МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Ижевский государственный технический университет
имени М.Т. Калашникова»
(ФГБОУ ВО «ИжГТУ имени М.Т. Калашникова)
Факультет «Информационные технологии»
кафедра «Информационные системы»

«ЦИФРОВОЙ СЛЕД СТУДЕНТА В ВЫСШЕМ УЧЕБНОМ ЗАВЕДЕНИИ»

(расчетно-пояснительная записка
к выпускной квалификационной работе бакалавра
по направлению 09.03.03 – «Прикладная информатика»)

Утверждаю
зав. кафедрой «Информационные системы»
д-р физ.-мат. наук, профессор М.М. Горохов

Руководитель работы
доцент кафедры «Информационные системы»
канд. физ.-мат. наук., доцент А.В. Корепанов

Выполнил
студент группы Б20-021-1 З.А. Чекалкин

ИЖЕВСК – 2024

УДК 004.891.2, 004.62

РЕФЕРАТ

Объем записки: 70 стр., 40 рис., 1 табл., 30 источников.

Ключевые слова: Машинное обучение, вероятность отчисления студента, цифровой след, анализ данных.

Данная работа посвящена исследованию и разработке методов прогнозирования вероятности отчисления студентов из высших учебных заведений на основе анализа их цифровых следов. В современном образовательном процессе студенты оставляют значительное количество цифровых данных, таких как оценки, средний балл аттестата, место проживания и другие характеристики. Эти данные могут быть использованы для раннего выявления студентов, находящихся в группе риска, и принятия соответствующих профилактических мер.

Основное внимание в работе уделяется применению методов машинного обучения и искусственного интеллекта для создания предиктивной модели, которая анализирует собранные данные и предсказывает вероятность отчисления студента. Таким образом, цель исследования заключается в повышении качества образования и снижении числа отчислений посредством использования передовых технологий анализа данных.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы объясняется тем, что современные высшие учебные заведения сталкиваются с проблемой значительного числа отчислений студентов, что негативно сказывается на их репутации и финансовом состоянии. В условиях цифровизации образовательного процесса вузы располагают большими объемами данных о своих студентах, которые могут быть использованы для раннего выявления студентов, находящихся в группе риска. Применение методов машинного обучения для анализа этих данных открывает новые возможности для повышения эффективности управления образовательным процессом и снижения числа отчислений.

Объектом исследования являются цифровые следы, оставляемые студентами в ходе их учебной деятельности.

Предметом исследования является использование методов машинного обучения для анализа цифровых следов студентов с целью предсказания вероятности их отчисления из высшего учебного заведения.

Целью данной дипломной работы является разработка и внедрение модели машинного обучения, которая анализирует цифровые следы студентов и предсказывает вероятность их отчисления, что позволит нашему университету своевременно выявлять студентов, находящихся в группе риска, и принимать меры для предотвращения их отчисления.

На защиту дипломной работы выносится:

1. Определение состава задач для работы.
2. Описание данных, используемых в исследовании.
3. Разработка и обучение модели машинного обучения.
4. Разработка программного приложения.
5. Оценка точности и надежности модели.

Практическая ценность данной дипломной работы заключается в нескольких ключевых аспектах:

1. Снижение числа отчислений: внедрение разработанной модели машинного обучения позволит университету своевременно выявлять студентов, находящихся в группе риска, и принимать превентивные меры. Это поможет снизить число отчислений, что позитивно скажется на репутации и финансовом состоянии учебного заведения.
2. Повышение успеваемости студентов: раннее выявление проблемных студентов даст возможность университету предоставить им необходимую поддержку, такую как дополнительные консультации, наставничество или психологическая помощь. Это поможет студентам улучшить свои академические результаты.
3. Оптимизация образовательного процесса: анализ цифровых следов студентов позволит выявить общие проблемные зоны в учебном процессе и принять меры для их устранения. Например, можно выявить сложные для большинства студентов темы или курсы и пересмотреть учебные программы или методы преподавания.
4. Улучшение планирования и управления ресурсами: модель машинного обучения может помочь в планировании ресурсов университета, таких как распределение преподавателей, организация дополнительных занятий или семинаров, что приведет к более эффективному использованию человеческих и материальных ресурсов.
5. Персонализация обучения: использование данных о студентах позволит разрабатывать индивидуальные учебные планы и программы, адаптированные к потребностям и возможностям каждого студента, что повысит качество образования и удовлетворенность студентов.
6. Развитие цифровой инфраструктуры университета: разработка и внедрение программного приложения для анализа данных и предсказания отчислений способствует развитию цифровых

технологий в образовательном процессе, что соответствует современным тенденциям цифровизации образования.

7. Повышение конкурентоспособности университета: университеты, внедряющие современные технологии и подходы к обучению, привлекают больше абитуриентов и повышают свою конкурентоспособность на рынке образовательных услуг.

Таким образом, практическая ценность работы заключается в создании эффективного инструмента для улучшения управления учебным процессом и повышения качества образования, что в конечном итоге способствует развитию высшего учебного заведения и улучшению его репутации. Работа содержит введение, три главы и заключение, изложенные на 70 страницах. В работу включены 40 рисунков, 1 таблица, и список литературы из 30 наименований. К работе приложены плакаты в виде электронной презентации, созданной в MS PowerPoint 2007. Презентация представлена на 10 слайдах и содержит 1 лист – постановка целей и задач для работы, 1 лист – разновидности цифрового следа и его роль в образовательной среде, 1 лист – машинное обучение в образовании, 2 листа – описание методов и средств, и используемых библиотек, 1 лист – этапы реализации проекта , 1 лист – предобработка, анализ и визуализация данных. 1 лист – создание модели машинного обучения, 1 лист – руководство пользователя, 1 лист – заключение.

Оглавление

ВВЕДЕНИЕ

В современных условиях высшее образование играет ключевую роль в развитии общества, обеспечивая подготовку квалифицированных специалистов для различных отраслей экономики. Однако одной из значительных проблем, с которыми сталкиваются высшие учебные заведения, является высокий уровень отчислений студентов. Это негативно влияет на репутацию вузов, их финансовое состояние, а также на будущее студентов, лишающихся возможности завершить образование и получить диплом.

С развитием цифровых технологий и повсеместной автоматизацией образовательных процессов, вузы стали располагать большим объемом данных о своих студентах. Эти данные включают академическую успеваемость, демографические характеристики и другие параметры, которые могут служить основой для анализа и прогнозирования. В условиях растущего объема данных возникает необходимость использования передовых методов анализа данных и машинного обучения для решения проблемы отчислений студентов.

Целью данной дипломной работы является разработка и внедрение модели машинного обучения для прогнозирования вероятности отчисления студентов на основе анализа их цифровых следов. В рамках исследования был собран и обработан набор данных, включающий различные характеристики студентов: пол, средний балл документа об образовании, факультет, присваиваемая квалификация, вид затрат, форма и условие освоения программы, а также результаты экзаменов и зачётов. На основе этих данных была построена и обучена модель машинного обучения, способная предсказывать вероятность отчисления студентов.

Постановка цели предопределила формулировку следующих задач:

1. Исследовать цифровые следы, их разновидности, а также роль в образовании.

2. Рассмотреть возможности применения машинного обучения в образовательной среде.

3. Проанализировать рынок похожих продуктов, которые уже сейчас работают в высших учебных заведениях мира.

4. Определить технических требований для будущей программы.

5. Рассмотреть, с помощью каких инструментов будет реализовываться проект, а также сравнить их.

6. Подготовить визуализацию данных, предоставленных для использования в проекте

7. Создать готовую модель машинного обучения, которую в будущем можно будет развивать и дообучать.

8. Подготовить рабочую программу для деканатов ВУЗа, с помощью которой будет производится достоверный прогноз для каждого студента.

9. Разработать руководство пользователя

Программа, разработанная в ходе исследования, принимает на вход excel файл с данными о студентах и использует обученную модель для предсказания вероятности их отчисления. На выходе программа генерирует новый excel файл, содержащий номера студентов и соответствующие предсказания. Это позволит высшему учебному заведению своевременно выявлять студентов, находящихся в группе риска, и предпринимать необходимые меры для предотвращения их отчисления.

Актуальность данной работы заключается в необходимости снижения уровня отчислений студентов, что способствует улучшению качества образовательного процесса и повышению репутации вуза. Применение методов машинного обучения для анализа образовательных данных открывает новые возможности для персонализации обучения и индивидуального подхода к каждому студенту, что соответствует современным тенденциям в образовательной среде.

1. СОВРЕМЕННЫЕ ТЕХНОЛОГИИ И ИХ ПРИМЕНЕНИЕ В ОБРАЗОВАТЕЛЬНОЙ СРЕДЕ

1.1. Разновидности цифрового следа и его роль в образовательной среде

Цифровой след — это совокупность данных, которые остаются после использования цифровых устройств и взаимодействия с цифровыми сервисами. Эти данные могут включать информацию о посещенных веб-страницах, сделанных покупках, отправленных электронных письмах, использованных приложениях, а также активность в социальных сетях и многое другое.

Цифровой след можно разделить на две основные категории:

1. Активный цифровой след. Это данные, которые пользователи оставляют намеренно. Например, публикации в социальных сетях, комментарии на форумах, заполнение форм на веб-сайтах и отправка электронных писем.
2. Пассивный цифровой след. Это данные, которые собираются без активного участия пользователя. Например, информация о местоположении, собираемая мобильными устройствами, данные о посещаемых веб-страницах, собранные с помощью cookies, и информация о времени использования приложений.

Цифровой след в различных сферах:

1. Электронная коммерция:
    - данные о покупках — информация о приобретенных товарах, предпочтениях и поведении покупателей;
    - платежные данные — информация о способах оплаты, используемых картами и транзакциях;
    - история просмотров — данные о просмотренных товарах и времени, проведенном на страницах продуктов;
2. Социальные сети:

- публикации и комментарии — контент, создаваемый пользователями, включая фотографии, видео и текстовые сообщения;
- взаимодействия — лайки, репосты, комментарии и подписки.
- местоположение — данные о геолокации, собранные с помощью мобильных устройств;

3. Здравоохранение:
    - медицинские записи — история болезни, результаты анализов и данные о лечении;
    - телемедицина — данные о взаимодействиях между пациентами и врачами через онлайн-консультации;
    - фитнес-трекеры — информация о физической активности, сна и общем состоянии здоровья, собранная носимыми устройствами;
4. Финансовый сектор:
    - транзакционные данные — информация о банковских операциях, переводах и платежах;
    - кредитная история — данные о кредитах, платежах и задолженностях;
    - инвестиции — информация о сделках с ценными бумагами и инвестиционных портфелях;
5. Образовательная среда:
    - успеваемость — оценки, результаты экзаменов и зачётов;
    - активность — посещаемость занятий, участие в дополнительных мероприятиях и учебных проектах;
    - взаимодействие — активность в образовательных платформах, взаимодействие с преподавателями и другими студентами;
    - поведение — использование онлайн-ресурсов, время, затраченное на выполнение заданий и участие в онлайн-дискуссиях;

В образовательной среде цифровой след становится важным инструментом для улучшения качества обучения и управления учебным процессом. Рассмотрим подробнее, как цифровой след проявляется в этой сфере:

1. Академическая успеваемость:
    - оценки — все оценки за выполненные задания, тесты, экзамены и проекты хранятся в электронных системах управления обучением;
    - средний балл — сводные данные об успеваемости студентов за весь период обучения;
2. Посещаемость и вовлеченность:
    - посещение занятий — электронные системы фиксируют присутствие студентов на лекциях, семинарах и лабораторных занятиях;
    - участие в мероприятиях — информация о посещении дополнительных учебных и внеклассных мероприятий, таких как конференции, семинары и клубы;
3. Взаимодействие с учебными материалами:
    - использование онлайн-ресурсов — данные о том, какие учебные материалы были просмотрены и как долго студент работал с ними.
    - дискуссии с преподавателями — участие в онлайн-дискуссиях, процесс и ответы на форуме;
4. Коммуникация и взаимодействие:
    - электронная почта и мессенджеры — сообщения, отправленные преподавателям и другим студентам;
    - системы управления обучением (LMS) — взаимодействие студентов с платформами, такими как Moodle, Blackboard и другие, где фиксируется активность, участие в курсах и выполнение заданий;
5. Другие параметры:

- личная информация — пол, возраст, место проживания, гражданство и другая демографическая информация;
- организационные параметры — факультет, специализация, форма и условие обучения (очная, заочная, бюджетная или платная основа);

Использование цифрового следа в образовательной среде позволяет решать множество задач:

1. Цифровой след помогает отслеживать академическую успеваемость студентов, выявлять закономерности и тенденции в обучении, а также определять проблемные области, требующие внимания.
2. На основе анализа цифрового следа можно разрабатывать индивидуальные учебные планы, адаптированные под потребности и способности каждого студента.
3. Анализ данных о студентах позволяет предсказывать вероятность отчисления и выявлять студентов, находящихся в группе риска. Это помогает вузам своевременно предпринимать меры по поддержке таких студентов.
4. Анализ взаимодействия студентов с учебными материалами и платформами позволяет улучшать содержание курсов, делать их более интерактивными и интересными.
5. Цифровой след помогает администрациям вузов эффективно управлять учебным процессом, планировать расписание, распределять ресурсы и принимать обоснованные решения на основе данных.

На рисунке 1 изображено использование цифрового следа в повседневной жизни и образовательной среде.

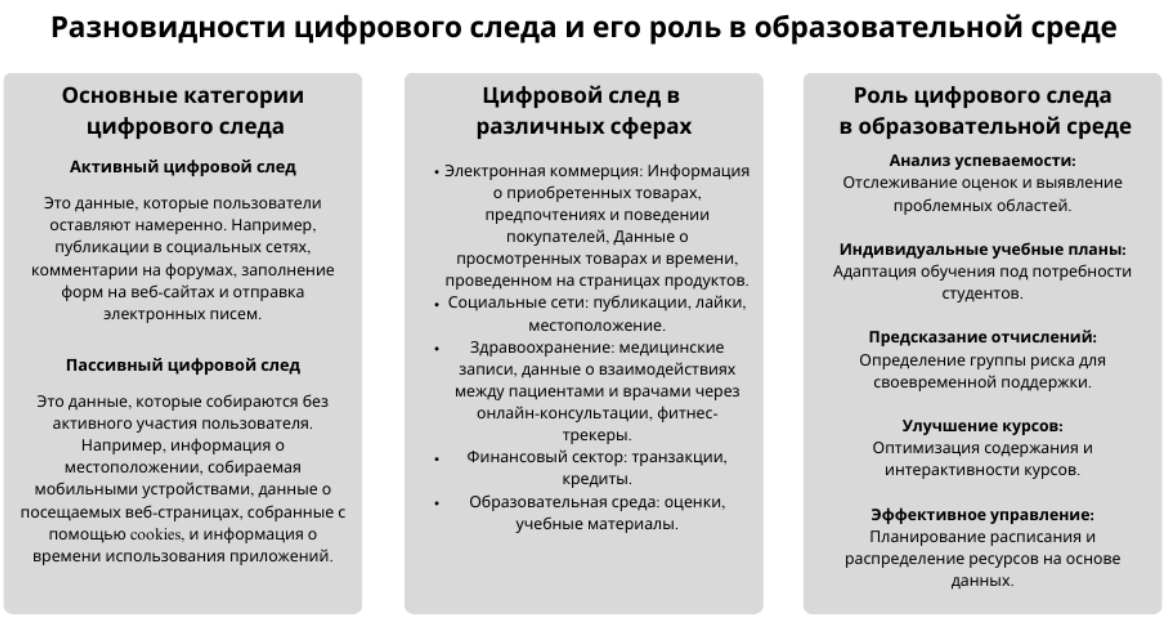


Рисунок 1 – Возможности цифрового следа

1.2. Машинное обучение в образовании

Машинное обучение (ML) — это направление искусственного интеллекта (ИИ), сосредоточенное на создании систем, которые обучаются и развиваются на основе получаемых ими данных. Искусственный интеллект — это широкий термин, который включает в себя компьютерные системы, имитирующие человеческий интеллект.

Сегодня компьютеры работают бок о бок с человеком. Каждый раз, когда мы пользуемся банковскими услугами, делаем покупки в Интернете или общаемся в социальных сетях, алгоритмы машинного обучения помогают сделать это взаимодействие удобнее, эффективнее и безопаснее. Машинное обучение и связанные с ним технологии быстро развиваются: их сегодняшние возможности — это только вершина айсберга.

Приступая к работе с машинным обучением, разработчики полагаются на свои знания в области статистики, теории вероятностей и математического анализа, чтобы успешно создавать модели, способные обучаться со временем. Разработчики также могут принимать решения о

том, будут ли их алгоритмы контролируемыми или нет. Они могут заранее настраивать модель в проекте, а затем разрешать модели учиться без дальнейшего вмешательства.

Часто трудно провести границу между разработчиком и исследователем данных. Иногда разработчики синтезируют данные из модели машинного обучения, а исследователи данных участвуют в разработке решений для конечного пользователя. Сотрудничество между этими двумя дисциплинами может повышать ценность и полезность проектов машинного обучения.

Виды машинного обучения:

1. Машинное обучение с учителем. Специалисты по работе с данными предоставляют алгоритмам размеченные и определенные обучающие данные для оценки корреляций. Демонстрационные данные определяют, как входные данные, так и выходные данные алгоритма. Например, изображения рукописных цифр аннотируются, чтобы указать, какому числу они соответствуют. Система обучения с учителем может распознавать кластеры пикселей и фигур, связанных с каждым числом, при наличии достаточного количества примеров. Со временем система распознает написанные от руки цифры, стабильно различая числа 9 и 4 или 6 и 8. Основное преимущество машинного обучения с учителем — это простота и легкость структуры. Такая система полезна при прогнозировании возможного ограниченного набора результатов, разделении данных на категории или объединении результатов двух других алгоритмов машинного обучения. В нашем случае будет использоваться машинное обучение с учителем, так как у нас уже есть информация о студентах, которые до сих пор учатся, а также о студентах, которые уже были отчислены.
2. Машинное обучение без учителя. Алгоритмы обучения без учителя обучаются на неразмеченных данных. Такие алгоритмы просматривают новые данные, пытаясь установить значимые связи

между входными и заранее определенными выходными данными. Они могут выявлять закономерности и классифицировать данные. Например, алгоритмы без учителя могут группировать новостные статьи с разных новостных веб-сайтов в общие категории, такие как спорт, криминал и т. д. Они могут использовать обработку естественного языка для понимания смысла и эмоций в статье. В розничной торговле обучение без учителя помогает найти закономерности в покупках клиентов и предоставить результаты анализа данных, такие как: покупатель, скорее всего, купит хлеб, если также купит масло. Обучение без учителя полезно для распознавания образов, обнаружения аномалий и автоматического группирования данных по категориям. Эти алгоритмы также можно использовать для автоматической очистки и обработки данных для дальнейшего моделирования. Ограничение этого метода состоит в том, что он не может дать точных прогнозов и самостоятельно выделять конкретные типы выходных данных.

3. Обучение с подкреплением. Обучение с подкреплением — это метод, в котором значения вознаграждения привязаны к различным шагам, которые должен пройти алгоритм. Цель модели — накопить как можно больше призовых баллов и в конечном итоге достичь конечной цели. Большая часть практического применения обучения с подкреплением за последнее десятилетие была связана с видеоиграми. Передовые алгоритмы обучения с подкреплением добились впечатляющих результатов в классических и современных играх, часто значительно превосходя ручные аналоги. Хотя этот метод лучше всего работает в неопределенных и сложных средах данных, он редко применяется в бизнес-контексте. Это неэффективно для четко определенных задач, и предвзятость разработчиков может повлиять на результаты, поскольку специалисты по работе с данными разрабатывают награды, что может влиять на результаты.

На рисунке 2 показана схема разделов машинного обучения.

