

Predicting Digital Literacy Levels in Higher Education: A LightGBM Model Integrating Feature Selection for Improved Accuracy

Arman Haqqi Anna Zili¹, Selly Anastassia Amellia Kharis²✉

¹ Mathematics, Faculty of Mathematics and Science, Universitas Indonesia, Indonesia

² Mathematics, Faculty of Science and Technology, Universitas Terbuka, Indonesia

✉ **Corresponding Author** : Selly Anastassia Amellia Kharis (e-mail: selly@ecampus.ut.ac.id)

Article Information	ABSTRACT
Article History Received : September 18, 2025 Revised : October 08, 2025 Published : October 20, 2025	Digital literacy has become an essential skill in higher education, particularly in online and distance learning settings. This study explores the use of Light Gradient Boosting Machine (LightGBM) to classify digital literacy levels among 10,393 students at Universitas Terbuka. To improve both efficiency and clarity of interpretation, feature selection was carried out using SelectKBest, which reduced the dataset to 33 predictors. The final model, evaluated through stratified 5-fold cross-validation, achieved an accuracy of 0.964 and a weighted F1-score of 0.964. The results show that limiting the number of features did not weaken predictive performance, while also making it easier to identify which aspects of digital literacy are most influential. Interestingly, the strongest predictors were not only technical skills but also ethical behavior, digital citizenship, and online communication. These findings highlight that digital literacy is multidimensional and that effective assessment tools must account for social and behavioral factors alongside technical competence. Taken together, applying feature selection with LightGBM offers an effective way to assess digital literacy in higher education. The method achieves strong predictive accuracy while keeping the model interpretable, giving universities clearer guidance for shaping interventions and curricula in online learning contexts.
Keywords: <i>Classification; Digital Literacy; Machine Learning; LightGBM.</i>	

INTRODUCTION

Digital literacy is increasingly recognized as a core requirement for university students in the 21st century (Martínez-Bravo et al., 2022). The term is no longer confined to knowing how to operate digital tools; it now also involves evaluating information critically, managing it effectively, and applying it responsibly in both academic and social settings. Within the context of higher education, these competencies have evolved into fundamental skills that support learning activities, particularly as institutions adopt digital platforms, online assessments, and technology-mediated collaboration as integral components of academic practice (Falloon, 2020). A growing body of research suggests that students with stronger digital literacy skills are often better positioned to succeed academically, although the precise mechanisms linking these skills to performance remain debated (Li et al., 2025; Zakir et al., 2025).

For an open and distance learning institution such as Universitas Terbuka, digital literacy is especially critical. Success in this environment depends not only on students' ability to navigate online resources but also on their capacity to communicate in virtual spaces and to engage responsibly in digital communities. Prior reviews of online education have consistently pointed out that measuring learning outcomes carefully is essential, both for monitoring student

development and for maintaining the quality of instructional design (Wei et al., 2021). In this sense, reliable measures of digital literacy serve the dual purpose of supporting individual learners and strengthening institutional accountability. Moreover, as distance learning environments rely heavily on self-directed study and asynchronous interaction, digital literacy becomes a decisive factor in ensuring student persistence. For institutions like Universitas Terbuka, which serve a geographically dispersed and diverse student population, understanding and predicting digital literacy levels not only informs targeted interventions but also contributes to the broader mission of expanding equitable access to higher education (Kharis et al., 2025).

The use of machine learning (ML) in education has grown rapidly over the past decade (Kharis et al., 2023; Kharis & Zili, 2022). Researchers have applied ML methods to identify patterns in large educational datasets, with the aim of predicting student outcomes or detecting those at risk of failure (Márquez-Vera et al., 2013; Romero & Ventura, n.d.). Approaches have ranged from relatively simple techniques such as logistic regression to more complex ensemble-based models (Audrin & Audrin, 2022). Gradient Boosting Decision Trees (GBDT) are one family of models that has attracted particular attention for their balance of predictive performance and computational efficiency (Friedman, 2001; Ke et al., 2017). Recent applications in education data mining demonstrate that GBDT-based methods are particularly effective when dealing with heterogenous features, nonlinear relationships, and missing values, which are common challenges in student learning dataset.

