

УДК 004.942

DOI: [10.26102/2310-6018/2026.60.9.018](https://doi.org/10.26102/2310-6018/2026.60.9.018)

Аналитика обучения: прогнозирование успеваемости на основе данных об активности и эффективности студентов в системах управления электронными образовательными курсами

А.А. Даничев, Д.И. Ликсонова✉, В.Н. Шестаков

Сибирский федеральный университет, Красноярск, Российская Федерация

Резюме. В настоящем исследовании рассматривается задача повышения точности прогнозирования результатов обучения студентов вузов на основе данных о текущей успеваемости по дисциплинам из электронных образовательных курсов. Исходные данные представляют собой числовые значения – оценки и показатели активности по итогам каждой недели обучения, которые могут относиться к разным диапазонам шкал и быть сильно разреженными, что осложняет их дальнейшее использование в алгоритмах прогнозирования. Поэтому необходимо проводить качественную предобработку исходной информации для построения прогнозной математической модели результатов обучения. В качестве одного из методов для обработки данных предлагается использование процентилей, которые позволяют эффективно описывать распределение признаков. Перцентильный подход решает проблему сравнения студентов при различной сложности курсов, переводя абсолютные оценки в относительные ранги внутри группы обучающихся. В исследовании проводится сравнительный анализ методологий обработки данных, включая стандартный метод центрирования-нормирования и перцентильный подход. Для построения прогнозной модели используется метод непараметрической оценки функции регрессии Надарая-Ватсона. Приводятся фрагменты вычислительных экспериментов с разным количеством групп, которые показывает достаточно приемлемые результаты с практической точки зрения.

Ключевые слова: образовательные данные, аналитика обучения, прогнозирование успеваемости учащихся, интеллектуальный анализ образовательных данных, методы машинного обучения, предобработка данных, нормирование, перцентили, электронная образовательная среда.

Для цитирования: Даничев А.А., Ликсонова Д.И., Шестаков В.Н. Аналитика обучения: прогнозирование успеваемости на основе данных об активности и эффективности студентов в системах управления электронными образовательными курсами. *Моделирование, оптимизация и информационные технологии*. 2026;14(9). URL: <https://moitvvt.ru/ru/journal/article?id=2412>
DOI: 10.26102/2310-6018/2026.60.9.018

Learning analytics: predicting student performance based on student activity and performance data in e-learning course management systems

A.A. Danichev, D.I. Liksonova✉, V.N. Shestakov

Siberian Federal University, Krasnoyarsk, the Russian Federation

Abstract. This study examines the problem of increasing the accuracy of predicting the learning outcomes of university students based on data on current academic performance in disciplines from electronic educational courses. The source data are numerical values – grades and activity indicators based on the results of each week of training – which may relate to different scale ranges and be highly sparse, making their subsequent use in forecasting algorithms difficult. Therefore, it is necessary to conduct high-quality preprocessing of the source information to construct a predictive mathematical

model of learning outcomes. One proposed method for data processing is the use of percentiles, which effectively describe the distribution of features. The percentile method solves the issue of fairly comparing students enrolled in courses with differing difficulty levels by turning raw scores into relative rankings within a learner's peer group. This study presents a side-by-side comparison of two data processing techniques: the conventional centering-and-norming method versus the percentile-based strategy. A nonparametric estimator designed for the Nadaraya-Watson regression function is then employed to build the predictive model. Excerpts from computational experiments with varying numbers of groups are presented, yielding acceptable results from a practical perspective.

Keywords: educational data, learning analytics, student performance prediction, educational data mining, machine learning methods, data preprocessing, normalization, percentiles, electronic educational environment.

For citation: Danichev A.A., Liksonova D.I., Shestakov V.N. Learning analytics: predicting student performance based on student activity and performance data in e-learning course management systems. *Modeling, Optimization and Information Technology*. 2026;14(9). (In Russ.). URL: <https://moitvvt.ru/ru/journal/article?id=2412> DOI: 10.26102/2310-6018/2026.60.9.018

Введение

В настоящее время одно из ключевых направлений исследований связано с тем, как можно предсказывать академическую успеваемость студентов, опираясь на сведения об их активности и эффективности в рамках образовательных онлайн-платформ. Для этого задействуются алгоритмы машинного обучения, позволяющие прогнозировать итоговые показатели обучения.

Цифровые технологии все активнее внедряются в учебный процесс, поэтому возникает необходимость выяснить, каким образом поведение учащихся соотносится с их оценками и общим уровнем успеха. Преподаватели, а также организаторы учебной деятельности анализируют широкий спектр данных – от базовой демографии до степени вовлеченности в онлайн-занятия. Такой подход дает возможность выстраивать персонализированные траектории обучения, повышающие успеваемость и удерживающие студентов в программе.

Развитие предиктивного моделирования в образовательной сфере прошло путь от универсальных шаблонов («один вариант для всех») до сложных систем, использующих разнородные массивы информации. Как показали недавние работы [1], такие фоновые признаки, как материальное положение семьи или уровень образования родителей, заметно сказываются на точности прогнозов. Это лишний раз подтверждает: при интеллектуальном анализе образовательных данных и в аналитике обучения обязательно нужно учитывать контекст.

Среди широко применяемых методов машинного обучения – регрессионные модели и нейронные сети. С их помощью удастся выявлять неочевидные зависимости между учебной активностью и реальными достижениями студентов, а значит – глубже понимать индивидуальные особенности учебного поведения [2].

Тем не менее, пока сохраняются сложности с получением достаточно точных и универсальных прогностических моделей, которые одинаково хорошо работали бы в разных образовательных условиях. Внедрение цифровых инструментов добавило к понятию студенческой вовлеченности новые грани, что требует постоянной доработки алгоритмов под многообразные учебные контексты [3].

В целом применение прогностической аналитики в учебных заведениях – это серьезный шаг к управлению на основе данных. Учителя и кураторы получают возможность гибко настраивать меры поддержки и вмешательства с учетом реальных потребностей учеников. Развитие исследований в этой области сулит как повышение

академических результатов, так и более глубокое осмысление факторов, влияющих на успеваемость.

