

Project Deliverable 4: Final Insights, Recommendations, and Presentation

Bishal Thapa

MSCS-634 Advanced Big Data and Data Mining

Dr. Satish Penmatsa

Dec 10, 2025

### **Explanation of the dataset used and why it was chosen.**

For the final project for the Advanced Big Data and Data Mining class, I have selected a dataset focused on student performance and the various factors that influence academic outcomes. As a student, I am particularly interested in understanding which elements have the most significant impact on academic success. This was the primary reason I was drawn to analyzing this dataset. Apart from that, this dataset was chosen for several other reasons. First, it contains 20 columns, predominantly numerical, allowing for a wide range of calculations and analyses to explore the relationships between different factors and student performance. Additionally, the dataset includes records for over 5,000 students, exceeding the minimum requirement for this project and providing a robust sample size for statistical reliability.

Notably, the dataset encompasses a diverse range of academic, personal, and socioeconomic variables, including study hours, attendance, parental involvement, motivation levels, and family income. This variety enables the examination of complex patterns and interactions that affect exam scores. The dataset enables the application of multiple data mining techniques, including correlation analysis, classification, regression, and clustering, to identify key predictors of academic achievement. Moreover, the combination of numerical and categorical attributes offers valuable opportunities for practicing essential data preprocessing steps, feature selection, and model evaluation. Overall, this dataset serves as an ideal, real-world resource for effectively learning and applying data mining concepts.

### **Key insights from data preprocessing, EDA, and feature engineering.**

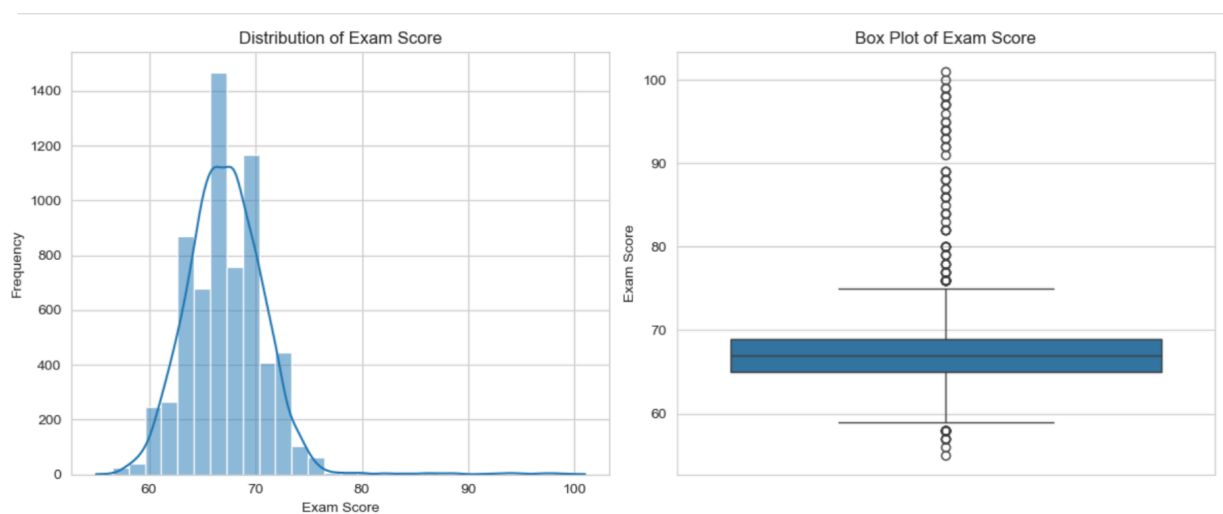
The dataset used for this analysis consists of 6,607 records and 20 features, providing a comprehensive view of factors influencing student exam performance. Initially, the dataset

contained 235 missing values, with the highest percentages in Parental\_Education\_Level (1.36%), Teacher\_Quality (1.18%), and Distance\_from\_Home (1.01%). These missing values were imputed using the mode for each respective feature, resulting in a complete dataset with no remaining missing values and maintaining the original shape of (6,607, 20).

Outlier detection was performed using the IQR method. Tutoring\_Sessions exhibited the highest proportion of outliers (430, 6.51%), followed by Exam\_Score (104, 1.57%) and Hours\_Studied (43, 0.65%). Features such as Attendance, Sleep\_Hours, Previous\_Scores, and Physical\_Activity showed no outliers. In total, 553 rows contained at least one outlier, representing 8.37% of the data. Exam scores ranged from 55 to 101, with a mean of 67.24 and a standard deviation of 3.89, indicating moderate variability across students.

