TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Разработка Web-приложения для выбора метода анализа и модели прогноза для поисковых нефтяных скважин

УДК 004.774:303.7:622.24

Студент

Группа	**ФИО**	**Подпись**	**Дата**
8ПМ1И	Филипас Иван Александрович		29.05.2023 г.

Руководитель ВКР

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
доцент ОИТ ИШИТР	Губин Е.И.	к.ф.-м.н.		29.05.2023 г.

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
доцент ОСГН ШБИП	Спицына Л. Ю.	к.э.н.		

По разделу «Социальная ответственность»

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
доцент ООД ШБИП	Антоневич О.А.	к.б.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
доцент ОИТ ИШИТР	Губин Е.И.	к.ф.-м.н.		

Томск – 2023

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ООП
по направлению 09.04.04 «Программная инженерия»

Код компетенции	Наименование компетенции
Универсальные компетенции	
УК(У)-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК(У)-2	Способен управлять проектом на всех этапах его жизненного цикла
УК(У)-3	Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК(У)-4	Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых) языке (-ах), для академического и профессионального взаимодействия
УК(У)-5	Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия
УК(У)-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки
Общепрофессиональные компетенции	
ОПК(У)-1	Способен самостоятельно приобретать, развивать и применять математические, естественно-научные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
ОПК(У)-2	Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач
ОПК(У)-3	Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями
ОПК(У)-4	Способен применять на практике новые научные принципы и методы исследований
ОПК(У)-5	Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем
ОПК(У)-6	Способен самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания

	и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности
ОПК(У)-7	Способен применять при решении профессиональных задач методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях
ОПК(У)-8	Способен осуществлять эффективное управление разработкой программных средств и проектов
Профессиональные компетенции	
ПК(У)-1	Способен к созданию вариантов архитектуры программного средства
ПК(У)-2	Способен разрабатывать и администрировать системы управления базам данных
ПК(У)-3	Способен управлять процессами и проектами по созданию (модификации) информационных ресурсов
ПК(У)-4	Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий
ПК(У)-5	Способен осуществлять руководство разработкой комплексных проектов на всех стадиях и этапах выполнения работ

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Губин Е.И.
(подпись) (дата)
(Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	**ФИО**
8ПМ1И	Филипасу Ивану Александровичу

Тема работы:

Разработка Web-приложения для выбора метода анализа и модели прогноза для поисковых нефтяных скважин.	
Утверждена приказом директора (дата, номер)	№ 37-58/с от 06.02.2023

Срок сдачи студентом выполненной работы:	10.06.2023

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или*	Объектом исследования является разработка Web-приложения для выбора метода анализа и модели прогноза для поисковых нефтяных скважин.

процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).	
Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).*	1. Обзор предметной области. 2. Описание исходных данных. 3. Описание проблемы исследования. 4. Подготовка данных. 5. Разработка web-приложения. 6. Работа над разделом по финансовому менеджменту, ресурсоэффективности и ресурсосбережения. 7. Работа над разделом по социальной ответственности.
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. Скриншоты исходных данных. 2. Скриншоты результатов работы алгоритмов. 3. Скриншоты экранов web-приложения 4. Матрица SWOT 5. Диаграмма Ганта.

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	**Консультант**
Основная часть	Доцент ОИТ ИШИТР, к.ф.-м.н., доцент Губин Е.И.
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Доцент ОСГН ШБИП, к.э.н., доцент Спицына Л.Ю.
Социальная ответственность	Доцент ООД ШБИП, к.б.н., доцент Антоневич О.А.
Английский язык	Доцент ОИЯ, к.п.н., доцент Уткина А.Н.

Названия разделов, которые должны быть написаны на русском и иностранном языках:

Development of a web application for selecting an analysis method and forecast model for exploratory oil wells

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	1.03.2023

Задание выдал руководитель ВКР:

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
доцент ОИТ ИШИТР	Губин Е.И..	к.ф.-м.н.., доцент		1.03.2023 г.

Задание принял к исполнению обучающийся:

Группа	**ФИО**	**Подпись**	**Дата**
8ПМ1И	Филипас Иван Александрович		1.03.2023 г.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа Информационных технологий и робототехники
Направление подготовки (ООП / ОПОП) 09.04.04 Программная инженерия
Уровень образования магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения весенний семестр 2022 /2023 учебного года

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Обучающийся:

Группа	ФИО
8ПМ1И	Филипас Иван Александрович

Тема работы:

Разработка Web-приложения для выбора метода анализа и модели прогноза для поисковых нефтяных скважин

Срок сдачи обучающимся выполненной работы:	10.06.2023 г.

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
10.06.2023	Основная часть	70
10.06.2023	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
10.06.2023	Социальная ответственность	10
10.06.2023	Раздел на английском языке	10

СОСТАВИЛ:
руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		

СОГЛАСОВАНО:
руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		

Задание принял к исполнению обучающийся:

Группа	ФИО	Подпись	Дата
8ПМ1И	Филипас Иван Александрович		1.03.2023 г.

Реферат

Работа содержит пояснительную записку на 80 страниц, 18 рисунков, 21 таблицу, 1 приложение.

Ключевые слова: web-приложение, набор данных, выбор метода анализа, регрессия, python.

Основная задача создаваемого приложения заключается в том, чтобы пользователи, которые хотят применять технологии больших данных для прогнозного анализа, могли использовать web-приложение для анализа и выбора наиболее подходящей прогнозной модели. Реализуемое приложение позволит пользователю загрузить дата-сет, который будет обработан программой, а затем, с помощью встроенных моделей, будет представлен краткий статистический анализ и какие методы были использованы, и какой из них более подходит под его конкретную задачу. В результате, пользователь получает простой и удобный способ для первичного анализа данных и выбора оптимальной прогнозной модели.

Содержание

Введение 11

Обозначения и сокращения 16

1. Разработка web-приложения для выбора оптимального метода прогнозного анализа перспективных нефтяных скважин. 17

1.1. Описание исходного дата-сета 17

1.2. Используемые алгоритмы обработки данных и модели прогнозного анализа 18

1.3. Структура приложения 22

1.4. Пользовательский интерфейс 25

1.5. Пользовательский сценарий 28

2. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение 33

2.1. Предпроектный анализ 33

2.1.1. Потенциальные потребители разработки 34

2.1.2. Технология QuaD 34

2.1.3. SWOT-анализ 35

2.1.4. Оценка готовности разработки к коммерциализации 38

2.2. Инициация проекта 39

2.3. Планирование управления разработкой 41

2.3.1. Иерархическая структура работ 41

2.3.2. План разработки 42

2.3.2.1. Разработка графика проведения разработки 43

2.3.3. Бюджет научного исследования 44

2.3.3.1. Материальные расходы 44

2.3.3.2. Основная заработная плата исполнителей темы 44

2.3.3.3. Дополнительная заработная плата исполнителей темы 45

2.3.3.4. Отчисления во внебюджетные фонды 46

2.3.3.5. Накладные расходы 47

2.3.3.6. Формирование бюджета затрат на разработку 47

2.3.3.7. Риски разработки....47
2.4. Экономическая эффективность....48
2.4.1. Интегральный показатель эффективности разработки....49
2.5. Выводы по разделу....50
3. Социальная ответственность....53
3.1. Правовые и организационные вопросы обеспечения безопасности....53
3.1.1. Специальные (характерные для проектируемой рабочей зоны) правовые нормы трудового законодательства....53
3.1.2. Организационные мероприятия при компоновке рабочей зоны....53
3.2. Производственная безопасность....54
3.2.1. Анализ потенциально возможных и опасных факторов, которые могут возникнуть на рабочем месте при проведении исследований....54
3.2.2. Недостаточная освещенность рабочей зоны....55
3.2.3. Повышенный уровень шума на рабочем месте....57
3.2.4. Отклонения параметров микроклимата....58
3.2.5. Поражение электрическим током....59
3.3. Экологическая безопасность при эксплуатации....59
3.3.1. Защита атмосферы....60
3.3.2. Защита гидросферы....61
3.3.3. Защита литосферы....62
3.4. Безопасность в чрезвычайных ситуациях при разработке проектного решения....63
3.5. Выводы по разделу....64
Заключение....65
Список публикаций....66
Список используемых источников....67
Приложение I....70
Development of a web application for selecting an analysis method and forecast model for exploratory oil wells....71

Введение

Процесс сбора и подготовки исходных данных, является одним из самых трудоемких и сложных этапов в анализе больших объемов данных, который порой занимает до 80% всего времени. Процесс подготовки исходных данных включает в себя следующие этапы [1]:

– перевод исходного файла из исходного формата .csv в формат .xlsx (либо sas, либо Python)

– проверку исходных данных на ошибки и описки («typo»);

– проверку исходных данных на пропущенные значения («missing»);

– проверку исходных данных на выбросы («outliers»);

– проверку исходных данных на наличие дублирующих строк (наблюдений);

– проверку исходных данных на мультиколлинеарность;

– трансформация исходных данных в цифровой формат («цифровизация»)

– выбор целевой переменной.

Полученная методика может быть реализована в программных пакетах Python, SAS, SAS Enterprise Miner, Excel.

Принимая во внимание, что большая часть реальных данных носит характер слабо структурированных или вовсе берётся из хранилища “озера данных”, вопрос корректности последних носит критический характер. Достаточно сказать, что при тщательной и корректной подготовке исходных данных удаётся почти на 18% повысить точность и предсказательную силу традиционных прогнозных моделей [2].

Каждый этап подготовительного анализа и обработки данных требует серьёзной вовлеченности аналитика в бизнес-сферу, однако некоторую часть работ можно делать в автоматическом режиме. Например, такие работы как проверка исходных данных на ошибки или описки, пропущенные значения, неверный формат данных – можно переложить на формальное исполнение в приведённом ниже приложении.

В качестве базовых методов обработки эксперты предлагают использовать следующие рекомендации для «чистки» исходных данных при наличии отсутствующих или ошибочных данных:

– если количество отсутствующих данных не превышает 5%, то при сохранении репрезентативности, эти строки с «missing» могут быть удалены;

– если количество отсутствующих данных превышает 50%, то данный атрибут возможно удалить из дальнейшего анализа;

– если количество отсутствующих данных находится в интервале 5% - 50%, то для численных атрибутов («numeric») возможно несколько вариантов замены отсутствующих значений: средним («mean»), медианой («median»), «ближайших соседей» и др. Для текстовых атрибутов («char») возможно использовать значения, наиболее часто встречающиеся либо «ближайших соседей».

Перед тем как приступать к подготовке («чистке») исходных данных желательно провести процедуру «описательная статистика» дата-сета для всех входящих переменных. Это поможет дать общее представление об исходных данных и определиться с входными данными и целевой переменной.

Обычно это включает следующие статистики [1]:

– для каждой численной входной переменной (атрибута- «Numeric»)
  – общее количество строк (наблюдений), включая отсутствующие («N»);
  – минимальное значение («Min»);
  – максимальное значение («Max»);
  – среднее квадратичное отклонение («St.div»);
  – среднее («Mean»);
  – медиана («Median»),
– для текстовой переменной (атрибута – «Char»):
  – общее количество строк (наблюдений), включая отсутствующие («N»);
  – частоту входящих параметров («Frequency») и их количество («N»).

Следующим этапом обработки дата-сета является формирование тренировочной и тестовой выборки. Их содержание и объем по отношению ко всему объему исходных данных обычно составляет 70:30 и обязательное требование репрезентативности. Если используется несколько обучающихся моделей и данных достаточно много, то добавляют валидирующую выборку в соотношении 60:20:20.

Так как процесс подготовки данных является крайне трудоёмким и требует наличия высококвалифицированного аналитика, в данной работе проведено исследование на предмет возможного упрощения и ускорения работы с дата-сетами посредством создания web-приложения.

Реализуемое приложение автоматизирует большую часть рутинной работы и позволит пользоваться технологиями анализа большему числу аналитиков, тем самым ускоряя и упрощая их работу.

Приложение автоматически обрабатывает пропущенные значения, выбросы данных, дублирующие строки, а также проверяет, и при необходимости, изменяет формат данных

(из буквенного формата переводит в числовой). Помимо этого, приложение устраняет мультиколлинеарность. Реализуемое приложение проводит автоматическое разбиение исходного дата-сета на две выборки: тренировочную и тестовую, в соотношении 70:30 соответственно.

В дополнение к описанному функционалу приложение включает в себя четыре варианта прогнозных моделей, которые основаны на логистической регрессии. В работе приведены три вида логистической регрессии:

– основанные на методе предсказания;
– основанные на методе регуляризации;
– с дополнительным коэффициентом регуляризации.

Следует подробнее остановиться на методах предсказания и регуляризации, которые использовались в исследовании.

Первый метод LBFGS [4] — это алгоритм оптимизации, который использует способ оптимизации Бройдена – Флетчера – Гольдфарба – Шанно, который принадлежит к квазиньютоновским методам. Решатель «lbfgs» рекомендуется использовать для небольших наборов данных, так как для больших наборов данных страдает его производительность [2].

Второй метод SAGA [5] – алгоритм представляет собой вариант, который поддерживает различные виды регуляризации. Это предпочтительный алгоритм для больших объемов данных, с большой плотностью измерений. [3].

Третий метод LIBLINEAR [6] – алгоритм, использующий метод координатного спуска и полагается на библиотеку из языка C ++, которая поставляется вместе с используемой в данном исследовании библиотекой scikit-learn. Алгоритм, реализованный в liblinear, решает проблему оптимизации с помощью декомпозиции по принципу «один против остальных» [2].

Четвертый метод, реализованный в приложении, это newton-cholesky [7]. Данный метод является хорошим выбором для тех случаев, в которых количество измерений в несколько раз больше, чем количество переменных в дата-сете. Использовать этот метод рекомендуется только в тех случаях, когда применяется логистическая регрессия и только для работы с категориальными переменными.

Так же в работе важной частью являются методы регуляризации или штрафы, принятые в англоязычной литературе. В рамках исследования были разобраны 4 разных варианта регуляризации: L1, L2, elastinet и отсутствие регуляризации. Регуляризации отвечает за то, чтобы модель не была переобучена или наоборот не была недостаточно

обучена. L1 по математике противоположен по смыслу L2, однако в отличии от L2 веса могут стать нулевыми, при очень большом значении коэффициента. Elastinet объединяет в себе две регуляризации.

Критерием оценки точности моделей выступает ROC\AUC кривая и параметр Accuracy (точность) [8]. Точность модели означает насколько предсказанный результат близок по значению с тем, что было заложено в обучающую выборку. ROC – график, который показывает эффективность модели машинного обучения при решении задачи классификации, отображая частоту истинных и ложных значений. AUC – площадь под кривой. ROC-кривая показывает соотношение истинно положительных результатов и ложноположительных.

Реализация в качестве Web-приложения была выбрана для удобства работы и уменьшения затрат для пользователя, а именно:

– все вычисления происходят на серверной стороне программы;

– не требуется скачивать приложение и иметь большие объёмы вычислительных мощностей, чтобы пользоваться данным приложением.

– Так же плюсами web-реализации будет то, что приложение:

– доступно с любого устройства, подключенного к интернету;

– интуитивно понятно для пользователя.

– В качестве основного языка разработки программного продукта был выбран язык Python, по причинам:

– этот язык является самым популярным для работы с большими данными;

– для этого языка написано большое количество библиотек и функций, которые позволяют эффективнее обрабатывать дата-сеты, чем другие языки программирования;

– преподавался в течении всей программы обучения.

Для работы и последующей обработки данных были выбраны библиотеки NumPy [9] и Pandas [10]. Они позволяют работать с большими данными и работать с дата-сетом как с одним объектом, так и обращаться к каждой отдельной ячейке.

Для регрессии была выбрана библиотека sklearn [10]. Данная библиотека позволяет быстро и просто использовать различные виды логистической и линейной регрессии, а также техники случайного леса и т.д.

В качестве инструмента реализации front-end’а был выбран Dash \ Plotly [3]. Данный инструмент позволяет, не изучая язык разметки и стилей, используя лишь код Python, легко создавать Web-приложения и интерфейс взаимодействия с ними. Dash \ Plotly состоит из двух основных частей. Часть Dash отвечает за разметку, различные компоненты и

элементы, кнопки и стилизацию внешнего вида страницы. Plotly отвечает за графики и отрисовку различных графических и интерактивных объектов.

Основным пользовательским сценарием при работе с Web-приложением предполагается следующий алгоритм:

– загрузка дата-сета;
– выбор настроек для дата-сета;
– выбор целевой функции;
– оценка качества обработки дата-сета;
– изучение результатов и выбор одного из предложенного метода анализа.

В качестве результатов анализа пользователю представлена описательная статистика загруженного им дата-сета, которая включает в себя следующие значения:

– количество измерений;
– среднее значение;
– среднеквадратичное отклонение;
– минимальное значение;
– максимальное значение;
– 25, 50 и 75 перцентили, а также тип данных.

