Techno-Science Research Journal V13 (1) (2025) P 59-66

Content list available at ITC

Techno-Science Research Journal

Journal Homepage: http://techno-srj.itc.edu.kh/

Techno-Science Research Journal

Utilizing Data Mining And AI To Enhance Cambodian High School Student Performance And Stakeholder Success

Chhunheng Seirey[1*], Sokkhey Phauk[1], Say Ol[1]

[1] *Department of Applied Mathematics and Statistics, Institute of Technology of Cambodia, Russian Federation Blvd., P.O. Box 86, Phnom Penh, Cambodia*

Received: 12 July 2024; Revised: 20 August 2024; Accepted: 02 September 2024; Available online: 30 April 2025

Abstract: *Despite significant improvement in education levels within Cambodia over recent decades, high school failure rates remain a persistent concern. A quick intervention through the usage of digital transformation is needed. Business Intelligence (BI) and Educational Data Mining (EDM), which offer powerful tools for extracting valuable insights from raw data, hold notable potential to address this challenge within the education domain. This research investigates the application of BI/EDM techniques to predict student achievement in high school students and analyzes effective features to help high school students reach their full potential. Data encompassing recent real-world factors such as student grades, demographics, socio-economic and school-related features was gathered through questionnaires. Most input features were modeled under various binary/five-level classification tasks. Additionally, eight data mining models (Multinomial Naïve Bayes, Decision Tree, Bootstrap Aggregating, Stochastic Gradient Descent, K-Nearest Neighbors, Random Forest, Support Vector Machines, and Neural Networks) were tested, alongside different input selection strategies (including and excluding semester grade). Evaluation metrics like Accuracy, Recall, Precision, and F-measure were used to assess models. The results demonstrate that highly accurate predictions of student achievement are achievable, particularly when incorporating grades from the first semester. Neural Network achieved the best result with an Accuracy of 93.40%, Precision of 95.45%, Recall of 96.55%, and F-Measure of 96.00%. Apriori algorithm is used to identify the relevant features and make a recommendation to the high school student based on the predicted result. While past performance emerges as a notable predictor, the explanatory analysis reveals the influence of other relevant features, including parental background, parent support, parent motivation, self-learning, self-motivation, learning environment, and even student health. This research led to the development of a more efficient analysis and reporting service that will streamline data analysis within the Learning Management System (LMS). It has the potential to significantly improve the quality of education and optimize educational resource management.*

Keywords: Business Intelligence in Education; Learning Management System; Educational Data Mining; Classification; Association Rule

1. INTRODUCTION

Education is a key factor for achieving long-term economic progress. Education is indeed the fourth goal of Sustainable Development Goals(SDGs). During the last decades, the Cambodian educational level has improved. However, the statistics keep the high student failure rate. In 2023, the baccalaureate exam (BaccII) had a failure rate of 27.11% and 2,300 students did not complete the BacII.

The rise of BI and DM [1] can be attributed to advancements in Information Technology, which have led to an explosion of data within businesses and organizations. This vast amount of data holds valuable information in the form of trends and patterns that can be leveraged to optimize decision-making and success. However, human experts are limited and overlook crucial details. Thus, the alternative is to use automated tools to analyze the raw data and extract insightful, high-level information for decision-makers.

In effect, several studies have addressed similar topics. Educational data mining prediction of students' academic performance using machine learning algorithms [2]. They used datasets from the academic achievement grades of 1854 students who took the Turkish language in a state University in Turkey during 2019-2020. They used algorithms such as Neural Network, Logistics Regression, Support Vector Machine, Random Forest, Naïve Bayes and evaluation metrics like accuracy, precision, recall, and F-criterion. The best models (i.e. Neural Network and Random Forest) result achieve a

