Open Educational Datasets from the DiSEA Project

Teodora Dogaru, Julian Fröhlich, Raj Waghela, Agathe Merceron, Petra Sauer
Berliner Hochschule für Technik, Berlin, Germany
{merceron,sauer}@bht-berlin.de

Abstract: These datasets were compiled in the DiSEA project: Digital Study Programmes: Analysis of Success and Dropout Factors, see https://disea-projekt.de/. This project has been funded by the German Ministry of Education and Research (Bundesministerium für Bildung und Forschung - BMBF) in the line "Studienerfolg und Studienabbruch II", grant number 16PX21001A.

File *mdl_tasks.csv*, the task dataset

Contains 616 347 rows of students' interactions with the learning platform Moodle. The users' interactions as stored in the Moodle-Logs have been transformed into tasks, see the following papers for explanations on this transformation:

- Dogaru, Teodora; Götze, Nora; Rotelli, Daniela; Berendsohn, Yoel; Merceron, Agathe; Sauer, Petra (2023). Task Definition in Big Sets of Heterogeneously Structured Moodle LMS Courses. *21. Fachtagung Bildungstechnologien (DELFI).* DOI: 10.18420/delfi2023-71
- Rotelli, Daniela & Monreale, Anna. (2022). Time-on-Task Estimation by data-driven Outlier Detection based on Learning Activities. in *Proceedings of the 12th International Conference on Learning Analytics and Knowledge (LAK22)*. ACM. 336-346. 10.1145/3506860.3506913.
- Rotelli, Daniela & Monreale, Anna. (2023). Processing and Understanding Moodle Log Data and Their Temporal Dimension. *Journal of Learning Analytics*. 1-23. 10.18608/jla.2023.7867.

Each row contains the following information:

- userid: The ID of the user who performed the task. This ID is unique and remains the same for this specific user across all courses.
- redef_comp: The type of task. Different activities can be described by the same task. For example, "Forum" is the task for both reading and writing a post in the forum. Such a task can be made up of several events.
- n_events: The number of events that make up the task.
- courseid: The ID of the Moodle course in which the task took place. If a 1, 0, or -1 is entered here, the task takes place outside of a course, e.g. in the dashboard or in the user settings. This ID refers to a course in a specific semester. If a course with the same content and the same name is offered again in a later semester, this new course is given its own new ID.
- duration: The total duration of the task.
- timecreated: The exact time at which the task was started.

The data covers 22 months, from 03.12.2021 to 24.09.2023. It contains information about the tasks of 549 different user IDs in 1243 courses.

The values that red_comp can take are briefly explained in the following table.

redef_comp	**Explanation**
Assignment	An assignment was viewed or handed in
Attendance	Own attendance in the course
BigBlueButton	Click the link to BBB-Meeting
Book	Interacted with a Book-Ressource
Calendar	Calendar visited or edited
Chat	Interaction with a chat
Choice	Interaction with a Moodle-Choice activity
Course_List	Overview of all courses
Course_Home	Course visited
Data	Clicked on files (e.g. zip folders)
Etherpad	Interaction with Etherpad
Feedback	Interaction with Feedback
Forum	Visited a forum or posted something
Glossary	Visited the Glossary in a course
Grades	Interacted with grades
Group	Viewed groups or acted in own group
Groupselect	Created own group or joined a group
H5P	Visited or interacted with interactive content created with H5P
Lesson	Learning activities of a Lesson completed or viewed
Page	Clicked an Internet or content page within Moodle
Participant_Profile	Visited another user's profile
Quiz	Quiz visited or completed
Resource	Accessed PDFs in Moodle
Scheduler	Schedule appointments with individual course participants or groups
URL	URL clicked in the course

User_Profile	Interacted with own user profile
Wiki	Wiki page visited or edited in a course
Workshop	Interaction with a peer-assessment tool
Yulinc	Click the link to Yulinc-Meeting

File connected_course_vector.csv, the dataset of Moodle activities connected with grades

The data represents students' learning behavior in the Moodle platform in the Media Informatics Online degree program offered in the VFH[1] - a network of German universities of applied sciences. The data is available in CSV format and contains a total of 1985 rows, with each row representing a student's behavior for one course over a semester.

