

МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ
ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ им. А. И. ГЕРЦЕНА»



Направление подготовки
09.03.01 Информатика и вычислительная техника

Направленность (профиль)
«Технологии разработки программного обеспечения»

Выпускная квалификационная работа

Использование машинного обучения для анализа эффективности цифровизации
учебного процесса

Обучающегося 4 курса
очной формы обучения
Воложанина Владислава Олеговича

Руководитель выпускной квалификационной
работы:
кандидат физико-математических наук, доцент
кафедры информационных технологий и
электронного обучения Власов Дмитрий
Викторович

Санкт-Петербург
2025

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ	3
INTRODUCTION	5
Цифровизация высшего образования и применение методов машинного обучения в образовательной аналитике	7
1.1 Цифровизация в высшем образовании: понятия, инициативы и современные инструменты	7
1.2 Основы и методы машинного обучения	11
1.3 Программные средства, платформы и методы кластеризации и классификации	14
1.4 Кластеризация и классификация в образовательной аналитике: методы, сравнительный анализ и программные средства	18
Вывод по главе №1	21
Кластеризация цифровых профилей студентов, классификация и генерация персонализированных рекомендаций	23
2.1 Разведочный анализ данных (EDA)	23
2.2 Предобработка данных и преобразование признаков	27
2.3 Кластеризация студентов (умар-признаки и матрица говера)	30
2.4 Оценка качества кластеризации	35
2.5 Интерпретация и профилирование кластеров по бинарным признакам	38
2.6 Полярные (radar) диаграммы профилей кластеров	41
2.7 Построение и оценка классификатора профиля	44
2.8 Генерация персонализированных рекомендаций	47
2.9 Анализ распределения цифровых профилей по факультетам и автоматическая интерпретация	50
Вывод по главе №2	53
ЗАКЛЮЧЕНИЕ	55
ЛИТЕРАТУРА	58

ВВЕДЕНИЕ

Цифровизация образовательного процесса является одной из приоритетных задач модернизации высшего образования. Несмотря на широкое внедрение электронных образовательных платформ, систем управления обучением и цифровых сервисов, эффективность их использования в большинстве случаев остаётся недостаточно изученной. Наличие цифровой инфраструктуры не гарантирует повышения качества образования без анализа того, как именно цифровые инструменты используются обучающимися и как они влияют на образовательные результаты.

Актуальность настоящего исследования заключается в необходимости применения методов машинного обучения для объективной оценки процессов цифровизации учебной деятельности. Традиционные подходы к анализу образовательных данных не позволяют выявлять скрытые закономерности в цифровом поведении студентов и не обеспечивают достаточного уровня индивидуализации образовательной поддержки. Использование алгоритмов машинного обучения предоставляет возможность автоматизированного анализа больших массивов данных, выделения цифровых профилей обучающихся и генерации персонализированных рекомендаций, что соответствует современным требованиям к качеству образовательной аналитики.

Практическая значимость работы заключается в возможности применения разработанных моделей для объективного анализа цифровизации учебного процесса и автоматизации поддержки студентов. Результаты исследования могут быть внедрены в деятельность вузов для своевременного выявления групп студентов с разным уровнем цифровой активности, разработки адресных рекомендаций и оптимизации образовательных стратегий на основе анализа больших данных. Это, в свою очередь, способствует не только повышению качества образования, но и формированию культуры использования современных цифровых сервисов в вузах.

Цель выпускной квалификационной работы — использование методов машинного обучения для анализа эффективности цифровизации учебного процесса на основе анкетных данных студентов.

Предметом исследования выступают процессы цифровизации высшего образования, методы обработки и анализа образовательных данных, а также алгоритмы построения цифровых профилей студентов и генерации персонализированных рекомендаций.

Для достижения поставленной цели решаются следующие задачи:

1. Изучить методы машинного обучения, применяемые в образовательной аналитике.
2. Осуществить сбор, предобработку и разведочный анализ анкетных данных студентов.
3. Построить модели кластеризации с целью выделения цифровых профилей обучающихся.
4. Обучить классификатор для автоматического определения цифрового профиля нового студента.
5. Разработать модуль генерации персонализированных рекомендаций на основе интеграции с GigaChat API.
6. Провести анализ распределения цифровых профилей по факультетам и интерпретировать полученные результаты.

Структура работы включает введение, первую главу, посвящённую теоретическим аспектам цифровизации и методам машинного обучения, и вторую главу, содержащую практическую реализацию: анализ данных, построение моделей кластеризации и классификации, а также модуль генерации рекомендаций. В завершении приведены выводы и список использованных источников.

В работе использованы современные научные публикации и статьи, посвящённые вопросам цифровизации образования, образовательной аналитики, машинного обучения и обработки больших данных.

INTRODUCTION

Digitalization of the educational process is one of the priorities of modernizing higher education. Despite the widespread introduction of electronic educational platforms, learning management systems, and digital services, the effectiveness of their use remains poorly understood in most cases. The presence of a digital infrastructure does not guarantee an improvement in the quality of education without analyzing exactly how digital tools are used by students and how they affect educational outcomes.

The relevance of this study lies in the need to use machine learning methods for an objective assessment of the processes of digitalization of educational activities. Traditional approaches to the analysis of educational data do not allow us to identify hidden patterns in the digital behavior of students and do not provide a sufficient level of individualization of educational support. The use of machine learning algorithms provides an opportunity for automated analysis of large amounts of data, highlighting digital profiles of students and generating personalized recommendations, which meets modern requirements for the quality of educational analytics.

The practical significance of the work lies in the possibility of applying the developed models for an objective analysis of the digitalization of the educational process and automation of student support. The research results can be implemented in the activities of universities to timely identify groups of students with different levels of digital activity, develop targeted recommendations and optimize educational strategies based on big data analysis. This, in turn, contributes not only to improving the quality of education, but also to the formation of a culture of using modern digital services in universities.

The purpose of the final qualification is to use machine learning methods to analyze the effectiveness of the digitalization of the educational process based on students' personal data.

The subject of the research is the processes of digitalization of higher education, methods of processing and analyzing educational data, as well as algorithms for building digital profiles of students and generating personalized recommendations.

To achieve this goal, the following tasks are being solved:

1. To study machine learning methods used in educational analytics.
2. To carry out the collection, preprocessing and exploratory analysis of students' personal data.
3. To build clustering models in order to identify digital profiles of students.
4. Train a classifier to automatically identify a new student's digital profile.
5. Develop a module for generating personalized recommendations based on integration with the GigaChat API.
6. To analyze the distribution of digital profiles by faculty and interpret the results obtained.

The structure of the paper includes an introduction, the first chapter devoted to the theoretical aspects of digitalization and machine learning methods, and the second chapter containing practical implementation: data analysis, building clustering and classification models, as well as a module for generating recommendations. In conclusion, the conclusions and the list of sources used are given.

The work uses modern scientific publications and articles on the issues of digitalization of education, educational analytics, machine learning and big data processing.

Цифровизация высшего образования и применение методов машинного обучения в образовательной аналитике

1.1 Цифровизация в высшем образовании: понятия, инициативы и современные инструменты

Цифровизация высшего образования подразумевает интеграцию информационно-коммуникационных технологий в учебный процесс, административную деятельность и инфраструктурное обеспечение университетов. Данный процесс не ограничивается переходом на дистанционные формы, а предполагает изменение всей архитектуры образовательной среды за счёт внедрения электронных систем, автоматизации процедур и использования программных продуктов для сопровождения обучения. Применение цифровых платформ позволяет университетам управлять образовательным контентом, отслеживать прогресс студентов, формировать индивидуальные образовательные траектории и обеспечивать взаимодействие между участниками образовательного процесса. В рамках цифровой трансформации реализуются проекты по оснащению вузов высокоскоростным интернетом, созданию облачных хранилищ, внедрению электронных журналов, автоматизации расписания и документооборота, что существенно упрощает администрирование и организацию учебной деятельности. Использование LMS-платформ (Learning Management Systems) стало обязательным стандартом: программные комплексы Moodle, Blackboard, Canvas и аналогичные решения интегрируются в структуру университета, предоставляя единое пространство для размещения учебных материалов, обмена заданиями, проверки знаний, мониторинга посещаемости и фиксации результатов.

Программные инициативы федерального и международного уровня способствуют продвижению цифровых инноваций в университетской среде. Примером служит стратегическая программа Министерства науки и высшего образования Российской Федерации, ориентированная на цифровую трансформацию вузов и рассчитанная до 2030 года, где основной акцент

делается на формирование индивидуальных образовательных маршрутов, развитие научно-исследовательских компетенций и внедрение практик использования аналитики данных для совершенствования образовательных решений. Сходные задачи отражены в ряде международных документов, включая рекомендации ЮНЕСКО, где подчеркивается роль цифровизации для обеспечения инклюзивности и повышения качества образования. Мировое образовательное сообщество рассматривает развитие цифровых навыков и гибких компетенций как фундамент современного образования, что требует от вузов пересмотра структуры учебных программ, а также перестройки методов преподавания и взаимодействия с обучающимися.

Реализация цифровизации охватывает ряд ключевых направлений. Современная цифровая инфраструктура включает в себя оснащение учебных корпусов компьютерными классами, доступом к высокоскоростному интернету, средствами защиты данных и облачными сервисами. Электронные образовательные ресурсы, представленные мультимедийными курсами, виртуальными лабораториями, электронными учебниками и видео-лекциями, обеспечивают доступность учебных материалов вне зависимости от физического присутствия в аудитории. LMS-платформы служат ядром управления образовательным процессом, интегрируя функции размещения материалов, проверки заданий, организации форумов и обсуждений, а также ведения электронной отчетности. Создание и поддержка электронных портфолио и цифровых идентификаторов обучающихся позволяет автоматизировать процесс признания образовательных достижений и облегчает академическую мобильность. Появление государственных и межуниверситетских цифровых платформ, таких как "Современная цифровая образовательная среда", обеспечивает стандартизацию доступа к онлайн-курсам, централизованный учет результатов и интеграцию с государственными системами идентификации.

Формирование новых моделей организации учебного процесса включает внедрение смешанного обучения, реализацию индивидуальных и гибких траекторий, массовых открытых онлайн-курсов и полностью дистанционных

программ. Перестройка содержания образовательных программ происходит с учетом запросов рынка труда и возможностей цифровых технологий. Актуализируются задачи персонализации обучения, где студенты получают возможность самостоятельно выбирать последовательность и содержание изучаемых дисциплин, а образовательные платформы обеспечивают автоматическую адаптацию под индивидуальные интересы и уровень подготовки.

Современные инструменты цифровизации охватывают все этапы образовательного цикла. Применение электронных учебников и автоматизированных систем проверки знаний упрощает контроль усвоения материала. Видеоконференции, мессенджеры, платформы для совместной работы (Zoom, Teams, Trello, Miro) интегрируются в учебный процесс и обеспечивают коммуникацию между преподавателями и студентами, а также внутри студенческих групп. Использование сервисов онлайн-опросов и тестирования (Kahoot, Quizlet) позволяет собирать обратную связь и организовывать интерактивное взаимодействие в режиме реального времени. Программное обеспечение для инженерных, математических, языковых и других дисциплин расширяет инструментарий преподавателей, способствуя освоению практических навыков.

В образовательной среде активно внедряются инновационные технологии, включая искусственный интеллект, аналитику больших данных, адаптивные образовательные платформы, а также виртуальную и дополненную реальность. Системы аналитики позволяют собирать данные о прогрессе студентов, выявлять закономерности и предлагать индивидуальные рекомендации. Интеллектуальные агенты и чат-боты обеспечивают автоматизированную поддержку обучающихся. AR/VR-технологии применяются для создания виртуальных лабораторий и тренажёров, которые позволяют воспроизводить профессиональные ситуации и формировать практические компетенции. Электронные сертификаты и дипломы с функцией верификации на основе блокчейна становятся стандартом для подтверждения квалификаций, облегчая

процессы трудоустройства и международного признания образовательных результатов.

Цифровизация сопровождается обновлением нормативно-правовой базы, разработкой стандартов и требований к электронным образовательным ресурсам, а также совершенствованием механизмов обеспечения информационной безопасности. В российской практике формируется система обязательной аккредитации электронных информационно-образовательных сред, что гарантирует соответствие используемых цифровых инструментов установленным критериям качества и доступности. Широкое внедрение цифровых технологий приводит к изменению роли преподавателя, который становится организатором и координатором образовательного процесса, сопровождающим студента на индивидуальном маршруте.

Развитие цифровых компетенций обучающихся и преподавателей рассматривается как стратегическая задача, что обуславливает появление специализированных программ повышения квалификации, образовательных интенсивов и курсов цифровой грамотности. Важной тенденцией становится распространение практик непрерывного образования, когда цифровые платформы обеспечивают доступ к учебным ресурсам на протяжении всей профессиональной деятельности. Актуальность цифровизации подтверждается аналитическими данными о росте числа пользователей онлайн-образовательных платформ и увеличении объёма цифрового контента, что свидетельствует о трансформации образовательных практик на всех уровнях системы.

Реализация столь масштабных преобразований невозможна без применения современных аналитических инструментов и технологий обработки данных. Особую роль в этом процессе начинают играть методы машинного обучения, позволяющие выявлять скрытые закономерности в образовательной среде и поддерживать принятие решений на основе больших массивов информации.

1.2 Основы и методы машинного обучения

Машинное обучение определяется как область искусственного интеллекта, фокусирующаяся на построении алгоритмов, которые извлекают зависимости и структуры из эмпирических данных и используют сформированные закономерности для решения практических задач, не предусматривающих явного программирования всех вариантов. Исходный набор данных включает объекты, представленные совокупностью признаков, и, при наличии разметки, целевыми переменными, которые подлежат прогнозированию либо категоризации. Система, обладающая способностью к машинному обучению, на этапе обучения анализирует примеры, оптимизирует внутренние параметры модели, минимизируя функцию ошибки, а затем демонстрирует способность к переносимости приобретённых знаний на новые входные данные. В определении, предложенном Томом Митчеллом, формализуется связь между качеством выполнения задачи, опытом и характеристиками модели, обучающейся на наборе примеров, что подчёркивает приоритет эмпирического подхода над детерминированным программированием.

