Высшее образование в России Vysshee obrazovanie v Rossii = Higher Education in Russia

ISSN 0869-3617 (Print), ISSN 2072-0459 (Online) http://vovr.elpub.ru

Прогностическое моделирование в высшем образовании: определение факторов академической успеваемости

Научная статья
DOI: 10.31992/0869-3617-2023-32-1-51-70

Гафаров Фаиль Мубаракович – канд. физ.-мат. наук, доцент, зав. кафедрой, ORCID: 0000-0003-4704-154X, fgafarov@yandex.ru

Руднева Яна Борисовна – канд. ист. наук, главный специалист Центра координации образовательных проектов, ORCID: 0000-0002-0068-4890, ya.rudneva@rambler.ru

Шарифов Умар Юсуфович – магистр, ORCID: 0000-0003-0840-7502, uysharifov@gmail.com
Казанский (Приволжский) федеральный университет, Казань, Россия
Адрес: 420000, г. Казань, ул. Кремлёвская, 35, Российская Федерация

Аннотация: *Несколько десятилетий в области интеллектуального анализа данных в образовании (EDM) прогнозирование успеваемости остаётся одной из самых популярных и обсуждаемых на международном уровне исследовательских тем. В частности, интеллектуальный анализ данных используется для прогнозирования таких образовательных результатов, как успеваемость, удержание, успех, удовлетворённость, достижения и процент отсева. В управленческой практике высших учебных заведений на основе оперативного прогноза разрабатываются и реализуются меры поддержки тех студентов, которые попадают в группу риска.*

Наше исследование направлено на обоснование модели прогнозирования досрочного выбытия студентов с использованием искусственной нейронной сети и анализ предикторов, повышающих точность прогнозировании успешного окончания российского университета. Эта работа позволит расширить международную практику компаративных исследований в высшем образовании.

В работе подтверждены уже существующие гипотезы о влиянии ряда факторов на прогнозирование академической успеваемости и выдвинуто предположение о необходимости проверки их универсальности или специфичности в конкретном высшем учебном заведении. Мы также доказали, что модель искусственной нейронной сети с определённым набором атрибутов может применяться в контексте отдельного высшего учебного заведения, независимо от специализации. Для определения потенциальной группы риска учащихся используется модель прогнозирования бинарной классификации. Общая точность прогноза нейронной сети с комбинированными данными достигает 88%. Для данной модели нейронной сети базовыми предикторами, влияющими на точность прогноза, являются совокупный средний уровень успеваемости (CGPA) и год поступления в университет.

Ключевые слова и фразы: *образовательная аналитика, факторы досрочного выбытия студентов, интеллектуальный анализ данных, искусственные нейронные сети, прогнозирование*

Для цитирования: *Гафаров Ф.М., Руднева Я.Б., Шарифов У.Ю.* Прогностическое моделирование в высшем образовании: определение факторов академической успеваемости // Высшее образование в России. 2023. Т. 32. № 1. С. 51–70. DOI: 10.31992/0869-3617-2023-32-1-51-70

Predictive Modeling in Higher Education: Determining Factors of Academic Performance

Original article
DOI: 10.31992/0869-3617-2023-32-1-51-70

Fail M. Gafarov – Cand. Sci. (Physics and Mathematics), Assoc. Prof., Head of Department. ORCID: 0000-0003-4704-154X, fgafarov@yandex.ru

Yana B. Rudneva – Cand. Sci. (History), Senior Specialist of the Centre for Coordination of Educational Project, ORCID: 0000-0002-0068-4890, ya.rudneva@rambler.ru

Umar Yu. Sharifov – Graduate Student, ORCID: 0000-0003-0840-7502, uysharifov@gmail.com
Kazan (Volga region) Federal University, Kazan, Russia
Address: 35, Kremlyovskaya str., Kazan, 420000, Russian Federation

Abstract. For several decades in the field of data mining in education (EDM), predictive learning has remained one of the most popular and internationally discussed research topics. Specifically, data mining is used to predict educational outcomes such as academic performance, retention, success, satisfaction, achievement and dropout rates. In the management practice of higher education institutions, on the basis of an operational forecast, measures are developed and implemented to support those students who fall into the risk group.

Our study is aimed at substantiating a model for predicting the early departure of students using an artificial neural network and analyzing predictors that increase the accuracy of predicting successful graduation from a Russian university. This work will expand the international practice of comparative research in higher education.

The paper confirms the already existing hypotheses about the influence of a number of factors on the prediction of academic performance and suggests the need to test their universality or specificity in a particular institution of higher education. We also proved that an artificial neural network model with a certain set of attributes can be applied in the context of a single higher education institution, regardless of specialization. To determine the potential risk group of students, a binary classification prediction model is used. The overall prediction accuracy of a neural network with combined data reaches 88%. For this neural network model, the basic predictors that affect the accuracy of the forecast are the cumulative average level of achievement (CGPA) and the year of admission to the university.

Keywords: educational analytics, student dropout factors, data mining, artificial neural networks, forecasting

Cite as: Gafarov, F.M., Rudneva, Ya.B., Sharifov, U.Yu. (2023). Predictive Modeling in Higher Education: Determining Factors of Academic Performance. *Vysshee obrazovanie v Rossii = Higher Education in Russia.* Vol. 32, no. 1, pp. 51-70, doi: 10.31992/0869-3617-2023-32-1-51-70 (In Russ., abstract in Eng.)

Введение

С 2020 года искусственный интеллект в образовании (AIEd) экспертами международного консорциума New Media определяется в качестве одной из ключевых технологических тенденций в сфере высшего образования[1]. Об этом свидетельствует достаточно высокий уровень научной публикационной активности в области исследования потенциальных возможностей применения искусственного интеллекта в высшем образовании, который сконцентрирован на четырёх основных направлениях: а) адаптивные системы и персонализация, б) оценка, в) профилирование и прогнозирование, г) интеллектуальные обучающие системы (ITS) [1].

Активное использование приложений AIEd для сбора, интерпретации и понимания больших систем образовательных данных является неотъемлемой частью интеллектуального анализа данных в образовании (EDM). Его практическое применение направлено на выявление скрытых закономерностей, которые необходимо учитывать в принятии стратегических решений на основе данных [2; 3]. Несколько десятилетий прогнозирование успеваемости остаётся одной из самых популярных и обсуждаемых исследовательских тем в области интеллектуального анализа данных. В частности, EDM используется для прогнозирования таких образовательных результатов, как успеваемость, удержание, успех, удовлетворённость, достижения и процент отсева [4]. В управленческой практике на основе оперативного прогноза разрабатываются и реализуются меры поддержки тех студентов, которые попадают в группу риска [1].

В России интеллектуальный анализ образовательных данных только начинает развиваться как отдельное направление. Среди причин можно выделить: медленные темпы развития онлайн-обучения, отсутствие большого массива образовательных данных, отсутствие практики принятия административных и педагогических решений на основе данных, ограниченное применением AIEd.

Эти причины оказывают прямое влияние на незначительно количество институциональных исследований российского высшего образования в области интеллектуального анализа данных, модели которого используются в качестве основы для автоматической адаптации, проводимой компьютерной системой [5]. Ещё одной причиной отсутствия интереса российских исследователей к разработке моделей прогнозирования с использованием методов AIEd может являться низкий, по сравнению с североамериканскими и европейскими университетами, уровень выбытия студентов из российских высших учебных заведений – 22% [6].

Есть ряд содержательных российских исследований с применением статистических алгоритмов, в которых анализируются связи между различными факторами и их влияние на успеваемость и удержание [6–11]. Однако данные по российскому высшему образованию практически не представлены в международном исследовательском поле, что осложняет развитие кросскультурных (компаративных) исследований, которые путём сравнения могут выявить общие и особенные тенденции в развитии национальных систем высшего образования.

Исследовательская работа, результаты которой представлены в данной статье, направлена на разработку модели прогнозирования досрочного выбытия студентов с использованием искусственной нейронной сети и анализ предикторов, повышающих точность прогнозировании успешного окончания студентами российских высших учебных заведений.

В работе для обучения и тестирования нейронной сети были использованы данные выпускников и досрочно выбывших студентов Казанского (Приволжского) феде-

[1] 2020 EDUCAUSE Horizon Report // Teaching and Learning Edition. URL: https://library.educause.edu/resources/2020/3/2020-educause-horizon-report-teaching-and-learning-edition (дата обращения: 18.07.2022).

рального университета за четыре полных цикла обучения в университете: 2012–2016, 2013–2017, 2014–2018, 2015–2019. Для определения потенциальной группы риска среди обучающихся нами была использована модель прогнозирования бинарной классификации. Общая точность прогноза нейронной сети с комбинированными данными составила 88%. Для данной модели нейронной сети базовыми предикторами, влияющими на точность прогноза, являются совокупный средний уровень успеваемости (CGPA) и год поступления в университет.