Широкое распространение цифровых технологий и онлайн-площадок для учебы усложнило задачу предсказания успеваемости. Научные данные говорят о том, что разные виды информации – от логов действий до опросных листов – играют важную роль при анализе поведения и учебных достижений в смешанной или полностью дистанционной среде [2].

В последнее время исследователи сосредоточились на применении продвинутых методов машинного обучения для предсказания академических результатов. Кто-то использовал регрессионные модели, а кто-то проверял, насколько эффективны сверточные нейронные сети для тех же целей. Общая задача таких моделей – установить надежную связь между действиями студентов и их итоговой успеваемостью [4, 5]. Проблема соотношения поведения при обучении с успеваемостью остается актуальной, поскольку существующие модели часто не поддаются обобщению в различных образовательных контекстах [6].

Появляется все больше литературы, изучающей взаимосвязь между вовлеченностью студентов в онлайн-обучение и академической успеваемостью, последовательно показывающей, что более высокий уровень вовлеченности коррелирует с улучшением результатов [7, 8].

В метаанализе систематических обзоров направлений, методов и алгоритмов анализа данных в сфере образования, самыми популярными темами исследования стали: успеваемость студентов, поддержка и эффективность работы преподавателей, отсеб студентов, поведение и особенности учащихся, электронная образовательная среда, рекомендации по дисциплине [9]. Среди методов самыми частотными оказались регрессия, классификация, кластеризация, статистический анализ, прогнозирование, визуализация. А вот алгоритмы анализа образовательных данных с убывающей частотностью: дерево решений, наивный байесовский классификатор, нейронная сеть, SVM, ассоциативные правила, логистическая регрессия, случайный лес, KNN.

В [10] описаны типичные наборы образовательных данных, их возможные источники и наиболее распространенные методы анализа данных в образовании, а также приводятся некоторые прогнозы относительно будущих тенденций анализа образовательных данных (EDM – Education Data Mining), особенно в отношении его интеграции в технологии искусственного интеллекта и тенденций развития, таких как дополненная и виртуальная реальность.

Еще один обзор [11] преследует цель систематически изучить текущее состояние области EDM с акцентом на методы глубокого обучения, которые используются в четырех основных образовательных сценариях: отслеживание знаний, анализ поведения учащихся, прогнозирование академической успеваемости и создание персонализированных рекомендаций. Также представлен анализ доступных наборов данных и инструментов для обработки данных в EDM.

Кроме того, поскольку образовательные учреждения все чаще используют стратегии, основанные на данных, роль аналитики обучения будет расти, позволяя педагогам создавать более захватывающие и индивидуализированные учебные процессы.

Эффективное использование аналитики данных в контексте электронного обучения позволяет преподавателям выявлять пробелы в обучении, адаптировать вмешательства и способствовать улучшению образовательных практик, в конечном итоге повышая успеваемость учащихся в условиях все более сложной среды обучения.

Цитируемые исследования и обзоры дают возможность сориентироваться в наборах данных и методах, однако результаты конкретных экспериментов в

значительной степени определяются комбинацией «данные – метод». Уникальность данных этой статьи заключается в прогнозировании успеваемости как в небольших академических группах, так и на объемных потоковых дисциплинах. Дополнительными, редко встречающимися в других исследованиях особенностями являются понедельное прогнозирование и ассортимент применяемых методов нормирования данных.

Таким образом, целью настоящего исследования является повышение точности прогнозирования успеваемости на основе данных об активности и текущей успеваемости студентов, доступных в системах управления образовательными электронными курсами Moodle. Среди задач следует выделить основные, а именно: проанализировать прогностический потенциал активности студентов в электронных курсах вуза; провести обработку исходных данных; найти оптимальные статистики для анализа активности студентов; объединить активность и текущую успеваемость для прогнозирования успешности окончания курсов.

Материалы и методы

В настоящей работе качество исходных данных является ключевым фактором, определяющим эффективность модели прогнозирования успеваемости, включающей элементы активности студентов. Поэтому этап их подготовки и предварительной обработки является критически важным компонентом в исследовании.

В качестве примера рассмотрим реальные данные о текущей успеваемости студентов в течение семестра по дисциплине «Иностранный язык (в рамках унифицированной рабочей программы – Часть 2)», которые были получены из автоматизированной системы Института космических и информационных технологий Сибирского федерального университета (ИКИТ СФУ) (Таблица 1).

Таблица 1 – Фрагмент исходных данных по текущей успеваемости по дисциплине «Иностранный язык»

Table 1 – A fragment of the initial data on the current academic performance in the Foreign Language discipline

Код студента	Текущая успеваемость по дисциплине «Иностранный язык» в баллах																	
	Номер недели																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
222558	–	–	4	5	5	7	8	19	24	24	31	35	35	35	40	60	75	75
222560	–	–	4	4	4	4	4	10	10	11	18	18	18	18	18	20	37	54
222567	–	1	1	1	1	1	1	3	5	11	11	14	14	14	14	14	14	24

Анализ Таблицы 1 позволяет увидеть ряд характерных проблем в исходных данных. Во-первых, набор содержит пропущенные значения, причиной которых могут быть как технические сбои при импорте данных их системы, так и отсутствие учебной активности. Во-вторых, наблюдаются длинные последовательности идентичных оценок, что может указывать на систематические пропуски занятий студентами или на некорректное ведение электронного журнала преподавателями. В-третьих, присутствуют аномалии в виде скачков, когда оценка на последующей неделе оказывается ниже, чем на предыдущей. Такая ситуация может быть следствием пересмотра правил оценивания или добавления новых элементов в курс.

Приведем пример данных по текущей активности студентов, где также могут встречаться пропуски и скачки (Таблица 2).

