

УДК 004.62/37.012.4

АНАЛИЗ РЕЗУЛЬТАТОВ ТЕСТИРОВАНИЯ БОЛЬШОЙ ВЫБОРКИ ОБУЧАЮЩИХСЯ С ПОМОЩЬЮ ИНСТРУМЕНТОВ DATA SCIENCE

А. Э. Бойко, аспирант, ассистент кафедры менеджмента и маркетинга РХТУ им. Д. И. Менделеева, Москва, Россия;

orcid.org/0009-0007-4006-4403, e-mail: boiko.a.e@muctr.ru

Т. В. Савицкая, д.т.н., профессор кафедры кибернетики химико-технологических процессов РХТУ им. Д. И. Менделеева, Москва, Россия;

orcid.org/0000-0001-9818-2196, e-mail: savitskaia.t.v@muctr.ru

Д. С. Лопаткин, к.э.н., доцент, заведующий кафедрой менеджмента и маркетинга РХТУ им. Д. И. Менделеева, Москва, Россия;

orcid.org/0000-0003-4260-4710, e-mail: lopatkin.d.s@muctr.ru

*Ввиду интенсификации процессов построения дистанционных курсов отечественными вузами с использованием LMS Moodle, где осуществляется сбор, хранение, мониторинг, и обработка образовательных данных, формирование единого подхода к реализации дескриптивной и диагностической учебной аналитики, способной удовлетворить задачи системы менеджмента качества (СМК) по части представления данных тестирования для целей внутреннего и внешнего контроля качества образования становится актуальной задачей. **Целью работы** является расширение возможностей мониторинга, представления и визуализации образовательных данных из LMS Moodle, связанных с различными формами тестирования обучающихся с помощью инструментов data science для оптимизации образовательных процессов и повышения качества обучения. **Результаты:** в работе рассмотрен процесс подготовки и дальнейшей обработки и визуализации данных о диагностическом тестировании обучающихся второго курса РХТУ им. Д.И. Менделеева (более 900 человек) по двум дисциплинам. Рассмотрены различные варианты агрегации данных: по группам на факультете, потоке, внутри образовательной программы, на уровне всего вуза. Предложены различные варианты визуализации результатов сопоставления успеваемости групп и фактической/списочной численности тестируемых; варианты и параметры визуализации для анализа «структуры» распределения баллов учебных групп на выбранном факультете и др. **Практическая значимость:** результаты исследования помогут преподавателям и администраторам LMS Moodle принимать более обоснованные решения, основанные на данных, а предложенные инструменты могут стать отправной точкой для реализации базовых адаптивных подходов в обучении.*

Ключевые слова: образовательная аналитика, система управления обучением, LMS Moodle, анализ данных, Python, библиотека Seaborn, библиотека Matplotlib, дескриптивная аналитика, диагностическая аналитика, высшее образование.

DOI: 10.21667/1995-4565-2023-85-103-116

Введение

Совершенствование реализации отдельных дисциплин, или даже целых образовательных программ, выработка адаптивных подходов к обучению, мониторинг успеваемости студентов, их вовлеченности, формирование индивидуальных образовательных траекторий в современном мире невозможны без сбора, хранения, мониторинга, обработки образовательных данных, осуществляющихся на уровне систем управления обучением (LMS).

В Российской Федерации вузы активно строят дистанционные курсы на базе LMS Moodle. Россия входит в топ 10 стран мира по числу сайтов, использующих Moodle в официальных образовательных структурах (37 %). Для сравнения в США – 34 %, в Италии – 32 %. Moodle особенно популярна в российских университетах, например, среди прочих ее используют Московский государственный университет имени М. В. Ломоносова (МГУ), Национальный исследовательский ядерный университет «МИФИ», Новосибирский Государствен-

ный Университет (НГУ) и Чувашский государственный университет (ЧГУ) [1]. В РХТУ им. Д. И. Менделеева также используется LMS Moodle.

Ввиду интенсификации использования LMS Moodle отечественными вузами для реализации смешанного формата обучения и построения дистанционных курсов, формирование единого подхода к реализации дескриптивной и диагностической учебной аналитики, способной удовлетворить задачи системы менеджмента качества (СМК) по части представления данных тестирования для целей внутреннего и внешнего контроля качества образования становится **актуальной задачей**. Вместе с этим расширение возможностей мониторинга, представления и визуализации образовательных данных, связанных с различными формами тестирования обучающихся, обуславливает необходимость практического внедрения учебной аналитики в образовательные процессы вузов.

В период ограничительных мероприятий в связи с предупреждением распространения COVID-19 большинство базовых дисциплин в РХТУ им. Д.И. Менделеева были оперативно перенесены полностью в дистанционный формат, что ввиду очень сжатых сроков могло сказаться на качестве обучения; поэтому по окончании адаптации ряда базовых дисциплин к дистанционному формату было принято решение о проведении диагностического контроля знаний обучающихся второго курса (математика, общая и неорганическая химия) и третьего курса (органическая и физическая химия).

Стоит отметить, что работ, посвященных процессу обработки образовательных данных из LMS Moodle практически нет, поскольку пользователи чаще всего обходятся встроенными в систему функциями, предположительно ввиду специфики применения самой платформы для реализации курсов на небольшое количество обучающихся. В [2] в кратких выводах для аналитического обзора отмечено, что в большинстве работ рассматриваются небольшие выборки обучающихся, что не ставит исследователей перед проблематикой совмещения, агрегации и отбора данных при работе с Moodle.

В данной работе рассмотрен процесс обработки образовательных данных для получения дескриптивной аналитики в разрезе факультетов и групп с использованием данных из LMS Moodle. Работа больше тяготеет к области «аналитики обучения», под которой подразумевается измерение, сбор, анализ и представление данных об обучающихся и образовательной среде, в которой оно происходит, с целью их понимания и оптимизации [3]; однако реализуется уже ставшими классическими в «интеллектуальном анализе образовательных данных» (educational data mining, EDM) и в науке о данных инструментами библиотек высокоуровневого языка программирования Python.

Теоретическая часть

Встроенных в Moodle аналитических инструментов достаточно, чтобы анализировать успеваемость на уровне группы, сопоставлять результаты нескольких групп между собой, анализировать статистику ответов (коэффициент сложности, дискриминации, стандартного отклонения, предполагаемого и действующего веса оценки, вероятности случайного угадывания), однако в базовой версии системы отсутствует возможность агрегации данных на уровне факультета (потенциально – на уровне одной образовательной программы), соответственно – получения статистики ответов по факультетам, или нескольким выбранным группам. Подобные ограничения побуждают на поиск удобных инструментов расширенной визуализации (помимо гистограмм) для сопоставления результатов нескольких учебных групп в рамках нескольких факультетов (или направлений подготовки).

В случае работы с небольшим количеством групп такие инструменты, действительно, могут быть излишни, однако для кафедр, реализующих базовые дисциплины для всего вуза (к примеру, на втором курсе обучения РХТУ им. Д. И. Менделеева на всех факультетах насчитывается 56 учебных групп, больше тысячи обучающихся) такие инструменты могут стать отправной точкой для реализации базовых адаптивных подходов.

Для решения поставленной задачи целесообразно рассмотреть возможность применения подхода, реализуемого с использованием инструментов EDM.

