MAIM Final Project

Student Performance Insights: From Cleaning to Clustering to Classification

Habiba Abdullah Said Hammad -2203167

Abstract:

This study investigates students' performance in two subjects, Mathematics and Portuguese, using behavioral and demographic data from 649 students. The analysis explores factors influencing academic outcomes, with particular focus on study time, prior failures, parental education, and absences. Results indicate that attendance plays a critical role in determining final grades, while prior failures and parental education further contribute to performance disparities. Cluster analysis revealed four main student groups: Diligent Students, Quiet Achievers, Average Performers, and Skippers. Based on these findings, several intervention strategies are proposed, including automated attendance monitoring, tailored workshops on time management and study skills, family educational support initiatives, and academic recovery programs for students with prior failures. Furthermore, automated classification tools could be developed to assign students to risk clusters and enable personalized interventions. Overall, this project highlights the importance of attendance, family background, and study habits in shaping students' academic success, while demonstrating how data-driven approaches can guide early detection and targeted support.

Problem and Value:

Problem

Students' performance is influenced by multiple behavioral, social, and demographic factors. However, schools often lack systematic ways to identify at-risk students early. As a result, warning signs such as high absence rates and prior academic failures—both strong predictors of poor outcomes—are frequently overlooked until it is too late to intervene effectively. This leads to academic trajectories that could have been improved or even reversed with timely support.

Value

By analyzing real student data, this project identified the factors that most strongly predict academic success or failure. These insights enable schools to design targeted interventions such as attendance monitoring, study workshops, and family support programs. Early identification not only saves resources but also helps students before they fall behind, improving overall academic outcomes. Moreover, the data-driven approach provides objective, evidence-based support for educational policy and decision-making, ensuring that interventions are both strategic and impactful.

Dataset:

Source:

The dataset used in this study is the *Student Performance dataset* collected from two Portuguese secondary schools and made available through the UCI Machine Learning Repository. It contains academic records and socio-demographic information for a total of 649 students across two subjects: Mathematics (395 students) and Portuguese (654 students). For this analysis, the two subject datasets were combined to focus on factors influencing overall academic outcomes.

Schema:

The dataset includes 649 rows (students) and 33 attributes, covering:

- **Demographic Factors**: age, gender, family background, parental education, and parental occupation.
- **Behavioral Factors**: travel time, extracurricular activities, daily alcohol consumption, study time, and absences.
- **Academic Performance**: prior grades (G1 and G2), past failures, and final grade (G3).

Both categorical and numerical attributes are present, which require preprocessing before model training.

Limits:

The dataset has several limitations. First, it was collected from only two Portuguese schools, which may reduce its generalizability to other contexts or countries. Second, some attributes (such as alcohol consumption or free time) are self-reported, introducing potential bias or inaccuracy. Finally, the strong correlation between prior grades (G1, G2) and the final grade (G3) posed a risk of data leakage, requiring careful handling during the modeling process.

Methods:

Data Pre-Processing:

After inspecting the dataset metadata and variables, the data was saved in CSV format to ensure reproducibility. Duplicate rows and missing values were checked, with both found to be absent. Outlier detection was then performed using the Interquartile Range (IQR) method, supported by boxplot visualizations. The feature *absences* showed the most extreme outliers, which were addressed through capping: values above 30 were replaced with the 95th percentile threshold. This reduced the influence of unrealistic absence values while preserving valid variation. Other variables, including *family relationships*, *free time*, and *daily alcohol consumption*, also displayed outliers; however, these were not considered irrational and were therefore retained. The feature *failures*, an ordinal variable (0–3), appeared unusual in boxplots but was consistent with its categorical nature. Outliers were also checked in the target variable (final grade), which was confirmed to lie within the valid 0–20 range.

