Predicting Student Performance Using Socio-Economic Features and Hybrid Machine Learning Models

Abdulrahman Mohammed Alawbathani
Department of Computer Science
GIU
Cairo, Egypt
abdullrahmanvi7@gmail.com

Waleed Essam Elsenawy
Department of Computer Science
GIU
Cairo, Egypt
waleed.elsenawy@gmail.com

***Abstract*—Predicting how students will perform academically is a complex task, especially when taking socio-economic factors into account. While these factors clearly influence educational outcomes, they're often overlooked in traditional assessment methods. In this study, we propose a hybrid machine learning approach that combines Random Forest and Gradient Boosting to make more accurate and meaningful predictions. Using the UCI Student Performance Dataset, which includes a wide range of socio-economic, academic, and behavioral data, we trained our model to classify students into three performance levels: Low, Medium, and High.**

The data was carefully preprocessed, including binary mapping, one-hot encoding, and target binning, before applying our hybrid model through a soft voting ensemble. The final model achieved 88.61% accuracy and a macro-averaged F1-score of 0.89. It performed reliably across all three classes, with strong results for identifying students at academic risk. Subgroup evaluations showed that the model handled socio-economic features effectively, though some performance gaps were observed between more and less advantaged student groups.

Overall, this research demonstrates that hybrid models can successfully incorporate socio-economic data to improve performance prediction and offer valuable insights for more inclusive, data-driven educational support.

***Index Terms*—Student performance prediction, socio-economic data, educational data mining, hybrid machine learning, fairness analysis, Random Forest, Gradient Boosting, ensemble methods.**

I. Introduction

Socio-economic factors are traits that reflect the social and economic status of an individual, such as income, education, and employment. These factors have a significant influence on various life outcomes, including academic performance. Understanding how these elements interact with a student's learning environment can provide valuable insights into their academic success or struggles. Machine learning is one of the subfields of AI that enables the automated recognition of patterns in data, allowing the computer to make predictions with minimal human intervention. In recent years, machine learning has started playing a bigger role in education, helping schools and educators make better use of the data they collect. By analyzing student data, these models can spot patterns that might not be obvious at first glance. This makes it easier to identify students who may be falling behind, modify learning to individual needs, and give teachers and school leaders better tools to support their students. In this study, our aim is to address this challenge by developing a hybrid machine learning model, combining random forest and gradient boosting, trained on the UCI Student Performance Dataset to predict academic outcomes based on socio-economic indicators.

The challenge which the research wants to address is the accuracy of student performance prediction based on socio-economic factors. While these factors are known to impact academic performance, traditional assessment methods often overlook them. This study seeks to build a machine learning model using these socio-economic indicators and obtain insightful predictions. Yet, challenges such as limited access to high-quality data, potential for algorithmic bias, and ensuring that the model operates reliably across diverse student groups must be addressed.

Understanding how socio-economic factors influence student performance is essential for creating more equitable educational systems. Research has consistently shown that students from higher socio-economic backgrounds tend to achieve better academic outcomes due to increased access to resources, parental involvement, and educational opportunities [1]. Accurately predicting performance based on these factors can help identify students who may be at risk of underachievement, allowing for targeted support and interventions. Additionally, tools like the Equivalized Household Income Indicator (EHII) offer innovative ways to estimate socio economic status even when direct income data is unavailable, making predictive models more scalable and inclusive [2]. By incorporating socio-economic indicators into performance prediction models, we can work toward narrowing the achievement gap and supporting students who are often overlooked by traditional evaluation methods.

