Predictive Analytics in Education

Suharsh Sandhu
Faculty of Engineering
University of Western Ontario
London, Canada
ssand3@uwo.ca

Reuben Anchan
Faculty of Engineering
University of Western Ontario
London, Canada
ranchan@uwo.ca

Luxshan Jeyaranjan
Faculty of Engineering
University of Western Ontario
London, Canada
ljeyaran@uwo.ca

Sehaj Naangal
Faculty of Engineering
University of Western Ontario
London, Canada
snaangal@uwo.ca

***Abstract*—Predicting academic achievement is pivotal for educational institutions aiming to enhance learning outcomes and tailor individualized support. This research examines a diverse set of predictive variables encompassing demographics, historical academic performance, and socioeconomic indicators to evaluate the proficiency of machine learning algorithms in forecasting academic success. In particular, the study focuses on the comparative efficacy of Random Forest and Gradient Boosting models. Rigorous data preprocessing and model training have been conducted on a vast dataset of student profiles to identify the most salient features influencing educational attainment. The application of these advanced analytical techniques has not only expanded the existing corpus of educational data mining but also highlighted the potential of machine learning in educational settings. Our findings contribute to a nuanced understanding of the interplay between student characteristics and academic results, offering a solid foundation for the creation of strategic interventions. Such interventions are essential for bolstering the performance of students at risk, thereby nurturing an equitable educational landscape where each learner's potential can be fully realized.**

I. INTRODUCTION

Predicting the academic achievement of students is a critical endeavor for educational institutions aiming to enhance learning outcomes and provide tailored support. "Many students are being left behind by an educational system that some people believe is in crisis. Improving educational outcomes will require efforts on many fronts" [1]. This research focuses on analyzing various predictive features—including demographics, academic performance metrics, and socioeconomic factors—to evaluate the effectiveness of machine learning algorithms in forecasting academic outcomes. "Student success plays a vital role in educational institutions, as it is often used as a metric for the institution's performance. Early detection of students at risk, along with preventive measures, can drastically improve their success" [2]. By examining these factors, our goal is to determine the most accurate methods for predicting student success.

"Machine Learning (ML), Collaborative Filtering (CF), Recommender Systems (RS) and Artificial Neural Networks (ANN) are the main computational techniques that process this information to predict students' performance, their grades or the risk of dropping out of school" [3]. In this study, we explore the performance of advanced machine learning techniques, specifically Random Forest Regression and Gradient Boosting Regression, to identify which model most effectively predicts academic success. Our analysis involves a comprehensive preprocessing of the dataset, which includes student demographics, academic records, and socioeconomic backgrounds from various schools. This preprocessing step is crucial for identifying relevant features that significantly impact academic performance. In summary, this study endeavors to harness the power of machine learning to refine our predictive capabilities in education, ultimately paving the way for more personalized and effective interventions to foster student success.

We aim to evaluate these models based on their ability to accurately predict student outcomes and uncover the underlying patterns that contribute to academic success or challenges. By doing so, we intend to provide insights that can guide educational institutions in implementing targeted interventions and support mechanisms.

Ultimately, this research seeks to leverage the predictive power of machine learning to enhance our understanding of the factors that influence academic achievement. Through detailed analysis and model comparison, we aspire to contribute to the development of more personalized educational strategies that promote student success, thereby addressing the broader challenge of improving educational outcomes in our schools.

II. RELATED WORK

Previous research in the field of educational data analysis has explored various factors influencing student performance, including socioeconomic status and parental educational background. One notable study by Shikah and Hala investigated the predictive power of socioeconomic factors on academic achievement using a Random Forest implementation. Their research focused on analyzing the impact of variables such as low-income household status and parental education level on student performance. By leveraging Random Forest methodology, they demonstrated the potential of machine learning algorithms in predicting student outcomes based on socioeconomic indicators [4]. However, while their study provided valuable insights into the predictive capabilities of Random Forest models, there remains a research gap in understanding the nuanced interactions between socioeconomic factors and academic success.

Yet, this area of research is far from fully explored. The interplay between socioeconomic indicators and student outcomes, particularly how these factors collectively impact academic performance, remains underexamined. For example, studies similar to those by **(Insert later)** on the role of external environmental factors [5], and the work of **(Insert later)** on the predictive accuracy of various machine learning techniques in educational contexts [6], highlight the breadth of research in this domain.

Building on these foundations, our study aims to delve deeper into the predictive capabilities of Random Forest and Gradient Boosting models, focusing on a broader spectrum of socioeconomic factors. By doing so, we contribute to the burgeoning literature on the application of machine learning in education, providing valuable insights that could inform targeted interventions aimed at enhancing the academic trajectories of students from diverse backgrounds.

III. METHODS

The selection of features for this study was guided by the hypothesis that socioeconomic backgrounds, as well as school enrollment numbers, have a substantial impact on student achievement. Specifically, we chose to examine the number of students enrolled in a school (`Enrolment`), the percentage of children from low-income households (`Low-Income Households`), and the percentage of students whose parents do not hold post-secondary degrees (`Parents with No Degrees`). These factors were posited to correlate with the academic achievement across elementary to secondary school levels, although we only used elementary schools for the data as other schools had substantially less data. This data was then quantified using average performance metrics in core subject areas such as reading, writing, and mathematics.

A. Dataset Analysis

The exploration of the dataset aimed to identify key factors influencing student performance across schools in Ontario, Canada. We utilized data sourced from the Board School Identification Database (BSID) and the Ontario School Information System (OnSIS) named “School Information and Student Demographics” [7]. The dataset contains 4916 different schools located in Ontario, Canada. The variable dictionary is shown in Table 1.

