EDA Report – High School Alcoholism and Academic Performance

# A brief description of the dataset and summary of attributes

The dataset, High School Alcoholism and Academic Performance was sourced from Kaggle. The data was collected through a survey conducted among high school students in Portugal and includes biographical and behavioral data of students relating to alcohol consumption and their academic population such as school, age, family dynamics, structure, school attendance, grades, etc. These features may prove to be valuable in understanding the impact of alcohol consumption on academics as well as identifying potential causes of teenage alcohol consumption.

## SIZE, AND Quality

The dataset contains 31 features (columns) and 649 data points (rows). Feature data types include integers as well as objects (strings). Some features appear to have missing values which will be handled appropriately for proper data analysis. Many of these features contain non-numeric data which will be encoded to numeric values. Refer to Appendix A, Screenshots 1, and 2 for more details.

## List of Features

Following is the list of features in the dataset and descriptions with valid values.

|  |  |
| --- | --- |
| Feature | Description |
| School | Student's school (Gabriel Pereira, Mousinho da Silveira) |
| Gender | Student's sex (Female, Male) |
| Age | Student age (numeric: 15 to 22) |
| Housing\_Type | Type of student's residential address (Urban, Rural) |
| Family\_Size | Family size (Up to 3, Above 3) |
| Parental\_Status | Parents' cohabitation status (Separated, Living Together) |
| Mother\_Education | Mother's education level (Primary School, Lower Secondary School,  High School, Higher Education) |
| Father\_Education | Father's education level (Primary School, Lower Secondary School,  High School, Higher Education) |
| Mother\_Work | Mother's job (Teacher, Health, Services, Homemaker, or other) |
| Father\_Work | Father's job (Teacher, Health, Services, Homemaker, or other) |
| Reason\_School\_Choice | Reason for choosing this school (Course Preference, Near Home, Reputation, Other) |
| Legal\_Responsibility | Student's guardian (Mother, Father, Other) |
| Commute\_Time | Travel time from home to school (Up to 15 min, 15 to 30 min, 30 min to 1h, More than 1h) |
| Weekly\_Study\_Time | Weekly study time (Up to 2h, 2 to 5h, 5 to 10h, More than 10h) |
| Extra\_Educational\_Support | Extra educational support (Yes, No) |
| Parental\_Educational\_Support | Family educational support (Yes, No) |
| Private\_Tutoring | Private classes on subjects related to the course (Yes, No) |
| Extracurricular\_Activities | Performs extracurricular activities (Yes, No) |
| Attended\_Daycare | Attended daycare (Yes, No) |
| Desired\_Graduate\_Education | Desire to pursue a degree (Yes, No) |
| Has\_Internet | Internet access at home (Yes, No) |
| Is\_Dating | Are you in a romantic relationship (Yes, No) |
| Good\_Family\_Relationship | Quality of family relationships (Very Poor, Poor, Fair, Good, Excellent) |
| Free\_Time\_After\_School | Free time after school (Very Low, Low, Moderate, High, Very High) |
| Time\_with\_Friends | Time with friends (Very Low, Low, Moderate, High, Very High) |
| Alchohol\_Weekdays | Alcohol consumption on the workday (Very Low, Low, Moderate, High, Very High) |
| Alchohol\_Weekends | Alcohol consumption on the weekend (Very Low, Low, Moderate, High, Very High) |
| Health\_Status | Current health status (Very Poor, Poor, Fair, Good, Very Good) |
| School\_Absence | Number of school absences (numeric: from 0 to 32) |
| Grade\_1st\_Semester | First-semester grade (numeric: from 0 to 20) |
| Grade\_2nd\_Semester | Second-semester grade (numeric: from 0 to 20) |

# INITIAL PLAN FOR DATA EXPLORATIoN

1. Examine the size, dimensions, type of contents, and inconsistencies within the dataset.
2. Clean the data by handling any missing values or outliers.
3. Perform necessary transformations, by encoding and scaling the data. It's worth testing different approaches for this step based on insights from step 4.
4. Visualize data using various tools and look for features that may indicate correlations between alcohol consumption, academic performance, and student information.
5. Utilize visual indicators and descriptive statistics to formulate hypotheses.
6. Perform trivial hypotheses testing and evaluate the validity of the assertions.
7. Identify areas for improvement and further development.

# Data Cleaning and feature engineering

## notes on approach

1. Some data points do not have values for Father’s Education and Mother’s Education. While numerical values may be estimated from other data, it is not possible to accurately “guess” their education levels. Thus, these data points are removed from the dataset when considering the impact of these features.
2. Most of the features in this dataset have non-numeric values. These values have been encoded for further use.

# Quality of data

While the dataset hardly contains any missing fields/values, the sample size is small and restricted to students of only two schools, Gabriel Pereira, and Mousinho da Silveira. This may potentially introduce some biases in analyzing high school students’ attitudes toward alcohol consumption and produce skewed results.

## REQUEST FOR ADDITIONAL DATA

Expand the survey to multiple schools in multiple locations across Portugal.

Parent/guardian alcohol consumption. Self-reported average drinking age within the family.

Define and collect metrics on the quality of education and student’s family finances (income, and assets).

# Appendix A: SCREENSHOTS

## 1) First 5 rows of data

A screenshot of a computer screen

Description automatically generated

A computer screen shot of a computer program

Description automatically generated

## 2) FEATURE INFORMATION

A screenshot of a computer program

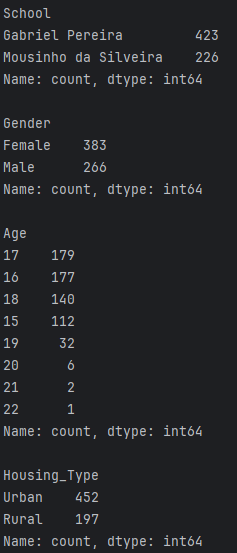
Description automatically generated

## 3) Statistical information about numeric attributes (Age, Absence, GRADES)

A screenshot of a computer screen

Description automatically generated

## 4) VALUE COUNTS

 A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

# REFERENCES

P. Cortez e A. Silva. Usando a Mineração de Dados para Prever o Desempenho do Aluno do Ensino Médio. Em A. Brito e J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, abril de 2008, EUROSIS, ISBN 978-9077381-39-7.