

AN INTERPRETABLE MACHINE LEARNING FRAMEWORK FOR PREDICTING STUDENT BURNOUT RISK LEVEL IN AI-ASSISTED LEARNING ENVIRONMENTS

by

Isizoh A. N¹, Ifeoma Lucia Nwosu², Dimson Ifeyinwa C.³

1,2,3: Dept. of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka, Nigeria.

ABSTRACT

The rapid adoption of generative AI (Gen AI) tools in higher education has raised concerns about their role in student burnout. However, most predictive models linking AI usage to wellbeing outcomes operate as black boxes, limiting their practical application by educators and institutions. This study presents an interpretable machine learning (ML) framework for predicting student burnout risk in AI-assisted learning environment. It combines supervised classification with SHAP-based explainability to determine which AI-usage and traditional academic behaviors drive risk. We trained and compared RF, DT, LR, XGBoost, and LightGBM classifiers on a dataset of 50,000 students, which included Gen AI usage patterns, prompt engineering skill and traditional study behaviors, as well as a burnout label. The top-performing model, LightGBM, had an F1-score of 0.639 and a ROC-AUC score of 0.717. SHAP-based interpretability revealed that Gen AI dependency per amount of overall study time was one of the top predictors of burnout risk, implying that subjective reliance on AI tools carries a stronger predictive signal than overall time spent using only these tools. These findings suggest that interpretable ML frameworks can help with the early identification of at-risk students in AI-integrated learning environments, as well as provide institutions with actionable feature insights to guide interventions and policies around Gen AI use.

Keywords/Phrases: Machine learning, Artificial intelligence, LightGBM, Academic burnout

I. INTRODUCTION

In the last few years, the learning style of university students have changed significantly with the introduction of Gen AI tools such as ChatGPT, Gemini and Meta AI. With these tools, students are able to utilize natural language prompts to address routine academic activities such as information retrieval, course assignment solutions and exam preparations. However, the proliferation and over-dependence on these tools in academic settings have brought up concerns regarding their psychological

effect of their excessive use in the context of student burnout (W. Wang et al., 2026).

Academic burnout is a term that refer to scenarios during which students experience a high level of anxiety, depression and feelings of diminished personal accomplishments (Yeskuatov et al., 2025). This is often as a result of long-term exposure to high-stress levels within the academic environment. Notably, burnout can also be influenced by other factors such as competitive academic environments, family pressure and societal expectations. These situations expose affected students to several behavioral, psychological and emotional vulnerabilities, often leading to exhaustion, increased absenteeism and student dropout rates (W. Wang et al., 2026). Recently, an increasing number of studies have explored the relationship between digital technology use and student burnout in different academic settings (Liu et al., 2025; Y. Wang & Xu, 2026). These have confirmed a positive correlation between excessive technological use, dependence and academic burnout.

In response to the issue of academic burnout, researchers have explored the need of utilizing machine learning (ML) algorithms for the early detection of burnout for at-risk students. The use of machine learning (ML) is being examined because to its promising methodological approach to analyzing and identifying nonlinear correlations connected with stress, anxiety, and related mental health disorders among university students (Andrade-Girón et al., 2026). Recent review papers have provided a detailed insight on the diversity and effectiveness of supervised ML classifiers for student mental health data (Lialiou & Maglogiannis, 2025; Turney et al., 2026). However, these studies identified that the related studies have relied on survey-based data related to traditional study habits, academic and demographic factors. However, in the specific context of Gen AI technology use, the research is still lacking due to a number of factors. First, few studies have utilized supervised ML algorithms to explain the intricate relationship between Gen AI usage patterns and academic burnout. Second, few studies have utilized

explainable ML algorithms to offer actionable insights into how sustained Gen AI usage patterns, behavioral and institutional factors influence student burnout after predictive modeling. Based on these findings, this study answers the following research question: Can supervised ML models effectively predict academic burnout among university students based on a variety of characteristics (academic and AI tool usage), while also providing interpretable explanations? Thus, the aim of this study is to design and develop a supervised, explainable ML framework for predicting student burnout risk levels under Gen AI augmented study behaviors.