Технологический цикл машинного обучения включает этапы сбора и структурирования обучающих данных, отбора информативных признаков, выбора и настройки алгоритма, процедуры оптимизации параметров на основании критерия качества, а также тестирования и оценки полученных моделей на независимой выборке. На практике для предотвращения переобучения и контроля обобщающей способности моделей используются методы кросс-валидации, регуляризации, фильтрации признаков и увеличение обучающего множества. Контроль качества работы моделей осуществляется с использованием метрик, отражающих способность алгоритма обобщать закономерности вне обучающей выборки.

Современная система классификации методов машинного обучения строится на принципе наличия или отсутствия априорной информации о правильных ответах в обучающей выборке. Класс супервизированных алгоритмов (обучение с учителем) предполагает работу с размеченными

данными, где для каждого объекта известна целевая переменная, подлежащая предсказанию. Типичными задачами в этой парадигме выступают классификация и регрессия. Классификация предполагает разделение объектов на конечное множество дискретных категорий на основе анализа признаков, при этом процесс обучения строится на сопоставлении реальных классов и прогнозируемых меток. Алгоритмы классификации варьируются от простых моделей, например, логистической регрессии и дерева решений, до сложных ансамблевых и нейросетевых архитектур, способных выявлять сложные, нелинейные зависимости между входными переменными и целевой переменной. Задача регрессии заключается в прогнозировании количественных показателей на основании признаков, где модель формирует функциональную зависимость между признаковым пространством и непрерывной переменной.

В рамках несупервизированного подхода к обучению используются данные, лишённые информации о целевых переменных или классах. Алгоритмы этой группы анализируют только распределение и структуру признаков объектов, осуществляя поиск внутренних закономерностей, паттернов и сегментов. Классическая задача в этом контексте — кластеризация, подразумевающая разбиение множества объектов на кластеры по критерию максимального сходства внутри группы и максимального различия между различными группами. Алгоритмы кластеризации различаются методами формирования сегментов: итеративные методы, такие как k-средних и Fuzzy C-Means, строят разбиение на фиксированное количество групп с возможностью мягкого (размытого) членства для FCM; иерархические подходы, такие как агломеративная кластеризация, организуют данные в виде дерева вложенных кластеров и допускают анализ структуры на разных уровнях детализации; плотностные алгоритмы формируют кластеры как области высокой плотности в пространстве признаков; вероятностные методы строят модели, основанные на предположении о распределении данных. Кластеризация находит применение для сегментации пользователей, выявления профилей и предварительного анализа структуры данных. В ряде случаев используются методы снижения

размерности, такие как метод главных компонент, t-SNE, UMAP, для упрощения анализа, визуализации и подготовки к последующей кластеризации, что позволяет представить многомерное пространство признаков в двумерном или трёхмерном виде без существенной потери информации о структуре выборки.

Результаты кластерного анализа формируют основу для перехода к задачам супервизированного обучения, где автоматически определённые профили или сегменты используются в качестве целевой переменной для последующего обучения классификаторов. На этапе классификации модели обучаются на выборке с присвоенными метками и верифицируются на тестовых данных по ряду метрик. Наиболее часто используются точность, полнота, F1-мера и значения макро- и микро-усреднённых показателей, особенно в условиях несбалансированных классов. Для объяснения решений и интерпретации значимости признаков в современных реализациях применяется анализ важности признаков, визуализация структуры дерева решений или оценка вклада переменных в ансамблевых моделях.

В современной образовательной аналитике машинное обучение становится инструментом выявления паттернов цифрового поведения, автоматизации профилирования обучающихся, а также предсказания индивидуальных образовательных траекторий. В процессе реализации задач на основе анкетных данных студентов в исследовании использовались и супервизированные, и несупервизированные методы: этап кластеризации позволил выделить типовые профили цифровых практик, а построение классификаторов обеспечило возможность автоматического определения профиля для новых студентов на основании их анкетных ответов. Процесс построения моделей сопровождался процедурами стандартизации признаков, отбора переменных с высокой информативностью и исключения скоррелированных либо слабо вариативных признаков, что повысило устойчивость и качество итоговых решений. Были использованы методы предотвращения переобучения и реализованы алгоритмы интерпретации

результатов, включая визуализацию структуры кластеров в пространстве UMAP и анализ важности признаков для выбранных моделей.

Для успешной реализации указанных методов на практике необходима интеграция соответствующих программных средств и платформ, которые обеспечивают обработку, хранение и анализ образовательных данных, а также построение и внедрение аналитических моделей. Следующий раздел посвящён рассмотрению современных инструментов, библиотек и программных решений, применяемых для кластеризации, классификации и анализа цифровых профилей в образовательной среде.

1.3 Программные средства, платформы и методы кластеризации и классификации

В рамках анализа образовательных данных и построения цифровых профилей студентов широко используются программные средства, разработанные с учётом специфики учебных задач, масштабов выборок и особенностей источников информации. Наиболее универсальным и применимым инструментом для реализации алгоритмов машинного обучения в образовательной среде является библиотека Scikit-Learn, включающая комплекс модулей для решения задач классификации, регрессии, кластеризации, отбора признаков, автоматизации поиска гиперпараметров и построения пайплайнов обработки. В среде научных и прикладных исследований Scikit-Learn используется для построения предиктивных моделей на основе анкетных, событийных и лог-файловых данных, анализа успеваемости, выявления факторов риска и прогнозирования образовательных траекторий. Реализация деревьев решений, логистической регрессии, метода опорных векторов, ансамблевых алгоритмов и модулей для визуализации структуры данных позволяет проводить широкий спектр исследований в области учебной аналитики.

Обработка текстовых, графических и событийных данных, а также построение моделей высокой сложности и глубины осуществляется посредством библиотек глубокого обучения, таких как TensorFlow, PyTorch, Keras. Данные

фреймворки обеспечивают возможность построения, обучения и внедрения нейронных сетей с различной архитектурой для решения задач анализа ответов в свободной форме, автоматической проверки письменных работ, обработки изображений, предсказания вероятности досрочного отсева в онлайн-курсах, а также сегментации обучающихся на основе многоуровневых признаковых структур. Применение глубоких моделей повышает точность решения комплексных аналитических задач и обеспечивает обработку больших массивов неструктурированных данных.

Для аналитиков, не обладающих компетенциями в программировании, разработаны визуальные конструкторы анализа данных, примером которых выступает среда Orange. С помощью графического интерфейса и инструментов визуализации Orange используется в образовательных проектах для кластеризации студентов, построения дендрограмм, анализа коммуникаций и создания интерактивных отчётов по результатам учебной деятельности. Дополнительный набор методов для автоматизации анализа и экспериментальной проверки алгоритмов реализован в пакете Weka, который находит применение в исследованиях по Educational Data Mining и позволяет выполнять полный цикл от импорта данных до построения сложных моделей и анализа результатов в визуальной форме.

Внедрение аналитических модулей и методов машинного обучения в образовательную инфраструктуру осуществляется через современные платформы управления обучением, среди которых Moodle, Canvas, Blackboard. Указанные системы интегрируют специализированные расширения для мониторинга успеваемости, анализа рисков и раннего выявления студентов с потенциальными трудностями. На основе данных LMS строятся автоматические отчёты для преподавателей и администраторов, а также формируются уведомления о снижении активности или вероятности неуспешного прохождения курса. Для агрегации, хранения и анализа событийных данных с различных платформ применяются внешние системы учебной аналитики,

осуществляющие интеграцию информации из LMS, электронных тестов, опросников, форумов и прочих цифровых следов обучающихся.

Для формирования отчётов, мониторинга и визуализации результатов образовательной аналитики используются инструменты бизнес-аналитики, такие как Power BI, Tableau, Qlik. С помощью этих платформ создаются интерактивные панели мониторинга, отражающие динамику академических показателей, посещаемости, активности в онлайн-курсах и результаты анкетирования. В образовательных организациях данные инструменты применяются для оперативного принятия управленческих решений и выстраивания адаптивных стратегий поддержки студентов.

Работа с большими массивами событийных данных, поступающих с платформ массового открытого онлайн-обучения, ведётся посредством параллельных систем обработки, таких как Apache Spark. Для сбора и структурирования событийных данных об образовательной активности студентов используются стандарты Experience API и хранилища учебных записей (Learning Record Store), что позволяет организовать масштабируемый процесс хранения и подготовки данных для последующего машинного анализа.

Разработка адаптивных систем и индивидуализированных рекомендаций по обучению реализуется на основе специализированных библиотек и фреймворков для построения рекомендательных моделей. Среди них — TensorFlow Recommenders, LensKit, Surprise, которые используются для автоматического подбора образовательных траекторий, рекомендаций по выбору дисциплин, формированию индивидуальных учебных планов. Интеграция таких моделей в образовательные платформы позволяет учитывать индивидуальные предпочтения, уровень подготовки и динамику освоения материала каждым студентом.

Методы кластеризации в образовательной аналитике играют центральную роль в автоматическом выделении однородных групп студентов по совокупности параметров учебной активности, цифровых компетенций, стилей взаимодействия с ресурсами и результатов промежуточного тестирования. В

отличие от традиционных методов сегментации, опирающихся на заранее определённые признаки или формальные критерии, кластерный анализ выявляет внутренние паттерны, сегменты и профили, отражающие реальные различия в образовательных практиках, вовлечённости, стилях освоения материала. Использование кластеризации оправдано при анализе многомерных данных, где необходимо учесть широкий спектр характеристик, включая посещаемость, участие в обсуждениях, сроки сдачи заданий, взаимодействие с LMS, а также использование дополнительных сервисов и ресурсов. В практике аналитических исследований зафиксированы примеры применения k-средних для выделения сегментов студентов по активности во внешкольной работе, Fuzzy C-Means — для обнаружения профилей с размытыми границами между группами, агломеративной кластеризации — для построения иерархических структур и анализа динамики групп по мере изменения образовательной политики.

Кластеризация используется для диагностики академических рисков и раннего выявления студентов, испытывающих трудности с освоением программ, а также для выделения центров сетевой активности и изолированных групп. В образовательных исследованиях по моделям инверсного класса и смешанного обучения анализ активности на цифровых платформах позволяет выделять группы с разной степенью вовлечённости, на основании чего преподаватели и администрация разрабатывают адресные меры поддержки и корректируют содержание учебных курсов. Визуализация результатов кластеризации реализуется в форме тепловых карт, графов, интерактивных дашбордов, которые используются для принятия решений по персонализации образовательного процесса.

Методы построения профилей и индивидуализация сопровождения обучающихся опираются на результаты кластерного анализа и классификации, а также на средства визуализации ключевых образовательных индикаторов. В учебной аналитике полученные сегменты служат основой для дальнейшего внедрения персонализированных рекомендаций, разработки траекторий развития компетенций и адаптации учебных стратегий в зависимости от

характеристик группы. Данные аналитические подходы становятся инструментом управления качеством образовательных программ и повышения эффективности работы образовательных организаций.

Детальное рассмотрение методов кластеризации и классификации, их сравнительный анализ и особенности программной реализации в контексте образовательной аналитики позволяют сформировать целостное представление о возможностях и ограничениях современных подходов. В следующем разделе будет проведён анализ применения этих методов для выявления цифровых профилей студентов и генерации персонализированных рекомендаций на основе эмпирических данных.

1.4 Кластеризация и классификация в образовательной аналитике: методы, сравнительный анализ и программные средства

В исследовании, посвящённом анализу анкет студентов и оценке уровня цифровизации образовательной среды, используется комплексная стратегия применения современных методов машинного обучения. Обработка массивов анкетных данных студентов осуществляется с использованием алгоритмов кластерного анализа, классификации, а также средств автоматической генерации индивидуализированных текстовых рекомендаций. Кластеризация служит основным инструментом для обнаружения скрытых сегментов в данных без необходимости предварительной категоризации наблюдений. В контексте образовательной аналитики данный метод позволяет выделять естественные группы студентов, обладающих сходными характеристиками цифровой активности, предпочитаемыми форматами работы с образовательными ресурсами и типичными паттернами взаимодействия с электронными платформами вуза. Применение кластерного анализа обеспечивает профилирование обучающихся на основании эмпирически выявленных моделей поведения, что служит базой для построения дифференцированных сценариев поддержки и развития цифровых компетенций.

Классификация, как метод обучения с учителем, реализуется на следующем этапе после формирования цифровых профилей с помощью

кластеризации. Данная категория алгоритмов применяется для построения моделей, способных автоматически определять класс или категорию новых наблюдений на основе обучающей выборки, в которой каждая запись снабжена меткой. В образовательных исследованиях классификация используется для автоматизированного прогнозирования академической успешности, диагностики рисков отсева, идентификации студентов с потенциальной потребностью в поддержке, а также для решения задачи отнесения нового респондента к одному из ранее сформированных кластерных профилей цифровизации. В рамках рассматриваемого проекта механизм классификации позволил реализовать технологию автоматического определения цифрового профиля студента на основании его индивидуальных ответов в анкете, что обеспечивает возможность масштабируемого и объективного профилирования при появлении новых данных.

Генерация индивидуальных рекомендаций осуществляется с привлечением моделей генеративного искусственного интеллекта, обеспечивающих формирование текстовых советов на естественном языке с учётом совокупности признаков цифрового поведения каждого студента. Персонализация советов достигается за счёт анализа результатов кластеризации, выявленных в анкетных данных закономерностей и автоматической передачи описания профиля в языковую модель. В проекте реализована интеграция с API отечественной генеративной языковой модели GigaChat, что позволяет формировать содержательные и релевантные рекомендации для студентов с различными цифровыми профилями без привлечения экспертов и ручного анализа.

Комплексная реализация методов анализа данных и построения моделей машинного обучения осуществляется с использованием современного инструментария языка Python. Для работы с табличными данными применяется библиотека Pandas, которая предоставляет средства для чтения, очистки, агрегации и трансформации структурированных массивов, полученных на основе результатов анкетирования студентов. Модули библиотеки Scikit-learn

используются для построения и тестирования алгоритмов машинного обучения, в том числе кластеризации (k-means, иерархическая агломерация, нечеткие средние), классификации (решающие деревья, случайный лес, многослойные перцептроны), а также для проведения оценки качества моделей по стандартным метрикам. Для визуализации данных, промежуточных результатов обработки, анализа структуры кластеров и проверки корректности моделей используется библиотека Matplotlib, предоставляющая инструменты для построения различных видов графиков, диаграмм распределения, карт признаков и прочих средств аналитического представления результатов.

