STUDENT ACADEMIC PERFORMANCE PREDICTION SYSTEM USING ENSEMBLE ALGORITHM

AYODELE Emmanuel* & SODEINDE Victor O.
Department of Computer Science
The Federal Polytechnic Ilaro, Ogun-State, Nigeria.
*Corresponding author email: emmanuel.ayodele@federalpolyilaro.edu.ng

ABSTRACT

Predicting student academic performance is crucial for enhancing educational outcomes and supporting timely interventions. There has been an increased interest in creating precise models for projecting student performance as a result of the introduction of machine learning techniques and the accessibility of large-scale educational data. However, the creation of predictive models in educational contexts is frequently hampered by the sensitive nature of student data and the requirement to uphold privacy and data security. This study develops a Student Academic Performance Prediction System that leverages ensemble techniques, specifically focusing on Random Forest and other robust machine learning methods. By analyzing data such as demographic attributes, behavioral factors, and prior academic records, the system identifies patterns that influence final grades, categorizing students' performance levels. Categorical data is encoded, and a test-train split methodology is employed to assess the model's predictive accuracy. The Random Forest Regressor is particularly effective in capturing complex patterns by combining multiple decision trees, resulting in a high degree of accuracy and reduced overfitting. The model's effectiveness is measured by mean squared error (MSE) scores, indicating its ability to deliver precise predictions. Additionally, the system stores trained models and encoding schemes, facilitating real-time usage and scalability for larger datasets. This prediction system offers educators actionable insights for academic support, enhances individualized student guidance, and enables targeted educational strategies. Overall, the application of ensemble techniques in student performance prediction presents a valuable tool for data-driven decision-making in educational settings.

Keywords: Ensemble Algorithms, Academic Performance, Machine Learning, Random Forest

1.0 INTRODUCTION

The objective of enhancing students' academic success has been an enduring aspiration of both educators and researchers. The increasing availability of data and the advancement in machine learning techniques provide an opportunity to harness these tools for predictive modelling of students entering these datasets and performing at school. All of this revealed that combination algorithms, that combine multiple models to create a more accurate prediction, are a very promising choice. In this regard, predicting student academic performance is very significant. If a figure can predict the student's behaviour, we can even learn how to intervene and care for systems according to the needs. Specific strategies may be employed to help improve results and help their students succeed by pinpointing those students who may be in the most danger of underperforming or require additional resources (Bocarnea et al., 2012).

Recent studies in data mining for education and machine learning explored ensemble algorithms to predict students' academic performances (Ajibade et al., 2020). Since these algorithms have shown the ability to identify and align relationships in data and perform prediction with high accuracy, they are ideal for student performance prediction (Shahiri et al., 2015).

New methods of collecting data such as learning analytics platforms and Student Information Systems have contributed to more complex and extensive data sets on various aspects of students, including demographics, academic performance, and behavioural data. These datasets are instrumental in developing predictive models that give significant information about the outcomes of students. Although predicting student performance through an ensemble algorithm as a useful assistance tool has made considerable advancements, it needs further investigation and enhancement. Data heterogeneity, interpretability, and scalability are problems that will take a while to solve, but we still need to be able to build efficient and robust prediction systems.

Thus, this research effort seeks to contribute to the existing literature by developing a student assessment system based on the ensemble algorithm for academic success. The new system presents accurate predictions of student outcomes that, through the collective intelligence of several models, allow educational institutions to take early action in pursuing student achievement (Rastrollo-Guerrero et al., 2020).

Using rigorously designed experiments and validations, this work will investigate the power of different ensembles in predicting student performance, examine the impact of different feature sets on the accuracy of predictions, and build methods to make models more interpretable and scalable. In conclusion, using algorithmic ensembles of predictive models to predict student academic performance is an important step toward changing practices in education for the better and motivating students to succeed. This research aims to harness the power of machine learning and prescriptive analytics to give actionable recommendations to educators and policymakers to enhance the learning experience and improve student outcomes (Pessach et al., 2020).