**Figure 1**

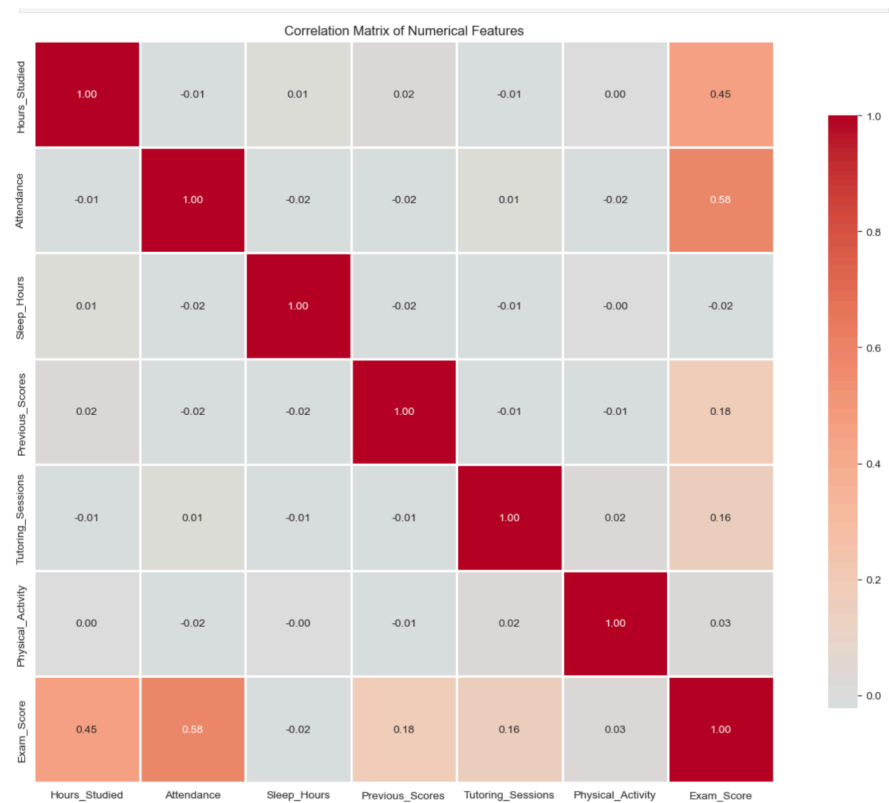
*Exam score distribution and Box Plot*



Correlation analysis revealed that attendance (0.581) and hours studied (0.445) are the strongest predictors of exam performance, followed by previous scores (0.175) and tutoring sessions (0.157). Other factors, including physical activity and sleep hours, showed minimal

correlation. Categorical features also displayed modest effects: higher parental involvement and motivation levels were associated with slightly higher scores, while students without learning disabilities performed marginally better.

**Figure 2**  
*Correlation Matrix of Numerical Features*



Overall, the analysis highlights that academic engagement and study habits are the most influential factors for exam success. The careful handling of missing values and identification of outliers ensures data quality, providing a solid foundation for predictive modeling and further exploration of factors impacting student performance.

**Results from regression:**

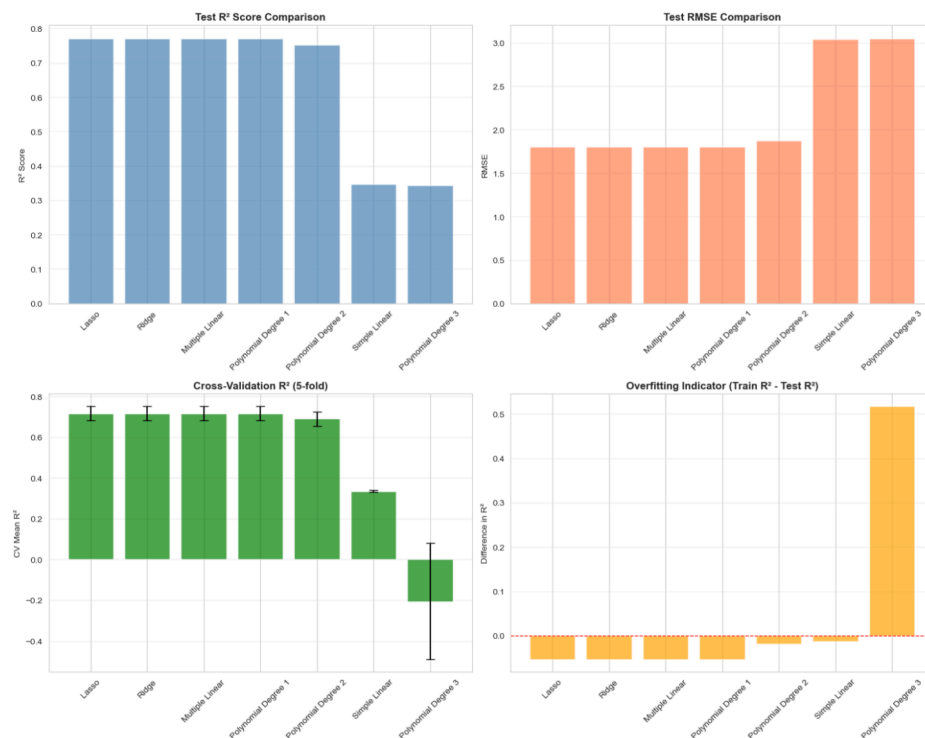
Before building the regression models, the dataset was cleaned by filling in a small number of missing values and checking for duplicate entries. The objective of the analysis was to predict