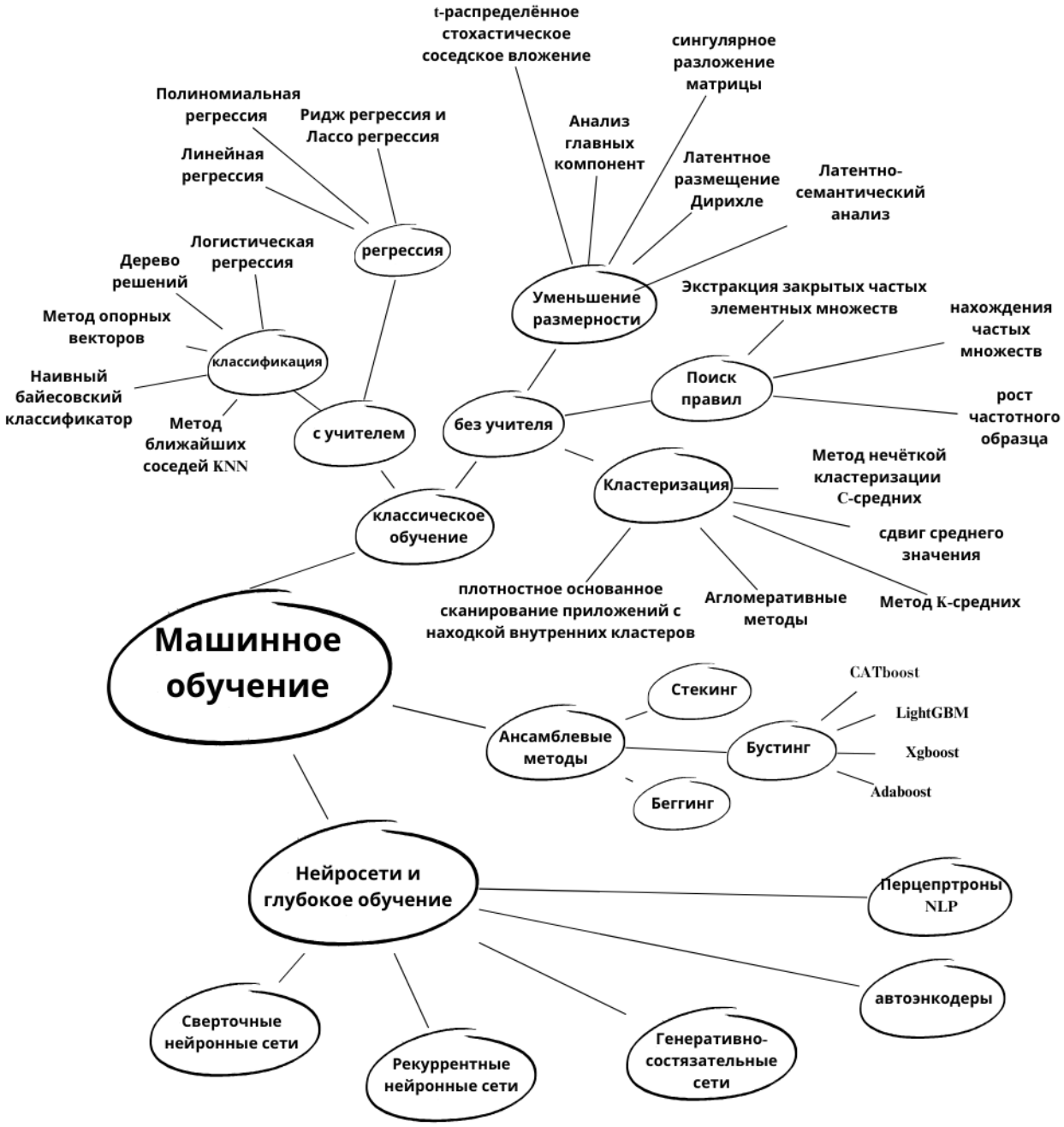


Рисунок 2 – Разделы машинного обучения

В современном мире образование становится все более персонализированным благодаря использованию технологий машинного обучения. Машинное обучение позволяет создавать адаптивные образовательные платформы, которые анализируют данные обучающихся и

автоматически адаптируют учебный материал под их потребности, уровень знаний и темп обучения. Это позволяет студентам получать индивидуализированные задания, рекомендации и обратную связь, что способствует более эффективному усвоению материала.

С помощью машинного обучения можно создавать уникальные учебные планы для каждого обучающегося на основе их интересов, целей и уровня подготовки. Это позволяет студентам более глубоко погружаться в изучаемый материал и развиваться в соответствии с их потребностями.

Технологии машинного обучения позволяют анализировать большие объемы данных об обучающихся и выявлять закономерности в их обучении. На основе этих данных можно предсказывать успеваемость студентов, выявлять проблемные области и предлагать индивидуализированные методы помощи.

Внедрение технологий машинного обучения в образовательную среду открывает новые возможности для улучшения обучения, однако сопровождается и вызовами. Среди преимуществ можно выделить повышение эффективности обучения, повышение мотивации студентов и расширение доступа к образованию. Однако существуют вызовы, такие как необходимость в качественных данных, вопросы приватности и этичности использования данных, а также необходимость обучения педагогов работе с новыми технологиями.

Машинное обучение в образовании представляет собой быстро развивающуюся область, которая обещает революционизировать способы обучения и привнести инновации в образовательный процесс. В будущем можно ожидать более широкого использования адаптивных образовательных платформ, интеграции виртуальной и дополненной реальности в учебный процесс, а также развития интеллектуальных систем поддержки принятия решений для преподавателей и администраторов.

1.3. Анализ рынка

Существуют несколько систем и программных решений, которые уже используются вузами для анализа данных студентов и предсказания вероятности их отчисления. Эти системы помогают университетам улучшить успеваемость студентов и своевременно выявлять тех, кто находится в группе риска.

Civitas Learning — это компания, предоставляющая решения для аналитики и повышения успеваемости студентов в образовательных учреждениях. Они используют методы анализа данных и машинного обучения, чтобы помочь учебным заведениям улучшить успеваемость, повысить уровень вовлеченности и уменьшить количество отчислений.

Основная деятельность Civitas Learning сосредоточена на аналитике успеваемости студентов. С помощью своих инструментов они могут отслеживать прогресс студентов, прогнозировать риск отчисления и выявлять тех, кто нуждается в дополнительной поддержке. Это позволяет учебным заведениям принимать обоснованные решения для своевременного вмешательства и поддержки студентов.

Кроме того, компания предлагает интеллектуальные решения для повышения вовлеченности студентов в учебный процесс. Платформа предоставляет персонализированные рекомендации, аналитику вовлеченности и помогает в разработке более эффективных учебных планов. Это помогает вузам понять, какие факторы влияют на вовлеченность студентов и как её можно улучшить.

Civitas Learning предлагает единую платформу для сбора и анализа данных о студентах, которая интегрируется с различными системами управления обучением. Платформа позволяет собирать данные из разных источников, создавать отчёты и визуализации, что упрощает процесс анализа и принятия решений на основе данных.

Компания также предоставляет поддержку и консультирование учебным заведениям, помогая им эффективно использовать аналитику данных. Это включает обучение персонала, консультации по анализу данных и помощь в разработке стратегий, на основе полученных прогнозов.

Примеры использования технологий Civitas Learning включают прогнозирование и предотвращение отчислений, повышение вовлеченности студентов и оптимизацию учебных планов. Используя машинное обучение и интеграцию с системами управления обучением, Civitas Learning помогает учебным заведениям достигать лучших результатов для своих студентов.

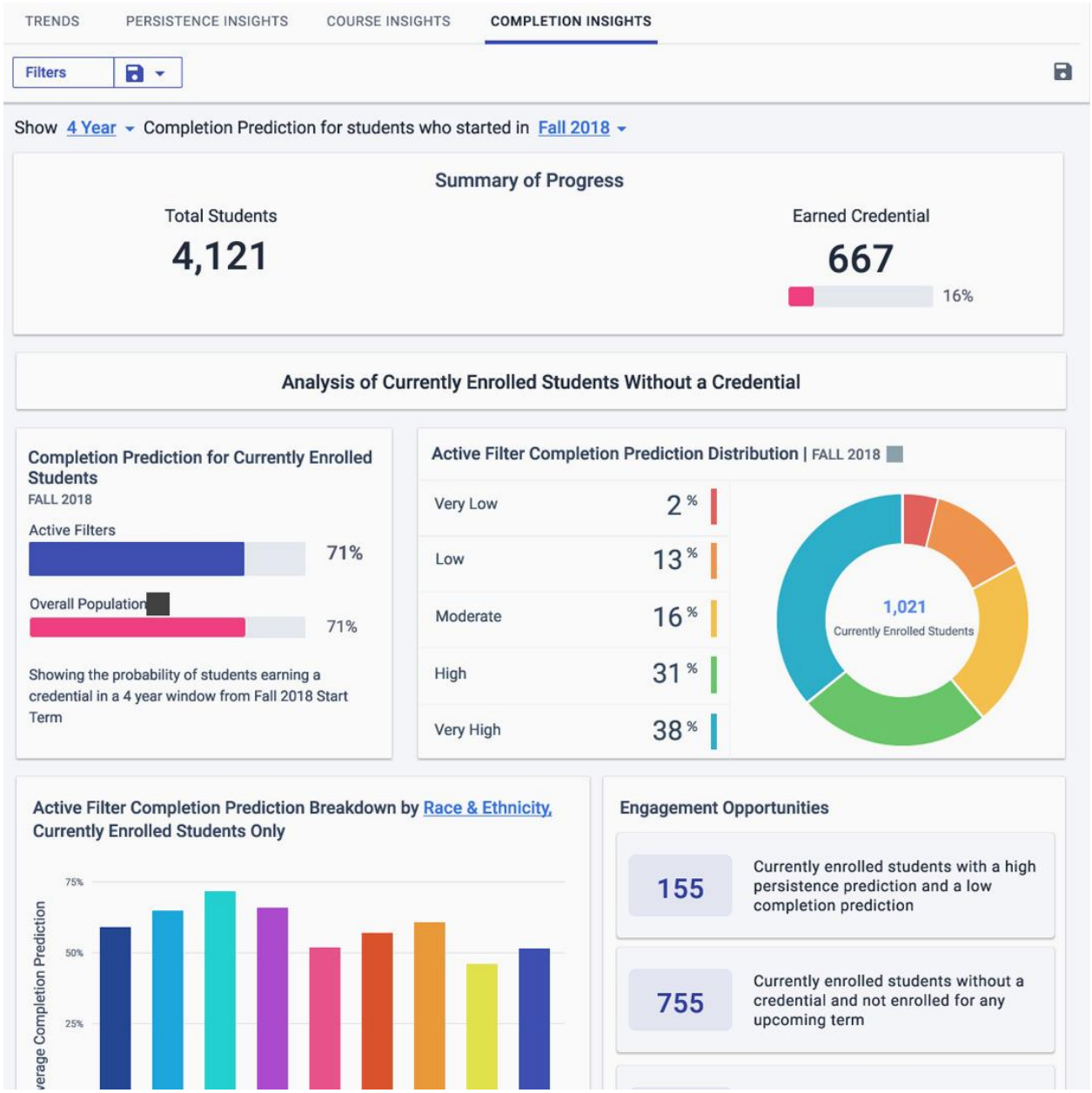


Рисунок 3 – Интерфейс платформы Civitas Learning

Navigate360 — ведущая CRM-система высшего образования. Navigate360, которой доверяют более 850 колледжей и университетов, и которая обслуживает более 10 миллионов студентов, представляет собой мощную технологию, которая объединяет администраторов, преподавателей, сотрудников и студентов в сеть для совместной работы, которая поддерживает весь процесс обучения студентов. Navigate360 основан на более чем десятилетнем исследовании успехов студентов, отточенном на основе миллиардов взаимодействий студентов, а теперь оно также усилено искусственным интеллектом. Партнеры Navigate360 отмечают рост числа выпускников от 3% до 15%. На рисунке 4 изображен графический интерфейс платформы Navigate360.

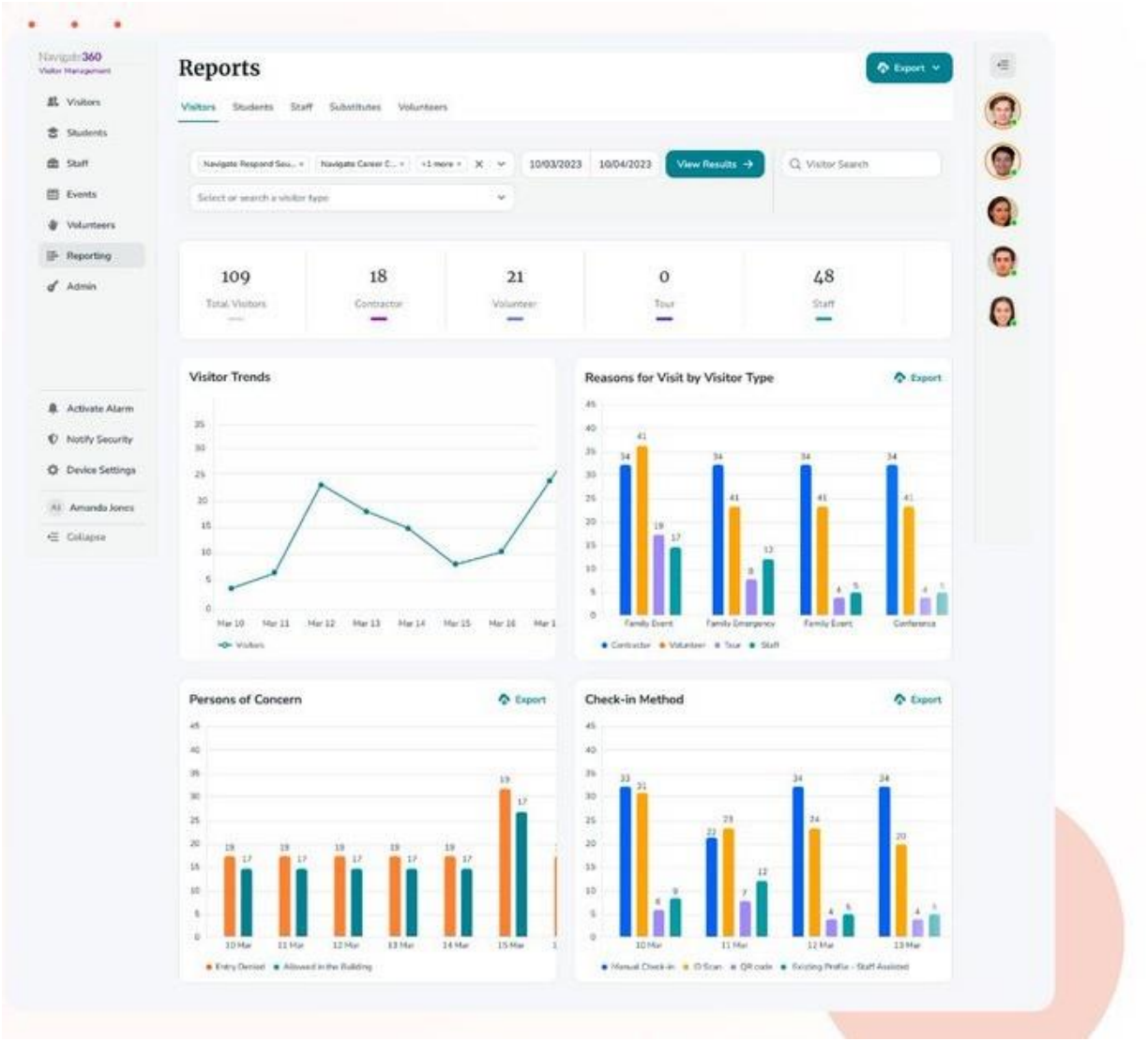


Рисунок 4 – Интерфейс платформы Navigate360

Еще одна популярная платформа для повышения успеваемости студентов — это Starfish by Hobsons. Эта система помогает почти 500 колледжам и университетам масштабировать свои усилия по обеспечению успеха студентов, позволяя большему числу учащихся достигать своих академических и жизненных целей.

Starfish предоставляет надежные данные, которые помогают определить проблемные области и возможности в рамках курсов и групп студентов. Платформа объединяет эти данные с действенными стратегиями, что позволяет вузам более эффективно поддерживать своих студентов. Система выявляет студентов, находящихся в группе риска, и связывает их с ценными ресурсами, такими как консультации и академическая поддержка.

Одна из ключевых функций Starfish — это создание комплексных карьерных и академических планов, которые помогают студентам минимизировать потерю мотивации и максимизировать их успех. Платформа не только помогает учебным заведениям отслеживать успеваемость студентов, но и строит планы по их дальнейшему развитию и карьерным достижениям.

На рисунке 5 изображен графический интерфейс платформы Starfish by Hobsons, демонстрирующий, как данные и стратегии интегрируются для поддержки студентов в их образовательном пути.

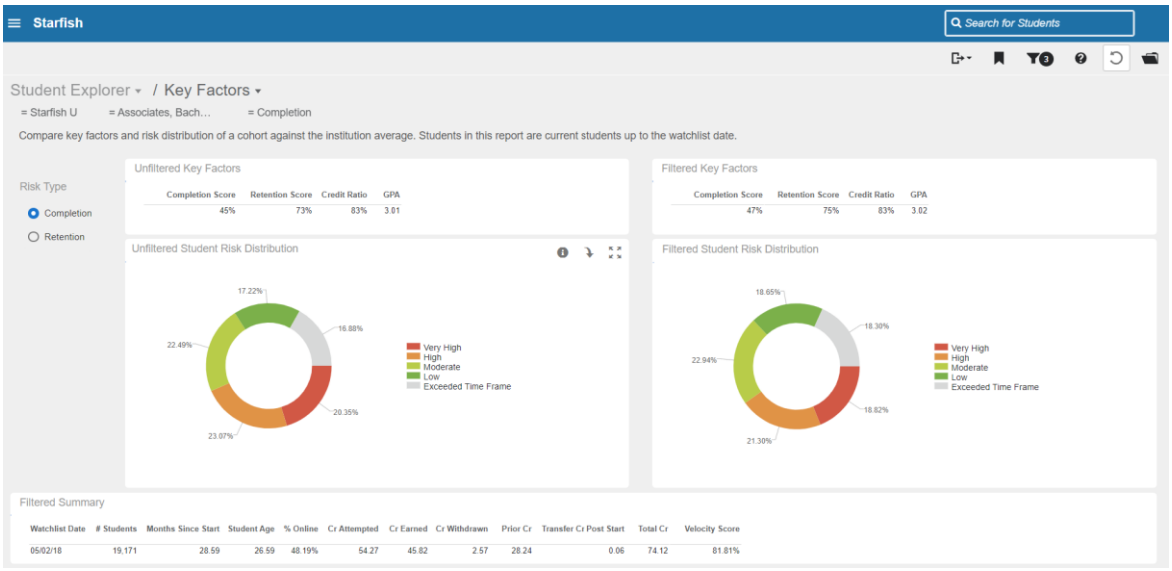


Рисунок 5 – Интерфейс платформы Starfish by Hobsons

Эти программы и системы представляют собой современные решения для управления учебным процессом, которые помогают вузам не только улучшать академическую успеваемость студентов, но и предсказывать возможные риски отчисления. Они обеспечивают:

1. Раннее выявление студентов в группе риска: Системы используют аналитику данных для определения студентов, которые могут испытывать трудности, и предоставляют инструменты для раннего вмешательства.
2. Персонализированные рекомендации: Использование машинного обучения и искусственного интеллекта позволяет создавать индивидуальные учебные планы и рекомендации, адаптированные под нужды каждого студента.
3. Управление взаимодействием и поддержка: Инструменты для коммуникации между студентами и преподавателями помогают улучшать взаимодействие и предоставлять необходимую поддержку.
4. Оптимизация учебного процесса: Анализ данных позволяет вузам эффективно планировать расписание, распределять ресурсы и улучшать содержание курсов.

Таким образом, использование таких систем становится неотъемлемой частью образовательного процесса, способствуя повышению качества образования и улучшению результатов обучения студентов.

1.4. Технические требования

Для проекта с учетом всех пожеланий заказчика, были сформулированы следующие технические требования:

1. Поддержка формата данных. Программа должна быть способна читать данные из Excel файлов.

2. Эффективная обработка данных. Система должна обрабатывать большие объемы данных эффективно и быстро, особенно если данные содержат большое количество строк.
3. Совместимость с моделью машинного обучения. Программа должна быть способна работать с моделью машинного обучения, которая была разработана и обучена заранее. Это включает в себя загрузку модели из файла и применение ее к новым данным для предсказания вероятности отчисления студента.
4. Простота использования. Интерфейс пользователя должен быть интуитивно понятным и легким в использовании, даже для людей без технического образования.
5. Руководство пользователя. Для пользователя сделать подробное руководство со всеми функциями и возможностями программы.

2. АРХИТЕКТУРА ПРОГРАММНО-ИНСТРУМЕНТАЛЬНЫХ СРЕДСТВ

2.1. Методы и средства разработки

Анализ данных и машинное обучение (МО) требуют мощных инструментов, которые могут обрабатывать большие объемы данных, выполнять сложные вычисления и поддерживать разнообразные библиотеки и фреймворки. Выбор правильного языка программирования является одним из ключевых решений при разработке любого проекта, так как он влияет на эффективность, удобство разработки, производительность и поддерживаемость. Рассмотрим наиболее популярные языки программирования, используемые для анализа данных и машинного обучения, и их особенности:

Python является одним из самых популярных языков программирования для анализа данных и машинного обучения благодаря своим многочисленным преимуществам. Одним из главных плюсов Python является его богатый набор библиотек и фреймворков, таких как pandas, NumPy, Matplotlib, Seaborn, scikit-learn, TensorFlow и PyTorch, что делает его мощным инструментом для анализа данных и машинного обучения (оценка 5). Простота и читаемость кода Python также способствуют его популярности, так как он имеет простой и понятный синтаксис, что облегчает обучение и ускоряет процесс разработки (оценка 5). Более того, Python поддерживается большим и активным сообществом, которое предоставляет множество учебных материалов и регулярные обновления, что делает его удобным и надежным выбором (оценка 5). Однако производительность Python несколько уступает другим языкам, особенно для вычислительно интенсивных задач (оценка 3). На рисунке 6 изображены возможности визуализации данных на Python с его многофункциональными библиотеками.

