Predicting and Understanding Student Academic Performance Using Machine Learning Techniques

George Mason University — Can Ngyuen — AIT 582

Harinipriya Vasu
Master's in Data Analytics
George Mason University
Fairfax,VA
hvasu@gmu.edu

Jerry Vishal Joseph Arockiasamy
Master's in Data Analytics
George Mason University
Fairfax,VA
jarocki@gmu.edu

Nithish Bilasunur Manjunatha Reddy
Master's in Data Analytics
George Mason University
Fairfax,VA
nbilasun@gmu.edu

Nithiya Varsha Markandan Rajendren
Master's in Data Analytics
George Mason University
Fairfax,VA
nmarkan@gmu.edu

Pritham Mahajan
Master's in Data Analytics
George Mason University
Fairfax , VA
pmahajan@gmu.edu

Rakesh Somukalahalli Venkatappa
Master's in Data Analytics
George Mason University
Fairfax,VA
rsomuka@gmu.edu

Prudhvi Raj Ananthu
Master's in Data Analytics
George Mason University
Fairfax, VA
pananthu@gmu.edu

***Abstract*—The academic success of students is influenced by a complex interplay of socioeconomic, psychological, and behavioral factors. This research leverages machine learning techniques, including Support Vector Regression, Random Forest, and Gradient Boosting, to analyze a diverse dataset encompassing demographic, academic, and social variables. Key findings highlight the critical role of attendance, study habits, and resource accessibility in predicting exam scores, while factors such as parental education and family income demonstrate minimal impact. Through hyperparameter tuning and advanced feature engineering, models achieved moderate predictive accuracy, with Gradient Boosting yielding a Mean Squared Error (MSE) of 5.79 and R-squared (R^2) of 0.59. Parental involvement exhibited limited explanatory power (R^2 = 0.03), suggesting the need for additional variables to capture variability in academic performance. This study provides actionable insights for educators and policymakers to design targeted interventions, such as improving attendance and resource distribution, fostering equitable educational opportunities, and enhancing student outcomes through data-driven approaches. Future research should incorporate longitudinal and multimodal data to refine models and address the multifaceted challenges of academic achievement.**

***Index Terms*—Machine Learning, Student Academic Performance, Predictive Modeling, Attendance, Study Habits, Educational Equity, Resource Accessibility, Parental Involvement, Gradient Boosting, Intervention Strategies**

I. Introduction

The academic success of students in secondary education has become a focal point for educators, parents, and policymakers alike, given its critical role in shaping personal and professional development by equipping students with essential skills, fostering critical thinking, and opening pathways to higher education and career opportunities. Despite increased attention, a persistent achievement gap exists, often influenced by a multitude of factors spanning socioeconomic, behavioral, and psychological domains. Socioeconomic status (SES), characterized by variables such as family income, parental involvement, and resource availability, is a key determinant in shaping academic outcomes. Research has consistently highlighted the influence of these factors, with students from lower SES backgrounds facing challenges like limited access to resources, reduced parental support, and higher absenteeism rates (Chinyoka & Naidu, 2013; Jeynes, 2007). Addressing these disparities is crucial for fostering equity in education and enhancing societal advancement.

Machine learning (ML) techniques offer a robust methodology for analyzing complex, multifaceted datasets, enabling researchers to identify significant predictors of academic performance. These methods are particularly well-suited for handling large datasets and capturing intricate relationships among variables, which are often too complex for traditional analytical approaches. Previous studies have demonstrated the efficacy of ML models, such as Random Forest, Gradient Boosting, and Support Vector Regressors, in accurately forecasting student outcomes while uncovering insights into the underlying factors affecting performance (Kotsiantis et al., 2004; Gray et al., 2014). By leveraging these techniques, this study seeks to explore the relationships between academic performance and a diverse set of variables, including attendance, study habits, parental education, and access to educational resources.

This research aims to achieve three primary objectives: (1) develop predictive models to estimate exam scores accurately by analyzing key factors, and (2) provide actionable insights to support educational policy by identifying influential variables and offering practical recommendations. These findings will not only contribute to the existing literature but also offer practical recommendations for educators and policymakers to enhance academic outcomes and bridge achievement gaps. The implications of this work extend beyond individual academic success, fostering long-term societal benefits such as improved economic stability and social equity.

Through a combination of quantitative and qualitative methodologies, this study will analyze a comprehensive dataset of secondary school students, leveraging the strengths of both approaches to provide a holistic understanding of the factors influencing academic performance. The dataset includes diverse features such as hours studied, attendance, parental involvement, family income, and psychological factors like motivation and self-control. Preliminary analysis highlights the critical role of attendance and study habits in predicting performance, aligning with existing research that emphasizes these factors as strong predictors (Kotsiantis et al., 2004; Gray et al., 2014). For instance, Kotsiantis et al. (2004) demonstrated the predictive power of attendance in educational outcomes, while Gray et al. (2014) emphasized the significant influence of structured study habits. However, other factors, such as parental education level, exhibit minimal impact in this analysis, contrasting earlier studies like Jeynes (2007), which underscored its importance. This deviation suggests that contextual or dataset-specific variables may moderate these relationships, highlighting the need for further investigation into local or demographic influences. By integrating machine learning and data-driven insights, this study offers a novel perspective on understanding and addressing the multifaceted challenges of academic achievement.

