ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ

УДК 004.89 Научная статья
DOI: 10.35330/1991-6639-2024-26-5-94-106
EDN: LTGJVQ

Основные направления интеллектуального анализа данных в сфере образования

Н. А. Попова[✉1], Е. С. Егорова[2]

[1]Пензенский государственный университет
440026, Россия, г. Пенза, ул. Красная, 40
[2]Пензенский государственный технологический университет
440039, Россия, г. Пенза, проезд Байдукова/ул. Гагарина, 1а/11

Аннотация. Интеллектуальный анализ данных в сфере образования становится все более популярным, и многие высшие учебные заведения все чаще применяют его для повышения своей конкурентоспособности. В последние годы было проведено множество исследований по анализу образовательных данных по различным учебным темам и с использованием различных методов и алгоритмов. Поэтому было бы полезно иметь краткий обзор наиболее используемых методов и подходов. С этой целью был произведен анализ зарубежных и отечественных трудов для выявления самых актуальных направлений исследований, важных методов и алгоритмов в области анализа образовательных данных в современных вузах. Для составления обзора была предложена методология систематизированного анализа, состоящая из 5 этапов. Были выявлены наиболее используемые темы, методы, алгоритмы и установлена взаимосвязь между ними. Научная новизна обзора заключается в определении актуальных задач исследований в области анализа образовательных данных в вузах и обнаружении перспективных методов и алгоритмов исследований.

Ключевые слова: Data Mining, интеллектуальный анализ образовательных данных, мета-анализ, бизнес-аналитика

Поступила 24.09.2024, одобрена после рецензирования 09.10.2024, принята к публикации 11.10.2024

Для цитирования. Попова Н. А., Егорова Е. С. Основные направления интеллектуального анализа данных в сфере образования // Известия Кабардино-Балкарского научного центра РАН. 2024. Т. 26. № 5. С. 94–106. DOI: 10.35330/1991-6639-2024-26-5-94-106

MSC: 68T09 Original article

Main directions of data mining in the field of education

N.A. Popova[✉1], E.S. Egorova[2]

[1]Penza State University
440026, Russia, Penza, 40 Krasnaya street
[2]Penza State Technological University
440039, Russia, Penza, 1a/11 Baidukova passage/Gagarina street

Abstract. Data mining in education is becoming increasingly popular and many educational institutions are increasingly applying it to improve their competitiveness. Many studies have been conducted recently on educational data analysis on different educational topics with the use of different methods and algorithms. Therefore, it would be useful to have a brief overview of the most used

methods and approaches. For this purpose, foreign and domestic works were analyzed to identify the most relevant research directions, important methods and algorithms in the field of educational data analysis in modern higher education. A systematic analysis methodology consisting of 5 stages was proposed to compile the review. Widely used topics, methods, algorithms were identified and the relationship between them was established. The scientific novelty of the overview lies in identifying the current research challenges in the field of educational data analysis in higher education and discovering promising research methods and algorithms.

Keywords: Data Mining, educational data mining, meta-analysis, business intelligence

Submitted 24.09.2024, approved after reviewing 09.10.2024, accepted for publication 11.10.2024

For citation. Popova N.A., Egorova E.S. Main directions of data mining in the field of education. *News of the Kabardino-Balkarian Scientific Center of RAS*. 2024. Vol. 26. No. 5. Pp. 94–106. DOI: 10.35330/1991-6639-2024-26-5-94-106

Введение

В последние годы в связи с развитием технологий произошли значительные изменения в сфере образования, процессе получения знаний обучающимися, в процессе обучения преподавателями и роли образовательных учреждений [1]. В современной системе высшего образования цифровизация привела к появлению большого объема данных, которые играют важную роль в реформировании и развитии образования, при этом возникла необходимость в более специальной области интеллектуального анализа данных – Анализе образовательных данных (Educational Data Mining) (АОД). АОД стал неотъемлемой частью процесса преподавания и обучения, и его главная цель – понимание особенностей обучения и максимальная его оптимизация. Эта область интеллектуального анализа является междисциплинарным предметом, поскольку охватывает различные области, такие как статистика, информатика и педагогика [14]. АОД включает типовые этапы Data Mining: сбор, предварительная обработка, преобразование, построение модели, обучение, тестирование и развертывание [2]. В то же время одним из основных направлений его деятельности является разработка методов, которые помогут исследовать уникальные типы данных [2, 3], понять студентов и контекст обучения [3]. Подобную стратегию используют самые современные университеты, которые хотят модернизировать свою тактику управления и облегчить процесс принятия решений [10, 12]. Эти учебные заведения сосредоточены на оптимизации своей работы и результативности для повышения своей конкурентоспособности. Помимо самого вуза, АОД также может предоставлять информацию для поддержки и помощи каждому студенту [12]. Благодаря разнообразным приложениям мы наблюдаем быстрый рост популярности этой области исследований [2, 15].