В дополнение к вышеописанной статистике пользователь получит две тепловые карты корреляции параметров: первая карта – до устранения мультиколлинеарности, вторая – после.

Обозначения и сокращения

Дата-сет — это обработанный и структурированный массив данных. В нём у каждого объекта есть конкретные свойства: признаки, связи между объектами или определённое место в выборке данных.

Web-приложение – это клиент-серверное приложение, в котором клиент взаимодействует с веб-сервером при помощи браузера. Логика веб-приложения распределена между сервером и клиентом, хранение данных осуществляется, преимущественно, на сервере, обмен информацией происходит по сети.

Регуляризация – это метод добавления некоторых дополнительных ограничений к условию с целью решить некорректно поставленную задачу или предотвратить переобучение. Чаще всего эта информация имеет вид штрафа за сложность модели.

Регрессия - математическое выражение, отражающее связь между зависимой переменной y и независимыми переменными x.

1. Разработка web-приложения для выбора оптимального метода прогнозного анализа перспективных нефтяных скважин.

В данной главе рассмотрены основные этапы разработки инструментов создания web-приложения, и анализ результатов его работы.

1.1. Описание исходного дата-сета

Исходными данными для проекта был выбран дата-сет, содержащий информацию о нефтедобывающих скважинах. На рисунке 1 представлена описательная статистика дата-сета. Всего дата-сет состоит из 185 строк, все данные представлены в числовом формате. Дата-сет содержит 16 переменных, включая одну целевую функцию.

Name	N	Min	Max	Mean	Median	Missing	Outliers	N	St.dev
L, м	185	170	750	477	480	0	0	185	186,1505
N, шт	185	3	11	5	5	0	25	185	2,287207
H pay, м	185	1	17	6	5	0	8	185	3,337318
Рпл, атм	185	0	461	229	165	0	0	185	112,5954
k, мД	185	0	13	3	2	0	4	185	2,032852
μ oil, сП	185	0	3	2	3	0	0	185	1,084404
Az, градусы	185	0	352	196	219	0	0	185	103,5732
Mprop, тонн	185	0	102	30	10	0	0	185	29,21837
Xf, м	185	32	253	98	86	0	0	185	46,57008
Height, м	185	9	69	24	17	0	4	185	14,32293
Width, м	185	0,0006	0,0089	0,0029	0	0	4	185	0,00114
Рзаб, атм	185	0	230	79	75	0	15	185	30,84568
dP, атм	185	0	400	150	99	0	0	185	109,8156
WLPR, м3/сут	185	9	299	74	65	0	5	185	49,59974
WOPR, м3/сут	185	0	123	35	28	0	5	185	26,00035
WWPR, м3/сут	185	1	262	21	8	0	5	185	45,57683

Рисунок 1 – Описательная статистика исходного дата-сета

В представленной таблице имеются следующие статистические параметры:

- N – количество не нулевых записей;
- Min \ max – Минимальное и максимальное значение параметра;
- Mean – средне-арифметическое значение параметра;
- Median – медианное значение параметра;
- Missing – количество пустых значений;
- Outliers – Количество значений, которые вне основной группы.
- St.dev – среднеквадратичное отклонение по параметрам.

В таблице 1 представлены размерность и описание каждой переменной исходного дата-сета.

Таблица 1 – Описание переменных из дета-сета

Параметр	Единица измерения	Описание
L	м	Радиус зоны дренирования
N	шт	Количество стадий ГРП
H pay	м	Мощность продуктивного пласта
Рпл	атм	Пластовое давление
k	мД	Проницаемость зоны дренирования
μ oil	сП	Вязкость нефти
Az	градусы	Азимут распространения трещины
Mprop	тонн	Масса проппанта
Xf	м	Полудлина трещины (длина одного крыла)
Height	м	Высота трещины зоны дренирования
Width	м	Ширина трещины зоны дренирования
Разб	атм	Забойное давление
dP	атм	Депрессия (разница пластового и забойного давлений)
WLPR	м3/сут	Дебит жидкости скважины
WOPR	м3/сут	Дебит нефти скважины
WWPR	м3/сут	Дебит воды скважины

В качестве целевой функции для данного дата-сета был выбран параметр WOPR, т.к. именно количество нефти будет определять, выгодно или не выгодно добывать нефть в данной скважине. Целевой функцией для данного дата-сета является Дебит нефти. Его значения выше медианного принимаются как “1”, или экономически эффективная зона добычи. Значения ниже медианного как “0”, или экономически неэффективная зона для добычи нефти.

1.2. Используемые алгоритмы обработки данных и модели прогнозного анализа

Используемые алгоритмы обработки данных, реализованные в данной работе, можно разделить на два вида. Первый вид направлен на упрощение и автоматизацию подготовительного процесса и включает в себя:

- проверку исходных данных на пропущенные значения («missing»);
- проверку исходных данных на выбросы («outliers»);
- проверку исходных данных на наличие дублирующих строк (наблюдений);
- проверку исходных данных на мультиколлинеарность;

Первым действием идёт отбрасывание пропущенных значений. Для этого выполняется метод dropna() библиотеки Pandas. Этот метод проходит по каждому значению в дата-сете и убирает те строки, в которых есть нулевые значения.

Проверка исходных данных на выбросы более трудоёмкая задача и заключается в том, чтобы отбросить только излишне высокие и низкие значения. В рамках приложения высокие и низкие значения подразумеваются как значения, которые сильно выше или сильно ниже основной группы значений. Для этого было решено использовать квантили. В математической статистике значение, которое заданная случайная величина не превышает с фиксированной вероятностью. Реализация представлена на рисунке 2.

```
def outliers(df,x):

    q_low = df[x].quantile(0.01)
    q_hi  = df[x].quantile(0.99)
    df_filtered = df[(df[x] < q_hi) & (df[x] > q_low)]

    return df_filtered
```

Рисунок 2 – Реализация удаления выбросов по заданному параметру

На рисунке изображена функция, в которую передаётся два параметра – исходный дата-сет и параметр, по которому происходит удаление выбросов. В функции удаляются все значения, которые больше, чем 99% значений и меньше, чем 1% значений от всего дата-сета.

Следующей частью проверки будет исключение дублирующих строк, для этого существует специальная встроенная функция в библиотеку Pandas . drop_duplicates. Функция сравнивает все строки между собой, и, при необходимости, удаляет одну из не уникальных строк.

Последним этапом будет удаление мультиколлинеарности. На рисунке 3 изображена функция, которая обрабатывает каждую строку дата-сета, и сравнивает её с заданным значением порога. Если значение порога будет меньше, чем коэффициент, посчитанный для данного параметра, то данный параметр удаляется из дата-сета. Величина порога выбирается пользователем в интерфейсе приложения.

```
def calculate_vif_(X, thresh):
    X = X.assign(const=1)  # faster than add_constant from statsmodels
    variables = list(range(X.shape[1]))
    dropped = True
    while dropped:
        dropped = False
        vif = [variance_inflation_factor(X.iloc[:, variables].values, ix)
               for ix in range(X.iloc[:, variables].shape[1])]
        vif = vif[:-1]  # don't let the constant be removed in the loop.
        maxloc = vif.index(max(vif))
        if max(vif) > thresh:
            print('dropping \'' + X.iloc[:, variables].columns[maxloc] +
                  '\' at index: ' + str(maxloc))
            del variables[maxloc]
            dropped = True
```

Рисунок 3 – Расчёт коэффициента для удаления мультиколлинеарности

Проведя вышеописанные действия, мы увеличили точность прогнозных моделей, согласно экспертам, на 6.9% [4]. Данный показатель можно увеличить ещё, добавив проверки на изменения формата данных с текстовых в числовые и проверку на опечатки и описки. Однако в работе эта проверка не представляется возможной, так как ее должен выполнять аналитик, потому что знание тонкостей предметной области недоступно при разработке данного приложения и удаление описок и изменения формата данных может повлечь за собой уменьшение точности прогнозной модели.

На данном этапе реализовано методология прогнозного анализа, основанная на использовании логистической регрессии. Модели логистической регрессии отличаются между собой следующим:

1) метод предсказания;
2) метод регуляризации;
3) дополнительный коэффициент регуляризации, отвечающий за качественное влияние метода регуляризации.

Первый метод логистической регрессии основан на алгоритме LBFGS. LBFGS — это алгоритм оптимизации, который приближает алгоритм Бройдена – Флетчера – Гольдфарба – Шанно, который принадлежит к квазиньютоновским методам. Решатель «lbfgs» рекомендуется использовать для небольших наборов данных, но для больших наборов данных страдает его производительность [2].

Второй метод логистической регрессии основан на алгоритме SAGA. SAGA – алгоритм представляет собой вариант, который поддерживает различные виды регуляризации. Это предпочтительный алгоритм для больших объемов данных, с большой плотностью измерений. [3].

Третий метод логистической регрессии основан на алгоритме LIBLINEAR. LIBLINEAR – алгоритм использующий метод координатного спуска и полагается на библиотеку из языка C ++, которая поставляется вместе с используемой в данном исследовании библиотекой scikit-learn. Алгоритм, реализованный в liblinear, решает проблему оптимизации с помощью декомпозиции по принципу «один против остальных», поэтому отдельные двоичные классификаторы обучаются для всех классов [2].

Четвертый метод, реализованный в приложении это newton-cholesky. Данный метод является хорошим выбором для тех случаев, в которых количество измерений в несколько раз больше, чем количество переменных в дата-сете. Использовать этот метод рекомендуется только в тех случаях, когда применяется логистическая регрессия и для работы с категориальными переменными.

Так же в работе важной частью являются методы регуляризации [11]. В рамках исследования были разобраны 4 разных варианта регуляризации: L1 [12], L2 [12], elastinet [13] и отсутствие регуляризации. Регуляризации отвечает за то, чтобы модель не была переобучена или наоборот недостаточно обучена. L1 по математике противоположен по смыслу L2, однако в отличии от L2 веса могут стать нулевыми, при очень большом значении коэффициента. Elastinet объединяет в себе две регуляризации. Пример реализации одной из моделей представлен на рисунке 4

```
X_train,X_test,y_train,y_test = train_test_split (x,y,test_size=0.5,random_state=22)
log_regression = LogisticRegression(penalty='elasticnet', l1_ratio= 0.4,  solver= 'saga', max_iter = 100)
log_regression.fit(X_train,y_train)
y_pred = log_regression.predict(X_test)
fpr, tpr, thresholds = roc_curve(y_test, y_pred)
fig = px.area(
    x=fpr, y=tpr,
    title=f'ROC Curve (AUC={auc(fpr, tpr):.4f})',
    labels=dict(x='False Positive Rate', y='True Positive Rate')
)
fig.add_shape(
    type='line', line=dict(dash='dash'),
    x0=0, x1=1, y0=0, y1=1
)

fig.update_yaxes(scaleanchor="x", scaleratio=1)
fig.update_xaxes(constrain='domain')
accuracy = str(metrics.accuracy_score(y_test, y_pred))
a = 'For SAGA method of predictioning and elasticnet (L1_ratio= 0.4) normalization the Accuracy is: ' + accuracy
return fig, a
```

Рисунок 4- Логистическая регрессия методом предсказания SAGA

Таким образом приложение выполняет функции автоматической подготовки дата-сета для дальнейшей работы аналитика, что уменьшает трудоёмкость, позволяя выполнять данную работу менее квалифицированному специалисту.

1.3. Структура приложения

Структура приложения древовидная и представлена на рисунке 5.

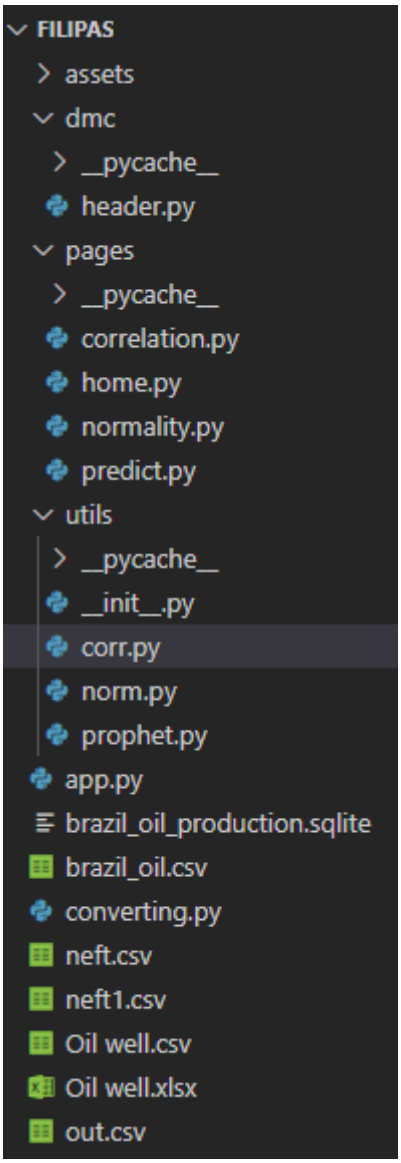

Рисунок 5 – Структура приложения

На рисунке изображены папки: assets, dmc, pages, utils. А также файлы app.py и различные дата-сеты, которые участвовали в проработке моделей и тестировании приложения.

Папка assets и dmc – системные и содержат в себе библиотеки и другие файлы, необходимые для корректной работы приложения, отрисовки графиков и запуска приложения на выделенном сервере. Папки pages и utils содержат в себе код для отрисовки пользовательского интерфейса и алгоритмов обработки данных и моделей прогнозирования соответственно.

Файл app.py является начальной точкой работы web-приложения и запускает выделенные локальный сервер для работы с моделями и алгоритмами через

пользовательский интерфейс. В файле headers.py содержится информация о шапке web-приложения.

Структура страницы состоит из пользовательского интерфейса и алгоритмов, которые реагируют на действия пользователя и обрабатывают дата-сет, предоставленный пользователем. На рисунках 6-7 представлен пример кода, который отвечает за отрисовку пользовательского интерфейса.

```
layout = dmc.Container(
    children=[
        html.Br(),
        dmc.Group(
            children=[
                dcc.Dropdown(···
                dcc.Dropdown(
                        ['Just Enter',','],
                        placeholder='Select line sep',
                        id = 'next_line'
                    ),

            ],
            grow=True,
            position='center',
            spacing='xl',

        ),
        html.Br(),
        dmc.Group(
            children=[
                dcc.Upload(
                            id='upload',
                            children=html.Div([
                                'Drag and Drop or ',
                                html.A('Select Files')
                            ]),
                            style={···
                            multiple=False
```

Рисунок 6 – Пример кода главной страницы приложения

```
html.Br(),
dash_table.DataTable(id = 'DF', style_table={'overflowX': 'auto'}),
html.Br(),
dmc.Group(
    children=[
        dcc.Dropdown(
                id = 'target'
            ),
        dcc.Dropdown(
                id = 'corr'
            ),
        dmc.Button(
            'Calculate and save',
            id = 'cas',
            style={'color':'White', 'background': '#0033FF', 'border':'1'}
        )

    ],
    grow=True,
    position='center',
    spacing='xl',
),
html.Br(),
dbc.Alert(id='tbl_out'),
html.Br(),
dcc.Graph(id = 'gr_before', responsive= True, style={
                    'height': '700px'
```

Рисунок 7 – Пример кода главной страницы приложения

На рисунках 6-7 представлен код, который с помощью библиотеки Dash преобразуется в код страницы, а именно в JavaScript, html и css. Пример работы данного фрагмента кода представлен на рисунке 8.