II. Related Work

The academic performance of students is influenced by a myriad of factors, encompassing educational delivery modes, socioeconomic status, parental involvement, classroom environment, peer influence, psychological factors, and study habits. Understanding these determinants is crucial for developing strategies to enhance educational outcomes.

A. Online Education and Academic Performance

Online education has emerged as a pivotal mode of learning, expanding access and convenience for students from diverse socioeconomic and geographical backgrounds. Misopoulos et al. [1] highlight that successful online programs leverage engaging tools and robust support systems, such as accessible online faculty, to foster student interaction and motivation. Additionally, the affordability of online education appeals to students concerned with the return on their investment. However, the effectiveness of online learning is contingent upon students' self-discipline and digital literacy, underscoring the importance of personal motivation for academic success in virtual environments.

B. Socioeconomic Status and Academic Achievement

Socioeconomic status (SES) plays a significant role in shaping academic achievement. Munir et al. [2] demonstrate that students from higher SES backgrounds generally perform better academically due to enhanced opportunities, such as higher-quality education, access to well-funded schools, extracurricular activities, and greater parental support. The education level and involvement of parents further contribute to this disparity, as more educated parents are better equipped to provide academic assistance and foster positive attitudes toward learning. Consequently, SES not only influences participation in enriching activities but also exacerbates the achievement gap between high- and low-SES students.

C. Parental Involvement

Parental involvement is another critical factor influencing student academic performance. Jeynes [3] conducts a meta-analysis revealing that active parental support—such as assisting with homework, attending school events, and maintaining communication with teachers—positively correlates with higher student achievement. This involvement varies across different socioeconomic and cultural contexts, with families from higher SES backgrounds typically offering more consistent academic support. Effective parental engagement not only boosts academic success but also enhances students' self-esteem and resilience, contributing to more favorable educational outcomes.

D. Peer Group Influence

Peer groups exert considerable influence on students' academic performance. Filade et al. [5] explore how peer relationships can have both positive and negative effects. Positive peer influence can enhance academic motivation and achievement, while negative peer influence may lead to the neglect of academic priorities. The nature of the peer group plays a crucial role; students associated with academically oriented groups tend to perform better, whereas those in groups that devalue academic success may experience poorer performance outcomes.

E. Psychological Factors

Psychological factors, including motivation, self-efficacy, and stress, are integral to students' academic performance. Filade et al. [5] discuss how intrinsic motivation, driven by personal interest and enjoyment in learning, positively correlates with academic success. Self-efficacy—the belief in one's academic capabilities—enhances engagement and the effectiveness of study practices. Conversely, academic stress can negatively impact well-being and performance, particularly during high-pressure periods like examinations. Implementing stress management techniques and support systems is therefore essential for maintaining optimal academic performance and overall student well-being.

F. Study Habits and Time Management

Effective study habits and time management skills are fundamental to academic success. Mashburn [6] analyzes undergraduate students at a Historically Black College and University (HBCU), finding that good study habits—such as maintaining a regular study schedule, employing active learning techniques, and taking effective notes—significantly enhance academic performance. Additionally, strong time management allows students to balance academic responsibilities with other commitments, reducing stress and improving overall performance. Conversely, poor study habits and inadequate time management are associated with lower academic achievement, highlighting the necessity of fostering these skills to support student success in higher education.

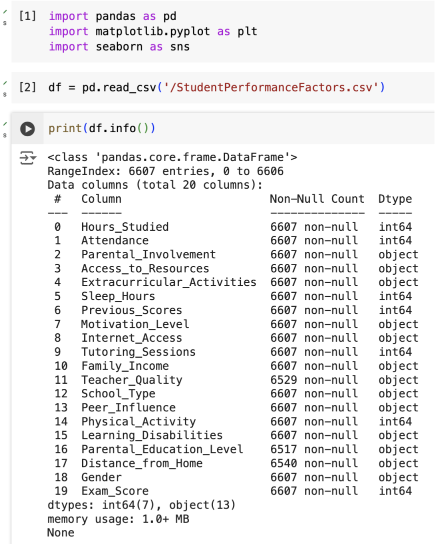
III. Primary Process Explained

The process of understanding and predicting academic performance among secondary school students involves a comprehensive analysis of various factors. This study begins by formulating targeted research questions aimed at dissecting the interplay between socioeconomic, behavioral, and psychological influences on academic outcomes. The journey from identifying these research questions to drawing actionable insights involves several critical stages.

The primary step in this research involves defining the key questions to guide the investigation. These include exploring the relationship between exam scores and family income, assessing the impact of parental involvement, and understanding how extracurricular activities and teacher actions contribute to academic success. This step is essential to frame the scope of the study and ensure a focused approach toward addressing specific gaps in existing literature.

Following the formulation of research questions, the study employs a dual-method approach, integrating both quantitative and qualitative analyses. Quantitative analysis plays a pivotal role, utilizing advanced statistical and machine learning models to unravel the patterns and predictors of academic achievement. Techniques such as regression analysis and algorithms like Random Forest and Gradient Boosting are applied to analyze the influence of variables like Hours Studied, Attendance, and Access to Resources. These models not only offer high predictive accuracy but also provide valuable insights into the relative importance of different factors.

