Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий – РТФ

Школа бакалавриата

ОТЧЕТ

По проекту
«Разработка образовательных материалов и проектов в сфере Data Science»

по дисциплине «Проектный практикум»

Заказчик: Ильинский Александр Дмитриевич

Куратор: Ильинский Александр Дмитриевич

Студенты команды 21/ЛКП-4391-2025

Гиззатуллина Р.А.

Яруллина А.Ш.

Демидова И. В.

Екатеринбург, 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Работа с данными, прогнозирование и аналитика в современном мире являются неотъемлемой частью любых бизнес-процессов, научных и медицинских исследований, общественных и государственных реформ. Именно данные лежат в основе принятия стратегических решений, оптимизации ресурсов, улучшения качества пользовательского опыта и повышения конкурентоспособности компаний. Без глубокого анализа данных невозможно эффективно управлять компаниями, исследованиями и другими процессами, принимать обоснованные решения и разрабатывать стратегии развития.

Data Science решает данные вопросы, предоставляя инструменты и методы для извлечения ценной информации из больших объемов данных. Эта область позволяет не только понимать текущие тенденции, но и предсказывать будущие события, позволяя грамотно выстраивать стратегии.

В связи с широким распространением технологий искусственного интеллекта и машинного обучения в различных сферах жизни спрос на квалифицированных специалистов в этой области постоянно растет. Владение навыками Data Science открывает широкие перспективы для карьерного роста и участия в передовых проектах, направленных на решение актуальных задач в различных отраслях, от финансов и медицины до образования и государственного управления.

Целью данного учебного проекта является разработка практических проектов в области Data Science и формирование базовых компетенций, необходимых для успешной работы в этой сфере.

Для достижения поставленной цели необходимо выполнить следующие задачи:

1) изучить базовые теоретические основы машинного обучения;
2) приобрести практические навыки работы с данными, включая их сбор, очистку, анализ и визуализацию, создание моделей;

3) ознакомиться с основными инструментами и средами разработки для создания ML-решений. Например, библиотеки Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn в Python, платформы Jupyter Notebook, Google Colab;
4) решить несколько лабораторных работ, позволяющих применить на практике изученные ранее библиотеки;
5) принять участие в соревнованиях, в рамках которых реализовать несколько проектов на основе реальных данных, в рамках которых применить предобработку данных, ресерч-анализ, выбор наилучших моделей.

В результате выполнения проектов ожидается реализация двух продуктов: аналитического отчета по данным о пассажирах Титаника, решающего задачу предсказания выживания конкретного пассажира и содержащего гипотезы, новые признаки, простые модели машинного обучения и подробные выводы о полученных результатах, и аналитического отчета по данным о пользователях Spotify, решающего задачу предсказания популярности трека и содержащего гипотезы, новые признаки, модели различных семейств. Кроме того, предполагается освоение основных функций и принципов, применяющихся в работе в сфере Data Science.

Созданные в рамках изучаемой дисциплины проекты могут применяться для демонстрации практического применения методов Data Science в решении реальных задач, обобщенного анализа исторических данных, анализа реальных данных музыкального сервиса Spotify, рекомендации музыки пользователям, анализа музыкальных трендов.

1 Независимое выполнение лабораторных работ

1.1 Лабораторная работа №1. Основы Python и NumPy

В рамках образовательного трека «Разработка образовательных материалов и проектов в сфере Data Science» заказчиком были даны 4 лабораторные работы обязательные к выполнению перед переходом к следующему этапу – совместным проектам. Цели и задачи, поставленные в рамках лабораторных работ, были общими для всех участников и предоставляли возможность освоить необходимые для дальнейшей работы с данными инструменты.

Цель: удостовериться в знании Python или изучить его, ознакомиться с библиотекой NumPy и ее функциями.

Задачи:

1) решить ряд задач, направленных на понимание базовых концепций Python;
2) решить ряд задач, направленных на изучение работы с библиотекой NumPy;
3) загрузить решения на удаленный репозиторий GitHub.

Для выполнения данной лабораторной работы использовалась следующая литература: [1].

Каждый из участников выполнил все задания в ноутбуках, предоставленных заказчиком.

