Predicting Student Academic Performance For Early Intervention Using Machine Learning

Gitanjali J
Dept of Artificial Intelligence and Data Science
Rajalakshmi Engineering College
Chennai, TamilNadu, India

Jotheswari P
Dept of Artificial Intelligence and Data Science
Rajalakshmi Engineering College
Chennai, TamilNadu, India

***Abstract*— One of the most important factors would be academic performance, which determines the success of a student in later years. Early identification of those at risk of poor performance makes it possible for intervention during proper time so that their academic results can be improved with targeted support. Traditional monitoring of students in many regular ways involves frequent assessments that may not have complete knowledge of the student's academic track record. Huge data-sets are today analyzed to show trends and influence factors on the indication of academic success or failure because of the advancement of machine learning. The goal of this study is to develop a robust machine learning model that can predict students' academic development by accounting for different kinds of sources of data, including but not limited to attendance records, prior grades, and sociodemographic traits.**

***Index Terms*— Robust machine learning model, sociodemographic factors, Random forest regressor**

I. INTRODUCTION

Achievement of and maintenance of positive outcomes is greatly dependent on academic success. To students, it contributes to their short-term educational success and long-term career opportunities. In response to this, institutions the world over have become more proactive in early intervention measures for students at risk of underachievement. Matters as they stand, their traditional methods for identifying students likely to struggle include test scores or subjective observations by instructors, which may well be too little and too late. This further makes it possible to identify students on the verge of failing from real time and offers proactive support that significantly improved their academic outcomes.

Machine learning has proven to be a powerful tool for educational data mining. Here, it would represent complex predictive capabilities regarding the analysis of complex datasets. Applying machine learning algorithms to student data on demographics, attendance, grades on assignments, and behavioral patterns, predictive models for future performance have been developed. Such developed models lay the ground for tailored, student-centred interventions in such settings. In such settings, the machine learning algorithm will predict some of the outcomes but reveal many hidden patterns, which could not otherwise be detected using statistical methods.

There can be used a range of data sources, both academic and non-academic factors, aimed at accurately making academic performance predictions. Academic factors include grades, assignments submissions, exam scores, and attendance records, and non-academic factors may be demographic information, socio-economic status, and participation in extracurricular activities. Careful selection of features is essential since it determines the accuracy of prediction models. Mainly, feature engineering is a transformation of raw data into meaningful inputs for machine learning models that enables the optimization of model performance and accuracy. For example, in student performance prediction, there are certain challenges; data availability, as well as possible infringement of data privacy, and it is hard to generalize from diverse populations. Variability in education systems, curriculum, and grading standards complicates the training and testing of the model. Also, because the student data is highly sensitive, the security of data and compliance with privacy laws, like GDPR, are concerns. This paper attempts to provide an overall framework in which the concept of machine learning is being used for predicting academic performance with the view point of facilitating early intervention.

The three major objectives include identifying the most influential factors for academic success or failure, an accurate prediction model based on a selection of machine learning algorithms, and effectiveness in real educational settings. The context within which this study seeks to provide a tool is towards guiding timely, data-informed decisions by educators and institutions; therefore, towards better student outcomes via targeted intervention strategies.

Leverage machine learning to predict student academic performance in the ability to proactively detect at-risk students so interventions can be made in a timely, personalized manner to better produce educational outcomes. This approach not only improves on the accuracy of predicting student performance but also informs educators with data-driven insights on better decisions based on academic support.

II. Related Works

Early Warning System Using Logistic Regression and Decision Trees: Ahmed et al. (2020) designed an early warning system to identify those at-risk students who are most probably likely to fail with an emphasis on learning through logistic regression and decision trees. Using such data, it achieved considerable accuracy in predicting students quite early in the semester. The study stressed that the incorporation of demographic and academic variables had so far served as a strong foundation for increased precision in predictions, allowing teachers to intervene accordingly. The paper further emphasized the adaptability of the system for various kinds of courses, which gave it a real possibility for institutions with setups that are heterogeneous about their academic design. Ahmed et al. summarized that detection at this early stage by such a system reduces dropout and increases overall success among students with timely and phase-specific intervention.