Among these, Light Gradient Boosting Machine (LightGBM) represents a more recent innovation. Its use of histogram-based binning and leaf-wise tree growth enables faster training and lower memory use than many earlier boosting algorithms (Ke et al., 2017). While applications of LightGBM have been documented in fields such as healthcare and student performance prediction (Wang & Wang, 2020), there remains relatively little work applying it to digital literacy specifically. Much of the existing literature focuses on broader constructs such as “digital competence” or on predicting academic success in general (Ilomäki et al., 2016). Despite broad recognition of digital literacy as a multidimensional construct, relatively few studies have attempted to classify it directly using efficient machine learning approaches such as LightGBM (Audrin & Audrin, 2022).

This study seeks to address that gap. We evaluate the potential of LightGBM, in combination with feature selection, to classify digital literacy levels among students at Universitas Terbuka. By doing so, the research not only contributes to the technical discussion of model performance but also offers practical insights into which aspects of digital literacy matter most in an online university setting. This research is particularly valuable for open and distance learning institutions, where understanding the relative importance of skills such as digital ethics, digital safety, digital culture, and digital skills can inform the design of targeted interventions and support services. In this way, the study bridges methodological advancement in education data mining with actionable implications for policy and practice in digital higher education.

RESEARCH METHODS

This study employed a machine learning approach to classify students’ digital literacy levels based on survey and performance data collected at Universitas Terbuka. The data consisting of four aspects of digital literacy, namely digital skill, digital safety, digital culture, and digital ethics (Permatasari et al., 2025). In addition, the data included questions related to students’ profiles. The methodological process included data preparation, feature selection, model choice, validation, and performance evaluation. Each step was designed with the dual aim of ensuring predictive accuracy and maintaining interpretability.

The methodological framework followed five main stages: (1) data preparation, (2) feature selection, (3) model selection, (4) validation, and (5) performance evaluation. Each stage was designed to ensure both predictive accuracy and interpretability. Figure 1 presents the overall workflow of the analytical procedure.

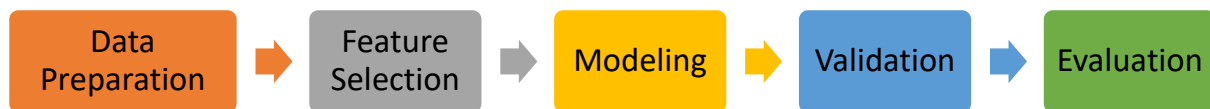


Figure 1. Methodological workflow of the study

Data Preparation

The initial dataset contained 10,396 student records and 41 features. After removing the very small number of cases in the “low” literacy category, the final working dataset comprised 10,393 students. Among them, 490 (4.71%) were classified as “satisfactory,” 6,635 (63.84%) as “high,” and 3,268 (31.44%) as “very high.” This imbalance across classes was addressed by assigning class weights during training so that minority categories were not overshadowed by the majority group.

The dataset included both numerical and categorical variables. While LightGBM can natively handle categorical inputs, the feature selection method chosen for this study required numerical features only. Therefore, categorical variables were temporarily excluded during feature selection but later reintroduced when training the final LightGBM model. Missing data were minimal and were either implicitly handled by the model or addressed in earlier preprocessing steps.

Feature Selection

To reduce dimensionality and highlight the most informative predictors, we used the SelectKBest method with an ANOVA F-test (`f_classif`). This statistical filter ranks features according to the strength of their association with the target variable (Kuhn & Johnson, 2013). The optimal number of features (k) was determined through empirical performance evaluation combined with theoretical considerations. Specifically, classification performance saturation was identified at $k = 33$, beyond which additional features provided negligible improvements in accuracy and F1-score. This ensured that the final model balanced efficiency and interpretability without sacrificing accuracy. The top 33 features were ultimately retained, many of which related to students’ self-reported skills, online interaction habits, and information management practices.

