УДК 378
DOI 10.54835/18102883_2021_30_5

# ИНСТРУМЕНТ РАЗВИТИЯ ОБРАЗОВАТЕЛЬНЫХ ПРОЦЕССОВ: DATA SCIENCE

**Хайруллина Эльмира Робертовна**,
доктор педагогических наук, декан,
Факультет дизайна и программной инженерии,
Elm.khair@list.ru

Казанский национальный исследовательский технологический университет,
Россия, 420015, г. Казань, ул. Карла Маркса, 68.

Рассматриваются и изучаются последние тенденции в области науки о данных в образовании, чтобы рассмотреть современные направления и вклад в эпоху смарт-образования. Это включает в себя набор тщательно отрецензированных рукописей мирового класса, в которых рассматриваются и подробно описываются самые современные рамочные и методические исследовательские проекты в области науки о данных, применяемые в образовании, с использованием различных подходов, таких как слияние информации, мягкие вычисления, машинное обучение, Интернет вещей и др. На основе этого систематического обзора мы сформулировали некоторые рекомендации и предложения для исследователей, практиков и ученых по улучшению качества исследований в этой области.

**Ключевые слова:** Педагогическая технология, наука о данных, технология Data Science, образовательные системы.

## Введение

Термин Data Science (DS) относится к междисциплинарной области, которая включает в себя ряд методов, процессов и систем, направленных на извлечение знаний из данных. DS, которая является дисциплиной, очень связанной с вычислительной техникой, доказала свое применение в самых разных областях, особенно в образовании [1]. В образовательной среде происходит множество процессов, связанных с обучением, и в учебных заведениях постоянно генерируется большое количество потенциально значимых для обучения данных. Чтобы извлечь знания из этих данных для лучшего понимания процессов, связанных с обучением, использование подхода DS представляется полезным и необходимым [2].

Наука о данных лежит на стыке статистики и информатики и применяется в конкретных областях, таких как астрономия, лингвистика, медицина, психология или социология. Идея этой науки состоит в том, чтобы использовать большие данные для решения неразрешимых проблем, например, как медицинские работники могут создавать персонализированные лекарства на основе генов пациента или как предприятия могут делать прогнозы покупок на основе поведения клиентов. Для ответственного применения науки о данных, являющейся мощным инструментом, требуется обучение тому, как ее использовать и как понимать ее последствия.

Применение DS в области образования может представлять большой интерес для заинтересованных сторон (студентов, преподавателей, учебных заведений, ...), поскольку извлеченные знания из образовательных данных будут полезны для решения таких образовательных проблем, как повышение успеваемости студентов, высокий уровень оттока в учебных заведениях, задержки в обучении и так далее. Существует ряд дисциплин, связанных с Data Science, таких как Educational Data Mining и Learning Analytics, и все они важны для данного специального выпуска [3].

## Материалы и методы исследования

Цель данной статьи – представить материалы исследований по применению методов DS для извлечения знаний, представляющих интерес для участников образовательного процесса, при условии, что анализируемые данные представляют определённый образовательный процесс, а извлечённые знания используются для улучшения этого процесса. Мы рассматривали работы, включающие обсуждение реализации программных и/или аппаратных подходов, которые также фокусируются на последствиях для улучшения любого учебного процесса. Приоритет был отдан работам, которые демонстрируют сильное обоснование в теории обучения и/или строгий дизайн образовательных исследований. Мы рассматривали исследования, посвященные

высшему и последующему образованию любого типа (электронное обучение, смешанное и традиционное образование). Все принятые работы включают исчерпывающую проверку и содержат исключительно новые идеи в данной области.

Исследование, представленное в статье «Multilayered-Quality Education Ecosystem (MQEE): An Intelligent Education Modal for Sustainable Quality Education») (Verma A., и др.) [4], направлено на выявление некоторых скрытых параметров, которые влияют на экосистему качественного образования (Quality Education Ecosystem). Академическая неосведомленность, неучастие, неудовлетворенность и непонятность – вот основные рассматриваемые факторы. Для изучения влияния этих параметров на качество образования на уровне учебного заведения выдвигается ряд гипотез и проводятся опросы. Метод двунаправленной взвешенной суммы используется для получения точных и достоверных результатов анализа граничной стоимости исследования. Связь между параметрами недоученности и качеством образования иллюстрируется с помощью корреляционных диаграмм и диаграмм рассеяния. Академическое безделье, скрытый и непреднамеренный рудимент, который влияет на QEE, также определен, намечен и исследован в этой работе.