[*] Corresponding author: Chhunheng Seirey
E-mail: schhunheng.it@gmail.com; Tel: +855-96 323 4760

classification accuracy of 70-75% and prediction of students at high risk of failure. In 2021, An artificial intelligence approach to monitor student performance and devise preventive measures [3]. They used datasets from student academic records for a course taught et al. Buraimi University collects (BUC), Sultanate of Oman. They used algorithms such as ANN, Naïve Bayes, Association rules, and evaluation metrics like accuracy, recall, precision, and F-measure. The ANN model result achieves an accuracy of over 86%, and the recommendation module will automatically send personalized recommendations to the students in accordance with their current status. Using Data Mining to Predict Secondary School Student Performance [4]. They used datasets from secondary data in Portugal on Mathematics and Portuguese. They used algorithms such as Decision Tree, Random Forest, Support Vector Machine, Artificial Neural Network, and evaluation metrics like MAE, r-squared, accuracy, and recall. The best models (i.e. Artificial Neural Network, Decision Tree) give the result achieve a high predictive accuracy 95-97%, analysis of the important features, and prediction of students at high risk of failure. A hybrid DL model using a combination of Convolutional Neural Networks(CNNs) and Recurrent Neural Networks (RNNs) was proposed by [5] to predict student performance and discover the primary factor with the highest association with student performance. RNN was used to obtain the semantic connection between features, whereas CNN was used to collect the local dominant features and alleviate the curse of dimensionality. The results of the trials showed that the hybrid CNN-RNN prediction model outperformed the existing DL models, with an accuracy of roughly 79.23% when using data from the Kaggle repository. To predict student's success on the final exam, [6] suggested a Deep Convolutional Neural Network model called (DCNN) model that used data from the Institute of Science, Trade & Technology (ISTT) and included about 2844 records of 158 students. The model used a total of 18 data features and achieved an accuracy of about 98.33% accuracy. A study by [7] examined how clickstream data can be used to predict student performance. The study utilized the most important indicators of how well students would perform from the OULA dataset with weekly and monthly time intervals; these indicators included clicks on the homepage, related sites, quizzes, and content. It was found that the optimal method for predicting student performance was weekly-based click count aggregation in the form of panel data, along with the LSTM model. Because 1D-CNN is inadequate for handling sequential data, LSTM has achieved a greater prediction accuracy of up to 89.25% compared to 1D-CNN.In panel data with a matrix structure that denotes student click behaviors, LSTM is excellent for learning the features. [8] proposed a Sequential Engagement Based Academic Performance Prediction Network (SEPN). The network contained a Sequential Predictor (SP) and an Engagement Detector (ED). The SP has an LSTM structure and learns about the interaction between the demographic and engagement features. Based on weekly correlation data, the ED created a matrix from the everyday actions of the students to detect patterns of student engagement by taking advantage of the Convolutional Neural Network (CNN). An experiment on a real dataset (OULAD) was conducted, and the results demonstrated that SEPN performed better in prediction accuracy than the ML models when involving the ED mechanism.

In this work, we will analyze recent real-world data from Cambodian high school students in different provinces such as Kandal, Siem Reap, Battambang, and Phnom Penh city. Input from physical and online surveys on the Google form platform allowed the collection of several demographics, academic performance, socioeconomic status, school-related features, and student behavior [9]. Leveraging machine learning techniques, particularly neural networks with association rule learning, we can uncover key variables that significantly impact educational success or failure. The overarching objective of this research is to investigate the effectiveness of BI and EDM techniques [10] to achieve three purposes:

i. Analyze important factors affecting high school students
ii. Provide a powerful asset that uses machine learning and artificial intelligence in education for guidance and success of all stakeholders in the education system
iii. Early warning systems for students at risk of failure aim to guide education decision-makers and schools on failure prevention

2. DATASET

The dataset was gathered from high school Cambodian student information relevant to their family and education experience such as academic performance, demographics, socioeconomic status, behavioral and school-related features**.** There are 37 features designed for online and physical surveys using Google Forms. This research project is gaining with the participation of 422 enthusiastic students.

