabilities: A comparison of two methods," J. Educ. Psychol., vol. 76, no. 6, pp. 1106–1118, 1984.

[6] A. Klašnja-Milićević, B. Vesin, M. Ivanović, and Z. Budimac, "E-Learning personalization based on hybrid recommendation strategy and learning style identification," Comput. Educ., vol. 56, no. 3, pp. 885–899, 2011.

[7] King Abdulaziz University, Saudi Arabia, Jeddah, A. Mueen, B. Zafar, and U. Manzoor, "Modeling and predicting students' academic performance using data mining techniques," Int. J. Mod. Educ. Comput. Sci., vol. 8, no. 11, pp. 36–42, 2016.

[8] I. A. Abu Amra and A. Y. A. Maghari, "Students performance prediction using KNN and Naïve Bayesian," in 2017 8th International Conference on Information Technology (ICIT), 2017, pp. 909–913.

[9] H. Almarabeh, "Analysis of students' performance by using different data mining classifiers," International Journal of Modern Education and Computer Science, vol. 9, pp. 9–15, 2017.

[10] H. Abu Mousa and Ashraf Y. A. Maghari,"School Students' Performance Predication Using Data Mining Classification", 2017, https://api.semanticscholar.org/CorpusID:210955062

[11] A. A. Butt, E. T. R. Babar, M. Menekse, and A. Alhaddad, Board 62: Work in progress: A Comparative Analysis of Large Language Models and NLP Algorithms to Enhance Student Reflection Summaries. 2024. doi: 10.18260/1-2–47060.

[12] Yass Khudheir Salal and Sanjar Abdullaev and Mukesh Kumar, "Educational Data Mining : Student Performance Prediction in Academic"

[13] A. I. Adekitan and E. Noma-Osaghae, "Data mining approach to predicting the performance of first year student in a university using the admission requirements," Educ. Inf. Technol., vol. 24, no. 2, pp. 1527–1543, 2019.

[14] K. Agarwal, E. Maheshwari, C. Roy, M. Pandey, and S. S. Rautray, "Analyzing student

performance in engineering placement using data mining," in Lecture Notes on Data Engineering and Communications Technologies, Singapore: Springer Singapore, 2019, pp. 171–181.

[15] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," Heliyon, vol. 5, no. 2, p. e01250, 2019.

[16] O. W. Adejo and T. Connolly, "Predicting student academic performance using multi-model heterogeneous ensemble approach," J. Appl. Res. High. Educ., vol. 10, no. 1, pp. 61–75, 2018.

[17] M. Adil, F. Tahir, and S. Maqsood, "Predictive analysis for student retention by using neuro-fuzzy algorithm," in 2018 10th Computer Science and Electronic Engineering (CEEC), 2018, pp. 41–45.

[18] A. A. Sabourin, J. C. Prater, and N. A. Mason, "Assessment of mental health in doctor of pharmacy students," Curr. Pharm. Teach. Learn., vol. 11, no. 3, pp. 243–250, 2019.

[19] Y. Hou, J. Xu, Y. Huang, and X. Ma, "A big data application to predict depression in the university based on the reading habits," in 2016 3rd International Conference on Systems and Informatics (ICSAI), 2016, pp. 1085–1089.

[20] A. Hasanbasic, M. Spahic, D. Bosnjic, H. H. Adzic, V. Mesic, and O. Jahic, "Recognition of stress levels among students with wearable sensors," in 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), 2019, pp. 1–4.

[21] S. Ghimire et al., "Explainable artificial intelligence-machine learning models to estimate overall scores in tertiary preparatory general science course," Computers and Education: Artificial Intelligence, vol. 7, no. 100331, p. 100331, 2024.

[22] "Welcome to the SHAP documentation — SHAP latest documentation," Readthedocs.io. [Online]. Available: https://shap.readthedocs.io/en/latest/. [Accessed: 14-Jan-2025].

[23] Y. Jang, S. Choi, H. Jung, and H. Kim, "Practical early prediction of students' performance using machine learning and eXplainable AI," Educ. Inf. Technol., vol. 27, no. 9, pp. 12855–12889, 2022.