В [4] описан подход к формированию фрейма данных (датасета) из нескольких источников, он представлен на рисунке 1.

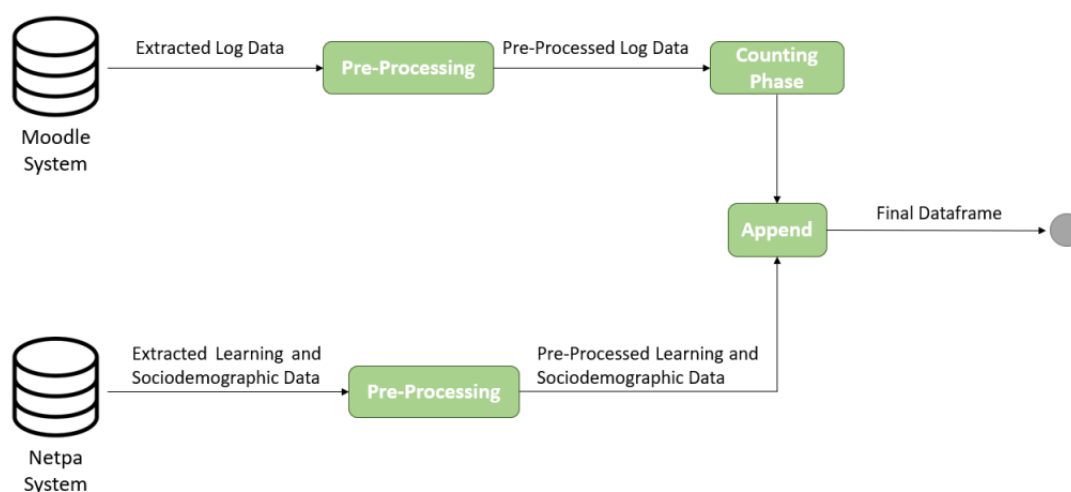


Рисунок 1 – Способ формирования фрейма данных для задач EDA из двух источников [3]

Figure 1 – Method of forming a data frame for EDA tasks from two sources

Автор использует выгрузки «цифровых следов» из базы данных системы Moodle и «Nepta System», содержащей социо-демографические данные, которые подвергает «препроцессингу» (очистка, замена дублей, изменение форматов, нормализация и стандартизация – в случае применения различных моделей машинного обучения для решения задач кластеризации, прогнозирования и др.).

Отечественным аналогом «Nepta System» нередко выступает система 1С «Студент» (которая используется в РХТУ им. Д.И. Менделеева), или данные из электронной информационно-образовательной среды (ЭИОС).

К примеру, в РХТУ им. Д.И. Менделеева для формирования и обновления списка групп обучающихся в LMS «Moodle», используются данные из 1С «Студент», которые ежедневно синхронизируются методом «коhort» для формирования т.н. «глобальных групп». Глобальные группы используются для организации миграции (перевод, отчисление, зачисление) обучающихся в системе.

На рисунке 2 в первой строке приведен пример «логирования» такой операции (пользователь с id «1» отчислил пользователя с id «a» с курса id «b» методом «коhort»). Существенным недостатком подобного решения для аналитика является удаление оценки обучающегося в случае перевода его на другой курс (вторая строка: пользователь с id «0» удалил оценку с id «c» для пользователя с id «d» для элемента оценивания с id «e»); в случае несвоевременной выгрузки результатов тестирования высока вероятность потерять все данные. Для восстановления оценки обучающегося его будет необходимо снова зачислить на курс, что в ручном режиме является затруднительной задачей.

The user with id '1' enrolled the user with id '25317' using the enrolment method 'cohort' in the course with id '11204'.
The user with id '0' deleted the grade with id '3492118' for the user with id '17357' for the grade item with id '9488'.

Рисунок 2 – Примеры логов LMS Moodle при синхронизации обучающихся с 1С «Студент» методом коhort и соответствующего удаления оценки студента с курса

Figure 2 – Examples of LMS Moodle logs when synchronizing students with 1С «Student» by cohort method and the corresponding removal of the student's assessment from the course

В работе [4] формирование фрейма данных происходит из логов системы для анализа поведения пользователей на курсе, однако для решения задач расширенного описательного анализа результатов массового тестирования (более тысячи обучающихся), вместо логов мо-

гут быть использованы выгрузки результатов тестирования в формате файлов с расширением. csv, – это файлы «оценки», «ответы», и «Export with groups and dates» (такой экспорт оценок появляется при установке одноименного плагина).

Аккумулируя все данные из форм трёх файлов, можно получить информацию о: личных данных (фамилия, имя, эл. почта, факультет, учебная группа), состоянии теста и времени его прохождения, детализации результатов тестирования (см. табл. 1), где 0 – отсутствие указанных в строке данных в выгрузке, 1 – наличие.

Таблица 1 – Данные из файлов «оценки», «ответы», и «Export with groups and dates»

Table 1 – Data from the «results: scores», « results: answers», «Export with groups and dates» files

Данные из файлов	Результаты: ответы	Результаты: оценки	Оценки: экспорт «Export with groups and dates»
Фамилия	1	1	1
Имя	1	1	1
Отчество	0	0	0
Адрес электронной почты	0	0	1
Отдел (факультет/кафедра)	0	0	1
Группа (учебная)	0	0	1
Состояние теста	1	1	0
Тест начат	1	1	0
Тест завершен	1	1	0
Оценка за тест общая	1	1	1
Текстовая формулировка каждого вопроса теста	1	0	0
Ответ обучающегося по каждому вопросу/выбранный вариант ответа	1	0	0
Правильный ответ по каждому вопросу	1	0	0
Балл обучающегося по каждому вопросу	0	1	0

Однако указанные данные распределены по всем трём файлам – в формах «ответы» и «оценки» отсутствует личная информация об обучающихся, в форме «Export with groups and dates» отсутствует информации о состоянии теста и детализация вопросов, а форма «ответы» не содержит баллов по каждому вопросу, что в случае оценивания в ситуации с разным удельным весом вопросов может стать существенным недостатком.

Кроме того, ввиду децентрализованного заполнения данных секретарями кафедр в 1С «Студент» отсутствует единообразие заполнения поля «Отдел (факультет/кафедра)», в связи с чем эти данные необходимо править вручную, кроме этого, списочная численность групп может формироваться недостоверно ввиду несвоевременного проведения данных о переводах/отчислениях.

Для получения общего фрейма данных необходимо использовать методы объединения данных (merge, join – для SQL/pandas, ВПР/ГПР (Vlookup / Hlookup – вертикальный/горизонтальный просмотр, поисковая функция) – для MS. excel), однако в указанных файлах отсутствует единая индексация, в связи с чем единственным релевантным индексом остается ФИО, в котором в нашем случае отсутствует отчество. Файлы «оценки», «ответы» в случае выгрузки в один день всё-таки имеют одинаковые индексы, поэтому они объединяются в первую очередь, а затем объединяются с файлом «Export with groups and dates». Вопрос индексации в случае объединения файлов по разным дисциплинам в единый фрейм данных стоит еще более остро и требует дополнительной проработки.