Table 1

Data Dictionary

Variable	**Variable Explanation**
Enrolment	Number of students enrolled in a school
Low-Income Households	Percentage of School-Aged Children who live in Low-Income Households
Parents with No Degrees	Percentage of Students Whose Parents Have No Degree, Diploma or Certificate
Grade 3 Students - Reading	Percentage of Grade 3 Students Achieving the Provincial Standard in Reading
Grade 3 Students - Writing	Percentage of Grade 3 Students Achieving the Provincial Standard in Writing
Grade 3 Students - Mathematics	Percentage of Grade 3 Students Achieving the Provincial Standard in Mathematics

B. Data Preprocessing

Data preprocessing was a critical initial step to ensure the quality and utility of the data for analysis. The process began with data cleaning, which involved removing or imputing missing values and handling non-numeric entries such as 'SP' and 'N/R'. Percentage values were standardized by converting them into a consistent numeric format. Furthermore, normalization of enrollment figures was conducted to facilitate comparability across schools, regardless of their size. Feature engineering played a crucial role, with the creation of composite metrics like 'Mean Achievement', calculated as the average of scores in reading, writing, and mathematics. These deliberate and careful preprocessing actions ensured the dataset was primed for the subsequent application of machine learning algorithms, allowing for an accurate assessment of how socioeconomic factors impact academic success.

C. Feature Selection

Feature selection is a fundamental process in machine learning that involves identifying the most significant predictors for a given model. For this project, the feature selection was guided by both theoretical understanding and empirical analysis. The choice of features was driven by the objective to examine the impact of socioeconomic factors on student performance.

The features selected for this study included the percentage of school-aged children living in low-income households and the percentage of students with parents who have no degree, diploma, or certificate. These variables were chosen due to their strong theoretical foundation in the literature as significant predictors of academic achievement. To capture the school's environment, the enrollment number was also included, offering insights into the school's size and its potential effect on student outcomes.

The final set of features was determined using a combination of statistical techniques and model-based importance evaluations. Correlation analysis helped in understanding the linear relationship between features and the target variable, while the models’ intrinsic feature importance methods, such as those provided by RandomForestRegressor and GradientBoostingRegressor, gave a non-linear perspective on feature relevance.

By employing this hybrid approach to feature selection, the study aimed to balance the inclusion of theoretically relevant predictors with those empirically shown to have the most substantial impact on the prediction of academic achievement. This method ensured a robust and interpretable model that could provide valuable insights into the factors influencing student success.

D. Model Selection

The Random Forest model was selected due to its proficiency in handling complex datasets that include a mix of categorical and numerical variables. “This method is particularly adept at predictive tasks due to its remarkable accuracy” [8]. This ensemble learning method operates by constructing a multitude of decision trees at training time and outputting the average prediction of the individual trees for regression tasks. Its inherent design to build multiple trees and make decisions based on the majority voting or averaging of their predictions contributes to its robustness against overfitting, a common pitfall in machine learning models. By harnessing randomness in both feature selection and data points, it ensures diversity in the constructed trees, which results in a more generalized model. The fundamental mechanism of Random Forest involves each tree in the ensemble making a prediction, with the final output being determined by averaging these individual predictions, reflecting the consensus among a "forest" of decision-making models.

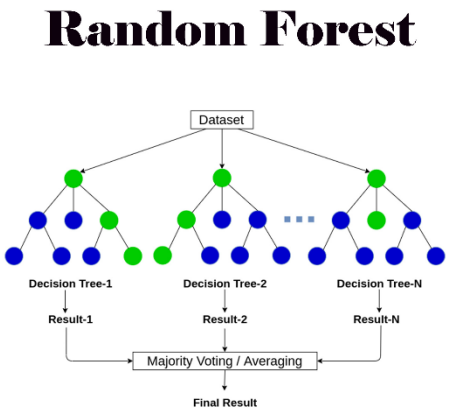

Fig. 1. Random Forest Model

Gradient Boosting was another choice for this study, well-known for its power and flexibility as a predictive modeling technique. This approach builds an additive model in a forward stage-wise fashion, where subsequent models are trained to predict the errors of the prior models combined. Each new model, therefore, improves on the shortcomings of the combined ensemble of existing models, effectively reducing bias and variance [9]. It employs gradient descent algorithm to minimize errors in a model by adjusting feature weights, making it particularly adept at handling various types of data and distributions. The model's strength lies in its capability to capture complex non-linear patterns by focusing training on areas where previous models have made errors, ensuring that the ensemble of weak predictive models converge to a strong learner.

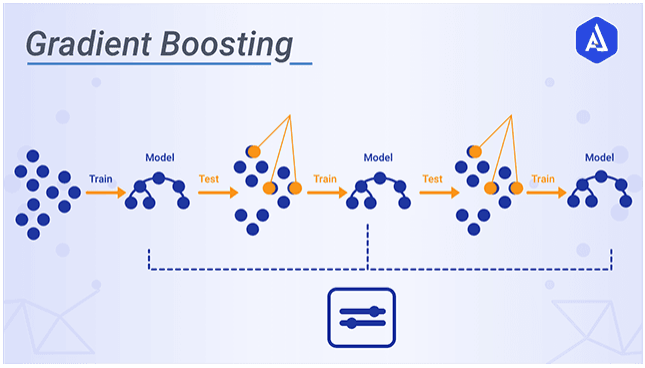

Fig. 2. Gradient Boosting Model

E. Model Training

The model training phase of our study was designed to optimize the predictive performance while ensuring that the model generalized well to new, unseen data. To this end, we employed an 80/20 split of the dataset, with 80% allocated to training the model and the remaining 20% reserved for testing. This approach provided a substantial amount of data for the model to learn from, encompassing a wide range of scenarios within the scope of the features and target variables. Simultaneously, it ensured that a representative sample was set aside to rigorously evaluate the model’s predictions. The Random Forest and Gradient Boosting models were iteratively trained on the training set, allowing them to learn complex relationships between the socioeconomic factors, enrollment data, and student achievement. This training process involved constructing numerous decision trees in the case of the Random Forest and sequentially improved trees for Gradient Boosting. Each model's parameters were fine-tuned during this phase to balance bias and variance, leading to a robust model capable of capturing the intricacies of the data while avoiding overfitting. The test set, then, served as a proving ground to assess the model's accuracy and reliability in predicting student achievement, which is critical for its application in real-world educational settings.