Table 1. The preprocessed student-related variables

Attribute	Description (Data type)
sex	Student's sex (binary: female or male)
age	Student's age (numeric: from 14 to 19)
grade	Student's grade (numeric: from 10 to 12)
address	Student's home address (binary: rural or urban)
school	Student's school (nominal: public school, private school, or other)
Medu	Mother's education (nominal: Illiteracy, Primary school, Secondary school, High school, Bachelor, Master or PhD)
Mjob	Mother's job (nominal: Housewife, Farmer, Minister, Private staff, Own business, other)

Fedu	Father's education (nominal: Illiteracy, Primary school, Secondary school, High school, Bachelor, Master or PhD)
Fjob	Father's job (nominal: No job, Farmer, Minister, Private staff, Own business, other)
guardian	Student's guardian (nominal: parents, father, mother or other)
sibling	Number of siblings (nominal: 1,2,3,4,>4)
Fincome	Family income (numeric: from 1 – very low to 5 – very high)
Fcare	Guardian care about their children (numeric: from 1 – very low to 5 – very high)
Fsup	Family educational support and spending time with their children (numeric: from 1 – very low to 5 – very high)
Fmotivation	Family motivate their children during student face struggle (numeric: from 1 – very low to 5 – very high)
Fresponse	Family response to children require and gives value to education (numeric: from 1 – very low to 5 – very high)
Fcharactor	Family interest in their children's characteristics (numeric: from 1 – very low to 5 – very high)
Fpraise	Family give praise (numeric: from 1 to 5)
Fencourage	Family encourage the children to share their feelings and problems (numeric: from 1 – very low to 5 – very high)
Henvironment	The home environment makes learning easier (numeric: from 1 – very low to 5 – very high)
Hdistance	Distance from home to school (numeric: 0- <=20min, 1- >20min, 2->45min)
homework	Students always do their homework (numeric: from 1 – very low to 5 – very high)
Rlesson	Students review lessons and prepare for exams (numeric: from 1 – very low to 5 – very high)
Smotivation	Students always push themselves to study hard to get a good result (numeric: from 1 – very low to 5 – very high)
internet	Internet access at home (binary: yes or no)
higher	Wants to take higher education (binary: yes or no)
table	The table is comfortable for self-study at home. (binary: yes or no)
computer	The student uses the computer for extra self-learning (binary: yes or no)
freetime	Free time after school (numeric: from 1 – very low to 5 – very high)
goout	Going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	Weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	Weekday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	Current health status (numeric: from 1 – very low to 5 – very high)
Slearning	Self-learning (numeric: from 1 – very low to 5 – very high)
absence	Student's absence (numeric: from 1 – very low to 5 – very high)

Table 2. The target output

Attribute	Description (Data type)
S1	The average score of the first semester (numeric: from 0 to 50)
S2	The average score of the second semester (numeric: from 0 to 50)

Table 3. Distribution of categorical variables

Attribute	Unique	Top	Freq
sex	2	Male	256
address	2	Rural	334
guardain	4	Parents	362
school	3	Public school	366
Medu	6	Primary school	172
Fedu	6	Primary school	129
Mjob	6	Housewife	176
Fjob	6	Farmer	210
internet	2	Yes	364
computer	2	No	263
higher	2	Yes	378
table	2	Yes	359

Table 4. Summary statistics of numerical variables

Attribute	mean	std	min	max
sibling	3.28	1.22	1	6
health	3.58	0.96	1	5
Fincome	2.79	0.74	1	5
Fcare	3.75	0.98	1	5
Fsup	4.06	0.98	1	5
Fmotivation	4.30	0.91	1	5

Fresponse	3.54	1.02	1	5
Fcharactor	3.64	1.12	1	5
Fencourage	3.57	1.09	1	5
Fpraise	3.85	1.09	1	5
Henvironment	4.58	1.01	1	5
Hdistance	0.63	0.67	0	2
Slearning	3.45	1.07	1	5
freetime	3.22	0.93	1	5
homework	3.96	1.04	1	5
absence	2.28	1.12	1	5
Rlesson	3.94	1.02	1	5
Smotivation	4.08	0.95	1	5
goout	2.69	1.09	1	5
Dalc	1.19	0.56	1	5
S1	36.53	8.15	14	47.9
S2	37.04	8.14	10	48