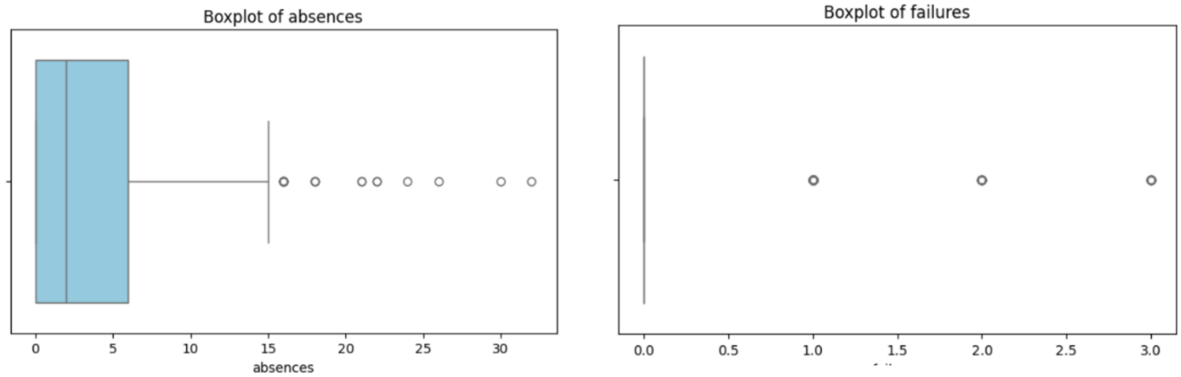

FIGURE 2

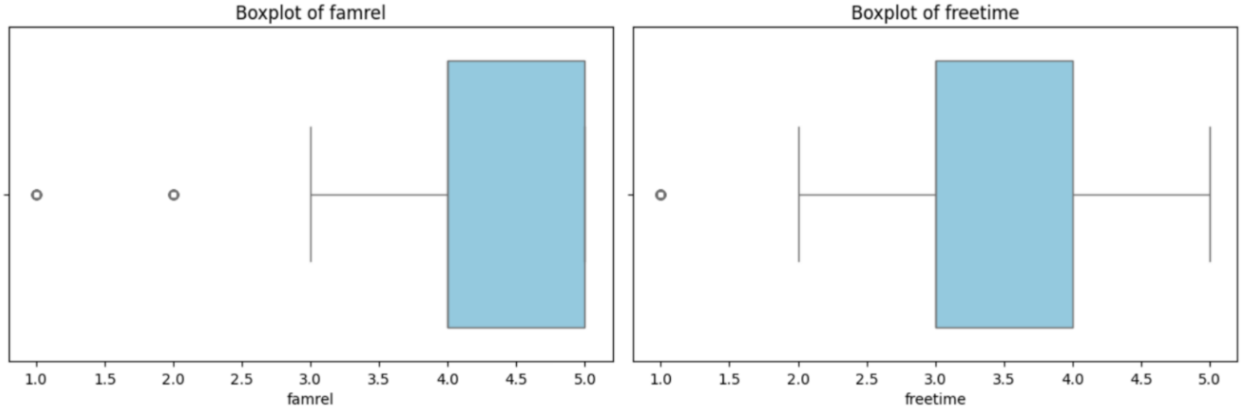

FIGURE 3

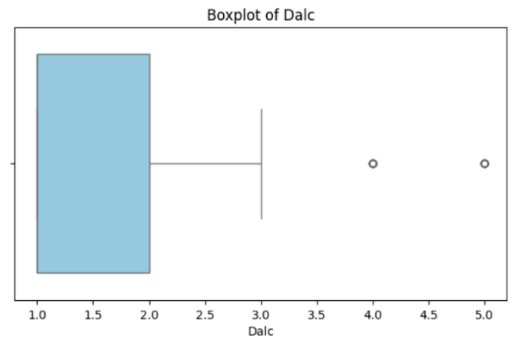

FIGURE 4

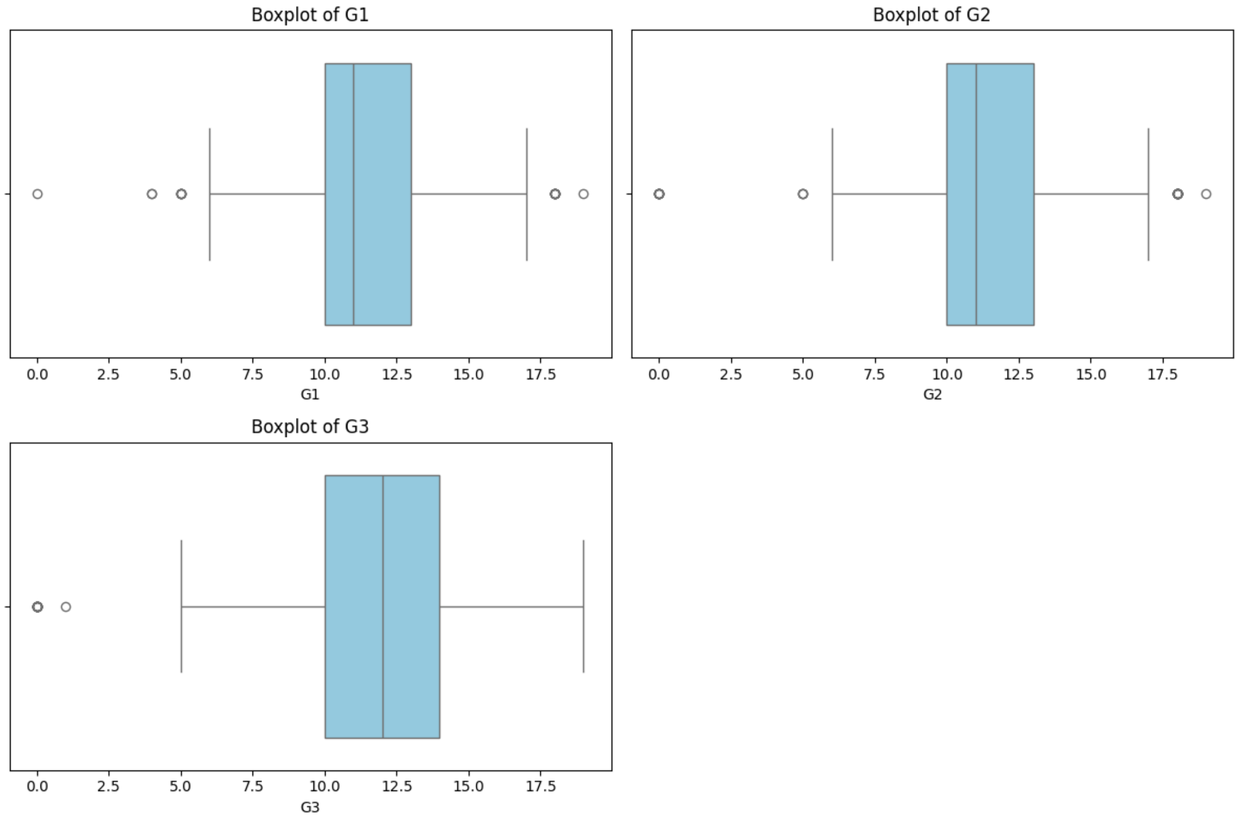

FIGURE 5

Data transformation steps included one-hot encoding of categorical and binary features and standardization of numerical variables using Standard Scaler. Feature engineering was also performed:

(1) attendance rate was calculated to replace raw absence values.

(2) students were grouped into three risk levels based on attendance.

(3) the average grade across G1–G3 was computed as an additional feature.

To address data leakage risks from prior grades, two datasets were created: df_with (including prior grades G1 and G2) and df_without (excluding prior grades). Corresponding target variables were defined for both classification (risk group) and regression (final grade prediction).

EDA and Visualization:

Exploratory analysis was conducted to better understand the dataset and the relationships among variables. Descriptive statistics were first generated for both numerical and categorical features. A correlation heatmap was then constructed to examine linear associations between numerical variables. The results revealed a strong positive correlation between the final grade (G3) and the prior grades (G1 and G2), confirming the risk of data leakage if these features were included in modeling. A strong negative correlation was also observed between past failures and final grades, indicating that students with more prior failures tended to achieve lower outcomes. In contrast, study time showed only a weak positive correlation with final grades, while absences displayed very weak or no correlation.

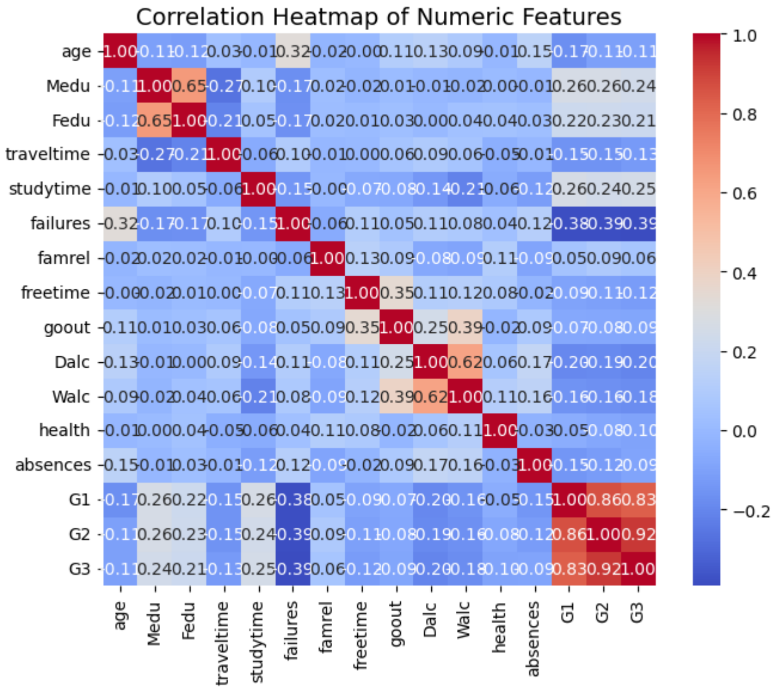

FIGURE 6

Group comparisons were also performed across several dimensions, including study time, absences, school support, failures, sex, and lifestyle score. The most significant findings were observed for absences and failures, both of which showed clear associations with academic performance. Other comparisons, such as study time and sex, revealed only weak or negligible effects.