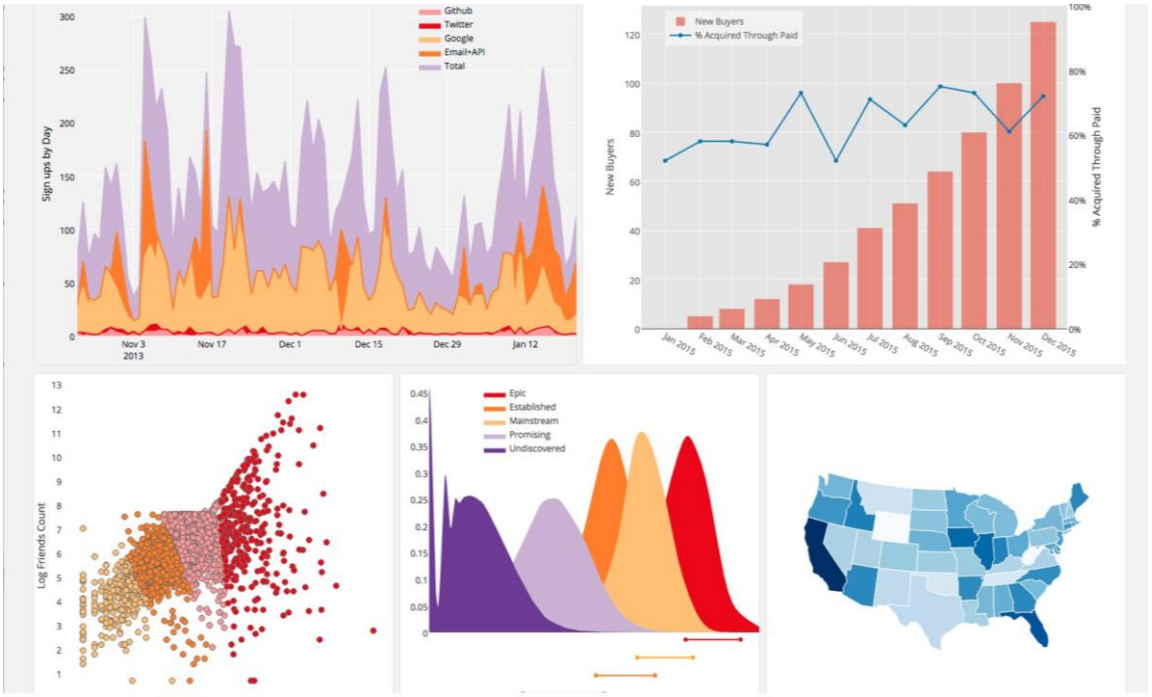


Рисунок 6 – Пример визуализации данных на Python

R — это специализированный язык программирования для статистических вычислений и анализа данных. Одним из ключевых преимуществ R являются мощные статистические и графические возможности, предлагающие широкий спектр статистических методов и инструментов для визуализации данных (оценка 5). Comprehensive R Archive Network (CRAN) содержит тысячи пакетов для различных задач анализа данных и машинного обучения, таких как ggplot2 для визуализации и caret для машинного обучения (оценка 4). R также широко используется в академической среде и научных исследованиях, что способствует быстрому внедрению новейших статистических методов (оценка 4). Однако R несколько сложнее в освоении по сравнению с Python и имеет меньшую масштабируемость (оценка 3). На рисунке 7 показаны возможности визуализации данных на языке R.

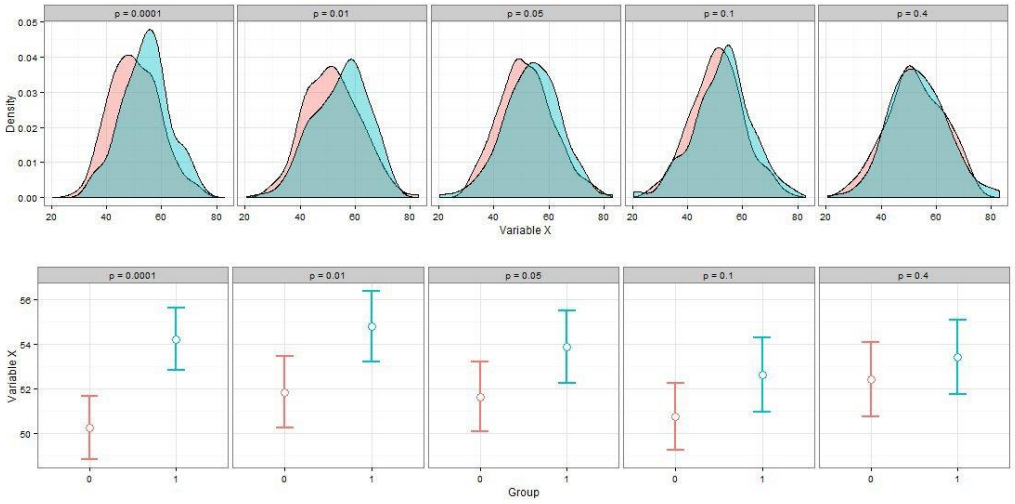


Рисунок 7 – Пример визуализации данных на R

Julia — это язык программирования, созданный для высокопроизводительных вычислений и анализа данных. Одним из главных плюсов Julia является её высокая производительность, близкая к C и Fortran, что делает его идеальным для вычислительно интенсивных задач (оценка 5). Julia также легко взаимодействует с Python, R и другими языками, что делает его гибким инструментом для анализа данных (оценка 4). Однако Julia имеет более ограниченное количество библиотек и фреймворков по сравнению с Python и R (оценка 3), а сообщество и поддержка ещё развиваются (оценка 3). На рисунке 8 изображен пример визуализации данных на Julia.

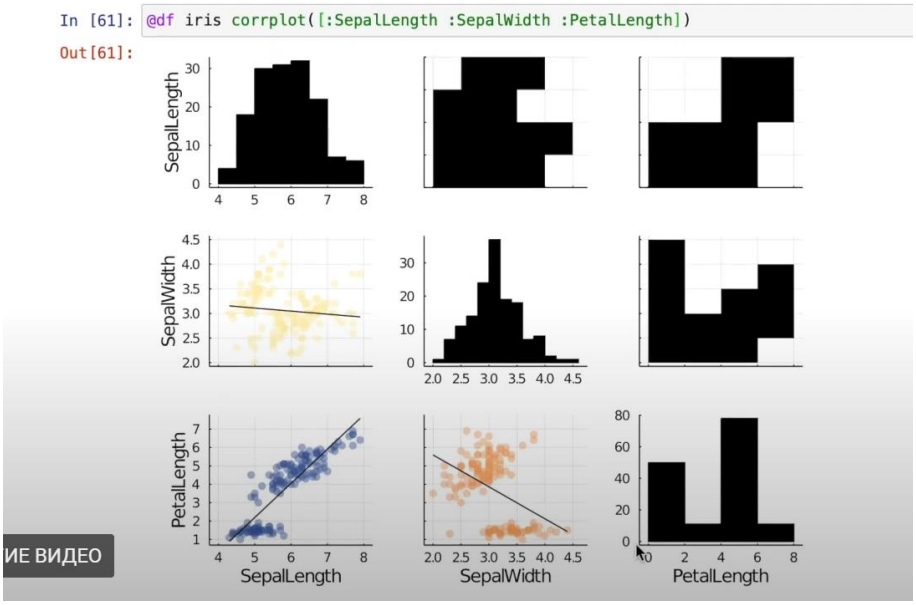


Рисунок 8 – Пример визуализации данных на Julia

Java, хотя и менее популярен в анализе данных по сравнению с Python и R, имеет свои сильные стороны. Java известен своей производительностью и может быть использован для разработки масштабируемых решений для анализа данных (оценка 5). Существуют библиотеки для машинного обучения, такие как Weka и DL4J (Deeplearning4j), которые поддерживают анализ данных и машинное обучение (оценка 3). Однако Java не так удобен и прост в использовании, как Python или R (оценка 3), а визуализация данных на Java также ограничена (оценка 3). На рисунке 9 показаны возможности визуализации график на языке Java.

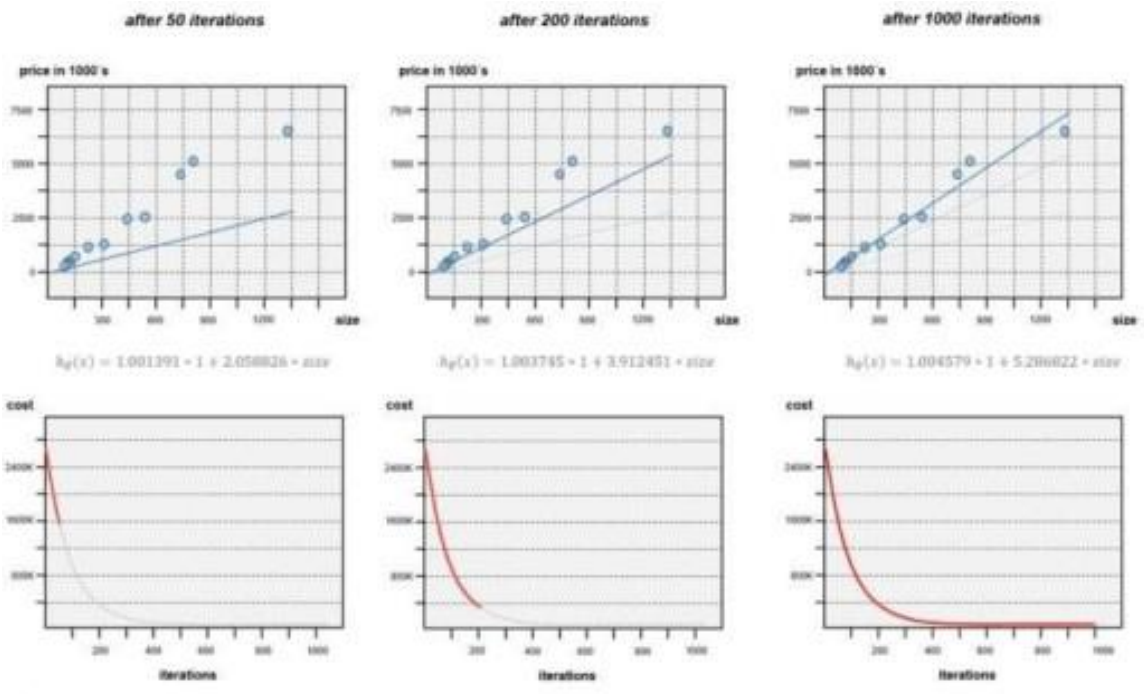


Рисунок 9 – Пример визуализация данных на Java

Scala часто используется вместе с Apache Spark, мощным инструментом для обработки больших данных. Одним из главных плюсов Scala является интеграция с Apache Spark, что делает его важным для анализа больших данных (оценка 5). Scala сочетает в себе объектно-ориентированное и функциональное программирование, что повышает его гибкость и мощность (оценка 4). Однако Scala имеет более сложный синтаксис по сравнению с Python и R (оценка 3), а количество библиотек и фреймворков несколько

ограничено (оценка 3). На рисунке 10 изображен пример визуализации на языке Scala.

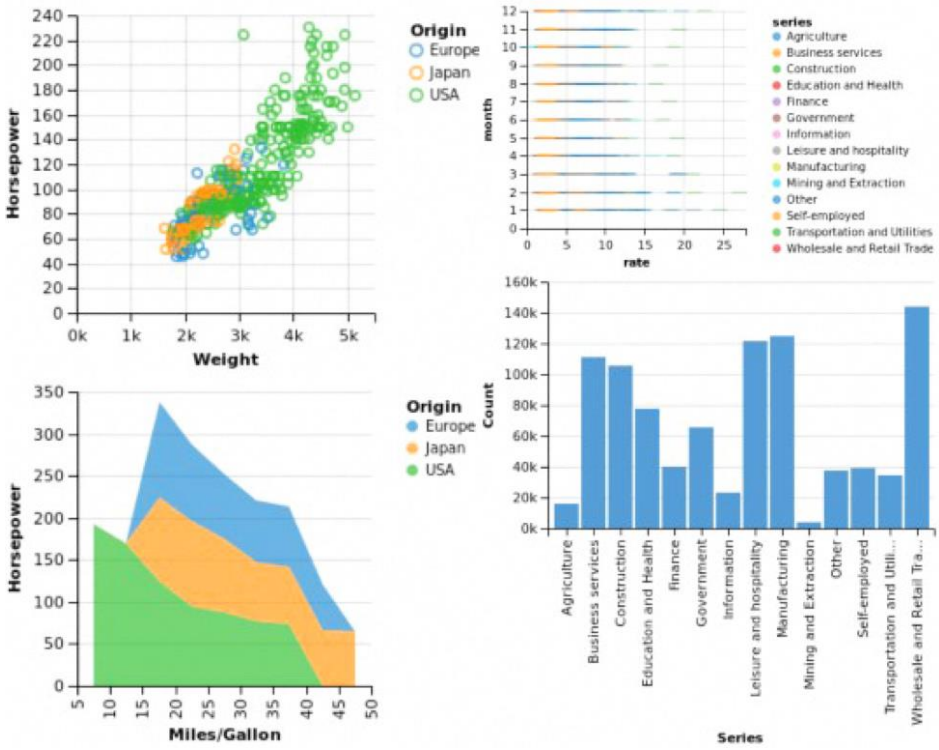


Рисунок 10 – Пример визуализация данных на Scala

MATLAB — это высокоуровневый язык программирования и среда для численных вычислений, широко используемая в инженерии и научных исследованиях. MATLAB предлагает мощные инструменты для анализа данных, визуализации и машинного обучения (оценка 4). Он также имеет обширные библиотеки, такие как Statistics and Machine Learning Toolbox, что облегчает работу с данными (оценка 3). Однако MATLAB менее удобен для масштабируемых решений и имеет более ограниченное сообщество по сравнению с Python и R (оценка 3). Простота и читаемость кода MATLAB также уступают Python (оценка 3). На рисунке 11 изображены возможности визуализации данных на языке MATLAB.

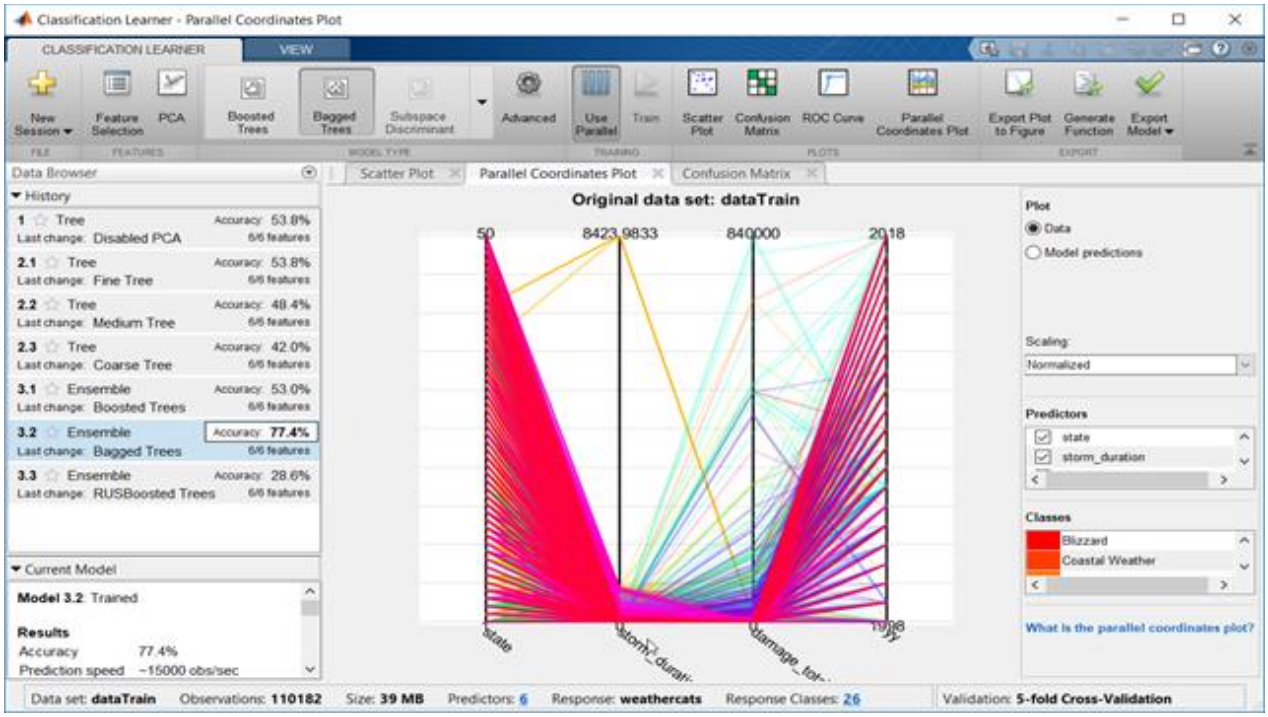


Рисунок 11 – Пример визуализация данных на MATLAB

На основе всех плюсов и минусов языков, представленных выше, была создана таблица, в которое указаны все преимущества и недостатки каждого из языков (табл. 1).

Таблица 1 – Сравнение языков программирования

Признак	Python	R	Julia	Java	Scala	MATLAB
Библиотеки и фреймворки	5	4	3	3	3	3
Простота и читаемость кода	5	4	4	3	3	3
Сообщество и поддержка	5	4	3	4	3	3
Производительность	3	3	5	4	4	4
Визуализация данных	4	5	4	2	3	4
Совместимость	4	5	4	4	4	3
Масштабируемость	4	3	4	5	5	3

Выбор правильного языка программирования для анализа данных и машинного обучения обеспечивает доступ к необходимым инструментам и библиотекам, что значительно ускоряет процесс разработки и улучшает качество конечного продукта. Разные языки программирования предлагают различные преимущества, и их выбор может зависеть от специфики проекта, уровня подготовки команды разработчиков, а также от особенностей данных и задач, которые предстоит решить.

В проекте по анализу цифрового следа студентов и предсказанию вероятности их отчисления выбор языка программирования будет определять не только технические возможности, но и удобство использования разработанного инструмента другими пользователями, такими как преподаватели и администрация учебного заведения. Правильный выбор языка программирования обеспечит возможность эффективного анализа данных, построения точных моделей машинного обучения и разработки интуитивно понятного графического интерфейса. Таким образом, осознанный выбор языка программирования является фундаментальным этапом, от которого зависит успех всего проекта.

Для данного проекта решено было выбрать язык Python. Этот выбор обусловлен не только широкой популярностью языка в сфере анализа данных и машинного обучения, но и его удобством, простотой в изучении и развертывании. Python обладает обширным экосистемой библиотек, что позволяет эффективно работать с данными и создавать точные прогнозные модели. Благодаря своей гибкости и распространенности, Python обеспечит легкость в разработке и дальнейшем масштабировании проекта.

Выбор правильных библиотек является ключевым аспектом при разработке проекта по анализу данных и машинному обучению. Каждая библиотека предоставляет свои уникальные инструменты и функциональность, которые могут значительно упростить разработку и повысить качество конечного результата. Рассмотрим некоторые из наиболее

широко используемых библиотек для Python, которые будут задействованы в нашем проекте:

1. Pandas – это мощная библиотека для работы с данными, которая предоставляет инструменты для чтения, обработки и анализа табличных данных. Основные структуры данных в pandas – это DataFrame и Series, которые позволяют легко выполнять операции по фильтрации, группировке и агрегации данных. pandas также обладает широким набором методов для работы с пропущенными значениями и временными рядами.
2. NumPy – это основная библиотека для вычислений с массивами в Python. Она предоставляет мощные средства для работы с многомерными массивами и выполнения математических операций. NumPy обеспечивает эффективное хранение и операции с данными, что делает его ключевым инструментом для многих приложений в анализе данных и научных вычислениях.
3. Scikit-learn – это библиотека машинного обучения, которая предоставляет простой и эффективный способ реализации различных алгоритмов обучения с учителем и без учителя. С ее помощью можно легко создавать и оценивать модели классификации, регрессии, кластеризации и многие другие. scikit-learn также содержит инструменты для предобработки данных, выбора моделей и оценки их качества.
4. Matplotlib - это базовая библиотека визуализации данных в Python, которая предоставляет широкие возможности для создания различных типов графиков и диаграмм. Matplotlib может использоваться как для создания простых графиков, таких как линейные графики и гистограммы, так и для создания сложных многопанельных диаграмм и 3D-графиков. Он предлагает гибкие настройки для кастомизации внешнего вида графиков, включая шрифты, цвета, размеры и многое

другое. Благодаря своей гибкости и мощным возможностям, Matplotlib является популярным выбором для визуализации данных в Python.

5. Seaborn - это библиотека визуализации данных для Python, которая строит на основе Matplotlib и предоставляет более высокоуровневый интерфейс для создания красивых и информативных графиков. Она часто используется для визуализации статистических моделей и данных. Seaborn упрощает создание сложных графиков, таких как ящики с усами, гистограммы с ядерной оценкой плотности, точечные графики с линейными регрессиями и т. д. Благодаря своему стильному оформлению и дружелюбному API, Seaborn стал популярным выбором для визуализации данных в научных исследованиях и прикладной аналитике.
6. Pickle – это стандартный модуль в Python, предназначенный для сериализации и десериализации объектов. С помощью pickle можно сохранять сложные объекты, такие как модели машинного обучения, словари, списки и другие структуры данных, в файл и загружать их обратно при необходимости. Это делает pickle удобным инструментом для сохранения промежуточных результатов и моделей, что позволяет экономить время на повторную обработку данных или обучение моделей при повторных запусках программы.
7. Tkinter – это библиотека для создания графических пользовательских интерфейсов (GUI) на Python, которая упрощает процесс разработки интерфейсов, делая его доступным даже для начинающих программистов. Tkinter предоставляет простой и интуитивно понятный API, позволяющий быстро создавать окна, диалоги, кнопки, текстовые поля и другие элементы управления. Благодаря совместимости с различными графическими библиотеками, такими как Qt, WxPython и Remi, PySimpleGUI обеспечивает гибкость и кросс-платформенность, что делает его идеальным выбором для разработки приложений с графическим интерфейсом.

