ISSN 0234-0453 • ИНФОРМАТИКА И ОБРАЗОВАНИЕ • 2023 • Том 38 № 4

DOI: 10.32517/0234-0453-2023-38-4-28-43

ПРОГНОЗИРОВАНИЕ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ СТУДЕНТОВ С ИСПОЛЬЗОВАНИЕМ ИНСТРУМЕНТОВ МАШИННОГО ОБУЧЕНИЯ

Ю. Ю. Якунин[1] ✉, **В. Н. Шестаков**[1], **Д. И. Ликсонова**[1], **А. А. Даничев**[1]

[1] ***Сибирский федеральный университет, г. Красноярск, Россия***

✉ *yakuninyy@mail.ru*

Аннотация

Цифровые помощники все больше проникают в разные области деятельности человека, в том числе и в образовательную. Сегодня это уже не просто автоматизированные системы или веб-приложения, поддерживающие и автоматизирующие некоторые процессы, включая учебный. Сейчас это более интеллектуальные и более автономные системы. Цифровые помощники в жизни студента играют особую роль, в некотором смысле замещая деканат, наставника, тьютора, представителей других служб университета и другие элементы образовательной инфраструктуры. Цифровая поддержка студента важна и полезна, особенно на первом курсе, в период адаптации обучающегося к среде высшего образования, существенно отличающегося от школьного. Именно в этот момент происходит наибольшее количество отчислений студентов по причине академической неуспеваемости. Цифровой помощник в виде мобильного приложения, умеющий спрогнозировать результаты обучения и вовремя проинформировать об этом, по мнению авторов, может оказать важную поддержку для студента и помочь ему сориентироваться и скорректировать свое поведение в случае угрозы негативного результата.

Для решения задач по созданию прогнозной модели результатов обучения студентов и соответствующего мобильного приложения, а также для проведения предпроектного исследования были использованы следующие методы и инструменты математической статистики: метод k-средних, метод корреляции Кендалла, критерий Фридмана с апостериорным критерием Дарбина—Коновера, линейная регрессия, логистическая регрессия, категориальный байесовский классификатор, метод случайного леса, нейронная сеть (многослойный перцептрон), непараметрическая оценка функции регрессии Надарая—Ватсона, STATISTICA 10.0 и Jamovi 2.2.5, библиотеки Python.

В результате исследования создана математическая модель прогнозирования результатов обучения по дисциплинам на основе текущей успеваемости в электронных образовательных курсах. Точность модели зависит от недели обучения, на которой она применяется, и достигает 92,6 %. На ранних этапах (например, для 7-й недели) точность составляет не ниже 85 % и варьируется в зависимости от контингента студентов и дисциплин.

В результате исследования разработано мобильное приложение, реализующее прогнозную модель и другие сопутствующие функции для информирования студента об ожидаемых успехах в обучении.

Созданная прогнозная модель основана на текущих данных об успеваемости, получаемых из электронных курсов, и способна делать точный прогноз, что позволяет применять ее на практике в онлайн-режиме и через мобильное приложение информировать обучающегося.

Ключевые слова: прогнозирование успеваемости обучающихся, прогнозирование результатов обучения, анализ образовательных данных, метод k-средних, непараметрическая ядерная оценка регрессии Надарая—Ватсона, машинное обучение, классификация обучающихся по успеваемости.

Для цитирования:

Якунин Ю. Ю., Шестаков В. Н., Ликсонова Д. И., Даничев А. А. Прогнозирование результатов обучения студентов с использованием инструментов машинного обучения. *Информатика и образование.* 2023;38(4):28–43. DOI: 10.32517/0234-0453-2023-38-4-28-43

PREDICTING STUDENT PERFORMANCE USING MACHINE LEARNING TOOLS

Yu. Yu. Yakunin[1] ✉, **V. N. Shestakov**[1], **D. I. Liksonova**[1], **A. A. Danichev**[1]

[1] ***Siberian Federal University, Krasnoyarsk, Russia***

✉ *yakuninyy@mail.ru*

ISSN 0234-0453 • INFORMATICS AND EDUCATION • 2023 • Volume 38 No 4

Abstract

Digital assistants are increasingly penetrating various areas of human activity, including education. Today, they are no longer just automated systems or web applications that support and automate certain processes, including educational processes. Now they are more intelligent and more autonomous systems. Digital assistants play a special role in a student's life, in a sense replacing the dean's office, mentor, tutor, representatives of other university services and other elements of educational infrastructure. The digital support for the student is important and useful, especially in the first year during his adaptation to the environment of higher education, which is significantly different from the school one. It is at this point that the largest amount of students dropouts occurs due to academic failure. According to the authors, a digital assistant in the form of a mobile application that can predict learning outcomes and inform about it in time, can provide important support for the student and help him/her orient and adjust his/her behavior in case of a threat of a negative result.

To solve the problems of creating a predictive model of student learning outcomes and a mobile application that implements it, as well as to conduct a pre-project study, the following methods and tools of mathematical statistics were used: *k*-means method, Kendall correlation method, Friedman' test with Durbin—Conover posterior test, linear regression, logistic regression, categorical Bayesian classifier, random forest method, neural network (multilayer perceptron), non-parametric estimation of the Nadaraya—Watson regression function, STATISTICA 10.0 and Jamovi 2.2.5, Python libraries.

As a result of the study, a mathematical model for predicting learning outcomes in disciplines based on current performance in e-learning courses was created. The accuracy of the model depends on the week of training in which it is applied and reaches 92,6 %. In the early stages (e. g., for week 7), the accuracy is at least 85 % and varies depending on the contingent of the student population and disciplines.

As a result of the study, a mobile application was developed that implements a predictive model and other related functions to inform the student about his/her estimated educational success.

The created predictive model is based on current performance data obtained from electronic courses and is capable of making accurate predictions, which allows it to be applied in practice online and through the mobile application to inform students.

Keywords: predicting student performance, predicting learning outcomes, educational data mining (EDM), *k*-means method, Nadaraya—Watson kernel regression, machine learning, students' performance classification.

For citation:

Yakunin Yu. Yu., Shestakov V. N., Liksonova D. I., Danichev A. A. Predicting student performance using machine learning tools. *Informatics and Education.* 2023;38(4):28–43. (In Russian.) DOI: 10.32517/0234-0453-2023-38-4-28-43

1. Введение

Исследования, направленные на поиск решений, позволяющих повысить эффективность образовательной деятельности в высших учебных заведениях, показали многоаспектность решаемой проблемы. Отдельным и тоже многогранным направлением этих исследований является область управления мотивацией студентов к обучению. Изучение этой проблемы показывает, что чем больше студент проинформирован об учебном процессе, чем больше он в него вовлечен, тем выше его мотивация и интерес к самому процессу обучения, что неизбежно приводит к повышению вероятности его личного успеха в учебе.

В современном мире для того, чтобы привлечь и заинтересовать студентов, можно предложить им специальные мобильные приложения, содержащие всю необходимую информацию об обучении, а также механизмы захвата внимания и долгосрочного удержания интереса при периодическом использовании такого приложения. Кроме прочего функционала, приложение позволяет получать прогноз результатов промежуточной аттестации в зависимости от данных о текущей успеваемости студента в семестре.

Доступные функции анализа образовательных данных и, в частности, прогнозирования успеваемости обучающихся расширяются благодаря возможности систематически собирать, хранить и автоматизированно обрабатывать большие объемы данных о студентах, применяя математические методы. Эти объемы особенно возросли в период пандемии COVID-19, когда электронные курсы стали использоваться намного интенсивнее, чем раньше.

2. Современные методы прогнозирования успеваемости обучающихся. Обзор актуальных исследований

За последние годы значительно увеличилось количество научных статей о прогнозировании успеваемости обучающихся с помощью методов машинного обучения. Проведенный **Y. A. Alsariera и коллегами в работе «Assessment and evaluation of different machine learning algorithms for predicting student performance»** [1] систематический обзор выявил около 40 публикаций о попытках применения методов анализа к образовательным данным с целью прогнозирования успеваемости студентов. При этом наборы обрабатываемых данных разнообразны и включают в себя:

- демографические данные;
- академические данные;
- личные данные;
- коммуникационные данные;
- психологические данные и др.

Разные исследователи применяют для анализа такие методы как:

- нейронные сети (*англ.* Neural Network, NN);
- машина опорных векторов (*англ.* Support Vector Machine, SVM);
- дерево решений (*англ.* Decision Tree);
- наивный байесовский метод (*англ.* Naive Bayes Classifier);
- *k*-ближайших соседей (*англ.* *k*-nearest neighbors algorithm, *k*-NN);
- линейная регрессия (*англ.* Linear Regression).

Наибольшую среднюю эффективность показали нейронные сети и дерево решений.

ISSN 0234-0453 • ИНФОРМАТИКА И ОБРАЗОВАНИЕ • 2023 • Том 38 № 4

2.1. Модель классификации

Большая часть исследований посвящена построению модели классификации на два класса (успех/неуспех) на основе разного количества и качества признаков.

В статье G. B. Brahim «Predicting student performance from online engagement activities using novel statistical features» [2] модель прогнозирования предусматривает извлечение в общей сложности 86 статистических признаков, которые были семантически разделены на три широкие категории на основе различных критериев:

- тип активности;
- временная статистика;
- количество периферийной активности.

Использовались пять популярных классификаторов:

- метод случайного леса (*англ.* Random Forest, RF);
- машина опорных векторов;
- наивный байесовский метод;
- логистическая регрессия (*англ.* Logistic Regression, LR);
- многослойный перцептрон (*англ.* Multilayered Perceptron, MLP).

