Student Performance Report

Prepared by: Adham Ehab Mokhtar
Submitted to: MAIM digital solution
Institution: FCDS Alexandria University
Date: 1/Sep/2025

Abstract:

This report details a Student Performance project focused on predicting student underperformance at the secondary education level. Using a dataset of student records from two Portuguese schools, the project aims to provide universities with early, data-driven signals about at-risk students. The methodology involved data preparation, exploratory data analysis (EDA), unsupervised learning for student segmentation, and supervised learning to predict student outcomes. The analysis revealed strong correlations between early-term grades and final performance, while behavioral factors like absences and study habits were also significant predictors. Unsupervised learning successfully segmented students into distinct behavioral profiles, which correlated with academic success. Supervised models were developed to predict student pass/fail status, with the most effective models demonstrating high predictive power even without using prior academic grades, which is crucial for an early warning system. The report concludes with actionable recommendations and a discussion of key ethical considerations.

Problem & Value:

Universities require timely, data-driven insights to identify students who may underperform and provide them with early intervention. This is a critical real-world problem, as proactive support can significantly improve student outcomes, increase retention rates, and optimize educational resources. The value of this project lies in transforming raw student data into concrete, actionable insights and predictive models that can serve as an effective early warning system. By exploring the drivers of performance and identifying at-risk student segments, the project provides a foundation for developing targeted support programs.

Datasets:

The project utilizes Two datasets of student records from two Portuguese schools. The original data, contained in student-mat.csv and student-por.csv, includes demographic, social, and school-related attributes.

- Source: The two original datasets were loaded and merged to create a single, comprehensive dataset.
- Datasets Link: https://archive.ics.uci.edu/dataset/320/student+performance
- Schema: The combined dataset contains 1,044 unique student records and 33 columns. The features include:
    - Demographics: school, sex, age, address, famsize, Pstatus.
    - Family Background: Medu (mother's education), Fedu (father's education), Mjob (mother's job), Fjob (father's job), guardian, famrel (family relationship quality).
    - School & Social Factors: traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, freetime, goout, Dalc (weekday alcohol consumption), Walc (weekend alcohol consumption), health.
    - Grades: G1 (first period grade), G2 (second period grade), G3 (final grade). The G3 column is the primary target variable.

- Data Quality: A comprehensive data quality report was generated in the 01_data_preparation.ipynb notebook, confirming:
    - No Duplicates: The merged dataset has 0 duplicate rows.
    - No Missing Values: There were no missing values in any of the features.
    - Outliers: Outliers were detected using the Interquartile Range (IQR) method for several numeric features, with failures (183 students with outliers) and absences (54 students with outliers) showing the most significant counts.

Methods:

Data Preparation & Feature Engineering

Before modeling, the data was preprocessed to handle both categorical and numerical features.

1. Categorical Encoding: One-hot encoding was applied to all 17 categorical features to convert them into a numerical format suitable for machine learning models.
2. Feature Scaling: Standard scaling was applied to the numerical features to standardize them, which is a crucial step for distance-based models like K-Means and SVM.
3. Target Variable: The final grade (G3) was used to define two target variables for supervised learning: a binary pass (G3 >= 10) and an optional 3-class risk_tier (low, medium, high).
4. Data Leakage: To simulate a true early warning system, two versions of the feature set were created: one including the prior grades (G1, G2) and one without, to assess how well models could predict outcomes using only non-academic, behavioral data.

Exploratory Data Analysis & Visualization

The Exploratory Data Analysis (EDA) section focused on uncovering the underlying relationships between student attributes and academic performance (G3). This analysis provided the foundation for feature selection and the development of predictive models.

