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
Программное обеспечение и модели анализа предикторов пациентов с заболеваниями, передаваемыми клещами

УДК 004.65:004.451:004.744:616.99

Студент

Группа	ФИО	Подпись	Дата
8ПМ0И1	Сафронов Василий Сергеевич		20.06.2022 г.

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Аксёнов С.В.	к.т.н.		20.06.2022 г.

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОСГН ШБИП	Меньшикова Е. В.	к.ф.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	Савельев А.О.	к.т.н.		

Томск – 2022

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
_____ _____ Савельев А.О.
(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
8ПМ0И1	Сафронову Василию Сергеевичу

Тема работы:

Программное обеспечение и модели анализа предикторов пациентов с заболеваниями, передаваемыми клещами	
Утверждена приказом директора (дата, номер)	№ 145-46/с от 25.05.2022

Срок сдачи студентом выполненной работы:	15.06.2022

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 процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	Объектом исследования настоящей работы являются данные пациентов, страдающих инфекционными заболеваниями, передаваемыми иксодовыми клещами. Предметом исследования является процесс разработки программного обеспечения и моделей анализа предикторов пациентов с заболеваниями, передаваемыми клещами. Для анализа доступна электронная таблица, содержащая сведения о 201 пациенте с клещевыми инфекциями по 168 признакам.

	Деперсонализированные данные предоставлены Сибирским государственным медицинским университетом.
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).*	Аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; анализ данных пациентов с клещевыми инфекциями; построение классификаторов диагнозов пациентов с клещевыми инфекциями; определение наиболее важных для классификации признаков; разработка интерактивного дашборда; обсуждение результатов выполненной работы; заключение работы. Дополнительно должны быть разработаны следующие разделы: финансовый менеджмент, ресурсоэффективность и ресурсосбережение; социальная ответственность; раздел на иностранном языке.
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. Схема структуры исходных данных 2. UML-диаграмма вариантов использования. 3. Скриншоты элементов дашборда.

Консультанты по разделам выпускной квалификационной работы

(с указанием разделов)

Раздел	Консультант
Основная часть	Доцент ОИТ ИШИТР, к.т.н., Аксёнов С. В.
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Доцент ОСГН ШБИП, к.ф.н., доцент Меньшикова Е. В.
Социальная ответственность	Доцент ООД ШБИП, к.б.н., доцент Антоневич О. А.
Английский язык	Доцент ОИЯ, к.ф.н., доцент Степура С. Н.

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

Data Preprocessing Methods

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	1.03.2022

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Аксёнов С. В.	к.т.н.		1.03.2022

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ0И1	Сафронов Василий Сергеевич		1.03.2022

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
Уровень образования магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения весенний семестр 2021 /2022 учебного года

Форма представления работы:

Магистерская диссертация

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	15.06.2022

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
10.06.2022	Основная часть	70
10.06.2022	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
10.06.2022	Социальная ответственность	10
10.06.2022	Английский язык	10

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Аксёнов С. В.	к.т.н.		

СОГЛАСОВАНО:
Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Савельев А. О.	к.т.н.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ПМ0И1	Сафронову Василию Сергеевичу

Школа	**ИШИТР**	**Отделение школы (НОЦ)**	Информационных технологий
Уровень образования	Магистратура	**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*	*Материальные затраты – 41 019 руб.;* *Затраты на электроэнергию – 1 090,60 руб.;* *Оклад руководителя – 37 700,00 руб.;* *Оклад инженера – 19 200,00 руб.*
2. *Нормы и нормативы расходования ресурсов*	*Тариф на электроэнергию 5,8 кВт/ч*
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	*Налог во внебюджетные фонды 27,1%* *Районный коэффициент – 1,3* *Накладные расходы – 70%*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала НТИ*	*Анализ перспективности технических решений посредством QuaD-анализа; Диаграмма Исикавы; SWOT-анализ.*
2. *Разработка устава научно-технического проекта*	*Определение цели научно-исследовательского проекта, требований к проекту, описание заинтересованных сторон проекта, рабочей группы.*
3. *Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок*	*Планирование этапов работы, определение календарного графика проведения исследования.* *Определение рисков научно-исследовательского проекта, оценка вероятности риска и потерь.* *Расчет бюджета затрат на проведение исследования.*
4. *Определение ресурсной, финансовой, экономической эффективности*	*Описание потенциального эффекта научно-исследовательского проекта.*
Перечень графического материала *(с точным указанием обязательных чертежей):*	
1. Оценочная карта технологии «QuaD»; 2. Диаграмма Исикавы; 3. Матрица SWOT; 4. Диаграмма Ганта; 5. Бюджет затрат; 6. Реестр рисков	

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОСГН ШБИП	Меньшикова Екатерина Валентиновна	к.ф.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ0И1	Сафронов Василий Сергеевич		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО		
8ПМ0И1	Сафронов Василий Сергеевич		
Школа	**Инженерная школа информационных технологий и робототехники**	**Отделение (НОЦ)**	**Информационных технологий**
Уровень образования	магистратура	**Направление/ специальность**	09.04.04 Программная инженерия

Тема ВКР:

Программное обеспечение и модели анализа предикторов пациентов с заболеваниями, передаваемыми клещами	
Исходные данные к разделу «Социальная ответственность»:	
Введение – Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика) и области его применения. – Описание рабочей зоны (рабочего места) при разработке проектного решения/при эксплуатации	*Объект исследования:* Алгоритм визуализации и анализа данных пациентов *Область применения:* медицинские и образовательные учреждения *Рабочая зона:* офис *Размеры помещения:* ширина - 7 м, длина – 5 м, высота – 4 м *Количество и наименование оборудования рабочей зоны:* Стол компьютерный, стул компьютерный, персональная электронно-вычислительная машина (ПЭВМ): ноутбук. *Рабочие процессы, связанные с объектом исследования, осуществляющиеся в рабочей зоне:* Разработка и создание программного обеспечения
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности <u>при разработке проектного решения:</u> – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	– Трудовой Кодекс Российской Федерации; – Федеральный закон от 28 декабря 2013 г. N 426-ФЗ «О специальной оценке условий труда»; – СП 2.2.3670-20 «Санитарно-эпидемиологические требования к условиям труда»; – ГОСТ 21889-76 Система «Человек-машина». Кресло человека-оператора; – ГОСТ 22269-76 Рабочее место оператора. Взаимное расположение элементов рабочего места; – ГОСТ 12.2.032-78 «Рабочее место при выполнении работ сидя»;
2. Производственная безопасность <u>при разработке проектного решения</u>: – Анализ выявленных вредных и опасных производственных факторов – Расчет уровня опасного или вредного производственного фактора	**Опасные факторы:** 1. Производственные факторы, связанные с электрическим током, вызываемым разницей электрических потенциалов, под действие которого попадает работающий. **Вредные факторы:** 1. Монотонность труда, вызывающая монотонию; 2. Длительность сосредоточенного наблюдения; 3. Отсутствие или недостаток естественного света; 4. Умственное перенапряжение, в том числе

	вызванное информационной нагрузкой; 5. Перенапряжение анализаторов, в том числе вызванное информационной нагрузкой; 6. Производственные факторы, связанные с электромагнитными полями, неионизирующими ткани тела человека; 7. Производственные факторы, связанные с аномальными микроклиматическими параметрами воздушной среды на местонахождении работающего. **Требуемые средства коллективной и индивидуальной защиты от выявленных факторов:** изолирующие устройства и покрытия, оградительные устройства, экранирующие устройства, заземляющие устройства, основная изоляция, защитные оболочки, безопасное расположение токоведущих частей, размещение их вне зоны досягаемости частями тела, конечностями, ограничение напряжения, защитное отключение. **Расчет:** расчет системы искусственного освещения.
3. Экологическая безопасность <u>при разработке проектного решения</u>	**Воздействие на селитебную зону:** отсутствует. **Воздействие на литосферу:** утилизация люминесцентных ламп, компьютеров и другой оргтехники. **Воздействие на гидросферу:** отсутствует. **Воздействие на атмосферу:** отсутствует.
4. Безопасность в чрезвычайных ситуациях <u>при разработке проектного решения</u>	**Возможные ЧС**: Техногенные аварии (взрыв, пожар) **Наиболее типичная ЧС:** пожар.
Дата выдачи задания для раздела по линейному графику	

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
8ПМ0И1	Сафронов Василий Сергеевич		

РЕФЕРАТ

Выпускная квалификационная работа 132 с., 26 рис., 33 табл., 61 источников, 2 прил.

Ключевые слова: медицинские данные, классификация, машинное обучение, клещевые инфекции, интерактивный дашборд.

Объектом исследования являются данные пациентов, страдающих инфекционными заболеваниями, передаваемыми иксодовыми клещами.

Цель работы – разработка программного обеспечения и моделей анализа предикторов пациентов с заболеваниями, передаваемыми клещами.

В процессе исследования проводились аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; анализ данных пациентов с клещевыми инфекциями; построение классификаторов диагнозов пациентов с клещевыми инфекциями; определение наиболее важных для классификации признаков; разработка интерактивного дашборда.

В результате исследования проведены анализ данных пациентов с клещевыми инфекциями; разработаны модели классификации диагнозов пациентов с клещевыми инфекциями; определены наиболее важные для классификации признаки, разработан интерактивный дашборд.

Область применения: здравоохранение, образование.

Экономическая эффективность/значимость работы – применение программного обеспечения позволит специалистам проводить анализ данных пациентов с клещевыми инфекциями посредством таблиц и графиков, что сократит время работы специалиста, затрачиваемое на данную процедуру.

Оглавление

Введение

В настоящее время в медицинских учреждениях активно внедряются и применяются различные медицинские информационные системы (МИС), данные системы служат в качестве инструментов для сбора и хранения данных. Однако уровень развития информационных технологий позволяет не ограничиваться данным функционалом. Массивы данных могут быть использованы для поиска скрытых закономерностей с помощью алгоритмов машинного и глубокого обучения.

МИС позволяют получать отчеты по введенным данным, однако, помимо важной для анализа информации выгружаются и персональные данные пациентов, такие отчеты не могут быть использованы для проведения анализа. Интеграцию с МИС реализовать достаточно сложно, поэтому целесообразней разработать отдельное программное обеспечение, которое позволит вносить данные, на основе которых будет производиться обучение моделей машинного обучения для дальнейшего использования их в качестве инструмента для оказания помощи специалисту в постановке диагноза, а также для визуального анализа собранных данных, чтобы наглядно оценивать, сравнивать между собой отдельные случаи [1], получать некоторые статистические сведения, чтобы иметь общее представление о наборе данных.

Так как в настоящее время наиболее популярными являются кроссплатформенные программные продукты, не требующие предварительной установки, оптимальным решением будет создать веб-интерфейс [2].

Отделением инфекционных заболеваний Сибирского государственного медицинского университета были предоставлены деперсонализированные данные пациентов с инфекциями, передаваемыми клещами в виде электронной таблицы. Таблица состоит из 9 листов, в которых представлены сведения о пациентах и проводимых исследованиях.

Целью работы является повышение эффективности работы врача инфекциониста при анализе данных пациентов с клещевыми инфекциями путем разработки программного обеспечения и моделей анализа предикторов пациентов с заболеваниями, передаваемыми клещами.

В рамках данной работы проводится подготовка табличных данных пациентов с инфекциями, передаваемыми клещами, к анализу. Также предложен подход классификации диагнозов пациентов с клещевыми инфекциями, разработан интерактивный веб-интерфейс.

Объектом исследования являются данные пациентов, страдающих инфекционными заболеваниями, передаваемыми иксодовыми клещами. Предметом исследования является процесс разработки программного обеспечения и моделей анализа предикторов пациентов с заболеваниями, передаваемыми клещами.

Методы исследования – поиск литературы и источников, анализ информационных материалов, сравнение, консультация со специалистами, методы машинного обучения, методы визуализации.

В работе использованы различные методические материалы и интернет-ресурсы. Работа будет реализована на языке программирования Python, веб-фреймворке Dash.

1 Обзор литературы

Научные исследования на стыке таких направлений, как анализ данных и медицина широко распространены в настоящее время. Здравоохранение генерирует большие массивы разных форматов данных. Аналитика же, в свою очередь, с помощью своих методов занимается описанием, изучением характеристик данных, извлечением информации об особенностях и взаимосвязях составляющих элементов набора данных, выявлением причинно-следственных связей, прогнозированием значений для предстоящих событий [3].

По данным Всемирной организации здравоохранения от трансмиссивных заболеваний ежегодно умирает более 700 тыс. человек. К трансмиссивным относят болезни, возбудителями которых являются паразиты, бактерии и вирусы. Клещи являются переносчиками следующих заболеваний: геморрагическая лихорадка Крым-Конго, болезнь Лайма, возвратный тиф, риккетсиальные заболевания, клещевой энцефалит и туляремия [4]. На территории Российской Федерации наиболее распространены клещевой энцефалит и иксодовый клещевой боррелиоз (болезнь Лайма).

Исследователи из разных регионов России, а также из зарубежных стран занимаются поиском и изучением факторов, оказывающих влияние на заболеваемость, течение и исход болезней, передаваемых клещами.

Учеными из Красноярска был выявлен общий предиктор (интерлейкин-8) для прогноза исхода инфекционного процесса (выздоровление или хронизация) при различных клинических формах клещевой инфекции, что позволяет своевременно оптимизировать этиопатогенетическую терапию заболевания [5].

Представители Архангельской области выявили взаимосвязь между исходом клещевого энцефалита и возрастом, половой принадлежностью пациентов, формой заболевания, наличием таких осложнений, как отёк и

набухание головного мозга, внутрибольничная пневмония и тромбоэмболия легочной артерии [6].

Зарубежными исследователями проводились исследования иксодовых клещей по их внутренним биологическим признакам [7], а также укусов насекомых для выявления укусов клещей [8].

Несколько работ посвящено анализу данных из реестра пациентов MyLymeData, разработанного LymeDisease.org (Калифорния, США). Так, Джонсон Л., Шапиро М. и др., а также Вендроу Д., Хэддок Д. и др. с помощью методов машинного обучения и статистического анализа занимались выявлением зависимости результатов лечения от индивидуальных особенностей пациентов [9, 10].

Ученые из Соединенных Штатов с помощью разработанных классификаторов машинного обучения показали, что больных пациентов можно отличить от здоровых и от пациентов с COVID-19, однако пациентов с ранними случаями ИКБ, у которых могли развиться стойкие симптомы после лечения, выделить не удалось [11].

Как выяснилось, объектом рассмотренных исследований являются данные пациентов, а также сведения о переносчиках заболеваний. В рамках текущего исследования планируется изучение данных пациентов, собранных в электронные таблицы. Как и все форматы данных, табличные данные имеют свои особенности при работе с ними.

Первым этапом анализа при работе с любого рода данными является их предварительная обработка. Существует большое количество методов для обнаружения и устранения аномальных значений в данных. Также принцип обработки данных зависит от их типа. Так, например, категориальные переменные необходимо перевести в воспринимаемый алгоритмами машинного обучения вид, для этого используются методы кодировки данных. Не менее важным является процесс приведения числовых признаков к одной шкале, для этого применяются методы стандартизации и нормализации.

При отборе признаков, которые в дальнейшем будут участвовать в обучении моделей следует учитывать наличие зависимостей между предикторами и целевой переменной, а также предикторами между собой.

Модели машинного обучения с помощью дополнительных инструментов, таких как SHAP, например, помимо предсказаний могут также предоставлять информацию о вкладе каждого отдельного предиктора в результат.

2 Объект и методы исследования

2.1 Объект исследования

Объектом исследования настоящей работы являются данные пациентов, страдающих инфекционными заболеваниями, передаваемыми иксодовыми клещами. Предметом исследования является процесс разработки программного обеспечения и моделей анализа предикторов пациентов с заболеваниями, передаваемыми клещами.

Набор данных сформирован на основе архивных историй болезни пациентов инфекционной клиники Сибирского государственного медицинского университета (СибГМУ) Минздрава России за период с 2009 по 2020 гг. Исходный набор данных представлен в виде электронной таблицы MS Excel, имеющей структуру, представленную на рисунке 1.

Электронная таблица включает в себя 9 листов, в таблице 1 представлена информация о количестве записей и признаков в каждом из листов.

Таблица 1 – Размерность содержимого листов электронной таблицы

Название листа	Количество	
	Записи (строки)	Признаки (столбцы)
Общие сведения	201	82
Температурный лист	101	4
Общий анализ крови	194	21
Биохимический анализ крови	96	24
Иммуноферментный анализ	197	11
Анализ спинномозговой жидкости	112	8
Электрокардиограмма	24	6
Окулист	26	5
Ультразвуковое исследование	6	7

Общее количество признаков – 168, в приложении Б представлена таблица с описанием признаков из исходного набора данных.

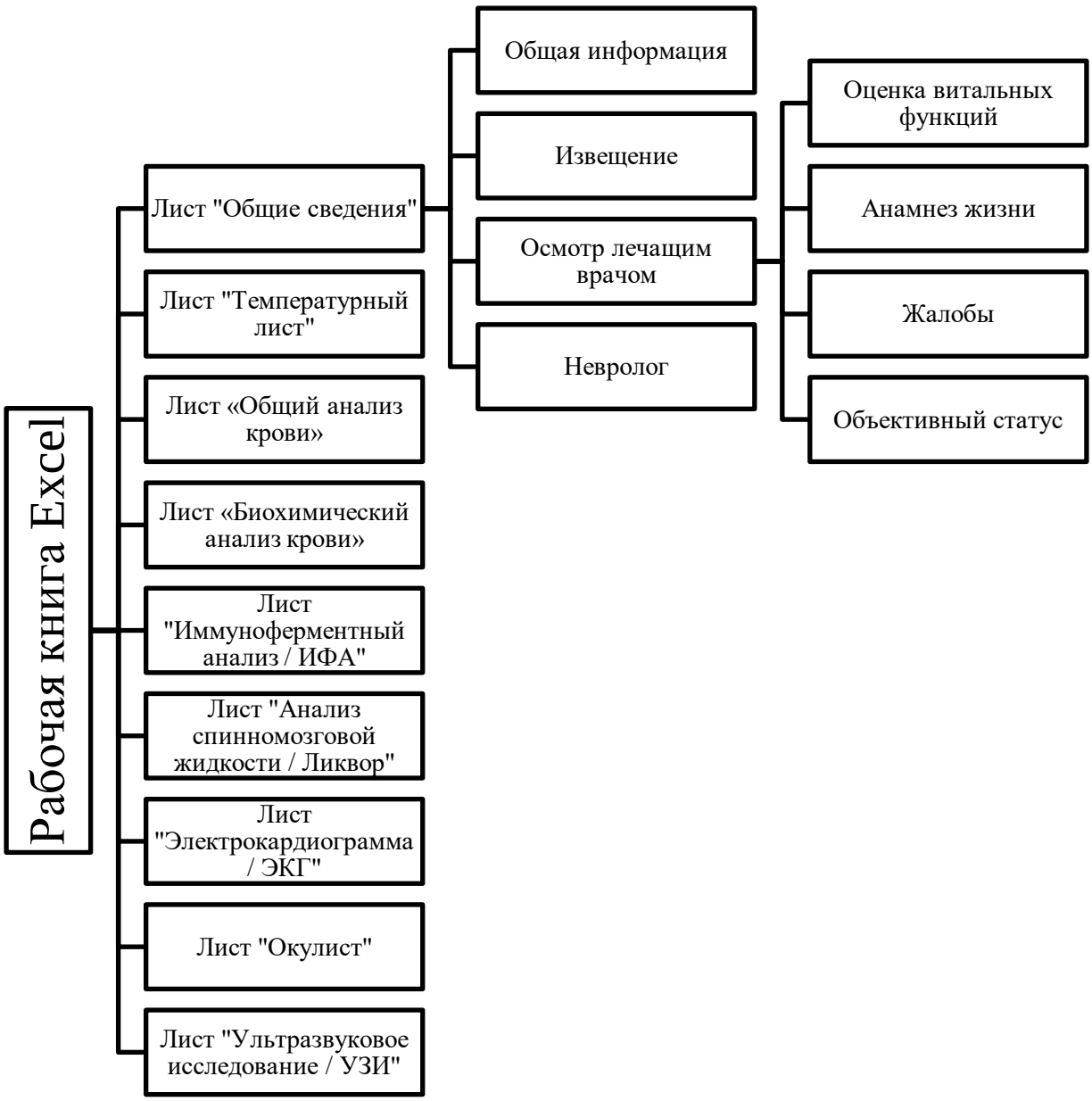


Рисунок 1 – Структура электронной таблицы с данными о пациентах

2.2 Методы исследования

2.2.1 Методы предварительной обработки данных

В процессе предобработки данных производится их подготовка к анализу, в результате которой они приводятся в соответствие с требованиями, определяемыми спецификой решаемой задачи. Предобработка является важнейшим этапом анализа данных, и если она не будет выполнена, то дальнейший анализ в большинстве случаев невозможен из-за того, что

аналитические алгоритмы просто не смогут работать или результаты их работы будут некорректными.

Предобработка данных включает два направления: очистку и оптимизацию.

Очистка производится с целью исключения различного рода факторов, снижающих качество данных и мешающих работе аналитических алгоритмов. Она включает обработку дубликатов, противоречий и фиктивных значений, восстановление и заполнение пропусков, сглаживание, подавление шума и редактирование аномальных значений. Кроме этого, в процессе очистки восстанавливаются нарушения структуры, полноты и целостности данных, преобразуются некорректные форматы.

Оптимизация данных как элемент предобработки включает снижение размерности, выявление и исключение незначащих признаков. Основное отличие оптимизации от очистки в том, что факторы, устраняемые в процессе очистки, существенно снижают точность решения задачи или делают работу аналитических алгоритмов невозможной. Проблемы, решаемые при оптимизации, адаптируют данные к конкретной задаче и повышают эффективность их анализа.

2.2.1.1 Методы обнаружения и устранения выбросов

В статистике и аналитике данных существует такое понятие, как выбросы. Под выбросами понимают значения, значительно отличающиеся от основной массы значений набора данных. Причинами возникновения выбросов могут быть сбои работы оборудования, человеческий фактор, случайность, уникальные явления и т.д. Очень важно обнаружить и оценить выбросы, чтобы повысить качество анализа. Существует несколько методов обнаружения выбросов [12]:

1. Экстремальный анализ данных. При таком анализе не применяются какие-либо специальные статистические методы. Алгоритм данного метода следующий:

- Визуализировать данные, используя диаграммы (рассеяния или размаха) и гистограммы для нахождения экстремальных значений;
- Задействовать, например Гауссовское распределение, и найти значения, чье стандартное отклонение отличается в 2-3 раза от математического ожидания или в полтора раза от первого либо третьего квартилей.

2. Аппроксимирующий метод, заключающийся в применении кластеризующих методов:

- Использовать метод кластеризации для определения кластеров в данных;
- Идентифицировать и отметить центроиды каждого кластера;
- Соотнести кластеры с экземплярами данных, находящимися на фиксированном расстоянии или на процентном удалении от центроида соответствующего кластера.

3. Проецирующие методы:

- Использовать один из проецирующих методов, например, метод главных компонент или самоорганизующиеся карты Кохонена или проекцию Саммона, для суммирования обучающих данных в двух измерениях;
- Визуализировать отображение;
- Использовать критерий близости от проецируемых значений или от вектора таблицы кодирования для идентифицирования выбросов.

