International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2025): 8.036

Impact of AI Based Learning on Student Performance Prediction in Higher Education Using Machine Learning

Prathamesh Deepak Salke[1], Dr. Ayesha Siddique[2]

[1]MCA (Computer Science), JSPM University, Wagholi, Pune, India
Email: *prathameshsalke8094[at]gmail.com*

[2]Associate Professor, Department of Computer Science, JSPM University, Pune

Abstract: ***Research on what influences student academic achievement has been ongoing for a while now; however, little has been documented about the impact of the use of artificial intelligence (AI). In this paper, we seek to bridge this gap by employing an empirical analysis of the effects of AI adoption using two machine learning methods, namely, Random Forest and Light Gradient Boosting Machine (LightGBM), on 8,000 student profiles. There were 22 variables representing students' performance, their learning habits, AI usage, and demographics. The two models are trained to classify students based on their performance as either High, Medium, or Low. According to experimental results, LightGBM outperforms Random Forest in terms of classification accuracy at 83.25%, and F1-Score at 0.8332 against 82.05% and 0.8106, respectively. Both algorithms have excellent prediction capabilities with ROC-AUC score above 0.93. Moreover, cross-validation tests confirm the accuracy, robustness, and consistency of the two models. In addition to their predictive prowess, both models demonstrate the importance of traditional academic variables like examination scores, assignments, and study habits. On the other hand, AI variables are secondary predictors in this case. This implies that AI has a positive impact on academic performance when combined with study habits.***

Keywords: Artificial Intelligence in Education, Educational Data Mining, Student Performance Prediction, Machine Learning, Random Forest, LightGBM

1. Introduction

One of the most revolutionary technologies in many different areas like health care, finance, transport, and education is Artificial Intelligence (AI) [1]. When talking about its role in the sphere of education, AI technologies like intelligent tutoring systems, automatic assessment tools, and adaptive learning technologies are becoming more popular in order to make the learning process more personalized and accessible [2], [3].

At the same time, in recent years, there appeared numerous applications that allow students to utilize AI technologies to do assignments, generate summaries, solve complicated problems, and even search for study materials [4]. Such applications can be very helpful but at the same time raise the issue of how it influences the independence and development of critical thinking skills of students.

One of the key revolutionary technologies is Artificial Intelligence (AI), which has found wide usage in many fields, such as health, finance, transport, and education [1]. When applied to the field of education, artificial intelligence has been successfully used for such applications as intelligent tutoring systems, automated feedback systems, and adaptive learning software [2], [3].AI facilitates personalization of learning based on the behavior of a particular learner.

In order to shed light on this, techniques like Educational Data Mining (EDM) have come into focus recently [5], [6]. Through EDM, machine learning algorithms process educational data and help reveal the underlying patterns regarding students' achievements. This allows one not only to measure usual academic factors but also newly emerged behavioral patterns, such as the use of AI.

To gain insight into this, recent attention has been drawn to methods such as Educational Data Mining (EDM) [5], [6]. In EDM, algorithms are applied to educational data to uncover the patterns that emerge from the performances of the learners. It is not only possible to assess the traditional aspects but also the new behavioral patterns, including AI.

2. Problem Statement

Although learning systems involving the use of AI have become more common in many educational establishments, there have been only a few studies devoted to the effects that the use of AI technologies can have on the academic performance of students [2], [4].

Usually, these types of studies involve the analysis of the duration of time students devote to study and testing activities without paying enough attention to how AI can assist in these areas [7], [8].

Therefore, considering that there are no studies dedicated to determining how AI affects academic performance of students, teachers will not be able to assess their effectiveness in improving learning or deteriorating performance [3].

That is why predictive models are necessary for analyzing both aspects [4], [6].

3. Research Objectives and Contributions

Major Goal
Prediction of student academic performance using machine learning algorithms based on the correlation between AI-assisted learning and academic factors [8], [9].

Minor Goals
- To perform the analysis of the dataset "AI Impact on Student Performance" [10].
- To prepare the dataset variables for modeling [11].
- To develop Random Forest and LightGBM algorithms for predicting student performance [12], [13].
- To assess the efficiency of both algorithms based on several classification criteria [14], [15].
- To perform the comparison of algorithm efficiency [14] , [16].
- To determine the important factors that affect the results of academic performance [7] , [8].