[24] A. Ul Hussna, I. Immami Trisha, I. Jahan Ritun, and M. G. Rabiul Alam, "COVID-19 impact on students' Mental Health: Explainable AI and Classifiers," in 2021 International Conference on Decision Aid Sciences and Application (DASA), 2021, pp. 847–851.

[25] Kaggle.com. [Online]. Available: https://www.kaggle.com/datasets/thedevastator/medical-student-mental-health. [Accessed: 14-Jan-2025].

[26] "UCI machine learning repository," Uci.edu. [Online]. Available: https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success. [Accessed: 15-Jan-2025].

[27] Kaggle.com. [Online]. Available: https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data. [Accessed: 15-Jan-2025].

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[29] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," IEEE Access, vol. 5, pp. 15991–16005, 2017.

[30] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods," J. Big Data, vol. 11, no. 1, 2024.

[31] "The effects of Anxiety and depression on your physical health," Advanced Psychiatry Associates, 29-Jul-2024.

[32] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," Data (Basel), vol. 7, no. 11, p. 146, 2022.

[33] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods," Int. J. Database Theory Appl., vol. 9, no. 8, pp. 119–136, 2016.

A Appendix A: Data sets feature names, abbreviations and Data types

Table 4: Details of Medical Students' mental health dataset

Feature	Abbreviation	Type
age	Age of participants	numerical
year	Year of study of the participant	nominal
sex	Gender of the participant	nominal
glang	Language spoken by the participant	nominal
part	Does participant have a partner	nominal
job	Does participant have paid job	nominal
stud_h	Hours of study per week of the participant	numerical
health	Self-reported health status of the participant	nominal
psyt	Psychological distress score of the participant	nominal
jspe	Job satisfaction score of the participant	numerical
qcae_cog	Cognitive empathy score of the participant	numerical
qcae_aff	Affective empathy score of the participant	numerical
amsp	Academic motivation score of the participant	numerical
erec_mean	Empathy rating score mean of the participant	numerical
cesd	Center for Epidemiologic Studies Depression scale of the participant	numerical
stai_t	State-Trait Anxiety Inventory scale of the participant	numerical
mbi_ex	Maslach Burnout Inventory-Exhaustion scale of the participant	numerical
mbi_cy	Maslach Burnout Inventory - Cynicism Scale of the participant	numerical
mbi_ea	Maslach Burnout Inventory - Professional Efficacy Scale of the participant	numerical

Table 5: Details of students' dropout and academic success dataset

Feature	Abbreviation	Type
Marital status	The marital status of the student	nominal
Application mode	The method of application used by the student	nominal
Application order	The order in which the student applied	numeric
Daytime/evening attendance	Student attends classes during the day or in the evening	nominal
Previous qualification	Student's qualification before enrolling in higher education	nominal
Nationality	The nationality of the student	nominal
Mother's qualification	qualification of the student's mother	nominal
Father's qualification	qualification of the student's father	nominal
Mother's occupation	occupation of the student's mother	numeric
Father's occupation	occupation of the student's father	numeric
Displaced	Whether the student is a displaced person	numeric
Educational special needs	Whether the student has any special educational needs	numeric
Debtor	Whether the student is a debtor	numeric
Tuition fees up to date	Whether the student's tuition fees are up to date	nominal
Gender	The gender of the student	nominal
Scholarship holder	Whether the student is a scholarship holder	nominal
Age at enrollment	The age of the student at the time of enrollment	nominal

Table 6: Details of students' academic performance dataset

Feature	**Abbreviation**	**Type**
Gender	Student's gender	nominal
Nationality	Student's nationality	nominal
PlaceofBirth	Student's Place of birth	nominal
StageID	Section student belongs	nominal
GradeID	Grade student belongs	nominal
SectionID	Classroom student belongs	nominal
Topic	Course topic	nominal
Semester	School year semester	nominal
Relation	Parent responsible for student	nominal
raisedhands	How many times the student raise hand on classroom	numeric
VisITedResources	How many times the student visits content	numeric
AnnouncementsView	How many times the student checks announcements	numeric
Discussion	How many times the student participate on discussion groups	numeric
ParentAnsweringSurvey	Parent answered the surveys which are provided from school or not	nominal
ParentschoolSatisfaction	The Degree of parent satisfaction from school	nominal
StudentAbsenceDays	The number of absence days for each student	nominal
Class	Level of student depending upon obtained marks	nominal