Интеграция генеративных языковых моделей достигается посредством использования API GigaChat, что обеспечивает автоматическую генерацию текстовых рекомендаций по развитию цифровых навыков студентов, обладающих разными профилями цифровизации. В процессе построения такой системы индивидуальных рекомендаций программное обеспечение формирует структурированный запрос к языковой модели, включающий описание профиля, типичные характеристики цифрового поведения, потенциальные слабые и сильные стороны, выявленные на предыдущих этапах анализа. Полученный от языковой модели текст сохраняется и используется для предоставления студенту в индивидуальном отчёте, а также для накопления статистики по наиболее востребованным образовательным стратегиям.

В аналитической части работы программная реализация включает последовательное применение кластеризации для выделения профилей студентов, обучение классификатора для автоматического определения профиля на новых данных, визуализацию структуры данных и кластеров с помощью графиков, а также генерацию персонализированных рекомендаций на основе профиля, полученного в ходе анализа. Весь цикл анализа реализован в среде Python с использованием открытых библиотек и интеграции с внешними языковыми сервисами.

Вывод по главе №1

В первой главе последовательно изложены подходы к анализу образовательных данных с применением методов машинного обучения, акцентировано внимание на современных методах профилирования студентов в условиях цифровой трансформации высшей школы. Рассматривается понятие цифровизации в образовании как процесс интеграции информационно-коммуникационных технологий на всех уровнях организации учебной и административной деятельности, приводятся особенности отличия цифровизации от традиционных форм дистанционного обучения. Раскрывается структура программных инициатив и стратегических документов в сфере цифровой трансформации образования, описывается влияние государственной политики и международных докладов на внедрение цифровых инструментов, а также формирование новых моделей образовательных траекторий.

Дается характеристика ключевых категорий цифровых образовательных инструментов, включая системы управления обучением, электронные образовательные ресурсы, коммуникационные и коллаборационные сервисы, специализированные приложения для поддержки учебного процесса, инновационные технологии на базе искусственного интеллекта, расширенной и виртуальной реальности, а также электронные сертификаты и механизмы подтверждения образовательных достижений. Описывается эволюция инструментов цифровизации, анализируются технологические и организационные предпосылки для повышения доступности, гибкости и индивидуализации учебной среды.

Определяются базовые принципы машинного обучения и методологические этапы построения моделей, раскрываются особенности обучения с учителем и без учителя, формулируются задачи классификации и кластеризации, приводятся ключевые алгоритмы построения и оценки моделей, затрагиваются вопросы предотвращения переобучения и оценки обобщающей способности. Описываются методы снижения размерности, поиска аномалий и анализа структуры данных, приводится классификация алгоритмов

кластеризации, включая итеративные, иерархические, плотностные, графовые и вероятностные подходы. Детализируется применение методов кластеризации и классификации для сегментации обучающихся, анализа профилей и автоматизации диагностики на больших массивах анкетных данных.

Проводится анализ программного инструментария для образовательной аналитики: описываются функциональные возможности библиотек Scikit-learn, Pandas, Matplotlib, рассматриваются среды визуального моделирования и бизнес-аналитики, такие как Orange, Tableau, PowerBI, анализируются современные решения для параллельной обработки событийных данных, интеграции систем учебной аналитики с LMS и платформами массового онлайн-обучения. Отдельное внимание уделяется появлению фреймворков для генерации персонализированных рекомендаций, интеграции языковых моделей и искусственного интеллекта в образовательную среду.

Кластеризация цифровых профилей студентов, классификация и генерация персонализированных рекомендаций

2.1 Разведочный анализ данных (EDA)

В рамках разведочного анализа анкетных данных студентов были проведены комплексные мероприятия, направленные на выявление структуры, оценку качества и изучение закономерностей исходной выборки. Начальный этап исследования включал импорт и первичную загрузку данных из файла Excel, содержащего ответы респондентов. Для корректной работы с числовыми характеристиками осуществлялась автоматическая конвертация строковых представлений чисел в числовой формат с целью предотвращения ошибок дальнейших статистических операций.

```
print("[EDA] Загрузка данных...")
df = pd.read_excel(INPUT_PATH)
print(f"Строк: {df.shape[0]}, Столбцов: {df.shape[1]}")
df.head().to_csv(os.path.join(OUTPUT_DIR, "head_students.csv"), index=False, encoding="utf-8-sig")
✓ 0.0s
[EDA] Загрузка данных...
Строк: 711, Столбцов: 23
```

Рисунок 2.1. Приведение строковых чисел к числовым типам

В ходе анализа структуры данных был выполнен поиск столбцов, отражающих факультетальную принадлежность студента. На основании найденного столбца построена визуализация распределения студентов по факультетам, демонстрирующая неоднородность представительства различных подразделений в выборке.

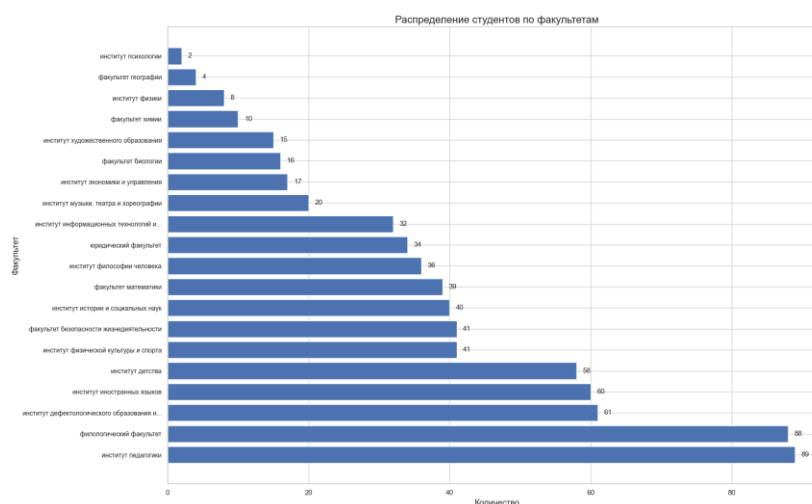


Рисунок 2.2. Распределение студентов по факультетам

Оценка полноты данных осуществлялась посредством анализа пропусков по всем признакам. Для признаков, обладающих пропущенными значениями, были построены гистограммы, отражающие долю пропусков по каждому из столбцов, что позволяет выявить потенциальные проблемные участки структуры анкеты.

По результатам анализа выявлено, что в имеющейся выборке отсутствуют существенные пропуски, что свидетельствует о высоком качестве сбора данных.

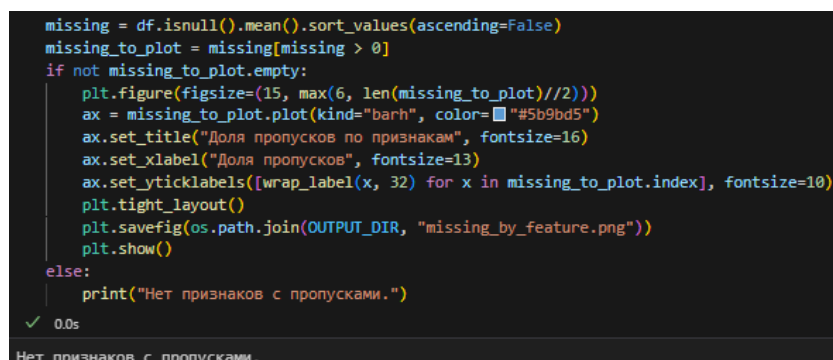


Рисунок 2.3. Доля пропусков по признакам анкеты

Следующим этапом стала идентификация бинарных признаков, имеющих два дискретных значения (например, «да/нет», «использую/не использую»). Для повышения корректности анализа все подобные ответы были приведены к единой числовой шкале (0/1) посредством применения словаря соответствий. Были выделены признаки, полностью удовлетворяющие бинарной природе, и произведён их количественный анализ.

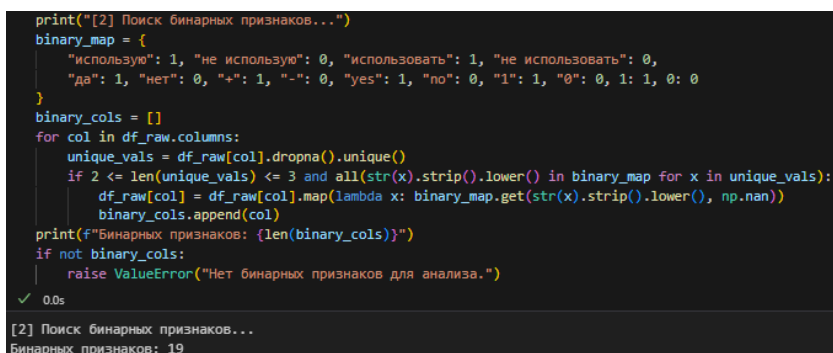


Рисунок 2.4. Определение и перевод бинарных признаков анкеты

Был выполнен отдельный анализ малоинформативных признаков, чьи значения практически не различаются между респондентами.

В ходе анализа не было выявлено бинарных признаков с выраженным дисбалансом распределения, все признаки признаны потенциально информативными для дальнейшего анализа.

```
bin_stats_df = pd.DataFrame(bin_stats)
uninformative = bin_stats_df[(bin_stats_df["frac_ones"] < 0.05) | (bin_stats_df["frac_ones"] > 0.95)]
n_uninformative = uninformative.shape[0]

print(f"Малоинформативных бинарных признаков (доля < 5% или > 95%): {n_uninformative}")
if n_uninformative > 0:
    print(uninformative)
    uninformative.to_csv(os.path.join(OUTPUT_DIR, "uninformative_binary_features.csv"), index=False)
else:
    print("Нет бинарных признаков с экстремально перекошенным распределением.")
✓ 0.0s
```

Малоинформативных бинарных признаков (доля < 5% или > 95%): 0
Нет бинарных признаков с экстремально перекошенным распределением.

Рисунок 2.5. Малоинформативные бинарные признаки с выраженным дисбалансом

Структура взаимосвязей между бинарными признаками была исследована с помощью boxplot-графика, а также корреляционной матрицы. Были определены пары признаков с максимальными абсолютными значениями корреляции, что может свидетельствовать о дублировании информации или выраженной сопряженности некоторых аспектов цифрового поведения студентов.

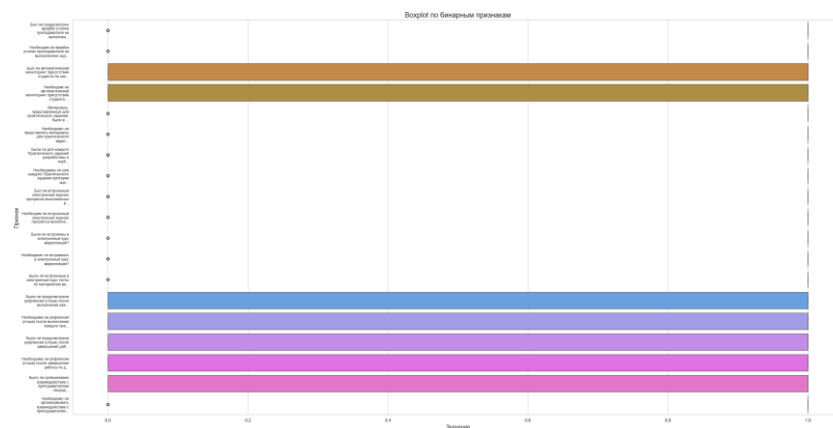


Рисунок 2.6. Boxplot по бинарным признакам анкеты

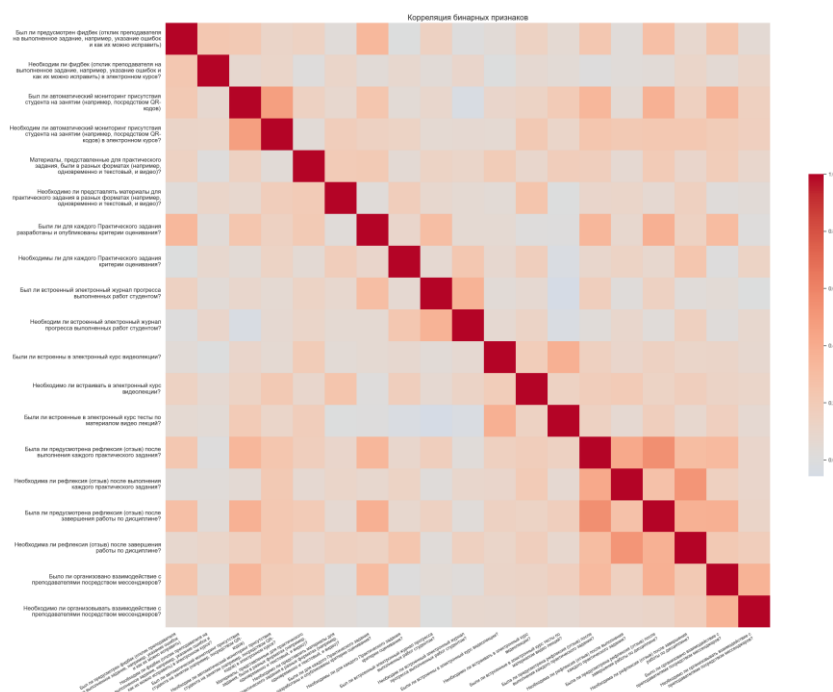


Рисунок 2.7. Матрица корреляции бинарных признаков

Для визуализации сбалансированных бинарных признаков были построены столбчатые диаграммы.

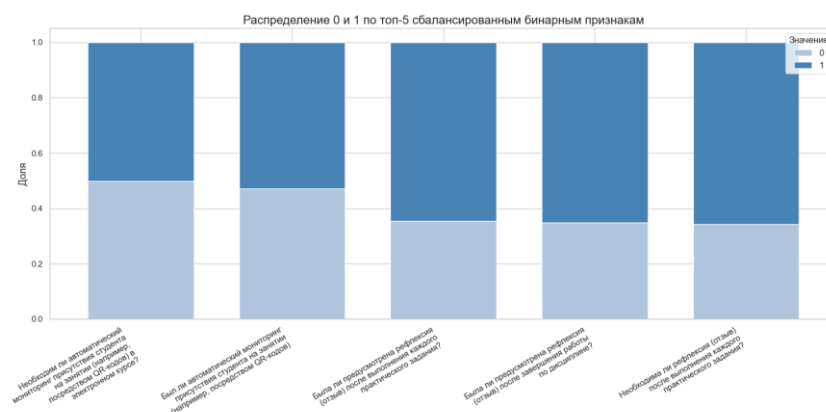


Рисунок 2.8. Диаграмма по сбалансированным бинарным признакам

В завершение была изучена взаимосвязь между бинарными признаками и факультетальной принадлежностью посредством расчёта коэффициента ассоциации Крамера (Cramér's V), что позволило количественно оценить силу связи между характеристиками цифрового поведения студентов и их учебным подразделением. Для наглядности результатов дополнительно строились столбчатые диаграммы, отражающие средние значения бинарных признаков по факультетам.