II. LITERATURE SURVEY

The use of ML algorithms has been explored in related research for early-time prediction of burnout risk in students especially at the university level. These intelligent algorithms are preferred due to their ability to learn the intricate and non-linear relationships present in the data to facilitate tailored intervention strategies (Andrade-Girón et al., 2026). For instance, the work by Endahti (2025) evaluated the comparative performance of two ML models (Random Forest and Logistic Regression) for predicting the likelihood of student burnout based on academic, lifestyle and demographic factors. Features such as student's age, gender, study habits and CGPA were collected for the primary objective of predicting the presence or absence of depression in students. The data was standardized and encoded with the StandardScaler and OneHotEncoder methods to allow for more nuanced analysis. Their results demonstrated that Random Forest (RF) recorded a higher performance in terms of accuracy and F1-score while Logistic Regression (LR) showed a marginal advantage in Precision and ROC-AUC score.

The research by Andrade-Girón et al. (2026) emphasized the use of ensemble and boosting ML algorithms for student burnout risk prediction. This included ML models such as AdaBoost, Extreme Gradient Boosting (XGBoost), RF and LightGBM amongst others. Their assessment of the performance of models indicated that XGBoost achieved an average accuracy and ROC-AUC score of 89.55% and 98.59% respectively after subsequent cross-validation. Another related study by (Joshi et al., 2021) further demonstrated the superiority of ensemble learning models over ML models. Their research presented CatBoost as the best-performing with an accuracy of 92.27%. These studies illustrate the capacity of ML algorithms, especially ensemble ML algorithms, for analysis of complex datasets and

identification of significant predictors of burnout risk across student populations.

The systematic review by Lialiou & Maglogiannis (2025) further provides a detailed insight on the diversity and effectiveness of various ML approaches for student mental health data. However, most of the traditional ML algorithms reported a limitation based on the "black-box" nature of the algorithms. This limits the interpretability of the model and makes it difficult to understand the precise reasoning behind its decision-making process. To address this, recent research by Lau & Abdul Rahman (2026) explores the utilization of explainable AI methods.

Several other research and psychological studies have been published that presents the correlation between the dependence on Gen AI technology usage and academic burnout. These tools through a process of aggregation and analysis of vast amounts of human knowledge has significantly enhanced the accessibility and availability of educational resources (Dong et al., 2025). However, the compulsive and excessive use of these tools during their learning activities results in reduced thinking, problem-solving and cognitive processing (Gu et al., 2026; W. Wang et al., 2026). This subsequently leads to a new form of learning problem centered on AI tool dependence. Such dependence further promotes the occurrence of negative academic and psychological consequences such as burnout and anxiety. The study by Lan et al. (2026) found that students who are highly reliant on utilizing AI tools often develop compulsive and addictive patterns which undermine the learning autonomy and inherent motivation, thereby leading increasing their risk of academic burnout. Another research by Dong et al. (2025) further confirmed that AI technology dependence and its subsequent information overload was a key contributor of students' learning burnout.

Based on these findings, this study raises the following hypotheses:

H1: Ensemble ML models will outperform traditional ML models in predictive accuracy of academic burnout among university students.

H2: AI usage variables significantly impact the prediction of academic burnout in students more than traditional variables like study hours and course of study.

H3: AI usage variables would be among the top-ranked predictive features based on the explainability ML framework utilized.

III. RESEARCH METHODOLOGY

This section presents the detailed analysis of the methodological framework which employs the use of ML algorithms for the automated prediction of possible student burnout risk based on certain input features. Figure 1 visualizes the step-by-step process of the adopted methodology carried out in this study. The raw dataset was first downloaded from Kaggle. An exploratory data analysis (EDA) was then carried out to ascertain the distribution of each column in the dataset. After data preprocessing was employed to handle missing and null data values, encoding the categorical features and

engineering new features based on the initial correlation analysis. The processed data was then split into two parts – 80% for the training data and the remaining 20% for the testing data. The training data was used to train the various ML models under consideration while the test data was used to verify the efficiency of the trained models. After each training, a model is assessed, optimized with hyperparameter tuning and sent back for processing stage for retraining. The final satisfactory model is evaluated on the test dataset. Finally, a post-hoc explainable framework, using SHAP algorithm, is deployed to identify how the various dataset features contributed to the ML model predictions.