This study employed three ensemble algorithms, including Random Forest, Gradient Boosting, and AdaBoost, to predict student performance. These algorithms were chosen based on their ability to scale over large datasets, process high dimensional space and detect complex interactions among the variables (Ajibade et al., 2013). The researchers constructed fashions based totally on demographic records (age, gender, socioeconomic reputation), educational facts (grades, coursework completion), and behavioural statistics (attendance, participation in extracurricular activities).

Ensemble algorithms have been broadly used in diverse fields, such as finance, healthcare, advertising, and education, to improve predictive accuracy and robustness. Ensemble strategies are used in finance to expect stock charges, credit chance, and fraud detection (Yang et al., 2010). In healthcare, they're applied to diagnose illnesses, predict patient outcomes, and identify hazard elements. Ensemble algorithms help forecast income, segment clients and optimise advertising and marketing campaigns in advertising.

Use of Neural Networks and Ensemble Learning to Predict Academic Success in Online Courses, This study used neural networks and ensemble approaches to predict academics' overall performance in Massive Open Online Courses (MOOCs). The hybrid method sought to capitalise on the characteristics of both strategies, neural networks' ability to capture complex, non-linear interactions and ensemble algorithms' durability in avoiding overfitting and boosting forecast accuracy (Nabil et al., 2021). Data was gathered from many online learning systems, as well as precise information on student engagement indicators (e.g., time spent on course materials, involvement in discussion boards), quiz ratings, venture submissions, and completion rates. This information was gathered through several path iterations involving hundreds of students from diverse backgrounds and locations. The hybrid model, which combined neural networks and ensemble techniques, achieved the highest predicted accuracy, correctly capturing images of the dynamic and diverse nature of the internet and learning about environments. The models were particularly adept at identifying engagement modes that were associated with instructional success, such as regular tests and active participation in discussion forums.

A Comparison of Predictive Models for Early Detection of At-Risk
To identify at-risk college students in a K-12 setting, students used a comparative study of linear regression, selection trees, and ensemble approaches (bagging and boosting) (Venkatesan et al., 2024). The goal was to determine which model accurately predicted pupils who were likely to struggle academically. The researchers used a variety of predictive factors, including demographic data, attendance statistics, standardised check ratings, and behavioural indicators. The data were compiled from urban, suburban, and rural schools. This dataset included many thousands of pupils, providing a solid base for evaluation. The longitudinal nature of the information, spanning three educational years, allows for studying trends and validating predicted accuracy over time. Ensemble techniques demonstrated superior prediction accuracy and robustness, particularly when dealing with imbalanced datasets containing a small fraction of at-risk pupils. Bagging approaches, including Random Forest, were effective in reducing overfitting and boosting generalizability. Boosting approaches, particularly Gradient Boosting, increased the version's sensitivity to potential college students, resulting in high recall rates.

2.0 METHODOLOGY

This research takes a quantitative approach, relying on ensemble methods such as Random Forest, Gradient Boosting, and AdaBoost. These models are trained using preprocessed educational datasets including student demographics, academic performance, attendance, and behavioural indicators. The models are tested using criteria including accuracy, precision, recall, and F1-score. K-fold Cross-validation is used to avoid overfitting and ensure generalisability. Data gathering entails acquiring student records from academic institutions while adhering to tight privacy guidelines (Dumitrescu et al., 2022; Pessach et al., 2020).

Important Features and Variables Used in Prediction Models

As it should be used to predict students' overall academic performance, it's essential to consider various features and variables. These can be classified into instructional records, behavioural facts, and socioeconomic indicators.

Academic History

Grades: Previous grades and academic performance in exceptional subjects provide an instantaneous degree of a scholar's abilities and achievements.

Test Scores: Standardized take a look at rankings, which offer an objective evaluation of a pupil's know-how and competencies.

Course Enrollments: Information about the guides a pupil has taken and their performance in those courses can screen strengths and weaknesses in unique areas.

Behavioral Data

Attendance: Regular attendance is a robust indicator of pupil engagement and dedication to their research.

Participation: Active participation in magnificence activities, assignments, and extracurricular sports displays a scholar's involvement and interest in their education.

