HW10 Report

Project title: “Mastering Success: How to be a good student”

Team members: Anna Sulg, Johan Veenpere, Hendrik Aruoja

Task 1. Setting up

Link of the repository : GitHub

Task 2. Business understanding

As students ourselves we know how sometimes despite putting in what feels like efficient effort, your test scores end up indicating otherwise. Our goal is to analyze the effect of different factors on a student’s performance in school, develop a predictive model that students can use to learn the potential test score they would achieve based on their lifestyle, studying habits and effort. We also want to give students personalised tips based on data analysis on how to improve their test scores. We will conduct a deeper dive into the data and connections between different factors. We wish to determine the main factors that contribute to a student's success, debunk or back up some common suggestions for improving scores and making our work easily accessible to all who are interested in gaining insight on how to be a better student.

Assessing your situation:

Inventory of resources - As human resources we have ourselves (3 CS students), who do the whole work. For assistance we can turn to our lab TA and course organizers. On hardware side, we have 3 laptops, atleast one desktop and access to Google Colab cloud environment for running long-term training.

Requirements, assumptions and constraints - This project must be finished by presentation day on Dec 13. The poster must be submitted 4 days before the poster session due to printing taking time. To complete the project we need atleast two datasets, to which we have access and have already downloaded. We do not have any legal and security obligations.

Risks and Contingencies - A team member falls ill: Other team members will do the other’s task(s).

Terminology - We do not have any special terminology.

Costs and benefits - The only real cost is time. This project cost 90 man-hours divided between three team members. Benefits are: learn to cooperate on data science projects, course credit, basic knowledge about data science, presentation and science communication skills.

Defining your data-mining goals

Data-mining goals - We plant to create a model to predict a student's performance based on data about their study habits, sleep schedule etc. If we have time we would like to make the model usable on a website to demo our work. Our main deliverable is the final report and presentation which is in the form of a poster presented at the poster session. Other deliverables are jupyter notebooks showcasing our methods and the analysis process, and the cleaned datasets.

Data-mining success criteria - Success criteria are a working model with 95% accuracy and a sufficient number of graphs which are clear and visually appealing.

Task 3. Data understanding

We followed the standard four step process for the CRISP-DM data understanding part.