Несмотря на множество систематических обзоров по АОД, определить наиболее распространенные методы анализа может быть непросто, и часто требуются рекомендации по выбору оптимальных подходов. Поэтому систематизация методов и подходов к управлению данными может улучшить использование АОД в будущем. Целью данного исследования является составление систематического обзора трудов по интеллектуальному анализу данных в высшем образовании для определения актуальных направлений, методов и алгоритмов. Задача заключается в том, чтобы проанализировать наиболее распространенные методы и возможности их применения к конкретной предметной области – ВУЗу. Для этого был проведен мета-анализ – систематизированный обзор 15 отобранных трудов по анализу образовательных данных.

Методология систематизированного анализа

Для проведения систематизированного анализа была разработана методология, включающая в себя пять этапов: вопросы исследования, источники данных, ключевые слова, критерии включения и исключения, извлечение данных (рис. 1).

Рис. 1. Систематизированные шаги методологии исследования

Fig. 1. Systematic steps of the research methodology

1). Определение вопросов исследования – характеризуется определением того, на какие вопросы будут даны ответы.

2). Выбор источников данных – источники должны быть достоверными и качественными.

3). Определение ключевых слов – слова должны быть репрезентативными для рассматриваемого исследования, и следует использовать несколько их комбинаций, чтобы охватить как можно больше публикаций.

4). Определение критериев включения и исключения – критерии позволяют фильтровать полученные публикации, и их определение должно быть строгим, чтобы исключить ненужные исследования и включить важные исследования, поэтому необходимо знать изучаемую область, чтобы критерии были точными и убедительными.

5) Анализ и извлечение информации – публикации фильтруются и анализируются для извлечения релевантной информации, что может быть успешным только в случае правильного выполнения предыдущих шагов, также данный шаг должен быть согласован с целями исследования, а также с шагом 4).

На первом этапе был проведен поиск исследований по выбранным источникам данных и извлечение названий, в результате чего было получено 85 исходных исследований. На втором этапе на основе анализа названий были определены потенциальные исследования, из которых было исключено 50, в результате чего 35 исследований были использованы на следующем этапе. Последний этап состоял из чтения аннотации, введения и результатов с использованием критериев включения и исключения. 20 исследований были исключены, а 15 использовались для извлечения релевантной информации. Большинство проанализированных исследований были опубликованы в 2023 году (n=6) и в 2020 году

(n=5), за ними следуют 2021 и 2024 годы – по 2 исследования. На рисунке 2 показано, как осуществлялся процесс отбора.

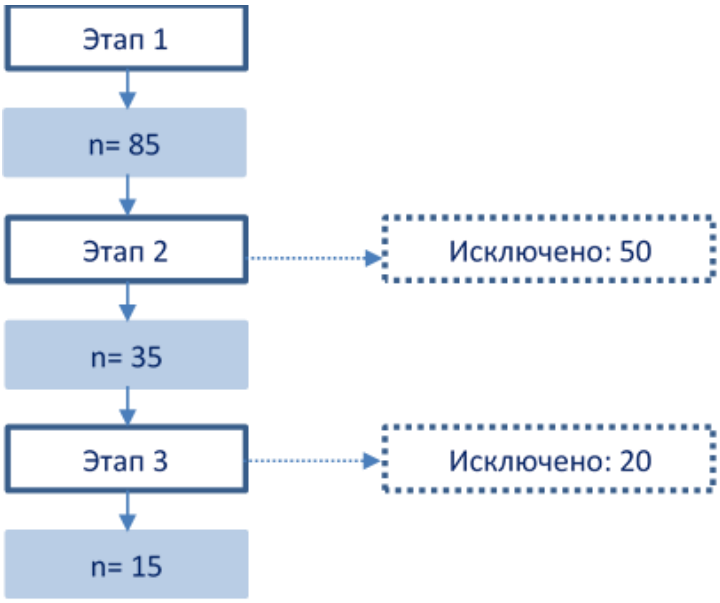

Рис. 2. *Этапы отбора научных трудов*

Fig. 2. *Stages of scientific papers selection*

Исследовательские вопросы – были выделены четыре исследовательских вопроса: ИВ1: Каковы основные темы исследований в области АОД? ИВ2: Каковы наиболее используемые методы в АОД? ИВ3: Каковы наиболее используемые алгоритмы в АОД? ИВ4: Какова связь между темами и методами, а также темами и алгоритмами, наиболее используемыми в АОД?