Было продемонстрировано, что участники понимают, как работать в Python, также были освоены основные функции библиотеки NumPy, а именно: усвоена работа с массивами, матрицами и математическими функциями NumPy.

1.2 Лабораторная работа №2. Pandas и корреляционный анализ

Цель: ознакомиться с библиотекой Pandas, разобраться с визуализацией и корреляционным анализом.

Задачи:

1) освоить библиотеку Pandas;
2) научиться подгружать данные из CSV файла и обрабатывать их;
3) обработать DataFrame;
4) изучить, как использовать Python для расчета корреляции;
5) создать диаграммы рассеяния;
6) построить матрицу корреляции при помощи Pandas;
7) визуализировать корреляционные данные;
8) загрузить решения на удаленный репозиторий GitHub.

Для выполнения данной лабораторной работы использовалась следующая литература: [2], [3].

Участниками были решены оба ноутбука в соответствии с условиями задач.

В результате выполнения лабораторной работы были усвоены основные функции библиотеки Pandas для работы с данными: DataFrame, чтение данных, фильтрация, сортировка, группировка и агрегация. Для работы с данными использовался датасет Titanic.

Проведен корреляционный анализ данных и визуализация с использованием Pandas Matplotlib. Использовался набор данных, содержащий 40 студентов-правшей с вводного курса по Психологии из университета Southwestern, прошедших 4 субтеста для расчета шкалы интеллекта взрослых по Векслеру (Приложение А).

1.3 Лабораторная работа №3. Линейная модель и градиентный спуск. Деревья, KNN

Цель: написать собственные классы линейных моделей, градиентный спуск для их обучения, провести их тестирование на примерах реальных данных.

Задачи:

1) дописать код для моделей;
2) загрузить данные из предоставленных файлов;
3) обработать данные;
4) подобрать оптимальные параметры обучения;
5) обучить модели;
6) проверить точность классификатора;
7) загрузить решения на удаленный репозиторий GitHub.

Участниками команды были выполнены все задачи, поставленные в данной лабораторной работе.

Для выполнения данной лабораторной работы использовалась следующая литература: [4].

В ноутбуке «Готовые модели» были обучены модели на датасете классификации из предоставленного ноутбука, сравнены результаты. На основе baseline были доделаны предсказывающие модели.

1.4 Лабораторная работа №4. Ансамбли, полносвязные нейронные сети.

Цель: решить задачи, описанные в параграфе 1.3 Лабораторная работа №3. Линейная модель и градиентный спуск. Деревья, KNN, используя ансамбли или полносвязные нейронные сети.

Задачи:

1) загрузить данные и обработать их;

2) подобрать оптимальные параметры обучения;

3) обучить модели;

4) проверить точность классификатора;

5) загрузить решения на удаленный репозиторий GitHub.

Для выполнения данной лабораторной работы использовалась следующая литература: [3], [5].

Все обученные модели (Random Forest, Gradient Boosting, MLP) показали хорошие результаты, достигнув ROC AUC выше 0.8 на тестовых данных.

На основе полученных результатов, для данной задачи можно рекомендовать Random Forest, как модель, показавшую наилучший результат.

2 Командное участие в соревнованиях

2.1 Соревнование «Титаник»

2.1.1 Постановка цели и общих задач

После лабораторных работ 1–4 началась командная работа, в ходе которой необходимо было выполнить два соревнования на данных, предоставленных заказчиком.

В рамках первого соревнования заказчик поставил перед участниками следующую цель: определить, выживет или нет пассажир на Титанике.

Для достижения поставленной цели были сформулированы общие задачи от заказчика, в дальнейшем более подробно разобранные командой и распределенные между ее участниками, проведен анализ требований, составлен календарный план.

Команда определила для себя следующие задачи:

1) провести предобработку данных, проверить их на выбросы;
2) построить матрицу корреляции;
3) составить гипотезы, основываясь на корреляции колонок с таргетом – выживаемостью;
4) создать новые признаки, провести корреляцию с ними;
5) построить графики и применить EDA для доказательства или опровержения гипотез;
6) провести эксперименты с моделями: линейные, деревья, модификации градиентного бустинга, нейронные сети;
7) выбрать лучшую модель и сделать итоговый вывод.