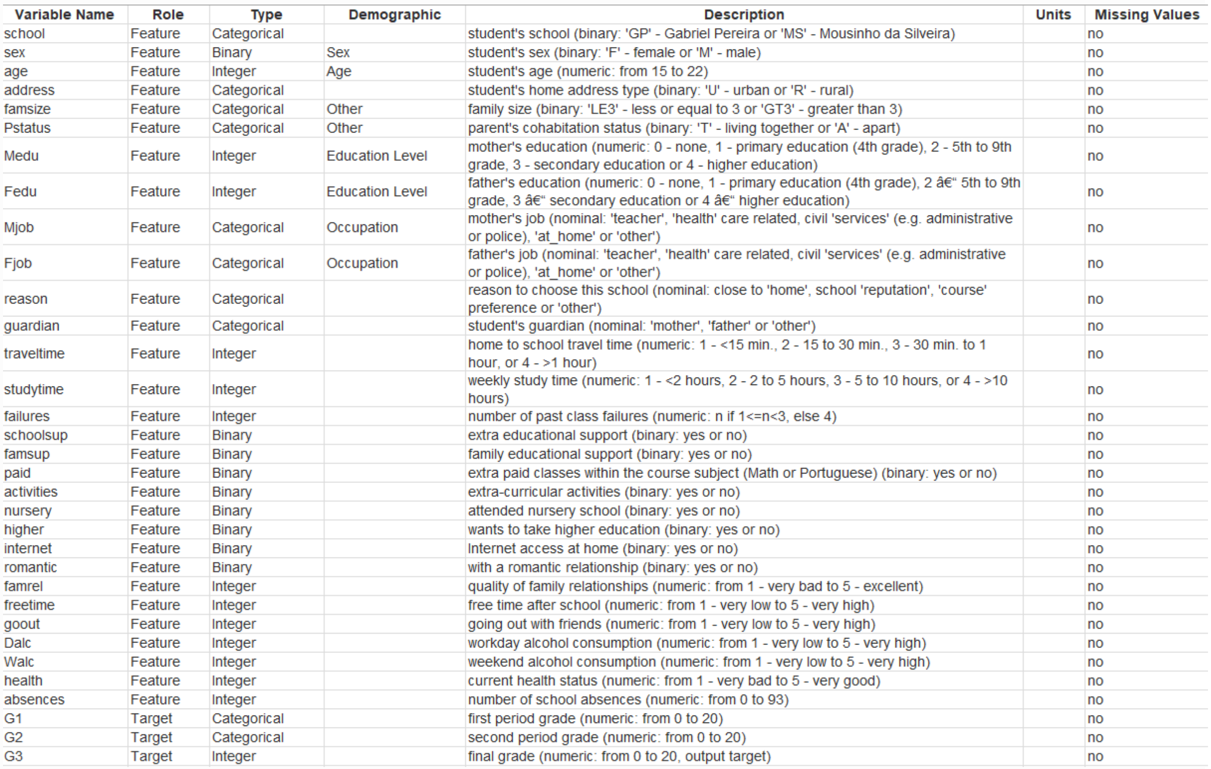
Gathering data

Our project requires data about how different factors affect student performance. Internet provides many sources of research where student performance data has been gathered . Our first choice was the machine learning community Kaggle and it had a few possible candidates freely available for this project. Each of them had their own flows. Some datasets were too small, some did not have enough features. Finally we found a good additional source of data from UC Irvine Machine Learning repository.

The data had to be sufficient in order to perform statistical or machine learning tasks on it for the result to be precise enough. We made the decision to combine the results of different datasets for the project to add complexity to the analysis.

Describing data

The datasets that we use come with sufficient descriptions of the variables used that are easily understandable using common sense. Example (Dataset 1):



