УДК 004.93 DOI 10.55421/3034-4689_2026_29_1_129

А. Р. Ильин, Р. М. Хусаинов

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ОЦЕНКИ РИСКОВ АКАДЕМИЧЕСКОЙ НЕУСПЕВАЕМОСТИ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

Ключевые слова: машинное обучение, классификационные алгоритмы, предсказательные модели, метки, обучение с подкреплением, анализ влияния признаков, интеллектуальная система, прогнозирование, предсказательные модели.

В статье рассмотрены современные подходы к прогнозированию академических рисков студентов на основе методов машинного обучения. Проведен сравнительный анализ технологий, включая ансамблевые методы (Random Forest, Gradient Boosting, AdaBoost) и другие алгоритмы классификации. В рамках исследования разработана методология оценки ключевых факторов, влияющих на успеваемость, таких как учебная активность (StudyTimeWeekly), посещаемость (Absences), уровень вовлеченности родителей (ParentalSupport) и внеучебная деятельность. Для проведения исследования использован набор данных о 2392 студентах, прошедший комплексную предобработку, включая анализ корреляций, который выявил влияние GPA и пропусков, и стратифицированное разделение на обучающую и тестовую выборки. Реализована сравнительная оценка моделей по метрикам классификации: Accuracy, Precision, Recall и F1-score. По результатам исследования выявлена высокая эффективность ансамблевых алгоритмов, среди которых метод AdaBoost имеет наивысшую результативность с показателями точности 92,48 %, F1-score 92,21 % и ROC-AUC 93,81 %. Анализ матрицы ошибок подтвердил сбалансированность модели с минимальным количеством ложных срабатываний (38) и пропусков риска (32). Оценка важности признаков показала роль GPA (0.689), а также существенное влияние времени самостоятельной подготовки и количества пропусков, что обеспечивает интерпретируемость модели. Предложены дальнейшие пути развития интеллектуальной системы, включая создание интерактивного веб-приложения, расширение датасета, внедрение механизмов адаптивной калибровки и интеграцию в системы управления обучением (LMS) для практического внедрения в образовательный процесс с целью раннего выявления студентов группы риска и оптимизации образовательных траекторий.

A. R. Ilyin, R. M. Khusainov

INTELLIGENT RISK ASSESSMENT SYSTEM FOR ACADEMIC FAILURE BASED ON MACHINE LEARNING

Keywords: machine learning, classification algorithms, predictive models, labels, reinforcement learning, feature impact analysis, intelligent system, forecasting, predictive models.

This article examines modern approaches to predicting students' academic risk using machine learning methods. A comparative analysis of these technologies is conducted, including ensemble methods (Random Forest, Gradient Boosting, AdaBoost) and other classification algorithms. The study developed a methodology for assessing key factors influencing academic performance, such as academic activity (StudyTimeWeekly), attendance (Absences), parental involvement (ParentalSupport), and extracurricular activities. The study utilized a dataset of 2,392 students that underwent comprehensive preprocessing, including correlation analysis to identify the impact of GPA and absences, and stratified separation into training and test sets. A comparative evaluation of the models was implemented using the following classification metrics: Accuracy, Precision, Recall, and F1-score. The study revealed the high efficiency of ensemble algorithms, with the AdaBoost method demonstrating the highest performance with an accuracy of 92.48 %, an F1-score of 92.21 %, and a ROC-AUC of 93.81 %. Confusion matrix analysis confirmed the model's balance, with a minimal number of false positives (38) and high-risk missed errors (32). An assessment of feature importance revealed the role of GPA (0.689), as well as the significant influence of self-study time and the number of missed errors, ensuring the model's interpretability. Further development paths for the intelligent system are proposed, including the creation of an interactive web application, dataset expansion, the implementation of adaptive calibration mechanisms, and integration into learning management systems (LMS) for practical implementation in the educational process to early identify at-risk students and optimize educational trajectories.

Введение

Современная образовательная система сталкивается с необходимостью разработки эффективных инструментов для раннего выявления студентов, находящихся в зоне академического риска [1]. Традиционные методы оценки успеваемости, основанные на ручном анализе академических показателей, имеют недостатки, включая субъективность присвоенных оценок, запаздывание реакции преподавательского состава и невозможность учета множества внешних (внеучебных) факторов, влияющих на учебные результаты. Автоматизация процессов с использованием машинного обучения открывает новые возможности для проектирования и создания интеллектуальных систем прогнозирования академической успеваемости, что позволит увеличить скорость и точность при принятии предупреждающих мер [2]. Актуальность исследования определена возрастающей потребностью образовательных учреждений в решениях прогнозной аналитики, обеспечивающих заблаговременное выявление контингента студентов с рисками академической неуспеваемости и реализацию превентивных мер для оптимизации их образовательных результатов.

