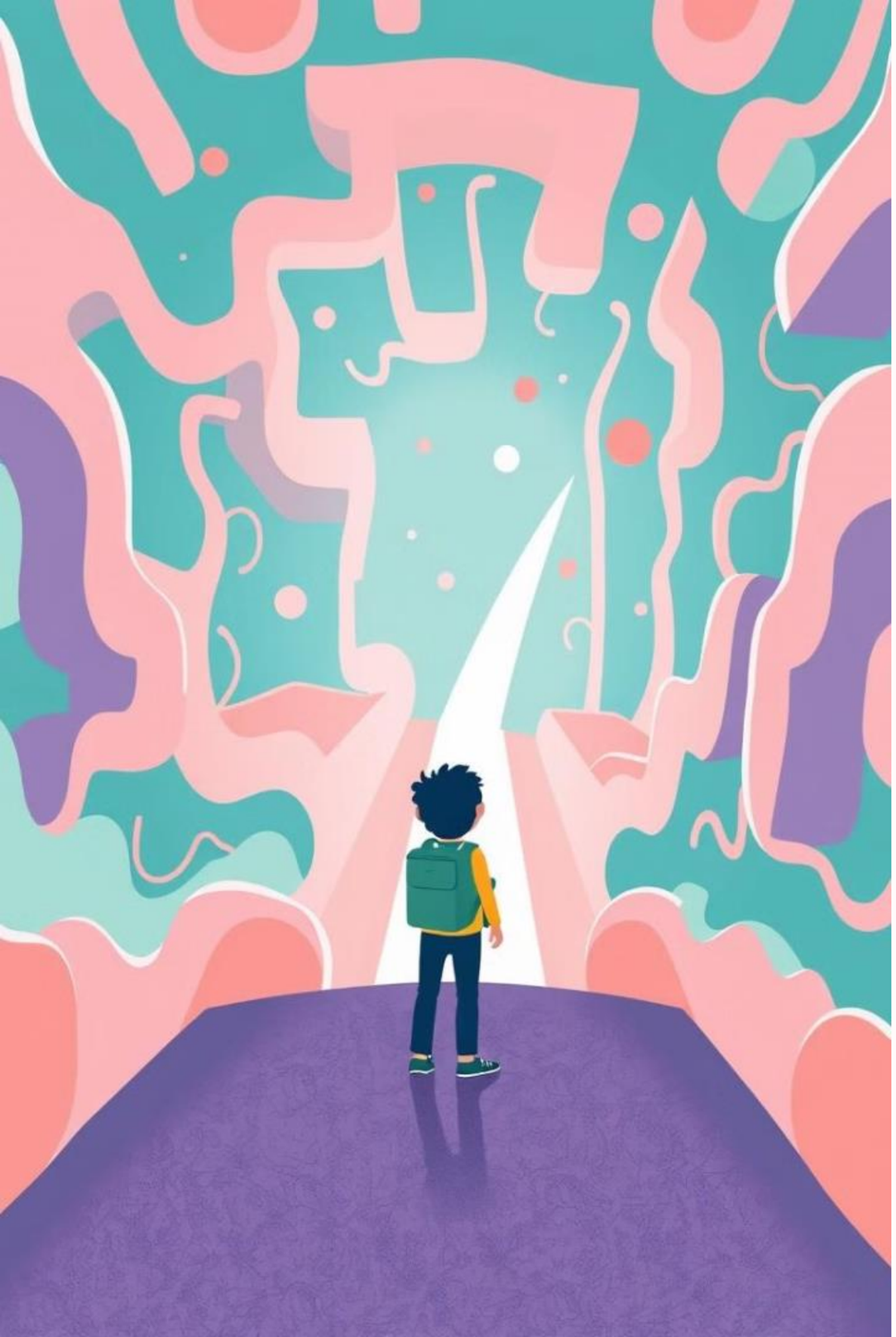
Predicting Student Academic Performance with Machine Learning

Rakesh Somukalahalli Venkatappa

Website link - https://mason.gmu.edu/~nmarkand/



Introduction

- 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.
- The dataset comprises 6,607 records with 20 features, including variables like attendance, study habits, parental involvement, and resource access.
- Academic success in secondary education is critical for personal and professional growth.
- Socioeconomic status (SES) often influences academic outcomes, with factors like family income, parental participation, and resource access impacting student performance.
- The complexity of this problem is further influenced by other variables, such as sleep patterns and teacher quality. This research examines a comprehensive set of factors to understand and potentially mitigate these performance gaps.

Problem Statement

- We see that there is an achievement gaps persist across socioeconomic backgrounds, with students from low-income households often facing barriers like limited resources, higher absenteeism, and reduced parental support.