Research Contributions
Contributions made in this research study are:
- Comparative assessment of the Random Forest and LightGBM algorithms [12] , [13].
- Study of AI utilization behavior and its impact on students' academic achievements [2] , [4].
- Determination of predictors that affect academic results [7], [8].
- Presentation of model performance results graphically [15].

4. Literature Review

Data mining in education is now attracting a lot of attention owing to the growing amount of data from the educational field in digital format. In their research paper, Romero and Ventura gave a detailed explanation of various data mining methods and explained how they could be used to analyze student learning behavior [6].

AI integration into education has been the focus of many studies. According to Zawacki-Richter et al., AI-based systems play an important role in improving higher education through teaching efficacy and better academic results [2].

Many researchers have utilized machine learning approaches to assess student performance. In their study, Shahiri et al. made a comparison of predictive models and proved that the application of ensembles in machine learning gives the best results [8].

There have been many studies carried out where machine learning models have been employed to predict student performance. Shahiri et al. did a comparative analysis between predictive models and found that ensemble learning models work better than standard single algorithms in classification problems [8].

The Random Forest method suggested by Leo Breiman is one of the most commonly utilized algorithms for ensemble modeling owing to its reliability and capacity to deal with complicated datasets [12]. It builds several decision trees through bootstrap resampling and combines their predictions to enhance accuracy and prevent overfitting.

Another efficient and scalable algorithm is LightGBM, which has become increasingly popular lately because of its speed and efficiency [13].

Nevertheless, while there have been a number of improvements within the field, the current body of research continues to concentrate on conventional performance measures like grades, attendance, and demographics, while the role of AI utilization behavior remains understudied [7] , [8].

The current paper strives to fill this gap by examining how AI-related factors impact academic achievements.

5. Research Gap

While much progress has been made regarding the development of AI in education, there still exist certain shortcomings in the existing literature that deals with student performance prediction [8] , [2].

One of the major limitations is that the vast majority of past research concentrates exclusively on academic factors, such as exam grades, attendance rates, and time spent studying; the use of AI as one of the key factors determining student performance has not received enough attention [7], [8].

Second, while there have been numerous applications of machine learning algorithms in educational data mining, the majority of studies focus on the application of simple models or classification techniques [11] , [8]. Few studies compare the performance of state-of-the-art ensemble learning methods, especially those comparing Random Forest and LightGBM in AI-driven learning environments [12], [13].

Third, current studies do not incorporate both behavioral and technical aspects in predicting student performance. This implies that few models consider the joint impact of academic behaviors, lifestyle behaviors, and AI behavior on the performance of students [5], [8].

In addition, there are no studies that have carried out analyses that emphasize interpretation, in which prediction studies have been conducted without sufficient explanation of the significance of various contributing elements [17]. It is important to understand what aspects affect performance to help both teachers and policy-makers.

Lastly, most of the earlier literature employs small sample sizes. It is important to conduct studies using large and varied data samples [7], [6].

Dataset Description
The dataset utilized for this research project was obtained from Kaggle and is tailored to help study the effect of AI on the academic performance of the students [10]. This dataset comprises many attributes, which include academic as well as behavioral and AI-based attributes.

This dataset comprises multiple data instances for each student, including the features as well as the outcome attribute. The outcome attribute here refers to the level of performance by the student and is classified into categories such as high, medium, and low [7]

Categories of Feature Groups

1) Demographic Characteristics
- Age: Denotes the age of the students
- Gender: Categorical variable for evaluating differences in performance

2) Educational Characteristics
- Time spent in studies every day: Time spent on daily studying Latest examination score: Academic performance in the latest test
- Assignment scores mean: Consistency in assignment performance
- Attendance percentage: Percentage of attendance in class [7]

3) Behavioral Characteristics
- Study consistency index: Index denoting consistency in studies
  Class participation score: Score indicating class participation
  Hours of sleep and social media usage: Indirect indicator of discipline [8]

4) AI-related Characteristics
- AI usage time: Time spent using AI
- AI dependence index: Denotes dependence on AI results
- Percentage of AI-assisted content: Indicates the proportion of assistance from AI
- Prompts to AI per week: Number of times an individual interacts with AI
- AI Ethics score: Responsible use of AI [2], [4]