2.2. Этапы реализации проекта

Этапы реализации проекта:

1. Сбор данных: Начальным этапом проекта является сбор данных от учебного управления ИжГТУ имени М. Т. Калашникова. Эти данные предоставляются в формате Excel (.xlsx) и включают подробную информацию об успеваемости студентов, их академических задолженностях, а также сведения о направлениях и факультетах. Важно, чтобы данные были полными и структурированными, что обеспечит точность и надежность последующего анализа.
2. Предобработка данных: На этом этапе данные из файлов Excel будут загружены и обработаны с использованием библиотеки pandas. Операции предобработки включают удаление пропущенных значений, обработку дубликатов и исправление возможных ошибок в данных. Также будут выполнены операции по преобразованию данных в формат, удобный для последующего анализа и использования в моделях машинного обучения.
3. Анализ данных и визуализация: С использованием библиотек NumPy и pandas будет проведен анализ данных для выявления ключевых факторов, влияющих на вероятность отчисления студентов. Этот этап включает в себя статистический анализ данных, визуализацию распределения оценок, анализ академических задолженностей и других параметров, что поможет выделить важные тренды и закономерности.
4. Моделирование и машинное обучение: На основе подготовленных данных будет разработана и обучена модель машинного обучения с использованием библиотеки scikit-learn. Данные будут разделены на тренировочные и тестовые выборки для оценки качества модели. Будет проведена настройка гиперпараметров моделей для достижения оптимальной производительности и точности предсказаний.

5. Интерфейс пользователя: Для удобства взаимодействия с системой будет разработан графический пользовательский интерфейс с использованием библиотеки Tkinter. Этот интерфейс позволит пользователям загружать новые данные, запускать анализ, просматривать результаты предсказаний и получать визуализации в удобном и понятном виде.
6. Сохранение и загрузка моделей: Для сохранения и последующей загрузки обученных моделей будет использоваться библиотека pickle. Это обеспечит возможность сохранения текущих состояний моделей и их повторное использование без необходимости повторного обучения при новых данных или сценариях использования.

Этот структурированный подход гарантирует эффективную обработку данных от их сбора до анализа и использования в предсказательной модели. Такой подход делает систему надежной и удобной для интеграции в образовательные учреждения, где анализ данных может играть ключевую роль в принятии решений и оптимизации образовательного процесса.

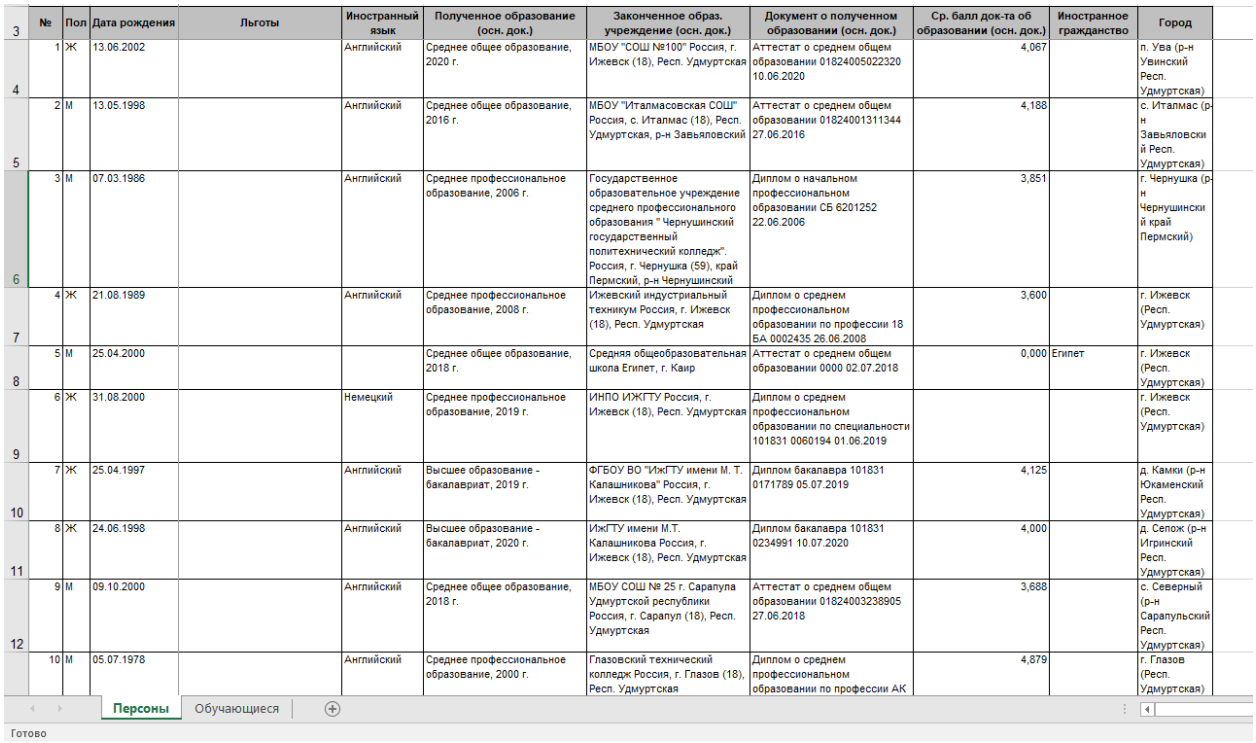
2.3. Предобработка, анализ и визуализация данных

Для реализации проекта были предоставлены данные учебным управлением. Все данные были представлены в формате Excel (.xlsx) и включали два основных документа с информацией о студентах.

Первый документ содержал подробную информацию об успеваемости студентов, включая их академические оценки, результаты экзаменов и курсовых работ. Также были предоставлены данные о текущих задолженностях по предметам, что является важным фактором при анализе вероятности отчисления.

Второй документ включал информацию о направлениях и факультетах, к которым относятся студенты. Эти данные позволяют учитывать контекст и специфику каждого направления при анализе и моделировании.

Пример данных изображен на рисунке 12 и 13.

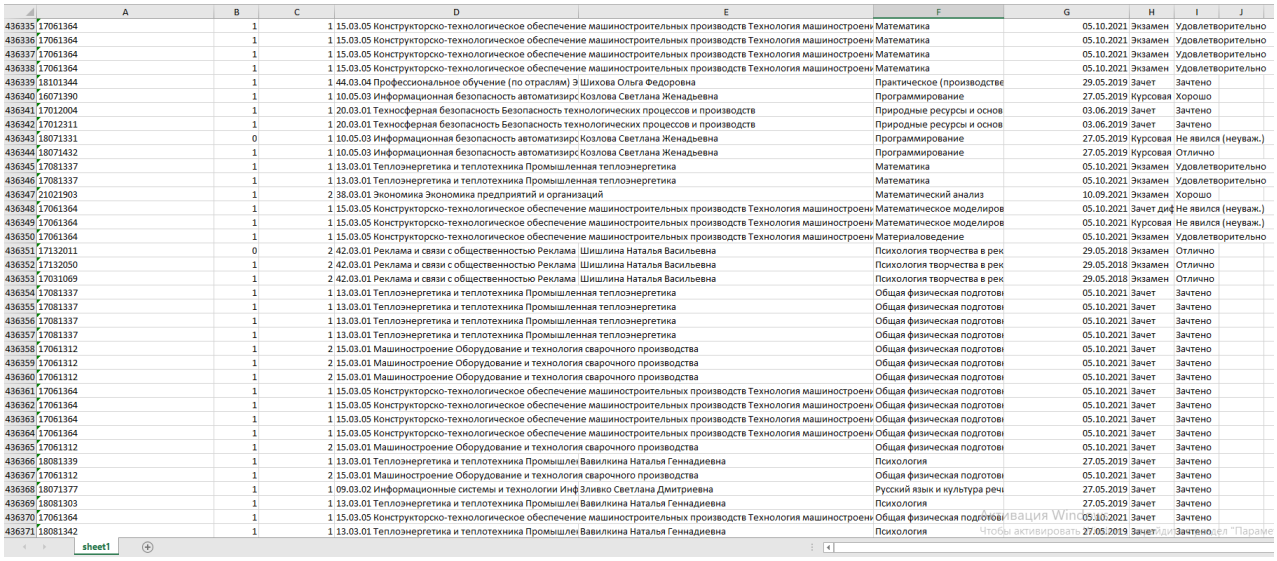


№	Пол	Дата рождения	Льготы	Иностранный язык	Полученное образование (осн. док.)	Законченное образ. учреждение (осн. док.)	Документ о полученном образовании (осн. док.)	Ср. балл док-та об образовании (осн. док.)	Иностранное гражданство	Город
1	Ж	13.06.2002		Английский	Среднее общее образование, 2020 г.	МБОУ "СОШ №100" Россия, г. Ижевск (18), Респ. Удмуртская	Аттестат о среднем общем образовании 01824005022320 10.06.2020	4,067		п. Ува (р-н Увинский Респ. Удмуртская)
2	М	13.05.1998		Английский	Среднее общее образование, 2016 г.	МБОУ "Италмасовская СОШ" Россия, с. Италмас (18), Респ. Удмуртская, р-н Завьяловский	Аттестат о среднем общем образовании 01824001311344 27.06.2016	4,188		с. Италмас (р-н Завьяловский Респ. Удмуртская)
3	М	07.03.1986		Английский	Среднее профессиональное образование, 2006 г.	Государственное образовательное учреждение среднего профессионального образования " Чернушинский государственный политехнический колледж". Россия, г. Чернушка (59), край Пермский, р-н Чернушинский	Диплом о начальном профессиональном образовании СБ 6201252 22.06.2006	3,851		г. Чернушка (р-н Чернушинский край Пермский)
4	Ж	21.08.1989		Английский	Среднее профессиональное образование, 2008 г.	Ижевский индустриальный техникум Россия, г. Ижевск (18), Респ. Удмуртская	Диплом о среднем профессиональном образовании по профессии 18 БА 0002435 26.06.2008	3,600		г. Ижевск (Респ. Удмуртская)
5	М	25.04.2000			Среднее общее образование, 2018 г.	Средняя общеобразовательная школа Египет, г. Каир	Аттестат о среднем общем образовании 0000 02.07.2018	0,000	Египет	г. Ижевск (Респ. Удмуртская)
6	Ж	31.08.2000		Немецкий	Среднее профессиональное образование, 2019 г.	ИНПО ИЖГТУ Россия, г. Ижевск (18), Респ. Удмуртская	Диплом о среднем профессиональном образовании по специальности 101831 0060194 01.06.2019			г. Ижевск (Респ. Удмуртская)
7	Ж	25.04.1997		Английский	Высшее образование - бакалавриат, 2019 г.	ФГБОУ ВО "ИжГТУ имени М. Т. Калашникова" Россия, г. Ижевск (18), Респ. Удмуртская	Диплом бакалавра 101831 0171789 05.07.2019	4,125		д. Камки (р-н Юкаменский Респ. Удмуртская)
8	Ж	24.06.1998		Английский	Высшее образование - бакалавриат, 2020 г.	ИжГТУ имени М.Т. Калашникова Россия, г. Ижевск (18), Респ. Удмуртская	Диплом бакалавра 101831 0234991 10.07.2020	4,000		д. Сепож (р-н Игринский Респ. Удмуртская)
9	М	09.10.2000		Английский	Среднее общее образование, 2018 г.	МБОУ СОШ № 25 г. Сарапула Удмуртской республики Россия, г. Сарапул (18), Респ. Удмуртская	Аттестат о среднем общем образовании 01824003238905 27.06.2018	3,688		с. Северный (р-н Сарапульский Респ. Удмуртская)
10	М	05.07.1978		Английский	Среднее профессиональное образование, 2000 г.	Глазовский технический колледж Россия, г. Глазов (18), Респ. Удмуртская	Диплом о среднем профессиональном образовании по профессии АК	4,879		г. Глазов (Респ. Удмуртская)

Персоны | Обучающиеся

Готово

Рисунок 12 – Личные данные студентов



	A	B	C	D	E	F	G	H	I	J
436335	17061364	1	1	15.03.05 Конструкторско-технологическое обеспечение машиностроительных производств	Технология машиностроени	Математика	05.10.2021	Экзамен	Удовлетворительно	
436336	17061364	1	1	15.03.05 Конструкторско-технологическое обеспечение машиностроительных производств	Технология машиностроени	Математика	05.10.2021	Экзамен	Удовлетворительно	
436337	17061364	1	1	15.03.05 Конструкторско-технологическое обеспечение машиностроительных производств	Технология машиностроени	Математика	05.10.2021	Экзамен	Удовлетворительно	
436338	17061364	1	1	15.03.05 Конструкторско-технологическое обеспечение машиностроительных производств	Технология машиностроени	Математика	05.10.2021	Экзамен	Удовлетворительно	
436339	18101344	1	1	44.03.04 Профессиональное обучение (по отраслям)	Э Шихова Ольга Федоровна	Практическое (производстве	29.05.2019	Зачет	Зачтено	
436340	16071390	1	1	10.05.03 Информационная безопасность автоматизирс	Козлова Светлана Женадьевна	Программирование	27.05.2019	Курсовая	Хорошо	
436341	17012004	1	1	20.03.01 Техносферная безопасность	Безопасность технологических процессов и производств	Природные ресурсы и основ	03.06.2019	Зачет	Зачтено	
436342	17012311	1	1	20.03.01 Техносферная безопасность	Безопасность технологических процессов и производств	Природные ресурсы и основ	03.06.2019	Зачет	Зачтено	
436343	18071331	0	1	10.05.03 Информационная безопасность автоматизирс	Козлова Светлана Женадьевна	Программирование	27.05.2019	Курсовая	Не явился (неуваж.)	
436344	18071432	1	1	10.05.03 Информационная безопасность автоматизирс	Козлова Светлана Женадьевна	Программирование	27.05.2019	Курсовая	Отлично	
436345	17081337	1	1	13.03.01 Теплоэнергетика и теплотехника	Промышленная теплоэнергетика	Математика	05.10.2021	Экзамен	Удовлетворительно	
436346	17081337	1	1	13.03.01 Теплоэнергетика и теплотехника	Промышленная теплоэнергетика	Математика	05.10.2021	Экзамен	Удовлетворительно	
436347	21021903	1	2	38.03.01 Экономика	Экономика предприятий и организаций	Математический анализ	10.09.2021	Экзамен	Хорошо	
436348	17061364	1	1	15.03.05 Конструкторско-технологическое обеспечение машиностроительных производств	Технология машиностроени	Математическое моделиров	05.10.2021	Зачет диф	Не явился (неуваж.)	
436349	17061364	1	1	15.03.05 Конструкторско-технологическое обеспечение машиностроительных производств	Технология машиностроени	Математическое моделиров	05.10.2021	Курсовая	Не явился (неуваж.)	
436350	17061364	1	1	15.03.05 Конструкторско-технологическое обеспечение машиностроительных производств	Технология машиностроени	Материаловедение	05.10.2021	Экзамен	Удовлетворительно	
436351	17132011	0	2	42.03.01 Реклама и связи с общественностью	Реклама	Шишлина Наталья Васильевна	Психология творчества в рек	29.05.2018	Экзамен	Отлично
436352	17132050	1	2	42.03.01 Реклама и связи с общественностью	Реклама	Шишлина Наталья Васильевна	Психология творчества в рек	29.05.2018	Экзамен	Отлично
436353	17031069	1	2	42.03.01 Реклама и связи с общественностью	Реклама	Шишлина Наталья Васильевна	Психология творчества в рек	29.05.2018	Экзамен	Отлично
436354	17081337	1	1	13.03.01 Теплоэнергетика и теплотехника	Промышленная теплоэнергетика	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436355	17081337	1	1	13.03.01 Теплоэнергетика и теплотехника	Промышленная теплоэнергетика	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436356	17081337	1	1	13.03.01 Теплоэнергетика и теплотехника	Промышленная теплоэнергетика	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436357	17081337	1	1	13.03.01 Теплоэнергетика и теплотехника	Промышленная теплоэнергетика	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436358	17061312	1	2	15.03.01 Машиностроение	Оборудование и технология сварочного производства	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436359	17061312	1	2	15.03.01 Машиностроение	Оборудование и технология сварочного производства	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436360	17061312	1	2	15.03.01 Машиностроение	Оборудование и технология сварочного производства	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436361	17061364	1	1	15.03.05 Конструкторско-технологическое обеспечение машиностроительных производств	Технология машиностроени	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436362	17061364	1	1	15.03.05 Конструкторско-технологическое обеспечение машиностроительных производств	Технология машиностроени	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436363	17061364	1	1	15.03.05 Конструкторско-технологическое обеспечение машиностроительных производств	Технология машиностроени	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436364	17061364	1	1	15.03.05 Конструкторско-технологическое обеспечение машиностроительных производств	Технология машиностроени	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436365	17061312	1	2	15.03.01 Машиностроение	Оборудование и технология сварочного производства	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436366	18081339	1	1	13.03.01 Теплоэнергетика и теплотехника	Промышленная	Вавилкина Наталья Геннадиевна	Психология	27.05.2019	Зачет	Зачтено
436367	17061312	1	2	15.03.01 Машиностроение	Оборудование и технология сварочного производства	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436368	18071377	1	1	09.03.02 Информационные системы и технологии	Инф	Зливко Светлана Дмитриевна	Русский язык и культура реч	27.05.2019	Зачет	Зачтено
436369	18081303	1	1	13.03.01 Теплоэнергетика и теплотехника	Промышленная	Вавилкина Наталья Геннадиевна	Психология	27.05.2019	Зачет	Зачтено
436370	17061364	1	1	15.03.05 Конструкторско-технологическое обеспечение машиностроительных производств	Технология машиностроени	Общая физическая подготови	05.10.2021	Зачет	Зачтено	
436371	18081342	1	1	13.03.01 Теплоэнергетика и теплотехника	Промышленная	Вавилкина Наталья Геннадиевна	Психология	27.05.2019	Зачет	Зачтено

sheet1

Рисунок 13 – Данные об оценках по каждому предмету

В данных документах была представлены такие данные как:

- атрибут «Номер зачетной книжки»;
- атрибут «Пол»;
- атрибут «Дата рождения»;
- атрибут «Льготы»;
- атрибут «Полученное образование (осн. док.)»;

- атрибут «Законченное образовательное учреждение»;
- атрибут «Документ о полученном образовании»;
- атрибут «Средний балл документа об образовании»;
- атрибут «Гражданство»;
- атрибут «Город»;
- атрибут «Факультет»;
- атрибут «Направление»;
- атрибут «Вид затрат (бюджет/договор)»;
- атрибут «Форма освоения (очная/заочная)»;
- атрибут «Условия освоения (Полный срок/ускоренная программа)»;
- атрибут «Оценки по каждому предмету»;

Большое количество данных для модели означает наличие значительного объема информации, которая доступна для использования при обучении модели машинного обучения. Это может включать в себя большое количество наблюдений или примеров, а также большое количество признаков или характеристик, которые описывают каждое наблюдение.

Иметь большое количество данных для модели обычно является преимуществом, поскольку это позволяет модели лучше обобщать и делать более точные прогнозы. Большой объем данных обычно позволяет модели выявлять более сложные зависимости и паттерны в данных, что может привести к более точным и устойчивым прогнозам на новых данных.

Предобработка данных для машинного обучения представляет собой процесс подготовки и очистки данных перед тем, как они будут использованы для обучения модели. Она включает в себя ряд шагов, направленных на улучшение качества данных и подготовку их для успешного обучения модели. Вот основные этапы предобработки данных:

1. Удаление дубликатов: идентификация и удаление повторяющихся записей в данных, чтобы избежать искажений при обучении модели.

2. Обработка пропущенных значений: заполнение или удаление пропущенных значений в данных, чтобы предотвратить искажения в процессе обучения модели.
3. Преобразование категориальных переменных: преобразование текстовых или категориальных переменных в числовой формат, чтобы их можно было использовать в моделировании.
4. Масштабирование признаков: нормализация или стандартизация значений признаков, чтобы обеспечить одинаковый масштаб и улучшить производительность модели.
5. Обработка выбросов: идентификация и коррекция выбросов в данных, которые могут искажать результаты модели.
6. Инженерия признаков: создание новых признаков, на основе существующих данных, чтобы улучшить способность модели к выделению закономерностей.
7. Разделение данных на обучающую и тестовую выборки: разделение данных на две независимые выборки - для обучения модели и для её тестирования, чтобы оценить её производительность на новых данных.

Для предобработки данных использовалась библиотека pandas. С ее помощью загружаются данные в dataframe — основной тип данных в pandas, вокруг которого строится вся работа. Его можно представить в виде обычной таблицы с любым количеством столбцов и строк. Также с помощью этой библиотеки мы объединим данные из двух excel файлов.

