

АНАЛИЗ БОЛЬШИХ ДАННЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ ФАКТОРОВ УСПЕШНОЙ АДАПТАЦИИ ИНОСТРАННЫХ СТУДЕНТОВ

Погодин И.М., аспирант, Российский институт стандартизации, г. Москва

Садыхбеков А.Д., аспирант, Московский городской педагогический университет, г. Москва

В статье представлена разработка аналитической системы больших данных (АСБД) на основе искусственного интеллекта для раннего выявления иностранных студентов из группы риска академической неуспеваемости. Система использует двухэтапный итеративный подход, включающий методы машинного обучения (ML) и глубокого обучения (DL) для анализа данных о взаимодействии студентов с системой управления обучением (LMS). На первом этапе применяются алгоритмы машинного обучения (J48, Random Forest, OneR, DecisionStump, NBTree) с использованием ансамблевой техники, на втором - алгоритмы глубокого обучения (LSTM, MLP, Sequential Model) после балансировки данных. Модель учитывает различные факторы, включая языковую подготовку, культурную адаптацию и академическую успеваемость студентов. Разработанная система позволяет своевременно выявлять проблемные ситуации в обучении иностранных студентов и предоставлять целенаправленную поддержку для повышения эффективности их образовательного процесса.

Ключевые слова: большие данные, машинное обучение, глубокое обучение, иностранные студенты, академическая успеваемость, прогнозирование, система поддержки принятия решений.

Цитирование: Погодин И.М., Садыхбеков А.Д. Анализ больших данных для прогнозирования факторов успешной адаптации иностранных студентов // Информационно-экономические аспекты стандартизации и технического регулирования. 2025. № 4(85). С. 39–45.

ВВЕДЕНИЕ

Раннее прогнозирование успеваемости иностранных студентов имеет жизненно важное значение для образовательных учреждений, поскольку оно позволяет своевременно вмешиваться и оказывать целенаправленную поддержку для преодоления языковых, культурных и академических барьеров. Различные исследования изучали методы интеллектуального анализа данных, направленные на прогнозирование академической успеваемости иностранных студентов на разных уровнях: на уровне степени, года обучения, курса и экзаменов, с учетом их уникальных потребностей и вызовов [1]. Примечательно, что прогнозирование на уровне отдельных экзаменов часто исключается из-за временного характера результатов одного экзамена, хотя для иностранных студентов эти показатели могут быть особенно важны для оценки адаптации [1]. Успех иностранных студентов служит важным показателем качества образовательных учреждений и их способности к интернационализации, включая в себя не только академические достижения, но и культурную адаптацию, языковую компетенцию, социальную интеграцию и результаты после окончания обучения [1].

Ранее авторами исследовалась эффективность применения аналитики больших данных для мониторинга успеваемости и социально-экономической адаптации иностранных сту-

дентов, изучающих русский язык в онлайн-формате. Это исследование показало, что такие данные позволяют не только прогнозировать успеваемость студентов, но и влиять на их адаптацию в новой социальной и культурной среде [2, 3].

Цель данного исследования заключается в разработке модели аналитической системы больших данных (далее – АСБД) на основе искусственного интеллекта для раннего выявления иностранных студентов из группы риска академической неуспеваемости и своевременного принятия мер по повышению показателей их успеваемости и адаптации.

Методы машинного обучения (ML) играют важную роль в прогнозировании академической успеваемости иностранных студентов. Алгоритмы, такие как случайные леса, логистическая регрессия и другие методы ML, особенно эффективны при анализе множества факторов, влияющих на успех иностранных учащихся, включая языковые барьеры, культурную адаптацию и академическую подготовку [4]. Интеграция методов глубокого обучения, включая сверточные нейронные сети (CNN) и рекуррентные нейронные сети (RNN), позволяет точнее выявлять паттерны успеваемости иностранных студентов и учитывать их уникальные потребности в обучении [4]. Эти передовые методы помогают образовательным учреждениям своевременно отслеживать прогресс иностранных студентов и предоставлять целенаправленную поддержку [5].