Таблица 2 – Фрагмент исходных данных по активностям студентов в курсе
Table 2 – A fragment of the initial data on student activities in the course

Код студента	Текущая активность по дисциплине «Иностранный язык»																	
	Номер недели																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
222558	–	–	–	6	6	7	7	9	9	11	11	13	13	23	23	31	31	31
222560	–	3	3	6	6	8	8	9	9	9	9	11	11	21	21	26	26	26
222567	–	3	3	3	4	5	5	5	5	5	5	7	7	11	11	25	25	25

За последние три года на электронном курсе обучалось 1479 студентов в 64 группах. При этом для разных групп максимальное количество баллов, набранных студентом за семестр, варьировалось от 35 до 86 (Таблица 3, Рисунок 1).

Таблица 3 – Максимальные баллы за семестр в группах
Table 3 – Maximum grades per semester in the groups

Год	Число групп	Минимум	Среднее	Максимум
2020	21	70	75,38	86
2021	23	54	74,52	82
2022	20	35	58,05	68

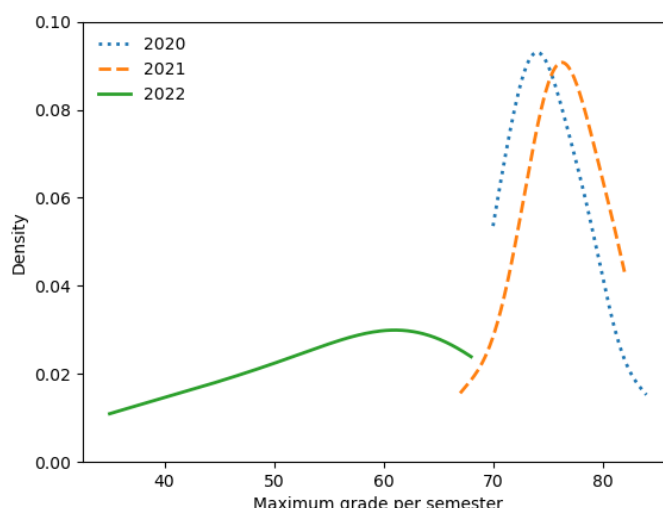


Рисунок 1 – Плотность распределения максимальной оценки за семестр
Figure 1 – The distribution density of the maximum grade per semester

Столь большой разброс баллов может объясняться различным начальным уровнем владения иностранным языком и различным количеством часов, выделенных на изучение дисциплины. В результате на одном и том же электронном курсе для получения зачета студентам некоторых групп достаточно выполнить меньшее число заданий.

При выборе подходящей модели прогнозирования результатов обучения на основе текущей успеваемости в семестре были апробированы и исследованы несколько подходящих, по мнению авторов, моделей и подходов: кластеризация k-средних, линейная регрессия, логистическая регрессия, категориальный байесовский классификатор, метод случайного леса, нейронная сеть (многослойный перцептрон).

В итоге для прогнозирования результатов обучения применяется модель, подробное описание которой приводится в работе [12, 13], на базе непараметрической

оценки функции регрессии [14], которая показала лучшие результаты по сравнению с другими методами. Она принимает следующий вид:

$$\hat{y}(x) = \frac{\sum_{i=1}^s y^i \prod_{j=1}^n \Phi\left(\frac{x_j - x_j^i}{gamma}\right)}{\sum_{i=1}^s \prod_{j=1}^n \Phi\left(\frac{x_j - x_j^i}{gamma}\right)}, \quad (1)$$

где i – номер студента; j – номер недели; s – количество студентов (объем обучающей выборки); n – количество недель для текущего прогноза (размерность входного вектора x); x^i, y^i – значения из обучающей выборки; $gamma$ – параметр размытости ядра. В качестве колоколообразной функции $\Phi\left(\frac{x_j - x_j^i}{gamma}\right)$ использовались следующие виды ядер: «Cosine similarity», «Linear kernel», «RBF kernel», «Laplacian kernel», «Polynomial kernel», «Sigmoid kernel» [7], среди них лучший результат показала «Sigmoid kernel».

Для получения прогнозных результатов успеваемости студентов необходимо обучение модели на данных предыдущего года. А также объединение в обучающую выборку результатов разных групп. Рассмотрим эффективность прогноза в зависимости от способа нормирования данных. Все преобразования данных выполняются по показателям одного семестра для каждой группы в отдельности.

Рассмотрим способы нормирования исходных данных.

1. Без нормирования. Для алгоритмов нечувствительных к масштабу данных, например, деревья решений и ансамбли на их основе (Random Forest, Gradient Boosting).

2. Преобразование данных от 0 до 1 (MinMaxScaler).

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}, \quad (2)$$

где x – вектор исходных данных за неделю обучения, z – вектор преобразованных данных.

3. Центрирование и нормирование данных на основе математических ожиданий и дисперсий (StandardScaler – стандартизирует признаки, удалив среднее значение и масштабировав их до единичной дисперсии).

$$z = \frac{x - u}{s}, \quad (3)$$

где u – среднее значение x , s – выборочная дисперсия x .

4. Преобразование данных от 0 до 1 с помощью процентильных рангов (PercentileRankScaler – процентильный ранг, указывает на его относительное положение в наборе данных).

z = количество значений ниже заданного показателя / общее количество значений.

Таким образом, помимо оценок успеваемости для прогнозирования результатов обучения можно использовать данные по активностям студента в электронном курсе, которые обозначают следующие показатели: 1 – сколько раз за неделю студент производил какие-либо действия на курсе; 2 – сколько дней он ни разу не производил никаких действий.

Следует отметить, что в формуле (1) после добавления показателей активности и дней отсутствия в качестве переменной x^i будет использоваться вектор значений, составленный из соответствующих нормированных значений этих показателей.

