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:
Халдин Д.А.

Екатеринбург, 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Современный мир стремительно движется в сторону цифровизации и автоматизации процессов, что приводит к росту спроса на специалистов в области Data Science. Однако качественное обучение в этой сфере требует не только теоретических знаний, но и практических навыков работы с реальными данными, инструментами анализа и алгоритмами машинного обучения. В связи с этим разработка образовательных материалов и проектов в сфере Data Science становится ключевым фактором подготовки квалифицированных кадров.

1. Цель и задачи проекта

Основной целью данного проекта является создание образовательных материалов и практических заданий, направленных на освоение ключевых аспектов Data Science: от базовых библиотек для анализа данных до сложных алгоритмов машинного обучения.

В рамках проекта решались следующие задачи:

1) Освоение и демонстрация работы с библиотеками анализа данных (NumPy, Pandas) – выполнение лабораторных работ по обработке и исследованию данных.

2) Изучение методов машинного обучения – реализация линейных моделей, градиентного спуска, ансамблевых методов и нейронных сетей.

3) Применение знаний на практике – участие в соревнованиях Kaggle (Titanic, Spotify) для решения задач классификации и прогнозирования.

2. Актуальность и важность проекта

Data Science – одна из самых востребованных областей в IT, применяемая в финансах, медицине, маркетинге и многих других сферах. Однако недостаток структурированных практических материалов затрудняет обучение начинающих специалистов. Данный проект позволяет:

– систематизировать знания по ключевым инструментам Data Science;

– предоставить готовые учебные материалы с примерами кода и разборами задач;

– сформировать навыки решения реальных задач через участие в соревнованиях.

3. Область применения

Разработанные материалы могут быть использованы:

– образовательных курсах по Data Science и машинному обучению;

– для самостоятельного изучения основ анализа данных;

– как основа для дальнейшего углубленного изучения сложных моделей ML и AI.

4. Ожидаемые результаты и планируемые достижения

По итогам проекта были выполнены:

– 4 лабораторные работы, охватывающие NumPy, Pandas, линейные модели, градиентный спуск, ансамбли и нейронные сети;

– два соревнования Kaggle (Titanic, Spotify), позволившие применить изученные методы на реальных данных.

В результате проекта были закреплены навыки обработки данных, построения моделей машинного обучения и их оптимизации, а также создана база учебных материалов, которая может быть полезна как начинающим, так и более опытным специалистам в области Data Science.

ОСНОВНАЯ ЧАСТЬ

1. Организация учебного процесса и выполнение лабораторных работ

В рамках проекта обучение строилось на основе структурированных учебных материалов, которые включали теоретические основы и практические задания. Для каждой лабораторной работы устанавливался четкий дедлайн, что позволяло систематизировать процесс изучения и контролировать прогресс.

Выполненные работы:

1) Лабораторная работа №1 (NumPy): изучение основных операций с массивами, методов линейной алгебры и генерации данных.

2) Лабораторная работа №2 (Pandas и корреляции): освоение обработки табличных данных, фильтрации, агрегации и анализа взаимосвязей между признаками.

3) Лабораторная работа №3 (Линейная модель и градиентный спуск. Деревья, KNN): реализация логистической регрессии и дерева решений, затем и градиентного бустинга и knn, исследование методов оптимизации.

4) Лабораторная работа №4 (Ансамбли/нейронные сети): знакомство с алгоритмами Random Forest, Gradient Boosting и построение простых нейросетевых архитектур MLP.

После выполнения лабораторных работ были проведены два соревнования (Titanic и Spotify), где применялись изученные методы для решения задач бинарной классификации и предсказания.

2. Структура учебных материалов и ноутбуков

Каждая лабораторная работа оформлялась в виде Google Colab Notebook, что обеспечивало наглядность и удобство воспроизведения кода. Ноутбуки включали:

– Теоретическую часть – краткое объяснение методов и математических основ.

– Практическую реализацию – пошаговый разбор кода с комментариями.

– Анализ результатов – визуализации, сравнение моделей, выводы.

Такой подход позволял не только закрепить материал, но и создать удобные конспекты для дальнейшего использования.

3. Методология разработки и тестирование

Работа велась по итеративному принципу:

1) Изучение теории – ознакомление с материалами перед выполнением задания.

2) Написание кода – реализация алгоритмов и проверка их работы на учебных данных.

3) Тестирование и отладка – анализ ошибок, оптимизация гиперпараметров, улучшение точности моделей.

4) Финализация решения – оформление ноутбука, подготовка отчета.

Основные сложности и их устранение:

– В задачах машинного обучения (особенно в градиентном спуске) важным этапом была отладка вычислений, чтобы избежать расходимости или переобучения.