Прогнозное моделирование в образовании иностранных студентов основывается на выявлении ключевых индикаторов, указывающих на вероятность их академического успеха. Критически важные показатели включают не только стандартные данные о посещаемости, оценках и результатах тестов, но и специфические факторы: уровень владения языком обучения, степень культурной адаптации, социальную интеграцию в новой среде, а также академические привычки, сформированные в родной стране [6]. Анализируя эти факторы, образовательные учреждения могут предоставлять целенаправленную поддержку иностранным студентам, обеспечивая своевременное вмешательство в области языковой подготовки, культурной адаптации и академической интеграции, что в конечном итоге способствует успешному завершению их образовательной программы.

В то время как существующая литература подчеркивает важность выявления значимых атрибутов для прогнозирования успеваемости иностранных студентов, остаются проблемы, связанные с точностью и понятностью моделей, особенно в контексте культурных и языковых различий. Исследования указывают на необходимость усовершенствованных прогностических моделей, которые не только сохраняют баланс между точностью и способностью четко объяснять свои результаты, а также учитывают специфические факторы, влияющие на успеваемость иностранных студентов [7]. Будущие исследования должны сосредоточиться на доработке этих моделей и изучении гибридных подходов, объединяющих статистические методы с методами машинного обучения для улучшения прогностических возможностей при сохранении прозрачности для педагогов и лиц, принимающих решения, с особым вниманием к межкультурным аспектам образования [4, 8].

АРХИТЕКТУРА И КОМПОНЕНТЫ АНАЛИТИЧЕСКОЙ СИСТЕМЫ БОЛЬШИХ ДАННЫХ

В основе разработки и обучения модели АСБД для анализа успеваемости иностранных студентов лежит верхнеуровневая структура системы, представленная на рис. 1. Данная модель определяет основные компоненты и принципы функционирования системы АСБД с учетом специфики иностранных учащихся, включая источники данных, алгоритмы обработки и конечные результаты.

Входные данные поступают в систему из двух основных источников:

- **Данные о взаимодействии иностранных студентов с LMS (Learning Management System)¹** включают в себя информацию о времени активности, количестве выполненных упражнений, языковых паттернах взаимодействия, использовании переводчиков и других па-

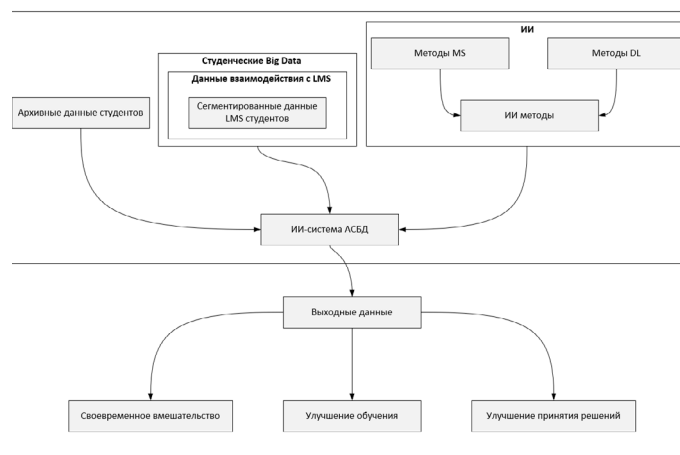


Рис. 1. Верхнеуровневая структура системы АСБД

раметрах поведения студентов в системе. Эти данные подвергаются сегментации и предварительной обработке с учетом языковых и культурных особенностей.

- **Исторические данные иностранных студентов** содержат накопленную информацию о результатах языковых тестов, академической успеваемости, культурной адаптации и других параметрах, необходимых для построения и настройки моделей прогнозирования.
- **Ядром системы** является блок ИИ-система АСБД, в котором осуществляется анализ и обработка данных с использованием методов машинного обучения (ML) и глубокого обучения (DL) [9]. Эти методы адаптированы для работы с многоязычными данными, что особенно важно для анализа и прогнозирования академической успеваемости иностранных студентов. В качестве алгоритмов машинного обучения используются модели, такие как Random Forest, J48, OneR, DecisionStump и NBTree. Методы глубокого обучения включают в себя многослойные перцептроны (MLP) и сети с долговременной краткосрочной памятью (LSTM)², особенно эффективные при анализе последовательных языковых данных. АСБД применяет ансамблевые методы, технику кросс-валидации и алгоритмы балансировки для повышения точности моделей [10, 11].