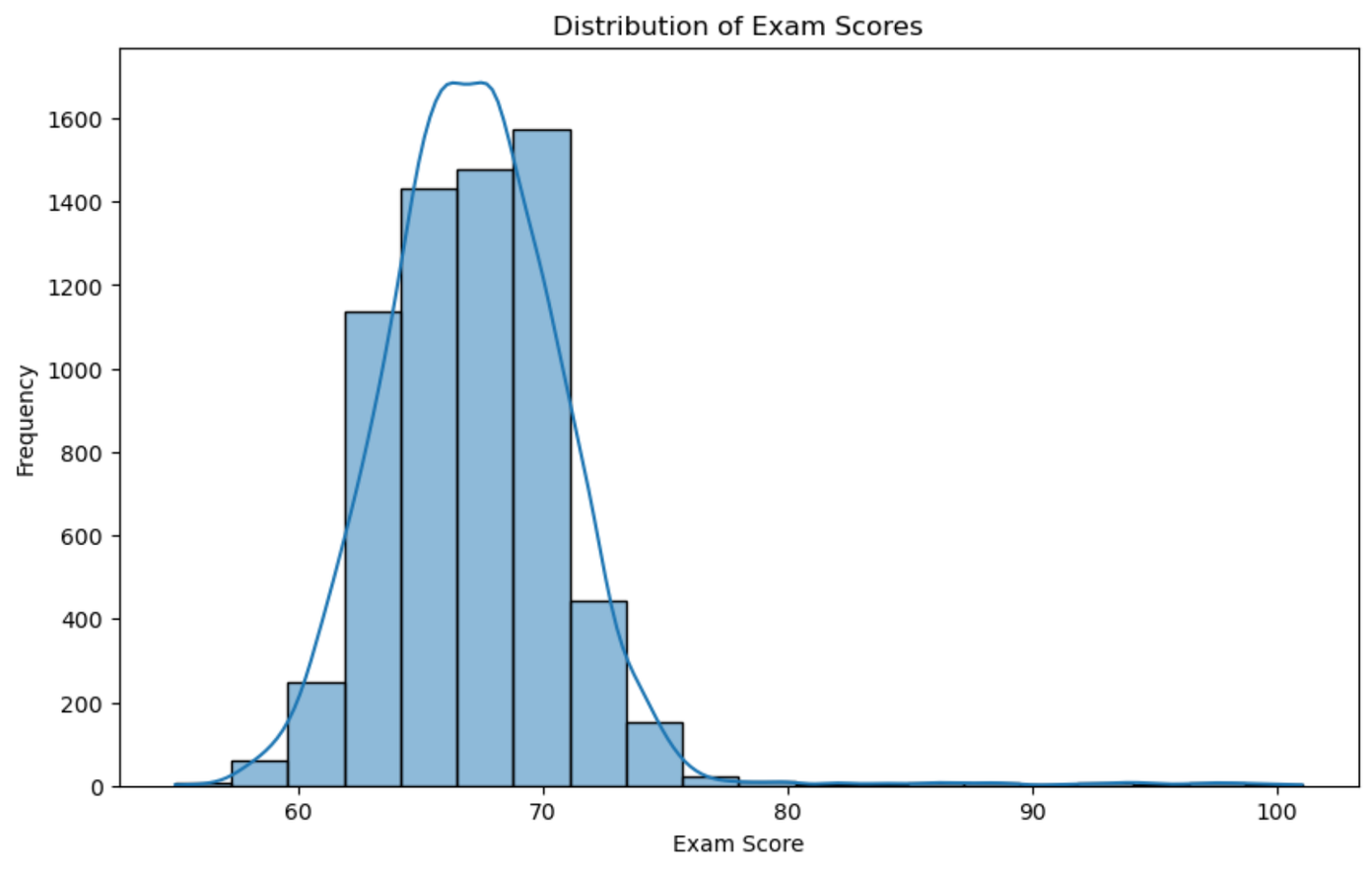
- This study examines how factors such as attendance, study habits, family income, and parental involvement influence academic performance in secondary school students.

- By analyzing these variables, the research aims to uncover actionable insights to address performance disparities and support equitable educational outcomes.

Significance

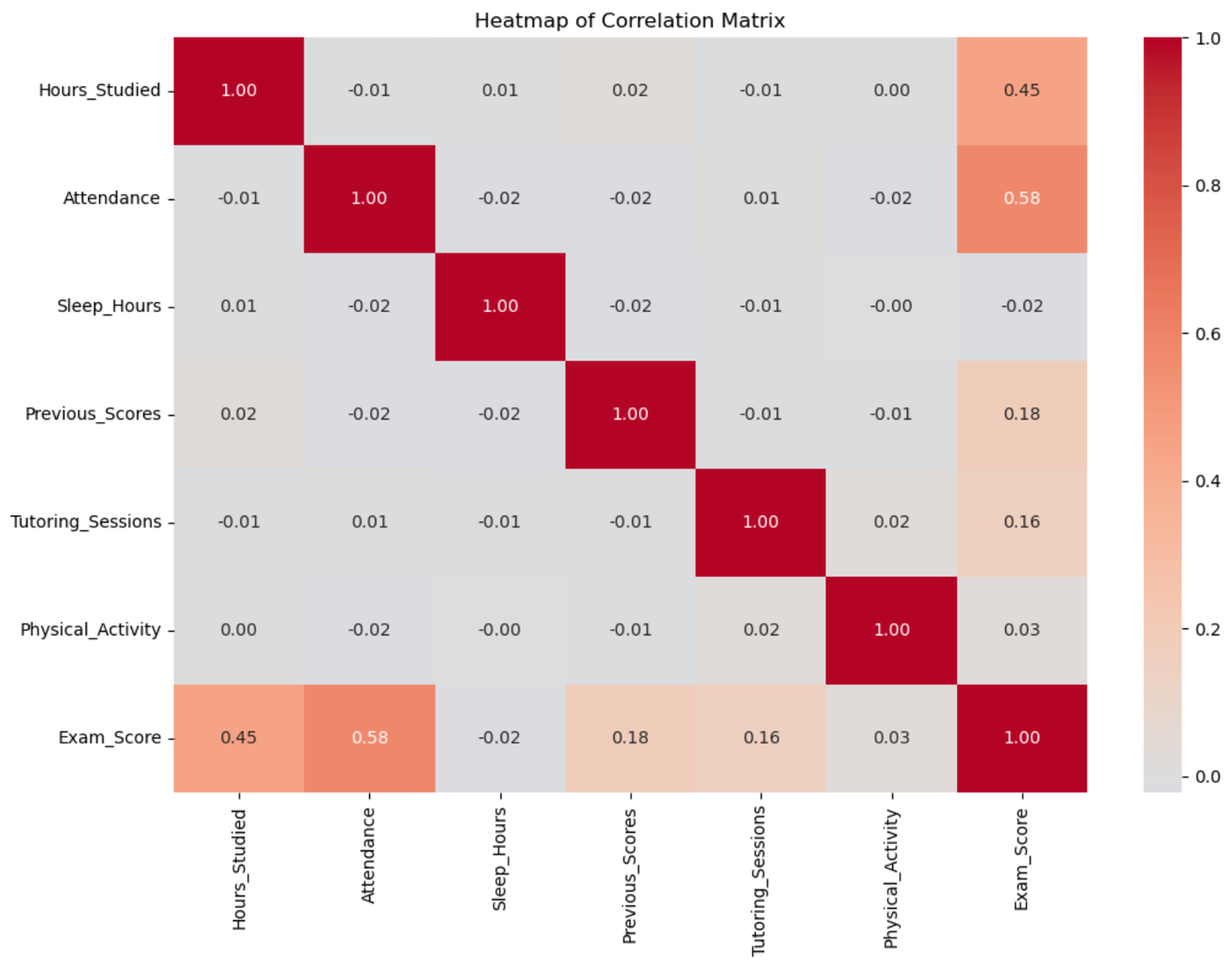
- This Research examines if there is any connection that exists between secondary school pupils' academic achievement (exam scores) and family income.
- This explores the impact of parental participation on students' academic performance.
- Also, it investigates the role of extracurricular activities and resource availability in shaping a student's academic success.
- This analyzes the impact of teacher actions and sleep habits on students' performance and motivation.

