Моделирование и анализ данных
2026, *16*(3), 7—29.
https://doi.org/10.17759/mda.2026160301
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2026, *16*(3), 7—29.
https://doi.org/10.17759/mda.2026160301
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

Научная статья | Original paper

УДК 004.85:519.23:37.016

Методы поиска закономерностей на основе кластерного анализа и больших данных образовательных учреждений среднего образования

Г.А. Чайкин ✉, **И.С. Блеканов**
Санкт-Петербургский государственный университет
Санкт-Петербург, Российская Федерация
✉ st061320@student.spbu.ru

Резюме

Контекст и актуальность. В условиях цифровизации образования накапливаются большие массивы данных об учебной деятельности, содержащие закономерности, недоступные традиционным статистическим методам, которые ограничиваются агрегированными показателями успеваемости. Глубокое обучение позволяет извлекать скрытые закономерности из структуры данных. **Цель.** Разработать и сравнить подходы к построению векторных представлений учащихся на основе государственных экзаменационных результатов для кластерного анализа и выявления скрытых паттернов. **Гипотеза.** Большие данные экзаменационных результатов содержат паттерны, не сводимые к агрегированным показателям успеваемости, а нейросетевые методы способны выделить их в виде векторных представлений, пригодных для кластеризации. **Методы и материалы.** Использованы данные экзаменов ГИА-9 и ГИА-11 за 2023–2025 годы (11 предметов, около 5 000 учащихся, около 500 000 взаимодействий). Сравнивались три подхода: базовый на агрегированных оценках (Score stats), вариационный автокодировщик (VAE) и маскированный автокодировщик (MAE), в том числе с контрастивной функцией потерь. Кластеризация выполнялась алгоритмом HDBSCAN со снижением размерности (PCA); качество оценивалось внутренними метриками (коэффициент силуэта, DBCV), долей выбросов и визуализацией t-SNE. **Результаты.** Score stats даёт устойчивое разбиение, отражающее тип

экзамена, набор предметов и успеваемость. VAE достигает высоких метрик при сильном сжатии представления, но его кластеры определяются длиной последовательности взаимодействий. MAE без контрастивной функции потерь не справляется с задачей; с её добавлением метрики выходят на уровень Score stats, а кластеры группируют учащихся по характерным сочетаниям совместно выбираемых экзаменов. **Выводы.** MAE с контрастивной функцией потерь извлекает из больших образовательных данных скрытые характеристики учащихся, не сводимые к агрегированным показателям. Разработанные подходы применимы к данным государственных экзаменов для категоризации учащихся и поддержки принятия решений.

Ключевые слова: анализ образовательных данных, образовательные данные, кластеризация, глубокое обучение, обучение представлений, автокодировщик

Для цитирования: Чайкин, Г.А., Блеканов, И.С. (2026). Методы поиска закономерностей на основе кластерного анализа и больших данных образовательных учреждений среднего образования. *Моделирование и анализ данных, 16*(3)*,* 7—29. https://doi.org/10.17759/mda.2026160301

Methods for discovering patterns based on cluster analysis and big data of secondary education institutions

G.A. Chaikin ✉, I.S. Blekanov
Saint Petersburg State University, Saint Petersburg, Russian Federation
✉ st061320@student.spbu.ru

Abstract

Context and relevance. The digitalization of education generates large volumes of learning-activity data containing patterns inaccessible to traditional statistical methods limited to aggregated performance indicators. Deep learning can extract latent patterns from data structure. **Objective.** To develop and compare student's vector representation approaches from examination results for cluster analysis and latent--pattern discovery. **Hypothesis.** Big data on examination results contain patterns not reducible to aggregated performance indicators, and neural-network methods can capture them as vector representations suitable for clustering. **Methods and materials.** Data from the GIA-9 and GIA-11 examinations for 2023–2025 were used (11 subjects, approximately 5,000 students and 500,000 interactions). Three approaches were compared: a baseline on aggregated scores (Score stats), a variational autoencoder (VAE), and a masked autoencoder (MAE), including a version trained with a contrastive loss. Clustering was performed with HDBSCAN after dimensionality reduction (PCA); quality was assessed with internal metrics (silhouette

coefficient, DBCV) and t-SNE visualization. **Results.** Score stats yields a stable partition reflecting examination type, subject set, and performance. VAE achieves high internal-metric values under aggressive dimensionality reduction, but its clusters are determined by the interaction-sequence length. MAE without a contrastive loss does not produce meaningful clusters; with it, the metrics reach the Score stats level, and the clusters group students by characteristic combinations of jointly chosen examinations. **Conclusions.** MAE with a contrastive loss extracts latent student characteristics, not reducible to aggregated indicators, from big educational data. The approaches are applicable to state examination data for categorizing students and supporting decision-making.

Keywords: educational data mining, educational data, clustering, deep learning, representation learning, autoencoder

For citation: Chaikin, G.A., Blekanov, I.S. (2026). Methods for discovering patterns based on cluster analysis and big data of secondary education institutions. *Modelling and Data Analysis, 16*(3)*,* 7—29. (In Russ.). https://doi.org/10.17759/mda.2026160301

Введение

В условиях цифровизации образования и накопления больших массивов данных об учебной деятельности актуальной становится задача автоматического анализа и категоризации учащихся на основе их образовательных результатов. Выявление скрытых закономерностей в поведении и академических достижениях обучающихся открывает возможности для построения адаптивных образовательных траекторий, обнаружения групп риска и повышения эффективности образовательного процесса в целом.

Анализ образовательных данных (Educational Data Mining, EDM) направлен на извлечение полезной информации из образовательных данных для повышения качества обучения, анализа поведения обучающихся и поддержки принятия решений (Barbeiro et al., 2024; Dutt, Ismail, Herawan, 2017). EDM опирается на методы машинного обучения, статистики и рекомендательных систем, а чаще всего решаются задачи прогнозирования успеваемости, анализа поведения обучающихся, рекомендации материалов и ассистирования преподавателя (Barbeiro et al., 2024).

Задачи анализа образовательных данных можно решать средствами глубокого обучения (Hernández-Blanco et al., 2019). В частности, трансформерные архитектуры на основе BERT используются для построения векторных представлений учащихся по данным их взаимодействия с платформой (Scarlatos, Brinton, Lan, 2022), маскированные автокодировщики — для восстановления пропущенных ответов и отслеживания динамики знаний (Freire, Curi, 2024), а мультимодальные нейросетевые модели — для автоматической диагностики когнитивных состояний обучающихся непосредственно в процессе учебной деятельности (Юрьева, 2025).

Одним из ключевых инструментов в области EDM является кластеризация учащихся — разбиение обучающихся на группы на основе их образовательных

характеристик, траекторий освоения материала и результатов выполнения заданий. Полученные кластеры позволяют выявлять типичные стратегии обучения, формировать персонализированные рекомендации и поддерживать принятие решений в интеллектуальных образовательных системах и аналитических платформах. В отличие от задач предсказания метки неизвестны: признаковое описание требуется разбить на группы так, чтобы внутри кластера объекты были схожи, а между кластерами — различались.