Результаты обработки в АСБД предоставляют широкие возможности для повышения эффективности образовательного процесса иностранных студентов. Основные выходные данные включают:

- **Своевременные интервенции** – раннее выявление проблемных ситуаций в обучении иностранных студентов, включая языковые барьеры и трудности культурной адаптации, и предоставление персонализированных рекомендаций по их устранению.
- **Улучшение результатов обучения студентов** – использование прогнозов для персонализации обучения и адап-

¹ Система управления обучением (LMS) – программное приложение, которое помогает организациям администрировать, документировать, отслеживать, составлять отчеты, автоматизировать и поставлять учебные курсы, программы обучения.

² LSTM – сети с долговременной краткосрочной памятью (от англ. Long Short Term Memory).

тации образовательных стратегий под культурные и языковые особенности иностранных студентов.

- **Улучшение принятия решений** – предоставление достоверных данных и аналитических выводов для преподавателей и администрации образовательных учреждений с учетом специфики работы с иностранными студентами.

ПРОЕКТИРОВАНИЕ И РАЗРАБОТКА МОДЕЛИ

В данном исследовании для разработки прогностической модели АСБД, направленной на выявление студентов с риском академической неуспеваемости, применены два подхода: модели машинного обучения и глубокого обучения. На основе данных о взаимодействии студентов с LMS алгоритмы были реализованы в итеративном режиме, что позволило сравнивать их эффективность с использованием матрицы ошибок. Для оценки качества работы моделей на каждом этапе анализировались метрики точности, чувствительности и специфичности. По итогам циклической проверки наилучший алгоритм будет выбран в качестве основы для финальной версии системы, обеспечивающей непрерывное улучшение ее прогностической способности.

АСБД создается посредством серии задач, состоящих из сбора данных, предварительной обработки данных, анализа данных с помощью алгоритмов ИИ, оценки и успешного принятия решений. Все эти задачи адаптированы для разработки и оценки прогностических моделей на основе машинного и глубокого обучения. Рабочий процесс обучения системы на основе ИИ проиллюстрирован на рис. 2–4.

Описание работы модели:

1. Сбор данных и создание вектора размерности

- **Начало процесса** – Инициализация работы системы.
- **Набор данных студентов** – Получение исходных данных в формате CSV.
- **Агрегация файлов данных** – Объединение наборов для создания единого вектора признаков.
- **Создание и проверка вектора размерности** – Формирование вектора с параметрами (таблица).

2. Предварительная обработка данных

- **Очистка переменных и экземпляров** – Удаление некорректных данных.
- **Проверка и удаление отрицательных значений** – Фильтрация недопустимых показателей.
- **Проверка и удаление пустых значений** – Обработка отсутствующих данных.
- **Масштабирование данных** – Нормализация значений.
- **Преобразование категориальных переменных** – Кодирование нечисловых данных.