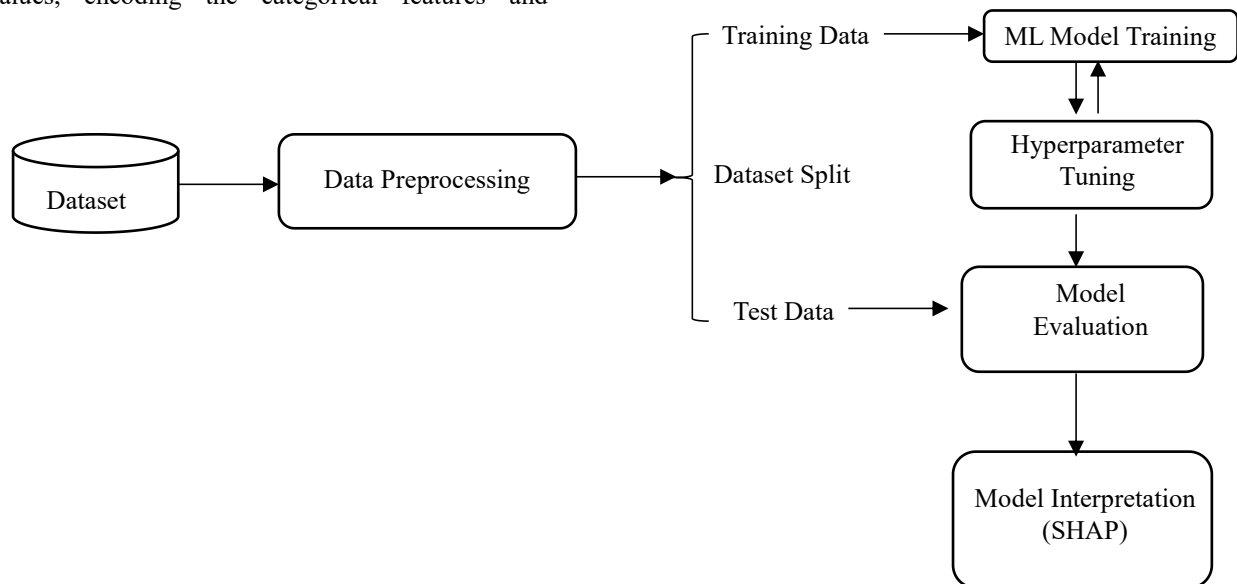


Figure 1: The methodological framework of this study

3.1 Data Source

The data for this study is an open- access dataset accessible on Kaggle (Jadon, 2026). The dataset captures how Gen AI tool usage shapes students' academic performance, psychological and burnout risk across 50,000 students. The dataset is structured in a tabular format and encompasses 16 feature variables of which three variables (Post_Semester_GPA, Skill_Retention_Score and

Burnout_Risk_Level) were suggested target variables. Since this study focuses on burnout risk assessment, the other target variables were dropped and only a 14 feature was imported for analysis. Table 1 presents the detailed categorization of these variables, which form the basis of this study's prediction models. The final target variable (Burnout_Risk_Level) is further classified into three ordinal categories: Low, Medium and High Risk.

Table 1: Variables from the Gen AI Assisted Learning and Student Burnout Dataset

Category	Feature	Description	Data Type
Academic	Student ID	A unique identifier for each student record	Int64
	Major Category	Field of Study in University	Object
	Year of study	Academic Level of Study	Object
	Semester GPA	Grade Point Average recorded at the start of the semester (range of 0.00 – 4.00)	Float64
	Traditional Study Hours	Number of hours spent studying weekly without using AI study tools	Float64
Institutional	Institutional AI policy	The institution's stance on AI usage	Object