После того, как задачи были сформулированы, каждый из участников команды выбрал то, чем хотел бы заниматься в рамках данного соревнования. Таким образом, задачи были разделены внутри команды.

2.1.2 Анализ требований и составление календарного плана

Для успешного выполнения соревнования и анализа данных о пассажирах Титаника необходимо уметь применять функции из ряда библиотек: Python, Pandas, NumPy, Scikit-learn, Matplotlib. И обладать знаниями и навыками в: EDA, Feature Engineering, основах машинного обучения, основных моделях: линейные, деревья, градиентный бустинг, нейронные сети.

Для каждой модели должны быть проверены несколько наборов данных: тестовые (30% от имеющихся данных) и тренировочные (70% от имеющихся данных).

Конечное решение должно включать в себя EDA, графики, матрицу корреляции, ресерч-анализ, лучшую модель, итоговый вывод.

Итоговый вывод должен включать в себя обоснование выбора лучшей модели, описание наиболее влияющих на выживаемость пассажира признаков.

На основе поставленных задач и предъявленных требований был составлен календарный план (таблица 1) для своевременного выполнения соревнования.

Таблица 1 – Календарный план соревнования «Титаник»

Задача	Приоритет	Кол-во часов	Ответственный	Статус
Загрузка и обработка данных	Высокий	2	Демидова И. В.	Выполнено
Обработка пропущенных значений и выбросов	Высокий	2	Демидова И. В.	Выполнено
Создание новых признаков	Средний	2	Гиззатуллина Р.А.	Выполнено

Продолжение таблицы 1

EDA	Высокий	4	Гиззатуллина Р.А, Яруллина А.Ш.	Выполнено
Feature Engineering	Средний	4	Демидова И. В, Гиззатуллина Р. А, Яруллина А. Ш.	Выполнено
Описание промежуточных выводов	Высокий	2	Яруллина А.Ш.	Выполнено
Моделирование	Высокий	8	Демидова И. В.	Выполнено
Выбор лучшей модели	Высокий	4	Демидова И. В.	Выполнено
Написание отчета	Высокий	6	Яруллина А.Ш., Гиззатуллина Р.А.	Выполнено

2.1.3 Вклад участника 1: Гиззатуллина Регина Айратовна

Перед участником 1 стояли следующие задачи:

1) сформулировать гипотезы, не зависящие напрямую от выживаемости, для выявления дополнительных признаков;
2) создать новые признаки;
3) провести корреляцию новых признаков с таргетом – выживаемостью;
4) помочь в экспериментах с моделями машинного обучения для выявления лучшей модели;
5) помочь в сборе и структурировании информации для написания отчета по текущей итерации.

Была выполнена работа по EDA и ресерч-анализу. Составлены различные гипотезы для определения направления исследования, созданы новые признаки, построены графики на основе некоторых гипотез для

изучения распределения признаков, собрана информация у участников команды об их успехах и сложностях в работе.

В ходе работы участник столкнулся с некоторыми сложностями при создании новых признаков и корреляции их с таргетом: многие признаки оказывались избыточными или имели низкое влияние на таргет, возникали противоречия. Например, при попытках выстроить зависимость между классом и стоимостью билета часто возникала такая проблема, что у пассажиров третьего класса был билет стоимость выше среднего.

Для решения возникших проблем был проведен углубленный анализ и проверка данных на выбросы.

2.1.4 Вклад участника 2: Яруллина Алина Шамилевна

Перед участником 2 стояли следующие задачи:

1) сформулировать основные гипотезы, не учитывающие новые признаки;
2) построить графики на основе данных, необходимых для подтверждения или опровержения гипотез;
3) проанализировать полученные результаты и написать подробные выводы для каждой из гипотез;
4) помочь в экспериментах с моделями машинного обучения для выявления лучшей модели;
5) написать отчет по текущей итерации.

Была выполнена работа по EDA и ресерч-анализу. Составлены первичные гипотезы, построены графики на основе приведенных гипотез, проанализированы имеющиеся данные и сформулированы выводы, написана часть отчета.

Сложность задач участника заключалась в грамотной и понятной визуализации данных для наиболее точного анализа. Многие гипотезы требовали рассмотрения больших групп пассажиров по некоторым признакам.