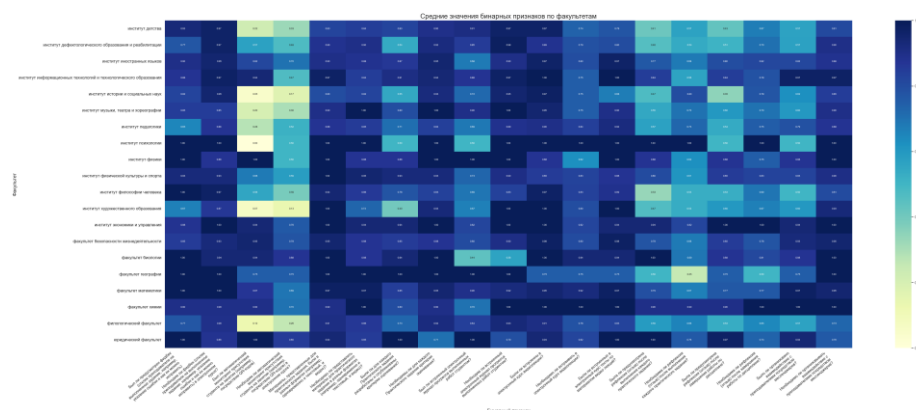


Рисунок 2.9. Тепловая карта бинарных признаков по факультетам

В результате проведённого этапа детального анализа выявлены ключевые структурные и содержательные характеристики исходной выборки, определены признаки с максимальной информативностью, обнаружены коррелирующие и малоинформативные переменные, а также количественно оценено влияние факультетской принадлежности на цифровое поведение студентов.

На основании результатов разведочного анализа следующим логическим шагом стала предобработка и преобразование признаков, позволяющие подготовить исходные данные к применению методов машинного обучения. Эти процедуры направлены на очистку, стандартизацию, устранение пропусков и приведение признаков к унифицированному формату, что обеспечивает корректность и воспроизводимость последующего анализа цифровых профилей студентов.

2.2 Предобработка данных и преобразование признаков

На этапе подготовки данных к дальнейшему анализу и моделированию выполнен ряд процедур, направленных на очистку, стандартизацию и преобразование исходной таблицы анкетных данных. Первоначальная загрузка данных осуществлялась из Excel-файла, содержащего ответы студентов, с последующим формированием структуры рабочей директории для хранения промежуточных и итоговых файлов.

В ходе первичной обработки данных были исключены все столбцы, содержащие временные метки, так как они не несут информативной нагрузки для последующего анализа. Для строковых признаков с пропущенными значениями

производилась замена на стандартное значение "нет", а для числовых — заполнение средним арифметическим по признаку. После очистки из таблицы удалялись строки, полностью состоящие из пропусков, а также признаки, не имеющие достаточного заполнения (менее 30% валидных значений).

```
df = df[[col for col in df.columns if "время" not in col.lower()]]
for col in df.columns:
    if df[col].dtype == object:
        df[col] = df[col].fillna("нет")
    else:
        df[col] = df[col].fillna(df[col].mean())
df = df.dropna(how='all')
df = df.dropna(axis=1, thresh=0.3 * len(df))
```

Рисунок 2.10. Удаление временных признаков, заполнение пропусков, фильтрация строк и столбцов

Для повышения интерпретируемости и совместимости с последующими этапами были стандартизированы бинарные признаки анкеты. Осуществлялось автоматическое приведение текстовых ответов, таких как "да", "нет", "использую", "не использую" и других, к единой числовой шкале 0/1 с помощью отображающего словаря.

```
binary_map = {
    "использую": 1, "не использую": 0,
    "использовать": 1, "не использовать": 0,
    "да": 1, "нет": 0,
    "Да": 1, "Нет": 0,
    "yes": 1, "no": 0,
    "Yes": 1, "No": 0,
    "1": 1, "0": 0,
    "1": 1, "0": 0,
}
binary_cols = []
for col in df.columns:
    if df[col].dtype == object:
        unique = df[col].dropna().unique()
        if all(str(x).strip().lower() in binary_map for x in unique):
            df[col] = df[col].map(lambda x: binary_map.get(str(x).strip().lower(), float("nan"))).astype(float)
            binary_cols.append(col)
print(f"[INFO] Найдено бинарных признаков: {len(binary_cols)} - {' '.join(binary_cols)}")
```

Рисунок 2.11. Преобразование бинарных признаков к числовому формату

Для организации корректного анализа факультетальных различий все обнаруженные столбцы, содержащие сведения о факультете или институте, были унифицированы в единую переменную "Факультет". При наличии дублирующих столбцов они исключались из итоговой таблицы. На основе полученного набора выделялись только признаки, значения которых ограничены множеством {0, 1}, что позволяет сформировать отдельную таблицу бинарных признаков для последующего кластерного анализа.

```

binaries = []
for col in df.columns:
    vals = set(df[col].dropna().unique())
    if vals <= {0, 1}:
        binaries.append(col)
bin_cols = binaries.copy()
if 'Факультет' in df.columns and 'Факультет' not in binaries:
    bin_cols.append('Факультет')
df_binaries = df[bin_cols]
df_binaries.to_csv(os.path.join(PREPROC_OUTPUT_DIR, "students_binaries.csv"), index=False)
print(f"[INFO] Бинарных признаков для ML-процессов: {len(binaries)} (с факультетом: {len(bin_cols)})")
✓ 0.0s
[INFO] Бинарных признаков для ML-процессов: 19 (с факультетом: 20)

```

Рисунок 2.12. Формирование набора бинарных признаков

Для числовых признаков осуществлялось масштабирование с использованием стандартного нормировщика (StandardScaler). Признаки с крайне низкой дисперсией (менее 0.01) исключались посредством отбора по порогу дисперсии (VarianceThreshold), что обеспечивает удаление почти константных столбцов и уменьшение размерности признакового пространства.

После отбора по дисперсии для последующего анализа использовано 20 числовых признаков с наибольшей вариативностью.

```

numeric_df = df.select_dtypes(include=[np.number])
scaler = StandardScaler()
X_scaled = scaler.fit_transform(numeric_df)
selector = VarianceThreshold(threshold=0.01)
X_selected = selector.fit_transform(X_scaled)
selected_features = numeric_df.columns[selector.get_support()]
pd.DataFrame(X_selected, columns=selected_features).to_csv(
    os.path.join(PREPROC_OUTPUT_DIR, "selected_features.csv"), index=False
)
with open(os.path.join(PREPROC_OUTPUT_DIR, "selected_feature_names.txt"), "w", encoding="utf-8") as f:
    f.write('\n'.join(selected_features))
print(f"[INFO] Числовых признаков после отбора: {len(selected_features)}")
✓ 0.0s
[INFO] Числовых признаков после отбора: 20

```

Рисунок 2.13. Масштабирование и отбор числовых признаков

Понижение размерности высоко размерного пространства реализовывалось посредством алгоритма UMAP, который позволяет перейти к двумерному представлению данных, сохраняя наиболее значимые топологические и кластерные структуры. Полученные двумерные координаты добавлялись к основной таблице.

Построенная визуализация двумерного пространства UMAP показала наличие выраженных компактных групп студентов, что указывает на потенциальную кластеризуемость исходных данных.

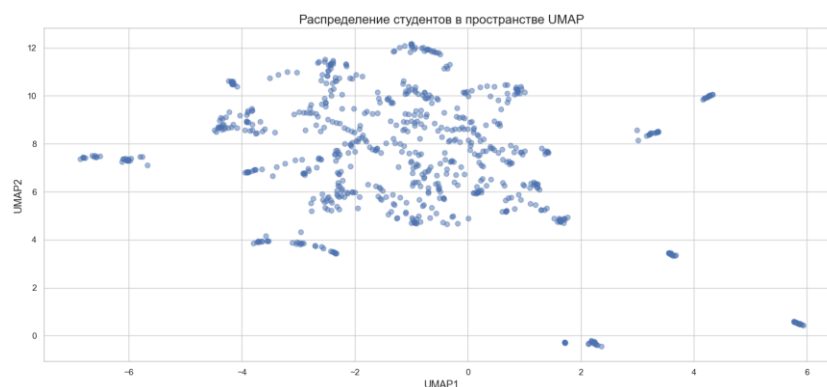


Рисунок 2.14. Визуализация двумерного пространства UMAP

Итоговые таблицы, включающие как исходные, так и бинарные и редуцированные данные, сохранялись на диск для обеспечения воспроизводимости дальнейших этапов анализа и возможности повторного использования в задачах моделирования.

На данном этапе выполнены все процедуры по приведению данных к формату, оптимальному для последующих методов кластеризации и построения предиктивных моделей.

На следующем этапе анализа акцент смещается на выявление однородных групп студентов с близкими цифровыми характеристиками. Для корректной работы алгоритмов кластеризации в условиях бинарных и категориальных данных использовалась матрица расстояний Говера, а также двумерные представления, полученные с помощью UMAP. Реализация различных методов кластеризации позволяет получить целостную картину цифровых профилей обучающихся, выявить устойчивые группы и обосновать дальнейшее построение персонализированных рекомендаций и моделей автоматической классификации.

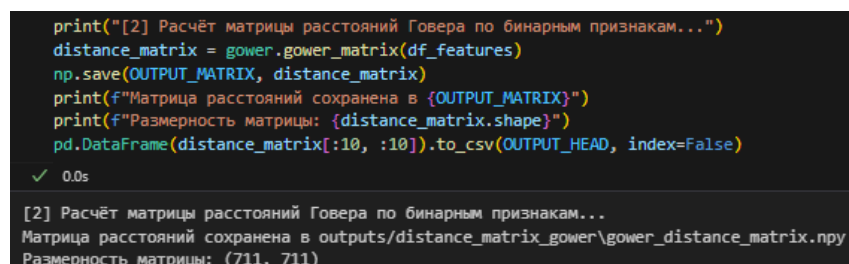
2.3 Кластеризация студентов (умар-признаки и матрица говера)

После завершения этапа предобработки данных следующим логическим

В ходе реализации кластерного анализа цифровых профилей студентов применялись методы, учитывающие специфику бинарных и категориальных данных. Ключевым этапом явилось формирование матрицы расстояний Говера, позволяющей корректно измерять попарные различия между объектами с

бинарными признаками. Для расчёта матрицы использовались только бинарные признаки, выделенные на предыдущем этапе подготовки данных.

В результате расчёта матрицы Говера были получены значения, характеризующие степень различия между каждым студентом по всем бинарным признакам. Это позволило количественно оценить цифровое сходство или различие студентов, выявить группы с минимальными и максимальными различиями.



```
print("[2] Расчёт матрицы расстояний Говера по бинарным признакам...")
distance_matrix = gower.gower_matrix(df_features)
np.save(OUTPUT_MATRIX, distance_matrix)
print(f"Матрица расстояний сохранена в {OUTPUT_MATRIX}")
print(f"Размерность матрицы: {distance_matrix.shape}")
pd.DataFrame(distance_matrix[:10, :10]).to_csv(OUTPUT_HEAD, index=False)
✓ 0.0s

[2] Расчёт матрицы расстояний Говера по бинарным признакам...
Матрица расстояний сохранена в outputs/distance_matrix_gower\gower_distance_matrix.npy
Размерность матрицы: (711, 711)
```

Рисунок 2.15. Расчёт и сохранение матрицы попарных расстояний Говера для бинарных признаков

Для наглядной иллюстрации паттернов схожести и различий между студентами по бинарным характеристикам строилась тепловая карта верхней части матрицы (100×100), где каждое значение отражает степень различия между двумя студентами: от полного совпадения (0) до максимального различия (1).

Тепловая карта показала наличие как плотных, так и разрежённых областей — то есть, в данных присутствуют как компактные группы схожих профилей, так и отдельные аномалии. Это свидетельствует о наличии явных кластеров среди студентов по цифровым признакам.

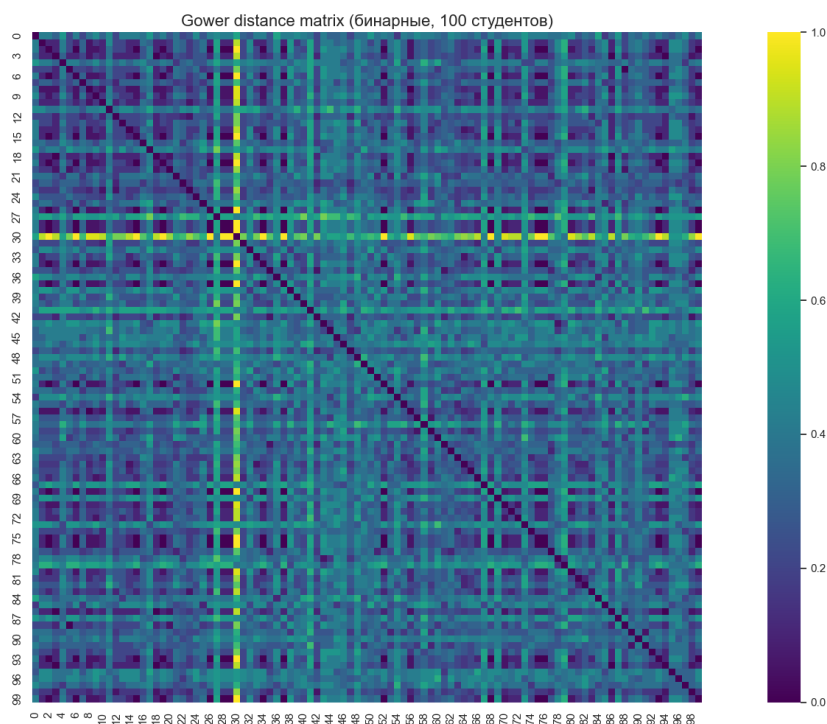


Рисунок 2.16. Фрагмент тепловой карты матрицы расстояний Говера

Следующим шагом стала реализация кластеризации, для чего были использованы два типа признаков: двумерные координаты, полученные посредством алгоритма UMAP, и сама матрица расстояний Говера. По UMAP-признакам выполнялись два вида кластеризации — агломеративная (иерархическая) и нечеткая c-means, что обеспечивало как жёсткое, так и мягкое разбиение на кластеры. Каждый подход проверялся на числе кластеров 4, а оптимальность определялась по метрикам силуэта и индексу Дэвиса–Болдуина.