AI usage	Gen AI study hours	Number of hours spent studying weekly while using AI study tools	Float64
	Primary Use Case	The study scenario for which the student uses AI	Object
	Prompt Engineering Skill	A self-accessed rating of well the student crafts effective prompts	Object
	Tool diversity	The number of distinct AI study tools used	Int64
	Paid Subscription	A variable indicating whether a student has access to a paid AI subscription or not	Bool
	AI dependency	A self-accessed metric score of how the student depends on the use of AI tools	Int64
Psychological	Anxiety Level	A self-accessed metric representing the students' anxiety on a scale of 1 - 10	Int64
	Burnout Risk	A post-semester assessed burnout risk level	Object

3.2 Data Preprocessing and Encoding

In order to prevent inconsistent transformations during the algorithm training, categorical labels were encoded accordingly. The dataset features such as Major_Category, Institutional_Poily, Primary_Use_Case and Paid_Subscription were transformed with the LabelEncoder() encoder function, while the other categorical columns such as Burnout_Risk_Level was transformed using the OrdinalEncoder() function as the order of the categorical features were considered important. After subsequent preprocessing and exploratory analysis of the dataset, the whitespaces in the feature names were removed and validity of the values in the dataset were further ensured.

3.3 Correlation Analysis and Feature Engineering

Correlation analysis is a statistical technique used to determine how a change in one feature affects the other features in the dataset. The analysis shows that the features Perceived_AI_Dependency and Weekly_GenAI_Hours have a strong positive correlation with the Burnout_Risk_Level feature. Other dataset features have low correlation. The Pre_Semester_GPA and Traditional_Study_Hours features show a negative correlation in the context of the classification problem. Based on the analysis, features such as AI_Study_Ratio, Traditional_Ratio, Anxiety_per_StudyHour, Anxiety_per_AIHour, and AI_Dependency_per_Study_Ratio were engineered for the dataset using equations (1) to (5).

$\text{AI_Study_Ratio} = \frac{\text{Weekly_GenAI_Hours}}{(\text{Weekly_GenAI_Hours} + \text{Traditional_Study_Hours})}$	(1)
$\text{Traditional_Ratio} = \frac{\text{Traditional_Study_Hours}}{(\text{Weekly_GenAI_Hours} + \text{Traditional_Study_Hours})}$	(2)
$\text{Anxiety_per_AIHour} = \frac{\text{Anxiety_Level}}{\text{Weekly_GenAI_Hours}}$	(3)
$\text{Anxiety_per_StudyHour} = \frac{\text{Anxiety_Level}}{\text{Traditional_Study_Hours}}$	(4)
$\text{AI_Dependency_per_Study_Ratio} = \text{AI_Study_Ratio} \times \text{Perceived_AI_Dependency}$	(5)

3.4 Data Partitioning

The final dataset was split into two subsets: a training set comprising of 80% of the observations, which was used for model development and initial comparison and an independent test dataset, which was reserved for final evaluation. The dataset splitting process effectively preserved the relative distribution of the target variable classes within both subsets.

3.5 ML Predictive Models and Hyperparameter Tuning

Conventional and ensemble machine learning algorithms were compared during the model evaluation process. RF, XGBoost, DT and LightGBM were the main models that were tested. The computational environment documented the availability of additional libraries, including LR as a base classifier. All of the algorithms were implemented through a unified pipeline that included preprocessing, scaling and estimation. This configuration ensured that these processes were

limited to the training data in each partition, preventing data leakage. Hyperparameter tuning was limited to the inner loop of the nested cross-validation, which utilized StratifiedKFold with five partitions. The evaluation metric used was F1 weighted score, which is appropriate for multiclass classification because it weights performance based on class frequency. GridSearchCV was used to tune all models' hyperparameters and a random_state of 42 was chosen throughout the training process to ensure consistency. Hyperparameters such as the number of estimators, learning rate, maximum depth, number of leaves and classifier-specific parameters were examined in relation to the algorithm.

3.6 Explainability Methods

Post hoc explainability methods were employed to provide an insight into how the various features of the dataset contributed to the prediction made by each model. The features of the best-performing classifier are further verified using the SHAP analysis framework applied. This is to ascertain which data features contribute to the model's

prediction after the subsequent cross-validation and evaluation of the models.