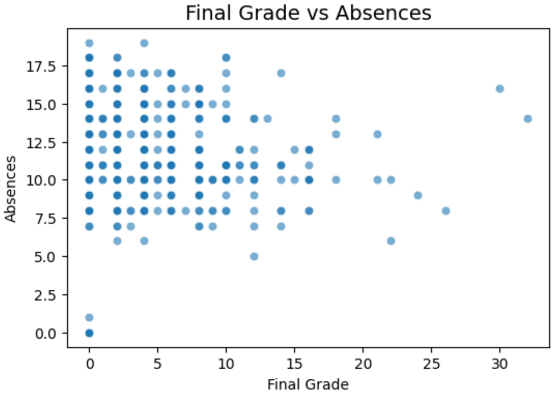

FIGURE 7

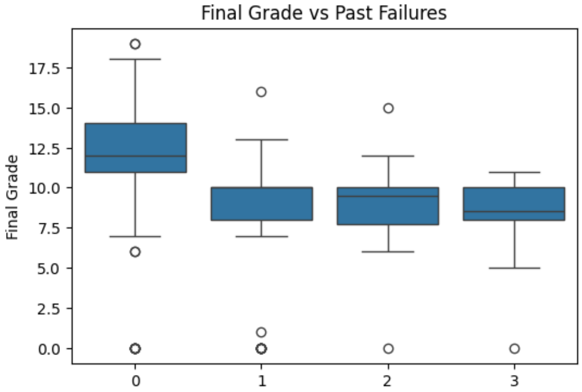

FIGURE 8

To further illustrate the data distribution and highlight potential imbalances, visualizations were generated for final grades, absences, and study time. These plots provided a clearer picture of how student outcomes were spread across the dataset and reinforced the importance of considering absence rates and prior failures as key indicators of risk.

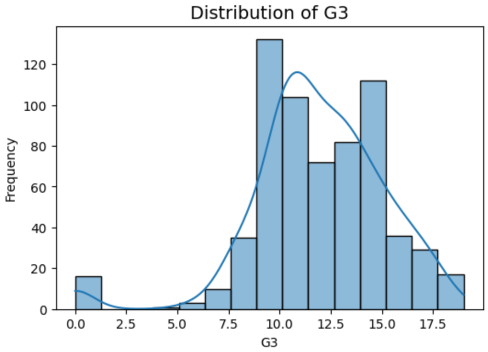

FIGURE 9

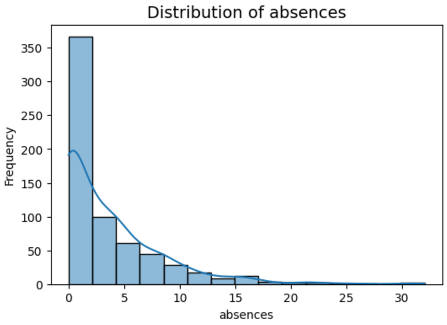

FIGURE 10

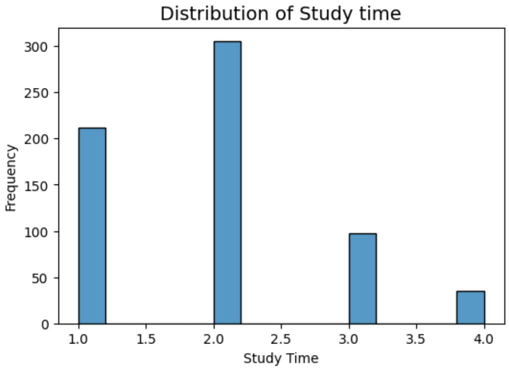

FIGURE 11

Unsupervised Learning:

To identify distinct profiles of students, K-means clustering was applied using the **df_without** dataset, which excluded academic features to prevent leakage from prior grades. The optimal number of clusters was evaluated using both the Elbow Method and the Silhouette Score. The results suggested different solutions: the elbow method indicated six clusters, while the silhouette score suggested four. To resolve this discrepancy, clustering was performed with both K=4 and K=6, and the results were compared.