Для бинарных признаков посредством матрицы расстояний Говера применялась агломеративная кластеризация с предвычисленной матрицей расстояний и средним типом связи. Для каждого разбиения также вычислялся коэффициент силуэта.

Итоговая таблица с признаками, кластерными метками и профилями сохранялась для последующего анализа и визуализации.

Оптимальное число кластеров для большинства алгоритмов оказалось равным четырём, что подтверждается анализом силуэта. Это указывает на наличие в выборке четырёх устойчивых цифровых профилей.

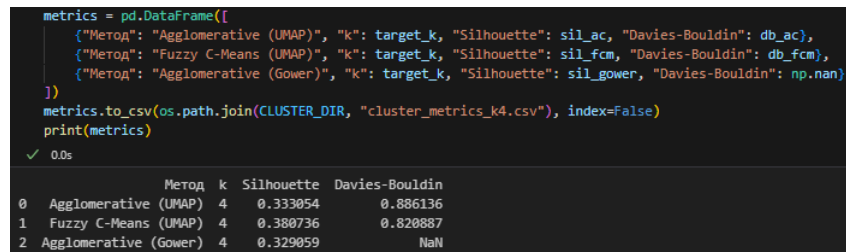


Рисунок 2.17. Результаты агломеративного, Fuzzy C-Means и Говера кластеризации с оценкой метрик качества

Для визуальной интерпретации результатов кластеризации строились диаграммы рассеяния студентов в пространстве UMAP, окрашенные по полученным профилям и меткам кластеров, а также линии зависимости silhouette score от числа кластеров для каждого метода.

Визуализация показала, что кластеры хорошо отделяются на плоскости UMAP, а профили сгруппированы достаточно компактно, что подтверждает валидность разбиения.

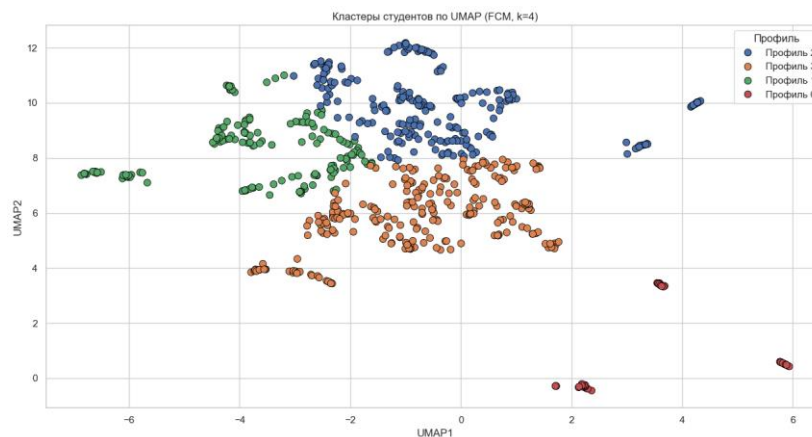


Рисунок 2.18. Визуализация кластеров по UMAP

Отдельные кластеры по матрице Говера могут содержать студентов из разных областей пространства UMAP, что отражает другую “грань” цифрового поведения — не только по совокупности признаков, но и по их индивидуальным сочетаниям.

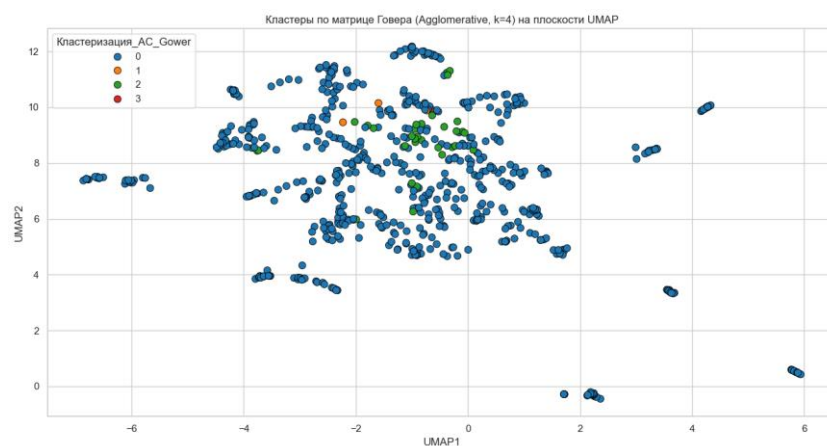


Рисунок 2.19. Визуализация кластеров по матрице Говера на плоскости UMAP

Кластеры, выделенные разными методами, частично совпадают, но в ряде случаев группы отличаются по составу, что демонстрирует разноплановость цифровых профилей. Это обогащает последующую интерпретацию и позволяет учитывать специфику разных подходов в построении рекомендаций.

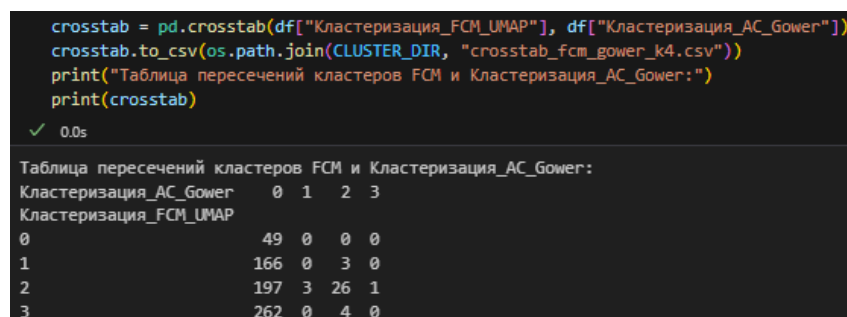


Рисунок 2.20. Пересечение кластеров FCM и Говера (таблица сопряжённости)

Результаты этого этапа обеспечили основу для дальнейшей интерпретации цифровых профилей студентов, их сравнительного анализа и построения моделей автоматической классификации.

Для объективной верификации полученных кластеров необходим формальный анализ их качества с использованием стандартных метрик, позволяющих количественно оценить как внутригрупповую сплочённость, так и различимость между кластерами. Следующий этап исследования посвящён комплексной оценке качества кластеризации с помощью коэффициента силуэта и индекса Дэвиса–Болдуина, а также анализу согласованности между результатами различных методов. Это позволит подтвердить устойчивость и

интерпретируемость выделенных цифровых профилей и обосновать их использование на следующих этапах моделирования и генерации рекомендаций.

2.4 Оценка качества кластеризации

Оценка качества разбиения студентов на кластеры осуществлялась с использованием стандартных метрик кластеризации, обеспечивающих формальное измерение разделимости и однородности полученных групп. Были выбраны коэффициент силуэта (Silhouette Score) и индекс Дэвиса–Болдуина (Davies-Bouldin Index), позволяющие комплексно оценивать как внутреннюю сплочённость кластеров, так и их различимость друг от друга. Анализ охватывал разбиения, полученные агломеративной кластеризацией и методом Fuzzy C-Means по пространству UMAP, а также агломеративной кластеризацией по матрице расстояний Говера.

Для кластеров, полученных в пространстве UMAP, рассчитывались обе метрики — Silhouette Score и Davies-Bouldin Index. Для кластеров, полученных по матрице Говера, рассчитывался только коэффициент силуэта с использованием предвычисленной матрицы расстояний.

Метрики Silhouette Score и Davies-Bouldin Index были рассчитаны для всех трёх методов кластеризации. Это позволило формально оценить, насколько хорошо выделяются кластеры в каждом варианте разбиения.

```

metrics = []
for method, col in label_cols.items():
    if col not in df.columns:
        print(f"ОШИБКА! Нет колонки '{col}' в df. Пропускаю.")
        continue
    labels = df[col].values
    if "Gower" in method:
        sil = silhouette_score(gower_matrix, labels, metric='precomputed')
        db = np.nan
    else:
        sil = silhouette_score(X_umap, labels)
        db = davies_bouldin_score(X_umap, labels)
    metrics.append({
        "Метод": method,
        "Метка": col,
        "Число кластеров": len(np.unique(labels)),
        "Silhouette Score": round(sil, 3),
        "Davies-Bouldin Index": round(db, 3) if not np.isnan(db) else "-"
    })

metrics_df = pd.DataFrame(metrics)
metrics_df.to_csv(os.path.join(OUTPUT_DIR, "clustering_quality_metrics.csv"), index=False)
display(metrics_df)

```

	Метод	Метка	Число кластеров	Silhouette Score	Davies-Bouldin Index
0	Agglomerative (UMAP)	Кластеризация_AC_UMAP	4	0.333	0.886
1	Fuzzy C-Means (UMAP)	Кластеризация_FCM_UMAP	4	0.381	0.821
2	Agglomerative (Gower)	Кластеризация_AC_Gower	4	0.329	—

Рисунок 2.21. Расчёт метрик качества кластеризации для каждого метода

Визуальное сравнение значений Silhouette Score между методами представлено в виде столбчатой диаграммы. Чем выше значение Silhouette, тем более отчётливо выделены кластеры.

Максимальное значение Silhouette Score (0.381) показал метод Fuzzy C-Means (UMAP), что свидетельствует о наибольшей отчётливости и разделимости кластеров среди всех рассмотренных подходов. Это значит, что студенты, попавшие в один и тот же кластер данным методом, обладают наиболее схожими цифровыми профилями, а различия между кластерами выражены наиболее ярко.

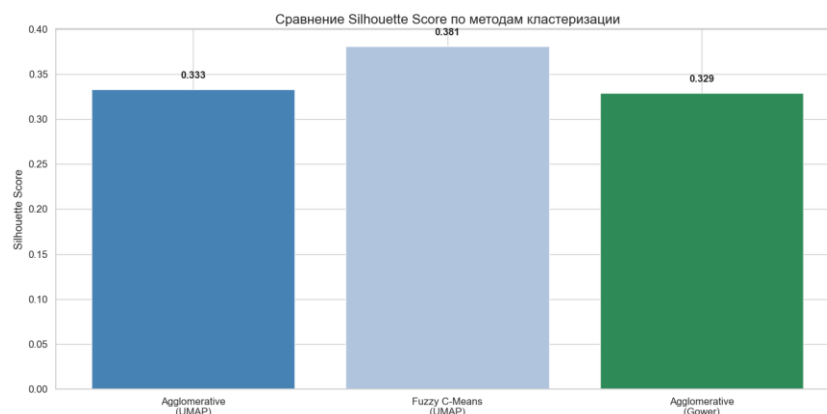


Рисунок 2.22. Сравнение Silhouette Score для разных методов кластеризации

Для методов, поддерживающих индекс Дэвиса–Болдуина, значения также были отображены на соответствующей диаграмме. Более низкое значение этой метрики свидетельствует о более качественном разбиении.

Более низкие значения Davies-Bouldin Index (0.821 у FCM и 0.886 у Agglomerative по UMAP) также подтверждают приемлемое качество разбиения: кластеры компактны внутри и хорошо отделены друг от друга. Отсутствие этого индекса для метода по Говеру связано с особенностями работы метрики на предвычисленных матрицах расстояний.

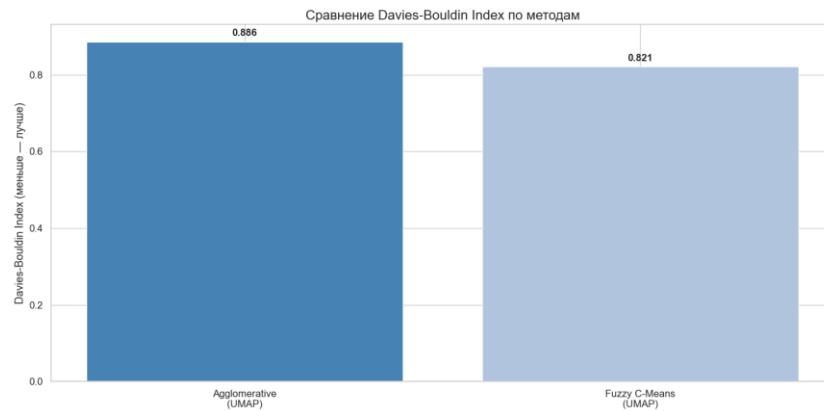


Рисунок 2.23. Сравнение Davies-Bouldin Index для разных методов кластеризации

Для анализа согласованности между разными алгоритмами проводился кросс-анализ кластеров по FCM (UMAP) и по Говеру — строилась таблица сопряжённости, отображающая, сколько объектов попало в пересечение между каждым парой кластеров двух методов.

Кросстаб позволяет проанализировать, как совпадают или различаются результаты разбиения по двум методикам: часть студентов стабильно попадает в одни и те же кластеры, а часть — распределяется по-разному. Это подтверждает устойчивость выделенных профилей, но также демонстрирует, что различные алгоритмы "видят" структуру цифрового поведения студентов по-разному.

Таблица пересечений кластеров FCM и Кластеризация_AC_Gower:				
Кластеризация_AC_Gower	0	1	2	3
Кластеризация_FCM_UMAP				
0	49	0	0	0
1	166	0	3	0
2	197	3	26	1
3	262	0	4	0

Рисунок 2.24. Таблица пересечений кластеров FCM и Говера

Результаты оценки метрик свидетельствуют о приемлемом разделении кластеров по всем методам, однако максимальное значение Silhouette Score наблюдалось для агломеративной кластеризации по UMAP. Наблюдение за значениями индекса Дэвиса–Болдуина также подтвердило хорошее качество кластеризации. Кросстаб помог идентифицировать пересечения и различия разбиений, а также подтвердил устойчивость найденных профилей при использовании различных подходов.

Следующим этапом исследования стала содержательная интерпретация и профилирование выявленных кластеров студентов на основе анализа распределения бинарных признаков. Комплексное рассмотрение ключевых характеристик цифрового поведения внутри каждого профиля позволяет выявить существенные различия между группами, а также определить уникальные особенности каждого цифрового профиля. Проведённая интерпретация служит основой для последующего назначения содержательных наименований профилей, а также для формирования персонализированных рекомендаций на следующих этапах исследования.