EDA - Distribution of Exam Scores



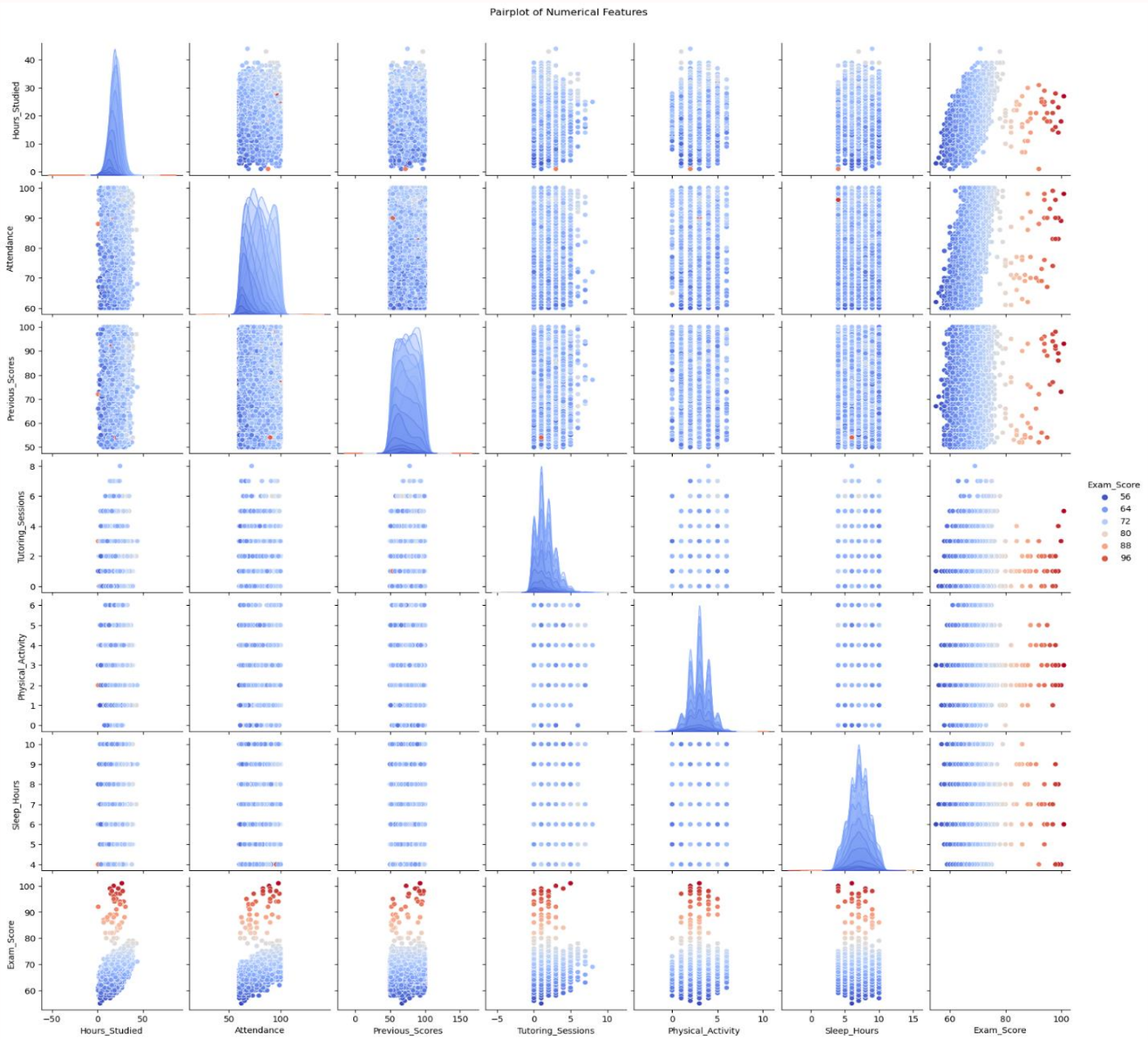
- The exam score distribution offers an overview of student performance variability, highlighting central tendencies, skewness, and potential outliers through a histogram and KDE overlay.
- The normal distribution, with most students clustered around a score of 70, suggests baseline performance and minimal outlier effects.
- This visualization underscores the need to explore factors such as attendance and study hours that may influence score variations.