II. Related Work

Several studies have explored the use of machine learning to predict student academic performance by incorporating socio-

economic and other related factors. A hybrid model based on ensemble learning was introduced to improve prediction accuracy using socio-economic features such as parental education, job, family income, and internet access, achieving strong results with an accuracy of 93.12% [3]. Another study applied multiple machine learning algorithms, including Artificial Neural Networks and Support Vector Machines, to evaluate high school performance across four performance levels, using academic, behavioral, and some socio-economic attributes collected from Iranian students [4]. A large-scale study involving over 5000 students used health, behavioral, and socio-economic data, including parental income and education, to classify student performance into three categories, with Logistic Regression and Na¨ıve Bayes delivering the highest accuracy scores [5]. A stacked ensemble model was also tested on university students, using a small but focused feature set that included income, parental education, employment status, and household size, and achieved an accuracy of over 85% [6]. Another study emphasized the predictive value of a wide variety of inputs, including socio-economic indicators like parental background and household assets, and found these factors to be among the most influential in predicting final grades [7]. Building directly on socio-economic data, a recent model used 18 socio-economic features, including guardian education, income, financial aid access, and regional background, and demonstrated that these variables significantly affect student performance, with Support Vector Machine reaching 90% accuracy [8]. A statistical study in Bangladesh applied ANCOVA and OLS regression on undergraduate data and found that variables like family income, parental education, study hours, and prior results significantly influenced CGPA [9]. A global-scale machine learning comparison using the PISA dataset tested algorithms across over 600,000 students and found that reading habits, digital access, and parental education were among the strongest predictors, with Gradient Boosted Trees achieving the highest accuracy of 74.17% [10]. Another study used the UCI dataset to predict student grades using academic and socio-economic data, with Random Forest reaching 90.6% accuracy in binary classification and showing strong results with variables like internet access and parental background [11]. A school-level study in Pennsylvania linked environmental and socio-economic factors such as crime rate and population to academic performance, with neural networks performing best at 60% accuracy [12]. A systematic review analyzed 56 studies and found that socio-economic, academic, and behavioral features are widely used in predictive modeling, with Decision Trees, SVMs, and ensemble methods 1 being the most common approaches [13]. Lastly, another hybrid model that combined mul- tiple classifiers achieved 98.7% accuracy on data from 1,227 students, with socio-economic and family attributes playing a key role in the model's success [14].

While these studies show that machine learning can be effective in predicting student performance, many of them share a few common limitations. Most focus mainly on improving overall accuracy, without looking closely at how their models perform for different types of students—especially those from less advantaged backgrounds. In many cases, socio-economic factors are included in the data but not treated as a central part of the prediction process. Also, although some studies use ensemble methods, very few combine different algorithms in a way that takes advantage of their individual strengths. These gaps highlight the need for models that not only aim for high accuracy, but also take socio-economic context seriously and work fairly across diverse student groups. Our work responds to this need by exploring a model that aims to do both.

III. Methodology

This study proposes a hybrid machine learning model to predict student academic performance using socio-economic, academic, and behavioral features. The goal is to classify students into three performance levels (Low, Medium, High) based on their final grades. The approach combines Random Forest and Gradient Boosting models in an ensemble structure to improve prediction accuracy and robustness.

A. Dataset

The dataset used in this study is the UCI Student Performance Dataset [15], which contains 395 student records from Portuguese secondary schools. Each record includes 33 attributes related to student demographics, family background, socio-economic status, school support, behavioral habits, and academic performance. The target variable is the final grade (`G3`), a score from 0 to 20. For multiclass classification, this variable is discretized into three categories: Low (0–9), Medium (10–13), and High (14–20), following structures seen in prior work.

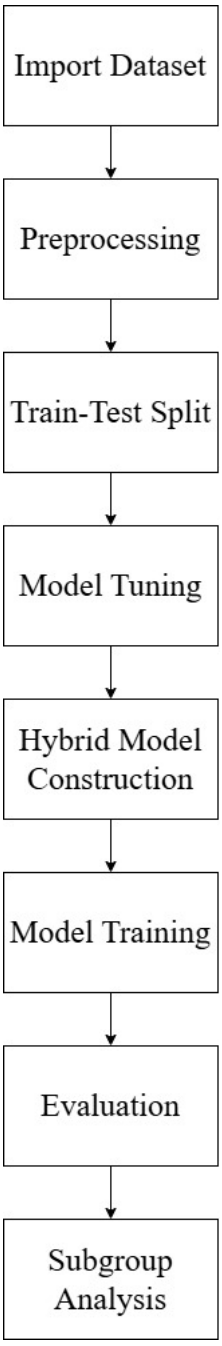
B. Data Preprocessing

Before training the model, the data is preprocessed through the following steps:

- **Target Binning:** The continuous `G3` scores are binned into three performance categories.
- **Encoding Categorical Features:** All nominal features (e.g., job types, school support) are converted into numerical representations using one-hot encoding, while binary data (like columns that contain yes/no values) were converted into numerical representations using .map()
- **Missing Values:** The dataset contains no missing values, so no imputation was necessary.