Each row contains the following information:

- student: the ID of the student who took the course. This ID is unique and remains the same for this specific student across all courses.
- course_id: The ID of the course in Moodle. This ID is unique and remains the same for this particular course across all students.
- course_title: The name of the course according to the study regulations.
- grade: The grade with which the course was completed. Possible grades are: 1.0, 1.3, 1.7, 2.0, 2.3, 2.7, 3.0, 3.3, 3.7,4.0, 5.0 and -1.0. The grade -1.0 means that the student has taken the course but there is no final grade. In Germany, 1.0 is the best possible mark, 4.0 is pass and 5.0 is fail.
- semester: The semester in which the course took place. 20221 is the summer semester of 2022, 20222 is the winter semester of 2022 and 20231 is the summer semester of 2023.
- university: The university at which the course was taken. The names of the two universities for whom we got the grades of the final exam were anonymized with A and B.
- w1_d1 to w23_d7: These columns represent the activity, i.e. the number of task events from the task dataset on one day, of a student for the specific course over a semester. w1_d1 is week 1, day 1 and w23_d7 is week 23, day 7. The start of a course is marked by the first activity that took place in the course.

The data covers a period of three semesters, from summer semester 2022 to summer semester 2023, and contains information on the behavior of 382 different user IDs in a total of 133 courses. Only courses that all students must complete according to the study and examination regulations were considered.

The list includes the following modules:

[1] www.vfh.de

- Semester 1: Computer Architecture and Operating Systems, Introduction to Computer Science, Principles of Mathematics, Principles of Programming 1, Communication, Leadership and Self-Management, Media Design 1.
- Semester 2: Principles of Programming 2, Media Design 2, Human-Computer Communication, Principles of Computer Networks, Relations and Functions, Theoretical Informatics.
- Semester 3: Algorithms and Data Structures, Computer Graphics, Database Management Systems, Multimedia Technology, Project Management, Web Programming.
- Semester 4: Introduction to Scientific Project Work, Principles of IT Security, Internet Applications for Mobile Devices, Internet Server Programming, IT Law, Software Engineering.
- Semester 5: Patterns and Frameworks.
- Semester 6: Business Administration.

Files
vektor_bachelor_medieninformatik_sem1.csv,
vektor_bachelor_medieninformatik_sem12.csv,
vektor_bachelor_medieninformatik_sem123.csv,
vektor_bachelor_medieninformatik_sem1234.csv, the dataset of academic performance of students aggregated into global features - Online Study Program “Media Informatics”

These files contain the vectors representing the academic performance of students from the “Media Informatics” online degree at two universities in a German network of universities of applied sciences (Virtuelle Fachhochschule (VFH), see https://www.vfh.de/). The degree comprises six semesters of full-time study and includes six compulsory modules per semester for the first four semesters.

As this is an online degree, most students study part-time as they often have other commitments, such as full-time work or caring for relatives. This means that they usually take fewer than six modules per semester, which differs from the original planning in the module handbook.

Students in this online degree have the following characteristics:

- They generally study longer than face-to-face students to complete their degree.
- They typically take fewer than six modules per semester.
- The number of students dropping out is higher than in traditional face-to-face degrees.

The data for these vectors was extracted between the summer semester 2014 and the summer semester 2023. In order to track the development of students throughout their studies, four vectors were calculated, each representing the students at the end of the 1st, 2nd, 3rd, and 4th semesters.

Description of the vector at the end of the 1st semester.

The vector at the end of the 1st semester represents each student with two demographic features (gender and university), four features that reflect their performance in the first semester, and a label.

1. gender: m = male, f = female.

2. university: A or B
3. no_passed_courses_1: Number of exams passed in the 1st semester.
4. no_failed_courses_1: Number of exams failed in the 1st semester.
5. avg_grade_passed_courses_1: Average grade of all passed exams in the 1st semester.
6. avg_grade_all_courses_1: Average grade of all exams (passed or failed) in the 1st semester.

In addition, the vector contains a label that reflects the student's degree status:

7. label: 1 = graduate (the student has graduated) and 0 = dropout (the student has not graduated).

Special case: No modules taken.