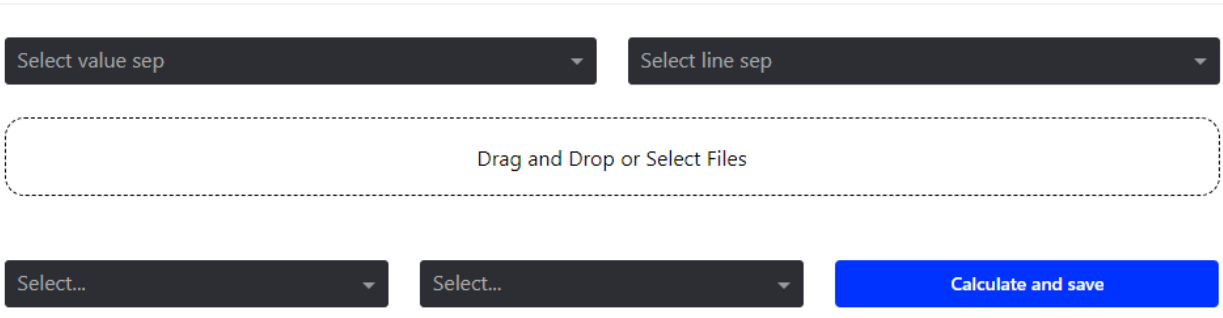

Рисунок 8 – Пример работы Dash по преобразованию кода Python в интерфейс приложения

При выполнении действий на пользовательском интерфейсе исполняется одна из функций, прописанных в файлах из папки pages. В каждой функции прописан вызов стороннего алгоритма, реализованного в одном из файле в папках utils. Пример вызова алгоритма из функции реакции на пользовательском интерфейсе представлен на рисунке 9.

```python
@callback(
    Output(component_id='DF', component_property='data'),
    Output(component_id='DF', component_property='columns'),
    Output('target', 'options'),
    Output('corr', 'options'),
    Input('upload', 'contents'),
    State('next_line', 'value'),
    State('next_value', 'value'),
    prevent_initial_call=True
)
def update_map(df, line_sep, value_sep):
    import io
    import base64
    import pandas as pd
    #print(df, line_sep)
    content_type, content_string = df.split(',')
    decoded = base64.b64decode(content_string).decode('utf-8')
    decoded = io.StringIO(decoded)
    df = pd.read_csv(decoded, sep = value_sep)
    df.to_csv('out.csv', index=False, sep = ';')
    df = df.head()
    columns = [{'name': col, 'id': col} for col in df.columns]
    tar = [col for col in df.columns]
    data = df.to_dict(orient='records')
    a = [20.0, 11.0 , 5.0]
    return data, columns, tar, a
```

Рисунок 9 – Обработка действий пользователя

Таким образом реализована древовидная структура проекта и обработка действий пользователя в приложении.

1.4. Пользовательский интерфейс

Пользовательский интерфейс является одной из важнейших частей приложения, так как позволяет в удобной форме обрабатывать дата-сет и в дальнейшем выбирать метод анализа, который лучше подойдёт для дата-сета пользователя.

Интерфейс представлен двумя страницами: для обработки дата-сета и подготовки для дальнейшего анализа, результаты анализа и предложения выбора метода анализа. На рисунках 10-12 представлен первый экран приложения. На 10 рисунке представлен стартовый экран, на 11 и 12 – экраны после добавления и обработки дата-сета.

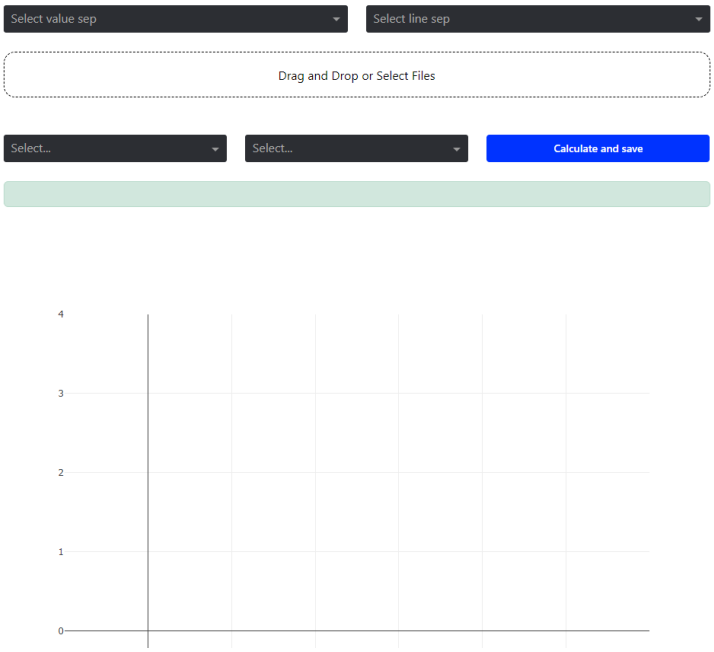

Рисунок 10 – Стартовый экран приложения

Рисунок 11 – Заполненный экран приложения с исходным графиком

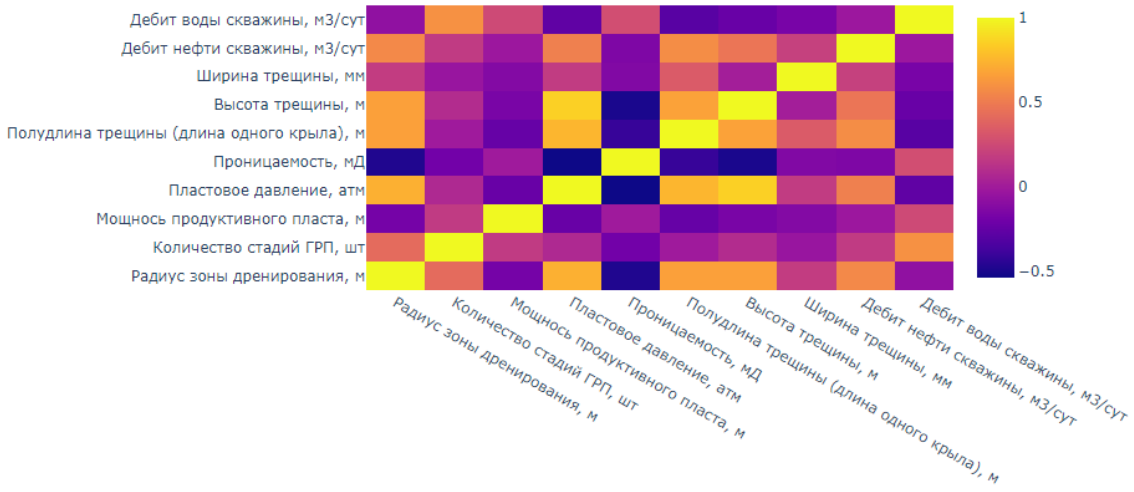

Рисунок 12 – Тепловая карта дата-сета после применения фильтра

На стартовом экране, в верхней части, отображаются селекторы для выбора и настройки формата заполнения дата-сета, выбора разделителя. Затем идёт элемент, в котором можно выбрать или перенести в него файл, который содержит дата-сет. Следом отображаются два селектора, в которых пользователь выбирает целевую переменную и силу фильтрации мультиколлинеарности.

Нажимая на кнопку “Calculate and Save” пользователь вызывает алгоритм обработки, отчистки и подготовки дата-сета для дальнейшего использования в приложении. Так же по нажатию на кнопку выполняется построение двух графиков: до и после устранения мультиколлинеарности.

После загрузки дата-сета пользователю отображается первые 5 строк, чтобы он мог проверить правильность и корректность обработки дата-сета. Успешная загрузка дата-сета отображается сообщением в зелёной рамке, после которого можно перейти на второй экран, где уже будут отображаться различные методы анализа. На рисунках 13 и 14 представлены стартовый и финальный экраны приложения в части предсказательной силы различных вариаций логистической регрессии.

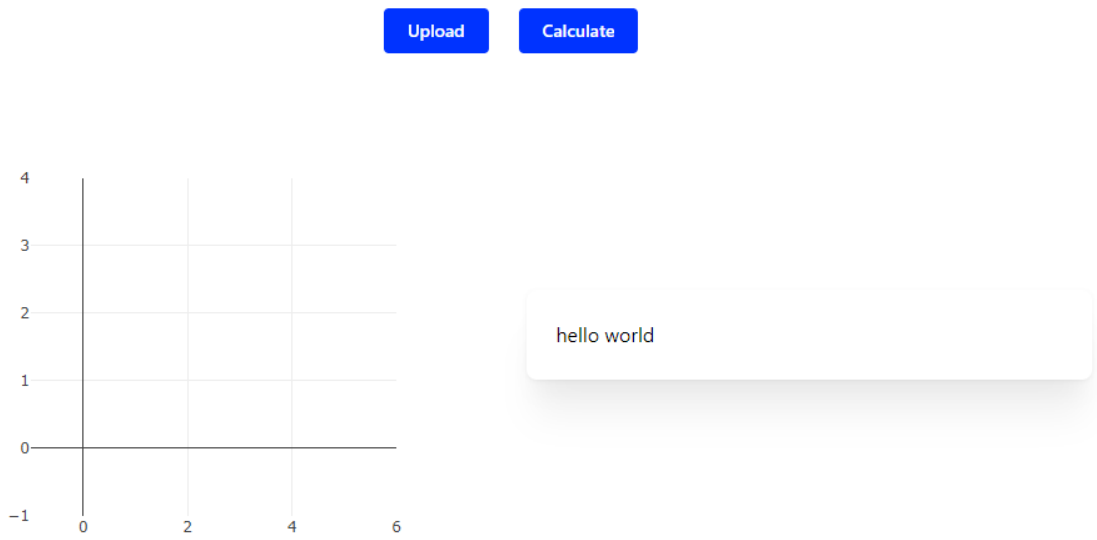

Рисунок 13 – Стартовый экран приложения в части вариантов логистической регрессии

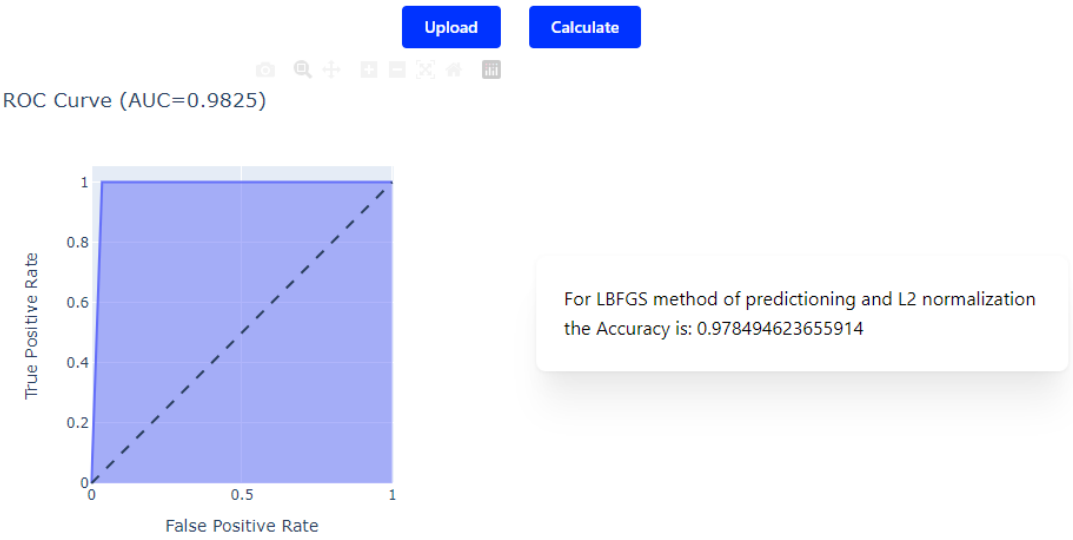

Рисунок 14 – Пример финального экрана приложения в части вариантов логистической регрессии (ROC-анализ)

Таким образом был разобран пользовательский интерфейс приложения, описаны экраны и способы взаимодействия с приложением.

1.5. Пользовательский сценарий

Основным пользовательским сценарием является [14]:

1) загрузка дата-сета;
2) выбор настроек для дата-сета;
3) выбор целевой функции;
4) оценка качества обработки дата-сета;
5) изучение результатов анализа и выбор одного из предложенного метода анализа.

На данном этапе, пользовательский сценарий заканчивается на моменте выбора метода анализа, однако в дальнейшем предполагается расширение функционала до момента, чтобы пользователь мог вносить новые данные и получать на них предсказание на заранее обученной модели.

Алгоритм обработки и движения данных представлен на рисунке 15.

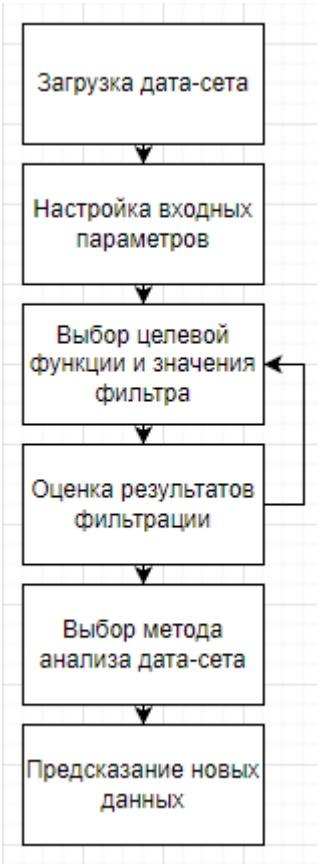

Рисунок 15 – Основной пользовательский сценарий и порядок обработки данных

На рисунке 16 изображен пользовательский интерфейс. Пронумерованы все элементы в порядке взаимодействия пользователя с приложением.

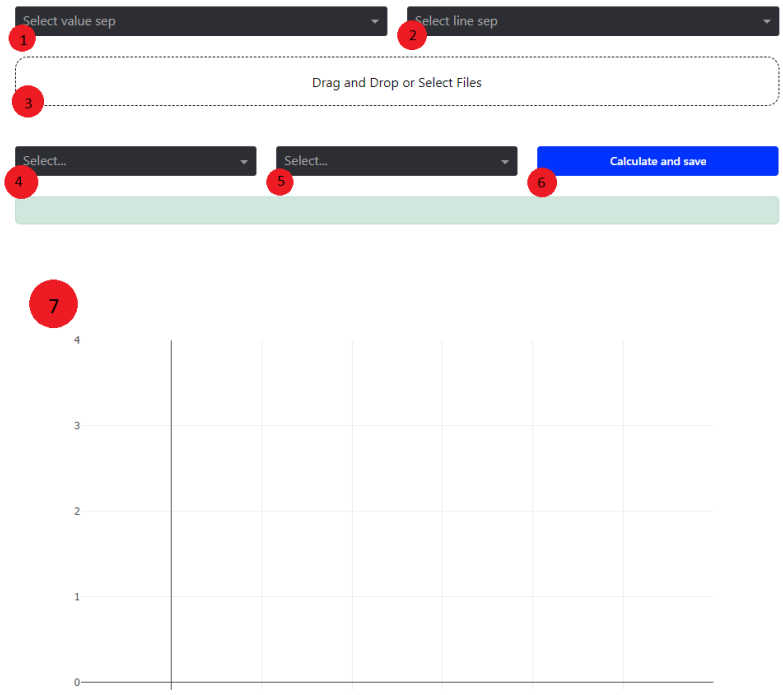

Рисунок 16 – Пользовательский интерфейс

На стартовом экране отображены основные действия пользователя:

1. выбор символа, разделяющего значения.
2. выбор символа, обозначающего окончание строки.
3. загрузка дата-сета через клик мышью или перенести файл в поле.
4. выбор целевой функции.
5. выбор порогового значения для устранения мультиколлинеарности.
6. подтверждение выбора и отображение графиков коллинеарности в блоке №7.

Таким образом расписан основной пользовательский сценарий, в котором участвует интерфейс web-приложения и используются все алгоритмы подготовки данных, создании и обучения прогнозных моделей.

ЗАДАНИЕ ДЛЯ РАЗДЕЛА

«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ПМ9И	Филипасу Ивану Александровичу

Школа	ИШИТР	Отделение школы (НОЦ)	ОИТ
Уровень образования	Магистр	**Направление/специальность**	09.04.04 Программная инженерия

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	*Стоимость материальных ресурсов определялась согласно прейскурантам компаний* *Оклад научного руководителя - 48600 р.* *Оклад исполнителя - 24800 р.*
2. *Нормы и нормативы расходования ресурсов*	*Накладные расходы – 34,8%* *Районный коэффициент 30%;*
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	*Отчисления на уплату во внебюджетные фонды 30,2%*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала НТИ*	*Анализ потенциальных потребителей результатов исследования, оценка качества и перспективности проекта по технологии QuaD, SWOT-анализ*
2. *Разработка устава научно-технического проекта*	*Инициация проекта: определение заинтересованных сторон проекта, целей и результатов проекта*
3. *Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок*	*План проекта, определение трудоемкости выполнения работ, разработка графика проведения научного исследования, расчет бюджета разработки*

4. *Определение ресурсной, финансовой, экономической эффективности*	*Описание потенциального эффекта*
Перечень графического материала *(с точным указанием обязательных чертежей):*	
1. *Оценочная карта для QuaD-анализа разработки* 2. *Иерархическая структура работ проекта* 3. *Диаграмма Ганта* 4. *Матрица SWOT* 5. *График проведения и бюджет НТИ* 6. *Бюджет затрат* 7. *Потенциальные риски*	

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
доцент	Спицына Любовь Юрьевна	к.э.н.		

Задание принял к исполнению студент:

Группа	**ФИО**	**Подпись**	**Дата**
8ПМ9И	Филипас Иван Александрович		

2. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

Диссертация посвящена разработке web-приложения для увеличения эффективности малого и среднего бизнеса посредством предоставления удобного и полезного инструмента для анализа и прогнозирования данных.

Актуальность данного раздела заключается в важности понимания коммерческой составляющей научно-технических проектов и оценки коммерческой ценности разработки.