Model Choice

LightGBM was selected as the classifier due to its strong track record in handling large, complex datasets efficiently (Bukonla, 2025). The model constructs an ensemble of decision trees, where each tree improves on the errors of its predecessors (Yadalam et al., 2024). Two design features make it particularly effective: (1) histogram-based binning, which accelerates the search for split points while reducing memory use, and (2) leaf-wise growth, which tends to converge faster than level-wise alternatives. Competing frameworks such as XGBoost and CatBoost also offer high performance, but LightGBM was preferred here for its speed and scalability in the context of large educational datasets.

Validation Strategy

To assess generalization performance reliably, the study used stratified 5-fold cross-validation. Stratification preserved the proportion of “satisfactory,” “high,” and “very high” classes in each fold, which is especially important for imbalanced datasets. Each data subset served once as the validation set while the remaining folds were used for training. This approach ensured that every student record contributed to both training and testing, reducing the risk of overfitting and providing a fairer estimate of model performance (Ca & Fr, 2004).

Evaluation Metrics

Model performance was evaluated with several commonly used classification metrics. Accuracy was reported as a basic measure, but given the imbalanced nature of the dataset,

additional emphasis was placed on precision, recall, and the F1-score. Weighted averages were used to reflect the distribution of students across classes and to provide a more balanced view of the model's performance (Sokolova & Lapalme, 2009).

RESULTS AND DISCUSSION

This study set out to evaluate how well LightGBM could classify students' digital literacy levels, with a particular focus on the role of feature selection. The target variable was divided into three categories: "satisfactory," "high," and "very high"—and class weights were applied to account for the imbalance across these groups.

Impact of Feature Selection

The experiments showed a clear relationship between the number of features included and model performance. As Table 1 and Figure 2 illustrate, accuracy and weighted F1-score improved steadily as more features were added, reaching their peak when 33 predictors were retained. Beyond this point, additional features did not yield measurable gains, and performance essentially plateaued.

Tabel 1. Comparison of model performance based on the number of features

Number of features selected	Accuracy	F1-score
26	0.942076	0.942637
27	0.946406	0.946852
28	0.948523	0.948917
29	0.955451	0.955666
30	0.955451	0.955644
31	0.957952	0.958133
32	0.96286	0.963057
33	0.964303	0.964492
34	0.964303	0.964492
35	0.964303	0.964492
36	0.964303	0.964492

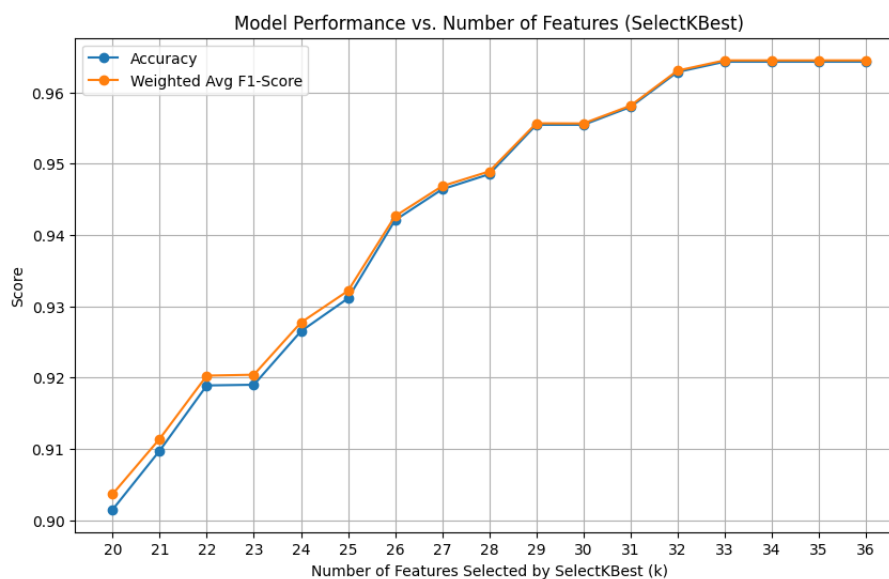


Figure 2. Evaluation of model performance based on the number of features.