Источники данных – для проведения исследования были выбраны следующие источники данных: Google Scholar (https://scholar.google.com), КиберЛенинка (https://cyberleninka.ru) и eLIBRARY.RU (https://www.elibrary.ru). Включение этих трех источников данных было сделано для того, чтобы охватить максимальное количество публикаций, имеющих прямые связи между цитатами и цитируемыми статьями. КиберЛенинка и eLIBRARY.RU являются двумя наиболее используемыми базами данных в научных исследованиях, а Google Scholar предоставляет обширную академическую информацию с несколькими типологиями документов (монографии, диссертации, доклады и сообщения конференций).

Определение ключевых слов – использовались следующие ключевые слова для поиска: «интеллектуальный анализ образовательных данных» [название] ИЛИ «анализ образовательных данных» [название] ИЛИ «educational data mining» [название] И «обзор» [весь документ]. Используемые источники данных не делают различий между типами регистров, поэтому все поисковые запросы были выполнены в нижнем регистре. Эти ключевые слова были выбраны для получения результатов, связанных с целями исследования. Ключевые слова «educational data mining» и «анализ образовательных данных» были обязательными для поиска, так как они являются предметом нашего исследования, что в сочетании с ключевым словом «обзор» позволило нам объединить последние тенденции изучаемой области и таким образом создать полученный систематизированный обзор.

Критерии включения и исключения – авторами были определены следующие критерии включения: дата проведения в период с января 2020 по сентябрь 2024 года и ответы на один или несколько исследовательских вопросов. Критериями исключения были: недоступность, когда исследование недоступно (неработающий URL или ограничения учреждения-источника), дублирование исследования (повторное проведение одного и того же исследования) и узкоспециализированность темы исследования, когда применялся один метод или алгоритм к определенному аспекту или области анализа образовательных данных.

Извлечение информации – исследования были отобраны в соответствии с ключевыми словами поиска, определенными на шаге 3. Для дальнейшего анализа были отобраны 15 исследований, опубликованных в 2020, 2021, 2023 и 2024 годах.

Результаты

Результаты исследований распределены по четырем категориям: темы, методы, алгоритмы и взаимосвязь между ними. В результатах указано количество исследований, которое представляет собой сумму исследований из проанализированных публикаций. Поскольку в данной работе используются обзорные публикации, а они имеют разные диапазоны поиска, считаем, что использование количества исследований позволит лучше понять каждое измерение. Таким образом, количество упоминаний по исследованиям было суммировано, и результат представлен в виде общего количества ключевых слов, k.

1. Темы

Основные найденные темы исследований представлены в таблице 1 в порядке убывания общего количества. Из-за большого количества результатов были рассмотрены только темы, на которые ссылались более 10 раз. «Успеваемость учащихся» относится к тому, насколько хорошо студент успевает в учебе, и обычно измеряется с помощью оценок, стандартных тестов и других форм оценивания. Это наиболее изучаемая тема (k=426), которая упоминалась в 12 проанализированных публикациях. «Поддержка/эффективность работы преподавателей» предоставляет педагогам новые и новаторские способы использования и анализа результатов работы учителей. Эта тема изучалась 174 раза и упоминалась в 3 проанализированных публикациях. «Отсев учащихся» происходит, когда учащиеся покидают вуз до завершения своего образования. Эта тема использовалась 121 раз и упоминалась в 9 публикациях. «Поведение/особенности учащихся» – это способы взаимодействия и коммуникации учащихся. Эта тема изучалась 108 раз и упоминалась в 4 публикациях. «Электронная образовательная среда» – это физическая и психологическая среда, в которой происходит обучение, охватывающая все аспекты образовательного процесса. Эта тема использовалась 75 раз и упоминалась в 4 публикациях. «Рекомендации по дисциплине/ направлению» – это предложение или совет, который дается студенту или частному лицу относительно наилучшей образовательной программы, основанной на их интересах, способностях, карьерных целях или других факторах. Эта тема изучалась 44 раза и упоминалась в 5 публикациях.