Целью данного раздела является анализ совокупности факторов, которые определяют коммерческую привлекательность разработки, ее перспективность и успешность.

Основными задачами является оценка перспективности разработки, ее готовности к коммерциализации, выявление потенциальных угроз, а также расчет стоимости и составление графика проведения работ.

2.1. Предпроектный анализ

Задача предпроектного анализа заключается в том, чтобы кратко, но структурированно и понятно, описать основные характеристики разрабатываемого продукта.

Web-приложение позволит представителям малого и среднего бизнеса увеличить производство посредством уменьшения времени на анализ данных. Основная пользовательская история, следующая:

1) Загрузить набор данных
2) Выбрать способ его чтения (тип файла, какой знак отделяет значения и т.д.)
3) Затем выбрать переменную, которую планируется предсказывать, основываясь на тех данных, что загрузил пользователь.
4) Выбор метода анализа, который более эффективен для пользователя.
5) Построение дополнительных графиков и объяснение результатов анализа для пользователя.

Web-интерфейс выбран в качестве реализации программы потому что пользователю не нужно будет скачивать отдельное приложение, а так же все вычисления будут происходить на серверах, что будет снижать нагрузку на устройство и уменьшит требования к вычислительным мощностям пользователя.

Прямых конкурентов у данной разработки нет, однако есть компании, которые предлагают услуги по анализу данных. Приложение же позволит пользователям сохранить конфиденциальность своих данных. Так же приложение может работать с различными предприятиями, что увеличивает охват рынка.

2.1.1. Потенциальные потребители разработки

Целевой рынок для продукта – это любые предприятия малого и среднего бизнесов, которые могут использовать технологии больших данных для предсказания и анализа. Основной сегмент, в который целится данная разработка, средних размеров предприятие, которое обладает достаточным объемом данных, чтобы на их основе было возможно обучить модель и предсказывать полезные данные.

2.1.2. Технология QuaD

Технология QuaD позволяет оценить перспективность разработки на рынке и целесообразность вложения средств в научно-исследовательский проект. Так как прямых конкурентов у разрабатываемого приложения нет, следует тщательно изучить все возможные перспективы работы. Результаты оценки, проведенной в табличной форме, представлены в таблице 2.

Таблица 2 – QuaD-анализ разработки

Критерии оценки	Вес критерия	Средний балл	Макси-мальный балл	Относительное значение (3/4)	Средневзвешенное значение (5x2)
1	2	3	4	5	6
Производительность	0,07	80	100	0,8	0,04
Отказоустойчивость	0,17	90	100	0,9	0,153
Унифицированность	0,1	70	100	0,7	0,07
Безопасность	0,05	80	100	0,8	0,04
Потребность в ресурсах памяти	0,13	95	100	0,95	0,1235
Функциональная мощность	0,1	75	100	0,75	0,075
Простота эксплуатации	0,03	40	100	0,4	0,012
Масштабируемость	0,06	75	100	0,75	0,045
Конкурентоспособ-ность продукта	0,07	50	100	0,5	0,035
Перспективность рынка	0,07	85	100	0,85	0,0595
Цена	0,1	40	100	0,4	0,04
Финансовая эффективность научной разработки	0,07	80	100	0,8	0,056
Итого	1				0,749

По результатам оценки качества и перспективности можно утверждать, что перспективность текущей разработки выше среднего. Улучшить данную разработку можно

путем повышения качества пользовательского интерфейса и снижения уровня сложности эксплуатации.

2.1.3. SWOT-анализ

SWOT-анализ разработанного протокола представляет собой двухэтапный комплексный анализ разработки. Анализ проводится в несколько этапов. Первый этап заключается в описании сильных и слабых сторон, с выявлением возможностей и угроз для реализации проекта. Результаты первого этапа анализа представлены в таблице 3.

Таблица 3 – Матрица первого этапа SWOT разработки

	Сильные стороны разработки: С1. Обработка больших объемов данных С2. Перспектива улучшения обучаемой модели С3. Гибкость разработки С4. Разработка на основе реальных данных С5. Универсальность разработки	**Слабые стороны разработки:** Сл1. Недостаточный уровень универсальности Сл2. Техническая сложность Сл3. Отсутствие достаточного кол-ва данных у пользователя Сл4. Высокая стоимость разработки
Возможности: В1. Потребность в автоматизации прогноза данных для среднего и малого бизнесов. В2. Востребованность инструментов автоматизации процесса обработки данных В3. Большой и свободный рынок для внедрения данного приложения.		

Угрозы: У1. Устаревание предустановленного ПО на серверах потенциального заказчика У2. Изоляция российского сегмента сети Интернет У3. Изменения требований к инструментам интерпретации потенциальным заказчиком У4. Рост требований к вычислительным мощностям оборудования У5. Возможные проблемы с безопасностью данных и коммерческой тайной.		

Второй этап – определение сильных сторон проекта и построении интерактивной матрицы. В рамках данного этапа необходимо построить интерактивную матрицу проекта. Её использование помогает разобраться с различными комбинациями взаимосвязей областей матрицы SWOT. Интерактивная матрица сильных сторон представлена в таблице 4.

Таблица 4 – Сильные стороны проекта

Возможности проекта		C1	C2	C3	C4	C5
	B1	0	+	+	+	+
	B2	+	+	+	+	+
	B3	0	0	-	-	-

Дальнейшим этапом будет построение финальной матрицы SWOT анализа, в котором будут собраны как сильные и слабые стороны, возможности и угрозы проекта, так и связи между сторонами проекта и угрозами, и возможностями. Финальная матрица SWOT-анализа представлена в таблице 5

Таблица 5 – Матрица SWOT разработки

	Сильные стороны разработки: С1. Обработка больших объемов данных С2. Перспектива улучшения обучаемой модели С3. Гибкость разработки С4. Разработка на основе реальных данных С5. Универсальность разработки	**Слабые стороны разработки:** Сл1. Недостаточный уровень универсальности Сл2. Техническая сложность Сл3. Отсутствие достаточного кол-ва данных у пользователя Сл4. Высокая стоимость разработки
Возможности: В1. Потребность в автоматизации прогноза данных для среднего и малого бизнесов. В2. Востребованность инструментов автоматизации процесса обработки данных В3. Большой и свободный рынок для внедрения данного приложения.	Данная разработка является востребованным, гибким, универсальным и удобным инструментом прогнозного анализа, а также, при дальнейшей разработке и улучшении – классификации.	При необходимости масштабирования проекта есть перспектива усложнения ядра системы, требующего дополнительных финансовых и временных затрат.
Угрозы: У1. Устаревание предустановленного ПО на серверах потенциального заказчика У2. Изоляция российского сегмента сети Интернет У3. Изменения требований к инструментам интерпретации потенциальным заказчиком У4. Рост требований к вычислительным мощностям оборудования У5. Возможные проблемы с безопасностью данных и коммерческой тайной.	Изменение требований – наиболее вероятная угроза, так как приложение было разработано и протестировано лишь на наборе данных, связанных с нефтью. Остальные угрозы наименее вероятные, т.к. при разработке ПО было предусмотрена сложность в освоении ПО представителями бизнеса, которые не сталкивались с анализом больших данных. Вычислительные мощности будут наращиваться в зависимости от количества пользователей приложения.	Техническая сложность разработки может повлечь за собой внесение ряда изменений в алгоритмы работы разработанных инструментов интерпретации. Невозможно заранее предугадать способ применения данного приложения и, соответственно, поле деятельности конкретного предпринимателя. Система может некорректно работать на данных, которые до этого не были протестированы.

Данная разработка обладает рядом возможностей в условиях низкой вероятности возникновения угроз. Разработка спроектирована таких образом, что сильные стороны предусматривают изменение требований к анализу, однако разработка изначально предусматривает наличие у пользователя необходимого количества данных для работы приложения, количество данных зависит от конкретного случая и именно поэтому предотвратить данную угрозу проекту становится крайне тяжело.

2.1.4. Оценка готовности разработки к коммерциализации

Одной из важных задач в ходе выполнения данного раздела является оценка готовности разработки к коммерциализации. Оцениваемыми параметрами являются как научная, так и коммерческая составляющая.

Таблица 6 представляет собой бланк оценки степени готовности разработки к коммерциализации.

Таблица 6 – Бланк оценки степени готовности разработки к коммерциализации

№ п/п	Наименование	Степень проработанности разработки	Уровень имеющихся знаний у разработчика
1.	Определен имеющийся научно-технический задел	4	4
2.	Определены перспективные направления коммерциализации научно-технического задела	3	4
3.	Определены отрасли и технологии (товары, услуги) для предложения на рынке	1	3
4.	Определена товарная форма научно-технического задела для представления на рынок	2	2
5.	Определены авторы и осуществлена охрана их прав	3	2
6.	Проведена оценка стоимости интеллектуальной собственности	3	3
7.	Проведены маркетинговые исследования рынков сбыта	2	2
8.	Разработан бизнес-план коммерциализации научной разработки	1	1
9.	Определены пути продвижения научной разработки на рынок	3	4
10.	Разработана стратегия (форма) реализации научной разработки	5	5
11.	Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок	3	3

12.	Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот	4	4
13.	Проработаны вопросы финансирования коммерциализации научной разработки	4	4
14.	Имеется команда для коммерциализации научной разработки	2	3
15.	Проработан механизм реализации разработки	5	5
	ИТОГО БАЛЛОВ:	45/75	49/75

Поскольку данная разработка является индивидуальным проектом для уникального научного эксперимента, не предполагающем дальнейший выход на рынок, коммерциализация данного продукта не является целесообразной. Однако, при должном внимании к данной работе, а также соответствующих инвестициях существует вероятность коммерческого применения данной разработки. По результатам оценки можно утверждать, что данный проект еще не готов к коммерциализации, главным образом, с точки зрения сбыта разработки и финансирования коммерциализации.

2.2. Инициация проекта

В рамках инициации разработки формулируются цели и ожидаемые результаты работы. Также определяются заинтересованные стороны разработки и возможные ограничения. Заинтересованные в данной разработке стороны представлены в таблице 7.

Таблица 7 – Заинтересованные стороны разработки

Заинтересованные стороны	**Ожидания заинтересованных сторон**
Представители малого и среднего бизнеса	Существенное сокращение временных затрат на анализ данных и прогноз.
Data-инженеры	Новый удобный инструмент для подготовки данных и выбора метода анализа.
ТПУ	Новое и полезное исследование, для увеличения рейтинга ВУЗа
Магистрант	Качественное приложение, которое будет являться доказательством полученных знаний

Цели и результат проекта отображены в таблице 8.

Таблица 8 – Цели и результат разработки

Цели разработки:	Разработка web-приложения для увеличения эффективности малого и среднего бизнеса посредством предоставления удобного и полезного инструмента для анализа и прогнозирования данных.
Ожидаемые результаты разработки:	Удобное и эффективное web-приложение, которое позволит сократить время, требуемое для принятия решений
Критерии приемки результата разработки:	Адекватность результатов анализа Web-приложения. Подтверждение эффективности приложение посредством консультаций с представителями бизнеса.
Требования к результату разработки:	**Требования:**
	Сокращение времени, требуемого для принятия решения предпринимателями
	Скорость и точность анализа
	Возможность адаптации к данным из других сфер бизнеса
	Удобный и понятный пользовательский интерфейс

В таблице 9 представлена рабочая группа разработки, определена роль и основные функции каждого участника в разработке.

Таблица 9 – Рабочая группа проекта

п/п	***ФИО, основное место работы, должность***	***Роль в разработке***	***Функции***	***Трудозатраты, час.***
1	*Губин Евгений Иванович, ТПУ ОИТ ИШИТР, доцент*	*Научный руководитель*	*Утверждение основных разделов, выдача заданий к исполнению, координирование деятельности исполнителя*	*29*
2	*Филипас Иван Александрович, ТПУ ОИТ ИШИТР, магистрант гр.8ПМ1И*	*Исполнитель*	*Исполнение поставленных задач*	*228*
ИТОГО:				*257*

Необходимо обозначить ограничения и допущения для проекта, чтобы лучше и четче понимать, в каких рамках будет реализована работа (таблица 10).

Таблица 10 – Ограничения проекта

Фактор	Ограничения / Допущения
Бюджет проекта	Не более 750 000 рублей
Источник финансирования	Средства ТПУ
Сроки проекта	
Дата утверждения плана управления проектом	06.09.2022
Дата завершения проекта	27.05.2023
Прочие ограничения и допущения	Отсутствуют

Данный раздел отражает тот факт, что выполняемая работа имеет довольно большой объем. Заинтересованные стороны проекта ожидают достаточно высококачественные результаты, которые необходимо достичь исполнителю.

2.3. Планирование управления разработкой

2.3.1. Иерархическая структура работ

Иерархическая структура работ для данной разработки представляет собой детализацию укрупненной структуры работ, продемонстрированной на рисунке 17.

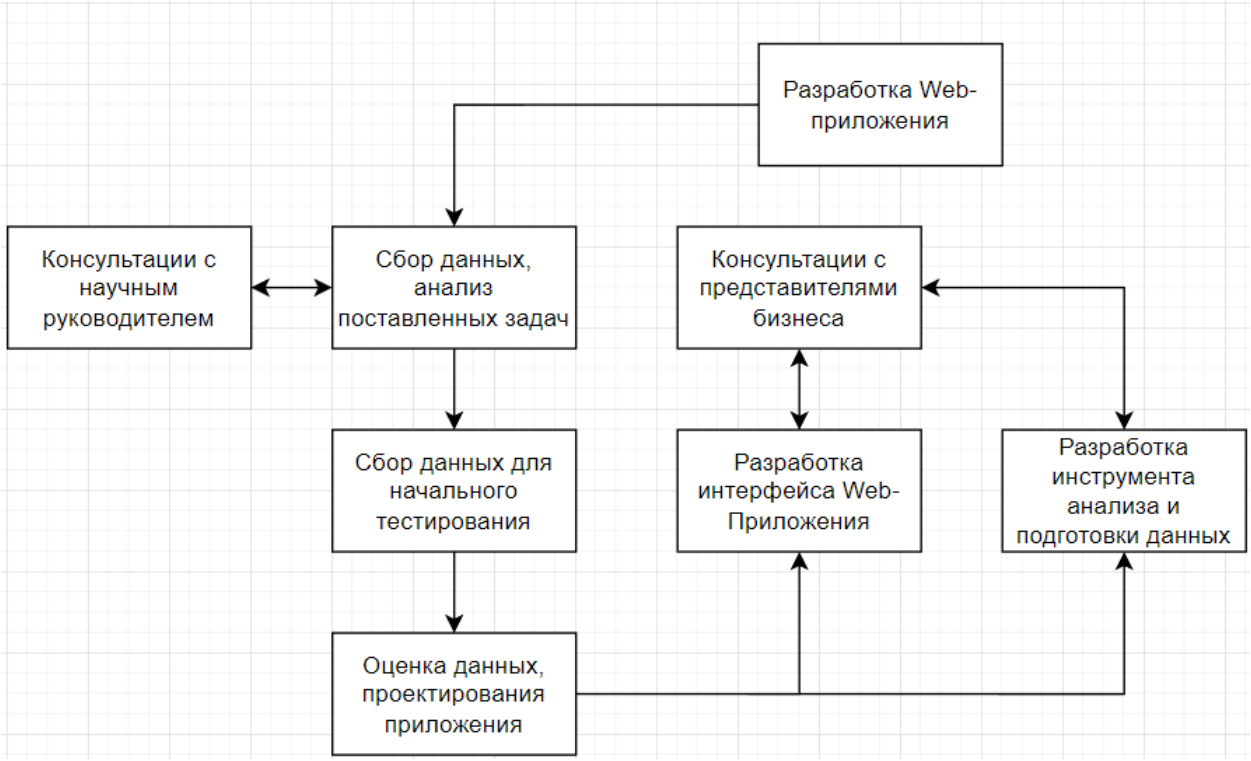

Рисунок 17 - Иерархическая структура работ по проведению разработки

Основная часть работы фокусируется на реализации инструментов и корректировки их работы на основе оценки точности их работы с представителями бизнеса.