each student's Exam\_Score. To achieve this, several regression approaches were built and compared, including simple linear, multiple linear, and polynomial models. A major part of the process involved feature engineering. Starting from the original 20 variables, five additional features were created to enhance the models and provide more meaningful inputs. This resulted in a total of 25 features used for analysis. The new engineered features were:

- Study\_Effectiveness:  $\text{Hours\_Studied} * \text{Attendance} * \text{Motivation\_Level}$
- Preparation\_Score:  $\text{Previous\_Scores} + (\text{Tutoring\_Sessions} * 5)$
- Health\_Wellness:  $\text{Sleep\_Hours} * \text{Physical\_Activity}$
- Resource\_Advantage:  $\text{Access\_to\_Resources} * \text{Family\_Income}$
- Parental\_Support:  $\text{Parental\_Involvement} * \text{Parental\_Education\_Level}$

**Figure 3**

*Comparison of different Regression Models*



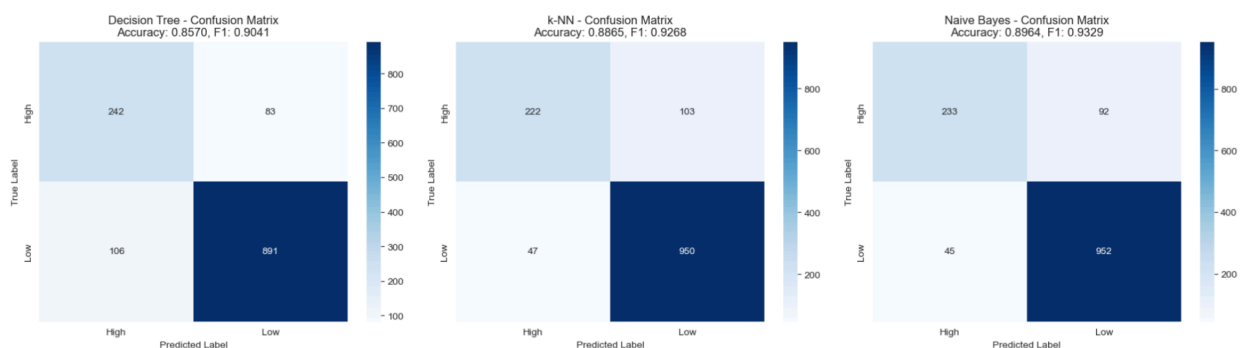
The model results highlighted that the number and quality of features were the strongest drivers of performance. A model using only one feature performed poorly, while models using all 24 usable features achieved a dramatic 121% improvement in accuracy. Surprisingly, the most complex model, Polynomial Regression (Degree 3), performed no better than the simplest approach, likely due to overfitting.

The best-performing models included Lasso, Ridge, and standard Multiple Linear Regression, all of which explained roughly 77% of the variance in exam scores and produced predictions with an average error of only  $\pm 1.8$  points. Because their performance was nearly identical, the regularization methods (Lasso and Ridge) did not provide a significant advantage over standard multiple regression. Overall, the regression models results show that no single factor strongly determines a student's exam score; instead, the score is influenced by a combination of many variables. This is why multiple linear regression greatly outperformed single-factor models. Among the top models, Lasso Regression stands out as a strong and reliable choice, capable of explaining about 77% of the variation in exam outcomes while maintaining high predictive accuracy.