F. Model Evaluation

When evaluating the predictive models in this study, we utilized Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). MSE measures the average squared difference between estimated values and the actual value, providing insight into the variance of the model's errors [10]. It is especially useful for emphasizing the impact of larger errors but can be sensitive to outliers. MAE, by contrast, offers an average of the absolute errors, reflecting the typical magnitude of errors without undue influence from outliers, which can be pertinent when predicting student performance where factors can vary widely.

R^2 complements these metrics by indicating the proportion of dependent variable variance that the model explains. It is a scale-free score that allows comparison across models and contexts. In essence, R^2 assesses the model’s explanatory power, while MSE and MAE provide different perspectives on prediction accuracy and error size. Collectively, these metrics offer a robust framework for evaluating model performance, crucial for subsequent analysis and decision-making processes in educational assessments.

G. *Hyperparameter Tuning*

The final stage in refining our models was hyperparameter tuning, a crucial step to enhance model performance by searching for the optimal combination of parameters that govern the model's learning process. For the Random Forest, this included tuning the number of decision trees, the maximum depth of the trees, and the minimum number of samples required to split a node. The Gradient Boosting model's tuning involved adjusting the learning rate, the number of boosting stages, and the tree-specific parameters similar to those of the Random Forest. To conduct this process systematically, we utilized grid search and cross-validation techniques, ensuring a comprehensive exploration of the hyperparameter space.

Grid search cross-validation was instrumental in this phase, where various sets of hyperparameters were tested exhaustively against a cross-validation framework to determine which combinations produced the best results based on predefined scoring metrics such as R^2 and MSE. The cross-validation method not only helped in mitigating the overfitting problem by using different subsets of the dataset for training and validation but also provided a more generalized performance estimate. The optimal set of hyperparameters was then selected based on their ability to maximize the performance of the model on the validation set. The incorporation of this rigorous hyperparameter tuning process contributed to the development of a robust predictive model, tailored to deliver accurate predictions of student academic achievement.

IV. RESULTS

A. *Feature Importance*

Feature importance is a crucial aspect of the analysis as it gives us an insight into the relative impact of each predictor variable on the model's predictions. The graphs provided illustrate the importance of different features as evaluated by the RandomForest and Gradient Boosting regression models.

In the Random Forest model, ‘Enrolment’ appears to be the most significant feature, followed by the ‘Percentage of School-Aged Children Who Live in Low-Income Households’ and ‘Percentage of Students Whose Parents Have No Degree, Diploma or Certificate’. This indicates that the number of students enrolled has a substantial influence on mean achievement scores, potentially reflecting the resources available or the community's educational engagement.

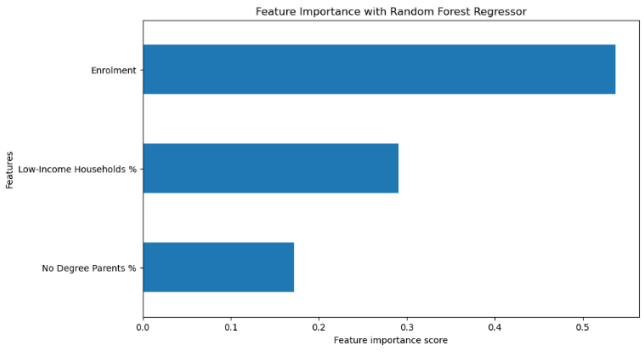

Fig. 3. Feature Importance with Random Forest Regressor

Conversely, the Gradient Boosting model prioritizes ‘Percentage of School-Aged Children Who Live in Low-Income Households’ as the most influential feature, followed by ‘Percentage of Students Whose Parents Have No Degree, Diploma or Certificate’, with ‘Enrolment’ being the least impactful. This variation from the Random Forest model highlights the importance of socioeconomic status and parental education in influencing academic achievement, suggesting that these factors may play a more significant role than school size in determining student outcomes.

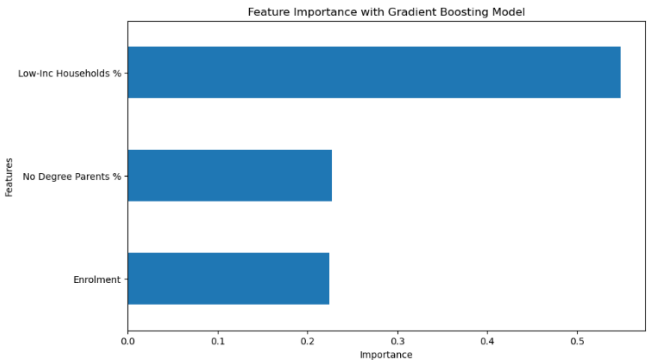

Fig. 4. Feature Importance with Gradient Boosting Model

The discrepancies between the two models highlight different interpretations of the data and highlight the need for careful consideration of model selection based on the context and the goal of the analysis. These insights are vital in guiding educational policy and targeted interventions for student support services.

B. *Mean Achievement Distribution*

The distribution of mean achievement scores is pivotal in understanding the overall performance trends of students. As illustrated in the histogram, the scores are approximately normally distributed, with the bulk of the data congregating around the 0.6 to 0.8 range. This central clustering suggests that a significant proportion of students are performing near the provincial standard on average, indicating a moderate level of achievement across the schools in Ontario. However, the presence of tails on both ends of the distribution highlights the existence of outliers, with some schools falling significantly below or above the average. Such a distribution is valuable for identifying not only the general trends in academic performance but also the extremities that may require additional investigation or resource allocation. Understanding this distribution lays the groundwork for further analysis, allowing for targeted interventions to support schools and students who may be underperforming, as well as those who are excelling.

