УДК 004.93 *DOI 10.55421/3034-4689_2026_29_2_137*

К. Г. Осанов, Р. М. Хусаинов

НЕЙРОСЕТЕВАЯ МОДЕЛЬ ПРОГНОЗИРОВАНИЯ УСПЕВАЕМОСТИ СТУДЕНТОВ

Ключевые слова: машинное обучение, нейронные сети, прогнозирование успеваемости, образовательные данные, нейросетевая модель, академическая успеваемость.

В статье рассмотрены вопросы применения методов машинного обучения и нейронных сетей для прогнозирования уровня академической успеваемости студентов на основе анализа образовательных данных. Актуальность исследования обусловлена ростом объемов цифровых данных в современных образовательных системах и необходимостью разработки интеллектуальных инструментов поддержки принятия решений, направленных на повышение качества обучения и раннее выявление студентов, находящихся в зоне академического риска. В качестве объекта исследования являются данные, характеризующие учебную и внеучебную деятельность студентов, включая показатели посещаемости, учебной нагрузки, социально-демографические характеристики и результаты предыдущего обучения. Предметом исследования является процесс прогнозирования уровня академической успеваемости студентов с использованием нейросетевых моделей классификации. В работе проведен анализ существующих методов прогнозирования, применяемых в задачах образовательной аналитики, и обоснован выбор многослойной полносвязной нейронной сети как наиболее подходящего инструмента для обработки табличных данных. Реализация модели выполнена с использованием библиотеки scikit-learn. Особое внимание уделено этапам предобработки данных, включая обработку пропусков, кодирование категориальных признаков, масштабирование числовых параметров и формирование классов целевой переменной. В рамках экспериментальных исследований проведено сравнение качества работы нейросетевой модели при различном количестве выходных классов целевой переменной. Оценка эффективности выполнялась с использованием метрик классификации, таких как Accuracy, Precision, Recall и F1-Score. Полученные результаты показали, что трехклассовая классификация обеспечивает оптимальный баланс между точностью прогнозирования и практической интерпретируемостью результатов. В перспективе целесообразно использование нейросетевой модели для прогнозирования успеваемости на различных этапах обучения.

K. G. Osanov, R. M. Khusainov

NEURAL NETWORK MODEL FOR PREDICTING STUDENTS' ACADEMIC PERFORMANCE

Keywords: machine learning, neural networks, academic performance prediction, educational data, neural network model, academic achievement.

This article examines the application of machine learning and neural network methods to predicting students' academic performance based on the analysis of educational data. The relevance of the study is driven by the growing volume of digital data in modern educational systems and the need to develop intelligent decision support tools aimed at improving the quality of education and early identification of students at academic risk. The object of the study is data characterizing students' academic and extracurricular activities, including attendance indicators, study load, socio-demographic characteristics, and previous learning outcomes. The subject of the study is the process of predicting students' academic performance using neural network classification models. The paper analyzes existing forecasting methods used in educational analytics and substantiates the choice of a multilayer fully connected neural network as the most suitable tool for processing tabular data. The model is implemented using the scikit-learn library. Particular attention is paid to the stages of data preprocessing, including gap handling, coding of categorical features, scaling of numerical parameters, and the formation of target variable classes. In the experimental studies, the performance of the neural network model was compared with different numbers of output classes of the target variable. Performance was assessed using classification metrics such as Accuracy, Precision, Recall, and F1-score. The results showed that the three-class classification provides an optimal balance between prediction accuracy and the practical interpretability of the results. The neural network model may be useful for predicting academic performance at various stages of learning.

Введение

В настоящее время современные образовательные учреждения сталкиваются с задачей своевременного выявления студентов, испытывающих трудности в обучении [1]. Переход к цифровым формам образования, внедрение электронных журналов, онлайн-платформ и систем дистанционного обучения привел к росту объема данных, характеризующих учебную деятельность учащихся. Эти данные могут использоваться для прогнозирования успеваемости студентов и своевременного принятия корректирующих мер.

Традиционные способы оценки учебной успеваемости, основанные на ручном анализе данных, имеют ряд недостатков: субъективность выставления оценок, а также невозможность комплексного учета большого числа факторов, включая внеучебные. В результате возникает необходимость применения методов интеллектуального анализа данных, которые позволяют повысить точность и скорость диагностики

Использование методов искусственного интеллекта, в частности нейронных сетей, предоставляет мощный инструмент для анализа образовательных данных [2]. Нейросетевые модели способны учитывать сложные нелинейные зависимости между факторами успеваемости студентов, что делает их особенно перспективными в задачах образовательной аналитики. Применение подобных моделей позволяет заранее выявлять риски академической неуспеваемости студентов, определять ключевые факторы риска, формировать индивидуальные рекомендации

и тем самым повышать общую результативность учебного процесса.

Цель и задачи исследования

Цель исследования: разработать нейросетевую модель для прогнозирования успеваемости обучающегося на основе ключевых факторов.