Внедрение интеллектуальных систем прогнозирования академических рисков особенно актуально

для высших учебных заведений, где даже незначительные просчеты в выявлении студентов группы риска могут привести к серьезным академическим последствиям и снижению общего качества образования [3]. Машинное обучение предоставляет возможность анализировать многомерные образовательные данные, выявлять скрытые закономерности и формировать точные прогнозы в режиме, близком к реальному времени.

Проблема прогнозирования академической успеваемости исследуется в рамках таких научных направлений, как образовательная аналитика, предиктивное моделирование и интеллектуальный анализ образовательных данных. За последнее десятилетие появилось значительное количество исследований, посвященных применению методов машинного обучения для классификации студентов, прогнозирования отсева и оптимизации образовательных траекторий [4-6].

Несмотря на обилие теоретических разработок, практическая реализация подобных систем в реальных образовательных процессах затруднена из-за качества исходных данных, необходимости обеспечения конфиденциальности образовательной информации и потребности в адаптации алгоритмов под специфику конкретного учебной программы и целевой группы обучающихся.

Цель и задачи исследования

Целью работы является разработка интеллектуальной бета-системы для оценки рисков академической неуспеваемости на основе методов машинного обучения.

Задачи исследования:

1. Провести сравнительный анализ современных алгоритмов машинного обучения, применяемых в образовательной аналитике.

2. Исследовать особенности данных и определить ключевые предикторы (признаков) академической успеваемости.

3. Разработать, протестировать и сравнить комплекс моделей для классификации студентов по уровню академического риска.

4. Спроектировать архитектуру интеллектуальной системы прогнозирования академических рисков.

5. Реализовать прототип системы и провести его валидацию на новом примере данных.

Предметом исследования являются методы и алгоритмы машинного обучения, применяемые для прогнозирования академической успеваемости студентов.

Объектом исследования является размеченный набор данных (датасет), который содержит признаки, характеризующие обучающихся для последующей оценки академических рисков.

Теоретической основой исследования являются научные работы в области машинного обучения, образовательной аналитики и предиктивного моделирования в образовании.

Описание датасета

При проведении исследования использован датасет «Students Performance Dataset» из репозитория Kaggle [7]. Этот датасет специально разработан для задач классификации оценочных показателей на базе факторах, влияющих на успеваемость учащихся, что позволяет провести эффективные образовательные исследования, осуществлять прогнозное моделирование и выполнять статистическую обработку данных в области педагогической науки.

Представленный набор данных является синтезированным (синтетическим), который содержит комплексную информацию о 2392 студентах старших классов, включая демографические характеристики, учебные привычки, вовлеченность родителей, внеучебную активность и академические показатели. Целевая переменная GradeClass категоризирует успеваемость студентов, предоставляя надежную основу для образовательных исследований, прогнозного моделирования и статистического анализа.

Основные характеристики датасета:

- Количество студентов: 2392 учащихся;
- Количество признаков: 14 переменных;
- Целевая переменная: 5 классов успеваемости;
- Размер датасета: ~350 КБ (CSV формат);
- Тип данных: синтетические, сгенерированные для исследовательских целей;
- Временной период: данные за один учебный год;
- Географический охват: условный регион.

Структура данных:

Идентификационные данные: StudentID – уникальный числовой идентификатор студента (от 1001 до 3392).

Демографические характеристики:

- Age – возраст студентов в диапазоне от 15 до 18 лет;
- Gender – пол студента (0 – мужской, 1 – женский).
- Ethnicity – этническая принадлежность:
- 0: Европеоидная раса;
- 1: Афроамериканцы;
- 2: Азиаты;
- 3: Другие.

ParentalEducation – уровень образования родителей:

- 0: Без образования;
- 1: Среднее образование;
- 2: Неоконченное высшее;
- 3: Бакалавриат;
- 4: Магистратура и выше.

Учебная активность:

- StudyTimeWeekly – еженедельное время самостоятельной подготовки (0-20 часов);
- Absences – количество пропущенных занятий за учебный год (0-30 дней);
- Tutoring – занятия с репетитором (0 – нет, 1 – да).

Семейная поддержка (ParentalSupport – уровень вовлеченности родителей в образовательный процесс):

- 0: Отсутствует;
- 1: Низкий;
- 2: Умеренный;
- 3: Высокий;
- 4: Очень высокий.

Внеучебная деятельность:

- Extracurricular – участие во внеклассных мероприятиях (0 - нет, 1 - да);

- Sports – занятия спортом (0 – нет, 1 – да);
- Music – музыкальная деятельность (0 – нет, 1 – да);
- Volunteering – волонтерская активность (0 – нет, 1 – да).