Для одной гипотезы могло потребоваться несколько различных графиков, учитывающих один признак. Например, при рассмотрении зависимости пола и возраста от выживаемости пассажиры были разделены на несколько категорий: мужчины, женщины, дети, женщины 0–30 лет, мужчины 0–30 лет, все пассажиры 0–30 лет, женщины 30–70 лет, мужчины 30–70 лет, все пассажиры 30–70 лет.

Для решения возникшей сложности потребовалось потратить больше времени и лучше структурировать данные для получения лучших выводов.

2.1.5 Вклад участника 3: Демидова Ирина Вадимовна

Перед участником 3 стояли следующие задачи:

1) загрузить и обработать данные;
2) описать модели машинного обучения и подобрать лучшие параметры;
3) выбрать лучшую модель для заданной задачи.

Была выполнена загрузка данных и их обработка: удалены выбросы и незначащие столбцы, заменены некорректные значения (Sex, Embarked) на числовые, добавлены новые признаки, удалены незначащие или малозначащие признаки, написаны и обучены на основных и тестовых данных основные модели машинного обучения: линейная, деревья решений, градиентный бустинг, нейронные сети, выбрана лучшая из них – линейная модель.

В ходе работы у участника возникали сложности с настройкой гиперпараметров моделей машинного обучения. Для решения проблемы были применены методы автоматической настройки гиперпараметров: Gird Search, библиотеки оптимизации (GirdSearchCV, scikit-learn), оценка производительности каждой комбинации гиперпараметров на кросс-валидации для выбора наилучшей конфигурации.

2.2 Соревнование «Spotify»

2.2.1 Постановка цели и общих задач

В рамках второго соревнования заказчик поставил перед участниками следующую цель: спрогнозировать, от чего будет зависеть популярность трека в Spotify.

Для достижения поставленной цели были сформулированы общие задачи от заказчика, в дальнейшем более подробно разобранные командой и распределенные между ее участниками, проведен анализ требований, составлен календарный план.

Команда определила для себя следующие задачи:

1) провести загрузку и предобработку данных, проверить их на выбросы;
2) построить матрицу корреляции;
3) составить гипотезы, основываясь на корреляции колонок с таргетом – популярностью;
4) создать новые признаки, провести корреляцию с ними;
5) построить графики и применить EDA для доказательства или опровержения гипотез;
6) описать модели машинного обучения: простая, линейная, деревья, градиентный бустинг, нейронные сети;
7) оценить кросс-валидацию для каждой модели;
8) выбрать лучшую модель;
9) подвести вывод.

После того, как задачи были сформулированы, каждый из участников команды выбрал то, чем хотел бы заниматься в рамках данного соревнования. Таким образом, задачи были разделены внутри команды.

2.2.2 Анализ требований и составление календарного плана

Для успешного выполнения соревнования и прогнозирования популярности треков в Spotify необходимо уметь применять функции из ряда библиотек: Python, Pandas, NumPy, Scikit-learn, Matplotlib. И обладать знаниями и навыками в: EDA, Feature Engineering, основах машинного обучения, основных моделях: линейные, деревья, градиентный бустинг, нейронные сети.

Для каждой модели должны быть проверены несколько наборов данных: тестовые (30% от имеющихся данных) и тренировочные (70% от имеющихся данных).

Конечное решение должно включать в себя EDA, графики, матрицу корреляции, ресерч-анализ, лучшую модель, итоговый вывод.

Итоговый вывод должен включать в себя обоснование выбора лучшей модели, описание наиболее влияющих на популярность трека признаков.

На основе поставленных задач и предъявленных требований был составлен календарный план (таблица 2) для своевременного выполнения соревнования.