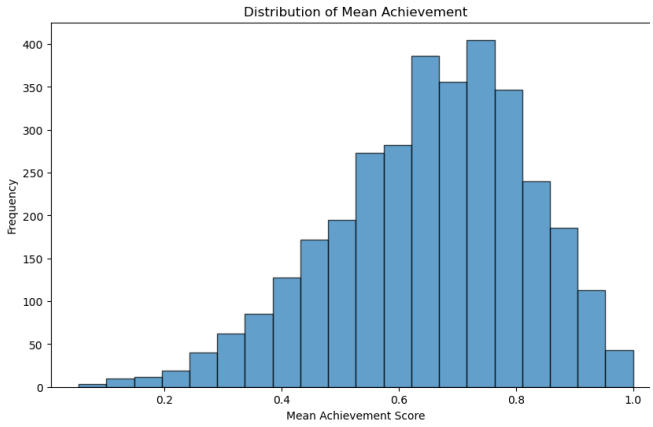

Fig. 5. Distribution of Mean Achievement

C. Correlation Matrix Analysis

The correlation matrix provides a visual and statistical summary of the potential relationships between different variables [11]. In this study, the heatmap reveals interesting dynamics between socioeconomic factors, school enrollment, and mean achievement. A notable negative correlation is observed between the percentage of school-aged children living in low-income households and mean achievement, suggesting that as the percentage increases, mean achievement tends to decrease. This is mirrored by the correlation with the percentage of students whose parents have no degree, diploma, or certificate, indicating that both these socioeconomic factors have a substantial impact on academic performance.

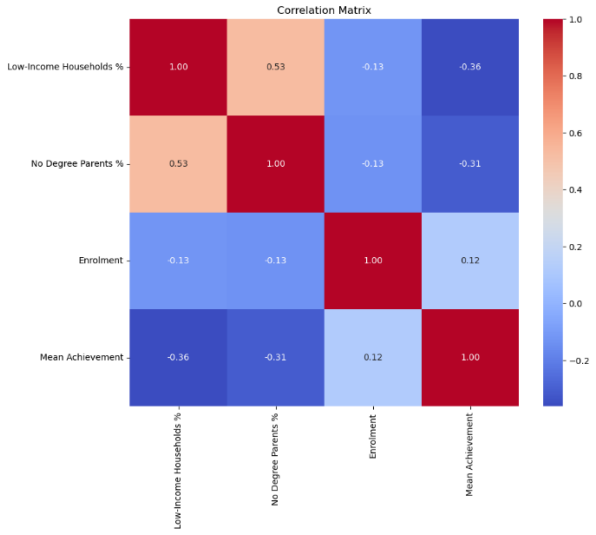

Fig. 6. Correlation Matrix

Conversely, the correlation between enrollment and mean achievement is positive, albeit weak, indicating that larger school sizes might have a slightly positive or neutral effect on mean achievement. The correlations between enrollment and socioeconomic indicators are also weak, suggesting that the size of the school does not significantly influence these factors. These insights are critical in informing policy decisions, as they underline the importance of socioeconomic context on educational success and point to potential areas for targeted interventions. The correlation matrix, thus, serves as a helpful tool in the analysis, guiding further exploratory and predictive modelling.

D. Evaluation Scores

The evaluation scores for the models provide insight into their performance and predictive accuracy. For the Random Forest model, the R^2 score is notably low at -0.1406, indicating that the model fails to capture the variance in the data and could be performing worse than the mean predictor. The Mean Absolute Error (MAE) stands at 0.1483, and the Mean Squared Error (MSE) at 0.0329, which are metrics indicating on average how much the predictions deviate from the actual values.

```
R^2 Score: -0.1406
Mean Absolute Error (MAE): 0.1403
Mean Squared Error (MSE): 0.0329
```

Fig. 7. Evaluation Scores for Random Forest Regressor

In contrast, the Gradient Boosting model exhibits a high R^2 score of 0.8919, suggesting that it has a strong predictive capacity, capturing a significant portion of the variance in the data. The MAE is slightly lower at 0.1258, along with an MSE of 0.0262, both suggesting more precise predictions than the Random Forest model.

```
R^2 Score: 0.0919
Mean Absolute Error (MAE): 0.1258
Mean Squared Error (MSE): 0.0262
```

Fig. 8. Evaluation Scores for Gradient Boosting Model

These evaluation scores are pivotal in model selection, as they directly reflect the ability of the models to predict unseen data accurately. A higher R^2 score and lower error metrics, like MAE and MSE, are generally desirable, indicating a model with better fit and prediction capabilities. The marked difference between the two models in these scores could stem from their inherent algorithmic approaches, with Gradient Boosting potentially providing a better fit for this particular dataset due to its sequential correction of errors.

E. Partial Dependence Plots

The Partial Dependence Plots (PDPs) for the Random Forest model provide a visual representation of the relationship between the target variable, mean achievement, and the selected features. For the percentage of school-aged children in low-income households and the percentage of students whose parents lack higher education, the PDPs display a relatively flat line, suggesting a consistent average prediction across different percentages. This could imply that within the scope of this model, changes in these socioeconomic factors do not have a significant impact on the predicted mean achievement. The plot for enrolment shows some variability, which may indicate that student numbers in schools have a more complex and variable relationship with mean achievement.