Академические показатели:

- GPA – средний балл успеваемости по шкале от 2.0 до 4.0

Целевая переменная (GradeClass) – классификация успеваемости на основе GPA:

- 0: 'A' (GPA ≥ 3.5) – отличная успеваемость;
- 1: 'B' (3.0 ≤ GPA < 3.5) – хорошая успеваемость;
- 2: 'C' (2.5 ≤ GPA < 3.0) – удовлетворительная успеваемость;
- 3: 'D' (2.0 ≤ GPA < 2.5) – низкая успеваемость;
- 4: 'F' (GPA < 2.0) – неудовлетворительная успеваемость.

Преимущества синтетических данных:

- Сохранение конфиденциальности персональной информации студентов;
- Отсутствие этических ограничений на использование и распространение.

Обработка данных

В рамках исследования использованы актуальные методы анализа данных (EDA, feature engineering), алгоритмы машинного обучения (классификация, ансамблевые методы), а также облачный сервис (Google Colab) для разработки прототипа системы. Для оценки результатов применены метрики точности прогнозирования и показатели классификационной эффективности (уверенности) [8].

Разработанная система может быть внедрена в образовательные процессы дополнительного профессионального образования, интегрирована в рамках высших учебных заведений для автоматизированного мониторинга успеваемости обучающихся, что позволит повысить эффективность учебного процесса, снизить уровень академических задолженностей и оптимизировать ресурсы преподавательского состава [9].

В рамках исследования реализована комплексная методология предобработки, включающая последовательные этапы обработки образовательных данных [10].

1. Загрузка и первичный анализ данных:

- импорт данных из CSV-файла, содержащего информацию о 2392 студентах;
- проведение эксплораторного анализа данных (Exploratory Data Analysis - EDA);
- анализ метаинформации: размерность датасета (2392 × 14), типы данных, статистические характеристики;
- визуализация распределения целевой переменной GradeClass, представленная на рисунке 1, для анализа сбалансированности классов [11].

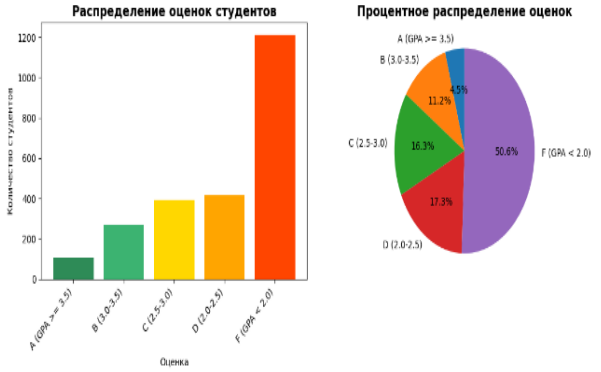

Рис. 1 – Распределение оценок студентов

Fig. 1 – Distribution of student grades

Анализ распределения целевой переменной (GradeClass), категоризирующей академическую успеваемость студентов, выявил выраженную негативную асимметрию в данных. Наиболее значимым наблюдением является доминирование категории «F», соответствующей низкому уровню успеваемости (GPA < 2.0). Представленная категория охватывает 1211 студента, что составляет 50.6 % от всей выборки. Таким образом, каждый второй студент в датасете демонстрирует неудовлетворительные академические результаты.

Распределение студентов по остальным категориям является более сбалансированным, однако также смещено в сторону нижней части шкалы. Категория «D» (2.0 ≤ GPA < 2.5) включает 414 человек (17.3 %), а категория «C» (2.5 ≤ GPA < 3.0) — 391 человека (16.3 %). В совокупности на эти две средние категории приходится 33.6 % студентов.

Группы с высокой успеваемостью наименее многочисленны, так категория «B» (3.0 ≤ GPA < 3.5) объединяет 269 студентов (11.2 %), а высшая категория «A» (GPA ≥ 3.5) — 107 человек, что соответствует всего 4.5% от общей численности выборки.

Распределение академических достижений является резко несбалансированным. Подавляющая половина выборки (50.6%) сконцентрирована в зоне неудовлетворительной успеваемости (GradeClass 'F').

Наблюдается дефицит высоких академических результатов. Доля студентов с оценками «B» и «A» в сумме не превышает 15.7 %, что указывает на существование системных факторов, ограничивающих достижение высоких показателей GPA.

2. Анализ корреляционных зависимостей:

- Построение матрицы корреляций Пирсона для количественной оценки взаимосвязей между признаками;
- Идентификация наиболее значимых предикторов академической успеваемости;
- Выявление сильной отрицательной корреляции между GPA и GradeClass (-0.783);
- Обнаружение умеренной положительной корреляции Absences с целевой переменной: 0.729, представленной на рисунке 2.