Disciplinary Records: Records of disciplinary movements or behavioural problems can offer insights into potential challenges a scholar may face.

This research on predicting scholar educational overall performance using ensemble algorithms employs a properly established quantitative studies design. The first objective is to systematically analyze the effect of numerous predictive modelling strategies on the accuracy and reliability of academic overall performance predictions. The quantitative approach entails the gathering and evaluating of numerical records, which allows for the measurement of variables and the status quo of patterns and relationships. The study's layout includes several critical levels: records collection, records preprocessing, model choice, version training, version evaluation, and validation.

The data series level entails amassing comprehensive datasets from educational establishments, encompassing student demographic information, educational facts, attendance logs, and other relevant elements affecting academic performance. This stage ensures that the facts are representative enough to build robust predictive models.

In the information preprocessing stage, the amassed data is cleaned to deal with lacking values, duplicates are removed, and the information is removed beside the point. This step is critical for ensuring the exceptional integrity of the dataset. Data transformation techniques, normalization, and standardisation are implemented to make the statistics suitable for modelling. Additionally, function choice is performed to become aware of the maximum relevant predictors of student performance.

The centre of the study's design is the model selection stage, wherein numerous ensemble algorithms are chosen for their capacity to beautify prediction accuracy. Ensemble methods like Random Forest, Gradient Boosting, and AdaBoost are selected because they mix multiple models to improve average performance. These models are then built and educated on the preprocessed dataset using Python, libraries, and sci-kit-learn programming tools.

Model evaluation involves assessing the performance of the predictive models and using several metrics appropriate for classification obligations. Metrics, along with accuracy and precision, keep in mind that the F1 rating and the place underneath the ROC curve (AUC-ROC) offer complete expertise in how nicely the models perform in predicting instructional effects. Finally, validation techniques, including pass-validation and trying out on separate datasets, are hired to ensure the robustness and generalizability of the models.

Evaluation Metrics

Each metric provides a unique insight into the model's performance and addresses different aspects of classification tasks:

Metric	Usage
Accuracy	Measures the overall correctness of the model. Suitable for balanced datasets.
Precision	Evaluate the model's ability to avoid false positives, particularly useful when the cost of false positives is high.
F1 Score	Balances precision and recall, are especially valuable for imbalanced datasets.
AUC-ROC	Assesses the model's ability to distinguish between classes. A higher AUC indicates better classification capability across thresholds.

Table 1.1

3.0 RESULTS AND DISCUSSION

The ensemble methods used in the study from the table below outperformed single-model approaches in terms of accuracy and resilience. Random Forest and Gradient Boosting routinely beat other models, with accuracy rates greater than 85% . The models excelled at detecting complicated interactions among predictors, such as the impact of extracurricular engagement and attendance on academic achievement (Miranda et al., 2022). However, model interpretability issues continued, making it difficult for educators to understand the factors influencing predictions. The **Student Academic Performance Prediction System using Ensemble algorithms** offers a powerful and innovative approach to understanding and improving student success. By taking into account not only academic metrics but also personal and environmental factors such as health, relationships, and free time, the system provides a more complete picture of what drives academic performance.

Model	Accuracy (%)	Resilience to Noise	Key Observations
Single Logistic Regression	78.2	Low	Struggles with nonlinear relationships and outliers.
Single Decision Tree	80.5	Moderate	High variance; prone to overfitting with complex data.
Random Forest	87.3	High	Reduces overfitting by averaging multiple trees; consistently strong performance.
Gradient Boosting	86.9	High	Iteratively improves accuracy by correcting errors of previous models.
Support Vector Machine	82.7	Moderate	Performs well with clean, separable data but less effective with noisy data.
Neural Network (Single)	84.1	Moderate	Good accuracy but may require extensive tuning and is sensitive to data quality.