Для достижения поставленной цели потребовалось решить следующие задачи:

1) анализ предметной области и существующих подходов к прогнозированию академической успеваемости студентов;

2) анализ и предобработка исходных данных, влияющих на успеваемость студентов;

3) формирование вариантов разбиения целевой переменной на классы успеваемости;

4) построение нейросетевой модели прогнозирования успеваемости студентов;

6) тестирование нейросетевой модели при различном количестве классов выходной переменной;

7) проведение сравнительного анализа результатов, определения оптимального количество классов.

Объект исследования: данные, характеризующие учебную и внеучебную деятельность студентов, включая показатели посещаемости, учебной нагрузки, социально-демографические характеристики.

Предмет исследования: нейросетевая модель для прогнозирования итоговой успеваемости.

Анализ методов и моделей прогнозирования успеваемости

Прогнозирование успеваемости студентов относится к классу задач интеллектуального анализа данных и является важным направлением образовательной аналитики. Основная цель прогнозирования заключается в выявлении студентов, находящихся в зоне академического риска, а также в оценке факторов, оказывающих наибольшее влияние на учебные результаты. Особенностью данной задачи является многофакторный характер данных, включающих как количественные, так и категориальные признаки, а также наличие нелинейных и слабоформализуемых зависимостей между ними.

Ниже представлены основные подходы, применимые к этой задаче, с анализом их преимуществ и недостатков в контексте образовательных данных.

Линейная и логистическая регрессия

Линейная регрессия предполагает, что целевая переменная может быть описана как линейная комбинация входных признаков. Модель минимизирует ошибку прогнозирования, чаще всего – среднеквадратичную, и позволяет оценить вклад каждого признака.

Логистическая регрессия, в отличие от линейной, использует логистическую функцию активации. Она преобразует линейную комбинацию признаков в вероятность, что делает ее удобной для задач бинарной классификации [3]. Этот метод позволяет использовать интерпретируемые коэффициенты, анализируя влияние каждого фактора.

Метод k-ближайших соседей (k-NN)

Метод k-NN (k-ближайших соседей) относится к алгоритмам, не строящим явную модель. Он определяет результат на основе расстояний между объектами в пространстве признаков [4]. Чем ближе объект к известным наблюдениям, тем выше вероятность совпадения их выходных значений. Метод интуитивно понятен, но чувствителен к масштабированию данных и плохо работает в пространствах высокой размерности.

Деревья решений

Деревья решений формируют правила вида «если-то», последовательно разделяя данные на основе наиболее информативных признаков. Разбиение выполняется для минимизации меры неопределенности (энтропии или индекса Джини). Такой подход позволяет выявлять сложные структурные зависимости между данными и представляет модель в виде древовидной структуры.

Метод опорных векторов (SVM)

Метод опорных векторов – один из наиболее теоретически обоснованных и мощных алгоритмов классического обучения. Его особенность заключается в обнаружении такой разделяющей гиперплоскости, которая максимизирует расстояние (зазор) между классами [5]. Если данные не являются линейно разделимыми, применяется мягкий зазор – подход, допускающий ошибки при классификации в обмен на более устойчивое разделение данных.

Использование ядерных функций (радиально-базисной, полиномиальной, сигмоидной и др.) позволяет SVM проецировать данные в пространства большей размерности, где становится возможным построить линейную разделяющую поверхность даже для сложных нелинейных зависимостей. Благодаря этому SVM часто демонстрирует высокую точность и устойчивость к переобучению.

Ансамблевые методы прогнозирования

Ансамблевые методы машинного обучения представляют собой подходы, основанные на объединении нескольких базовых моделей с целью повышения точности и устойчивости прогнозирования. Особенность ансамблевого обучения заключается в том, что совокупность относительно простых моделей может обеспечить более качественные результаты по сравнению с одной, даже более сложной моделью. Основными подходами к построению ансамблей являются bagging и boosting.

Bagging (Bootstrap Aggregating) основан на обучении нескольких моделей одного типа на различных подвыборках исходного датасета, сформированных методом бутстрэппинга. Наиболее известным представителем данного подхода является метод случайный лес (random forest), в котором используется ансамбль деревьев решений. Random forest отличается устойчивостью к переобучению, способностью работать с большим числом признаков и относительной нечувствительностью к масштабированию данных, что делает его популярным инструментом анализа образовательных показателей.

Boosting предполагает последовательное обучение моделей, при котором каждая последующая модель фокусируется на объектах, ошибочно классифицированных предыдущими. Одним из наиболее распространенных алгоритмов данного класса является градиентный бустинг (gradient boosting), который

строит ансамбль моделей путем минимизации функции потерь с использованием градиентного спуска.

Нейронные сети

Помимо вышеизложенных методов для прогнозирования успеваемости студентов используют нейронные сети. Упомянуты множество исследований про возможность использования нейронных сетей для решения данной задачи, однако информации о реальном внедрении таких прогнозных моделей не встречается [6-9].

Нейронные сети – это математические модели, построенные по принципу организации и функционирования биологических нейронных сетей. Нейронные сети не программируются в привычном смысле этого слова, они обучаются. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными данными и выходными, а также выполнять обобщение. После обучения сеть способна предсказать значение некой последовательности на основе нескольких предыдущих значений.