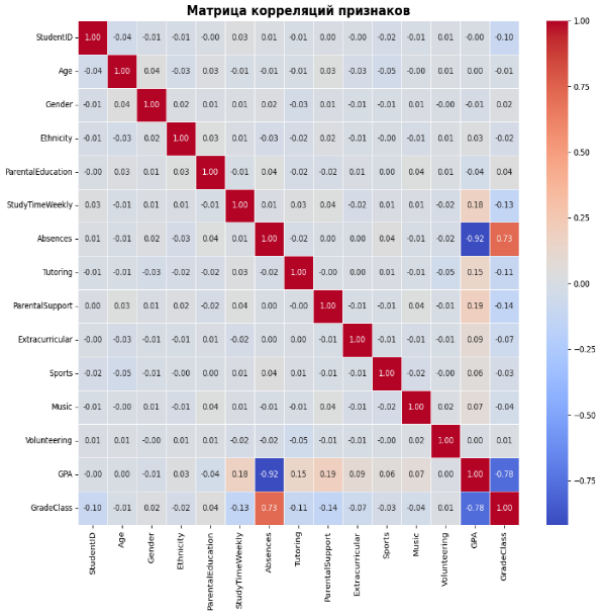

Рис. 2 – Матрица корреляций признаков

Fig. 2 – Feature correlation matrix

Анализ матрицы корреляций, позволяет идентифицировать силу и направление линейных взаимосвязей между анализируемыми признаками и целевой переменной. Результаты демонстрируют резкую дифференциацию прогностической силы переменных.

Наиболее значимым наблюдением является наличие двух признаков с исключительно сильной положительной корреляцией с академической успеваемостью: GPA (коэффициент корреляции r = 0.783) и Absences (r = 0.729). Данные значения указывают на практически детерминированную прямую связь, где текущий совокупный средний балл является прямым предиктором итоговой категории успеваемости, а количество пропусков занятий — одним из ключевых негативных факторов.

Группа признаков со слабой, но статистически значимой положительной корреляцией включает ParentalSupport (r = 0.137) и StudyTimeWeekly (r = 0.134). Это позволяет сделать вывод о том, что поддержка родителей и время, уделяемое самостоятельным занятиям, оказывают умеренное положительное влияние на образовательный результат.

Остальные переменные, такие как Tutoring (r = 0.112), Extracurricular (r = 0.070), ParentalEducation (r = 0.041), а также демографические и социальные характеристики (Ethnicity, Gender, Age), демонстрируют крайне слабую корреляцию (r < 0.1), что позволяет предположить их несущественное прямое влияние на целевую переменную в рамках линейной модели.

3. Исключение нефункциональных признаков:

- Удаление атрибута StudentID как идентификационной переменной, не несущей прогностической ценности;
- Сохранение только релевантных признаков, влияющих на академические результаты;

4. Сегментация данных на обучающую и тестовую выборки:

- разделение на обучающую (80 % данных: 15,870 окон из 19,837) и тестовую (20 % данных) выборки;
- стратификация: применяется для сохранения пропорций классов в обеих выборках;
- фиксация random_state = 42 для обеспечения воспроизводимости результатов;
- формирование обучающей выборки размером 1913 наблюдений и тестовой - 479 наблюдений.

5. Нормализация числовых признаков:

- применение StandardScaler для стандартизации числовых признаков;
- трансформация данных к распределению с нулевым математическим ожиданием и единичной дисперсией;
- раздельное масштабирование для различных категорий моделей:

Масштабированные данные: Logistic Regression, SVM, K-Nearest Neighbors;

- немасштабированные данные: Decision Tree, Random Forest, Gradient Boosting, AdaBoost, Naive Bayes;
- обучение scaler исключительно на тренировочных данных для предотвращения data Leakage [12].

6. Подготовка целевой переменной:

- использование категориального представления GradeClass с пятью классами успеваемости;
- семантическое кодирование: {0: 'A', 1: 'B', 2: 'C', 3: 'D', 4: 'F'};
- сохранение соответствий для интерпретации результатов классификации.

7. Валидация качества предобработки:

- контроль размерностей полученных выборок;
- проверка сохранения стратификации при разделении;
- визуализация распределения ключевых признаков до и после масштабирования;
- анализ выбросов и аномальных значений в числовых признаках.

8. Особенности обработки образовательных данных:

- учет специфики синтетического датасета с гарантированной полнотой и отсутствием пропусков;
- сохранение интерпретируемости признаков для педагогического анализа;
- оптимизация обработки для работы с разнотипными данными: демографические, академические, поведенческие.