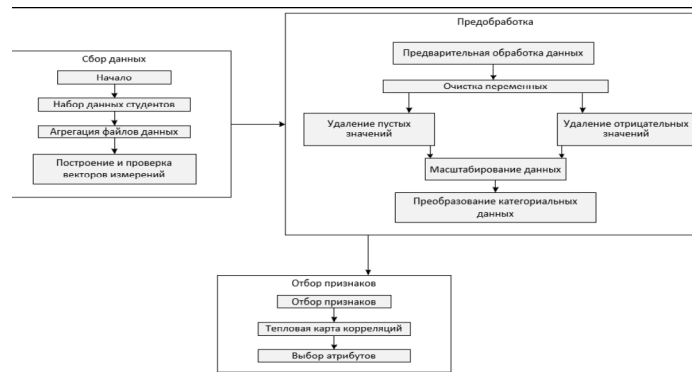


Рис. 2. Подготовка и предварительный анализ данных

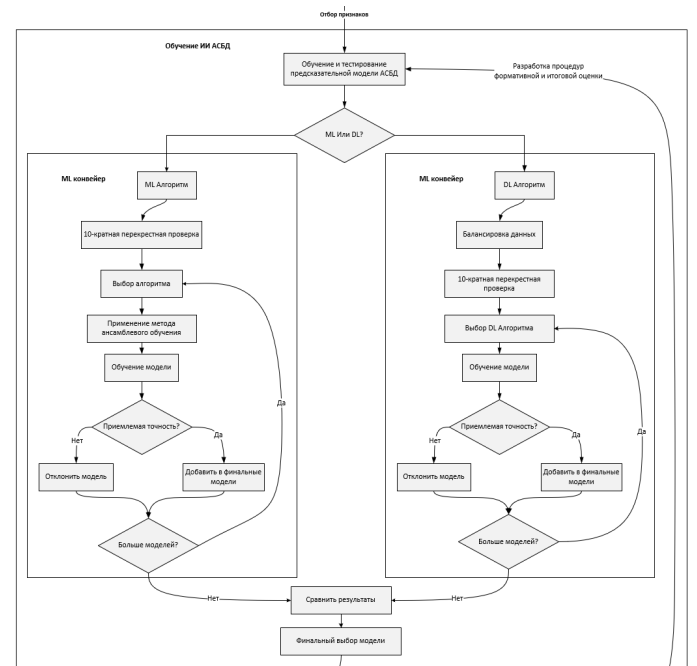


Рис. 3. Обучение ИИ АСБД

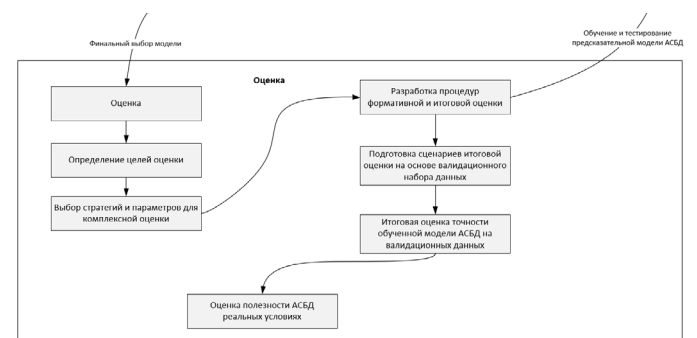


Рис. 4. Оценка и итоговый выбор

Результат: Подготовленный вектор размерности с обработанными данными.

3. Отбор признаков

- **Построение тепловой карты корреляций** – Анализ взаимосвязей между признаками.
- **Выбор значимых атрибутов** – Определение ключевых признаков для модели.

Результат: Оптимизированный вектор признаков.

4. Алгоритмы ИИ – Обучение и тестирование прогностической модели

4.1. Первая итерация (Алгоритмы ML)

1. **Настройка 10-кратной кросс-валидации** – Разделение данных для обучения и проверки.

2. **Формирование списка алгоритмов ML** – Подготовка набора методов машинного обучения.

3. **Выбор алгоритма ML из списка:**

- J48 (Дерево решений)
- Случайный лес
- OneR (Простое правило)
- DecisionStump (Пень решений)
- NBTree (Наивный Байес)

4. **Применение ансамблевого метода** – Усиление результатов обучения.

5. **Обучение модели** – Тренировка на подготовленных данных.

6. **Формативная оценка** – Проверка точности на тестовой выборке.

7. **Приемлемая точность?**

- **Да** → Включение в финальный набор моделей
- **Нет** → Исключение модели

8. **Есть ли другие ML модели?**

- **Да** → Повторить выбор алгоритма
- **Нет** → Перейти к сравнению результатов

4.2. Вторая итерация (Алгоритмы DL)

1. **Балансировка и аугментация данных** – Улучшение качества набора данных.