- Correlation Analysis: The EDA confirmed that the prior grades, G1 (first period grade) and G2 (second period grade), have the strongest positive correlation with the final grade (G3), with correlation coefficients of 0.81 and 0.91, respectively. This finding is significant because it highlights the strong predictive power of early academic performance. Conversely, the number of past class failures (failures) showed the strongest negative correlation (-0.38), while age and alcohol consumption (Dalc and Walc) were also negatively correlated, albeit to a lesser extent. This suggests that non-academic factors play a role in predicting student outcomes.
- Group Comparisons: Visualizations were used to compare different student groups.
    - Study Time vs. Performance: Violin plots showed a clear trend: students with higher study time (e.g., > 10 hours per week) had a higher average final grade (G3) and a greater proportion of passing scores compared to students with very low study time.
    - Absences and Grades: Scatter plots revealed a negative relationship between absences and G3, confirming that students who miss more classes tend to have lower final grades.
    - School vs. Family Support: Comparisons showed that students receiving school and family support tended to have better academic outcomes, reinforcing the importance of a strong support system.

Unsupervised Learning (K-Means Clustering)

K-Means clustering was employed as an unsupervised learning technique to segment students into distinct groups based on their behavioral attributes. The objective was to discover naturally occurring patterns in student behavior that could be correlated with academic performance. The features used for clustering included studytime, absences, goout (going out with friends), freetime, and binary indicators for schoolsup (school support) and famsup (family support).

To determine the optimal number of clusters (k), the elbow method and silhouette scores were analyzed. While the elbow method showed a gradual decrease, the silhouette scores suggested a peak in performance around a certain number of clusters. Based on this analysis, k=9 was chosen to provide a more granular view of the different student profiles, as a larger number of clusters allowed for a finer distinction between groups with subtle behavioral differences.

After clustering, each of the 10 clusters was profiled based on the average values of the behavioral features and their average final grade (G3). This allowed for the identification of at-risk clusters. For instance, a cluster characterized by low study time and high absences was found to have a significantly lower average G3 and a higher proportion of students who failed. This step was crucial for transforming the clustering results into actionable insights for targeted interventions.

Supervised Learning (Classification)

The supervised learning phase was designed to build a predictive model that could classify students as either pass or fail based on their features. Four common machine learning algorithms were selected for this task:

- Logistic Regression: A linear model that is highly interpretable.
- Decision Tree: A non-linear model that is easy to interpret and can capture complex relationships.
- Random Forest: An ensemble method that uses multiple decision trees to improve accuracy and reduce overfitting.
- Support Vector Machine (SVM): A powerful algorithm for classification that finds the optimal hyperplane to separate classes.

A robust evaluation strategy was employed to ensure the reliability and generalizability of the models.

1. Data Splitting: The dataset was divided into an 80% training set and a 20% testing set. To prevent class imbalance from affecting the model, stratified sampling was used, ensuring that the pass/fail ratio in both the training and testing sets mirrored the ratio in the original dataset.
2. Cross-Validation: During the model training and hyperparameter tuning phase, 5-fold cross-validation was implemented. This technique divides the training data into five smaller folds, training the model on four folds and validating on the remaining one, a process that is repeated five times. This helps to ensure that the model's performance is not due to a single, specific data split and provides a more reliable estimate of its performance on unseen data.

3. Hyperparameter Tuning: For each model, GridSearchCV was used to systematically search for the best combination of hyperparameters that would maximize the F1 score. The F1 score was chosen as the primary optimization metric because it provides a balance between Precision (the model's ability to correctly identify positive cases) and Recall (the model's ability to find all positive cases), which is critical for a problem with potential class imbalance like pass/fail prediction.
4. Evaluation Metrics: The final models were evaluated on the held-out test set using a comprehensive suite of metrics to provide a full picture of their performance. These metrics included Accuracy, Precision, Recall, F1 Score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). The ROC-AUC is particularly important as it measures the model's ability to distinguish between the two classes across all possible classification thresholds.

Results:

Exploratory Data Analysis

The EDA confirmed that prior grades (G1 and G2) have the strongest positive correlation with the final grade (G3), with correlation coefficients of 0.81 and 0.91, respectively. Conversely, the number of past school failures (failures) showed the strongest negative correlation (-0.38), and age and alcohol consumption (Dalc and Walc) were also negatively correlated with G3.