Результаты

Для обоснования необходимости предварительной обработки исходных данных обратимся к результатам расчетов, выполненных на основе образовательных данных за

2020–2022 годы. В Таблице 4 приведены ключевые показатели: точность прогнозирования успешного / неуспешного результата обучения, а также доля критических ошибок. Под критической ошибкой здесь понимается ситуация, при которой благоприятный прогноз выдается студенту, впоследствии показывающему неблагоприятный результат по итогам сессии. Т. е. студент будет считать, что его задолженности и низкая активность на электронных курсах – это нормально, ведь система будет выдавать положительный прогноз. Так как в работе используется 3 класса (1 класс – зона риска, крайне неудовлетворительный результат обучения; 2 класс – зона повышенного внимания, нежелательный результат обучения; 3 класс – зона успеха, допустимый и/или желательный результат обучения), то критически важно не допустить ложноположительный прогноз для 3-го кластера [12].

Таблица 4 – Точность прогноза на текущий год по обучению на предыдущем
Table 4 – Accuracy of the forecast for the current year based on previous training

Данные	Точность Асс		Доля критических ошибок	
	2020/2021	2021/2022	2020/2021	2021/2022
Без нормирования	0,82	0,82	1,00	1,00
MinMaxScaler	0,84	0,84	–	0,03
StandardScaler	0,84	0,84	–	–
PercentileRankScaler	0,80	0,80	–	0,06

Анализ результатов Таблицы 4 показывает, что без нормирования исходных данных модель не работает (доля критических ошибок равна 1).

Сравнение диапазонов обучающей и тестовой выборки (Рисунки 2, 3) показало: для периода 2020/2021 годов они близки, тогда как для 2021/2022 годов наблюдаются существенные расхождения. При работе с корректными данными выбор способа нормирования практически не влияет на результат (все методы дают схожую точность, как это отмечено на Рисунке 2).

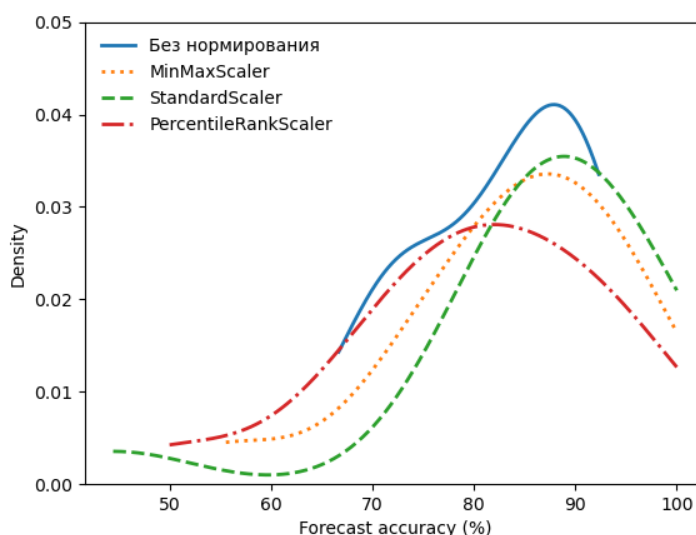


Рисунок 2 – Плотность распределения точности прогноза для групп 2021 года
Figure 2 – Distribution density of forecast accuracy for 2021 groups

На «плохих» данных (2021/2022) ситуация меняется: нормирование процентильным рангом обеспечивает меньшую дисперсию точности прогноза по группам при сохранении средней точности на уровне других методов (Рисунок 3).

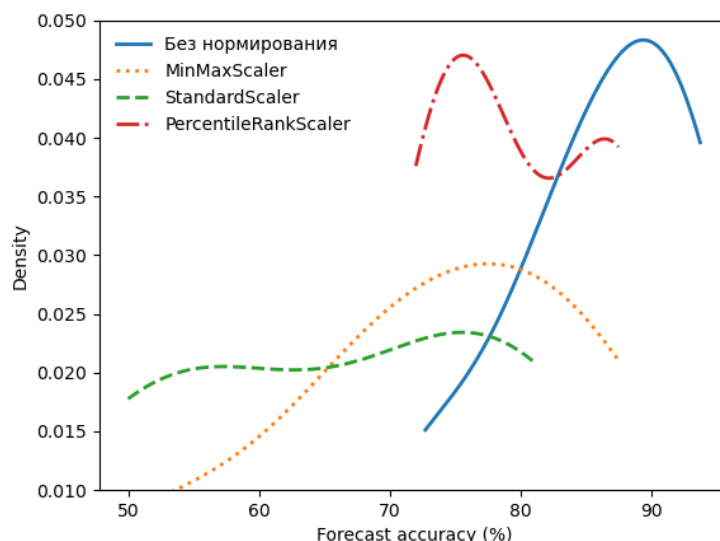


Рисунок 3 – Плотность распределения точности прогноза для групп 2022 года
Figure 3 – Distribution density of forecast accuracy for 2022 groups

Далее приведем примеры обработки исходных данных по дисциплине «Управление в организационных системах» с помощью процентильного ранга за 2024 учебный год (Таблицы 5, 6, 7). Именно на этих обработанных данных приведены дальнейшие результаты моделирования. При этом студенты, числившиеся в группе, но не посещавшие курс и не получившие ни одной оценки, были исключены из рассмотрения.

Таблица 5 – Фрагмент исходных данных по текущей успеваемости
Table 5 – A fragment of the initial data on current academic performance

Код студента	Успеваемость по дисциплине «Управление в организационных системах»																	
	Номер недели																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
168581	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0,2	0,2	0,2
168126	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
168279	–	–	–	–	–	0,9	0,9	0,9	0,9	0,8	0,8	0,5	0,5	0,5	0,5	0,5	0,5	0,5
168922	–	–	–	–	–	–	–	–	–	0,7	0,7	0,4	0,4	0,5	0,5	0,3	0,3	0,3
169301	–	–	–	–	–	0,8	0,8	0,8	0,8	0,6	0,6	0,3	0,3	0,2	0,2	0,1	0,1	0,1
168990	–	–	–	–	–	–	–	–	–	–	–	0,5	0,5	0,7	0,7	0,8	0,8	0,8
168378	–	–	–	–	–	–	–	–	–	–	–	0,4	0,4	0,3	0,3	0,5	0,5	0,5
169660	–	–	–	–	–	–	–	–	–	–	–	–	–	0,3	0,3	0,4	0,4	0,4
169784	–	–	–	–	–	0,9	0,9	0,9	0,9	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
169170	–	–	–	–	–	–	–	–	–	0,8	0,8	0,9	0,9	0,9	0,9	0,9	0,9	0,9
168483	–	–	–	–	–	–	0,7	0,7	0,7	0,5	0,5	0,7	0,7	0,6	0,6	0,7	0,7	0,7
169132	–	–	–	–	–	–	–	–	–	–	–	0,8	0,8	0,8	0,8	0,5	0,5	0,5