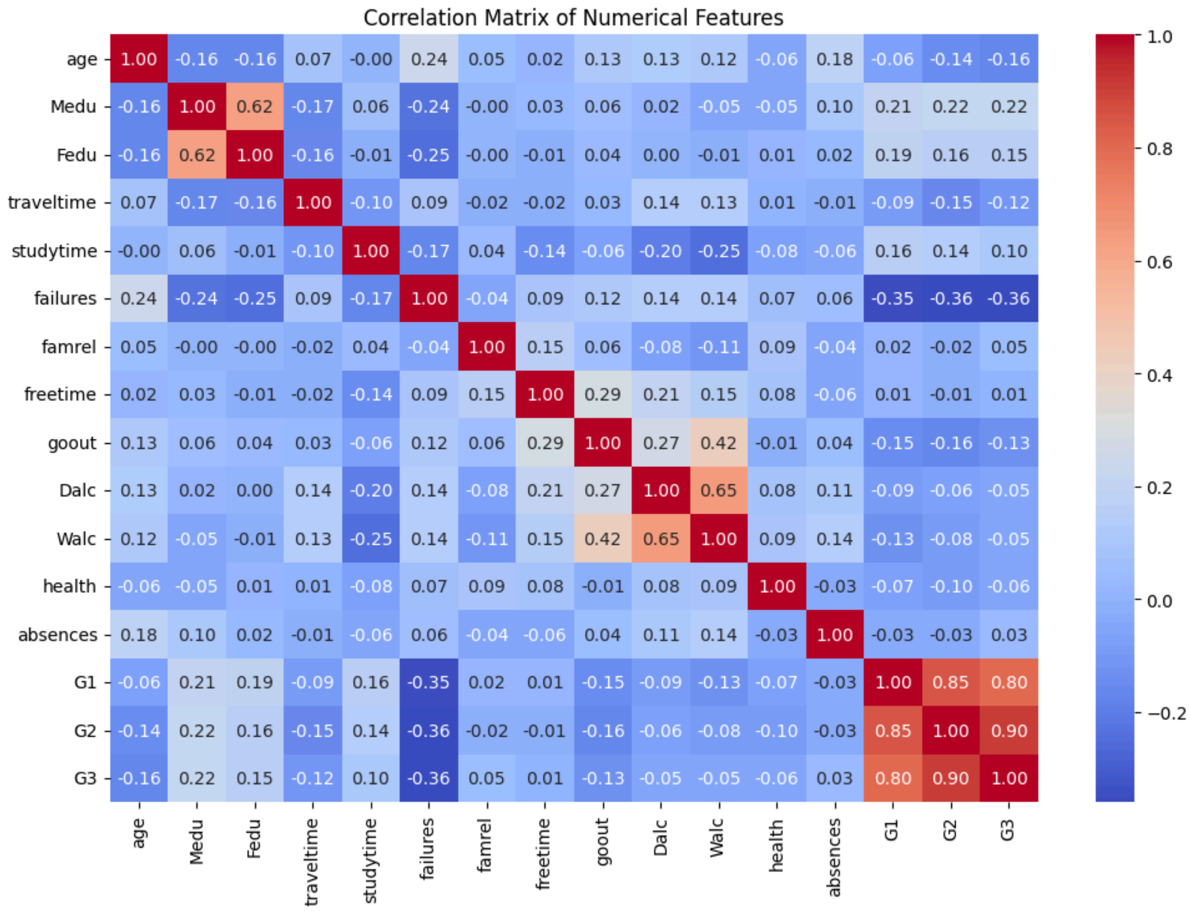
Variable Name	Role	Type	Demographic	Description	Units	Missing Values
school	Feature	Categorical		student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)		no
sex	Feature	Binary	Sex	student's sex (binary: 'F' - female or 'M' - male)		no
age	Feature	Integer	Age	student's age (numeric: from 15 to 22)		no
address	Feature	Categorical		student's home address type (binary: 'U' - urban or 'R' - rural)		no
famsize	Feature	Categorical	Other	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)		no
Pstatus	Feature	Categorical	Other	parent's cohabitation status (binary: 'T' - living together or 'A' - apart)		no
Medu	Feature	Integer	Education Level	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)		no
Fedu	Feature	Integer	Education Level	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)		no
Mjob	Feature	Categorical	Occupation	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')		no
Fjob	Feature	Categorical	Occupation	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')		no
reason	Feature	Categorical		reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')		no
guardian	Feature	Categorical		student's guardian (nominal: 'mother', 'father' or 'other')		no
traveltime	Feature	Integer		home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)		no
studytime	Feature	Integer		weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)		no
failures	Feature	Integer		number of past class failures (numeric: n if 1<=n<3, else 4)		no
schoolsup	Feature	Binary		extra educational support (binary: yes or no)		no
famsup	Feature	Binary		family educational support (binary: yes or no)		no
paid	Feature	Binary		extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)		no
activities	Feature	Binary		extra-curricular activities (binary: yes or no)		no
nursery	Feature	Binary		attended nursery school (binary: yes or no)		no
higher	Feature	Binary		wants to take higher education (binary: yes or no)		no
internet	Feature	Binary		Internet access at home (binary: yes or no)		no
romantic	Feature	Binary		with a romantic relationship (binary: yes or no)		no
famrel	Feature	Integer		quality of family relationships (numeric: from 1 - very bad to 5 - excellent)		no
freetime	Feature	Integer		free time after school (numeric: from 1 - very low to 5 - very high)		no
goout	Feature	Integer		going out with friends (numeric: from 1 - very low to 5 - very high)		no
Dalc	Feature	Integer		workday alcohol consumption (numeric: from 1 - very low to 5 - very high)		no
Walc	Feature	Integer		weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)		no
health	Feature	Integer		current health status (numeric: from 1 - very bad to 5 - very good)		no
absences	Feature	Integer		number of school absences (numeric: from 0 to 93)		no
G1	Target	Categorical		first period grade (numeric: from 0 to 20)		no
G2	Target	Categorical		second period grade (numeric: from 0 to 20)		no
G3	Target	Integer		final grade (numeric: from 0 to 20, output target)		no

Exploring the data

Two different datasets have been chosen for the project and the goal is to extract different knowledge from these datasets. Also it is possible to validate and compare the results in similar features.

1. The first dataset “Student Performance in Exams” (Dataset 1) https://www.kaggle.com/datasets/spscientist/students-performance-in-exams
   - less variables (21) but more rows (6600) .
2. The second dataset “Student Performance” (Dataset 2) https://archive.ics.uci.edu/dataset/320/student+performance
   - More variables (33) and a merged dataset of 1000 rows.