2.5 Интерпретация и профилирование кластеров по бинарным признакам

Для разработанной автоматизированной системы генерации и проверки заданий в качестве входных данных требуются фамилия, имя обучающегося и цифра для выбора действия: «1» для генерации файла задания и «2» для его проверки.

В целях содержательной интерпретации и профилирования выделенных кластеров студентов был проведён комплексный анализ распределения бинарных признаков внутри каждого профиля. Исходные данные включали для каждого студента информацию о принадлежности к определённому кластеру, факультету, а также о значениях бинарных характеристик цифрового поведения.

Для выявления наиболее характерных различий между кластерами был рассчитан абсолютный размах долей для каждого признака — разница между максимальным и минимальным средним значением признака по всем профилям. Признаки с максимальным размахом были определены как ключевые для интерпретации различий между группами студентов. Визуализация этих признаков представлена в виде barplot-графика.

Barplot наглядно демонстрирует, какие бинарные признаки (например, “Был ли предусмотрен фидбек...”, “Необходима ли рефлексия...”) сильнее всего различают кластеры студентов. Кластеры выраженно отличаются по долям

студентов с этими признаками, что указывает на ключевые особенности каждого цифрового профиля.

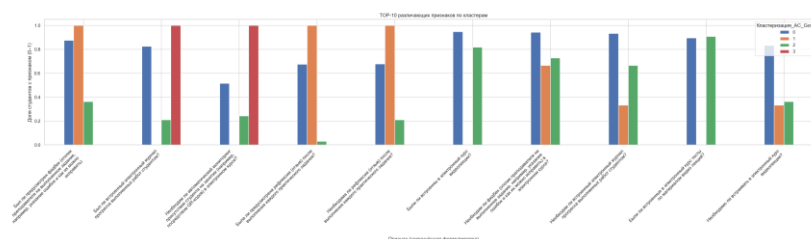


Рисунок 2.25. Barplot: топ-10 различающих бинарных признаков между кластерами

Для комплексной визуализации структурных различий между профилями студентов был построен радиальный график (radar plot), отражающий доли выраженности топовых признаков по каждому кластеру. Подобная форма визуализации позволяет сразу увидеть "силуэт" каждого цифрового профиля.

Радиальная диаграмма позволяет сразу увидеть “силуэт” и уникальные черты каждого цифрового профиля: например, один кластер может иметь максимальные значения по большинству признаков цифровизации, другой — только по отдельным аспектам, третий — минимальные значения почти по всем признакам.



Рисунок 2.26. Радиальная диаграмма (radar plot) по топ-10 различающих признаков профилей кластеров

Для подготовки интерпретируемых текстовых описаний каждого профиля был сформирован автоматический шаблон. В нем для каждого кластера перечислялись наиболее выраженные и наименее выраженные признаки, что облегчает содержательную интерпретацию полученных групп и может служить основой для назначения содержательных названий цифровых профилей.

Шаблон облегчает содержательную интерпретацию выделенных групп: сразу видно, какие признаки максимально выражены в профиле, а какие встречаются редко. Это помогает вручную присвоить “человеческие” имена профилям, такие как “Цифровые энтузиасты”, “Скептики”, “Консерваторы” и др.

Шаблон для содержательной интерпретации профилей кластеров

Кластер 0:

Наиболее выраженные признаки (выделяют этот кластер):

- Были ли встроены в электронный курс видеолекции?: 0.95
- Необходим ли фидбек (отклик преподавателя на выполненное задание, например, указание ошибок и как их можно исправить) в электронном курсе?: 0.94
- Необходим ли встроенный электронный журнал прогресса выполненных работ студентом?: 0.93
- Были ли встроенные в электронный курс тесты по материалам видео лекций?: 0.89
- Был ли предусмотрен фидбек (отклик преподавателя на выполненное задание, например, указание ошибок и как их можно исправить): 0.88
- Необходимо ли встраивать в электронный курс видеолекции?: 0.83
- Был ли встроенный электронный журнал прогресса выполненных работ студентом?: 0.82
- Необходима ли рефлексия (отзыв) после выполнения каждого практического задания?: 0.68
- ...
- Необходимо ли встраивать в электронный курс видеолекции?: 0.00

Рисунок 2.27. Автоматически сформированный шаблон интерпретации профилей кластеров

Лучшую интерпретируемость и качественное разбиение профилей студентов показал метод Agglomerative Clustering по матрице Говера. На barplot и radar plot видно, что кластеры существенно различаются по ряду бинарных признаков анкеты: выделены наиболее значимые для каждого профиля.

Для дальнейшей детализации и наглядного сравнения цифровых профилей студентов на следующем этапе исследования были построены полярные (radar) диаграммы. Этот подход позволяет визуализировать средние значения бинарных

признаков по каждому кластеру в едином графическом формате, что существенно облегчает выявление уникальных и общих черт между профилями. Ниже приведён подробный анализ построения и интерпретации полярных диаграмм, отражающих специфику цифрового поведения студентов в рамках выделенных кластеров.

2.6 Полярные (radar) диаграммы профилей кластеров

Для визуализации различий между цифровыми профилями студентов использовались полярные (radar) диаграммы, которые позволяют отразить средние значения бинарных признаков для каждого выделенного кластера на единой круговой сетке. Такой подход предоставляет интуитивно понятный способ сравнения ключевых характеристик профилей, выявленных в процессе кластерного анализа.

На предварительном этапе осуществлялось приведение исходных ответов к единому бинарному формату (0/1) на основании словаря сопоставления, который покрывает типовые варианты ответа («да/нет», «использую/не использую», «+/-» и аналогичные). Были автоматически выделены все признаки, удовлетворяющие бинарному критерию.

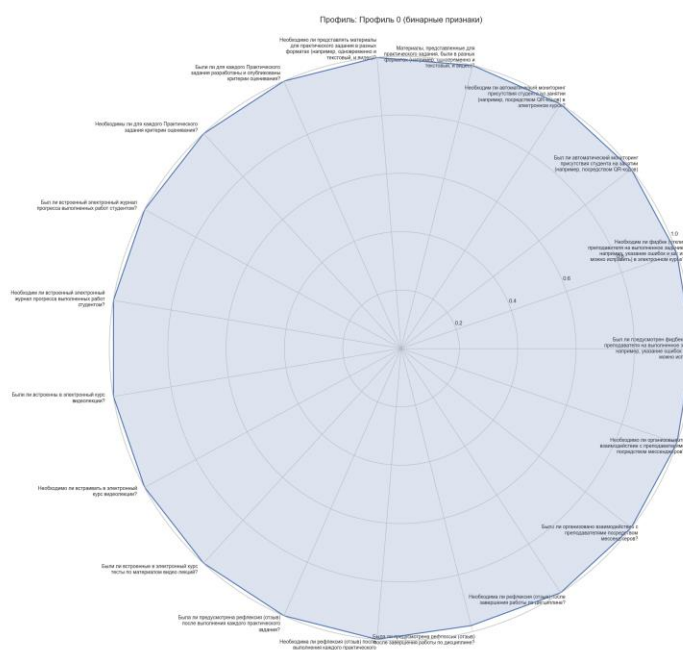
```
print("[2] Поиск бинарных признаков...")
binary_map = {
    "использую": 1, "не использую": 0, "использовать": 1, "не использовать": 0,
    "да": 1, "нет": 0, "+": 1, "-": 0, "yes": 1, "no": 0, "1": 1, "0": 0, 1: 1, 0: 0
}
binary_cols = []
for col in df_raw.columns:
    unique_vals = df_raw[col].dropna().unique()
    if 2 <= len(unique_vals) <= 3 and all(str(x).strip().lower() in binary_map for x in unique_vals):
        df_raw[col] = df_raw[col].map(lambda x: binary_map.get(str(x).strip().lower(), np.nan))
        binary_cols.append(col)
print(f"Бинарных признаков: {len(binary_cols)}")
if not binary_cols:
    raise ValueError("Нет бинарных признаков для анализа.")
✓ 0.0s
[2] Поиск бинарных признаков...
Бинарных признаков: 19
```

Рисунок 2.28. Определение бинарных признаков в исходных данных

На следующем этапе метки профиля, полученные по итогам кластеризации, были совмещены с исходной таблицей бинарных признаков.

Для каждого профиля (кластера) были рассчитаны средние значения бинарных признаков — это отражает долю студентов в кластере, отмечающих тот или иной признак. Далее, для каждой группы строилась индивидуальная

Профиль 0 характеризуется практически максимальными значениями по большинству бинарных признаков цифрового поведения (почти все значения близки к 1). Это свидетельствует о том, что в этот кластер попали студенты с самой широкой цифровой вовлечённостью: у них реализованы почти все цифровые сервисы и активности, связанные с обучением. Такой профиль можно трактовать как «цифровые энтузиасты» или «максимально вовлечённые».



Профиль 1 отличается высокой выраженностью рефлексии и обратной связи, однако по ряду других цифровых инструментов значения существенно ниже, чем у профиля 0. Это может быть группа студентов, для которых критично наличие рефлексии и взаимодействия с преподавателем, но не обязательны все цифровые сервисы.

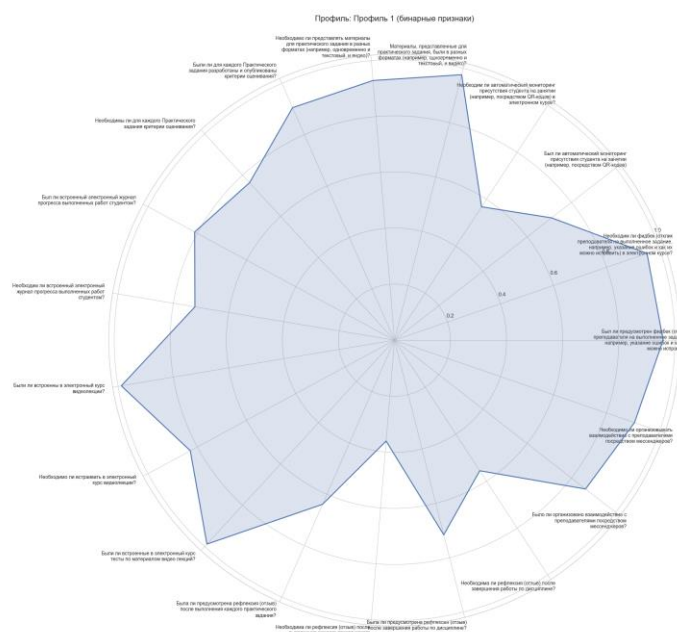


Рисунок 2.30. Полярная (radar) диаграмма: профиль 1

Профиль 2 имеет сбалансированное распределение по большинству признаков, без выраженного перекаса к высоким или низким значениям. Это студенты с умеренной цифровой вовлечённостью: им характерна частичная реализация цифровых сервисов, но нет явного лидерства по большинству признаков.

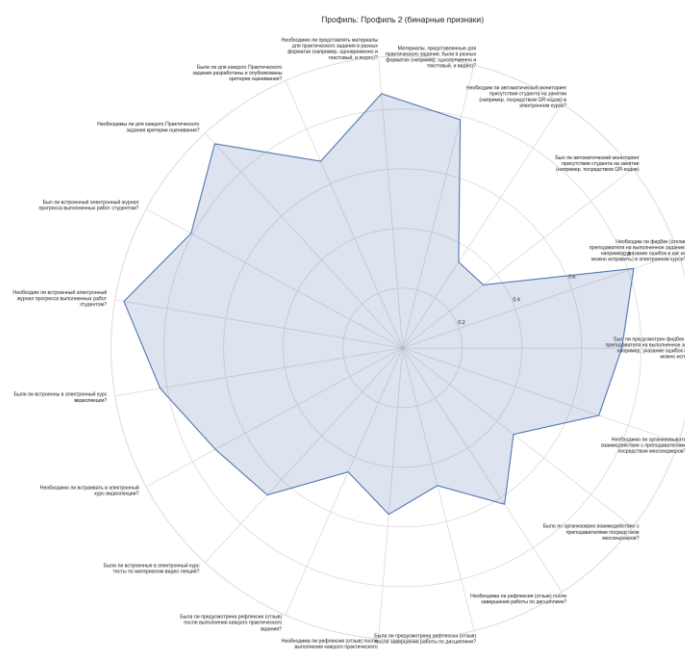


Рисунок 2.31. Полярная (radar) диаграмма: профиль 2

Профиль 3, напротив, показывает очень низкие значения почти по всем признакам — большинство осей диаграммы расположены близко к центру. Это может быть группа скептиков или минимально вовлечённых студентов, для которых цифровизация учебного процесса практически не реализована.

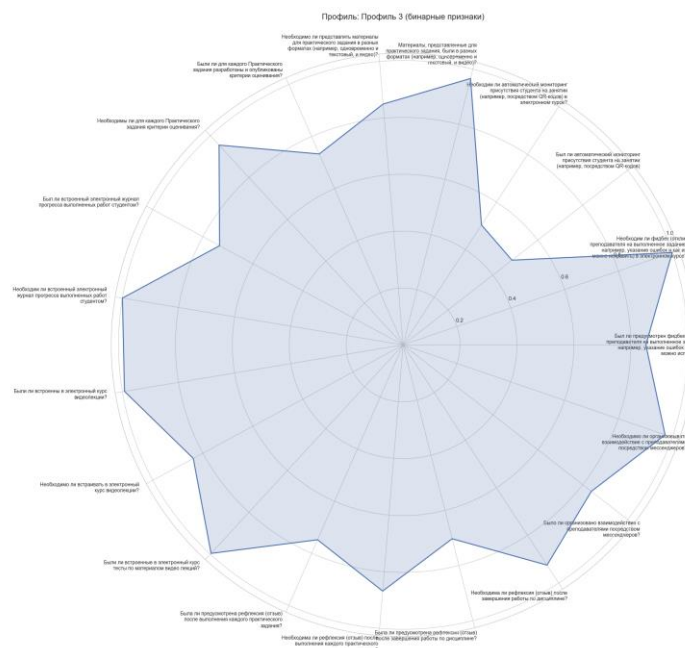


Рисунок 2.32. Полярная (radar) диаграмма: профиль 3

Каждая из полученных диаграмм даёт наглядное представление о том, по каким признакам цифрового поведения тот или иной кластер отличается от остальных. Значения, максимально удалённые от центра, соответствуют наиболее характерным чертам данного профиля. Это облегчает интерпретацию профилей и выявление ключевых различий между группами студентов.