IV. RESULTS AND DISCUSSION

This section presents the empirical findings obtained using the methodological framework discussed in the previous section. The analysis presents the results of the exploratory data analysis of the dataset, the comparative performance of the evaluated ML models on the test dataset and the complementary interpretability results based on the explainability methods utilized.

4.1 Descriptive Statistics of the Dataset

Table 2 depicts the descriptive analysis of some key numeric variables using minimum, maximum, mean, and standard deviation (SD) in the dataset. The SD, which represents the difference in the data value dispersion in each numeric column, indicate that absence of outlier values. Furthermore, the minimum and maximum values further provide a detailed overview on the range of scores being present in each data column.

Table 2: Descriptive analysis of the Gen AI Assisted Learning and Student Burnout Dataset

Feature	Count	Mean	SD	Min	25%	50%	75%	Max
Pre_Semester_GPA	50,000	3.146	0.479	1.183	2.834	3.210	3.521	3.998
Weekly_GenAI_Hours	50,000	8.428	8.269	0.000	2.390	5.800	11.720	40.00
Traditional_Study_Hours	50,000	11.209	5.156	1.000	7.560	11.180	14.710	35.860
Perceived_AI_Dependency	50,000	3.505	1.821	1.000	2.000	3.000	5.000	10.000
Anxiety_Level	50,000	4.271	2.144	1.000	3.000	4.000	6.000	10.000

A preliminary correlation analysis, as seen in Figure 2, revealed that features such Weekly_GenAI_Hours, Pre_Semester_GPA, Year_of_Study, Traditional_Study_Hours and Perceived_AI_Dependency were highly related to burnout risk level classification. This serves as the backbone for the engineered features presented in section 3.3. The engineered feature AI_Dependency_per_Study_Ratio reported a 0.45

correlation while Traditional_Ratio recorded a -0.42 correlation with Burnout_Risk_Level target variable. These correlation scores suggested that the dependence on Gen AI tools may be related to high student burnout risks. Additionally, the negative correlation in Traditional_Ratio further suggest that the more time spent studying with minimal Gen AI tool usage significantly reduces a student's exposure to risk.