Таблица 2 – Календарный план соревнования «Spotify»

Задача	Приоритет	Кол-во часов	Ответственный	Статус
Загрузка и обработка данных	Высокий	2	Демидова И. В., Гиззатуллина Р.А.	Выполнено
EDA	Высокий	9	Демидова И. В, Гиззатуллина Р. А., Яруллина А. Ш.	Выполнено

Продолжение таблицы 2

Обработка пропущенных значений и выбросов	Средний	1	Демидова И. В.	Выполнено
Feature Engineering	Средний	4	Демидова И. В., Яруллина А. Ш.	Выполнено
Моделирование	Высокий	10	Демидова И. В., Гиззатуллина Р.А.	В процессе
Выбор лучшей модели	Высокий	4	Демидова И. В., Яруллина А.Ш.	В процессе
Написание отчета	Высокий	6	Яруллина А. Ш., Гиззатуллина Р.А.	В процессе

2.2.3 Вклад участника 1: Гиззатуллина Регина Айратовна

Перед участником 1 стояли следующие задачи:

1) загрузить данные, проанализировать имеющиеся признаки и создать новые;
2) сформулировать гипотезы о популярности треков;
3) провести корреляцию новых признаков с таргетом – популярностью;
4) визуализировать данные;
5) помочь в подборе гиперпараметров моделей машинного обучения;
6) помочь в сборе и структурировании информации для написания отчета по текущей итерации.

Была выполнена загрузка данных, созданы новые признаки и проведена их корреляция с таргетом, сформулированы гипотезы. Были построены графики для дальнейшего анализа зависимостей между признаками другими участниками команды. С помощью автоматической настройки гиперпараметров были подобраны и записаны лучшие параметры для

используемых моделей. Была собрана и структурирована информация о работе каждого участника, успехах и возникших сложностях.

Как и при первом соревновании большую часть времени участник потратил на анализ признаков и их визуализацию. Поскольку датасет, предоставленный для соревнования «Spotify», содержал намного больше признаков, чем датасет «Титаник», то и времени на обработку данных, составление графиков, исключение ненужных колонок и гипотез ушло намного больше.

Планируется провести больше экспериментов с гиперпараметрами моделей машинного обучения.

2.2.4 Вклад участника 2: Яруллина Алина Шамилевна

Перед участником 2 стояли следующие задачи:

1) построить графики на основе данных, необходимых для подтверждения или опровержения гипотез;
2) проанализировать полученные результаты и написать подробные выводы для каждой из гипотез;
3) выбрать лучшую модель и обосновать сделанный выбор;
4) сделать итоговый вывод;
5) написать отчет по текущей итерации.

Были построены и проанализированы графики различных принципов и их зависимостей, подтверждены и опровергнуты некоторые гипотезы на основе сделанных выводов, выбрана лучшая модель по результатам кросс-валидации – нейронная сеть, написана часть отчета, связанная с соревнованием «Spotify».

Планируется провести больше экспериментов с моделями на различных данных и поработать над выводами: дополнить их, описать зависимости. Кроме того, необходимо закончить главы отчета, содержащие информацию об архитектуре программного продукта и методологии.

2.2.5 Вклад участника 3: Демидова Ирина Вадимовна

Перед участником 3 стояли следующие задачи:

1) загрузить данные, обработать пропущенные значения и избавиться от выбросов;
2) сформулировать гипотезы;
3) описать модели машинного обучения и подобрать лучшие параметры;
4) определить параметры для выбора лучшей модели машинного обучения для заданной цели;
5) сделать итоговый вывод.

Были загружены данные, обработаны пропущенные значения и выбросы, сформулированы гипотезы о влиянии различных признаков на популярность трека, описаны модели машинного обучения: простая, линейная, дерево, градиентный бустинг, нейронные сети, для каждой модели выделены лучшие параметры и проведена кросс-валидация, найдена лучшая модель.

Планируется добавить новые признаки для более глубокого анализа, подвести итоговый вывод, учитывающий новые признаки и возможные новые зависимости.

3 Архитектура программного продукта

Так как основная задача данного проекта – это участие в соревнованиях по анализу данных, где нужно проводить EDA, Feature Engineering и создавать модели для предсказания целевой переменной, архитектура продукта представляет собой скорее организацию кода и самого рабочего процесса, чем архитектуру ПО.

Таким образом, проект организован в виде набора Jupyter Notebooks с разделением на этапы (Приложение А).

Каждое соревнование представляет из себя файл с расширением ipynb, который имеет определенную структуру, состоящую из ячеек кода и блоков комментариев, разделенных на следующие модули:

1) загрузка данных:

— загрузка данных их CSV файлов;

— проверка структуры данных, типов данных;

— проверка пропущенных значений.