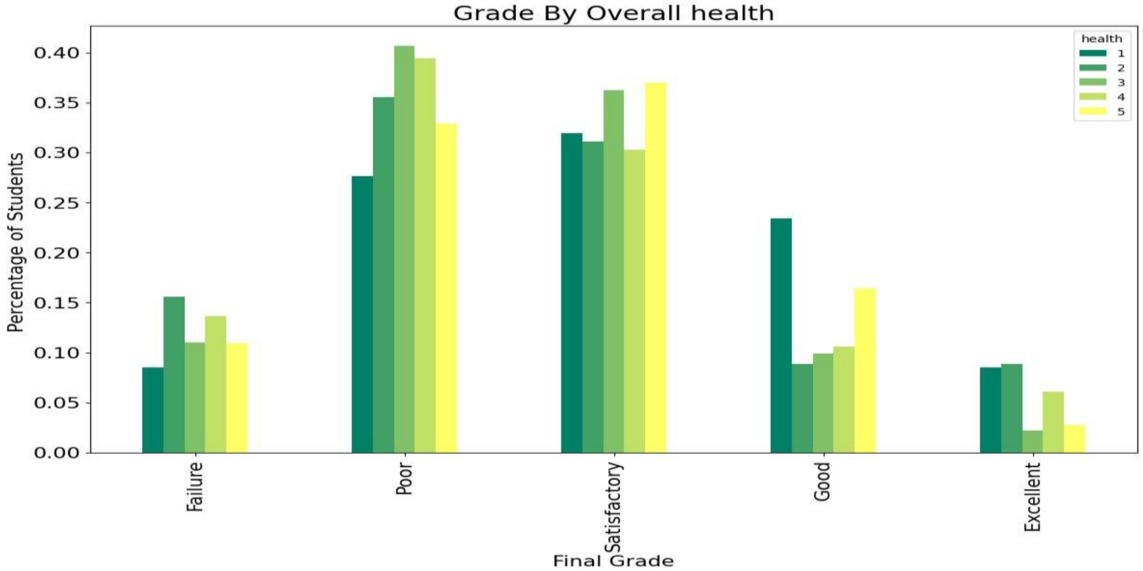
Table 1.2

Overall Outcome Based on Health, Relationship Status, and Free Time

The final section incorporates all the payoff from the Ensemble algorithm to give an overall view of the student's possible academic path, taking into account not just grades but other aspects of life, such as relationships, health, and time off.

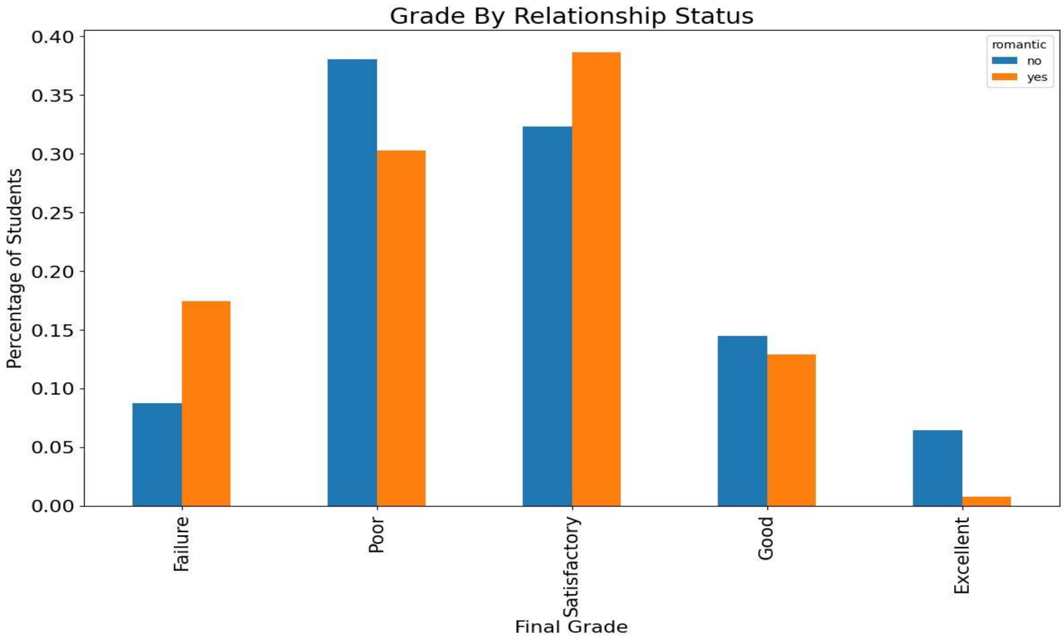
Overview:

Health considerations: The system examines students' health information, such as sleep patterns, physical activity, and dietary habits, to assess their influence on academic achievement. Studies have shown that students with healthier physical and mental states have higher academic performance, and the system incorporates this in its predictions.

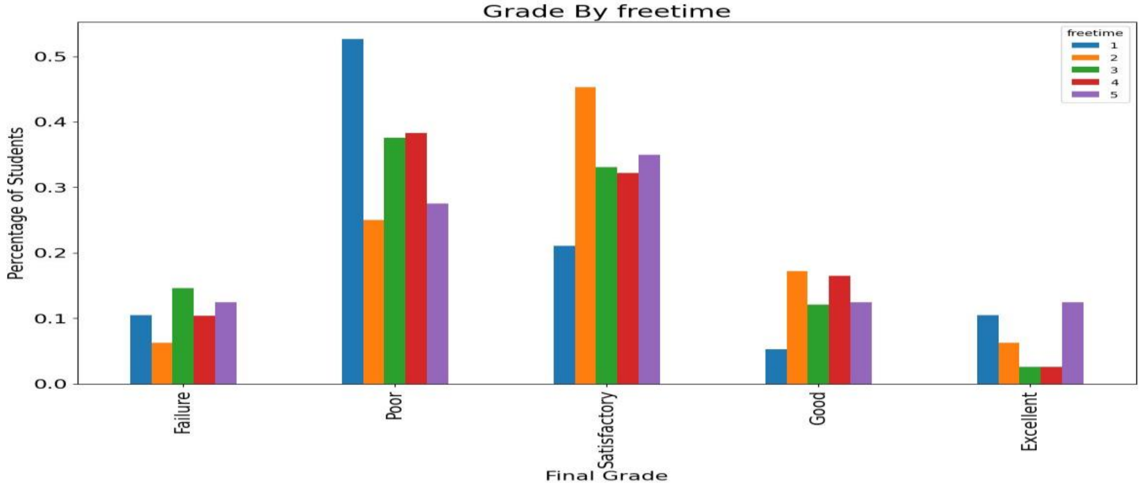


Graph representing health status to grade

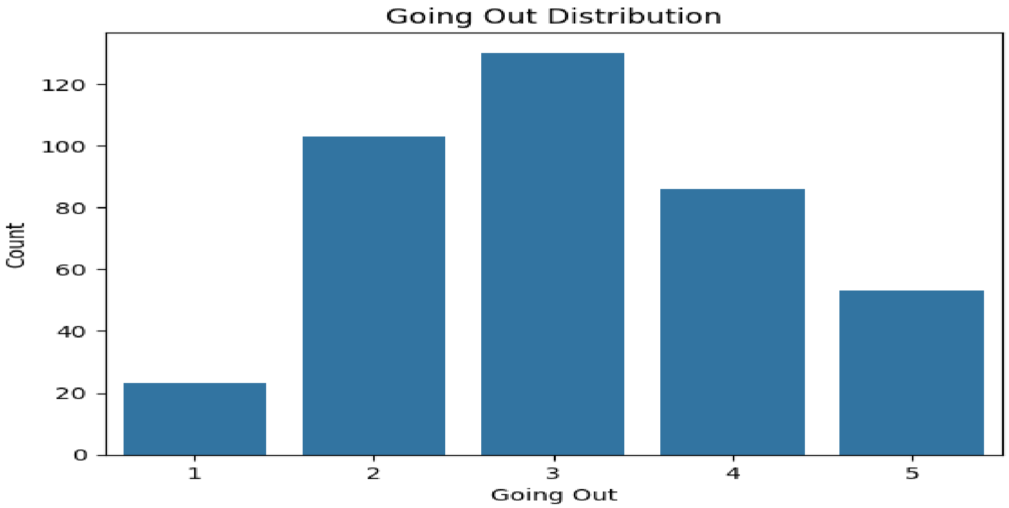
Relationship Status: Emotional and social well-being significantly influences student performance. The system analyzes data on romantic relationships, relationships with peers and family dynamics to determine their effect on academic achievement. For instance, a student with a robust support system from peers is likely to do better.