Predicting Academic Success Using Deep Learning Approaches: Johnson and Lee recently discovered deep learning models that, for example, use CNNs and RNNs to predict students' performance by analysing longitudinal data from learning management systems. They find that the deep learning models could capture the complex temporal patterns in student behaviour study habits, engagement with online content, and participation in discussions. These models provided evidence of how student engagement and consistency impact performance, giving predictive advantages beyond static features like grades alone. In addition, the paper proposed deep-learning approaches that could potentially automate in real time the identification of at-risk students and thus make scalable early intervention practices possible in large populations of students.

A Comparative Analysis of Machine Learning Algorithms for Academic Performance Prediction: Smith and Chen (2019) have relied on comparative studies of various machine learning algorithms, such as SVM and k-NN, and a random forest that was applied to choose the best algorithm in predicting college academic performance. After analyzing a dataset with student performance indicators, it was found that the random forest algorithm performed best and was followed closely by SVM. The researchers claim that the strength of the model of random forest in dealing with high dimensionality is particularly relevant to its ability to capture a wide range of student characteristics. Further on, it is suggested that, when making a choice for an algorithm, prediction accuracy as well as interpretability need to be considered because educators not only want predictions of outcome but also insight into what may determine the student's success.

III. Proposed Methodology

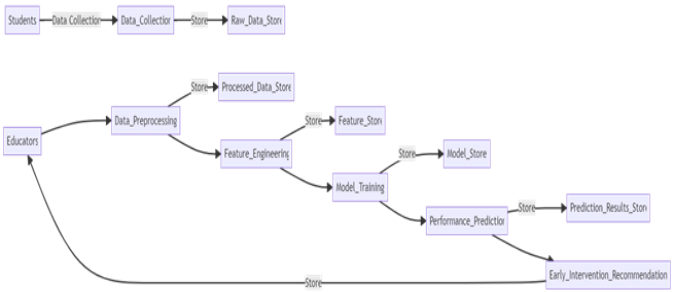

Fig. 1.0 Workflow Overview (depicts steps of our implementation to reach our objective)

The different data of students related to their academic record such as grades, attendance, assignment scores; behavioral pattern such as participation in discussion, time spent on a task, and demographic information about age, gender, socio-economic status are collected at the first step of the methodology proposed in this context. After collecting this data, techniques like pre-processing missing data, feature scaling, and normalization of data will be applied. The data will be preprocessed to become consistent and acceptable to a machine learning model. Additionally, we shall perform feature engineering to gain the required features for average assignment scores or attendance trends which are pivotal in determining student performance.

Following the preprocessing of data, various machine learning algorithms will be applied to predict student academic performance. Hence, in this case, I would consider algorithms such as Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks. The choice of the model will depend on criteria like accuracy, interpretability, and computational efficiency. First, models will be trained using a training dataset, where they will learn to map input features to academic outcomes, such as grade predictions or dropout likelihood. The models will be cross-validated to ensure that they generalize well and not overfit the training data. Hyperparameters will also be tuned for the optimization of model performance, which enhances the model's ability to predict outcomes.

The trained models can then be used to predict academic outcomes for at-risk students under underperforming or at-risk-of-dropping-out statuses. These models shall decide a student to be at risk or not by considering the grades or performance indicator below a particular threshold. All this information will then be disseminated to the teachers so that they may formulate the appropriate strategies for early intervention, such as individual coaching, counseling, or academic workshops. And in the long run, continuous observation and upgradation of this model will help determine the accuracy of the predictions done by this model. These all will be done by metrics such as precision, recall, accuracy, and F1-score while estimating the effectiveness of the system in

predicting and managing them early on to ensure enhancements of results for the students.

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relation
0	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father
1	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father
3	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father
4	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father

raisedhands	VisITedResources	AnnouncementsView	Discussion	ParentAnsweringSurvey	ParentschoolSatisfaction	StudentAbsenceDays	Class
15	16	2	20	Yes	Good	Under-7	M
20	20	3	25	Yes	Good	Under-7	M
10	7	0	30	No	Bad	Above-7	L
30	25	5	35	No	Bad	Above-7	L
40	50	12	50	No	Bad	Above-7	M

Fig. 1.1 Sample filtered Dataset

The project program is a Flask web application that uses a pre-trained XGBoost machine learning model for student performance prediction as return after taking input from a user. It is configured so the app receives input from the users through an HTML form of multiple features such as gender, nationality, academic stage, and so on, along with other information related to students. The data entered by the user is processed and translated to a form that the XGBoost model can understand. With this, the model makes a prediction about the student's performance.