2) EDA - исследовательский анализ данных:

— модуль для проведения EDA;

— код для построения графиков;

— вычисление статистических характеристик;

— написание комментариев и выводов, полученных в результате анализа.

3) Feature Engineering:

— модуль для создания новых признаков на основе существующих.

4) Обучение и оценка моделей:

— модуль для обучения моделей машинного обучения;

— обучение различных моделей;

— оценка производительности моделей.

5) Выбор модели:

— Выбор лучшей модели на основе результатов кросс-валидации.

Компоненты взаимодействуют последовательно, используя входные данные одного компонента в качестве входных данных для другого.

Для реализации данного проекта был использован следующий технологический стек:

1) язык программирования – Python;
2) основные библиотеки:

— Pandas. Для обработки и анализа структурированных данных;

— NumPy. Для выполнения численных расчетов и операций с массивами;

— Scikit-learn. Для построения и оценки моделей машинного обучения;

— Matplotlib. Для визуализации данных и результатов анализа.

3) инструменты:

— Jupyter Notebook. Для интерактивной разработки и удобного и понятного документирования кода.

— Google Colab. Для совместного написания кода и его выполнения в облачной среде.

— GitHub. Для контроля версий.

Архитектура проекта построена на следующих принципах:

1. Модульность. Код разделен на отдельные модули для удобства разработки и отладки.
2. Документирование. Каждый модуль содержит подробные комментарии и описание, что облегчает понимание кода и его дальнейшее сопровождение.
3. Совместная работа.

Основной целью архитектуры является создание структурированного и воспроизводимого документа машинного обучения, который позволит эффективно исследовать данные, разрабатывать новые признаки, обучать и оценивать модели, а также обеспечивать воспроизводимость полученных результатов.

4 Методология разработки и процесс разработки

В основу разработки проектов в рамках соревнований был положен экспериментальный подход, предполагающий выдвижение гипотез о влиянии различных признаков и моделей на качество прогнозирования. Данные гипотезы последовательно проверялись посредством проведения экспериментов.

Для организации процесса разработки использовался Agile-подход, характеризующийся итеративной структурой работы. Работа была разбита на короткие итерации, в рамках которых команда проводила регулярные встречи для обсуждения достигнутого прогресса, выявления возникающих проблем и координации дальнейших действий.

Процесс разработки включал следующие этапы:

1. Планирование итерации. На данном этапе определялись цели и задачи текущей итерации, производилось распределение задач между участниками и оценивалось время, необходимое для их выполнения.
2. Выполнение задач. На этом этапе выполнялись задачи, определенные на этапе планирования. Для контроля версий и обеспечения совместной работы использовалась среда Google Colab и GitHub.
3. Тестирование и оценка. Данный этап включал проверку корректности кода и результатов экспериментов, оценку производительности разработанных моделей и анализ возникающих ошибок.
4. Рефлексия. На заключительном этапе проводился анализ результатов итерации, определялись возможности для улучшения процесса разработки и осуществлялось планирование следующей итерации.

Для управления процессом разработки использовались следующие инструменты: GitHub (для хранения итогового кода), Google Colab (для контроля версий, совместной работы и отслеживания изменений) и Telegram (для оперативной связи и обмена информацией между участниками команды).

ЗАКЛЮЧЕНИЕ

В ходе выполнения данного проекта был проведен ряд лабораторных работ, направленных на изучение основных инструментов для разработки проектов в сфере Data Science. Освоены навыки, необходимые для обработки и анализа данных и создания и обучения моделей, что позволило сформировать базовые компетенции в этой области. Результатом работы стала реализация двух проектов: прогнозирования выживаемости пассажиров Титаника и популярности треков на платформе Spotify. В каждом из проектов успешно применены методы EDA, Feature Engineering, а также осуществлен подбор гиперпараметров для достижения наилучших результатов моделей, что демонстрирует эффективное применение знаний, полученных на этапе выполнения лабораторных работ.