Методологическое обоснование выбранных подходов

Стратифицированное разделение обеспечивает репрезентативность выборок для многоклассовой задачи с неравномерным распределением. Стандартизация признаков важна для алгоритмов, основанных на градиентных методах и метриках расстояния. Исключение идентификаторов предотвращает переобучение моделей на артефактных закономерностях [13].

Разработанный конвейер предобработки обеспечивает воспроизводимость результатов, соответствие требованиям различных алгоритмов машинного обучения и сохранение семантической интерпретируемости образовательных данных для последующего

анализа факторов академической успеваемости. В таблице 1 представлен сравнительный анализ моделей машинного обучения.

На основании комплексного сравнительного анализа моделей машинного обучения выявлены их принципиальные различия в функциональности, применимости и производительности. Установлено, что ансамблевые методы, в особенности бустинговые алгоритмы, имеют наивысшую прогностическую эффективность для задач образовательной аналитики, что подтверждается их способностью к последовательному улучшению предсказаний и устойчивостью к переобучению. Модель AdaBoost, имеющая максимальную точность и адаптивность за счет механизма коррекции ошибок классификаторов, выбрана в качестве базовой для практической реализации. В таблице 2 представлены результаты расчета метрик классификации моделей.

Таблица 1 – Сравнительный анализ моделей

Table 1 – Comparative analysis of models

Модель машинного обучения	Применение в образовании	Преимущества	Недостатки
AdaBoost	Прогнозирование академических рисков, классификация студентов по уровню успеваемости.	Наивысшая точность (92.48%), устойчивость к переобучению за счет комбинирования множества слабых learners, адаптивное увеличение весов ошибочно классифицированных объектов.	Чувствительность к шуму в данных, требовательность к вычислительным ресурсам.
Cradint Boosting	Раннее выявление студентов группы риска, оптимизация образовательных траекторий.	Высокая точность (91.86%), хорошая обобщающая способность.	Сложность настройки гиперпараметров, длительное время обучения.
Random Forest	Классификация успеваемости, анализ важности факторов влияния.	Устойчивость к переобучению за счет агрегирования предсказаний множества деревьев.	Требовательность к памяти, менее точный чем бустинговые методы.
Decision Tree	Визуализация факторов успеваемости, быстрый прототипинг.	Простота интерпретации, Способность выявлять нелинейные зависимости.	Склонность к переобучению, неустойчивость к малым изменениям данных.
Logistic Regression	Бинарная и много-классовая классификация успеваемости.	Высокая интерпретируемость, быстрая работа с большими данными.	Предполагает линейную зависимость, требует масштабирования признаков.
SVM	Классификация студентов по академическим показателям.	Эффективность в высокоразмерных пространствах.	Чувствительность к выбору ядра, требовательность к вычислительным ресурсам.
KNN	Классификация по аналогии с похожими студентами	Простота реализации, не требует обучения	Чувствительность к масштабу данных, низкая эффективность с большими выбор-ками

Таблица 2 – Результаты эффективности моделей машинного обучения

Table 2 – Results of machine learning model effectiveness

Модель машинного обучения	Accuracy	Precision	Recall	F1-Score
AdaBoost	0.9248	0.9262	0.9248	0.9221
Cradint Boosting	0.9186	0.9182	0.9186	0.9152
Random Forest	0.9123	0.9152	0.9123	0.9048
Decision Tree	0.8685	0.8680	0.8685	0.8652
Logistic Regression	0.8017	0.7919	0.8017	0.7918
SVM	0.7954	0.7981	0.7954	0.7880
Naive Bayes	0.7537	0.7857	0.7537	0.7488
KNN	0.6451	0.6333	0.6451	0.6345

Проведенный сравнительный анализ восьми алгоритмов машинного обучения для задачи прогнозирования академической успеваемости позволяет сформулировать следующие выводы:

1. Доминирование ансамблевых методов.

Ансамблевые алгоритмы продемонстрировали наивысшую эффективность, заняв три первых позиции в рейтинге. AdaBoost показал максимальную точность (92.48 %) и сбалансированные метрики Precision (92.62 %) и Recall (92.48 %), Gradient Boosting и Random Forest показали сопоставимые результаты с точностью 91.86 % и 91.23 % соответственно.

Преимущество ансамблевых методов обусловлено их способностью комбинировать слабые классификаторы и снижать variance ошибки

2. Высокая устойчивость бустинговых алгоритмов.

Модели AdaBoost и Gradient Boosting характеризуются минимальным стандартным отклонением при кросс-валидации (0.0059 и 0.0037), что свидетельствует об их стабильности и надежности на различных подвыборках данных.

3. Эффективность нелинейных моделей.