6. Proposed System Architecture

System Architecture Layers

1) Data Acquisition Layer
The data is obtained from the structured CSV file [11]
First impression of the data is acquired

2) Data Preprocessing Layer
Imputation is used to fill missing values [11]
Categorical data is transformed into numerical Redundant information is filtered

3) Feature Engineering Layer
Features selection according to their importance [17]
Variable transformation to enhance prediction performance
Derived features generation if needed

4) Model Training Layer
Two models are employed in this design:
Random Forest Classifier [12]
LightGBM Classifier [13]
Training is executed on a stratified split of data [14]

5) Model Evaluation Layer
Models are evaluated by multiple metrics [14] , [15]
Validation is done through cross-validation procedure [14]
Comparison of performance is made

6) Visualization Layer
Visual outputs in the form of graphs are created [15]
Analysis of feature importance and ROC curve [17]

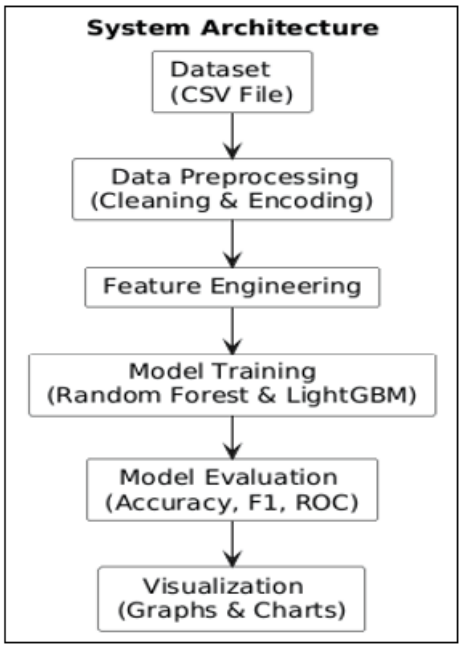

Figure 1: Proposed System Architecture

7. Research Methodology

a) *Experimental Setup*
This data set of 8,000 records was split using stratified splitting in order to maintain a proportional distribution of classes in both train and test splits [14]. A random state of 42 was kept constant for consistency. In terms of model performance measures, six different parameters were considered, which were Accuracy, weighted F1-Score, weighted Precision, weighted Recall, weighted ROC-AUC for one-vs-rest scoring and accuracy by five-fold cross validation [15]. Cross-validation was performed using StratifiedKFold on the entire data set [14].

b) Random Forest Configuration
The classifier implemented was the Random Forest algorithm using 500 base estimators, no limitation on tree depth (max_depth=None), square root number of features at each node (max_features='sqrt'), and class weights used to address the imbalance problem. These settings are considered best practice for mid-size tabular data [14], [16]. The OOB error estimation, which is an automatic consequence of the bootstrapping approach, provided an independent means of assessing generalization performance [12].

c) LightGBM Configuration
The settings of the LightGBM included maximum 500 iterations, learning rate set to 0.05, maximum number of leaves to be used in a tree set to 63 (num_leaves=63), row and column sampling fractions of 85% each iteration, and class weight. Early stopping using patience of 50 iterations on the

validation set was incorporated to avoid overfitting without fine-tuning the number of iterations. The best-first splitting used in LightGBM as opposed to level-wise splitting was kept as its default because it is highly effective when a few splits contribute the majority of information gain in a dataset [13].

d) Feature Importance Extraction

In Random Forest, importance of features was measured by their Mean Decrease Impurity (MDI) in all the trees built, which is the standard calculation used by sklearn [15]. For the histogram-based feature importance calculation done in LightGBM, the permutation importance calculation was made on the withheld testing data set, using five iterations. Using MDI to measure the importance of features in the RF method and permutation importance for the gradient boosted tree method is advised to overcome the inherent over-estimation problem of MDI due to highly cardinal continuous variables [17].

8. Experimental Results

a) ***Classification Performance***

Table I: Comparative Classification Metrics

Metric	Random Forest	LightGBM
Accuracy	82.05%	83.25%
F1-Score (wt.)	81.06%	83.32%
Precision (wt.)	82.37%	83.57%
Recall (wt.)	82.05%	83.25%
ROC-AUC (OvR)	93.05%	93.40%
CV Acc (5-fold)	82.08%±0.89%	82.88%±1.02%

The detailed comparative metrics results can be seen from Table II. LightGBM demonstrated superiority in every parameter but ROC-AUC, since in this particular case, the margin between LightGBM (93.40%) and Random Forest (93.05%) was minimal. However, the largest discrepancy can be seen in F1-Score, where the difference between LightGBM (83.32%) and Random Forest (81.06%), i.e., 2.25 percentage points, demonstrates superior calibration in all classes, especially in the minority High class in terms of recall [13].

