Gathering Data

Outline Data Requirements

The project requires data that captures student demographics, academic performance, study habits, and extracurricular activities to analyze the factors influencing students' grades (target variable: GradeClass).

Verify Data Availability

The dataset contains 2,392 entries and includes the following fields:

- Student ID
- Age, Gender, Ethnicity
- Parental Education
- Study Time Weekly, Absences, Tutoring
- Parental Support
- Participation in extracurricular activities (sports, music, volunteering)
- GPA (Grade Point Average) and GradeClass (target variable).

Define Selection Criteria

The dataset was selected due to its comprehensive details on student performance and associated factors. Features with high relevance to academic success, such as GPA, GradeClass, and study habits are included. The dataset is complete with no missing key variables.

Describing Data

The dataset contains 15 columns and 2,392 rows. Key characteristics include:

- Demographics: Age (15–18), Gender (binary), Ethnicity (4 categories).
- Study Habits: StudyTimeWeekly (0–20 hours), Absences (0–29), Tutoring (binary).
- Parental Factors: Education level (5 categories, 1-4) and support level (5 categories, 1-4).
- Extracurriculars: Sports, music, and volunteering (all binary).
- Performance: GPA (0.0–4.0) and GradeClass (0.0-4.0) explained under.

GradeClass: Classification of students' grades based on GPA:

- 0: 'A' (GPA >= 3.5)
- 1: 'B' (3.0 <= GPA < 3.5)
- 2: 'C' (2.5 <= GPA < 3.0)
- 3: 'D' (2.0 <= GPA < 2.5)
- 4: 'F' (GPA < 2.0)

Exploring Data

Exploratory analysis reveals:

- Age Distribution: Students are distributed fairly evenly across ages 15–18.
- Gender: Nearly equal proportions of male (0) and female (1) students.
- Parental Education: Most parents have completed high school or some college.
- Study Habits:
    - Weekly study time is very event distributed between 0-20 hours.
    - Absences don’t differ a lot from each other.
    - Few students (approximately 30%) receive tutoring.
- Parental Support: Leaned towards moderate and high support levels.
- Extracurriculars: Participation is moderate, with sports being the most popular.
- Academic Performance:
    - GPA is common between 0.0 and 2.0, indicating most students do not get a high enough grade(E - A) to pass.
    - GradeClass distribution suggests more students fail (Grade F) compared to those excelling (Grade A).

Verifying Data Quality

The dataset quality is verified through:

1. Completeness:
    - There are no missing values in essential fields.
    - All 15 columns have data for all rows.
2. Accuracy:
    - Numerical fields (e.g., GPA, StudyTimeWeekly) fall within expected ranges.
    - Categorical fields (e.g., Gender, GradeClass) match documented codes.
3. Consistency:
    - GPA aligns with GradeClass labels.
    - No outliers in numeric fields (e.g., no GPA below 0.0 or above 4.0).

https://github.com/mattiasminejev/andmeteadus.git