В. И. Токтарова и Ю. А. Пашкова в работе «Предиктивная аналитика в цифровом образовании: анализ и оценка успешности обучения студентов» [3] использовали следующие признаки:

- год поступления;
- курс;
- направление подготовки;
- группа;
- пол;
- средний балл аттестата;
- баллы текущей аттестации (на входном контроле по дисциплине, за выполнение лабораторных работ, домашних заданий, контрольных работ, защиту рефератов);
- баллы промежуточной аттестации (зачеты по дисциплине или ее разделам, зачет по курсовой работе, экзамен, контроль остаточных знаний);
- итоговый балл и итоговая оценка;
- количество кликов по компонентам дисциплины.

Были получены следующие результаты применения моделей прогнозирования:

- линейная регрессия (точность 63 %);
- логистическая регрессия (80 %);
- глубокие нейронные сети (82 %);
- дерево решений (89 %).

A. Almasri, E. Celebi, R. S. Alkhawaldeh в статье «EMT: Ensemble meta-based tree model for predicting student performance» [4] применили различные методы машинного обучения, используя возможности их гибкого сочетания. Исходными данными для анализа являлись:

- средний балл обучающегося;
- промежуточные результаты (оценки по дисциплинам, тестам, посещаемость);
- социально-демографические данные (пол, возраст, доходы, семейное положение);
- внеклассные занятия;
- психометрические данные;
- посещение подготовительных курсов для университета;
- взаимодействие в социальных сетях.

В. А. Шевченко в работе «Прогнозирование успеваемости студентов на основе методов кластерного анализа» [5] для прогноза использовала кластеризацию методом k-средних. Обрабатываются данные об оценке входных знаний по дисциплине, оценке знаний по первой теме дисциплины, числе пропусков на момент прогноза.

Авторы статьи **«Significance of non-academic parameters for predicting student performance using ensemble learning techniques»** [6] **D. Aggarwal, S. Mittal, V. Bali** сравнивают модели с различным набором признаков:

- с использованием только академических параметров;
- с использованием академических и демографических параметров.

В статье L. M. A. Zohair «Prediction of student's performance by modelling small dataset size» [7] изучены методы KNN и LDA (*англ.* Latent Dirichlet Allocation, латентное размещение Дирихле) для малых выборок. В двух статьях **Р. Б. Куприянова** и **Д. Ю. Звонарева** — **«Повышение качества модели прогнозирования образовательных результатов студентов университетов»** [8] и **«Разработка модели прогнозирования образовательных результатов обучающихся для университетов»** [9] — прогнозирование успеваемости основано на алгоритмах градиентного бустинга (*англ.* Gradient Tree Boosting, Gradient Boosting Machine, GBM) над решающими деревьями и линейной регрессии.

Кроме этого, встречаются исследования, связанные с классификацией по шаблонам поведения обучающихся в электронных курсах.

В работе «Predicting student performance using sequence classification with time-based windows» [10] **G. Deeva и соавторы** изучают шаблоны поведения, выражающиеся в последовательности событий пользователя курса с учетом времени поведения (недели обучения). **С. В. Русаков, О. Л. Русакова и К. А. Посохина в статье «Нейросетевая модель прогнозирования группы риска по успеваемости студентов первого курса»** [11] с помощью нейронной сети составили портрет (шаблон) студента — кандидата на попадание в группу риска.

В работе R. H. Ali «Educational data mining for predicting academic student performance using active classification» [12] методы интеллектуального анализа образовательных данных (*англ.* Educational Data Mining, EDM) используются для обнаружения шаблона, чтобы улучшить образовательный процесс и добиться высокой эффективности всех образова-

ISSN 0234-0453 • INFORMATICS AND EDUCATION • 2023 • Volume 38 No 4

тельных элементов. Для классификации признаков применялись четыре метода:

- метод случайного леса;
- метод распространения меток (*англ.* Label Distribution Protocol, LDP);
- логистическая регрессия;
- многослойный перцептрон.

2.2. Нейронные сети

Отдельно стоит рассмотреть исследования с применением нейронных сетей для решения задачи прогнозирования.

Метод, предлагаемый **S. Li и T. Liu в статье «Performance prediction for higher education students using deep learning»** [13], использует глубокую нейронную сеть для прогнозирования путем извлечения информативных данных в виде признаков с соответствующими весами. Несколько обновленных скрытых слоев применяются для автоматического проектирования нейронной сети.

S. Poudyal, M. J. Mohammadi-Aragh, J. E. Ball в статье «Prediction of student academic performance using a hybrid 2D CNN model» [14] построили гибридную 2D-модель по архитектуре сверточных нейронных сетей CNN, объединив две разные 2D-модели CNN для прогнозирования успеваемости. Достигнутая точность выше, чем у классических моделей, таких как k-ближайших соседей, наивный байесовский метод, деревья решений и логистическая регрессия.

S. Sood, M. Saini в статье «Hybridization of cluster-based LDA and ANN for student performance prediction and comments evaluation» [15] использовали гибридный подход, состоящий из линейного дискриминантного анализа на основе кластера и искусственной нейронной сети (*англ.* Artificial Neural Network, ANN). Благодаря этому они предоставили потенциальным студентам мотивационные комментарии и видеорекомендации, с помощью которых они могут выбрать правильный предмет, чтобы снизить вероятность отчисления.

M. Tsiakmaki и коллеги в работе «Transfer learning from deep neural networks for predicting student performance» [16] предложили метод трансфертного обучения, который позволяет обучать глубокую сеть, применяя набор данных прошлого курса (исходный курс), и повторно использовать его в качестве отправной точки для набора данных нового курса (целевой курс).

Y. Liu и др. проводят эксперименты по сравнению алгоритмов глубокого обучения (включая LSTM (*англ.* Long Short-Term Memory)) и 1D-CNN с традиционными подходами к машинному обучению на основе выборки 5 341 студента и данных об их поведении при кликах из Open University Learning Analytics. Выделены четыре из двенадцати учебных инструментов, которые являются критически важными и влияют на успеваемость студентов. Результаты исследования представлены **в статье «Predicting student performance using clickstream data and machine learning»** [17].

2.3. Комбинированные и иные методы

В работе R. Conijn, A. Van den Beemt, P. Cuijpers «Predicting student performance in a blended MOOC» [18] изучается подвыборка студентов смешанных массовых открытых онлайн-курсов (МООК) для выпускников, цель обучения которых — завершение курса в кампусе. Для прогнозирования успеваемости обучающихся совокупная частота действий, частота конкретных элементов курса и порядок действий были проанализированы с использованием корреляций, множественной регрессии и анализа процессов. Все агрегированные частоты активности МООК оказались положительно связаны с оценками экзамена в кампусе. Однако эта взаимосвязь менее ясна при контроле прошлой деятельности. В целом 65 % конкретных предметов курса показали значительную корреляцию с итоговой оценкой на экзамене. Студенты, успешно сдавшие курс, распределили свое обучение на большее количество дней по сравнению с теми, кто не прошел курс. Небольшая разница была обнаружена в порядке действий в рамках МООК между сдавшими и не прошедшими испытания студентами.

А. Е. Шухман, Д. И. Парфенов, Л. В. Легашев, Л. С. Гришина в статье «Анализ и прогнозирование успеваемости обучающихся при использовании цифровой образовательной среды» [19] показывают результаты изучения возможностей прогнозирования успеваемости (среднего балла за сессию) методом машинного обучения алгоритмом градиентного бустинга LightGBM. В этой работе, проведенной в Оренбургском государственном университете, использовались следующие входные данные:

- накопленный средний балл за предыдущий период обучения;
- средний балл по итогам первого рубежного контроля;
- средний балл по итогам второго рубежного контроля (12 недель с начала семестра);
- процент пропусков;
- год рождения;
- пол;
- место проживания (город/село);
- уровень предыдущего образования;
- семейное положение;
- результаты вступительных испытаний;
- является ли обучающийся сиротой;
- является ли обучающийся инвалидом;
- проживает ли обучающийся в общежитии.

Среднее отклонение предсказанного среднего балла от реального составило 0,18.

Актуальными представляются исследования, связанные с изучением поведения обучающихся в электронных курсах. **F. Qiu и коллеги в статье «Predicting students' performance in e-learning using learning process and behaviour data»** [20] предлагают структуру прогнозирования эффективности электронного обучения на основе классификации поведения (*англ.* Behaviour-Based Classification of the

ISSN 0234-0453 • ИНФОРМАТИКА И ОБРАЗОВАНИЕ • 2023 • Том 38 № 4

E-Learning Performance, BCEP), которая выбирает особенности поведения в электронном обучении, использует слияние признаков с данными о поведении в соответствии с моделью его классификации, чтобы получить значения признаков класса для каждого типа поведения, и, наконец, создает предиктор эффективности обучения на основе машинного обучения.

Предпринимаются попытки привлечь в массив анализируемых данных индивидуальные психологические характеристики обучающихся. **Е. Е. Котова в работе «Прогнозирование успешности обучения в интегрированной образовательной среде с применением инструментов онлайн-аналитики»** [21], помимо успеваемости, предлагает для прогноза использовать данные, полученные с помощью психологического инструментария (когнитивно-стилевой потенциал). **I. Al-Kindi, Z. Al-Khanjari в статье «Tracking student performance tool for predicting students EBPP in online courses»** [22] прогнозируют такие параметры, как вовлеченность обучающихся, поведение, личность и успеваемость, с помощью инструмента отслеживания успеваемости студентов, который получает данные непосредственно из журналов Moodle любых выбранных курсов.

Ряд исследований доведены до программной реализации моделей в виде приложения. Так, **в Томском политехническом университете** изучили возможность создания прогнозной модели успеваемости студентов* на основе следующих данных:

- форма обучения;
- квалификация;
- курс;
- специальность;
- академический отпуск (действующий): да/нет;
- всего часов пропусков в семестре;
- всего часов аудиторных занятий в семестре.

Классификация проведена следующими методами:

- метод логистической регрессии;
- метод опорных векторов;
- метод случайного леса (показал лучшие прогнозные результаты).

С. Б. Пахирко создал приложение для прогнозирования успеваемости студентов, работающее на мобильных устройствах с ОС Android. Для прогнозирования результатов экзамена на основе данных контрольных недель использовался метод линейной регрессии, для прогнозирования сдачи зачета применялись методы бинарной классификации**.