Рассмотрим более подробно этапы обнаружения выбросов при экстремальном анализе данных:

1. Вычисление медианы Q_2 (величины, находящейся в середине набора данных). В случае, если число значений в наборе данных нечетное, то медианой является значение, до и после которого расположено одинаковое

количество значений. Если же количество значений кратно двум, то медиана рассчитывается, как среднее арифметическое двух средних значений.

2. Вычисление нижнего квартиля Q_1 (величины, ниже которой располагается $\frac{1}{4}$ значений из набора данных, т.е. половина значений, лежащих ниже медианы. Если до медианы находится четное количество значений, то, как и на предыдущем шаге, рассчитывается среднее арифметическое двух средних значений.

3. Вычисление верхнего квартиля Q_3 (величины, выше которой располагается $\frac{1}{4}$ значений из набора данных, т.е. половина значений, лежащих выше медианы. Процесс расчета верхнего квартиля аналогичен расчету нижнего квартиля.

4. Вычисление межквартильного размаха (диапазона). Данный показатель представляет собой расстояние между верхним и нижним квартилями и рассчитывается следующим образом: $Q_3 - Q_1$

5. Нахождение внутренних границ значений в наборе данных. Значения, находящиеся за пределами внутренних границ считаются незначительными выбросами, а лежащие за пределами внешних границ считаются значительными выбросами. Внутренние границы определяются следующим образом:

5.1. $(Q_3 - Q_1) * 1{,}5$

5.2. Нижняя внутренняя граница = $Q_1 - (Q_3 - Q_1) * 1{,}5$

5.3. Верхняя внутренняя граница = $Q_3 + (Q_3 - Q_1) * 1{,}5$

6. Нахождение внешних границ значений в наборе данных. Внешние границы определяются следующим образом:

6.1. $(Q_3 - Q_1) * 3$

6.2. Нижняя внешняя граница = $Q_1 - (Q_3 - Q_1) * 3$

6.3. Верхняя внешняя граница = $Q_3 + (Q_3 - Q_1) * 3$

После нахождения выбросов необходимо решить, как поступить с такими значениями в наборе данных: исключить или же оставить. Основным

фактором, влияющим на данное решение, является причина возникновения данной аномалии. Выбросы, появившиеся в результате допущения ошибки, исключаются. Выбросы, связанные с новой информацией или тенденцией, в наборе данных оставляют, из-за ошибочного исключения выбросов могут быть упущены какая-либо неизвестная ранее тенденция или открытие [13].

Помимо рассмотренного выше, выбросами также считают значения, отличающиеся от основной массы не только величиной, но и типом данных. Такие ситуации могут возникать, если при заполнении большого количества признаков специалист допустил опечатку, вместо числа ввел буквенные значения, по ошибке заполнил соседнее поле и т.д. В таком случае, часто выбросы подвергаются исключению, если восстановить значения невозможно.

2.2.1.2 Методы работы с пропущенными значениями

Нередко набор данных может содержать пропущенные значения. Причины наличия пропусков в данных могут быть различными: поле для заполнения из-за невнимательности заполняющего было пропущено, либо поле осталось пустым из-за отсутствия данных для заполнения (к примеру, пациенту не назначалось определенное исследование, поэтому записи о его результатах отсутствуют). Также, если обнаруженные выбросы было решено исключить, то вместо них появились пропущенные значения.

Существует несколько способов по устранению пропущенных значений [14]:

Заполнение нулём (0). Если числовой признак принимает, как отрицательные, так и положительные значения, заполнение 0 будет свидетельствовать о нейтральности заполненного случая. Данный способ применим и к категориальным признакам, когда значение 0 характеризует отсутствие данного признака у отдельного наблюдения.

Заполнение статистическими показателями. Применяются среднее арифметическое значение и медиана. Если признак не имеет выбросы, то используется среднее значение, иначе используется медиана, т.к. этот показатель устойчив к выбросам.

Заполнение значениями, найденными вспомогательными моделями. На основе значений других признаков и известных значений признака, имеющего пропущенные значения, обучается модель машинного обучения и пропущенные значения предсказываются.

Введение индикаторных переменных. Пропущенные значения заменяются нулями и добавляется новый признак, принимающий значение 1 для наблюдений с пропусками и 0 для наблюдений без пропусков, или наоборот.

Заполнение значением соседнего наблюдения. Данный способ применяется при заполнении временных рядов, когда последующие значения сильно связаны с предыдущими.

Исключение из набора данных наблюдения. Если набор данных достаточно большой и пропусков немного, то удаление наблюдения не окажет значительное влияние на дальнейший анализ. Также наблюдение удаляется, если пропуски присутствуют в большом количестве признаков.

Исключение из набора данных признака. Если признак содержит большое количество пропусков, восстановить которые невозможно, такой признак удаляется из набора данных.

2.2.1.3 Методы преобразования данных

Существует два основных типа данных: категориальные и числовые (Рисунок 2) [15].

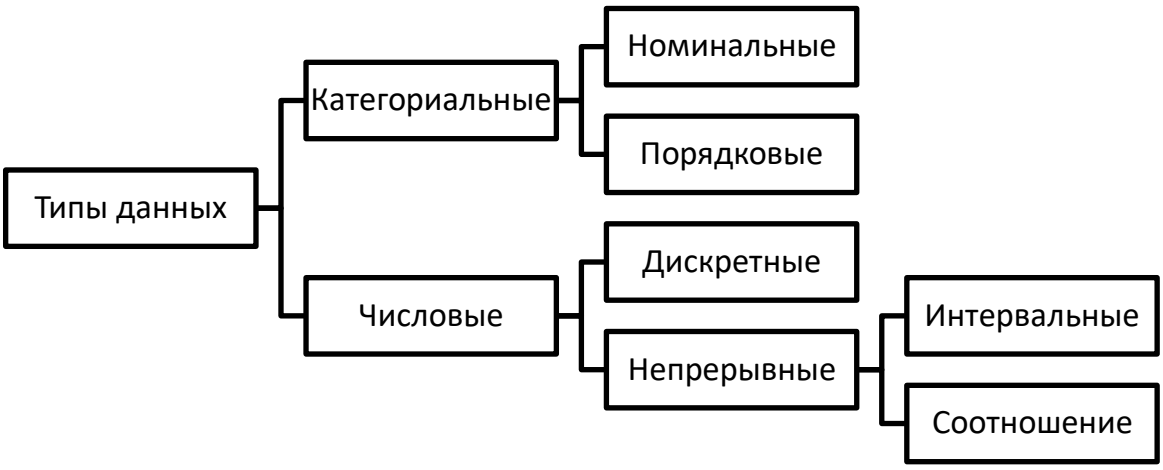


Рисунок 2 – Классификация данных по типам

Номинальные значения представляют собой дискретные единицы и используются для обозначения переменных, не имеющих количественного значения, номинальные данные не имеют порядка. Примером номинальных данных может быть пол (женский, мужской).

Порядковые значения представляют собой дискретные и упорядоченные единицы. Пример порядковых данных – ступени образования (дошкольное, начальное общее, основное общее, среднее общее или профессиональное, высшее).

Дискретные данные – тип данных, которые не могут быть измерены, но могут быть подсчитаны. Примером может быть количество голов в 100 монетах. Проверить, являются ли данные дискретными, можно ответив на следующие два вопроса: можно ли их посчитать и можно ли их разделить на меньшие и меньшие части?

Непрерывные данные представляют измерения, поэтому их значения не могут быть подсчитаны, но они могут быть измерены. Примером может быть рост человека.

Интервальные значения представляют упорядоченные единицы, имеющие одинаковую разницу. Пример – температура воздуха (-10°C, -5°C, 0°C, +5°C, +10°C).

Значения отношения также являются упорядоченными единицами, которые имеют одинаковую разницу, однако отношения имеют абсолютный ноль, т.е. возможно отсутствие того или иного признака. Примером могут быть рост, вес, длина и т.д.

2.2.1.3.1 Методы работы с категориальными признаками

Что касается категориальных признаков, то чаще всего они представляются в виде текстовых значений, которые требуют перекодировки в числа.

Учитывая тот факт, что значения порядковых признаков являются неравноценными между собой, т.е. порядок имеет значение, оптимальным является поставить в соответствие каждому значению его порядковый номер в отсортированном списке всех возможных значений. В качестве примера можно представить преобразование европейских размеров следующим образом: M – 1, L – 2, XL – 3.

При работе с номинальными признаками данный подход не применим, каждое отдельное значение является равноценным по отношению к другим и ставить в соответствие им порядковые номера некорректно. Так, например, в признаке «цвет» нельзя поставить следующие соответствия: синий – 0, зеленый – 1, красный – 2, т.к. синий не меньше зеленого, а красный не больше синего. Поэтому для номинальных признаков используется метод под названием прямое кодирование. Суть данного метода заключается в том, что для каждого уникального значения в столбце номинального признака создается фиктивный признак, принимающий значение 1, если конкретному наблюдению свойственно наличие данного признака, а иначе – 0. В таблице 2 представлены исходные значения и результат прямого кодирования номинального признака «Цвет» [16].

Таблица 2 – Пример прямого кодирования номинального признака

Цвет	Цвет_зеленый	Цвет_красный	Цвет_синий
зеленый	1	0	0
красный	0	1	0
синий	0	0	1

2.2.1.3.2 Методы масштабирования данных

Большинство алгоритмов машинного обучения и оптимизации показывают лучшие результаты, если признаки находятся в одинаковых шкалах. Процесс приведения признаков к одинаковой шкале называется масштабированием. Масштабирование применяется к данным числового типа. Существует два общих подхода к приведению разных признаков к одинаковой шкале: нормализация и стандартизация [16].

Нормализация означает приведение признаков к диапазону [0, 1], к каждому признаковому столбцу применяется минимаксное масштабирование, где нормализованное значение $x_{norm}^{(i)}$ из образца $x^{(i)}$ можно вычислить следующим образом:

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}},$$

где $x^{(i)}$ – отдельно взятый образец, x_{min} – наименьшее значение в признаковом столбце, x_{max} – максимальное значение в признаковом столбце.

При помощи стандартизации признаковые столбцы центрируются в нулевом среднем значении, т.е. равном 0, с единичным стандартным отклонением, т.е. равным 1, в результате чего признаковые столбцы принимают вид нормального распределения. Процедура стандартизации может быть выражена следующей формулой:

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x},$$

где μ_x – эмпирическое среднее отдельно взятого признакового столбца, σ_x – стандартное отклонение.

2.2.2 Методы машинного обучения

Ввиду того, что существует большое разнообразие методов и способов обучения искусственного интеллекта, далее будут рассмотрены основные из них, а также их возможности, ограничения и сферы применения.

Обучение с учителем. Данный способ является оптимальным, если известно, чему будет обучаться машина. Компьютер знакомится с обучающей выборкой данных, параметры варьируются до тех пор, пока на выходе не будут получены ожидаемые результаты. Затем выполняется проверка обучения машины с помощью прогнозирования результатов для контрольных данных, с которыми компьютер сталкивается впервые. Обучение с учителем применяется для задач классификации и прогнозирования. Данный способ может быть полезен в сфере медицины при решении следующих задач: классификация вида заболевания; определение наиболее целесообразного способа лечения; предсказание длительности и исхода заболевания; оценка риска осложнений; поиск синдромов (наиболее характерных для данного заболевания совокупностей симптомов) [17].

Обучение без учителя. В данном случае компьютер занимается изучением набора данных и поиском скрытых связей между различными переменными. Данный способ применяется для объединения данных в группы (кластеры) на основании их статистических свойств. В медицине используется кластеризация симптомов, заболеваний, препаратов [18].

Обучение с частичным привлечением учителя. Данный способ объединяет обучение с учителем и без него. Разметив часть набора данных, учитель дает машине понять, каким образом кластеризовать оставшийся массив данных. Этот способ машинного обучения распространен для анализа медицинских изображений, таких как сканы компьютерной томографии или

МРТ. Опытный рентгенолог может разметить небольшое подмножество сканов, на которых выявлены опухоли и заболевания. А машина, в свою очередь может извлечь информацию из небольшой доли размеченных данных и улучшить точность предсказаний по сравнению с моделью, обучающейся исключительно на неразмеченных данных. [19].

Обучение с подкреплением. При данном способе машина взаимодействует с окружением и получает “вознаграждение”, в случае правильного выполнения задания. При автоматизированном подсчёте вознаграждений машина может обучаться самостоятельно. Данный способ может применяться для улучшения процесса отслеживания состояния пациентов во время операции [20], а также рекомендаций по более эффективной тактике лечения для текущего состояния пациента [21].

Глубинное обучение. Данный способ машинного обучения может проходить как без учителя, так и с подкреплением. При глубинном обучении частично имитируются принципы обучения людей — используются нейронные сети для все более подробного уточнения характеристик набора данных. Используются для медицинской диагностики, анализа медицинских изображений. Глубинные нейронные сети применяются, в частности, для ускорения скрининга больших объемов данных при поиске лекарственных средств. Такие нейросети способны обрабатывать множество изображений за короткое время и извлечь больше признаков, которые модель в конечном счете запоминает [22].

2.2.2.1 Метод дерева решений

Метод дерева решений строится на основе решающих правил, упорядоченных в древовидную иерархическую структуру, вида «если, то».

Данный метод основан на процессе рекурсивного разбиения исходного множества объектов на подмножества, предварительно отнесенные к заданным классам. С помощью решающих правил производится разбиение,

значения атрибутов проверяются по заданному условию. Структура дерева решений представлена на рисунке 3 [23].

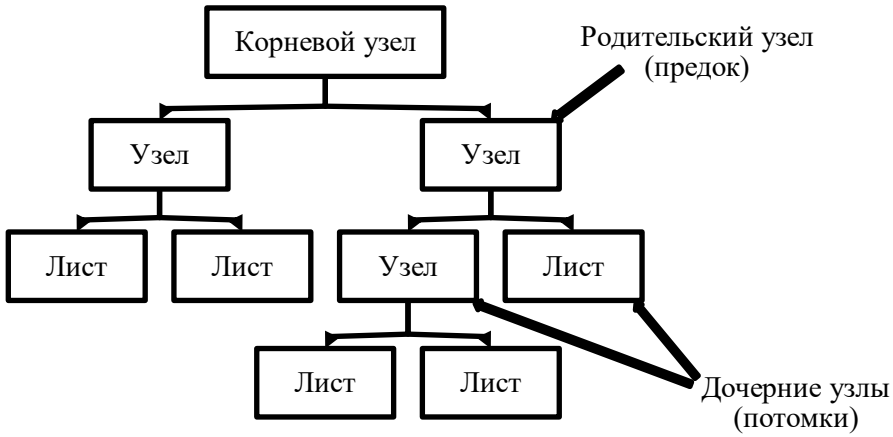


Рисунок 3 – Структура дерева решений

Выделяют два основных элемента структуры – узлы и листья. Узлы содержат решающие правила и удовлетворяющие им подмножества наблюдений. В листьях располагаются классифицированные деревом наблюдения. Каждый лист относится к одному из классов и объекту присваивается соответствующая метка класса.

В узлах указываются разбивающие содержащиеся в нем наблюдения правила, а листья, в свою очередь, помечаются меткой класса, объекты которого попали в данный лист. Если класс, определенный деревом, совпадает с целевым классом, то объект является распознанным, иначе - нераспознанным. Самый верхний узел имеет название корневой узел, он содержит весь обучающий и рабочий наборы данных.

Дерево решений относится к линейным классификаторам, объекты разбиваются в двумерном пространстве линиями (в многомерном пространстве - плоскостями).

На рисунке 4 представлено дерево, решающее задачу классификации объектов на три класса по двум атрибутам [23].

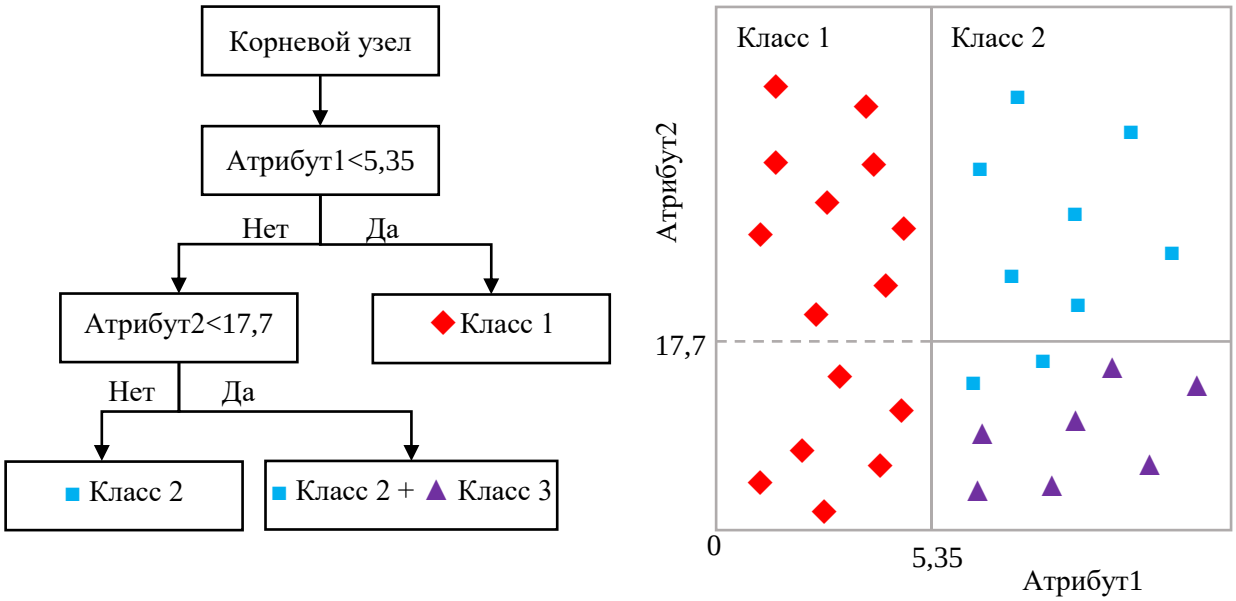


Рисунок 4 – Пример классификации методом дерева решений

Красные ромбы представляют объекты, относящиеся к 1 классу, синие квадраты – 2, а фиолетовые треугольники – 3. Признаковое пространство разделено на 3 подмножества с помощью линий. Данные подмножества соответствуют трем возможным исходам классификации. В классе треугольников находятся нераспознанные примеры – квадраты, т.е. попавшие в подмножество примеры, относящиеся к другому классу.

Алгоритм может разбивать данные до тех пор, пока все объекты не будут отнесены к верному классу, однако это может привести к усложнению дерева. Увеличение числа ветвлений, узлов и листьев снижает интерпретируемость модели. Поэтому деревья ограничивают даже за счет потери точности, процесс упрощения деревьев реализуется с помощью методов ранней остановки и отсечения ветвей.

Деревья решений могут быть дихотомичными (бинарными), т.е. иметь два потомка в узле, и полихотомичными - иметь более двух потомков в узле. Более простыми в построении и интерпретации являются дихотомичные деревья.

В качестве преимуществ данного метода следует выделить следующие:

– Высокая объясняющая способность за счет формирования правил практически на естественном языке;

– Работает с числовыми и категориальными признаками;

– Не требуют большой предобработки, в т.ч. не требуют нормализации, создания фиктивных переменных, могут работать с пропущенными значениями;

– Возможна работа с большими объемами данных.

К недостаткам метода дерева решений можно отнести:

– Неустойчивость, т.е. при небольших изменениях в данных, результаты классификации могут значительно измениться;

– Построение оптимального дерева не гарантировано, т.к. на каждом этапе находится локально оптимальное решение;

– Склонность к переобучению.

2.2.2.2 Метод логистической регрессии

Метод логистической регрессии, также известный как метод логит-регрессии, в математической статистике представляет собой применение логистической функции для моделирования зависимости выходной бинарной переменной от набора входных переменных.

Логистическая регрессия является разновидностью множественной регрессии, целью которой является изучение связи между независимыми переменными (регрессорами, предикторами) и зависимой переменной. В общем виде регрессия используется для работы с непрерывными данными, а логистическая регрессия, в свою очередь, применима для выходных переменных, принимающих исключительно два значения.

Данный метод широко применим т.к. многие задачи анализа данных решаются с помощью бинарной классификации или могут быть сведены к ней.

Так, например, логистическая регрессия позволяет оценить вероятность наступления или ненаступления того или иного события: пациент болен или здоров, заемщик оплатил кредит или срок оплаты был пропущен и т.д.

В общем виде регрессионные модели можно представить следующей формулой:

$$y = F(x_1, x_2, \dots x_n)$$

Если рассматривается задача с пациентом, переменная y задается со значениями 1 и 0, где 1 означает, что пациент болен, а 0 – здоров.

Однако множественная регрессия не учитывает, что переменная отклика бинарная, и алгоритмом будут предсказываться значения отличные от 1 и 0. Таким образом, множественной регрессией будут проигнорированы ограничения на диапазон значений для y.

Задача регрессии может быть сформулирована также следующим образом: вместо поиска бинарных значений переменных предсказывается непрерывная переменная со значениями, входящими в диапазон [0, 1]. Это можно реализовать, применив регрессионное уравнение следующего вида:

$$p = \frac{1}{1+e^{-y}},$$

где p – вероятность наступления интересующего события; e – основание натурального логарифма 2,7; y – стандартное уравнение регрессии.

Зависимость между вероятностью наступления события и величиной y представлена на рисунке 5.

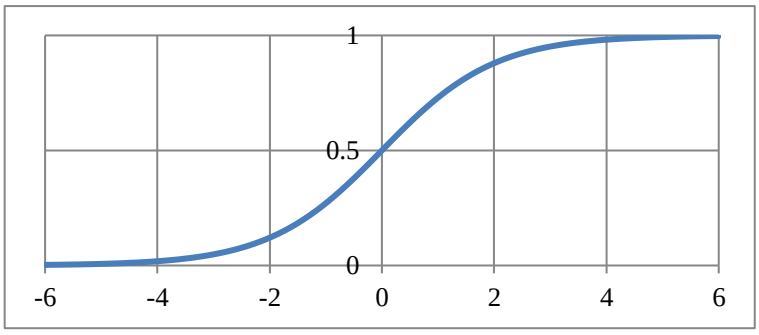


Рисунок 5 – График зависимости вероятности события и величины $\boldsymbol{y}$

Преобразование следующего вида называется логистическим (логит-) преобразованием:

$$P' = \log_e(\frac{P}{1-P})$$

Для нахождения коэффициентов логистической регрессии существует несколько способов, одним из которых является метод максимального правдоподобия. Данный способ используется в статистике для получения оценок параметров генеральной совокупности по выборочным данным [24].

2.2.2.3 Метод случайного леса

Суть метода случайного леса заключается в применении совокупности (ансамбля) деревьев решений, каждое из которых по отдельности дает достаточно низкое качество классификации, а в совокупности за счет их большого числа получается более высокий результат. Данный метод применяется для задач классификации, в таком случае принимается решение голосованием по большинству, и регрессии, ответы деревьев усредняются.

Метод случайного леса основывается на так называемой мудрости толпы. Эффективность работы случайного леса определяется следующим правилом: «Большое число относительно некоррелированных деревьев, работающих совместно, будет превосходить любую из их отдельных составляющих» [25].