3. METHODOLOGY

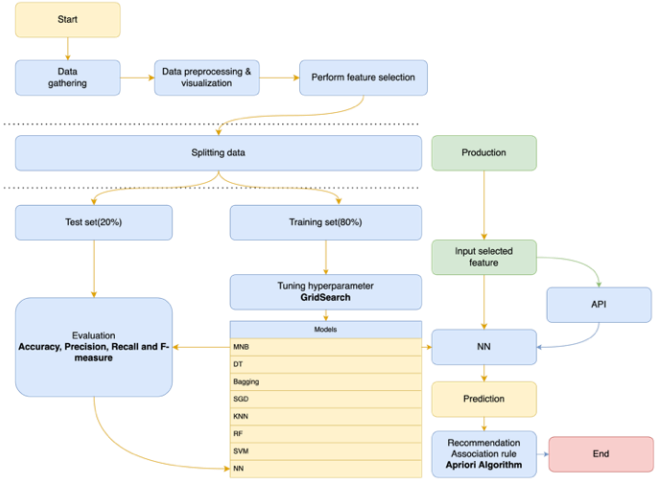

Fig. 1. Diagram research workflow

Establishing a robust research methodology is foundational to the success of any experiment. To conduct experiments effectively, it's imperative to have a well-defined roadmap for the research process. This roadmap typically involves several key steps performed in a clear and sequential order. The research problem needs to be clearly defined. This often involves utilizing advanced technologies such as Machine Learning (ML) and Artificial Intelligence (AI) to analyze the potential of high school student learning outcomes [11]– [13]. These technologies can help identify research objectives and gaps in the current knowledge landscape.

3.1 Data Preprocessing

Data preprocessing is a fundamental step in educational data mining (EDM) models [14]. In this stage, four critical steps are applied: handling missing values, data cleaning, standardization, and one-hot encoding. Additionally, box plots are employed to detect and address outliers [15].

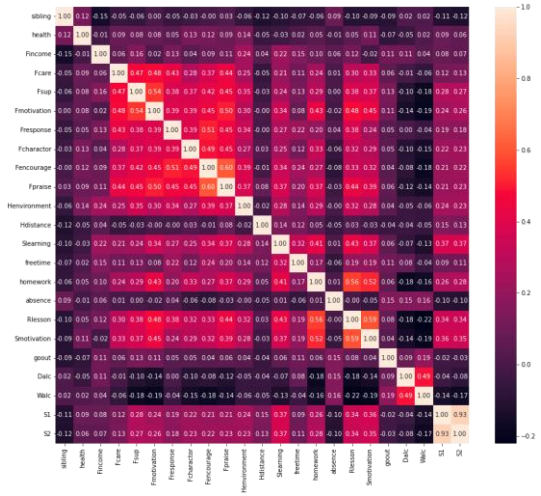

Fig. 2. The heatmap graph shows the relation between numerical variables

Our analysis reveals strong correlations between student factors in academic performance, Family factors (Fpraise, Fsup, Fencouragement, Fcare), and student factors (homework, Rlession, Smotivation).

To indicate a student's status, the target of prediction is the S1, and S2 are split into three categories “Poor”, “Medium”, and “Good”. The S1 and S2 range from 0 to 50. Students with scores between 0 and 30 are classified as poor. Students scoring in the 30 to 40 range are labeled as medium, and those scoring in the 40 to 50 range are labeled as good. It’s simply in the condition below,

a) if S1 >= 40 then “Good”;
b) else if S1 >= 30 then “Medium”;
c) else then “Poor”;

3.2 Feature Selection

Feature selection is one of the crucial stages of the entire process beginning with data collection and ending with modeling. It reduces overfitting, improves accuracy, and reduces training time [16]. It significantly combats overfitting, boosts accuracy, and streamlines training time. Information gain is used to select important features in the dataset.

$$\text{Entropy}(S) = -\sum_{i=1}^{C} P_i \log_2(P_i) \quad \text{(Eq. 1)}$$

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$ (Eq. 2)

Where,

- S denotes a set of training input
- C denotes the number of class labels in the S
- P_i is the proposition of the i-th class,
- *Values(A)* is the all possible values for attribute A, and S_v is the subset of *S* for which attribute A has value v.