В нашем исследовании нашли подтверждение уже существующие гипотезы о влиянии ряда факторов на прогнозирование академической успеваемости и было выдвинуто предположение о необходимости проверки их универсальности или специфичности в конкретном высшем учебном заведении. Мы также доказали, что модель искусственной нейронной сети с определённым набором атрибутов может применяться в контексте отдельного высшего учебного заведения, независимо от специализации.

Прогностическое моделирование в высшем образовании с использованием искусственных нейронных сетей (обзор)

Систематический обзор научной публикационной активности показывает явно растущее количество исследований в области прогнозирования успеваемости, а также растущее разнообразие используемых алгоритмов машинного обучения [12]. Исследования в этой области направлены на выявление функций, которые можно использовать для прогнозирования, на определение алгоритмов, которые могут улучшить прогнозы, и на количественную оценку аспектов успеваемости студентов. В большинстве исследований изучается общая точность, чувствительность и прецизионность моделей.

Одним из эффективных алгоритмов, применяемых в области анализа данных в образовании, в аналитических обзорах методов машинного обучения признаны нейронные сети [13]. В качестве основы для построения прогностической модели нами были использованы полносвязные трёхслойные нейронные сети с прямыми связями. По сути, большая часть исследований основана на классической архитектуре искусственной нейронной сети с прямыми связями, с одним или двумя скрытыми слоями. Полносвязная нейронная сеть с 11 входными переменными, двумя слоями скрытых нейронов и одним выходным слоем достигла точности прогнозирования 84,8% в задаче прогнозирования успеваемости учащихся [14]. В ещё одной работе [15] представлен нейросетевой подход для прогнозирования досрочного выбытия учащихся или удлинения периода их обучения на основе девяти категориальных и числовых входных переменных (рейтинг средней школы, качество средней школы, стандартизованные результаты тестов (по математике и английскому языку), оценки преподавателей средней школы, оценка внеклассной деятельности, уровень образования отца, уровень образования матери и время, прошедшее после окончания средней школы). Нейронная сеть с прямой связью с 50 скрытыми нейронами и функциями активации скрытых нейронов с гиперболическим тангенсом смогла точно предсказать успешность выпуска и достигла лучших показателей с точностью классификации, превышающей 95%.

С помощью нейронной сети с прямой связью исследовалась сложная нелинейная связь между когнитивными и психологическими переменными, влияющими на академическую успеваемость учащихся средней школы [16]. Интересный подход с использованием метода переноса обучения предложен для задачи прогнозирования успеваемости студентов на курсах бакалавриата [17]. Авторы использовали метод переноса обучения для обучения глубокой нейронной сети на наборе данных прошлого курса и повторно использовали его в качестве отправной точки для дообучения на наборе данных нового связанного курса.

Исследователем из Саудовской Аравии Х.А. Менгашем были разработаны четыре модели прогнозирования с использованием четырёх хорошо известных методов интеллектуального анализа данных (искусственная нейронная сеть, дерево решений, метод опорных векторов и наивный байесовский метод) для прогнозирования успеваемости абитуриентов перед их приёмом [18]. В этом сравнительном исследовании модель на основе искусственных нейронных сетей достигла уровня точности около 79,22%. Автор также отмечает, что после применения новой системы вступительного отбора, основанного на результатах этого исследования, количество студентов, получивших отличные или очень хорошие оценки за первый год обучения, увеличилось на 31%.

В российской научно-исследовательской среде также возрастает интерес к прогностическому моделированию академической успеваемости с использованием нейронных сетей [19; 20]. В 2020 году в Екатеринбурге состоялась Международная научно-практическая конференция «Цифровизация образования: история, тенденции и перспективы DETP 2020», в рамках которой обсуждались практики российских исследователей по применению нейронных сетей для обработки образовательных данных [21–23; 32].

Стоит отметить, что в последние годы наметилась тенденция перехода от разработки моделей прогнозирования успешности студентов, не зависящих от конкретного учебного контекста, к исследованиям по прогнозированию результатов по одной дисциплине или курсу [24]. В первом случае исследователи получали несогласованные и противоречивые результаты при переносе моделей с одного курса на другой, во втором – ограничение возможности обобщения результатов. И в том, и в другом случае становится актуальной проблема переноса моделей прогнозирования в разные контексты и создания унифицированной модели, подходящей для тиражирования.

Предикторы академической успеваемости, выявленные с использованием статистических методов и методов машинного обучения (обзор)

Теоретические концепты о факторах досрочного выбытия студентов формировались в рамках психологии образования и социологии образования. Среди основных концепций нужно упомянуть теорию культурного капитала Й. Бергера, культурную теорию Г. Ку и П. Лава, теорию интеграции В. Тинто [6], модель саморегулируемого обучения Ф. Винна и Э. Хэдвин [4]. Теоретические модели устойчивых паттернов поведения студентов учитывают внутренние и внешние условия регуляция обучения и, следовательно, успеваемости. К внутренним относятся социально-психологические характеристики студентов, к внешним – образовательная среда (инфраструктура, кадры, академическая культура, образовательная политика) [25–29]. Среди часто упоминаемых факторов, влияющих на прогноз академической успеваемости учащихся, указываются предыдущие академические достижения, демографические характеристики учащихся, активность электронного обучения, психологические характеристики и окружающая среда [4].

Однако в области интеллектуального анализа данных в образовании до сих пор остаются дискуссионными вопросы о методах обработки данных, релевантных для построения прогностических моделей, о взаимосвязях факторов, которые определяют учебную производительность. В частности, ряд исследователей утверждают, что комбинированные (демографические и институциональные) данные дают лучшие результаты, чем данные из одной категории [4]. Их оппоненты указывают на то, что ограничение данных, используемых для классификатора, может дать лучшие результаты даже при меньшем размере выборки [30].

Исследование, представленное в этой статье, направлено на дальнейшее изучение предикторов, определяющих успешное оконча-

ние университета, и повышающих точность прогнозирования учебной производительности. В частности, на основе анализа более ранних исследований мы определили ряд параметров для включения в модель прогнозирования, чтобы эмпирически 1) подтвердить или опровергнуть наличие переменных, не зависящих от учебного контекста и особенностей национального образования, а также 2) исследовать новые взаимосвязи для дополнения существующих. Поэтому здесь мы отдельно остановимся на обзоре уже констатируемых и доказанных связей между факторами, влияющими на академическую успеваемость.

Связь между баллами на вступительных/выпускных экзаменах и успеваемостью в университете. В зарубежных исследованиях указываются два основных фактора, а именно предыдущая академическая успеваемость и демографические данные студентов (69% научных работ). Более 40% исследований указывают в качестве важного фактора предыдущую академическую успеваемость, в том числе результаты средней школы, стандартизированного выпускного или вступительного теста [4]. Многие исследователи считают, что именно первый год обучения является определяющим для успеваемости на всех последующих курсах, доказывая устойчивую связь между результатами средней школы и производительностью в первый год обучения в высшем учебном заведении.

Вместе с тем исследователи фиксируют устойчивую связь между академической успеваемостью и баллами на вступительных экзаменах. Средняя оценка коэффициента корреляции между баллами по вступительным тестам в североамериканские университеты (SAT и ACT) и показателем академической успешности студента за весь период учёбы находится в интервале от 0,35 до 0,46 с учётом стандартной ошибки. Таким образом, вступительные экзамены предсказывают 12–25% вариации оценок в вузе [31].

В России в высшие учебные заведения приём производится на конкурсной основе по результатам стандартизированных тестов (100-балльная шкала оценивания) по предметным областям – Единый государственный экзамен (ЕГЭ). Чем выше балл ЕГЭ, тем выше шанс попасть в престижное учебное заведение [10]. Для зачисления на образовательную программу необходимо сдать не менее трёх стандартизированных тестов – один обязательный тест на знание государственного (русского) языка и два – по выбору учебного заведения. Как правило, российские университеты устанавливают в качестве вступительных экзаменов один тест по основной дисциплине специализации и один тест по профильной дисциплине. Общий балл ЕГЭ для зачисления определяется путём суммирования результатов трёх испытаний.

В отличие от SAT и ACT в ЕГЭ совмещены выпускной и вступительный экзамены, он также выполняет функцию школьной оценки [11]. Совместный учёт результатов вступительных экзаменов и средней школьной оценки в одной модели значительно повышает успешность предсказания академических достижений. Российскими исследователями с применением статистических алгоритмов показано, что от 13 до 30% успеваемости на 1-м курсе объясняются суммарным баллом ЕГЭ. ЕГЭ по математике и русскому языку почти на всех направлениях оказываются лучшими предикторами успеваемости, чем ЕГЭ по профильным предметам.