Таблица 6 – Фрагмент исходных данных по текущей активности
Table 6 – A fragment of the initial data for the current activity

Код студента	Активность по дисциплине «Управление в организационных системах»																	
	Номер недели																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
168581	–	–	–	–	–	–	–	–	–	–	–	0,2	0,2	0,3	0,3	0,6	0,5	0,5
168126	1,0	1,0	1,0	0,6	0,6	0,5	0,5	0,5	0,6	0,5	0,5	0,3	0,3	0,2	0,2	0,5	0,5	0,5

Таблица 6 (продолжение)
Table 6 (continued)

168279	–	0,2	0,2	0,2	0,1	0,2	0,2	0,2	0,2	0,2	0,2	0,2	0,1	0,1	–	–	–	–	–
168922	–	–	–	0,3	0,3	0,2	0,2	0,3	0,3	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,4	0,3	0,3
169301	–	0,2	0,2	0,3	0,3	0,4	0,4	0,3	0,3	0,3	0,3	0,3	0,3	0,3	0,5	0,5	0,2	0,2	0,2
168990	0,9	0,9	0,8	0,6	0,6	0,5	0,5	0,5	0,5	0,6	0,6	0,7	0,7	0,8	0,8	0,7	0,7	0,7	0,7
168378	–	0,5	0,5	0,5	0,5	0,8	0,7	0,7	0,7	0,7	0,7	0,5	0,5	0,4	0,4	0,7	0,7	0,7	0,7
169660	–	0,7	0,6	0,5	0,5	0,5	0,5	0,3	0,3	0,3	0,3	0,5	0,5	0,5	0,5	0,3	0,4	0,4	0,4
169784	–	0,5	0,9	0,8	0,8	1,0	1,0	1,0	1,0	1,0	1,0	0,9	0,9	0,9	0,9	0,9	0,9	0,9	0,9
169170	–	0,7	0,6	1,0	1,0	0,8	0,7	0,9	0,9	0,9	0,9	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
168483	0,8	0,5	0,5	0,8	0,8	0,7	0,7	0,7	0,7	0,7	0,7	0,8	0,8	0,7	0,7	0,5	0,6	0,5	0,5
169132	–	0,2	0,2	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	–	–	0,1	0,1	0,1	0,1	0,1	0,1

Таблица 7 – Фрагмент исходных данных по дням отсутствия
Table 7 – A fragment of the initial data on the days of absence

Код студента	Дни отсутствия по дисциплине «Управление в организационных системах»																	
	Номер недели																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
168581	–	–	–	–	–	0,1	0,1	0,1	0,1	0,2	0,2	0,4	0,4	0,3	0,3	0,2	0,5	0,6
168126	0,9	0,9	0,9	0,9	0,9	0,8	0,7	0,6	0,8	0,6	0,6	0,4	0,4	0,6	0,6	0,6	0,8	0,8
168279	–	0,4	0,4	0,4	0,4	0,4	0,4	0,3	0,3	0,4	0,4	0,2	0,2	0,2	0,2	0,2	0,2	0,2
168922	–	–	–	0,2	0,1	0,2	0,2	0,3	0,3	0,4	0,4	0,4	0,4	0,3	0,3	0,2	0,2	0,2
169301	–	0,4	0,4	0,4	0,4	0,4	0,4	0,6	0,6	0,4	0,4	0,4	0,4	0,6	0,6	0,6	0,6	0,6
168990	0,9	1,0	0,9	1,0	1,0	1,0	1,0	0,9	0,9	0,9	0,9	0,9	1,0	0,9	0,9	0,9	0,9	0,8
168378	–	0,4	0,4	0,4	0,4	0,4	0,4	0,3	0,3	0,4	0,4	0,4	0,4	0,3	0,3	0,2	0,2	0,2
169660	–	0,4	0,4	0,4	0,4	0,2	0,2	0,2	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1
169784	–	0,4	0,8	0,8	0,8	0,8	0,9	0,9	0,9	0,9	0,9	0,9	0,9	1,0	1,0	0,9	1,0	1,0
169170	–	0,4	0,4	0,4	0,4	0,2	0,2	0,3	0,3	0,6	0,6	0,7	0,7	0,6	0,6	0,9	0,9	0,8
168483	0,9	0,9	0,8	0,8	0,8	0,4	0,7	0,6	0,6	0,6	0,6	0,8	0,8	0,8	0,6	0,6	0,7	0,7
169132	–	0,4	0,4	0,2	0,4	0,4	0,4	0,3	0,3	0,2	0,2	0,2	0,2	0,3	0,3	0,2	0,2	0,2

В вышеприведенных таблицах пропущенные значения заполнялись нулем, поскольку отсутствие оценки интерпретируется как нулевой результат.

Для прогнозирования исходная выборка была разделена на обучающую (70 %) и тестовую (30 %) подвыборки. Обучение модели производилось на обучающей выборке, все приведенные значения R^2 и MSE рассчитаны на тестовой выборке, что обеспечивает несмещенную оценку обобщающей способности модели.

Для начала обратимся к фрагменту вычислительного эксперимента, выполненному на основе данных Таблицы 5, содержащей сведения о текущей успеваемости студентов (Рисунок 4). На Рисунке 4 представлены прогнозируемые результаты обучения по оценкам успеваемости.