В данном случае имеется 2 excel файла, в одном из который содержится 2 листа, с информацией о студентах. На рисунках 14-16 показаны первоначальные данные о студентах и их оценках.

```
[3]: df_grades.head(5)
```

[3]:

	student_number	student_status	student_compensation_type	student_program	staff_fio	subject	date	form_control	mark
0	21071352	1	1	09.03.01 Информатика и вычислительная техника ...	NaN	Иностранный язык	NaN	Зачет	NaN
1	21051359	1	1	17.05.02 Стрелково-пушечное, артиллерийское и ...	NaN	Иностранный язык	NaN	Зачет	NaN
2	20091445	1	1	08.03.01 Строительство Промышленное и гражданс...	NaN	Иностранный язык	NaN	Экзамен	NaN
3	21051354	1	1	17.05.02 Стрелково-пушечное, артиллерийское и ...	NaN	Иностранный язык	NaN	Зачет	NaN
4	21071017	1	2	09.03.01 Информатика и вычислительная техника ...	NaN	Иностранный язык	NaN	Зачет	NaN

Рисунок 14 – Первоначальные данные о оценках

```
[4]: df_persons_2017.head(5)
```

[4]:

	№	Пол	Дата рождения	Льготы	Иностранный язык	Полученное образование (осн. док.)	Законченное образ. учреждение (осн. док.)	Документ о полученном образовании (осн. док.)	Ср. балл док-та об образовании (осн. док.)	Иностранное гражданство	Город
0	1	Ж	13.06.2002	NaN	Английский	Среднее общее образование, 2020 г.	МБОУ "СОШ №100" Россия, г. Ижевск (18), Респ. ...	Аттестат о среднем общем образовании 018240050...	4.067	NaN	п. Ува (р-н Увинский Респ. Удмуртская)
1	2	М	13.05.1998	NaN	Английский	Среднее общее образование, 2016 г.	МБОУ "Италмасовская СОШ" Россия, с. Италмас (1...	Аттестат о среднем общем образовании 018240013...	4.188	NaN	с. Италмас (р-н Завьяловский Респ. Удмуртская)
2	3	М	07.03.1986	NaN	Английский	Среднее профессиональное образование, 2006 г.	Государственное образовательное учреждение сре...	Диплом о начальном профессиональном образовани...	3.851	NaN	г. Чернушка (р-н Чернушинский край Пермский)
3	4	Ж	21.08.1989	NaN	Английский	Среднее профессиональное образование, 2008 г.	Ижевский индустриальный техникум Россия, г. Иж...	Диплом о среднем профессиональном образовании ...	3.600	NaN	г. Ижевск (Респ. Удмуртская)
4	5	М	25.04.2000	NaN	Английский	Среднее общее	Средняя	Аттестат о среднем			

Рисунок 15 – Первоначальные данные с информацией о студентах (лист 1)

5]:

	№	Пол	Дата рождения	Полученное образование (док. студ.)	Законченное образ. учреждение (док. студ.)	Документ о полученном образовании (док. студ.)	Ср. балл док-та об образовании (док. студ.)	Личный номер	Зачетная книжка	Состояние	Курс	Группа	Формирующее п
0	1	Ж	13.06.2002	Среднее общее образование, 2020 г.	МБОУ "СОШ №100" Россия, г. Ижевск (18), Респ. ...	Аттестат о среднем общем образовании 018240050...	4.067	200695	20071489	активный	2	Б20-791-1	Информати вычислител техника (Инст
1	2	М	13.05.1998	Среднее общее образование, 2016 г.	МБОУ "Италмасовская СОШ" Россия, с. Италмас (1...	Аттестат о среднем общем образовании 018240013...	4.188	203013	20092127	активный	2	Б20-191-1з	Информати вычислител техника (Инст
2	3	М	07.03.1986	Среднее профессиональное образование, 2006 г.	Государственное образовательное учреждение сре...	Диплом о начальном профессиональном образовани...	3.851	193580	19042199	активный	3	Б19-831-1зу	Приборостроитель (Факул
3	4	Ж	21.08.1989	Среднее профессиональное образование, 2008 г.	Ижевский индустриальный техникум Россия, г. Иж...	Диплом о среднем профессиональном образовании ...	3.600	172802	17062039	отчислен	3	Б17-721-1з	Современ технол машинострое автом
4	5	М	25.04.2000	Среднее общее образование, 2018 г.	Средняя общеобразовательная школа Египет, г. Каир	Аттестат о среднем общем образовании 0000 02.0...	0.000	212298	21561314	активный	1	Б21-310-10	Инст международ образовател прогр

Рисунок 16 – Первоначальные данные с информацией о студентах (лист 2)

Следующим шагом будет удаление всех ненужных столбцов, а также пересекающихся с другими датафреймами столбцов. Это такие столбцы, как например «student_status», «student_compensation_type», «student_program»,

«staff_fio», «subject», «date», «дата рождения», «льготы», «иностранный язык», «законченное образ. учреждение (осн. док.)».

Далее создаются новые категориальные столбцы под названиями «Гражданство_категория» и «Город_категория». Это сделано для того, чтобы каждый столбец имел два варианта значения: либо у человека российское гражданство, либо нет, аналогично с городом. Например, студент приехал из другого города для поступления в университет, или же этот студент постоянно проживает в Ижевске.

Также изменится представление оценок в датафрейме. В нем будет содержаться подсчет каждой оценки, таких как «отлично», «хорошо», «удовлетворительно», «зачет» и другие.

Все столбцы датафрейма изображены на рисунке 17. Эти признаки уже будут участвовать в машинном обучении.

```
Пол                                              object
Ср. балл док-та об образовании (док. студ.)     float64
Формирующее подр.                                object
Присваиваемая квалификация                       object
Вид затрат                                       object
Форма освоения                                   object
Условие освоения                                 object
Целевик                                          object
Выбыл                                             int64
Зачтено                                           int64
Не должен сдавать                                 int64
Не допущен                                        int64
Не зачтено                                        int64
Не явился (неуваж.)                               int64
Не явился (уваж.)                                 int64
Неудовлетворительно                               int64
Неявка по неизвестной причине                     int64
Отлично                                           int64
Удовлетворительно                                 int64
Хорошо                                            int64
Status                                            int32
Город_Категория                                   int32
Гражданство_Категория                             int32
dtype: object
```

Рисунок 17 – Все столбцы готового датафрейма

Также в столбце «Status», раннее называвшимся «Состояние», хранится информация по каждому студенту о том, активен он, отчислен или, например, находится в академическом отпуске. Все возможные варианты переменной отображены на рисунке 18.

```
[28]: df_merged['Состояние'].value_counts()

[28]: Состояние
      активный              4897
      отчислен               984
      отп.акад.б.посещ       191
      отп.по.ух.за.реб.        6
      отп.по.бер.и.род         1
      Name: count, dtype: int64
```

Рисунок 18 – Все возможные варианты переменной “Состояние”

В данном случае необходимо привести данные столбца только к двум возможным вариантам – «активен» или «отчислен». Также можно сразу закодировать эти категориальные данные для удобства, представив их в виде 1 для «активен» и 0 для «отчислен».

В датафрейме также необходимо заполнить пустые ячейки в столбцах «Ср. балл док-та об образовании (док. студ.)», «Город_Категория» и «Гражданство_Категория». Для среднего балла аттестата было принято решение заполнить пропуски средним значением, а пустые ячейки в столбцах «Город_Категория» и «Гражданство_Категория» заменить на «Ижевск» и «Россия» соответственно.

Для создания модели машинного обучения нужны обработанные данные. Однако нужно закончить обработку данных до конца. Для того, чтобы начать обучать модель, требуется закодировать все категориальные переменные.

Кодирование категориальных переменных в задачах машинного обучения является важным этапом и требуется по нескольким причинам:

1. Необходимость числовых данных: многие алгоритмы машинного обучения работают только с числовыми данными. Категориальные переменные, такие как тип образования (например, бакалавр, магистр), форма занятости (бюджет, платно), или регион проживания (город, село), представлены в виде текстовых или категориальных значений, которые не могут быть напрямую использованы моделями.
2. Избегание неверных интерпретаций: если не проводить кодирование, модель может неправильно интерпретировать категориальные значения как числовые, что может привести к искажению результатов и неправильным выводам.
3. Улучшение производительности модели: кодирование категориальных переменных позволяет моделям эффективнее использовать информацию о взаимосвязях между данными. Например, при использовании методов кодирования, таких как порядковое кодирование или кодирование с использованием гиперплоскостей, сохраняется структура данных, что может улучшить точность и скорость обучения моделей.

Существует несколько методов кодирования категориальных переменных:

1. Порядковое кодирование (Ordinal Encoding). Присваивает каждому уникальному значению категориальной переменной целочисленное значение. Этот метод подходит, когда значения могут быть упорядочены по какому-то признаку, например, по уровню образования.
2. One-Hot Encoding. Создает бинарные (дамми) переменные для каждого уникального значения категориальной переменной. Этот метод подходит, когда значения категориальной переменной не имеют порядка или, когда их количество не очень большое.
3. Target Encoding. Использует целевую переменную для вычисления среднего значения целевой переменной для каждого уникального

значения категориальной переменной. Этот метод может помочь модели лучше улавливать зависимости между категориальными переменными и целевой переменной.

Кодирование категориальных переменных зависит от конкретной задачи, характеристик данных и выбранного алгоритма машинного обучения. Правильный выбор метода кодирования помогает обеспечить корректность и эффективность работы моделей машинного обучения. В нашем случае будем использовать Ordinal Encoding.

Также важно привести числовые переменные в вид, когда среднее значение будет равно 0, а стандартное отклонение будет равно 1. Приведение числовых переменных к такому виду имеет несколько важных причин в контексте подготовки данных для моделей машинного обучения:

1. Улучшение сходимости алгоритмов: многие алгоритмы машинного обучения лучше работают на данных, которые имеют стандартное нормальное распределение или близкое к нему. Приведение переменных к такому виду помогает алгоритмам быстрее и лучше сходиться к оптимальным решениям
2. Предотвращение доминирования признаков: если признаки имеют различные масштабы (например, один признак варьируется от 0 до 1, а другой от 100 до 1000), это может привести к проблеме, где признак с большими числовыми значениями будет доминировать в процессе обучения модели. Приведение всех признаков к одному масштабу (стандартному отклонению 1) помогает предотвратить такое доминирование.
3. Интерпретируемость модели: приведение признаков к стандартному виду (среднее 0, стандартное отклонение 1) упрощает интерпретацию влияния каждого признака на целевую переменную. Это делает модель более понятной для аналитиков и экспертов предметной области, которые могут лучше понять, как каждый признак влияет на результаты предсказаний.

4. Улучшение стабильности и репрезентативности результатов: когда данные находятся в одном масштабе, модель становится более стабильной и предсказания более репрезентативными. Это особенно важно при работе с данными, где признаки измеряются в разных единицах или имеют различные диапазоны значений.

На рисунке 19 изображены закодированные и приведенные к среднему значению 0 и стандартному отклонению 1. Такие данные готовы в машинному обучению

```
df1 = pd.DataFrame(X_processed)
df1.head(30)
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	-2.903101e-01	-0.053468	-0.639867	-0.056321	-0.305835	-0.349814	-0.682798	-0.112411	-0.213084	-0.055664	0.268887	-0.782109	-0.545176	-0.979000	0.191941
1	-5.271105e-03	-0.053468	-1.131131	-0.056321	-0.305835	-0.349814	-0.202654	-0.112411	-0.213084	-0.055664	-0.807495	-0.193251	-1.059925	-0.979000	0.191941
2	-1.390419e+00	-0.053468	-0.557989	-0.056321	-0.305835	-0.349814	-0.202654	-0.112411	-0.213084	-0.055664	-0.499958	-0.389537	-0.716759	1.021451	0.191941
3	-4.790131e-01	-0.053468	-0.557989	-0.056321	-0.305835	1.062217	-0.442726	-0.112411	2.491176	-0.055664	-0.038651	0.199321	0.484322	1.021451	0.191941
4	-4.481416e-01	-0.053468	-0.476112	-0.056321	-0.305835	-0.349814	-0.682798	-0.112411	-0.213084	-0.055664	-0.038651	-0.782109	-0.373593	-0.979000	0.191941
5	-1.183118e+00	-0.053468	1.570823	-0.056321	-0.305835	1.062217	-0.322690	-0.112411	-0.213084	-0.055664	-0.346189	2.162181	0.999071	-0.979000	0.191941
6	1.622514e+00	-0.053468	2.307719	-0.056321	-0.305835	-0.349814	-0.682798	-0.112411	-0.213084	-0.055664	0.422656	0.984465	2.371734	-0.979000	0.191941
7	-1.820281e-13	-0.053468	-0.230480	-0.056321	-0.305835	-0.349814	-0.682798	-0.112411	-0.213084	-0.055664	-0.499958	-0.782109	0.141156	-0.979000	0.191941
8	2.444325e-01	-0.053468	0.588294	-0.056321	-0.305835	-0.349814	-0.682798	-0.112411	-0.213084	-0.055664	0.115118	-0.389537	0.655905	-0.979000	0.191941
9	-1.536798e-01	-0.053468	-1.049253	-0.056321	-0.305835	-0.349814	-0.442726	-0.112411	-0.213084	-0.055664	-0.807495	-0.193251	-0.716759	1.021451	0.191941
10	-4.481416e-01	-0.053468	0.588294	-0.056321	-0.305835	-0.349814	-0.682798	-0.112411	-0.213084	-0.055664	0.268887	-0.782109	1.170653	1.021451	0.191941
11	1.419924e+00	-0.053468	0.424539	-0.056321	-0.305835	-0.349814	-0.682798	-0.112411	-0.213084	-0.055664	2.114114	-0.585823	1.513819	1.021451	0.191941
12	-7.850059e-01	-0.053468	1.570823	-0.056321	-0.305835	-0.349814	-0.562762	-0.112411	-0.213084	-0.055664	0.268887	2.358467	0.827488	1.021451	0.191941
13	1.200844e+00	-0.053468	-1.294886	-0.056321	0.366316	-0.349814	1.597889	-0.112411	-0.213084	-0.055664	-0.807495	-0.782109	-1.059925	1.021451	-5.209944
14	-1.820281e-13	-0.053468	-0.721744	-0.056321	-0.305835	-0.349814	0.037419	-0.112411	-0.213084	-0.055664	-0.807495	-0.193251	-0.373593	1.021451	-5.209944
15	-4.790131e-01	-0.053468	-0.312357	-0.056321	-0.305835	1.062217	1.597889	-0.112411	-0.213084	-0.055664	-0.807495	0.591893	-0.716759	1.021451	-5.209944
16	-4.790131e-01	-0.053468	0.015153	-0.056321	-0.305835	-0.349814	1.477853	-0.112411	-0.213084	-0.055664	-0.653726	0.395607	-0.373593	1.021451	-5.209944
17	-1.820281e-13	-0.053468	-0.967376	-0.056321	-0.305835	-0.349814	0.877672	-0.112411	-0.213084	-0.055664	-0.807495	-0.389537	-0.888342	1.021451	-5.209944
18	-1.820281e-13	-0.053468	-0.476112	-0.056321	-0.305835	1.062217	0.397527	-0.112411	-0.213084	-0.055664	-0.807495	0.199321	-0.716759	1.021451	-5.209944
19	-4.790131e-01	-0.053468	-0.803621	-0.056321	-0.305835	1.062217	1.717925	-0.112411	-0.213084	-0.055664	-0.807495	-0.193251	-0.888342	1.021451	-5.209944
20	-1.820281e-13	-0.053468	-0.721744	-0.056321	-0.305835	-0.349814	0.397527	-0.112411	-0.213084	-0.055664	-0.807495	-0.193251	-0.716759	1.021451	-5.209944
21	-1.820281e-13	-0.053468	-0.557989	-0.056321	0.366316	-0.349814	0.037419	-0.112411	-0.213084	-0.055664	-0.807495	0.199321	-0.888342	1.021451	-5.209944
22	-1.820281e-13	-0.053468	-1.294886	-0.056321	-0.305835	-0.349814	-0.442726	-0.112411	2.491176	-0.055664	-0.807495	-0.585823	-0.716759	1.021451	-5.209944
23	-1.820281e-13	-0.053468	-1.294886	-0.056321	-0.305835	-0.349814	1.117744	-0.112411	-0.213084	-0.055664	-0.807495	-0.389537	-1.059925	1.021451	-5.209944

Рисунок 19 – Полностью обработанные данные.

Для визуализации данных и результатов предсказаний использовались библиотеки matplotlib и seaborn. С их помощью были созданы различные графики и диаграммы, которые позволили наглядно представить распределение данных, важность признаков и результаты предсказаний модели. Эти визуализации сыграли ключевую роль в анализе данных и интерпретации работы модели, предоставляя ясное и интуитивное понимание ключевых аспектов исследования.

На рисунке 20 изображено распределение среднего балла аттестата среди всех студентов ВУЗа.

```
import matplotlib.pyplot as plt
import matplotlib as mpl
import seaborn as sns
import warnings; warnings.filterwarnings(action='once')
```

```
df['Ср. балл док-та об образовании (док. студ.)'].plot(kind='hist')
plt.title('Распределение среднего балла по Ижгту')
```

```
Text(0.5, 1.0, 'Распределение среднего балла по Ижгту')
```

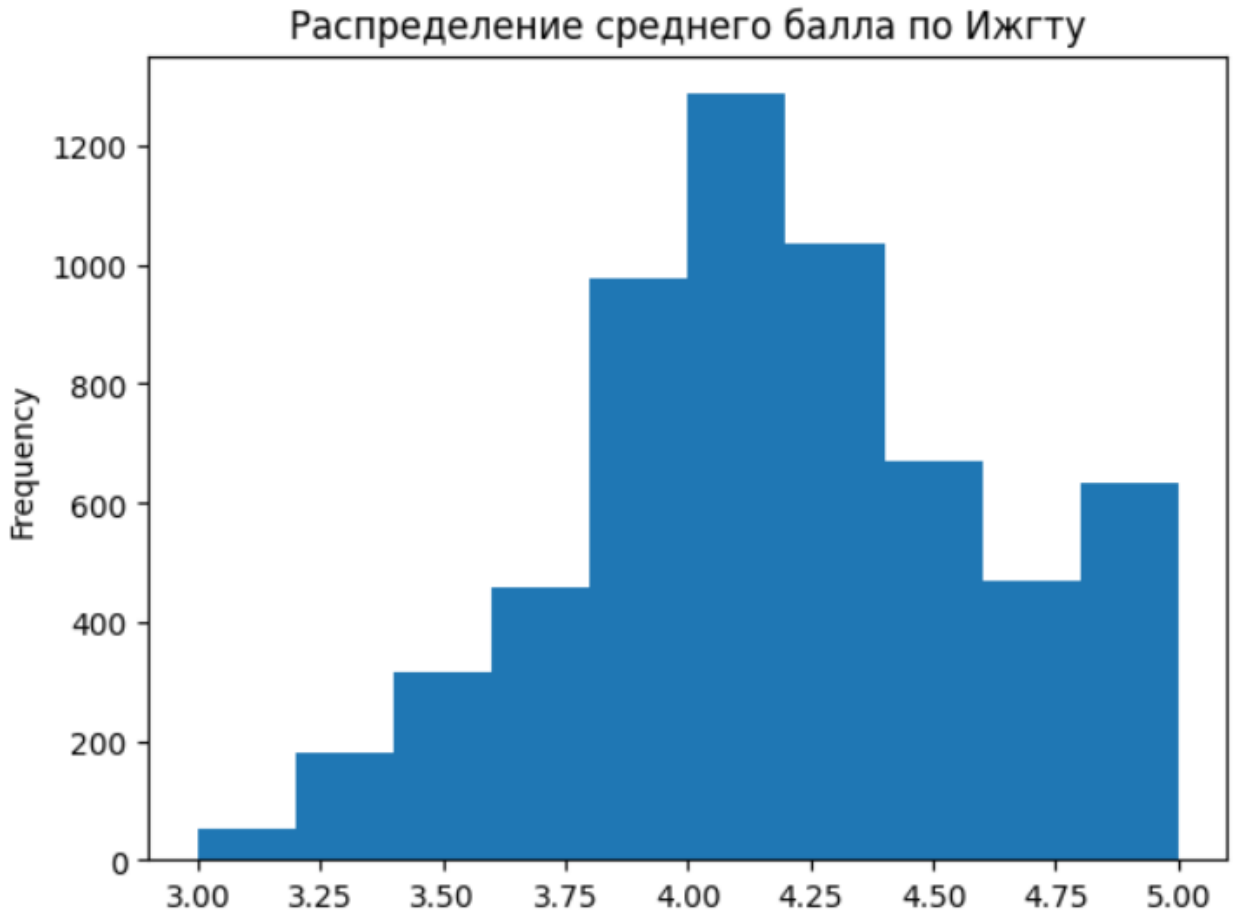


Рисунок 20 – распределение среднего балла аттестата среди всех студентов ВУЗа

Использование boxplot (ящик с усами) для визуализации распределения среднего балла по полу имеет несколько преимуществ:

1. Идентификация медианы и межквартильного диапазона: boxplot позволяет легко увидеть медиану (центральную тенденцию) и

межквартильный диапазон (распространение данных), что дает представление о типичных значениях среднего балла для разных полов.

2. Выявление выбросов: ящик с усами помогает выявлять выбросы, которые отображаются как точки за пределами усов. Это позволяет обнаружить аномалии в данных, которые могут повлиять на анализ.
3. Сравнение групп: boxplot удобно использовать для сравнения распределений среднего балла между различными группами, в данном случае между мужчинами и женщинами. Можно легко увидеть различия в центральных тенденциях и разбросе данных.

На рисунке 21 изображено распределение среднего балла аттестата среди мужчин и девушек ВУЗа с помощью boxplot.