Несмотря на доказанное прогностическое влияние прошлой успеваемости на успеваемость в университете в предыдущем нашем исследовании на данных 14 724 учащихся, окончивших Казанский (Приволжский) федеральный университет, мы обнаружили, используя несколько методов машинного обучения, что точность прогнозирования при удалении такой переменной, как ЕГЭ, снижается несущественно [32]. Таким образом, наше предыдущее исследование подтвердило гипотезу о том, что наряду с результатами ЕГЭ в качестве предикторов учебной производительности в университете

необходимо использовать дополнительные переменные.

Средние показатели успеваемости в высшем учебном заведении (GPA и CGPA). Академическая успеваемость основана на среднем показателе успеваемости (GPA) или совокупном среднем уровне успеваемости (CGPA), которые представляют собой системы оценок, используемых в зарубежных высших учебных заведениях для определения шкалы оценок студентов. GPA рассчитывается на один семестр или на год, тогда как CGPA рассчитывается на всю продолжительность курса.

За последние пять лет в исследованиях, направленных на прогнозирование академической успеваемости, можно наблюдать два основных подхода. Классификация, в которой CGPA является целевой переменной для задачи многоклассовой или бинарной классификации (буквенная оценка, или общий рейтинг, или категория бинарного класса «прошёл / не прошёл»). Что касается другого подхода, это регрессия, при которой прогнозируется численное значение CGPA [4].

Для измерения учебной производительности в нашем исследовании использовались количественные значения GPA1, GPA2 и CGPA в 100-балльной шкале. Значение CGPA рассчитывалось как среднее значение успеваемости за весь период обучения путём деления общей суммы набранных баллов по всем дисциплинам с учётом учебной производительности в семестре и итоговой оценки (max=100) на общее количество оценённых дисциплин. Аналогично рассчитывались значения GPA1 (средний показатель успеваемости за первый семестр обучения) и GPA2 (средний показатель успеваемости за второй семестр обучения). Значения GPA1 и GPA2 применялись в модели для проверки гипотезы о том, что первый год учёбы является определяющим для успеваемости на всех последующих курсах.

Цель нашего исследования состоит не в том, чтобы установить точное значение успеваемости каждого учащегося, а в том, чтобы предложить модель бинарного прогнозирования (обучение / досрочное выбытие) для определения потенциальной группы риска учащихся. Это окажет поддержку административным центрам высших учебных заведений в разработке дифференцированных действий по отношению к неоднородным группам студентов, а также более эффективному распределению ресурсов [29].

Оплата обучения. В российских университетах специализации, востребованные в экономике, бюджетируются государством. Бюджетное место гарантирует студенту отсутствие оплаты за обучение и выплату финансовой поддержки в виде стипендии за академическую успешность. Количественное распределение бюджетных мест по специализациям устанавливается государственным заказом. Приоритет зачисления на бюджетные места получают студенты, имеющие высокие баллы ЕГЭ. Другая часть студентов самостоятельно оплачивает обучение в университете, не получая финансовой поддержки. При обучении с полной занятостью и с самостоятельной платой за обучение студент либо подрабатывает, либо ищет финансовую поддержку у родителей.

Д. Делён пришёл к выводу, что наиболее важными факторами отсева учащихся, кроме их прошлых и настоящих академических достижений, является получение финансовой помощи на оплату обучения [33]. Чтобы проверить эту гипотезу, в нашем исследовании мы используем фактор оплаты за обучение как один из базовых предикторов.

Специализация. Диапазон отсева сильно колеблется в зависимости от специализации: отсев студентов на специальностях STEM, включая инженерные и технологические, значительно выше, чем на других специальностях, и достигает, например, в США 60% [34]. Эта проблема выходит за рамки национальных систем образования и является мировой тенденцией. Это подтверждает обзор публикационной активности, в котором показано, что большинство дисциплин, задействованных в статьях об искусственном

интеллекте в образовании, происходят из компьютерных наук и STEM [1].

Исследования факторов отсева и удержания на специальностях STEAM часто имеют схожие результаты и демонстрируют, что показатели предварительной подготовки, такие как средний балл средней школы (HSGPA) и баллы ACT или SAT, в сочетании с CGPA, дают статистически значимые устойчивые модели [34]. На примере инженерных и технологических специальностей доказана значимость средней оценки успеваемости и такого фактора как время, затраченное на получение степени [29]. Целый ряд исследователей сосредоточились на прогнозировании успеваемости по физике как дисциплине, вызывающей проблемы у студентов [30; 34; 35], и других дисциплинах STEM, которые преподаются на первом курсе [36].

В российских высших учебных заведениях в первые 2,5 года обучения доля студентов, отчисливших с инженерных направлений подготовки (25%), значительно превышает аналогичный показатель на других направлениях (19%) [37]. Российские исследования показывают, что факторами отсева могут быть рейтинг университета (элитный/массовый), низкий балл ЕГЭ по математике, слабая мотивация и спад интереса к специальностям STEM [37; 38].

В нашем исследовании на данных разнопрофильного федерального университета специализация будет рассматриваться как важная переменная, влияющая на общую точность и чувствительность модели нейронной сети. Чтобы доказать или опровергнуть влияние школьной подготовки по STEAM дисциплинам на риск досрочного выбытия из российского университета, в качестве переменных мы исследуем результаты по каждому из стандартизированных тестов ЕГЭ.

Набор данных

Для такого классификатора, как нейронная сеть, имеет значение полный набор данных, который не содержит пропущенных значений и ошибок [4], поэтому на первом этапе исследования были созданы (с использованием языка программирования Python) программные модули для очистки данных от неполной, недостоверной или ошибочной информации и для выбора обучавшихся с аналогичными характеристиками. Далее данные собирались таким образом, чтобы информация по заданным параметрам (предикторам), которая содержится в университетском хранилище данных, агрегировалась в уникальной записи каждого выпускника / досрочно выбывшего.

База исследуемых данных формировалась в соответствии со следующими ограничениями: (1) выпускники / досрочно выбывшие Казанского (Приволжского) федерального университета, (2) базовое высшее образование (бакалавриат), (3) наличие результатов ЕГЭ по трём дисциплинам и (4) наличие полных данных об учебной успеваемости/неуспеваемости в университете за полный четырёхлетний цикл обучения в информационной системе университета. База содержит данные о выпускниках / досрочно выбывших в четырёх полных циклах обучения в университете: 2012–2016, 2013–2017, 2014–2018, 2015–2019.

Для обобщённого описания когорты студентов были использованы термины «graduate» и «attrition»: «graduate» – это доля студентов, переходящих с одного года обучения на другой, «attrition», соответственно, означает студентов, не перешедших на следующий год обучения, то есть выбывших до полного срока окончания образовательной программы [6].

В исследовании для проверки гипотезы о возможности переноса модели в разные контексты мы использовали три выборки предварительно очищенных и сбалансированных данных: data set 1_ University (все направления подготовки), data set 2_ Management, Economics and Finance, data set 3_ Engineering, Technology and Computer Science. Под контекстом в данном случае понимается интенсивность досрочного выбытия студентов в зависимости от направления подготовки. Для

Таблица 1

Репрезентативность выборочной совокупности

Table 1

Representativeness of the sample

	Data set 1_ University	Data set 2_ Management, Economics and Finance	Data set 3_ Engineering, Technology and Computer Science
Генеральная совокупность / выборочная совокупность	43 677/26 910	9513/6902	9614/6550
Из них:			
«Graduate», чел.	34 908/21 095	8216/5716	6651/4802
«Attrition», чел.	8769/5815	1297/1186	2963/1748

Казанского (Приволжского) федерального университета характерна общероссийская тенденция: отсев студентов на специальностях STEM, включая инженерные и технологические (data set 3_ Engineering, Technology and Computer Science), значительно выше, чем на других специальностях. В то же время отсев по направлениям подготовки, связанным с экономикой, финансами и управлением, является одним из самых низких (data set 2_ Management, Economics and Finance). Таким образом, были сформированы два дополнительных набора данных с различной динамикой досрочного выбытия студентов.

В data set 1_ University (все направления подготовки) были созданы 26 910 уникальных записей «graduate»/ «attrition». Принадлежность к специализации кодировалась в виде укрупнённых блоков: 1) медицина, биология (855 человек); 2) естественные науки (1992 человек); 3) гуманитарные, поведенческие, социальные науки и образование (4464 человек); 4) филология и журналистика (3520 человек); 5) менеджмент, экономика и финансы (6902 человек) 6) инженерные и компьютерные науки (6550 человек), 7) юриспруденция (2627 человек). Данные выборки о специализации отражают пропорции распределения специализаций в университете. Кроме того, данная выборка уравновешена по циклам обучения (в среднем 20–25% уникальных записей «graduate»/ «attrition» на каждый цикл).