Часть деревьев может быть неправильной, но большинство будет правильным и в результате чего совокупность деревьев может следовать в правильном направлении. Предпосылками успешного прогнозирования можно считать некоторый осмысленный сигнал в признаках (чтобы модели были точнее случайного угадывания), а также слабую корреляцию между прогнозами (и ошибками) отдельных деревьев.

В качестве преимуществ данного метода можно выделить следующие:

– Эффективная обработка данных с большим числом признаков и классов;

– Нечувствительность к монотонным преобразованиям значений признаков;

– Применимость как к дискретным, так и непрерывным признакам;

– Поддерживает методы оценки значимости отдельных признаков;

– Параллелизуемость и масштабируемость;

– Гибкость и высокая точность.

В качестве недостатков можно выделить следующие:

– Большой размер обученных моделей;

– Выполнение алгоритма занимает больше времени, чем выполнение неансамблевых методов;

– Увеличение объема оказывает отрицательное влияние на интуитивное понимание.

2.2.2.4 Метод градиентного бустинга

Градиентный бустинг – это техника для выполнения задач машинного обучения с учителем, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений. Принцип действия данного метода представлен на рисунке 6.

Бустинг строит модели из отдельных «слабых учеников» итеративным способом. Отдельные модели строятся не на случайных подмножествах данных и функций, а на согласованной основе, придавая больший вес экземплярам с неправильными предсказаниями и высокими ошибками. Общая идея заключается в том, что случаи, которые трудно предсказать правильно («сложные» случаи), будут сконцентрированы в процессе обучения, так что модель учится на ошибках прошлого. Тренировку каждого ансамбля на подмножестве обучающего набора также называют

стохастическим градиентным бустингом, который может помочь улучшить обобщаемость модели.

Градиент используется для минимизации функции потерь. На каждом этапе обучения создается слабый ученик, и его прогнозы сравниваются с правильным результатом, который является ожидаемым.

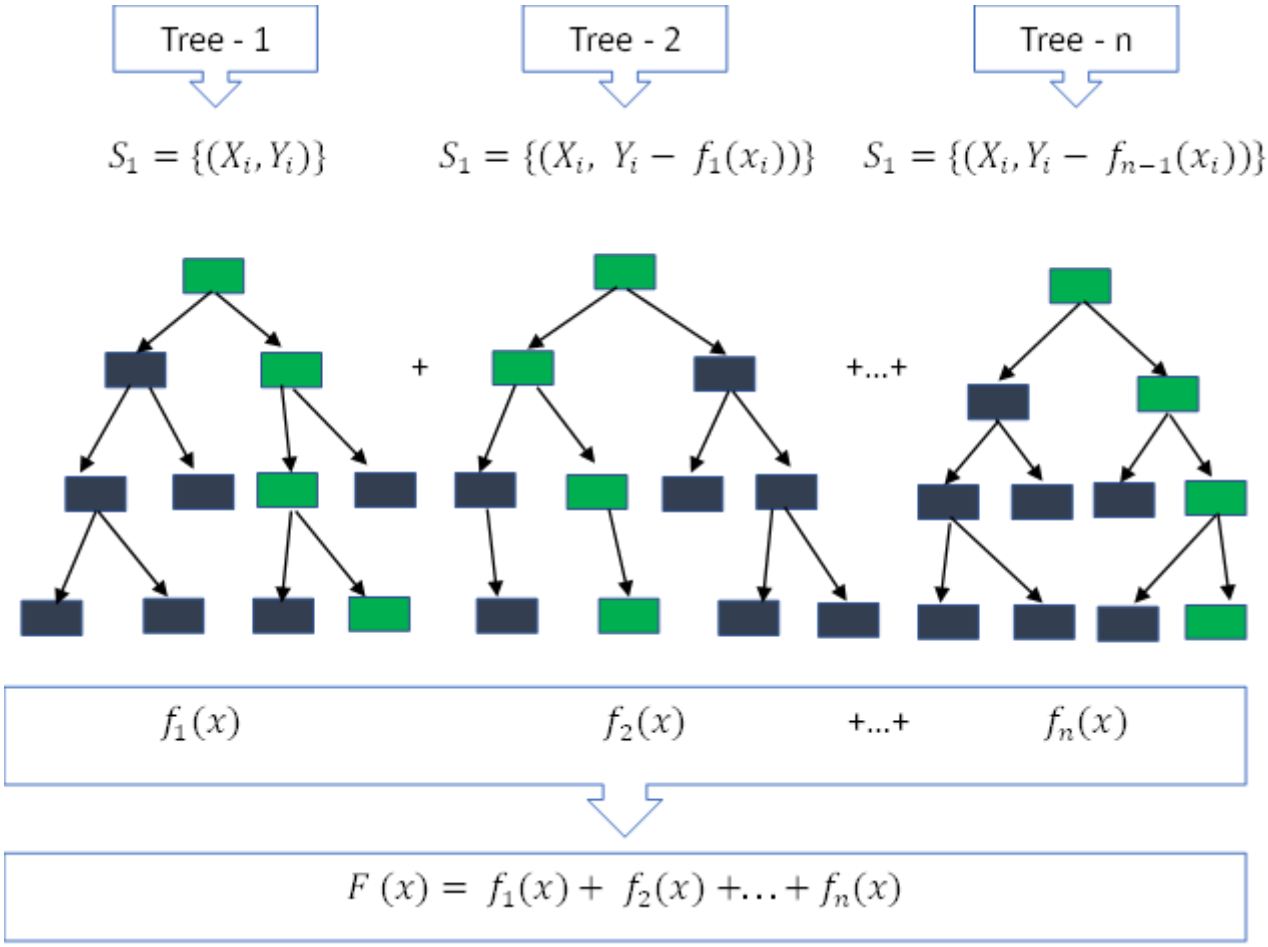


Рисунок 6 – Принцип работы метода градиентного бустинга

Расстояние между предсказанием и истиной представляет частоту ошибок модели. Эти ошибки используются для расчета градиента. Градиент представляет собой в основном частную производную функции потерь - поэтому она описывает крутизну функции ошибок. Градиент можно использовать для определения направления, в котором нужно изменить параметры модели, чтобы (максимизировать) уменьшить ошибку в следующем раунде обучения, «снижая градиент».

2.2.3 Методы оценки качества и интерпретации работы моделей

2.2.3.1 Матрица ошибок и простые оценки

Матрица несоответствий, также известная как матрица ошибок (таблица 3), представляет собой квадратную таблицу, содержащую значения количества истинно положительных, истинно отрицательных, ложноположительных и ложноотрицательных исходов классификации [16].

Таблица 3 – Матрица ошибок бинарной классификации

		Спрогнозированный класс	
		1 (P)	0 (N)
Фактический класс	1 (P)	Истинно положительные (TP)	Ложно-отрицательные (FN)
	0 (N)	Ложно-положительные (FP)	Истинно отрицательные (TN)

К истинно положительным исходам (TP) относят объекты, фактически относящиеся к положительному классу (1) и классифицированные как положительный класс (0).

Ложноположительные исходы (FP) – это объекты, фактически принадлежащие к отрицательному классу (0) и классифицированные как положительный класс (1).

К истинно отрицательным исходам (TN) относят объекты, фактически относящиеся к отрицательному классу (0) и классифицированные как отрицательный класс (0).

Ложноотрицательные исходы (FN) – это объекты, фактически принадлежащие к положительному классу (1) и классифицированные как отрицательный класс (0).

Для многоклассовой классификации матрица ошибок строится по такому же принципу, что и для бинарной классификации (таблица 4) [26].

Таблица 4 – Матрица ошибок многоклассовой классификации

		Спрогнозированный класс			
		1	2	…	n
Фактический класс	1	T_1	F_{12}	…	F_{1n}
	2	F_{21}	T_2	…	F_{2n}
	…	…	…	…	…
	n	F_{n1}	F_{n2}	…	T_n

В таком случае TP, TN, FP и FN рассчитываются относительно каждого отдельного класса (i) следующим образом:

$$TP_i = T_i,$$

где TP_i – доля истинно положительных исходов для класса i.

$$FP_i = \sum_{j=1}^{n} F_{j,i},$$

где FP_i – доля ложноположительных исходов для класса i, n – общее количество классов.

$$FN_i = \sum_{j=1}^{n} F_{i,j},$$

где FN_i – доля ложноотрицательных исходов для класса i.

$$TN_i = All - TP_i - FP_i - FN_i,$$

где TN_i – доля истинно отрицательных исходов для класса i, All – общее количество всех исходов.

На основе данных четырех показателей рассчитываются метрики качества работы моделей классификации.

Для задач медицины используются такие метрики, как специфичность (specificity) и чувствительность (sensitivity), рассчитываются они следующим образом:

$$Sensitivity = TPR = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{FP + TN}$$

Чувствительность представляет собой отношение верно классифицированных объектов к общему числу элементов, принадлежащих к данному классу. Специфичность показывает отношение верного отнесения классификатором объектов, не относящихся к данному классу. Так, например, при классификации пациентов на больных и здоровых, с помощью чувствительности можно оценить долю обнаруженных больных пациентов, а с помощью специфичности – долю здоровых пациентов, отнесенных классификатором к таковым.

Помимо простых числовых метрик также существуют графические методы представления результатов классификации. Один из таких методов – ROC-кривые, он заключается в визуализации зависимости таких метрик, как истинно положительный (TPR, true positive rate) и ложноположительный (FPR, false positive rate) показатели. TPR равен чувствительности, а FPR рассчитывается по следующей формуле:

$$FPR = \frac{FP}{FP + TN}$$

На рисунке 7 представлены ROC-кривые.

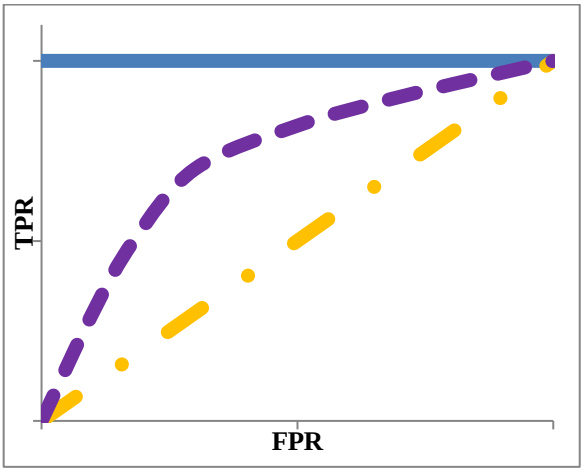


Рисунок 7 – График ROC-кривых

По горизонтальной оси откладываются значения ложноположительного показателя, а по вертикальной оси – истинно

положительного. Синей сплошной линией показана ROC-кривая для идеальной работы алгоритма классификации, оранжевой штрих-пунктирной линией – худшая работа алгоритма, а фиолетовой штриховой линией - типичная.

Для численной оценки алгоритма по ROC-кривой применяется величина, равная площади под кривой (AUC, area under curve). Идеальной работе алгоритма соответствует значение AUC 1, а худшей – 0,5.

2.2.3.2 Важность и степень влияния признаков

Одним из подходов отбора релевантных признаков из набора данных является определение важности признаков. Используя такой ансамблевый метод классификации, как случайный лес, можно измерить важность каждого отдельного признака как усредненное уменьшение неоднородности, вычисленное на основе всех деревьев решений леса.

Существуют разные способы расчета важности признаков, наибольшее распространение получили такие методы, как важность Джини и важность признаков на основе перестановки.

Важность Джини используется для расчета неоднородности узла дерева, а важность признака – это уменьшение неоднородности узла, взвешенное по количеству объектов, достигающих этого узла из общего числа объектов.

После расчета важностей признаков может быть построен график (рисунок 8), в котором в порядке убывания степени важности располагаются признаки, оказывающие наибольшее влияние на результат классификации. Значение важностей признаков нормализованы, поэтому сумма значений каждого отдельного признака равна 1 [16].

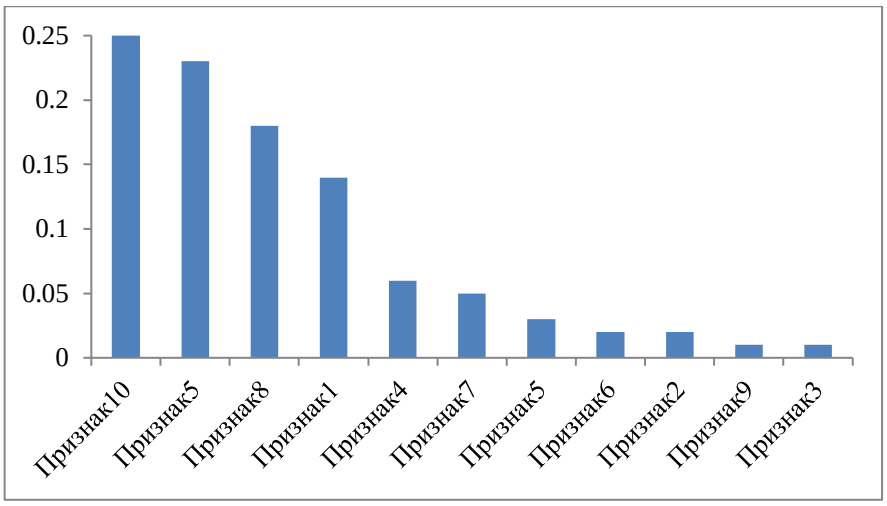


Рисунок 8 – График важностей признаков

Помимо стандартной важности признаков, рассчитываемой исключительно на основе случайного леса, существует более универсальный инструмент для оценки важности признаков в различных моделях машинного обучения. Данный метод получил название SHAP (Shapley Additive Explanation Values), что в переводе с английского означает аддитивные объяснения Шепли и предназначен для объяснения индивидуальных прогнозов на основе модели машинного обучения.

Данный метод в анализ данных пришел из теории игр, суть которой заключается в выборе оптимальной стратегии с учётом представлений о других участниках, их возможных поступках и ресурсах. Определить наилучшее распределение выигрыша между игроками можно с помощью вектора Шепли. Данный вектор – это распределение, выигрыш каждого игрока в котором равен его среднему вкладу в общее благосостояние при определенном механизме формирования коалиции [27]:

$$\Phi(v)_i = \sum_{K \ni i} \frac{(k-1)!(n-k)!}{n!} (v(K) - v(K \backslash i)),$$

где n – количество игроков, k – количество участников коалиции K.

Таким образом, рассмотренные положения теории игр можно применить к интерпретации моделей машинного обучения следующим образом:

– В качестве игры выступает результат обучения модели классификации;

– Выигрыш – это разница между математическим ожиданием результата на всех существующих примерах и полученным результатом на данном примере;

– В качестве вкладов игроков в игру выступает влияние каждого значения признака на результат.

При расчёте вектора Шепли коалиции формируются из ограниченного набора признаков. Для формирования коалиций не убирают «лишние» признаки, а заменяют их на случайные значения из «фонового» набора данных. Усреднённый результат модели со случайными значениями признака эквивалентен результату модели, в которой этот признак вообще отсутствует.

На основе рассчитанных значений Шепли стоится график важности признаков, используемых в модели (рисунок 9). Полученный график интерпретируется следующим образом:

– значения слева от вертикальной оси относятся к негативному классу (0), а справа – к положительному классу (1) по матрице ошибок;

– толщина линии напрямую зависит от количества точек наблюдения;

– красные точки соответствуют наибольшим значениям признаков, а синие – наименьшим.

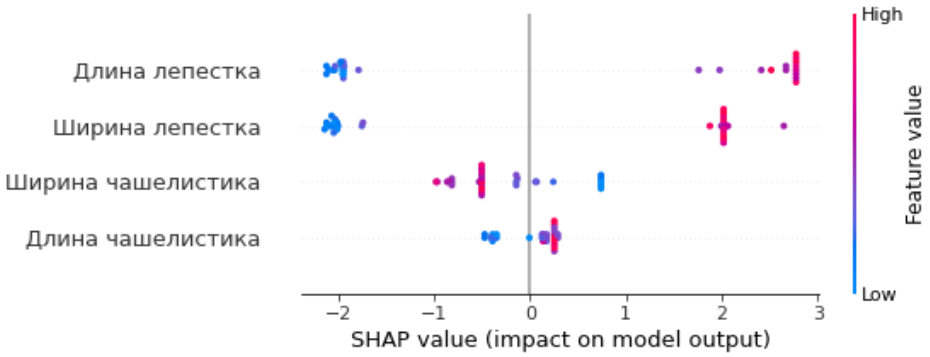


Рисунок 9 – SHAP-график важности признаков

График на рисунке 9 демонстрирует важность признаков классификации популярного набора данных «Ирисы Фишера». Однако, в стандартной задаче существует три класса, в данном же случае модель производит бинарную классификацию на разноцветные и виргинские ирисы. В качестве отрицательного класса выступает разноцветный ирис, а в качестве положительного – виргинский.

Таким образом, при большей длине лепестка ирисы относятся к виргинскому классу, та же зависимость наблюдается с шириной лепестка и длиной чашелистика. Что касается ширины чашелистика то, чем больше данный показатель, тем выше вероятность отнесения ириса к классу разноцветный.

В данном разделе приведено описание объекта исследования, а именно данных пациентов с клещевыми инфекциями. Также приведен обзор методов предварительной обработки, машинного обучения и интерпретации результатов классификации.

3 Расчеты и аналитика

В рамках данного исследования будет разработано программное обеспечение, варианты использования которого представлены на рисунке 10. При работе с интерактивным дашбордом пользователь может изучить особенности и структуру исходных данных с помощью таблицы данных и графиков зависимостей между переменными. Также есть возможность ознакомиться с результатами работы моделей машинного обучения посредством визуализированных ROC-кривых и графиков важностей признаков.

Рисунок 10 – UML-диаграмма вариантов использования приложения

Для разработки и создания программного обеспечения, реализующего функционал, описанный выше, были определены следующие этапы:

1. Загрузка, предварительный анализ и обработка исходного набора данных;
2. Определение целевой переменной и предикторов, деление набора данных на тренировочную и тестовую выборки;

3. Подбор оптимальных гиперпараметров и обучение моделей классификации;

4. Оценка качества работы моделей, определение важности признаков;

5. Создание интерактивного дашборда на основе результатов проведенного анализа.

3.1 Выбор программного обеспечения и инструментов разработки

Существует большое количество средств для решения аналитических задач и разработки программных средств анализа и визуализации. Для анализа данных широко используются следующие инструменты: язык программирования Python, язык программирования R, электронные таблицы Excel, программные продукты Apache Software Foundation и т.д. Каждый из инструментов имеет свои достоинства и недостатки, а также специализируется на решении определенного круга задач.

Продукты Microsoft Office и Apache Software Foundation являются лицензионными, что предполагает платное использование, а также узконаправленными. MS Excel ограничен в функционале, скорость работы снижается при увеличении размеров анализируемых данных. Apache Spark специализируется на анализе больших объемов информации, применение данного инструмента для решения задачи текущего исследования нецелесообразно. Среди языков программирования Python и R большинство разработчиков отдает предпочтение Python, он имеет множество встроенных постоянно развивающихся библиотек, синтаксис достаточно прост, существует большое количество документации и т.д. Был выбран язык программирования Python версии 3.9.6 с использованием библиотек (модулей), информация о которых представлена в таблице 5. [28].

Таблица 5 – Характеристика используемых инструментов языка программирования Python

№	Название библиотеки (модуля)	Версия	Описание
1	pandas	1.3.2	Библиотека с открытым исходным кодом, предоставляющая высокопроизводительные и простые в использовании структуры данных и инструменты анализа данных.
2	numpy	1.21.5	Библиотека, предоставляющая объект многомерного массива, различные производные объекты и набор процедур для быстрых операций с массивами.
3	matplotlib	3.4.3	Комплексная библиотека для создания статических, анимированных и интерактивных визуализаций.
4	sklearn	0.24.2	Библиотека для решения задач машинного обучения, прогнозного анализа данных.
5	scikitplot	0.3.7	Библиотека, добавляющая функции построения графиков объектам scikit-learn.
6	xgboost	1.4.2	Распределенная библиотека, реализующая алгоритмы машинного обучения, обеспечивающая высокую эффективность, гибкость и портативность.
7	shap	0.40.0	Аддитивные объяснения Шэпли, библиотека, реализующая теоретико-игровой подход для объяснения результатов моделей машинного обучения.
8	dash	2.2.0	Фреймворк Python для создания веб-приложений. Разработано компанией Plotly.
9	plotly	5.4.0	Интерактивная библиотека визуализации данных.

Для создания веб-приложения был использован веб-фреймворк Dash Plotly, на языке программирования Python [29]. Веб-фреймворк Dash позволяет программистам Python разрабатывать приложения для анализа данных и интерактивные информационные панели.

Одним из основных преимуществ использования Dash является то, что данный фреймворк позволяет создавать интерактивные данные, аналитику, веб-приложения и интерфейсы, используя Python, не требуя от разработчика углубленных знаний HTML, CSS или JavaScript.

3.2 Загрузка, предварительный анализ и предобработка данных

Исходный набор данных хранится в электронной таблице Excel, включающей 9 листов. Лист «Общие сведения» содержит общую информацию о пациенте, оценку витальных функций, анамнез жизни, жалобы, а также объективный статус. «Температурный лист» содержит сведения о ежедневных измерениях температуры, давления и пульса. «Общий анализ крови», «Биохимический анализ крови», «Иммуноферментный анализ», «Анализ спинномозговой жидкости», «Электрокардиограмма», «Окулист», «Ультразвуковое исследование» содержат результаты соответствующих анализов, исследований и осмотров.

Данные из всех листов электронной таблицы были считаны в объекты «DataFrame» с помощью библиотеки «pandas». После чего была произведена проверка данных на наличие пропущенных значений с помощью «df.isnull().sum()». Ориентируясь на количество записей на листе «Общие сведения», было определенно количество пропущенных записей на остальных листах. В таблице 6 представлены результаты проверки.

Оказалось, что в большинстве листов пропущено больше 40% значений, поэтому для дальнейшей работы были оставлены два листа «Общие сведения» и «Общий анализ крови». Что касается листа «Иммуноферментный анализ», он тоже был исключен, так как диагноз

определяется по результатам данного исследования, т.е. между диагнозом и признаками из данного листа существует известная зависимость.

Таблица 6 – Пропущенные записи на листах электронной таблицы

Название листа	Количество пропущенных записей
Общие сведения	0
Температурный лист	100
Общий анализ крови	7
Биохимический анализ крови	105
Иммуноферментный анализ	4
Анализ спинномозговой жидкости	89
Электрокардиограмма	177
Окулист	175
Ультразвуковое исследование	195

Из оставшихся двух листов был сформирован общий набор данных по 194 пациентам, информация о 7 пациентах, сведения об общем анализе крови которых отсутствовала, было решено также исключить из дальнейшей работы. В полученном наборе данных была проведена проверка на наличие пропущенных значений в признаках. В таблице 7 представлены признаки, которые было решено исключить, и соответствующие им количества пропущенных значений.

