Group B15 report

KAGGLE- STUDENT ALCOHOL CONSUMPTION

Otto Bruno Koobakene, Liivika Koobakene

Link to repository:
https://github.com/ottobrunok/IDS-project

Task 2

Background

Students from all over the world are affected by their social and economical status as it plays a great role in their academic achievements. One aspect of social life among secondary school students is social gatherings which often include alcohol consumption. Unfortunately alcohol consumption affects overall health in adolescence and is often related to lack of focus, brain development issues, memory and has negative consequences for psychological, physical and social health. This affects students' academic performance as well, which has a larger impact on overall career success later in life. Skipping or failing classes could lead to the student being expelled and therefore lead to the student not acquiring education that could help them get higher paying jobs in their future careers.

The data in our dataset was obtained in a survey in secondary education of two Portuguese schools. The data attributes include students' alcohol consumption, grades, demographic, social and school related achievements and it was collected by a university student using questionnaires and grade reports. Datasets are provided regarding the performance in the subjects of Portuguese language and Mathematics.

We are also planning to carry out a survey among our fellow university students in Estonia. These datasets will help us get a better understanding and overview of the effects of alcohol consumption on the grades as well as its relation to socio-economic attributes. It will also help to predict students' final grade considering their alcohol consumption.

Business goals

The main goal of our project is to determine if and how many different socio-economic factors, alcohol consumption and absence from class affect the students’ final grade. If we could determine the main causes for lower final grades and skipping class, educational institutions and social assistance programs could use that information to spread awareness about the impact of those factors to the future of students. They could also implement different restrictions or incentives for students to avoid skipping class and spend more time studying.

Business success criteria

We aim to rank the impact of different socio-economic factors, skipping class and alcohol consumption on the final grade. Therefore we could determine the lead cause for lower final grades. We would consider our project to be successful if our prediction algorithm could predict the final grade of a student with a 75% accuracy, depending on the socio-economic factors, number of skipping class and habits of alcohol consumption.

Inventory of resources

We have student alcohol consumption, grades and socio-economic factors data from Kaggle, last updated in 2016, from Portugal secondary schools.
We will gather similar data from our fellow university students in Estonia.
Other resources include us as project members, software we will use(Jupyter Notebook and Python 3 with its packages) and our personal computers as hardware.

Requirements, assumptions, and constraints

After we conduct a survey among Estonian university students, we will make sure that the gathered data is stored securely and is not shared with anyone.

Risks and contingencies

Main risk we are fearing during our project is that we might not be able to gather enough data from our fellow students. Also the accuracy of the gathered data, as grades are a fixed value that can be looked up, but alcohol consumption might not be remembered or measured accurately. Also the accuracy of the results, as some aspects might be subjective and it might be hard to analyze the direct impact of only alcohol consumption.

Terminology

Socio-economic factors - Social and economic factors, such as income, education, employment and community safety.

Failures - number of past class failures

Study time - weekly study time in hours

Alcohol consumption- weekly alcohol consumption defined in a scale from 1-5(none to very high)

Costs and benefits

Our project does not have any direct monetary costs and benefits. However, if the results and findings are used by social assistance programs or educational institutes to spread awareness and apply incentives or restrictions, it could lead to some students changing their habits, passing classes and acquiring a good education that could lead to a higher paying job.

Data-mining goals

Our goal is to use existing data, gather our own data by doing an online survey and sharing it with fellow students/friends. In the end the goal is to go through existing Portuguese survey data, select the attributes that are relevant to our project which is to train a prediction model to show how alcohol consumption affects students' final grade. We might not use all the attributes that are available in the survey dataset for that, but use only the ones that seem to be related to academic achievement like grade reports and absences.

Data-mining success criteria

When conducting the survey, we would consider it a success if we get at least 20 participants.

Task 3

Gathering data

We mainly use socio-economic factors, workday and weekend alcohol consumption, number of absences and grade report data. The existing dataset is already available to us on Kaggle, however we will also gather new data with our survey among our fellow Estonian university students.
Within the existing dataset, we will make a selection of necessary fields and for getting weekly alcohol consumption data, we will combine weekday and weekend consumption values in the dataset.

Describing data

We have an existing dataset from Kaggle - https://www.kaggle.com/uciml/student-alcohol-consumption.
It has survey data from 382 students from a secondary school in Portugal. The survey was conducted by a university student from Portugal.
The data fields in bold are the ones that are necessary for our analysis.

school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex - student's sex (binary: 'F' - female or 'M' - male)
age - student's age (numeric: from 15 to 22)
address - student's home address type (binary: 'U' - urban or 'R' - rural)
famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian - student's guardian (nominal: 'mother', 'father' or 'other')
traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures - number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup - extra educational support (binary: yes or no)
famsup - family educational support (binary: yes or no)
paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities - extra-curricular activities (binary: yes or no)
nursery - attended nursery school (binary: yes or no)
higher - wants to take higher education (binary: yes or no)
internet - Internet access at home (binary: yes or no)
romantic - with a romantic relationship (binary: yes or no)
famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime - free time after school (numeric: from 1 - very low to 5 - very high)
goout - going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health - current health status (numeric: from 1 - very bad to 5 - very good)
absences - number of school absences (numeric: from 0 to 93)
G1 - first period grade (numeric: from 0 to 20)
G2 - second period grade (numeric: from 0 to 20)
G3 - final grade (numeric: from 0 to 20, output target)

Exploring data

Looking at the data, we have found some questionable data. For example, there are 16 students in the Portuguese language course, who have 0 absences from class, but also got a final grade of 0/20, which might be true but is very unlikely and does not make sense.
There are several data fields that are not of interest to us. For example: father's job, with a romantic relationship, home to school travel time etc.

Verifying data quality

Overall, the quality of the data seems fine. The source of the data seems to be reliable, acquired from 2 students at University of Minho. We believe the data is good enough to support our goals, after we select the necessary data fields.

Task 4

The tasks are distributed between 2 team members, many of the tasks will be carried out and discussed together, estimated working time for both members is around 30-35 hours.

Project plan including asks, methods and tools:

1. Making the online survey. (Effort estimation: 2 hours)(Using: Google Forms)
2. Sharing the link to the survey among fellow university students. (Effort estimation: 1 hour)
3. Analyzing and cleaning the existing data from Kaggle and our survey, extracting only the attributes that are necessary for our project. (Effort estimation: 4 hours)(Using: Python and Jupyter notebook)
4. Choosing the model for the training and test dataset. (Effort estimation: 2 hours)
5. Deciding which algorithms to use to train the models. (Effort estimation: 6 hours)
6. Training the models and picking the most accurate one. (Effort estimation: 9 hours)
7. Predicting the student’s final grade based on his/her weekly alcohol consumption using the most accurate model. (Effort estimation: 2 hours)
8. Predicting the student’s final grade based on his/her socio-economic factors using the most accurate model. (Effort estimation: 2 hours)
9. Predicting the student’s final grade based on his/her total absences using the most accurate model. (Effort estimation: 2 hours)
10. Analyzing the prediction results. (Effort estimation: 2 hours)
11. Picking out relevant information. (Effort estimation: 2 hours)
12. Visualizing relevant results. (Effort estimation: 2 hours)

13. Preparing, recording and editing the project video. (Effort estimation: 8 hours)(Using: ShotCut)
14. Making the poster. (Effort estimation: 6 hours)