Структурная схема нейронной сети представлена на рисунке 1.

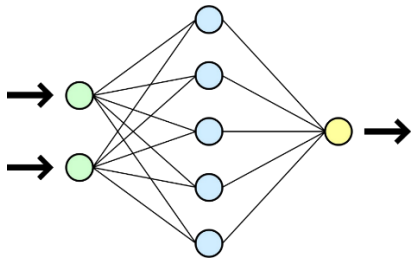

Рис. 1 – Обобщенная схема нейронной сети

Fig. 1 – Generalized diagram of a neural network

Нейронная сеть состоит из нейронов, слоев и синапсов. Нейроны изображены в виде узлов разного цвета. Все узлы одного цвета относятся к одному слою нейронной сети. Синапсы – это линии, которые связывают нейроны одного слоя с нейронами другого слоя. Синапсы имеют один параметр – вес. Каждый нейрон выполняет определенную математическую функцию, таким образом на вход ему подается множество значений, а на выходе формируется только одно. Следовательно, на выходе получается определенное значение, которое выдала уже обученная нейронная сеть.

Каждый из вышерассмотренных подходов обладает своими преимуществами и недостатками, что делает необходимым их сравнительный анализ при выборе модели для практического применения.

Линейные модели, такие как линейная и логистическая регрессия, отличаются простотой реализации и высокой интерпретируемостью результатов. Они позволяют наглядно оценить вклад каждого признака в формирование прогноза и хорошо работают на данных с линейными зависимостями. Данные методы плохо справляются с нелинейными взаимосвязями, характерными для образовательных процессов, и чувствительны к мультиколлинеарности признаков, что ограничивает их применение в сложных реальных задачах.

Метод k-ближайших соседей основан на предположении о схожести объектов в пространстве признаков. Прогноз формируется на основе анализа значений целевой переменной у ближайших по расстоянию наблюдений. Основными преимуществами данного метода являются простота реализации, отсутствие этапа обучения в классическом понимании и способность учитывать локальные закономерности в данных. Метод k-ближайших соседей крайне чувствителен к масштабированию признаков, выбору метрики расстояния и параметра k, а также плохо масштабируется при увеличении объема выборки, что ограничивает его применение для больших образовательных наборов данных.

Метод опорных векторов представляет собой гибкий подход, способный моделировать как линейные, так и нелинейные зависимости за счет использования ядерных функций. Он отличается высокой обобщающей способностью и устойчивостью к переобучению при корректной настройке гиперпараметров. Метод опорных векторов характеризуется высокой вычислительной сложностью, чувствительностью к выбору ядра и параметров модели, а также ограниченной применимостью к большим объемам данных.

Деревья решений являются наглядным и интуитивно понятным методом прогнозирования, позволяющим работать как с числовыми, так и с категориальными признаками без необходимости предварительного масштабирования данных. Они способны выявлять сложные нелинейные зависимости и легко интерпретируются, что является важным преимуществом в задачах образовательной аналитики. К основным недостаткам деревьев решений относятся склонность к переобучению, чувствительность к шуму и нестабильность структуры дерева при незначительных изменениях обучающей выборки.

Ансамблевые методы, такие как случайный лес и градиентный бустинг, направлены на повышение качества прогнозирования за счет объединения множества базовых моделей, чаще всего деревьев решений. Эти методы обладают высокой точностью, устойчивостью к переобучению и способностью эффективно работать с разнородными признаками. Усложнение структуры моделей приводит к увеличению вычислительных затрат, времени обучения и снижению интерпретируемости результатов.

Нейросетевые модели позволяют автоматически выявлять сложные нелинейные зависимости между факторами успеваемости и целевой переменной, адаптироваться к различным структурам данных и достигать высокой точности прогнозирования. Вместе с тем нейронные сети требуют значительных вычислительных ресурсов, больших объемов обучающих данных и тщательной настройки архитектуры.

С учетом особенностей задачи прогнозирования успеваемости студентов, структуры исходного датасета и целей исследования, в качестве основного инструмента прогнозирования выбран тип нейронной сети – многослойный персептрон. Нейронные сети обладают высокой гибкостью, способны автоматически выявлять сложные нелинейные зависимости между признаками и адаптироваться к различным структурам данных. Это делает их особенно эффективными для задач образовательной аналитики, где

на итоговый результат влияет совокупность множества факторов различной природы.

Дополнительным аргументом в пользу выбора нейросетевой модели является возможность проведения серии экспериментов с различным числом классов целевой переменной, что позволяет исследовать влияние структуры выходного параметра на качество прогнозирования. Использование нейронной сети также создает основу для дальнейшего расширения модели, включая настройку архитектуры, изменение числа слоев и нейронов, а также применение методов регуляризации.

Несмотря на широкое распространение нейросетевых моделей в задачах образовательной аналитики, большинство существующих исследований ориентировано либо на бинарную классификацию успеваемости (низкая или высокая), либо на достижение максимальных значений отдельных метрик качества без учета практической интерпретируемости результатов. Научная новизна предлагаемого подхода заключается в исследовании влияния способа формирования целевой переменной на качество и практическую полезность нейросетевого прогнозирования академической успеваемости студентов. В отличие от существующих работ, в статье проводится сравнительный анализ нейросетевой модели при различном количестве выходных классов, что позволяет оценить компромисс между точностью классификации и интерпретируемостью результатов для задач образовательного мониторинга.