Таблица 7 – Пропущенные значения в исключенных признаках

Название признака	Количество пропусков
Дата выписки	100
Дата присасывания клеща	109
Дата начала заболевания	100
Температура при поступлении	64
Пульс в мин	110
Систолическое давление	105
Диастолическое давление	106
Скорость оседания эритроцитов	47
Гематокрит	141
Средний объем эритроцитов	141
Среднее содержание гемоглобина	141

Средняя концентрация гемоглобина	141
Ширина распределения эритроцитов	141
Ширина распределения эритроцитов	142
Средний объем тромбоцита	143
Тромбоциты	115
Ширина распределения тромбоцитов	142
Тромбокрит	142
Количество крупных тромбоцитов	184
Отношение объема крупных тромбоцитов	143

Также из дальнейшей работы были исключены следующие признаки:

1. № медицинской карты – 100 пропущенных значений, вместо него создан признак «id» (для каждого пациента уникальное число, соответствующее его порядковому номеру в наборе данных).

2. Дата поступления (госпитализации) – в разных форматах (у 98 пациентов указан исключительно год, у 96 – дата в формате «дд.мм.гггг»). Показатель исключен, т.к. в каждом году встречались все диагнозы, а отследить зависимость между датой и диагнозом по всем пациентам невозможно.

3. Форма КЭ – признак, напрямую связанный с целевой переменной.

4. Форма ИКБ – признак, напрямую связанный с целевой переменной.

5. Поражение органов при ИКБ – признак, напрямую связанный с целевой переменной.

6. 2-х волновое течение КЭ – признак, напрямую связанный с целевой переменной.

7. Осложнения КЭ Отек и набухание головного мозга / судороги – признак, напрямую связанный с целевой переменной.

8. Осложнения КЭ Парезы – признак, напрямую связанный с целевой переменной.

9. Дата введения антибиотика – признак, напрямую связанный с целевой переменной.

10. Дата введения иммуноглобулина – признак, напрямую связанный с целевой переменной.

11. План лечения (Этиотропный препарат) – признак, напрямую связанный с целевой переменной.

Пропущенные значения также были найдены в признаках «Вес» и «Рост», на основе заполненных данных был рассчитан индекс массы тела и создан категориальный признак, принимающий значение 1, если рассчитанный показатель ниже 18,5 или выше 25, и 0, если показатель входит в диапазон или индекс не был рассчитан.

Оставшиеся признаки имеют категориальный тип и если обнаруживались пропуски, то они заменялись на значение 0, что говорит об отсутствии данного признака у того или иного пациента.

Также пропущенные значения присутствовали и в некоторых признаках общего анализ крови:

1. Эозинофилы / EOS% – 16 пропущенных значений
2. Лимфоциты / LYM% – 7 пропущенных значений
3. Базофилы / BAS% – 122 пропущенных значения

Данные пропуски были восстановлены на основе непропущенных значений других признаков с помощью следующей формулы [30]:

$$EOS\%+ NEU\%+ LYM\%+MON\%+BAS\%=100$$

Таким образом, в результате чистки данных сформировался набор с 71 признаком по 194 пациентам. Не все категориальные признаки принимают закодированные значения (0 и 1), а именно, сопутствующие заболевания и локализация присасывания клеща. Значения данных признаков представляют собой слова или списки слов (если у одного пациента более одного сопутствующего заболевания или укусы клеща были в нескольких местах). Поэтому было решено составить список уникальных значений признаков, на основе которого создать новые категориальные признаки, принимающие значение 1, если у пациента присутствует то или иное сопутствующее заболевание или присасывание было на той или иной области тела. В результате кодировки данных двух признаков число признаков увеличилось с 71 до 127.

Проверке на наличие выбросов были подвержены числовые признаки, а именно показатели анализа крови. Для обнаружения выбросов был рассчитан межквартильный размах. Значения, оказавшиеся меньше разности первого квартиля и полутора межквартельного размаха, и больше суммы третьего квартиля и полутора межквартельного размаха, были рассмотрены и, оказалось, что они входят в пределы допустимых значений показателей крови и не были приняты за выбросы.

3.3 Деление данных на обучающую и тестовую выборки

В качестве целевой переменной был выбран признак «Диагноз», соответственно, остальные признаки будут выступать в качестве независимых переменных (предикторов).

Прежде чем приступить к обучению моделей, набор данных был разделен на две выборки: обучающую и тестовую. Первая выборка предназначена для обучения моделей классификации, а вторая – для оценки качества работы моделей классификации.

Было решено разделить набор в соотношении 0,7 к 0,3 на тренировочную и тестовую выборки, соответственно. Для разбиения данных был использован метод «sklearn.model_selection.train_test.split()», принимающий в качестве параметров зависимую и независимые переменные, а также размер тестовой выборки.

3.4 Построение классификаторов

3.4.1 Выбор моделей классификации и поиск оптимальных гиперпараметров

Для классификации было решено использовать следующие модели и соответствующие им функции:

Дерево решений – «sklearn.tree.DecisionTreeClassifier»;

Логистическая регрессия – «sklearn.linear_model.LogisticRegression»;

Случайный лес – «sklearn.ensemble.RandomForestClassifier»;

Градиентный бустинг – «xgboost.XGBClassifier».

Для каждой модели был проведен подбор оптимальных гиперпараметров с помощью исчерпывающего поиска по сетке «sklearn.model_selection.GridSearchCV». В качестве метрики качества на вход функции был подан параметр точность (accuracy), также было указано, что перекрестная проверка будет проводиться пятикратно.

В таблице 8 представлены гиперпараметры и принимаемые ими значения, из которых проводился поиск оптимальных.

Таблица 8 – Гиперпараметры классификаторов для выбора оптимальных с помощью сеточного поиска

Классификатор	Гиперпараметр	Принимаемые значения
Дерево решений	Критерий расщепления (criterion)	энтропия (entropy), коэффициент Джини (gini)
	Максимальная глубина дерева (max_depth)	3 – 10
	Пороговое значение критерия расщепления для разделения данных в узле (min_impurity_decrease)	0.3, 0.2, 0.1
Логистическая регрессия	Инверсия коэффициента регуляризации C	0.001, 0.01, 0.1, 1.0, 100.0, 1000.0
	Функция штрафа (penalty)	L1-регуляризация, L2-регуляризация, ElasticNet
	Алгоритм оптимизации (solver)	алгоритм Ньютона – сопряженных градиентов (newton-cg), алгоритм BFGS с ограниченной памятью (lbfgs), LIBLINEAR оценка линейной комбинации признаков (liblinear)
Случайный лес	Количество оценщиков (n_estimators)	5-19
	Критерий расщепления	энтропия (entropy),

	(criterion)	коэффициент Джини (gini)
	Минимальное количество примеров в узле для расщепления (min_samples_split)	2-6
	Максимальная глубина одного дерева в модели (max_depth)	3-10, не ограничена
XGBoost	Алгоритм бустинга (booster)	gbtree - настройка на основе слабых учеников, gblinear - настройка с использованием линейной регрессии с L1 и L2 сжатием, dart - метод обучения с использованием прореживания
	Уменьшение размера шага η (eta)	0.1, 0.3, 0.5, 0.7
	Количество деревьев градиентного бустинга (n_estimators)	5-19
	Коэффициент L2-регуляризации λ (lambda)	0.01, 0.1, 0.0, 1.0, 10.0
	Коэффициент L1-регуляризации α (alpha)	0.01, 0.1, 0.0, 1.0, 10.0
	Максимальная глубина дерева (max_depth)	3-10

3.4.2 Построение классификаторов с оптимальными гиперпараметрами

После того, как для всех моделей классификации были найдены оптимальные гиперпараметры, модели классификации были обучены на тренировочном наборе. На вход обученным моделям подавались неразмеченные данные из тестового набора.

Затем на основе предсказанных и реальных меток тестового набора были рассчитаны такие показатели, как чувствительность и специфичность, с помощью функций «sklearn.metrics.recall_score», «sklearn.metrics.make_scorer (recall_score, pos_label=0)», соответственно.

Также проведена визуализация ROC-кривых и SHAP-важностей признаков с помощью «scikitplot.metrics.plot_roc()», «shap.summary_plot (shap.TreeExplainer().shap_values())», соответственно.

В данном разделе были определены основные этапы работы по созданию программного обеспечения для анализа данных пациентов с клещевыми инфекциями.

4 Результаты

4.1 Классификация диагнозов пациентов с клещевыми инфекциями

Для моделей классификации были найдены следующие оптимальные значения гиперпараметров:

- `DecisionTreeClassifier(criterion='entropy', max_depth=3, min_impurity_decrease=0.1);`
- `LogisticRegression(C=0.1, penalty='l2', solver='newton-cg');`
- `RandomForestClassifier(criterion='entropy', max_depth=6, min_samples_split=3, n_estimators=12);`
- `XGBClassifier(alpha=0.01, booster='gbtree', eta=0.7, reg_lambda=0.01, max_depth=3, n_estimators=5).`

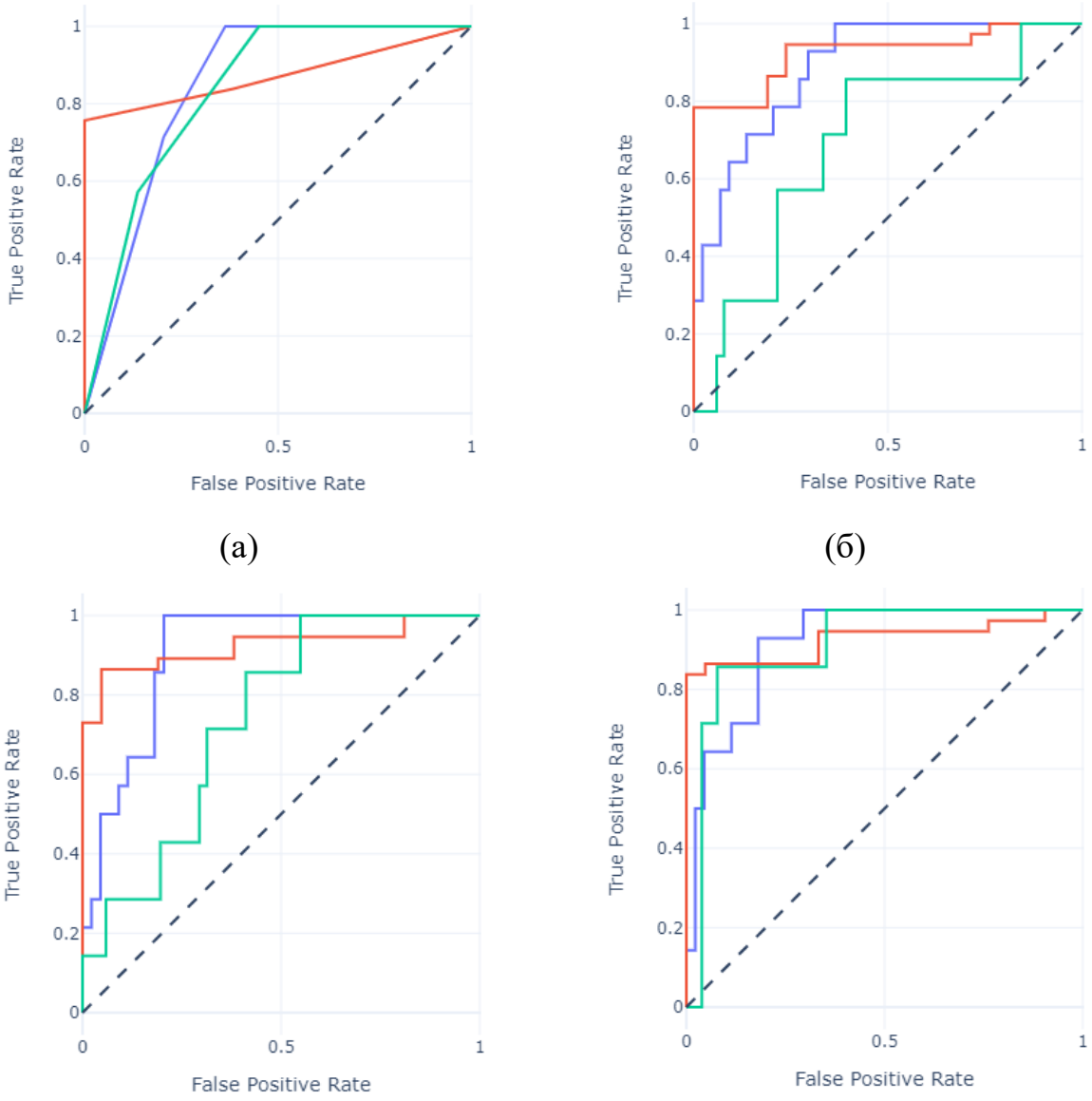
Для оценки качества работы моделей были рассчитаны значения чувствительности и специфичности. В таблице 9 представлены результаты расчетов.

Таблица 9 – Средние значения и отклонения метрик качества

Алгоритм	Чувствительность, отклонение	Специфичность, отклонение
Дерево решений	0,67±0,05	0,7±0,07
Логистическая регрессия	0,75±0,05	0,77±0,05
Случайный лес	0,81±0,07	0,79±0,04
Градиентный бустинг	0,77±0,04	0,78±0,05

Следует отметить, что лучший результат по обеим метрикам показал такой алгоритм машинного обучения, как случайный лес. В 81% случаев данным алгоритмом были верно отнесены пациенты к заболеванию, которое действительно было диагностировано у них. В 79 % случаев случайный лес верно определил отсутствие принадлежности пациентов к тому или иному диагнозу.

Также для дерева решений, логистической регрессии, случайного леса и градиентного бустинга были построены ROC-кривые, представленные на рисунке 11 (а), (б), (в) и (г), соответственно. Кривые красного цвета соответствуют диагнозу клещевой энцефалит, синего – иксодовому клещевому боррелиозу, зеленого – микст-инфекции, соответственно.



(а) (б)

(в) (г)

Рисунок 11 – ROC-кривые результатов классификации различными методами

По графикам можно сделать вывод, что наилучшие результаты по выявлению всех трёх диагнозов, клещевого энцефалита, иксодового

клещевого боррелиоза и микст-инфекции, показал градиентный бустинг. В таблице 10 представлены значения площадей под ROC-кривыми для каждого отдельного диагноза.

Таблица 10 – Площади под ROC-кривыми для отдельных классов

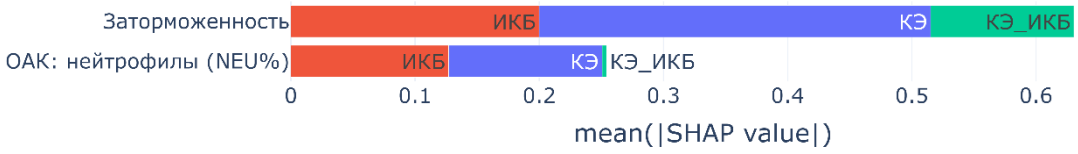
	Клещевой энцефалит (КЭ)	Иксодовый клещевой боррелиоз (ИКБ)	Микст-инфекция КЭ и ИКБ
Дерево решений	0,87	0,85	0,83
Логистическая регрессия	0,93	0,89	0,69
Случайный лес	0,92	0,91	0,74
Градиентный бустинг	0,93	0,92	0,91

Как было отмечено ранее, наилучший результат по выявлению трех диагнозов показал градиентный бустинг, затем идет дерево решений. Что касается логистической регрессии и случайного леса, то данными алгоритмами плохо определяется такой диагноз, как микст-инфекция клещевого энцефалита и иксодового клещевого боррелиоза.

4.2 Важности признаков

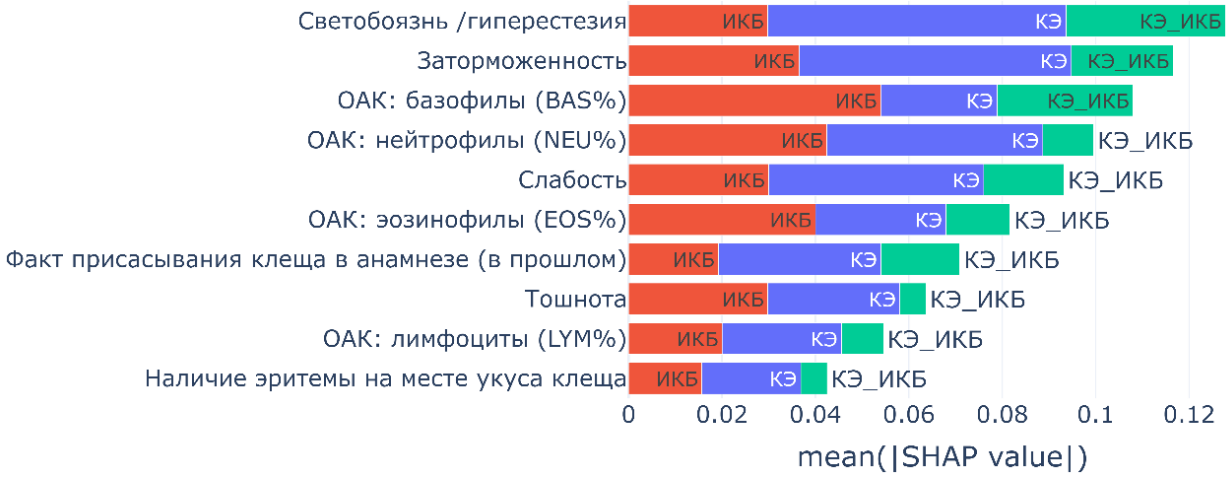
Для определения степени влияния отдельных предикторов на исход классификации были построены графики SHAP-важности признаков для дерева решений, логистической регрессии, случайного леса и градиентного бустинга, представленные на рисунке 12 (а), (б), (в) и (г), соответственно.

В данном случае классификация является многоклассовой, поэтому графики представлены в виде столбчатых диаграмм с накоплением. Длина столбиков красного цвета соответствуют средним значениям показателя Шепли признаков, оказывающих влияние на определение иксодового клещевого боррелиоза, синего цвета – клещевого энцефалита и зеленого – микст-инфекцци ИКБ и КЭ.



(а)

(б)



(в)

(г)

Рисунок 12 – SHAP-важности признаков при классификации диагнозов клещевых инфекций

Для всех четырех алгоритмов наиболее важным признаком при классификации пациентов по диагнозу является наличие такого клинического проявления, как заторможенность и процентное содержание в крови таких клеток, как нейтрофилы. Также в числе первых для логистической регрессии, случайного леса и градиентного бустинга находится наличие или отсутствие светобоязни или гиперестезии и слабости.

4.3 Разработка дашборда

Поскольку веб-приложение предназначено как для анализа данных, так и для интерпретации работы моделей машинного обучения, было решено разделить его на два условных блока: элементоы визуализации исходных данных и интерпретации результатов работы моделей машинного обучения.

В верхней части блока интерфейса для представления данных расположена таблица, отображающая общую структуру загруженного файла (рисунок 13).

Таблица содержит следующие функции:

1. Пагинация страниц при просмотре таблицы для лучшей производительности.
2. Сортировка данных по одному или нескольким столбцам.
3. Фильтрация данных по столбцам.
4. Отображение данных в виде гистограммы при выборе столбца.

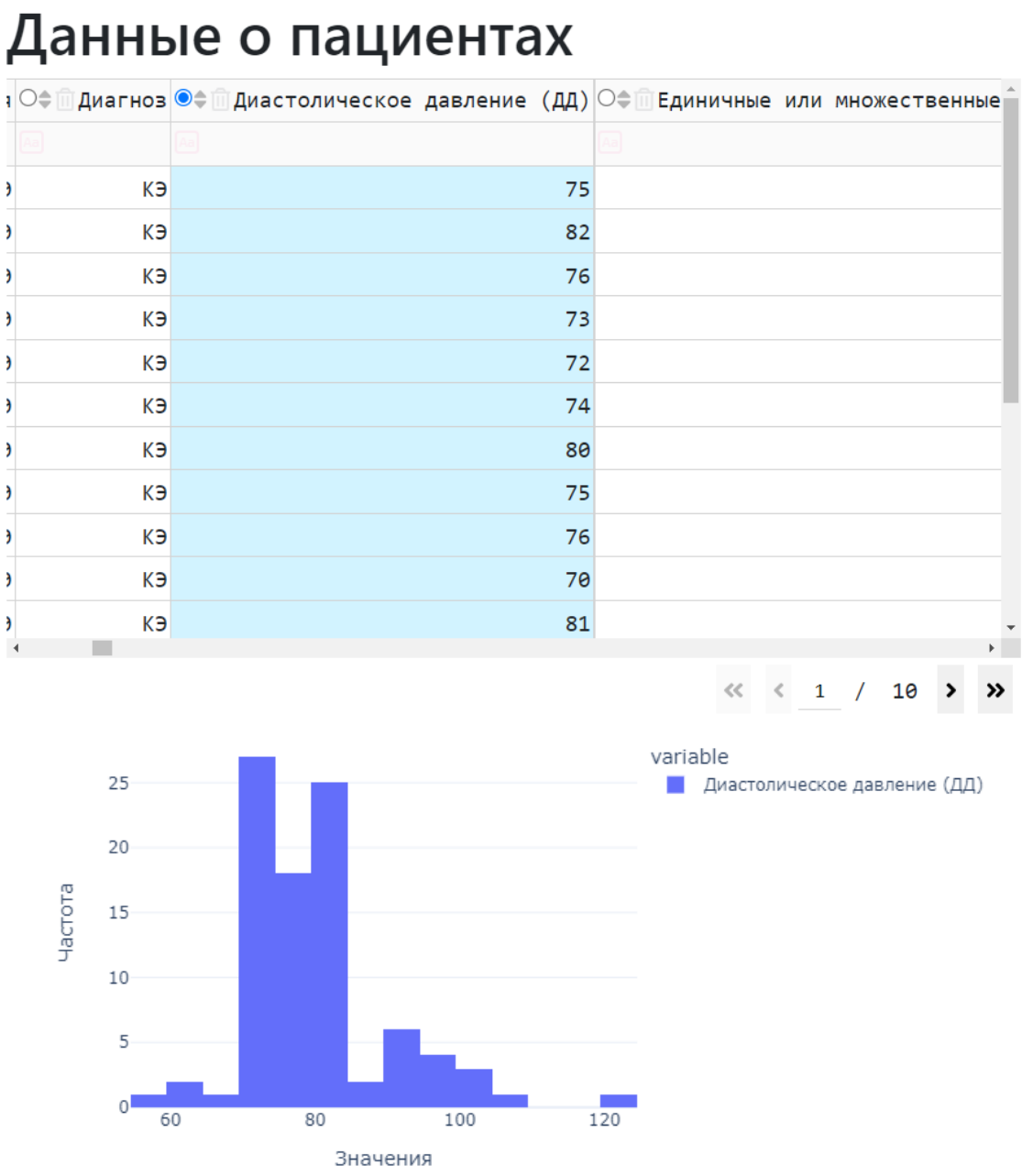


Рисунок 13 – Распределение значений выбранного столбца

Также был реализован линейный график зависимости количества пациентов от даты поступления для анализа динамики заболеваемости, цветом указан диагноз (рисунок 14).