Экспериментальные исследования

Ввиду необходимости определения точной списочной численности обучающихся для целей дескриптивного анализа необходимо воспользоваться традиционным для вузов файлом

«контингента», который потребуется на этапе анализа по факультетам для сопоставления списочной и фактической численности тестируемых.

На рисунке 3 представим процесс подготовки и дальнейшей обработки и визуализации данных в целях расширенного описательного анализа результатов диагностического контроля обучающихся второго курса РХТУ им. Д.И. Менделеева, реализуемый в интерактивной облачной среде для работы с кодом с использованием библиотек Python: pandas и специализированных библиотек для визуализации seaborn и matplotlib.



Рисунок 3 – Процесс обработки образовательных данных по дисциплинам: «Математика», «Общая и неорганическая химия» (ОНХ)

Figure 3 – «Mathematics», «General and inorganic chemistry» disciplines: educational data processing

В результате препроцессинга, с учетом ручного удаления дублирующих строк и результатов тестирования обучающихся, у которых в файле «Export with groups and dates» отсутствовали значения в столбце «Отдел (факультет/кафедра)» и «группа», из исходных 879 попыток по математике и 815 попыток по ОНХ, был получен фрейм данных размером 21*790 (790 студентов, имеющих завершённые попытки теста по математике и/или по ОНХ), столбцы которого содержат следующие данные: фамилия + имя обучающегося, время на выполнение тестирования по математике в секундах, собственная или корпоративная электронная почта, факультет, группа, суммарный балл по математике из ста, полученный балл по каждому из 7-и вопросов, входящих в тест, время на выполнение тестирования по ОНХ в секундах, суммарная оценка по общей и неорганической химии (из 8-ми баллов, всего шесть вопросов, первый и пятый вопрос по два балла), полученный балл по каждому из 6-и вопросов, входящих в тест по ОНХ.

Потенциально пропуски данных являются поводом для применения различных методов их заполнения (среднее значение, мода, медиана) – для количественных данных; чтобы заполнить категориальные данные, в нашем случае – группу или факультет, потребуется развёртывание нейросети/алгоритма классификации, что само по себе является поводом для отдельной работы.

В файле «Export with groups and dates» располагаются столбцы, указанные в таблице 1, в них содержится около 3500 строк, соответствующих номинальному количеству обучающихся как со второго, так и с третьего курса Университета (оба учебных курса для удобства обработки данных тестировались внутри одного курса в LMS Moodle).

Сам процесс объединения вышеуказанных файлов в единый фрейм данных производится ввиду удобства манипуляции с ними функционалом библиотеки pandas¹, а также наглядных

¹ Pandas – это библиотека, написанная для языка программирования Python для манипулирования данными и их анализа. В частности, она предлагает структуры данных и операции для манипулирования числовыми таблицами и временными рядами. Одна из таких структур называется Data frame (фрейм данных), которая пред-

инструментов визуализации библиотек seaborn² и matplotlib³: countplot, barplot, histplot, heatmap, и др., уже ставших классическими при работе с любыми данными.

Указанные библиотеки для визуализации удобны в первую очередь тем, что не требуют ручного подсчета частот и в качестве аргумента функции принимают на вход массивы данных формата pd.array и pd.series, которые можно извлекать непосредственно из массива данных, с помощью функций для манипуляции данными: pandas.DataFrame.groupby⁴, pandas.DataFrame.aggregate⁵, pandas.DataFrame.query⁶, pandas.DataFrame.filter⁷ и/или функции для подсчета частот pandas.Series.value_counts⁸.

Указанные функции в контексте обработки образовательных данных в целях дескриптивного анализа используются в первую очередь для формирования срезов данных.

Начнём описание процесса обработки результатов с верхнеуровневого анализа для всей выборки в целом.

Одним из первых аспектов дескриптивного анализа, для которого не имеет смысла делать срезы – время тестирования (столбец получен подсчётом разницы между временем окончания тестирования и временем начала тестирования из формы «оценки» или «ответы»), см. рисунок 4. Для общей и неорганической химии (ОНХ) (6 заданий в тесте) оптимальное время на тестирование составляет 30 минут (большинство тестируемых завершает работу до 1800 секунд). По математике наибольшая плотность тестируемых приходится на 2800 секунд, что составляет примерно 47 минут (8 заданий в тесте). Графики были построены с помощью функции sns.distplot⁹, куда в качестве аргумента передается весь столбец с данными о времени тестирования.

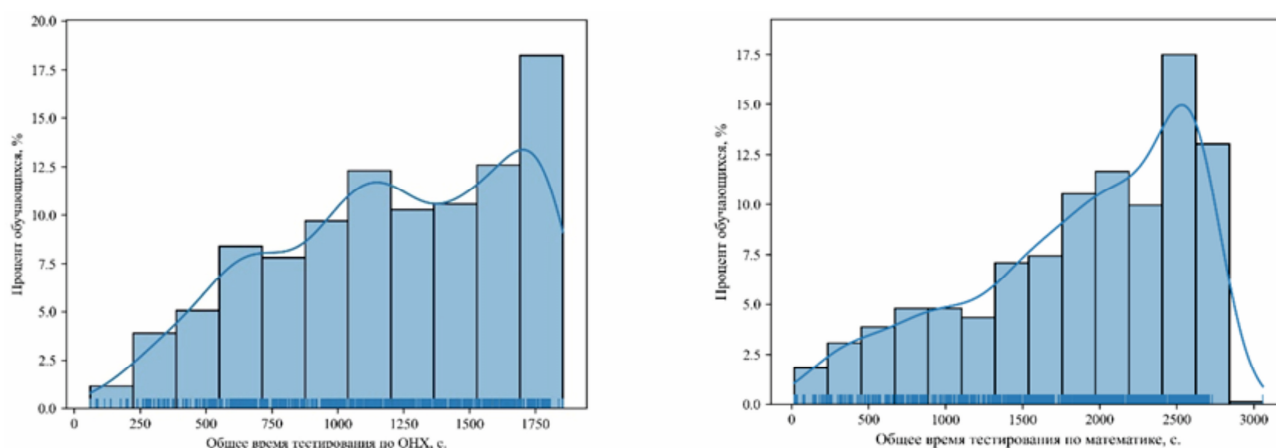


Рисунок 4 – Распределение времени тестирования обучающихся в целом по вузу по математике и ОНХ

Figure 4 – Distribution of testing time of students in general at the university in mathematics and ONX

Классическим инструментом «разведочного анализа данных» (exploratory data analysis) на уровне всей выборки также могут стать «тепловая карта» (корреляционная матрица, sea-

ставляет собой двумерную структуру данных, т.е. данные представлены в табличном виде по строкам и столбцам, как электронная таблица или таблица SQL (Pandas, 2021).

² Seaborn – это библиотека визуализации данных на Python. Она предоставляет высокоуровневый интерфейс для изображения наглядной и информативной статистической графики (Seaborn, 2021).

³ Matplotlib – библиотека на языке программирования Python для визуализации данных двумерной и трёхмерной графикой.