3.2 Machine Learning Model

This is the core iterative process that drives experimentation with machine learning and deep learning models, including Multinomial Naïve Bayes (MNB), Decision Trees (DT), Bootstrap Aggregating (Bagging), Stochastic Gradient Descent (SGD), K-Nearest Neighbors (KKN), Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN). GridSearchCV is a powerful automated tool used in machine learning to find the optimal combination of hyperparameters through an exhaustive grid search that leads to the best model performance.

Table 5. Optimal hyperparameters of ML models

Model	Optimal Hyperparameters
MNB	alpha = 0.3434
DT	max_dept=9 ; min_samples_split=6; min_samples_leaf=3; max-features="sqrt" ; criterion= "entropy"
RF	n_estimators="20" ; max_depth=8 ; min_samples_split=2; min_sample_leaf=2; max_features ="sqrt"
Bagging	n_estimators =150
SGD	alpha=0.2627; learning_rate="optimal"; loss ="hinge"
KNN	n_neighbors=16; metric ="euclidean"
SVM	C=2.03 ; kernel ="rbf" ; gamma="auto"

Table 6. Optimal hyperparameters of NN model

	Optimal Hyperparameters of NN Model
Activation function	Relu, Softmax
Loss function	Sparse categorical cross-entropy loss
Dropout rate	0.01
Hidden Layer	4
Epoch	40
Batch size	25

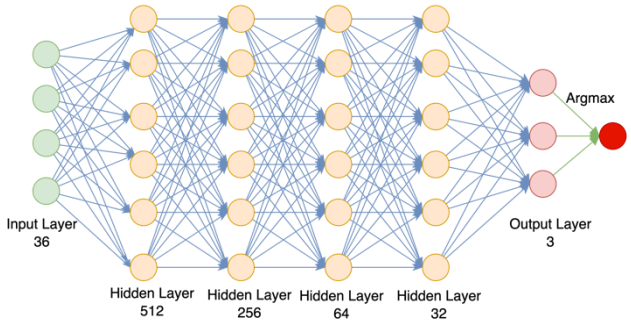

Fig. 3. Neural Network Architecture

3.3 Evaluation Metrics

Evaluation and classification metrics (PCC) used to evaluate the performance of all proposed models are accuracy, precision, recall, and F-measure [17].

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN)$$ (Eq. 3)

$$\text{Precision} = TP/(TP + FP)$$ (Eq. 4)

$$\text{Recall} = TP/(TP + FN)$$ (Eq. 5)

$$F - measure = 2 \times (Precision \times Recall)/(Precision + Recall)$$ (Eq. 6)

Where,

- TP denotes the model correctly identifies a positive case
- TN denotes the model correctly identifies a negative case
- FP denotes the model incorrectly identifies a negative as positive
- FN denotes the model incorrectly identifies a positive as negative.

3.4 Association Rule

Association rule mining is used to uncover hidden relations between input features to recommend educational stakeholders [18]. The Apriori algorithm is applied in this research.