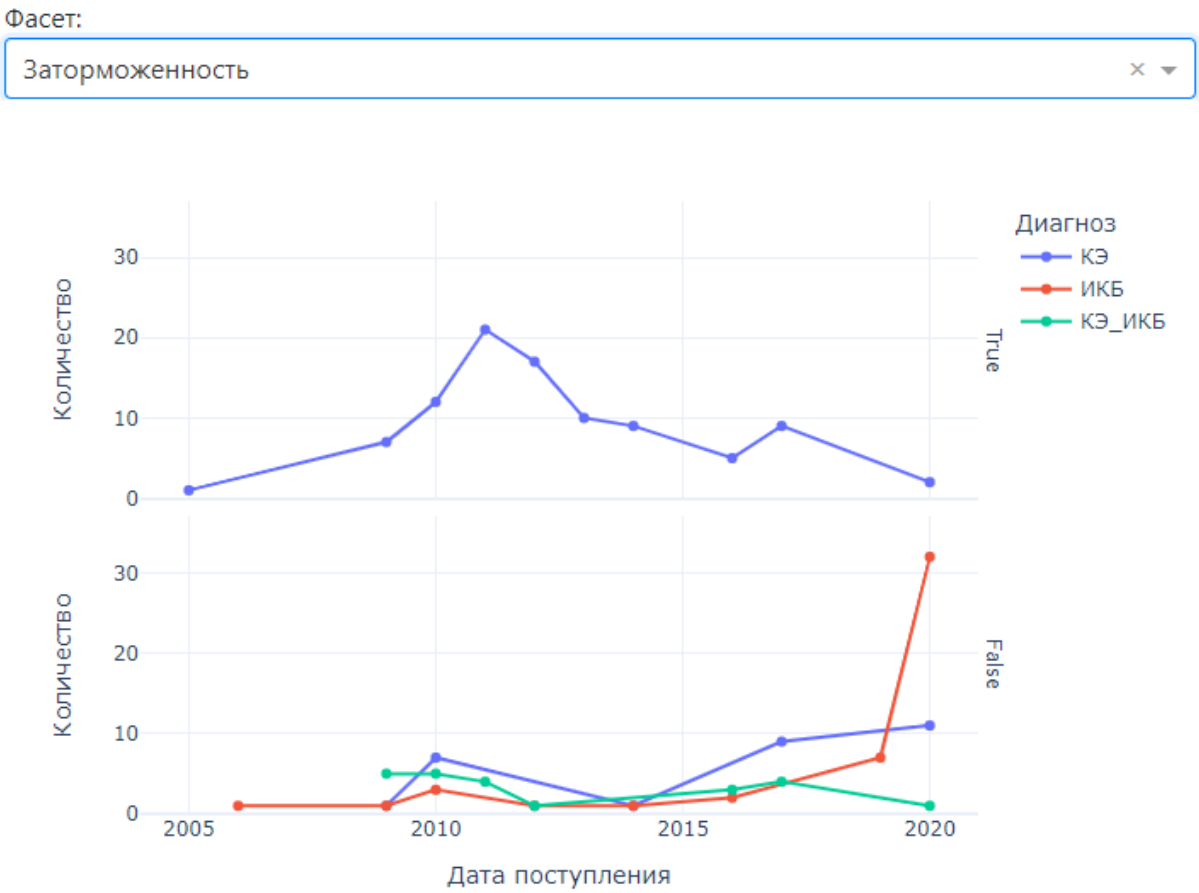


Рисунок 14 - Динамика заболеваемости у мужчин и женщин по диагнозам

Дополнительно указывается категориальная переменная для формирования фасета. Фильтрация категориальных признаков выполняется автоматически.

Также была реализована диаграмма размаха для визуализации числовых признаков. На рисунке 15 представлен график размаха возраста в зависимости от половой принадлежности для каждого из диагнозов.

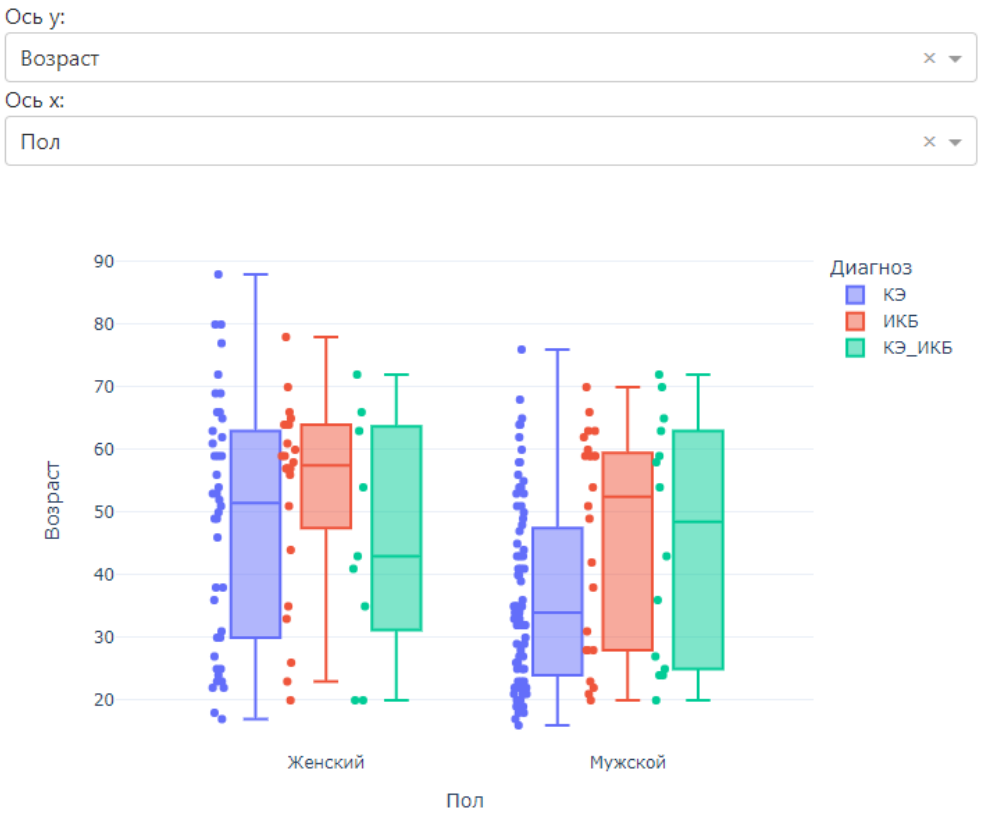


Рисунок 15 - Диаграмма размаха возраста по диагнозу и полу

Для визуализации данных о локализации присасывания клеща была использована диаграмма разброса (рисунок 16). Цветом и размером точек указано количество укусов в процентах. Вместо стандартного размещения точек на координатной плоскости было решено использовать схематическое изображение тела человека. Расположение точек на векторном силуэте человека спереди и сзади отображает локализацию присасывания вне зависимости от правой или левой сторон.

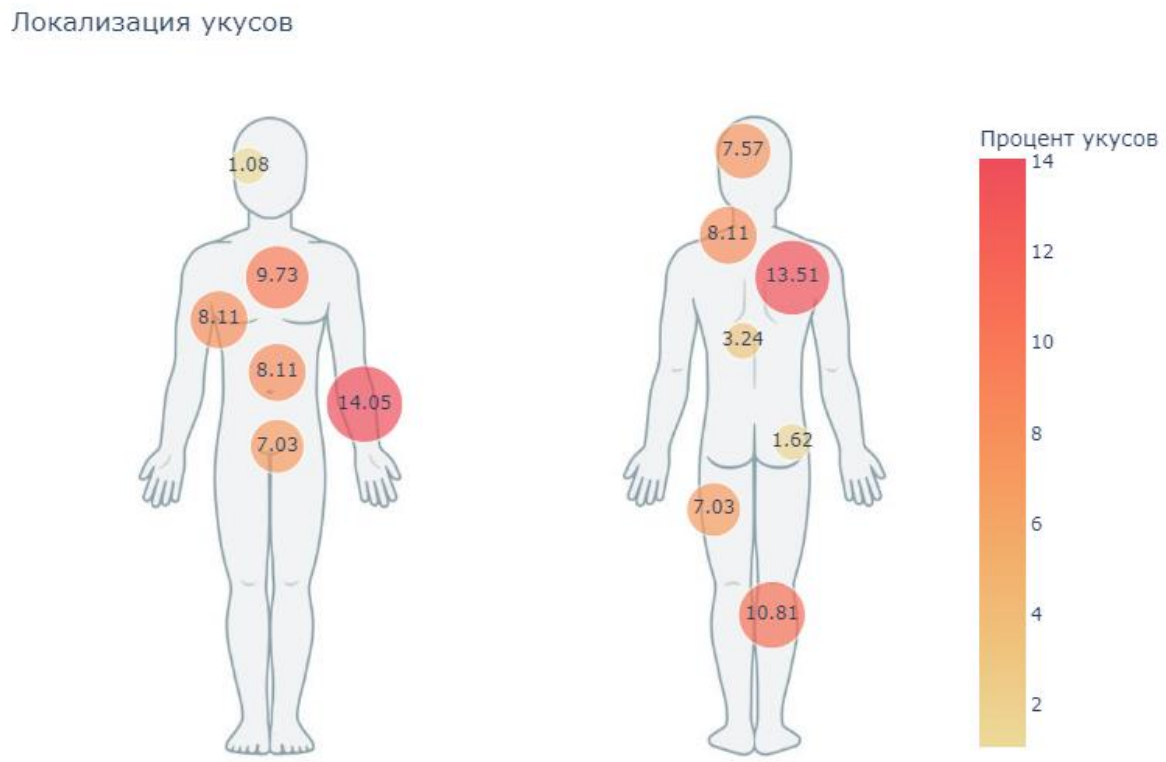


Рисунок 16 – Локализация укусов клещей

Таким образом, разработанный блок интерфейса позволяет более подробно изучить структуру исходных данных: определить характер распределения значений признаков, оценить динамику заболеваемости в зависимости от значений категориальных переменных по диагнозам, определить процентное соотношение укусов в зависимости от их локализации.

Равное соотношение классов в тренировочной и тестовой выборках является важным аспектом при обучении моделей классификации.

Результаты разделения данных на тренировочную и тестовую выборки были представлены в виде круговой диаграммы с применением px.sunburst (рисунок 17). Данный тип графика удобен для отображения иерархических данных.

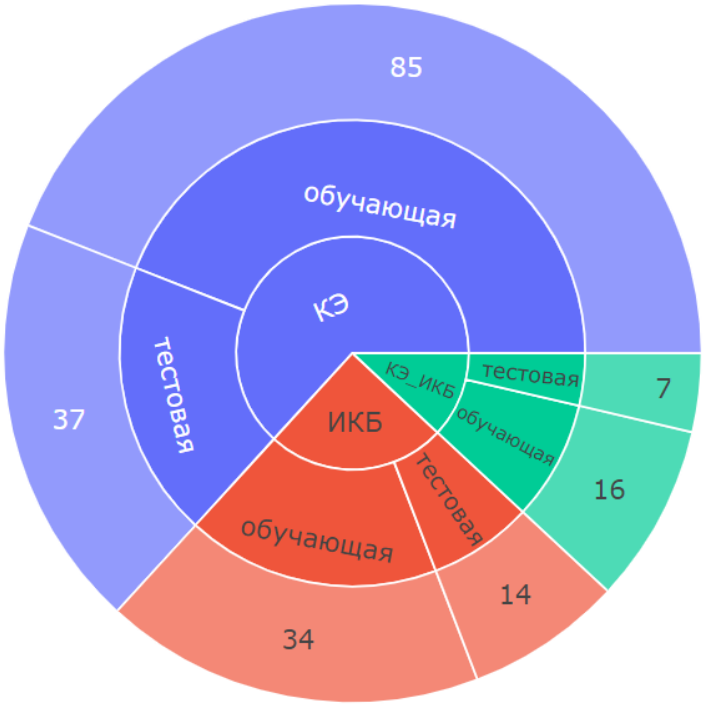


Рисунок 17 – Визуализация соотношения тренировочной и тестовой выборок по диагнозам

При выборе корневой переменной график отображает данные потомков (рисунок 18).

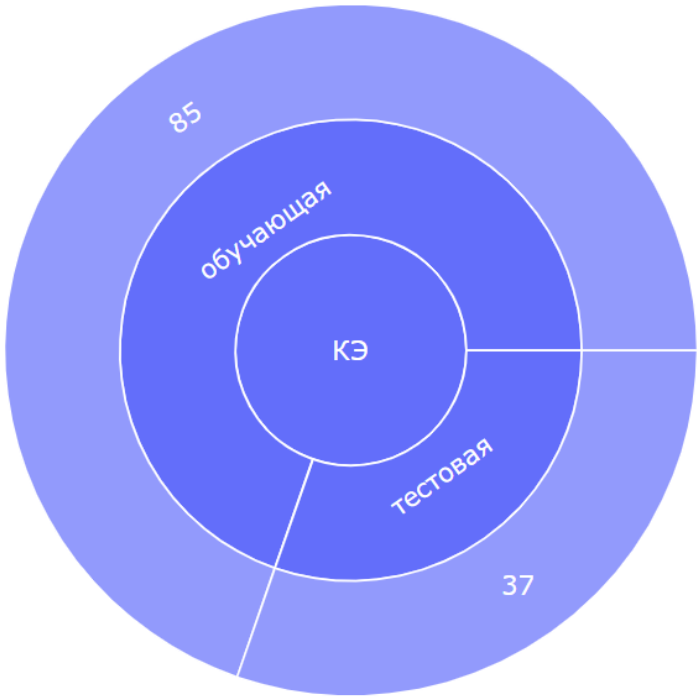


Рисунок 18 – Разделение тестовой и тренировочной выборки класса КЭ

Для визуализации результатов обучения моделей классификации использовался график Multiclass ROC Curve (рисунок 19).

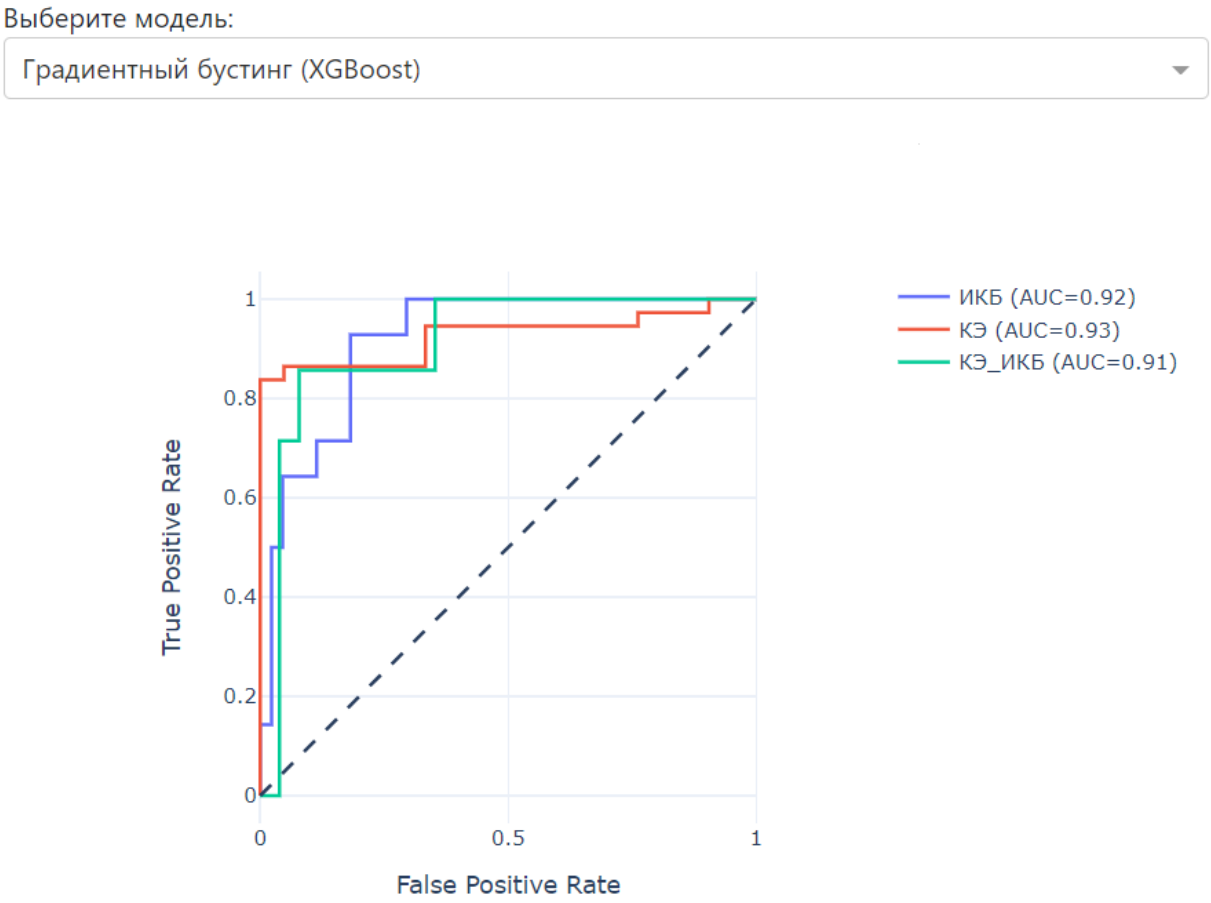


Рисунок 19 – Результат работы модели градиентного бустинга для многоклассовой классификации

Существует множество моделей машинного обучения, которые очень точны и высокоэффективны при прогнозировании. Одно из ограничений этих моделей, заключается в том, что трудно объяснить качество получаемых ими результатов. Всегда необходимо сделать результаты модели более объяснимыми. Для этого был использован инструмент «SHAP (SHAPley Additive exPlanations)», который может помочь сделать результаты моделей машинного обучения более объяснимыми (рисунок 20).

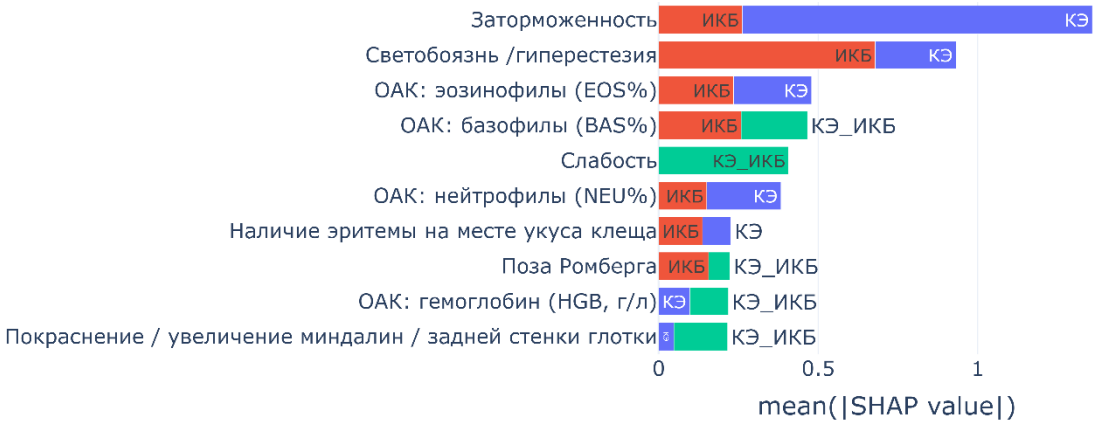


Рисунок 20 – Вклад каждой функции в прогноз по диагнозам

Таким образом, разработанный на данном этапе блок интерфейса позволяет оценить результаты работы классификации данных с помощью различных методов классификации: качество работы моделей с помощью ROC-кривых, степень важности влияния предикторов на целевую переменную с помощью SHAP.

В данном разделе представлены результаты предварительной обработки данных, обучения классификаторов, разработки интерактивного дашборда для анализа данных пациентов с клещевыми инфекциями.

5 Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

Настоящая работа посвящена разработке алгоритма визуализации и анализа данных пациентов с инфекциями, передающимися клещами, а также интеграции данного алгоритма в программный веб-интерфейс. Данное программное обеспечение может найти свое применение в деятельности медицинских работников и исследователей в сфере медицины. Функциональные возможности разработанного приложения позволят сократить время, а также повысить качество анализа данных из медицинских документов, увеличить точность принятия решений о способе лечения пациентов на основе прогнозных значений, полученных обученными моделями машинного обучения.

В качестве потенциальных потребителей результатов научно-исследовательского проекта могут выступать организации здравоохранения, медицинский персонал, а также представители научного сообщества, заинтересованные в продолжении исследования или в использовании его технических результатов.

С целью эффективного использования научного потенциала проекта необходимо уделять внимание, как разработке, так и проведению её анализа с точки зрения экономических требований. В качестве задач данного раздела выступают следующие:

– Определить перспективность научно-исследовательского проекта с помощью технологии «QuaD»;

– Проанализировать причины отсутствия или низкого уровня анализа медицинскими специалистами данных для повышения качества диагностики заболеваний, используя диаграмму Исикавы;

– Исследовать внешнюю и внутреннюю среды проекта с помощью SWOT-анализа;

– Провести инициацию научно-исследовательского проекта;

– Составить план научно-исследовательского проекта, провести планирование бюджета;

– Провести анализ рисков.

5.1 Предпроектный анализ

5.1.1 Технология «QuaD»

В виду того, что разработки, решающие задачи, подобные задачам данного исследования используются исключительно в рамках конкретного образовательного или медицинского учреждения и не распространяются за его пределы, анализ конкурентных технических решений невозможен ввиду отсутствия открытых данных о наличии и свойствах подобных решений.

Одним из инструментов измерения характеристик качества новой разработки и ее перспективности на рынке является технология «QuaD» [31], которая позволяет принимать решение о целесообразности вложения денежных средств в научно-исследовательский проект. В основе данной технологии лежит нахождение средневзвешенного значения показателя качества и перспективности научной разработки $П_{ср}$ (1).

$$П_{ср} = B_i \cdot Б_i, \quad (1)$$

где B_i – вес показателя; $Б_i$ – средневзвешенного значение показателя.

Если значение показателя $П_{ср}$ варьируется в диапазоне от 100 до 80, то разработку принято считать перспективной. В случае, когда данный показатель входит в диапазон от 79 до 60, то перспективность разработки выше среднего. Средняя, ниже среднего и низкая перспективность считается при значениях показателя от 69 до 40, от 39 до 20 и от 19 и ниже, соответственно. Результаты расчетов средневзвешенных значений показателей представлены в таблице 11.

Таблица 11 – Оценочная карта технологии «QuaD»

Критерий оценки	Вес критерия	Баллы	Максимальный балл	Относительное значение	Средневзвешенное значение
Энергоэффективность	0,1	100	100	1	0,1
Помехоустойчивость	0,1	90	100	0,9	0,09
Надежность	0,15	95	100	0,95	0,1425
Унифицированность	0,05	70	100	0,7	0,035
Время выполнения алгоритма	0,05	100	100	1	0,05
Пользовательский интерфейс	0,1	100	100	1	0,1
Безопасность	0,1	100	100	1	0,1
Потребность в ресурсах памяти	0,1	75	100	0,75	0,075
Функциональная мощность	0,05	80	100	0,8	0,04
Простота эксплуатации	0,1	100	100	1	0,1
Качество интеллектуального интерфейса	0,05	80	100	0,8	0,04
Прозрачность кода	0,05	100	100	1	0,05
Итого	1				0,9225

В виду того, что у рассматриваемой разработки наиболее важные критерии имеют высокие показатели, что в сумме составляет 0,9225, разработка является перспективной.

5.1.2 Диаграмма Исикавы

Диаграмма причины-следствия Исикавы позволяет графически проанализировать и сформировать причинно-следственные связи, это инструментальное средство для систематического определения причин проблемы и последующего графического представления [32].

Проблемной областью анализа является отсутствие или низкий уровень анализа медицинскими специалистами данных для повышения качества диагностики заболеваний. На рисунке 21 представлена причинно-следственная диаграмма Исикавы.