At the optimal condition (using 33 features), the model achieved an accuracy of 0.964 and a weighted F1-score of 0.964—identical to the results obtained when using the full feature set. This suggests that a carefully chosen subset of predictors is sufficient for robust classification, reducing complexity without compromising accuracy. The small drop in performance compared with the maximum observed values (less than 1%) is negligible in practical terms, confirming the effectiveness of the SelectKBest method in eliminating redundant or noisy variables.

Classification Performance

The normalized confusion matrix (Figure 3) provides further insight. The final model correctly identified 96% of students in the “high” literacy category, 91% in the “satisfactory” group, and 97% in the “very high” group. These results indicate that LightGBM handled all three categories well, though some overlap between “satisfactory” and “high” was apparent. This overlap may reflect the reality that the boundary between these two groups is not always sharp, as both share many characteristics in terms of basic tool use and online engagement.

Gabungan Confusion Matrix yang Dinormalisasi (dengan fitur penting (SelectKBest, k=33) dan bobot kelas)

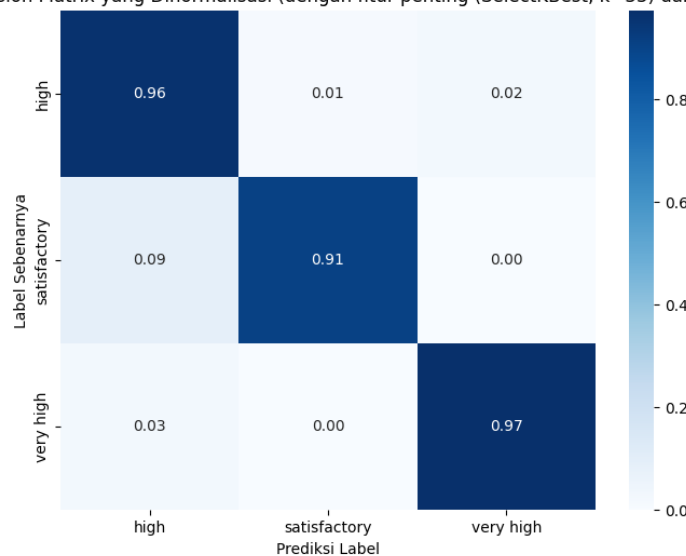


Figure 3. Normalized confusion matrix of model performance.

Interpretation and Implications

The symbols not previously mentioned in the equation must be explained in the following text. The findings highlight two important points. First, advanced boosting methods such as LightGBM can deliver strong performance in classifying nuanced competencies like digital literacy. Second, feature selection plays a valuable role in simplifying the model while keeping performance nearly unchanged. For institutions, the findings indicate that predictive tools can be developed to remain both accurate and computationally efficient, an important consideration when models are applied at scale in large online learning environments. Beyond algorithmic performance, the feature selection results provide valuable insights into the determinants of digital literacy. Notably, the top predictors extended beyond technical competencies to include ethical behavior and online communication practices—for example, avoiding plagiarism, respecting digital privacy, and maintaining appropriate communication online.

The feature selection process also offers insight into what matters most for digital literacy. The strongest predictors were not limited to technical proficiency but included indicators of ethical behavior and digital citizenship. These findings align with multidimensional framework of digital literacy, such as those proposed by Deschênes (2024) and He & Zhu (2017), which conceptualize digital literacy as a composite of technical, cognitive, and socio-ethical dimensions. The

prominence of ethical and communicative indicators suggests that students' ability to navigate digital spaces responsibly is as critical as their operational skills.