Деревья решений и их ансамбли существенно превзошли линейные методы. Разрыв между Decision Tree (86.85 %) и Logistic Regression (80.17 %) составляет 6.68 процентных пункта. Это указывает на наличие сложных нелинейных зависимостей в образовательных данных

4. Превосходство Random Forest в метрике ROC-AUC

Несмотря на третье место по точности, Random Forest показал наивысшее значение ROC-AUC (94.16 %), что демонстрирует его exceptional ability в различении классов и устойчивость к дисбалансу

5. Ограничения дистанционных методов.

K-Nearest Neighbors показал наихудший результат (64.51 %), что объясняется:

- высокой размерностью пространства признаков;
- наличием зашумленных данных;
- неоптимальностью евклидовой метрики для данной предметной области.

6. Конкурентность вероятностных подходов.

Naive Bayes, несмотря на относительно низкую точность (75.37 %), показал достойные значения ROC-AUC (92.74 %), что делает его перспективным для задач с требованием к скорости работы.

7. Сбалансированность метрик у лучших моделей.

Три алгоритма демонстрируют сходные значения Accuracy, Precision, Recall и F1-Score, что указывает на отсутствие существенного дисбаланса между типами ошибок и устойчивость классификации.

8. Практические рекомендации.

Для внедрения в образовательные системы рекомендуется:

- AdaBoost - как эталонная модель с максимальной точностью и стабильностью;
- Random Forest - для задач, требующих высокой интерпретируемости и анализа важности признаков;
- Gradient Boosting - при необходимости баланса между точностью и скоростью работы.

Полученные результаты подтверждают эффективность применения современных методов машинного обучения для прогнозирования академической успеваемости и обосновывают выбор ансамблевых алгоритмов в качестве базовых для построения интеллектуальных образовательных систем.

Анализ модели AdaBoost

На рисунке 3 представлена матрица ошибок модели AdaBoost.

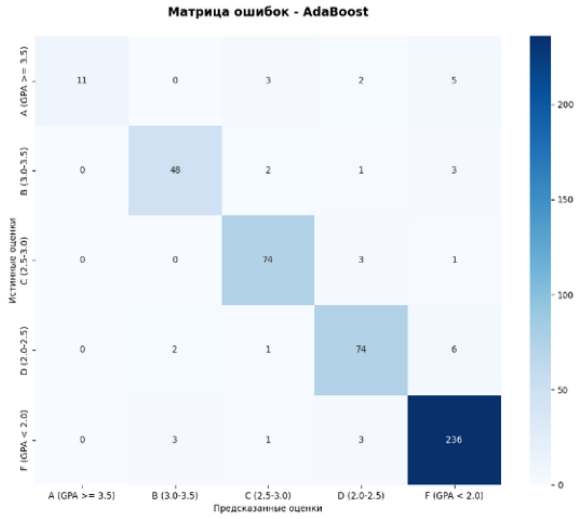

Рис. 3 – Матрица ошибок AdaBoost

Fig. 3 – AdaBoost error matrix

Анализ матрицы ошибок модели AdaBoost подтверждает ее высокую диагностическую эффективность в рамках задачи бинарной классификации студентов на группы академического риска [14]. Модель демонстрирует исключительно низкий уровень ошибок II рода (False Negative), составляющий всего 32 случая, что является важным для системы прогнозирования неуспеваемости, так как позволяет минимизировать количество студентов группы риска, оставшихся без внимания кураторов. При этом количество ошибок I рода (False Positive), когда успешным студентам ошибочно присваивается категория риска, также невелико (38 случаев), что указывает на сбалансированность модели и снижает риск необоснованного административного вмешательства. Данные показатели, в совокупности с ранее зафиксированными значениями Accuracy (0.9248), F1-Score (0.9221) и ROC-AUC (0.9381), убедительно свидетельствуют о том, что модель AdaBoost адекватно отражает дисбаланс классов в данных, обеспечивая распознавание истинных случаев академической неуспеваемости. Следовательно, ее реализация в качестве состава интеллектуальной системы оценки рисков является статистически обоснованной и позволит обеспечить целевое применение превентивных мер поддержки.

Для обеспечения практической ценности интеллектуальной системы оценки рисков критически важен не только высокий показатель точности, но и интерпретируемость модели, позволяющая выделить ключевые факторы, влияющие на академическую

неуспеваемость. Анализ важности признаков, представленный на рисунке 4, позволяет перейти от абстрактного прогноза к конкретным, измеримым индикаторам риска и сформировать обоснованные рекомендации для превентивного вмешательства.