Под генеральной совокупностью мы понимаем общую численность «graduate»/ «attrition» Казанского (Приволжского) федерального университета за четыре полных цикла обучения в университете: 2012–2016, 2013–2017, 2014–2018, 2015–2019. В data set 1_ University (все направления подготовки) выборочная совокупность полностью очищенных данных составила 61,6% от генеральной совокупности, из них выпускников 60,4%, выбывших досрочно – 66,3%. В data set 2_ Management, Economics and Finance выборочная совокупность полностью очищенных данных составила 72,5% от генеральной совокупности, из них выпускников 69,6%, выбывших досрочно – 91,4%. В data set 3_ Engineering, Technology and Computer Science выборочная совокупность полностью очищенных данных составила 68,1% от генеральной совокупности, из них выпускников 72,2%, выбывших досрочно – 59%. Таким образом, выборочную совокупность по всем трём наборам данных в когортах «graduate»/ «attrition» можно считать репрезентативной (*Табл. 1*).

Предварительный анализ полной неочищенной базы данных «attrition» (8769 человек за указанный период) показал, что в Казанском (Приволжском) федеральном университете доминируют четыре причины досрочного выбытия: «академическая неуспеваемость» (44%), «по собственному желанию» (27%), «в связи с расторжением договора на обучение по неуважительной причине» (11%), «в связи с переводом в другое образовательное учреждение» (9%), 8% приходятся на другие причины (здоровье, дисциплинарные нарушения и т. д.). Данные причины классифицируются в административных целях. Более детальный анализ CGPA выявил значительное число студентов с низкими баллами успеваемости в категориях «по собственному желанию»

Таблица 2

Характеристика выборочной совокупности

Table 2

Characteristics of the sample

	Data set 1_ University	Data set 2_ Management, Economics and Finance	Data set 3_ Engineering, Technology and Computer Science
Доля в выборке студентов, обучающихся в условиях полной занятости (очное обучение), %	96	94	99
Доля в выборке студентов, самостоятельно оплачивающих обучение по договору (контрактная форма), %	56	73	28
Гендерное распределение в выборке, %	мужчины – 38 женщины – 62	мужчины – 33 женщины – 67	мужчины – 63 женщины – 37
Распределение в выборке по предыдущему образованию, %:			
Лицей, гимназия	28	25	31
Школа	67	68	64
Второе профессиональное образование	5	7	5

и «в связи с расторжением договора на обучение по неуважительной причине». Нами было принято решение для сохранения когорты «attrition» не исключать данные студентов, отчисленных по этим причинам.

Данные *таблицы 2* подтверждают, что Data set 2_ Management, Economics and Finance и Data set 3_ Engineering, Technology and Computer Science не равновесны ещё по двум признакам: гендерная принадлежность и бюджетная/контрактная форма обучения. Это создаёт дополнительные условия для проверки гипотезы о возможности создания универсальной модели прогнозирования с использованием искусственной нейронной сети без существенной потери точности на различных направлениях подготовки. Нейронная сеть, обученная и протестированная на выборке Data set 1_ University, была использована для набора данных Data set 2_ Management, Economics and Finance и Data set 3_ Engineering, Technology and Computer Science.

Модель искусственной нейронной сети

Нейронные сети являются одним из наиболее сложных и ресурсоёмких методов машинного обучения, которые используются как в задачах регрессии, так и в задачах классификации и кластеризации. Имеются различные архитектуры нейронных сетей. В настоящее время наиболее часто используемыми являются нейронные сети с прямыми связями, рекуррентные, свёрточные и графовые нейронные сети. Вне зависимости от архитектуры всей сети отдельные вычислительные элементы (нейроны) нейронных сетей выполняют взвешенное суммирование входного вектора данных и применяют к этой сумме нелинейную функцию, которая называется функцией активации нейрона. Основой обучения моделей, построенных на нейронных сетях, является алгоритм обратного распространения ошибки. Это итерационный алгоритм, с помощью которого пошагово выполняется настройка всех параметров (весов и смещений) всех нейронов сети. В задачах, связанных с обработкой табличных данных, в основном используются нейронные сети с прямыми связями, поэтому мы выбрали именно этот тип. Количество входов нейронной сети (размерность первого слоя) всегда равна количеству входных параметров. Между входным и выходным слоем нейронная сеть может содержать один или несколько скрытых слоёв. В нашей работе мы использовали трёхслойную нейронную сеть с 46 нейронами на входном слое, с функцией

Таблица 3

Характеристики используемой искусственной нейронной сети (ИНС)

Table 3

The parameters of used neural network

	Входной слой	Скрытый слой	Выходной слой
Функция активации	ReLU	ReLU	Sigmoid
Количество нейронов	46	10	1
Оптимизатор	Adam		
Количество эпох при обучении	1000		
Функция потерь	Бинарная перекрёстная энтропия		

активации ReLU (линейный выпрямитель). Скрытый слой содержит 10 нейронов также с функцией активации ReLU (*Табл. 3*). В данной работе мы решаем задачу бинарной классификации, поэтому на выходе имеется только один нейрон с сигмоидальной функцией активации. Нейронная сеть обучалась в течение 1000 эпох с использованием оптимизатора Adam, функция потерь – бинарная перекрёстная энтропия.

Следующим шагом стал отбор данных для обучения сети. Изначальный набор данных был не сбалансирован, так как количество данных, относящихся к когорте «graduate», оказалось намного больше, чем в «attrition». Этот недостаток может негативно повлиять на корректность оценки качества обученной нейронной сети [4]. Поэтому для каждого отдельного эксперимента по обучению нейронной сети подготавливались сбалансированные наборы данных из общей выборки. По когорте «attrition» мы каждый раз включали весь набор (5815 человек), а для «graduate» каждый раз случайным образом брали выборку с количеством данных, равным количеству данных «attrition». Таким образом, общий набор данных, использованных в каждом эксперименте, по двум когортам составил 11 630 человек (для Data set 1_ University), из которых 50% «graduate» и 50% «attrition». На этапе обучения нейронной сети производилось деление исходной выборки на два подмножества: обучающая выборка (training set), которая используется для обучения нейронной сети, и тестовая выборка (test set). На основе тестовой выборки вычислялись характеристики, описывающие её точность (Accuracy, Precision, Recall, Specificity, F1).

Далее данные из 11 630 уникальных записей (Data set 1_ University) были рандомно распределены в training set (70% данных) и test set (30% данных). Данная процедура (подготовка сбалансированного набора данных, обучение нейронной сети, вычисление характеристик точности) проводилась 10 раз, и далее вычислялись средние значения характеристик, описывающих точность по всем 10 экспериментам, для достижения статистической достоверности результатов.

Благодаря данному алгоритму все имеющиеся данные по «graduate» и «attrition» были использованы в экспериментах по обучению нейронных сетей по всем трём наборам данных: Data set 1_ University, Data set 2_ Management, Economics and Finance и Data set 3_ Engineering, Technology and Computer Science. Общий сбалансированный набор данных Data set 2_ Management, Economics and Finance составил 2372 человек, Data set 3_ Engineering, Technology and Computer Science – 3496.

Количественные и качественные предикторы (переменные) досрочного выбытия

Для проверки выдвинутых гипотез в обучении и тестировании модели нейронной сети мы использовали четыре группы количественных и качественных факторов: 1) факторы, связанные с университетом, – общий средний балл успеваемости за все полные семестры, общий средний балл успеваемости за 1-й и 2-й семестр, год поступления в университет, специализация, форма обучения (очная/заочная); 2) факторы, связанные

Таблица 4

Количественные и качественные факторы, используемые в качестве входных переменных искусственной нейронной сети

Table 4

Quantitative and qualitative factors used as neural network input variables

Предиктор	Описание	Количественная/ Категориальная
CGPA	Общий средний балл успеваемости за все полные семестры	100-балльная шкала
GPA1	Общий средний балл успеваемости за 1-й семестр	100-балльная шкала
GPA2	Общий средний балл успеваемости за 2-й семестр	100-балльная шкала
ЕГЭ	Общий средний балл по трём вступительным/выпускным стандартизированным тестам	100-балльная шкала
ЕГЭ1	Балл стандартизированного вступительного/выпускного теста по основной учебной дисциплине	100-балльная шкала
ЕГЭ2	Балл стандартизированного вступительного/выпускного теста по профильной учебной дисциплине	100-балльная шкала
ЕГЭ3	Балл стандартизированного вступительного/выпускного теста по русскому языку	100-балльная шкала
ЕГЭ1_ база	Основная учебная дисциплина	Категориальная
ЕГЭ2_ профиль	Профильная учебная дисциплина	Категориальная
Предыдущий уровень образования	Школа Лицей/гимназия Второе профессиональное образование	Категориальная
Год поступления	Дата начала обучения в университете	Категориальная
Специализация	Направление подготовки	Категориальная
Форма обучения	Очная/заочная	Категориальная
Оплата	Самостоятельная оплата обучения (контракт)/ за счёт поддержки государства (бюджет)	Категориальная
Пол	Мужской/Женский	Категориальная

с доуниверситетским обучением, – общий средний балл по трём вступительным/выпускным стандартизированным тестам ЕГЭ, баллы по каждому стандартизированному вступительному/выпускному тесту ЕГЭ, по основной и профильной учебным дисциплинам ЕГЭ, уровень предыдущего образования; 3) пол и 4) оплата обучения (*Табл. 4*).