Как видно из Рисунка 4, модель (1) описывает исходные данные с коэффициентом детерминации $R^2 = 0,66$ на тестовой выборке, что свидетельствует об удовлетворительной аппроксимации и позволяет использовать модель (1) для предварительного прогнозирования. С учетом значительного разброса исходных данных такой результат считается допустимым. Среднеквадратичная ошибка (MSE) при этом составляет 0,105, что указывает на невысокую величину ошибки прогноза в абсолютном выражении.

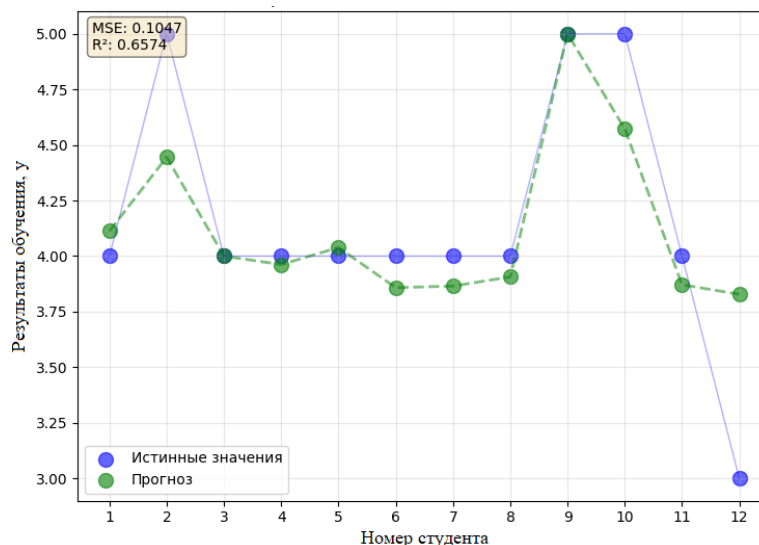


Рисунок 4 – Прогнозируемые результаты обучения по оценкам успеваемости
Figure 4 – Predicted learning outcomes based on academic performance

Далее приведем результаты прогнозирования, полученные с привлечением дополнительных данных об академической активности студентов и количестве дней пропусков, которые содержатся в Таблицах 6 и 7 (Рисунок 5). На Рисунке 5 отражены результаты прогнозирования по всем трем таблицам (5, 6 и 7) совместно.

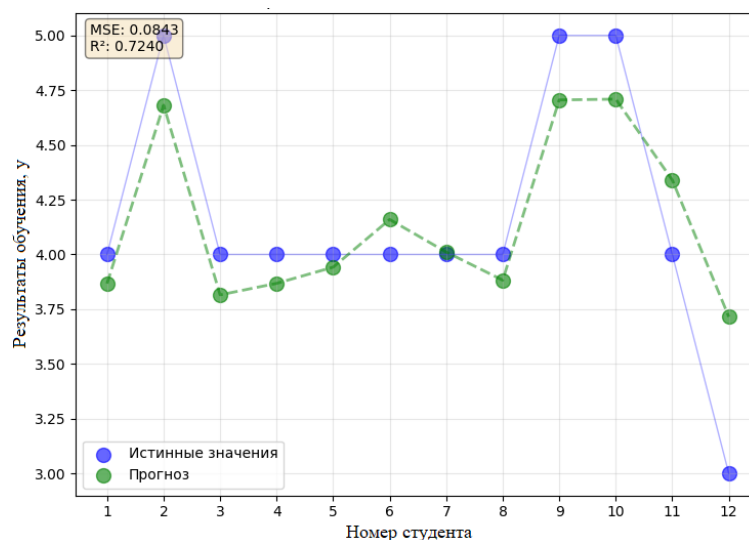


Рисунок 5 – Результаты прогнозирования по всем таблицам 5, 6 и 7
Figure 5 – Prediction results for all tables 5, 6, and 7

Из Рисунка 5 видно, что при включении дополнительных показателей коэффициент детерминации на тестовой выборке повышается до $R^2 = 0,72$, что указывает на приемлемое качество результатов: модель объясняет 72 % вариации результирующего показателя. Это позволяет использовать ее для анализа влияния образовательной активности на успеваемость и выявления студентов, относящихся к группе риска. Также характеризуется небольшой ошибкой моделирования $MSE = 0,084$.

При агрегировании информации из Таблиц 5, 6 и 7 точность прогнозирования превышает показатель, достигнутый на данных, состоящих только из оценок.

Таким образом, применение процентильных рангов для приведения данных к единому диапазону $[0, 1]$ дает приемлемые результаты.

Обсуждение

Полученные в ходе исследования результаты демонстрируют, что выбор метода предобработки данных существенно влияет на качество прогнозирования успеваемости студентов при использовании непараметрической функции регрессии Надарая-Ватсона. Наиболее важным выводом является то, что процентильное преобразование исходных данных несколько превосходит как использование необработанных данных, так и стандартную нормализацию методом центрирования и нормирования, но есть ограничения, на которые будет указано далее.

Преимущество процентильного подхода можно объяснить несколькими факторами. Во-первых, как было отмечено при анализе исходных данных (Таблица 1), образовательные данные характеризуются высокой разреженностью, наличием пропусков и аномальных значений. Процентильные ранги, в отличие от абсолютных значений, устойчивы к выбросам и не требуют предположений о распределении признаков. Во-вторых, перевод оценок в относительные ранги внутри группы обучающихся решает проблему сравнения студентов при различной сложности курсов или разных преподавателей – студент, получивший 70 баллов на сложном курсе, может иметь более высокий процентильный ранг, чем студент со 100 баллами на простом курсе. В-третьих, приведение данных к единому диапазону $[0, 1]$ через процентиля делает возможным корректное вычисление ядерных функций в многомерном пространстве признаков, что критически важно для метода Надарая-Ватсона.

Далее рассмотрим ситуацию в рамках одной дисциплины, которую вел один преподаватель. В выборку вошли 677 студентов из 36 групп. На Рисунке 6 представлено соответствующее неоднородное распределение данных.