### **Result for classification, clustering, and association rule mining.**

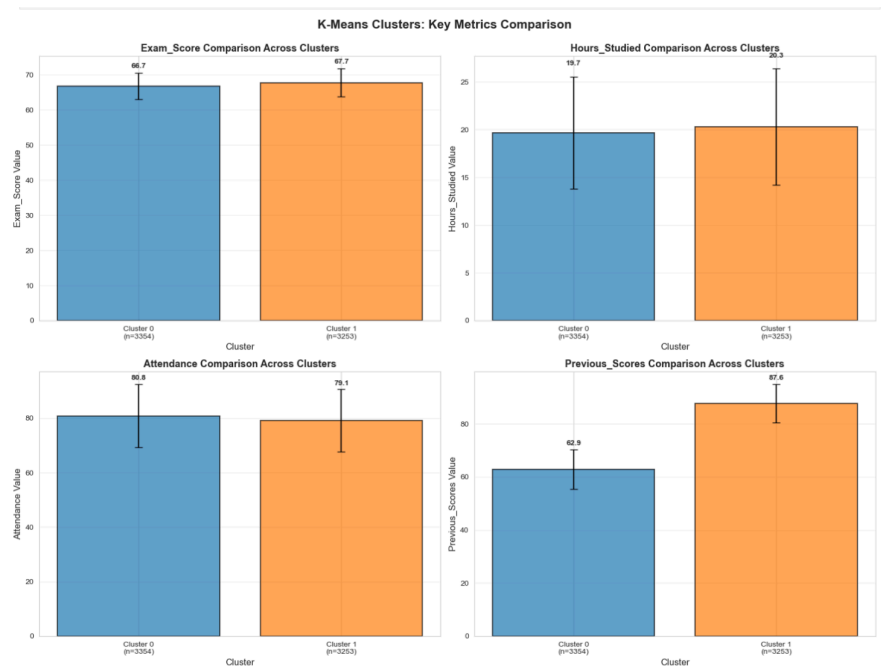
**Figure 4**

*Confusion Matrix of Different Classification models:*



The Tuned Naïve Bayes model emerged as the top-performing classification model, achieving an accuracy of 0.9281 and an F1 score of 0.9536. This makes it a reliable tool for predicting whether a student is likely to achieve high or low exam scores. By analyzing historical or current student data, such as past scores, attendance, and demographics, schools can identify students at risk of underperformance well before exams. This predictive capability enables targeted interventions, including remedial classes, tutoring, counseling, or parent engagement, allowing institutions to shift from reactive to proactive support and maximize the impact of academic assistance programs.

**Figure 5**  
*K-means clustering details*

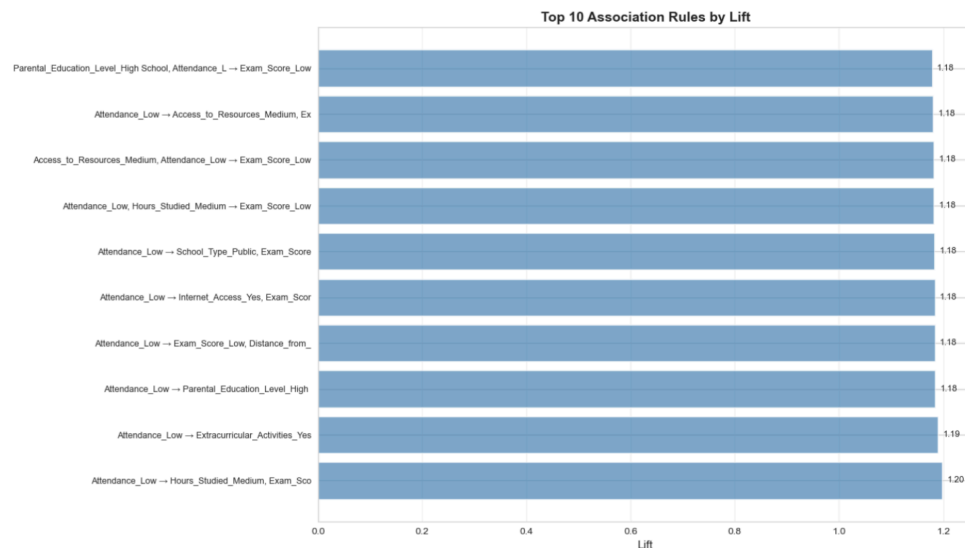


K-Means clustering revealed two distinct student groups primarily differentiated by previous scores and family income. Cluster 0, with lower scores and lower family income, contains only 20% high performers, suggesting the need for foundational skill support and resource assistance. Cluster 1, with higher scores and medium family income, performs better

overall, but 70.7% of students still score low, indicating the need for enrichment, motivation, and strategies to maintain momentum. These clusters provide a data-driven basis for resource allocation, ensuring equitable distribution of scholarships, tutoring, and counseling.

