Project Name: **Project B1: UT STUDENT PROFILING**
Repository: **https://github.com/mbz4/IDS-PROJECT-2021-B01**

Topic: **UT-CS Student profiling based on Moodle log data**
Team: **Matevž Zorec, Farnaz Baksh**

Dataset B01 - Student Profiling (SIZE ~54MB): complete log files of activities of roughly 342 students + instructors, at the Computer Programming course in Moodle during fall semester 2020, totalling about 1M entries, split into 20 spreadsheets. (we already have this data in our storage) [logs_course-code_date&time(when downloaded from Moodle)-lab-group]

Dataset Grade Results (SIZE KB 141): complete gradebook of the course at the end of the fall 2020 semester (we already have this data in our storage) [02-05 Hinded(Results).xlsx]

Goal 1: identify students who may be struggling based on their early activities (from activity logs)

Goal 2: predict the student' final grades in the course (from activity logs)

Goal 3: discover common activity patterns of the students and create profiles of typical students (from activity logs).

Resources:
[1] https://docs.moodle.org/311/en/Logs

[2] This is a topical project from the Educational domain from this spreadsheet: https://docs.google.com/spreadsheets/d/1M0JmYdonHGMyjYuwx0xq2CO7AqNoQ7OMsuGzktoAZ8g/edit#gid=323246102

Business Understanding

NB! Don't forget to mention your project title and team members at the beginning of the report.

Developing a business understanding within CRISP-DM consists of four tasks: identifying your business goals, assessing your situation, defining your data-analysis, data-mining or machine learning goals and producing your project plan. For this exercise, please, develop a business understanding of your project. According to CRISP-DM, you should report the following:

- **Identifying your business goals**

Focusing on developing a solution one day which may help students with learning impairments cope better when it comes to learning online, this project represents an important milestone on that journey. Understanding the factors well, when it comes to online learning is important as it lays a foundation for continued understanding of difficulties and common issues in dealing with online learning for the learning impaired.

Predicting student performance using moodle log data. ⇒ the business goal is to see what impacts student performance the most from just behavioral data in interacting with this online learning platform. Additionally, predict what the final student grade would be based on early efforts students logged on the platform.

    - *Background*

Moodle is an online learning platform commonly used by schools globally. It has a logging system which tracks a variety of different events. This system can generate and output csv log files that can be used in further, such as this project. Moodle also ships their own log analysis toolkit but for this project, we are to develop our own.

    - *Business goals*

The University would like to streamline and perhaps automate detecting students who may be struggling early on in the semester so as to improve their overall performance. This represents a key strategic optimisation solution.

    - *Business success criteria*

a) Predict student performance issues
b) Recognize if a student is struggling based on just a few early submission scores
c) Prepare a map of most common indicators of students that are struggling and need help

If we were to successfully train a model which can predict student outcome and performance as well as provide a map of most common indicators of student performance decrease, we would have succeeded in meeting our business success criteria.

- **Assessing your situation**

Presently, we are time constricted and have to modify our methodologies further as well as prepare this homework.

    - Inventory of resources

We have adequate computing resources available (including a relatively capable laptop GPU) but do not have prerequisite packages installed to utilise it, to supposedly run model training on the GPU (up to 40x time save). We do not have adequate time to fully immerse ourselves in this project and therefore the final outcome may be only partially meeting initial requirements.

    - Requirements, assumptions, and constraints

We require more time to commit and assume that within 1 week we will have to output significant progress for 4 different course projects simultaneously, including this one. This situation was not predictable at the start of the semester. Additionally, we are constrained in the fact neither of us speaks Estonian while the dataset is provided in Estonian (at least column names), the dataset provider did equip us with a translation and meaning interpretations, thankfully.

We would appreciate any support and time TAs could spare when we hit hurdles we can't cope with.

    - Risks and contingencies

A key risk is that the dataset analysis is too time consuming or that perhaps we make misleading assumptions. A key contingency in this case is an abundance of sanity checks as well as communicating with the dataset provider.

    - Terminology

Dataset, model training, data analysis, hyper parameter tuning, elastic net, communication, lackoftimeomgimpanickingrn.

    - Costs and benefits

This will cost us both several nights of sleep and deteriorate further our overall mental and physical well-being after an already exhausting breakfree non-stop grind semester.

In retrospect, we stand to gain a multitude of benefits: we learned basic data science and machine learning concepts and even applied them in a structured project setting. This is amazing and fantastic value.

- **Defining your data-mining goals**

We need to understand our data thoroughly.

    - Data-mining goals

Understand data structure, hierarchy, file organization. In this phase, we have to extract a set of features we identified as key in training the most successful model possible.

- Data-mining success criteria

If we evaluate the accuracy and AUC score of our trained model to be acceptable enough, we can also consider that our data mining has been successful.

Please, follow this given structure and cover all these aspects in your report. Consult this PDF-file with a chapter on Embracing the Data-Mining Process for more information on each of the deliverables. Keep the report concise and feel free to state that some aspect is not relevant in your project. If your project is not meant to benefit a 'business', then please specify who will benefit from the project and perform business understanding from their perspective. For instance, this could be either one or multiple individuals, organizations, or societies. Please focus on the goals that you plan to directly contribute to, not on the generic goals (like making the world a better place).

