ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

Отчет о программном проекте на тему:

Применение методов машинного обучения для предсказания результатов экзаменов

Выполнил студент:
группы БПМИИ229, 3 курса
Аглиев Камиль Марселевич

Принял руководитель проекта:
Тараканов Александр Александрович
Доцент
Факультета компьютерных наук НИУ ВШЭ

Москва
2025

Содержание

Аннотация

Настоящее исследование посвящено разработке и применению методов машинного обучения для прогнозирования академической успеваемости студентов Национального исследовательского университета «Высшая школа экономики». Основной целью работы является создание эффективного инструмента для идентификации студентов, относящихся к группе риска по показателю вероятности неуспешной сдачи экзаменационных испытаний.

В рамках исследования были реализованы и сравнительно проанализированы различные алгоритмы машинного обучения, включая логистическую регрессию и градиентный бустинг. Особое внимание уделено вопросам предобработки образовательных данных, включая обработку пропущенных значений, дублей и отбор признаков.

Результатом работы стала прогностическая модель, демонстрирующая высокую точность (F1 = 0.798) в задаче бинарной классификации (сдача/несдача экзамена).

Практическая значимость исследования заключается в возможности интеграции разработанной модели в существующие образовательные платформы университета для автоматического мониторинга успеваемости.

Ключевые слова:

прогнозирование успеваемости, машинное обучение, образовательная аналитика, логистическая регрессия, градиентный бустинг, предиктивная аналитика.

1 Введение

1.1 Постановка задачи

В рамках образовательного процесса Национального исследовательского университета «Высшая школа экономики» все студенты обязаны пройти три экзаменационных испытания по цифровым компетенциям:

- Цифровая грамотность
- Программирование
- Анализ данных

Подготовка к данным экзаменам осуществляется посредством прохождения специализированных онлайн-курсов в системе Smart LMS. Однако эмпирические наблюдения свидетельствуют о недостаточной эффективности текущей системы подготовки.

Основными факторами, способствующими снижению результативности, являются:

- Отсутствие системы промежуточного контроля знаний, интегрированной в итоговую оценку
- Недостаточный уровень педагогического сопровождения учебного процесса
- Неадекватная оценка студентами сложности экзаменационных требований

В связи с обозначенными проблемами, администрацией университета была инициирована разработка предиктивной аналитической системы на основе методов машинного обучения. Целью данного проекта является создание модели, способной:

- Прогнозировать вероятность успешной сдачи экзаменов
- Идентифицировать студентов группы академического риска
- Формировать индивидуальные траектории подготовки

Научная новизна исследования заключается в разработке комплексного подхода, объединяющего:

- Анализ академической успеваемости
- Оценку цифрового следа в LMS
- Прогностическое моделирование на основе ансамбля алгоритмов

Практическая значимость работы определяется возможностью интеграции разработанного решения в образовательный процесс НИУ ВШЭ для:

- Своевременного выявления студентов, нуждающихся в дополнительной поддержке

- Оптимизации распределения педагогических ресурсов
- Повышения общей успеваемости по цифровым компетенциям

В перспективе разработанная система может быть расширена за счет:

- Интеграции с другими образовательными платформами
- Реализации механизма автоматизированных рекомендаций
- Разработки мобильного интерфейса для студентов

1.2 Анализ существующих решений

Современные исследования подтверждают эффективность методов машинного обучения для прогнозирования успеваемости студентов. Анализ литературных источников позволяет выделить ключевые закономерности:

- Основные используемые данные:
  - Академическая история (оценки, посещаемость)
  - Активность в электронных образовательных системах
  - Демографические характеристики
- Наиболее распространенные методы:
  - Градиентный бустинг (XGBoost, CatBoost)
  - Логистическая регрессия
  - Методы ансамблирования
- Типичные результаты:
  - Точность прогнозирования: 75-90%
  - Средние показатели AUC-ROC: 0.70-0.85

Основные ограничения существующих подходов связаны с особенностями образовательных данных (неполнота, изменчивость) и необходимостью баланса между точностью и интерпретируемостью моделей. Тем не менее, результаты исследований демонстрируют принципиальную возможность построения эффективных прогностических систем.