Simultaneously, qualitative analysis complements the numerical findings by delving into the lived experiences of students, parents, and educators. This phase involves conducting interviews and focus group discussions to uncover nuanced challenges and barriers that numbers alone cannot reveal. For example, while quantitative data might highlight the correlation between low family income and poor academic performance, qualitative insights shed light on how financial stress impacts a student's motivation and concentration.



```
[1] import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
```

```
[2] df = pd.read_csv('/StudentPerformanceFactors.csv')
```

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6607 entries, 0 to 6606
Data columns (total 20 columns):
 #   Column                      Non-Null Count  Dtype 
---  ------                      --------------  ----- 
 0   Hours_Studied               6607 non-null   int64 
 1   Attendance                  6607 non-null   int64 
 2   Parental_Involvement        6607 non-null   object
 3   Access_to_Resources         6607 non-null   object
 4   Extracurricular_Activities  6607 non-null   object
 5   Sleep_Hours                 6607 non-null   int64 
 6   Previous_Scores             6607 non-null   int64 
 7   Motivation_Level            6607 non-null   object
 8   Internet_Access             6607 non-null   object
 9   Tutoring_Sessions           6607 non-null   int64 
 10  Family_Income               6607 non-null   object
 11  Teacher_Quality             6529 non-null   object
 12  School_Type                 6607 non-null   object
 13  Peer_Influence              6607 non-null   object
 14  Physical_Activity           6607 non-null   int64 
 15  Learning_Disabilities       6607 non-null   object
 16  Parental_Education_Level    6517 non-null   object
 17  Distance_from_Home          6540 non-null   object
 18  Gender                      6607 non-null   object
 19  Exam_Score                  6607 non-null   int64 
dtypes: int64(7), object(13)
memory usage: 1.0+ MB
None
```

Fig. 1. Snapshot of Dataset

IV. Dataset

The dataset, which was sourced via Kaggle, includes academic, social, and demographic data gathered from two schools. It offers insights on student behaviors and is a useful tool for comprehending the elements affecting academic success. It is perfect for exploratory data analysis and predictive modeling because of its wide range of characteristics, which include hours studied, attendance, socioeconomic aspects, and more.

This dataset offers a comprehensive perspective on the factors influencing students' academic performance. The dataset enables researchers to examine complex correlations between socioeconomic conditions, individual behaviors, and educational resources by merging numerical and categorical data. A thorough framework for determining performance predictors is provided by variables like teacher quality, internet access, and family participation, which broaden the scope of the investigation.Additionally, adding a variety of characteristics, such as physical activity and sleep duration, adds special perspectives that are frequently missed in traditional academic performance research.

The diverse features of the dataset also offer a chance to investigate subgroup analyses and temporal trends. Features like motivation level and peer influence can disclose psychological and social dynamics, whereas family wealth and resource accessibility, for example, can assist highlight differences across various socioeconomic groups. Educational institutions and policymakers can create focused initiatives to improve learning outcomes and address systemic injustices by utilizing this dataset.

Additionally, for prediction purposes, the dataset makes it possible to integrate sophisticated machine learning models. Key determinants of academic achievement include past test scores and study hours, although moderating impacts of other factors, such as internet availability and instructor caliber, can be investigated. Mixed-methods research approaches are also supported by the existence of both qualitative and quantitative data, which enables a more nuanced understanding of the factors influencing educational success.

The dataset is robust since it covers both extrinsic (like family income and school type) and intrinsic (like motivation level and learning difficulties) components. This two-pronged strategy guarantees that the analysis fully captures the intricacy of factors influencing academic success. The dataset's usefulness goes beyond academia as well, as findings can inform curriculum design, resource allocation, and support systems for students from a variety of backgrounds in the actual world of educational policy-making.

A. Data Preprocessing

1) Handling Missing Values: The dataset was thoroughly examined for missing values in critical columns such as teacher quality, parental education level, and distance from home. Missing values in categorical variables were filled with placeholders like ”Unknown,” while missing values in numerical columns were imputed using statistical measures like the mode or median. In cases where significant portions of data were missing, rows were removed to maintain the dataset's reliability and accuracy.

2) Removing Duplicate Rows: During data cleaning, duplicate rows were identified and removed to ensure each observation appeared only once. This step was crucial to maintain the integrity of the dataset and avoid biases that could misrepresent trends. With duplicates removed, the analysis became more precise, ensuring accurate insights into the factors influencing academic performance.