3. Цель и задачи исследования

Анализ литературных источников показал, что тема прогнозирования успеваемости студентов является актуальной и многие исследователи пытаются решить эту задачу разными методами и с разных точек зрения. Существенное влияние на постановку задачи прогнозирования и выбор методов оказывают исходные данные, которые используются для обучения моделей прогнозирования и их применения. Кроме этого, определение самой функции расчета результата обучения, успеваемости или успешности также играет важную роль для постановки и решения задачи прогнозирования обучения.

Целью настоящего исследования является разработка системы прогнозирования результатов обучения студентов на основе данных текущей успеваемости в электронных образовательных курсах (ЭОК), способной в реальном времени через мобильное приложение информировать каждого студента Института космических и информационных технологий Сибирского федерального университета (ИКИТ СФУ) о вероятном результате его обучения в текущем семестре по каждой изучаемой дисциплине.

Задачи исследования:

- формализовать понятие результата обучения, используя доступные актуальные исходные данные, для применения его в прогнозной модели;
- провести предпроектное исследование методом кластерного анализа для подтверждения гипотезы о возможности разделения студентов на классы в зависимости от успешности результатов обучения;
- провести анализ методов прогнозирования, выбрать и исследовать наиболее эффективный метод для данной задачи прогнозирования;
- реализовать прогнозную модель в мобильном приложении для использования студентами в процессе обучения.

Сформулированные выше цель и задачи, а также доступные в режиме онлайн актуальные исходные данные определили характер исследования и повлияли на выбор методов прогнозирования.

4. Материалы и методы

Общая выборка исходных данных сформирована на основе данных по 2 130 студентам из 89 учебных групп, обучающимся в общей сложности по 526 электронным образовательным курсам.

Для демонстрации примеров и результатов исследований в данной статье использованы данные двух выборок:

* *Зяблецев П. А.* Прогнозная модель для оценки успеваемости студентов университета по итогам текущего обучения. Магистерская диссертация по направлению подготовки 09.04.04 «Программная инженерия». Томск: Томский политехнический университет; 2020. https://earchive.tpu.ru/bitstream/11683/61074/1/TPU930128.pdf

** *Пахирко С. Б.* Система прогнозирования успеваемости студентов с использованием методов интеллектуального анализа данных. Выпускная квалификационная работа бакалавра по направлению 02.03.03 «Математическое обеспечение и администрирование информационных систем. Профиль “Технологии программирования”». https://nauchkor.ru/pubs/sistema-prognozirovaniya-uspevaemosti-studentov-s-ispolzovaniem-metodov-intellektualnogo-analiza-dannyh-5f4d0605cd3d3e0001ce9b0f

- 126 студентов (7 учебных групп), обучающихся по дисциплине «Дискретная математика» с формой контроля «экзамен»;
- 102 студента (4 учебные группы), обучающихся по дисциплине «Математический анализ» с формой контроля «зачет».

Сбор, группировка и предварительная обработка исходных данных производились в электронных таблицах Microsoft Excel. Статистическая обработка данных исследования проведена с использованием программ STATISTICA 10.0 и Jamovi 2.2.5.

Для кластеризации наблюдений использовался метод k-средних с настройкой измерения минимальных расстояний (в STATISTICA).

Для изучения связи между количественными переменными, одна из которых имела порядковую природу, использовался метод корреляции Кендалла (в STATISTICA).

Для сравнения парных номинальных переменных применялся критерий Фридмана с апостериорным критерием Дарбина—Коновера (в Jamovi).

Исследовались возможности применения следующих методов для задачи прогнозирования результатов обучения:

- кластеризация k-средних;
- линейная регрессия;
- логистическая регрессия;
- категориальный байесовский классификатор;
- метод случайного леса;
- нейронная сеть (многослойный перцептрон).

Для прогнозирования результатов обучения студентов использовалась непараметрическая оценка функции регрессии по наблюдениям типа Надарая—Ватсона.

5. Результаты исследования

Формализуем понятие **«результат обучения»**. Во-первых, оно непосредственно связано с оценкой в промежуточной аттестации (сессии). Во-вторых, поскольку прохождение промежуточной аттестации студентами может быть растянуто во времени, в течение которого обучающиеся имеют возможность пересдачи после неудачных попыток, то факты пересдач также должны быть учтены. В-третьих, результат обучения должен по-разному учитываться для дисциплин с разными формами контроля (зачет, экзамен). Таким образом, формулу расчета результата обучения можно представить, как показано в (1).

$$y = \begin{cases} CR \cdot 2 + (1 - RE/2), \text{ если ФК } = \text{ зачет;} \\ SC + (1 - RE/2), \text{ если ФК } = \text{ экзамен,} \end{cases} \quad (1)$$

где:

ФК — форма контроля;

$CR = \{0,1\}$ — зачет/незачет;

$SC = \{0,3,4,5\}$ — экзаменационная оценка;

$RE = \{0,1,2\}$ — количество пересдач.

Интерпретация значений результатов обучения y приведена в таблице 1.

Таблица 1 / Table 1

Интерпретация значений результатов обучения

Interpretation of learning outcome values

Для дисциплин с ФК = зачет				
CR	*RE*	*y*	Класс	Интерпретация
0	2	0	1	Зона риска. Крайне неудовлетворительный результат обучения по дисциплине
0	1	0,5		
0	0	1		
1	2	2	2	Зона повышенного внимания. Нежелательный результат обучения по дисциплине
1	1	2,5		
1	0	3	3	Зона успеха. Допустимый и/или желательный результат обучения
Для дисциплин с ФК = экзамен				
SC	*RE*	*y*	Класс	Интерпретация
0	2	0	1	Зона риска. Крайне неудовлетворительный результат обучения по дисциплине
0	1	0,5		
0	0	1		
3	2	3	2	Зона повышенного внимания. Нежелательный результат обучения по дисциплине
3	1	3,5		
3	0	4		
4	2	4		
4	1	4,5		
4	0	5	3	Зона успеха. Допустимый и/или желательный результат обучения
5	2	5		
5	1	5,5		
5	0	6		

5.1. Предпроектные исследования

В качестве показателей текущей учебной деятельности студентов в течение семестра, исходя из доступных данных в эксплуатируемых автоматизированных системах в ИКИТ СФУ [23] и корреляционного анализа зависимостей, были взяты баллы текущей успеваемости за каждую неделю обучения. В таблице 2 представлены фрагменты исходных выборок текущей успеваемости студентов по дисциплинам:

- дискретная математика (с ФК = экзамен);
- математический анализ (с ФК = зачет).

Для дискретной математики баллы текущей успеваемости представлены начиная со 2-й недели, для математического анализа — с 3-й. Баллы текущей успеваемости принимают значения в диапазоне [0, 100].

Приведенное в таблице 1 распределение значений результатов обучения по условным классам произведено эмпирическим путем. Для научного подтверждения или опровержения такого распределения был проведен ряд исследований, и прежде всего — *кластеризация*, которая проводилась для каждой недели с использованием текущей успеваемости за весь предшествующий период, т. е. с учетом истории обучения на предыдущих неделях, а также с учетом итогового результата обучения по дисциплине (y). Таким образом, было получено по 17 и 16 вариантов кластеризации для каждой дисциплины соответственно.

В соответствии с настройками параметров (количество кластеров, метод Варда) кластерный анализ объединил студентов в три кластера. Первый кластер — студенты с низким результатом по дисциплине, второй — со средним, третий — с высоким. Полученные кластеры в основе своей соответствуют эмпирическим классам из таблицы 1 с учетом погрешностей границ между кластерами и между классами. Это обстоятельство позволяет сделать предварительный вывод, что ***классы результатов обучения зависят от текущей успеваемости студентов в семестре***.

Для подтверждения указанного вывода были проведены дополнительные исследования, результаты которых приведены в таблице 3 и которые показывают корреляционные связи между результатом обучения (y) и кластерами, с одной стороны, а с другой — между результатом обучения (y) и баллами текущей успеваемости.

Связь результата обучения (y) экзаменационной дисциплины «Дискретная математика» с кластерами заметно возрастает на 7-й неделе ($r = 0{,}52$) и плавно увеличивается к последней неделе. Связь результата обучения (y) с баллами текущей успеваемости ведет себя аналогично: на 7-й неделе $r = 0{,}71$ и далее растет.

Связь результата обучения (y) зачетной дисциплины «Математический анализ» с кластерами заметно возрастает на 5-й неделе ($r = 0{,}51$), но уменьшается с 8-й недели и далее увеличивается только с 15-й. Связь результатов обучения (y) с баллами заметно растет лишь с 13-й недели.

Резкое изменение корреляции на определенной неделе показывает усиление связи текущей успеваемости и результата обучения (y), что позволяет сделать предположение о целесообразности применения прогнозной модели начиная именно с этой недели. Проведение апостериорного теста Дарбина—Коновера для сравнения кластеризации по всем парам недель показывает схожие результаты, что подтверждает сделанное предположение.