Для обучения и тестирования модели нейронной сети категориальные параметры подавались one-hot-кодированием, количественные – сырыми баллами (100-балльная шкала).

Результаты

Оценка качества и универсальности модели искусственной нейронной сети. В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются определённые метрики. Перед переходом к самим метрикам необходимо ввести важную концепцию для описания этих метрик в терминах ошибок классификации – confusion matrix (матрица ошибок).

Допустим, что у нас есть два класса и алгоритм, предсказывающий принадлежность каждого объекта одному из классов, тогда матрица ошибок классификации будет выглядеть следующим образом (*Табл. 5*).

В библиотеке sklearn.metrics существует метод – confusion_matrix, который возвращает матрицу ошибок. На вход подаётся два вектора – вектор фактических значений и вектор классифицированных значений. С помощью матрицы ошибок были выявлены результаты основных метрик оценки качества нейронной сети в задачах классификации: Accuracy, Recall, Precision, Specificity, F1 Score.

Модель на Data set 1_ University демонстрирует высокую общую точность прогноза (Accuracy = 0,88), которая практически совпадает со средневзвешенным значением точности и отзыва (F1 Score = 0,87) из-за

Таблица 5

Матрица ошибок

Table 5

Confusion matrix

Категория		Экспертная оценка	
		Положительная	Отрицательная
Оценка системы	Положительная	Истинно-положительный (True Positive – TP)	Ложно-положительный (False Positive – FP)
	Отрицательная	Ложно-отрицательный (False Negative – FN)	Истинно-отрицательный (True Negative – TN)

Таблица 6

Показатели производительности обученной ИНС на тестовом наборе данных

Table 6

Trained neural network's performance metrics on the test neural network

Показатели качества	Значения
Accuracy	0,88
Recall	0,83
Precision	0,91
Specificity	0,93
Fall-out (FPR)	0,07
F1 Score	0,87

сбалансированного набора данных в training set и test set. F1 Score учитывает как ложные срабатывания, так и ложные отрицательные результаты. Доля найденных объектов искомого класса («attrition») к общему числу объектов класса («attrition») составила 0,83 (Recall). Более низкие значения Recall относительно остальных показателей можно объяснить наличием в данных студентов, причины досрочного выбытия которых не связаны с учебной успеваемостью (*Табл. 6*).

Для проверки устойчивости результатов мы также провели оценку точности модели, используя метод кросс-валидации. В этом случае весь набор данных базовой модели (Data set 1_ University) был разбит на 10 частей (фолдов) с равным количеством уникальных записей в каждой. Далее было проведено 10 итераций обучения нейронной сети, во время каждой из которых один фолд выступал в роли тестового набора, а остальные – в роли обучающего набора, т. е. в каждой итерации нейронная сеть обучалась на 9 фолдах и тестировалась на одном фолде, последовательно смещавшемся на каждом шаге. Среднее значения точности по всем итерациям составило 0,875, что не отличается существенно от результатов, полученных методом разбивки набора данных на обучающую и тестовую выборку.

При высоких значениях базовых показателей качества прогноза нейронной сети, данная модель демонстрирует несущественный разброс значений при десятикратном создании, обучении и тестировании.

Доля верных срабатываний классификатора к общему числу объектов за пределами класса «graduate» оказалась наиболее высокой – Specificity = 0,93. Последнее значение – Specificity – в исследовательской части данного проекта представляется более значимым. С точки зрения оперативного принятия решения о корректировке образовательной траектории студента ситуация ошибочного прогнозирования потенциального выпускника с невысоким CGPA в качестве «attrition» (Recall) некритична по сравнению с ситуацией, когда в категорию «graduate» ошибочно попадают студенты с высоким риском выбытия из-за низкой успеваемости (Specificity).

При пороге 25% (широкий отбор) доля правильных классификаций в категории «attrition» (Recall) значительно понижается (72% верных классификаций), но доля верно отнесённых ИНС случаев в категорию «graduate» достигает 98%, «graduate» классифицируются с ошибкой 2%.

При пороге 75% (узкий отбор) растёт доля правильных классификаций в категории «attrition» (Recall), но снижается доля правильных классификаций в категории «graduate»

Таблица 7

Значения показателей производительности ИНС для разных порогов классификации

Table 7

The values of performance metrics for different thresholds

	Порог 0,25	Порог 0,5	Порог 0,75
Accuracy	0,85	0,88	0,85
Recall	0,72	0,83	0,92
Precision	0,98	0,91	0,80
Specificity	0,98	0,93	0,78
FPR	0,01	0,07	0,22
F1 Score	0,83	0,87	0,86

Таблица 8

Показатели производительности ИНС для разных наборов данных

Table 8

Neural network's performance metrics for different datasets

Показатели качества	Data set 1_ University	Data set 2_ Management, Economics and Finance	Data set 3_ Engineering, Technology and Computer Science
Accuracy	0,88	0,86	0,88
Recall	0,83	0,90	0,91
Precision	0,91	0,82	0,83
Specificity	0,93	0,90	0,92
Fall-out (FPR)	0,07	0,1	0,08
F_1 Score	0,87	0,85	0,87

(Specificity), как следствие – 22% «graduate» ИНС классифицирует как «attrition». Precision значительно понижается (*Табл. 7*).

Таким образом, при использовании полного набора комбинированных данных предлагаемая модель искусственной нейронной сети демонстрирует высокие значения показателей качества прогноза.

Результаты обучения и тестирования на Data set 2_ Management, Economics and Finance и Data set 3_ Engineering, Technology and Computer Science говорят о том, что перенос основной модели прогнозирования на разные специальности в разнопрофильном университете возможен без потери общей высокой точности прогноза. На малых специализированных выборках выросла доля правильных классификаций в категории «attrition» (Recall) (*Табл. 8*).

Оценка важности и взаимосвязи переменных. Включение всех собранных данных в анализ может привести к результатам ниже оптимального прогноза, особенно в случае избыточности данных или зависимости данных. Таким образом, важно определить, какие переменные важны или должны быть включены в анализ [4].

Для проверки значимости переменных нами был использован метод прямого отбора (Forward selection) факторов: начинаем с пустого набора признаков, а затем итеративно добавляем признаки, обеспечивающие наилучший прирост Accuracy.

Полученные результаты подтверждают ряд гипотез об устойчивых связях между факторами, уже констатированных в предыдущих исследованиях. Так, значения CGPA и GPA1 являются наиболее важными переменными, влияющими на общую точность прогноза ИНС – Accuracy. Вместе с тем наши данные показывают, что в комбинации с CGPA и GPA1 значительное влияние на прогноз оказывает год поступления в университет. Последовательное добавление остальных факторов в ряде случаев несущественно понижает точность прогноза (*Табл. 9*).

Выводы

Для прогнозирования отсева студентов разработаны и обучены довольно простые

Таблица 9

Значимость параметров

Table 9

The variable's significance

	Accuracy	Precision	Recall	Specificity	F1
CGPA	0,801	0,847	0,724	0,875	0,781
Год поступления	0,806	0,853	0,729	0,879	0,786
GPA1	0,865	0,904	0,808	0,919	0,853
Специализация	0,871	0,907	0,822	0,919	0,862
GPA2	0,877	0,915	0,825	0,927	0,868
ЕГЭ2_ профиль	0,870	0,905	0,820	0,918	0,860
Пол	0,877	0,915	0,824	0,927	0,868
Форма обучения	0,872	0,904	0,826	0,916	0,863
ЕГЭ1	0,877	0,914	0,826	0,925	0,868
ЕГЭ3	0,877	0,911	0,830	0,922	0,868
ЕГЭ	0,877	0,902	0,834	0,914	0,869
Предыдущий уровень образования	0,876	0,906	0,832	0,917	0,867
ЕГЭ1_ база	0,875	0,913	0,822	0,926	0,865
ЕГЭ2	0,874	0,899	0,838	0,909	0,867
Оплата	0,877	0,906	0,835	0,918	0,869

модели-классификаторы, построенные на основе полносвязных трёхслойных нейронных сетей с прямыми связями. На вход нейронных сетей в качестве входных параметров подавались вектора, содержащие 15 числовых и категориальных факторов, характеризующих отдельного студента. Модели обучалась на решение задачи бинарной классификации (студент успешно завершит обучение / студент не завершит обучение). Для проверки гипотезы о возможности переноса модели в разные контексты были использованы три выборки данных, различающиеся направлением подготовки студентов. Оценка важности влияния отдельных предикторов на результат прогноза выполнена на основе метода исключения переменных. Точность прогноза обученных нейронных сетей оценена на основе параметров Accuracy, Recall, Precision, Specificity, F1 Score на тестовой выборке, а также построена ROC-кривая.