3) Checking and Converting Data Types: The data types of each column were carefully reviewed to ensure correctness. Numerical columns, such as attendance and sleep hours, were converted to appropriate numeric types to enable statistical calculations. Categorical columns, like teacher quality and family income, were standardized to ensure consistency. These standardized categorical variables were encoded for compatibility with machine learning models, enhancing the dataset's usability for predictive analysis.

4) Ensuring Data Quality: Quality checks were conducted to address inconsistencies in data, such as variations in categorical labels. For instance, similar entries like ”Internet” and ”internet access” were standardized. These corrections ensured the dataset was accurate, consistent, and ready for further analysis. By resolving these issues, the dataset was prepared for exploratory data analysis (EDA) and model development, enabling reliable insights into the factors affecting student performance.

B. Exploratory Data Analysis

The exploratory analysis aimed to uncover patterns, relationships, and trends within the dataset. Key visualizations and their interpretations include:

1) Distribution of Exam Scores: The distribution of exam scores provides an initial overview of student performance variability across the dataset. By visualizing scores with a histogram and KDE overlay, we can determine the central tendency, identify any skewness, and detect outliers. This visualization is foundational for understanding the general academic performance of students and setting benchmarks for further analysis. It also establishes the need to explore underlying factors that might explain variations in scores.

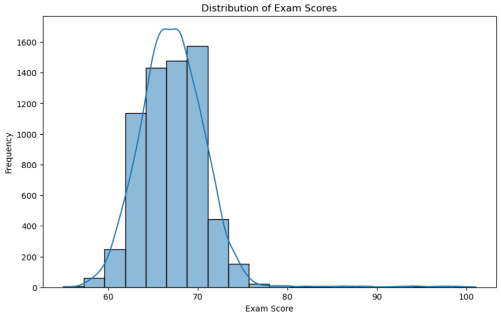


Fig. 2. Distribution of Exam Scores

This normal distribution suggests that most students perform near the average, with a concentration around a score of 70. The bell-shaped curve allows us to assume a baseline performance and emphasizes that outlier effects are minimal. This foundational understanding reinforces the importance of examining features with strong potential to impact scores, like attendance and hours studied.

2) Correlation Heatmap: The correlation heatmap helps identify relationships between exam scores and other numerical features, such as attendance, hours studied, and physical activity. Visualizing correlations in this way allows us to quickly pinpoint which variables are most closely linked to academic performance. Strongly correlated factors are ideal candidates for inclusion in predictive models, as they have the potential to drive meaningful predictions.

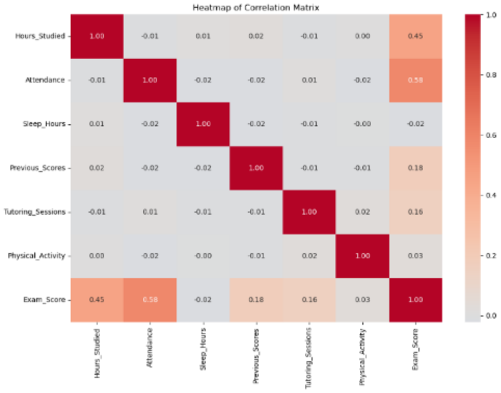


Fig. 3. Correlation Heatmap

The heatmap shows correlations between exam scores and other numerical features in the dataset. Attendance and hours studied stand out, with attendance having the strongest positive correlation (0.58) and hours studied showing a moderate positive correlation (0.45). These findings indicate that increased attendance and study hours have a substantial impact on academic outcomes. Other variables, like physical activity and sleep hours, demonstrate weak or negligible correlations, emphasizing that they have minimal influence on performance.

3) Pairplot of Key Variables: The pairplot provides a comprehensive look at interactions between multiple variables, such as hours studied and attendance, in relation to exam scores. By using color gradients to represent different exam score levels, we can visually assess patterns and clusters that may indicate performance trends. This plot allows us to explore feature interactions and check if any non-linear patterns emerge.

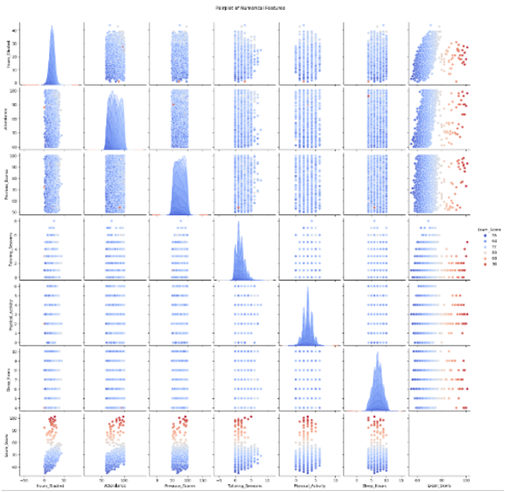


Fig. 4. Pairplot of key variable

The positive relationships between hours studied, attendance, and exam scores reinforce that these variables are strong performance indicators. The clustering of higher scores with greater study hours and higher attendance shows that these factors are essential for prediction. This pairplot analysis further supports focusing on these key variables in model training.

4) Relationship between sleep hours and participation in extracurricular activities among students: The histogram illustrates the relationship between sleep hours and participation in extracurricular activities among students. It compares the distribution of sleep hours for those who participate in extracurricular activities and those who do not. Students not involved in extracurricular activities tend to have a peak sleep duration of around 7 hours, with a higher proportion sleeping consistently in the 7–9 hour range. On the other hand, students involved in extracurricular activities display a wider spread of sleep hours, ranging between 6 and 7 hours for many, but with more variability overall. Both groups have fewer students at the extremes of 4 or 10 hours of sleep, indicating that such sleep patterns are uncommon. Overall, participation in extracurricular activities appears to affect sleep consistency, as those not involved have more uniform sleep patterns, whereas participants show greater variation, likely due to balancing activities with academic demands.

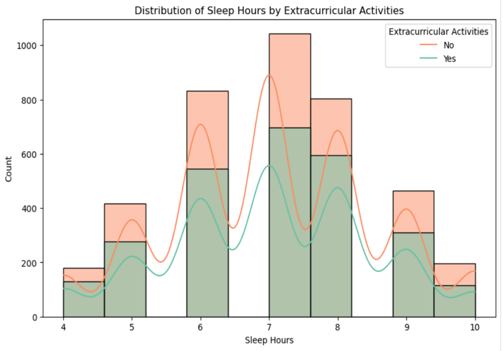


Fig. 5. Sleep hours Vs Extracurricular activities