Кластеризация в EDM используется как инструмент категоризации учащихся, оценки их текущей и будущей успеваемости, аналитики и выявления скрытых, не наблюдаемых в данных явно характеристик (Dutt et al., 2015). Кластерные методы выявляют поведенческие и академические различия учащихся, что позволяет учреждениям предоставлять адаптируемую поддержку (Lu et al., 2025).

На уровне критериального оценивания и мониторинга качества образования кластеризация применяется для автоматизации экспертных процедур. Критериальное оценивание можно строить на кластеризации допустимых объектов по трём классам качества (низкое, среднее, высокое) с последующим распознаванием принадлежности нового объекта; результаты согласуются с традиционной балльной шкалой (Пак, Клунникова, 2022). Кластерный анализ применим к показателям мониторинга деятельности вузов (образовательная, научная, международная и финансово-экономическая составляющие) (Айназаров, Востроknутов, 2025).

Вместе с тем в качестве входных данных для кластеризации можно использовать не только агрегированные показатели успеваемости, но и принципиально иные структуры — последовательности взаимодействий учащегося с заданиями, временные ряды попыток, составы сдаваемых предметов. Такие данные, однако, требуют предварительной обработки: прежде чем применять классические алгоритмы кластеризации, их необходимо привести к единому векторному представлению фиксированной размерности. Для этой цели естественно использовать модели обучения представлений, которые кодируют исходные данные учащегося — вне зависимости от их формата и длины — в компактный вектор, пригодный для последующей кластеризации.

Построение информативных векторных представлений является ключевым этапом перед непосредственно самой кластеризацией ввиду того, что разные векторные представления могут содержать в себе разные интерпретируемые и неинтерпретируемые закономерности.

Маскированные автокодировщики (MAE) и трансформеры активно применяются для табличных и последовательных данных с пропусками. Модификация MAE, восстанавливающая пропущенные ответы на задания, предложена в (Freire, Curi, 2024), а в формате временных рядов MAE изучается в (Li et al., 2023; Shrivastava, Rameshan, Agnihotri, 2026). Эффективность показана для двухэтапного подхода: предобучение на восстановлении замаскированной последовательности с последующим использованием кодировщика как модели векторного представления на сторонней задаче. Аналогично BERT строит векторные представления учащихся по пользовательским

логам платформ MOOC (Scarlatos, Brinton, Lan, 2022), пригодные и для сторонних задач вроде прогнозирования результатов обучения.

Вариационные автокодировщики (VAE) применяются для родственных задач: восстановления пропусков в табличных данных наряду с GAN (Choi, Lam, Mendes, 2025), генерации синтетических образовательных данных, сохраняющих статистические свойства исходного набора (Kostopoulos et al., 2025), а также моделирования психометрических характеристик учащихся (Paaßen et al., 2022).

Отдельно стоит выделить методы глубокой кластеризации, где построение представлений и кластеризация решаются совместно (постановка и категоризация таких методов, например, представлена в (Lu et al., 2024)). Среди примеров можно выделить кластеризацию траекторий попыток решений заданий (Saïdi, Durand, Flouvat, 2025), двухэтапный подход на основе VAE и гауссовских смесей (Guo et al., 2025), а также автокодировщик для оценки уровня усвоения учебного материала (Zhao, Dong, 2024).

Наиболее современные подходы применяют большие языковые модели. Среди примеров можно выделить обогащение признаков и попарные ограничения для полуавтоматической кластеризации текстов (Viswanathan et al., 2024), сведение кластеризации к классификации через генерацию кандидатов-меток (Huang, He, 2025), а в анализе образовательных данных — построение метрики сходства вопросов и кластеризация заданий для выделения проверяемых навыков (Wei, Carvalho, Stamper, 2025).

Рассмотренные подходы к построению представлений, как правило, ориентированы на специфические типы данных: поведенческие логи платформ MOOC, последовательности действий в системах программирования, развёрнутые текстовые описания заданий. Это ограничивает их прямое применение к данным государственных экзаменов. В настоящей работе мы ставим задачу иначе: нас интересует единый подход к обработке данных, при котором модель предобучается на объединённых данных разных форматов и формирует векторные представления, отражающие паттерны учащихся в рамках общей выборки. Такой подход воспроизводит логику, зарекомендовавшую себя в обработке естественного языка (Devlin et al., 2019) и в задачах с временными рядами (Ansari et al., 2024; Woo et al., 2024): предобучение на разнородных источниках данных позволяет модели улавливать общие закономерности.

Для решения поставленной задачи предлагается подход к формированию векторных представлений учащихся, при котором единая модель предобучается на объединённых экзаменационных данных разных форматов (ГИА-9 и ГИА-11) и совместно обрабатывает разнородные признаки заданий. Исследуется применимость нейросетевых архитектур обучения представлений с последующей кластеризацией полученных векторов. Проводится сравнительный анализ представленных подходов, анализ влияния отдельных компонент методов на качество кластеризации, а также оценка интерпретируемости и устойчивости выделяемых групп учащихся.

Основные вклады данной работы заключаются в следующем:

- Разработка подхода к формированию векторных представлений учащихся на основе разнородных образовательных данных.

- Исследование применимости методов глубокого обучения представлений для задачи кластеризации учащихся на данных экзаменационных результатов разных форматов одновременно (ГИА-9 и ГИА-11).
- Сравнительный анализ предлагаемых методов.
- Анализ влияния компонент методов на качество кластеризации и устойчивость выделяемых групп учащихся.
- Проведение численных экспериментов и оценка интерпретируемости кластеров.

Материалы и методы

Постановка задачи

В данной работе рассматривается задача анализа образовательных данных с целью выявления скрытых характеристик на основе их учебной активности. Поскольку искомые группы заранее неизвестны, эта задача естественным образом сводится к задаче кластеризации, позволяющей разбить обучающихся на группы со схожими паттернами без предварительной разметки.

Пусть данные по каждому учащемуся заданы множеством $X = \{x_1, x_2, \dots, x_n\}$, где каждый объект .. соответствует студенту (или его учебной траектории) и описывается многомерным набором признаков, включающим поведенческие, временные и/или текстовые характеристики.

Необходимо построить отображение:

$$f: X \to Z,$$

где $Z = \{z_1, z_2, \dots, z_n\}$ — пространство векторных представлений, содержащие в себе полезную информацию.

На основе полученных представлений требуется решить задачу кластеризации:

$$g: Z \to C,$$

где $C = \{c_1, c_2, \dots, c_k\}$ — множество кластеров, отражающих группы обучающихся со схожими паттернами поведения.

Данные

В работе используются деперсонализированные данные экзаменационных результатов ГИА-9 и ГИА-11, представленные в виде разрозненных таблиц для каждого предмета. В исходных таблицах строки соответствуют взаимодействию учащегося с экзаменом, а столбцы содержат задания. Каждая таблица соответствует конкретному предмету и классу. Всего в данных содержится 11 предметов по каждому типу экзамена. Всего данных взаимодействия в районе 500 тысяч. После объединения и фильтрации получается совокупность последовательностей взаимодействий. Причём идентификаторы учащихся позволяют восстановить временные интервалы

сдач нескольких экзаменов (учащиеся сдают несколько предметов в рамках одного года). То есть табличные данные можно, по крайней мере частично, отсортировать в формате последовательности взаимодействий в зависимости от даты сдачи по конкретному предмету, а значит, в качестве входных данных в неявном виде будет задан и порядок сдачи экзаменов. В дальнейшем эти последовательности используются в качестве признаков для моделей.