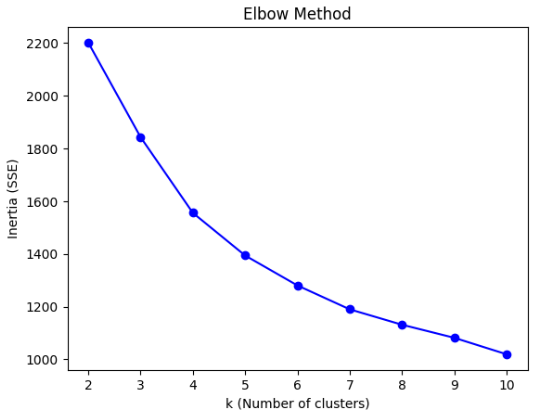

FIGURE 11

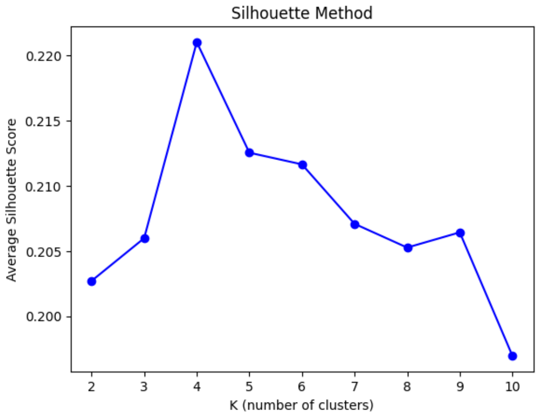

FIGURE 12

The six-cluster solution produced overlapping and repeated groups, reducing interpretability. In contrast, the four-cluster solution yielded more distinct and meaningful student profiles. Based on this, the four-cluster solution was selected. For each cluster, the average final grade was calculated using group-by aggregation, and the results were visualized. This analysis revealed four distinct student groups with varying performance levels, which later informed the identification of risk categories such as “Diligent Students,” “Quiet Achievers,” “Average Performers,” and “Skippers.”

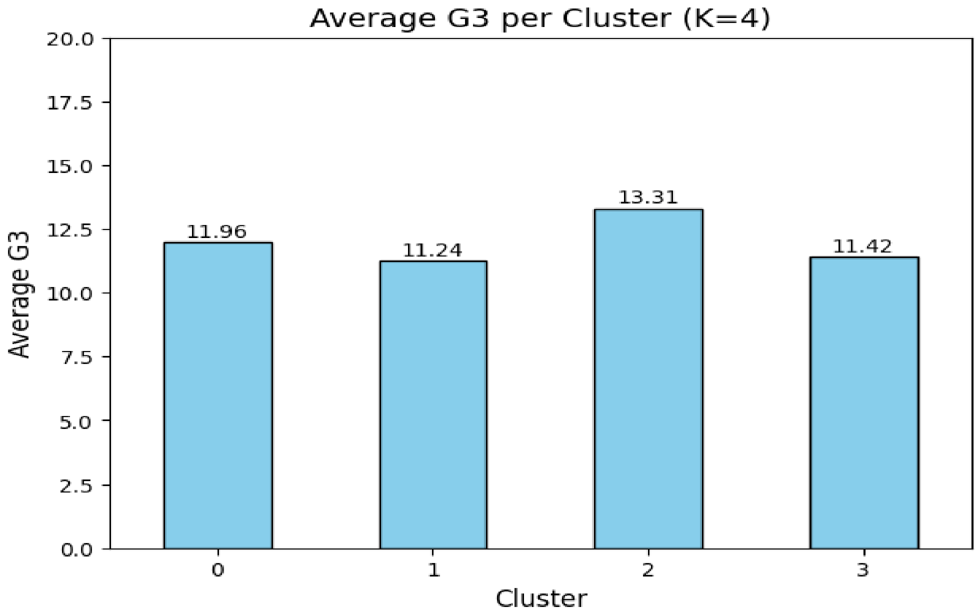

FIGURE 13

Supervised Learning:

For the supervised learning tasks, the target variable was defined as the three-class risk group created during data transformation. Three classification algorithms were implemented: **Logistic Regression**, **Decision Tree**, and **Random Forest**. Each model followed the same pipeline, which consisted of:

1. Splitting the dataset into training (80%) and testing (20%) subsets.
2. Performing hyperparameter tuning using **Grid Search** with cross-validation to select the best parameters.
3. Training the optimized model and evaluating performance using multiple classification metrics.

To address potential data leakage, models were trained separately on two datasets: **df_with** (including prior grades) and **df_without** (excluding prior grades). A helper function was used to streamline training, generate evaluation metrics, and store results for comparison.

The three models were then compared to assess predictive performance and interpretability. Detailed results, including metrics and model-specific insights, are reported in the Results section.