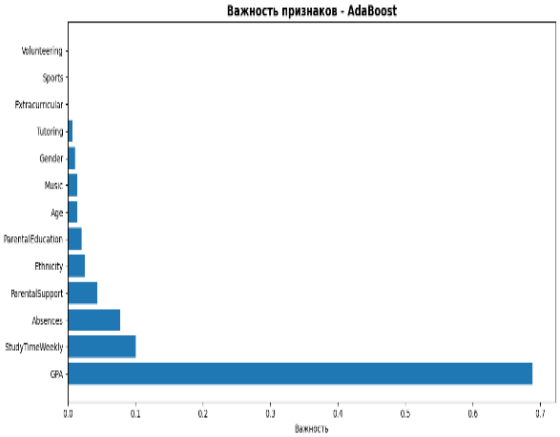

Рис. 4 – Важность признаков модели AdaBoost

Fig. 4 – Importance of AdaBoost model features

По результатам анализа важности признаков в модели AdaBoost выявлены факторы, влияющие на прогнозирование академических рисков, что придает разработанной системе свойство интерпретируемости и практической направленности. Абсолютно доминирующим предиктором является совокупный средний балл (GPA), чья важность (0.689) почти на порядок превышает значимость остальных факторов, что подтверждает его роль как ключевого индикатора текущего академического состояния студента. Существенный, но значительно меньший вклад вносят такие управляемые параметры, как время самостоятельной подготовки (StudyTimeWeekly, 0.101) и количество пропусков занятий (Absences, 0.078), что указывает на зоны для потенциального адресного педагогического вмешательства. При этом такие контекстные признаки, как уровень образования родителей (ParentalEducation, 0.020) или пол (Gender, 0.010), оказывают статистически незначимое влияние на прогноз, что позволяет сфокусировать ресурсы образовательного учреждения на объективных и динамических показателях успеваемости. Таким образом, система не только выполняет точную классификацию, но и предоставляет аналитическую основу для разработки конкретных мер поддержки, направленных на улучшение академических результатов и снижение риска отчисления.

В результате проведенного исследования разработана и протестирована интеллектуальная система оценки рисков академической неуспеваемости на основе машинного обучения. В качестве базового алгоритма выбран ансамблевый метод AdaBoost, продемонстрировавший наивысшую эффективность с показателями точности 92,48 %, F1-score 92,21 % и ROC-AUC 93,81 %, что подтверждает его надежность для решения задач образовательной аналитики.

По анализу матрицы ошибок выявлена сбалансированность модели с минимальным количеством ложноположительных (38 случаев) и ложноотрицательных срабатываний (32 случая), что особенно важно для минимизации рисков как необоснованного вмешательства, так и пропуска реальных случаев академической неуспеваемости.

Корреляционный анализ и оценка важности признаков позволили идентифицировать ключевые детерминанты академических рисков [15]. Установлено, что совокупный средний балл (GPA) является абсолютно доминирующим предиктором с корреляцией -0.783 и важностью 0.689, что подтверждает его роль как основного индикатора академического состояния. Существенное, хотя и менее выраженное влияние оказывают время самостоятельной подготовки (StudyTimeWeekly) и количество пропусков занятий (Absences), что указывает на потенциальные направления для педагогического вмешательства.

Полученные результаты свидетельствуют о практической применимости разработанной системы для поддержки принятия решений в образовательных учреждениях [16]. Система обеспечивает не только точное прогнозирование академических рисков, но и содержательную интерпретацию факторов, влияющих на успеваемость, что позволяет разрабатывать целевые превентивные меры и оптимизировать распределение ресурсов для поддержки студентов группы риска.

Заключение

Проведенное исследование позволило разработать и протестировать интеллектуальную систему оценки академических рисков, основанную на методах машинного обучения. Анализ современных подходов к прогнозированию успеваемости показал эффективность применения ансамблевых алгоритмов, среди которых метод AdaBoost продемонстрировал наивысшую результативность с показателями точности 92,48 %, F1-score 92,21 % и ROC-AUC 93,81 %.

Преимуществом разработанной системы является ее сбалансированность, подтвержденная анализом матрицы ошибок - количество ложноположительных (38) и ложноотрицательных (32) срабатываний минимизировано, что обеспечивает эффективное распределение ресурсов образовательного учреждения.

Перспективные направления развития системы включают:

- создание интерактивного веб-приложения на платформе Streamlit;
- расширение базы данных до 5000-10000 наблюдений с диверсификацией выборки;
- внедрение механизмов адаптивной калибровки моделей;
- интеграцию с системами управления обучением (LMS);
- разработку автоматизированной системы формирования индивидуальных образовательных траекторий.

Практическая реализация системы возможна в различных областях образовательного процесса:

- поддержка принятия решений при разработке учебных планов;
- раннее выявление студентов, нуждающихся в педагогической поддержке;

- оптимизация распределения ресурсов преподавательского состава;

- профориентационная поддержка на основе анализа академических предпочтений.