Таблица 1. Основные темы исследования по анализу образовательных данных

Table 1. The main research topics on the educational data analysis

Темы исследования	k	Ссылки
Успеваемость студентов	426	[1–3, 5, 7–10, 12–15]
Поддержка/эффективность работы преподавателя	174	[1, 10, 12]
Отсев студентов	121	[1, 2, 4, 5, 9, 11, 12, 14, 15]
Поведение/особенности учащихся	108	[1, 7, 13, 15]
Электронная образовательная среда	75	[5, 7, 9, 14]
Рекомендации по дисциплине/направлению	44	[2, 9, 12–14]

2. Методы

Основные найденные методы анализа образовательных данных в обзорных публикациях представлены в таблице 2 в порядке убывания общего количества исследований. Для лучшего понимания добавлено краткое описание методов. «Регрессия»: моделирует взаимосвязь между зависимой переменной и одной или несколькими независимыми переменными. Цель данного метода – разработать математическую модель, которая может быть использована для прогнозирования зависимой переменной на основе значений независимых переменных. «Классификация»: используется для прогнозирования категориального результата на основе набора входных переменных. Цель данного метода – разработать модель, которая может быть использована для отнесения новых наблюдений к предопределенным классам. «Кластеризация»: используется для группировки похожих объектов или точек данных в кластеры. Цель данного метода состоит в том, чтобы разделить набор точек данных на отдельные группы таким образом, чтобы точки данных внутри каждой группы были как можно более похожими, а точки данных в разных группах – как можно более разными. «Статистический анализ»: раздел математики, занимающийся сбором, анализом, интерпретацией, представлением и организацией данных. Данный метод предоставляет инструменты и методы для обобщения, описания и вывода данных. «Прогнозирование»: оценка исхода будущего события на основе анализа прошлых тенденций и закономерностей. «Визуализация»: процесс создания графического представления данных, облегчающий их понимание, изучение и анализ. Цель данного метода – преобразовать необработанные данные в осмысленные идеи, шаблоны и взаимосвязи, которые можно передавать другим пользователям.

Из 15 проанализированных исследований в 8 упоминались используемые методы. «Регрессия» является наиболее часто используемой (k=306) и упоминалась в 5 проанализированных публикациях. «Классификация» использовалась 262 раза и упоминалась в 6 проанализированных публикациях. «Кластеризация» использовалась 218 раз и упоминалась в 7 проанализированных публикациях. «Статистический анализ» использовался 177 раз и упоминался в 2 проанализированных публикациях. «Прогнозирование» использовалось 86 раз и упоминалось в 3 проанализированных публикациях. Термин «Визуализация» был использован 63 раза и упомянут в 3 проанализированных публикациях.

Таблица 2. Методы интеллектуального анализа образовательных данных

Table 2. Methods of intellectual analysis of educational data

Методы	k	Ссылки
Регрессия	306	[1, 7, 9–11]
Классификация	262	[4, 7, 9–11, 14]
Кластеризация	218	[1, 4, 7, 9–11, 14]
Статистический анализ	177	[1, 9]
Прогнозирование	86	[11, 13, 14]
Визуализация	63	[1, 9, 11]

3. Алгоритмы

Основные найденные алгоритмы анализа образовательных данных в обзорных публикациях представлены в таблице 3 в порядке убывания общего количества исследований. Для лучшего понимания добавлено краткое описание алгоритмов. «Дерево решений» (Decision tree): используется как для задач классификации, так и для задач регрессии. Это древовидная модель, представляющая ряд решений и их возможные последствия. Данный алгоритм легко понять, визуализировать и интерпретировать, что делает данный алгоритм полезным инструментом для изучения сложных данных, выявления закономерностей и взаимосвязей в данных. «Наивный байесовский классификатор» (Naive Bayes): вероятностный алгоритм, основанный на теореме Байеса. Алгоритм вычисляет вероятность принадлежности точки данных к каждому классу на основе ее характеристик, а затем выбирает класс с наибольшей вероятностью для прогнозирования. «Нейронная сеть» (Neuronal Network): модель, основанная на структуре и функциях человеческого мозга. Данный алгоритм состоит из множества взаимосвязанных узлов обработки, называемых искусственными нейронами, которые работают сообща для решения сложных задач. Он обладает высокой гибкостью и может обрабатывать сложные и нелинейные взаимосвязи в данных. «Метод опорных векторов» (SVM): используется для классификации и регрессионного анализа. Данный алгоритм работает путем нахождения гиперплоскости в многомерном пространстве, которая наилучшим образом разделяет данные на различные классы. Гиперплоскость выбирается таким образом, чтобы максимально увеличить поле, представляющее собой расстояние между гиперплоскостью и ближайшими точками данных, известными как опорные векторы. Данный алгоритм может обрабатывать линейные и нелинейные взаимосвязи в данных. «Ассоциативные правила» (Associative Rules): выявлять взаимосвязи и закономерности. Цель алгоритма – найти связи между элементами. Правила обнаруживаются с помощью алгоритмов, которые идентифицируют часто используемые наборы элементов и затем генерируют правила на основе этих наборов. Затем правила оцениваются с использованием показателей, которые измеряют силу и значимость ассоциаций. «Логистическая регрессия» (Logistic Regression): статистический метод для задач бинарной классификации, целью которого является предсказание одного из двух возможных результатов. Данный алгоритм моделирует взаимосвязь между зависимой переменной и одной или несколькими независимыми переменными путем подгонки логистической кривой к данным. Логистическая кривая представлена логистической функцией, которая сопоставляет прогнозируемый результат с вероятностью от 0 до 1, что указывает на вероятность того, что результат относится к одному из двух классов. Затем прогнозируемый класс определяется путем