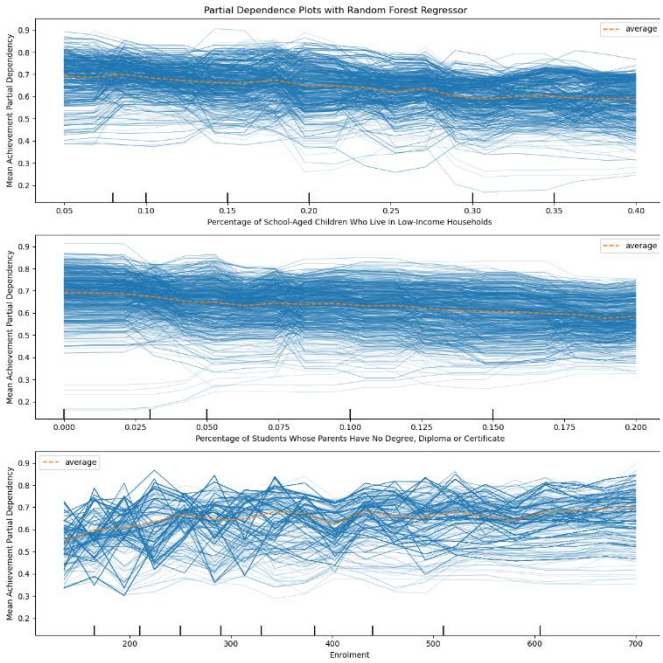

Fig. 9. Partial Dependence Plot for Random Forest Regressor

The PDPs for the Gradient Boosting model, in contrast, show a different pattern. The plots for both socioeconomic variables display a more distinct trend, indicating a potential impact of these factors on the mean achievement. This might suggest that the Gradient Boosting model has identified a more nuanced relationship, possibly capturing more complex interactions between these features and the target variable. The plot for enrolment again shows variability, consistent with the Random Forest model, but with a different pattern that might reveal how this feature's relationship to achievement is captured differently by Gradient Boosting.

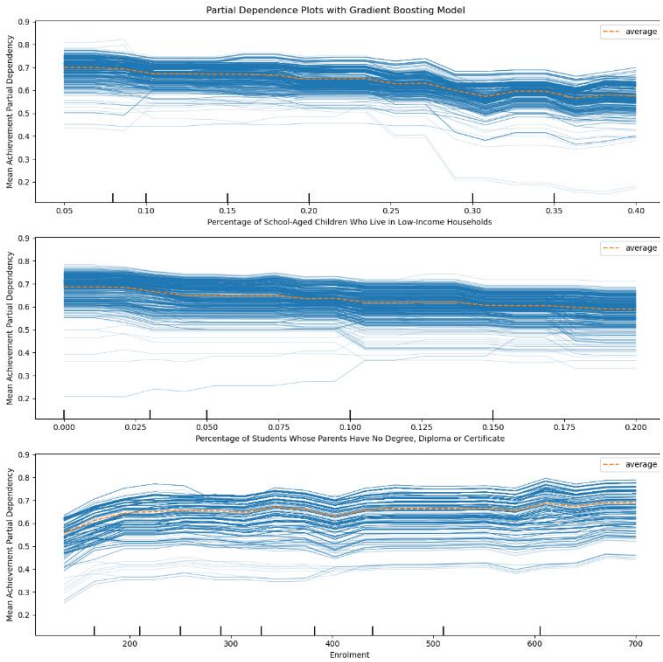

Fig. 10. Partial Dependence Plot for Gradient Boosting Model

Comparing the results of both PDPs, it's evident that each model processes the features differently. Random Forest appears to treat the socioeconomic factors as having a uniform effect across their range, whereas Gradient Boosting identifies more specific areas where these features influence mean achievement. The enrolment feature shows variability in both models, indicating its impact on the mean achievement is not straightforward and likely depends on other contextual factors. This comparison underscores the importance of model selection in predictive analytics, as different models can lead to different interpretations and insights, which are crucial for making informed decisions in educational planning and policy-making.

F. Model Evaluation Plots

For the Random Forest model focusing on school enrollment, the R^2 score suggests a moderate fit, improving as more data is used in training. The MSE and MAE scores decrease with additional data, indicating that the model is benefiting from more examples and is making fewer errors in prediction.