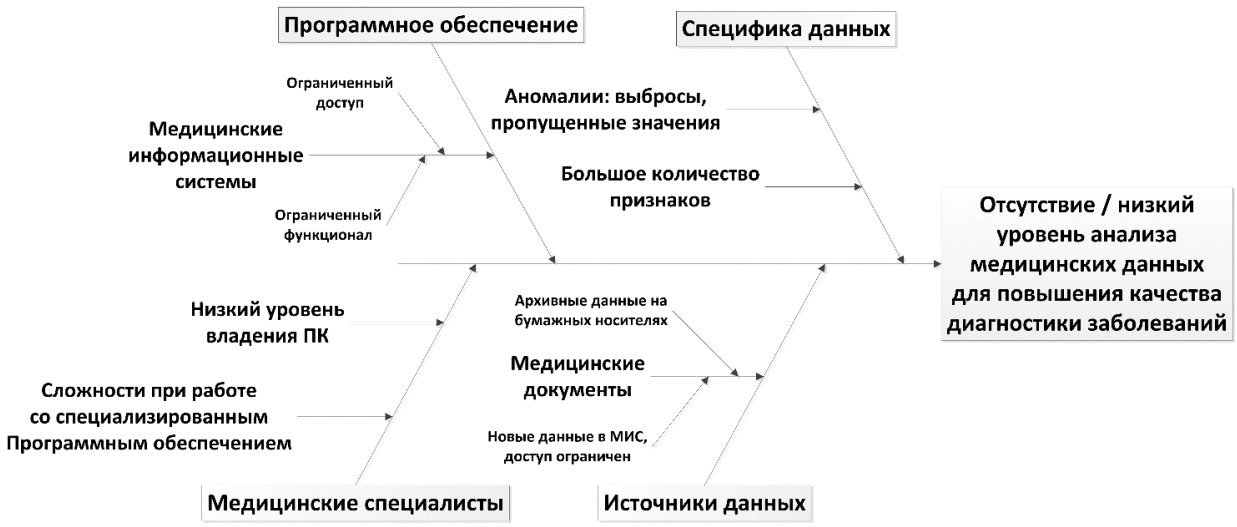


Рисунок 21 – Причинно-следственная диаграмма

Источниками данной проблемы могут быть медицинские специалисты, обладающие низким уровнем владения персональным компьютером (ПК), а также не имеющие опыта работы с инструментами анализа данных. Проблемы возникают также из-за источников данных, а именно медицинских документов, которые частично хранятся на бумажных носителях, а с внедрения медицинских информационных систем (МИС) представлены в виде отчетов по формам, сформированным на основе внесенных в МИС данных. Доступ к таким данным из-за наличия персональных данных пациентов ограничен. Помимо ограниченного доступа к собранным сведениям в МИС существует также проблема ограниченности функционала таких систем: данные вносятся в систему, формируются печатные формы, в дальнейшем собранные данные не используются. Непосредственно сами данные также выступают в качестве источника проблемы – они содержат множество аномалий, т.е. пропущенных значений и выбросов. Специалисты, заполняющие документы могут допустить ошибки. Не все значения могут быть заполнены, к примеру, потому что пациенту не было назначено то или иное обследование. Очень часто размеры сформированных наборов медицинских данных по количеству признаков значительно большие, однако

по количеству записей – наоборот, особенно после чистки данных очень малы.

5.1.3 SWOT-анализ

SWOT — Strengths (сильные стороны), Weaknesses (слабые стороны), Opportunities (возможности) и Threats (угрозы) — представляет собой комплексный анализ научно-исследовательского проекта. SWOT-анализ применяют для исследования внешней и внутренней среды проекта. В таблице 12 представлены результаты матрицы SWOT. исследования внешней и внутренней среды проекта.

Таблица 12 - SWOT-анализ

	Сильные стороны научно-исследовательского проекта: С1. Поддержка русского языка. С2. Обработка производится онлайн. С3. Простой интерфейс. С4. Не используется мощности клиентского оборудования.	**Слабые стороны научно-исследовательского проекта:** Сл1. Решается узкий круг задач. Сл2. Требуется подключение к Интернету. Сл3. Зависимость от имеющихся данных.
Возможности: В1. Выявление скрытых закономерностей в данных. В2. Сокращение времени на постановку диагноза.	Веб-сервис поможет медицинским специалистам провести визуальный и интеллектуальный анализ данных для последующей постановки диагноза. При этом медицинскому специалисту достаточно иметь доступ в Интернет с любого компьютера в независимости от его вычислительных мощностей, т.е. на старых компьютерах веб-сервис будет исправно работать.	Возможность внедрения новых функциональных возможностей. Получение дополнительных данных для улучшения работы алгоритма.
Угрозы: У1. Отключение Интернета. У2. Хакерские атаки на	Внедрение дополнительного слоя защиты. Улучшение текущего	Создание прототипа в виде десктопного приложения.

сервер. У3. Появление новых разработок с оптимизированными решениями.	решения за счет получения новых данных	

Еще особенностью разрабатываемой системы является поддержка русского языка, доступность за счет подключения к сети Интернет и специализация на узкой задаче — визуализации и интеллектуальном анализе зависимостей между признаками данных пациентов с клещевыми инфекциями. Также благодаря независимости от вычислительных мощностей клиентского технического оборудования веб-сервис может запускаться на старых компьютерах.

Недостатком разрабатываемого решения является зависимость от Интернета. Поэтому в те медицинские и образовательные учреждения, которые его не имеют или временно его лишились, не смогут получить доступ к веб-сервису. Более того, сервер, на котором происходят вычисления, может быть подвергнут хакерской атаке. Если первую угрозу нельзя избежать, то со второй можно бороться с помощью усиления защиты исходного кода.

5.2 Инициация научно-исследовательского проекта

Группа процессов инициации включает процессы, выполняемые для определения нового научно-исследовательского проекта или новой фазы существующего. В рамках процессов инициации определяются цели, а также заинтересованные стороны проекта.

5.2.1 Цели и результаты научно-исследовательского проекта

В данном разделе приводится информация о заинтересованных сторонах научно-исследовательского проекта, иерархии целей проекта и

критериях достижения целей. Заинтересованные стороны проекта – это лица или организации, которые активно участвуют в проекте или интересы которых могут быть затронуты как положительно, так и отрицательно в ходе исполнения или в результате завершения проекта. Информация по заинтересованным сторонам научно-исследовательского проекта представлена в таблице 13.

Таблица 13 – Заинтересованные стороны научно-исследовательского проекта

Заинтересованные стороны научно-исследовательского проекта	Ожидания заинтересованных сторон
Отделение информационных технологий ТПУ	Научные публикации Защита магистерской диссертации
Медицинские специалисты	Инструмент для анализа и визуализации закономерностей и зависимостей между предикторами пациентов с клещевыми инфекциями
Пациенты	Сокращение длительности ожидания постановки диагноза, получение более качественных консультаций врачей
Научное сообщество	Алгоритм визуализации и анализа данных пациентов с клещевыми инфекциями

В таблице 14 представлена информация об иерархии целей научно-исследовательского проекта и критериях их достижения.

Таблица 14 – Иерархия целей научно-исследовательского проекта и критерии их достижения

Цели проекта:	Разработать программное обеспечение визуализации и анализа данных пациентов с инфекциями, передающимися клещами
Ожидаемые результаты проекта:	Программное обеспечение визуализации и данных пациентов с инфекциями, передающимися клещами
Требования к результату проекта:	Реализована загрузка данных из файлов
	Реализована визуализация зависимостей между значениями признаков
	Реализован выбор предикторов и целевой переменной
	Реализовано обучение моделей машинного обучения
	Реализована оценка качества работы моделей и визуализация данных результатов
	Реализована возможность построения прогнозов на новых данных
	Бесперебойная работа программных модулей проекта
	Формализованное описание работы программных модулей проекта

5.2.2 Организационная структура научно-исследовательского проекта

На данном этапе работы определяется состав рабочей группы научно-исследовательского проекта, определяются роли каждого участника в данном проекте. В таблице 15 определены участники научно-исследовательского проекта и их роли.

Таблица 15– Рабочая группа научно-исследовательского проекта

№ п/п	ФИО, основное место работы, должность	Роль в проекте	Функции	Трудо-затраты, час.
1	Аксёнов Сергей Владимирович, ОИТ ИШИТР ТПУ, доцент	Научный руководитель	Составление научных задач, контроль выполнения проекта, проверка разработки, проверка документации	165
2	Сафронов Василий Сергеевич, магистрант ОИТ ИШИТР ТПУ	Инженер	Проектирование, реализация	672
ИТОГО				837

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

5.2.3 Ограничения и допущения проекта

Ограничения проекта — это все факторы, которые могут послужить ограничением степени свободы участников команды проекта, а также «границы проекта» — параметры проекта или его продукта, которые не будут реализованных в рамках данного проекта. Ограничения проекта представлены в таблице 16.

Таблица 16 – Ограничения проекта

Фактор	Ограничения/ допущения
Бюджет проекта	459 329,79 руб.
Источник финансирования	НИ ТПУ
Сроки проекта:	4 месяца
Дата утверждения плана управления проектом	31.01.2022
Дата завершения проекта	31.05.2022

Таким образом, максимальный бюджет настоящего проекта установлен в сумме 459 329,79 рублей, а сроки выполнения составляют с 31 января по 31 мая.

5.3 Планирование управления научно-исследовательским проектом

5.3.1 План научно-исследовательского проекта

При планировании научно-исследовательского проекта были построены календарный план и диаграмма Ганта (рисунок 22). Календарный план включает информацию об этапах работы, длительности их выполнения и датах начала и завершения выполнения.

Диаграмма Ганта является наиболее удобным и наглядным способом отслеживания выполнения проектной работы [33]. Данная диаграмма представляет собой ленточный график, на котором работы представляются протяженными во времени отрезками, характеризующимися датами начала и окончания выполнения данных работ.

Из диаграммы Ганта наглядно видны границы этапов научно-исследовательского проекта, длительность выполнения которого составила 4 месяца. Дата начала выполнения проекта – 31 января, дата окончания – 31 мая 2022 года, общее количество дней – 87, суббота и воскресенье – выходные. Также на диаграмме представлена информация об исполнителях каждого отдельного этапа работы.

№	Название задачи	Начало	Окончание	Длительность	2022
					фев мар апр май июн июл авг сен окт
1	Составление и утверждение темы	31.01.2022	02.02.2022	3д	Аксёнов С.В.
2	Разработка календарного плана	03.02.2022	04.02.2022	2д	Аксёнов С.В.; Сафронов В.С.
3	Подбор и изучение литературы	07.02.2022	18.02.2022	10д	Аксёнов С.В.; Сафронов В.С.
4	Изучение предметной области	21.02.2022	04.03.2022	10д	Сафронов В.С.
5	Изучение специфики данных	07.03.2022	11.03.2022	5д	Сафронов В.С.
6	Подготовка данных для анализа	14.03.2022	18.03.2022	5д	Сафронов В.С.
7	Обучение моделей классификации и оценка качества их работы	21.03.2022	25.03.2022	5д	Сафронов В.С.
8	Подбор способов визуализации	28.03.2022	29.03.2022	2д	Сафронов В.С.
9	Разработка веб-интерфейса	30.03.2022	15.04.2022	13д	Сафронов В.С.
10	Интеграция обученных моделей с интерфейсом	18.04.2022	19.04.2022	2д	Сафронов В.С.
11	Согласование выполненной работы с научным руководителем	20.04.2022	26.04.2022	5д	Аксёнов С.В.; Сафронов В.С.
12	Выполнение раздела «Финансовый менеджмент»	27.04.2022	10.05.2022	10д	Сафронов В.С.
13	Выполнение раздела «Социальная ответственность»	11.05.2022	24.05.2022	10д	Сафронов В.С.
14	Подведение итогов, оформление работы	25.05.2022	31.05.2022	5д	Сафронов В.С.

Рисунок 22 – Диаграмма Ганта

5.3.2 Бюджет научно-исследовательского проекта

В процессе планирования бюджета научно-исследовательского проекта необходимо обеспечить полное и достоверное отражение всех видов расходов, которые связаны с его выполнением. При формировании бюджета научно-исследовательского проекта используется следующая группировка затрат по статьям:

– материальные затраты научно-исследовательского проекта;
– затраты на электроэнергию;
– затраты на специальное оборудование для научных (экспериментальных) работ;
– основная заработная плата исполнителей;
– дополнительная заработная плата исполнителей;
– отчисления во внебюджетные фонды (страховые отчисления);
– накладные расходы.

5.3.2.1 Расчет материальных затрат

К материальным затратам относятся затраты на приобретаемые сырье и материалы, канцелярские принадлежности и другие материальные ценности, расходуемые непосредственно в процессе выполнения работ над объектом проектирования. Сюда же относятся специально приобретенное оборудование, инструменты и прочие объекты, относимые к основным средствам, стоимостью до 40 000 руб. включительно. Материальные затраты на проведение данного научно-исследовательского проекта представлены в таблице 17.

Таблица 17 – Материальные затраты

Наименование	Единица измерения	Количество	Цена за ед., руб.	Затраты, руб.
Ручка шариковая	шт.	2	35	70
Тетрадь в клетку, 24 листа	шт.	1	40	40
Бумага А4, 500 листов	пачка	1	320	320
Картридж черный PG-10	шт.	1	1599	1599
Ноутбук ASUS X750JB	шт.	1	38 990	38 990
Итого, руб.				41 019

Материальные расходы составили 41 019 рублей.

5.3.2.2 Расчет затрат на электроэнергию

Работа выполнялась в компьютерном классе КЦ ТПУ. Затраты на электроэнергию, потраченную на работу используемого оборудования, рассчитываются по формуле:

$$С_{эл.об} = Р_{об} \times t_{об} \times Ц_э, \quad (2)$$

где $Р_{об}$– мощность, потребляемая оборудованием, кВт; $t_{об}$ – время работы оборудования, час; $Ц_э$ – тариф на 1 кВт·ч (для ТПУ – 5,8 руб./кВт/ч).

Время работы оборудования вычисляется на основе длительности работы $T_{РД}$ по количеству рабочих дней из рисунка 2 расчета, что продолжительность рабочего дня равна 8 часов:

$$t_{об} = Т_{РД} \times К_t, \quad (3)$$

где $К_t \leq 1$ – коэффициент использования оборудования по времени, равный отношению времени его работы в процессе выполнения проекта к $Т_{РД}$.

Мощность, потребляемая оборудованием, определяется по формуле:

$$Р_{об} = Р_{ном} \times К_с, \quad (4)$$

где $Р_{ном}$ – номинальная мощность оборудования, кВт; $К_с \leq 1$ – коэффициент нагрузки, зависящий от средней степени использования номинальной мощности.

Расчет затрат на электроэнергию для технологических целей приведен в таблице 18.

Таблица 18– Затраты на электроэнергию технологическую

Наименование оборудования	Время работы оборудования $t_{об}$, час	Потребляемая мощность $Р_{об}$, кВт	$С_{эл.об}$, руб
Ноутбук	87·8·0,9	0,3	1089,94
Струйный принтер	1	0,1	0,66
Итого:			1090,60

5.3.2.3 Заработная плата исполнителей

Заработная плата рассчитывается из суммы основной и дополнительной заработной платы научного руководителя и исполнителя.

Формула для расчета основной заработной платы следующая:

$$З_{осн} = З_{дн} \cdot Т_р \cdot (1 + К_{пр} + К_д) \cdot К_р, \quad (7)$$

где $З_{дн}$ – среднедневная заработная плата, руб.; $К_{пр}$ – премиальный коэффициент (30% от $З_{дн}$); $К_д$ – коэффициент доплат и надбавок составляет 0,3; $К_р$ – районный коэффициент (для Томска 1,3); $Т_р$ – продолжительность работ, выполняемых работником, раб. дни.

Среднедневная заработная плата рассчитывается по формуле:

$$З_{дн} = \frac{З_м \cdot М}{F_д}, \quad (8)$$

где $З_м$ – оклад работника за месяц, руб.; $М$ – количество месяцев работы без отпуска в течение года: при отпуске в 24 рабочих дня $М$ = 11,2, 5-дневная неделя; $F_д$ – действительный годовой фонд рабочего времени научно-технического персонала, раб. дн.

Баланс рабочего времени представлен в таблице 19.

Таблица 19 – Баланс рабочего времени

Показатели рабочего времени	Количество дней
Календарные дни	365
Нерабочие дни (выходные, праздничные)	118
Потери рабочего времени (отпуск, невыходы по болезни)	24
Действительный годовой фонд рабочего времени	223

Для расчета заработной платы инженера возьмем оклад, равный 19 200,00 руб., для зарплаты руководителя – 37 700,00 руб. В таблице 20 представлен расчет основной заработной платы.

Таблица 20– Расчет основной заработной платы

Исполнители	$З_{дн}$, руб.	$К_{пр}$	$К_д$	$К_р$	$Т_р$	$З_{осн}$, руб.
Инженер	964,30	0	0	1,3	84	105 301,56
Руководитель	1893,45	0,3	0,3	1,3	20	78 767,52

В дополнительную зарплату входят суммы выплат, предусмотренные трудовым кодексом, например, оплата ежегодных и дополнительных отпусков, оплата времени, связанного с выполнением государственных и общественных обязанностей и т.д. Запланируем дополнительную заработную плату в размере 15% от основной зарплаты исполнителей. В таблице 21 представлен расчет затрат на заработную плату исполнителей.

Таблица 21– Расчет затрат на заработную плату исполнителей

Исполнители	$З_{осн}$, руб.	$З_{доп}$, руб.	$З_{зп}$, руб.
Инженер	105 301,56	15 795,23	121 096,80
Научный руководитель	78 767,52	11 815,13	90 582,65

5.3.2.4 Отчисления во внебюджетные фонды (страховые отчисления)

Размер страховых отчислений составляет 27,1% от суммы основной и дополнительной заработной платы исполнителя. В таблице 22 представлен расчет суммы страховых отчислений.

Таблица 22– Расчет суммы страховых отчислений

Исполнители	$З_{зп}$, руб.	Страховые отчисления
Инженер	121 096,80	32 817,23
Научный руководитель	90 582,65	24 547,90
Итого	211 679,45	57 365,13

5.3.2.5 Накладные расходы

К накладным расходам относятся расходы проекта, которые не учтены в предыдущих статьях. Величина накладных расходов рассчитывается как 70 % от суммы основной и дополнительной заработной платы, работников. $(121\,096{,}80 + 90\,582{,}65) \cdot 0{,}7 = 148\,175{,}62$

5.3.2.6 Формирование бюджета затрат научно-исследовательского проекта

В таблице 23 представлена информация о бюджете научно-исследовательского проекта.

Таблица 23 – Бюджет научно-исследовательского проекта

Наименование	Сумма, руб.	%
Материальные затраты	41 019,00	8,93
Затраты на электроэнергию	1 090,60	0,24
Затраты на основную заработную плату	184 069,08	40,07
Затраты на дополнительную заработную плату	27 610,36	6,01
Страховые взносы	57 365,13	12,49
Накладные расходы	148 175,62	32,26
Общий бюджет	459 329,79	100

Таким образом, большая часть расходов по проекту относится к категории затрат на основную заработную плату исполнителей, что составляет 40,07 %.

5.3.3 Риски научно-исследовательского проекта

При разработке научно-исследовательского проекта следует понимать и учитывать возможные риски. Риски в реализации проекта включают в себя возможные неопределенные события, которые могут возникнуть в проекте и вызвать последствия, которые повлекут за собой нежелательные эффекты. В таблице 24 представлены результаты оценки рисков проекта. Для каждого из них даны рекомендации по смягчению их воздействия.

Таблица 24– Реестр рисков

№	Риск	Потенциальное воздействие	Вероятность наступления (1-5)	Влияние риска (1-5)	Уровень риска	Способы смягчения риска	Условия наступления
1	Кадровый риск	Отсутствие заинтересованных исполнителей проекта	3	5	Существенный риск	Повышение мотивации исполнителей проекта	Потеря интереса исполнителей к деятельности проекта
2	Технический риск	Потеря файлов проекта	2	5	Существенный риск	Регулярное создание резервных копий файлов проекта	Отказ используемого оборудования

3	Доступ к данным	Отсутствие данных для работы системы	2	5	Существенный риск	Заключение официального договора на доступ к данным	Отсутствие доступа к данным истории

Из анализа реестра рисков можно заключить, что первым и вторым типами рисков обладает практически каждый проект. Риск же потери доступа к данным во время выполнения данного научно-исследовательского проекта существенен для реализации, однако маловероятен.

5.3.4 Описание потенциального эффекта

В результате проделанной в рамках раздела работы, можно сделать выводы о том, что на данный момент невозможно оценить экономический эффект разработки до её внедрения. Однако согласно оценке научно-исследовательского проекта по технологии «QuaD» разработка считается перспективной. Основные работы в рамках данного научно-исследовательского проекта проводились в период с 31 января по 31 мая 2022 года. Команда проекта состоит из инженера и научного руководителя. Общий бюджет научно-исследовательского проекта составил 459 329,79 руб. Основная часть затрат приходится на заработную плату исполнителей проекта. Все рассмотренные риски научно-исследовательского проекта являются существенными для его реализации, однако вероятность их наступления достаточно мала.

6 Социальная ответственность

Настоящая работа посвящена разработке алгоритма визуализации и анализа данных пациентов с инфекциями, передающимися клещами, а также интеграции данного алгоритма в программный веб-интерфейс. Данное программное обеспечение может найти свое применение в деятельности медицинских работников и исследователей в сфере медицины. Функциональные возможности разработанного приложения позволят сократить время, а также повысить качество анализа данных из медицинских документов, увеличить точность принятия решений о способе лечения пациентов на основе прогнозных значений, полученных обученными моделями машинного обучения.

Работа выполнялась с использованием персональной электронно-вычислительной машины (ПЭВМ) на базе Отделения информационных технологий Инженерной школы информационных технологий и робототехники Томского политехнического университета (ТПУ), данные для работы были предоставлены Отделением инфекционных заболеваний Сибирского государственного медицинского университета.

В разделе будут рассмотрены опасные и вредные факторы, оказывающие влияние на производственную деятельность инженера-программиста. Исследовано рабочее место программиста и помещение, в котором оно находится. Разработка осуществлялась в компьютерном классе Кибернетического центра ТПУ.

6.1 Правовые и организационные вопросы обеспечения безопасности

6.1.1 Специальные правовые нормы трудового законодательства

Основным документом, регулирующим отношения между работником и работодателем, является Трудовой Кодекс Российской Федерации (ТК РФ) [34].

Согласно трудовому кодексу РФ, нормальная продолжительность рабочего времени не может превышать 40 часов в неделю. Также работнику в течение рабочего дня (смены) должен предоставляться, перерыв на отдых и питание продолжительностью от 30 минут до 2 часов, который не включается в рабочее время. Если продолжительность ежедневной работы не превышает 4 часов, то указанный перерыв может не предоставляться. Всем работникам предоставляются выходные дни (еженедельный непрерывный отдых).

Федеральный закон «О специальной оценке условий труда» [35] регламентирует проведение специальной оценки в случае, если деятельность работников предприятия предусматривает непрерывную работу за компьютерными системами. По проведенной оценке, устанавливаются гарантии и компенсации работникам согласно ТК РФ.