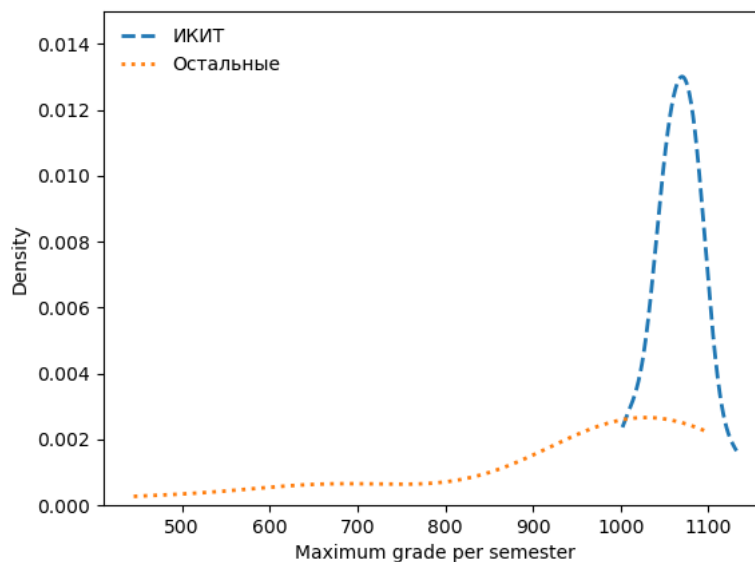


Рисунок 6 – Неоднородное распределение данных
Figure 6 – Heterogeneous data distribution

Особенностью данной дисциплины является дистанционный асинхронный формат и отсутствие четких временных границ для изучения материала в течение семестра. Поэтому группы крайне неоднородны как по итоговому количеству набранных баллов (Рисунок 6), так и по количеству недель активного обучения.

В Таблице 8 приведены результаты эксперимента, в котором модель обучалась на каждой группе по отдельности, после чего предсказывала результат для других групп. Ожидалось, что модель, успешно обученная по одной группе, будет корректно предсказывать результаты обучения по другим группам.

Успешное обучение или предсказание модели: точность больше 0,8 и критическая ошибка меньше 0,06.

Ложно обученная модель: результаты группы, по которой происходило обучение, не смогли предсказать другие модели.

Случайно предсказанные результаты: для группы не удалось обучить модель, но другая успешная модель дала успешное предсказание.

Успешно предсказано по неуспешно обученным: аномальная ситуация.

Последние два случая не должны происходить. Однако особенностью данных является то, что абсолютное большинство студентов увеличивает оценки в течение семестра и по итогу успешно сдает предмет. В результате даже плохо обученная модель прогнозирования может случайно выдать хороший результат для другой группы. Главная же причина неудачи обучения моделей – это студенты, которые плохо учились, но успешно сдали сессию. Во всех аномальных случаях в обучающих группах такие студенты были, а в тестовых – отсутствовали. И задача состоит в том, чтобы понять, при каком способе нормирования данных мы можем больше доверять результатам.

Причиной неудовлетворительного обучения моделей в этом примере является расхождение ± 2 недели между моментом выполнения работ студентами и проверкой их преподавателем для разных групп. Корректировка исходных данных приводит к результатам строки с PercentileRankScale*, где можно увидеть отсутствие случаев неадекватного поведения модели.

Таблица 8 – Успешность перекрестного обучения и прогноза групп одной дисциплины
Table 8 – Success of cross-training and prediction of same-discipline groups

	Успешно обучено (доля ложно обученных), %	Успешно предсказано (доля случайно предсказанных), %	Успешно предсказано по неуспешно обученным, %
Без нормирования	8 (0)	8 (0)	8
MinMaxScaler	17 (50)	14 (40)	31
StandardScaler	78 (61)	39 (21)	0
PercentileRankScaler	44 (31)	39 (21)	25
StandardScaler + PercentileRankScaler	39 (29)	39 (29)	0
StandardScalerUnion	78 (29)	58 (5)	0
PercentileRankScale*	31 (0)	36 (15)	0

Анализ Таблицы 8 показывает, что по количеству успешных прогнозов StandardScaler и PercentileRankScaler равны. Недостаток StandardScaler – это 61 % ложно обученных моделей, неспособных дать прогноз для других групп. Т. е. в этих случаях при попытке прогноза будут ошибки. Недостаток PercentileRankScaler – это 25 % успешных предсказаний на тестовых данных для заведомо некорректно обученных моделей. Можно предположить, что четверть прогнозов мы просто угадываем и действительно обученных моделей не 44 %, а 19 % и успешных предсказаний не 39 %, а 14 %. При совместном использовании обоих способов нормирования данных результаты улучшаются. Наилучшие результаты (StandardScalerUnion) достигаются для

StandardScaler при объединении всех успешно обученных групп в одну обучающую выборку.

Таким образом, для специфических данных возможна ситуация, когда успешное обучение модели на исторических данных гарантированно приводит к некачественным прогнозам. Поэтому необходима разработка некоторого критерия, который позволит по историческим данным оценивать корректность применения разработанной методики. Таким критерием может служить факт расхождения точности обучения по разным методам нормирования.