EDA - Correlation Heatmap



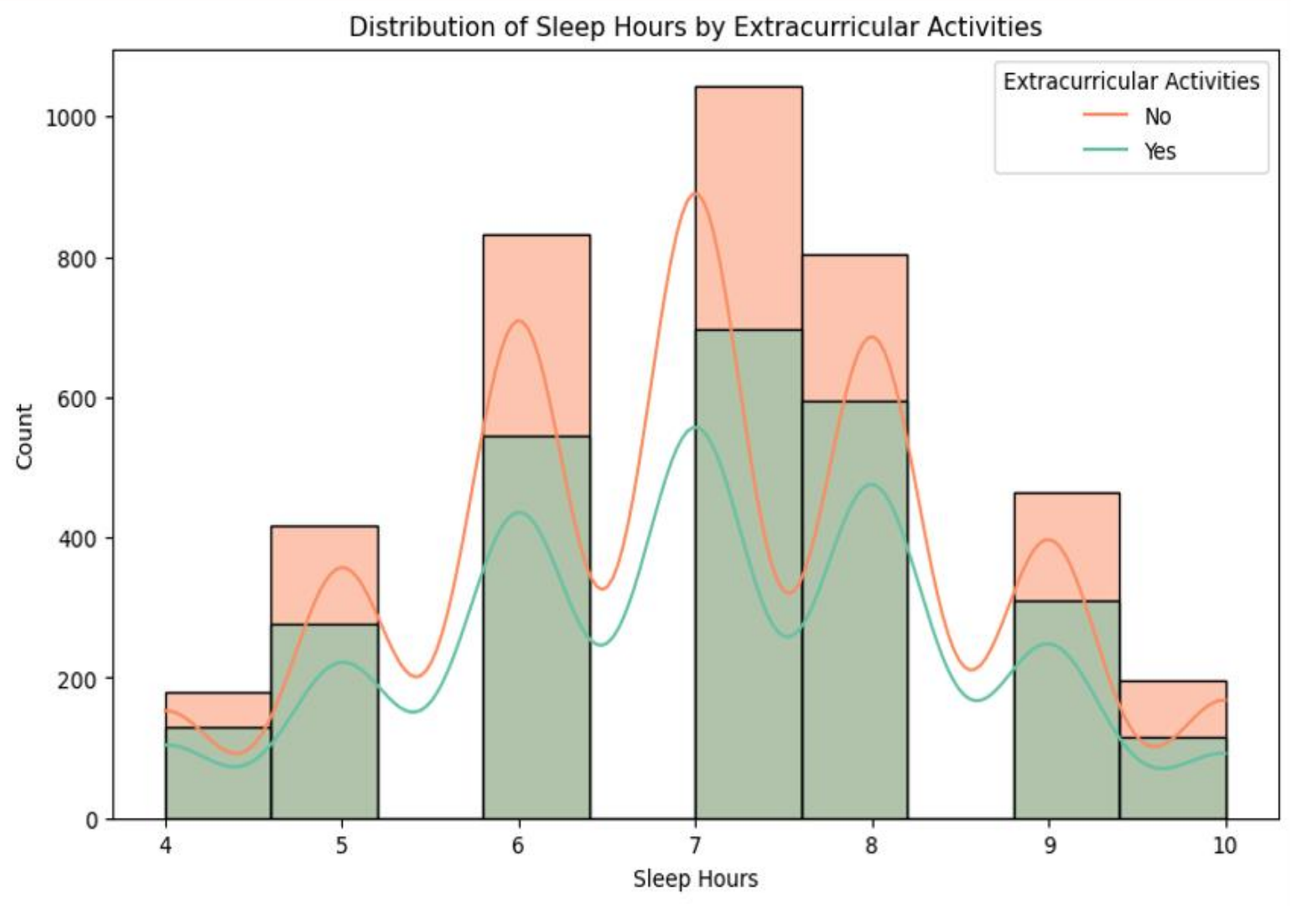
- The correlation heatmap reveals key relationships between exam scores and other numerical features like attendance, hours studied, and physical activity.
- Attendance shows the strongest positive correlation (0.58) with exam scores, followed by hours studied (0.45).
- These findings highlight the significant impact of attendance and study hours on academic performance.
- In contrast, physical activity and sleep hours show weak or negligible correlations, suggesting minimal influence on performance.
- These strongly correlated factors are ideal candidates for predictive models.