5) Impact of Peer Influence and Physical Activity on Exam Scores: The chart illustrates how Peer Influence Types (Negative, Neutral, and Positive) interact with Physical Activity Levels to affect students' average exam scores. By visualizing this relationship using a grouped bar chart with distinct color gradients, we can identify trends and variations in academic performance based on social and physical factors. This visualization provides a foundation for exploring how supportive peer environments and active lifestyles contribute to exam success. It also emphasizes the need to examine the compounded effects of social and physical activity dynamics on academic outcomes. This analysis shows that students with Positive Peer Influence consistently achieve higher scores across all Physical Activity Levels. Conversely, students with Negative Peer Influence tend to score lower, highlighting the impact of discouraging peer environments. Higher Physical Activity Levels (5 and 6) correspond to better performance across all peer groups, while inactivity (Level 0) is linked to the lowest scores. These findings underscore the importance of promoting active lifestyles and fostering positive peer relationships to improve academic outcomes. This foundational understanding highlights the value of interventions targeting both physical and social factors to optimize student success.

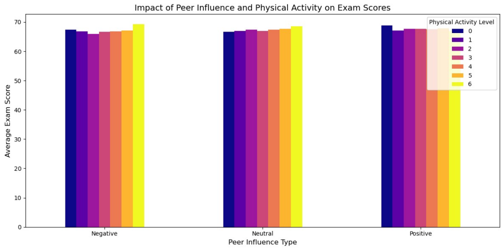


Fig. 6. Impact of Peer Influence and Physical Activity on Exam Scores

6) Influence of Internet Access and Motivation Level on Student Performance: The bar chart (Figure 1) highlights the relationship between motivation levels, Internet access, and exam performance. Students with higher motivation levels

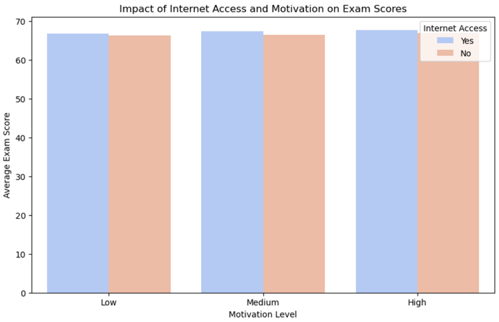


Fig. 7. Impact of Internet Access Vs Motication Level

consistently achieve better scores, regardless of Internet access. For students with low motivation, those with Internet access score slightly higher (average score of 67) compared to those without (average score of 66). In the medium motivation category, there is no difference in scores, with both groups achieving an average score of 67, suggesting that Internet access has minimal impact for this group. However, for highly motivated students, Internet access provides a slight edge, increasing their scores from 67 (without Internet) to 68 (with Internet). This emphasizes that motivation is the key driver of performance, while Internet access plays a supporting role, especially for students with low or high motivation.

7) Impact of Peer Influence and School Type on Exam Scores: The analysis presented in Figure 1 explores the relationship between peer influence (Positive, Negative, Neutral) and exam scores across public and private schools. The boxplot provides a detailed comparison, where the y-axis represents exam scores (ranging from 0 to 100), and the x-axis categorizes the type of peer influence. Public schools are depicted in blue, and private schools in orange, with each boxplot summarizing the interquartile range (IQR), the median, and the spread of exam scores for each category.

Across all types of peer influence, students in private schools consistently demonstrate slightly higher median scores than those in public schools. However, this difference is marginal, suggesting that school type alone may not be a significant determinant of academic performance. Both school types show a wide range of scores, with public schools exhibiting greater variability, as reflected by the longer whiskers. This variability could point to disparities in resources, teaching quality, or student demographics within public schools.

When examining peer influence, students experiencing Positive peer dynamics tend to have slightly higher median scores compared to those under Negative or Neutral influence. This trend is more prominent in private schools, where the Positive influence category demonstrates better concentration and slightly higher performance. Conversely, Negative influence results in wider score variability, though the median scores remain relatively consistent across school types. Neutral influence, on the other hand, shows balanced distributions with medians similar to those of Negative influence, implying a limited impact on performance.

In conclusion, as highlighted in Figure 1, neither peer influence nor school type emerges as a dominant factor independently. Positive peer dynamics show a modest benefit for exam performance, particularly in private schools. However, the larger score variability in public schools underscores the need to address systemic disparities. Enhancing peer relationships and providing equitable resources in public schools could help reduce this variability and improve overall academic outcomes.

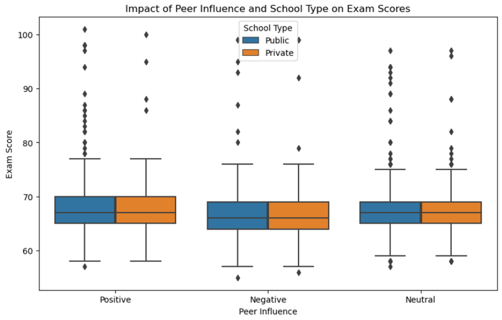


Fig. 8. Peer Influence Vs School type on Exam Scores

V. Higher-Level Framework

Research Question 1: How Do Extracurricular Activities and Resources Impact Academic Success?

Objective

This study looks at how extracurricular activities and access to resources affect students' exam scores. The goal is to see if these factors help predict academic performance and how strong their impact is.

Methodology

1) Data Preparation:
   - Two key features, `Extracurricular Activities` and `Access to Resources`, were prepared by converting them into numerical values to make them easier to use in the model.
   - Exam scores were set as the target variable, which we aim to predict.
2) Splitting the Data:
   - The data was divided into two parts: 80% was used to train the model, and 20% was used to test its accuracy on unseen data.
3) Model Used:
   - A simple machine learning model called Support Vector Regressor (SVR) was used. It tries to find the best relationship between the given features and exam scores.