Rule: X →Y (Eq. 7)

Where X and Y are itemsets.
Rule evaluation metrics

$$\text{Support} = \frac{\text{Frequency(X,Y)}}{N}$$ (Eq. 8)

$$\text{Confidence} = \frac{\text{Frequency(X,Y)}}{\text{Frequency(X)}}$$ (Eq. 9)

$$\text{Lift} = \frac{\text{Support}}{\text{Support(X)} \times \text{Support(Y)}}$$ (Eq. 10)

Where,

- N denotes the total number of transactions
- Support is the fraction of transactions that contain both X and Y
- Confidence is a measure of how often items Y appear in transactions that contain X
- Lift is a measure of the popularity of an item

Table 7. Association Rule

Min support	Min confidence	Min Lift	Min length
0.04	0.10	5	6

3.5 Learning Management System Architecture

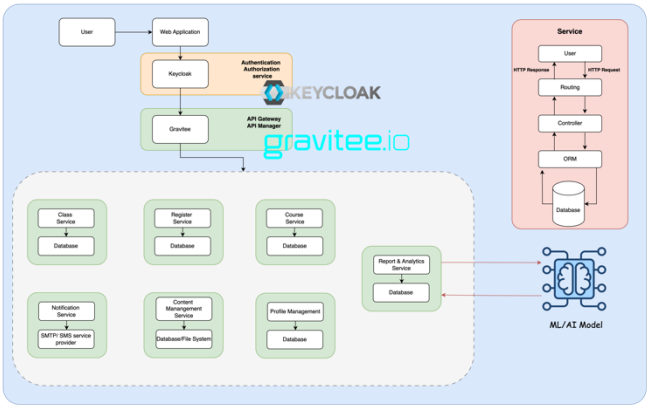

Fig. 4. LMS and ML/AI integration architecture

The design and development of an intelligent learning system is the subject of this study. The system will be able to forecast student development, personalize learning experiences, and enable early intervention by integrating an ML model into an API. This will ultimately lead to an improvement in student outcomes. The backend utilizes microservices architecture, with Flask build reporting and analytics services that integrate with ML/AL models, and other services built with Node.js and Java. For the frontend, we implemented Vite.js.

4. RESULTS AND DISCUSSION

The investigation into applying BI and EDM techniques for predicting student achievement in Cambodian high schools yielded promising results. The study analyzed real-world data of Cambodian students. Eight data mining models were evaluated alongside various input selection strategies.

Configuration of the input into the ML/AI model and target output.

- **A** – with all variables from Table 1 to predict S1 (the output)
- **B** – with all variables from Table 1 with S1 to predict S2 (the output)

Table 8. Classification results with input A (PCC values, in %; **bold** – best model)

Model	Accuracy	Precision	Recall	F-measure
MNB	55.29	65.67	74.57	69.84
DT	55.29	59.70	78.43	67.79
RF	52.94	67.16	71.42	69.23
Bagging	64.70	73.13	80.32	76.56
SGD	52.94	67.16	71.42	69.23
KNN	60.00	68.65	77.96	73.01
SVM	62.35	71.64	78.68	75.00
NN	**90.56**	**93.90**	**93.90**	**93.90**

Table 9. Classification results with input B (PCC values, in %; **bold** – best model)

Model	Accuracy	Precision	Recall	F-measure
MNB	68.23	74.28	85.24	79.38
DT	72.94	78.57	87.30	82.70
RF	80.00	87.14	88.40	87.76
Bagging	84.70	87.14	93.84	90.37
SGD	68.23	0.71	87.71	78.40
KNN	88.23	90.00	95.45	92.64
SVM	87.05	90.00	94.02	91.97
NN	**93.40**	**95.45**	**96.55**	**96.00**

A dataset of 422 responses formed the foundation of this study. Initial experiments yielded promising results using a classic neural network architecture. Future research will explore strategies for expanding data collection through integration with the Learning Management System (LMS). Given the limitations of the basic neural network, advanced deep learning models will be investigated to unlock the full potential of the dataset.

By analyzing a dataset of high school students, we can identify the important elements influencing their achievement.