Results:

Table 1 summarizes the key performance metrics (Accuracy and Macro F1) across the three models. As expected, models trained on **df_with** (which includes prior grades) achieved substantially higher performance, with both Logistic Regression and Decision Tree reaching an accuracy of 92.3% and strong F1 scores. In contrast, when prior grades were excluded (**df_without**), performance dropped significantly across all models. Random Forest achieved the best accuracy (66.2%) in this setting, though the overall F1 scores remained modest, reflecting the difficulty of classification without strong academic predictors

Dataset	**Model**	**Accuracy**	**Macro CV score**	**Macro Precision**	**Macro Recall**	**Macro F1**
df_with	Logistic Regression	0.923	0.905	0.946	0.863	0.891
df_with	Random Forest	0.915	0.918	0.931	0.854	0.886
df_with	Decision Tree	0.923	0.912	0.936	0.870	0.899
df_without	Logistic Regression	0.600	0.486	0.475	0.426	0.435
df_without	Random Forest	0.662	0.415	0.721	0.397	0.387
df_without	Decision Tree	0.546	0.495	0.425	0.425	0.424

Table (1)

Table 2 presents the best hyperparameter configurations selected via grid search for each model. These tuned values helped optimize model performance and ensured fair comparisons across algorithms.

Dataset	Model	Best Params
df_with	Logistic Regression	{'C': 100}
df_with	Random Forest	{'max_depth': None, 'min_samples_split': 2, 'n_estimators': 100}
df_with	Decision Tree	{'criterion': 'gini', 'max_depth': 5, 'min_samples_split': 10}
df_without	Logistic Regression	{'C': 100}
df_without	Random Forest	{'max_depth': None, 'min_samples_split': 5, 'n_estimators': 100}
df_without	Decision Tree	{'criterion': 'entropy', 'max_depth': None, 'min_samples_split': 10}

Table (2)

Model Interpretability

Logistic Regression coefficients were plotted for each risk group, providing insights into how demographic, behavioral, and family-related factors influenced classification outcomes. For example, features such as failures and study time showed clear directional effects on the predicted risk category.

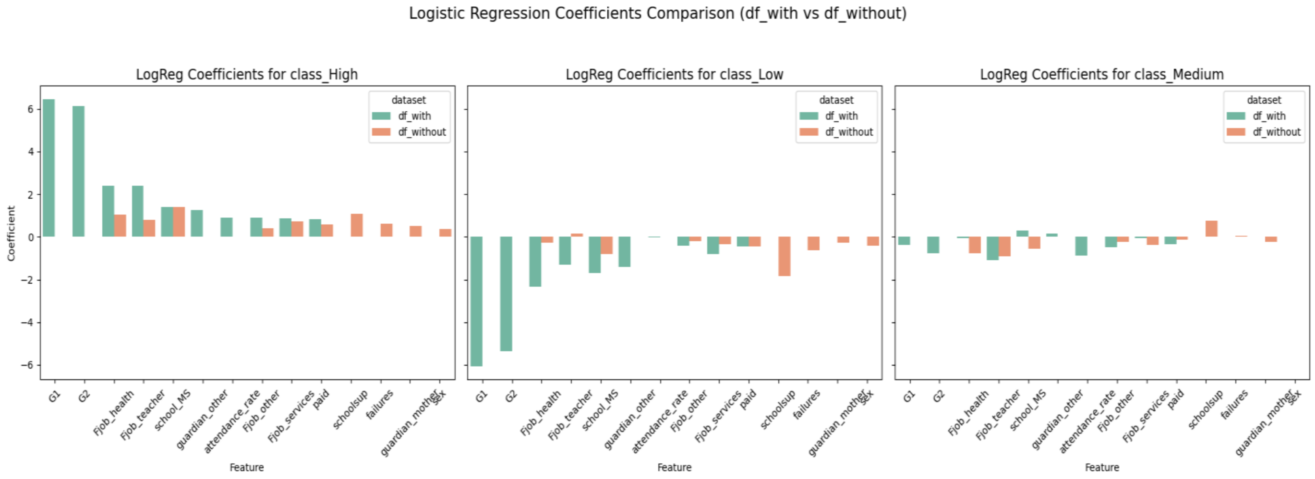

FIGURE 14

Random Forest feature importance highlighted key predictors such as failures, study time, and parental education, aligning with findings from exploratory analysis.