   - After training, the model was tested to predict exam scores for new data.
4) Visualization:

- A graph (Fig. 6) was created to compare actual exam scores (on the x-axis) with predicted scores (on the y-axis).
- A blue line was added to show the overall trend, and a shaded area around the line indicates how confident the model is.

Results

1) General Trend:
   - The blue line in Fig. 6 shows a positive trend, meaning higher actual exam scores are generally matched by higher predicted scores.
   - However, the orange dots are not perfectly aligned with the line, indicating the model's predictions are not always exact.
2) Spread of Scores:
   - The spread of orange dots around the blue line shows that the model is not fully capturing all factors that influence exam scores.
   - Some clusters of dots suggest that students with similar extracurricular activities and access to resources have similar scores.
3) Confidence in Predictions:
   - The shaded area around the blue line shows the model's confidence. A narrow band means the model is confident; a wider band means it is less certain.
4) Impact of Features:
   - The results show that extracurricular activities and access to resources influence exam scores.
   - However, the results also suggest that other factors, like hours studied or teacher quality, may have a strong impact too.

Insights

- Key Finding: Students involved in extracurricular activities and with good access to resources tend to perform better in exams.
- Model Limitations: The model does not fully explain all variations in scores. Adding more features could improve accuracy.
- Takeaway: While extracurricular activities and resources matter, other factors should also be explored to better understand academic success.

The findings of this study are shown in Fig. 6.

Research Question 2: What Effect Does Parental Participation Have on Students' Academic Performance?

Objective

To assess how parental involvement impacts students' academic performance, this analysis examines the relationship between the level of parental involvement and students' exam scores.

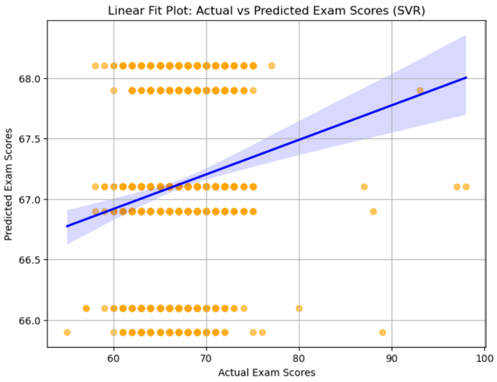


Fig. 9. Linear Fit Plot: Actual vs Predicted Exam Scores (SVR).

Methodology

1) Data Preparation:
   - `Parental_Involvement` was the only feature used to predict `Exam_Score`.
   - Categorical levels of parental involvement (e.g., "Low," "Medium," "High") were converted into numerical values using `pd.get_dummies`. This step allowed the model to process the data effectively.
2) Model Used:
   - A Random Forest Regressor was applied to model the relationship between parental involvement and exam scores.
   - Model performance was evaluated using:
     - Mean Squared Error (MSE): 13.76. This measures the average difference between predicted and actual scores. A lower value indicates better accuracy.
     - R-squared (R^2): 0.03. This shows that parental involvement explains only a small portion of the variation in exam scores, suggesting that additional factors likely contribute significantly.

Graph Analysis

1) Performance Analysis:
   - The MSE value of 13.76 indicates moderate prediction accuracy. A lower MSE would suggest a closer match between predicted and actual scores.
   - The R^2 value of 0.03 is relatively low, meaning parental involvement alone explains only a small portion of the variability in exam scores. This highlights the potential influence of other factors.
2) Scatter Plot Insights:
   - The scatter plot compares actual exam scores with predicted scores.
   - The red dashed line represents perfect predictions, where the actual and predicted scores match. Points closer to the line indicate accurate predictions, while points farther away show greater errors.
3) Variability of Predictions:

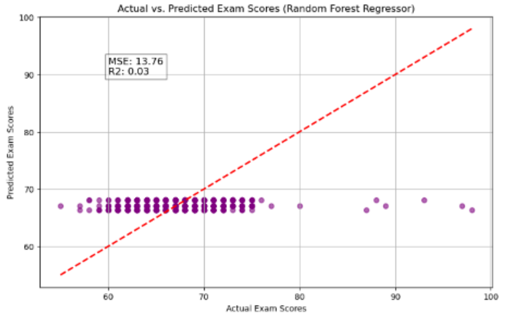


Fig. 10. Scatter Plot: Predicted vs. Actual Exam Scores

- The scatter plot shows a spread of points around the red line, highlighting variability in the effect of parental involvement on exam scores.
- The inconsistent pattern suggests that some students benefit more significantly from parental participation, while others do not show as strong a correlation.

Conclusions

- **Key Finding:** Parental involvement plays a role in predicting exam scores but has a limited impact, as shown by the low R^2 value.
- **Model Validity:** The Random Forest Regressor provides moderate accuracy in predicting scores based on parental involvement, but the variability in predictions indicates that other factors should be considered.
- **Implications:** While parental involvement is important, a comprehensive approach including additional features (e.g., socio-economic status, teacher quality, study habits) is necessary to better understand academic performance.

Research Question 3: How Does Tuning a Gradient Boosting Regressor Impact the Accuracy and Predictive Power in Forecasting Exam Scores?

Detailed Explanation

1. Research Context:

- This study investigates how Gradient Boosting Regression, with hyperparameter tuning, predicts exam scores.
- The objective is to optimize key parameters to balance predictive accuracy and model generalizability.

2. Methodology:

1) Dataset Split:
   - The dataset was divided into training and testing sets to evaluate the model on unseen data, simulating real-world performance.
2) Gradient Boosting Regressor:
   - Builds an ensemble of decision trees, correcting errors iteratively.
   - Handles both linear and non-linear regression tasks effectively.
3) Grid Search with Cross-Validation:
   - Optimized three hyperparameters:
     – `n_estimators`: Number of boosting stages.