⁴ <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.groupby.html>

⁵ <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.aggregate.html>

⁶ <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.query.html>

⁷ <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.filter.html>

⁸ https://pandas.pydata.org/docs/reference/api/pandas.Series.value_counts.html

⁹ <https://seaborn.pydata.org/generated/seaborn.distplot.html>

born.heatmap¹⁰) корреляции. Общая карта корреляции для всего фрейма данных не показала связи между ответами обучающихся по разным дисциплинам, однако отбор столбцов, связанных с тестированием по одной дисциплине всё – таки наводит на некоторые гипотезы см. рис. 5. Тепловые карты построены с помощью инструмента sns.heatmap, в которые в качестве аргумента передаются фреймы данных с нужными столбцами, отобранные с помощью функции pandas.DataFrame.iloc¹¹.

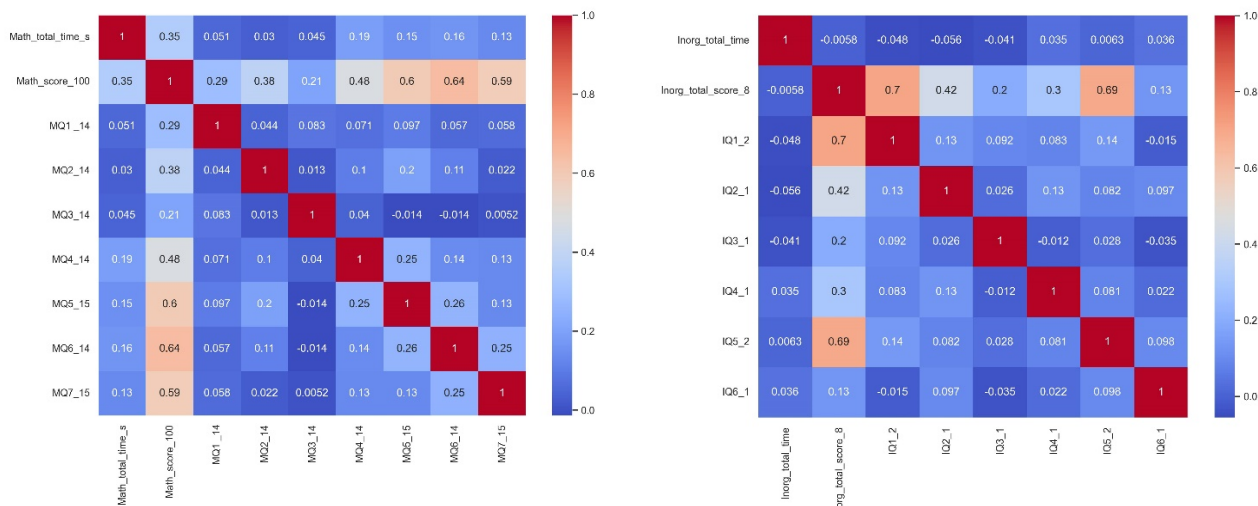


Рисунок 5 – «Тепловая карта» корреляции каждого вопроса дисциплины и времени тестирования с итоговым баллом по математике (слева) и ОНХ (справа)
Figure 5 – «Heat map» of the correlation of each discipline question and test time with the final score in mathematics (left) and ONX (right)

По ОНХ совершенно отсутствует взаимосвязь между временем тестированием и суммарным баллом, по математике незначительная корреляция между временем тестирования и итоговым баллом всё-таки присутствует. Также heatmap по ОНХ очевидным образом выделяет два вопроса: первый и пятый, которые оцениваются по два балла. Предположительно – вопросы, которые в большей степени коррелируют с итоговым баллом могут быть с более высоким значением коэффициента сложности, однако это уже вопрос диагностической аналитики и во многом зависит от структуры теста и способа его организации. В нашем случае тест организован по конкретным темам учебного материала, для которых в базе тестов существует минимум по 20 вопросов, поэтому метод конструирования теста, где обучающему выдается случайный из категории вопрос принципиально не подходит для диагностической аналитики ввиду необходимости валидации указанных тестов на базовую «психометрику». Несмотря на это, корреляционная матрица может послужить опорой для анализа связи между темами учебного материала.

В данной работе мы смещаем акцент с описания полученных результатов тестирования на сам процесс обработки данных, поэтому перейдем к обработке результатов по факультетам.

Важным аспектом описательной статистики является сопоставление генеральной совокупности обучающихся и фактической численности тестируемых как по факультетом (рис. 6), так и по группам (рисунок 7).

На рисунках 6 и 7 по оси абсцисс представлены аббревиатуры факультетов и номера учебных групп соответственно.

Для определения списочной численности обучающихся был задействован контингент, который также конвертируется в pandas.DataFrame для передачи значений в функцию по-

¹⁰ <https://seaborn.pydata.org/generated/seaborn.heatmap.html>

¹¹ <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.iloc.html>

строения гистограмм matplotlib.pyplot (см. рис. 6.). Подробнее о построении сгруппированной столбчатой гистограммы (Grouped bar chart with labels) [5].

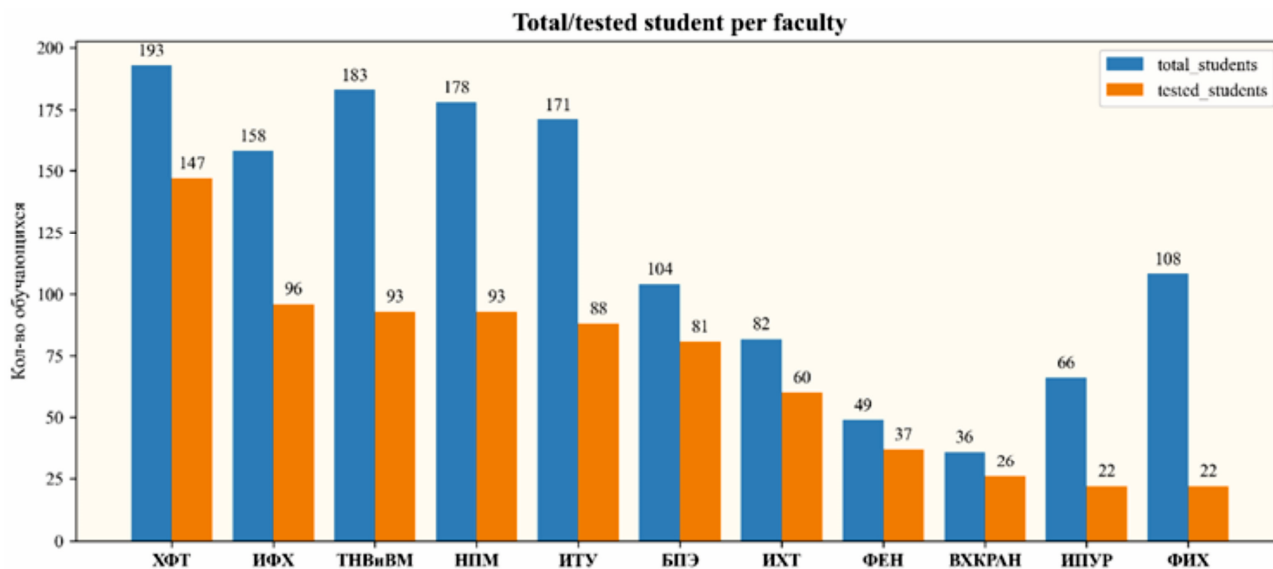


Рисунок 6 – Сопоставление генеральной совокупности обучающихся и фактической численности тестируемых по факультетам
Figure 6 – Total/tested students comparison by faculty

По гистограмме визуально определяем, что наименьшее соотношение total/tested students относится к факультетам ФИХ и ИПУР, также на факультетах ТНВиВМ, НПМ, ИТУ тестирование прошли примерно половина обучающихся.