2. **Настройка 10-кратной кросс-валидации** – Подготовка к обучению.

3. **Формирование списка алгоритмов DL** – Подбор методов глубокого обучения.

4. **Выбор алгоритма DL из списка:**

- Многослойный перцептрон
- LSTM (сети с долговременной краткосрочной памятью)
- Последовательная модель

5. **Обучение модели** – Тренировка на подготовленных данных.

6. **Формативная оценка** – Проверка точности.

7. **Приемлемая точность?**

- **Да** → Включение в финальный набор
- **Нет** → Исключение модели

8. **Есть ли другие DL модели?**

- **Да** → Повторить выбор алгоритма
- **Нет** → Перейти к сравнению

4.3. Сравнение и итоговый выбор

- Анализ результатов всех моделей.
- Выбор оптимальной модели для создания системы поддержки принятия решений.

5. Оценка прогностической модели

1. **Определение целей оценки** – Постановка критериев эффективности.

2. **Выбор стратегий оценки** – Определение методологии тестирования.

3. **Разработка процедур оценки** – Создание методов проверки.

4. **Подготовка валидационных сценариев** – Формирование тестовых случаев.

5. **Итоговая оценка точности** – Проверка на валидационном наборе.

6. Оценка полезности АСБД в реальных условиях.

После применения моделей машинного и глубокого обучения, результаты прогнозирования сравниваются с реальными данными успеваемости студентов, что позволяет точно оценить их предсказательную силу [10]. Также производится финальная проверка точности на валидационном наборе с целью минимизировать количество ложных прогнозов.

Набор данных включает 16 характеристик, из которых наиболее значимыми для анализа адаптации иностранных студентов являются: время бездействия во время активности,

Перечень идентификаторов набора данных

№	ИДЕНТИФИКАТОРЫ НАБОРА ДАННЫХ	КРАТКОЕ ОПИСАНИЕ
1	session_Id	Числовое значение для номера сессии
2	student_Id	Числовое значение для номера иностранного студента
3	exercise	Числовое значение для номера упражнения в определенной сессии
4	activity	Текстовое значение для сокращения категорий активности
5	start_time	Значение даты для времени начала активности
6	end_time	Значение даты для времени окончания активности
7	idle_time	Числовое значение времени бездействия во время активности (важно для анализа трудностей с языком)
8	mouse_wheel	Числовое значение количества прокруток колесика мыши
9	mouse_wheel_click	Числовое значение количества кликов колесиком мыши
10	mouse_click_left	Числовое значение количества левых кликов мыши
11	mouse_click_right	Числовое значение количества правых кликов мыши
12	keystroke	Числовое значение общего количества нажатий клавиш в определенной активности
13	mouse_movement	Числовое значение расстояния, пройденного движениями мыши
14	final_total	Числовое значение итоговых оценок иностранного студента в конце семестра
15	language_level	Оценка уровня владения языком обучения (от A1 до C2)

количество кликов мышью и нажатий клавиш. Эти показатели позволяют определить, насколько активно студент взаимодействует с материалами курса и насколько эффективно происходит его погружение в учебный процесс. Перечень данных представлен в таблице.

Набор данных предварительно обрабатывается и нормализуется, а признаки отбираются с помощью корреляционного анализа для построения размерного вектора, включающего категоризированные признаки. Набор данных состоит из нескольких документов в формате CSV, содержащих данные о сессиях и студентах. Дополнительный CSV-документ содержит итоговые оценки каждого студента, посещавшего занятия в конце семестра, и показывает результат студента.

