Dataset Summary: Parental Involvement in Student Performance

Malek Frikha, Mathis Lesgourgues, Bastien Mousse, Amine Ould Hocine, Samuel Povoa

Overview of the Dataset

The dataset titled *Students Performance Dataset* (Kaggle link) provides comprehensive information on 2,392 high school students. It includes variables covering demographics, study habits, parental involvement, extracurricular participation, and academic performance. The dataset is made for analyzing factors that influence student performance and offer opportunities for predictive modeling. Below is the description of the main variables of this dataset.

1. Demographic Details

- **StudentID**: A unique identifier for each student, ranging from 1001 to 3392.
- **Age**: Age of students, ranging from 15 to 18 years.
- **Gender**: Gender is a binary variable, where 0 = Male and 1 = Female.
- **Ethnicity**: Categorical variable labeled as:
  - 0: Caucasian
  - 1: African American
  - 2: Asian
  - 3: Other

2. Parental Education and Support

- **ParentalEducation**: Education level of parents, labeled as:
  - 0: None
  - 1: High School
  - 2: Some College
  - 3: Bachelor's Degree
  - 4: Higher (Postgraduate degrees)
- **ParentalSupport**: Level of parental support, coded as:
  - 0: None
  - 1: Low
  - 2: Moderate
  - 3: High
  - 4: Very High

3. Study Habits

- **StudyTimeWeekly**: Weekly study time in hours, ranging from 0 to 20 hours.
- **Absences**: Number of absences during the school year, ranging from 0 to 30.
- **Tutoring**: Tutoring status, where 0 = No and 1 = Yes.

4. Extracurricular Activities

- **Extracurricular**: Participation in extracurricular activities, where 0 = No and 1 = Yes.
- **Sports**: Participation in sports activities, where 0 = No and 1 = Yes.
- **Music**: Participation in music activities, where 0 = No and 1 = Yes.
- **Volunteering**: Involvement in volunteering activities, where 0 = No and 1 = Yes.

5. Academic Performance

- **GPA**: Grade Point Average on a scale of 2.0 to 4.0, which reflects academic performance influenced by study habits and parental involvement.
- **GradeClass**: Classification of student grades based on GPA:
  - 0: 'A' (GPA ≥ 3.5)
  - 1: 'B' ($3.0 \leq$ GPA < 3.5)
  - 2: 'C' ($2.5 \leq$ GPA < 3.0)
  - 3: 'D' ($2.0 \leq$ GPA < 2.5)
  - 4: 'F' (GPA < 2.0)

Target variable

For prediction and inference we will use GradeClass variable instead of GPA. The choice of GradeClass as the target variable is practical and aligns with educational goals. Predicting categorical grades is more interpretable and suitable for educational interventions compared to raw GPA values. Additionally, this allows for better analysis of what distinguishes higher-achieving students from lower-achieving ones, facilitating targeted recommendations for students in different performance brackets.

Dataset Usage and Attribution

According to the kaggle description of this dataset,it was created by Rabie El Kharoua for educational and research purposes. This dataset is described as particularly useful for data science projects aimed at educational analysis.

It was chosen for its comprehensive scope, which covers various potential influences on student performance. It is particularly valuable in educational research, as it enables an analysis of both individual behaviors (study habits) and external factors (parental support, extracurricular participation).

Conclusion

This dataset provides a rich basis for studying factors associated with student performance and parental involvement. By using GradeClass as a categorical target variable, we can more effectively understand and communicate insights relevant to education, supporting efforts to improve student outcomes through data-driven recommendations.