Можно дискутировать с тезисом о том, что машинное обучение нацелено исключительно на прогнозирование результатов обучения, а не на исследование причинно-следственных связей [39]. Любой алгоритм интеллектуального анализа данных зависим не только от технических характеристик, но и от полноценного набора данных и в этом смысле EDM для прогноза академической успеваемости оперирует такими категориями, как «значимая» или «незначимая» переменная. Подтверждение значимости такого фактора, как год поступления в университет, означает, что влияние на учебный опыт студента оказывают эффекты взаимодействия «преподаватель-студент» или более широко – «система обучения-студент» (изменения организационных условий обучения, состава студенческих групп, состава преподавателей, содержания учебных программ).

Дискуссионным остаётся вопрос о возможности тиражирования данной модели прогнозирования на другие университеты. Этот вопрос нуждается в дополнительном экспериментальном подтверждении, так как у каждого учебного заведения своя политика распределения оценок, что затрудняет приравнивание одного и того же общего среднего балла успеваемости за все полные семестры к разным университетам [40]. Процесс инфляции оценок за успеваемость может оказывать влияние на точность прогнозирования.

При наличии трёх базовых предикторов – общий средний балл успеваемости, год поступления, общий средний балл успеваемости за 1-й семестр – влияние таких факторов, как

ЕГЭ и оплата обучения, оказалось минимальным. Как нам кажется, данный вывод тоже может быть связан со спецификой рассматриваемого университета и этот вывод нуждается в дополнительном подтверждении с использованием данных других университетов.

Таким образом, наше исследование не только подтвердило ряд уже существующих гипотез о влиянии ряда факторов на прогнозирование академической успеваемости, но и выдвинуло предположение о необходимости проверки их универсальности или специфичности в конкретном высшем учебном заведении. Вместе с тем мы также доказали, что модель искусственной нейронной сети с определённым набором атрибутов может применяться в контексте отдельного высшего учебного заведения, независимо от специализации.

Литература

1. *Zawacki-Richter O., Marín V.I., Bond M., Gouverneur F.* Systematic review of research on artificial intelligence applications in higher education – where are the educators? // International Journal of Educational Technology in Higher Education. 2019. Vol. 16. Art. no. 39. P. 1–27. DOI: 10.1186/s41239-019-0171-0
2. *Romero C., Ventura S.* Data mining in education // WIREs Data Mining Knowledge Discovery. 2013. Vol. 3. No. 1. P. 12–27. DOI: 10.1002/widm.1075
3. *Baepler P., Murdoch C.J.* Academic analytics and data mining in higher education // International Journal Schoolarship of Teaching & Learn. 2010. Vol. 4. No. 2. P. 1–9. DOI: 10.20429/ijsotl.2010.040217
4. *Alyahyan E., Düştegör D.* Predicting academic success in higher education: literature review and best practices // International Journal of Educational Technology in Higher Education. 2020. Vol. 17. Art. no. 3. P. 1–21. DOI: 10.1186/s41239-020-0177-7
5. *Viberg O., Hatakka M., Bälter O., Mavroudi A.* The current landscape of learning analytics in higher education // Computers in Human Behavior. 2018. Vol. 89. P. 98–110. DOI: 10.1016/j.chb.2018.07.027
6. *Горбунова Е.В.* Выбытия студентов из вузов: исследования в России и США // Вопросы образования. 2018. № 1. С. 110–131. DOI: 10.17323/1814-9545-2018-1-110-131
7. *Груздев И.А., Горбунова Е.В., Фрумин И.Д.* Студенческий отсев в российских вузах: к постановке проблемы // Вопросы образования. 2013. № 2. С. 67–81. DOI: 10.17323/1814-9545-2013-2-67-81
8. *Терентьев Е.А., Груздев И.А., Горбунова Е.В.* Суд идёт: дискурс преподавателей об отсеве студентов // Вопросы образования. 2015. № 2. С. 129–151. DOI: 10.17323/1814-9545-2015-2-129-151
9. *Валеева Д.Р., Докука С.В., Юдкевич М.М.* Разрыв дружеских связей при академическом неуспехе: социальные сети и пересдачи у студентов // Вопросы образования. 2017. № 1. С. 8–24. DOI: 10.17323/1814-9545-2017-1-8-24
10. *Богданов М.Б., Малик В.М.* Как сочетаются социальное, территориальное и гендерное неравенства в образовательных траекториях молодёжи России? // Мониторинг общественного мнения: экономические и социальные перемены. 2020. № 3. С. 391–421. DOI: 10.14515/monitoring.2020.3.1603
11. *Хавенсон Т.Е., Соловьева А.А.* Связь результатов Единого государственного экзамена и успеваемости в вуз // Вопросы образования. 2014. № 1. С. 176–199. DOI: 10.17323/1814-9545-2014-1-176-199
12. *Hellas A., Ihantola P., Petersen A., Ajanovski V.V., Gutica M., Hynninen T., Knutas A., Leinonen J., Messom C., Liao S.N.* Predicting academic performance: a systematic literature review // In: Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018 Companion). Association for Computing Machinery. New York, USA. 2018. P. 175–199. DOI: 10.1145/3293881.3295783
13. *Raju D., Schumacker R.* Exploring student characteristics of retention that lead to graduation in higher education using data mining models // Journal of College Student Retention: Research, Theory and Practice. 2015. Vol. 16. No. 4. P. 563–591. DOI: 10.2190/CS.16.4.e
14. *Lau E.T., Sun L., Yang Q.* Modelling, prediction and classification of student academic performance using artificial neural networks // SN Applied Sciences. 2019. Vol. 1. Art. no. 982. DOI: 10.1007/s42452-019-0884-7
15. *Lesinski G., Corns S., Dagli C.* Application of an Artificial Neural Network to Predict

Graduation Success at the United States Military Academy // Procedia Computer Science. 2016. Vol. 95. P. 375–382. DOI: 10.1016/j.procs.2016.09.348