EDA - Pair Plot of Key Variables



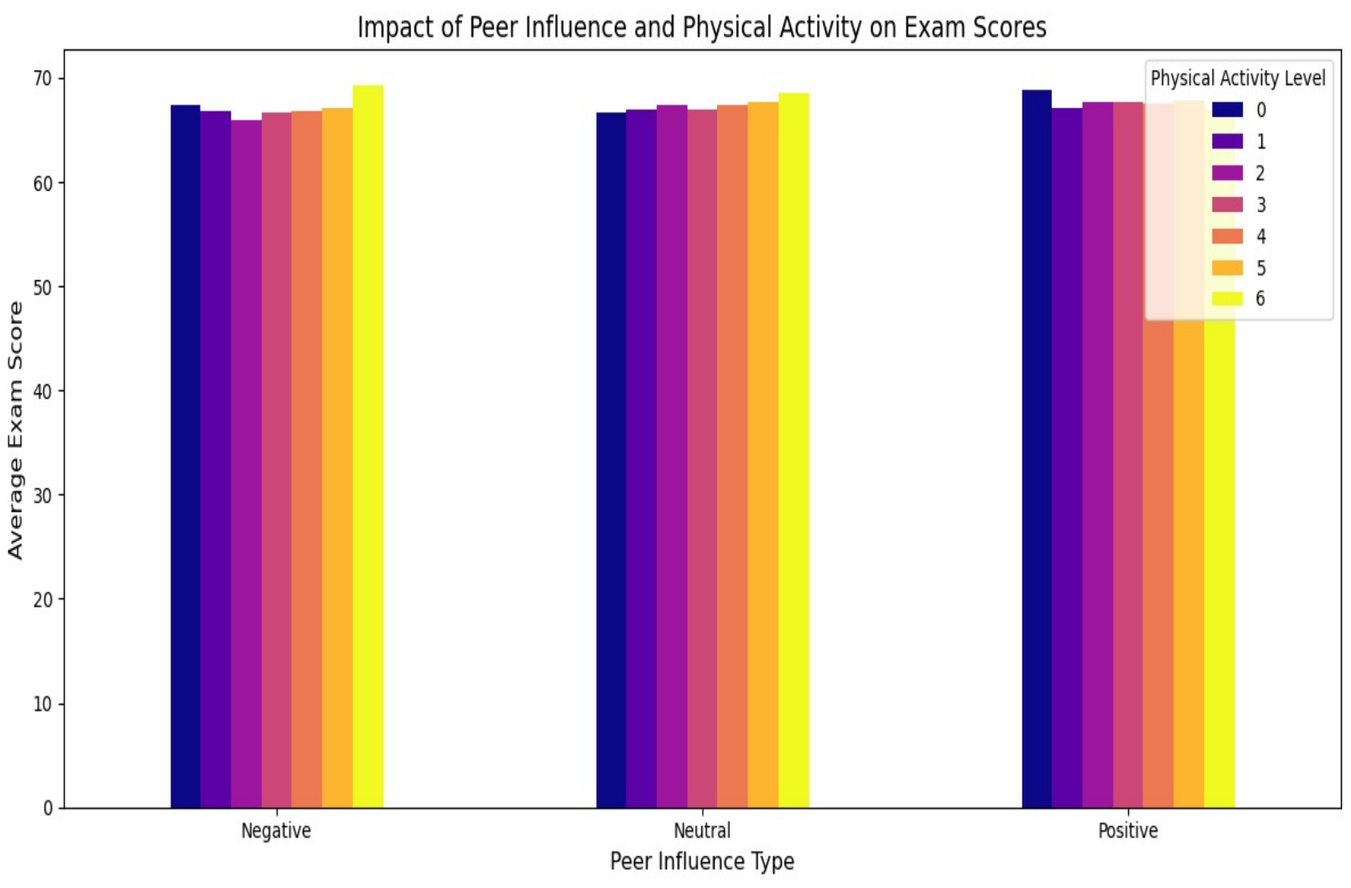
- The pair plot visualizes the interactions between variables like hours studied, attendance, and exam scores, using color gradients to indicate different score levels.
- It reveals positive relationships between hours studied, attendance, and exam scores, with higher scores clustering around greater study hours and attendance.
- These patterns confirm that attendance and study hours are strong performance indicators and essential for predictive modeling, reinforcing their importance in model training.

EDA - Relationship between sleep hours and participation in extracurricular activities among students:



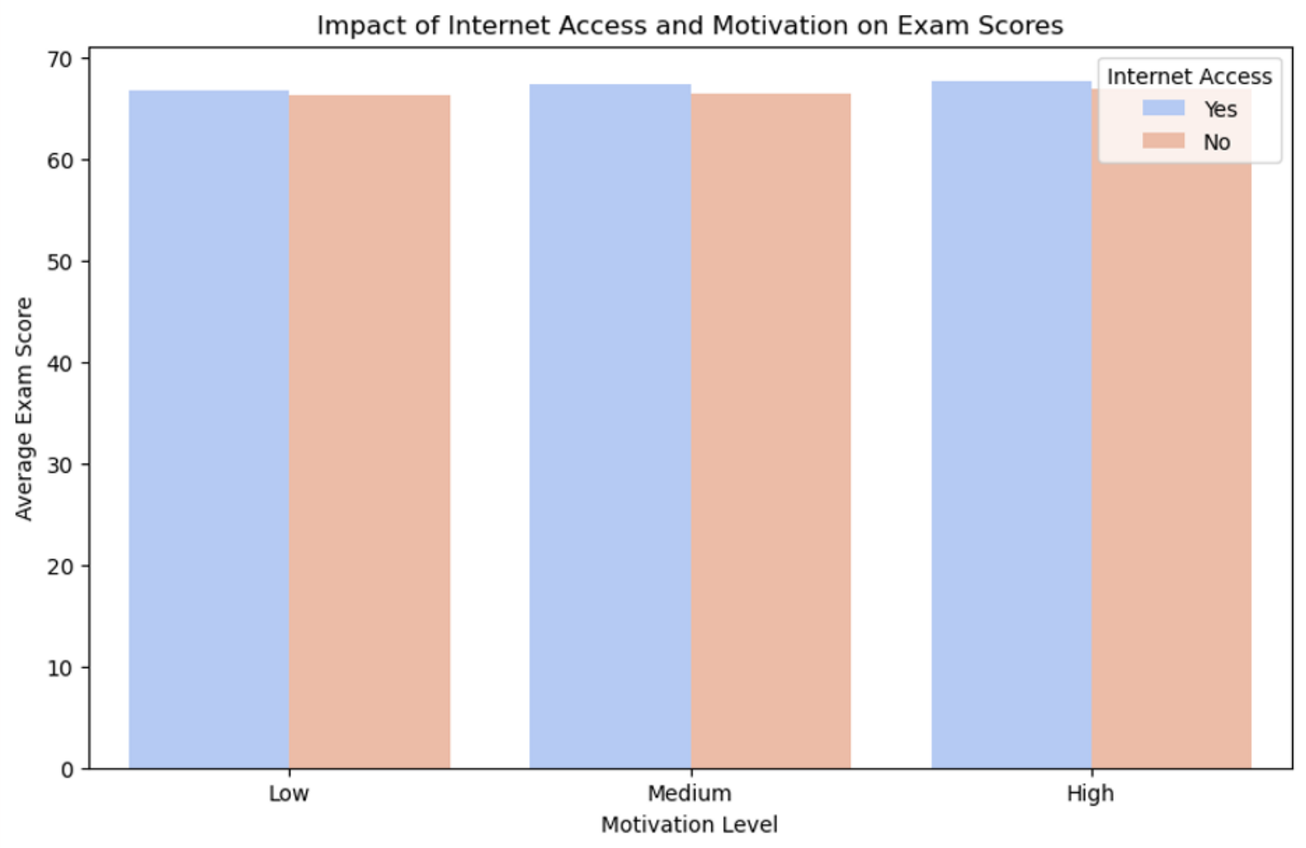
- The histogram compares sleep hours between students involved in extracurricular activities and those not involved.
- Students not in extracurricular activities tend to sleep consistently around 7–9 hours, with a peak at 7 hours.
- In contrast, extracurricular participants show more variability in sleep hours, with many sleeping between 6–7 hours.
- Both groups have fewer students with extreme sleep patterns (4 or 10 hours).
- Overall, extracurricular participation affects sleep consistency, with non-participants showing more uniform patterns, while participants experience greater variation due to balancing activities and academics.