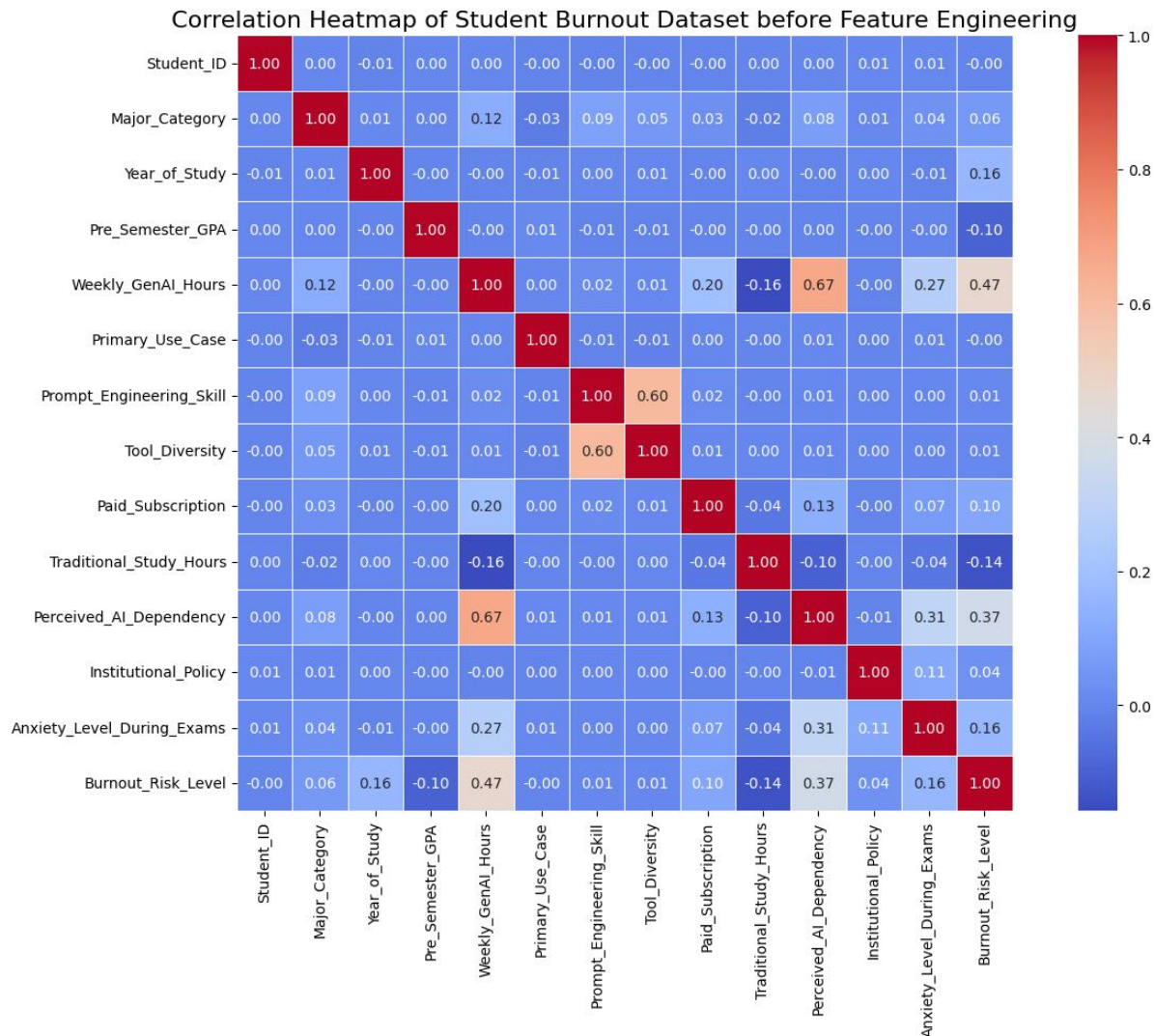


Figure 2: Correlation analysis of the student burnout dataset

4.2 Model Performance

The performance of the selected models is assessed using key performance metrics that allow for the evaluation of robustness and stability of the supervised ML algorithms. For this purpose, the 5 ML models (LR, RF, DT, XGBoost and LightGBM) are compared using accuracy, weighted F1 and ROC-AUC scores. Table 3 presents the comparative analysis of the models' performance. The comparison revealed significant differences in

predictive accuracy and generalizability among the various algorithms. The linear model Logistic Regression (LR) serve as a reliable baseline model, with an accuracy score of 0.5361. These figures show that linear ML models can capture linear assumptions in a dataset but are limited in their ability to model nonlinearities in the data. Its weighted F1 score of 0.535984 indicates that the model has moderate classification performance for capturing borderline or ambiguous cases.

Table 3: Comparative analysis of ML model performance

ML Model	Accuracy	Weighted F1 Score	ROC-AUC
DT	0.4432	0.444092	0.578144
RF	0.5223	0.524601	0.699625
LR	0.5361	0.535984	0.714656
XGBoost	0.6298	0.632110	0.714707
LightGBM	0.6374	0.638979	0.716859

In contrast, ensemble-based algorithms, such as XGBoost and Light GBM, perform significantly better across all metrics. LightGBM achieves a weighted F1 score of 0.6389, demonstrating the model's balanced trade-off between precision and recall, which is especially important in educational settings where under-identifying at-risk students can result in biased behavioral interventions. Furthermore, its ROC-AUC score of 0.7169 indicates how well the model distinguishes between

true positive and false positive classifications across all predictive categories. These findings support hypothesis H1 by demonstrating that ensemble ML models outperform traditional ML models. Figure 3 depicts the confusion matrices of each evaluated model, which show the actual and predicted numeric differences between each predictive model. The observed performance gains are also particularly relevant for real-world deployments, especially in scenarios requiring early intervention.