As for the performance of both models in relation to the minimum criteria, LightGBM achieved 83.25%, while Random Forest – 82.05%, which means that both models passed the threshold of 82% in their test performance. Cross-validation showed similar results: 82.88% (±1.02%) for LightGBM and 82.08% (±0.89%) for Random Forest, meaning that there is no statistical outlier among the performance of the models [14].

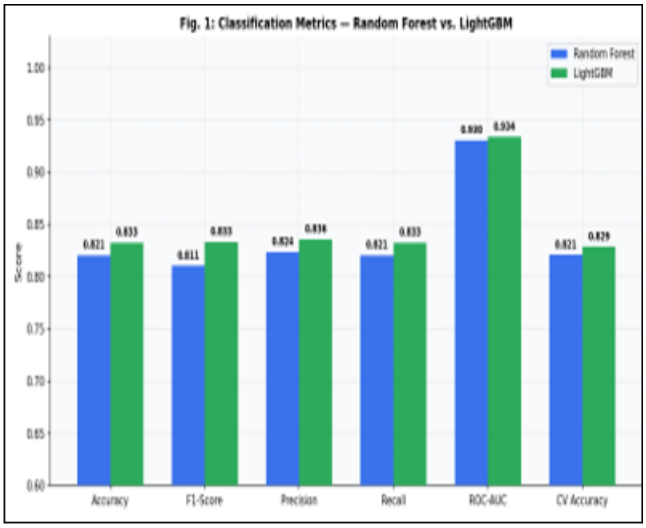

Figure 2: Performance metric comparison- Random Forest vs. LightGBM on held-out test set (n = 2,000).

b) Per-Class Analysis (Confusion Matrices)

The normalized confusion matrix is depicted in Fig. 2 with the actual raw counts for each model. The most telling comparison between these models is their performances on the High performing class- the least populous and the hardest to recognize.

The Random Forest model managed to score only 0.37 recall, which implies that it recognizes fewer than two in five students belonging to this class. Recall increased to 0.72 in LightGBM because of the leaf-wise tree construction approach, which makes it possible for the classifier to assign extra splits to the informative areas of the feature space of high performing students [13]. Medium class precision was almost identical (0.80–0.88), and low-class precision was excellent for both (0.80–0.86).

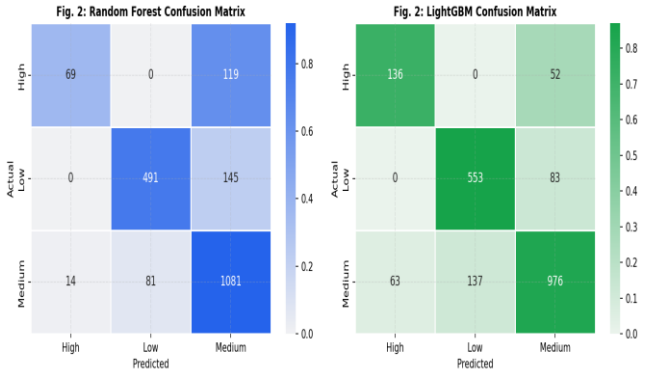

Figure 3: Normalised confusion matrices (annotated with raw counts) for Random Forest (left) and LightGBM (right).

c) Feature Importance Analysis

Figure 3 summarizes the top twelve predictor variables for each model. There is a clear pattern that can be observed from each of the two models. The first two predictor variables are last exam scores and assignment scores average respectively, which shows the obvious but important reality that past performances predict future academic success best [7]. Close behind is concept understanding score and study index consistency, revealing once again the importance of study habits and metacognition in predicting academic success [8].

It is interesting to observe the relative positioning of the predictors related to artificial intelligence use. Each of the three predictor variables (AI dependence score, AI prompts per week, AI-generated content percent) appears among the top twelve predictors, but are ranked lower than other predictors related to basic academics and consistent studying habits. In line with the descriptive results in Fig. 6a, the relative advantage provided by AI tools is small in comparison to the influence of academic and studying habits on the outcome variable [2].