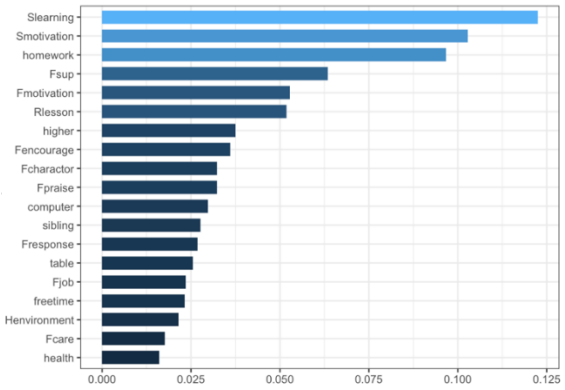

Fig.5. Important features of high school students with input A

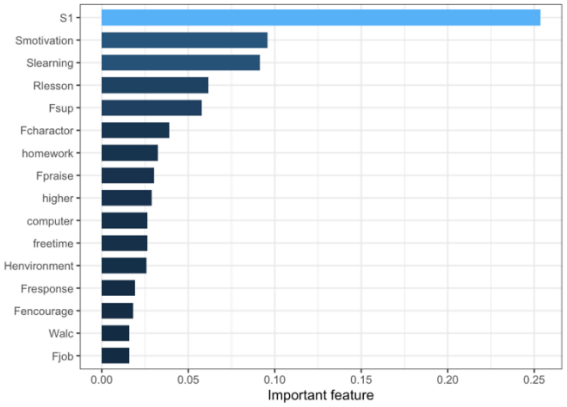

Fig. 6. Important features of high school students with input B

R, along with the ggplot2 package, is a powerful combination for creating the diagrams. These factors include family factors, student factors, and family factors are crucial factors for improving student achievement.

How families support learning

- Family interest in their children's education: When parents take an active interest in their children's learning, it shows support and encourages them to strive for success.
- Guardian care and support: A caring environment provided by parents or guardians fosters a child's well-being and helps them focus on learning.
- Motivational support during challenges: Parents who encourage and motivate their children during difficulties build resilience and a positive learning attitude.
- Educational support and quality time: Engaging with their children's education by offering help, spending quality time, and creating a stimulating learning environment are essential for success.

Student initiative and engagement

- Independent learning: Taking the initiative to learn beyond what's taught in class, such as through research, exploration, or pursuing personal interests related to the subject matter.
- Students review lessons and prepare for exams: Dedicating time to review lessons, practice problems, and prepare effectively for exams using techniques that promote deeper understanding and retention.
- Strong Work Ethic and Goal Setting: Demonstrating a consistent commitment to studying by setting personal goals for improvement and developing a disciplined approach to learning.
- Good health significantly impacts academic achievement. Please provide a solution. Help students develop skills to manage their workload effectively, reducing stress and allowing for healthy habits.

EDM unearthed insights hidden within the data. These association rules were derived from the most important features, suggesting strong relationships between them. This is significant because it allows us to move beyond simply having data to truly understand the connections that drive outcomes. These insights can be harnessed to improve decision-making, personalize experiences, or predict future trends.

5. CONCLUSIONS

Educational Data Mining (EDM) and AI have the potential to revolutionize education, enhancing learning outcomes and enabling personalized education experiences. By leveraging the power of data analytics, institutions can gain insights into student performance, personalize learning paths, identify at-risk students, make evidence-based decisions, and create more effective education environments. The eight machine learning models are examined. The obtained results show that the Neural Network (NN) model has achieved good results with an Accuracy of 93.40%, Precision of 95.45%, Recall of 96.55%, and F-Measure of 96.00%. Incorporating Generative AI into this framework is an exciting next step. Generative AI can create personalized learning materials, practice problems, or feedback tailored to individual student needs. This personalized approach, informed by data analysis and empowered by AI, has the potential to significantly enhance learning outcomes and create a more engaging and effective educational experience for all students.

6. ACKNOWLEDGMENTS

I am appreciative of Dr. PHAUK Sokkhey and Mr. OL Say for their great guidance and support during this research. Their knowledge and support were crucial to my academic

development, and I am grateful to have had the opportunity to collaborate with them. In this study, student and school professional participation has proven to be crucial. Their valuable contribution is sincerely acknowledged.