2.3.2. План разработки

Чтобы отразить ключевые события по ведению разработки, необходимо составить календарный план. План представлен в таблице 11. Таблица 11 – Календарный план разработки

Код работы	Название	Длительность, дни	Дата начала работ	Дата окончания работ	Состав участников
1	Выбор научного руководителя магистерской работы	1	01.09.22	01.09. 22	Исполнитель
2	Составление и утверждение темы магистерской работы	2	03.09. 22	04.09. 22	Научный руководитель
3	Составление календарного плана- графика выполнения магистерской работы	2	05.09. 22	06.09. 22	Исполнитель и Научный руководитель
4	Выявление требований к разработке	7	07.09. 22	14.09. 22	Исполнитель и Научный руководитель
5	Подбор и изучение литературы по теме магистерской работы	25	15.09. 22	13.10. 22	Исполнитель
6	Анализ предметной области	15	15.10. 22	31.10. 22	Исполнитель
7	Подготовка данных	30	01.11. 22	05.12. 22	Исполнитель
8	Настройка и обучение инструментов	80	06.12. 22	08.03.23	Исполнитель
9	Тестирование	20	09.03.23	01.04.23	Исполнитель
10	Анализ полученных результатов	4	02.04.23	05.04.23	Исполнитель и Научный руководитель

11	Согласование выполненной работы снаучным руководителем	4	06.04.23	10.04.23	Исполнитель и Научный руководитель
12	Выполнение других частей работы (финансовый менеджмент, социальная ответственность)	30	11.04.23	15.05.23	Исполнитель
13	Подведение итогов, оформление работы	10	16.05.23	27.05.23	Исполнитель и Научный руководитель
Итого:		230			

2.3.2.1. Разработка графика проведения разработки

На рисунке 18 представлена диаграмма Ганта с планом выполнения работ, где И (зелёный) – Исполнитель, НР (красный) – научный руководитель.

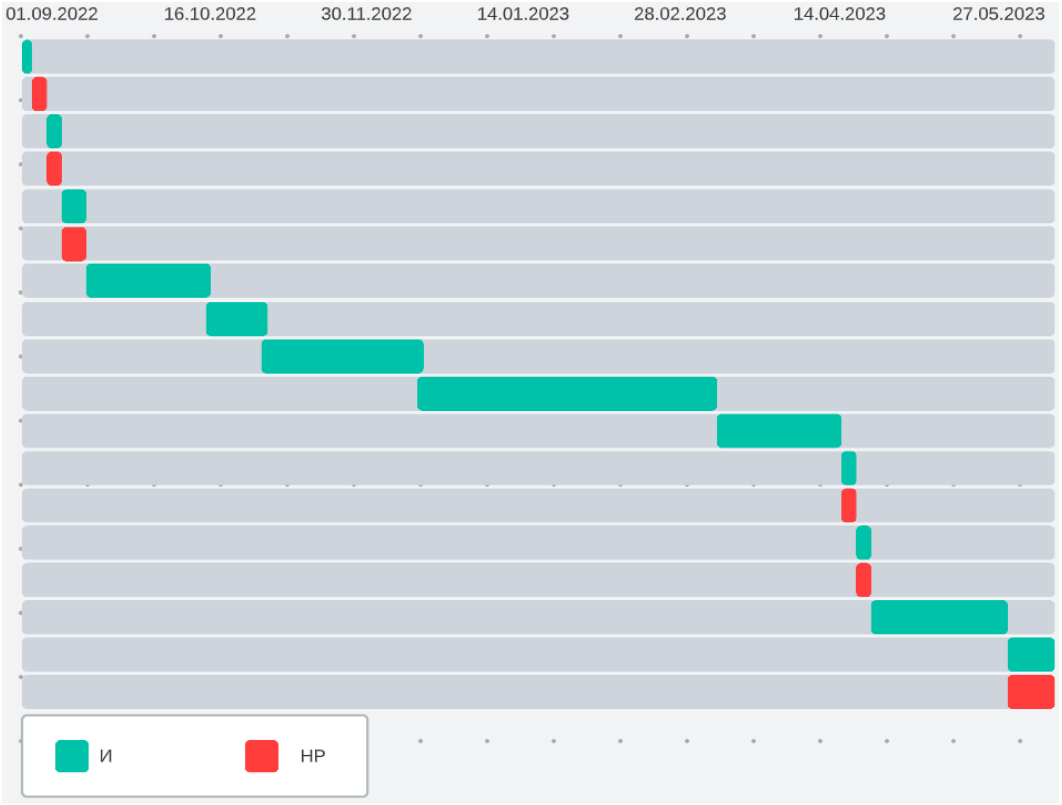

Рисунок 18 - График проведения разработки

2.3.3. Бюджет научного исследования

При планировании бюджета научного исследования должно быть обеспечено полное и достоверное отражение всех видов планируемых расходов, необходимых для его выполнения. Для данной разработки бюджет состоит из следующих пунктов:

- затраты на материалы;
- основная заработная плата исполнителей;
- дополнительная заработная плата исполнителей;
- отчисления во внебюджетные фонды;
- накладные расходы.

2.3.3.1. Материальные расходы

Для проведения исследования какие-либо специальные материалы и комплектующие не приобретались. Единая сумма на канцелярские принадлежности составляет 2500 рублей.

2.3.3.2. Основная заработная плата исполнителей темы

Исполнителями темы выступают научный руководитель и исполнитель. Баланс рабочего времени для 6-дневной недели, по которой учитывается рабочее время преподавателей и студентов, представлен в таблице 12.

Таблица 12 – Баланс рабочего времени

Показатели рабочего времени	***Научный руководитель***	***Исполнитель***
Календарное число дней	365	365
Количество нерабочих дней -выходные дни -праздничные дни	118	118
Потери рабочего времени-отпуск-невыходы по болезни	48	48
Действительный годовой фонд рабочего времени	199	199

Заработная плата рассчитывается из суммы заработной платы исполнителя и научного руководителя исходя из трудоемкости каждого этапа и занятости каждого из них на данном этапе. Расходы по статье заработной плате рассчитываются по следующей формуле:

$$С_{зп} = З_{осн} + З_{доп},$$

где $З_{осн}$ - основная заработная плата; $З_{доп}$ - дополнительная заработная плата. Основная заработная плата одного работника рассчитывается по формуле:

$$З_{осн} = З_{дн} + Т_р,$$

где $З_{дн}$ - среднедневная заработная плата; $Т_р$ - продолжительность работ, выполняемых работником.

Среднедневная заработная плата рассчитывается по следующей формуле:

$$З_{дн} = \frac{З_м \cdot М}{Ф_д},$$

где $З_{дн}$ - месячный должностной оклад работника; М - количество месяцев без отпуска в течение года, при отпуске 48 рабочих дней М = 10,4 месяца, 6 - дневная неделя;

$Ф_д$ - действительный годовой фонд рабочего времени научно-технического персонала.

Месячный должностной оклад работника вычисляется по следующей формуле:

$$З_{дн} = З_б \cdot (к_{пр} + к_д) \cdot к_р,$$

Где

$к_{пр}$ - премиальный коэффициент оплаты труда;

$З_б$ - базовый должностной оклад;

$к_д$ - коэффициент доплат и надбавок (15-20%);

$к_р$ - районный коэффициент (1,3 для Томска).

Для расчета основной заработной платы исполнителя возьмем оклад, равный окладу 24800 рублей. Для расчета основной заработной платы научного руководителя возьмем оклад равный 48600 рублей. В таблице 13 приведены расчеты основной заработной платы.

Таблица 13 - Расчеты основной заработной платы

Исполнители	$З_б$	$к_{пр}$	$к_д$	$к_р$	$З_м$	$З_{дн}$	$Т_р$	$З_{осн}$
Исполнитель	24800	0.3	0.2	1.3	44640	2 333	86	200 633
Научный руководитель	48600	0.3	0.2	1.3	22828	4 572	10	45 718

2.3.3.3. Дополнительная заработная плата исполнителей темы

В данную статью включается сумма выплат, предусмотренных законодательством о труде, например, оплата очередных и дополнительных отпусков; оплата времени, связанного с выполнением государственных и общественных обязанностей; выплата вознаграждения за выслугу лет и т.п. (в среднем — 12 % от суммы основной заработной

платы). Дополнительная заработная плата рассчитывается исходя из 10 –15 % от основной заработной платы, работников, непосредственно участвующих в выполнение темы:

$$З_{доп} = к_{доп} \cdot З_{осн},$$

где $З_{доп}$ - дополнительная заработная плата;

$к_{доп}$ - коэффициент дополнительной заработной платы (на стадии проектирования принимается равным 0,15);

$З_{осн}$ - основная зарплата.

В таблице 14 приведены основной и дополнительной заработной платы исполнителей.

Таблица 14 - Заработная плата исполнителей

Заработная плата	**Научный руководитель**	**Исполнитель**
Основная зарплата	45 718	200 633
Дополнительная зарплата	6 857	30 095
Зарплата исполнителя	52 575	230 728
Итого по статье $С_{зп}$	283 303	

Таким образом, зарплата научного руководителя за период исполнения проекта составляет 52 575 рублей, исполнителя — 230 728 рублей. Всего расходов по статье заработной платы — 283 303 рублей.

2.3.3.4. Отчисления во внебюджетные фонды

В данной статье расходов отражаются обязательные отчисления во внебюджетные фонды по установленным законодательством Российской Федерации нормам органам государственного социального страхования (ФСС), пенсионного фонда (ПФ) и медицинского страхования (ФФОМС) от затрат на оплату труда работников.

Величина отчислений во внебюджетные фонды определяется по следующей формуле:

$$С_{внеб} = к_{внеб} \cdot (З_{осн} + З_{доп}),$$

где $к_{внеб}$ - коэффициент отчислений за уплату во внебюджетные фонды (пенсионный фонд, фонд обязательного медицинского страхования и т.д.).

Общие тарифы страховых взносов в 2023 году в ИФНС: 22% - страхование по временной, 5,1% - медицинское страхование, 2,9% - страхование по временной нетрудоспособности.

Таким образом, отчисления во внебюджетные фонды, исходя из всех вышеперечисленных взносов, составляют:

$$С_{внеб} = 0,3 \cdot 283\,303 = 84\,991 \text{ рублей.}$$

2.3.3.5. Накладные расходы

При выполнении проекта могут возникнуть косвенные издержки - накладные расходы, возникающие дополнительно к основным затратам, например, на консультационные услуги, оплату коммунальных услуг, расход на услуги связи (телефон, Интернет).

Расчет накладных расходов осуществляется по формуле:

$$С_{накл} = к_{накл} \cdot (З_{осн} + З_{доп}),$$

где $к_{накл}$ - коэффициент накладных расходов. Величину коэффициента накладных расходов можно взять в размере 70%.

Сумма накладных расходов составляет:

$$С_{накл} = 0,16 \cdot 283\,303 = 198\,312 \text{ рублей}$$

2.3.3.6. Формирование бюджета затрат на разработку

После выполнения всех расчетов можно определить плановую себестоимость проекта. В таблице 15 представлен бюджет затрат.

Таблица 15 - Бюджет затрат

Наименование	**Сумма, руб**	**Удельный вес, %**
Затраты на материалы	2500	0,4
Затраты на основную заработную плату	246 351	43,2
Затраты на дополнительную заработную плату	36 952	6
Страховые взносы	84 991	15
Накладные расходы	198 312	34.8
Общий бюджет	569 006	100

Исходя из вышеприведенного расчета бюджета следует, что наибольшая его часть приходится на основную и дополнительную заработную платы (49,2%). Стоит также отметить, что расходы на страховые взносы (15%) составляют значительную часть расходов. Затраты на материалы и накладные расходы составляют существенную долю (35%).

2.3.3.7. Риски разработки

Проведение любого научно-исследовательского проекта сопряжено с возникновением различных рисков. Предварительное определение рисков помогает своевременному принятию мер по предотвращению возникновения угроз или минимизации их последствий. В таблице 16 приведена оценка рисков с рекомендациями по смягчению их воздействия.

Таблица 16 - Реестр рисков

№	Риск	Потенциальное воздействие	Вероятность наступления	Влияние риска	Уровень риска	Способы смягчения риска	Условия наступления
1	Несоответствие разработанной и требуемой функциональности	Недостаточная функциональность может привести к неконкурентоспосо-бности решения	2	4	средний	Создание прототипов, разработка сценариев использования, участие потенциальных пользователей	Неправильно поставлены задачи, неполный анализ качества разработки и ее перспективности на рынке
2	Постоянный поток изменений требований	Задержки выполнения работ	3	2	высокий	Установка ограничений для внесения изменений, итеративность разработки (внесения изменений в следующих итерациях)	Ошибки при постановке задачи
3	Технологическое отставание	Неконкурентоспособность устройства	1	2	низкий	Технический анализ, анализ стоимости, прототипирование	Недостаточная оценка существу ющих аналогов
4	Недостаточная производлительность	Неконкурентоспособность устройства	1	3	средний	Проведение сравнительного тестирования, прототипирование	Ошибки при постановке задачи, недостаточный анализ качества разработки и ее перспективности на рынке

2.4. Экономическая эффективность

Показатели экономической эффективности проекта учитывают финансовые последствия для предприятия или компании, которая реализует данный проект.

Определение эффективности происходит на основе расчета интегрального финансового показателя, который рассчитывается следующим образом:

$$I = \frac{Ф_р}{Ф_{max}},$$

где $Ф_р$ - стоимость исполнения работ; $Ф_{max}$ - максимально допустимая стоимость исполнения проекта.

Общий бюджет проекта составил 569 006 рублей. Исходя из ограничений, накладываемых на проект, максимальный бюджет не должен превышать 750000 рублей. Таким образом, значение финансового показателя составляет:

$$I = \frac{569\,006}{750000} = 0{,}758$$

Значение финансового показателя составляет 0,758, что свидетельствует об эффективном использовании финансовых ресурсов.

2.4.1. Интегральный показатель эффективности разработки

Интегральный показатель финансовой эффективности научного исследования получают в ходе оценки бюджета затрат трёх или более вариантов исполнения научного исследования. Расчёт ведётся по формуле:

$$I = \frac{\Phi_{pi}}{\Phi_{max}},$$

Где – I – интегральный финансовый показатель разработки, Φ_{pi} - стоимость исполнения работ; Φ_{max} - максимально возможная стоимость исполнения проекта.

В качестве первого аналога возьмём вариант, при котором проект будет направлен на конкретную сферу предпринимательства, например на нефтегазовую промышленность. Это позволит уменьшить нагрузку на тестирование, так как не нужно будет проверять данные из различных сфер бизнеса.

Второй аналог будет заключать в себя большие затраты на выборку данных и тестировании различных сфер предпринимательства. Это позволит увеличить надежность и удобство эксплуатации, а также увеличит эффективность разрабатываемого продукта.

Сравнительная оценка характеристик всех вариантов исполнения проекта представлена в таблице 17

Таблица 17 - сравнительная оценка характеристик вариантов исполнения проекта

Критерий	Вес	Весовой коэфицент параметра	Текущая реализация	Аналог 1	Аналог 2
Способствует росту производительности труда пользователя		0,1	4	3	5
Удобство в эксплуатации		0,15	3	4	4
Помехоустойчивость		0,15	4	5	3

Энергосбережение	0,2	5	4	2
Надежность	0,25	4	4	3
Материалоёмкость	0,15	5	4	2
ИТОГО	1	25	24	24

Текущая реализация, $I_{\text{ТР}}$= 0,1 * 4 + 0,15 * 3 + 0,15 * 4 + 0.2 * 5 + 0,25 * 4 + 0,15 * 5 = 4,2

Аналог 1, I_{A1}= 0,1 * 3 + 0,15 * 4 + 0,15 * 5 + 0.2 * 4 + 0,25 * 4 + 0,15 * 4 = 4,05

Аналог 2, I_{A2}= 0,1 * 5 + 0,15 * 4 + 0,15 * 3 + 0.2 * 2 + 0,25 * 3 + 0,15 * 2 = 3

Исходя из расчётов получается, что Текущая реализация является самой эффективной, т.к. при первом аналоге уменьшается целевой рынок, следовательно, уменьшается потенциальный доход, а также требуется дополнительно время для разработки и отладки. Второй аналог позволит предоставить более качественный продукт, однако время, необходимое для улучшения эффективности продукта несоизмеримо выше чем у текущей реализации проекта.

2.5. Выводы по разделу

Результаты оценки востребованности проекта можно считать положительными, поскольку были выявлены потенциальные потребители разрабатываемого решения. В результате анализа конкурентоспособности выяснилось, что разработанный проект обладает достаточными конкурентными преимуществами благодаря новизне web-приложения и универсальности. Однако выявлен также высокий риск, связанный с одним из главных преимуществ – универсальности. Существует вероятность, что из-за высокой универсальности приложение не добьётся достаточно эффективного результата, чтобы превзойти обычный человеческий анализ. Для устранения данной угрозы предприняты консультации и демонстрации представителям малого и среднего бизнесов. Проведенный SWOT-анализ показал весьма высокую перспективность разработки.

Продолжительность проекта составила 118 календарных дней, а общий бюджет составил 569 006 рублей, что успешно укладывается в ограничения. Разработанный реестр рисков отражает потенциальные пути преодоления внешних и внутренних рисков, что способствует успешной реализации проекта.