5.2. Исследование прогнозных моделей

При выборе подходящей модели прогнозирования результатов обучения на основе текущей успеваемости в семестре были апробированы и исследованы

Таблица 2 / Table 2

Пример исходных данных по текущей успеваемости и результатам обучения

Example of initial data on current academic performance and learning outcomes

Код студента	Текущая успеваемость по дисциплине «Дискретная математика» в баллах																	*SC*	*RE*	*y*
	Номер недели																			
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18			
221323	0	3	7	7	14	16	16	16	16	16	16	16	16	16	16	17	17	0	2	0
221602	3	7	7	7	11	16	20	24	24	27	35	35	43	43	43	47	51	4	0	5
Код студента	Текущая успеваемость по дисциплине «Математический анализ» в баллах																	*CR*	*RE*	*y*
	Номер недели																			
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18			
221954	—	2	2	2	2	21	12	12	12	12	12	19	31	31	37	37	67	1	0	3
222533	—	2	2	2	2	2	9	9	9	9	9	20	20	28	28	54	77	1	0	3
220791	—	2	2	0	5	5	8	15	15	15	15	15	15	20	20	20	20	0	1	0,5

Таблица 3 / Table 3

Коэффициенты корреляции Кендалла результата обучения (*y*) с кластерами и баллами текущей успеваемости по неделям обучения

Kendall's correlation coefficients of learning outcomes (*y*) with clusters and current performance scores by week of study

Дисциплина	«Дискретная математика»																
Неделя	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Кластеры	0,51	0,45	0,48	0,42	0,39	0,52	0,53	0,54	0,53	0,54	0,54	0,54	0,55	0,55	0,55	0,57	0,57
Баллы	0,26	0,48	0,39	0,46	0,56	0,71	0,73	0,74	0,74	0,74	0,76	0,76	0,77	0,76	0,77	0,77	0,78
Дисциплина	«Математический анализ»																
Неделя	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Кластеры	—	0,24	0,25	0,51	0,51	0,47	0,32	0,26	0,23	0,23	0,23	0,23	0,22	0,37	0,39	0,45	0,41
Баллы	—	0,21	0,26	0,27	0,22	0,29	0,29	0,28	0,29	0,29	0,29	0,39	0,40	0,44	0,43	0,47	0,53

несколько подходящих, по мнению авторов, моделей и подходов:

- линейная регрессия;
- логистическая регрессия;
- байесовский классификатор;
- кластеризация k-средних;
- классификация методом случайного леса;
- непараметрическая оценка функции регрессии;
- нейронные сети.

Для оценки качества прогнозирования результатов обучения разными методами применялась **матрица неточностей** (*англ.* Confusion Matrix) CM_{3x3} [24]. В ней строки соответствуют истинным классам, столбцы — прогнозу, на пересечении — количество совпадающих случаев. Выбор матрицы неточностей обоснован тем, что не все ошибки, допускаемые прогнозной моделью, одинаковы, это зависит от прогнозируемого класса (см. табл. 1). Так, ошибка прогнозирования класса 1 и класса 2 по смыслу не является критичной, с точки зрения применения прогнозной модели на практике, где студенту выдается прогноз его результата обучения на основе его текущей успеваемости. Если в прогнозной модели происходит ошибка для этих классов, то студенту сообщится информация о том, что его ждет негативный результат обучения (класс 1) или нежелательный (класс 2). Таким образом, действия студента будут скорректированы на усиление обучения по дисциплине с неверно прогнозируемым результатом, что должно привести к повышению вероятности успешного освоения данной дисциплины. И, напротив, если неверно спрогнозировать класс 3, т. е. сообщить студенту, что у него прогнозируется успешный результат обучения по дисциплине, а в действительности будет провал, то такая ошибка является критичной и неприемлемой для исследуемой модели. На такой ошибке и был сделан акцент, помимо классических методов расчета при выполнении сравнительного анализа методов прогнозирования.

Подсчитаем общую точность Acc и для каждого класса $c = 1,2,3$ точность Pr_c и полноту Rc_c по следующим формулам:

$$Acc = \frac{\sum_{i=1}^{3} CM_{i,i}}{\sum_{i,j=1}^{3} CM_{i,j}},\quad Pr_c = \frac{CM_{c,c}}{\sum_{i=1}^{3} CM_{c,i}},\quad Rc_c = \frac{CM_{c,c}}{\sum_{i=1}^{3} CM_{i,c}}. \qquad (2)$$

Как отмечено выше, критической ошибкой прогноза является случай, когда дается прогноз успешной сдачи сессии (класс 3), но реальный результат соответствует классу 1. Такую ошибку рассчитаем по формуле:

$$err_{3,1} = \frac{CM_{1,3}}{\sum_{i=1}^{3} CM_{3,i}}. \qquad (3)$$

Примеры расчетов по матрице неточностей для 18-й недели обучения для непараметрической регрессии и кластеризации методом k-средних приведены в таблице 4. Расчеты выполнялись на основании данных по 2 130 студентам различных групп и курсов. Для всех прогнозируемых результатов обучения объем выборки составлял более 10 точек, и по всем дисциплинам выполнялось корректное ведение электронного журнала преподавателями.

На основании формул расчета точности Acc и ошибки прогноза класса 3 вместо класса 1 ($err_{3,1}$) было проведено сравнение прогнозных моделей на обобщенных исходных данных. Результаты приведены на рисунках 1 и 2 соответственно.

Линейная регрессия и кластеризация k-средних показали очень низкую точность прогноза (см. рис. 1) и далее не рассматривались. Нейронные сети, логистическая регрессия и байесовский классификатор не смогли обучиться на исходных выборках реальных данных. Эти модели почти всегда предсказывали самый высокий класс 3, что часто приводило к ложным прогнозам успешной сдачи сессии (см. рис. 2).

ISSN 0234-0453 • ИНФОРМАТИКА И ОБРАЗОВАНИЕ • 2023 • Том 38 № 4

Таблица 4 / Table 4

Примеры расчетов матрицы неточностей для прогнозных моделей

Examples of Confusion Matrix calculations for predictive models

Непараметрическая регрессия								Кластеризация методом k-средних							
Факт	Прогноз			Численная оценка				Факт	Прогноз			Численная оценка			
Класс, c	1	2	3	Pr_c, %	Rc_c, %	Acc, %	$err_{3,1}$, %	Класс, c	1	2	3	Pr_c,%	Rc_c, %	Acc, %	$err_{3,1}$, %
1	74	174	60	57	24	69	3	1	148	85	75	21	48	53	3,8
2	38	393	226	41	60			2	208	213	236	30	32		
3	17	398	1557	84	79			3	359	404	1209	80	61		

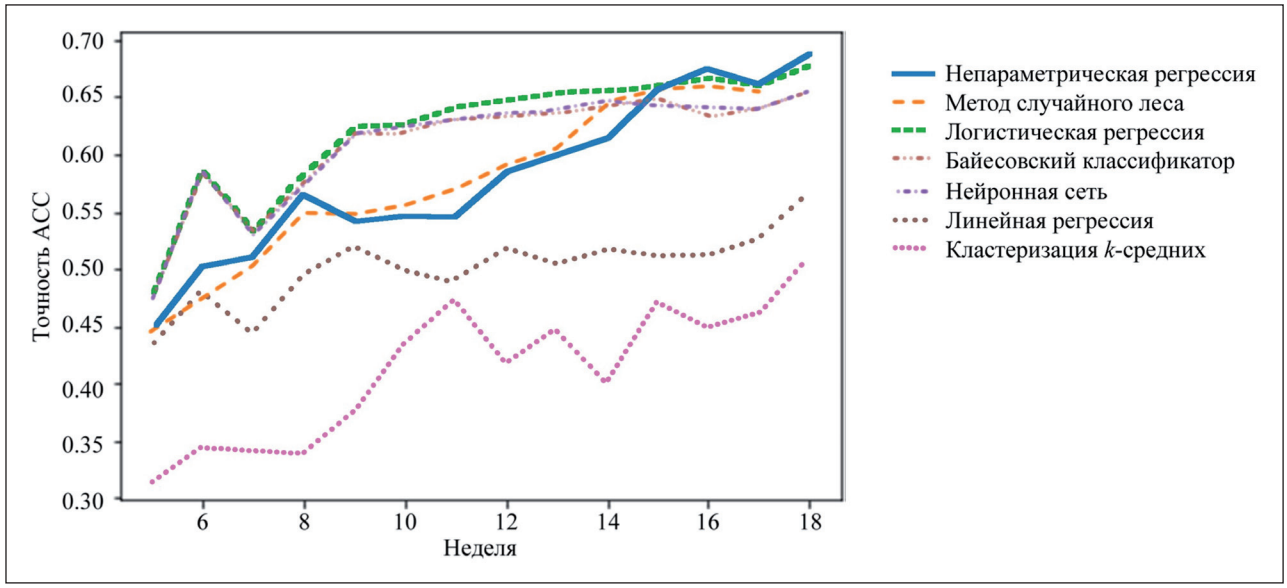

Рис. 1. Точность прогноза при использовании различных алгоритмов

Fig. 1. Prediction accuracy by using various algorithms

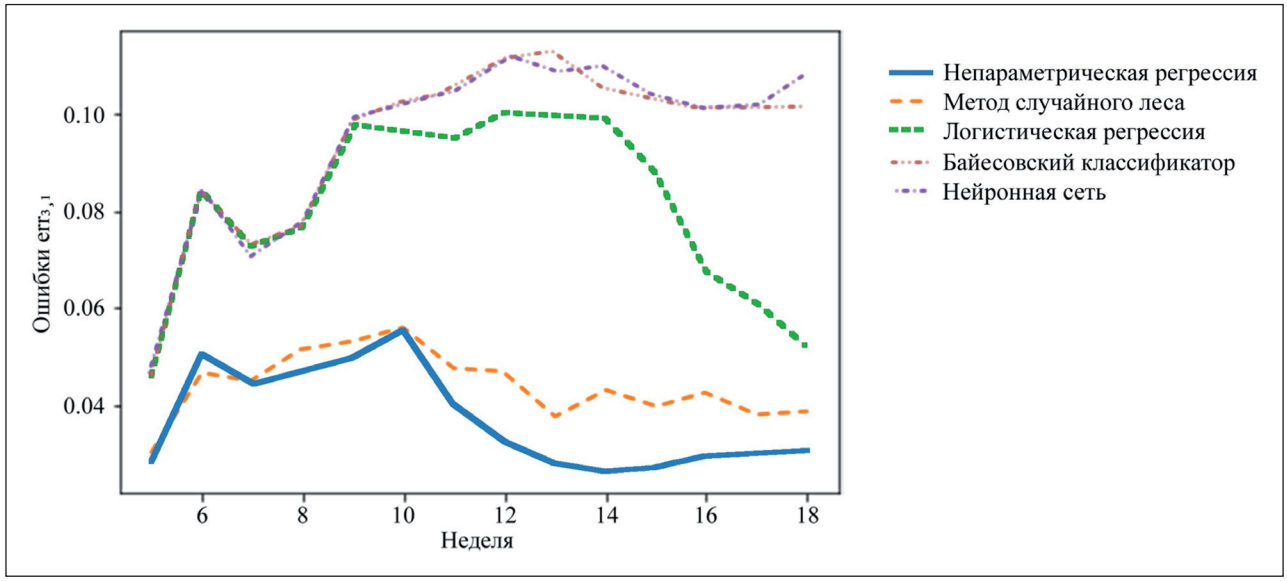

Рис. 2. Ошибочные предсказания класса 3 для класса 1

Fig. 2. Class 3 mispredictions for Class 1

ISSN 0234-0453 • INFORMATICS AND EDUCATION • 2023 • Volume 38 No 4

Классификация методом случайного леса и непараметрическая оценка функции регрессии показали наилучший результат. При этом с 11-й недели непараметрическая регрессия реже дает ошибочные предсказания класса 3.