Полученные профили легли в основу дальнейшей практической задачи — автоматизации процесса отнесения новых студентов к одному из выявленных цифровых кластеров. Для этого на следующем этапе исследования была реализована процедура построения и оценки моделей классификации, позволяющих предсказывать профиль студента на основе его анкетных данных.

2.7 Построение и оценка классификатора профиля

Для решения задачи автоматической классификации цифрового профиля студента была построена серия моделей, призванных предсказывать принадлежность к кластеру на основе исходных признаков анкеты. В качестве

целевой переменной выступала метка профиля, определённая по результатам Fuzzy C-Means кластеризации.

В качестве исходных переменных использовались числовые и бинарные признаки анкеты, за исключением признаков, непосредственно связанных с результатами кластеризации и координатами в пространстве UMAP. После отбора и масштабирования признаков применялся фильтр по порогу дисперсии, что позволило исключить малозначимые переменные.

Множество данных было разделено на обучающую и тестовую выборки с сохранением пропорций классов, что обеспечивает корректную оценку качества моделей.

Для задачи мультиклассовой классификации обучались три типа моделей: решающее дерево, случайный лес и многослойный персептрон (MLP). Для каждой модели рассчитывались метрики качества (precision, recall, f1-score) и средний F1_macro по 5-кратной кросс-валидации.

На основании результатов кросс-валидации наилучшее качество классификации цифровых профилей показал алгоритм RandomForest. Эта модель достигла наивысших значений F1_macro, демонстрируя стабильные показатели точности на различных подвыборках данных.

```
>>> Обучение модели DecisionTree...
      precision    recall  f1-score   support

0       0.99      0.99      0.99      135
1       0.00      0.00      0.00       0
2       0.86      0.86      0.86       7

 accuracy          0.98      142
 macro avg          0.62      142
weighted avg          0.99      142

DecisionTree: accuracy=0.979, F1_macro_CV=0.577±0.119

>>> Обучение модели RandomForest...
      precision    recall  f1-score   support

0       0.99      1.00      0.99      135
1       0.00      0.00      0.00       0
2       1.00      0.71      0.83       7

 accuracy          0.99      142
 macro avg          0.66      142
weighted avg          0.99      142

RandomForest: accuracy=0.986, F1_macro_CV=0.634±0.217

>>> Обучение модели MLP...
      precision    recall  f1-score   support

0       0.96      0.96      0.96      135
1       0.00      0.00      0.00       0
2       0.17      0.14      0.15       7

 accuracy          0.92      142
 macro avg          0.37      142
weighted avg          0.92      142

MLP: accuracy=0.923, F1_macro_CV=0.454±0.163
```

Рисунок 2.33. Качество классификаторов: отчёты и кросс-валидация по F1_macro

Наилучшее качество демонстрировал случайный лес (RandomForest), что подтверждалось высокими значениями $F1_macro$ и устойчивостью метрик на кросс-валидации. Для данной модели дополнительно анализировалась важность признаков — строились Barplot-графики для топ-15 наиболее значимых переменных, что даёт возможность объяснить, какие характеристики анкеты вносят наибольший вклад в определение профиля.

Barplot иллюстрирует, что наиболее значимыми для предсказания профиля оказываются такие признаки, как наличие электронного журнала прогресса, видеолекций и организация обратной связи с преподавателем. Это подтверждает, что именно эти элементы образовательного процесса наиболее чувствительны к различиям между профилями студентов.

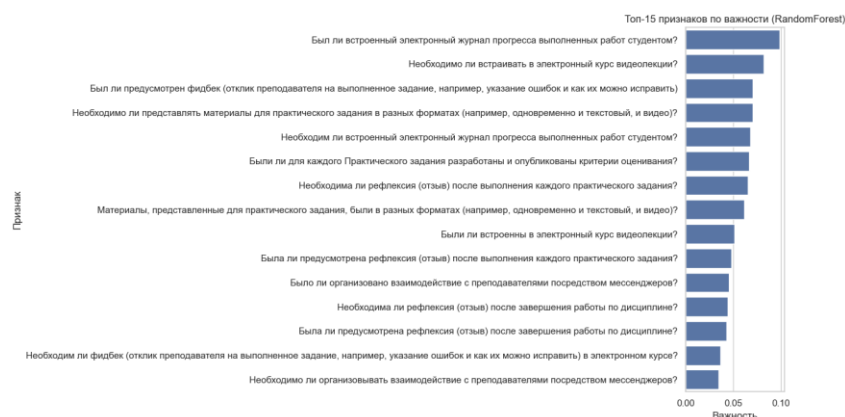


Рисунок 2.34. Топ-15 признаков по важности (RandomForest)

Полученные модели и конфигурации признаков сохранялись для последующего использования на этапе генерации персонализированных рекомендаций. Проведённая серия экспериментов продемонстрировала высокую предсказательную способность построенных моделей, а также выявила наиболее значимые характеристики, определяющие цифровой профиль студента.

Высокие показатели качества построенных классификационных моделей и выявление ключевых признаков, определяющих цифровой профиль студента, открывают возможность их практического применения для адресной поддержки обучающихся. На следующем этапе работы эти результаты были интегрированы в модуль автоматической генерации персонализированных рекомендаций. Такой

модуль позволяет не только оперативно определять профиль нового студента на основании его анкетных данных, но и формировать индивидуальные рекомендации с использованием возможностей искусственного интеллекта.

2.8 Генерация персонализированных рекомендаций

На заключительном этапе реализован модуль автоматической генерации персонализированных цифровых рекомендаций для студентов на основе их анкетных данных и выделенного цифрового профиля. Алгоритм строится на связке предсказания профиля с помощью обученного классификатора и последующей генерации текстовых советов с использованием API GigaChat, что обеспечивает как автоматизацию аналитики, так и индивидуальный подход в образовательной поддержке.

Процесс начинается с получения access token через OAuth-авторизацию в системе GigaChat. Токен необходим для безопасного обращения к API генерации текста.

На данном этапе выполняется получение защищенного токена доступа через протокол OAuth для использования API GigaChat. Этот токен необходим для безопасного взаимодействия с моделью искусственного интеллекта при формировании рекомендаций.

```
def get_gigachat_token(auth_key):
    TOKEN_URL = "https://ngw.devices.sberbank.ru:9443/api/v2/oauth"
    headers = {
        "Content-Type": "application/x-www-form-urlencoded",
        "Accept": "application/json",
        "RqUID": "01b9786e-c315-43ef-b955-33008e01c884",
        "Authorization": f"Basic {auth_key}",
    }
    data = {"scope": "GIGACHAT_API_PERS"}
    resp = requests.post(TOKEN_URL, headers=headers, data=data, verify=False)
    resp.raise_for_status()
    return resp.json()['access_token']
```

Рисунок 2.35. Получение OAuth токена для GigaChat API

Следующий шаг — загрузка обученной модели классификации, масштабировщика, селектора признаков и списка отобранных признаков. Профиль нового студента определяется автоматически по данным анкеты.

Для нового студента данные анкеты приводятся к формату, необходимому для работы классификатора. После обработки и масштабирования признаков

обученная модель автоматически предсказывает профиль (кластер) студента, что позволяет перейти к этапу персонализированной генерации рекомендаций.

```
scaler, selector, model, selected_features, le = joblib.load(MODEL_PATH)
selected_features = list(selected_features)

new_student = pd.read_excel(NEW_FORM_PATH)
new_student.columns = [str(c).strip() for c in new_student.columns]

missing = set(selected_features) - set(new_student.columns)
if missing:
    print(f"[!] В анкете не хватает признаков: {missing}")
    for col in missing:
        new_student[col] = 0

student_X = new_student[selected_features].copy()
student_X = student_X.iloc[[0]]
student_X = scaler.transform(student_X)
student_X = selector.transform(student_X)
```

Рисунок 2.36. Автоматическое определение цифрового профиля нового студента

Для генерации персональных рекомендаций формируется детализированный prompt, в котором отражаются как тип профиля, так и индивидуальные особенности использования цифровых инструментов. Затем запрос отправляется в модель GigaChat с использованием токена.

Сформированный prompt содержит как результаты анкеты, так и краткое описание цифрового профиля студента. Этот prompt отправляется в API GigaChat, который на основе искусственного интеллекта генерирует структурированные и индивидуализированные рекомендации по развитию цифровых компетенций.

```
def make_prompt(student_row, profile_label):
    text = f"Анкета студента:\n{student_row.to_string(index=True)}\n\nПрофиль кластера: {profile_label}\n"
    text += "Сформируй индивидуальные рекомендации по цифровому развитию для этого студента. Формулируй понятно и структурировано, желательно в виде списка."
    return text

prompt = make_prompt(new_student.iloc[0], profile_label)
try:
    ACCESS_TOKEN = get_gigachat_token(AUTH_KEY)
    HEADERS = {"Authorization": f"Bearer {ACCESS_TOKEN}" }
    API_URL = "https://gigachat.devices.sberbank.ru/api/v1/chat/completions"
    payload = {"model": "GigaChat", "messages": [{"role": "user", "content": prompt}]}
    response = requests.post(API_URL, json=payload, headers=HEADERS, timeout=30, verify=False)
    response.raise_for_status()
    recommendations = response.json().get("choices", [{}])[0].get("message", {}).get("content", "")
except Exception as e:
    recommendations = f"[ОШИБКА ПРИ ЗАПРОСЕ К GigaChat]\n{str(e)}"
print(recommendations)
```

Рисунок 2.37. Вызов GigaChat API и получение индивидуальных цифровых рекомендаций

Полученные рекомендации сохраняются как в текстовый, так и PDF-формат для последующего использования студентом или преподавателем.

Генерируемые рекомендации автоматически сохраняются в двух форматах — txt и PDF. Это обеспечивает удобство последующего хранения, передачи и печати материалов как для студента, так и для преподавателя.

```
with open(OUTPUT_TXT, "w", encoding="utf-8") as f:
    f.write(recommendations)

pdf = FPDF()
pdf.add_page()
pdf.add_font('DejaVu', '', FONT_PATH, uni=True)
pdf.set_font("DejaVu", size=12)
for line in recommendations.split('\n'):
    pdf.multi_cell(0, 10, line)
pdf.output(OUTPUT_PDF)

print(f"\nГотово! Рекомендации сохранены в {OUTPUT_TXT} и {OUTPUT_PDF}")
if missing:
    print("\n[ВНИМАНИЕ] В анкете не было следующих признаков, они заполнены нулями для предсказания:")
    for m in missing:
        print(" -", m)
```

Рисунок 2.38. Сохранение рекомендаций в текстовый и PDF-формат

Финальный результат работы модуля — индивидуальные цифровые рекомендации, созданные на основании профиля студента и его анкетных данных. Пример типичных рекомендаций включает советы по развитию конкретных цифровых навыков, рекомендации по использованию электронных платформ и развитию медиаграмотности. Такой подход позволяет обеспечивать адресную поддержку каждому обучающемуся, повышая эффективность цифровизации образовательного процесса.

Вот несколько рекомендаций по цифровому развитию для данного студента:

1. **Развитие навыков анализа мультимедийного контента**

Поскольку студент отмечает, что материалы для практических заданий представлены в различных форматах (текст, видео), рекомендуется углубить навыки анализа информации из разнообразных источников. Можно рекомендовать курсы по медиаграмотности, работе с различными типами данных и анализу информации.

2. **Активное использование возможностей платформы Moodle**

Учитывая, что студент уже знаком с платформой Moodle, полезным будет изучить её более глубоко. Это поможет лучше организовать учебный процесс, использовать все возможности платформы для самоподготовки и взаимодействия с преподавателем. Рекомендуется пройти онлайн-курсы по управлению обучением на платформе Moodle.

3. **Повышение эффективности работы с обратной связью**

Важно научиться правильно воспринимать и применять полученную от преподавателей обратную связь. Стоит уделить внимание курсам по саморефлексии и улучшению коммуникативных навыков, чтобы эффективно взаимодействовать с преподавателями и получать максимальную пользу от обратной связи.

Рисунок 2.39. Рекомендации для студента

Данный этап завершает автоматизированный цикл анализа: от первичной анкеты и распознавания профиля до генерации персональных рекомендаций с

помощью искусственного интеллекта. Технологическое решение интегрирует классификацию и генерацию текста, позволяя обеспечить адресную цифровую поддержку каждому студенту на основе объективных данных.

Однако для повышения управленческой и педагогической ценности результатов возникает задача анализа распределения сформированных профилей внутри образовательной организации, выявления особенностей их структуры по факультетам и построения интерпретируемых описаний для практического применения. Следующий этап работы направлен на комплексную оценку распространённости различных цифровых профилей среди студентов разных факультетов, а также автоматическую генерацию содержательных характеристик выявленных групп на основе информативных признаков анкеты.

2.9 Анализ распределения цифровых профилей по факультетам и автоматическая интерпретация

Для углублённого анализа цифровых профилей студентов по факультетам реализован автоматизированный модуль, позволяющий не только выявить различия между подразделениями, но и получить интерпретируемые характеристики каждого профиля на основании числовых и бинарных признаков анкеты.

Анализ распределения профилей между факультетами строился на группировке студентов по признаку факультета и подсчёте процентного соотношения каждого цифрового профиля в пределах отдельного факультета. Визуализация выполнялась в виде *stacked barplot*, наглядно показывающего пропорции цифровых профилей по всем факультетам.

Каждый столбец на графике соответствует отдельному институту. Цветовая гамма показывает распределение профилей — чем выше процент конкретного цвета в столбце, тем больше студентов данного профиля на соответствующем факультете. Такой подход позволяет быстро выявить факультеты с доминированием определённых цифровых профилей, что может быть использовано для последующей адресной поддержки.

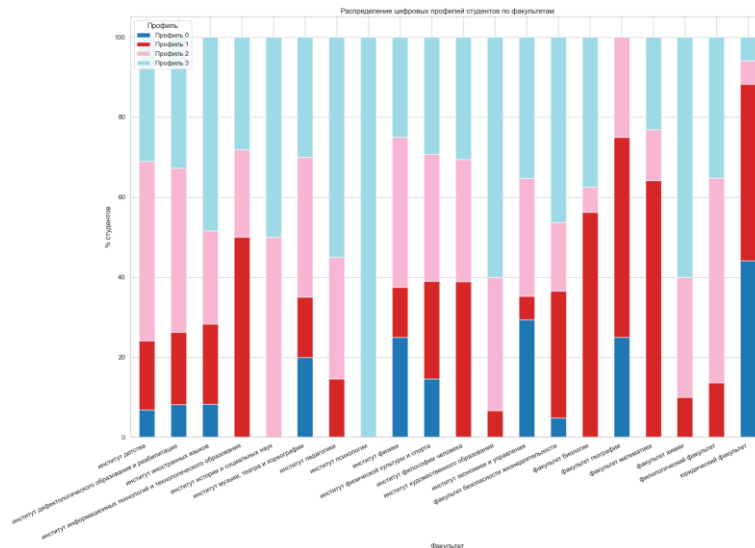


Рисунок 2.40. Распределение цифровых профилей студентов по факультетам (stacked barplot)

Для дальнейшей детализации профилей по признакам формировался список числовых переменных, не включающих технические столбцы и идентификаторы. Это обеспечивает анализ только информативных характеристик студентов.