If a student has registered in a semester but has not taken any modules, the following values will appear in the vectorial representation:

1. gender
2. university
3. no_passed_courses_1: 0
4. no_failed_courses_1: 0
5. avg_grade_all_courses_1: 5.1 (specific value indicating a missing value)
6. avg_grade_passed_courses_1: 5.1
7. label

This special case makes it possible to represent students who have registered in a semester but have not taken any modules.

Description of the vector at the end of the 2nd semester.

In order to take the history into account, the values of the 1st semester were repeated and combined with the new values of the 2nd semester. This makes it possible to track the students' history over time.

The vector at the end of the 2nd semester represents each student by 10 features and a label, including eighth features that reflect their performance and progress in the first and second semesters.

The 10 features + label of the vector at the end of the 2nd semester are:

1. gender: see description above.
2. university: see description above.
3. no_passed_courses_1: Number of passed exams in the 1st semester.
4. no_failed_courses_1: Number of failed exams in the 1st semester.
5. avg_grade_all_courses_1: Average grade of all exams passed in the 1st semester.
6. avg_grade_passed_courses_1: Average grade of all exams (passed or failed) in the 1st semester.
7. no_passed_courses_2: Number of passed exams in the 2nd semester.
8. no_failed_courses_2: Number of failed exams in the 2nd semester.
9. avg_grade_all_courses_2: Average grade of all exams passed in the 2nd semester.
10. avg_grade_passed_courses_2: Average grade of all exams (passed or failed) in the 2nd semester.
11. label: see description above.

Extension of the vector for the 3rd and 4th semesters.

For the 3rd and 4th semesters, we proceeded in the same way as for the 2nd semester. At the end of the 3rd semester, a student is thus represented by 14 features and by 18 features at the end of the 4th semester plus the label.

Vector at the end of the 3rd semester

At the end of the 3rd semester, students are represented by 14 features that reflect their performance and progress in the first, second, and third semesters. These features include:

- The six features from the 1st semester
- The four academic features from the 2nd semester
- The four academic features from the 3rd semester
- The label is kept.

Vector at the end of the 4th semester

At the end of the 4th semester, students are represented by 18 features that reflect their performance and progress in the first, second, third, and fourth semesters. These features include:

- The six features from the 1st semester
- The four academic features from the 2nd semester
- The four academic features from the 3rd semester
- The four academic features from the 4th semester
- The label is kept.

Files sem1.csv, sem2.csv, sem3.csv, sem4.csv, the dataset of academic performance of students aggregated into global features - Online Study Program “Business Informatics”

These files contain the vectors representing the academic performance of students from the “Business Informatics” online degree at one university in a German network of universities of applied sciences (Virtuelle Fachhochschule (VFH), see https://www.vfh.de/). The general conditions given above for the “media Informatics” online degree hold here too.

The data for these vectors was extracted between the winter semester of 2007 and the summer semester of 2023. In order to track the development of students throughout their studies, four vectors were calculated, each representing the students at the end of the 1st, 2nd, 3rd, and 4th semesters.

Description of the vector at the end of the 1st semester.

The vector at the end of the 1st semester represents each student with four features that reflect their performance in the first semester and a label.

1. no_passed_courses_1: Number of exams passed in the 1st semester.
2. no_failed_courses_1: Number of exams failed in the 1st semester.
3. no_unattempted_exams_1: Number of courses enrolled in but without a grade (student did not sit the exam) in the 1st semester.

4. avg_grade_all_courses_1: Average grade of all exams (passed or failed) in the 1st semester.

In addition, the vector contains a label that reflects the student's degree status:

5. label: 1 = graduate (the student has graduated) and 0 = dropout (the student has not graduated).

Special case: No modules taken.

If a student has registered for a semester but has not taken any modules, the following values will appear in the vectorial representation:

1. no_passed_courses_1: 0
2. no_failed_courses_1: 0
3. no_unattempted_exams_1: 0
4. avg_grade_all_courses_1: 5.1 (specific value indicating a missing value)

This special case makes it possible to represent students who have registered in a semester but have not taken any modules.

Extension of the vector for the 2nd, 3rd, and 4th semesters.

For the 2nd, 3rd, and 4th semesters, we proceeded in the same way as for the study program “Media Informatics” and took into account the history. At the end of the 2nd semester, a student is thus represented by eight features and the label, by 12 features and a label at the end of the 3rd semester, and by 16 features at the end of the 4th semester plus the label.