Для подготовки единого набора данных разрозненные таблицы по предметам и классам упорядочиваются по идентификаторам учащихся и объединяются в единую структуру, где каждая строка соответствует учащемуся, а столбцы — последовательности заданий.

В табличных данных результатов экзаменов последовательная компонента представлена порядком заданий и ответом в рамках экзаменационной работы. В качестве признаков используются категориальные идентификаторы заданий и результаты по каждому заданию. Вместе с этим для экзаменов всегда предоставляются документы с описанием экзаменов, которые содержат дополнительные данные. В отдельных текстовых документах содержатся метаданные для каждой таблицы с результатами экзаменов в соответствии с каждым предметом. Эти данные предварительно обрабатываются и интегрируются в основной набор данных. Для каждого задания добавляются следующие характеристики:

- Уровень задания: Базовый (Б) или Повышенный (П).
- Максимальный балл за задание, используемый для нормализации и бинаризации оценок.
- Текстовое описание задания: краткое описание, что позволяет использовать не только идентификатор задания, но и его текстовое содержание в виде векторных представлений.

Таким образом, каждая запись в выборке содержит не только результаты выполнения заданий, но и дополнительные признаки, важные для построения информативных представлений навыков и заданий. При этом выборка организована в виде последовательностей взаимодействий учащихся с заданиями.

Методы построения векторных представлений

Для построения векторных представлений учащихся, подаваемых на вход алгоритмам кластеризации, рассматриваются три подхода. В отличие от первого, второй и третий подходы принимают на вход последовательности данных и агрегируют их в единый вектор фиксированной размерности.

Базовый метод

В качестве базового подхода рассматривается явное формирование признакового пространства учащегося на основе результатов экзаменов.

Каждый учащийся представляется в виде вектора фиксированной размерности:

$$x_i \in \mathbb{R}^d, i = \overline{1,n},$$

где каждая компонента соответствует нормированному показателю успешности по определённому аспекту экзаменационной деятельности. Структура признаков включает предметные баллы по 11 предметам, нормированные баллы по отдельным частям экзамена и распределение результатов по уровням сложности (Б, П, В).

Ключевой особенностью данного подхода является интерпретируемость признаков.

Вариационный автокодировщик

Второй подход основан на использовании модели Variational Autoencoder (VAE) с трансформерными кодировщиком и декодировщиком, которая применялась в ранее рассмотренных работах. VAE кодирует входные данные в распределение скрытого пространства, из которого декодировщик восстанавливает исходные данные.

В отличие от базового метода, входом модели является последовательность взаимодействий учащегося с экзаменами:

$$S_i = (e_1, e_2, \ldots, e_T).$$

После обучения кодировщик используется для получения вектора учащегося фиксированной размерности:

$$z_i \in \mathbb{R}^d.$$

В отличие от базового метода, VAE учитывает последовательную структуру данных и моделирует нелинейные зависимости между признаками.

Маскированный автокодировщик с [CLS]-агрегацией

Третий подход является основным предлагаемым методом и основан на идее модели Masked Autoencoder (MAE), адаптированной к данным экзаменационных результатов.

В отличие от VAE, задача формулируется как восстановление замаскированных ответов последовательности взаимодействий учащегося.

Пусть дана последовательность:

$$S_i = (e_1, e_2, \ldots, e_T),$$

где $e_t = (q_t, a_t)$ — взаимодействие учащегося с платформой на шаге t, q_t — вопрос, а a_t — ответ учащегося на этот вопрос.

В процессе обучения случайным образом выбирается подмножество ответов, которые заменяются специальным токеном [*MASK*]. Модель обучается восстанавливать исходные значения путем кодирования видимой части последовательности через кодировщик и обработки всей части последовательности (вывод кодировщика вместе с замаскированной частью исходной последовательности) через декодировщик.

Дополнительно для видимой части последовательности вводится специальный токен [*CLS*] для получения единого векторного представления учащегося:

$$[CLS] \rightarrow ([CLS], e_1, e_2, \ldots, e_T).$$

После обработки через кодировщик итоговое представление извлекается как:

$$h_i = h_{[CLS]},$$

где $h_{[CLS]}$ является вектором фиксированной размерности и уже не зависит от длины последовательности.

Дополнительно вводится контрастивная функция потерь:

$$L_{contrastive} = -log \frac{exp\left(sim\left(h_i, h_j\right)/\tau\right)}{\sum_{k=1}^{2N} 1_{[k \neq i]} exp\left(sim\left(h_i, h_k\right)/\tau\right)}.$$

Положительные пары формируются из двух независимых случайных масок одной последовательности без дополнительной аугментации.

Итоговая функция потерь будет выглядеть следующим образом.

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{con}$$

где λ — коэффициент в интервале (0,1).

Предлагаемый подход учитывает последовательную структуру данных. Структурой векторных представлений можно управлять через долю маскирования и добавление контрастивной функции потерь.

В рамках экспериментальной части предполагается исследовать:

- Влияние доли маскирования на качество кластеризации;
- Влияние контрастивной функции потерь на разделимость кластеров;

Метрики оценки кластеризации

В данной работе в качестве метода кластеризации будет использован алгоритм HDBSCAN. Данный алгоритм основан на плотностном подходе и позволяет выделять кластеры произвольной формы, а также автоматически определять объекты, не принадлежащие ни одному кластеру (выбросы). Результаты кластеризации зависят от настроенных гиперпараметров, не только модели векторного представления, но и алгоритма кластеризации. Поэтому для каждого подхода проводятся эксперименты для оценки влияния компонент на качество кластеризации.

Перед применением HDBSCAN ко всем описанным выше векторным представлениям для снижения размерности используется метод главных компонент (Principal Component Analysis, PCA). Число компонент определяется пороговым значением доли объяснённой дисперсии, которое задано равным 95%. Это также позволяет оценить, сколько компонент необходимо для сохранения заданной доли информации. Кроме того, PCA позволяет устранить мультиколлинеарность, что особенно полезно для первого метода, где в роли компонент выступают оценки по предметам.

Для качественного анализа структуры кластеров применяется метод t-SNE, который проецирует векторные представления учащихся в двумерное пространство для оценки компактности и разделимости кластеров, а также выявления пересечений и неоднородных областей. Совместное использование HDBSCAN и t-SNE позволяет не только количественно оценить качество кластеризации, но и визуально интерпретировать структуру характеристик учащихся в скрытом пространстве.