Это обеспечивает, что при построении профиля используются только реально характеризующие цифровое поведение признаки, без технических и дублирующих переменных.

```
technical_cols = [
    "ID", "Кластер_AC", "Кластер_FCM", "UMAP1", "UMAP2", "Факультет", "Профиль"
]
feature_cols = [
    col for col in df.columns
    if col not in technical_cols and pd.api.types.is_numeric_dtype(df[col])
]

if not feature_cols:
    print("[!] Нет числовых признаков для построения radar plot.")
else:
    print(f"[INFO] Для анализа используются признаки: {feature_cols}")
```

Рисунок 2.41. Выбор информативных числовых признаков для профилирования

По каждому профилю (кластеру) строился radar plot (полярная диаграмма), отражающая средние значения по топ-10 наиболее различающих признаков. Такой подход позволяет увидеть силуэт и уникальные особенности каждого цифрового профиля.

Полярная диаграмма позволяет визуально сравнить "силуэт" каждого профиля по ключевым признакам: чем дальше точка от центра по конкретной оси, тем более выражен соответствующий признак у студентов данного профиля.

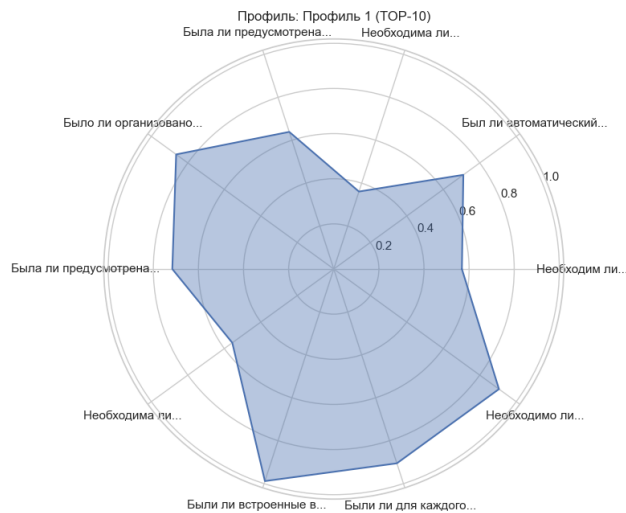


Рисунок 2.42. Полярная (radar) диаграмма цифрового профиля по факультету и признакам

Для автоматизации интерпретации были сгенерированы текстовые описания каждого профиля на основе средних значений признаков. Описания содержат общую численность студентов, топовые признаки профиля, а также указание на характерные сильные и слабые стороны.

Описания включают в себя численность студентов в профиле, список ключевых характеристик, а также выделяют сильные и слабые стороны. Такой подход позволяет быстро подготовить содержательные аналитические выводы для отчёта, презентации или принятия управленческих решений.

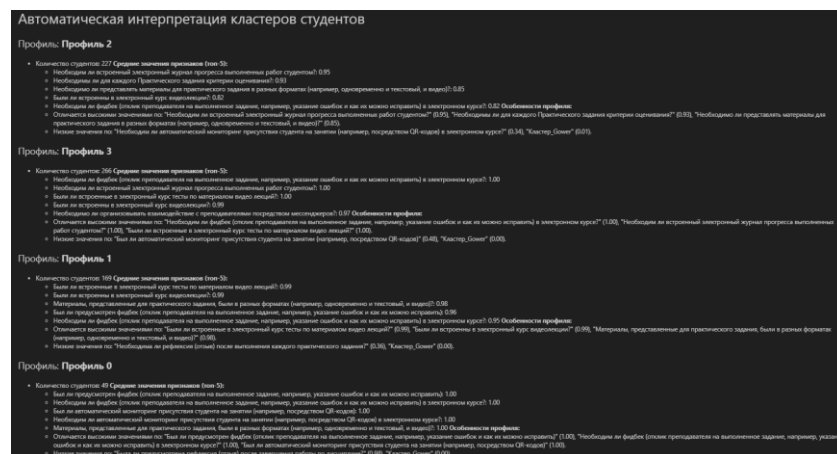


Рисунок 2.43. Автоматическая текстовая интерпретация профилей студентов

Этот этап завершает аналитику распределение и характеристика цифровых профилей студентов по факультетам становится полностью прозрачным, автоматическая генерация текстовых интерпретаций позволяет быстро получать содержательные выводы для отчёта, принятия решений или дальнейших образовательных инициатив.

Вывод по главе №2

В ходе реализации комплексной главы, посвящённой анализу анкетных данных студентов средствами методов машинного обучения и современных подходов образовательной аналитики, был осуществлён полный цикл обработки, структурирования и содержательной интерпретации исходной выборки.

На этапе разведочного анализа данных (EDA) была детально исследована структура и полнота анкет, выявлены все ключевые переменные, а также обеспечена корректная обработка бинарных и числовых признаков. Анализ пропусков, перевод всех релевантных ответов к единому формату и стандартизация структуры данных позволили сформировать качественный и целостный набор для последующего моделирования.

В процессе предобработки и инженерии признаков реализованы все необходимые процедуры по очистке, масштабированию и отбору информативных переменных, что обеспечило минимизацию шума и избыточности в данных. Применение специальных техник для унификации факультетальных данных и бинаризации признаков позволило получить репрезентативный и интерпретируемый массив для дальнейшего анализа.

Кластерный анализ цифровых профилей студентов был построен на использовании как классических методов (UMAP, Agglomerative Clustering), так и специализированных подходов для работы с бинарными признаками (расстояние Говера). Благодаря расчёту метрик качества разбиения (Silhouette Score, Davies–Bouldin Index), проведён объективный выбор наилучших методов кластеризации, а итоговые профили получили качественную и количественную интерпретацию. Визуализация результатов с помощью barplot, radar plot и

тепловых карт позволила не только выявить структуру данных, но и отразить уникальные характеристики каждого цифрового профиля.

Построенные модели автоматической классификации профилей на основе исходных признаков анкеты продемонстрировали высокую точность и устойчивость, что подтверждается результатами кросс-валидации и анализом важности переменных (feature importance). Эти результаты легли в основу для последующей автоматической генерации индивидуальных рекомендаций с применением языковой модели GigaChat. Интеграция модуля предсказания профиля и генерации текстовых советов обеспечила адресную цифровую поддержку для каждого студента, опираясь на его индивидуальные особенности цифрового поведения.

Анализ распределения цифровых профилей по факультетам позволил выявить значимые различия между подразделениями, что может быть использовано для принятия решений по развитию цифровых компетенций и совершенствованию образовательной среды в вузе. Автоматическая генерация интерпретаций профилей обеспечила прозрачность и содержательность итоговых выводов.

В целом, проведённый комплексный анализ продемонстрировал возможности сквозной автоматизации и интерпретации образовательных данных — от первичной очистки и профилирования до генерации персонализированных рекомендаций. Полученные результаты не только подтверждают высокое качество реализованных методов, но и открывают перспективы для дальнейшего развития персонализированной поддержки студентов в условиях цифровой трансформации высшего образования.

ЗАКЛЮЧЕНИЕ

Выпускная квалификационная работа посвящена применению методов машинного обучения для анализа эффективности цифровизации учебного процесса. В ходе исследования проведён аналитический обзор современных подходов к цифровизации высшего образования, рассмотрены стратегические инициативы в области образовательной аналитики и проанализирован инструментарий, используемый для решения задач, связанных с формированием цифровых компетенций и индивидуализацией образовательных траекторий на основе эмпирических данных. Сформулированы методологические основания для выявления цифровых профилей студентов, основанные на обработке анкетных данных.

В рамках эмпирической части работы выполнен сбор и разведочный анализ анкетных данных 711 студентов. Осуществлена структуризация массива данных, определены ключевые признаки цифрового поведения, выявлены закономерности, отражающие специфику цифровизации в образовательной среде. Проведена комплексная предобработка исходных данных, включающая стандартизацию, очистку, преобразование и отбор информативных признаков, что обеспечило формирование набора для построения аналитических моделей. В итоговую выборку включены 18 бинарных и 20 числовых признаков, прошедших фильтрацию по информативности.

Для выявления и интерпретации цифровых профилей студентов реализованы и сравнены различные методы кластеризации, среди которых UMAP, агломеративная кластеризация, Fuzzy C-Means и кластеризация на основе матрицы расстояний Говера. В результате применения данных подходов установлено, что наиболее устойчивое разбиение достигается при выделении четырёх профилей, причём оптимальные показатели качества получены методом Fuzzy C-Means по значениям Silhouette Score и Davies-Bouldin Index. Анализ структуры полученных кластеров осуществлялся на основании ключевых различающих бинарных признаков. Выполнена визуализация данных с помощью

barplot и radar plot, сформированы содержательные описания профилей, отражающие специфику цифрового поведения и вовлечённости студентов.

В ходе исследования обучен автоматический классификатор профиля студента на основе алгоритма RandomForest, продемонстрировавший высокие значения точности при определении принадлежности нового респондента к одному из выделенных цифровых профилей. Оценка качества классификации проведена на основе кросс-валидации по метрике F1_macro, что подтверждает устойчивость и воспроизводимость полученной модели. Реализован программный модуль автоматической генерации индивидуальных цифровых рекомендаций для студентов, основанный на интеграции обученного классификатора с языковой моделью GigaChat. Система формирует рекомендации в зависимости от профиля цифрового поведения и обеспечивает их сохранение в различных форматах для последующего использования.

Проведён анализ распределения цифровых профилей студентов по факультетам, что позволило установить выраженные различия между подразделениями, определить характерные особенности цифрового поведения в рамках отдельных факультетов и обозначить направления для адресного развития цифровых компетенций обучающихся. Сформированы автоматические текстовые интерпретации цифровых профилей с учётом численности, сильных и слабых сторон, специфики выраженности ключевых признаков.

Выполненное исследование продемонстрировало возможность построения комплексной системы анализа и интерпретации цифровых данных в образовательной среде, охватывающей все этапы: от сбора и структурирования данных до автоматической генерации персонализированных рекомендаций для обучающихся. Научная и практическая значимость работы заключается в формировании инструментария, способного обеспечивать объективную поддержку принятия решений, направленных на совершенствование процессов цифровизации, повышение адресности образовательной поддержки и оптимизацию стратегий развития цифровых компетенций в вузе.

Расширение представленной методологии возможно за счёт интеграции поведенческих и событийных данных, поступающих с образовательных платформ и LMS, а также за счёт изучения динамики цифровых профилей студентов в долгосрочном временном разрезе. Представляет интерес задача внедрения разработанной аналитической системы в деятельность образовательных организаций для апробации и формирования банка лучших практик цифровой поддержки обучающихся, а также создание средств для самостоятельной оценки и обратной связи по развитию цифровых компетенций на основе автоматизированных моделей профилирования.

ЛИТЕРАТУРА

1. Алпатов, А. В. Применение машинного обучения для анализа образовательных результатов студентов вузов // Информационные и математические технологии в науке и управлении. – 2023.
2. Брейман, Л. Случайные леса // Машинное обучение. – 2001.
3. Виттен, И. Х., Франк, Э. Интеллектуальный анализ данных: Практическое руководство по применению методов машинного обучения / И. Х. Виттен, Э. Франк. – 2-е изд. – Морган Кауфманн, 2005.
4. Жайн, А. К., Мерти, М. Н., Флинн, П. Дж. Обзор методов кластеризации данных // Обзоры ACM Computing Surveys. – 1999.
5. Коробков, Н. Что такое цифровизация образования и зачем она нужна [Электронный ресурс] // Блог платформы SkillSpace. – 2021. – Режим доступа: <https://skillspace.ru/blog/chto-takoe-cifrovizaciya-obrazovaniya-i-zachem-ona-nuzhna/> (дата обращения: 20.05.2025).
6. Малышев, В. В., Сливкин, С. С., Рукавишников, В. С., Базаркин, Е. В. Применение методов машинного обучения для построения рекомендательной системы отбора анкет абитуриентов // Научный вестник НГТУ. – 2017.
7. Научно-исследовательский институт развития образования РЭУ им. Г. В. Плеханова. Основные тренды цифровизации высшего образования: результаты мониторинга тенденций развития высшего образования в мире и в России. – Вып. 1. – М.: РЭУ им. Г. В. Плеханова, 2021. – 46 с.
8. Флах, П. Машинное обучение / П. Флах. – М.: ДМК Пресс, 2015. – 400 с.
9. Хайкин, С. Нейронные сети: полный курс / С. Хайкин. – 2-е изд. – М.: Вильямс.
10. Итоги eSTARS: высшее электронное [Электронный ресурс] // IQ — портал НИУ ВШЭ. – URL: <https://iq.hse.ru/news/423670621.html> (дата обращения: 20.05.2025).

11. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – 2nd ed. – Springer, 2009.
12. Mitchell, T. Machine Learning. – New York: McGraw-Hill, 1997.
13. UNESCO. Digital learning and transformation of education [Electronic resource]. – URL: <https://www.unesco.org/en/digital-education> (accessed: 20.05.2025).
14. UNESCO. UNESCO spotlights how digital learning can promote equity in low-resource contexts [Electronic resource]. – 31.03.2025. – URL: <https://www.unesco.org/en/articles/unesco-spotlights-how-digital-learning-can-promote-equity-low-resource-contexts> (accessed: 20.05.2025).
15. UNESCO. Digital Learning Week 2025 – AI and the future of education: Disruptions, dilemmas and directions (Concept note) [Electronic resource]. – URL: <https://www.unesco.org/en/weeks/digital-learning> (accessed: 20.05.2025).

ПРИЛОЖЕНИЕ А

Репозиторий готового продукта

Ссылка: https://github.com/rikigg/VKR-practice/blob/main/vkr_final.ipynb