Student Performance Dataset Information

📊 Dataset Overview

Source

- **Repository**: UCI Machine Learning Repository
- **Dataset ID**: 320
- **License**: CC BY 4.0
- **Citation**: Cortez and Silva, 2008

Description

This dataset approaches student achievement in **secondary education** of two Portuguese schools. The data attributes include student grades, demographic, social and school-related features collected using school reports and questionnaires.

Key Details

- **Subjects**: Mathematics (mat) and Portuguese language (por)
- **Records**: 649 students
- **Features**: 33 variables (30 features + 3 targets)
- **Missing Values**: ✕ None
- **Tasks**: Binary/five-level classification and regression

Important Note on Target Variables

⚠ **Data Leakage Warning**: The target attribute `G3` has a **strong correlation** with `G2` and `G1`:

- **G1**: 1st period grade
- **G2**: 2nd period grade
- **G3**: Final year grade (3rd period) - **Main Target**

It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful for early intervention.

📋 Variable Documentation

🏫 School & Demographics

Variable	Type	Description	Values
`school`	Categorical	Student's school	GP (Gabriel Pereira) or MS (Mousinho da Silveira)
`sex`	Binary	Student's gender	F (female) or M (male)
`age`	Integer	Student's age	15 to 22 years
`address`	Categorical	Home address type	U (urban) or R (rural)

Family Background

Variable	Type	Description	Values
famsize	Categorical	Family size	LE3 (≤3 members) or GT3 (>3 members)
Pstatus	Categorical	Parent cohabitation status	T (together) or A (apart)
Medu	Integer	Mother's education level	0-4 (none to higher education)
Fedu	Integer	Father's education level	0-4 (none to higher education)
Mjob	Categorical	Mother's job	teacher, health, services, at_home, other
Fjob	Categorical	Father's job	teacher, health, services, at_home, other
guardian	Categorical	Student's guardian	mother, father, other

School-Related Factors

Variable	Type	Description	Values
reason	Categorical	Reason for school choice	home, reputation, course, other
traveltime	Integer	Home to school travel time	1 (<15min) to 4 (>1hour)
studytime	Integer	Weekly study time	1 (<2h) to 4 (>10h)
failures	Integer	Past class failures	0-4 (4 means ≥3 failures)
schoolsup	Binary	Extra educational support	yes or no
famsup	Binary	Family educational support	yes or no
paid	Binary	Extra paid classes	yes or no
activities	Binary	Extra-curricular activities	yes or no
nursery	Binary	Attended nursery school	yes or no
higher	Binary	Wants higher education	yes or no
internet	Binary	Internet access at home	yes or no

Social & Personal

Variable	Type	Description	Values
romantic	Binary	In romantic relationship	yes or no
famrel	Integer	Family relationship quality	1 (very bad) to 5 (excellent)
freetime	Integer	Free time after school	1 (very low) to 5 (very high)
goout	Integer	Going out with friends	1 (very low) to 5 (very high)
Dalc	Integer	Workday alcohol consumption	1 (very low) to 5 (very high)

Variable	Type	Description	Values
`Walc`	Integer	Weekend alcohol consumption	1 (very low) to 5 (very high)
`health`	Integer	Current health status	1 (very bad) to 5 (very good)
`absences`	Integer	School absences	0 to 93

🎯 Target Variables (Grades)

Variable	Type	Description	Values
`G1`	Integer	First period grade	0 to 20
`G2`	Integer	Second period grade	0 to 20
`G3`	Integer	**Final grade (Main Target)**	0 to 20

🔍 Data Analysis Considerations

Feature Categories

- **Demographic**: `sex`, `age`, `address`, `famsize`, `Pstatus`
- **Socioeconomic**: `Medu`, `Fedu`, `Mjob`, `Fjob`, `guardian`
- **Academic**: `school`, `reason`, `traveltime`, `studytime`, `failures`
- **Support Systems**: `schoolsup`, `famsup`, `paid`, `activities`, `nursery`, `higher`, `internet`
- **Behavioral**: `romantic`, `famrel`, `freetime`, `goout`, `Dalc`, `Walc`, `health`, `absences`

Data Quality

- ✅ **No missing values** across all 649 records
- ✅ **Consistent data types** and encoding
- ✅ **Well-documented** variable definitions
- ✅ **Balanced representation** across schools

Research Applications

1. **Early Warning Systems**: Predict G3 using only demographic and early behavioral indicators
2. **Intervention Targeting**: Identify at-risk student profiles
3. **Factor Analysis**: Understand drivers of academic performance
4. **Policy Insights**: Inform educational support programs

📚 References

- Cortez, P., & Silva, A. M. G. (2008). *Using data mining to predict secondary school student performance*. Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008).
- UCI ML Repository: https://archive.ics.uci.edu/dataset/320/student+performance