Данные контингента также могут быть использованы для сопоставления списочной численности групп и количества тестируемых в каждой группе. Несмотря на то, что в LMS Moodle имеется возможность просматривать статистику по группам, вручную открывать 59 групп и формировать отдельную таблицу не видится целесообразным решением. Ввиду этого построим гистограмму seaborn.countplot¹², упорядоченную по алфавиту (ввиду специфики определения названий учебных групп: за каждым факультетом закрепляется определенный перечень букв «А» – ВХК РАН, «Е» – ФЕН, «И»–ИХТ, «К» – ИТУ и т.д.), что позволит визуально определить не только мало задействованные в тестировании группы, но и определить уровень участия групп в диагностическом тестировании в структуре факультета. Указанную гистограмму приведем в двух вариантах: с количеством тестируемых и долей тестируемых от списочной численности по оси ординат (см. рисунок 7). Кроме того, на рисунке 7 обнаруживаем две учебные группы (ПР-24 и Ц-25), обучающиеся из которых совсем не приняли участие в тестировании.

Функции groupby и aggregate позволяют получить среднее значение (mean), стандартное отклонение (std), и дисперсию (var), в разрезе факультетов: как по математике, так и по ОНХ, что производится ровно в две строчки кода (представлено на рисунке 8).

Табличное представление результатов агрегации базовой статистики по факультетам для целей облегчения восприятия целесообразно представить не с помощью сгруппированных столбчатых гистограмм, а с помощью функции seaborn.barplot¹³, которая автоматически рассчитывает стандартное отклонение и представляет его в виде продолжающейся линии, соразмерно среднему баллу, кроме того, в аргументах функции можно указать порядок столбцов, которые предварительно сортируются по возрастанию среднего балла (см. рисунок 9).

¹² <https://seaborn.pydata.org/generated/seaborn.countplot.html>

¹³ <https://seaborn.pydata.org/generated/seaborn.barplot.html>

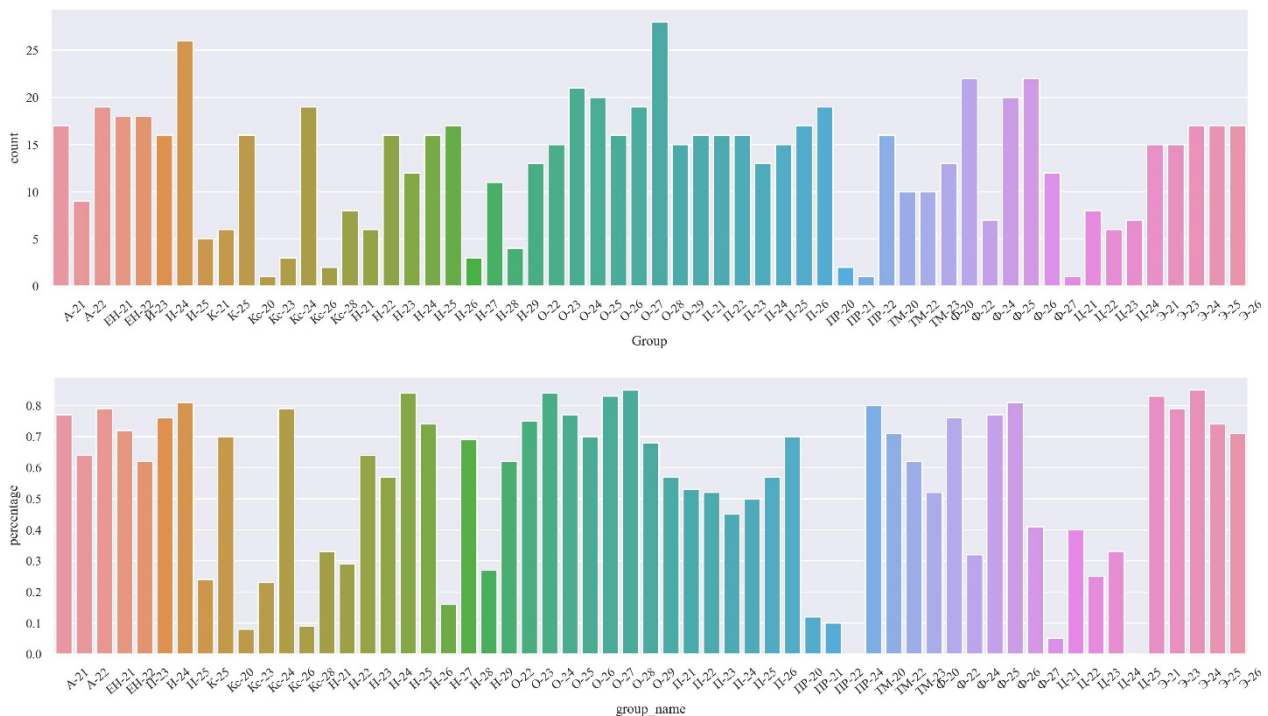


Рисунок 7 – Упорядоченные столбчатые гистограммы seaborn.barplot и seaborn.countplot: количество тестируемых по группам (сверху); доля участвующих в тестировании от списочной численности группы (снизу)

Figure 7 – Grouped bar chart of seaborn.barplot and seaborn.count plot: the number of tested student by group (top); the percentage of tested students of list quantity of the group (bottom)

```
stats=df1.groupby('Faculty', as_index=False) \
    .aggregate({'Math_score_100': ['mean','std','var'],
               'Inorg_total_score_8':['mean','std','var']})
```

stats

	Faculty	Math_score_100			Inorg_total_score_8		
		mean	std	var	mean	std	var
0	БПЭ	74.493827	15.330626	235.028086	5.125000	2.043185	4.174603
1	ВХКРАН	63.384615	22.339341	499.046154	5.045455	2.419706	5.854978
2	ИПУР	68.318182	21.866230	478.132035	5.523810	2.112322	4.461905
3	ИТУ	70.113636	17.704102	313.435214	5.549296	1.940462	3.765392
4	ИФХ	66.343750	21.367870	456.585855	5.297297	1.963570	3.855609
5	ИХТ	67.583333	23.134311	535.196328	5.255319	1.835260	3.368178
6	НПМ	66.688172	23.986399	575.347359	5.178082	2.090604	4.370624
7	ТНВиВМ	69.215054	20.325530	413.127162	5.053333	2.026136	4.105225
8	ФЕН	74.918919	17.418666	303.409910	5.466667	1.907035	3.636782
9	ФИХ	61.136364	20.910912	437.266234	5.933333	1.387015	1.923810
10	ХФТ	74.455782	22.101333	488.468922	4.852174	1.836489	3.372693

Рисунок 8 – Табличное представление результатов агрегации базовой статистики по факультетам