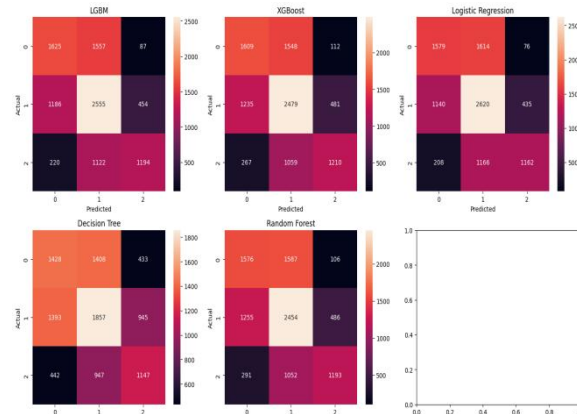


Figure 3: Confusion matrices of the evaluated ML algorithms

4.3 Feature Importance

To better understand the reasoning behind the best-performing model predictions, the SHAP algorithm was used to calculate feature importance values across all predictions. Figure 4 shows the feature importance ranking of the best-performing model for predicting student burnout risk levels. The findings show that the features are ranked in descending order of progression, implying that the topmost features contribute more to the model's predictive power than the bottom features. According to the graph, the number of weekly hours spent using a Gen AI tool has a greater impact than other variables, implying that changing this feature may have a greater impact on other variables. However, this feature has a greater impact on students rated class 2 (high risk) and class 0 (low risk) than on those rated class 1 (medium risk). A similar feature importance is observed in the second topmost feature, AI_Dependency_per_Study_Hour, indicating that the student's reliance on the tool over the overall study time is a more significant contributor to burnout than the total amount of time spent studying (Traditional_Study_Hours). Notably, four of the top ten features are AI usage variables, which have a significant impact on predicting student burnout risks. These findings further support hypotheses H2 and H3.

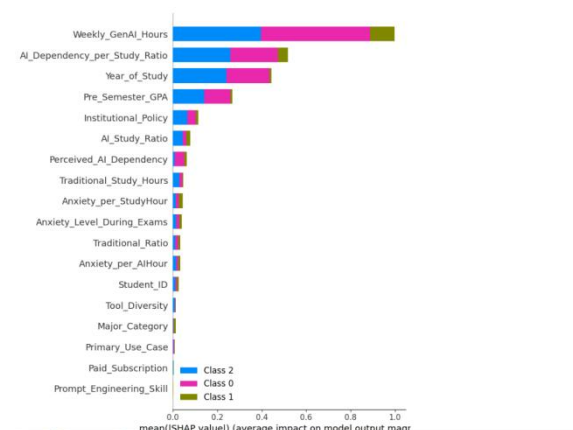


Figure 4: SHAP feature Importance Plot for the LightGBM Model

V. CONCLUSION

This study presents an interpretable machine learning framework for predicting student burnout risk levels in Gen AI-assisted study environments. It combines supervised machine learning classification algorithms such as RF, DT and XGBoost with SHAP-based explainability to determine which AI-related and traditional academic behaviors contribute to burnout risk classification. Among the five evaluated algorithms, LightGBM had the best predictive performance (F1 = 0.638979, ROC-AUC = 0.716859), which is consistent with the general advantage of ensemble tree methods on structured

tabular data. The number of hours spent using Gen AI tools weekly, as well as the AI tool dependency per study hour, were found to be among the best predictors of burnout risk, implying that the frequency with which these tools are used is an informative signal for identifying at-risk students. These findings can be used to guide institutions that are deploying AI-informed early-warning systems with the goal of reducing the rate of student burnout and improving the institutional environment.

However, these findings should be interpreted with caution due to several limitations. First, the dataset is based on self-reported measures of dependency and anxiety. This limits the causal interpretation of the features' relationship to burnout. Second, the findings are drawn from a single dataset and have not been evaluated or generalized to other institutional contexts. Future research should validate this framework using data from additional institutions and expand the analysis to a cost-sensitive formulation that emphasizes missed high-risk cases.