ЗАДАНИЕ ДЛЯ РАЗДЕЛА

«СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	**ФИО**		
8ПМ1И	Филипасу Ивану Александровичу		
Школа		**Отделение (НОЦ)**	
Уровень образования	Магистратура	**Направление/специальность**	09.04.04 «Программная инженерия»

Тема ВКР:

Разработка Web-приложения для выбора метода анализа и модели прогноза для поисковых нефтяных скважин	
Исходные данные к разделу «Социальная ответственность»:	
Введение – Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения – Описание рабочей зоны (рабочего места) при разработке проектного решения	Объект исследования: алгоритмы анализа и интерпретации больших данных Область применения: производство по добыче нефти Рабочая зона: рабочий стол и персональный компьютер
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности при разработке проектного решения: – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	ГОСТ 12.2.032-78 ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования. СП 2.2.3670-20 "Санитарно-эпидемиологические требования к условиям труда” "Трудовой кодекс Российской Федерации" от 30.12.2001 N 197-ФЗ
2. Производственная безопасность при разработке проектного решения: – Анализ выявленных вредных и опасных факторов – Обоснование мероприятий по снижению воздействия	Вредные факторы: - недостаточная освещенность рабочей зоны; - повышенный уровень шума на рабочем месте - отклонение параметров микроклимата; Опасные факторы: - повышенное значение напряжения в электрической цепи
3. Экологическая безопасность при эксплуатации	– Воздействие на литосферу: ликвидация всех замазученных участков, прежде всего, в водоохранных зонах рек и озер; вырубка

	лесов; выбор специальных мест для захоронения отходов (например, отработанные карьеры); – Воздействие на гидросферу: Особое отрицательное воздействие на химический состав водоемов при эксплуатации объектов нефтедобычи оказывают разливы нефти, химических реагентов и вод с высокой минерализацией. При попадании нефти в водоемы на поверхности воды образуется пленка, препятствующая воздушному обмену; – Воздействие на атмосферу: Эксплуатация объектов нефтедобычи связана с выделением загрязняющих веществ в атмосферный воздух.
4. Безопасность в чрезвычайных ситуациях при разработке проектного решения:	Возможные ЧС: - пожар, - землетрясение Наиболее типичная ЧС: - пожар
Дата выдачи задания для раздела по линейному графику	**25.05.2023**

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент	Антоневич Ольга Алексеевна	к.б.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ1И	Филипас Иван Александрович		

3. Социальная ответственность

Разработанный в рамках магистерской диссертации проект является онлайн-инструментом для анализа и выбора метода прогнозирования для конкретного дата-сета пользователя. Web-приложение позволит удаленно планировать, ставить, исполнять и контролировать задачи в проекте, а также отслеживать результаты их выполнения. Представляет интерес для различных проектных организаций.

Разработка Web-приложения велась с использованием компьютерной техники. Использование средств вычислительной техники, накладывает целый ряд вредных факторов на человека, что впоследствии снижает производительность его труда и может привести к существенным проблемам со здоровьем сотрудника. Данный раздел посвящен анализу вредных и опасных факторов производственной среды как для разработчиков, так и для пользователей.

3.1. Правовые и организационные вопросы обеспечения безопасности

3.1.1. Специальные (характерные для проектируемой рабочей зоны) правовые нормы трудового законодательства

Большую часть рабочего времени специалист проводит за ноутбуком. Трудовым кодексом не предусмотрены специальные перерывы для таких видов работ. Однако есть регламентированные перерывы, описанные в СП 2.2.3670-20 «Санитарно-эпидемиологические требования к условиям труда».

Во время рабочего дня специалиста должно быть несколько перерывов суммарно от 50 до 90 минут. Также есть типовая инструкция по охране труда при работе на персональном компьютере (ТОИ Р-45-084-01) [15]. В ней регулируется время непрерывной работы за компьютером – не более 2х часов.

Вышеприведённые нормативы следует учитывать при формировании внутреннего трудового распорядка и перерывов в работе за компьютером. А также необходимо учитывать законы из «Трудовой кодекс Российской Федерации» от 30.12.2001 N 197-ФЗ [16].

3.1.2. Организационные мероприятия при компоновке рабочей зоны

Рабочее место специалиста представляет собой кабинет на втором этаже двухэтажного здания. Площадь кабинета 20 м2. Рабочий стол с ноутбуком расположен сбоку от окна таким образом, что свет от окна падает слева. В помещении отсутствуют силовые кабели и выводы, а также не создаётся дополнительных помех работе ноутбука. Рабочее место в помещении организации должно соответствовать требованиям ГОСТ 12.2.032-78 [17], а также СП 2.2.3670-20 [18]. Оптимальные размеры поверхности стола 1600 х 1000 кв мм. Под столом пространство для ног с размерами по глубине 650 мм.

Рабочий стол должен также иметь подставку для ног, расположенную под углом 15 градусов к поверхности стола. Длина подставки 400 мм, ширинка – 350 мм. Удаленность клавиатуры от края стола должна быть не более 300 мм, что обеспечит удобную опору для предплечий. Расстояние между глазами сотрудника и экраном видеодисплея должно составлять 40-80 см. Помимо этого, рабочий стол должен быть устойчивым, иметь однотонное неметаллическое покрытие, не обладающее способностью накапливать статическое электричество. Рабочий стул должен иметь дизайн, исключающий онемение тела из-за нарушения кровообращения при продолжительной работе на рабочем месте.

3.2. Производственная безопасность

Проводимые в практической части работы подразумевает использование электронной вычислительной машины (ЭВМ). С точки зрения социальной ответственности целесообразно рассмотреть вредные и опасные факторы, которые могут возникать при анализе информационного массива данных в электронном формате, а также требования по организации рабочего места.

3.2.1. Анализ потенциально возможных и опасных факторов, которые могут возникнуть на рабочем месте при проведении исследований

Для выбора факторов использовался ГОСТ 12.0.003-2015 «Опасные и вредные производственные факторы. Классификация» [19]. Перечень опасных и вредных факторов, характерных для проектируемой производственной среды представлен в виде таблицы 18.

Таблица 18 – Опасные и вредные факторы при выполнении исследований

Факторы (по ГОСТ 12.0.003-2015)	Нормативные документы
Недостаточная освещенность рабочей зоны	СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания [20] »
Повышенный уровень шума на рабочем месте	СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» [20]
Неудовлетворительные микроклимат	СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» [20]

	ГОСТ 30494-2011. Здания жилые и общественные. Параметры микроклимата в помещениях [21]
Поражение электрическим током	ГОСТ Р 12.1.019-2009 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты [22]

3.2.2. Недостаточная освещенность рабочей зоны

Источником такого вредного фактора может быть отсутствие естественного освещения или неправильная планировка освещения искусственного. Причиной может быть несвоевременная замена ламп и дросселей.

При неудовлетворительном освещении снижается производительность труда и увеличивается количество допускаемых работником ошибок.

В действительности на рабочем месте специалиста отдела ЭЛКСС и АВС имеется освещение естественное (боковое одностороннее) и искусственное.

Искусственное освещение в помещениях для эксплуатации ЭВМ осуществляется системой общего равномерного освещения. В случаях работы с документами следует применять системы комбинированного освещения (к общему освещению дополнительно устанавливаются светильники местного освещения, предназначенные для освещения зоны расположения документов).

Освещенность на поверхности стола в зоне размещения рабочего документа должна быть 300-500 лк. Освещение не должно создавать бликов на поверхности экрана. Освещенность поверхности экрана не должна быть более 300 лк. В таблице 19 показатели естественного, искусственного и совмещенного освещения помещений жилых зданий.

Таблица 19 – Нормируемые показатели естественного, искусственного и совмещенного освещения помещений жилых зданий

Помещение	Рабочая поверхность и плоскость нормирование КЕО и освещенности (Г	Естественное освещение		Совмещенное освещение		Искусственное освещение		
		КЕО ен, %		КЕО ен, %				
		При верхнем или комбинированном освещении	При боковом освещении	При верхнем или комбинированном освещении	При боковом освещении	Освещенность рабочих поверхностей, лк	Показатель дискомфорта, М, не более	Коэффициент пульсации, Кп, не более

	– горизонтальная) и высота плоскости над полом, м							
Кабинеты	Г – 0,0	3,0	1,0	1,8	0,6	300	-	-

В качестве источников света применяются светодиодные светильники или металлогалогенные лампы (используются в качестве местного освещения).

Проведём расчёт общего равномерного освещения горизонтальной рабочей поверхности. Параметры рабочего кабинета: ширина 4 метра, длина 5 метра, высота потолков 2,9 метра. В кабинете расположено 6 потолочных светильников для фальш-потолка «Армстронг» с 4 люминесцентными лампами в каждом. Для расчёта используем метод светового потока, учитывающий световой поток, отраженный от потолка и стен. Расчётный световой поток, лм, группы светильников с люминесцентными лампами рассчитывается по формуле:

$$\text{Фл. расч} = \frac{\text{Ен} * S * Z * K}{N * n * \eta}$$

Где Ен – нормированная минимальная освещенность, лк;

S – площадь кабинета (20 кв. м)

Z – коэффициент минимальной освещенности; Z = 1,15 для люминесцентных ламп;

K – коэффициент запаса (для люминесцентных ламп 1,1)

N – число светильников;

n – количество ламп в светильнике;

η - коэффициент использования светового потока ламп.

Коэффициент использования светового потока является табличной характеристикой и зависит от типа ламп, коэффициентов отражения светового потока от стен ρС и потолка ρП и индекса помещения. Стены помещения оклеяны белыми обоями и выкрашены в светлые тона, потолок в светло-серый. Для таких условий коэффициенты принимаются по 30%.

Вычислим индекс помещения:

$$I = \frac{S}{h(A+B)} = 0{,}76$$

Где S – площадь кабинета;

h – высота кабинета;

A, B – длина и ширина кабинета.

Для найденных I и коэффициентов отражения, а также типу светильников найдём значение $\eta = 34.5\%$

Вычислим расчётный световой поток для одной лампы:

- для Ен = 200 лм, Фл.расч = 797 лм.

- для Ен = 400 лм, Фл.расч = 1584 лм.

Освещенность в допустимых нормах, так как используются лампы OSRAM L18W/640 G13 со световым потоком 1200 лм каждая.

3.2.3. Повышенный уровень шума на рабочем месте

Источником такого вредного фактора могут быть автомобили и другие транспортные средства, проезжающие рядом со зданием и окном. Также сам рабочий ноутбук издаёт шум во включенном состоянии.

Чем длительнее воздействие шума на человека, тем негативнее он влияет на физическое и психическое здоровье. Длительное воздействие шума, уровень которого равен 68-92 дБ, становится причиной возникновения определенных заболеваний нервной системы.

Характер шума – широкополосный с непрерывным спектром более 1 октавы. В таблице 20 указаны предельные допустимые уровни звукового давления из СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» [20].

Таблица 20 – Предельно допустимые уровни звукового давления, уровни звука и эквивалентные уровни звука для основных наиболее типичных видов трудовой деятельности и рабочих мест

Вид трудовой деятельности, рабочее место	Уровни звукового давления, дБ, в октавных полосах со среднегеометрическими частотами, Гц								Уровни звука и эквивалентные уровни звука (дБА)
Творческая деятельность, руководящая работа с повышенными требованиями, научная деятельность, конструирование и проектирование, программирование,	31,5	63	125	250	500	1000	2000	4000	
	86	71	61	54	49	45	42	40	50

преподавание и обучение, врачебная деятельность. Рабочие места в помещениях дирекции, проектно-конструкторских бюро, расчетчиков, программистов вычислительных машин, в лабораториях для теоретических работ и обработки данных, приема больных в здравпунктах.									

Уровень шума в рабочем кабинете специалиста отдела ЭЛКСС и АВС не более 80 дБА и соответствует нормам.

В качестве дополнительных мер по уменьшению влияния шума предлагается: заменить окна на более современные, со стеклопакетами, устранить щели и зазоры в стояках труб отопления, повесить шторы.

3.2.4. Отклонения параметров микроклимата

Причиной плохого микроклимата могут быть:

- нарушения в работе систем отопления,
- нарушения в работе систем вентиляции, кондиционирования и сплит-систем,
- загрязненность помещения

Параметры микроклимата оказывают непосредственное влияние на тепловое состояние человека. Недостаточная влажность, в свою очередь, может негативно отражаться на организме, становясь причиной пересыхания и растрескивания кожи и слизистой, а также последующего заражения болезнетворными микроорганизмами.

Длительное воздействие высокой температуры при повышенной влажности может привести к гипертермии или накоплению теплоты и перегреву организма, а пониженные показатели температуры, особенно при повышенной влажности, могут быть причиной гипотермии или переохлаждения.

Все параметры микроклимата прописаны в СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» [20] Работа по проектированию соответствует категории Iб. Для данной категории характерны следующие параметры микроклимата:

– Влажность должна быть в диапазоне 15-17%.

– Температура должна быть в диапазоне от 19 градусов до 28 градусов.

– Движение воздуха должно быть в диапазоне от 0,1 до 0,2.

Для создания и автоматического поддержания в рабочем кабинете специалиста отдела ЭЛКСС и АВС независимо от условий оптимальных значений температуры, влажности, чистоты и скорости движения воздуха используется водяное отопление, в теплое время года применяется кондиционирование воздуха.

В кабинетах проводится ежедневная влажная уборка и систематическое проветривание после каждого часа работы на ЭВМ.

3.2.5. Поражение электрическим током

Электрический ток оказывает на организм человека термическое, электролитическое и биологическое действие. Термическое действие тока проявляется в ожогах отдельных участков тела, а также в нагреве до высоких температур других органов. Электролитическое действие тока проявляется в разложении органических жидкостей, вызывая значительные нарушения их физико-химического состава. Биологическое действие тока проявляется в раздражении и возбуждении живых тканей организма, а также в нарушении внутренних биоэлектрических процессов.

Согласно ПУЭ-7 [23], действующим в РФ помещение кабинета специалиста отдела ЭЛКСС и АВС относится к категории без повышенной опасности поражения электрическим током.

Защита высоковольтных и низковольтных цепей от коротких замыканий и перегрузок, также при отказах оборудования должна предотвращать появления пожароопасных режимов и повреждений оборудования.

Внутренние поверхности шкафов, ниш электрощитов, аппаратных отсеков не должны распространять пламя.

Помещение, в котором проводилась исследовательская работа, полностью соответствует требованиям к электрооборудованию. Никаких мер по устранению несоответствий требований проводиться не будет. Тем не менее, в обязательном порядке проводится вводный, плановый и внеочередные инструктажи со специалистами.

3.3. Экологическая безопасность при эксплуатации

Основными типами антропогенных воздействий на природу, являются:

– нефтяное и химическое загрязнение окружающей среды вследствие несовершенства технологии, аварийных разливов и несоблюдение природоохранных требований;

– загрязнение атмосферы от испарений нефтепродуктов при их нагреве для проведения исследований;

– загрязнение природной среды промышленными, бытовыми и лабораторными отходами.

И как следствие от вышеотмеченных воздействий на природу:

– сокращение ареалов редких видов растений, площадей, занятых ягодниками, лекарственными растениями и другими ценными видами флоры;

– нарушение лесов и нерациональный расход древесины при обустройстве передвижных поселков, временных дорог, промплощадок и др.;

– сокращение рыбных запасов вследствие загрязнения поверхностных вод, нарушения гидрологического режима при строительстве и эксплуатации месторождений.

Общими мерами по охране окружающей среды являются:

– сокращение потерь нефти и газа; повышение герметичности и надежности нефтепромыслового оборудования;

– оптимизация процессов сжигания топлива при одновременном снижении образования токсичных продуктов сгорания.

3.3.1. Защита атмосферы

Основными причинами аварий являются:

– некачественное строительство, ремонт нефтепромыслового оборудования;

– -еханические повреждения;

– несоблюдение техники безопасности.

Основные мероприятия по охране атмосферного воздуха от загрязнений:

– защита оборудования от коррозии;

– применение оборудования заводского изготовления;

– разработанный план действий при аварийной ситуации;

– ликвидация аварий должна осуществляться аварийной службой.

Чистота атмосферного воздуха обеспечивается путем сокращения абсолютных выбросов газов и обезвреживанием выбросов, содержащих вредные вещества (Таблица 21).

Таблица 21 – Вредные вещества

№	Наименование загрязняющих веществ	ПДК м.р. в воздухе населенных мест, мг/м3	Класс опасности	Параметры выбросов	
				г/сек	т/год
1	Двукись азота	0.085	2	0.078	1.230
2	Окись углерода	5.000	4	0.220	4.88
3	Углеводороды	300	4	9.140	298.8
4	Сажа	0.15	3	0	2
5	Метанол	1	3	0.041	1.290

3.3.2. Защита гидросферы

Особое отрицательное воздействие на химический состав водоемов при эксплуатации объектов нефтедобычи оказывают разливы нефти, химических реагентов и вод с высокой минерализацией, а также утилизация остатков химических реагентов. При попадании нефти в водоемы на поверхности воды образуется пленка, препятствующая воздушному обмену.