Figure 8 – Tabular representation of basic statistics aggregation by faculty

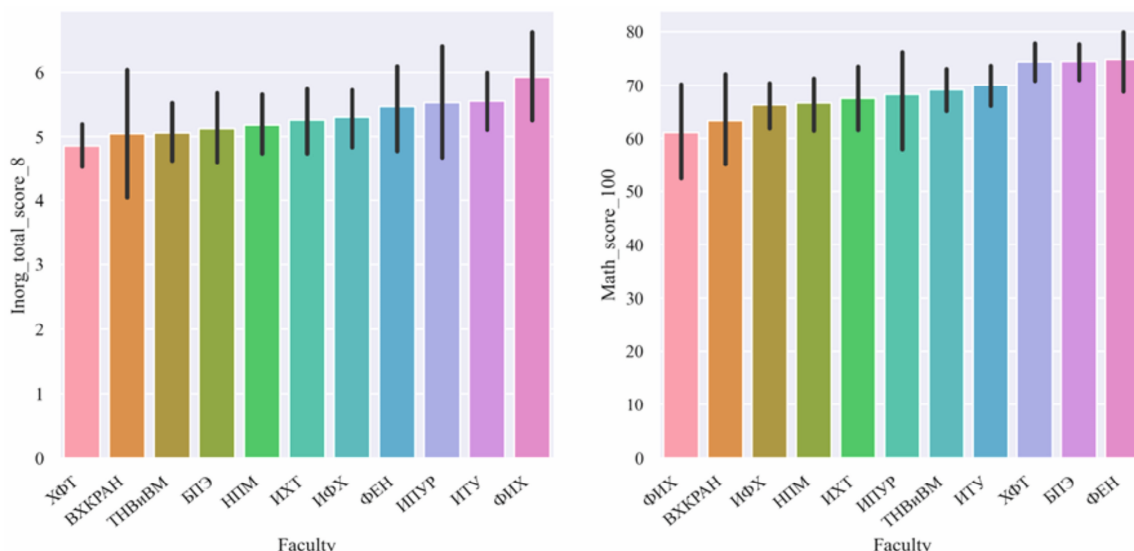


Рисунок 9 – Сопоставление среднего балла по факультетам по ОНХ (слева) и математике (справа) с помощью sns.barplot

Figure 9 – Comparison of the average ONX (left) and mathematics (right) score by faculties using sns.barplot

Для изучения структуры баллов по факультетам применим наиболее интересный с визуальной точки зрения график библиотеки seaborn – «histplot»¹⁴ с аргументом `multiple = «stack»`, который визуализирует распределение баллов по факультетам, суммируя доли обучающихся, попавших в заданный интервал оценок, как показано на рисунке 10. На рисунке видно, что по всему вузу 39 % обучающихся завершили тестирование по математике с результатом более 70 баллов, при этом большая часть обучающихся с наиболее высоким баллом приходится на факультет ХФТ (примерно 6 % от всей численности обучающихся), что несколько больше доли обучающихся этого же факультета, приходящейся на оценку «удовлетворительно», 4 %; обратная ситуация наблюдается, к примеру, на факультете ТНВиВМ, где всего 1,5 % обучающихся факультета справились с тестом на самый высокий балл и 4 % обучающихся на оценку «удовлетворительно».

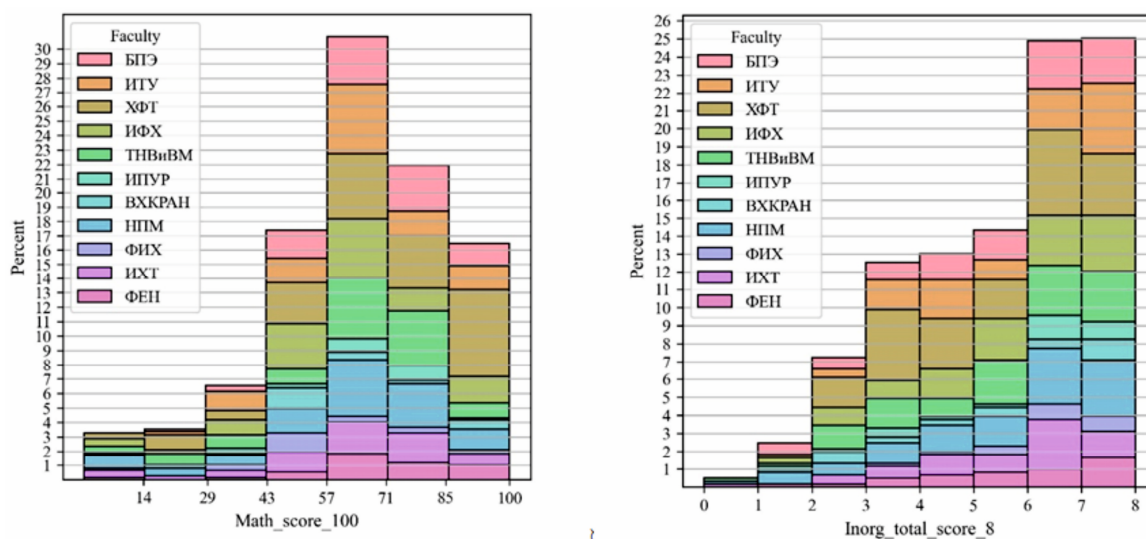


Рисунок 10 – Анализ «структуры» распределения баллов по математике (слева) и ОНХ (справа) с помощью seaborn.histplot с аргументом `multiple = «stack»`

Figure 10 – Analysis of the distribution «structure» of mathematics (left) and ONX (right) scores using seaborn.histplot with the argument `multiple = «stack»`

¹⁴ <https://seaborn.pydata.org/generated/seaborn.histplot.html>

Аналогичный подход можно применить для изучения структуры баллов по группам внутри факультета. На рисунке 10 представим такое распределение баллов по математике (слева) и ОНХ (справа) для групп на факультете ХТФ. Здесь по математике наибольшее количество баллов набрали группы О-28, О-25 и О-26; по ОНХ – группы О-22, О-29, О-24 и О-25.

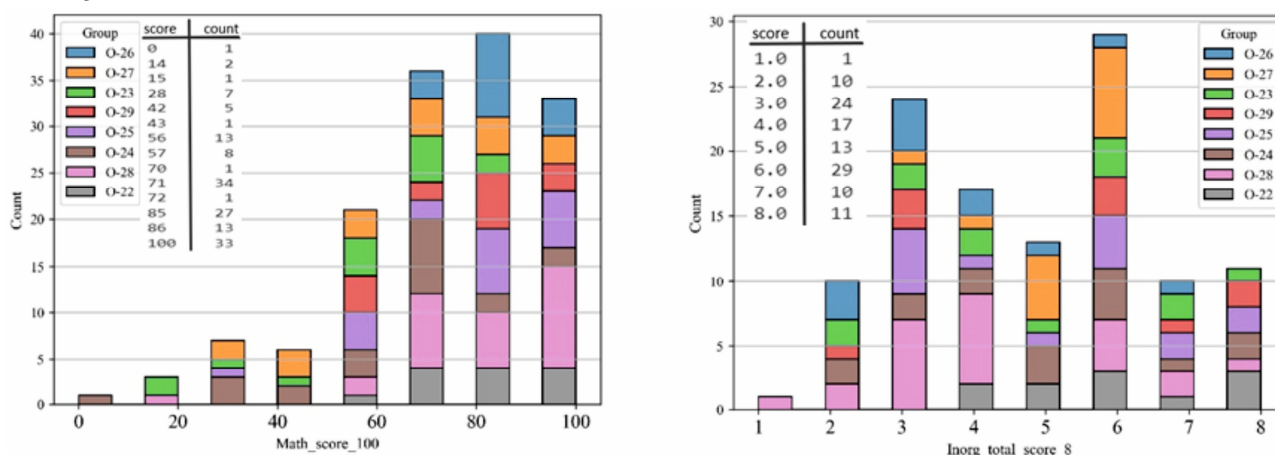


Рисунок 11 – Анализ «структуры» распределения баллов по математике (слева) и ОНХ (справа) на факультете ХФТ