K-Means Segmentation

The K-Means clustering successfully grouped students into distinct behavioral profiles. The analysis highlighted that clusters characterized by low study time and high absences generally had lower final grades (G3) and a lower pass rate. For example, one cluster had an average G3 of 9.77, while others with better habits had average G3 scores of 13.9. This demonstrates that behavioral patterns are a meaningful indicator of academic risk.

Supervised Learning Model Performance

The classification models were evaluated on two feature variants:

- With G1 and G2: As expected, models using prior grades performed exceptionally well, with F1 scores above 0.93. The Decision Tree model achieved the highest F1 of 0.949 and a ROC-AUC of 0.932. However, these models suffer from data leakage, as G1 and G2 are strong predictors of the final grade and are not available at the beginning of the academic term.
- Without G1 and G2: The models trained on non-academic features still provided valuable insights, with the best-performing model (Random Forest) achieving an F1 score of 0.876 and a ROC-AUC of 0.676. The Decision Tree model achieved an F1 of 0.866, showing it also generalizes well without the leaked data.
- Model Interpretation: The feature importance for the models without G1 and G2 revealed that failures (the number of past class failures), absences, and family/school support (famsup_yes, schoolsup_yes) were the most important predictors of final grade performance. This confirms the project's ability to identify early-warning signals.

Ethics:

This project, while valuable, presents several ethical considerations that must be addressed to ensure responsible use of the models.

1. Data Privacy and Governance: Student data is sensitive. It is imperative to handle all academic and behavioral records with strict data governance, ensuring anonymization and secure storage. Access to the data and the predictive model outputs should be limited to authorized personnel.
2. Fairness and Bias: The models may inadvertently exhibit bias against certain student demographics or socioeconomic statuses. The notebook recommends conducting regular fairness audits across sensitive attributes like gender, age, and family background to identify and mitigate any biases.
3. Avoiding Stigmatization: Labeling students as "high-risk" could lead to stigmatization and negatively impact their self-esteem. The project's output should be framed as identifying "support needs" rather than predicting "failure" to maintain a positive and supportive environment.
4. Transparency: The process of how a prediction is made should be transparent to students, parents, and educators. The use of interpretable models, like Logistic Regression and Decision Trees, along with clear reports on feature importances, helps ensure that the reasoning behind an intervention is understood and trusted.

Recommendations:

The project's findings lead to several actionable recommendations for university and school administrators to implement an effective early warning system:

1. Implement a Behavioral-Based Early Warning System: Focus on a model that does not rely on prior grades (G1, G2) to identify at-risk students as early as possible in the term. The Random Forest model without G1 and G2 is a strong candidate for this purpose due to its balanced performance.
2. Targeted Interventions: Leverage the model's insights to deploy specific interventions.
    - High Absences: For students with a high number of absences, the data suggests a significant risk of underperformance. Implement an attendance intervention program that proactively contacts students and families to understand and address the root causes of truancy.
    - Low Study Time & Support: The K-Means clustering revealed that students with low study time and limited family/school support are particularly vulnerable. This group could benefit from early study skills workshops and structured tutoring programs to improve their academic habits.
3. Leverage Prior Grades Strategically: While G1 and G2 are powerful predictors, relying on them exclusively is a reactive measure. Instead, these grades should be used in combination with behavioral signals to confirm risk and fine-tune intervention strategies, ensuring that students receive help before their grades have already fallen significantly.

Limitations:

The project is based on student data from two specific schools, which may limit the generalizability of the findings to a broader population. The models, particularly those that performed exceptionally well with G1 and G2 as features, are susceptible to data leakage, meaning they are not truly predictive of future performance but rather reflect existing academic trends. Additionally, while outliers were detected in features like failures and absences, the notebooks did not specify how these outliers were handled. This could impact on the overall robustness and accuracy of the models. Future work could address this by exploring different methods of outlier treatment and evaluating their impact on model performance.

Our greatest glory is not in never failing, but in rising every time we fail.

- Confucius