For institutions like Universitas Terbuka, these insights underscore the importance of designing interventions that go beyond tool training. Digital literacy curricula should explicitly incorporate modules on digital ethics, responsible participation, and critical communication practices to address the full spectrum of competencies required for effective online learning.

Limitations and Future Directions

While the results are encouraging, some caveats are worth noting. The dataset came from a single institution and was largely based on self-reported survey responses, which are subject to individual bias. Future studies should incorporate behavioral data, such as log files from learning management systems, to provide a more objective complement to self-report measures. In addition, applying the model to data from multiple universities would test its generalizability. A longitudinal design could also capture how digital literacy develops over time, making it possible to predict not only current levels but also likely trajectories.

CONCLUSION

This study examined the extent to which LightGBM, when combined with feature selection, can classify digital literacy levels among university students accurately and efficiently. The findings show that the model performed strongly even when the feature set was reduced to 33 predictors, reaching a weighted F1-score of 0.964. This suggests that careful feature selection can simplify the model without sacrificing predictive power, which is an important consideration for large-scale educational contexts. Beyond the technical results, the study also reinforces the idea that digital literacy should not be viewed solely as technical proficiency. Features linked to ethical conduct, digital citizenship, and online communication were among the most influential predictors, highlighting the multidimensional nature of digital literacy. For institutions such as Universitas Terbuka, this has practical meaning: interventions and curricula should extend beyond tool training and place greater emphasis on responsible, thoughtful, and collaborative engagement in digital spaces. At the same time, the results should be interpreted with some caution. The dataset came from a single institution and relied largely on self-reported survey responses, which may not fully capture students' real digital practices.

Future work could benefit from incorporating behavioral data, such as activity logs from learning management systems, and from testing the model across multiple institutions to check its generalizability. Longitudinal studies would also help to track how students' digital literacy develops over time and how predictive models might support that growth. Overall, the integration of LightGBM with feature selection shows promise as a practical approach for assessing digital literacy in higher education. By balancing accuracy with interpretability, this approach can inform not only the development of predictive tools but also the design of targeted educational strategies that support students in becoming capable and responsible participants in digital learning environments.

ACKNOWLEDGEMENTS

The author would like to express sincere gratitude to the Lembaga Penelitian dan Pengabdian kepada Masyarakat Universitas Terbuka (LPPM UT) for their support and facilities provided during the implementation of this research.