В работе «Improving prediction of students' performance in intelligent tutoring systems using attribute selection and ensembles of different multimodal data sources») [5] авторы намерены предсказать успеваемость студентов университета, используя различные источники показателей и мультимодальных данных из интеллектуальной обучающей системы. Они собрали и предварительно обработали данные 40 студентов из различных мультимодальных источников: стратегии обучения из системных журналов, эмоции из видеозаписей мимики, распределение и фиксацию внимания из отслеживания глаз, а также результаты тестов на знание предмета. Их цель – проверить, можно ли улучшить предсказание, используя выбор атрибутов и ансамбли классификации.

В работе «Automated text detection from big data scene videos in higher education» (Manasa Devi M. и др.) [6] использовали новый подход к очистке видеокадров для подачи нейросетевой модели на основе сети предложения регионов (region proposal network) с конволюционными нейронными сетями путем поиска соответствующих соотношений якорей для извлечения кандидатов на текст. Обученная модель с извлеченными кадрами предсказывает для тестовых видео. Предложенный метод был оценен на эталонном наборе данных ICDAR Video text и нескольких общедоступных тестовых наборах данных, что позволило достичь высокого показателя запоминания.

В статье «Improve teaching with modalities and collaborative groups in an LMS: an analysis of monitoring using visualisation techniques») (Sáiz-Manzanares и др.) [7] основной целью является проверка эффективности трех форм преподавания (все они используют онлайн-обучение на основе проектов OPBL и Flipped Classroom и отличаются использованием виртуальных лабораторий и интеллектуального персонального помощника IPA) на поведение в Moodle и успеваемость студентов с учетом варианта формата совместной группы. Использовались как количественные, так и качественные методы исследования. Что касается количественного анализа, были обнаружены различия в поведении студентов в Moodle и в результатах обучения в зависимости от методов обучения, включающих виртуальные лаборатории. Аналогичным образом качественное исследование проанализировало модели поведения, найденные в каждой совместной группе в трех изученных модальностях обучения.

Исследование «Fuzzy-based Active Learning for Predicting Student Academic Performance using autoML: a step-wise approach» (Tsiakmaki и др.) [8], представляет метод нечеткого активного обучения для прогнозирования академической успеваемости студентов, который модульно сочетает в себе методы autoML. Было проведено множество экспериментов, показавших эффективность предложенного метода для точного прогнозирования студентов, подверженных риску неуспеваемости. Эти данные могут быть полезны для поддержки учебного процесса и более широкого изучения науки об образовании.

В статье «Peer Assessment Using Soft Computing Techniques» (Pinargote-Ortega и др.) [9], был применен сценарий оценки коллег в Техническом университете Манаби (Эквадор). Студенты и преподаватели оценивают некоторые работы с помощью рубрикаторов, выставляют числовой балл и текстовый отзыв, обосновывающий причины, по которым был выставлен такой числовой балл. Интерес

представляет сценарий выявления неточностей между обеими оценками. Предлагается модель с использованием методов «мягких» вычислений для обнаружения неточностей и снижения нагрузки на преподавателя в процессе исправления.

Авторы статьи «A Novel Automated Essay Scoring Approach for Reliable Higher Educational Assessments» [10] представляют нейросетевую модель на основе трансформатора для повышения эффективности автоматической оценки эссе с использованием Bi-LSTM (Bidirectional Long Short-Term Memory) и языковой модели RoBERTa на основе набора данных ASAP (Automated Student Assessment Prize) от Kaggle. Предлагаемая модель использует модель Bi-LSTM над предварительно обученной языковой моделью RoBERTa для решения проблемы связности в эссе, которая игнорируется традиционными методами оценки эссе, включая традиционные конвейеры обработки естественного языка, методы на основе глубокого обучения, смесь обоих. Сравнение экспериментальных результатов по оценке эссе с человеческими оценщиками показывает, что предложенная модель превосходит существующие методы оценки эссе по показателю QWK (Quadratic Weighted Kappa).