```
# Распределение среднего балла по полу
plt.figure(figsize=(10, 6))
sns.boxplot(x='Пол', y='Ср. балл док-та об образовании (док. студ.)', data=df)
plt.title('Распределение среднего балла по полу')
plt.show()
```

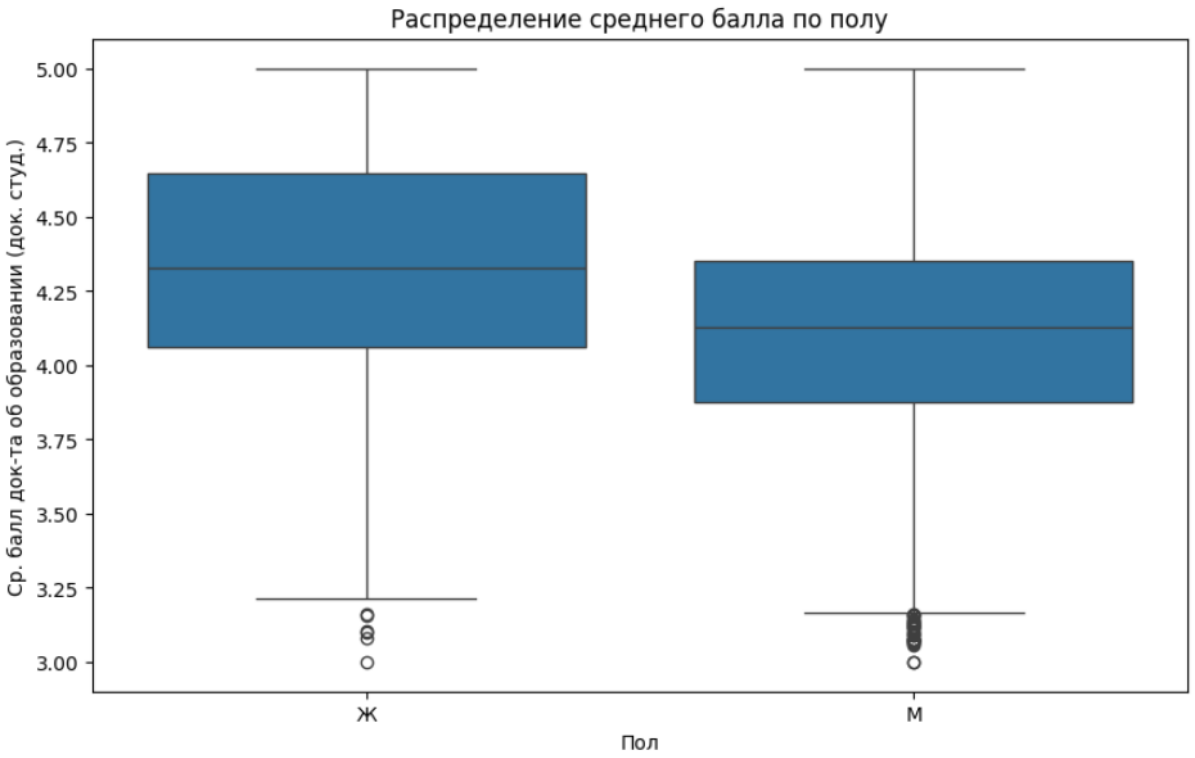


Рисунок 21 – Распределение среднего балла аттестата среди мужчин и девушек вуза

Объяснение Boxplot:

1. Медиана: горизонтальная линия внутри ящика представляет медиану данных.
2. Q1 и Q3: нижняя и верхняя границы ящика представляют первый (Q1) и третий (Q3) квартиль, соответственно.
3. Усы: линии, выходящие за пределы ящика, представляют данные, которые находятся в пределах 1.5 межквартильного размаха от Q1 и Q3.
4. Выбросы: точки, расположенные за пределами усов, являются выбросами или экстремальными значениями данных.

На рисунке выше явно видно, что девушки, поступающие в наш ВУЗ более прилежно, относительно мужчин, учились в школе или техникуме. А также прослеживается, что выбросов по нижней границе также гораздо больше у мужчин.

На рисунке 22 изображен также график boxplot со средним баллом в зависимости от города.

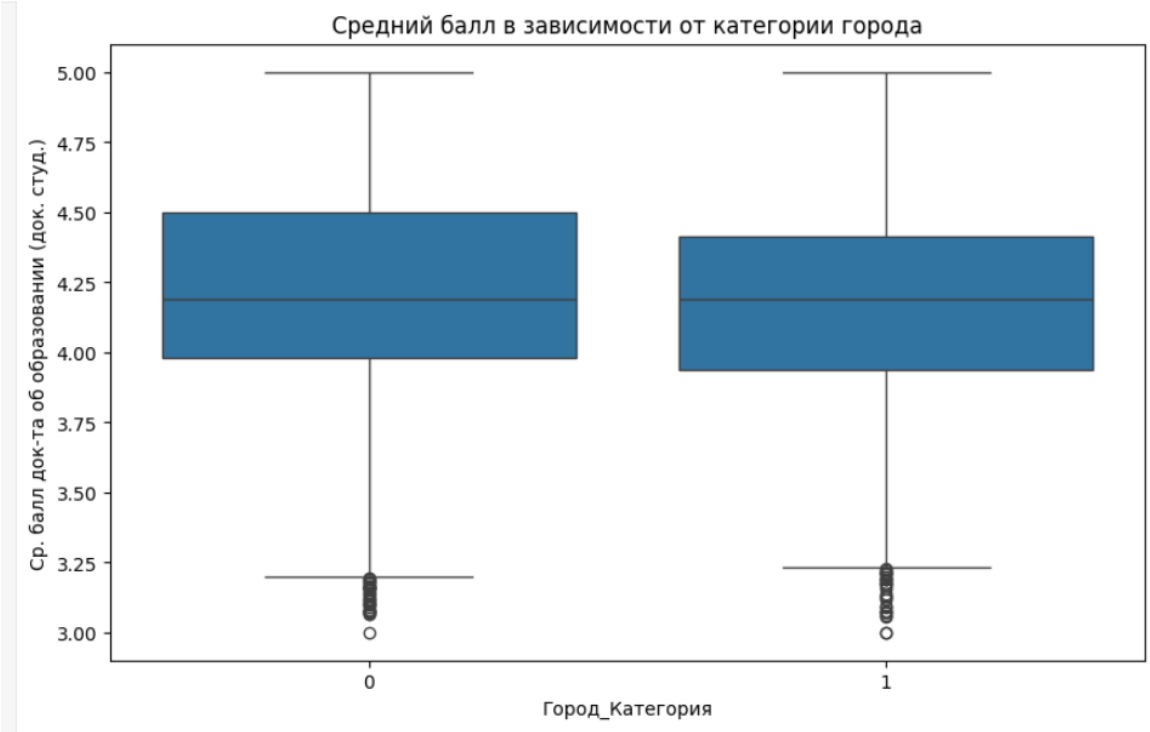


Рисунок 22 – Средний балл аттестата в зависимости от города

На рисунке видно, средний балл аттестата чуть выше у приезжих из других городов, но эта разница не велика.

На рисунке 23 изображен также график boxplot с отношением количества оценок «отлично» с видом затрат.

```
plt.title('Зависимость вида затрат и количества оценок "отлично"')
plt.xlabel('Вид затрат')
plt.ylabel('Количество оценок "отлично"')
plt.show()
```

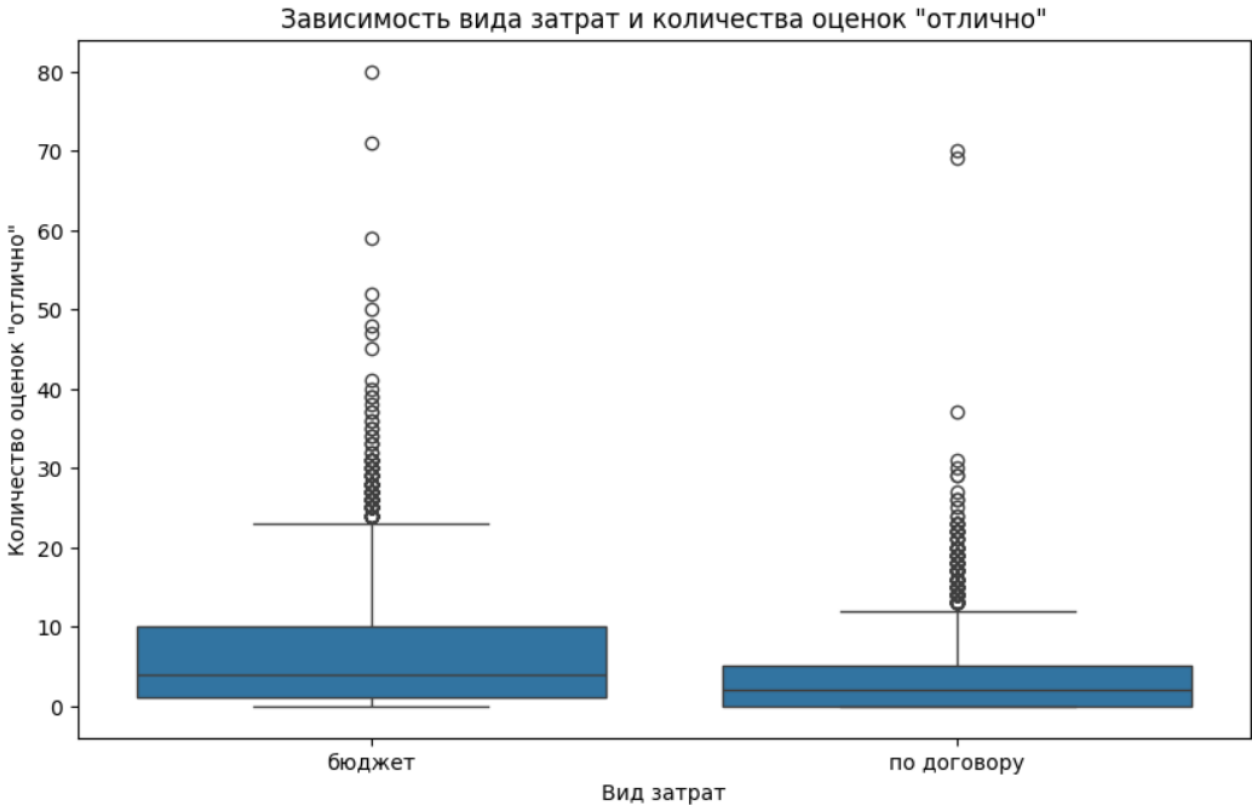


Рисунок 23 – Зависимость кол-во оценок «отлично» от «вида затрат»

На данном рисунке отчетливо видно, как студенты, которые учатся на бюджете, получают гораздо больше оценок «отлично», относительно студентов, которые учатся на платной основе.

На рисунке 24 изображен график, отображающий то, с каким средним баллом аттестата идут на каждую из форм обучения.

```
sns.boxplot(x='Форма освоения', y='Ср. балл док-та об образовании (док. студ.)', data=df)
plt.title('Зависимость среднего балла от формы освоения')
plt.show()
```

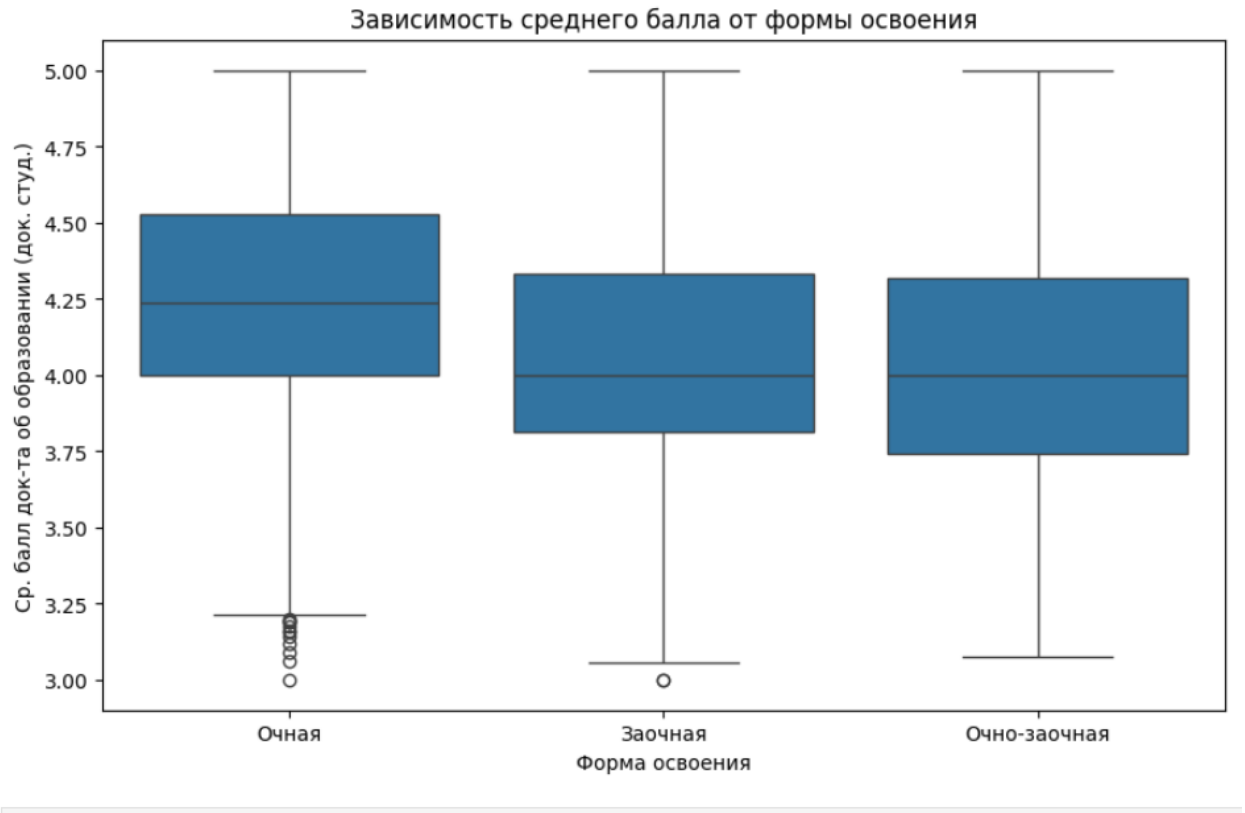


Рисунок 24 – Средний балл аттестата на каждой из форм освоения

На рисунке выше довольно ясно прослеживается, что на очную форму освоения студенты идут со средним баллом выше, чем по остальным формам.

График sns.scatterplot из библиотеки seaborn представляет собой инструмент визуализации, используемый для исследования взаимосвязи между двумя переменными. Вот основные особенности и характеристики этого графика:

1. Отображение точек: главная цель scatterplot - показать распределение точек данных по двум осям координат. Каждая точка представляет собой отдельное наблюдение из набора данных.
2. Простота интерпретации: график прост в интерпретации благодаря своей наглядности. Он позволяет быстро увидеть, есть ли какая-либо зависимость или корреляция между переменными.

3. Использование цвета и размера: возможность использовать дополнительные параметры для визуализации дополнительных измерений данных. Например, цвет и размер могут отражать дополнительные числовые или категориальные переменные, что позволяет добавить больше информации на графике.
4. Масштабирование и диапазон осей: важно учитывать, как масштабируются оси X и Y, чтобы не искажать визуальное восприятие данных.

На рисунке 25 изображен график, который помогает визуально оценить, есть ли какая-либо зависимость между количеством оценок "Отлично" и средним баллом студентов в зависимости от их пола. Также он позволяет сравнить распределение этих переменных среди разных групп студентов (по полу) на одном графике.

```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Отлично', y='Ср. балл док-та об образовании (док. студ.)', hue='Пол', data=df)
plt.title('Связь между средней успеваемостью и количеством оценок "Отлично" по полу')
plt.xlabel('Количество оценок "Отлично"')
plt.ylabel('Средний балл')
plt.legend(title='Пол')
plt.show()
```

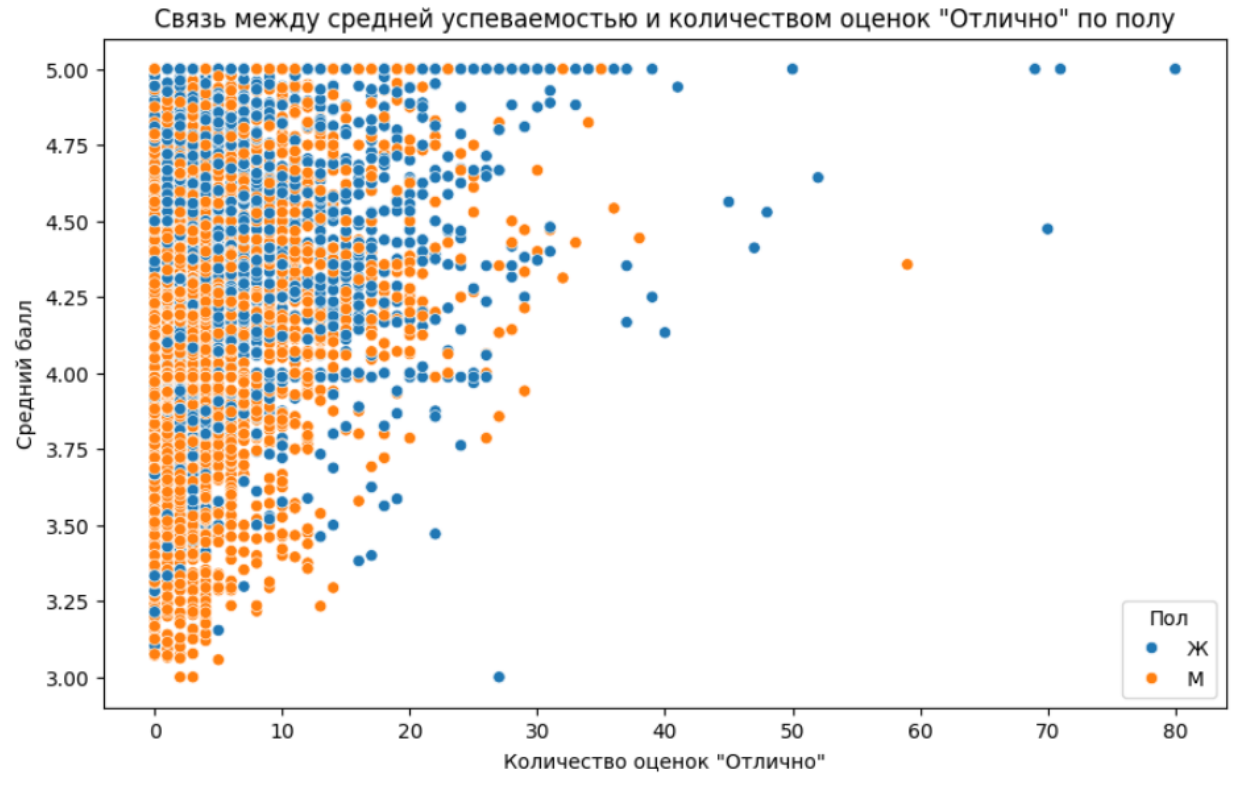


Рисунок 25 – Средний балл аттестата на каждой из форм освоения

На рисунке 26 изображена зависимость отчисления от формы освоения.