     – `learning_rate`: Controls each tree's contribution to the final prediction.
     – `max_depth`: Limits tree depth to prevent overfitting.
   - Used 5-fold cross-validation to ensure robust and unbiased hyperparameter tuning.
4) Hyperparameter Combinations:
   - Tested 27 combinations (3 values each for `n_estimators`, `learning_rate`, and `max_depth`), totaling 135 fits (5 folds × 27 combinations).
5) Optimal Parameters:
   - `learning_rate`: 0.1
   - `n_estimators`: 100
   - `max_depth`: 3
   - These parameters provided a good balance between accuracy and generalization.

3. Model Performance:

1) Metrics:
   - Mean Squared Error (MSE): 5.79
   - R-squared (R^2): 0.59
   - These metrics indicate moderate accuracy, with room for improvement.
2) Scatter Plot Analysis:
   - The scatter plot compares actual vs. predicted exam scores.
   - Most predictions align with the diagonal red line (perfect prediction line), reflecting reasonable accuracy.
   - Deviations occur for outliers and higher actual scores, indicating the model struggles with extreme values.

Interpretation of Results

- The tuned model predicts mid-range exam scores effectively, aligning closely with actual values.
- Hyperparameter tuning significantly improves model performance compared to untuned models.
- While the model performs well for mid-range predictions, variability in extreme values suggests additional factors could enhance accuracy.

Conclusion

This research provides a comprehensive analysis of the factors influencing student academic performance by leveraging machine learning techniques such as Support Vector Regression, Random Forest Regressor, and Gradient Boosting. The study highlights critical variables like attendance, study habits, and resource accessibility as significant predictors of academic success. Findings indicate that attendance and study hours have the strongest positive correlations with exam scores, aligning with previous literature on the importance of

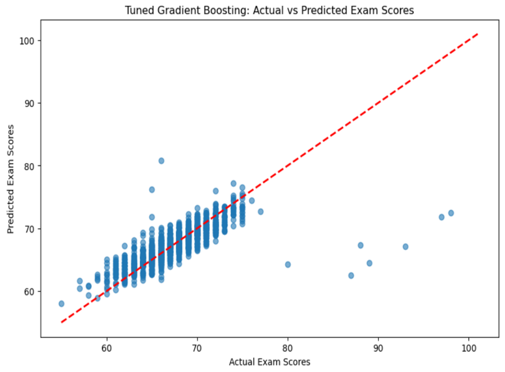


Fig. 11. Tuned Gradient Boosting: Actual vs. Predicted Exam Scores. MSE = 5.79, R² = 0.59.

consistent effort and presence in academic settings (Kotsiantis et al., 2004; Gray et al., 2014). Meanwhile, socioeconomic factors such as family income and parental education showed minimal direct impact, emphasizing the need to focus on actionable behavioral and resource-based interventions.

Advanced machine learning methodologies, including hyperparameter tuning for Gradient Boosting, demonstrated moderate accuracy in predicting exam scores (MSE = 5.79, R² = 0.59), with mid-range predictions performing well but variability in extreme values signaling room for improvement. Parental involvement, while impactful in some contexts, accounted for a limited portion of exam score variability (MSE = 13.76, R² = 0.03), suggesting additional influencing factors such as peer influence, psychological factors, and classroom environments should be incorporated in future studies (Jeynes, 2007; Filade et al., 2019).

Overall, this study underscores the utility of machine learning in identifying actionable insights for improving academic outcomes. The findings advocate for targeted interventions like promoting attendance, structured study habits, and equitable resource distribution to bridge achievement gaps and foster educational equity. Future research should incorporate a broader set of variables, including demographic and psychological factors, to further refine predictive models and address the multifaceted challenges of academic achievement.

Future work & Recommendations

Future research should explore a broader range of variables to better understand the multifaceted factors influencing academic performance. These variables could include psychological attributes such as stress levels, self-efficacy, and intrinsic motivation, as well as environmental factors like teacher quality, classroom engagement, and peer group dynamics. Investigating these dimensions can provide a more holistic understanding of how various internal and external factors contribute to student success. Additionally, future studies should incorporate longitudinal datasets to capture changes in academic performance over time, offering a clearer perspective on the sustained impact of interventions and life circumstances.

Expanding the dataset to include students from diverse geographic, socioeconomic, and cultural contexts is critical for improving model generalizability and identifying region-specific factors that may not be evident in the current dataset. For example, differences in education systems, cultural attitudes toward learning, and resource availability can significantly shape academic outcomes. Moreover, integrating multimodal data, such as behavioral data collected through learning management systems or psychometric assessments, can enhance the depth of analysis.

In terms of methodology, applying advanced machine learning models like neural networks, ensemble methods, or hybrid approaches could improve predictive accuracy, especially for non-linear relationships or high-dimensional datasets. Incorporating explainable AI (XAI) frameworks would further enable the identification of actionable insights, ensuring that predictions and recommendations are interpretable for educators and policymakers. Additionally, performing sensitivity analysis on feature importance could help identify critical intervention points that yield the highest impact on student performance.