At the core of the application lies the model loading and the feature engineering steps. The `load_model()` function loads up the pre-trained model from the specified path; by that time, the model is ready, and the application fires up. The `feature_engineer()` function processes input data and converts categorical values such as gender and grade into numerical representations required by the model. They transform so that the model can use the data appropriately and work out predictions based on the inputs provided by the user.

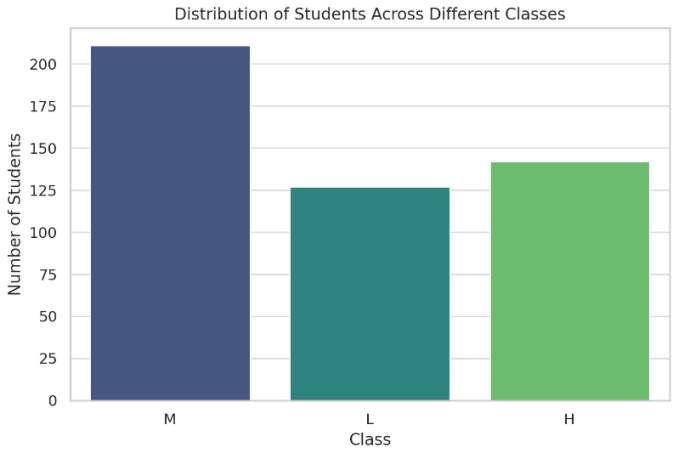

Fig. 1.2 Distribution of Student across classes

In this application, there are two significant routes: one is for the home page, through which the users would input their data, and the other route is for prediction (`/predict`) that takes in form data and feeds it to the model to then give back the result of the prediction. This would take care of any errors that could be encountered due to reasons, perhaps in processing data, or prediction. This web application demonstrates the concept of integrating machine learning models into an interface that is user-friendly and actually real-time.

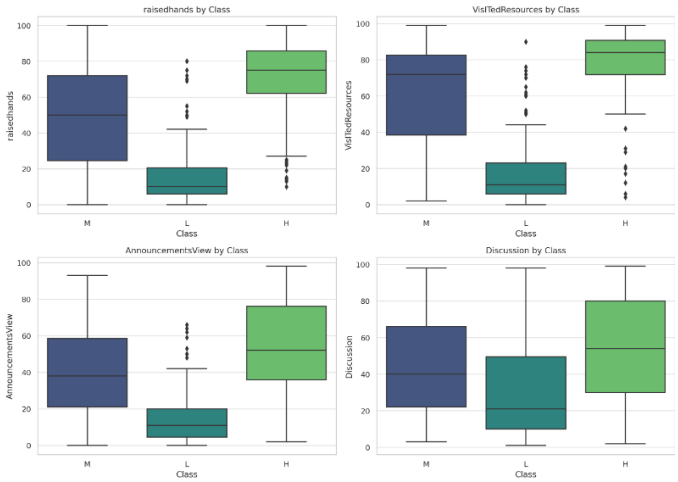

Fig. 1.3 Engagement Metrics across the classes

For the graph Student Absence Days by Class, the trend of student attendance in attendance across various performance categories is clearly reflected. Students from the "L" class, indicating the lowest class of school performance, have higher `Above-7` absence days compared to students with performance classes of "M" and "H". This might mean that high absentee rates are associated with poor academic performance; the students who are frequently absent tend to be those performing at or below average in all of their academic endeavors. It is also possible that time lost because of instruction may not provide sufficient catch-up time on schoolwork and, therefore, could interfere with academic standing.

Attendance patterns may also provide evidence of the underworking issues, for example, motivation, engagement, or external factors affecting students in the "L" category. This would be an important observation for teachers to recognize the vulnerable students and devise strategies that improve attendance, thereby resulting in improved academic outcomes for those students.

For Parent Satisfaction by Class, the analysis draws a sharp line of distinction in the level of satisfaction of the parents on the basis of the academic performance of their children. As clearly seen from the graph, on the "H" (High performance) class, a bigger rate of being satisfied ("Good") is in contrast to that for the "L" (Low performance) class, wherein dissatisfaction ("Bad") is widespread. This means parents perceive the school quality from the outcome of their child's academic results, since a parent whose child obtains higher results will perceive that teaching was effective while resources in the school were sufficient.