Так как данные для задачи кластеризации не содержат явных меток классов, для оценки качества используются внутренние метрики, основанные на компактности и разделимости кластеров. Среди используемых внутренних метрик — доля выбросов, коэффициент силуэта и DBCV. DBCV (Density-Based Clustering Validation) оценивает качество кластеризации с учётом плотности, что делает её пригодной для оценки кластеров произвольной формы и позволяет корректно учитывать выбросы, выделяемые алгоритмом кластеризации. Метрика принимает значения в диапазоне от –1 до 1, где более высокие значения соответствуют лучшему качеству кластеризации.

Постановка и реализация экспериментов

Для архитектур VAE и MAE для каждой комбинации доли маскирования и наличия/отсутствия контрастивной функции потерь выполнялась серия обучений с различными конфигурациями архитектуры модели (число блоков, размер скрытых слоёв и др.). Размер батча во всех экспериментах был зафиксирован равным 1024, а размерность выходного векторного пространства — равной 128 компонентам для всех архитектур, что обеспечивает сопоставимость VAE и MAE и единый вход для кластеризации; варьирование этой размерности заметно не влияло на результаты. Для архитектуры MAE, в которой применяется маскирование патчами (patch masking), размер патча (patch_size) был зафиксирован равным 10: длины последовательностей взаимодействий у учащихся существенно различаются — встречаются как короткие, так и длинные, и это значение подходит для всего диапазона. Из полученных моделей отбиралась та, которая достигает наилучшего значения функции потерь на тестовой выборке. Данный критерий позволяет зафиксировать конфигурацию модели и сосредоточиться на анализе влияния гиперпараметров алгоритма кластеризации HDBSCAN.

Обучение моделей было реализовано в единой программной среде, разработанной на основе библиотек PyTorch и Hydra. Обучающая выборка формируется из 80% учащихся, а оставшиеся 20% используются в качестве тестовой выборки. Разбиение выполняется случайным образом. Для численных экспериментов по задаче кластеризации используются реализованные методы из библиотеки scikit-learn. Все эксперименты выполнялись на виртуальной машине с образом Ubuntu 22.04, оборудованной видеокартой NVIDIA RTX 4090 24 ГБ.

Результаты

Численные эксперименты для Score stats

Как было указано ранее, Score stats представляет собой набор признаков, собранный на основе оценок в рамках конкретных частей экзамена по каждому предмету.

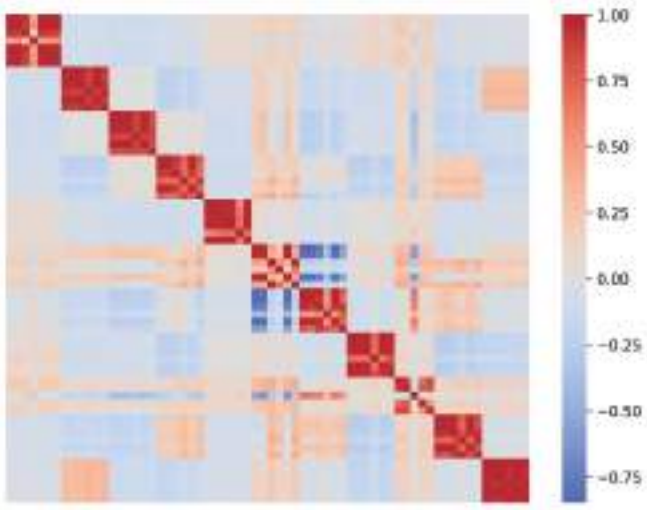

Рис. 1. Матрица корреляции признаков Score stats

Fig. 1. Score stats features correlation matrix

На рис. 1 представлена корреляция признаков. Как можно заметить, по диагонали наблюдается корреляция между признаками. На данных участках признаки соответствуют общему баллу, баллам по уровням сложности (Б, П, В) и частям экзамена по конкретному предмету.

Это уже говорит как минимум о линейной зависимости между показателями в рамках одного экзамена. Однако есть исключения — это признаки для предметов «Русский язык» и «Математика профильная» и «Математика базовая» (для результатов ГИА-9 по математике было решено не использовать отдельные признаки, а указать общие баллы учащихся как признаки для «Математика базовая»).

Решить проблему мультиколлинеарности можно путем исключения части признаков. Также можно решить проблему путем применения метода главных компонент, решающего не только задачу борьбы с корреляцией, но и задачу уменьшения изначальной размерности векторного пространства. Чтобы достичь компромисса между сокращением размерности и сохранением информативности данных, количество компонент не задавалось заранее, а определялось автоматически на основе доли объяснённой дисперсии (95%). В итоге удалось снизить размерность с 66 компонент до 15.

В таблице 1 представлены основные результаты внутренних метрик с учетом гиперпараметров алгоритма HDBSCAN (набор гиперпараметров, при которых метрики достигают лучших значений). Подробная таблица представлена отдельно, см. Приложение, табл. 1.

Таблица 1 / Table 1

Метрики качества кластеризации для Score stats

Clustering quality metrics for Score stats

Гиперпараметры / Hyperparameters (min_cluster_size / min_samples / method)	**Кол-во кластеров / Number of clusters**	**Доля выбросов / Share of emissions**	**Коэффициент силуэта / Silhouette coefficient**	**DBCV**
25 / 5 / eom	**34**	**0,166**	**0,300**	0,209
50 / 25 / eom	17	0,300	0,255	**0,361**

Анализ результатов показывает выраженную зависимость качества разбиения от гиперпараметров HDBSCAN. Параметр min_cluster_size оказывает наибольшее влияние на структуру разбиения. При малых значениях (25) алгоритм выделяет от 27 до 38 кластеров с относительно низкой долей выбросов (0,17–0,31), тогда как при больших значениях (200) число кластеров сокращается до 8–10, а доля выбросов возрастает до 0,42. Увеличение другого параметра min_samples и переход от метода eom к leaf повышают долю выбросов. Коэффициент силуэта достигает максимума (0,3) при наиболее дробном разбиении (25 / 5 / eom), однако такая конфигурация даёт чрезмерно большое число кластеров (34), что может затруднить содержательную интерпретацию. Метрика DBCV достигает максимума (0,361) при конфигурации 50 / 25 / eom с 17 кластерами и долей выбросов 0,3. В итоге в качестве рабочей конфигурации предпочтение отдано умеренным значениям min_cluster_size, обеспечивающим баланс между числом кластеров, долей выбросов и устойчивостью разбиения.

Анализ построенной визуализации t-SNE (см. рис. 2) показывает, что объекты формируют множество чётко разделённых плотных облаков точек, которым соответствуют кластеры, выделенные алгоритмом HDBSCAN; рассеянные между облаками точки отвечают объектам, отнесённым к выбросам. Изучение содержательного состава кластеров — оценок учащихся и предметов, по которым сдавались экзамены, — позволяет заключить, что кластеры группируют объекты по двум признакам: конкретному набору сдаваемых предметов и уровню успеваемости в целом или успеваемости отдельно по предметам. Первый признак проявляется за счёт доли экзаменов по конкретным сданным предметам и среднего балла, при этом стандартное отклонение нормированной итоговой оценки внутри кластера варьируется в пределах от 0,05 до 0,15, но также наблюдаются отдельные предметы внутри кластеров, в которых большое стандартное отклонение (больше 0,15), а среднее ниже

остальных предметов. Также наблюдаются целые кластера с такой особенностью по всем предметам. Кроме того, установлено, что большая часть кластеров (порядка 81%) относится к одному типу экзамена, то есть такие кластеры содержат результаты либо ГИА-9, либо ГИА-11. При этом признак типа экзамена явным образом в данных не задан, а оценки, выступающие компонентами векторного представления, нормированы делением на максимальные баллы (5 и 100 для ГИА-9 и ГИА-11 соответственно), что может свидетельствовать о неявном присутствии признака типа экзамена в данных. Дополнительно были проанализированы объекты, отнесённые к выбросам: каких-либо характерных особенностей у них не выявлено, эти точки представляют собой пограничные значения, расположенные на границах кластеров.