```
plt.legend(title='Статус', labels=['Отчислен', 'Не отчислен'])
plt.show()
```

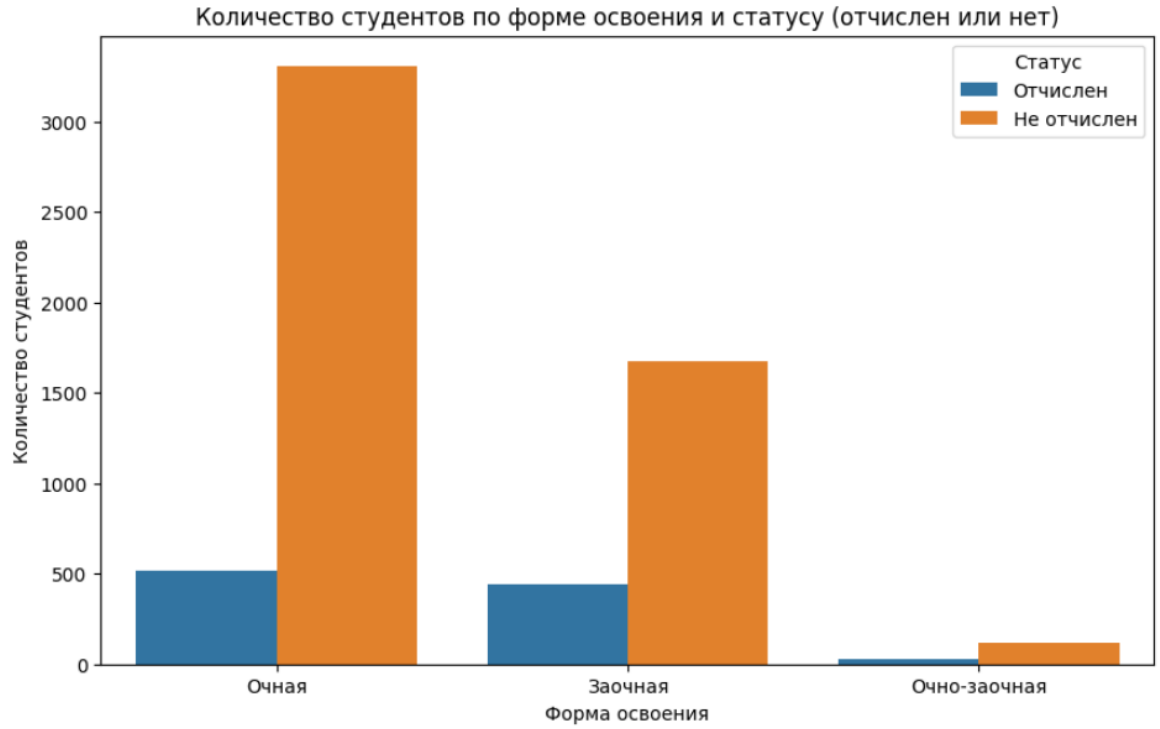


Рисунок 26 – Зависимость отчисления от формы освоения

На рисунке 27 показана соотношение мужчин и девушек ВУЗа.

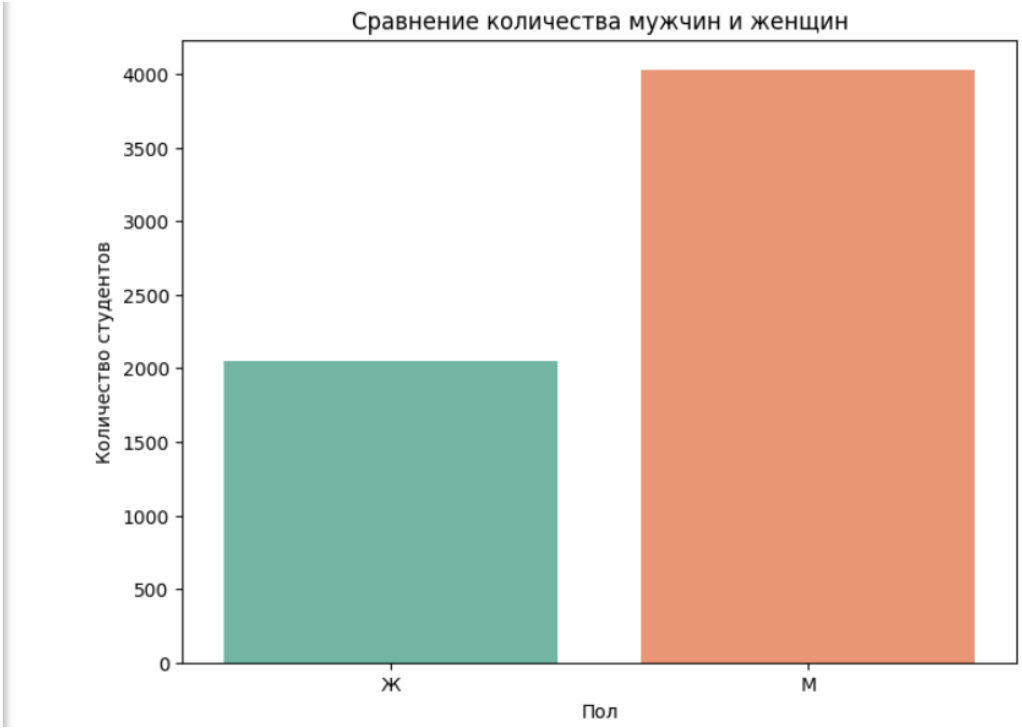


Рисунок 27 – Отношение мужчин и девушек в институте

3. РЕАЛИЗАЦИЯ ПРОЕКТА

3.1 Создание модели машинного обучения

Для обучения обязательно разделить выборку на тренировочные данные и данные, на которых будем тестировать точность модели. При разработке модели машинного обучения необходимо иметь отдельные наборы данных для обучения модели и для тестирования ее точности. Основные причины такого разделения:

1. Оценка качества модели: Использование отдельного набора данных для тестирования позволяет оценить, насколько хорошо модель обобщает знания на новых данных, которые она ранее не видела. Это помогает предотвратить переобучение (когда модель хорошо работает на тренировочных данных, но плохо обобщает на новые данные).
2. Выбор гиперпараметров: Оценка производится на тестовом наборе данных после настройки модели на тренировочных данных. Это позволяет выбирать лучшие гиперпараметры модели, минимизируя риск переобучения на тренировочных данных.
3. Объективная оценка: Тестовый набор данных представляет собой независимый контрольный экземпляр, который используется для объективной оценки реальной производительности модели.

Обычно данные разделяются в соотношении 70-30 или 80-20, где большая часть данных отводится под обучение, а меньшая — под тестирование.

Кросс-валидация — это метод оценки производительности модели, который позволяет получить более точную оценку качества модели, чем простое разделение на тренировочные и тестовые данные. Основная идея кросс-валидации заключается в том, чтобы делить данные на несколько частей (фолдов), обучать модель на одной части и оценивать ее на оставшихся частях.

С помощью кросс-валидации мы проверим несколько алгоритмов машинного обучения. Для этого воспользуемся массивом. На рисунке 28 изображен пример кода, в котором с помощью цикла выполняется перебор алгоритмов машинного обучения. На рисунке 29 изображена точность для каждого из алгоритмов на наших данных, а также стандартное отклонение.

```
for model in [
    LogisticRegression,
    DecisionTreeClassifier,
    KNeighborsClassifier,
    GaussianNB,
    SVC,
    RandomForestClassifier,
    XGBClassifier
]:
    cls = model()
    kfold = model_selection.KFold(n_splits=10, random_state = 42, shuffle=True)

    s = model_selection.cross_val_score(cls, X_processed, y, scoring='roc_auc', cv=kfold)
    print(f"{model.__name__:22} AUC: {s.mean():.3f} | STD: {s.std():.2f}")
```

Рисунок 28 – Код перебора алгоритмов машинного обучения

```
LogisticRegression     AUC: 0.871 | STD: 0.01
DecisionTreeClassifier AUC: 0.719 | STD: 0.02
KNeighborsClassifier   AUC: 0.821 | STD: 0.02
GaussianNB             AUC: 0.828 | STD: 0.01
SVC                    AUC: 0.874 | STD: 0.02
RandomForestClassifier AUC: 0.901 | STD: 0.01
XGBClassifier          AUC: 0.896 | STD: 0.02
```

Рисунок 29 – Точность и стандартное отклонение для каждого из алгоритмов

В данном случае AUC (Area Under the Curve) или площадь под кривой это метрика оценки качества бинарной классификации, которая измеряет общую производительность модели. Она представляет собой площадь под кривой ROC (Receiver Operating Characteristic), которая отображает зависимость между долей верных положительных классификаций (True Positive Rate, TPR) и долей ложных положительных классификаций (False

Positive Rate, FPR) при варьировании порога классификации. А стандартное отклонение измеряет разброс значений в выборке относительно их среднего значения. Оно показывает, насколько данные разбросаны относительно среднего значения. Чем выше стандартное отклонение, тем больше разброс данных, а чем меньше — тем ближе данные к среднему значению. В контексте оценки моделей машинного обучения через кросс-валидацию, стандартное отклонение играет важную роль:

1. Оценка устойчивости модели: Меньшее стандартное отклонение указывает на более устойчивую модель. Это означает, что модель показывает более однородные результаты на различных частях данных при кросс-валидации. Более высокое стандартное отклонение может свидетельствовать о том, что модель чувствительна к изменениям в данных и может менять свое поведение.
2. Точность оценки качества модели: Стандартное отклонение также отражает точность оценки качества модели. Маленькое стандартное отклонение означает, что результаты оценки (например, AUC в случае ROC-кривой) имеют меньшую дисперсию и более точно отражают реальное качество модели.

Именно по данным исследованиям было принято использовать алгоритм RandomForestClassifier в проекте.

Теперь перейдем к следующему этапу в создании модели машинного обучения. И это подбор гиперпараметров. Вот несколько ключевых причин, почему подбор гиперпараметров является важным этапом:

1. Оптимизация производительности модели: Гиперпараметры определяют внутренние настройки модели, которые не могут быть обучены самой моделью во время обучения. Например, в случае модели RandomForestClassifier, гиперпараметры включают количество деревьев (n_estimators) и глубину деревьев (max_depth). Подбор оптимальных значений этих параметров может существенно повлиять на качество и скорость работы модели.

2. Предотвращение переобучения или недообучения: Недостаточное или избыточное значение гиперпараметров может привести к переобучению или недообучению модели. Например, слишком большая глубина деревьев в модели случайного леса может привести к переобучению, когда модель слишком точно подстраивается под обучающие данные, но плохо обобщается на новые данные. Подбор оптимальных значений гиперпараметров помогает найти баланс между смещением и разбросом модели, что улучшает ее способность к обобщению.
3. Улучшение точности и обобщающей способности: Подбор гиперпараметров позволяет настроить модель таким образом, чтобы она достигала наилучшей возможной производительности на тестовых данных или новых наблюдениях. Оптимальные гиперпараметры обеспечивают максимальную точность предсказаний модели.
4. Экономия времени и ресурсов: Подбор гиперпараметров с использованием методов, таких как GridSearchCV или RandomizedSearchCV, позволяет автоматизировать процесс поиска оптимальных значений гиперпараметров. Это позволяет экономить время разработчика и вычислительные ресурсы, которые могут быть потрачены на ручной подбор.
5. Обоснованность выбора модели: Подбор гиперпараметров помогает систематически оценить различные конфигурации модели и выбрать ту, которая демонстрирует наилучшие результаты на основе заранее выбранных метрик оценки (например, точность, AUC).

В итоге, правильный подбор гиперпараметров позволяет создать модель машинного обучения, которая не только достигает высокой точности на обучающих данных, но и обладает хорошей способностью к обобщению на новые данные, что является ключевым аспектом в успешном

применении машинного обучения на практике. На рисунке 30 изображен код, с помощью которого мы перебираем варианты гиперпараметров.

```
from sklearn import ensemble

rf = ensemble.RandomForestClassifier()
params = {
    "max_features" : [0.1, 1, 2, 3, 4, 5, 6],
    "n_estimators" : [30, 100, 200, 300, 400],
    "criterion": ["gini", "entropy", "log_loss"],
    "max_depth": [3, 5, 7, 10],
    "min_samples_leaf": [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
    "min_samples_split": [2, 3, 5, 10],
    "random_state" : [42]
}
cv = model_selection.GridSearchCV(
    rf, params, n_jobs=-1).fit(X_train, y_train)
print(cv.best_params_)
```

Рисунок 30 – Код перебора гиперпараметров

В данном случае были подобраны такие гиперпараметры:

- max_depth = 10;
- min_samples_leaf = 1;
- n_estimators = 30;
- random_state = 42;
- criterion = 'gini';
- max_features = 6;
- min_samples_split = 3;

1. max_depth: максимальная глубина дерева решений. Ограничивает глубину дерева, чтобы предотвратить переобучение и улучшить обобщающую способность модели. На рисунке 31 показан пример глубины одного дерева со значением 4.

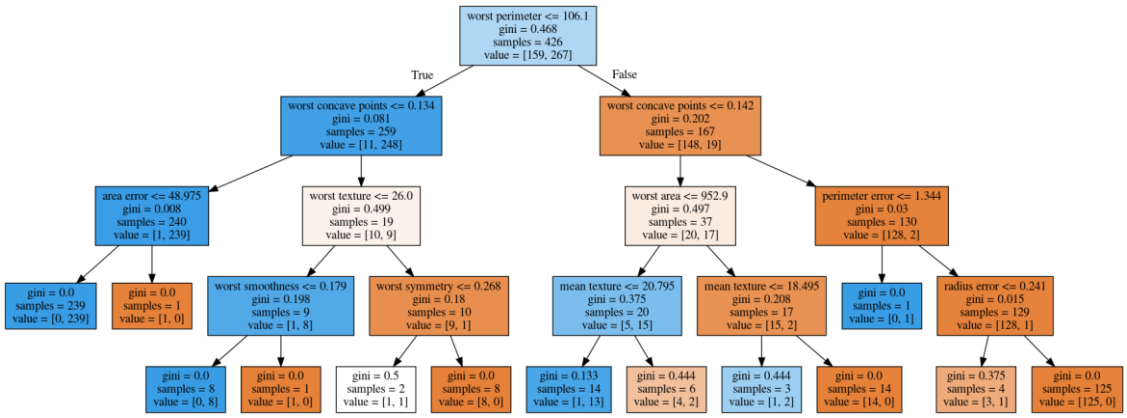


Рисунок 31 – Глубина одно дерева со значением 4

2. min_samples_leaf: минимальное количество образцов (наблюдений), требуемое для того, чтобы узел был листовым узлом. Этот параметр помогает предотвратить переобучение, устанавливая минимальный размер образцов в листовых узлах.
3. n_estimators: количество деревьев в ансамбле (например, случайного леса). Большее количество деревьев может улучшить обобщающую способность модели, но может увеличить время обучения и использовать больше ресурсов.
4. random_state: зерно случайности для воспроизводимости результатов. Устанавливает начальное состояние для случайной генерации чисел, что позволяет получать одинаковые результаты при каждом запуске модели.
5. criterion: функция для измерения качества разделения в дереве решений. 'gini' использует критерий Джини для измерения неоднородности узлов, что помогает оптимально разбивать данные.
6. max_features: максимальное количество признаков, рассматриваемых при разделении узла. Ограничивает количество признаков, что может улучшить скорость работы и уменьшить переобучение модели.
7. min_samples_split: минимальное количество образцов, необходимое для разделения внутреннего узла. Параметр контролирует, как далеко модель будет шагать в глубину перед тем, как разделить узел, что помогает предотвратить переобучение.

И теперь приступим к обучению нашей модели с использованием оптимальных гиперпараметров. На рисунке 32 отображается, точность модели, а также матрица ошибок.

```
rf2.fit(X_train, y_train)
```

```
RandomForestClassifier(max_depth=10, max_features=6, min_samples_split=3,
                       n_estimators=30, random_state=42)
```

```
rf2.score(X_test, y_test)
```

```
0.8898026315789473
```

```
from sklearn.metrics import confusion_matrix
y_pred = rf2.predict(X_test)
confusion_matrix(y_test, y_pred)
```

```
array([[ 111,  155],
       [  46, 1512]], dtype=int64)
```

Рисунок 32 – Код обучения и матрицы ошибок

Матрица ошибок (confusion matrix) в задачах машинного обучения используется для оценки качества работы алгоритма классификации. Она представляет собой таблицу, где строки представляют фактические классы, а столбцы предсказанные классы.

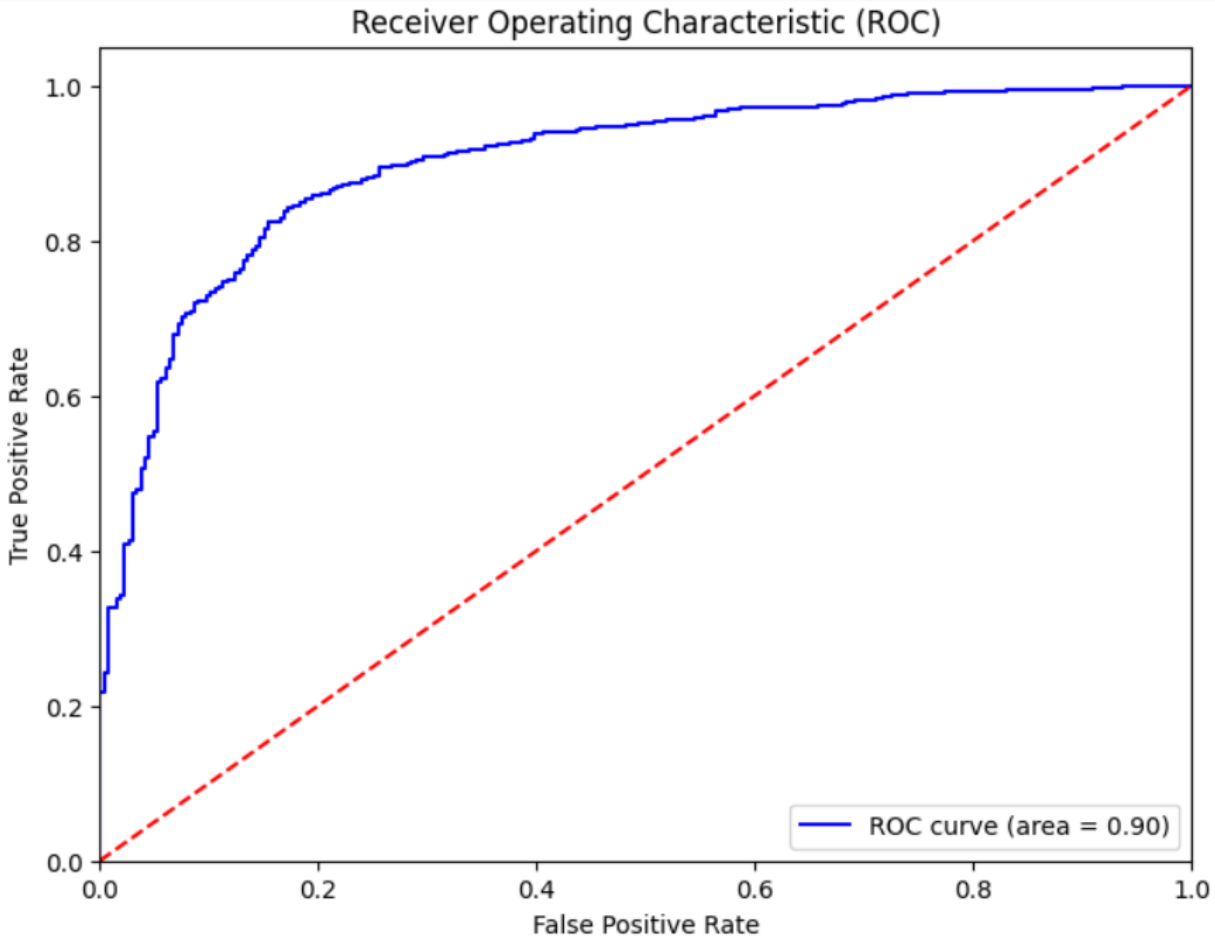
В матрице ошибок есть четыре основных элемента:

1. True Positive (TP): количество объектов, для которых истинный класс и предсказанный класс являются положительными.
2. True Negative (TN): количество объектов, для которых истинный класс и предсказанный класс являются отрицательными.
3. False Positive (FP): количество объектов, для которых истинный класс является отрицательным, а предсказанный класс — положительным (ошибка первого рода).

4. False Negative (FN): количество объектов, для которых истинный класс является положительным, а предсказанный класс — отрицательным (ошибка второго рода).

В этом случае 1623 студента идентифицированы были верно, однако 201 студента модель классифицировала неправильно.

Теперь построим график ROC-AUC, а также проверим результат метрики на тестовых данных. На рисунке 33 показан график ROC-AUC, а ниже результат в цифрах.




```
print(f'ROC AUC: {roc_auc:.2f}')
```

```
ROC AUC: 0.90
```

Рисунок 33 – график ROC-AUC.

Чем выше площадь под кривой, тем лучше производительность модели классификации.

На рисунке 34 и 35 информация о важности признаков при обучении.

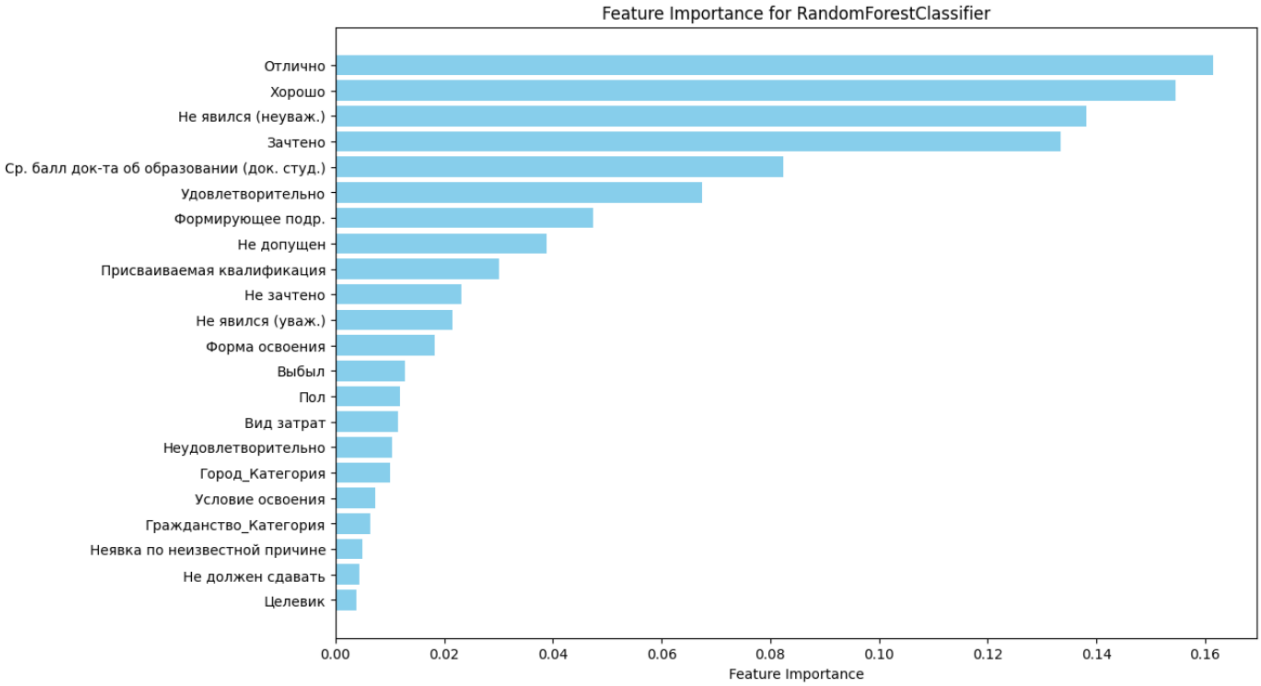


Рисунок 34 – График важности признаков

	Importance
Отлично	0.161417
Хорошо	0.154562
Не явился (неуваж.)	0.138106
Зачтено	0.133456
Ср. балл док-та об образовании (док. студ.)	0.082312
Удовлетворительно	0.067369
Формирующее подр.	0.047330
Не допущен	0.038805
Присваиваемая квалификация	0.030013
Не зачтено	0.023147
Не явился (уваж.)	0.021508
Форма освоения	0.018317
Выбыл	0.012795
Пол	0.011952
Вид затрат	0.011544
Неудовлетворительно	0.010350
Город_Категория	0.010093
Условие освоения	0.007313
Гражданство_Категория	0.006469
Неявка по неизвестной причине	0.004864
Не должен сдавать	0.004489
Целевик	0.003789

Рисунок 35 – Важность признаков

И последний этап – это сохранить модель с помощью библиотеки pickle. На рисунке 36 показан код сохранения модели.