16. *Adewale A.M., Bamidele A.O., Lateef U.O.* Predictive modelling and analysis of academic performance of secondary school students: Artificial Neural Network approach // International Journal of Educational Technology in Higher Education. 2018. Vol. 9. No. 1. P. 1–8. DOI: 10.5897/IJSTER2017.0415
17. *Tsiakmaki M., Kostopoulos G.K., Kotsiantis S., Ragos O.* Transfer Learning from Deep Neural Networks for Predicting Student Performance // SN Applied Sciences. 2020. Vol. 10. No. 6. Art. no. 2145. DOI: 10.3390/app10062145
18. *Mengash H.A.* Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems // IEEE Access. 2020. No. 8. P. 55462–55470. DOI: 10.1109/ACCESS.2020.2981905
19. *Русаков С.В., Русакова О.Л., Посохина К.А.* Нейросетевая модель прогнозирования группы риска по успеваемости студентов первого курса // Современные информационные технологии и ИТ-образование. 2018. Т. 14. № 4. С. 815–822. DOI: 10.25559/SITITO.14.201804.815-822
20. *Котова Е.Е.* Прогнозирование успешности обучения в интегрированной образовательной среде с применением инструментов онлайн аналитики // Компьютерные инструменты в образовании. 2019. № 4. С. 55–81. DOI: 10.32603/2071- 2340-2019-4-55-80
21. *Lapenok M.V., Patrusheva O.M., Hudyakova S.A.* Using Neural Network Mathematical Models to Solve Pedagogical Problems // International Scientific Conference "Digitalization of Education: History, Trends and Prospects" (DETP 2020), Russia. 2020. P. 22–26. DOI: 10.2991/assehr.k.200509.005
22. *Galimyanov A.F., Gafarov F.M., Muzafarova A.I.* Application of Big Data in Determining and Regulating Trends in Education // International Scientific Conference "Digitalization of Education: History, Trends and Prospects" (DETP 2020), Russia. 2020. P. 681–684. DOI: 10.2991/assehr.k.200509.121
23. *Prokopyev N.A., Vakhitov G.Z., Ustin P.N.* Indexing of Social Network Texts for Psychometric Model of Academic Success Prediction // International Scientific Conference "Digitalization of Education: History, Trends and Prospects" (DETP 2020), Russia. 2020. P. 810–815. DOI: 10.2991/assehr.k.200509.143
24. *Jovanović, J., Saqr, M., Joksimović, S., Gašević, D.* Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success // Computers & Education. 2021. No. 172. Art. no. 104251. DOI: 10.1016/j.compedu.2021.104251
25. *Araque F., Roldán C., Salguero A.* Factors influencing university dropout rates // Computers & Education. 2009. Vol. 53. No. 3. P. 563–574. DOI: 10.1016/j.compedu.2009.03.013
26. *Gray G., McGuinness C., Owende P., Hofmann M.* Learning Factor Models of Students at Risk of Failing in the Early Stage of Tertiary Education // Journal of Learning Analytics. 2016. Vol. 3. No. 2. P. 330–372. DOI: 10.18608/jla.2016.32.20
27. *Asif R., Merceron A., Ali S.A., Haider N.G.* Analyzing undergraduate students' performance using educational data mining // Computers & Education. 2017. Vol. 113. No. 1. P. 177–194. DOI: 10.1016/j.compedu.2017.05.007
28. *Lonn S., Koester B.* Rearchitecting Data for Researchers: A Collaborative Model for Enabling Institutional Learning Analytics in Higher Education // Journal of Learning Analytics. 2019. Vol. 6. No. 2. P. 107–119. DOI: 10.18608/jla.2019.62.8
29. *Miguéis V.L., Freitas A., Garciab P.J.V., Silva A.* Early segmentation of students according to theiracademic performance: A predictive modelling approach // Decision Support Systems. 2018. No. 115. P. 36–51. DOI: 10.1016/j.dss.2018.09.001
30. *Yang J., DeVore S., Hewagallage D., Miller P., Ryan Q.X., Stewart J.* Using machine learning to identify the most at-risk students in physics classes // Physical Review Physics Education Research. 2020. No. 16. Art. no. 020130. DOI: 10.1103/PhysRevPhysEducRes.16.020130
31. *Kuncel N.R., Hezlett S.A.* Standardized Tests Predict Graduate Students' Success // Science. 2007. Vol. 315. No. 5815. P. 1080–1081. DOI: 10.1126/science.1136618
32. *Gafarov F.M., Rudneva Ya.B., Sharifov U.Yu., Trofimova A.V., Bormotov P.M.* Analysis of Students' Academic Performance by Using Machine Learning Tools // International Scientific Conference "Digitalization of Education: History, Trends and Prospects" (DETP 2020). Russia, 2020. P. 574–579. DOI: 10.2991/assehr.k.200509.104 URL: https://www.researchgate.net/publication/341498648_Analysis_of_Students'_Academic_Performance_by_Using_Machine_Learning_Tools

33. *Delen D.* A comparative analysis of machine learning techniques for student retention management // Decision Support Systems. 2010. Vol. 49. No. 4. P. 498–506. DOI: 10.1016/j.dss.2010.06.003
34. *Zabriskie C., Yang J., DeVore S., Stewart J.* Using machine learning to predict physics course outcomes // Physical Review Physics Education Research. 2019. No. 15. Art. no. 020120. DOI: 10.1103/PhysRevPhysEducRes.15.020120
35. *Aiken J.M., Henderson R., Caballero M.D.* Modeling student pathways in a physics bachelor's degree program // Physical Review Physics Education Research. 2019. No. 15. Art. no. 010128. DOI: 10.1103/PhysRevPhysEducRes.15.010128
36. *Alkhasawneh R., Hobson R.* Modeling student retention in science and engineering disciplines using neural networks // IEEE Global Engineering Education Conference (EDUCON). Amman, 2011. P. 660–663. DOI: 10.1109/EDUCON.2011.5773209
37. *Шмелева Е.Д., Фрумин И.Д.* Факторы отсева студентов инженерно-технического профиля в российских вузах // Вопросы образования. 2020. № 3. С. 110–136. DOI: 10.17323/1814-9545-2020-3-110-136
38. *Смык А.Ф., Прусова В.И., Зиманов Л.Л., Солнцев А.А.* Анализ масштаба и причин отсева студентов в техническом университете // Высшее образование в России. 2019. Т. 28. № 6. С. 52–62. DOI: 10.31992/0869-3617-2019-28-6-52-62
39. *Tsai S.-C., Chen C.-H., Shiao Y.-T., Ciou J.-S., Wu T.-N.* Precision education with statistical learning and deep learning: a case study in Taiwan // International Journal of Educational Technology in Higher Education. 2020. Vol. 17. Art. no. 12. DOI: 10.1186/s41239-020-00186-2
40. *Thai-Nghe N., Janecek P., Haddawy P.* A comparative analysis of techniques for predicting academic performance // 37th Annual Frontiers In Education Conference – Global Engineering: Knowledge Without Borders, Opportunities Without Passports, Milwaukee, WI. 2007. T2G-7-T2G-12. DOI: 10.1109/FIE.2007.4417993

Благодарность. Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта № 19-29-14082.

Статья поступила в редакцию 25.08.22
Принята к публикации 28.11.22

References

1. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F. (2019). Systematic Review of Research on Artificial Intelligence Applications in Higher Education – Where Are the Educators? *International Journal of Educational Technology in Higher Education*. Vol. 16, art. no. 39, pp. 1-27, doi: 10.1186/s41239-019-0171-0
2. Romero, C., Ventura, S. (2013). Data Mining in Education. *WIREs Data Mining Knowledge Discovery*. Vol. 3, no. 1, pp. 12-27, doi: 10.1002/widm.1075
3. Baepler, P., Murdoch, C.J. (2010) Academic Analytics and Data Mining in Higher Education. *International Journal Schoolarship of Teaching & Learn*. Vol. 4, no. 2, pp. 1-9, doi: 10.20429/ijsotl.2010.040217
4. Alyahyan, E., Düştegör, D. (2020). Predicting Academic Success in Higher Education: Literature Review and Best Practices. *International Journal of Educational Technology in Higher Education*. Vol. 17, art. no. 3, pp. 1-21, doi: 10.1186/s41239-020-0177-7
5. Viberg, O., Hatakka, M., Bälter, O., Mavroudi, A. (2018). The Current Landscape of Learning Analytics in Higher Education. *Computers in Human Behavior*. Vol. 89, pp. 98-110, doi: 10.1016/j.chb.2018.07.027
6. Gorbunova, E.V. (2018). Elaboration of Research on Student Withdrawal from Universities in Russia and the United States. *Voprosy obrazovaniya = Educational Studies Moscow.* No. 1, pp. 110-131, doi: 10.17323/1814-9545-2018-1-110-131 (In Russ., abstract in Eng.)
7. Gruzdev, I.A., Gorbunova, E.V., Frumin, I.D. (2013). Academic Dismissal in Russian Higher Education Institutions: Defining the Problem. *Voprosy obrazovaniya = Educational Studies Moscow.* No. 2, pp. 67-81, doi: 10.17323/1814-9545-2013-2-67-81 (In Russ., abstract in Eng.)
8. Terent'ev, E.A., Gruzdev, I.A., Gorbunova, E.V. (2015). The Court is Now in Session: Professor Discourse on Student Attrition. *Voprosy obrazovaniya = Educational Studies Moscow.* No. 2, pp. 129-151, doi: 10.17323/1814-9545-2015-2-129-151 (In Russ., abstract in Eng.)
9. Valeeva, D.R., Dokuka, S.V., Yudkevich, M.M. (2017). How Academic Failures Break up Friendship Ties: Social Networks and Retake. *Voprosy obrazovaniya =Educational Studies Moscow*. No. 1, pp. 8-24, doi: 10.17323/1814-9545-2017-1-8-24 (In Russ., abstract in Eng.)