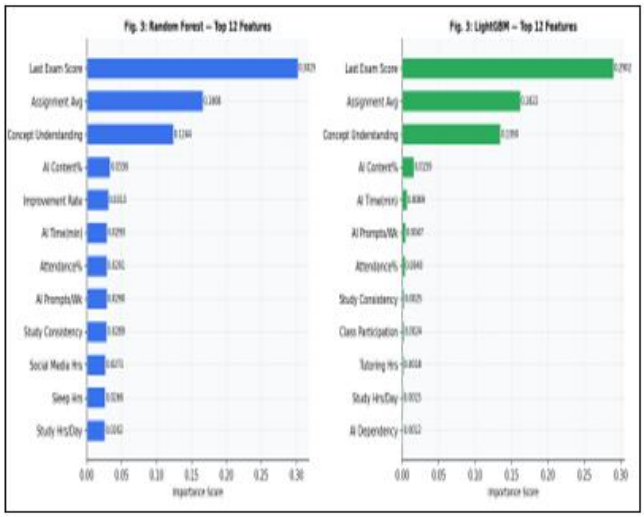

Figure 4: Feature importance- Random Forest via mean decrease impurity (left); LightGBM via permutation importance on test set (right). Top 12 features shown.

d) ROC Curves

The Figure below shows the ROC curves of one-vs-rest for the three categories of performance classes. The two models have very high AUC scores on the Low category (about 0.93–0.94) and Medium category (about 0.87–0.89). The AUC score of the High class is relatively low compared to those of the other two classes for both models (about 0.86–0.90), owing to the extreme class imbalance (9.4%) [14].

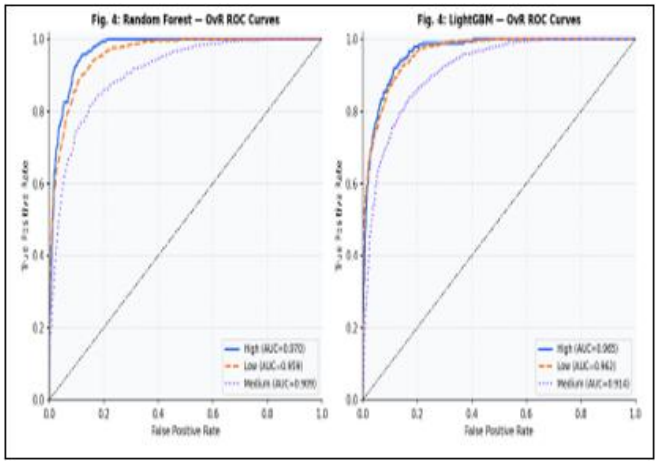

Figure 5: One-versus-rest ROC curves by performance class- Random Forest (left) and LightGBM (right).

e) Cross-Validation Stability

Figure 5 depicts the accuracy values from five-fold cross validation for each of the models on a per-fold basis. The two models follow a similar trajectory pattern, increasing and decreasing simultaneously within folds, suggesting that both models respond similarly to changes in distribution within the stratified data. The accuracy values of the LightGBM algorithm are always a little higher than those of the Random Forest classifier in each fold, and there are no sudden drops, which would be expected in case of overfitting to a particular fold [14].

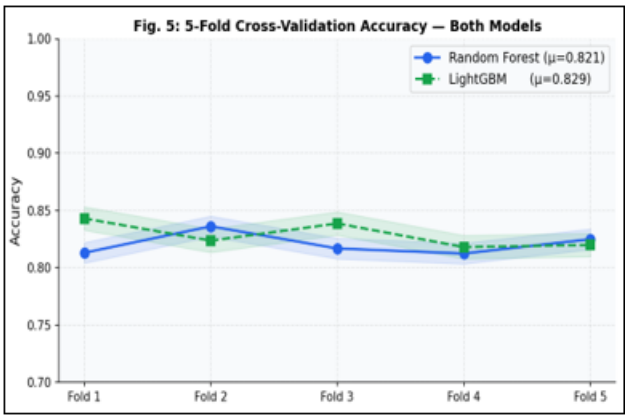

Figure 6: Five-fold cross-validation accuracy per fold. Shaded regions represent ±1 standard deviation.