We started to gather our initial ideas by producing a simplified correlation matrix of the most impactful variables of the datasets. Visualization (Dataset 1) :



Correlation Matrix of Numerical Features

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
age	1.00	-0.16	-0.16	0.07	-0.00	0.24	0.05	0.02	0.13	0.13	0.12	-0.06	0.18	-0.06	-0.14	-0.16
Medu	-0.16	1.00	0.62	-0.17	0.06	-0.24	-0.00	0.03	0.06	0.02	-0.05	-0.05	0.10	0.21	0.22	0.22
Fedu	-0.16	0.62	1.00	-0.16	-0.01	-0.25	-0.00	-0.01	0.04	0.00	-0.01	0.01	0.02	0.19	0.16	0.15
traveltime	0.07	-0.17	-0.16	1.00	-0.10	0.09	-0.02	-0.02	0.03	0.14	0.13	0.01	-0.01	-0.09	-0.15	-0.12
studytime	-0.00	0.06	-0.01	-0.10	1.00	-0.17	0.04	-0.14	-0.06	-0.20	-0.25	-0.08	-0.06	0.16	0.14	0.10
failures	0.24	-0.24	-0.25	0.09	-0.17	1.00	-0.04	0.09	0.12	0.14	0.14	0.07	0.06	-0.35	-0.36	-0.36
famrel	0.05	-0.00	-0.00	-0.02	0.04	-0.04	1.00	0.15	0.06	-0.08	-0.11	0.09	-0.04	0.02	-0.02	0.05
freetime	0.02	0.03	-0.01	-0.02	-0.14	0.09	0.15	1.00	0.29	0.21	0.15	0.08	-0.06	0.01	-0.01	0.01
goout	0.13	0.06	0.04	0.03	-0.06	0.12	0.06	0.29	1.00	0.27	0.42	-0.01	0.04	-0.15	-0.16	-0.13
Dalc	0.13	0.02	0.00	0.14	-0.20	0.14	-0.08	0.21	0.27	1.00	0.65	0.08	0.11	-0.09	-0.06	-0.05
Walc	0.12	-0.05	-0.01	0.13	-0.25	0.14	-0.11	0.15	0.42	0.65	1.00	0.09	0.14	-0.13	-0.08	-0.05
health	-0.06	-0.05	0.01	0.01	-0.08	0.07	0.09	0.08	-0.01	0.08	0.09	1.00	-0.03	-0.07	-0.10	-0.06
absences	0.18	0.10	0.02	-0.01	-0.06	0.06	-0.04	-0.06	0.04	0.11	0.14	-0.03	1.00	-0.03	-0.03	0.03
G1	-0.06	0.21	0.19	-0.09	0.16	-0.35	0.02	0.01	-0.15	-0.09	-0.13	-0.07	-0.03	1.00	0.85	0.80
G2	-0.14	0.22	0.16	-0.15	0.14	-0.36	-0.02	-0.01	-0.16	-0.06	-0.08	-0.10	-0.03	0.85	1.00	0.90
G3	-0.16	0.22	0.15	-0.12	0.10	-0.36	0.05	0.01	-0.13	-0.05	-0.05	-0.06	0.03	0.80	0.90	1.00

Data Quality

1. Dataset 1 descriptions file already has a field that states that none of the variables have missing values. A double check proved it to be true. No incorrect values were detected by manual inspection among the ranges of allowed values.
2. Dataset 2 contains about 300 rows with missing values, that were removed during the data cleaning process. We applied techniques to ensure that all the rest of the values within the fields fall into correct ranges.

Task 4. Planning your project

Project plan

Task nr	Description	Contribution hours	Methods / tools	Comments
1	HW10	Anna: 3	Dividing the	

		Johan: 3 Hendrik:	homework tasks equally	
2	Exploration of the data	Anna: 9 Johan: 5 Hendrik:	Analyzing our original dataset for key factors to success, patterns and more. Searching for and combining related datasets to provide further details regarding a factor.	
3	Data visualization	Anna: 10 Johan:1 Hendrik:	Using plotting techniques to test our hypothesis and to visualize our findings to present during poster session	
4	Machine learning models	Anna: 2 Johan:16 Hendrik:	Tools: Jupyter notebook, pandas, sklearn, plotnine	
5	Project poster	Anna: 6 Johan: 5 Hendrik:	Creating an application that can be used to access our machine learning model in order to get an estimate of your test scores. Designing a poster that helps to convey our goals, work process and result in a way that is easy to understand and engaging.	