В процессе предварительной обработки данных отрицательные, пустые или нулевые значения удаляются из набора данных. Размерный вектор строится путем агрегирования каждого признака для каждого студента и объединения их с итоговыми оценками студентов. Итоговые оценки студентов преобразуются в классификационные категориальные переменные, т.е. «Сдал» или «Не сдал». Соответствующие признаки отбираются из набора данных из 14 признаков и 1 категориальной переменной с использованием тепловой карты корреляции для определения положительной или отрицательной корреляции с итоговым результатом (final total). Например, тепловая карта показывает, что когда количество нажатий клавиш имеет положительную корреляцию с итоговым результатом, т.е. когда значение

«keystroke» высокое, то существует более высокая вероятность того, что итоговый результат (final total) также будет высоким. Этот преобразованный набор данных затем используется для обучения прогностической модели с помощью алгоритмов ML и DL для выявления студентов, находящихся в зоне риска неуспеваемости.

Для повышения точности прогностической классификации модели и разработки оптимальной версии АСБД используется двухэтапный итеративный подход. На первом этапе выполняется обучение модели с применением нескольких алгоритмов машинного обучения, которые затем оптимизируются с помощью ансамблевой техники адаптивного бустинга. На втором этапе происходит балансировка набора данных методом синтетической избыточной выборки меньшинства с последующим применением алгоритмов глубокого обучения для улучшения точности прогнозирования.

В первой итерации используются пять алгоритмов машинного обучения с учителем (J48, Random Forest, OneR, Decision Stump, NBTree), основанных на деревьях решений. Эти алгоритмы применяют последовательность условных операторов для создания точных и интерпретируемых прогнозов. Модель проходит валидацию с помощью 10-кратной перекрестной проверки, после чего выбирается наиболее эффективный вариант для интеграции в систему. Для работы с потоковыми данными в реальном времени планируется использование Apache Spark.

Вторая итерация фокусируется на улучшении качества данных путем корректировки распределения классов и расширения набора данных. Будут применены три алгоритма глубокого обучения (LSTM, MLP и SM), использующие многослойные нейронные сети. Это позволит значительно повысить точность классификации и снизить количество ложных прогнозов, что критически важно для корректного выявления студентов группы риска.

Важно отметить, что выбор алгоритмов был обусловлен особенностями задачи: для ML мы отдаем предпочтение методам, хорошо работающим с табличными данными и деревьями решений, а DL методы, такие как LSTM, идеально подходят для анализа последовательных данных, например, данных о взаимодействии студентов с LMS.

ЗАКЛЮЧЕНИЕ

В ходе проведенного исследования была разработана аналитическая система больших данных (АСБД), предназначенная для прогнозирования академической успеваемости и адаптации иностранных студентов. Основной целью работы было создание прогностической модели, использующей методы машинного и глубокого обучения для вы-

явления студентов, находящихся в группе риска, с целью последующего предоставления им своевременной поддержки.

Предложенная модель позволяет учитывать широкий спектр факторов, таких как языковая подготовленность, культурная адаптация и академическая подготовка студентов, что дает возможность более точно прогнозировать их успеваемость и адаптацию в образовательной среде. Модели машинного и глубокого обучения, использованные в этом исследовании, обеспечивают основу для построения прогностических систем, которые могут быть применены в разных образовательных контекстах.

Однако стоит отметить, что дальнейшие исследования должны включать тестирование и валидацию модели на реальных данных для оценки ее точности и эффективности в прогнозировании. Ожидается, что будущие шаги будут направлены на улучшение модели и на ее адаптацию с учетом новых данных и практических условий.

Таким образом, данное исследование предоставляет основу для дальнейших разработок в области использования больших данных для поддержки иностранных студентов.

