Dataset Analysis Report

Dataset: Students Performance Dataset (`students_performance_data.csv`)

---

1. Dataset Overview

The Students Performance Dataset contains academic, demographic, and extracurricular information related to student performance. The dataset includes features such as age, study time, absences, parental background, extracurricular involvement, and academic outcomes. The dataset was explored by inspecting the structure, column names, and representative sample values to understand the nature of each feature.

The dataset consists of multiple numerical and categorical variables, making it suitable for exploratory data analysis and supervised machine learning tasks.

---

2. Target Variable Identification

The **GradeClass** column was identified as the target variable, as it represents the academic performance category of students and follows an ordered structure. All remaining columns, except the StudentID identifier, were considered input features for machine learning models.

---

3. Machine Learning Suitability

The dataset is suitable for supervised machine learning, particularly classification tasks. It contains a balanced mix of numerical, categorical, ordinal, and binary features, allowing for feature engineering and preprocessing techniques such as scaling, encoding, and ordinal mapping. The dataset size is adequate for training and evaluating classification models, provided proper preprocessing is applied.

---

4. Data Quality Observations

- Column names and values are well-defined and interpretable.
- Binary features are consistently represented.
- Ordinal features require encoding while preserving order.
- Categorical features require appropriate encoding techniques.

- The GPA distribution should be examined for potential outliers that may influence model performance.
- The target variable GradeClass may exhibit class imbalance, which should be addressed during model training.

Conclusion

Overall, the Students Performance Dataset is well-structured and suitable for machine learning applications. Proper preprocessing and data quality checks will enable effective model development and reliable performance evaluation.