1.3 Обзор литературы

1.3.1 Статья №1

В статье "A Machine Learning Approach to Predictive Modelling of Student Performance" [6] разрабатывалась система прогнозирования академической успеваемости на основе данных 1044 португальских школьников. Основные результаты и выводы:

- Методология исследования:
  - Использованы 33 признака (социальные, поведенческие и академические)
  - Применены три алгоритма: SVM, Naive Bayes и MLP
  - Две постановки задачи: бинарная и многоклассовая классификация
- Ключевые результаты:
  - Наивысшая точность (91%) достигнута методом SVM для бинарной классификации
  - Для многоклассового прогноза лучший результат - 73% (SVM)
  - Наименее эффективным оказался алгоритм Naive Bayes (78% и 65% соответственно)
- Основные выводы:
  - Академическая история - наиболее значимый предиктор успеваемости
  - Социально-демографические факторы снижают точность прогноза
  - SVM демонстрирует лучшую эффективность для задач образовательного прогнозирования

Полученные результаты имеют важное значение для нашего исследования, подтверждая:

- Возможность построения точных прогностических моделей
- Ключевую роль академических показателей
- Эффективность SVM-подхода в образовательной аналитике

Ограничения исследования:

- Относительно небольшая выборка (1044 наблюдения)
- Специфика португальской образовательной системы
- Отсутствие данных о цифровой активности учащихся

1.3.2 Статья №2

В статье "A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques"[7] авторы провели систематический обзор литературы, связанной с предсказанием успеваемости студентов с использованием методов машинного обучения. В работе рассматривалось предсказание отчислений и выявление студентов, находящихся в группе риска. Авторы рассмотрели множество статей за 2009-2021 гг и выявили наиболее часто используемые для предсказания успеваемости студентов модели машинного обучения:

- Decision Tree
- Support Vector Machine (SVM)
- Naive Bayes
- Random Forest
- Neural Networks (NN)

Среди важных признаков в точности предсказания отчислений авторы выделяют историю оценок, посещаемость и уровень вовлеченности, а наиболее успешными моделями в предсказании отчислений стали Random Forest и Gradient Boosting.

В предсказании студентов, находящихся в группах риска, зачастую использовались данные университетов (оценки, посещаемость, рейтинги) и данные онлайн платформ (частота взаимодействия с онлайн платформой, количество просмотренных материалов и пройденных тестов). Самыми успешными моделями в этой области стали Gradient Boosting и Support Vector Machine.

1.3.3 Статья №3

В статье "Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review"[8] проведен комплексный анализ 346 научных работ (2010-2022 гг.) по прогнозированию успеваемости с помощью ML. Авторы применили PRISMA-методологию для отбора и классификации исследований, выделив ключевые тенденции, методы и проблемы в этой области.

Все работы были разделены на две основные категории:

- Прогнозирование успеваемости студентов на экзаменах, курсах или программах, а также выявление студентов, находящихся в группе риска по неуспеваемости.
- Прогнозирование отчисления или удержания студентов на курсе или в программе

Эффективность методов:

- Наивысший средний accuracy (86.4%) показали ансамбли Gradient Boosting + временные признаки

- Нейросети требуют в 3-5 раз больше данных для сопоставимой точности
- Простые модели (логистическая регрессия) часто превосходят сложные при малых выборках (<1,000 студентов)

Критические факторы успеха:

- Временные паттерны (регулярность подготовки) увеличивают точность на 17-23%
- Комбинация LMS-данных + академической истории дает лучшие результаты
- Добавление психометрических тестов улучшает прогноз лишь на 4-6%