Дополнительной особенностью исследования является ориентация на использование нейросетевой модели для анализа табличных образовательных данных с учетом этапов предобработки и ограничений, характерных для реальных образовательных информационных систем. Такой подход позволяет рассматривать нейросетевую модель не только как инструмент повышения точности прогнозирования, но и как практический элемент интеллектуальной поддержки принятия решений, направленной на раннее выявление студентов, находящихся в зоне академического риска.

Описание и характеристика исходных данных

В качестве исходных данных для проведения исследования использован открытый датасет Student Performance Factors, размещенный на платформе Kaggle [10]. Датасет предназначен для анализа совокупности факторов, влияющих на академическую успеваемость студентов, и содержит информацию о 6607 обучающихся. Набор данных включает 20 признаков, один из которых является целевой переменной, а остальные используются в качестве входных факторов модели.

Датасет представлен в табличном формате и содержит как количественные, так и категориальные признаки. Каждая строка соответствует описанию студента, а каждый столбец определяет его характеристику, потенциально влияющую на результаты обучения. Разнообразие представленных признаков делает данный набор данных пригодным для применения методов машинного обучения и нейронных сетей.

Демографические и социально-экономические характеристики

Gender – пол студента:

– male: мужчина;

– female: женщина.

Family_Income – уровень дохода семьи:

– low: низкий;

– medium: средний;

– high: высокий.

Parental_Education_Level – уровень образования родителей:

– high school: средняя школа;

– college: колледж;

– postgraduate: аспирантура.

Distance_from_Home – расстояние от места проживания до учебного заведения:

– near: близко;

– moderate: умеренно;

– far: далеко.

School_Type – тип образовательного учреждения:

– public: государственное;

– private: частное.

Учебная активность и академические показатели

Hours_Studied – количество часов, затрачиваемых на самостоятельное обучение.

Attendance – процент посещаемости учебных занятий.

Previous_Scores – баллы за предыдущие экзамены.

Tutoring_Sessions – количество занятий с репетитором в месяц.

Teacher_Quality – квалификация преподавателей:

– low: низкая;

– medium: средняя;

– high: высокая.

Access_to_Resources – доступность учебных материалов и образовательных ресурсов:

– low: низкая;

– medium: средняя;

– high: высокая.

Internet_Access – наличие доступа к сети Интернет (yes – да, no – нет).

Parental_Involvement – степень вовлеченности родителей в образовательный процесс:

– low: низкая;

– medium: средняя;

– high: высокая.

Психологические и мотивационные факторы

Motivation_Level – уровень учебной мотивации:

– low: низкий;

– medium: средний;

– high: высокий.

Peer_Influence – влияние окружения и сверстников на успеваемость:

– positive: положительное;

– neutral: нейтральное;

– negative: отрицательное.

Learning_Disabilities – наличие или отсутствие учебных трудностей (yes – да, no – нет).

Внеучебная деятельность

Extracurricular_Activities – участие во внеучебных мероприятиях (yes – да, no – нет).

Physical_Activity – среднее количество часов физической активности в неделю.

Sleep_Hours – среднее количество часов сна за ночь.

Целевая переменная

Итоговым показателем успеваемости студентов будет являться:

Exam_Score – итоговый балл за экзамен.

Предобработка данных

Исходный набор данных содержит как числовые, так и категориальные признаки, при этом в ряде атрибутов присутствуют пропущенные значения. Наличие пропусков может негативно сказаться на качестве обучения моделей машинного обучения, а в ряде случаев делает обучение невозможным. В ходе анализа данных выявлено, что пропуски присутствовали в следующих признаках: Teacher_Quality, Parental_Education_Level и Distance_from_Home. В связи с этим на этапе предобработки данных выполнена обработка пропущенных значений, а именно с использованием моды, поскольку признаки являются категориальными [11, 12]. Мода представляет собой значение признака, которое встречается наиболее часто, пропуски заполнялись часто встречающимися (модальными) значениями, что позволило сохранить структуру распределения категорий и избежать искажения данных.

Поскольку алгоритмы машинного обучения и нейросетевые модели работают исключительно с числовыми данными, категориальные признаки преобразованы в числовой формат. Для бинарных признаков, а это Gender, Internet_Access, Learning_Disabilities, School_Type, Extracurricular_Activities, использовалось бинарное кодирование, при котором каждому из возможных значений сопоставлялось числовое представление: yes – 1, no – 0; male – 0, female – 1; public – 0, private – 1. Для категориальных признаков с упорядоченными значениями применялось порядковое кодирование (Ordinal Encoding), позволяющее сохранить логическую иерархию категорий.

Кодированные значения Parental_Involvement, Motivation_Level, Teacher_Quality, Access_to_Resources, Family_Income выглядят следующим образом: low – 1, medium – 2, high – 0; для Peer_Influence: positive – 2, neutral – 1, negative – 0; для Distance_from_Home: near – 2, moderate – 1, far – 0; для Parental_Education_Level: postgraduate – 2, high school – 1, college – 0.