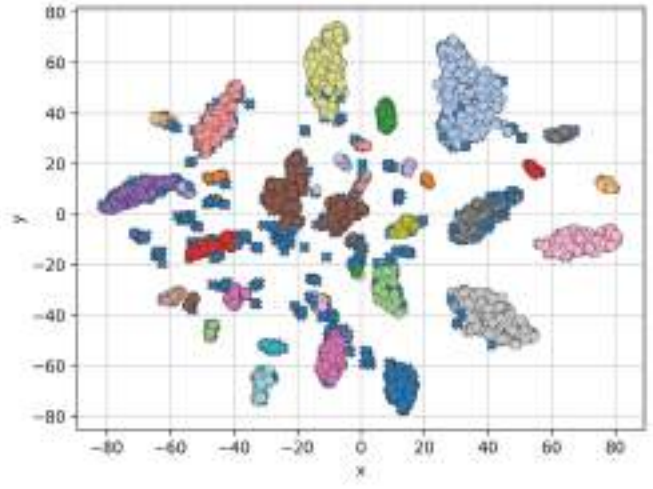

Рис. 2. Визуализация кластеров Score stats (25 / 5 / eom)
Fig. 2. Visualization of Score stats clusters (25 / 5 / eom)

Численные эксперименты для VAE

Выстроим похожим образом численные эксперименты для вариационного автокодировщика. В таблице 2 представлены основные результаты внутренних метрик с учетом гиперпараметров алгоритма HDBSCAN (приведены конфигурации, при которых достигается лучшее значение по каждой из метрик — доле выбросов, коэффициенту силуэта и DBCV). Подробные результаты приведены в соответствующем разделе Приложения (см. Приложение, табл. 2).

Таблица 2 / Table 2

Метрики качества кластеризации для VAE

Clustering quality metrics for VAE

Гиперпараметры / Hyperparameters (min_cluster_size / min_samples / method)	**Кол-во кластеров / Number of clusters**	**Доля выбросов / Share of emissions**	**Коэффициент силуэта / Silhouette coefficient**	**DBCV**
50 / 5 / eom	40	**0,048**	**0,624**	0,382
25 / 15 / eom	**53**	0,071	0,608	**0,576**

Данные результаты получены после уменьшения размерности векторного пространства. В итоге пространство уменьшилось до 4 компонент.

Для VAE сохраняется та же зависимость от гиперпараметров HDBSCAN, что и для Score stats: малые значения min_cluster_size (25) дают 53–64 кластера с долей выбросов 0,05–0,25, большие (200) — 12 кластеров, а увеличение min_samples и переход от eom к leaf повышают долю выбросов. Коэффициент силуэта достигает максимума (0,624) при 50 / 5 / eom, а DBCV — (0,576) при 25 / 15 / eom; как и ранее, их оптимумы не совпадают. По сравнению со Score stats VAE даёт более дробное разбиение и более высокие значения внутренних метрик, что свидетельствует о более выраженной кластерной структуре. В качестве рабочей конфигурации выбраны умеренные значения min_cluster_size.

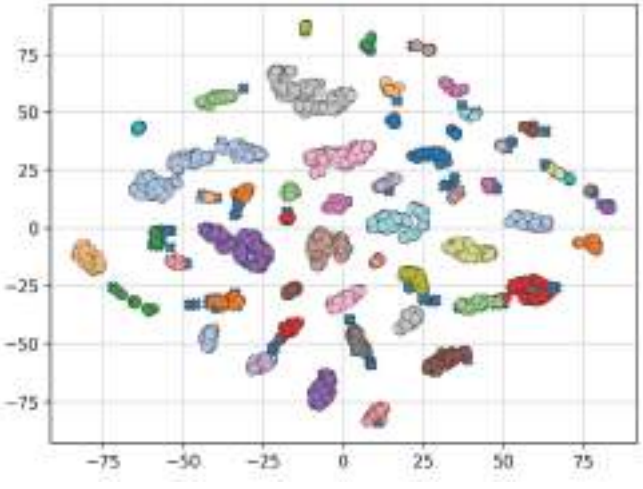

Рис. 3. Визуализация кластеров VAE (25 / 15 / eom)

Fig. 3. Visualization of VAE clusters (25 / 15 / eom)

Совместный анализ значений внутренних метрик (см. Приложение, таб. 2) и визуализации кластеров (см. рис. 3) показывает, что облака точек образуют не только крупные, но и более мелкие плотные группы, что согласуется с дробным разбиением — в рабочих конфигурациях выделяется порядка 53 кластеров при низкой доле выбросов (0,05–0,07). Около половины кластеров (≈51%) содержат результаты только одного типа экзамена, тогда как внутри кластеров — независимо от типа экзамена — устойчиво прослеживается сопоставление по длине исходной последовательности. В ряде кластеров чёткого паттерна по предметной области не наблюдается: для результатов ГИА-9 характерен один набор предметов, а для ГИА-11 — другой, причём средние баллы могут различаться (например, по предметам ГИА-9 относительные результаты оказываются выше, чем по ГИА-11). Доля выбросов заметно ниже, чем в базовом методе, однако стандартное отклонение балла внутри кластеров превышает соответствующие значения для Score stats и сильно варьируется (от 0,05 до 0,30). Таким образом, о кластерном разбиении именно по общей успеваемости и предметной области судить затруднительно: устойчивые закономерности проявляются по длине последовательности, тогда как при разделении произвольного кластера по типу экзамена чёткий паттерн по успеваемости и предметам выражен слабее.

Численные эксперименты для MAE

Эксперименты для MAE отличаются тем, что рассматриваются не только гиперпараметры кластеризации, но и доля маскирования и наличие контрастивной функции потерь при предобучении.

В таблице 3 представлены основные результаты экспериментов — конфигурации, при которых достигается лучшее значение по каждой из метрик (доле выбросов, коэффициенту силуэта и DBCV). Подробные результаты приведены в соответствующем разделе Приложения (см. Приложение, табл. 3).