5.3. Прогнозная модель на базе непараметрической оценки функции регрессии

В результате была выбрана модель, основанная на применении непараметрической оценки функции регрессии Надарая—Ватсона [25], как показавшая наибольшую адекватность решаемой задаче. Рассмотрим эту модель более подробно.

Введем n-мерный вектор входных переменных процесса обучения

$$x^i = \left(x_1^i,\ x_2^i,\ \ldots,\ x_j^i,\ \ldots,\ x_n^i\right),\ i = \overline{1,s},\ j = \overline{1,n},$$

где:

i — это номер студента;

j — номер недели;

s — объем обучающей выборки (количество студентов);

n — количество входных переменных (количество недель для текущего прогноза).

Выходом модели будем считать m-мерный вектор

$$y^i = \left(y_1^i,\ y_2^i,\ \ldots,\ y_k^i,\ \ldots,\ y_m^i\right),$$

где:

$k = \overline{1,m}$ — номер выходных переменных процесса, обозначающий номер результата обучения по дисциплине i-го студента в конкретной сессии конкретной учебной группы или объединения учебных групп в поток в рамках конкретного электронного обучающего курса.

Тогда общая непараметрическая оценка функции регрессии Надарая—Ватсона для многомерной системы примет следующий вид:

$$y_k^{i'}(x) = \frac{\sum\limits_{i=1}^{s} y_k^i \prod\limits_{j=1}^{n} \Phi\left(\dfrac{x_j - x_j^i}{c_s}\right)}{\sum\limits_{i=1}^{s} \prod\limits_{j=1}^{n} \Phi\left(\dfrac{x_j - x_j^i}{c_s}\right)}, \qquad (4)$$

где:

k — номер результата обучения по дисциплине в определенную сессию определенной учебной группы;

i' — номер студента, для которого рассчитывается прогнозное значение результата обучения;

x_j^i, y_k^i — значения переменных из обучающей выборки наблюдений;

x_j — текущая оценка студента на j-й неделе, для которого рассчитывается прогноз y;

c_s — параметр размытости ядра.

Колоколообразные функции Ф() и параметр размытости c_s удовлетворяют следующим условиям сходимости:

$$\Phi() < \infty; \qquad (5)$$

$$\int\limits_{\Omega(x)} \Phi\left(c_s^{-1}(x_j - x_j^i)\right)dx = 1; \qquad (6)$$

$$\lim_{s\to\infty} c_s^{-1}\Phi\left(c_s^{-1}(x_j - x_j^i)\right) = \delta(x_j - x_j^i); \qquad (7)$$

$$\lim_{s\to\infty} c_s = 0; \qquad (8)$$

$$\lim_{s\to\infty} s c_s^n = \infty. \qquad (9)$$

В качестве колоколообразной функции было использовано треугольное ядро, которое можно записать следующим образом:

$$\Phi\left(\frac{x_j - x_j^i}{c_s}\right) = \begin{cases} 1 - \dfrac{|x_j - x_j^i|}{c_s}, \text{ если } \dfrac{|x_j - x_j^i|}{c_s} < 1; \\ 0, \text{ если } \dfrac{|x_j - x_j^i|}{c_s} \geq 1. \end{cases} \qquad (10)$$

Отметим, что выбор вида колоколообразной функции несущественно влияет на точность построения модели. Здесь большую роль играют значения коэффициента размытости c_s.

Ошибку прогноза для предложенной модели будем определять по формуле относительной средней абсолютной ошибки [26]:

$$\delta_k = \frac{\sum\limits_{i=1}^{s} |y_k^i - \hat{y}_k^i|}{\sum\limits_{i=1}^{s} |y_k^i - \bar{y}_k|},\ k = \overline{1,m}, \qquad (11)$$

где:

y_k^i — значения выходных переменных из исходной выборки наблюдений;

$\hat{y}_k^i$ — прогнозируемые значения выходных переменных, которые были найдены по формуле (4);

$\bar{y}_k$ — средние значения выходов объекта из исходной выборки наблюдений.

5.4. Пример применения прогнозной модели

В качестве примера рассмотрим данные об обучении по одной из дисциплин — «Математический анализ». Начиная с 1-й недели обучения будем выполнять прогнозирование выходной переменной, а именно определять класс для студентов, в который они могут попасть со своими результатами обучения. Исходная выборка наблюдений для рассматриваемой дисциплины состояла из 102 наблюдений (студентов). Фрагмент выборки представлен в таблице 2. Производилось нормирование и центрирование исходных данных согласно следующей формуле:

$$\tilde{x}_j^i = \frac{x_j^i - \bar{x}_j}{\sigma x_j}, \qquad (12)$$

где σx_j — среднеквадратичное (стандартное) отклонение (СКО) по признаку, которое равно:

$$\tilde{x}_j^i = \frac{x_j^i - \bar{x}_j}{\sqrt{\dfrac{\sum\limits_{i=1}^{s}(x_j^i - \bar{x}_j)^2}{s-1}}}, \qquad (13)$$

где:

$i = \overline{1,s}$, $j = \overline{1,n}$, x_j^i — текущие значения входных переменных;

$\bar{x}_j$ — среднее значение по каждой входной переменной.

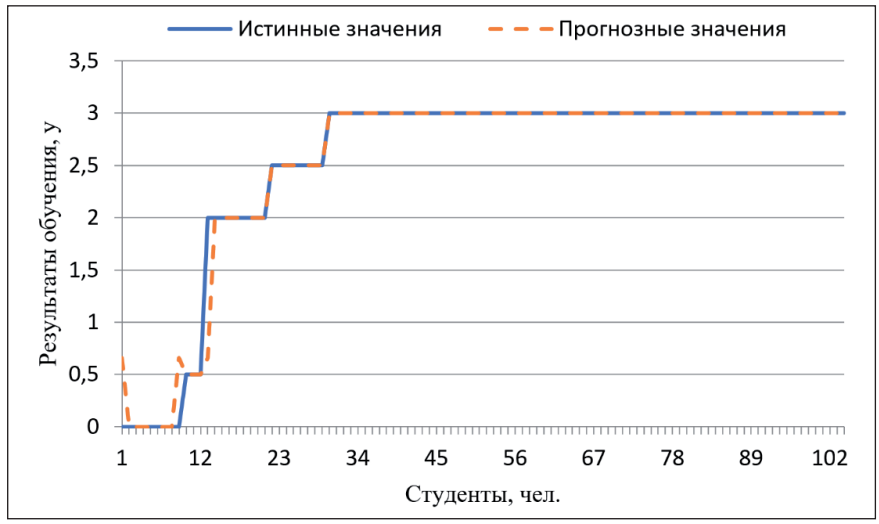

Рис. 3. Сравнение истинных и спрогнозированных значений выходной переменной на 18-й неделе обучения по дисциплине «Математический анализ»

Fig. 3. Comparison of true and predicted values of the output variable at the 18th week of study in the discipline "Mathematical Analysis"

Непараметрическая оценка функции регрессии Надарая—Ватсона для 18-й недели обучения примет следующий вид:

$$\hat{y}^{i'}(x) = \frac{\sum_{i=1}^{102} y^i \prod_{j=1}^{18} \Phi\left(\frac{x_j - x_j^i}{0{,}2}\right)}{\sum_{i=1}^{102} \prod_{j=1}^{18} \Phi\left(\frac{x_j - x_j^i}{0{,}2}\right)}, \qquad (14)$$

где на выходе рассматривается только одна переменная $\hat{y}^{i'}$ которая будет показывать прогноз результата обучения по дисциплине «Математический анализ» на 18-й неделе обучения, а на входе — значения оценок студентов с 1-й по 18-ю недели обучения.

На рисунке 3 показан график, на котором представлены истинные значения выходных переменных (y) и спрогнозированные значения ($\hat{y}$) для 18-й недели обучения.

Здесь сплошная синяя линия показывает истинные значения результатов обучения студентов, а пунктирная оранжевая — спрогнозированные значения, рассчитанные по формуле (14). По оси абсцисс расположены порядковые номера студентов начиная с 1 по 102. По оси ординат — результаты обучения. Для представленного результата использовались значения всех восемнадцати недель обучения; параметр размытости c_s был принят равным 0,2 (параметр был определен из многочисленных экспериментов и при этом показывал наименьшую ошибку прогноза). Здесь представлен точный прогноз, значения модели практически совпадают с истинными значениями. При этом:

- ошибка моделирования (δ), рассчитываемая по формуле (11), равна 0,02;
- ошибка прогнозирования класса 3 вместо класса 1 ($err_{3,1}$), рассчитываемая по формуле (3), равна 0,0 %;
- общая точность прогноза (Acc), рассчитываемая по формуле (2), равна 98,0 %.

Далее приведем график, на котором показана зависимость ошибки прогноза от недели обучения (рис. 4).