В результате выполнения данных процедур датасет приведен к формату, пригодному для последующего анализа.

Корреляционный анализ

С целью выявления факторов, оказывающих наибольшее влияние на уровень успеваемости студентов, проведен корреляционный анализ признаков. В качестве меры линейной зависимости между признаками использован коэффициент корреляции Пирсона, позволяющий количественно оценить степень и направление связи между переменными. На рисунке 2 представлена построенная матрица корреляций. Анализ матрицы корреляций, позволяет идентифицировать направление линейных взаимосвязей между анализируемыми признаками и целевой переменной Exam_Score [13]. Наиболее выраженные связи наблюдаются у Attendance (r = 0.58) и Hours_Studied (r = 0.45). Признаками со слабой, но статистически значимой положительной корреляцией являются Previous_Scores (r = 0.18) и Tutoring_Sessions (r = 0.16). Это позволяет сделать вывод, что процент посещаемости и время, затраченное на самостоятельное обучение, являются одними из ключевых факторов при прогнозировании успеваемости студентов, а баллы за предыдущие экзамены и занятия с репетитором умеренно положительно сказываются на академической успеваемости. Эти переменные отобраны для дальнейшего анализа и обучения нейросетевой модели.

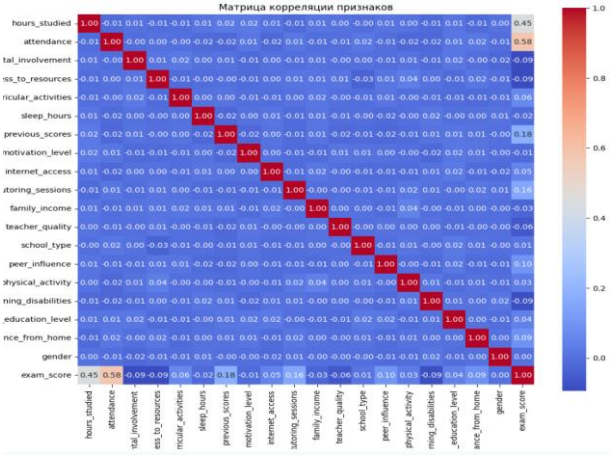

Рис. 2 – Вычисленная матрица корреляций признаков

Fig. 2 – Calculated matrix of feature correlations

Остальные факторы демонстрируют слабую корреляцию (r < 0.1), что позволяет предположить о их незначительном прямом влиянии на выходной параметр.

Формирование классов целевой переменной

В исходном наборе данных целевая переменная Exam_Score представлена в виде непрерывной числовой величины, однако для проведения серии экспериментов по определению количества градаций выходного параметра необходимо сформировать классы Exam_Score. В связи с этим выполнена дискретизация целевой переменной с формированием нескольких категориальных классов. Диаграмма распределения значений целевой переменной представлена на рисунке 3.

Из диаграммы видно, что основная масса значений лежит в диапазоне от 60 до 75, а пик около 67–70. Это свидетельствует о преобладании студентов со средним уровнем успеваемости, при этом наблюдаются как более низкие, так и более высокие значения экзаменационного балла. Основываясь на анализе распределения целевой переменной, а также с учетом педагогического смысла и интерпретации уровней академической успеваемости, в таблице 1 представлены варианты классификации целевой переменной.

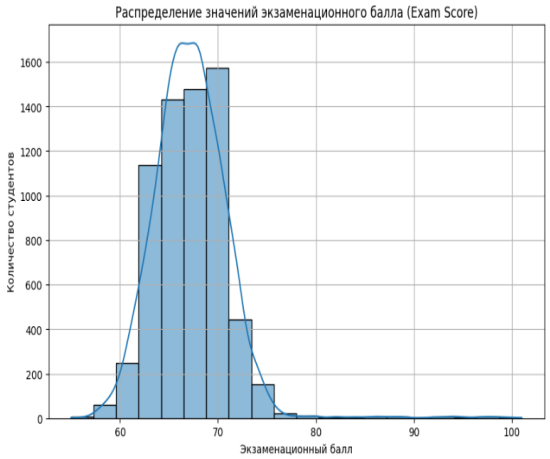

Рис. 3 – Диаграмма распределения значений выходного параметра

Fig. 3 – Output parameter value distribution diagram

Таблица 1 – Варианты градаций выходного параметра

Table 1 – Output parameter gradation options

Классы выходной переменной	Схема группировки
2 (Ранее выявление зоны риска)	0 – низкая успеваемость (exam_score ≤ 65) 1 – средняя и высокая (exam_score > 65)
3 (Разделение по уровню успеваемости)	0 – низкая успеваемость (exam_score ≤ 63) 1 – средняя успеваемость (exam_score: 63-70) 2 – высокая успеваемость (exam_score > 70)
4 (Более четкая градация знаний)	0 – низкая успеваемость (exam_score ≤ 62) 1 – ниже среднего (exam_score: 62-65) 2 – средняя успеваемость (exam_score: 65-70) 3 – высокая успеваемость (exam score > 70)

Разделение на выборки и масштабирование

После завершения этапов анализа, очистки и кодирования исходных данных выполнен переход к формированию обучающей и тестовой выборок, а также масштабированию числовых признаков. Данный этап является обязательным при построении моделей машинного обучения, особенно нейросетевых, поскольку позволяет обеспечить корректное обучение модели и повысить стабильность процесса оптимизации.