Manage Free Time: Time spent in free time and extracurricular activities significantly impacts academic performance. The system evaluates how pupils spend their time in free space whether productive or not and incorporates that into the total forecast. Students with structured free time, including study time and extracurricular activities or activities that foster intellectual growth, may have higher performance expectations.



Comprehensive Outcome Report: The study's final report gives a complete overview of the student's expected performance and outlines the impact of the variables above. The report is helpful not just for students but also for teachers and administrators, assisting them in recognizing the child's complete growth.



Interactive Features Students can revisit the page for evaluation and alter specific lifestyle elements (such as increasing their study time or adjusting their sleeping habits) to assess the ways these changes can increase their academic achievement. This interactive feature promotes active behaviour by students and allows them to control their learning journey.

4.0 CONCLUSION

The use of ensemble algorithms to predict student academic achievement is an effective approach to improving educational results. These algorithms outperform standard methods in terms of accuracy since they combine numerous models. This study highlights the potential of such systems in detecting at-risk individuals, allowing institutions to adopt preemptive interventions, and improving overall student achievement. It provides a road map for students to improve their academic performance, while it is a powerful tool for educators and administrators to assess student progress and identify those who may require more assistance.

Furthermore, the system's user-friendly design, extensive data input possibilities, and thorough evaluation procedure make it a versatile tool that can be applied to a variety of educational scenarios. It is capable of tackling the numerous issues faced by students from different backgrounds, assisting institutions in developing tailored methods to improve overall academic performance

REFERENCES

Ajibade, O., Connolly, T., & Adejo, W. (2020). Predicting student academic performance using multi-model heterogeneous ensemble approach. Journal of Applied Research in Higher Education, 10(1), 61-75. https://doi.org/10.1108/JARHE-10-2020-0023

Bocarnea, M. C. (Ed.). (2012). Online instruments, data collection, and electronic measurements. IGI Global. https://doi.org/10.4018/978-1-61350-504-5

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research, 297(3), 1178-1192. https://doi.org/10.1016/j.ejor.2021.06.029

Miranda, E. N., Barbosa, B. H. G., Silva, S. H. G., Monti, C. A. U., Tng, D. Y. P., & Gomide, L. R. (2022). Variable selection for estimating individual tree height using genetic algorithm and random forest. Forest Ecology and Management, 504, 119828. https://doi.org/10.1016/j.foreco.2021.119828

Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students' academic performance basedon courses' grades using deep neural networks. *IEEE Access*, *9*, 140731-140746.

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. Applied Sciences, 11(1), 237. https://doi.org/10.3390/app11010237

Pessach, D., Singer, G., Avrahami, D., Ben-Gal, H. C., Shmueli, E., & Ben-Gal, I. (2020). Employee recruitment: A prescriptive analytics approach via machine learning and mathematical programming. Decision Support Systems, 134, 113290.https://doi.org/10.1016/j.dss.2020.113290

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied sciences*, *10*(3), 1042.

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. Procedia Computer Science, 72, 414-422. https://doi.org/10.1016/j.procs.2015.12.157

Venkatesan, R. G., Karmegam, D., & Mappillairaju, B. (2024). Exploring statistical approaches for predicting student dropout in education: A systematic review and meta-analysis. *Journal of Computational Social Science*, *7*(1), 171-196.

Yang, G., Wang, Y., Zeng, Y., Gao, G. F., Liang, X., Zhou, M., ... & Murray, C. J. (2013). Rapid health transition in China, 1990–2010: findings from the Global Burden of Disease Study 2010. *The lancet*, *381*(9882), 1987-2015.