EDA - Impact of Peer Influence and Physical Activity on Exam Scores



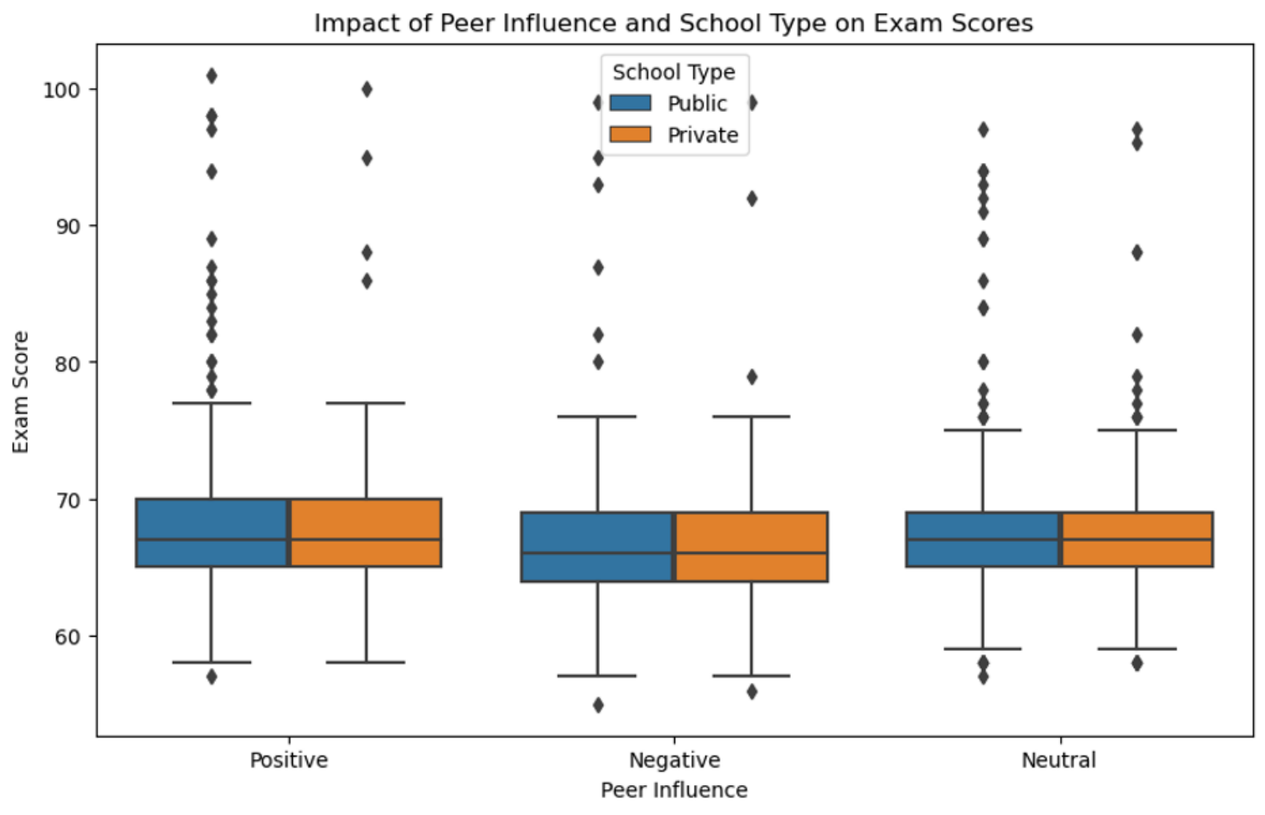
- The graph highlights the interaction between Peer Influence Types (Negative, Neutral, Positive) and Physical Activity Levels on students' exam scores.
- Students with Positive Peer Influence consistently achieve higher scores, while those with Negative Peer Influence score lower.
- Higher Physical Activity Levels (5 and 6) are linked to better performance across all peer groups, while inactivity (Level 0) correlates with the lowest scores.
- These findings stress the importance of fostering positive peer environments and promoting active lifestyles to improve academic outcomes, emphasizing the value of interventions targeting both social and physical factors.

EDA - Influence of Internet Access and Motivation Level on student performance.



- The bar chart shows the relationship between motivation, Internet access, and exam performance.
- Motivation is the primary driver of performance, with highly motivated students achieving the best scores
- Internet access has minimal impact overall but slightly benefits students with low or high motivation. Low-motivation students score 67 with Internet access vs. 66 without, while high-motivation students score 68 with Internet access vs. 67 without.

EDA - Impact of Peer Influence and School Type on Exam Scores



This Box plot explains:

1. **Impact of Peer Influence:** Positive peer influence is associated with slightly higher median exam scores compared to Negative and Neutral influences.