2 Сбор данных

2.1 Структура данных

Исследование основано на данных из информационных систем НИУ ВШЭ. Исходные данные представлены в 27 взаимосвязанных таблицах форматов CSV, ODS и XLSX, охватывающих период с 2019 по 2024 год. Основные категории информации включают:

- Общие сведения
  - Образовательная программа
  - Год поступления
  - Форма обучения
  - Кампус
- Академические данные
  - История оценок по всем дисциплинам
  - Результаты промежуточных аттестаций
  - Рейтинги успеваемости
- Данные цифрового следа
  - Статистика выполнения онлайн-курсов, тестов, квизов
  - просмотры лекций, семинаров, видео по предмету

2.2 Методы сбора и текущие ограничения

На текущем этапе исследования используется статичный снимок данных, который является выгрузкой всех таблиц с данными за определенный период. В будущем предполагается использование API для постоянного обновление таблиц и получения самых свежих данных, но в данный момент работаем с тем, что есть.

3 Обработка данных

3.1 Импорт и чтение исходных данных

Работа проводилась в среде Jupyter Notebook с использованием библиотеки Pandas. Входные данные представляли собой множество таблиц в различных форматах и кодировках, что потребовало индивидуального подхода к чтению каждого файла.

3.2 Обработка дубликатов и агрегация данных

В качестве уникального идентификатора студента использовалась корпоративная почта, она всегда будет разной внутри домена ВШЭ. Для таблиц с дублирующимися почтами применялась логика группировки и агрегации значений, к примеру из таблицы оценок по одному студенту можно брать среднее по всем дисциплинам, сколько раз он сдал экзамен, сколько раз не сдал.

3.3 Выделение данных о независимом экзамене

Так как модель должна четко понимать, что влияет на результат студента на независимом экзамене, то данных о нем должно быть соответствующе много, поэтому во всех таблицах с оценками в первую очередь было просмотрено нет ли данных о независимом экзамене, если они были, то они выделялись отдельно от остальных, так как имеют особую важность. К примеру в таблице 'Выгрузка оценок общая 3 и 4 курсы.xlsx' нужно было отдельно прописать фильтр на РАЗДЕЛ = 'Data Culture', чтобы получить желаемые оценки.

3.4 Объединение таблиц

После обработки дубликатов по почте основную таблицу с оценками по независимому экзамену нужно объединить со всеми остальными таблицами, так получилась таблица с 1500+ столбцами. Далее фильтруются столбцы и остаются только более менее полезные. Фильтр смотрел, что в названии столбца есть 'тест' или подобные слова. Набор этих слов увеличивался итеративно, смотря на слова из столбцов, которые были отфильтрованы.

3.5 Числовое представние данных

В процессе подготовки данных для обучения моделей машинного обучения было проведено комплексное преобразование всех признаков в числовой формат. Основная сложность заключалась в разнородности исходных данных - где-то были текстовые значения ("сдал"/"не сдал"), где-то дроби ("10/100"), а где-то категориальные описания. Для унификации была разработана специальная функция-конвертер, которая пыталась привести любое значение к числовому виду. Алгоритм ее работы был итеративным: при возникновении ошибки функция сообщала, с каким именно форматом данных она не смогла