REFERENCES

[1] M. A. Khder and I. A. Abu-AlSondos, "Business intelligence and data mining: Opportunities and future," European Journal of Business and Management, vol. 13, no. 11, pp. 1–9, Jun. 2021. [Online]. Available: www.iiste.org, doi: 10.7176/EJBM/13-11-01.

[2] M. Yağcı, "Educational data mining: Prediction of students' academic performance using machine learning algorithms," Smart Learn. Env., vol. 9, no. 1, pp. 1-19, Mar. 2022, doi: 10.1186/s40561-022-00192-z.

[3] I. Khan, A. R. Ahmad, N. Jabeur, and M. N. Mahdi, "An artificial intelligence approach to monitor student performance and devise preventive measures," *Smart Learn. Environ.*, vol. 8, no. 1, pp. 1–18, Sep. 2021, doi: 10.1186/s40561-021-00161-y.

[4] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," Universidade do Minho. [Online]. Available: http://www3.dsi.uminho.pt/pcortez/student.pdf. [Accessed: Apr. 17, 2024].

[5] X. Xiong, Y. Li, and Z. Zhang, "A hybrid CNN-RNN model for student performance prediction," *IEEE Access*, vol. 10, pp. 12345–12356, 2022, doi: 10.1109/ACCESS.2022.1234567.

[6] I. U. Sikder, M. N. Hossain, and A. K. M. N. Islam, "A deep convolutional neural network model for predicting student success," *J. Educ. Technol.*, vol. 15, no. 2, pp. 45–60, 2022, doi: 10.1016/j.jet.2022.123456.

[7] Y. Liu, S. Fan, S. Xu, and A. Sajjanhar, "Predicting student performance using clickstream data and machine learning," *Educ. Sci.*, vol. 13, no. 1, p. 17, Dec. 2022, doi: 10.3390/educsci13010017.

[8] L. Song, H. Zhang, and Y. Wang, "Sequential engagement based academic performance prediction network," *IEEE Trans. Learn. Technol.*, vol. 13, no. 4, pp. 789–802, 2020, doi: 10.1109/TLT.2020.1234567.

[9] S. M. F. D. Syed Mustapha, "Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods," *Appl. Syst. Innov.*, vol. 6, no. 5, p. 86, Sep. 2023, doi: 10.3390/asi6050086.

[10] W. Villegas-Ch, X. Palacios-Pacheco, and S. Luján-Mora, "A business intelligence framework for analyzing educational data," *Sustainability*, vol. 12, no. 14, p. 5745, Jul. 2020, doi: 10.3390/su12145745.

[11] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," *Proc. 1st Int. Conf. Soft Comput. Data Sci. (SCDS)*, 2015, pp. 414-422, doi: 10.1109/SCDS.2015.76.

[12] W. Malmia, S. H. Makatita, S. Lisaholit, and A. Azwan, "Problem-based learning as an effort to improve student learning outcomes," Int. J. Sci. Technol. Res, vol. 8, no. 9, pp. 1140-1143, Sep. 2019.

[13] M. Hooda, C. Rana, O. Dahiya, A. Rizwan, and R. S. Hossian, "Artificial intelligence for assessment and feedback to enhance student success in higher education," Mathematical Problems in Engineering, vol. May. 2022, pp. 1-19, 2022, doi: 10.1155/2022/5215722.

[14] C. Romero and S. Ventura, "Educational Data Mining: A Survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, no. 1, pp. 135–146, Jul. 2007, doi: 10.1016/j.eswa.2006.04.005.

[15] P. J. Rousseeuw and M. Hubert, "Anomaly Detection by Robust Statistics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 2, Mar./Apr. 2018, doi: 10.1002/widm.1236.

[16] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003. [Online]. Available: https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf.

[17] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 1–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.

[18] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 1993, pp. 207–216, doi: 10.1145/170035.170072.