Исходный датасет разделен на обучающую и тестовую выборки в соотношении 80 % и 20 % соответственно. Такое соотношение является общепринятым и обеспечивает достаточный объем данных как для обучения модели, так и для объективной оценки ее обобщающей способности [15, 16]. Разделение выполнено случайным образом с фиксированным значением параметра random_state, что обеспечивает воспроизводимость результатов экспериментов. Дополнительно использована стратификация по целевой переменной, позволяющая сохранить исходное распределение классов в обеих выборках.

Для корректной работы нейросетевой модели выполнена процедура масштабирования входных признаков. Масштабирование необходимо в связи с тем, что числовые признаки датасета имеют различные диапазоны значений. Отсутствие масштабирования может приводить к доминированию отдельных признаков и замедлению процесса обучения нейронной сети.

В данной работе для нормализации данных использован метод стандартизации, реализованный с помощью алгоритма StandardScaler. Данный метод преобразует значения признаков таким образом, что для каждого признака среднее значение становится равным нулю, а стандартное отклонение – единице. Стандартизация описывается выражением:

$$x' = \frac{x - \mu}{\sigma},$$

где x – исходное значение признака, μ – среднее значение признака в обучающей выборке, σ – стандартное отклонение.

Обучение нейросетевой модели выполнено на обучающей выборке, после чего полученные параметры преобразования применялись как к обучающим, так и к тестовым данным. Такой подход предотвращает утечку информации из тестовой выборки и обеспечивает корректность оценки качества модели [14].

Результатом данного этапа является подготовка масштабированных обучающей и тестовой выборок, пригодных для последующего обучения и тестирования нейросетевой модели. Проведенное масштабирование позволило обеспечить сопоставимость признаков, ускорить сходимость алгоритма обучения и повысить устойчивость модели к изменениям входных данных.

Разработка нейросетевой модели прогнозирования

Для решения поставленной задачи выбрана нейросетевая модель типа многослойного персептрона (Multilayer Perceptron, MLP). Данный тип модели широко применяется для задач классификации табличных данных и обладает преимуществами:

– способность моделировать сложные нелинейные зависимости;

– универсальность архитектуры;

– совместимость с предварительно обработанными числовыми признаками;

– возможность адаптации под различное количество выходных классов.

В отличие от сверточных и рекуррентных нейронных сетей, многослойный персептрон не требует специальной структуры входных данных, что делает его оптимальным выбором для анализа образовательных данных.

В рамках данной работы для реализации нейросетевой модели прогнозирования уровня успеваемости студентов использован классификатор MLPClassifier из библиотеки scikit-learn, представляющий собой многослойную полносвязную нейронную сеть прямого распространения [17].

Архитектура нейросетевой модели состоит из одного входного слоя, двух скрытых слоя с 64 и 32 нейронами и одного выходного слоя. Такая архитектура позволяет эффективно выявлять нелинейные зависимости между входными признаками и целевой переменной, не усложняя модель избыточным числом параметров. В качестве функции активации выбрана функция ReLU, которая широко применяется в задачах классификации благодаря вычислительной эффективности и способности ускорять процесс обучения. Для оптимизации весовых коэффициентов использован алгоритм Adam, обеспечивающий устойчивую и быструю сходимость модели. Максимальное число эпох обучения нейросетевой модели - 300, а фиксированное значение параметра random_state использовано для обеспечения воспроизводимости результатов эксперимента.

После определения архитектуры нейросетевой модели выполняется ее обучение с помощью метода fit, в который передаются масштабированные признаки обучающей выборки и соответствующие значения целевой переменной. В процессе обучения нейронная сеть настраивает внутренние веса для минимизации ошибки классификации [18]. Прогнозирование классов целевой переменной на тестовой выборке выполнено методом predict. Полученные результаты сравниваются с истинными значениями, после чего проводится оценка качества нейросетевой модели.

Проведение экспериментальных исследований

С целью обоснования выбора нейросетевой модели и оценки её эффективности в задаче прогнозирования академической успеваемости студентов был проведён ряд экспериментальных исследований. Эксперименты включали сравнительный анализ нейросетевой модели с базовыми алгоритмами машинного обучения, а также исследование влияния количества классов целевой переменной на качество прогнозирования.

На первом этапе экспериментальных исследований проведено сравнение нейросетевой модели с базовыми алгоритмами машинного обучения с целью обоснования выбора основного метода прогнозирования. В качестве алгоритмов сравнения были выбраны логистическая регрессия и случайный лес, представляющие соответственно линейный и ансамблевый подходы к решению задачи классификации. Сравнение проводилось для варианта разбиения целевой переменной на три класса, для всех моделей использовались одинаковые обучающая и тестовая выборки, а также единые процедуры предобработки данных. В таблице 2 представлены результаты сравнения метрик нейросетевой модели и алгоритмов машинного обучения.