Разработанные в рамках соревнований проекты в целом соответствуют поставленным требованиям заказчика: составлены гипотезы, отражены зависимости с целевой переменной (таргетом), обучены и выбраны наиболее подходящие модели для решения поставленных задач. Несмотря на достигнутые результаты, в процессе разработки были выявлены некоторые ограничения, связанные с недостаточностью признаков и неочевидными зависимостями. Для преодоления ограничений был проведен дополнительный анализ, благодаря которому были добавлены новые признаки и выявлены дополнительные зависимости. Реализация описанных функций позволила улучшить качество и точность проведенного анализа.

Кроме того, в процессе тестирования разработанных моделей были выявлены ошибки, связанные с некорректной обработкой выбросов. Модели показывали нереалистичные результаты или вовсе не могли обработать данные из-за некорректного типа значений. Решить возникшую проблему удалось, проанализировав имеющиеся данные и разработав более эффективный метод обработки выбросов, повышающий устойчивость моделей к аномальным данным.

Для дальнейшего улучшения уже созданных решений можно скорректировать и дополнить их следующим образом:

1. Добавить больше комплексных гипотез, охватывающих большее количество признаков и тем самым открывающих возможность увидеть новые зависимости.
2. Исследовать возможность использования ансамблевых методов машинного обучения для повышения точности прогнозирования.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Python for Data Science: Практика Numpy [Электронный ресурс] // Stepik. – URL: https://stepik.org/course/189476/promo (дата обращения: 29.03.2025).

2. Модуль Pandas [Электронный ресурс] // Яндекс Образование. – URL: https://education.yandex.ru/handbook/python/article/modul-pandas (дата обращения: 09.04.2025).

3. Рашка С. Python и машинное обучение [Электронный ресурс] / Себастьян Рашка, Вахид Мирджалили ; пер. с англ. Ю.Н. Артеменко. – СПб. : ООО «Диалектика», 2020. – 848 с. – С. 145-185, 273-313. – URL: https://psv4.userapi.com/s/v1/d/7S9_E2Y3VlvwfVKnOZD2libzl8m6v-4iPrLp3cyGp0rEw3Ma-6p9qWv_TkUGmkDhdhCJ5-sDYa32lh40h0MPyfC5auRBFFkk1Ex9Td56NqYt-xRcLC8qwQ/Python_i_mashinnoe_obuchenie_2020_Rashka_Mirdzhalili.pdf (дата обращения: 16.04.2025).

4. Линейные модели [Электронный ресурс] // Яндекс Образование. – URL: https://education.yandex.ru/handbook/ml/article/linear-models (дата обращения: 16.04.2025).

5. Николенко С. Глубокое обучение. Погружение в мир нейронных сетей [Электронный ресурс] / Сергей Николенко, Артем Кадурин, Евгения Архангельская. – СПб.: Питер, 2018. – 480 с. – С. 138-142. – URL: https://psv4.userapi.com/s/v1/d/rhkn7ayv38gxW_GcLhy4wKScBTBKYX0ACXjXGeTNa4X_FFJQ4A1Vo7R9mzAh144-GonNsVd4rXI_RpPZvAVV7BSPwgl351qTEQ-PNM2VOcVYl-WcfV2AiA/Glubokoe_obuchenie_pogruzhenie_v_mir_neyronnykh_setey_pdf.pdf (дата обращения: 18.04.2025).

ПРИЛОЖЕНИЕ А

(справочное)

Первые 15 строк датасета для лабораторной работы №3

Gender	FSIQ	VIQ	PIQ	Weight	Height	MRI_Count
Female	133	132	124	118	64.5	816932
Male	140	150	124	NA	72.5	1001121
Male	139	123	150	143	73.3	1038437
Male	133	129	128	172	68.8	965353
Female	137	132	134	147	65.0	951545
Female	99	90	110	146	69.0	928799
Female	138	136	131	138	64.5	991305
Female	92	90	98	175	66.0	854258
Male	89	93	84	134	66.3	904858
Male	133	114	147	172	68.8	955466
Female	132	129	124	118	64.5	833868
Male	141	150	128	151	70.0	1079549
Male	135	129	124	155	69.0	924059
Female	140	120	147	155	70.5	856472
Female	96	100	90	146	66.0	878897
Female	133	132	124	118	64.5	816932
Male	140	150	124	NA	72.5	1001121
Male	139	123	150	143	73.3	1038437
Male	133	129	128	172	68.8	965353
Female	137	132	134	147	65.0	951545