**Figure 6**

### *Association Rule Mining*



Association rule mining highlighted strong relationships between factors and performance. Rules such as  $\text{Attendance\_Low} \Rightarrow \text{Exam\_Score\_Low}$  (98.16% confidence) emphasize the critical role of attendance, while rules related to resource access reveal that limited availability, beyond internet connectivity, significantly affects performance. These insights support the implementation of policies for attendance monitoring and improved resource accessibility, enabling schools to address gaps and optimize student outcomes effectively.

### **Practical recommendations based on your findings.**

The study provides several actionable insights for improving student performance. Predictive models, such as the Tuned Naïve Bayes classifier, can identify students at risk of low exam scores, allowing schools to implement early interventions like mentoring, tutoring, or

supplemental classes. Linear regression analysis further complements these findings: models including Lasso, Ridge, and standard Multiple Linear Regression explained approximately 77% of the variance in exam scores, with predictions averaging an error of only  $\pm 1.8$  points. These results highlight that no single factor determines performance; rather, multiple variables collectively influence outcomes. Consequently, interventions should address several dimensions simultaneously, including past performance, attendance, resource access, and student effort.

Clustering and association rule analyses reveal which students require additional support, particularly those with lower previous scores or from lower-income families. Schools can use this information to allocate resources equitably, providing free tutoring, learning materials, or technology access to those who need it most. Attendance emerged as a critical predictor, with low attendance strongly linked to poor performance. Strict monitoring and prompt interventions, such as automated alerts or counseling, can help mitigate this risk.

Overall, these findings suggest that combining predictive classification, regression analysis, and data-driven clustering allows schools to make informed decisions. By targeting interventions where they are most needed and considering multiple contributing factors, institutions can improve learning outcomes, ensure equitable resource distribution, and support every student in achieving their academic potential.

### **Ethical considerations related to the project.**

The use of student data for predictive modeling raises important ethical concerns related to privacy, fairness, and bias. The dataset includes sensitive information such as family income, parental education, gender, school type, and previous scores, which could inadvertently reveal inequalities or be misused if not properly protected. Without careful handling, models may

reinforce existing disparities, for example by systematically underestimating the potential of students from lower-income families or flagging them disproportionately as “at-risk.”

Attendance, access to resources, and parental involvement are particularly sensitive predictors that could reflect broader socioeconomic factors rather than individual effort.

To address these concerns, schools should obtain informed consent from students and parents, anonymize or pseudonymize sensitive data, and ensure secure storage. Fairness can be promoted by regularly auditing model outcomes across demographic groups to detect bias and by incorporating fairness-aware methods such as reweighting or balanced sampling. Additionally, predictive outputs should be combined with human judgment so that educators contextualize the results before making decisions about interventions. By adopting these measures, schools can use predictive analytics responsibly, providing targeted support where it is needed most while minimizing the risk of discrimination or unfair treatment.

### **Summary**

To sum up, using this dataset we were able to analyze various factors affecting student performance. The Tuned Naïve Bayes model accurately predicted students at risk of low exam scores. Linear regression showed that multiple variables together explained about 77% of exam score variance. Clustering and association rules highlighted the influence of socioeconomic factors, attendance, and resource access. We also learned the importance of fairness, ensuring interventions do not reinforce existing inequalities.

## **References:**

- Han, J., Pei, J., & Kamber, H. (2018). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann. <https://doi.org/10.1016/C2015-0-01625>
- Wu, D. (2020). *Data mining with Python*. Routledge. <https://doi.org/10.1201/9780429285392>
- Kaushik, P., Sharma, K., Mahawar, M. K., Wasim, J., Dey, G., & Nibiya, S. A. (2024, December). Ethical Considerations in Data Mining and Analytics. In 2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N) (pp. 1516-1521). IEEE.
- Wu, D. (n.d.). *Student Success: Factors & Insights* [Data set]. Kaggle. <https://www.kaggle.com/datasets/anassarfraz13/student-success-factors-and-insights>