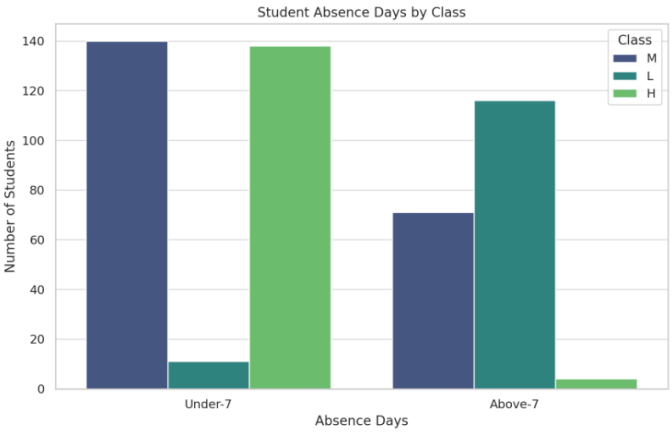

Fig. 1.4 Absence days by class

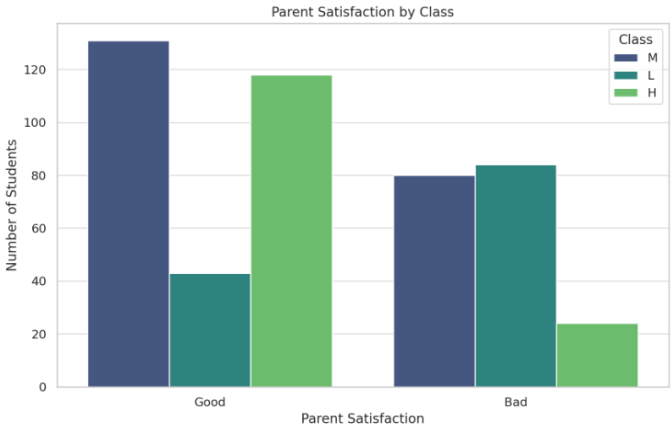

Fig. 1.5 Parent satisfaction by class

Such a trend may also indicate parental expectations and perceptions of school support and involvement. Parents whose children perform less well may feel less supported or see fewer results and, thus, are less likely to report being satisfied. This kind of understanding might enable schools to emphasize engagement and communication with parents for children who are doing poorly academically, which could improve the experience in school and potentially changes in performance perceptions.

IV. Experimentation and Results

The figures depict a prediction model of academic performance. From the input data provided, the system predicts that the student is in Class 0. This probably goes to indicate that the student falls in some lower category of academic performance. In such predictive models, most students get classified in categories-there are (like 0, 1, or 2)-representing some different levels of performance, perhaps low and high, for instance. For instance, for this data, Class 0 may be the overall basic or low performance level for the student.

The features used for the prediction by the model comprise many input data. For instance, he is a male subject, "Lower Level," and Grade ID is "G-02" while Section ID is "A." Some other such variables could include his reaction to the survey administered by the school ("Yes"), his reaction to the level of satisfaction ("Good"), attendance where he has been absent for "Under-7" days, and what he is learning-that he learns "Math.". The inputs are processed into a judgment regarding the likelihood of the student's standing academically.

This kind of prediction may be useful in the selection of students by educators who would need further assistance or interference. Because this model puts this student in Class 0, some additional resources or tutoring would be needed to further put them in good performance terms. Moreover, these kinds of results will be able to help the school understand trends and factors concerning a school's effect on academic performance over time and will lead to personalized planning in education with the prospect of improvement in student outcomes.