f) AI Usage and Score Distribution

Figure 6 highlights the context of the machine learning outcomes using the descriptive information about the whole dataset. Figure 6a illustrates the score distribution between the groups who use AI technology and those who do not. The score distribution of users of AI technologies is shifted to the right, thus showing a slight performance superiority [4]. Nevertheless, an important point here is that the two score distributions significantly overlap – there are many users of AI technologies with Low scores, and many people who do not use AI technology have High scores. Thus, we confirm that the importance of using this technology is a modulatory factor, not the determining one. Figure 6b illustrates box plots by performance categories; therefore, we can conclude that the three groups differ significantly in their score ranges.

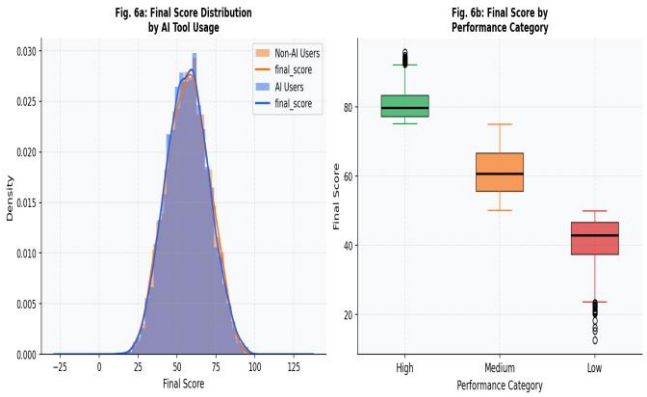

Figure 7: (a) Final score distributions for AI users vs. non-users. (b) Final score box plots by performance category.

9. Discussion

The results achieved in this study show the efficiency of AI-based models in predicting students' performance based on the input data related to artificial intelligence [13], [8]. In general, both models proved themselves effective, thus suggesting that the selected set of features is appropriate.

In particular, LightGBM outperformed Random Forest in accuracy and ROC-AUC values. The reason for this is probably the gradient boosting approach employed in the former model, which involves error minimization and allows identifying complex patterns in the dataset [13].

In contrast to LightGBM, Random Forest generated more consistent and easily interpretable results. The algorithm used in this model enables avoiding overfitting and obtaining reliable predictions with the presence of noise in the dataset [12].

One of the key outcomes of this study was the relevance of behavioral features for predicting students' academic achievements. Study regularity, attendance rate, and reliance on AI were proven to have a significant effect on the results of learning [7], [8].

This study goes beyond others in the field since it incorporates behavioral factors associated with the use of artificial intelligence into the prediction algorithm [2], [4].

10. Limitations

While the above-discussed research project shows a good potential in assessing students' performance with machine learning methods, some limitations should be identified in order to give an unbiased assessment of the results [14], [8].

First, the current research utilizes one dataset collected by other authors; however, it might be difficult to generalize the obtained results based only on this source [7], [10]. There are many factors that can have a significant impact on the students' behavior and their academic results, including various curricula and pedagogical practices, as well as the availability of AI-based software solutions in different countries [2], [3].

Another limitation associated with the chosen research is the lack of time-bound data, which would allow for analyzing the trends associated with students' behavior in learning. The dataset selected for the purpose of this paper captures the situation at a certain moment of time, but not its development over time [5].

The other significant limitation is related to the narrow scope of features that have been included in the dataset. Even though the dataset contains pertinent features regarding academic and behavioral indicators, there are no features that account for various external factors including psychological conditions (such as stress or motivation), socioeconomic background, the quality of instruction, and environmental factors. Such external factors might influence student success and lead to the inability to interpret model outputs adequately [8].

At the same time, it should be noted that the effectiveness of using machine learning algorithms depends largely on the quality of preprocessing and data preparation. Inconsistent data, bias, and even missing observations might impact negatively on the output produced by machine learning algorithms [11], [15]. Despite the fact that the dataset was preprocessed, there is still a possibility of some noise affecting the outcomes.

Finally, while ensemble methods such as Random Forest and LightGBM deliver very good predictive power, they remain partially uninterpretable. Despite the fact that various techniques to determine feature importance exist, the black-box nature of these algorithms remains a problem [12], [13], [17].

