**Data structure**

**Data\_output sheets of the google file**

**https://docs.google.com/spreadsheets/d/1ShUilbDAzeq2XgP9ALpVX2jTjYU9NBDB\_CNj0qsg4Uo/edit#gid=1647678448**

In the excel file, data relevant to students’ math tests have been collected. For each test, the following info have been saved: data of the student performing the test, questions asked and relevant student’s answers, timestamps of all the events in the prototype (timestamps of buttons’ pressing and timestamps of answers’ clicking). Each test is composed by 5 math questions and 4 satisfaction questions.

**One row of the excel table**:

in the first 4 columns there are info relevant to the student (name, surname, university, email); in the following 4 columns there are the indexes of the answers to the final satisfaction questionnaire (each answer can be an index between 0 and 4, where 0 is the first answer and so on). In the following 10 columns there are, alternating, question and relevant user’s answer ids: each id corresponds to the index of the question/answer in the data\_input sheet (indexes start from 0 and the first row of the data\_input sheet is the header so it is ignored when reading the file); for example question index 6 will be question at row 8 of the data\_input sheet and answer index 2 will be that in column D. Therefore, each answer can be an index between 0 and 4, where 0 is the first answer and so on, or None in case of no answer (skipped question), while the indexes of the questions can be a number between 0 to 39, without repetition in the same test’s session. These 10 columns values are initialized with -1.

In the following columns there are the timestamps of all the events in the prototype:

- timestamps of buttons’ pressing and timestamps of answers’ clicking.

**Timestamps’ structure**:

identifying\_number#type\_of\_event#timestamp.

Identifying numbers: -2 -> begin of login phase, 0->pressing of start test/take new test buttons, 1 -> event relevant to question 1 (it could be pressing of confirm/back/skip buttons and clicks on question 1 answers), 2 -> event relevant to question 2 (it could be pressing of confirm/back/skip buttons and clicks on question 2 answers), etc. till 5, 6 -> pressing of submit button or back button because the user doesn’t answer to all questions, 7-> pressing of the button to begin satisfaction questionnaire, e1-> event relevant to question 1 of satisfaction questionnaire (it could be pressing of confirm button or clicks on answers), e2-> event relevant to question 2 of satisfaction questionnaire (it could be pressing of confirm button or clicks on answers), e3-> event relevant to question 3 of satisfaction questionnaire (it could be pressing of confirm button or clicks on answers), e4-> event relevant to question 4 of satisfaction questionnaire (it could be pressing of confirm button or clicks on answers).

type\_of\_event: pressing of confirm/back/skip button, click on a question, pressing of start test/take new test button.

Timestamp: ms elapsed from January 1970 till the click of the button/answer.

Refer to “*partner\_output\_unige”* sheet, to verify the output generated launching our code.

In Sheet "unige\_data\_processed", you can find the tests performed by different students which are used to train the model.

In Sheet “Data\_input “, you can find the list of the 40 possible questions for the test, and, for each question, you have the 5 possible answers, the index of the correct answer, the difficulty of the question and the related keywords.

**user\_features\_df.pickle:**

Data from the informative questionnaire filled in by each student, already processed.

These are the info saved: mail, name, gender, year of birthday, country, high school, high school final grade, university course, university year, if he/she is a working student or not, individual or team work preferences, learning methodology, learning style, hobbies, love for math (1 to 5), student’s personal evaluation in math (1 to 5), home study hours, number of passed subjects, average of the positive grades in the previous semester, highest grade in the current semester, average grade in a Mathematics test in the current semester, highest grade in a Mathematics test in the current semester, average grade in a Linear Algebra test in the current semester, highest grade in a Linear Algebra test in the current semester, additional student's information.

**Prototype sheet of the google file**

**https://docs.google.com/spreadsheets/d/1uiH0bhM4z2nqERrlO1ReEp2yCo0nJ6nBBcM-kqrrNvA/edit#gid=0**

In this file, there are the statistics for each of the 40 questions: percentage of times the question has been correctly answered, average time spent on the question by all users, average number of times the question has been skipped, average number of clicks on the question’s answers. Between these columns, there is a column, which repeated itself, with the number of times each question is appeared in tests.

**Users sheet of the google file https://docs.google.com/spreadsheets/d/1uiH0bhM4z2nqERrlO1ReEp2yCo0nJ6nBBcM-kqrrNvA/edit#gid=0**

In this file, for each user (identified by the e-mail), it is registered the percentage of times the user has correctly answered a question out the total number of answered questions.