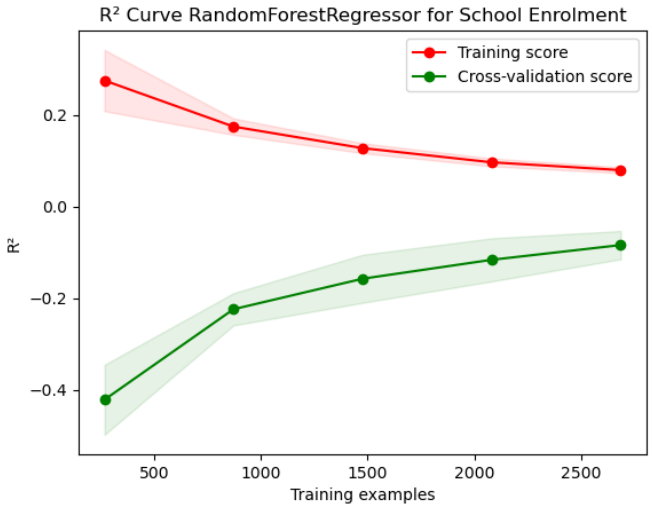

Fig. 11. R^2 Curve Random Forest Regressor for School Enrolment

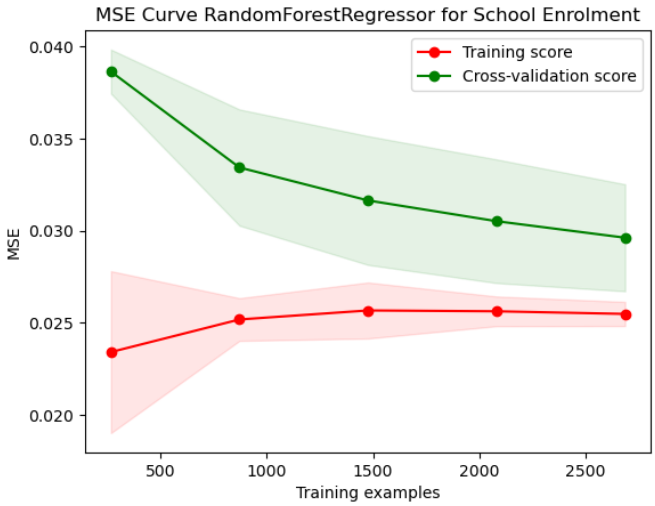

Fig. 12. MSE Curve Random Forest Regressor for School Enrolment

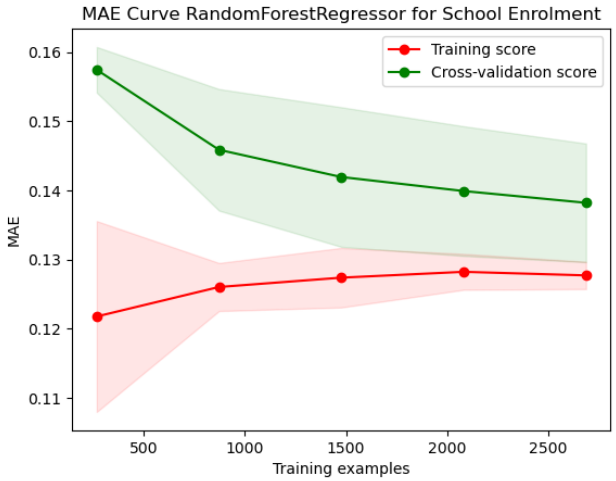

Fig. 13. MSE Curve Random Forest Regressor for School Enrolment

These metrics imply that the model is quite effective in using school enrollment data to predict academic success, with a stable performance across different data segments. The consistency across training and validation scores, especially with MAE, assures us that the predictions are reliable and not prone to overfitting, reinforcing the significance of enrollment numbers in understanding student achievement.

For the socioeconomic challenge features using the Random Forest model, the learning curves present a consistent trend across the R^2, MSE, and MAE metrics. The R^2 curve indicates an improvement in the model's ability to explain variance with an increase in training examples, although the cross-validation score remains relatively stable, suggesting the model's generalization is not substantially changing with more data. The MSE and MAE curves demonstrate a decrease in error rates as the number of training examples grows, which is an expected outcome, as more data typically enables the model to make more accurate predictions.

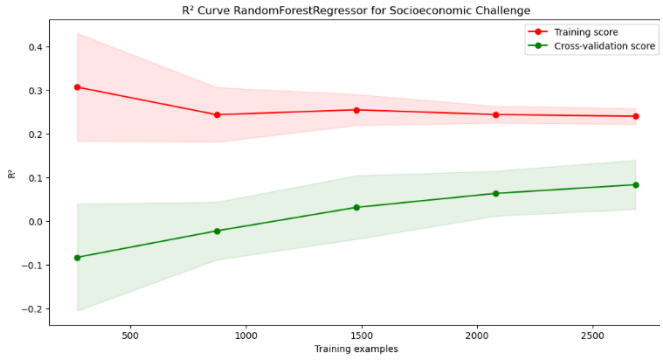

Fig. 14. R^2 Curve Random Forest Regressor for Socioeconomic Challenge

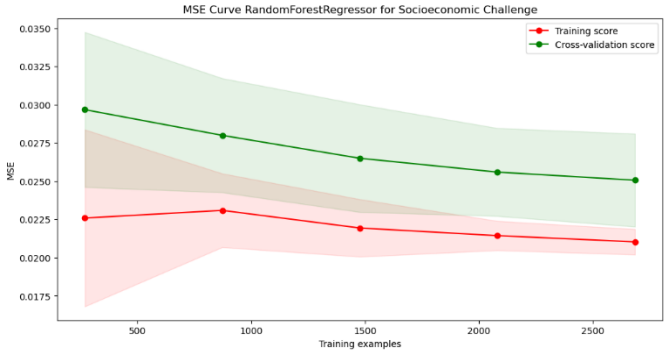

Fig. 15. MSE Curve Random Forest Regressor for Socioeconomic Challenge

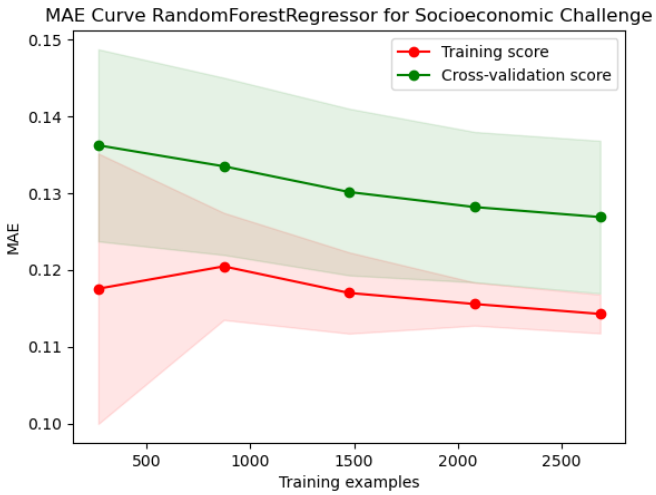

Fig. 16. MAE Curve Random Forest Regressor for Socioeconomic Challenge

When contextualizing these results, it is important to consider the implications for deploying this model in an educational setting. The learning curves suggest that the Random Forest model can provide reasonably stable and reliable predictions for the socioeconomic challenge features as the data size increases. However, the leveling off of cross-validation scores indicates a limit to the improvements that can be achieved simply by adding more training data. These insights can help educators and policymakers understand that while data volume is important, the quality of features and the intrinsic complexity of the dataset also play critical roles in predictive performance.