На рисунке 4 на оси абсцисс отмечены 18 недель обучения начиная со значений 1-й недели; затем вместе 1-я и 2-я недели; далее 1, 2 и 3-я недели и т. д. По оси ординат отмечена соответствующая ошибка прогноза по формуле (11). Исходя из представленного рисунка, можно сделать вывод, что с ростом недель обучения ошибка прогнозирования уменьшается.

Важной задачей является *определение номера недели обучения*, начиная с которой мы можем делать достоверные прогнозы. Исходя из представленного на рисунке 6 графика, можно увидеть, что ошибка существенно уменьшилась на 13-й неделе обучения и стала равна 0,03. Но это не является своевременной информацией ввиду того, что давать наиболее точный прогноз начиная с 13-й недели обучения уже поздно: студент может не успеть скорректировать свой подход к обучению. На 8-й неделе ошибка равна 0,12, что существенно выше по сравнению с 13-й неделей обучения, но тем не менее значительно ниже предыдущих недель. Поэтому проведем дополнительные исследования, направленные на определение недели обучения с приемлемой точностью прогнозирования предложенной модели.

При ошибке моделирования (δ), равной 0,03, доля неправильно спрогнозированных значений ($1–Acc$) составляет 4,9 %, а при 0,12 — 10,78 %. Здесь необходимо определиться: какой процент неправильного прогноза приемлем, с точки зрения применения модели на практике, когда мы реально информируем студентов о прогнозируемых результатах обучения в течение семестра? Будем считать, что *общий барьер ошибочных прогнозов, равный 15 %, приемлем* и для автоматического определения недели обучения, начиная с которой можно делать прогноз. Восполь-

ISSN 0234-0453 • INFORMATICS AND EDUCATION • 2023 • Volume 38 No 4

Рис. 4. Зависимость величины ошибки прогноза от числа недель обучения
Fig. 4. Dependence of the magnitude of the forecast error on the number of weeks of study

зуемся значениями градиентов между ошибками соседних недель. Таким образом, алгоритм определения недели будет отслеживать резкое изменение градиента и учитывать ограничение в 15 %.

Для рассматриваемого примера по дисциплине «Математический анализ» (см. табл. 2) прогноз возможен уже на 7-й неделе обучения, что видно на рисунке 5.

Здесь ошибка моделирования (δ), рассчитываемая по формуле (11), равна 0,24; ошибка прогнозирования класса 3 вместо класса 1 ($err_{3,1}$), рассчитываемая по формуле (3), равна 7,8 %; общая точность прогноза (*Acc*), рассчитываемая по формуле (2), равна 84,4 %.

Далее приведем результаты экспериментов, в которых в качестве исходной выборки наблюдений (обучающей выборки) использовались результаты обучения предыдущего учебного года, а прогнозируемые результаты рассчитывались для текущего учебного года. На рисунке 6 приведены реальные значения текущего учебного года (представленные сплошной красной линией) и прогнозируемые значения текущего учебного года на 7-й неделе обучения (представленные пунктирной фиолетовой линией).

Ошибка моделирования на 7-й неделе обучения составляет δ = 0,64, а процент неправильных прогнозных значений — 15 %, что является приемлемым результатом, с практической точки зрения. Здесь ошибка прогнозирования класса 3 вместо класса 1 ($err_{3,1}$) равна 5,9 %; точность (*Acc*) составляет 85,0 %. Для 13-й недели обучения ошибка моделирования (δ) равна 0,47; ошибка прогнозирования класса 3 вместо класса 1 ($err_{3,1}$) равна 4,4 %; общая точность прогноза (*Acc*) равна 89,6 %. Для 18-й недели обучения ошибка моделирования (δ) равна 0,42; ошибка прогнозирования класса 3 вместо класса 1 ($err_{3,1}$) равна 3,7 %; общая точность прогноза (*Acc*) равна 92,6 %. Описанные показатели моделирования проведенных экспериментов представлены в таблице 5.

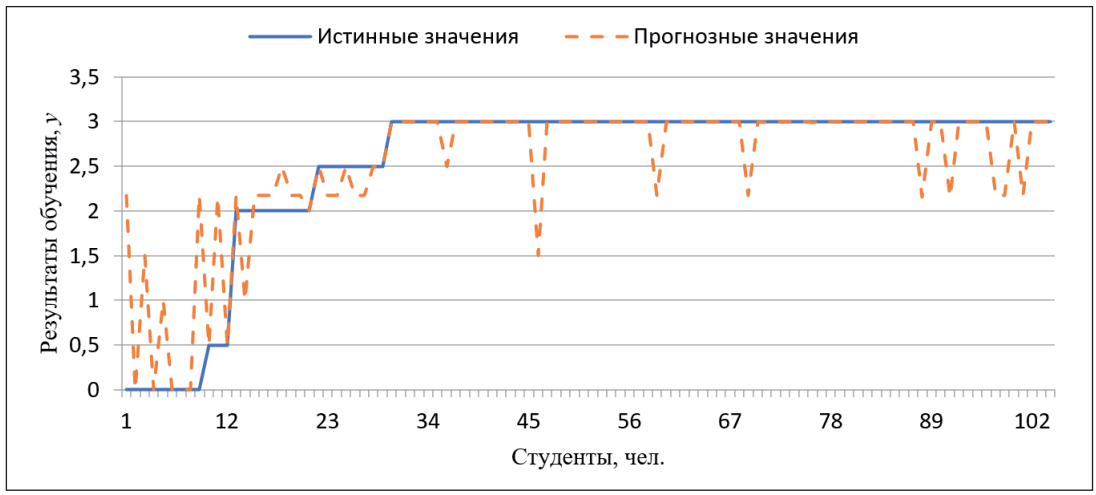

Рис. 5. Сравнение истинных и прогнозных значений выходной переменной на 7-й неделе обучения
Fig. 5. Comparison of true and predicted values of the output variable at the 7th week of studying

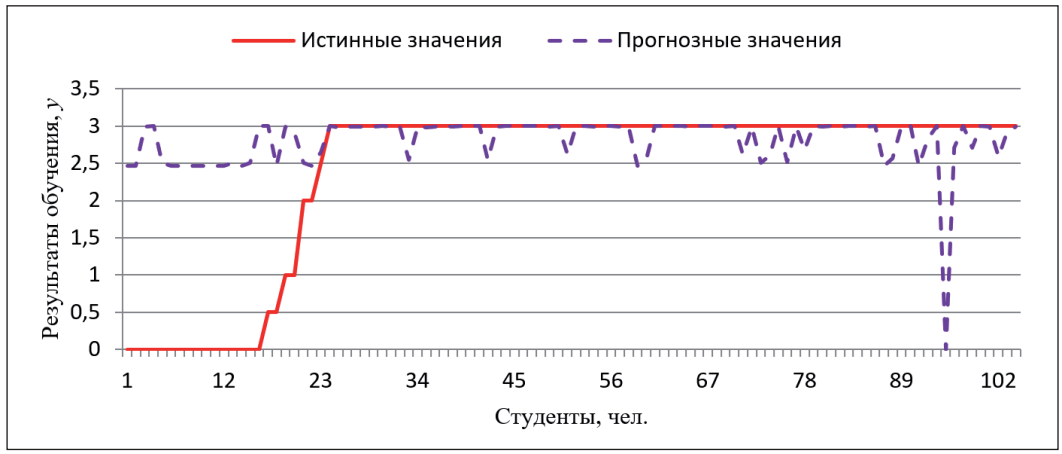

Рис. 6. Сравнение истинных и прогнозных значений выходной переменной на 7-й неделе обучения с использованием обучающей выборки предыдущего года обучения

Fig. 6. Comparison of true and predicted values of the output variable at the 7th week of study using the training sample of the previous year of study

Таблица 5 / Table 5

Индикаторы экспериментов по прогнозированию результатов обучения на разных неделях

Indicators of experiments on predicting learning outcomes in different weeks

Неделя обучения	Индикаторы		
	δ	*Acc*, %	$err_{3,1}$, %
7-я	0,64	85,0	5,9
13-я	0,47	89,6	4,4
18-я	0,42	92,6	3,7

Проведенные исследования позволили применить полученные результаты на практике и разработать **мобильное приложение «Студент СФУ»***, позволяющее прогнозировать результаты обучения и показывать прогноз в онлайн-режиме каждому студенту по каждой дисциплине (рис. 7).

Кроме того, данное приложение позволяет:

- информировать студентов о текущей ситуации по дисциплинам и их позициях в рейтингах в группе, на потоке;
- вести электронную зачетную книжку с информированием о предстоящих контрольных мероприятиях и незакрытых контрольных точках по дисциплинам;
- просматривать расписание занятий.

* Свидетельство о государственной регистрации программы для ЭВМ № 2022681221 Российская Федерация. Мобильное приложение «Студент СФУ»: № 2022666780: заявлено 13.09.2022: опубликовано 10.11.2022 / Якунин Ю. Ю., Шапошник С. С.; правообладатели: Якунин Ю. Ю., Шапошник С. С. Зарегистрировано в реестре программ для ЭВМ. https://new.fips.ru/registers-doc-view/fips_servlet?DB=EVM&DocNumber=2022681221&TypeFile=html

6. Обсуждение результатов

Ретроспективная кластеризация студентов по неделям обучения на потенциально успешно аттестованных, средних и имеющих проблемы с аттестацией по дисциплине позволила, во-первых, предварительно оценить возможности кластеризации оперативных данных (баллов в электронных курсах); во-вторых, апробировать поиск недели, с которой можно доверять прогнозу относительно аттестации студентов по дисциплине.

Кластеризация дала понять, что наблюдается прямая умеренная связь прогнозируемого результата обучения студента с текущими оценками в электронных образовательных курсах. Седьмая неделя является ключевой: начиная с нее обнаруживается рост связи кластеров и текущих оценок с результатом обучения. Чем ближе к последней неделе, тем точнее прогноз. Для дисциплин с формой контроля «зачет» связь прогноза результата обучения с фактом слабее и нестабильна. Вероятно, в аттестации по зачетной дисциплине действуют не учтенные в электронном курсе факторы.