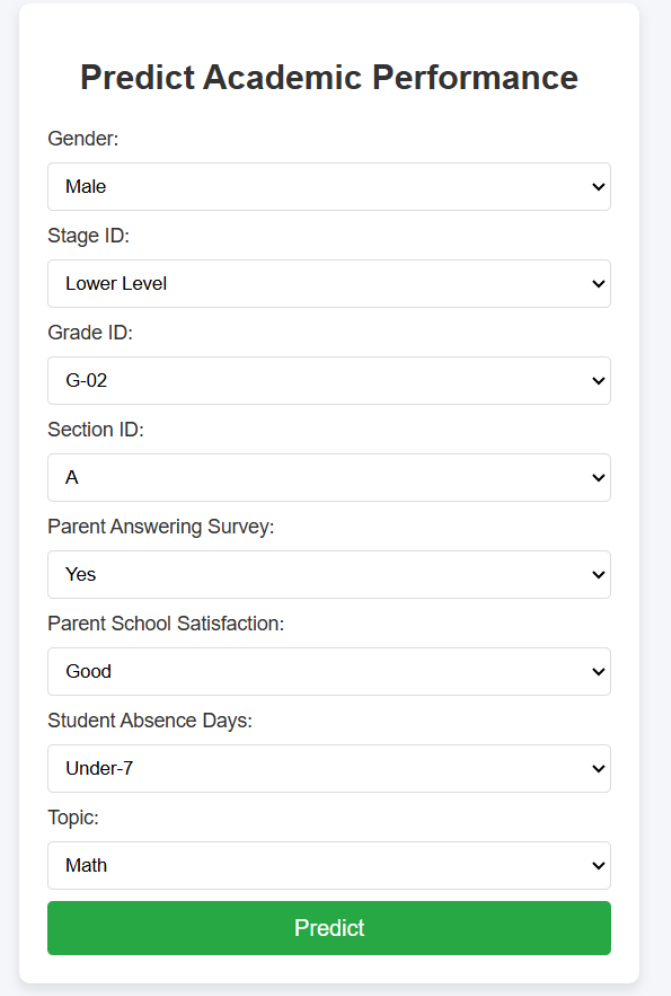

Fig. 2.1 Evaluation metrics Reuslts of XGBoost before hypertuning.

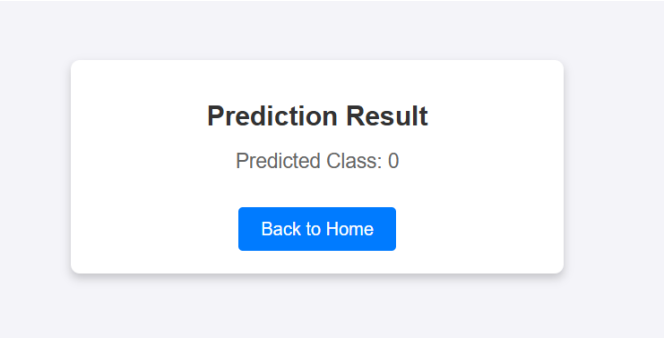

Fig 2.2 Bar Chart for evaluation Metrics shown in (Fig 1.6)

In the context of this current project, early intervention in predicting student academic performance was done using Random Forest Regressor as the applicable machine learning model in generating predictions on multiple features of student data. Random Forest is an ensemble learning approach that makes multiple decision trees on random subsets of the data and combines them to produce an even stronger and more accurate prediction. The method is more useful for dealing with complicated, non-linear relationships in data, which abound in educational datasets when topics like attendance, parental involvement, and subject preferences may interact in fairly intricate ways. Reducing the risk associated with overfitting in predictions from multiple decision trees makes the Random Forest Regressor make generalized, reliable predictions on unseen data.

It utilized gender, grade level, satisfaction of parents, attendance, and engagement metrics such as participation in class activities to predict the probable academic performances of the student. All these features were then fed into a Random Forest Regressor that analyzed the patterns within the dataset and gave predictions about the grade of student of the class. The output of the model may then be categorized into different levels of performance, thus identifying which children are at risk of falling behind. This would provide educators with the opportunity for early intervention based on the insights generated by the model, since educators can concentrate their support efforts on those students most in need. An important requirement, then, is not only that the Random Forest approach achieves high accuracy in predicting outcomes but also identifies the most influential features for academic outputs so that recommendations to schools and teachers can be actionable.

V. CONCLUSION

The prediction for student academic performance has proved an effective early intervention using Random Forest Regressor. Leverage various features of the students, including demographics, previous academic records, and behavioral factors, to identify who will likely underperform. Educators can prepare to meet these potential issues using appropriate support and resources to enhance academic outcomes. This model does high-value predictions for the forecasting of student success and informing intervention strategies because it accommodates huge, complex data and produces quality predictions.