REFERENCES

- Audrin, C., & Audrin, B. (2022). Key factors in digital literacy in learning and education: a systematic literature review using text mining. *Education and Information Technologies*, 27(6), 7395–7419. <https://doi.org/10.1007/s10639-021-10832-5>
- Bukonla, A. O. (2025). An Optimized Light-GBM Based Classification Model for Effective Classification of Loan Defaulter. *Dutse Journal of Pure and Applied Sciences*, 11(1b), 301–310. <https://doi.org/10.4314/dujopas.v11i1b.32>
- Ca, B. U., & Fr, Y. G. (2004). No Unbiased Estimator of the Variance of K-Fold Cross-Validation Yoshua Bengio Yves Grandvalet. In *Journal of Machine Learning Research* (Vol. 5).
- Deschênes, A. A. (2024). Digital literacy, the use of collaborative technologies, and perceived social proximity in a hybrid work environment: Technology as a social binder. *Computers in Human Behavior Reports*, 13. <https://doi.org/10.1016/j.chbr.2023.100351>
- Falloon, G. (2020). From digital literacy to digital competence: the teacher digital competency (TDC) framework. *Educational Technology Research and Development*, 68(5), 2449–2472. <https://doi.org/10.1007/s11423-020-09767-4>
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. In *Source: The Annals of Statistics* (Vol. 29, Issue 5).
- He, T., & Zhu, C. (2017). Digital informal learning among Chinese university students: the effects of digital competence and personal factors. *International Journal of Educational Technology in Higher Education*, 14(1). <https://doi.org/10.1186/s41239-017-0082-x>
- Ilomäki, L., Paavola, S., Lakkala, M., & Kantosalo, A. (2016). Digital competence – an emergent boundary concept for policy and educational research. *Education and Information Technologies*, 21(3), 655–679. <https://doi.org/10.1007/s10639-014-9346-4>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. <https://github.com/Microsoft/LightGBM>.
- Kharis, S. A. A., Arisanty, M., Permatasari, M., Robiansyah, A., & Sukatmi, S. (2025). Exploration of The Digital Literacy Level of New Students in Open and Distance Learning. *Paedagoria : Jurnal Kajian, Penelitian Dan Pengembangan Kependidikan*, 16(2), 259–266. <https://doi.org/10.31764>
- Kharis, S. A. A., Hertono, G. F., Irawan, S. R., Wahyuningrum, E., & Yumiati. (2023). Students' success prediction based on the Fuzzy K-Nearest Neighbor method in Universitas Terbuka. *Education Technology in the New Normal: Now and Beyond*, 212–218. <https://doi.org/10.1201/9781003353423-22>
- Kharis, S. A. A., & Zili, A. H. A. (2022). Learning Analytics dan Educational Data Mining pada Data Pendidikan. *Jurnal Riset Pembelajaran Matematika Sekolah*, 6.
- Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
- Li, F., Cheng, L., Wang, X., Shen, L., Ma, Y., & Islam, A. Y. M. A. (2025). The causal relationship between digital literacy and students' academic achievement: a meta-analysis. *Humanities and Social Sciences Communications*, 12(1). <https://doi.org/10.1057/s41599-025-04399-6>

- Márquez-Vera, C., Romero Morales, C., & Ventura Soto, S. (2013). Predicting school failure and dropout by using data mining techniques. *Revista Iberoamericana de Tecnologías Del Aprendizaje*, 8(1), 7–14. <https://doi.org/10.1109/RITA.2013.2244695>
- Martínez-Bravo, M. C., Chalezquer, C. S., & Serrano-Puche, J. (2022). Dimensions of Digital Literacy in the 21st Century Competency Frameworks. *Sustainability (Switzerland)*, 14(3). <https://doi.org/10.3390/su14031867>
- Permatasari, M., Kharis, A. A., Arisanty, M., Robiansyah, A., & Zubir, E. (2025). Validity and Reliability Testing of Student Digital Literacy Instrument in Distance Education. In *Science and Technology ISST* (Vol. 2024).
- Romero, C., & Ventura, S. (n.d.). *Educational Data mining and Learning Analytics: An updated survey*.
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Wang, Y., & Wang, T. (2020). Application of improved LightGBM model in blood glucose prediction. *Applied Sciences (Switzerland)*, 10(9). <https://doi.org/10.3390/app10093227>
- Wei, X., Saab, N., & Admiraal, W. (2021). Assessment of cognitive, behavioral, and affective learning outcomes in massive open online courses: A systematic literature review. *Computers and Education*, 163. <https://doi.org/10.1016/j.compedu.2020.104097>
- Yadalam, P. K., Thirukkumaran, P. V., Natarajan, P. M., & Ardila, C. M. (2024). Light gradient boost tree classifier predictions on appendicitis with periodontal disease from biochemical and clinical parameters. *Frontiers in Oral Health*, 5. <https://doi.org/10.3389/froh.2024.1462873>
- Zakir, S., Hoque, M. E., Susanto, P., Nisaa, V., Alam, M. K., Khatimah, H., & Mulyani, E. (2025). Digital literacy and academic performance: the mediating roles of digital informal learning, self-efficacy, and students' digital competence. *Frontiers in Education*, 10. <https://doi.org/10.3389/feduc.2025.1590274>