применения порогового значения к прогнозируемой вероятности. «Случайный лес» (Random Forest): используется для задач классификации и регрессии. Данный алгоритм состоит из нескольких деревьев решений, каждое из которых обучено на случайном подмножестве данных и случайном подмножестве признаков. Окончательный прогноз составляется путем объединения прогнозов всех деревьев решений. Цель алгоритма – уменьшить дисперсию в прогнозах отдельных деревьев путем объединения их выходных данных, что позволяет получить более надежную и стабильную модель. «K-ближайших соседей» (K-Nearest Neighbors или KNN)» (KNN): используется для классификации и регрессии. Данный алгоритм классифицирует точку данных на основе ее ближайших соседей, которые определяются путем вычисления расстояния между точками данных. После определения ближайших соседей точка данных присваивается классу с использованием механизма голосования большинством голосов. Класс с наибольшим числом соседей выбирается в качестве класса, к которому принадлежит точка данных.

Из 15 проанализированных исследований в 9 упоминались используемые алгоритмы. «Дерево решений» является наиболее часто используемым (k=346) и упоминалось в 8 проанализированных публикациях. «Наивный байесовский классификатор» использовался 149 раз и упоминался в 6 проанализированных публикациях. Термин «Нейронная сеть» использовался 112 раз и упоминался в 4 проанализированных публикациях. Термин «SVM» использовался 110 раз и упоминался в 6 проанализированных публикациях. «Ассоциативные правила» использовались 89 раз и упоминались в 4 проанализированных публикациях. «Логистическая регрессия» использовалась 68 раз и упоминалась в 5 проанализированных публикациях. «Случайный лес» также использовался 68 раз и упоминался в 4 проанализированных публикациях. «KNN» использовался 23 раза и упоминался в 4 проанализированных публикациях.

Таблица 3. Алгоритмы анализа образовательных данных

Table 3. Algorithms for analyzing educational data

Методы	k	Ссылки
Дерево решений	346	[1, 2, 7, 9, 11–14]
Наивный байесовский классификатор	149	[2, 7, 9, 11–13]
Нейронная сеть	112	[2, 7, 9, 14]
SVM	110	[2, 7, 9, 11–13]
Ассоциативные правила	89	[4, 7, 9, 14]
Логистическая регрессия	68	[2, 9, 11–13]
Случайный лес	68	[9, 11–13]
KNN	23	[2, 11–13]

4. Взаимосвязь тем с методами и алгоритмами

Связь между методами и темами исследований представлена в таблице 4. Наиболее связаны между собой «Кластеризация» – «Успеваемость студентов» и «Кластеризация» – «Отсев студентов» – по 5 исследований, затем следуют «Классификация» – «Успеваемость студентов», «Классификация» – «Отсев студентов» и «Регрессия» – «Успеваемость студентов» – по 4 исследования. Исследований, связывающих «Прогнозирование» – «Поддержка/эффективность работы преподавателя» или «Визуализация» – «Поддержка/эффективность работы преподавателя», нет. Остальные взаимосвязи представлены в таблице 4.

Таблица 4. Взаимосвязь между методами и темами исследований

Table 4. Relationship among research methods and topics

	Регрессия	Классификация	Кластеризация	Статистический анализ	Прогнозирование	Визуализация
Успеваемость студентов	4 [1,7,9,10]	4 [7,9,10,14]		2 [1, 9]	2 [13, 14]	2 [1, 9]
Поддержка/эффективность работы преподавателя	2 [1, 10]	1 [10]	2 [1, 10]	1 [11]		
Отсев студентов	3 [1, 9, 11]	4 [4,9,11,14]		2 [1, 9]	2 [11, 14]	3 [1, 9, 11]
Поведение/особенности учащихся	2 [1, 7]	1 [7]	2 [1, 7]	1 [1]	1 [13]	1 [1]
Электронная образовательная среда	2 [7, 9]	3 [7, 9, 14]	3 [7, 9, 14]	1 [9]	1 [14]	1 [9]
Рекомендации по дисциплине/ направлению	1 [9]	2 [9, 14]	2 [9, 14]	1 [9]	2 [13, 14]	1 [9]

Взаимосвязь между алгоритмами и темами исследований представлена в таблице 5. Наиболее тесно связаны «Дерево решений» – «Успеваемость студентов» – 7 исследований, затем следует «Дерево решений» – «Отсев студентов» – 6 исследований. Нет исследований, связывающих «Нейронная сеть» – «Поддержка/эффективность работы преподавателя», «Ассоциативные правила» – «Поддержка/эффективность работы преподавателя», «Случайный лес» – «Успеваемость студентов» и «KNN» – «Учебная среда».

