Progress check document

DSGP8

January 2024

1 Project idea

Our project idea is to investigate if family background and different factors such as academic grades or social life can affect the alcohol consumption of students. The question proposed is therefore how different features can affect alcohol consumption in secondary school students.

2 Data source

The data were obtained in a survey of students' maths and Portuguese language courses in secondary school. It contains a lot of interesting social, gender, and study information about students. The link is attached below: `https://data.world/databeats/student-alcohol-consumption`

3 Datasheet

3.1 Motivation

This dataset was created to test the correlation between alcohol consumption and social, gender, study time, and grade attributes for each student. Jonathan Ortiz created it at data.world. There was no organization since it was a survey of students in secondary school.

3.2 Composition

The instances represent the number of students and there are 1045 in total. The dataset contains all possible instances because it is a school survey and cannot be larger. It is separated into two subsets corresponding to students of Maths and Portuguese language courses.

Each feature description is explained in detail at the source page, written in a text file called *student*. The dataset does not contain any instances with missing values and errors but in the majority, there are categorical values.

The dataset is self-contained and it does not rely on other external resources, as it was created 8 years ago, it is highly guaranteed that will exist over time. There are no restrictions associated and it is shared with everyone. Also, does not contain confidential data since the names of students are not provided, so it is anonymous and it is not possible to identify each individual. The dataset identifies sub-populations such as gender (F/M), range of age (from 15 to 22), or school ("GP" - Gabriel Pereira or "MS" - Mousinho da Silveira).

3.3 Collection Process

As mentioned before, the data associated with each instance are acquired by survey responses of students in Portugal. There is neither information given regarding the procedures used to collect the data nor if the data was verified. Since the collection process was not managed by us, there are no further details that we can provide.

3.4 Preprocessing/cleaning/labelling

The preprocessing of the data was not complex, it was mainly about removing some features that are not considered important for our dataset such as address or internet access at home.

Two subsets of the dataset were merged into one to implement the analysis so that there are a larger amount of instances and more convincing. This process was implemented in Python, using the library *pandas*.

3.5 Uses

The dataset could be used to visualize the alcohol consumption of the number of secondary students, which are mostly minors. And how it affects their academic performances, by comparing their grades and alcohol consumption grouping by age. Dataset consumers do not need to know anything because no sensible information can lead to legal risks or financial harm.

3.6 Distribution

This dataset has no digital object identifier, so it was not distributed. But it has a **CC0 1.0 DEED** license, which the author dedicated the work to the public domain with no copyright. It is possible to copy, modify, and distribute even for commercial purposes.

3.7 Maintenance

There are no limits on the retention of the data associated with the instances and the dataset is not likely to be updated since the last update was 8 years ago.

4 Plan for the project

Week	Date	Tasks	Division of labour
W2	1.26 Fri	Discussion and EDA	All members get together to explore the dataset, delete unrelated features, test the feasibility of the dataset, and define our project question and goals.
W3	1.28 Mon	Conclude the EDA result and prepare the document for the progress check	Monica is responsible for clarifying the project question and goals. Meichen is in charge of showing the distribution of our dataset. Kanghui is responsible for showing the result of the chi-square test which illustrates project feasibility. Yingxin is in charge of writing the plan for the project and completing the progress check document with Monica.
W4	2.5 Mon	Do some research about causal inference and try models that are suitable	Monica and Yingxin are responsible for searching literature about causal inference and structure causal model. Meichen and Kanghui are in charge of trying some models like decision trees, random forest, and logistic regression to find whether they are suitable.
W5	2.12 Mon	Try to use structure causal model to the dataset	Yingxin and Kanghui are in charge of defining the causal structure and its maths function. Monica and Meichen are responsible for coding.
W6	2.19 Mon	If coding is not completed, keep coding. If done, try to improve it.	Have a short break and prepare for exams in other modules. Debug together and try to improve our model.
W7	2.26 Mon	Read the report example. Analyse the result and give some conclusions.	Read the report example together and then Yingxin is responsible for defining an outline of the report. All the members start writing the corresponding section of the report.
W8	3.4 Mon	Writing report	Keep searching pertinent literature and writing the report.
W9	3.11 Mon	Prepare for the midterm demo	Yingxin and Monica are responsible for the slides for the presentation while Meichen and Kanghui are in charge of making the poster.
W11	4.25 Thu	Report submission	Check the report one last time.