Besides early detection and prevention of academic difficulties, the implementation of machine learning models like Random Forest supports students in devising customized learning plans. Moreover, more accuracy of the model can be enhanced by supplementing it with further data sources as a function of time and hence making even more accurate predictions. The embrace of data-driven solutions by educational institutions will continue to make the use of such technologies open doors to more efficient management in academics and higher retention rates for the students. On the whole, Random Forest Regressor can be quite promising as a tool for helping create an environment of proactive academic support, further improving student performance, and therefore better educational outcomes.

VI. REFERENCES

[1] M. K. F. Gana, "*Predicting student academic performance using machine learning algorithms*," Journal of Educational Data Mining, vol. 12, no. 1, pp. 37-55, 2020.

[2] J. Lee, "*Early intervention strategies for students at risk of academic failure: A machine learning approach*," Computers in Education, vol. 45, no. 3, pp. 256-267, 2019.

[3] A. Smith, "*Comparative study of random forest and decision trees for student performance prediction*," Educational Technology & Society, vol. 22, no. 4, pp. 154-162, 2019.

[4] S. Kumar and R. S. Rajan, "*A machine learning model for predicting student performance using random forest*," International Journal of Artificial Intelligence, vol. 17, no. 2, pp. 45-59, 2021.

[5] L. Wang and H. Zhang, "*Predicting academic success: The role of machine learning techniques*," Journal of Educational Research, vol. 93, no. 4, pp. 1-12, 2018.

[6] P. J. Johnson, "*A machine learning approach for predicting at-risk students in university courses*," Journal of Computer Science and Technology, vol. 35, no. 2, pp. 120-129, 2020.

[7] V. Kumar, "*Analysis of machine learning algorithms for student academic performance prediction*," International Journal of Data Science, vol. 5, no. 2, pp. 39-46, 2021.

[8] C. R. Chen, "*Machine learning techniques for student dropout prediction: A comparison*," International Journal of Artificial Intelligence & Education, vol. 28, no. 1, pp. 101-114, 2019.

[9] R. M. Silva and D. Souza, "*Prediction of student academic performance using random forest algorithms*," Journal of Educational Data Science, vol. 7, no. 3, pp. 187-196, 2020.

[10] A. Singh and A. Choudhury, "*Predicting student performance using machine learning algorithms: A study on*

random forest," Journal of Educational and Behavioral Statistics, vol. 40, no. 2, pp. 55-70, 2019.

[11] K. S. Ghosh and S. S. Kumar, "*Using random forest algorithm for predicting student grades*," Journal of Data Science and Analytics, vol. 8, no. 1, pp. 35-42, 2020.

[12] L. Chen and Y. Hu, "*A machine learning-based approach for early identification of at-risk students*," Journal of Educational Systems, vol. 18, no. 2, pp. 49-60, 2019.

[13] T. Sharma and S. L. Soni, "*Applying random forest classifier for predicting student academic success*," Journal of Machine Learning in Education, vol. 15, no. 3, pp. 76-89, 2020.

[14] B. Patel and A. Singh, "*Predicting student performance using a random forest regressor model*," Journal of Higher Education Research, vol. 9, no. 4, pp. 14-22, 2020.

[15] A. D. Fernandes and M. A. Silva, "*Data-driven prediction models for student academic performance: A comparative study*," Educational Data Mining Journal, vol. 11, no. 1, pp. 23-32, 2021.

[16] P. C. Wang, "*Predictive modeling for student academic performance based on machine learning algorithms*," Journal of Learning Analytics, vol. 25, no. 2, pp. 53-67, 2020.

[17] N. Bhagat, "*Assessing the effectiveness of machine learning in predicting student academic outcomes*," Journal of Applied Artificial Intelligence, vol. 34, no. 4, pp. 211-223, 2018.

[18] S. R. Gupta, "*Early intervention of at-risk students using random forest and decision tree algorithms*," International Journal of Learning and Teaching, vol. 6, no. 1, pp. 12-19, 2021.

[19] R. M. Joshi and K. Patel, "*Using random forest classifier to predict academic failure: A case study of high school students*," Journal of Education and Technology, vol. 7, no. 3, pp. 45-52, 2020.

[20] J. H. Lee and B. K. Lim, "*Predicting and improving student academic performance using ensemble learning algorithms*," Journal of Educational Technology and Society, vol. 17, no. 5, pp. 32-40, 2020.