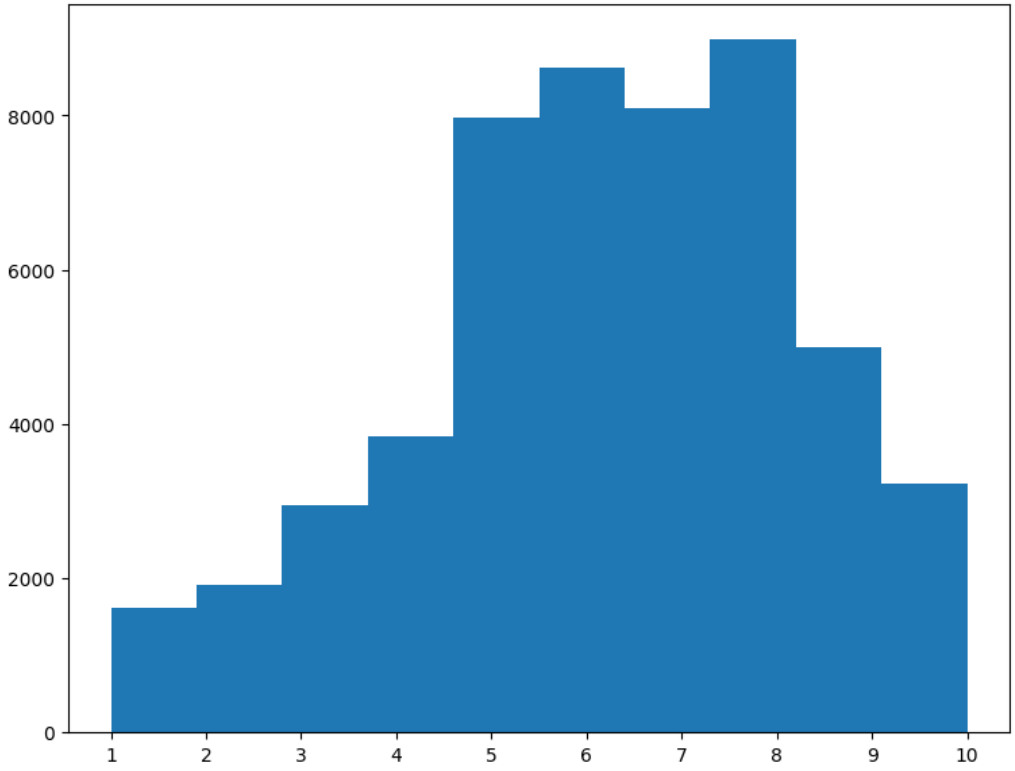
справиться, и после последовательно дописывалась обработка новых случаев.

4 Формирование признаков

После обработки и фильтрации сырый данных нужно собрать интерпретируемые признаки, с которыми уже будет работать модель, поэтому важно отбирать реально значимые и полезные признаки. Отобранные признаки можно разделить по группам:

4.1 Целевая переменная

В качестве целевой переменной я взял последнюю оценку студента по соответствующему предмету независимого экзамена



Распределение целевой переменной ожидаемое, больше смещено в сторону 5+. Имеем 6480 элемента класса 'не сдал' и 45720 элемента класса 'сдал'. Достаточно сильный дисбаланс классов.

4.2 Оценки

Так как их может быть много все оценки агрегируются:

- среднее
- дисперсия
- количество оценок меньше удовлетворительного порога

- доля оценок меньше удовлетворительного порога
- последняя оценка

4.3 Рейтинги

Получилось собрать 36 признаков связанных с каким либо рейтингом. Для каждого рейтинга считались статистики и фильтровались те, что имели маленькую дисперсию или большое количество пропусков. Примеры рейтингов:

- Перцентиль
- Сумма кредитов
- Место в кампусе

4.4 ЕГЭ

В полученных данных есть данные о ЕГЭ по Математике и Информатике. существует гипотеза, что по результатам ЕГЭ можно определить усердность студента и готовность к экзаменационной подготовке.

4.5 База знаний

Все пройденные тесты, видео, лекции фильтровались так, что оставлялись те, у которых был хотя бы 1% заполненности, иначе считалось, что это будет лишь привносить шум в данные. Теперь когда мы имеем для каждого студента длинный вектор из 448 числа можно применить SVD разложение и получить вектор меньшей размерности и уже его использовать как описание знаний студента. Было выбрано 30 компонент, что объяснили 0.736 дисперсии данных.