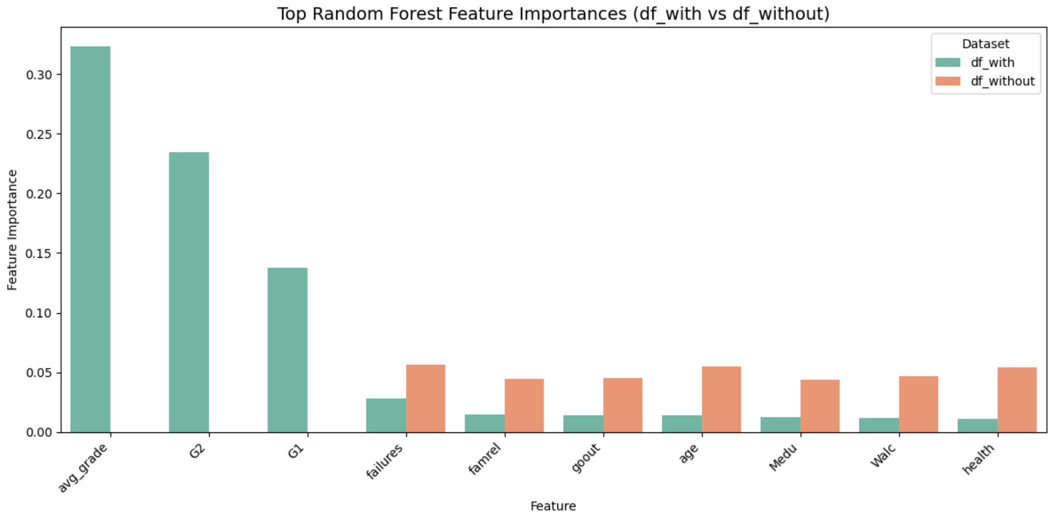

FIGURE 15

Decision Trees offered intuitive rule-based splits, though their simplicity also limited generalizability compared to ensemble methods.

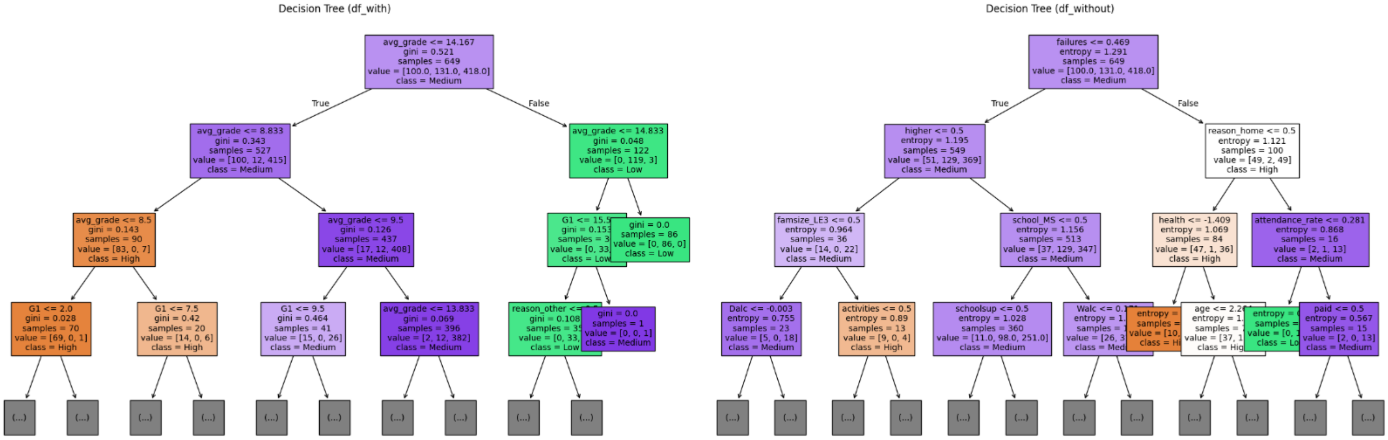
Decision Tree (df_with)
avg_grade <= 14.167 gini = 0.521 samples = 649 value = [100.0, 131.0, 418.0] class = Medium
True
False
avg_grade <= 8.833 gini = 0.343 samples = 527 value = [100, 12, 415] class = Medium
avg_grade <= 14.833 gini = 0.048 samples = 122 value = [0, 119, 3] class = Low
avg_grade <= 8.5 gini = 0.143 samples = 90 value = [83, 0, 7] class = High
avg_grade <= 9.5 gini = 0.126 samples = 437 value = [17, 12, 408] class = Medium
G1 <= 15.5 gini = 0.153 samples = 3 value = [0, 33, class = Low
gini = 0.0 samples = 86 value = [0, 86, 0] class = Low
G1 <= 2.0 gini = 0.028 samples = 70 value = [69, 0, 1] class = High
G1 <= 7.5 gini = 0.42 samples = 20 value = [14, 0, 6] class = High
G1 <= 9.5 gini = 0.464 samples = 41 value = [15, 0, 26] class = Medium
avg_grade <= 13.833 gini = 0.069 samples = 396 value = [2, 12, 382] class = Medium
reason_other <= gini = 0.108 samples = 3 value = [0, 33, class = Low
gini = 0.0 samples = 1 value = [0, 0, 1] class = Medium
(...)
Decision Tree (df_without)
failures <= 0.469 entropy = 1.291 samples = 649 value = [100.0, 131.0, 418.0] class = Medium
True
False
higher <= 0.5 entropy = 1.195 samples = 549 value = [51, 129, 369] class = Medium
reason_home <= 0.5 entropy = 1.121 samples = 100 value = [49, 2, 49] class = High
famsize_LE3 <= 0.5 entropy = 0.964 samples = 36 value = [14, 0, 22] class = Medium
school_MS <= 0.5 entropy = 1.156 samples = 513 value = [37, 129, 347] class = Medium
health <= -1.409 entropy = 1.069 samples = 84 value = [47, 1, 36] class = High
attendance_rate <= 0.281 entropy = 0.868 samples = 16 value = [2, 1, 13] class = Medium
Dalc <= -0.003 entropy = 0.755 samples = 23 value = [5, 0, 18] class = Medium
activities <= 0.5 entropy = 0.89 samples = 13 value = [9, 0, 4] class = High
schoolsup <= 0.5 entropy = 1.028 samples = 360 value = [11.0, 98.0, 251.0] class = Medium
Walc <= 0.1 entropy = 1. samples = 1 value = [26, 3 class = Medi
entropy = samples = value = [10, class = Hi
age <= 2.2 entropy = 1. samples = value = [37, 1 class = Hig
entropy = samples = value = [0, 1 class = Lo
paid <= 0.5 entropy = 0.567 samples = 15 value = [2, 0, 13] class = Medium
(...)