Полученные результаты показывают, что нейросетевая модель демонстрирует более высокие значения метрик качества по сравнению с логистической регрессией и лучшие результаты по сравнению со случайным лесом. Это подтверждает целесообразность использования нейросетевой модели в качестве основного инструмента прогнозирования академической успеваемости студентов.

Таблица 2 – Сравнение нейросетевой модели с базовыми алгоритмами

Table 2 – Comparison of the neural network model with basic algorithms

Модель	Accuracy	Precision	Recall	F1 score
Логистическая регрессия	0,75	0,78	0,75	0,74
Случайный лес	0,80	0,79	0,80	0,79
Нейросетевая модель	0,84	0,84	0,84	0,83

После обоснования выбора нейросетевой модели в качестве основного инструмента прогнозирования целесообразно исследовать влияние параметров постановки задачи на качество классификации. В частности, важным аспектом является выбор числа классов целевой переменной, от которого напрямую зависит как точность прогнозирования, так и практическая интерпретируемость результатов.

С целью анализа влияния количества классов целевой переменной на качество прогнозирования проведено тестирование нейросетевой модели. В рамках экспериментов рассмотрены варианты разбиения успеваемости на два, три и четыре класса. Для каждого варианта выполнено обучение нейросетевой модели с одинаковыми параметрами, после чего проведена оценка качества классификации. Результаты сравнения моделей представлены в таблице 3.

Таблица 3 – Результаты проведения экспериментов

Table 3 – Results of the experiments

№ теста	Количество выходных классов	Accuracy	Precision	Recall	F1 score
1	2	0,88	0,88	0,86	0,86
2	3	0,84	0,84	0,84	0,83
3	4	0,75	0,76	0,69	0,71

Проведенные экспериментальные исследования показали, что увеличение числа классов целевой переменной приводит к закономерному снижению качества классификации. Наилучшие значения метрик получены при использовании двух классов, однако такая классификация является слишком грубой с точки зрения педагогической интерпретации. Использование трех классов обеспечивает баланс между качеством прогнозирования и информативностью результатов. Таким образом, оптимальным решением является использование 3 выходных классов.

Важным аспектом практического применения моделей прогнозирования успеваемости, подтверждённым в ходе проведённых экспериментов, является

интерпретируемость полученных результатов, особенно в контексте образовательной аналитики. Несмотря на то, что нейросетевые модели традиционно рассматриваются как «чёрные ящики», в рамках данной работы интерпретируемость обеспечивается за счёт формулировки задачи в виде классификации с ограниченным числом педагогически осмысленных классов.

Каждый выходной класс нейросетевой модели соответствует определённому уровню академической успеваемости, что позволяет интерпретировать прогноз в терминах образовательной успеваемости (низкая, средняя и высокая) без необходимости анализа внутренних параметров модели. Такой подход делает результаты прогнозирования понятными для преподавателей и администраторов образовательных организаций и позволяет использовать модель в качестве инструмента поддержки принятия решений.

Дополнительно интерпретируемость модели обеспечивается предварительным анализом данных, включающим построение корреляционной матрицы и отбор наиболее значимых признаков. Это позволяет связать результаты прогнозирования с группами факторов, отражающими учебную активность, социальные и организационные характеристики обучающихся. Таким образом, нейросетевая модель не только обеспечивает высокую точность прогнозирования, но и сохраняет практическую интерпретируемость, необходимую для внедрения в реальные информационные системы образования.

Заключение

В ходе проведенного исследования решены следующие задачи:

1) проведен анализ основных методов и моделей прогнозирования, применяемых при аналитике образовательных данных;

2) выполнена подготовка исходного набора данных для его дальнейшего использования при обучении модели;

3) построена нейросетевая модель прогнозирования академической успеваемости студентов;

4) проведены обучение и тестирование разработанной модели;

5) проведен сравнительный анализ результатов работы модели при различном количестве выходных классов и определен наиболее эффективный вариант классификации.

С целью развития научного направления, связанного с прогнозированием академической успеваемости студентов на основе методов машинного обучения и нейронных сетей, целесообразно дальнейшее совершенствование разработанной модели за счет расширения исходного датасета редкими значениями целевой переменной для большей балансировки классов и, тем самым, для улучшения эффективности прогнозирования модели. Перспективным направлением является интеграция разработанного подхода в информационные системы образовательных учреждений для обеспечения поддержки принятия управленческих решений, а также адаптация модели для прогнозирования успеваемости на различных этапах обучения.