5 Обучение моделей

5.1 Разделение выборки

В данных было очень сложно собрать информацию о дате и в данный момент она отсутствует, поэтому предпологаем, что все знания имеют одно время. С механизмом регулярного обновления данных эта проблема должна рещиться. Стоит отметить, что во время реального применения модели все знания модели будут уже старыми, поэтому дата-лика в будущее не будет, он лишь может быть при замере качества на тестовой выборке. Датасет был поделен на обучение и тест в пропорции 70/30.

5.2 Базовые модели (baseline)

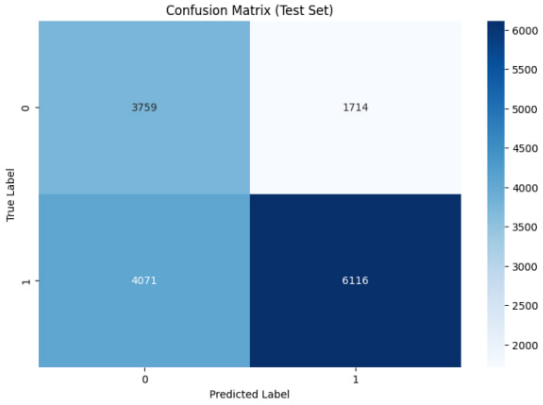
В рамках исследования были реализованы базовые модели машинного обучения, которые служат точкой отсчета для оценки эффективности более сложных алгоритмов. В качестве baseline-решений выбраны:

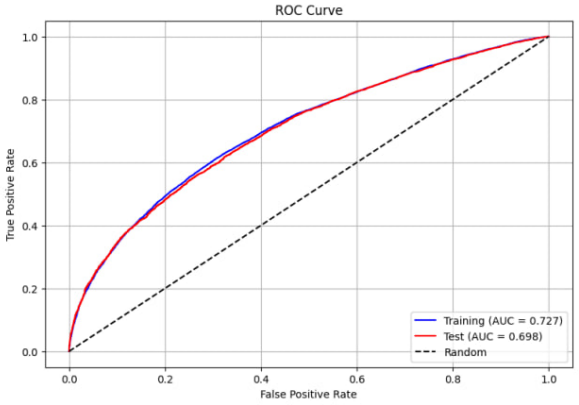
Линейная регрессия для задачи прогнозирования экзаменационной оценки:

- Все признаки предварительно нормализованы с использованием StandardScaler
- Полученное значение RMSE = 2.0 свидетельствует о значительной ошибке предсказания
- Данный результат отражает базовый уровень точности для регрессионной задачи

Логистическая регрессия для бинарной классификации (сдача/несдача экзамена):

- Метрики качества модели:
  - Precision = 0.895
  - Recall = 0.555
  - ROC-AUC = 0.698





- Показатели демонстрируют приемлемый базовый уровень классификации
- Наблюдается дисбаланс между precision и recall, характерный для простых моделей

Полученные результаты позволяют сделать следующие выводы:

- Baseline-модели обеспечивают разумный уровень предсказательной способности
- Значения метрик указывают на принципиальную решаемость поставленных задач
- Существует значительный потенциал для улучшения показателей за счет более сложных алгоритмов

Данные модели будут использоваться в качестве референсных значений при оценке эффективности продвинутых методов машинного обучения.

5.3 Улучшение результата

Для повышения качества предсказаний были исследованы алгоритмы градиентного бустинга, обладающие способностью выявлять сложные нелинейные зависимости в данных.
В работе использована реализация GradientBoostingRegressor для регрессии и GradientBoostingClas
для вероятностей из библиотеки sklearn.ensemble.