REFERENCES

- Andrade-Girón, D. C., Marin-Rodriguez, W. J., Zuñiga-Rojas, M. G., Neri-Ayala, A. C., Susanibar-Ramírez, E. T., & Aguilar-Luna-Victoria, M. A. (2026). Predicting Student Stress Using Machine Learning Ensemble Models: A Multi-Criteria Comparison with Explainable Artificial Intelligence Analysis. *AI*, 7(7), 268. <https://doi.org/10.3390/ai7070268>
- Dong, X., Wang, Z., & Han, S. (2025). Mitigating Learning Burnout Caused by Generative Artificial Intelligence Misuse in Higher Education: A Case Study in Programming Language Teaching. *Informatics*, 12(2), 51. <https://doi.org/10.3390/informatics12020051>
- Endahti, L. (2025). Predicting Student Depression Based on Academic, Lifestyle, and Demographic Factors Using Machine Learning. *Artificial Intelligence in Learning*, 1(3), 195. <https://doi.org/10.63913/ail.v1i3.15>
- Gu, C., He, B., & Zhang, X. (2026). The relationship between college students' AI technology dependence and learning burnout: A chain mediation analysis of technology acceptance and AI self-efficacy. *Frontiers in Psychology*, 17, 1843366. <https://doi.org/10.3389/fpsyg.2026.1843366>
- Jadon, L. (2026). *Impact of Ai on Students*. <https://www.kaggle.com/datasets/laveshjadon/ai-impact-on-students>
- Joshi, A., Saggar, P., Jain, R., Sharma, M., Gupta, D., & Khanna, A. (2021). CatBoost—An Ensemble Machine Learning Model for Prediction and Classification of Student Academic Performance. *Advances in Data Science and Adaptive Analysis*, 13(03n04), 2141002. <https://doi.org/10.1142/S2424922X21410023>
- Lan, Y., Liu, S., & Chen, H. (2026). Exploring the formation of learning burnout among college students in AI context: A serial mediation mechanism of AI dependence and addiction based on I-PACE model. *Frontiers in Computer Science*, 8, 1756441. <https://doi.org/10.3389/fcomp.2026.1756441>
- Lau, W. J., & Abdul Rahman, H. (2026). Predicting academic performance through machine learning: Integrating demographic, psychological, and behavioral predictors using explainable AI. *Asia Pacific Education Review*. <https://doi.org/10.1007/s12564-026-10124-y>
- Lialiou, P., & Maglogiannis, I. (2025). Students' Burnout Symptoms Detection Using Smartwatch Wearable Devices: A Systematic Literature Review. *AI Sensors*, 1(1), 2. <https://doi.org/10.3390/aisens1010002>
- Liu, C., Zhu, K., Zhang, Z., & Ding, Y. (2025). The impact of Chinese undergraduates' perceived acceptance of AI technology on academic achievement: The mediating role of critical thinking. *Frontiers in Psychology*, 16, 1727037. <https://doi.org/10.3389/fpsyg.2025.1727037>
- Turney, J., Young, T., Chauhan, D., Beecharry, R., & Mahmud, M. (2026). AI Use for Medical Students: Impact on Clinical Skill Acquisition and Retention. A Systematic Review. *Advances in Medical Education and Practice*, Volume 17, 1–13. <https://doi.org/10.2147/AMEP.S583763>

- Wang, W., Wu, Y., Fang, J., Yang, C., & Wen, L. (2026). When cognitive offloading becomes dependence: How AI dependence mediates the pathway from academic stress to burnout and anxiety. *BMC Psychology*, 14(1), 1125. <https://doi.org/10.1186/s40359-026-04911-w>
- Wang, Y., & Xu, S. (2026). Relationship between artificial intelligence tool usage experience and academic stress among college students: Mediating role of loneliness and moderating role of academic self-efficacy. *Acta Psychologica*, 263, 106220. <https://doi.org/10.1016/j.actpsy.2026.106220>
- Yeskuatov, E., Foo, L. K., & Chua, S.-L. (2025). Detecting Burnout Among Undergraduate Computing Students with Supervised Machine Learning. *Healthcare*, 13(23), 3182. <https://doi.org/10.3390/healthcare13233182>