– В соревнованиях ключевой задачей была корректная предобработка данных и выбор метрик, что потребовало дополнительного анализа.

4. Планирование и управление временем

Для эффективного выполнения работ использовался календарный план, в котором были зафиксированы:

– Дедлайны по лабораторным – равномерное распределение нагрузки.

– Время на участие в соревнованиях – дополнительный период для экспериментов с данными и моделями.

– Анализ прогресса – еженедельная проверка выполненных задач и корректировка при необходимости.

Такой подход позволил завершить проект в срок, получив не только выполненные задания, но и практический опыт в Data Science.

ЗАКЛЮЧЕНИЕ

1. Оценка соответствия поставленным целям

Проект полностью соответствует заявленным образовательным целям в сфере Data Science. В рамках выполнения работ был пройден полный цикл обучения: от теоретических основ до практической реализации ML-моделей. Особенно стоит отметить:

– Успешное освоение ключевых технологий (NumPy, Pandas, Scikit-learn)

– Практическое применение знаний в соревнованиях Kaggle

– Создание полноценных учебных материалов в виде Google-colab-ноутбуков

Проект достиг «оптимального» уровня результата по заявленной шкале, в том числе благодаря участию в соревнованиях и есть все шансы приближаться к абстрактном «повышенному» уровню.

2. Оценка качества результатов

Анализ выполненных работ показывает:

1) Полнота охвата тем:

– Рассмотрены все заявленные направления: от базовой обработки данных до сложных ML-алгоритмов

– Особенно ценным оказался сравнительный анализ разных подходов (линейные модели / деревья / KNN)

2) Практическая значимость:

– Созданные материалы действительно позволяют получить навыки, востребованные в индустрии

– Участие в соревнованиях дало опыт работы с реальными данными

3) Выявленные ограничения:

– Недостаточная глубина рассмотрения некоторых современных методов (например, нейронных сетей)

– Ориентация в основном на классические задачи, без учета последних трендов

3. Рекомендации по развитию

Для перехода на «повышенный» уровень планируется:

1) Расширение тематики:

– Добавить разделы по глубокому обучению

– Включить кейсы из реальных бизнес-задач

2) Улучшение методики:

– Разработать систему проверки знаний

– Добавить интерактивные элементы обучения

3) Практическое применение:

- Организовать работу над реальными проектами

- Ввести систему наставничества от опытных специалистов

4. Вывод

Проект успешно выполнил свою образовательную миссию, создав прочную основу для дальнейшего профессионального роста в области Data Science. Следующим логичным шагом должно стать углубление практического опыта через участие в реальных проектах.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Маккинни, У. Python и анализ данных / У. Маккинни. – Москва : ДМК Пресс, 2020. – 482 с.

2. Вандер Плас, Дж. Python для сложных задач: наука о данных и машинное обучение / Дж. Вандер Плас. – Санкт-Петербург : Питер, 2018. – 576 с.

3. Джеррард, П. Изучаем Pandas / П. Джеррард, М. Хейсман. – Москва : Эксмо, 2021. – 320 с.

4. Бринсон, Л. Машинное обучение с использованием Python / Л. Бринсон, Р. Бхаргава. – Москва : Вильямс, 2019. – 448 с.

5. Харрисон, М. Pandas: работа с данными / М. Харрисон. – Санкт-Петербург : БХВ-Петербург, 2022. – 304 с.Лугинина, И. Г. Химия и химическая технология неорганических вяжущих материалов : учеб. пособие для студентов вузов. В 3 ч. Ч. 1 / И. Г. Лугинина. – Белгород : Изд-во БГТУ, 2004. – 240 с.

ПРИЛОЖЕНИЕ

ПРИЛОЖЕНИЕ А

(обязательное)

https://github.com/denisbebrovich/Project

Репрозиторий с первой лабораторной работой.

ПРИЛОЖЕНИЕ В

(обязательное)

https://github.com/denisbebrovich/Project2

Репрозиторий со второй лабораторной работой.

ПРИЛОЖЕНИЕ С

(обязательное)

https://github.com/denisbebrovich/Project3_

Репрозиторий с третьей лабораторной работой.

ПРИЛОЖЕНИЕ D

(обязательное)

https://github.com/denisbebrovich/Project4.git

Репрозиторий со четвертой лабораторной работой.

ПРИЛОЖЕНИЕ Е

(обязательное)

https://github.com/denisbebrovich/Titanic

Репрозиторий с первым соревнованием Титаник

ПРИЛОЖЕНИЕ F

(обязательное)

https://github.com/denisbebrovich/Spotify

Репрозиторий со вторым соревнованием Спотифай