Regarding the GradientBoostingRegressor model applied to the school enrollment data, the learning curve for R^2 displays a consistent trend in training and cross-validation scores, indicating a reasonable match between the model's performance on the training and unseen data. Although the R^2 scores are not high, the model does not exhibit a high variance problem, as shown by the narrowing confidence intervals with more training examples. The MSE and MAE curves reveal a gradual decrease in error with increasing training examples, which denotes improving model performance with more data. The cross-validation scores for both error metrics plateau as the number of training examples increases, suggesting that additional data beyond a certain point does not lead to significant performance gains.

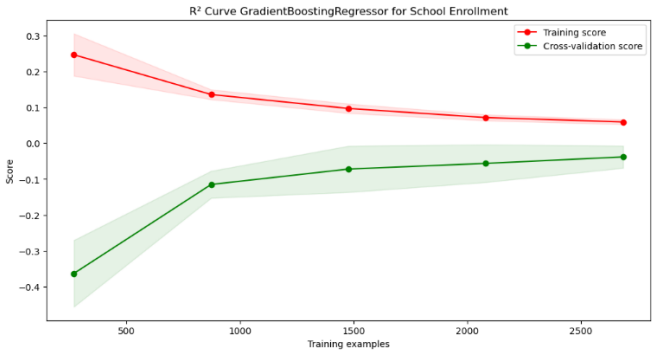

Fig. 17. R^2 Curve Gradient Boosting Regressor for School Enrolment

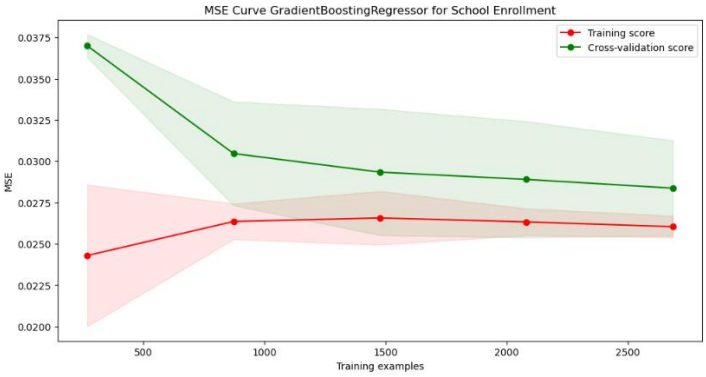

Fig. 18. MSE Curve Gradient Boosting Regressor for School Enrolment

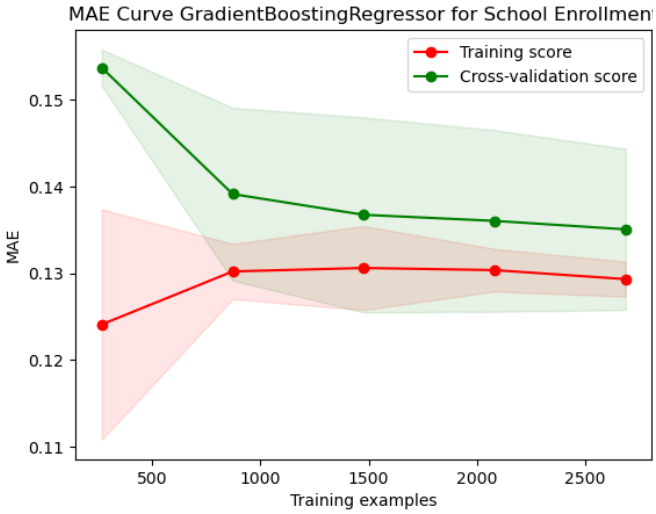

Fig. 19. MAE Curve Gradient Boosting Regressor for School Enrolment

In the overall context of predicting student success based on school enrollment, the GradientBoostingRegressor's learning curves point to a stable model with low variance but also low predictive power as evidenced by the R^2 scores. This suggests that while the model is learning the training data without overfitting, the feature of school enrollment alone does not have a strong linear relationship with the mean achievement of students. Further analysis with additional or alternative features, or the consideration of more complex models, might be necessary to improve the predictive capabilities for the target variable of mean achievement.

For the evaluation of the GradientBoostingRegressor model on the socioeconomic challenge, the learning curves reveal notable insights. The R^2 curve indicates that as the number of training examples increases, the cross-validation score improves, suggesting that the model generalizes well and benefits from more data. The MSE curve shows a consistent decrease in both training and cross-validation errors, highlighting the model's ability to minimize prediction errors as it learns from more data. The MAE curve follows a similar trend, with errors decreasing as the training examples grow, pointing to the model's robustness in handling outliers or errors in the data.