Таблица 3 / Table 3

Метрики качества кластеризации для MAE

Clustering quality metrics for MAE

коэффициент маскирования / mask ratio	Контрастивная функция потерь / Contrastive loss	Гиперпараметры / Hyperparameters (min_cluster_size / min_samples / method)	Кол-во кластеров / Number of clusters	Доля выбросов / Share of emissions	Коэффициент силуэта / Silhouette coefficient	DBCV
0,3	False	50 / 15 / eom	2	0,038	0,511	-0,261
0,3	False	50 / 25 / eom	2	0,039	0,512	-0,071
0,3	False	25 / 15 / leaf	**32**	0,664	-0,414	0,128
0,5	False	25 / 15 / eom	2	0,049	0,484	-0,292
0,5	False	25 / 15 / leaf	37	0,664	-0,424	0,137
0,7	False	25 / 15 / eom	2	0,027	**0,550**	-0,166
0,7	False	150 / 5 / leaf	6	0,664	-0,277	0,104
0,3	True	150 / 15 / eom	2	**0,015**	0,213	-0,134
0,3	True	100 / 50 / eom	2	0,030	0,198	0,242

коэффициент маскирования / mask ratio	Контрастивная функция потерь / Contrastive loss	Гиперпараметры / Hyperparameters (min_cluster_size / min_samples / method)	Кол-во кластеров / Number of clusters	Доля выбросов / Share of emissions	Коэффициент силуэта / Silhouette coefficient	DBCV
0,5	True	50 / 5 / eom	22	0,157	0,188	0,006
0,5	True	100 / 15 / eom	12	0,212	0,216	0,151
0,5	True	100 / 50 / eom	9	0,413	0,133	**0,243**
0,7	True	100 / 50 / eom	2	0,022	0,323	0,019
0,7	True	100 / 25 / eom	10	0,312	0,082	0,211

После снижения размерности пространство уменьшилось до 2 компонент для MAE без контрастивной функции потерь и до 7 компонент — с ней.

Результаты для MAE без контрастивной функции потерь оказываются заметно хуже, чем у остальных подходов. Метрика DBCV принимает отрицательные значения практически во всех конфигурациях и при всех долях маскирования: даже когда долю выбросов удаётся удержать на низком уровне (например, при mask ratio = 0,3 и конфигурации 50 / 15 / eom доля выбросов составляет 0,038, а коэффициент силуэта — 0,511), DBCV остаётся отрицательной (–0,261). Случаи, когда DBCV выходит в небольшую положительную область (метод leaf, до 0,137 при mask ratio = 0,5), достигаются лишь ценой резкого роста доли выбросов (до 0,66 и выше). Таким образом, в отличие от других методов, здесь не удаётся найти компромисс между кластерной разделимостью и долей выбросов: при низкой доле выбросов разбиение плотностно неустойчиво, а осмысленное разбиение сопровождается отнесением большей части точек к шуму. В подробных таблицах (см. Приложени, табл. 3) при ряде конфигураций наблюдаются пропуски: число кластеров и коэффициент силуэта не определены, доля выбросов равна 1, а DBCV — 0. Это означает, что HDBSCAN не справился с разбиением и отнёс все точки к выбросам.

Влияние доли маскирования на качество кластеризации проявляется прежде всего в устойчивости разбиения. Для MAE без контрастивной функции потерь наибольшая доля маскирования (mask ratio = 0,7) оказывается наименее устойчивой: именно при ней чаще всего наступает вырожденный режим, когда все точки помечаются выбросами. Доли маскирования 0,3 и 0,5 ведут себя схожим образом и дают сопоставимые, но по-прежнему отрицательные по DBCV результаты. Выраженной монотонной зависимости качества от доли маскирования при этом не прослеживается — её влияние оказывается слабее, чем влияние гиперпараметров HDBSCAN.

Добавление контрастивной функции потерь при предобучении качественно меняет картину. Вырожденных режимов и пропусков в подробных таблицах (см. Приложение, табл. 4) не возникает, а DBCV становится положительной в большинстве конфигураций при умеренных долях выбросов (0,15–0,5). При этом MAE с контрастивной функцией потерь демонстрирует те же закономерности, что и Score stats и VAE, — универсального набора гиперпараметров нет: коэффициент силуэта и DBCV достигают максимума при разных настройках. Влияние доли маскирования

остаётся слабым и немонотонным, а умеренные значения (0,3–0,5) обеспечивают наилучший баланс метрик.

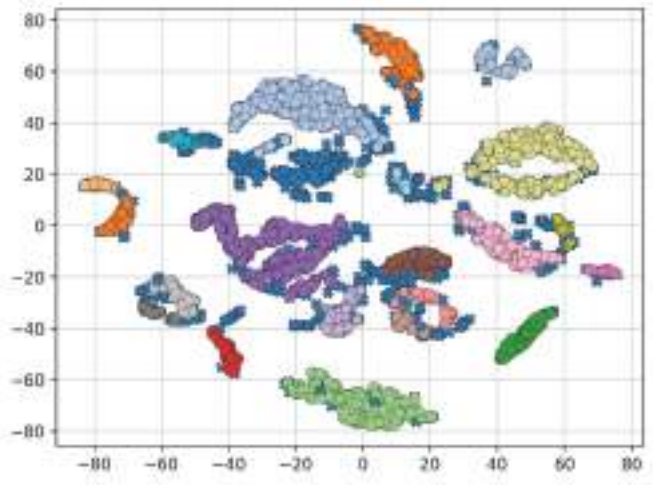

Рис. 4. Визуализация кластеров MAE (mask ratio=0,5, contrastive loss, 50 / 5 / eom)
Fig. 4. Visualization of MAE clusters (mask ratio=0,5, contrastive loss, 50 / 5 / eom)

Совместный анализ внутренних метрик (табл. 3, Приложение, табл. 4) и визуализации кластеров (см. рис. 4) показывает, что для MAE с контрастивной функцией потерь прослеживается более или менее чёткое разделение по предметной области и общей успеваемости. При этом длина исходной последовательности внутри кластера не обязательно оказывается однородной: наряду с кластерами, где длины близки, встречаются и такие, где отклонение по числу заданий достаточно велико (10 заданий и более). Как и для Score stats, средние баллы по предметам в рамках кластера варьируются слабо (до 0,15 по нормированной шкале), и отчётливо проявляется разделение по типу экзамена. Это объясняется тем, что в последовательностях используется категориальный признак, отвечающий за номер задания: он формируется путём группировки последовательностей по предметам и типу экзамена, благодаря чему предметная информация задаётся не отдельной компонентой вектора, а самой последовательностью заданий. В отличие от VAE, MAE удаётся эффективно воспользоваться входными признаками заданий.

Как и для Score stats, наблюдаются несколько кластеров с нестабильными результатами по конкретным предметам, либо с нестабильными результатами по всем предметам.

В то же время, в отличие от Score stats, где разбиение чётко определяется отдельными предметами, для MAE наблюдаются как кластеры с чётким предметным разделением, так и кластеры, где оно выражено слабее. Это может означать, что

в MAE кластер выражает не просто перечень присутствующих в нём предметов, а устойчивые группы совместно сдаваемых предметов: внутри кластера доминируют несколько характерных предметных комбинаций, на которые приходится основная масса наблюдений. Так, в одном из рассматриваемых кластеров преобладают последовательности (Математика базовая, Информатика, Русский язык, Английский язык) и (Математика базовая, Физика, Русский язык, Английский язык), охватывающие большинство учащихся, тогда как прочие сочетания встречаются единично. Тем самым кластеризация группирует учащихся по последовательным наборам совместно выбираемых экзаменов, а не по присутствию одного набора предметов.