10. Bogdanov, M.B., Malik,V.M. (2020). Social, Territorial and Gender Inequalities in Educational Trajectories of the Russian Youth. *Monitoring obshchestvennogo mneniya: economicheskiye i socialnye izmeneniya = Monitoring of Public Opinion: Economic and Social Changes*. No. 3, pp. 391-421, doi: 10.14515/monitoring.2020.3.1603 (In Russ., abstract in Eng.)
11. Khavenson, T.Ye., Solovyova, A.A. (2014). Studying the Relation Between the Unified State Exam Points and Higher Education Performance. *Voprosy obrazovaniya = Educational Studies Moscow*. No. 1, pp. 176-199, doi: 10.17323/1814-9545-2014-1-176-199 (In Russ., abstract in Eng.).
12. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., Liao, S.N. (2018). Predicting Academic Performance: a Systematic Literature Review. In: *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018 Companion). Association for Computing Machinery*. New York, USA, pp. 175-199, doi: 10.1145/3293881.3295783
13. Raju, D., Schumacker, R. (2015). Exploring Student Characteristics of Retention That Lead to Graduation in Higher Education Using Data Mining Models. *Journal of College Student Retention: Research, Theory and Practic*e. Vol. 16, no. 4, pp. 563-591, doi: 10.2190/CS.16.4.e
14. Lau, E.T., Sun, L., Yang, Q. Modelling, Prediction and Classification of Student Academic Performance Using Artificial Neural Networks. (2019). *SN Applied Sciences*. Vol. 1, art. no. 982, doi: 10.1007/s42452-019-0884-7
15. Lesinski, G., Corns, S., Dagli, C. (2016). Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy. *Procedia Computer Science*. Vol. 95, pp. 375-382, doi: 10.1016/j.procs.2016.09.348
16. Adewale, A.M., Bamidele, A.O., Lateef, U.O. (2018). Predictive Modelling and Analysis of Academic Performance of Secondary School Students: Artificial Neural Network Approach. *International Journal of Educational Technology in Higher Education*. Vol. 9, no. 1, pp. 1-8, doi: 10.5897/IJSTER2017.0415
17. Tsiakmaki, M., Kostopoulos, G.K., Kotsiantis, S., Ragos, O. (2020). Transfer Learning from Deep Neural Networks for Predicting Student Performance. *SN Applied Sciences*. Vol. 10, no. 6, art. no. 2145, doi: 10.3390/app10062145
18. Mengash, H. (2020). Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems. *IEEE Access*. No. 8, pp. 55462-55470, doi: 10.1109/ACCESS.2020.2981905
19. Rusakov, S.V., Rusakova, O.L., Posokhina, K.A. (2018). Neural Network Model of Predicting the Risk Group for the Accession of Students of the First Course. *Modern Information Technologies and IT-Education*. Vol. 14, no. 4, pp. 815-822, doi: 10.25559/SITITO.14.201804.815-822
20. Kotova, E.E. (2019). Prediction of Learning Success in an Integrated Educational Environment Using Online Analytics Tools. *Computer tools in education*. No. 4, pp. 55-80, doi:10.32603/2071-2340-2019-4-55-80
21. Lapenok, M.V., Patrusheva, O.M., Hudyakova, S.A. (2020). Using Neural Network Mathematical Models to Solve Pedagogical Problems. In: *International Scientific Conference "Digitalization of Education: History, Trends and Prospects" (DETP 2020)*, Russia. Pp. 22-26, doi: 10.2991/assehr.k.200509.005
22. Galimyanov, A.F., Gafarov, F.M., Muzafarova, A.I. (2020). Application of Big Data in Determining and Regulating Trends in Education. In: *International Scientific Conference "Digitalization of Education: History, Trends and Prospects" (DETP 2020)*, Russia. Pp. 681-684, doi: 10.2991/assehr.k.200509.121
23. Prokopyev, N.A., Vakhitov, G.Z., Ustin, P.N. (2020). Indexing of Social Network Texts for Psychometric Model of Academic Success Prediction. In: *International Scientific Conference "Digitalization of Education: History, Trends and Prospects" (DETP 2020)*, Russia. Pp. 810-815, doi: 10.2991/assehr.k.200509.143
24. Jovanović, J., Saqr, M., Joksimović, S., Gašević, D. (2021). Students Matter the Most in Learning Analytics: The Effects of Internal and Instructional Conditions in Predicting Academic Success. Computers & Education. No. 172. art. no. 104251, doi: 10.1016/j.compedu.2021.104251
25. Araque, F., Roldán, C., Salguero, A. (2009). Factors Influencing University Dropout Rates. *Computers & Education*. Vol. 53, no. 3, pp. 563-574, doi: 10.1016/j.compedu.2009.03.013
26. Gray, G., McGuinness, C., Owende, P., Hofmann, M. (2016). Learning Factor Models of Students at Risk of Failing in the Early Stage of Tertiary Education. *Journal of Learning Analytics*. Vol. 3, no. 2, pp. 330-372, doi: 10.18608/jla.2016.32.20

27. Asif, R., Merceron, A., Ali, S.A., Haider, N.G. (2017). Analyzing Undergraduate Students' Performance Using Educational Data Mining. *Computers & Education*. Vol. 113, no. 1, pp. 177-194, doi: 10.1016/j.compedu.2017.05.007
28. Lonn, S., Koester, B. (2019). Rearchitecting Data for Researchers: A Collaborative Model for Enabling Institutional Learning Analytics in Higher Education. *Journal of Learning Analytics*. Vol. 6, no. 2, pp. 107-119, doi: 10.18608/jla.2019.62.8
29. Miguéis, V.L., Freitas, A., Garciab, P. J.V., Silva, A. (2018). Early Segmentation of Students According to Their Academic Performance: A Predictive Modelling Approach. *Decision Support Systems*. No. 115, pp. 36-51, doi: 10.1016/j.dss.2018.09.001
30. Yang, J., DeVore, S., Hewagallage, D., Miller, P., Ryan, Q.X., Stewart, J. (2020). Using Machine Learning to Identify the Most At-risk Students in Physics Classes. *Physical Review Physics Education Research*. No. 16, art. no. 020130, doi: 10.1103/PhysRevPhysEducRes.16.020130
31. Kuncel, N.R., Hezlett, S.A. (2007). Standardized Tests Predict Graduate Students' Success. *Science*. No. 315, no. 5815, pp. 1080-1081, doi: 10.1126/science.1136618
32. Gafarov, F.M., Rudneva, Ya.B., Sharifov, U.Yu., Trofimova, A.V., Bormotov, P.M. (2020). Analysis of Students' Academic Performance by Using Machine Learning Tools. In: *International Scientific Conference* "*Digitalization of Education: History, Trends and Prospects*" *(DETP 2020),* Russia. Pp. 574-579, doi: 10.2991/assehr.k.200509.104 Available at: https://www.researchgate.net/publication/341498648_Analysis_of_Students'_Academic_Performance_by_Using_Machine_Learning_Tools
33. Delen, D. (2010). A Comparative Analysis of Machine Learning Techniques for Student Retention Management. *Decision Support Systems*. Vol. 49, no. 4, pp. 498-506, doi: 10.1016/j.dss.2010.06.003
34. Zabriskie, C., Yang, J., DeVore, S., Stewart, J. (2019). Using Machine Learning to Predict Physics Course Outcomes. *Physical Review Physics Education Research*. No. 15, art. no. 020120, doi: 10.1103/PhysRevPhysEducRes.15.020120
35. Aiken, J.M., Henderson, R., Caballero, M.D. (2019). Modeling Student Pathways in a Physics Bachelor's Degree Program. *Physical Review Physics Education Research*. No. 15, art. no. 010128, doi: 10.1103/PhysRevPhysEducRes.15.010128
36. Alkhasawneh, R., Hobson, R. (2011). Modeling Student Retention in Science and Engineering Disciplines Using Neural Networks. In: *IEEE Global Engineering Education Conference (EDUCON),* Amman. Pp. 660-663, doi: 10.1109/EDUCON.2011.5773209
37. Shmeleva, E.D., Froumin, I.D. (2020). Factors of Attrition among Computer Science and Engineering Undergraduates in Russia. *Voprosy obrazovaniya = Educational Studies Moscow.* No. 3, pp. 110-136, doi: 10.17323/1814-9545-2020-3-110-136 (In Russ., abstract in Eng.)
38. Smyk, A.F., Prusova, V.I., Zimanov, L.L., Solntsev, A.A. (2019). Study of the Scale and the Reasons of Student Dropout from Technical University. *Vysshee obrazovanie v Rossii = Higher Education in Russia*. No. 6, pp. 52-62, doi: /10.31992/0869-3617-2019-28-6-52-62 (In Russ., abstract in Eng.)
39. Tsai, S.-C., Chen, C.-H., Shiao, Y.-T., Ciou , J.-S., Wu, T.-N. (2020). Precision Education with Statistical Learning and Deep Learning: a Case Study in Taiwan. *International Journal of Educational Technology in Higher Education*. Vol. 17, art. no. 12, doi: /10.1186/s41239-020-00186-2
40. Thai-Nghe, N., Janecek, P., Haddawy, P. (2007). A Comparative Analysis of Techniques for Predicting Academic Performance. In: *37th Annual Frontiers In Education Conference – Global Engineering: Knowledge Without Borders, Opportunities Without Passports*, Milwaukee, WI. T2G-7-T2G-12, doi: 10.1109/FIE.2007.4417993

Acknowledgements. The reported study was funded by RFBR, project number 19-29-14082.

The paper was submitted 25.08.22
Accepted for publication 28.11.22