Остальные взаимосвязи можно увидеть в таблице 5.

Таблица 5. Взаимосвязь между алгоритмами и темами исследования

Table 5. The relationship among algorithms and research topics

	Дерево решений	Наивный байесовский классификатор	Нейронная сеть	SVM	Ассоциативные правила	Логистическая регрессия	Случайный лес	KNN
Успеваемость студентов	7 [1, 2, 7, 9, 12–14]	5 [2, 7 ,9 ,12, 13]	4 [2, 7, 9, 14]	5 [2, 7, 9, 12, 13]	3 [7, 9, 14]	4 [2, 9, 12, 13]	3 [9, 12, 13]	3 [2, 12, 13]
Поддержка/эффективность работы преподавателя	2 [1, 12]	1 [12]		1 [12]		1 [12]	1 [9]	1 [12]
Отсев студентов	6 [1, 2, 9, 11, 12, 15]	4 [2, 9, 11, 12]	3 [2, 9, 14]	4 [2, 9, 11, 12]	3 [4, 9, 14]	4 [2, 9, 11, 12]	4 [9, 11–13]	3 [2, 11, 12]
Поведение/ особенности учащихся	3 [1, 7, 13]	2 [7, 13]	1 [7]	2 [7, 13]	1 [7]	1 [13]	1 [13]	1 [13]
Электронная образовательная среда	3 [7, 9, 14]	2 [7, 9]	3 [7, 9, 14]	2 [7, 9]	3 [7, 9, 14]	1 [9]	1 [9]	
Рекомендации по дисциплине/ направлению	5 [2, 9, 12–14]	4 [2, 9, 12, 13]	3 [2, 9, 14]	4 [2, 9, 12, 13]	2 [9, 14]	4 [2, 9, 12, 13]	3 [9, 12, 13]	3 [2, 12, 13]

Обсуждение

После отбора и анализа 15 исследований можно получить ответы на четыре определенных исследовательских вопроса.

Вопрос 1: Каковы основные темы исследований в области АОД? Применение АОД может быть самым разнообразным. В ходе нашего исследования убедились, что тема «Успеваемость студентов» является наиболее изученной, и почти все проанализированные публикации ссылаются на нее. Эта тема изучалась почти в три раза больше, чем вторая по изученности тема, и почти в десять раз больше, чем наименее изученная тема. «Поддержка/эффективность работы преподавателя» – вторая наиболее изученная тема, за ней следует «Отсев студентов». «Поведение/особенности учащихся» – четвертая по количеству изученных тем, за ней следует «Электронная образовательная среда». Наименее изученной оказалась тема «Рекомендации по дисциплине/направлению».

Вопрос 2: Какие методы наиболее часто используются в АОД? Согласно нашим результатам наиболее популярным методом, используемым в АОД, является «Регрессия», за которой следуют «Классификация» и «Кластеризация». Определение «Статистический анализ» с помощью специального программного обеспечения заняло 4-е место, за ним следуют «Прогнозирование» и «Визуализация». В исследованиях [2, 3, 5, 8, 12, 15] не упоминались используемые методы.

Вопрос 3: Какие алгоритмы наиболее часто используются в АОД? Согласно нашим результатам наиболее популярным алгоритмом, используемым в АОД, является «Дерево решений», за которым следуют «Наивный байесовский классификатор» и «Нейронная сеть», которые использовались менее чем в половине случаев. «SVM» занял 4-е место по популярности, за ним следуют «Ассоциативные правила», «Логистическая регрессия», «Случайный лес» и «KNN». В исследованиях [3, 5, 8, 10, 15] не упоминались используемые алгоритмы.

Вопрос 4: Какова взаимосвязь между темами и методами, а также темами и алгоритмами, наиболее часто используемыми в АОД? Согласно нашим результатам наиболее схожими методами и темами исследования были «Кластеризация» – «Успеваемость студентов» и «Кластеризация» – «Отсев студентов», на которые ссылаются 5 исследований. «Классификация» – «Успеваемость учащихся», «Классификация» – «Отсев учащихся» и «Регрессия» – «Успеваемость учащихся» связаны в каждом из 4 исследований. «Успеваемость учащихся» является наиболее изучаемой темой, и то, что она является одной из наиболее связанных тем, подтверждает ее частое использование. Наиболее схожим алгоритмом и темой исследования было «Дерево решений» – «Успеваемость студентов» (7 исследований). «Дерево решений» – «Отсев студентов» – связано с 6 исследованиями. В данном случае «Успеваемость учащихся» наиболее тесно связана с алгоритмом «Дерева решений», где «Успеваемость учащихся» является наиболее распространенной темой, а «Дерево решений» – наиболее часто используемым алгоритмом.