6.1.2 Эргономические требования к рабочему месту оператора ПЭВМ

Главными элементами рабочего места программиста являются стол, кресло, дисплей, клавиатура и мышь. Основным рабочим положением является положение сидя. Нормативными положениями ГОСТ 12.2.032-78 ССБТ, ГОСТ 21889-76, ГОСТ 22269-76 предъявляются определенные требования к оснащению рабочего места, предусматривающего длительную работу за персональным компьютером (Таблица 25).

Таблица 25 – Нормы оборудования рабочих мест с ПЭВМ

Параметр		Значение
Высота перегородок, разделяющих рабочие места		Не менее 1,5 метров
Стол	Ширина рабочей поверхности	От 80 до 140 см
	Глубина рабочей поверхности	От 80 до 100 см
	Высота рабочей поверхности	72,5 см
Расстояние от глаз до монитора		От 60 до 70 см
Расстояние клавиатуры от края стола		От 10 до 30 см
Стул	Ширина поверхности	От 40 см
	Глубина поверхности	От 40 см
	Регулировка высоты поверхности	От 40 до 50 см
	Угол наклона вперед	До 15 градусов
	Угол наклона назад	До 5 градусов
	Высота опорной поверхности спинки	30 плюс/минус 2 см
	Ширина опорной поверхности спинки	От 38 см
	Радиус кривизны горизонтальной плоскости спинки	40 см
	Угол наклона спинки в вертикальной плоскости	Плюс/минус 30 градусов
	Регулировка расстояния спинки от переднего края сидения	От 26 до 40 см
	Длина подлокотников	От 25 см
	Ширина подлокотников	От 5 до 7 см
	Регулировка подлокотников по высоте над сиденьем	23 плюс/минус 3 см
	Внутреннее расстояние между подлокотниками	От 35 до 50 см
Подставка для ног	Ширина	От 30 см
	Глубина	От 40 см
	Регулировка по высоте	До 15 см
	Угол наклона опорной поверхности	До 20 градусов
	Высота бортика по переднему краю	1 см

Работающий на ПЭВМ должен сидеть прямо, опираясь в области нижнего края лопаток на спинку кресла, не сутулясь, с небольшим наклоном головы вперед (до 5 – 7 градусов). Предплечья должны опираться на поверхность стола, снимая статическое напряжение плечевого пояса и рук.

Средства отображения информации следует располагать в вертикальной плоскости под углом ±15◦ от нормальной линии взгляда и в горизонтальной плоскости под углом ±15◦ от сагитальной плоскости (рисунок 23).

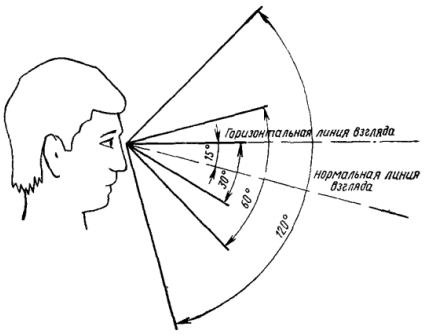


Рисунок 23 — Зона зрительного наблюдения в вертикальной плоскости

Взаимное расположение элементов рабочего места должно обеспечивать возможность осуществления всех необходимых движений и перемещений для эксплуатации и технического обслуживания оборудования [36].

Приборы на столе должны размещаться так, чтобы руки не скрещивались. Аварийные органы управления (кнопка выключения) должны располагаются в зоне досягаемости моторного поля.

Рабочие места в компьютерном классе Кибернетического центра (КЦ) ТПУ отвечают данным требованиям.

6.2 Производственная безопасность

6.2.1 Анализ вредных и опасных факторов, которые может создать объект исследования

Опасные и вредные производственные факторы, выявленные в рамках настоящей работы, представлены в таблице 26.

Таблица 26 – Опасные и вредные производственные факторы

Факторы (ГОСТ 12.0.003-2015)	Нормативные документы
Производственные факторы, связанные с аномальными микроклиматическими параметрами воздушной среды на местонахождении работающего	СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания

Производственные факторы, связанные с отсутствием или недостатком необходимого естественного освещения	СП 52.13330.2016 Естественное и искусственное освещение
Производственные факторы, обладающие свойствами психофизиологического воздействия на организм человека	МР 2.2.9.2311 – 07 «Профилактика стрессового состояния работников при различных видах профессиональной деятельности»
Производственные факторы, связанные с электрическим током, вызываемым разницей электрических потенциалов, под действие которого попадает работающий;	ГОСТ 12.1.030-81 ССБТ Защитное заземление, зануление. ГОСТ Р 12.1.019-2009 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты.
Производственные факторы, связанные с электромагнитными полями, неионизирующими ткани тела человека	ГОСТ Р 50948-2001 Средства отображения информации индивидуального пользования. Общие эргономические требования и требования безопасности

6.2.2 Производственные факторы, связанные с аномальными микроклиматическими параметрами воздушной среды на местонахождении работающего

Для обеспечения установленных норм микроклиматических параметров и чистоты воздуха на рабочих местах и в помещениях применяют вентиляцию. Общеобменная вентиляция используется для обеспечения в помещениях соответствующего микроклимата. Периодически должен вестись контроль за влажностью воздуха.

В летнее время при высокой уличной температуре должны использоваться системы кондиционирования. В холодное время года предусматривается система отопления. Для отопления помещений используются водяные системы центрального отопления.

Для категории тяжести работ 1а (работы с интенсивностью энергозатрат до 120 ккал/ч (до 139 Вт), производимые сидя и сопровождающиеся незначительным физическим напряжением) оптимальные и допустимые показатели температуры, относительной влажности и скорости движения воздуха в рабочей зоне производственных помещений должны соответствовать значениям, представленным в таблице 27 [37].

Таблица 27 – Оптимальные нормы микроклимата

Период года	Температура, °С	Температура поверхностей, °С	Относительная влажность, %	Скорость движения воздуха, м/с
Холодный	22-24	21-25	40-60	0,1
Теплый	23-25	22-26	40-60	0,1

Допустимые микроклиматические условия установлены по критериям допустимого теплового и функционального состояния человека на период 8-часовой рабочей смены. Они устанавливаются в случаях, когда по технологическим требованиям, технически и экономически обоснованным причинам не могут быть обеспечены оптимальные величины. В таблице 28 представлены допустимые величины показателей микроклимата на рабочих местах.

Таблица 28 – Допустимые величины показателей микроклимата

Период года	Температура воздуха, °С		Температура поверхностей, °С	Относительная влажность воздуха, %	Скорость движения воздуха, не более, м/с	
	Диапазон ниже оптимальных величин	Диапазон выше оптимальных величин			при температуре воздуха ниже оптимальной	при температуре воздуха выше оптимальной
Холодный	20,0 – 21,9	24,1 – 25,0	19,0- 26,0	15 – 75	0,1	0,1
Теплый	21,0 – 22,9	25,1 – 28,0	20,0 – 29,0	15 - 75	0,1	0,2

При обеспечении допустимых величин микроклимата на рабочих местах перепад температуры воздуха по высоте должен быть не более 3°С, перепад температуры воздуха по горизонтали, а также ее изменения в течение смены не должны превышать – 4°С. При этом абсолютные значения температуры воздуха не должны выходить за пределы оптимальных величин.

При температуре воздуха на рабочих местах 25°С и выше максимально допустимые значения относительной влажности воздуха не должны выходить за следующие пределы:

– 70% – при температуре воздуха 25°С,

– 65% – при температуре воздуха 26°С,

– 60% – при температуре воздуха 27°С,

– 55% – при температуре воздуха 28°С.

Для обеспечения установленных норм микроклиматических параметров и чистоты воздуха на рабочих местах и в помещениях применяется вентиляция, периодически ведется контроль влажности воздуха. В летнее время при высокой уличной температуре используется система кондиционирования. В холодное время года предусматривается система отопления. Для отопления помещений используются водяные системы центрального отопления.

Помещение, где выполнялась работа, было обследовано на соответствие данным требованиям. Результаты обследования приведены в таблице 29.

Таблица 29 – Результаты измерений параметров микроклимата

Рабочее место	Период года	Температура, °С	Относительная влажность воздуха, %	Скорость движения воздуха, не более, м/с
Аудитория КЦ ТПУ	Холодный (январь)	22,4	60	0
	Тёплый (июль)	23,0	59	0

Результаты измерений соответствуют допустимым значениям нормативов, следовательно, микроклимат помещений удовлетворяет требованиям санитарных норм и правил.

6.2.3 Производственные факторы, связанные с отсутствием или недостатком необходимого естественного освещения

В помещении при работе с ПК должно быть естественное и искусственное освещение. Естественное освещение обеспечивается через оконные проемы с коэффициентом естественного освещения не ниже 1,2% в зонах с устойчивым снежным покровом и не ниже 1,5% на остальной территории. Световой поток из оконного проема должен падать на рабочее место оператора с левой стороны.

Для определения приемлемого уровня освещенности в помещении необходимо следующее: определить требуемый уровень освещенности внешними источниками света; если требуемый уровень освещенности не приемлем, необходимо найти способ сохранения требуемого контраста изображения другими средствами.

Рекомендуемые соотношения яркостей в поле зрения между рабочими поверхностями не должно превышать 1:3 – 1:5, между рабочими поверхностями и поверхностями стен и оборудования – 1:10.

Освещенность на поверхности стола в зоне размещения документа должна быть 300 – 500 лк [38]. Допускается установка светильников местного освещения для подсветки документов. Местное освещение не должно создавать бликов на поверхности экрана и увеличивать освещенность экрана более 300 лк. Прямую блеклость от источников освещения следует ограничить. Яркость светящихся поверхностей (окна, светильники), находящихся в поле зрения, должна быть не более 200 кд/м2.

В таблице 30 представлены нормы на освещение для операторов поста управления [39].

Таблица 30 – Нормы на освещение для оператора

Характер зрительной работы	Разряд зрительной работы	Подразряд зрительной работы	Искусственное освещение		Естественное освещение, % при боковом направлении
			Освещенность при системе общего освещения, лк	Коэффициент пульсации, %	
Различие объектов высокой точности	Б	1	300	15	1

Ниже приведем расчет системы искусственного освещения для компьютерного класса КЦ ТПУ методом светового потока, предназначенного для расчета освещенности общего равномерного освещения горизонтальных поверхностей. Помещение, для которого будут производиться расчеты, прямоугольное со следующими параметрами: аудитория КЦ ТПУ имеет длину 5 м, ширину – 7 м, высоту – 4 м, количество ламп – 12 шт. Уровень

рабочей поверхности над полом должен составлять 0,8 м, установленная минимальная норма освещенности – 300 лк.

Формула для определения светового потока лампы накаливания или группы люминесцентных ламп светильника следующая [40]:

$$Ф = \frac{E_н \cdot S \cdot K_з \cdot Z \cdot 100}{n \cdot \eta}, \quad (1)$$

где $E_н$ – нормируемая минимальная освещенность [39]; S – площадь освещаемого помещения; $K_з$ – коэффициент запаса, учитывающий загрязнение светильника, пыли (для помещений с малым выделением пыли равен 1,5); Z – коэффициент неравномерного освещения (для люминесцентных ламп равен 1,1); n – количество светильников; η – коэффициент использования светового потока, %.

Коэффициент использования светового потока показывает, какая часть светового потока ламп попадает на рабочую поверхность. Он зависит от индекса помещения i, типа светильника, высоты светильников над рабочей поверхностью h и коэффициентов отражения стен, потолка и рабочей поверхности.

Формула для определения индекса помещения следующая [40]:

$$i = \frac{S}{h \cdot (A+B)}, \quad (2)$$

где h – расчетная высота подвеса светильников над рабочей поверхностью; A – длина помещения; B – ширина помещения.

Произведем расчеты высоты подвеса светильников над рабочей поверхностью для аудитории:

$$h_a = H - 0{,}8 = 4 - 0{,}8 = 3{,}2 \text{ (м)}$$

Рассчитаем индексы помещений:

$$i_a = \frac{5 \cdot 7}{3{,}2 \cdot (5+7)} = \frac{35}{3{,}2 \cdot 12} = 0{,}91$$

Найдем коэффициенты отражения поверхностей стен, потолка и рабочей поверхности [40].

Так как поверхности стен аудитории окрашены в светлый цвет, то коэффициенты отражения стен равны 50%, поверхности потолков аудитории также светлого цвета, коэффициенты отражения потолков равны 50%, рабочие поверхности средней светлости, коэффициенты отражения рабочих поверхностей равны 30%. Учитывая коэффициенты отражения поверхностей стен, потолка, рабочей поверхности и индекс помещения i, определяем значения коэффициентов $\eta_{\text{а}} = 36\%$, $\eta_{\text{к}} = 42\%$.

Подставив значения в формулу (1), рассчитываем световой поток одного источника света:

$$\Phi_a = \frac{300 \cdot 35 \cdot 1{,}5 \cdot 1{,}1}{12 \cdot 0{,}36} = 4010 \text{ (лм)}$$

По полученному световому потоку подбираем лампы, наиболее подходящими являются светодиодная лампа Smartbuy LED HP50W (4000 лм) и светодиодная лампа КОСМОС LED 30Вт E27 (2650 лм).

Выразим E:

$$E = \frac{(F \cdot N \cdot \eta)}{(k)}, \quad (3)$$

Рассчитаем норму освещенности:

$$\text{E}_{\text{а}} = \frac{(4000 \cdot 12 \cdot 0{,}36)}{(1{,}5 \cdot 35 \cdot 1{,}1)} = 299{,}2 \text{ (лк)}$$

Как видно из расчета, минимальная освещенность в пределах нормы [41].

Для того чтобы доказать, что использование люминесцентных ламп Smartbuy и КОСМОС является наиболее рациональным, рассчитаем необходимое количество светильников по формуле:

$$N = \frac{(E \cdot k \cdot S \cdot Z)}{(n \cdot \eta \cdot F)}, \quad (4)$$

где E – норма освещенности, 300 лк; k – коэффициент запаса, учитывающий старение ламп и загрязнение светильников, 1,5; S – площадь помещения; Z – коэффициент неравномерности освещения, 1,1; n – число рядов светильников, 2; η – коэффициент использования светового потока; F – световой поток, излучаемый светильником.

Рассчитаем количество ламп:

$$N_a = \frac{300 \cdot 1{,}5 \cdot 35 \cdot 1{,}1}{0{,}36 \cdot 4000} = 12 \text{ (шт.)}$$

В аудитории находятся три светильника, в каждом из которых по 4 лампы. Расчетное количество ламп соответствует их реальному количеству при норме освещенности 300 лк, что говорит о соблюдении норм по искусственному освещению [39].

6.2.4 Производственные факторы, обладающие свойствами психофизиологического воздействия на организм человека

Работа с ПЭВМ вызывает зрительную и умственную нагрузку на организм человека.

При умственной нагрузке необходима длительность сосредоточенного внимания, выраженная ответственность, плотность сигналов и сообщений в единицу времени по МР 2.2.9.2311 – 07 «Профилактика стрессового состояния работников при различных видах профессиональной деятельности» [42]. Оказывает угнетающее влияние на психическую деятельность ухудшаются функции внимания (объем, концентрация, переключение), памяти (кратковременной и долговременной), восприятия (появляется большое число ошибок).

При зрительной нагрузке необходима высокая координация сенсорных и моторных элементов зрительной системы. Вызывает головную боль, ухудшение зрения, астенопию – патологического состояния, связанного с быстрым переутомлением глаз.

Для устранения накопленной усталости и нагрузки на организм человека необходимо выполнять комплекс физических упражнений на координацию движений, концентрацию внимания, комплекс упражнений на глаз, использовать методику психической саморегуляции.

6.2.5 Производственные факторы, связанные с электрическим током, вызываемым разницей электрических потенциалов, под действие которого попадает работающий

Степень опасного воздействий на человека электрического тока зависит от:

- рода и величины напряжения и тока;
- частоты электрического тока;
- пути прохождения тока через тело человека;
- продолжительности воздействия на организм человека;
- условий внешней среды.

Согласно ПУЭ аудиторию КЦ–105 НИ ТПУ по степени опасности поражения электрическим током можно отнести к классу помещений без повышенной опасности.

Основными мероприятиями по защите от электропоражения являются:

- обеспечение недоступности токоведущих частей путем использования изоляции в корпусах оборудования;
- применение средств коллективной защиты от поражения электрическим током;
- защитного заземления, зануления [43];
- защитного отключения;
- использование устройств бесперебойного питания.

Технические способы и средства применяют раздельно или в сочетании друг с другом так, чтобы обеспечивалась оптимальная защита.

Электробезопасность должна обеспечиваться [44]:

- конструкцией электроустановок;
- техническими способами и средствами защиты;
- организационными и техническими мероприятиями.

6.2.6 Производственные факторы, связанные с электромагнитными полями, неионизирующими ткани тела человека

Электромагнитные поля, характеризующиеся напряженностями электрических и магнитных полей, наиболее вредны для организма человека. Основным источником этих проблем, связанных с охраной здоровья людей, использующих в своей работе автоматизированные информационные системы на основе персональных компьютеров, являются дисплеи (мониторы), они представляют собой источники наиболее вредных излучений, неблагоприятно влияющих на здоровье человека. Предельно допустимые значения излучений от ЭВМ [45] приведены в таблице 31.

Таблица 31 – Допустимые уровни электромагнитного поля

Наименование параметров		ВДУ ЭМП
Напряженность электрического поля	В диапазоне частот 5 Гц – 2 кГц	25 В/м
	В диапазоне частот 2 кГц – 400 кГц	2,5 В/м
Плотность магнитного потока	В диапазоне частот 5 Гц – 2 кГц	250 нТл
	В диапазоне частот 2 кГц – 400 кГц	25 нТл
Электростатический потенциал экрана видеомонитора		500 В

Для оценки соблюдения уровней электромагнитного поля необходим производственный контроль. В случае их превышения необходимо проводить организационно-технические мероприятия (защита временем, расстоянием, экранирование источника, замена оборудования, использование средств индивидуальной защиты).

6.3 Экологическая безопасность

Воздействие на селитебную зону: отсутствует.

Воздействие на гидросферу: отсутствует.

Воздействие на атмосферу: отсутствует.

Воздействие на литосферу. При проектировании и разработке алгоритма необходима ПЭВМ. В случае неисправной поломки или

нестабильности ПЭВМ утилизируется. Составляющие его электронные компоненты оказывают серьезное воздействие на литосферу при их утилизации.

Федеральный закон №89 от 1998 г. «Об отходах производства и потребления» запрещает юридическим лицам самовольно избавляться от опасных отходов [46]. Этим видом деятельности, согласно постановлению Правительства РФ №340 от 2002 г., могут заниматься только специализированные структуры. В их число входят и фирмы, которые занимаются утилизацией электронных отходов. Обращение с отходами регламентируется ГОСТ Р 53692-2009 «Ресурсосбережение. Обращение с отходами» [47].

Вышедшая из строя ПЭВМ и сопутствующая оргтехника относится к IV классу опасности (малоопасные отходы) и подлежит специальной утилизации (код отхода 4 81 206 11 52 4, класс опасности – 4) [48]. Для оказания наименьшего влияния на окружающую среду, необходимо проводить специальную процедуру утилизации ПЭВМ и оргтехники, при которой более 90% отправится на вторичную переработку и менее 10% будут отправлены на свалки. При этом она должна соответствовать процедуре утилизации, как это указано в этапах технологического цикла отходов [47].

Также негативное воздействие на литосферу оказывают люминесцентные лампы при их утилизации. Их эксплуатация требует осторожности и четкого выполнения инструкции по обращению с данным отходом (код отхода 4 71 101 01 52 1, класс опасности – 1 [48]). В данной лампе содержится опасное вещество ртуть в газообразном состоянии (предельно допустимая концентрация 0,0003 мг/м^3) [37]. При неправильной утилизации, лампа может разбиться и пары ртути могут попасть в окружающую среду. Вдыхание паров ртути может привести к тяжелому повреждению здоровья.

Для утилизации люминесцентных ламп действует Постановление Правительства РФ от 28.12.2020 N 2314 «Об утверждении Правил обращения

с отходами производства и потребления в части осветительных устройств, электрических ламп, ненадлежащие сбор, накопление, использование, обезвреживание, транспортирование и размещение которых может повлечь причинение вреда жизни, здоровью граждан, вреда животным, растениям и окружающей среде» [49]. Согласно постановлению, устанавливается порядок обращения с отходами производства и потребления в части осветительных устройств, электрических ламп, ненадлежащие сбор, накопление, использование, обезвреживание, транспортирование и размещение которых может повлечь причинение вреда жизни, здоровью граждан, вреда животным, растениям и окружающей среде.

6.4 Безопасность в чрезвычайных ситуациях

Наиболее характерной ЧС для помещений, оборудованных ПЭВМ, является пожар. Пожары на производстве возникают по определенным причинам, устранение которых составляет основу всех мероприятий по пожарной безопасности. Основные причины возникновения пожара:

• нарушение правил эксплуатации электрического оборудования, эксплуатация его в неисправном состоянии;

• применение неисправных осветительных приборов, электропроводки и устройств, дающих искрение, замыкание и т. п.;

• перегрузка электрических сетей.

Аудитория КЦ ТПУ, относится к категории В3 по пожароопасности, содержит вещества и материалы, способные гореть при взаимодействии с водой, кислородом воздуха или друг с другом [50].

Для минимизации возможности возникновения пожара необходимо проводить пожарную профилактику. Под пожарной профилактикой понимают комплекс организационных и технических мероприятий, направленных на обеспечение безопасности людей, на предотвращении пожара, ограничение его распространения, а также создание условий для

успешного тушения пожара. Для профилактики пожара чрезвычайно важна правильная оценка пожароопасности, определение опасных факторов и обоснование способов и средств предупреждения пожара и защиты от него.

Одно из условий обеспечения пожаробезопасности – ликвидация возможных источников воспламенения. Обогревание помещения открытыми электронагревательными приборами могут привести к пожару, т.к. в помещении находятся бумажные документы и справочная литература. Следовательно, использование открытого нагревательного прибора неприемлемо.

Пожарная безопасность объекта должна обеспечиваться системами предотвращения пожара и противопожарной защиты, в том числе организационно-техническими мероприятиями.

Пожарная защита должна обеспечиваться применением средств пожаротушения, а также применением автоматических установок пожарной сигнализации.

Должны быть приняты следующие меры противопожарной безопасности:

- обеспечение эффективного удаления дыма, т.к. в помещениях, имеющих оргтехнику, содержится большое количество пластиковых веществ, выделяющих при горении летучие ядовитые вещества и едкий дым;

- обеспечение правильных путей эвакуации;

- наличие огнетушителей и пожарной сигнализации;

- соблюдение всех противопожарных требований к системам отопления и кондиционирования воздуха.

Для тушения пожаров на участке производства необходимо применять углекислотные (ОУ-5 или ОУ-10) и порошковые огнетушители (например, типа ОП-10), которые обладают высокой скоростью тушения, большим временем действия, возможностью тушения электроустановок, высокой эффективностью борьбы с огнем.

Помещение (КЦ ТПУ) оборудовано пожарными извещателями, которые позволяют оповестить дежурный персонал о пожаре. В качестве пожарных извещателей в помещении устанавливаются дымовые фотоэлектрические извещатели типа ИДФ-1 или ДИП-1.

Выведение людей из зоны пожара должно производиться по плану эвакуации.