Литература

1. Агабабян Е.О., Юданова В.В., *Молодежь и научно-технический прогресс в современном мире*, 4-7 (2022).
2. Богатырева М.Р., Корчевская Е.А., *Молодость. Интеллект. Инициатива*, 18-19 (2023).
3. Моисеев В.Б., Зубков А.Ф., Деркаченко В.Н., *Информационные и телекоммуникационные технологии в образовании*, 6 (113), 169-173 (2010).
4. Будаева А.А, *XV Ежегодная Международная научно-техническая конференция Кавказского государственного технологического университета: Сборник докладов*, 9-16 (2018).
5. Al-Shehri H., *IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, IEEE, 1-4 (2017).
6. Ясинский И.Ф., *Вестник ИГЭУ*, 4, 1-4 (2007).
7. Прошкина Е.Н., Балашова И.Ю. *Технические науки: традиции и инновации: материалы III Международной научной конференции: Молодой ученый*, 24-28 (2018).
8. Харламова И.Ю, *Базис*, 1 (1), 57-59 (2017).
9. Русаков С.В., Русакова О.Л., Посохина К.А., *Современные информационные технологии и ИТ-образование»*, **14**, 4, 815- 822 (2018).
10. Student Perfomance Factors. – URL: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors (дата обращения 28.11.2025).
11. Губин Е.И., *Наука и бизнес: Пути развития*, 3 (105), 27-31 (2020).
12. Якунин Ю.Ю., *Информатика и образование*, **38**, 4, 28-43 (2023).
13. Рекуррентные нейронные сети и LSTM. – URL: https://nweb42.com/books/r-lang/rekurrentnye-neyronnye-seti-i-lstm/ (дата обращения 29.11.2025).
14. Asselman A., Khaldi M., Aammou S., *Interactive Learning Environments*, **31**, 6, 3360-3379 (2023).
15. Баклашов Д.М., Жашкова Т.В., Мартышкин А.И., *Современные информационные технологии*, 39 (39), 6-9 (2024).
16. Милованович Н.Г., Басс Н.В., *Тенденции развития науки и образования*, 107-4, 97-100 (2024).
17. Sckit-learn: Machine learning in Python. – URL: https://scikit-learn.org/stable/index.html (дата обращения 30.11.2025).
18. Сальникова Н.А., Реклер Е.Н., *Тенденции развития науки и образования*, 106-9, 95-97 (2024).

References

1. Agababyan E.O., Yudanova V.V., Youth and Scientific and Technological Progress in the Modern World, 4-7 (2022).
2. Bogatyreva M.R., Korchevskaya E.A., Youth. Intelligence. Initiative, 18-19 (2023).
3. Moiseev V.B., Zubkov A.F., Derkachenko V.N., Information and Telecommunication Technologies in Education, 6 (113), 169-173 (2010).
4. Budaeva A.A., XV Annual International Scientific and Technical Conference of the Caucasus State Technological University: Collection of Papers, 9-16 (2018).
5. Al-Shehri H., IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, 1-4 (2017).
6. Yasinsky I.F., ISEU Bulletin, 4, 1-4 (2007).
7. Proshkina E.N., Balashova I.Yu. Engineering Sciences: Traditions and Innovations: Proceedings of the III International Scientific Conference: Young Scientist, 24-28 (2018).
8. Kharlamova I.Yu., Basis, 1 (1), 57-59 (2017).
9. Rusakov S.V., Rusakova O.L., Posokhin K.A., Modern Information Technologies and IT Education, 14, 4, 815-822 (2018).

10. Student Performance Factors. – URL: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors (accessed 28.11.2025)
11. Gubin E.I., Science and Business: Development Paths, 3 (105), 27-31 (2020).
12. Yakunin Yu.Yu., Computer Science and Education, 38, 4, 28-43 (2023).
13. Recurrent Neural Networks and LSTM. – URL: https://nweb42.com/books/r-lang/rekurrentnye-neyronnye-seti-i-lstm/ (accessed 29.11.2025).
14. Asselman A., Khaldi M., Aammou S., Interactive Learning Environments, 31, 6, 3360-3379 (2023).
15. Baklashov D.M., Zhashkova T.V., Martyshkin A.I., Modern Information Technologies, 39 (39), 6-9 (2024).
16. Milovanovich N.G., Bass N.V., Trends in the Development of Science and Education, 107-4, 97-100 (2024).
17. Sckit-learn: Machine learning in Python. – URL: https://scikit-learn.org/stable/index.html (date of access 30.11.2025).
18. Salnikova N.A., Rekler E.N., Trends in the Development of Science and Education, 106-9, 95-97 (2024).

© К. Г. Осанов – магистрант кафедры Систем информационной безопасности (СИБ), Казанский национальный исследовательский технический университет им. А.Н. Туполева (КНИТУ им. А.Н. Туполева), Казань, Россия, kirillosanov@yandex.ru; **Р. М. Хусаинов** – ассистент кафедры СИБ, КНИТУ им. А.Н. Туполева, rumil_husainov98@mail.ru.

© K. G. Osanov – PhD-student of Information Security Systems (ISS) Department, Kazan National Research Technical University named after A.N. Tupolev (KNRTU named after A.N. Tupolev), Kazan, Russia, kirillosanov@yandex.ru; **R. M. Khusainov** – Assistant of the ISS Department, KNRTU named after A.N. Tupolev, rumil_husainov98@mail.ru.

Дата поступления рукописи в редакцию – 20.01.26
Дата принятия рукописи в печать – 02.02.26