Обсуждение результатов

Ответы на сформулированные исследовательские вопросы можно резюмировать следующим образом:

- Среди рассмотренных подходов к построению векторных представлений наиболее выраженную кластерную структуру даёт вариационный автокодировщик (VAE): он формирует более дробное разбиение при низкой доле выбросов и достигает наибольших значений внутренних метрик (коэффициента силуэта и DBCV), однако эти высокие метрики достигаются при низком числе компонент PCA (всего 4 компоненты). Базовый метод Score stats обеспечивает устойчивое, но менее выраженное разбиение с более скромными значениями метрик. Маскированный автокодировщик (MAE) без контрастивной функции потерь уступает остальным подходам — метрика DBCV принимает отрицательные значения практически во всех конфигурациях, и компромисса между кластерной разделимостью и долей выбросов достичь не удаётся; добавление же контрастивной функции потерь выводит MAE на уровень, сопоставимый со Score stats.
- Отдельные компоненты и признаки предложенной модели влияют на качество кластеризации неравномерно. Для MAE решающим компонентом оказывается контрастивная функция потерь: без неё разбиение неустойчиво и нередко вырождается (все точки относятся к выбросам), а с ней вырожденность исчезает и DBCV становится положительной в большинстве конфигураций. Доля маскирования влияет слабо и немонотонно — её влияние оказывается меньше влияния гиперпараметров HDBSCAN, при этом умеренные значения обеспечивают наилучший баланс метрик. Категориальный признак номера задания, формируемый группировкой последовательностей по предметам и типу экзамена, позволяет MAE кодировать предметную информацию самой последовательностью заданий, а не отдельной компонентой вектора; в отличие от VAE, MAE удаётся эффективно воспользоваться входными признаками заданий.
- Полученные векторные представления содержат скрытые характеристики, явным образом не заданные во входных данных. Для Score stats кластеры неявно кодируют тип экзамена: большинство кластеров относятся к одному типу (ГИА-9

либо ГИА-11), хотя этот признак в данных не задан, а оценки нормированы. Представления VAE кодируют преимущественно структурные характеристики последовательности — её длину, тогда как разделение по общей успеваемости и предметной области выражено слабее. Представления MAE с контрастивной функцией потерь отражают устойчивые группы совместно сдаваемых предметов и тип экзамена, а также успеваемость по предметам, группируя учащихся по характерным наборам совместно выбираемых экзаменов, а не по присутствию одного набора предметов.

Заключение

В работе исследовано влияние различных способов построения векторных представлений на качество кластеризации образовательных данных государственной итоговой аттестации (ГИА-9 и ГИА-11). Из разрозненных таблиц с результатами экзаменов и текстовых описаний заданий сформирован единый набор данных, пригодный для обучения моделей и последующей кластеризации обучающихся по выявленным паттернам.

Предложено и реализовано три подхода к построению векторных представлений: базовый метод на основе агрегированных статистик оценок (Score stats), вариационный автокодировщик (VAE) для последовательностей взаимодействий и маскированный автокодировщик (MAE) с [CLS]-агрегацией. Проведены численные эксперименты, в том числе по влиянию отдельных компонент на качество кластеризации. Наиболее выраженную кластерную структуру даёт VAE, базовый метод — устойчивое, но менее выраженное разбиение, а для MAE решающей оказывается контрастивная функция потерь. Кроме того, полученные представления содержат закономерности, явно не заданные во входных данных, — тип экзамена, структуру последовательности и устойчивые группы совместно сдаваемых предметов.

Перспективным направлением дальнейших исследований является добавление большего количества разнообразных признаков, в том числе полнотекстовых формулировок заданий, что позволит расширить исследование возможности поиска закономерностей в образовательных данных.

Ограничения. Следует отметить ограничения проведённого исследования. Во-первых, данные по каждому предмету представлены в табличном формате, поэтому порядок выполнения отдельных заданий внутри одного предмета неизвестен и не учитывается при построении векторных представлений. Во-вторых, влияние состава входных признаков на качество представлений для моделей VAE и MAE не исследовалось. В-третьих, оценка кластеризации опирается на внутренние метрики (коэффициент силуэта, DBCV, доля выбросов) и визуальный анализ; внешняя валидация по размеченным группам не проводилась ввиду отсутствия эталонной разметки, а сравнение с более широким кругом методов построения представлений остаётся направлением дальнейших исследований.

Limitations. It should be noted that the conducted study has a number of limitations. First, the data for each subject is presented in tabular format, so the order in which individual tasks within a single subject were solved is unknown and is not taken into account when constructing the vector representations. Second, the effect of the composition of the input features on the quality of the representations for the VAE and MAE models was not studied. Third, the evaluation of clustering relies on internal metrics (the silhouette coefficient, DBCV, the proportion of outliers) and visual analysis; external validation against labeled groups was not carried out owing to the absence of reference labeling, and a comparison with a broader range of representation-construction methods remains a direction for future research.

Список источников / References

1. Айназаров, Р.Р., Востроknутов, И.Е. (2025). Математическое моделирование кластеризации по результатам мониторинга деятельности образовательных организаций высшего образования. Computational Nanotechnology, 12(5), 118–128. https://doi.org/10.33693/2313-223X-2025-12-5-118-128
   Ainazarov, R.R., Vostroknutov, I.E. (2025). Mathematical modeling of clustering based on results of monitoring the activities of higher education institutions. Computational Nanotechnology, 12(5), 118–128. (In Russ.). https://doi.org/10.33693/2313-223X-2025-12-5-118-128
2. Пак, Н.И., Клунникова, М.М. (2022). Кластерный подход к критериальному оцениванию качества образовательного результата обучаемого. Вестник РУДН. Серия: Информатизация образования, 19(3), 196–207. https://doi.org/10.22363/2312-8631-2022-19-3-196-207
   Pak, N.I., Klunnikova, M.M. (2022). Cluster approach to criteria-based assessment of the quality of educational outcomes of learners. RUDN Journal of Informatization in Education, 19(3), 196–207. (In Russ.). https://doi.org/10.22363/2312-8631-2022-19-3-196-207
3. Юрьева, Н.Е. (2025). Искусственный интеллект в психодиагностике: когнитивные состояния в цифровой образовательной среде. Моделирование и анализ данных, 15(3), 47–55. https://doi.org/10.17759/mda.2025150303
   Yuryeva, N.E. (2025). Artificial intelligence in psychodiagnostics: cognitive states in a digital educational environment. Modelling and Data Analysis, 15(3), 47–55. (In Russ.). https://doi.org/10.17759/mda.2025150303
4. Ansari, A.F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S.S., Pineda Arango, S., Kapoor, S., Zschiegner, J., Maddix, D.C., Wang, H., Mahoney, M.W., Torkkola, K., Wilson, A.G., Bohlke-Schneider, M., Flunkert, V. (2024). Chronos: Learning the language of time series. Transactions on Machine Learning Research. https://doi.org/10.48550/arXiv.2403.07815
5. Barbeiro, L., Gomes, A., Correia, F.B., Bernardino, J. (2024). A review of educational data mining trends. Procedia Computer Science, 237, 88–95. https://doi.org/10.1016/j.procs.2024.05.083
6. Choi, W.C., Lam, C.T., Mendes, A.J. (2025). Comparison of data imputation performance in deep generative models for educational tabular missing data. In: Proceedings of the 18th International Conference on Educational Data Mining (pp. 133–142). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.15870169

7. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
8. Dutt, A., Aghabozrgi, S., Ismail, M.A.B., Mahroeian, H. (2015). Clustering algorithms applied in educational data mining. International Journal of Information and Electronics Engineering, 5(2), 112–116. https://doi.org/10.7763/IJIEE.2015.V5.513
9. Dutt, A., Ismail, M.A., Herawan, T. (2017). A systematic review on educational data mining. IEEE Access, 5, 15991–16005. https://doi.org/10.1109/ACCESS.2017.2654247
10. Freire, G.M., Curi, M. (2024). Masked autoencoder transformer for missing data imputation of PISA. In Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky (pp. 364–372). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64315-6_33
11. Guo, J., Fan, W., Amayri, M., Bouguila, N. (2025). Deep clustering analysis via variational autoencoder with Gamma mixture latent embeddings. Neural Networks, 183, 106979. https://doi.org/10.1016/j.neunet.2024.106979
12. Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. Complexity, 2019, 1306039. https://doi.org/10.1155/2019/1306039
13. Huang, C., He, G. (2025). Text clustering as classification with LLMs. In: Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (pp. 374–384). ACM. https://doi.org/10.1145/3767695.3769519
14. Kostopoulos, G., Fazakis, N., Kotsiantis, S., Dimakopoulos, Y. (2025). Enhancing semi-supervised learning in educational data mining through synthetic data generation using tabular variational autoencoder. Algorithms, 18(10), 663. https://doi.org/10.3390/a18100663
15. Li, Z., Rao, Z., Pan, L., Wang, P., Xu, Z. (2023). Ti-MAE: Self-supervised masked time series autoencoders. arXiv. https://doi.org/10.48550/arXiv.2301.08871
16. Lin, Y., Chen, H., Xia, W., Lin, F., Wang, Z., Liu, Y. (2025). A comprehensive survey on deep learning techniques in educational data mining. Data Science and Engineering. https://doi.org/10.1007/s41019-025-00303-z
17. Lu, Y., Li, H., Li, Y., Lin, Y., Peng, X. (2024). A survey on deep clustering: From the prior perspective. Vicinagearth, 1, 4. https://doi.org/10.1007/s44336-024-00001-w
18. Lu, Y., Yeom, S., Maktoubian, J., Rahman, M.M., Kim, S.-H. (2025). Improve student risk prediction with clustering techniques: A systematic review in education data mining. Education Sciences, 15(12), 1695. https://doi.org/10.3390/educsci15121695
19. Paaßen, B., Dywel, M., Fleckenstein, M., Pinkwart, N. (2022). Sparse factor autoencoders for item response theory. In: Proceedings of the 15th International Conference on Educational Data Mining (pp. 17–26). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.6853067
20. Saïdi, I., Durand, N., Flouvat, F. (2025). Analysis of students' attempts trajectories in learning programming. In: Proceedings of the 18th International Conference on Educational Data

Mining (pp. 66–76). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.15870155

21. Scarlatos, A., Brinton, C., Lan, A. (2022). Process-BERT: A framework for representation learning on educational process data. In: Proceedings of the 15th International Conference on Educational Data Mining (pp. 715–719). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.6853006
22. Shrivastava, A., Rameshan, R., Agnihotri, S. (2026). Robust representation learning in masked autoencoders. arXiv. https://doi.org/10.48550/arXiv.2602.03531
23. Viswanathan, V., Gashteovski, K., Lawrence, C., Wu, T., Neubig, G. (2024). Large language models enable few-shot clustering. Transactions of the Association for Computational Linguistics, 12, 321–333. https://doi.org/10.1162/tacl_a_00648
24. Wei, Y., Carvalho, P., Stamper, J. (2025). KCluster: An LLM-based clustering approach to knowledge component discovery. In: Proceedings of the 18th International Conference on Educational Data Mining (pp. 228–240). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.15870196
25. Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., Sahoo, D. (2024). Unified training of universal time series forecasting transformers. In: Proceedings of the 41st International Conference on Machine Learning (ICML 2024). https://doi.org/10.48550/arXiv.2402.02592
26. Zhao, M., Dong, X. (2024). Evaluation of deep clustering for assessing undergraduate understanding in ideological and political education: Data-driven analytics. In: Genetic and evolutionary computing (pp. 103–111). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-0068-4_10

Информация об авторах

Георгий Александрович Чайкин, аспирант, Санкт-Петербургский государственный университет (ФГБОУ ВО СПбГУ), Санкт-Петербург, Российская Федерация, ORCID: https://orcid.org/0009-0007-9956-9457, SPIN-код: 1443-9915, e-mail: st061320@student.spbu.ru

Иван Станиславович Блеканов, кандидат технических наук, доцент, заведующий кафедрой технологии программирования, Санкт-Петербургский государственный университет (ФГБОУ ВО СПбГУ), Санкт-Петербург, Российская Федерация, ORCID: https://orcid.org/0000-0002-7305-1429, SPIN-код: 7473-1900, e-mail: i.blekanov@spbu.ru

Information about the authors

Georgii A. Chaikin, Postgraduate Student, Saint Petersburg State University, St. Petersburg, Russian Federation, ORCID: https://orcid.org/0009-0007-9956-9457, SPIN-code: 1443-9915, e-mail: st061320@student.spbu.ru

Ivan S. Blekanov, Candidate of Science (Engineering), Associate Professor, Head of the Department of Programming Technology, Saint Petersburg State University, St. Petersburg, Russian Federation, ORCID: https://orcid.org/0000–0002–7305–1429, SPIN-code: 7473-1900, e-mail: i.blekanov@spbu.ru

Вклад авторов

Чайкин Г.А. — сбор, обработка и анализ данных; проведение исследования; визуализация результатов исследования; написание и оформление рукописи.

Блеканов И.С. — планирование исследования; сбор данных; контроль за проведением исследования.

Все авторы приняли участие в обсуждении результатов и согласовали окончательный текст рукописи.

Contribution of the authors

Georgii A. Chaikin — collection, processing and analysis of data; conduct of the study; visualization of the research results; writing and formatting the manuscript.

Ivan S. Blekanov — study design; data collection; supervision of the research process.
All authors participated in the discussion of the results and approved the final text of the manuscript.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest

The authors declare no conflict of interest.

Поступила в редакцию 11.06.2026
Поступила после рецензирования 10.08.2026
Принята к публикации 10.08.2026
Опубликована 30.09.2026

Received 2026.06.11
Revised 2026.08.10
Accepted 2026.08.10
Published 2026.09.30