Полученные результаты подтверждают практическую значимость разработанного решения и его готовность к внедрению в образовательных учреждениях для повышения эффективности управления академическими рисками.

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. Попова Н.А., Егорова Е.С., *Известия Кабардино-Балкарского научного центра РАН*, 2 (112), 18-29 (2023).
4. Львович Я.Е., Преображенский Ю.П., Ружицкий Е., *Вестник Воронежского института высоких технологий*, 2 (41), 96-99 (2022).
5. Большаков Н.И., Сидорова Е.В., *Математические методы в технологиях и технике*, 8, 66-71 (2023).
6. Милованович Н.Г., Басс Н.В., *Тенденции развития науки и образования*, 107-4, 97-100 (2024).
7. Keras: Deep Learning for humans — GitHub. URL: https://github.com/kerasteam/keras/ - (дата обращения 03.11.2025).
8. Основы генеративно-состязательных сетей. URL: https://habr.com/ru/articles/726254/ (дата обращения 06.11.2025).
9. Nayak P., *Education and Information Technologies*, **28**, 11, 14611-14637 (2023).
10. Якунин Ю.Ю, *Информатика и образование*, 38, 4, 28-43 (2023).
11. Nabil A., Seyam M., Abou-Elfetouh A., *IEEE Access*, 9, 140731-140746 (2021).
12. Asselman A., Khaldi M., Aammou S., *Interactive Learning Environments*, **31**, 6, 3360-3379 (2023).
13. Pallathadka H., *Materials today: proceedings*, **80**, 3782-3785 (2023).
14. Adnan M., *Ieee Access*, **9**, 7519-7539 (2021).
15. Рекуррентные нейронные сети и LSTM. URL: https://nweb42.com/books/r-lang/rekurrentnye-neyronnye-seti-i-lstm/ - (дата обращения 02.11.2025).
16. Бражникова С.С., *Психолого-педагогические исследования-Тульскому региону*, 335-341 (2023).

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. Intellect. Initiative*, 18-19 (2023).
3. Popova N.A., Egorova E.S., *News of the Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences*, 2 (112), 18-29 (2023).
4. Lvovich Ya.E., Preobrazhensky Yu.P., Ruzhitsky E., *Bulletin of the Voronezh Institute of High Technologies*, 2 (41), 96-99 (2022).
5. Bolshakov N.I., Sidorova E.V., *Mathematical Methods in Technology and Engineering*, 8, 66-71 (2023).
6. Milovanovich N.G., Bass N.V., *Trends in the Development of Science and Education*, 107-4, 97-100 (2024).
7. Keras: Deep Learning for humans — GitHub. URL: https://github.com/kerasteam/keras/ - (accessed 03.11.2025).
8. Fundamentals of generative adversarial networks. URL: https://habr.com/ru/articles/726254/ (accessed 06.11.2025).
9. Nayak P., *Education and Information Technologies*, **28**, 11, 14611-14637 (2023).
10. Yakunin Yu.Yu., *Informatics and Education*, 38, 4, 28-43 (2023).
11. Nabil A., Seyam M., Abou-Elfetouh A., *IEEE Access*, 9, 140731-140746 (2021).
12. Asselman A., Khaldi M., Aammou S., *Interactive Learning Environments*, **31**, 6, 3360-3379 (2023).
13. Pallathadka H., *Materials today: proceedings*, **80**, 3782-3785 (2023).
14. Adnan M., *Ieee Access*, **9**, 7519-7539 (2021).
15. Recurrent neural networks and LSTM. URL: https://nweb42.com/books/r-lang/rekurrentnye-neyronnye-seti-i-lstm/- (accessed on 02.11.2025).
16. Brazhnikova S.S., *Psychological and Pedagogical Research in the Tula Region*, 335-341 (2023).

© **А. Р. Ильин** – магистрант кафедры Систем информационной безопасности (СИБ), Казанский национальный исследовательский технический университет им. А.Н. Туполева (КНИТУ им. А.Н. Туполева), Казань, Россия, ilyin.alexander.of@bk.ru; **Р. М. Хусаинов** – ассистент кафедры СИБ, КНИТУ им. А.Н. Туполева, rumil_husainov98@mail.ru.

© **A. R. Ilyin** – Master-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, ilyin.alexander.of@bk.ru; **R. M. Khusainov** – Assistant of the ISS department, KNRTU named after A.N. Tupolev, rumil_husainov98@mail.ru.

Дата поступления рукописи в редакцию – 07.11.25.
Дата принятия рукописи в печать – 18.01.26.