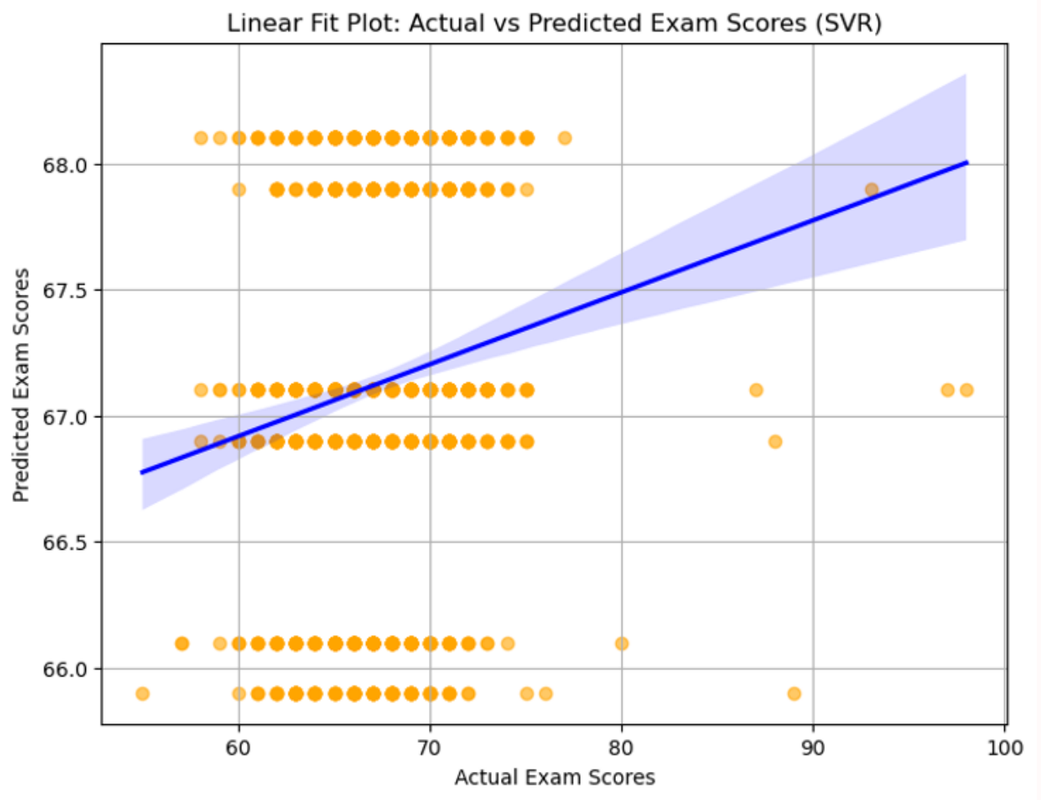
2. **School Type Comparison:** Public and Private schools show minimal differences in exam score distribution across all peer influence categories.

3. **Variability in Scores:** Negative and Neutral influences exhibit greater score variability compared to Positive influence.

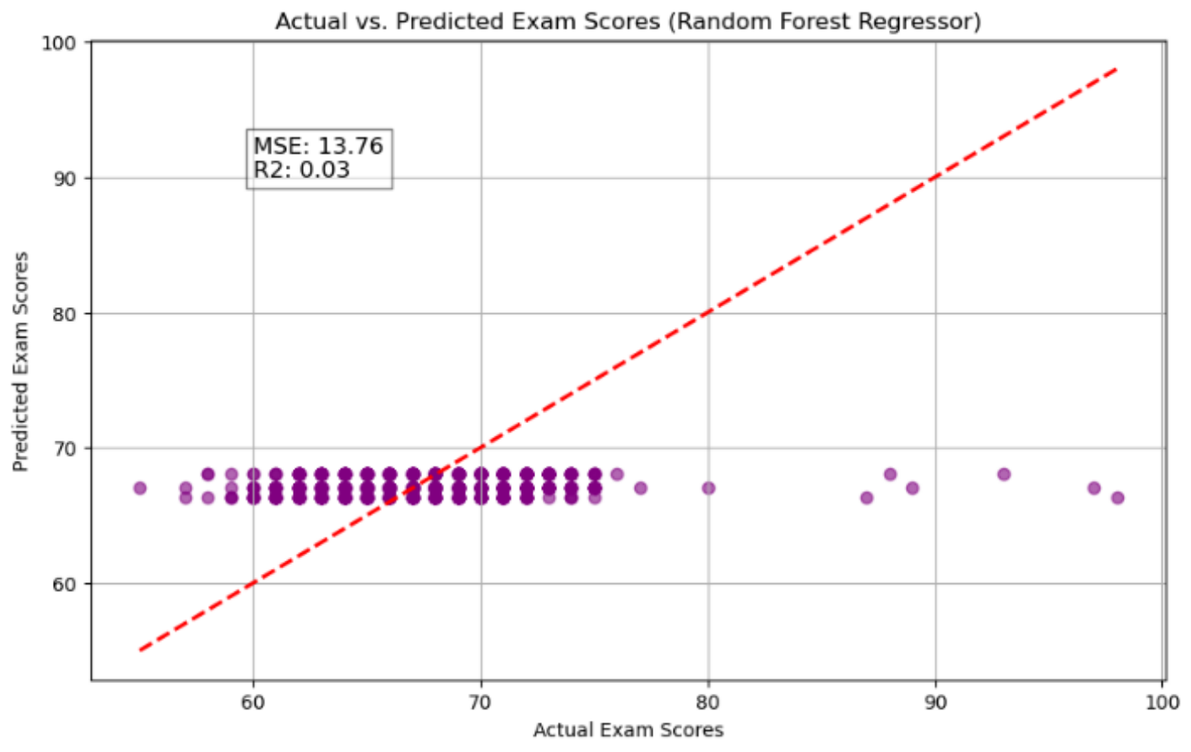
4. **Key Insight:** Peer influence plays a more notable role in performance than school type, suggesting the importance of social environments in education.

RESEARCH QUESTION 1: HOW DO EXTRACURRICULAR ACTIVITIES AND RESOURCES IMPACT ACADEMIC SUCCESS?

- The SVR model shows a positive relationship between extracurricular activities, access to resources, and exam scores.
- While the model captures some variability, the spread of points around the trend line indicates that it does not fully explain exam score differences.
- Clusters of points suggest that students with similar levels of resources and participation in extracurriculars tend to have comparable academic outcomes.
- The consistent confidence interval indicates stable but limited predictive accuracy across exam score ranges.
- These results highlight the importance of extracurricular activities and resources but suggest that additional factors are needed for a more comprehensive understanding of academic success.