Пути попадания токсичных загрязнений в природные воды:

– загрязнение грунтовых вод в результате отсутствия гидроизоляции технологических площадок;

– поступление нефти и химических реагентов в подземные воды в результате перетоков по затрубному пространству при некачественном цементировании скважины и ее не герметичности.

Мероприятия по рациональному использованию и охране водных ресурсов:

– установление и поддержание водоохранных зон;

– вынесение объектов из экологически уязвимых зон;

– герметизированная системы закачки химического реагента;

– отсыпка кустовых площадок с учетом поверхностной системы стока;

– сбор разлившихся нефтепродуктов в аварийную емкость с последующей перекачкой на установку подготовки нефти (УПН);

– осуществлять биологическую очистку хозяйственно-бытовых стоков.

3.3.3. Защита литосферы

Загрязнение почв нефтью и химическими реагентами приводит к значительному экологическому и экономическому ущербу: понижается продуктивность лесных ресурсов, ухудшается санитарное состояние окружающей среды.

При выборе площадок и трасс под строительство объектов основным критерием является минимальное использование лесов I и II групп, пойменной части рек и озер, а также обход кедровников, путей миграции животных и птиц. Принимается прокладка линейных сооружений (автодорог, трубопроводов, линий электропередач) в одном коридоре, что обеспечивает снижение площади занимаемых земель на 30-40%.

Согласно требованиям лесного хозяйства организации, выполняющие строительные работы обязаны:

– обеспечить минимальное повреждение почв, травянистой и моховой растительности;

– произвести очистку лесосек и ликвидировать порубочные остатки;

– не допускать повреждения корневых систем и стволов опушечных деревьев;

– не оставлять пни выше 1/3 диаметра среза, а при рубке деревьев больше 30 см - выше 10 см, считая высоту шейки корня.

Рекультивация нарушенных земель по трассам линейных трубопроводов носит природоохранное направление и выполняется в два этапа:

– технический этап рекультивации состоит из сбора пролитой нефти, срезки почвенно-растительного слоя толщиной 0.2-0.4 м и перемещения его во временные отвалы до начала строительных работ;

– биологический этап рекультивации включает дискование почвы боронами в один след, поверхностное внесение минеральных удобрений и посев многолетних трав механическим способом.

Предотвращение аварийных разливов нефти и химических реагентов обеспечивается:

– контролем давления в общем коллекторе и замерном сепараторе с сигнализацией предельных значений на замерных установках (ЗУ);

– в случае аварии на УПН автоматическим переключением потока нефти в аварийные емкости;

– аварийным отключением насосных агрегатов на УПН и узлах дозирования ингибиторов;

– закреплением трубопроводов на проектных отметках грузами и анкерами, препятствующими всплытию и порыву;

– прокладкой трубопроводов в кожухах через автомобильные дороги; контролем качества сварных швов трубопроводов методом радиографирования и магнитографирования и гидравлическое испытание на прочность и герметичность.

3.4. Безопасность в чрезвычайных ситуациях при разработке проектного решения

При проведении исследований наиболее вероятной ЧС является возникновение пожара в помещении организации. Требования к пожарной безопасности прописаны в ГОСТ 12.1.004-91 [24] и ГОСТ 12.4.026-2015 [25]. Пожарная безопасность должна обеспечиваться системами предотвращение пожара и противопожарной защиты, в том числе организационно-техническими мероприятиями. Основные источники возникновения пожара:

– Неработоспособное электрооборудование, неисправности в проводке, розетках и выключателях. Для исключения возникновения пожара по этим причинам необходимо вовремя выявлять и устранять неполадки, а также проводить плановый осмотр электрооборудования.

– Электрические приборы с дефектами. Профилактика пожара включает в себя своевременный и качественный ремонт электроприборов.

– Перегрузка в электроэнергетической системе и короткое замыкание в электроустановке.

– Под пожарной профилактикой понимается обучение пожарной технике безопасности и комплекс мероприятий, направленных на предупреждение пожаров. Пожарная безопасность обеспечивается комплексом мероприятий:

– Обучение, в т.ч. распространение знаний о пожаробезопасном поведении;

– Пожарный надзор, предусматривающий разработку государственных норм пожарной безопасности и строительных норм, а также проверку их выполнения;

– Обеспечение оборудованием и технические разработки.

В соответствии с ТР «О требованиях пожарной безопасности» для административного жилого здания требуется устройство внутреннего противопожарного водопровода.

Согласно ФЗ-123 [26], НПБ 104-03 «Проектирование систем оповещения людей о пожаре в зданиях и сооружения» [27] для оповещения о возникновении пожара в каждом

помещении должны быть установлены дымовые оптико-электронные автономные пожарные извещатели, а оповещение о пожаре должно осуществляться подачей звуковых и световых сигналов во все помещения с постоянным ли временным пребывание людей.

Помещения организации оборудованы средствами пожаротушения: огнетушителями ОУ-3, ОП-3 (предназначены для тушения любых материалов, предметов и веществ, применяется для тушения ПК и оргтехники)

В помещениях организации имеется пожарная автоматика, сигнализация. В случае возникновения загорания необходимо обесточить электрооборудования, отключить систему вентиляции, принять меры тушения (на начальной стадии) и обеспечить срочную эвакуацию студентов и сотрудников в соответствии с планом эвакуации.

3.5. Выводы по разделу

Значение всех производственных факторов на изучаемом рабочем месте соответствует нормам, которые также были продемонстрированы в данном разделе, за исключением фактора, обладающего свойствами психофизиологического воздействия на организм человека. Для минимизации влияния данного фактора на организм человека, достаточно соблюдать меры, приведенные в МР 2.2.9.2311 – 07 “Профилактика стрессового состояния работников при различных видах профессиональной деятельности” [28].

Категория помещения по электробезопасности, согласно ПУЭ, соответствует первому классу – «помещения без повышенной опасности».

Согласно правилам по охране труда при эксплуатации ЭВМ персонал должен обладать I группой допуска по электробезопасности. Присвоение группы I по электробезопасности производится путем проведения инструктажа, который должен завершаться проверкой знаний в форме устного опроса и (при необходимости) проверкой приобретенных навыков безопасных способов работы или оказания первой помощи при поражении электрическим током. Категория тяжести труда в лаборатории по СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» [20] относится к категории Iб (работы, производимые сидя, стоя или связанные с ходьбой и сопровождающиеся физическим напряжением).

Помещение рабочего места относится к категории помещения группы А, возможный класс пожара В. Рассмотренный объект исследования, оказывающий незначительное негативное воздействие на окружающую среду, относится к объектам III категории.

Заключение

В результате проведённого исследования было создано Web-приложение, которое выполняет функции автоматической подготовки дата-сета для дальнейшей работы аналитика, что уменьшает трудоёмкость, позволяя выполнять данную работу менее квалифицированному специалисту.

Разработан пользовательский интерфейс приложения, позволяющий воспользоваться методами автоматической подготовки дата-сета и алгоритмами анализа. Разобран основной пользовательский сценарий, в котором участвует интерфейс, используются все алгоритмы подготовки данных, создание и обучение прогнозных моделей.

Созданное web-приложение является уникальным аналогом таких программных продуктов как SAS, однако, в отличии от конкурента, позволяет пользоваться продуктом менее обученному специалисту, а также значительно дешевле в использовании и освоении.

Помимо разобранного дата-сета, представляющего собой информацию про нефтяные скважины, данным web-приложением можно с большой эффективностью пользоваться представителям различных сфер бизнеса, от маркетинга до рекламы.

Список публикаций

По результатам дипломной работы была опубликована статья в XX Международной научно-практической конференции студентов, аспирантов и молодых ученых «Молодежь и современные информационные технологии» (МСИТ).

Список используемых источников

[1] Методика подготовки больших данных для прогнозного анализа// Научная электронная библиотека URL: https://www.elibrary.ru/item.asp?id=43024834 (дата обращения: 31.05.2023).

[2] Влияние неподготовленных исходных данных на прогнозный анализ // Научная электронная библиотека URL: https://www.elibrary.ru/item.asp?id=50378380 (дата обращения: 31.05.2023).

[3] Документация к Dash\Plotly // plotly URL: https://dash.plotly.com/ (дата обращения: 31.05.2023).

[4] Optimizing Neural Networks with LFBGS in PyTorch // Личный блог Йоханеса Гаупта URL: https://johaupt.github.io/blog/pytorch_lbfgs.html (дата обращения: 31.05.2023).

[5] Logistic Regression on a Large Data Set // KoalaTea URL: https://koalatea.io/large-data-logistic-regression-sklearn/ (дата обращения: 31.05.2023).

[6] Logistic regression via LiblineaR // parsnip URL: https://parsnip.tidymodels.org/reference/details_logistic_reg_LiblineaR.html (дата обращения: 31.05.2023).

[7] Solving Logistic Regression with Newton's Method // TLP URL: https://thelaziestprogrammer.com/sharrington/math-of-machine-learning/solving-logreg-newtons-method (дата обращения: 31.05.2023).

[8] Глубокое погружение в ROC-AUC // PythonRu URL: https://pythonru.com/baza-znanij/sklearn-roc-auc (дата обращения: 31.05.2023).

[9] Документация к библиотеке Pandas // Pandas URL: https://pandas.pydata.org/ (дата обращения: 31.05.2023).

[10] Документация к библиотеке Numpy // numpy URL: https://numpy.org/ (дата обращения: 31.05.2023).

[11] Regularization in Machine Learning // Towards Data Science URL: https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a (дата обращения: 31.05.2023).

[12] L1 и L2 регуляризация // Python school URL: https://python-school.ru/blog/regularization-l1-l2/ (дата обращения: 31.05.2023).

[13] Elastic Net Regularization // Open Genus URL: https://iq.opengenus.org/elastic-net-regularization/ (дата обращения: 31.05.2023).

[14] Пользовательские сценарии: что это, как и для чего их нужно строить // Laba URL: https://l-a-b-a.com/blog/2651-polzovatelskie-scenarii-chto-eto-kak-i-dlya-chego-ih-nuzhno-stroit (дата обращения: 31.05.2023).

[15] Российская Федерация. Законы. Типовая инструкция по охране труда при работе на персональном компьютере. ТОИ Р-45-084-01 (с изм. и доп., вступ. в силу с 01.01.2021) — URL: https://files.stroyinf.ru/Index2/1/4293792/4293792052.htm (дата обращения 24.03.2022)

[16] Российская Федерация. Законы. Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 25.02.2022) (с изм. и доп., вступ. в силу с 01.03.2022) — URL: http://www.consultant.ru/document/cons_doc_law_34683/ (дата обращения 24.03.2022)

[17] ГОСТ 12.2.032-78 Система стандартов безопасности труда. Рабочее место при выполнении работ сидя. Общие эргономические требования. Режим доступа: https://allgosts.ru/13/180/gost_12.2.032-78 (дата обращения: 17.05.2023)

[18] СП 2.2.3670-20 «Санитарно-эпидемиологические требования к условиям труда» (с изменениями на 30 декабря 2022 года) Режим доступа: https://www.rospotrebnadzor.ru/files/news/SP2.2.3670-20_trud.pdf (дата обращения: 31.05.2023)

[19] ГОСТ 12.0.003-2015 Система стандартов безопасности труда. Опасные и вредные производственные факторы. Классификация Режим доступа: https://allgosts.ru/13/100/gost_12.0.003-2015 (дата обращения: 17.05.2023)

[20] СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания. Режим доступа: http://docs.cntd.ru/document/901704046 (дата обращения: 29.05.23)

[21] ГОСТ 30494-2011 «Здания жилые и общественные. Параметры микроклимата в помещениях» Режим доступа: https://rags.ru/gosts/gost/52219/ (дата обращения: 29.05.23)

[22] ГОСТ Р 12.1.019-2009 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты (с изменениями на 13 июля 2017 года) Режим доступа: https://ohranatruda.ru/ot_biblio/standart/191676/ (дата обращения: 31.05.2023)

[23] Правила устройства электроустановок (ПУЭ-7) Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_98464/ (дата обращения: 31.05.2023)

[24] ГОСТ 12.1.004-91 Система стандартов безопасности труда. Пожарная безопасность. Общие требования Режим доступа: https://docs.cntd.ru/document/9051953 (дата обращения: 31.05.2023)

[25] ГОСТ 12.4.026-2015 Система стандартов безопасности труда (ССБТ). Цвета сигнальные, знаки безопасности и разметка сигнальная. Назначение и правила применения. Общие технические требования и характеристики. Методы испытаний (с Поправками, с Изменением N 1) Режим доступа: https://tiflocentre.ru/download/gost-12-4-026-2015.pdf (дата обращения: 31.05.2023)

[26] Российская Федерация. Законы. Федеральный закон «Технический регламент о требованиях пожарной безопасности» от 22.07.2008 N 123-ФЗ (последняя редакция) Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_78699/ (дата обращения 01.06.2023)

[27] Приказ МЧС РФ от 20.06.2003 N 323 (ред. от 07.02.2008) «Об утверждении норм пожарной безопасности. Проектирование систем оповещения людей о пожаре в зданиях и сооружениях (НПБ 104-03)» (Зарегистрировано в Минюсте РФ 27.06.2003 N 4837) Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_43036/64bd720fc0f589ab6edcfc33c309562a06321645/ (дата обращения 01.06.2023)

[28] «МР 2.2.9.2311-07. 2.2.9. Состояние здоровья работающих в связи с состоянием производственной среды. Профилактика стрессового состояния работников при различных видах профессиональной деятельности. Методические рекомендации» (утв. Главным государственным санитарным врачом РФ от 18.12.2007) (вместе с «Методикой психической саморегуляции») Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_83834/ (дата обращения 05.06.2023)

Приложение I

(справочное)

Development of a web application for selecting an analysis method and forecast model for exploratory oil wells

Студент

Группа	ФИО	Подпись	Дата
8МП1И	Филипас Иван Александрович		01.06.2023

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е.И.	к.ф.-м.н.		01.06.2023

Консультант-лингвист отделения иностранных языков ШБИП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИЯ	Уткина А. Н.	к. филос. н		01.06.2023

Development of a web application for selecting an analysis method and forecast model for exploratory oil wells

The process of collecting and preparing the initial data is one of the most time-consuming and complex stages in the analysis of large amounts of data, which sometimes takes up to 80% of the total time. The process of preparing the initial data includes the following steps:

- translation of the original file from the original format .csv to format .xlsx (either sas or Python)
- checking the source data for errors and typos ("typo");
- checking the source data for missing values ("missing");
- checking the source data for outliers ("outliers");
- checking the source data for duplicate rows (observations);
- checking the source data for multicollinearity;
- transformation of the source data into a digital format ("digitalization")
- selection of the target variable.

The resulting technique can be implemented in Python, SAS, SAS Enterprise Miner, Excel software packages.

Taking into account that most of the real data is poorly structured or even taken from the storage of the “data lake”, the question of the correctness of the latter is critical. Suffice it to say that with careful and correct preparation of the initial data, it is possible to increase the accuracy and predictive power of traditional predictive models by almost 18%.

Each stage of the preparatory analysis and data processing requires serious involvement of the analyst in the business sphere, but some of the work can be done automatically. For example, such work as checking the source data for errors or typos, missing values, incorrect data format can be transferred to formal execution in the appendix below.

As basic processing methods, experts suggest using the following recommendations for "cleaning" the source data in the presence of missing or erroneous data:

- if the amount of missing data does not exceed 5%, then while maintaining representativeness, these lines with "missing" can be deleted;
- if the amount of missing data exceeds 50%, then this attribute can be removed from further analysis;
- if the amount of missing data is in the range of 5% - 50%, then for numerical attributes ("numeric") there are several options for replacing the missing values: average ("mean"), median ("median"), "nearest neighbors", etc. For text attributes ("char"), it is possible to use values that are most common or "nearest neighbors".

Before starting the preparation ("cleaning") of the source data, it is advisable to carry out the "descriptive statistics" procedure of the data set for all incoming variables. This will help to give a general idea of the source data and determine the input data and the target variable.

This usually includes the following statistics:

- for each numerical input variable (attribute- "Numeric")
    - total number of rows (observations), including missing ones ("N");
    - minimum value ("Min");
    - maximum value ("Max");
    - Mean square deviation ("St.div");
    - mean ("Mean");
    - median ("Median"),
- for a text variable (attribute – "Char"):
    - total number of rows (observations), including missing ones ("N");
    - frequency of incoming parameters ("Frequency") and their number ("N").

The next stage of processing the data set is the formation of a training-and-test sample. Their content and volume in relation to the entire volume of source data is usually 70:30 and a mandatory requirement of representativeness. If several learning models are used and there is a lot of data, then a validating sample is added in the ratio 60:20:20.