```
[64]: import pickle
      with open('random_forest_model.pkl', 'wb') as f:
          pickle.dump(rf2, f)

      print("Модель сохранена в 'random_forest_model.pkl'")
```

```
Модель сохранена в 'random_forest_model.pkl'
```

Рисунок 36 – Сохранение модели.

3.2 Руководство пользователя

1. Введение:

   1.1. Область применения.

   Система предназначена для образовательных учреждений и научных исследовательских организаций, занимающихся анализом данных об успеваемости студентов. Она предоставляет возможности для загрузки, предобработки, анализа данных и получения предсказаний о результатах студентов на основе исторических данных. Данная система помогает улучшить качество обучения за счет анализа успеваемости студентов и выявления закономерностей в их учебной деятельности.

   1.2. Краткое описание возможностей.

   Система предоставляет следующие возможности:

   - загрузка данных: пользователи могут загружать данные об успеваемости студентов из файлов формата Excel (.xlsx);
   - предобработка данных: включает очистку данных, кодирование категориальных переменных и заполнение пропусков для обеспечения качества данных перед анализом;

- анализ данных: система позволяет выполнять анализ данных с использованием статистических методов и визуализаций для выявления тенденций и закономерностей в успеваемости студентов;
- интерфейс пользователя: интуитивно понятный графический интерфейс позволяет пользователям легко загружать данные, выполнять их анализ и получать предсказания;

1.3. Уровень подготовки пользователя.

Для эффективного использования системы пользователю требуется следующий уровень подготовки:

- основные навыки работы с компьютером: пользователь должен быть знаком с операционными системами Windows, macOS или Linux;
- навыки работы с Excel: пользователь должен уметь работать с файлами формата Excel (.xlsx) для загрузки данных в систему;

2. Назначение и условия применения:

2.1 Виды деятельности, функции.

Система предназначена для выполнения следующих видов деятельности:

- предобработка данных об успеваемости студентов: включает очистку данных, кодирование категорий и заполнение пропусков;
- взаимодействие с пользователем через графический интерфейс: позволяет загружать данные, выполнять их анализ и получать предсказания;

2.2 Программные требования к системе.

Для корректной работы системы необходимы следующие программные средства:

- операционная система: Windows, macOS или Linux;

- дополнительное ПО: Microsoft Excel или аналог для работы с .xlsx файлами;

2.3 Аппаратные требования к системе.

Для корректной работы системы необходимы следующие аппаратные средства:

- процессор: Intel i3 или эквивалентный;
- оперативная память: минимум 8 ГБ;
- свободное место на диске: минимум 500 МБ;

3. Подготовка к работе:

3.1 Запуск системы.

Для запуска системы выполните следующие действия:

- убедитесь, что все необходимые файлы готовы к работе;
- дважды щелкните по исполняемому файлу (.exe) для запуска программы;
- подождите, пока программа запустится;

3.2 Проверка работоспособности системы.

Для проверки работоспособности системы выполните следующие действия:

- в первую форму загрузите excel файл с оценками студентов;
- во вторую форму загрузите excel файл с информацией о студентах;
- в третью форму загрузите модель машинного обучения в формате (.pkl);
- получите предсказания и убедитесь в корректности результатов;

4. Описание операций:

4.1. В появившемся окне заполнить 3 формы, указанные выше. Рисунок 37 показывает окно программы до заполнения полей. Рисунок 38 изображенный ниже, уже с заполненными полями.

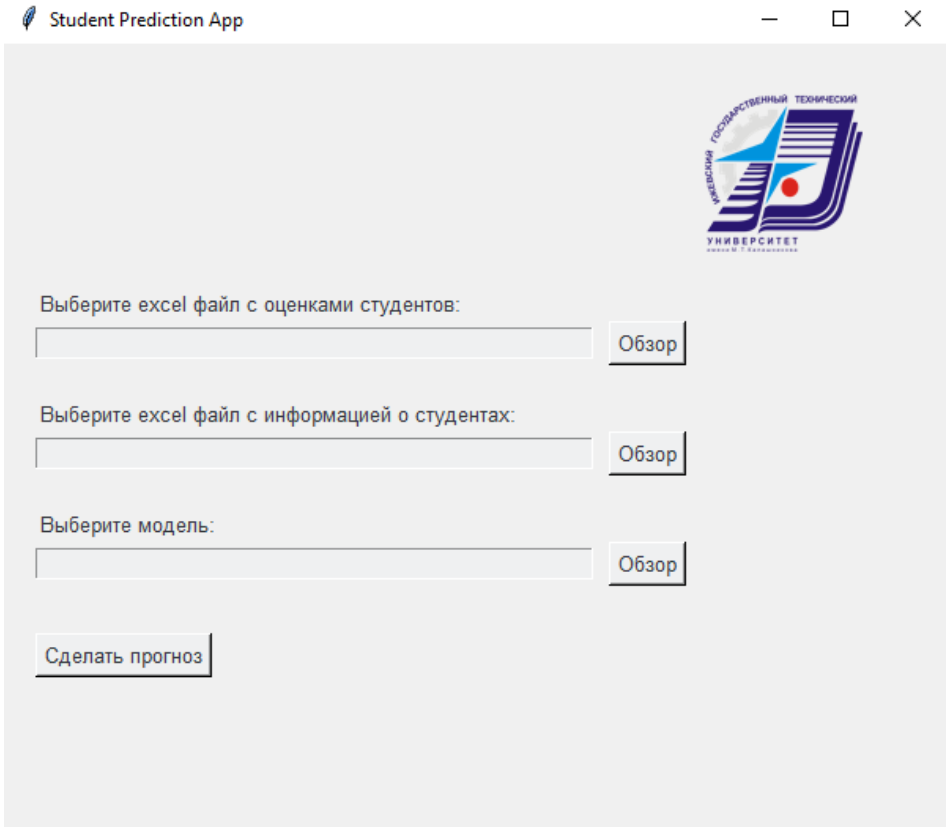


Рисунок 37 – Графический интерфейс программы до заполнения полей

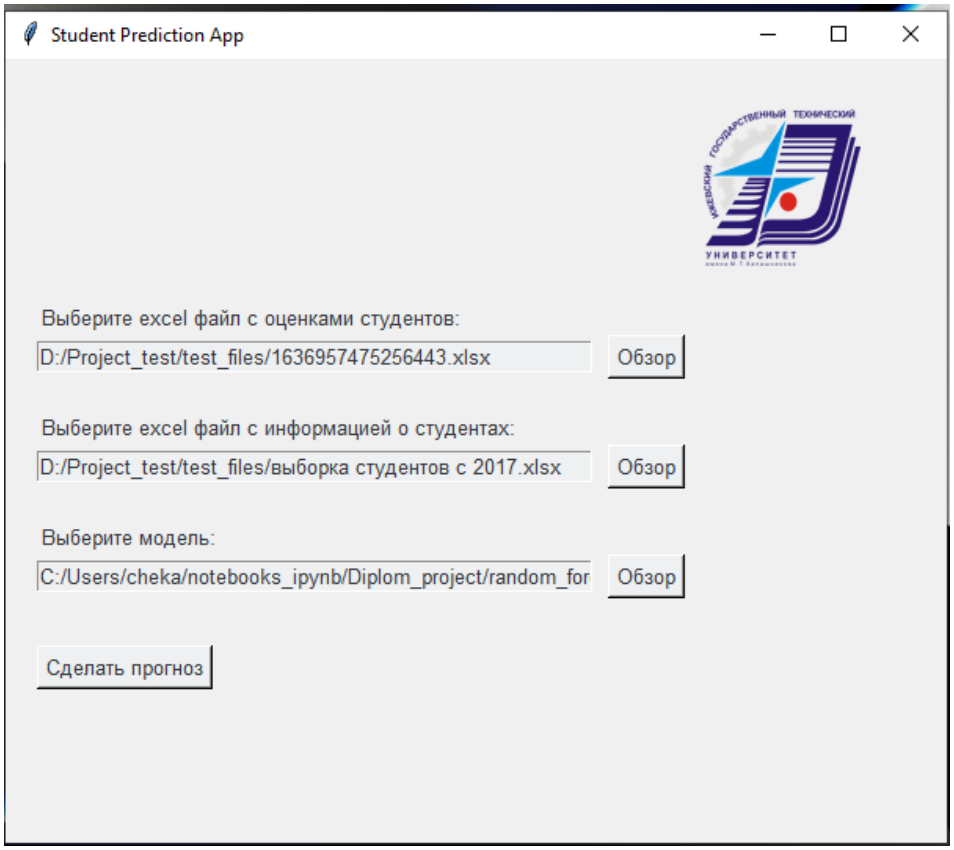


Рисунок 38 – Графический интерфейс программы с заполненными полями

4.2. Кнопка «Сделать прогноз». Данная функция выполняется около 5 минут. Когда программа будет выполнена, появится окно, изображенное на рисунке 39. Время выполнения – 5 минут.

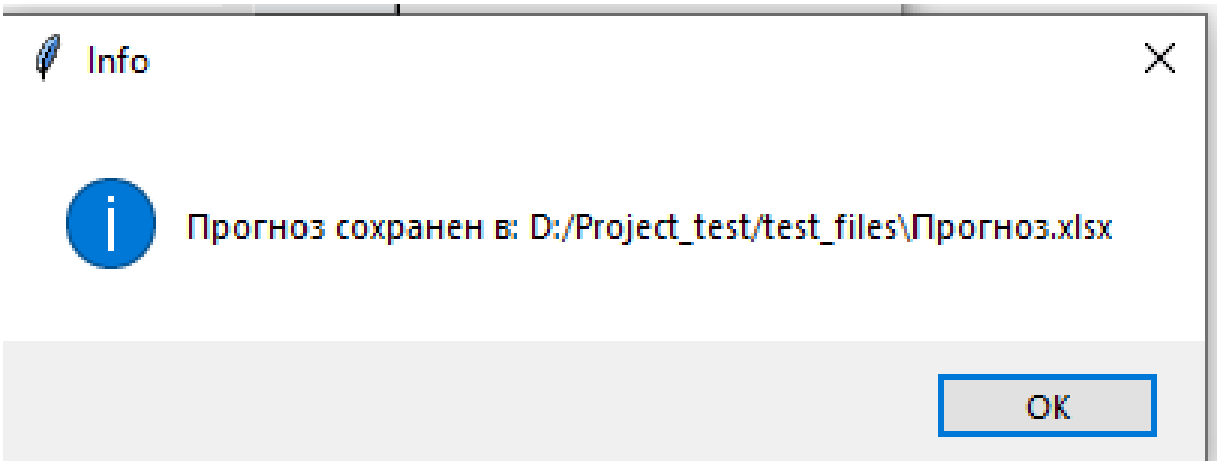


Рисунок 39 – Графический интерфейс программы с заполненными полями

4.3 Заключительный действия: файл с результатом выполнения программы будет создан и сохранен в ту директорию, где находился файл с оценками. Пример готового excel файла с прогнозом изображен на рисунке 40.

	A	B	C	D
1	student_number	Prediction		
2	20071489	1		
3	20092127	0		
4	17062039	1		
5	19022080	1		
6	20051308	1		
7	18061316	1		
8	21093041	1		
9	20051494	1		
10	19041369	1		
11	20072080	1		
12	19041317	1		
13	17382050	1		
14	17062089	1		
15	20061008	1		
16	20561010	1		
17	19061009	0		
18	19561027	1		
19				

Рисунок 40 – Пример excel файла с прогнозом

4.4 После завершения работы с программой необходимо закрыть, нажав на крестик. Время выполнения – 1 секунда.

5. Аварийные ситуации

5.1 Общие положения. При возникновении ошибок и сбоев в работе системы пользователи будут уведомлены о характере ошибки с помощью всплывающих окон. Эти уведомления помогут пользователям быстро идентифицировать и устранить проблемы, возникающие в процессе использования программы.

5.2 Возможные ошибки и их описание. Ниже приведены примеры возможных ошибок, которые могут возникнуть во время использования системы, а также описание каждой ошибки и рекомендации по их устранению:

- ошибка при обработке данных. Эта ошибка возникает, если система не может корректно обработать загруженные данные. Рекомендуемые действия: Убедитесь, что файлы данных соответствуют ожидаемому формату и содержат все необходимые данные. Проверьте целостность файлов и попробуйте загрузить их снова;
- пожалуйста, загрузите все файлы. Эта ошибка возникает, если пользователь не загрузил все необходимые файлы перед выполнением операции. Убедитесь, что все обязательные файлы загружены в систему. Проверьте, что файлы выбраны правильно и повторите попытку;
- ошибка при загрузке модели. Эта ошибка возникает, если система не может загрузить модель для предсказания. Проверьте, что файл модели существует и находится в указанном пути. Убедитесь, что файл модели не поврежден и соответствует ожидаемому формату;
- ошибка при предсказании. Эта ошибка возникает во время выполнения процесса предсказания, если система не может

корректно обработать данные или использовать модель. Проверьте правильность и целостность входных данных. Убедитесь, что модель загружена правильно и совместима с загруженными данными;

- ошибка при сохранении прогноза. Эта ошибка возникает, если система не может сохранить результаты предсказания в файл. Проверьте наличие достаточного места на диске для сохранения файла. Убедитесь, что путь для сохранения файла доступен и у вас есть необходимые права доступа для записи в этот путь;

5.3 Рекомендации по решению проблем. В случае возникновения ошибок и сбоев в работе системы, выполните следующие шаги для устранения проблемы:

- перезапуск системы: закройте приложение и запустите его снова. Это может помочь в случае временных сбоев;
- проверка файлов: убедитесь, что все файлы данных и модель загружены правильно и соответствуют ожидаемым форматам;
- свободное место на диске: убедитесь, что на диске достаточно свободного места для выполнения операций и сохранения результатов;
- права доступа: проверьте, что у вас есть все необходимые права доступа для чтения и записи файлов;
- обновление системы: убедитесь, что вы используете последнюю версию системы;

ЗАКЛЮЧЕНИЕ

В дипломной работе была разработана и внедрена модель машинного обучения для анализа цифровых следов студентов с целью предсказания вероятности их отчисления из высшего учебного заведения. Работа проведена в условиях активной цифровизации образовательного процесса, что позволило эффективно использовать большие объемы данных о студентах для повышения управляемости и качества образования.

В ходе работы были выполнены все поставленные задачи:

1. Исследованы цифровые следы, их разновидности, а также роль в образовании.
2. Рассмотрены возможности применения машинного обучения в образовательной среде.
3. Проанализирован рынок похожих продуктов, которые уже сейчас работают в высших учебных заведениях мира.
4. Определены технические требования для будущей программы.
5. Проведено рассмотрение и сравнение инструментов для реализации проекта.
6. Подготовлена визуализация данных, предоставленных для использования в проекте
7. Создана готовая модель машинного обучения, которую в будущем можно будет развивать и дообучать.
8. Подготовлена рабочая программа для деканатов ВУЗа, с помощью которой будет производится достоверный прогноз для каждого студента.
9. Разработано руководство пользователя.

СПИСОК ЛИТЕРАТУРЫ

1. Python For Beginners | Python.org. – URL: https://www.python.org/about/gettingstarted/ (дата обращения: 21.03.2024).

2. Язык программирования R: преимущества и недостатки. – URL: https://gb.ru/blog/yazyk-programmirovniya-r/ (дата обращения: 21.03.2024).

3. SAS Programming | Basics of SAS Programming Language | Edureka. – URL: https://www.edureka.co/blog/sas-programming/ (дата обращения: 22.03.2024).

4. The Python Standard Library – Python 3.12.3 documentation. – URL: https://docs.python.org/3/library/index.html (дата обращения: 24.03.2024).

5. PIP – менеджер пакетов в Python. Гайд по использованию. – URL: https://pythonchik.ru/okruzhenie-i-pakety/pip--menedzher-python-paketov (датаобращения: 25.03.2024).

6. Python | Модули. – URL: https://metanit.com/python/tutorial/2.10.php(дата обращения: 26.03.2024).

7. HeidiSQL Open Source Database Management Tool: MariaDB, MySQL, SQL Server. – URL: https://www.methodsandtools.com/tools/heidisql.php (дата обращения: 27.03.2024).

8. Типы данных SQL: какие бывают и как с ними работать. – URL: https://gb.ru/blog/tipy-dannykh-sql/ (дата обращения: 29.03.2024).

9. Буч, Г. Язык UML. Руководство пользователя. 2-е изд.: Пер. с англ. Мухин Н. / Г. Буч, Д. Рамбо, И. Якобсон. – Москва: ДМК Пресс, 2006– 496 с. – ISBN 5-94074-334-X. (дата обращения: 29.03.2024).

10. Что такое прототип: определение, функции, тонкости разработки. – URL: https://gb.ru/blog/chto-takoe-prototip/ (дата

обращения: 02.04.2024).

11. Почему Python такой популярный // TProger – сайт для программистов URL: https://tproger.ru/articles/pochemu-python-takoj-populjarnyj/ (дата обращения: 14.05.2024).

12. Топ 10 алгоритмов ML // Фонд развития онлайн-образования URL: https://eldf.ru/top10ml (дата обращения: 14.05.2024).

13. 19 полезных библиотек для Python // Хекслет URL: https://ru.hexlet.io/blog/posts/19-bibliotek-dlya-python (дата обращения: 14.05.2024).

14. Руководство по SQLite в Python // Python – скрипты, библиотеки, модули URL: https://pythonru.com/osnovy/sqlite-v-python (дата обращения: 14.05.2024).

15. Pip 24.0 // pypi URL: https://pypi.org/project/pip/ (дата обращения: 14.05.2024).

16. Scikit-learn Machine Learning in Python // scikit-learn URL: https://scikit-learn.org/stable/ (дата обращения: 14.05.2024).

17. Scikit-learn // github URL: https://github.com/scikit-learn/scikit-learn (дата обращения: 14.05.2024).

18. Seaborn: statistical data visualization // seaborn URL: https://seaborn.pydata.org/ (дата обращения: 14.05.2024).

19. Python Seaborn Tutorial For Beginners: Start Visualizing Data // datacamp URL: https://www.datacamp.com/tutorial/seaborn-python-tutorial (дата обращения: 14.05.2024).

20. Matplotlib Pyplot // w3schools URL: https://www.w3schools.com/python/matplotlib_pyplot.asp (дата обращения: 14.05.2024).

21. Proactively Impact Student Success // civitaslearning URL: https://www.civitaslearning.com/ (дата обращения: 05.05.2024).

22. Navigate360 Recruit, retain, and empower students with higher ed’s leading CRM, now powered by AI. // EAB URL:

https://eab.com/solutions/navigate360/ (дата обращения: 05.05.2024).

23. Starfish by Hobsons Expanded Impact on Higher Education Student Success During a Turbulent Year // PR newswire URL: https://www.prnewswire.com/news-releases/starfish-by-hobsons-expanded-impact-on-higher-education-student-success-during-a-turbulent-year-301192527.html (дата обращения: 04.05.2024).

24. Изучение и анализ данных с помощью Python // microsoft URL: https://learn.microsoft.com/ru-ru/training/modules/explore-analyze-data-with-python/ (дата обращения: 02.04.24).

25. Python и аналитика данных // itproger URL: https://itproger.com/news/python-i-data-analytics-chto-nuzhno-znat-i-primer-proekta (дата обращения: 02.04.24).

26. Экспресс-анализ данных на Python // habr URL: https://habr.com/ru/articles/729292/ (дата обращения: 02.04.24).

27. Машинное обучение в совершенствовании образовательной среды // cyberleninka URL: https://cyberleninka.ru/article/n/mashinnoe-obuchenie-v-sovershenstvovanii-obrazovatelnoy-sredy (дата обращения: 02.04.24).

28. 17 примеров применения машинного обучения в 5 отраслях бизнеса // cloud.vk URL: https://cloud.vk.com/blog/17-primerov-mashinnogo-obucheniya (дата обращения: 03.05.2024).

29. Машинное обучение для начинающих: основные понятия, задачи и сфера применения // proglib URL: https://proglib.io/p/mashinnoe-obuchenie-dlya-nachinayushchih-osnovnye-ponyatiya-zadachi-i-sfera-primeneniya-2021-08-29 (дата обращения: 03.05.2024).

30. Что такое машинное обучение? // sap URL: https://www.sap.com/central-asia-caucasus/products/artificial-intelligence/what-is-machine-learning.html (дата обращения: 03.05.2024).