Список использованных источников и литературы / References

1. Ifenthaler D., Gibson D.C. Implementation and Application of AI in Higher Education: Case Studies and Recommendations // *International Journal of Educational Technology in Higher Education*. 2020. Vol. 17, № 1. <https://doi.org/10.1186/s41239-020-0177-7>
2. Погодин И.М., Садыкбеков А.Д. Применение аналитики больших данных в онлайн-обучении иностранных студентов, изучающих русский язык в целях их адаптации // *Информационно-экономические аспекты стандартизации и технического регулирования*. 2025. № 1(82). С. 76–81. / Pogodin I.M. Sadykhbekov A.D.o. Application of Big Data Analytics in Online Education of Foreign Students Learning Russian for Adaptation Purposes. *Information-Economic Aspects of Standardization and Technical Regulation*, 2025, no. 1(82), pp. 76–81.
3. Бурый А.С., Погодин И.М. Оценка качества больших данных. Часть 1. Основные понятия и метрики // *Информационно-экономические аспекты стандартизации и технического регулирования*. 2024. № 3(78). С. 49–58 / Buryi A.S., Pogodin I.M. Ocenka kachestva bol'shih dannyh. Part 1. Osnovnye ponyatiya i metriki *Information-Economic Aspects of Standardization and Technical Regulation*, 2024, no. 3(78), pp. 49–58.
4. Delen D. Predicting Student Attrition with Data Mining Methods // *Journal of College Student Retention: Research // Theory & Practice*. 2011. T. 13, № 1, pp. 17–35.
5. Бурый А.С., Цаплина О.С. Генеративный искусственный интеллект цифрового университета // *Информационно-экономические аспекты стандартизации и технического регулирования*. 2024. № 5(80). С. 85–91. / Buryi A.S., Tsaplina O.S. Generativnyj iskusstvennyj intellekt cifrovogo universiteta. *Information-Economic Aspects of Standardization and Technical Regulation*, 2024, no. 5(80), pp. 85–91.
6. Borna M-R., Saadat H., Hojjati A.T., Akbari E. Analyzing click data with AI: implications for student performance prediction and learning assessment // *Frontiers in Education*. 2024. T. 9. Art. 1421479. <https://doi.org/10.3389/educ.2024.1421479>
7. AI Agents for Student Success Prediction: Complete Guide // *Rapid Innovation*. 2024. URL: <https://www.rapidinnovation.io/post/ai-agents-for-student-success-prediction> (Дата обращения: 25.03.2025).
8. Guanin-Fajardo J.H., Guaña-Moya J., Casillas J. Predicting Academic Success of College Students Using Machine Learning Techniques // *Data*. 2024. № 9. С. 60. <https://doi.org/10.3390/data9040060>
9. Wang K. Optimized ensemble deep learning for predictive analysis of student achievement // *PLoS ONE*. 2024. № 19(8). С. e0309141. DOI: <https://doi.org/10.1371/journal.pone.0309141>
10. 7 Metrics for Tracking Student Success with AI // *Quizcat*. 2024. URL: <https://www.quizcat.ai/blog/7-metrics-for-tracking-student-success-with-ai> (Дата обращения: 25.03.2025).
11. Baig T., Bandalai M.M., Bhutto M.Y., Gul H. The Impact of Artificial Intelligence on Adaptive Learning: A Comparative Study of Traditional and AI-Based Pedagogy // *Dialogue Social Science Review (DSSR)*. 2025. T. 3, № 3. С. 926–941.

BIG DATA ANALYSIS FOR PREDICTING FACTORS OF SUCCESSFUL ADAPTATION OF FOREIGN STUDENTS

Pogodin I.M., PhD student, Russian Standardization Institute, Moscow

Sadykhbekov A.D.o., PhD student, Moscow City Pedagogical University, Moscow

The article presents the development of a big data analytics system (BDAS) based on artificial intelligence for early identification of foreign students at risk of academic underperformance. The system employs a two-stage iterative approach, combining machine learning and deep learning methods to analyze data on student interactions with a learning management system (LMS). The first stage uses machine learning algorithms (J48, Random Forest, OneR, DecisionStump, NBTree) with an ensemble technique, while the second stage applies deep learning algorithms (LSTM, MLP, Sequential Model) after data balancing. The model accounts for various factors, including language preparation, cultural adaptation, and student academic performance. The developed system enables timely detection of problematic learning situations among foreign students and provides targeted support to enhance the effectiveness of their educational process.

Keywords: big data, machine learning, deep learning, foreign students, academic performance, forecasting, decision support system

For citation: Pogodin I.M., Sadykhbekov A.D.o. Big data analysis for predicting factors of successful adaptation of foreign students. Information and Economic Aspects of Standardization and Technical Regulation. 2025; 4(85): 39–45. (In Russ.).