Figure 11 – Analysis of the distribution «structure» of the mathematics (left) and ONX (right) scores on the Faculty of «Chemistry and Pharmacy»

Здесь вместо линий сетки, как это было сделано на предыдущем рисунке 10, вынесены таблицы с фактическим распределением частот баллов обучающихся на выбранном факультете, полученные с помощью функции `pandas.Series.value_counts`.

Такой подход кроме прочего позволяет продемонстрировать работу функции `seaborn.histplot` с параметрами «по умолчанию» (кроме увеличения значения параметра «bins» для получения большей точности и параметра `multiple = «stack»` для наглядности структуры баллов по группам в распределении).

Для сопоставления среднего балла по учебным группам на факультете предложим два подхода.

Первый – построение `seaborn.barplot`, аналогично рисунку 9, который позволяет сравнить средний балл по всем группам на факультете. На рисунке 12 представим сравнение среднего балла по факультету ХФТ.

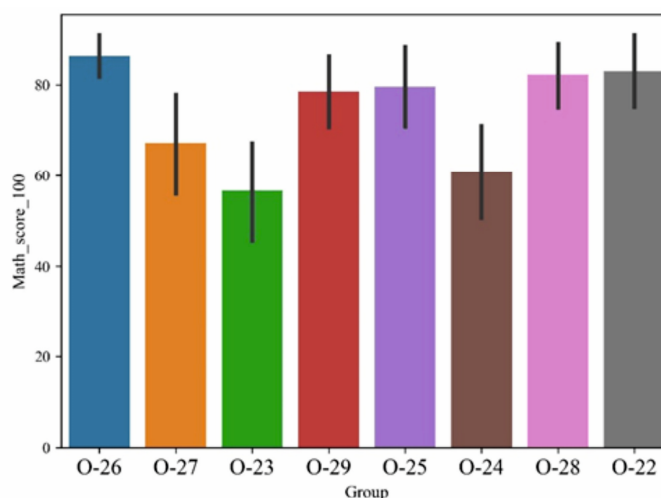


Рисунок 12 – Сопоставление результатов тестирования по математике на факультете ХФТ с использованием `seaborn.barplot`

Figure 12 – Mathematics testing score comparison on the Faculty of «Chemistry and Pharmacy» using `seaborn.barplot`

Подчеркнём, что основное преимущество данной библиотеки – не столько возможность строить наглядные графики как таковые, сколько наличие способов мгновенного отбора данных непосредственно из общего фрейма данных и корректность работы инфографики с параметрами «по умолчанию», что играет важную роль при повторном задействовании графиков и обработки образовательных данных как таковых.

Второй подход – это возможность реализации собственной функции, которая будет фильтровать исходный фрейм данных под конкретные учебные группы внутри выбранного факультета и для каждой строить отдельный график с помощью функции `matplotlib.pyplot.subplot`¹⁵, что продемонстрировано на рис. 13 для факультета ИТУ с результатами тестирования по математике.

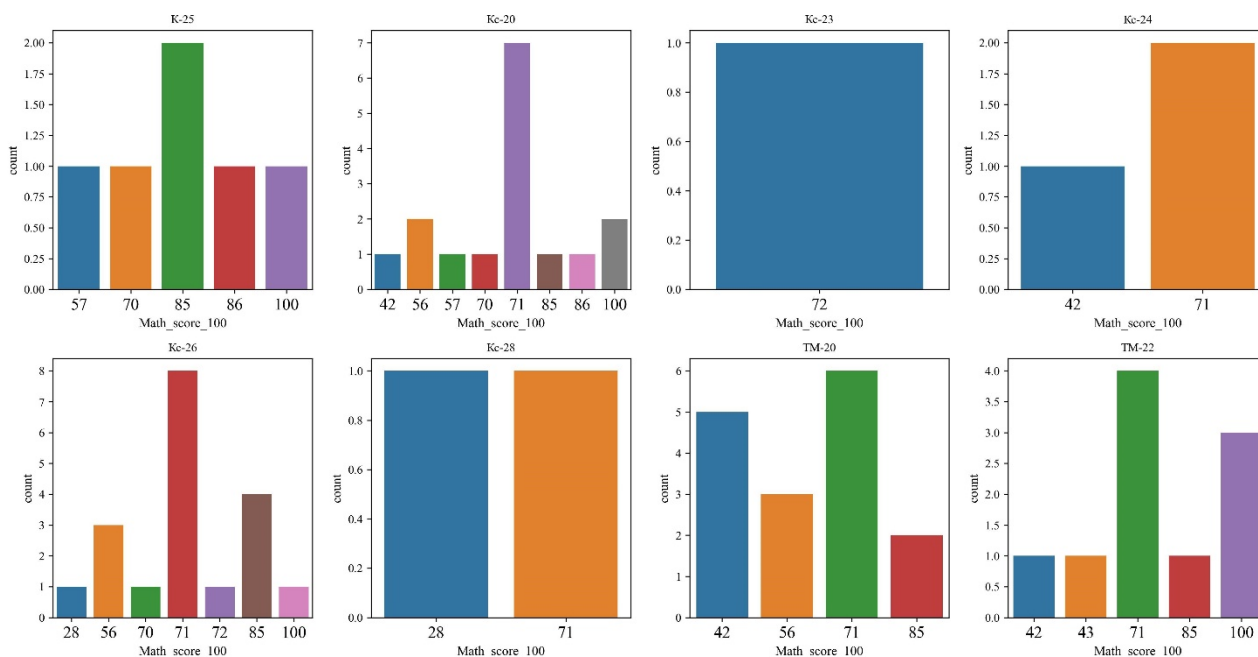


Рисунок 13 – Реализация функции, строящей гистограммы по количеству групп на факультете на примере факультета ИТУ с результатами тестирования по математике
Figure 13 – Implementation of histograms drawing function by the number of groups at the faculty on the example of ITU Faculty with the mathematics testing scores

Из такой «панели» графиков наглядно определяется количество групп на факультете, участвующих в тестировании, распределение баллов по группам, частоты значений баллов; что позволяет выделить группы, в недостаточной степени усвоивших материал (ТМ-20), и группы, для которых необходимо инициировать повторное тестирование ввиду недостаточного количества участников (Кс-23, Кс-24, Кс-28).

В данном случае функция итерировалась по элементам списка групп факультета, который был создан с помощью `pandas.Value_counts`. Такой подход позволяет быстро определить группы, которые, с одной стороны, в недостаточной степени усвоили материал, с другой стороны – были слабо задействованы в тестировании.