Попарное сравнение кластеров по неделям позволило оценить динамику смещения студентов из кластера в кластер. Для экзаменационной дисциплины с 7-й недели студенты относительно стабильны в занятом кластере, но кластер последней недели эту стабильность прерывает. Возможно, начисление баллов в конце курса носит нетипичный для курса характер. Для зачетной дисциплины стабильность кластеров наступает на неделю позже экзаменационной дисциплины, и миграция по кластерам свойственна пяти последним неделям курса.

Сделаем вывод, что прогнозировать результат обучения студентов можно, но нужно учитывать форму контроля: для экзамена качество прогноза лучше, чем для зачета. Седьмая неделя курса — временной ориентир для начала анализа баллов в курсе

ISSN 0234-0453 • INFORMATICS AND EDUCATION • 2023 • Volume 38 No 4

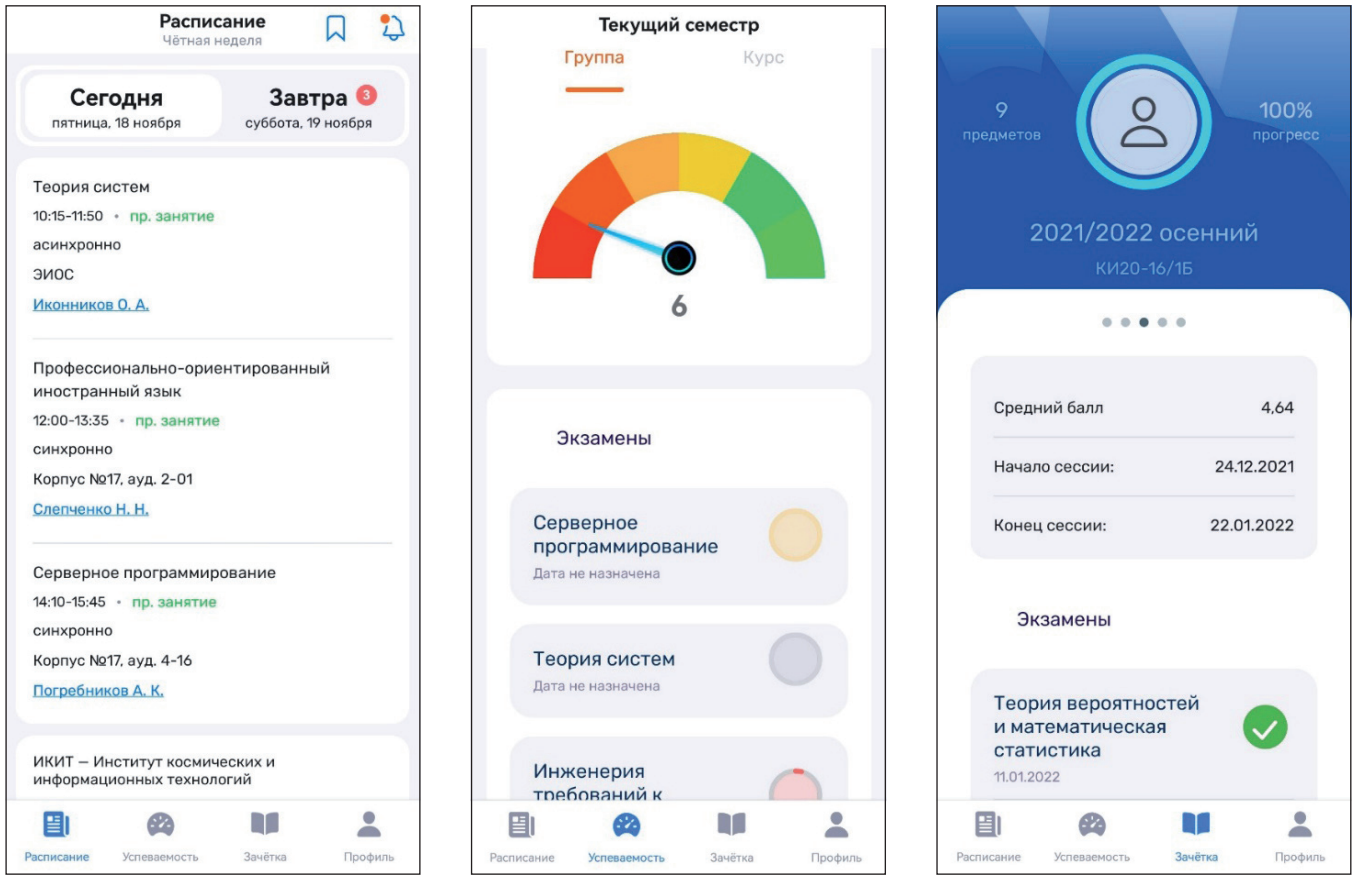

Рис. 7. Экранные формы мобильного приложения «Студент СФУ»
Fig. 7. Screen forms of the "Student of SibFU" mobile application

и прогноза. В зачетной дисциплине последние недели курсов менее прогнозируемы.

Для прогнозирования результата обучения была использована непараметрическая статистика, которая относится к классу локальных аппроксимаций. Отметим, что теорию непараметрических систем целесообразно использовать, когда априорная информация об исследуемом процессе мала, т. е. отсутствуют дополнительные данные, которые могли бы послужить основой для математической постановки задачи. В рассматриваемой задаче исследователь располагает только исходными оценками по дисциплине на каждой неделе обучения, что приводит к необходимости использовать данные методы.

Кроме того, очень многое зависит от исходной выборки наблюдений. При этом проводились эксперименты с малыми группами студентов (14–25 человек), которые показали приемлемые результаты. Но в случае, если до 7-й недели обучения оценивание студентов не производилось или были использованы оценки, близкие к нулю, то прогноз по таким дисциплинам не удавалось получить.

Следует обратить внимание на то, что результаты, полученные в данном исследовании, согласуются с работой [3], в которой авторы также используют для построения моделей прогнозирования средние баллы, баллы текущей и промежуточных аттестаций и итоговые оценки, или работой [5], где происходит построение моделей на основе оценки входных знаний по дисциплине, оценки знаний по первой теме дисциплины, числа пропусков на момент прогноза. Однако применение непараметрической оценки функции регрессии Надарая—Ватсона для прогнозирования результатов обучения студентов в литературе не встречается.

Таким образом, ***предложенный метод можно охарактеризовать как новый подход к прогнозированию результатов обучения с применением показателей учебного процесса по неделям***. Этот метод позволяет построить точный прогноз, который поможет студенту вовремя обратить внимание на дисциплины с низкой успеваемостью и предпринять меры по ее улучшению.

7. Заключение

В результате проведенного исследования была построена прогнозная модель результатов обучения студентов, которая помогает информировать обучающихся о прогнозируемых результатах обучения по каждой дисциплине на основании данных о текущей успеваемости из электронных образовательных курсов. Как показали результаты исследования, по большей части дисциплин прогноз можно произво-

дить начиная с 7-й недели обучения, что является хорошим результатом и позволяет сигнализировать студентам о возможных проблемах на ранней стадии, когда есть возможность исправить положение и пройти промежуточную аттестацию успешно.

Следует отметить, что по мере увеличения номера недели обучения возрастает точность прогнозной модели, но даже на ранних стадиях она составляет более 85 %. Кроме того, для критической ошибки ($err_{3,1}$), показывающей ошибочно класс 3 (успешный) вместо класса 1 (неуспешный), точность модели на ранних стадиях в среднем выше 94 %.

Полученные результаты позволили создать удобное мобильное приложение для студентов на базе данных, доступных из онлайн-курсов. Приложение позволяет студентам самостоятельно контролировать вероятные результаты обучения на каждой неделе и вовремя корректировать внимание и усилия по изучению дисциплин и другим внеучебным мероприятиям.

В развитие созданной модели и мобильного приложения предполагается реализовать:

- автоматический анализ пригодности данных для прогноза;
- учет дополнительных параметров для повышения точности модели, таких как:
  – оценки за курсовые проекты/работы;
  – посещаемость занятий;
  – активность в ЭОК;
- динамическое определение недели обучения, в которой можно делать прогноз с заданной точностью;
- интеллектуальный подбор и подготовку исторических данных для формирования обучающей выборки прогнозной модели для повышения ее точности.

Список источников / References

1. *Alsariera Y. A., Baashar Y., Alkawsi G., Mustafa A., Alkahtani A. A., Ali N. A.* Assessment and evaluation of different machine learning algorithms for predicting student performance. *Computational Intelligence and Neuroscience*. 2022;1:4151487. DOI: 10.1155/2022/4151487

2. *Brahim G. B.* Predicting student performance from online engagement activities using novel statistical features. *Arabian Journal for Science and Engineering*. 2022;47(8):10225–10243. DOI: 10.1007/s13369-021-06548-w

3. *Токтарова В. И., Пашкова Ю. А.* Предиктивная аналитика в цифровом образовании: анализ и оценка успешности обучения студентов. *Сибирский педагогический журнал*. 2022;(1):97–106. EDN: NSPMDR. DOI: 10.15293/1813-4718.2201.09

[*Toktarova V. I., Pashkova Y. A.* Predictive analytics in digital education: Analysis and evaluation of students' learning success. *Siberian Pedagogical Journal*. 2022;(1):97–106. (In Russian.) EDN: NSPMDR. DOI: 10.15293/1813-4718.2201.09]

4. *Almasri A., Celebi E., Alkhawaldeh R. S.* EMT: Ensemble meta-based tree model for predicting student performance. *Scientific Programming*. 2019:1–13. DOI: 10.1155/2019/3610248

5. *Шевченко В. А.* Прогнозирование успеваемости студентов на основе методов кластерного анализа. *Вестник Харьковского национального автомобильно-дорожного университета*. 2015;(68):15–18. EDN: UNSPAX