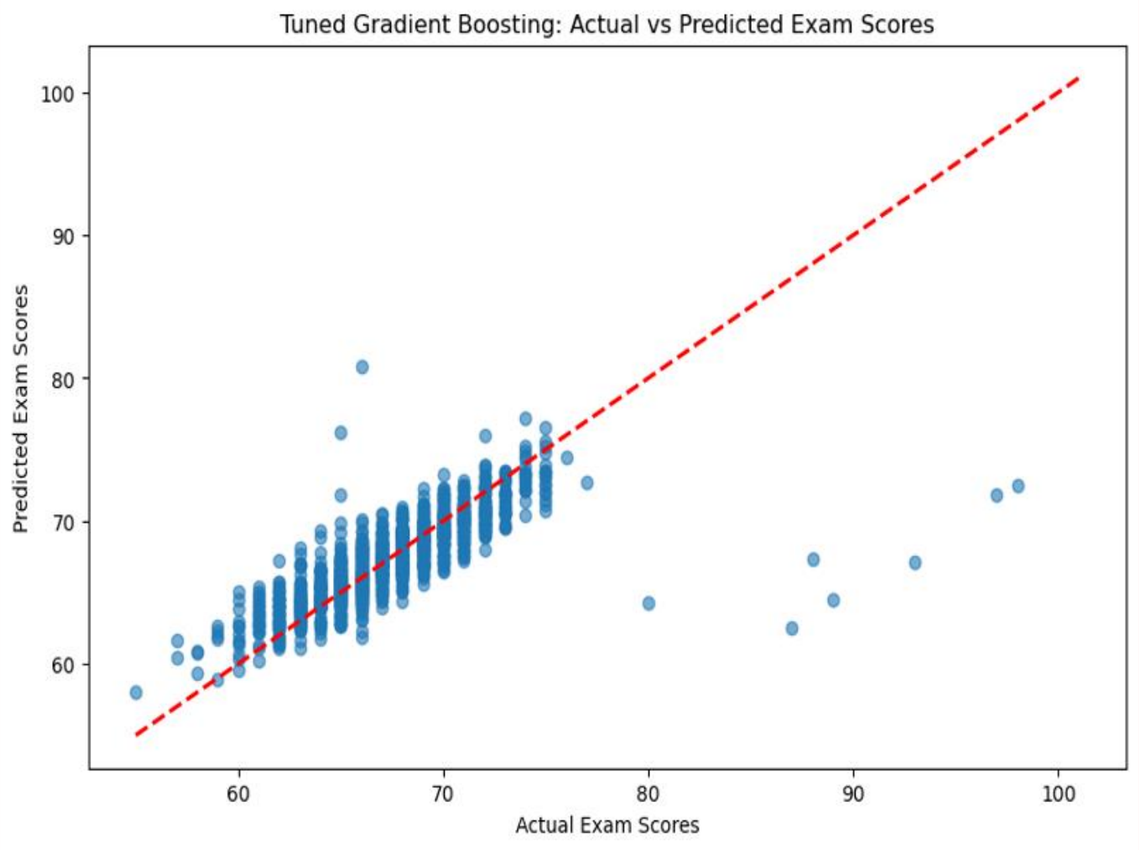
RESEARCH QUESTION 2: WHAT EFFECT DOES PARENTAL PARTICIPATION HAVE ON STUDENTS' ACADEMIC PERFORMANCE?



- The Random Forest Regressor demonstrates a reasonable ability to predict exam scores based on parental involvement, as indicated by the R^2 value and low MSE.
- Points close to the red dashed line highlight accurate predictions, while those farther away represent larger prediction errors, showing variability in the effect of parental involvement.
- The spread of points around the red line suggests that while parental involvement plays a role in academic performance, it does not fully account for all score variations.
- Dummy variables were created for Parental_Involvement levels (e.g., "Low," "Medium," "High") to enable the model to process categorical data effectively.
- The analysis indicates that parental involvement is significant but suggests that incorporating additional factors could improve the model's predictive performance for academic success.

RESEARCH QUESTION 3: HOW DOES TUNING A GRADIENT BOOSTING REGRESSOR IMPACT THE ACCURACY AND PREDICTIVE POWER IN FORECASTING EXAM SCORES?

- GridSearchCV identified the optimal hyperparameters (n_estimators, learning_rate, and max_depth), ensuring the Gradient Boosting Regressor effectively balances complexity and overfitting.
- The model achieved a low MSE, indicating that predicted exam scores are, on average, close to actual values, reflecting good accuracy.
- An R^2 value close to 1 demonstrates the model's strong ability to explain the variance in exam scores, affirming its predictive power.
- The scatter plot of actual vs. predicted exam scores shows a tight clustering of points along the red diagonal line, indicating the model's effectiveness in capturing patterns in the data.
- Deviations from the line highlight areas where predictions can be further improved through additional tuning or feature enhancements.



Conclusion

1. The study highlights critical variables like attendance, study habits, and resource accessibility as significant predictors of academic success.

2. Findings indicate that attendance and study hours have the strongest positive correlations with exam scores, aligning with previous literature on the importance of consistent effort and presence in academic settings (Kotsiantis et al., 2004; Gray et al., 2014).

3. The 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.

FUTURE WORK & RECOMMENDATIONS

1. Future research should include psychological factors (e.g., stress, self-efficacy, motivation), environmental factors (e.g., teacher quality, peer dynamics), and diverse datasets to improve model accuracy and generalizability.

2. Longitudinal studies and multimodal data (e.g., behavioral or psychometric data) can enhance analysis depth.

3. Key recommendations include prioritizing attendance, study habits, and resource accessibility through structured interventions like monitoring systems, workshops, and equitable resource distribution.

4. Schools should offer counseling services for stress management, while personalized learning paths can improve inclusivity. Parental involvement programs and policy reforms should address disparities, ensuring equitable and effective education for diverse student needs.

References

[1] Chinyoka, K., & Naidu, N. (2013). Influence of home-based factors on the academic performance of girl learners from poverty-stricken fam-ilies: 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.

Thank you!!