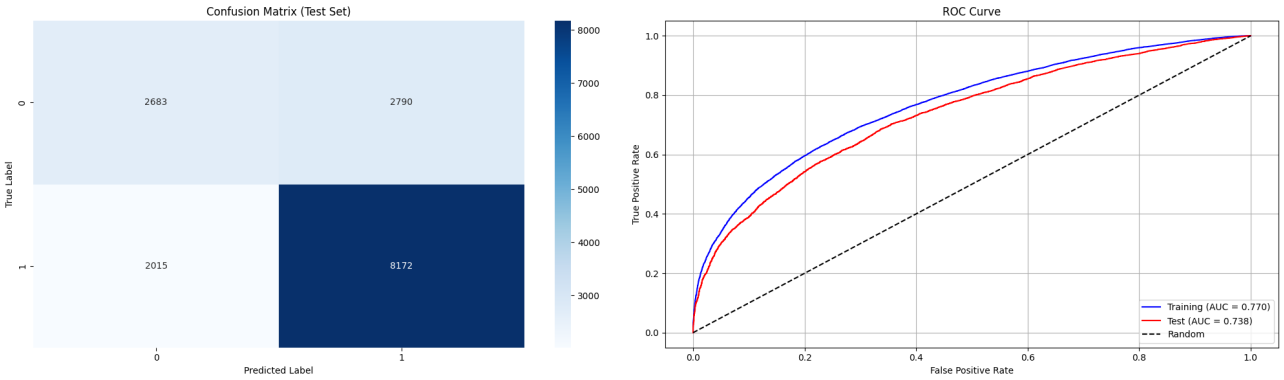
5.3.1 Регрессионная модель

- Целевая метрика: RMSE = 1.86
- Относительное улучшение по сравнению с baseline: 7

5.3.2 Регрессионная модель как классификатор

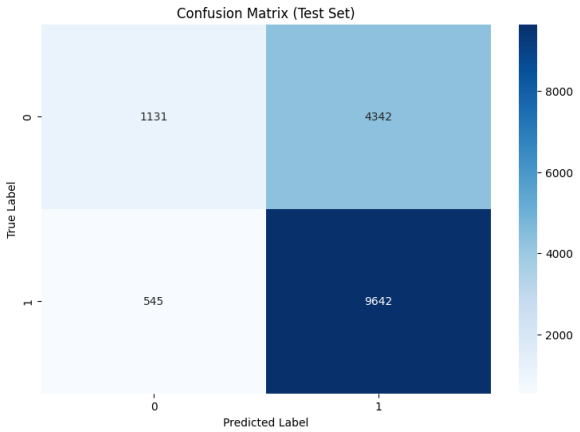
Выходы регрессионной модели можно превратить в классификацию задав порог для определения класса сдаст / не сдаст. Таким образом получил метрики:

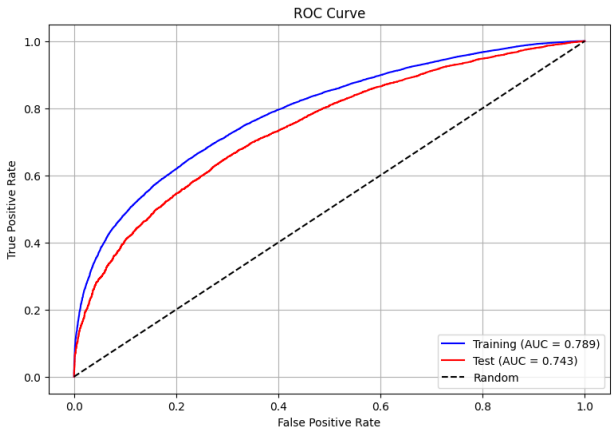
- Precision = 0.746
- Recall = 0.802
- ROC-AUC = 0.738
- Улучшение ROC-AUC на 5.7% относительно baseline



5.3.3 Классификационная модель

- ROC-AUC = 0.743
- Precision = 0.690
- Recall = 0.947
- F1-score = 0.798





5.3.4 Интерпретация результатов

- Высокий recall свидетельствует о способности модели идентифицировать 94.7% студентов группы риска, что для нас является самым главным
- Сравнительно низкий precision указывает на наличие ложноположительных срабатываний
- Значение F1-score демонстрирует сбалансированное качество классификации
- Модель показывает значимое улучшение по метрикам относительно baseline

