Predicting of Students' Droppput

Horizon Europe
Data Management Plan

01 March 2023

Data Management Plan created in Data Stewardship Wizard «ds-wizard.org»
using Common DSW Knowledge Model v2.4.4 (dsw:root:2.4.4).

<table>
<tr><th colspan="3">HISTORY OF CHANGES</th></tr>
<tr><th>Version</th><th>Publication date</th><th>Changes</th></tr>
<tr><td colspan="3"><i>There are no named versions</i></td></tr>
</table>

Contributors

The following contributors are related to the project of this DMP:

- **Yingyue Jiang**
  y.jiang.34@student.rug.nl
  Roles: *Contact Person*, *Data Collector*, *Editor*, *Researcher*
  Affiliation: *University of Groningen*

Projects

We will be working on the following project and for those are the data and work described in this DMP.

Predicting Student's Dropout

Acronym:	*PSD*
Start date:	*2023-03-01*
End date:	*2023-03-31*
Funding:	*Rijksuniversiteit Groningen: grant number not yet given*

Student dropout is a critical factor in the education system that have long been studied by researchers. Dropout rates have significant negative consequences, including lost educational opportunities, decreased earning potential, and an increased likelihood of unemployment. By using census data and administrative data, and analyzing a range of factors, including academic performance, behavior, and socio-economic status, we attempt to be effective in predicting students at risk of dropping out.

1. Data Summary

Data formats and types

We will be using the following data formats and types:

- **ipynb**

  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

2. FAIR Data

2.1. Making data findable, including provisions for metadata

- **higher education institution dataset** (not published)

There are no 'Minimal Metadata About ...' (MIA...) standards for our experiments. However, we have a good idea of what metadata is needed to make it possible for others to read and interpret our data in the future.

We will use an electronic lab notebook to make sure that there is good provenance of the data analysis.

The provenance will be captured using W3C PROV.

We made a SOP (Standard Operating Procedure) for file naming. We will be keeping the relationships between data clear in the file names. All the metadata in the file names also will be available in the proper metadata.

2.2. Making data accessible

We will be working with the philosophy *as open as possible* for our data.

All of our data can become completely open immediately.

Limited embargo will not be used as all data will be opened immediately.

Metadata will be openly available including instructions how to get access to the data. Metadata will not be available in a form that can be harvested and indexed.

For our produced data, conditions are as follows:

- **higher education institution dataset** (not published)

2.3. Making data interoperable

We will be using the following data formats and types:

- **ipynb**

  It is a standardized format.

We will be using the following standards (encodings, terminologies, vocabularies, ontologies):

- **python**

2.4. Increase data re-use

The metadata for our produced data will be kept as follows:

- **higher education institution dataset** (not published) – This data set will be kept available as long as technically possible. – The metadata will be available even when the data no longer exists.

As stated already in Section 2.2, all of our data can become completely open immediately.

We will be archiving data (using so-called *cold storage*) for long term preservation already during the project. The data are expected to be still understandable and reusable after a long time.

To validate the integrity of the results, the following will be done:

- We will run a subset of our jobs several times across the different compute infrastructures.
- We will run part of the data set repeatedly to catch unexpected changes in results.

3. Other research outputs

We use Data Stewardship Wizard for planning our data management and creating this DMP. The management and planning of other research outputs is done separately and is included as appendix to this DMP. Still, we benefit from data stewardship guidance (e.g. FAIR principles, openness, or security) and it is reflected in our plans with respect to other research outputs.

4. Allocation of resources

FAIR is a central part of our data management; it is considered at every decision in our data management plan. We use the FAIR data process ourselves to make our use of the data as efficient as possible. Making our data FAIR is therefore not a cost that can be separated from the rest of the project.

We will be archiving data (using so-called 'cold storage') for long term preservation already during the project.

None of the used repositories charge for their services.

Yingyue Jiang is responsible for finding, gathering, and collecting data.

To execute the DMP, no additional specialist expertise is required.

We do not require any hardware or software in addition to what is usually available in the institute.

5. Data security

Project members will not store data or software on computers in the lab or external hard drives connected to those computers. They can carry data with them on encrypted data carriers and password-protected laptops. All data centers where project data is stored carry sufficient certifications. All project web services are addressed via secure HTTP (https://...). Project members have been instructed about both generic and specific risks to the project.

The possible impact to the project or organization if information is lost is small. The possible impact to the project or organization if information is leaked is small. The possible impact to the project or organization if information is vandalised is small.

All personal data will be anonymized as early as possible.

The archive will be stored in a remote location to protect the data against disasters. The archive need to be protected against loss or theft. It is clear who has physical access to the archives.

6. Ethics

For the data we produce, the ethical aspects are as follows:

- **higher education institution dataset**
    - It does not contain personal data.
    - It does not contain sensitive data.

Data we collect

We will not collect any data connected to a person, i.e. "personal data".

7. Other issues

We use the Data Stewardship Wizard with its *Common DSW Knowledge Model* (ID: dsw:root:2.4.4) knowledge model to make our DMP. More specifically, we use the https://researchers.ds-wizard.org DSW instance where the project has direct URL: https://researchers.ds-wizard.org/projects/44f4d8fa-1a7c-4ac0-bce9-31e42d2769d2.

We will be using the following policies and procedures for data management:

- **RUG Policy**
  https://www.rug.nl/research/research-data-management/policy/ug-rdm/
  Its purpose is to ensure innovative research and research integrity.