[*Shevchenko V. A.* Prognostication of students progress on the basis of cluster analysis methods. *Bulletin of Kharkov National Automobile and Highway University*. 2015;(68):15–18. (In Russian.) EDN: UNSPAX]

6. *Aggarwal D., Mittal S., Bali V.* Significance of non-academic parameters for predicting student performance using ensemble learning techniques. *International Journal of System Dynamics Applications*. 2021;10(3):38–49. DOI: 10.4018/IJSDA.2021070103

7. *Zohair L. M. A.* Prediction of student's performance by modelling small dataset size. *International Journal of Educational Technology in Higher Education*. 2019;(16):1–18. DOI: 10.1186/s41239-019-0160-3

8. *Куприянов Р. Б., Звонарев Д. Ю.* Повышение качества модели прогнозирования образовательных результатов студентов университетов. *Информатика и образование*. 2021;36(9(328)):40–46. EDN: WAAWTP. DOI: 10.32517/0234-0453-2021-36-9-40-46

[*Kupriyanov R. B., Zvonarev D. Yu.* Improving the quality of the university students' academic performance prediction model. *Informatics and Education*. 2021;36(9(328)):40–46. (In Russian.) EDN: WAAWTP. DOI: 10.32517/0234-0453-2021-36-9-40-46]

9. *Куприянов Р. Б., Звонарев Д. Ю.* Разработка модели прогнозирования образовательных результатов обучающихся для университетов. *Искусственный интеллект и принятие решений*. 2021;(2):11–20. EDN: EENEFV. DOI: 10.14357/20718594210202

[*Kupriyanov R. B., Zvonarev D. Yu.* Development of the students' educational success prediction model for universities. *Artificial Intelligence and Decision Making*. 2021;(2):11–20. (In Russian.) EDN: EENEFV. DOI: 10.14357/20718594210202]

10. *Deeva G., De Smedt J., Saint-Pierre C., Weber R., De Weerdt J.* Predicting student performance using sequence classification with time-based windows. *Expert Systems with Applications*. 2022;(209):118182. DOI: 10.1016/j.eswa.2022.118182

11. *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;14(4):815–822. EDN: JWGGFH. DOI: 10.25559/SITITO.14.201804.815-822

[*Rusakov S. V., Rusakova O. L., Posokhina K. A.* Neural network model of predicting the risk group for the accession of students of the first course. *Modern Information Technologies and IT-Education*. 2018;14(4):815–822. (In Russian.) EDN: JWGGFH. DOI: 10.25559/SITITO.14.201804.815-822]

12. *Ali R. H.* Educational data mining for predicting academic student performance using active classification. *Iraqi Journal of Science*. 2022;63(9):3954–3965. DOI: 10.24996/ijs.2022.63.9.27

13. *Li S., Liu T.* Performance prediction for higher education students using deep learning. *Complexity*. 2021:1–10. DOI: 10.1155/2021/9958203

14. *Poudyal S., Mohammadi-Aragh M. J., Ball J. E.* Prediction of student academic performance using a hybrid 2D CNN model. *Electronics*. 2022;11(7):1–21. DOI: 10.3390/electronics11071005

15. *Sood S., Saini M.* Hybridization of cluster-based LDA and ANN for student performance prediction and comments evaluation. *Education and Information Technologies*. 2021;26(3):2863–2878. DOI: 10.1007/s10639-020-10381-3

16. *Tsiakmaki M., Kostopoulos G., Kotsiantis S., Ragos O.* Transfer learning from deep neural networks for predicting student performance. *Applied Sciences*. 2020;10(6):2145. DOI: 10.3390/app10062145

17. *Liu Y., Fan S., Xu S., Sajjanhar A., Yeom S., Wei Y.* Predicting student performance using clickstream data and machine learning. *Education Sciences*. 2023;13(1):1–17. DOI: 10.3390/educsci13010017

18. *Conijn R., Van den Beemt A., Cuijpers P.* Predicting student performance in a blended MOOC. *Journal of Computer Assisted Learning*. 2018;34(5):615–628. DOI: 10.1111/jcal.12270

19. *Шухман А. Е., Парфенов Д. И., Легашев Л. В., Гришина Л. С.* Анализ и прогнозирование успеваемости обучающихся при использовании цифровой образовательной среды. *Высшее образование в России*. 2021;30(8-9):125–133. EDN: QKRTNV. DOI: 10.31992/0869-3617-2021-30-8-9-125-133

[*Shukhman A. E., Parfenov D. I., Legashev L. V., Grishina L. S.* Analysis and forecasting students' academic performance using a digital educational environment. *Higher Education in Russia*. 2021;30(8-9):125–133. (In Russian.) EDN: QKRTNV. DOI: 10.31992/0869-3617-2021-30-8-9-125-133]

20. *Qiu F., Zhang G., Sheng X., Jiang L., Zhu L., Xiang Q., Jiang B., Chen P.* Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*. 2022;12(1):1–15. DOI: 10.1038/s41598-021-03867-8

21. *Котова Е. Е.* Прогнозирование успешности обучения в интегрированной образовательной среде с применением инструментов онлайн-аналитики. *Компьютерные инструменты в образовании*. 2019;(4):55–80. EDN: GSJKDN. DOI: 10.32603/2071-2340-2019-4-55-80

[*Kotova E. E.* Prediction of learning success in an integrated educational environment using online analytics tools. *Computer Tools in Education Journal*. 2019;(4):55–80. (In Russian.) EDN: GSJKDN. DOI: 10.32603/2071-2340-2019-4-55-80]

22. *Al-Kindi I., Al-Khanjari Z.* Tracking student performance tool for predicting students EBPP in online courses. *International Journal of Emerging Technologies in Learning (iJET)*. 2021;16(23):140–157. DOI: 10.3991/ijet.v16i23.25503

23. *Погребников А. К., Шестаков В. Н., Якунин Ю. Ю.* Влияние использования элементов персональной образовательной среды на успеваемость студентов и их мотивацию к обучению. *Информатика и образование*. 2020;(1(310)):42–50. EDN: LTZJIQ. DOI: 10.32517/0234-0453-2020-35-1-42-50

[*Pogrebnikov A. K., Shestakov V. N., Yakunin Yu. Yu.* The influence of using parts of personal learning environment on student performance and learning motivation. *Informatics and Education*. 2020;(1(310)):42–50. (In Russian.) EDN: LTZJIQ. DOI: 10.32517/0234-0453-2020-35-1-42-50]

24. *Lewis H. G., Brown M.* A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*. 2001;22(16):3223–3235. DOI: 10.1080/01431160152558332

25. *Надарая Э. А.* Непараметрическое оценивание плотности вероятностей и кривой регрессии. Тбилиси: Изд-во ТГУ; 1983. 194 с.

[*Nadaraya E. A.* Nonparametric estimation of probability density and regression curve. Tbilisi, Tbilisi University; 1983. 194 p. (In Russian.)]

26. *Hyndman R. J., Koehler A. B.* Another look at measures of forecast accuracy. *International Journal of Forecasting*. 2006;22(4):679–688. DOI: 10.1016/j.ijforecast.2006.03.001

Информация об авторах

Якунин Юрий Юрьевич, канд. тех. наук, доцент, зав. базовой кафедрой интеллектуальных систем управления, Институт космических и информационных технологий, Сибирский федеральный университет, г. Красноярск, Россия; *ORCID:* https://orcid.org/0000-0002-2330-2963; *e-mail*: yakuninyy@mail.ru

Шестаков Вячеслав Николаевич, канд. филос. наук, доцент кафедры современных образовательных технологий, Институт педагогики, психологии и социологии, Сибирский федеральный университет, г. Красноярск, Россия; *ORCID:* https://orcid.org/0000-0001-7737-2900; *e-mail*: vshestakov@sfu-kras.ru

Ликсонова Дарья Игоревна, канд. тех. наук, доцент базовой кафедры интеллектуальных систем управления, Институт космических и информационных технологий, Сибирский федеральный университет, г. Красноярск, Россия; *ORCID:* https://orcid.org/0000-0001-9663-6481; *e-mail*: liksonovadi@yandex.ru

Даничев Алексей Александрович, канд. тех. наук, доцент базовой кафедры интеллектуальных систем управления, Институт космических и информационных технологий, Сибирский федеральный университет, г. Красноярск, Россия; *ORCID:* https://orcid.org/0000-0001-9830-5205; *e-mail*: adanichev@sfu-kras.ru

Information about the authors

Yuriy Yu. Yakunin, Candidate of Sciences (Engineering), Docent, Head of the Basic Department of Intellectual Control Systems, School of Space and Information Technology, Siberian Federal University, Krasnoyarsk, Russia; *ORCID:* https://orcid.org/0000-0002-2330-2963; *e-mail:* yakuninyy@mail.ru

Viacheslav N. Shestakov, Candidate of Sciences (Philosophy), Associate Professor at the Department of Modern Educational Technologies, School of Education Science, Psychology and Sociology, Siberian Federal University, Krasnoyarsk, Russia; *ORCID:* https://orcid.org/0000-0001-7737-2900; *e-mail:* vshestakov@sfu-kras.ru

Darya I. Liksonova, Candidate of Sciences (Engineering), Associate Professor at the Basic Department of Intellectual Control Systems, School of Space and Information Technology, Siberian Federal University, Krasnoyarsk, Russia; *ORCID:* https://orcid.org/0000-0001-9663-6481; *e-mail:* liksonovadi@yandex.ru

Aleksey A. Danichev, Candidate of Sciences (Engineering), Associate Professor at the Basic Department of Intellectual Control Systems, School of Space and Information Technology, Siberian Federal University, Krasnoyarsk, Russia; *ORCID:* https://orcid.org/0000-0001-9830-5205; *e-mail:* adanichev@sfu-kras.ru

Поступила в редакцию / Received: 02.03.23.
Поступила после рецензирования / Revised: 31.05.23.
Принята к печати / Accepted: 06.06.23.