Закключение

Таким образом в работе был предложен подход к обработке большого количества данных о тестировании обучающихся в LMS Moodle, который по схожим принципам применяется для анализа «цифровых следов» пользователей, при этом целью данной работы был именно расширенный дескриптивный анализ, который может стать первым шагом к выстраиванию адаптивных механизмов и подходов, а также использоваться в качестве рабочего инструмента мониторинга за результатами тестирования обучающихся.

¹⁵ https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.subplot.html

Ввиду широкого распространения LMS Moodle в качестве системы сбора, хранения, мониторинга, обработки образовательных данных в российских вузах, а также перспективности применения этой системы для погружения в образовательные процессы за счет формирования «цифровых следов» и обширных возможностей представления и визуализации данных, предложенный подход может быть использован не только для процедур внутреннего контроля знаний обучающихся, но и для проведения углублённого анализа в рамках т.н. «отчётов о самообследовании», а также в процедурах внешнего и внутреннего контроля качества образования, осуществляющихся в системе менеджмента качества образования применительно к проектированию и осуществлению образовательной деятельности по программам высшего и среднего профессионального образования.

Ввиду наличия требований организаций, проводящих профессионально-общественную аккредитацию к представлению образовательных данных в разрезе образовательных программ, предложенный процесс обработки может быть усовершенствован путем добавления к фрейму данных кодов образовательных программ и формирования соответствующих «срезов» данных для построения инфографики, что не будет столь ресурсозатратным мероприятием ввиду широких возможностей автоматизации языка Python и используемых библиотек.

Предложенный подход к обработке данных в РХТУ им. Д. И. Менделеева в настоящий момент совершенствуется как по программно-алгоритмической, так и методической части для применения его заинтересованными подразделениями. Единоразово реализовав обработку образовательных данных в среде выполнения кода (к примеру, «Google colab»), можно автоматизировать процесс обработки результатов любого массового тестирования, загружая на вход три файла-выгрузки из LMS Moodle: «оценки», «ответы», и «Export with groups and dates» и контингент, а на выходе получая полноценную дескриптивную аналитику результатов тестирования.

Библиографический список

1. **Тербушева Е. А., Пиотровская К. Р.** Аналитический потенциал платформы Moodle для мониторинга качества персонифицированного обучения // Общество. Коммуникация. Образование. 2021. Т. 12. № 4. С. 19-34. DOI: 10.18721/JHSS.1 (дата обращения: 24.02.2023).
2. **Ширинкина, Е. В.** Интеллектуальный анализ образовательных данных / Е. В. Ширинкина // Вестник Тверского государственного университета. Серия: Экономика и управление. 2021. № 3(55). С. 179-188. DOI 10.26456/2219-1453/2021.3.179-188. EDN ZKOGBS (дата обращения: 24.02.2023).
3. Определение понятия «Учебная аналитика». ГОСТ Р 59895-2021. «Технологии искусственного интеллекта в образовании. Общие положения и терминология» [Электронный ресурс]: база ГОСТов allgosts.ru - Режим доступа: https://allgosts.ru/35/240/gost_r_59895-2021 (дата обращения: 24.02.2023)
4. **Rosário N. A. L.** Predicting academic performance-A practical study using Moodle log data and sociodemographic traits : дис. 2022 (дата обращения: 24.02.2023).
5. Grouped bar chart with labels [Электронный ресурс]: документация библиотеки matplotlib. Режим доступа: https://matplotlib.org/stable/gallery/lines_bars_and_markers/barchart.html (дата обращения: 24.02.2023).

UDC 004.62/37.012.4

DATA SCIENCE TOOLS FOR ANALYZING A LARGE SAMPLE OF STUDENTS TESTING DATA IN MOODLE

A. E. Boyko, post-graduate student, MUCTR, Moscow, Russia;

orcid.org/ 0009-0007-4006-4403, e-mail: boiko.a.e@muctr.ru

T. V. Savitskaya, Dr. Sc. (Tech.), professor, MUCTR, Moscow, Russia;

orcid.org/ 0000-0001-9818-2196, e-mail: savitskaia.t.v@muctr.ru

D. S. Lopatkin, Ph. D (Econ.), Head of the Department of Management and Marketing, MUCTR, Moscow, Russia;

orcid.org/ 0000-0003-4260-4710, e-mail: lopatkin.d.s@muctr.ru

*Due to the intensification of processes for building online-courses by national universities using LMS Moodle, where collection, storage, monitoring, and processing of educational data is carried out, the formation of a unified approach to the implementation of descriptive and diagnostic educational analytics that can meet the objectives of quality management system (QMS) in terms of the presenting testing data for the purposes of internal and external quality control of education becomes an urgent task. **The aim** is to expand the possibilities of monitoring, presenting and visualizing educational data from LMS Moodle related to various forms of testing students using data science tools to optimize educational processes and improve the quality of learning. Results: the paper considers the process of preparation and further processing and visualization of data on diagnostic testing of second-year students in D. Mendeleev University of Chemical Technology of Russia (more than 900 people) in two disciplines. Various options for data aggregation are considered: by groups at the faculty, heterogeneous class, within the educational program, at the level of entire university. Various options for visualizing the results of comparing the academic performance of groups and the test-ed/total quantity of students are proposed; visualization options and parameters for analyzing distribution "structure" of test scores in study groups at a selected faculty, etc. Practical significance: the results of the study will help teachers and administrators of LMS Moodle to make more informed decisions based on data; the tools proposed can become a starting point for the implementation of basic adaptive approaches in teach-in.*

Keywords: learning analytics, learning methodology system, LMS Moodle, data analysis, Python, Seaborn library, Matplotlib library, descriptive analytics, diagnostic analytics, higher education.

DOI: 10.21667/1995-4565-2023-85-103-116

References

1. Terbusheva E. A., Piotrovskaya K. R. Analiticheskij potencial platformy Moodle dlya monitoringa kachestva personificirovannogo obucheniya [Analytical potential of the LMS Moodle for monitoring the quality of personification]. *Obshchestvo. Kommunikaciya. Obrazovanie*, 2021, vol. 12, no. 4, pp. 19-34. DOI: 10.18721/JHSS.1 (accessed: 24.02.2023).
2. Shirinkina, E. V. Intellektual'nyĭ analiz obrazovatel'nykh dannyx [Educational data mining]. *Vestnik Tverskogo gosudarstvennogo universiteta. Seriya: Ėkonomika i upravlenie*, 2021, vol. 3(55), pp. 179-188. DOI 10.26456/2219 1453/2021.3.179 188. EDN ZKOGBS (accessed: 24 February 2023).
3. Opredelenie poniatii «Uchebnaia analitika». GOST R 59895 2021. «Tekhnologii iskusstvennogo intellekta v obrazovanii. Obshchie polozheniia i terminologiiia», baza GOSTov *allgosts.ru*, available at: https://allgosts.ru/35/240/gost_r_59895-2021
4. Rosário, N. A. L. Predicting academic performance – A practical study using Moodle log data and sociodemographic traits, Diss, 2022.
5. Grouped bar chart with labels. Matplotlib documentation, available at: https://matplotlib.org/stable/gallery/lines_bars_and_markers/barchart.html (accessed: 24.02.2023).