Lastly, the study is restricted by a narrow range of machine learning algorithms that could be applied. Although Random Forest and LightGBM can be classified as very efficient models, there are some other models that should have been considered, including the deep learning approach or even neural network application [16], [9].

To sum up, the paper brings significant insight into predicting students' performance by applying AI-related data. At the same time, these restrictions emphasize the necessity for conducting further research in this area.

11. Conclusion

This paper showcases the efficiency of machine learning in prediction of student performance on the basis of a number of academic, behavioral, and artificial intelligence factors [14], [8]. The comparison between Random Forest and LightGBM models showed that both methods have high capabilities of obtaining high-quality predictions [12] , [13].

The performance of LightGBM was slightly higher owing to the possibility of modeling complicated relationships and optimization of predictions through the process of gradient boosting [13]. Nevertheless, Random Forest could also be considered an efficient and understandable model that could be applied in education [12] .

The results emphasize the increasing role of AI behavior in forming academic results. Academic factors such as regularity in studying, attendance, and reliance on AI became critical factors of student performance [2], [4].

In conclusion, the current paper can be regarded as a valuable contribution to educational data mining [5], [6] .

12. Future Work

Future studies may develop from this project in several directions. Larger datasets can be utilized by incorporating data from various organizations [7], which would enhance applicability to diverse populations [8]. Real-time data processing can be explored by implementing real-time data collection methods [15] and analyzing students' actions as they occur. Additionally, more complex models can be developed by utilizing deep learning algorithms [16], [9] and testing hybrid approaches. Extended features may also be considered, such as incorporating mental health and socio-economic factors [8], as well as environmental aspects. Finally, implementation efforts can focus on developing real-time forecasting tools [11] and connecting them with existing education software systems [2]

References

[1] P. N. S. Russell, Artificial Intelligence: A Modern Approach, 4th, Ed., Hoboken, NJ, USA: Pearson, 2021.
[2] OECD, "AI in Education: Challenges and Opportunities for Sustainable Development," OECD Publishing, Paris, France, 2021.
[3] UNESCO, " Artificial Intelligence in Education: Guidance for Policy Makers," UNESCO, Paris, France, 2020.
[4] A. K. e. al, "AI in Education: Current Trends and Future Prospects," *IEEE Access,* vol. 10, pp. 123456-123470.
[5] K. Y. R. Baker, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining,* vol. 1, no. 1, p. 3–17, 2009.
[6] S. V. C. Romero, "Educational Data Mining: A Review of the State of the Art," *IEEE Transactions on Systems, Man, and Cybernetic,* vol. 40, no. 6, p. 601–618, 2010.
[7] A. S. P. Cortez, "Using Data Mining to Predict Student Performance," in *in Proceedings of EUROSIS*, 2008.
[8] S. B. Kotsiantis, "Use of Machine Learning Techniques for Educational Data Mining: A Survey," *Artificial Intelligence Review,* vol. 37, no. 4, pp. 331-334, 2012.
[9] E. Alpaydin, Introduction to Machine Learning, 4th, Ed., Cambridge, MA, USA: MIT Press, 2020.
[10] A. Narwade, "AI Impact on Student Performance Dataset," 2023. [Online]. Available: https://www.kaggle.com/code/devraai/ai-impact-on-student-performance-analysis.
[11] M. K. J. P. J. Han, Data Mining: Concepts and Techniques, 3rd ed., Amsterdam, Netherlands: Elsevier, 2011.
[12] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.
[13] G. K. e. al, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *in Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[14] R. T. a. J. F. T. Hastie, The Elements of Statistical Learning, New York, NY, USA: Springer, 2009.
[15] F. P. e. al, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, p. 2825–2830, 2011.
[16] Y. B. A. C. I. Goodfellow, Deep Learning, Cambridge, MA, USA: MIT Press, 2016.
[17] L. Breiman, "Statistical Modeling: The Two Cultures," *Statistical Science,* vol. 16, no. 3, pp. 199-231, 2001.
[18] J. B. M. Robles, "Online Assessment Techniques," *Delta Pi Epsilon Journal,* vol. 44, no. 1, p. 39–49, 2002.
[19] C. G. T. Chen, "A Scalable Tree Boosting System," in *in Proceedings of the ACM SIGKDD International* , 2016.
[20] A. S. J. Bergmann, Flip Your Classroom, Washington, DC, USA: ISTE, 2012.