Полученные результаты подтверждают эффективность градиентного бустинга для решения поставленной задачи. Дальнейшее улучшение модели возможно за счет:

- Оптимизации гиперпараметров (learning rate, глубина деревьев)
- Учета временных зависимостей в данных
- Применения методов балансировки классов
- Увеличение размера обучающей выборки
- Улучшение качества отбора признаков

6 Применение модели

6.1 Архитектура решения

Разработанное решение реализовано в виде модуля Python, содержащего следующие ключевые компоненты:

- **prepare_data** – функция предварительной обработки данных:
  - Автоматическое чтение и консолидация таблиц из различных источников

  - Нормализация и преобразование признаков
  - Обработка пропущенных значений и выбросов
  - Сохранение подготовленных данных в едином формате

- **train_model** – функция обучения модели:
  - Реализация кросс-валидации и настройки гиперпараметров
  - Визуализация кривых обучения и валидации
  - Расчет и сохранение ключевых метрик (Precision, Recall, ROC-AUC)
  - Экспорт обученной модели в файл .pkl

- **apply_model** – функция применения модели:
  - Загрузка сохраненной модели
  - Генерация прогнозов (вероятности и точечные оценки)
  - Форматирование результатов для последующего анализа

- **check_model** – функция валидации модели:
  - Сравнение предсказаний с фактическими результатами
  - Расчет и визуализация матрицы ошибок
  - Генерация отчета о качестве модели

- **get_red_zone** – функция идентификации рисков:
  - Фильтрация студентов с высокой вероятностью неуспешной сдачи
  - Ранжирование по уровню риска

7 Заключение

В данной курсовой работе было реализована модель машинного обучения, которая по данным о студенте и выбранному экзамену будет возвращать вероятность сдать экзамен и прогнозируемую оценку. Сейчас модели сохранены в репозитории на github, все функции их обработки и применения готовы, осталось лишь интегрировать их в сервис для сбора данных и прогона модели.

В перспективе можно развивать следующие направления:

- обучение модели, которая экспертна в любом экзамене
- рекомендательная система курсов, прохождение которых положительно влияет на оценку
- отображение оценку в онлайн формате для повышение мотивации

Список литературы

1. NumPy. NumPy Documentation. URL: https://numpy.org/doc/ (дата обр. 20.03.2025).
2. Pandas. Pandas Documentation. URL: https://pandas.pydata.org/docs/ (дата обр. 20.03.2025).
3. Scikit-learn. Machine Learning in Python. URL: https://scikit-learn.org/stable/ (дата обр. 20.03.2025).
4. Gradient Boosting. Gradient Boosting Explained. URL: https://en.wikipedia.org/wiki/Gradient_boosting (дата обр. 20.03.2025).
5. SVD Decomposition. Singular Value Decomposition. URL: https://en.wikipedia.org/wiki/Singular_value_decomposition (дата обр. 20.03.2025).
6. Hu Ng, Azmin Alias bin Mohd Azha, Timothy Tzen Vun Yap, Vik Tor Goh. A Machine Learning Approach to Predictive Modelling of Student Performance [version 2; peer review: 2 approved]. URL: https://pubmed.ncbi.nlm.nih.gov/35719314/ (дата обр. 20.03.2025).
7. Balqis Albreiki, Nazar Zaki, Hany Alashwal. A Systematic Literature Review of Student's Performance Prediction Using Machine Learning Techniques. URL: https://www.researchgate.net/publication/354661323_A_Systematic_Literature_Review_of_Student'_Performance_Prediction_Using_Machine_Learning_Techniques (дата обр. 20.03.2025).
8. Khalid Alalawi, Rukshan Athauda, Raymond Chiong. Contextualizing the current state of research on the use of machine learning for student performance prediction: A systematic literature review. URL: https://www.researchgate.net/publication/371754145_Contextualithe_current_state_of_research_on_the_use_of_machine_learning_for_student_performaprediction_A_systematic_literature_review (дата обр. 20.03.2025).