Since the process of data preparation is extremely time-consuming and requires a highly qualified analyst, in this paper a study was conducted on the possible simplification and acceleration of work with data sets by creating a web application.

The implemented application automates most of the routine work and will allow more analysts to use analysis technologies, thereby speeding up and simplifying their work.

The application automatically processes missing values, data outliers, duplicate strings, and also checks and, if necessary, changes the data format (converts from alphabetic format to numeric). In addition, the application eliminates multicollinearity. The implemented application automatically splits the original data set into two samples: training and test, in a ratio of 70:30, respectively.

In addition to the described functionality, the application includes four variants of predictive models that are based on logistic regression. The paper presents three types of logistic regression:

- based on the prediction method;
- based on the regularization method;
- with an additional regularization coefficient.

It is necessary to elaborate on the methods of prediction and regularization that were used in the study.

The first LBFGS method is an optimization algorithm that uses the Broyden—Fletcher–Goldfarb–Shanno optimization method, which belongs to quasi–Newtonian methods. The "lbfgs" solver is recommended for small datasets, as its performance suffers for large datasets.

The second method of the SAGA algorithm is a variant that supports various types of regularization. This is the preferred algorithm for large amounts of data, with a high density of measurements.

The third method of LIBLINEAR is an algorithm using the coordinate descent method and relies on a library from the C++ language, which is supplied together with the scikit–learn library used in this study. The algorithm implemented in liblinear solves the optimization problem with the help of decomposition on the principle of "one against the rest".

The fourth method implemented in the application is newton-cholesky. This method is a good choice for those cases in which the number of dimensions is several times greater than the number of variables in the data set. It is recommended to use this method only in cases where logistic regression is used and only for working with categorical variables.

Also in the work, an important part is the regularization methods or penalties adopted in the English-language literature. As part of the study, 4 different regularization options were analyzed: L1, L2, elastinet and lack of regularization. Regularization is responsible for ensuring that the model is not retrained or, conversely, is not insufficiently trained. L1 is mathematically opposite in meaning to L2, however, unlike L2, weights can become zero, with a very large coefficient value. Elastinet combines two regularizations.

The criterion for evaluating the accuracy of models is the ROC\AUC curve and the Accuracy parameter. The accuracy of the model means how close the predicted result is in value to what was included in the training sample. ROC is a graph that shows the effectiveness of the machine learning model in solving the classification problem by displaying the frequency of true and false values. AUC is the area under the curve. The ROC curve shows the ratio of true positive and false positive results.

The implementation as a Web application was chosen for simplicity of operation and cost reduction for the user, namely:

- all calculations take place on the server side of the program;
- you do not need to download the application and have large amounts of computing power to use this application.

Also, the advantages of the web implementation will be that the application:

- accessible from any device connected to the internet;

- intuitive for the user.

Python was chosen as the main language of software product development, for the following reasons:

- this language is the most popular for working with big data;
- a large number of libraries and functions have been written for this language, which make it possible to process data sets more efficiently than other programming languages;
- was taught throughout the entire training program.

NumPy and Pandas libraries were selected for work and subsequent data processing. They allow you to work with big data and work with a data set both with one object and to access each individual cell.

The sklearn library was chosen for regression. This library allows you to quickly and easily use various types of logistic and linear regression, as well as random forest techniques, etc.

Dash\Plotly was chosen as the front-end implementation tool. This tool allows you, without learning the markup language and styles, using only Python code, to easily create Web applications and an interface for interaction. Dash\Plotly consists of two main parts. The Dash part is responsible for the markup, various components and elements, buttons and styling of the appearance of the page. Plotly is responsible for the graphics and rendering of various graphical and interactive objects.

The following algorithm is assumed to be the main user scenario when working with a Web application:

- loading a data set;
- selection of settings for the data set;
- target function selection;
- evaluation of the quality of data set processing;
- study of the results and the choice of one of the proposed methods of analysis.

As the results of the analysis, the user is presented with descriptive statistics of the downloaded data set, which includes the following values:

- number of measurements;
- average value;
- standard deviation;
- minimum value;
- maximum value;
- 25, 50 and 75 percentiles, as well as the data type.

In addition to the statistics described above, the user will receive two heat maps of parameter correlation: the first map – before multicollinearity is eliminated, the second – after.

The user interface is one of the most important parts of the application, as it allows you to process the data set in a convenient form and then choose the analysis method that is best suited for the user's data set.

The interface is represented by two pages: for processing the dataset and preparing for further analysis, the results of the analysis and suggestions for choosing an analysis method. Figures 1-3 show the first screen of the application. Figure 1 shows the start screen, 2 and 3 – screens after adding and processing the data set.

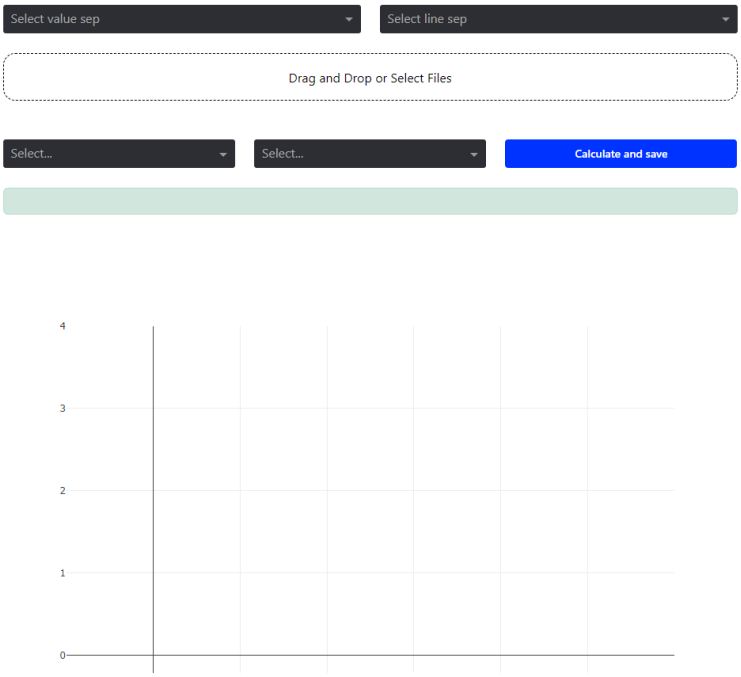

Figure 1 – Staring screen

Figure 2 – Screen with filled information

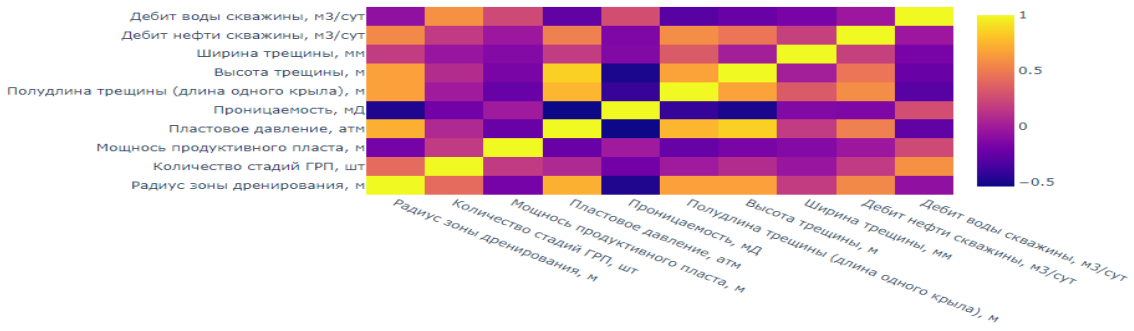

Figure 3 – Heatmap after filtering

On the start screen, in the upper part, selectors are displayed for selecting and configuring the format of filling the dataset, selecting the separator. Then there is an element in which you can select or transfer to it a file that contains a data set. Next, two selectors are displayed in which the user selects the target variable and the multicollinearity filtering force.

By clicking on the “Calculate and Save” button, the user invokes the algorithm for processing, cleaning and preparing the data set for further use in the application. Also, by clicking on the button, two graphs are plotted: before and after multicollinearity is eliminated.

After loading the dataset, the user is shown the first 5 lines so that he can check the correctness and correctness of processing the data set. The successful loading of the data set is displayed with a message in a green frame, after which you can go to the second screen, where various analysis methods will already be displayed. Figures 4 and 5 show the starting and final screens of the application in terms of the predictive power of various variations of logistic regression.

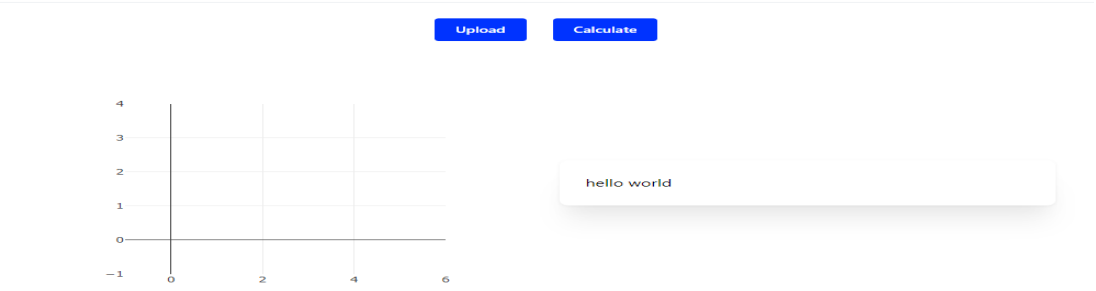

Figure 4 – Startup analysis screen of an application

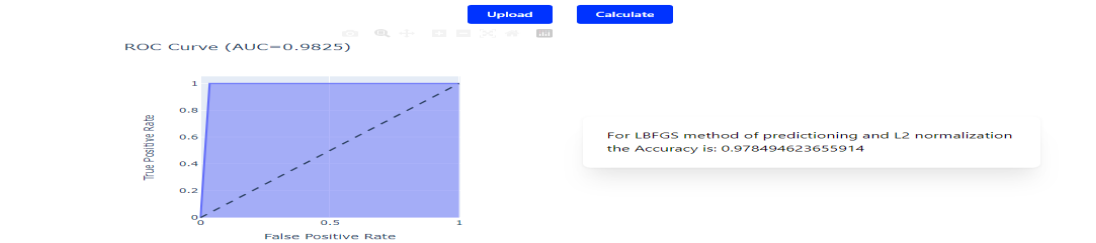

Figure 5 – ROC/AUC graph and results of analysis

The structure of the application is tree-like and is shown in Figure 6.

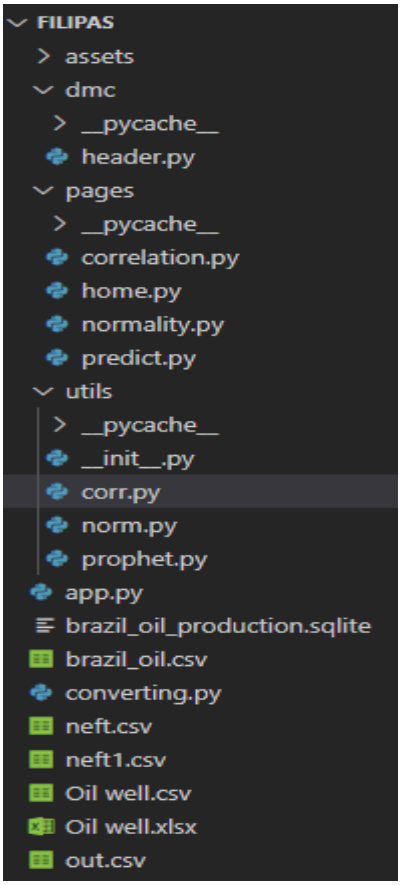

Figure 6 – Application structure

The picture shows the folders: assets, dmc, pages, utils. As well as files app.py and various data sets that participated in the development of models and testing of the application.

The assets folder and dmc are system folders and contain libraries and other files necessary for the correct operation of the application, drawing graphs and running the application on a dedicated server. The pages and utils folders contain code for rendering the user interface and data processing algorithms and forecasting models, respectively.

File app.py It is the starting point of the web application and launches a dedicated local server to work with models and algorithms through the user interface. In the headers file.py contains information about the header of the web application.

The structure of the page consists of a user interface and algorithms that respond to user actions and process the data set provided by the user. Figures 7-8 show an example of the code that is responsible for rendering the user interface.

Figure 7 – Main page code

Figure 8 – Main page code

Figures 7-8 show the code that is converted to the page code using the Dash library, namely JavaScript, html and css. An example of how this code snippet works is shown in Figure 9.

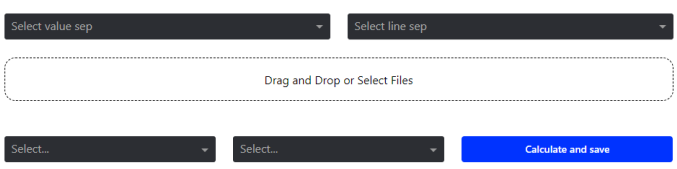

Figure 9 – Example of Dash work

When performing actions on the user interface, one of the functions prescribed in the files from the pages folder is executed. Each function contains a call to a third-party algorithm implemented in one of the files in the utils folders. An example of calling the algorithm from the reaction function on the user interface is shown in Figure 10.

```
@callback(
    Output(component_id='DF', component_property='data'),
    Output(component_id='DF', component_property='columns'),
    Output('target', 'options'),
    Output('corr', 'options'),
    Input('upload', 'contents'),
    State('next_line', 'value'),
    State('next_value', 'value'),
    prevent_initial_call=True
)
def update_map(df, line_sep, value_sep):
    import io
    import base64
    import pandas as pd
    #print(df, line_sep)
    content_type, content_string = df.split(',')
    decoded = base64.b64decode(content_string).decode('utf-8')
    decoded = io.StringIO(decoded)
    df = pd.read_csv(decoded, sep = value_sep)
    df.to_csv('out.csv', index=False, sep = ';')
    df = df.head()
    columns = [{'name': col, 'id': col} for col in df.columns]
    tar = [col for col in df.columns]
    data = df.to_dict(orient='records')
    a = [20.0, 11.0 , 5.0]
    return data, columns, tar, a
```

Figure 10 – Handling user's requests

The main user scenario is:

1) loading the dataset;
2) selecting settings for the data set;
3) target function selection;
4) evaluation of the quality of data set processing;
5) study of the results of the analysis and the choice of one of the proposed methods of analysis.

At this stage, the user scenario ends at the moment of choosing the analysis method, but in the future, it is planned to expand the functionality to the point where the user can enter new data and get a prediction on them on a pre-trained model.

Figure 11 shows the user interface. All elements are numbered in the order in which the user interacts with the application.

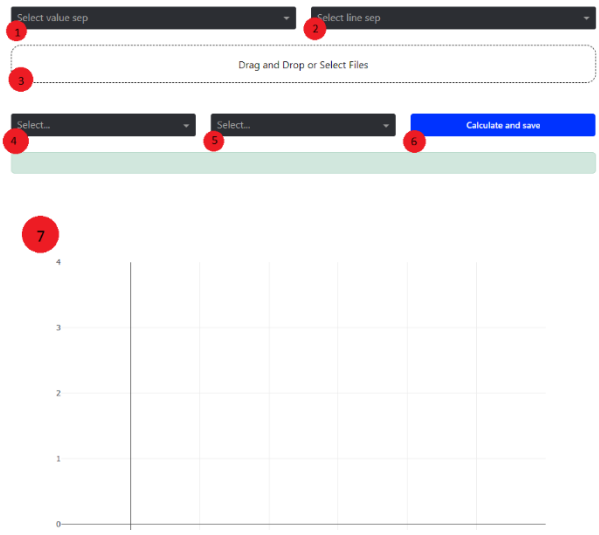

Figure 11 – User Interface

The main user actions are displayed on the start screen:

1) select the symbol separating the values.
2) select the character indicating the end of the line.
3) download the dataset via mouse click or transfer the file to the field.
4) selection of the objective function.
5) selection of the threshold value to eliminate multicollinearity.
6) confirmation of selection and display of collinearity graphs in block No. 7.

Thus, the main user scenario is painted, in which the web application interface participates and all algorithms for data preparation, creation and training of predictive models are used.

As a result of the conducted research, a Web application was created that performs the functions of automatic preparation of a data set for further work of the analyst, which reduces the complexity, allowing a less qualified specialist to perform this work.

The user interface of the application has been developed, which allows using the methods of automatic preparation of the data set and analysis algorithms. The main user scenario in which the interface participates is analyzed, all algorithms of data preparation, creation and training of predictive models are used.

The created web application is a unique analogue of such software products as SAS, however, unlike a competitor, it allows a less trained specialist to use the product, as well as much cheaper to use and master.

In addition to the disassembled data set, which is information about oil wells, this web application can be used with great efficiency by representatives of various business areas, from marketing to advertising.