Основной целью исследования «Personalized training model for organizing blended and lifelong distance learning courses and its effectiveness in Higher Education» [11] является повышение персонификации обучения в высшем образовании. Предлагаемая гибкая модель организации смешанного и дистанционного обучения в высшем образовании предполагает создание индивидуальной траектории обучения путем тестирования студентов перед началом обучения. На основании результатов обучения студент зачисляется на учебную траекторию. Учебная траектория состоит из обязательных и дополнительных модулей для обучения; дополнительные модули можно не изучать в случае успешного прохождения теста по ним. В статье рассматривается состав интеллектуальных обучающих систем: модель студента, модель обучения и модель интерфейса.

Авторы статьи «IoT Text Analytics in Smart Education and Beyond» (Mohammed A.H.K., и др.) [12] освещают основные компоненты аналитики IoT, а также дают всесторонний обзор используемых методов и приложений текстовой аналитики и сравнение используемых моделей и методов текстовой аналитики IoT в интеллектуальном образовании и многих других приложениях.

Наконец, в статье «A Framework to Capture the Dependency between prerequisite and Advanced Courses in Higher Education») [13] авторы предлагают новый алгоритм анализа графиков в сочетании со статистическим анализом для выявления зависимых отношений между результатами обучения по курсам (CLOs) предварительных и продвинутых курсов. Кроме того, построена новая модель для прогнозирования успеваемости студентов на продвинутых курсах на основе предварительных требований. Оценка доказывает, что предложенный алгоритм является точным, эффективным, действенным и применимым к реальным графам в большей степени, чем традиционный алгоритм.

**Обсуждения**

Для улучшения исследований в этой области был предложен ряд следующих рекомендаций:

В работах, отобранных для включения в данный специальный выпуск, описан ряд методов науки о данных для извлечения знаний из образовательных данных. Однако извлеченные знания применимы только к рассматриваемой проблеме. Желательно получить общие модели, которые можно применять в других сценариях.

Большинство исследований сосредоточено на анализе только одного источника образовательных данных. Однако в современных «умных» классах регистрируется множество различных мульти-источников и мульти-модальных данных, и может быть очень интересно обьединить эти данные для получения более содержательных и достоверных моделей.

Многие подходы DS генерируют модели, которые трудно интерпретировать, несмотря на то, что они могут давать очень точные результаты. Однако интерпретируемость иногда является требованием в образовании, поскольку она помогает понять процессы обучения и, следовательно, улучшить их путем вмешательства.

Современные образовательные модели разработаны на основе принципа повсеместности, особенно в случае чрезвычайных ситуаций, подобных той, что вызвана пандемией Ковид-19. В этом сценарии учащийся должен уметь регулировать свое обучение, что иногда

бывает непросто. Очень важно рассчитывать на инструменты для персонализированного обучения, которые адаптируются к каждому ученику в зависимости от его эмоций в определенный момент. В настоящее время перспективным направлением является использование виртуальных аффективных агентов.

**Заключение**

В эту статью вошли 10 избранных статей, представляющих важные достижения в области Educational Data Science – Наука о данных в сфере образования. Отобранные статьи включают интересные исследования о развитии этой области, работы о перспективных технологиях и выдающихся исследований теорий и методов, которые будут играть решающую роль в будущем этой дисциплины.

В заключение мы даем рекомендации о том, какие возможности может использовать зарождающаяся область EDS, чтобы оказать большее влияние на образование. Первая такая возможность заключается в том, что область EDS должна опираться на богатый набор традиций, которые лежат в основе исследований в области образования; в частности, на традиции гуманистических и социальных наук, которые переживают схожий формат развития, когда наука о данных и большие данные входят в их сферы и революционизируют их. Слияние этих областей выявляет напряженность и успехи, которые, возможно, EDS может перенять. Одним из потенциальных подходов к интеграции науки о данных и образования является «состязательное сотрудничество». Аналогичным образом проекты в области образовательной науки о данных должны стремиться к тому, чтобы вобрать в себя лучшее из методологических традиций, присущих другим дисциплинам, наряду с их теоретическими и концептуальными традициями.