Заключение

Таким образом, проведенное исследование подтверждает, что процентильный подход к предобработке образовательных данных является эффективным инструментом повышения точности прогнозирования успеваемости, однако возможны ситуации, когда данный подход значительно уступает классическому нормированию. Несмотря на отмеченные ограничения, полученные результаты имеют практическую ценность для разработки систем аналитики обучения и могут служить основой для дальнейших исследований в области интеллектуального анализа образовательных данных.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Zhang Y., Yun Y., An R., Cui J., Dai H., Shang X. Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. *Frontiers in Psychology*. 2021;12:698490. <https://doi.org/10.3389/fpsyg.2021.698490>
2. Angeioplastis A., Aliprantis J., Konstantakis M., Tsimpiris A. Predicting Student Performance and Enhancing Learning Outcomes: A Data-Driven Approach Using Educational Data Mining Techniques. *Computers*. 2025;14(3):83. <https://doi.org/10.3390/computers14030083>
3. Lu Ch., Cutumisu M. Online engagement and performance on formative assessments mediate the relationship between attendance and course performance. *International Journal of Educational Technology in Higher Education*. 2022;19(1):2. <https://doi.org/10.1186/s41239-021-00307-5>
4. Han L. Prediction and Analysis of Students' Behavior Based on Data Mining in Educational Administration. In: *Artificial Intelligence for Future Society: Proceedings of the International Conference on Artificial Intelligence for Society, 18–19 May 2024, Bhubaneswar, India*. Cham: Springer; 2024. P. 229–238. https://doi.org/10.1007/978-3-031-69457-8_22
5. Wang J., Yu Y. Machine learning approach to student performance prediction of online learning. *PLoS ONE*. 2025;20(1):e0299018. <https://doi.org/10.1371/journal.pone.0299018>
6. Al-Tameemi Gh., Xue J., Hadi Ali I., Ajit S. A hybrid machine learning approach for predicting student performance using multi-class educational datasets. *Procedia Computer Science*. 2024;238:888–895. <https://doi.org/10.1016/j.procs.2024.06.108>
7. Gadde S.S., Anand D., Sasidhar Babu N., Pujitha B.V., Sai Reethi M., Pradeep Ghantasala G.S. Performance Prediction of Students Using Machine Learning Algorithms. In: *Applications of Computational Methods in Manufacturing and Product Design: Select Proceedings of IPDIMS 2020, 02–03 December 2020, Rourkela, India*. Singapore: Springer; 2022. P. 405–411. https://doi.org/10.1007/978-981-19-0296-3_36
8. Duch D., May M., George S. Enhancing Predictive Analytics for Students' Performance in Moodle: Insight from an Empirical Study. *Journal of Data Science and Intelligent Systems*. 2026;4(1):84–96. <https://doi.org/10.47852/bonviewJDSIS42023777>

9. Попова Н.А., Егорова Е.С. Основные направления интеллектуального анализа данных в сфере образования. *Известия Кабардино-Балкарского научного центра РАН*. 2024;26(5):94–106. <https://doi.org/10.35330/1991-6639-2024-26-5-94-106>
Popova N.A., Egorova E.S. Main directions of data mining in the field of education. *News of the Kabardino-Balkarian Scientific Center of RAS*. 2024;26(5):94–106. (In Russ.). <https://doi.org/10.35330/1991-6639-2024-26-5-94-106>
10. Papadogiannis I., Wallace M., Karountzou G. Educational Data Mining: A Foundational Overview. *Encyclopedia*. 2024;4(4):1644–1664. <https://doi.org/10.3390/encyclopedia4040108>
11. Lin Y., Chen H., Xia W., Lin F., Wang Z., Liu Y. A Comprehensive Survey on Deep Learning Techniques in Educational Data Mining. *Data Science and Engineering*. 2025;10(4):564–590. <https://doi.org/10.1007/s41019-025-00303-z>
12. 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. <https://doi.org/10.32517/0234-0453-2023-38-4-28-43>
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 Russ.). <https://doi.org/10.32517/0234-0453-2023-38-4-28-43>
13. Ликсонова Д.И., Даничев А.А., Шестаков В.Н., Якунин Ю.Ю. Повышение точности модели прогнозирования успеваемости студентов вузов. *Информационные технологии*. 2025;31(4):215–224. <https://doi.org/10.17587/it.31.215-224>
Liksonova D.I., Danichev A.A., Shestakov V.N., Yakunin Yu.Yu. Increasing the accuracy of the model for predicting the performance of university students. *Information Technologies*. 2025;31(4):215–224. (In Russ.). <https://doi.org/10.17587/it.31.215-224>
14. Надарая Э.А. *Непараметрическое оценивание плотности вероятностей и кривой регрессии*. Тбилиси: Изд-во ТГУ; 1983. 194 с.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Даничев Алексей Александрович, кандидат технических наук, доцент, доцент базовой кафедры интеллектуальных систем управления, Институт космических и информационных технологий, Сибирский федеральный университет, Красноярск, Российская Федерация.

e-mail: adanichev@inbox.ru

ORCID: [0000-0001-9830-5205](https://orcid.org/0000-0001-9830-5205)

Alexey A. Danichev, Candidate of Engineering Sciences, Docent, Associate Professor at the Basic Department of Intelligent Control Systems, Institute of Space and Information Technologies, Siberian Federal University, Krasnoyarsk, the Russian Federation.

Ликсонова Дарья Игоревна, кандидат технических наук, доцент, заведующий базовой кафедрой интеллектуальных систем управления, Институт космических и информационных технологий, Сибирский федеральный университет, Красноярск, Российская Федерация.

e-mail: liksonovadi@yandex.ru

ORCID: [0000-0001-9663-6481](https://orcid.org/0000-0001-9663-6481)

Darya I. Liksonova, Candidate of Engineering Sciences, Docent, Head of the Basic Department of Intelligent Control Systems, Institute of Space and Information Technologies, Siberian Federal University, Krasnoyarsk, the Russian Federation.

Шестаков Вячеслав Николаевич, кандидат философских наук, доцент кафедры современных образовательных технологий, Институт педагогики, психологии и социологии, Сибирский федеральный университет, Красноярск, Российская Федерация.

e-mail: vshestakov@sfu-kras.ru

ORCID: [0000-0001-7737-2900](https://orcid.org/0000-0001-7737-2900)

Vyacheslav N. Shestakov, Candidate of Philosophical Sciences, Associate Professor at the Department of Modern Educational Technologies, Institute of Pedagogy, Psychology and Sociology, Siberian Federal University, Krasnoyarsk, the Russian Federation.

Статья поступила в редакцию 22.05.2026; одобрена после рецензирования 07.09.2026; принята к публикации 15.09.2026.

The article was submitted 22.05.2026; approved after reviewing 07.09.2026; accepted for publication 15.09.2026.