C. Model Architecture

We use a hybrid model combining Random Forest and Gradient Boosting classifiers, implemented via a soft voting ensemble. This decision is supported by consistent findings in the literature showing ensemble methods outperform individual classifiers in educational prediction tasks. Random Forest offers strong performance on structured data, handles high-dimensionality well, and provides feature importance insights, while Gradient Boosting captures complex patterns with iterative learning. Previous studies, such as those of Kumar et al. [3] and Vergaray et al. [6], demonstrate that hybrid approaches provide more reliable predictions than standalone models.



D. Evaluation Measures

To assess the model's performance in a multiclass setting, the following evaluation metrics are used:

- **Accuracy:** Measures the overall proportion of correct predictions.
- **Precision, Recall, and F1-Score:** Calculated for each class, and macro-averaged to handle any class imbalance.
- **Confusion Matrix:** Provides a visual summary of classification performance across the three classes.

IV. Experiment and Evaluation

A. Data Preprocessing

The original dataset from the UCI Student Performance Repository included 395 rows and 33 attributes, covering academic, demographic, behavioral, and socio-economic variables. The following preprocessing steps were performed:

- **Binary Mapping:** Features with binary values such as `yes/no` (e.g., `internet`, `schoolsup`, `higher`) and binary categories like `sex` (M/F), `school` (GP/MS), and `address` (U/R) were mapped to 1/0 using the `.map()` function.
- **One-Hot Encoding:** Categorical variables with more than two categories (e.g., `Mjob`, `Fjob`, `guardian`, `reason`) were encoded using `pd.get_dummies()` with `drop_first=True` to avoid multicollinearity.
- **Target Binning:** The original target variable `G3` (final grade from 0 to 20) was converted into a three-class categorical variable `performance` using `pd.cut()` with bins: Low (0–9), Medium (10–13), and High (14–20), and `include_lowest=True` to ensure values at the lower boundary were included.
- **Feature Inclusion:** The intermediate performance indicators `G1` and `G2` were retained as features due to their high predictive relevance. `G3` was excluded as it is the target.
- **Final Dataset:** The resulting dataset contained 41 numerical features and a single multiclass target variable.

B. Model Tuning and Hybrid Ensemble Construction

Two classifiers were individually trained and optimized using `GridSearchCV`:

- **Random Forest:** Tuned over the hyperparameters: `n_estimators` {100, 200}, `max_depth` {None, 10, 20}, `min_samples_split` {2, 5}.
- **Gradient Boosting:** Tuned over: `n_estimators` {100, 200}, `learning_rate` {0.05, 0.1}, `max_depth` {3, 5}.

Each model was optimized for macro-averaged F1-score using 3-fold cross-validation. The best estimators from each search were then merged using `VotingClassifier` with `voting='soft'` to form the final hybrid model. This setup allows the classifiers to contribute predictions based on class probabilities, which proved effective in balancing precision and recall.

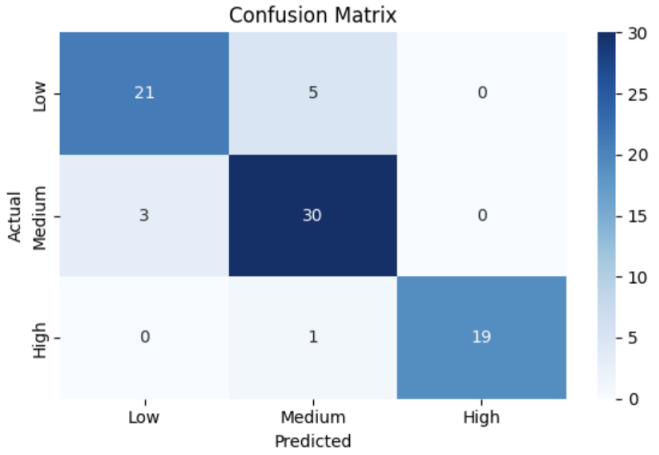
C. Model Performance

The final hybrid model achieved an accuracy of **88.61%** on the test set. The macro-averaged F1-score was **0.89**. Class-wise performance was as follows:

- **High:** Precision = 1.00, Recall = 0.95, F1 = 0.97
- **Medium:** Precision = 0.83, Recall = 0.91, F1 = 0.87
- **Low:** Precision = 0.88, Recall = 0.81, F1 = 0.84

The model showed strong capability in correctly identifying students at both performance extremes. The confusion matrix indicated minimal misclassification between Low and High

classes, with most errors occurring between adjacent categories.



D. Subgroup Analysis

To assess fairness and reliability, the model was evaluated on various socio-economic subgroups.

1) Gender:

- **Male:** Accuracy = 91.18%, Macro F1 = 0.92
- **Female:** Accuracy = 86.67%, Macro F1 = 0.87

The model performed consistently across genders but showed a 5% gap in macro F1-score, slightly favoring male students.

2) Internet Access:

- **With Internet:** Accuracy = 90.91%, Macro F1 = 0.91
- **Without Internet:** Accuracy = 76.92%, Macro F1 = 0.82

Students without home internet access showed a noticeable performance drop, especially in the "Low" and "Medium" categories, suggesting lower reliability for this group.

3) Family Educational Support:

- **With Support:** Accuracy = 90.0%, Macro F1 = 0.91
- **Without Support:** Accuracy = 86.21%, Macro F1 = 0.86

The model performed better for students who received family support, although the gap was moderate.

4) Mother's education (Medu):

- **High Medu (greater than or equal to 3):** Accuracy = 92.86%, Macro F1 = 0.92
- **Low Medu (less than or equal to 2):** Accuracy = 83.78%, Macro F1 = 0.87

A 9% difference in macro F1-score was observed, favoring students whose mothers had higher education levels.

5) Father's education (Fedu):

- **High Fedu (greater than or equal to 3):** Accuracy = 91.89%, Macro F1 = 0.91
- **Low Fedu (less than or equal to 2):** Accuracy = 85.71%, Macro F1 = 0.89

Similar to the maternal education results, the model performed more reliably for students with more educated fathers.

E. Interpretation and Implications

The model consistently performed well across most groups, but subgroup evaluations revealed some gaps in predictive accuracy and F1-score based on access to educational resources. These findings suggest that, while the model is generally strong and generalizable, more work is needed to further improve fairness.

Declaration of AI Use

AI tools were used during the research project to support and accelerate various tasks. These tasks included writing assistance, code optimization, formatting LaTeX content, and generating explanatory summaries for model interpretation and evaluation. All analytical decisions, model training, data preprocessing, and result interpretations were conducted and verified by the authors. The AI was used strictly as a supporting tool and not as a replacement for human reasoning or academic responsibility.

V. Conclusions and Future Work

The goal of this study was to build a model that not only predicts student performance accurately, but also makes use of socio-economic factors to do so in a meaningful way. The results show that this goal was met. By combining Random Forest and Gradient Boosting classifiers in a hybrid ensemble, the final model achieved an accuracy of 88.61% and a macro-averaged F1-score of 0.89. This is a strong outcome for a multi-class classification problem in education, where performance categories can often overlap.

A major reason behind this success was the model's ability to effectively use socio-economic indicators, such as parental education, internet access, and family support, alongside academic and behavioral data. These features helped the model not only improve overall accuracy, but also identify students at risk of underperforming. Including intermediate grade features (G1 and G2) further improved the model's ability to distinguish between performance categories.

More importantly, the subgroup evaluation helped show how the model performs across different types of students. Although it handled most groups well, it was slightly more accurate and consistent for students who had more resources or support, such as access to internet or parents with higher education. This finding is both expected and important, and shows that even though the model is effective, fairness needs to be kept in mind.