For practical applications, it is recommended that educators and institutions prioritize interventions targeting high-impact factors such as attendance, study habits, and resource accessibility. Schools should implement attendance improvement programs, such as monitoring systems with parental notifications, to encourage consistent student participation. Structured workshops and training sessions can be introduced to teach effective study strategies and time management skills, equipping students with tools to enhance their learning efficiency. Moreover, policymakers should allocate resources to bridge disparities in educational materials and technological access, particularly for underprivileged communities, to ensure equity in learning opportunities.

Parental involvement initiatives, such as parent-teacher communication platforms and workshops aimed at enhancing parental engagement in education, can help bridge gaps for students who lack sufficient academic support at home. To address psychological barriers, schools should offer counseling services and stress management programs, especially during exam periods, to improve mental well-being and academic outcomes. Additionally, tailoring teaching methodologies based on data-driven insights—such as offering personalized learning paths and adaptive curricula—can foster an inclusive and supportive learning environment.

In conclusion, future research should aim for a multidisciplinary approach that combines psychological, social, and academic dimensions with cutting-edge machine learning techniques. By addressing these aspects through targeted interventions and policy reforms, the education system can better support students' diverse needs, promote equitable learning opportunities, and enhance overall academic success.

Acknowledgment

We are deeply grateful to Professor Can Nguyen for his invaluable guidance, crucial advice, and unwavering support throughout the course of this study. His insightful feedback, constructive critiques, and constant encouragement have been instrumental in shaping the direction and outcomes of our work, helping us overcome challenges and achieve our objectives. We are also sincerely thankful to George Mason University for providing us with an exceptional academic environment, access to cutting-edge resources, and the opportunities needed to carry out this research. The infrastructure and support from the university have been vital in facilitating our exploration and analysis. Finally, we extend our heartfelt gratitude to our dedicated team members for their relentless hard work, collaboration, and commitment. Their combined efforts and shared vision were essential in ensuring the successful completion of this project. This endeavor would not have been possible without the collective support and contributions of all involved.

References

[1] Chinyoka, K., & Naidu, N. (2013). Influence of home-based factors on the academic performance of girl learners from poverty-stricken families: A case of Zimbabwe. Mediterranean Journal of Social Sciences, 4(14), 223-233. https://doi.org/10.5901/mjss.2013.v4n14p223

[2] Gray, G., McGuinness, C., & Owende, P. (2014). An application of supervised machine learning to predict student performance. Proceedings of the 2014 International Conference on Educational Data Mining (EDM), 21-30.

[3] Jeynes, W. H. (2007). The relationship between parental involvement and urban secondary school student academic achievement: A meta-analysis. Urban Education, 42(1), 82-110. https://doi.org/10.1177/0042085906293818

[4] Kotsiantis, S. B., Pintelas, P. E., & Athanasopoulos, G. A. (2004). Predicting students' performance in distance learning using machine learning techniques. Applied Artificial Intelligence, 18(5), 411-426. https://doi.org/10.1080/08839510490442053

[5] Misopoulos, F., Argyropoulou, M., & Tzavara, D. (2017). Exploring the factors affecting student academic performance in online Programs: a literature review. Online, 235–250. https://doi.org/10.1007/978-3-319-62776-218.

[6] Munir, J., Faiza, M., Jamal, B., Daud, S., & Iqbal, K. (2023). The impact of socio-economic status on academic achievement. Journal of Social Sciences Review, 3(2), 695–705. https://doi.org/10.54183/jssr.v3i2.308

[7] Jeynes, W. H. (2006). The relationship between parental involvement and urban Secondary school student academic achievement. Urban Education, 42(1), 82–110. https://doi.org/10.1177/0042085906293818

[8] Widiyawanti, S. B. & Wahyono. (2024). The Importance of Setting the Classroom Learning Environment to Optimize its Function as a Learning Resource. In J. Electrical Systems (Vols. 20–5s, pp. 1088–1092). https://pdfs.semanticscholar.org/1e9a/734d2d92714cfb6f8cb23b5fc8a5a8554c99.pdf

[9] B, Bello, A., Uwaoma, C., Anwanane, B., & Nwangburuka, K. (2019, June 10). Peer group influence on academic performance of undergraduate students in Babcock University, Ogun State. African Educational Research Journal - Net Journals. https://www.netjournals.org/zAERJ19010.html.

[10] Zou, S. (2020). Research on GDP forecast of Ji'an City based on ARIMA model. Open Journal of Social Sciences, 08(12), 353–365. https://doi.org/10.4236/jss.2020.812029

[11] Filade, B. A., Bello, A. A., Uwaoma, C. O., Anwanane, B. B., & Nwangburuka, K. (2019). Peer group influence on academic performance of undergraduate students. African Educational Research Journal, 7(2), 81–87. https://doi.org/10.30918/AERJ.72.19.010

[12] Mashburn, D. (2020). The effects of study habits on academic performance: An analysis of undergraduate students at a Historically Black College and University. Open Journal of Social Sciences, 8(12), 381–399. https://doi.org/10.4236/jss.2020.812029