Заключение, ограничения исследования и будущие исследования

В данной работе систематизирован обзор 15 исследований, связанных с АОД, показывающий, что «Успеваемость студентов» является самой популярной изучаемой темой, которая использовалась 426 раз в 12 проанализированных статьях. Термин «Дерево решений», представляющий собой алгоритм, используемый для классификации или регрессии,

был наиболее часто используемым, что могло быть связано с наиболее часто используемыми методом и темой в целом для анализа данных. Другими часто изучаемыми темами были «Поддержка/эффективность работы преподавателя», «Отсев студентов», «Поведение/особенности студентов» и «Рекомендации по дисциплине/направлению». Такие методы, как «Классификация», «Кластеризация», «Статистический анализ», «Прогнозирование» и «Визуализация данных», также упоминались, но реже, чем регрессия. Другими алгоритмами, представленными в проанализированных исследованиях, были «Наивный байесовский классификатор», «Нейронная сеть», «SVM», «Ассоциативные правила», «Логистическая регрессия», «Случайный лес» и «KNN». Более высокой взаимосвязью между методом и темой исследования являются «Кластеризация» – «Успеваемость студентов» и «Кластеризация» – «Отсев студентов», а между алгоритмом и темой исследования – «Дерево решений» – «Успеваемость студентов».

У этого исследования были некоторые ограничения. Выбор источников из базы данных и определенные критерии ограничивали сферу охвата, которая может не охватывать некоторые исследовательские работы. В исследованиях использовались разные термины, что иногда затрудняло сравнение и обобщение результатов. Были исследования из вторичных источников, в которых не хватало деталей. Некоторые исследования не дали ответов на все вопросы исследования.

Результаты работы могут дать ответы на поставленные исследовательские вопросы, послужить руководством для будущих исследований АОД и стать ценной отправной точкой в расширяющейся области интеллектуального анализа образовательных данных.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. Du X., Yang J., Hung J.-L., Shelton B. Educational data mining: a systematic review of research and emerging trends. *Information Discovery and Delivery*. 2020. No. 48(4). Pp. 225–236. DOI: 10.1108/idd-09-2019-0070

2. *Семенкина И. А., Прусакова П. В.* Направления исследований в области анализа образовательных данных в высшей школе: теоретический обзор // Педагогика. Вопросы теории и практики. 2023. Т. 8. № 7. С. 761–770. DOI: 10.30853/ ped20230111

Semenkina I.A., Prusakova P.V. Research directions in the field of educational data analysis in higher education: A theoretical review. *Pedagogika. Voprosy teorii i praktiki* [Pedagogy. Theory and Practice]. 2023. Vol. 8. No. 7. Pp. 761–770. DOI: 10.30853/ped20230111. (In Russian)

3. *Политов А. Ю., Акжигитов Р. Р., Сударикov К. А.* Анализ моделей и инструментов предиктивной аналитики для анализа образовательных данных // Инновации. Наука. Образование. 2021. № 28. С. 1055–1065. EDN: SSBBWE

Politov A.Yu., Akzhigitov R.R., Sudarikov K.A. Analysis of predictive analytics models and tools for analyzing educational data. *Innovatsii. Nauka. Obrazovanie* [Innovations. Science. Education]. 2021. No. 28. Pp. 1055–1065. EDN: SSBBWE. (In Russian)

4. Salal Ya. Kh., Abdullaev S.M. Monitoring of the education quality and implementing of individual learning: demonstration of approaches and educational data mining algorithms. *Izvestiya SFedU. Engineering Sciences*. 2020. No. 3(213). Pp. 112–122. DOI: 10.18522/2311-3103-2020-3-112-122

5. Bunkar K. Educational data mining in practice literature review. *Journal of Advanced Research in Embedded System*. 2020. Vol. 7. Pp. 1–7. DOI: 10.24321/2395.3802.202001

6. *Терентьев А. В.* Методы и алгоритмы интеллектуального анализа данных в образовании // Вестник науки. 2024. Т. 4. № 5(74). С. 1545–1550.