Future Work: In the future, it would be valuable to focus on closing the performance gap across socio-economic groups. Techniques like bias mitigation, group-specific thresholds, or even separate models per subgroup could help improve fairness. The model could also be tested on new datasets or in real-time school environments to check if it holds up under different conditions. Finally, adding explainability methods like SHAP or LIME would help teachers and administrators better understand the model's decisions and build trust in using such tools to support students. Although the model performed well, it is worth noting that the dataset was relatively small

compared to the number of features used. This raises the possibility that the model may be somewhat tailored to this specific data. To ensure that the approach holds up in broader contexts, future work should test it on a larger dataset.

REFERENCES

[1] J. Munir, M. Faiza, B. Jamal, S. Daud, and K. Iqbal, "The Impact of Socio-economic Status on Academic Achievement," *Journal of Social Sciences Review*, vol. 3, no. 2, pp. 695–705, 2023. [Online]. Available: https://doi.org/10.54183/jssr.v3i2.308

[2] L. Gagliardi et al., "A new approach to estimate household income in European birth cohort studies: The Equivalized Household Income Indicator (EHII)," *International Journal of Environmental Research and Public Health*, 2020.

[3] V. Kumar, S. Goel, and A. Kaur, "A hybrid machine learning model for student performance prediction using socio-economic features," *Heliyon*, vol. 8, no. 3, p. e11165, 2022.

[4] M. Zafari, A. Sadeghi-Niaraki, S. M. Choi, and A. Esmaeily, "A practical model for the evaluation of high school student performance based on machine learning," *Applied Sciences*, vol. 11, no. 23, p. 11534, 2021.

[5] R. Qasrawi, S. VicunaPolo, D. Abu Al-Halawa, S. Hallaq, and Z. Abdeen, "Predicting school children academic performance using machine learning techniques," *Advances in Science, Technology and Engineering Systems Journal*, vol. 6, no. 5, pp. 448–456, 2021.

[6] M. Vergaray, R. Palomino, H. E. Pérez, and M. Reyes, "Predicting academic performance using a multiclassification model: Case study," *International Journal of Emerging Technology and Advanced Engineering*, vol. 12, no. 5, pp. 23–30, 2022.

[7] M. Musso, E. Kyndt, E. Cascallar, and A. Sarra, "Identifying reliable predictors of educational outcomes through machine-learning predictive modeling," *Frontiers in Education*, vol. 5, p. 104, 2020.

[8] A. Muhammad, M. Imran, A. A. Alqarni, and A. S. Alzahrani, "The role of socioeconomic factors in improving the performance of students based on intelligent computational approaches," *Electronics*, vol. 12, no. 19, p. 1982, 2023.

[9] M. Rahman, "Impact of socio-economic factors on undergraduate students' academic performance in Bangladesh," *Asian Journal of Social Sciences and Legal Studies*, 2021.

[10] T. T. Ting, L. S. Hock, and O. M. Ikumapayi, "Educational big data mining: Comparison of multiple machine learning algorithms in predictive modelling of student academic performance," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 6, 2024.

[11] C. Dervenis, V. Kyriatzis, S. Stoufis, and P. Fitsilis, "Predicting students performance using machine learning algorithms," in *Proc. 6th Int. Conf. Algorithms, Computing and Systems (ICACS)*, 2022.

[12] S. Chen and Y. Ding, "A machine learning approach to predicting academic performance in Pennsylvania's schools," *Social Sciences*, vol. 12, p. 118, 2023.

[13] P. Balaji, S. Alelyani, A. Qahmash, and M. Mohana, "Contributions of machine learning models towards student academic performance prediction: A systematic review," *Applied Sciences*, vol. 11, p. 10007, 2021.

[14] A. Siddique et al., "Predicting academic performance using an efficient model based on fusion of classifiers," *Applied Sciences*, vol. 11, p. 11845, 2021.

[15] P. Cortez, "Student Performance," UCI Machine Learning Repository, 2008. [Online]. Available: https://doi.org/10.24432/C5TG7T