FIGURE 16

Finally, comparing **df_with** and **df_without** confirmed the issue of data leakage: prior grades dominated predictions when included, inflating performance. The more realistic scenario, excluding prior grades, showed lower predictive accuracy but offered a fairer reflection of which non-academic factors drive student outcomes.

Ethics:

The analysis relies on sensitive student data, which introduces important ethical responsibilities that include:

1. All attendance tracking and classification tools must ensure data privacy and secure storage and restricted access to authorized staff only as the goal is to support students and not penalize them.
2. Programs such as mentorship, family workshops and recovery initiatives must remain voluntary and inclusive to avoid labeling students.
3. Family engagement strategies should be culturally sensitive and accessible.
4. Participation in recovery or mentoring programs must remain confidential which reinforces trust and empowers students.

Recommendations:

Several recommendations can be made for schools and policy makers:

1. Implement an Automated Attendance Monitoring System that flags students with excessive absences early and trigger supportive interventions such as counseling and tutoring.
2. Launch Tailored Student Engagement Programs to ensure students do not plateau academically. It should include workshops on time management and structured mentorship for the "quiet achievers"
3. Strengthening Family Engagement by equipping parents and guardians with practical tools to support learning at home to amplify students' performance outcomes
4. Develop "Academic Fresh Start" Program to target students with prior failures to restore their confidence
5. Deploy a Cluster-Based Early Warning System that classifies students into risk groups within the first 4 weeks of the semester for targeted, early interventions.

Limitations:

Several limitations should be noted:

1. Limited Generalizability – The dataset comes from only two Portuguese schools, reducing applicability to other contexts or countries.
2. Self-Reported Features – Variables such as alcohol consumption and free time were self-reported, which may introduce bias or inaccuracies.
3. Risk of Data Leakage – Prior grades were highly correlated with final outcomes. Although this was mitigated by testing models with and without prior grades, it highlights the difficulty of isolating causal predictors.
4. Correlation vs. Causation – The models achieved high predictive accuracy but remain correlational. Features should be viewed as early warning signals rather than deterministic predictors of academic failure

Conclusion

This project analyzed student performance data to uncover behavioral, social, and demographic factors influencing academic outcomes. Through preprocessing, exploratory data analysis, and the application of both unsupervised and supervised learning techniques, key predictors of success and risk were identified, including absences, prior failures, study time, and parental education. The findings were translated into actionable recommendations such as attendance monitoring, family engagement programs, academic recovery initiatives, and predictive early warning systems. While the dataset and methods have limitations in terms of generalizability and causality, the results highlight the value of data-driven approaches for proactive student support. Overall, this study demonstrates how educational institutions can leverage analytics to move from reactive interventions toward early, personalized strategies that improve academic trajectories.