The report of task 2 should be 400-800 words.

Data Understanding

Data understanding within CRISP-DM consists of performing four tasks: gathering data, describing data, exploring data and verifying data quality. For this exercise please develop a data understanding of your project. Report the results of the tasks according to the following structure:

- Gathering data

We have acquired all project relevant data, we have it in our storage and we have already spent time assessing its structure and got acquainted with the moodle logging format as well as discussed in brief about how moodle themselves process log data and analyze it with the stock model they ship.

    - Outline data requirements

The data needs to be prepared such that key features are extracted. Parsing needs to be accurate and parsing functions may only be developed using copycat toy datasets.

    - Verify data availability

Data is readily available using UT owncloud service.

    - Define selection criteria

Our current assumption (pre thorough analysis) is that we will likely limit the dataset to the first month and use that in training a model that predicts students final grade.

- Describing data

This is the log data from the Computer Programming course in fall 2020. Since the log is big, it is split into 20 xlsx-files, according to lab groups.

We also have a complete student gradebook file and we will use it in training and verification of our trained models.The semester started on 31.08.2020, the students were added to Moodle on 1.09.2020, the semester officially ended on 31.01.2021. Therefore we will limit the dataset to the first month of the semester in order to train a model that predicts just the student grade based on early activities on moodle.

- Exploring data

* Aeg (Time)

* Kasutaja t�isnimi (User full name)

* M�jutatud kasutaja (Affected user)

* S�ndmuse kontekst (Event context)

- Ülesanne: Assignment submission tool without autochecker

* Komponent (Component) and Sündmuse nimi (Event name)

* Kirjeldus (Description)

* Päritolu (Origin)

* IP-aadress (IP address)

It is likely most of these will not be key features in the final model training solution.

- Verifying data quality

The data we have acquired appears to be auto generated and has no faults or errors contained within whatsoever. We may intervene and rename some of the provided estonian column names with translated english counterparts.

Consult the above-given book chapter to understand what is expected under all these deliverables. Take inspiration from when describing and exploring the data. As a result of this exercise, you should have gathered and understood the data. You should have decided which parts of the data you are potentially going to use and understood the meaning of all fields within these parts. Note that data cleaning is part of the data preparation step in CRISP-DM but you might choose to do some of it already during this task.

The report of task 3 should be 400-800 words.

Project Plan

- Make a detailed plan of your project with a list of tasks. There should be at least 5 tasks. Specify how many hours each team member is going to contribute to each task.

1) Student profiling using Moodle log data is not a new idea. It has been done before therefore we will spend time researching how others solved this problem before us. As a result we will try to collect as much inspiratory and exemplary material as we can, keeping it concise and clear to understand. **[Matevz: 6hrs] [Farnaz: 6hrs]**
2) In hoping to narrow down the key solutions we found relevant in our research from task 1, we will again examine the data we received to confirm what steps to take in task 2. It is highly likely that we will begin by creating an abstractable jupyter script which will parse thru a copied, subtracted test test dataset. The scraped info will be passed thru pandas as a dataframe and examined. We hope to gather key insight from this task about how the data we have on student moodle log

activities compares to findings from others in available research we covered. **[Matevz: 6hrs] [Farnaz: 6hrs]**

3) In task 2 we finished researching methods, and already started working with the dataset. In this task we will have to attempt to draw our first real assumptions about the data we have. Based on these assumptions and experience from research in tasks 1 & 2, we will setup a set of functions, each containing a script-standardised call for a different model training setup set. This setup will not be final as we have not yet decided on what the features and labels should be, but have already realized what the general outcome is (provided in project assignment instructions: predict student grade (0-100 ⇒ A-F))... The reason for putting the idea of using different models is that we want to perform hyper parameter tuning - we know already we will be relying almost solely on Random Forest Classifier, Decision Tree, TensorFlow or XGBoost for this project, based on research we have already conducted. [**Matevz: 7hrs] [Farnaz: 7hrs]**
4) At this point we have abstractified, parsed and setup several model functions to run the dataset thru. We must continue analysing the datasets and draw more conclusions. Additionally, for this task we should prepare the data for training, then train the models. After training the models, we also need a function which outputs for each student the predicted outcoming grade score or perhaps even the associated final grade. **[Matevz: 7hrs] [Farnaz: 7hrs]**
5) We must analyze the results, confirm our initial assumptions and check with task 1 research if our method and results are alright (sanity check). Then we must prepare a concise report and presentation video and submit it. **[Matevz: 7hrs] [Farnaz: 7hrs]**

- List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify.

+ Research
+ Parsing
+ Data cleaning
+ Analysis
+ Multiple model training
+ Abstraction
+ Elastic net analysis
+ Hyperparameter tuning
+ Models: Random Forest, Decision Tree, XGBoost, TensorFlow
+ Dataset toy set extraction
+ Common sense
+ Sanity checking
+ Feedback from TAs

The report of task 4 should be 100-300 words.