Эпистемологические методы экономистов порой вступают в противоречие с методами сообщества машинного обучения в области добычи данных [14]. Однако в высшем образовании методы EDS начинают открывать новые дополнительные перспективы в отношении институциональных данных и данных об уровне развития и достижений студентов. Поскольку эти данные становятся все более доступными, машинное обучение можно использовать для синтеза и формирования компетенций, через которые проходят студенты на протяжении своего обучения. Внедрение подходов науки о данных не означает, что области отказываются от заслуженных достижений и устоявшихся постулатов, то есть речь не идет о межобластной колонизации. Например, экспериментальные разработки из области экономики иногда могут быть использованы в качестве золотого стандарта для оценки эффектов вмешательств, основанных на EDS, в высшем образовании.

Другой пример: столетний опыт психометрических исследований предлагает множество подходов к моделированию, которые стоит использовать наряду с современными подходами машинного обучения. Психометрические подходы могут быть информативными для последующего развития методологии (с точки зрения особенностей, которые могут заслуживать внимания) и полезными в качестве эталонов для новых методов науки о данных. Такая перспектива показывает, что выигрыш от применения подходов машинного обучения зачастую относительно невелик (если он вообще существует).

Один из важнейших вопросов заключается в том, как наилучшим образом подготовить студентов к работе. С нашей точки зрения, необходимо четко сосредоточиться на проблемах, имеющих отношение к образованию и потенциально решаемых с учетом имеющихся у нас данных. Несмотря на относительный «взрыв» данных в образовании, у нас по-прежнему гораздо меньше данных (т. е. данных, богатых многими полями), чем в других областях, и это может ограничить применимость самых сложных алгоритмических подходов. Мы также должны работать над формированием у студентов этики ответственности. Хотя многие представители технологической сферы придерживаются этики «быстро двигаться и идти напролом», мы считаем, что такое отношение было бы крайне неуместным, учитывая характер образования (т. е. разнообразие заинтересованных сторон и осторожность, необходимую при решении вопросов, затрагивающих молодежь). Скорее, мы должны быть больше похожи на врачей с их мандатом Гиппократа (прежде всего, не навреди). Вопросы справедливости, например, не могут рассматриваться в конце, а должны быть главными с самого начала.

Вычислительные подходы, используемые в EDS, интересны тем, что они могут дать новое понимание старых проблем или позволить использовать новые виды данных и перспектив

в исследованиях в области образования. Однако эти данные и подходы не станут панацеей. Поведенческие науки в целом и наука об образовании в частности являются сложными. Следует ожидать, что большинство инноваций в области данных или вычислений приведут лишь к незначительному улучшению нашего понимания.