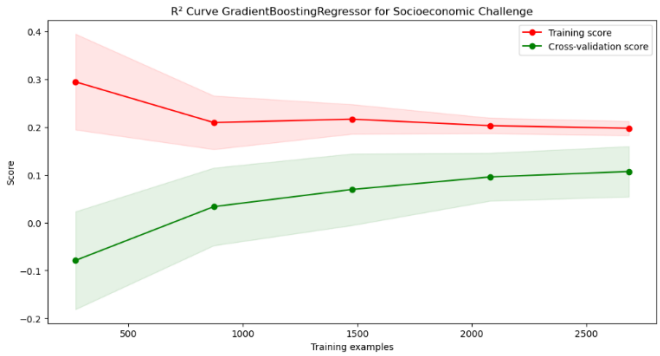

Fig. 20. R^2 Curve Gradient Boosting Regressor for Socioeconomic Challenge

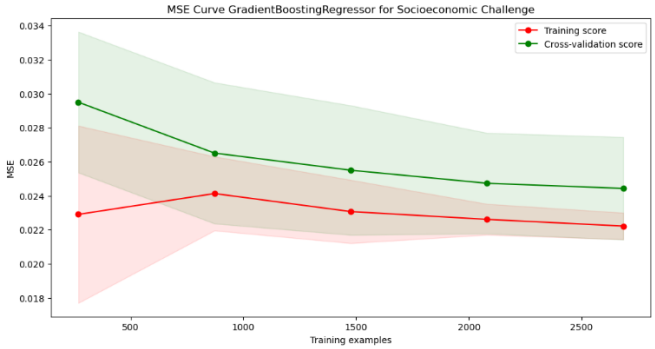

Fig. 21. MSE Curve Gradient Boosting Regressor for Socioeconomic Challenge

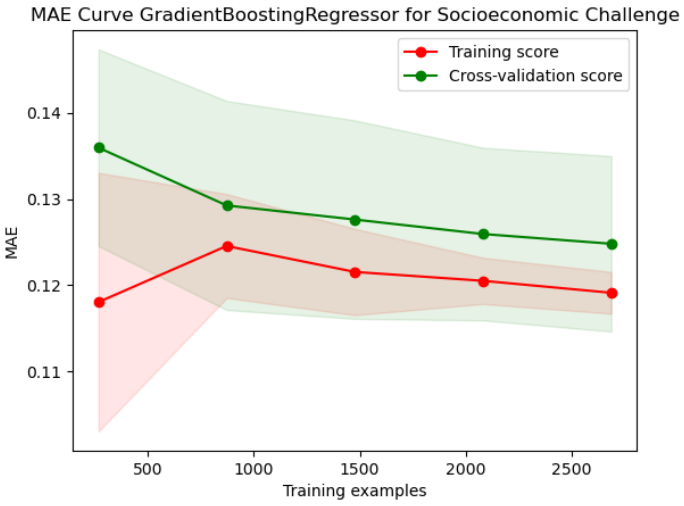

Fig. 22. MAE Curve Gradient Boosting Regressor for Socioeconomic Challenge

Connecting these evaluations back to the broader context, the socioeconomic challenge model's performance, as indicated by the R^2, MSE, and MAE, illustrates its capability in capturing the complex relationships between socioeconomic factors and student academic performance. Although the initial performance starts off less optimal with fewer data points, the increasing trend in model accuracy with more data emphasizes

the importance of rich datasets for training robust models in educational settings. These insights assist in understanding the critical socioeconomic factors that can influence educational outcomes, which can be vital for policy-making and targeted interventions.

V. CONCLUSION

The comprehensive analysis conducted in this study highlights the superior performance of the Gradient Boosting Regressor over the Random Forest Regressor in predicting student academic achievement based on various socioeconomic and enrollment factors. Through meticulous preprocessing, feature selection, and robust model evaluation, Gradient Boosting has demonstrated a more precise ability to capture the nuanced influences on educational outcomes. This outcome not only confirms the efficacy of Gradient Boosting for this type of prediction task but also highlights the potential of machine learning algorithms to unravel complex patterns within educational data. The insights gained pave the way for more informed and data-driven decision-making in educational strategies, enabling a focus on interventions that address the specific socioeconomic challenges affecting student performance.

REFERENCES

[1] J. Dunlosky, K. Rawson, and D. Willingham, "Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology," Psychological Science in the Public Interest, vol. 14, no. 1, pp. 4–58, Jan. 2013, doi: https://doi.org/10.1177/1529100612453266.

[2] E. Alyahyan and D. Düştegör, "Predicting Academic Success in Higher education: Literature Review and Best Practices," International Journal of Educational Technology in Higher Education, vol. 17, no. 1, Feb. 2020, doi: https://doi.org/10.1186/s41239-020-0177-7.

[3] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review," Applied Sciences, vol. 10, no. 3, p. 1042, Feb. 2020, doi: https://doi.org/10.3390/app10031042.

[4] S. Abdullah and H. F. Eid, "Utilizing random forest algorithm for early detection of academic underperformance in open learning environments," PeerJ, vol. 9, pp. e1708–e1708, Nov. 2023, doi: https://doi.org/10.7717/peerj-cs.1708.

[5] D.-P. Chen et al., "Exploration of the external and internal factors that affected learning effectiveness for the students: a questionnaire survey," BMC Medical Education, vol. 23, no. 1, Jan. 2023, doi: https://doi.org/10.1186/s12909-023-04035-4.

[6] Yavuz Selim Balcıoğlu and Melike Artar, "Predicting academic performance of students with machine learning," Information Development, Nov. 2023, doi: https://doi.org/10.1177/02666669231213023.

[7] Government of Ontario, "September 22, 2023," https://data.ontario.ca/. https://data.ontario.ca/dataset/school-information-and-student-demographics/resource/e0e90bd5-d662-401a-a6d2-60d69ac89d14 (accessed Apr. 07, 2024).

[8] D. Doz, M. Cotič, and Darjo Felda, "Random Forest Regression in Predicting Students' Achievements and Fuzzy Grades," Mathematics, vol. 11, no. 19, pp. 4129–4129, Sep. 2023, doi: https://doi.org/10.3390/math11194129.

[9] Tuychiev, "A guide to the gradient boosting algorithm," DataCamp, https://www.datacamp.com/tutorial/guide-to-the-gradient-boosting-algorithm (accessed Apr. 7, 2024).

[10] "Mean square error (MSE): Machine learning glossary: Encord," Encord, https://encord.com/glossary/mean-square-error-mse/#:~:text=The%20primary%20objective%20of%20the,align%20with%20the%20ground%20truth. (accessed Apr. 7, 2024).

[11] "Introduction to the correlation matrix," Built In, https://builtin.com/data-science/correlation-matrix (accessed Apr. 7, 2024).