Terentyev A.V. Methods and algorithms for data mining in education. *Vestnik nauki* [Science Bulletin]. 2024. Vol. 4. No. 5(74). Pp. 1545–1550. (In Russian)

7. Dol S. M., Jawandhiya P. M. Classification technique and its combination with clustering and association rule mining in educational data mining–A survey. *Engineering Applications of Artificial Intelligence*. 2023. Vol. 122. P. 106071. DOI: 10.1016/j.engappai.2023.106071

8. Rabelo A., Rodrigues M.W., Nobre C.N., Isotani S. Educational data mining and learning analytics: A review of educational management in e-learning. *Information Discovery and Delivery*. 2024. Vol. 52. No. 2. Pp. 149–163. DOI: 10.1108/IDD-10-2022-0099

9. Bošnjaković N., Đurđević Babić I. Systematic review on educational data mining in educational gamification. *Tech Know Learn*. 2023. Vol. 21. Pp. 5–19. DOI: 10.1007/s10758-023-09686-2

10. Salloum S.A., Elnagar A., Shaalan K., Alshurideh M. Mining in educational data: review and future directions. *Advances in Intelligent Systems and Computing*. 2020. Vol. 1153. Pp. 92–102. DOI: 10.1007/978-3-030-44289-7_9

11. Andrade T., Rigo S., Barbosa J. Active methodology, educational data mining and learning analytics: a systematic mapping study. *Informatics in Education*. 20201. Vol. 20. No. 2. Pp. 171–204. DOI: 10.15388/infedu.2021.09

12. Ampadu Y.B. Handling big data in education: a review of educational data mining techniques for specific educational problems. *AI, Computer Science and Robotics Technology*. 2023. No. 13. DOI: 10.5772/ACRT.17

13. *Попова Н. А., Егорова Е. С.* Data mining в образовании: прогнозирование успеваемости учащихся // Моделирование, оптимизация и информационные технологии. 2023. Т. 11. № 2(41). С. 9–10. DOI: 10.26102/2310-6018/2023.41.2.003

Popova N.A., Egorova E.S. Data mining in education: forecasting student performance. *Modelirovanie, optimizatsiya i informatsionnye tekhnologii* [Modeling, optimization and information technology]. 2023. Vol. 11. No. 2 (41). Pp. 9–10. DOI: 10.26102/2310-6018/2023.41.2.003. (In Russian)

14. *Попова Н. А., Егорова Е. С.* Интеллектуальный анализ образовательных данных для прогноза успеваемости студентов вуза // Известия Кабардино-Балкарского научного центра РАН. 2023. № 2(112). С. 18–29. DOI: 10.35330/1991-6639-2023-2-112-18-29

Popova N.A., Egorova E.S. Intelligent analysis of educational data to forecast university students' academic performance. *News of the Kabardino-Balkarian Scientific Center of RAS*. 2023. No. 2(112). Pp. 18–29. DOI: 10.35330/1991-6639-2023-2-112-18-29. (In Russian)

15. Kovalev S., Kolodenkova A., Muntyan E. Educational data mining: current problems and solutions. *2020 5th International Conference on Information Technologies in Engineering Education, Inforino 2020 – Proceedings*. Moscow, 14–17 апреля 2020 года. Moscow, 2020. P. 9111699. DOI: 10.1109/Inforino48376.2020.9111699

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Финансирование. Исследование проведено без спонсорской поддержки.

Funding. The study was performed without external funding.

Информация об авторах

Попова Наталия Александровна, канд. техн. наук, доцент, доцент кафедры «Математическое обеспечение и применение ЭВМ», Пензенский государственный университет;
440026, Россия, г. Пенза, ул. Красная, 40;
popov.tasha@yandex.ru, ORCID: https://orcid.org/0000-0001-9713-4897, SPIN-код: 9358-8567
Егорова Екатерина Сергеевна, канд. экон. наук, доцент кафедры «Прикладная информатика», Пензенский государственный технологический университет;
440039, Россия, г. Пенза, проезд Байдукова/ул. Гагарина, 1а/11;
katepost@yandex.ru, ORCID: https://orcid.org/0000-0002-0816-0944, SPIN-код: 5624-6036

Information about the authors

Nataliya A. Popova, Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of Mathematical Support and Computer Use, Penza State University;
440026, Russia, Penza, 40 Krasnaya street;
popov.tasha@yandex.ru, ORCID: https://orcid.org/0000-0001-9713-4897, SPIN-код: 9358-8567
Ekaterina S. Egorova, Candidate of Economic Sciences, Associate Professor of the Department of Applied Informatics, Penza State Technological University;
440039, Russia, Penza, 1a/11 Baidukova Passage/Gagarina street;
katepost@yandex.ru, ORCID: https://orcid.org/0000-0002-0816-0944, SPIN-код: 5624-6036