## СПИСОК ЛИТЕРАТУРЫ/REFERENCES


1. Klašnja-Milićević A., Ivanović M., Budimac Z. Data science in education: Big data and learning analytics. *Computer Applications in Engineering Education*, 2017, no. 25, pp. 1066–1078. DOI: https://doi.org/10.1002/cae.21844
2. Mitrofanova Y.S., Sherstobitova A.A., Filippova O.A. Modeling smart learning processes based on educational data mining tools. *Smart Education and e-Learning*, 2019, vol. 144, pp. 561–571. DOI: https://doi.org/10.1007/978-981-13-8260-4_49
3. Romero C., Ventura S. Educational data mining and learning analytics: an updated survey. *Wires Data Mining and Knowledge Discovery*, 2020, vol. 10, Iss. 3. DOI: https://doi.org/10.1002/widm.1355
4. Verma A., Singh A., Lughofer E., Xiaochun Cheng, Abualsaud Kh. Multilayered-quality education ecosystem (MQEE): an intelligent education modal for sustainable quality education. *Journal of Computing in Higher Education*, 2021, Iss. 3. Available at**:** https://www.springerprofessional.de/en/multilayered-quality-education-ecosystem-mqee-an-intelligent-edu/19404344 (accessed 21 May 2021).
5. Chango W., Cerezo R., Sanchez-Santillan M. Improving prediction of students' performance in intelligent tutoring systems using attribute selection and ensembles of different multimodal data sources. *Journal of Computing in Higher Education*, 2021, Iss. 33, pp. 614–634. DOI: https://doi.org/10.1007/s12528-021-09298-8
6. Manasa Devi M., Seetha M., Viswanadha Raju S. Automated text detection from big data scene videos in higher education: a practical approach for MOOCs case study. *Journal of Computing in Higher Education*, 2021, Iss. 33, pp. 581–613. DOI: https://doi.org/10.1007/s12528-021-09294-y
7. Sáiz-Manzanares M.C., Marticorena-Sánchez R., Rodríguez-Díez J.J. Improve teaching with modalities and collaborative groups in an LMS: an analysis of monitoring using visualisation techniques. *Journal of Computing in Higher Education*, 2021, Iss. 33, pp. 747–778. DOI: https://doi.org/10.1007/s12528-021-09289-9
8. Tsiakmaki M., Kostopoulos G., Kotsiantis S., Ragos O. Fuzzy-based active learning for predicting student academic performance. *Proc. of the 6th International Conference on Engineering & MIS 2020 (ICEMIS'20)*. New York, Association for Computing Machinery, 2020. Article No. 87, pp. 1–6. DOI: https://doi.org/10.1145/3410352.3410823
9. Pinargote-Ortega M., Bowen-Mendoza L., Meza J. Peer assessment using soft computing techniques. *Journal of Computing in Higher Education*, 2021, Iss. 33, pp. 684–726. DOI: https://doi.org/10.1007/s12528-021-09296-w
10. Beseiso M., Alzubi O.A., Rashaideh H. A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education*, 2021, Iss. 33, pp. 727–746. DOI: https://doi.org/10.1007/s12528-021-09283-1
11. Bekmanova G., Ongarbayev Y., Somzhurek B. Personalized training model for organizing blended and lifelong distance learning courses and its effectiveness in Higher Education. *Journal of Computing in Higher Education*, 2021, Iss. 33, pp. 668–683. DOI: https://doi.org/10.1007/s12528-021-09282-2
12. Mohammed A.H.K., Jebamikyous H.H., Nawara D. IoT text analytics in smart education and beyond. *Journal of Computing in Higher Education*, 2021, Iss. 33, pp. 779–806. DOI: https://doi.org/10.1007/s12528-021-09295-x
13. Raghda Fawzey Hriez, Ghazi Al-Naymat. A framework to capture the dependency between prerequisite and advanced courses in higher education. *Journal of Computing in Higher Education*, 2021, Iss. 33, pp. 1–38. DOI: 10.1007/s12528-021-09292-0
14. Aljawarneh S., Lara J.A. Data science for analyzing and improving educational processes. *Journal of Computing in Higher Education*, 2021, Iss. 33, pp. 545–550. DOI: https://doi.org/10.1007/s12528-021-09299-7


Дата поступления: 17.08.2021 г.

UDC 378
DOI 10.54835/18102883_2021_30_5

# DATA SCIENCE AS A TOOL FOR EDUCATIONAL PROCESS DEVELOPMENT

**Elmira R. Khairullina**,
Dr. Sc., dean, Faculty of Design and Software Engineering, Elm.
khair@list.ru

Kazan National Research Technological University,
68, K. Marx street, Kazan, 420015, Russia.

The latest trends in data science in education are reviewed and studied to consider current trends and contributions to the era of smart education. This includes a set of carefully peer-reviewed, world-class manuscripts that review and detail cutting-edge, framework and methodological data science research projects applied in education using various approaches such as information fusion, soft computing, machine learning, Internet of Things, etc. Based on this systematic review, we have formulated some recommendations and suggestions for researchers, practitioners and scientists to improve the quality of research in this area.

**Key words:** Pedagogical technology, data science, data science technology, educational systems.

Received: 17 August 2021.