План эвакуации представляет собой заранее разработанный план (схему), в которой указаны пути эвакуации, эвакуационные и аварийные выходы, установлены правила поведения людей, порядок и последовательность действий в условиях чрезвычайной ситуации [41].

Согласно Правилам противопожарного режима в Российской Федерации (с изменениями на 21 мая 2021 года) (п. 5) в зданиях и сооружениях (кроме жилых домов) при единовременном нахождении на этаже 50 и более человек руководитель организации (объекта) организует разработку планов эвакуации людей при пожаре, которые размещаются на видных местах.

Ответственность за нарушение Правил пожарной безопасности, согласно действующему федеральному законодательству, несет руководитель объекта.

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 [42].

Категория помещения по электробезопасности, согласно ПУЭ, соответствует первому классу – «помещения без повышенной опасности» [51].

Согласно правилам по охране труда при эксплуатации электроустановок 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 [52].

Категория тяжести труда в аудитории КЦ ТПУ по СанПиН 1.2.3685-21 "Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания" относится к категории Iа (работы, производимые сидя и сопровождающиеся незначительным физическим напряжением с интенсивностью энергозатрат до 120 ккал/час (до 139 Вт)) [37].

Аудитория КЦ ТПУ, относится к категории В3 по пожароопасности. Характеристика веществ и материалов, находящихся в помещении: горючие и трудногорючие жидкости, твердые горючие и трудногорючие вещества и материалы (в том числе пыли и волокна), вещества и материалы, способные при взаимодействии с водой, кислородом воздуха или друг с другом только гореть. [53].

Рассмотренный объект, оказывающий незначительное негативное воздействие на окружающую среду, относится к объектам III категории [54].

Заключение

На протяжении выполнения данной работы поставленные задачи были выполнены. А именно, была проведена подготовка табличных данных пациентов с инфекциями, передаваемыми клещами, к анализу, а также построены классификаторы диагнозов пациентов с клещевыми инфекциями, разработан интерактивный веб-интерфейс.

Предварительно загруженные данные были проверены на наличие пропущенных значений и выбросов. Признаки и записи, содержащие пропущенные значения, которые невозможно было восстановить исключены из дальнейшего анализа. Пропущенные значения в бинарных категориальных признаках были заполнены нулями, что соответствует отсутствию у пациента того или иного признака, пропуски в числовых признаках анализов крови рассчитаны на основе лейкоцитарной формулы. Текстовые категориальные признаки были закодированы методом прямого кодирования.

Набор данных был разделен на тренировочное и тестовое подмножества. Было построено несколько моделей для определения диагноза пациентов с клещевыми инфекциями. Для выбранных моделей был проведен подбор оптимальных параметров. После чего проведена оценка качества работы моделей классификации с помощью метрик качества и ROC-кривых. Чувствительность дерева решений, логистической регрессии, случайного леса и градиентного бустинга составила 0,67, 0,75, 0,81 и 0,77, соответственно. Значения же специфичности для данных алгоритмов, соответственно, 0,7, 0,77, 0,79 и 0,78. Таким образом, наилучший результат показала модель случайного леса. Что касается значений площадей под ROC-кривыми, то наилучшие результаты по трем классам, а именно, клещевому энцефалиту, иксодовому клещевому боррелиозу и микст-инфекции, показали градиентный бустинг и дерево решений. При классификации диагнозов пациентов с помощью логистической регрессии и случайного леса, микст-инфекция определяется слабее, чем два других диагноза.

На основе проведенного анализа данных был реализован интерактивный веб-интерфейс, позволяющий изучить особенности структуры набора данных и результаты анализа.

Также разработаны следующие разделы: «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение», «Социальная ответственность», а также раздел на иностранном языке (английский) – «Data Preprocessing Methods», приведенный в Приложении А.

Список использованных источников и литературы

1. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP: учебное пособие / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. – Санкт-Петербург: БХВ-Петербург, 2007. – 384 с.
2. Topflight // How to develop a dashboard web app for your website [Электронный ресурс]. – URL: https://topflightapps.com/ideas/how-to-create-a-dashboard-web-application/#1 (дата обращения: 26.02.2022).
3. Анализ данных // Аналитическая культура [Электронный ресурс]. – URL: https://www.mann-ivanov-ferber.ru/assets/files/bookparts-new/analiticheskaya-kultura/analiticheskaya-kultura-mail_stamped.pdf (дата обращения: 10.02.2022).
4. Трансмиссивные болезни // Всемирная организация здравоохранения [Электронный ресурс]. – URL: https://www.who.int/ru/news-room/fact-sheets/detail/vector-borne-diseases (дата обращения: 6.01.2022).
5. Миноранская Н.С., Сарап П.В., Андронова Н.В., Миноранская Е.И. Клинико-лабораторные предикторы прогноза исходов иксодовых клещевых боррелиозов // Вестник Российской академии медицинских наук. – 2015. Т. 70. № 3. С. 378-385.
6. Лодыгина У.В., Веселова А.Н., Лысанова А.И., Воробьева Ю.С. Клещевой энцефалит: факторы, определяющие исход // Бюллетень Северного государственного медицинского университета. – 2018. № 1 (40). С. 247-249.
7. Yang L.H., Han B.A. Data-driven predictions and novel hypotheses about zoonotic tick vectors from the genus Ixodes // BMC Ecology – 2018. – Vol. 18. № 7. DOI: 10.1186/s12898-018-0163-2. PMID: 29448923.
8. Pfeifer L.M., Valdenegro-Toro M. Automatic Detection and Classification of Tick-borne Skin Lesions using Deep Learning // arxiv.org. 2020. Дата обновления: 23.11.2020. – URL: https://arxiv.org/abs/2011.11459 (дата обращения 07.02.2022).

9. Johnson L., Shapiro M., Stricker R.B., Vendrow J., Haddock J., Needell D. Antibiotic Treatment Response in Chronic Lyme Disease: Why Do Some Patients Improve While Others Do Not? // Healthcare (Basel). – 2020 Oct 3;8(4):383. DOI: 10.3390/healthcare8040383. PMID: 33022914

10. Vendrow J., Haddock J., Needell D., Johnson L. Feature Selection on Lyme Disease Patient Survey Data [Электронный ресурс] // arxiv.org. 2020. Дата обновления: 24.08.2020. – URL: https://arxiv.org/abs/2009.09087 (дата обращения 14.02.2022).

11. Clarke D.J.B., Rebman A.W., Bailey A., Wojciechowicz M.L., Jenkins S.L., Evangelista J.E., Danieletto M., Fan J, Eshoo M.W., Mosel M.R., Robinson W., Ramadoss N., Bobe J., Soloski M.J., Aucott J.N., Ma'ayan A. Predicting Lyme Disease From Patients' Peripheral Blood Mononuclear Cells Profiled With RNA-Sequencing // Frontiers in immunology. – 2021 Mar 8;12:636289. DOI: 10.3389/fimmu.2021.636289. PMID: 33763080

12. Выброс // Викиконспекты [Электронный ресурс]. – URL: https://neerc.ifmo.ru/wiki/index.php?title=%D0%92%D1%8B%D0%B1%D1%80%D0%BE%D1%81#.D0.9C.D0.B5.D1.82.D0.BE.D0.B4.D1.8B_.D0.BE.D0.B1.D0.BD.D0.B0.D1.80.D1.83.D0.B6.D0.B5.D0.BD.D0.B8.D1.8F_.D0.B2.D1.8B.D0.B1.D1.80.D0.BE.D1.81.D0.BE.D0.B2 (дата обращения: 15.02.2022).

13. Как вычислить выбросы // wikiHow [Электронный ресурс]. – URL: https://ru.wikihow.com/%D0%B2%D1%8B%D1%87%D0%B8%D1%81%D0%BB%D0%B8%D1%82%D1%8C-%D0%B2%D1%8B%D0%B1%D1%80%D0%BE%D1%81%D1%8B (дата обращения: 20.05.2022).

14. Обработка пропусков в данных // Loginom [Электронный ресурс]. – URL: https://loginom.ru/blog/missing (дата обращения: 10.04.2022).

15. Типы данных в статистике // Машинное обучение [Электронный ресурс]. – URL: https://www.machinelearningmastery.ru/data-types-in-statistics-347e152e8bee/ (дата обращения: 15.05.2022).

16. Рашка С. Python и машинное обучение // пер. с англ. А. В. Логунова. - М.: ДМК Пресс, 2017. - 418 с.

17. Обучение с учителем // Машинное обучение [Электронный ресурс]. – URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9E%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%D1%81_%D1%83%D1%87%D0%B8%D1%82%D0%B5%D0%BB%D0%B5%D0%BC (дата обращения: 11.05.2022).

18. Кластеризация // Skill factory [Электронный ресурс]. – URL: https://blog.skillfactory.ru/glossary/klasterizacziya-klasternyj-analiz/ (дата обращения: 21.05.2022).

19. Обучение нейросети с учителем, без учителя, с подкреплением — в чем отличие? Какой алгоритм лучше? // Neurohive [Электронный ресурс]. – URL: https://neurohive.io/ru/osnovy-data-science/obuchenie-s-uchitelem-bez-uchitelja-s-podkrepleniem/ (дата обращения: 13.04.2022).

20. Машинное обучение применили для помощи анастезиологам // Neurohive [Электронный ресурс]. – URL: https://neurohive.io/ru/gotovye-prilozhenija/mashinnoe-obuchenie-primenili-dlya-pomoshhi-anasteziologam/ (дата обращения: 14.02.2022).

21. Модель обучили находить оптимальную схему лечения // Neurohive [Электронный ресурс]. – URL: https://neurohive.io/ru/papers/model-obuchili-nahodit-optimalnuju-shemu-lecheniya/ (дата обращения: 18.04.2022).

22. Машинное обучение: методы и способы // OSP – Гид по технологиям цифровой трансформации [Электронный ресурс]. – URL: https://www.osp.ru/cio/2018/05/13054535 (дата обращения: 12.04.2022).

23. Дерево решений // Loginom [Электронный ресурс]. – URL: https://wiki.loginom.ru/articles/decision-trees.html (дата обращения: 1.03.2022).

24. Логистическая регрессия (Logistic Regression) // Loginom [Электронный ресурс]. – URL: https://wiki.loginom.ru/articles/logistic-regression.html (дата обращения: 3.03.2022).

25. Как работает случайный лес? // Nuancesprog [Электронный ресурс]. – URL: https://nuancesprog.ru/p/6160/ (дата обращения: 8.03.2022).

26. Оценка качества в задачах классификации // Университет ИТМО [Электронный ресурс]. – URL: https://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D1%86%D0%B5%D0%BD%D0%BA%D0%B0_%D0%BA%D0%B0%D1%87%D0%B5%D1%81%D1%82%D0%B2%D0%B0_%D0%B2_%D0%B7%D0%B0%D0%B4%D0%B0%D1%87%D0%B0%D1%85_%D0%BA%D0%BB%D0%B0%D1%81%D1%81%D0%B8%D1%84%D0%B8%D0%BA%D0%B0%D1%86%D0%B8%D0%B8 (дата обращения: 8.04.2022).

27. Интерпретируй это: метод SHAP в Data Science // Чернобровов Алексей [Электронный ресурс]. – URL: https://chernobrovov.ru/articles/interpretiruj-eto-metod-shap-v-data-science.html (дата обращения: 16.04.2022).

28. Основные инструменты анализа данных. Откройте для себя список из 14 лучших программ и инструментов анализа // Xmldatafeed [Электронный ресурс]. – URL: https://xmldatafeed.com/osnovnye-instrumenty-analiza-dannyh-otkrojte-dlya-sebya-spisok-iz-14-luchshih-programm-i-instrumentov-analiza/ (дата обращения: 25.04.2022).

29. Choosing a Better Framework // Tutorials point [Электронный ресурс]. – URL: https://www.tutorialspoint.com/python_web_development_libraries/python_web_development_libraries_choosing_a_better_framework.htm (дата обращения: 22.04.2022).

30. Долгов В. В., Меньшиков В. В. Клиническая лабораторная диагностика. Национальное руководство //М.: ГЭОТАР-Медиа. – 2016. – С. 688.

31. Криницына З. В., Видяев И. Г. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение: учебно-методическое пособие / З. В. Криницына, И. Г. Видяев; Томский политехнический университет. − Томск: Изд-во Томского политехнического университета, 2014. – 73 с.

32. Диаграмма Исикавы [Электронный ресурс]: сайт. – URL: https://up-pro.ru/encyclopedia/diagramma-isikavy/ (дата обращения: 22.05.2022).

33. История диаграммы Ганта [Электронный ресурс] / Юлия Челянова и Евгений Пикулев. — Электрон. текстовые дан. — Режим доступа: http://gibtech.ru/blog/discus?entry_id=177 (дата обращения: 22.05.2022).

34. «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).

35. Федеральный закон «О специальной оценке условий труда» от 28.12.2013 N 426-ФЗ.

36. ГОСТ 12.2.032-78 «Рабочее место при выполнении работ сидя».

37. СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания.

38. ГОСТ Р 50923-96 Дисплеи. Рабочее место оператора.

39. СП 52.13330.2016 Естественное и искусственное освещение.

40. Безопасность жизнедеятельности. Расчёт искусственного освещения. Методические указания к выполнению индивидуальных заданий для студентов дневного и заочного обучения всех направлений и специальностей ТПУ. – Томск: Изд. ТПУ, 2008. – 20 с.

41. ГОСТ Р 12.2.143-2009 Система стандартов безопасности труда. Системы фотолюминесцентные эвакуационные. Требования и методы контроля.

42. МР 2.2.9.2311 – 07 «Профилактика стрессового состояния работников при различных видах профессиональной деятельности».

43. ГОСТ 12.1.030-81 ССБТ Защитное заземление, зануление.

44. ГОСТ Р 12.1.019-2009 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты.

45. ГОСТ Р 50948-2001 Средства отображения информации индивидуального пользования. Общие эргономические требования и требования безопасности.

46. Федеральный закон №89 от 1998 г. «Об отходах производства и потребления». Глава III, ст. № 9. — 1988. — С. 39.

47. ГОСТ Р 53692-2009 Ресурсосбережение. Обращение с отходами. Этапы технологического цикла отходов — введ. впервые 15.12.2009. — Москва: Стандартинформ, 2011. — С. 20.

48. Федеральный классификационный каталог отходов (с изменениями на 4 октября 2021 года) [Электронный ресурс]. – 2021. – Режим доступа: http://www.consultant.ru/document/cons_doc_LAW_218071/, свободный.

49. Постановление Правительства РФ от 28.12.2020 N 2314 «Об утверждении Правил обращения с отходами производства и потребления в части осветительных устройств, электрических ламп, ненадлежащие сбор, накопление, использование, обезвреживание, транспортирование и размещение которых может повлечь причинение вреда жизни, здоровью граждан, вреда животным, растениям и окружающей среде».

50. N 123-ФЗ от 22.07.2008 (ред. от 30.04.2021) Технический регламент о требованиях пожарной безопасности.

51. Правила устройства электроустановок [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/1200030216, свободный.

52. Правила по охране труда при эксплуатации электроустановок [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/573264184, свободный.

53. СП 12.13130.2009 Определение категорий помещений, зданий и наружных установок по взрывопожарной и пожарной опасности.

54. Критерии отнесения объектов, оказывающих негативное воздействие на окружающую среду, к объектам I, II, III и IV категорий [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/573292854, свободный.

Приложение А (справочное)

Data Preprocessing Methods

Студент

Группа	ФИО	Подпись	Дата
8ПМ0И1	Сафронов Василий Сергеевич		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Аксёнов Сергей Владимирович	к.т.н.		

Консультант – лингвист отделения иностранных языков ШБИП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Степура Светлана Николаевна	к.ф.н		

Introduction

Data preprocessing is the process of preparing the source data for analysis. The data is brought into compliance with the requirements that are determined by the specifics of the task being solved. Data preprocessing is the most important stage of data analysis. In most cases, the analysis cannot be carried out if the data has not been preprocessed. Analytical algorithms simply will not be able to work with such data or the results of the analytical algorithms will be incorrect.

Data preprocessing includes the following two directions: cleaning and optimization.

Cleaning is carried out in order to exclude various factors that reduce the quality of data and interfere with the work of analytical algorithms. It includes processing duplicates, contradictions and fictitious values, restoring and filling in missing values, smoothing, noise suppression and editing abnormal values. In addition, during the cleaning process, violations of the structure, completeness and integrity of data are restored, incorrect formats are converted.

Data optimization as an element of preprocessing includes dimensionality reduction, identification and exclusion of insignificant features. The main difference between optimization and cleaning is that the factors eliminated during the cleaning process significantly reduce the accuracy of solving the problem or make the work of analytical algorithms impossible. The problems solved during optimization adapt the data to a specific task and increase the efficiency of their analysis.

Outliers Detection

In statistics and data analytics, there is such a term as outliers. Outliers are values that differ significantly from most of the values in the dataset. The reasons for the appearance of outliers may be equipment failures, human factors, randomness, unique phenomena, etc. It is very important to detect and evaluate outliers in order to improve the quality of the analysis. Next, several methods for detecting outliers will be considered [6].

1. Extreme data analysis. No special statistical methods are used in this analysis. The algorithm of this method is as follows:

– Visualize data using charts (scatterplot or boxplot) and histograms to find extreme values;

– Use, for example, a Gaussian distribution, and find values whose standard deviation differs 2-3 times from the mathematical expectation or one and a half times from the first or third quartiles.

2. The approximating method, which consists in the use of clustering methods. The algorithm is as follows:

– Use the clustering method to identify clusters in the data;

– Identify and mark the centroids of each cluster;

– Correlate clusters with data instances located at a fixed distance or percentage distance from the centroid of the corresponding cluster.

3. Projecting methods. The algorithm is as follows:

– Use one of the projection methods, for example, the principal component method or Kohonen self-organizing maps or Sammon projection, to summarize the training data in two dimensions;

– Visualize data display;

– Use the proximity criterion from the projected values or from the vector of the coding table to identify outliers.

The simplest and most visual way to detect outliers is to visualize the values that a particular feature takes. The distribution of values can be visualized using

scatterplot and boxplot. Figure 1 shows a graph of the relationship between two variables. Here, an outlier is a point that is located far from other points.

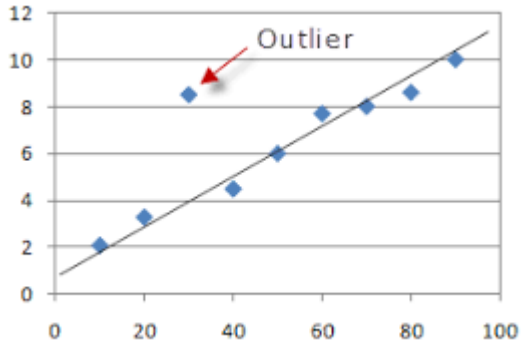


Figure 1. Outliers on a scatterplot

Figure 2 shows the boxplot structure. Outliers are those values that are located above the sum of the third quartile with 1.5 interquartile range or those values that are located below the difference of the first quartile and 1.5 interquartile range.

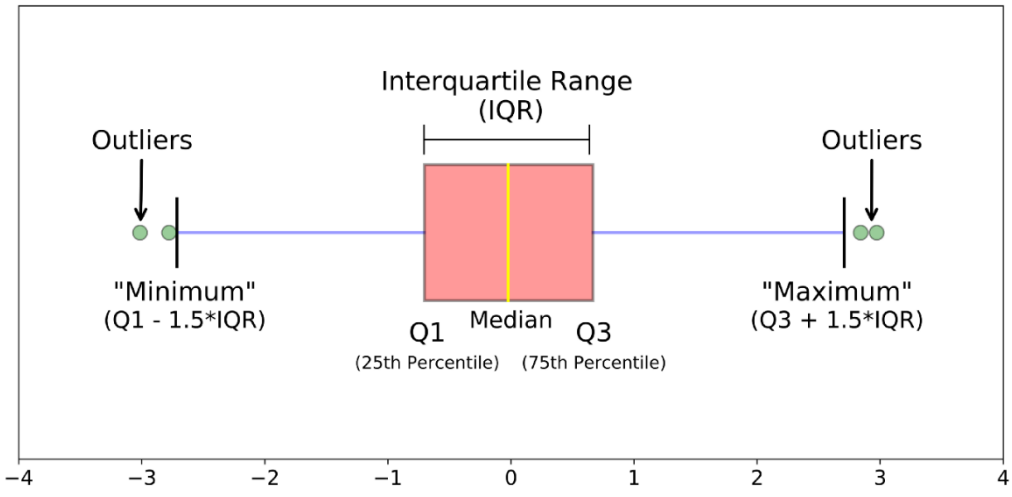


Figure 2. Boxplot structure [5]

Next, the stages of detecting outliers in extreme data analysis will be discussed in more detail:

1. Calculation of the median value Q_2 (the value located in the middle of the dataset). If the number of values in the dataset is odd, then the median is the value before and after which the same number of values are located. If the number

of values is even, then the median is calculated as the arithmetic mean of the two values in the middle.

2. Calculation of the lower quartile Q_1 (the value below which 0.25 values from the dataset are located, i.e. half of the values lying below the median). If there is an even number of values before the median, then, as in the previous step, the arithmetic mean of the two values in the middle is calculated.

3. Calculation of the upper quartile Q_3 (the value above which 0.25 values from the dataset are located, i.e. half of the values lying above the median). The process of calculating the upper quartile is similar to the calculation of the lower quartile.

4. Calculation of the interquartile range. This indicator represents the distance between the upper and lower quartiles and is calculated as follows: $Q_3 - Q_1$

5. Defining the internal boundaries of values in a dataset. Values outside the internal boundaries are considered minor outliers, and those outside the external boundaries are considered significant outliers. The internal boundaries are defined as follows:

5.1. $(Q_3 - Q_1) * 1{,}5$

5.2. Lower inner boundary = $Q_1 - (Q_3 - Q_1) * 1{,}5$

5.3. Upper inner boundary = $Q_3 + (Q_3 - Q_1) * 1{,}5$

6. Defining the outer bounds of values in a dataset. The outer boundaries are defined as follows:

6.1. $(Q_3 - Q_1) * 3$

6.2. Lower outer boundary = $Q_1 - (Q_3 - Q_1) * 3$

6.3. Upper outer boundary = $Q_3 + (Q_3 - Q_1) * 3$

After detecting outliers, it is necessary to decide what to do with such values in the dataset. Outliers can be excluded or left in the dataset. The main factor influencing this decision is the cause of this anomaly. Outliers resulting from an error are excluded. Outliers associated with new information or a trend are left in

the data set, due to erroneous exclusion of outliers, some previously unknown trend or discovery may be missed [6].

In addition to the above, outliers are also considered values that differ from most values not only in magnitude, but also in data type. Such cases may occur if, when filling in a large number of signs, a specialist made a typo, entered letter values instead of a number, filled in an adjacent field by mistake, etc. In this case, outliers are often excluded if it is impossible to restore the values.

Missing Values

It is not uncommon for a data set to contain missing values. The reasons for missing values in the data may be different. The field to fill in was skipped due to the inattention of the fill-in, or the field was left empty due to the fact that there is no data to fill in. For example, the patient was not assigned a specific study, so there are no records of its results. Also, if it was decided to exclude the detected outliers, then the missing values appeared instead.

In order to understand how to properly process the missing values, it is necessary to determine the mechanisms of their formation. There are the following three mechanisms for generating missing values: MCAR, MAR, MNAR [1].

MCAR (Missing Completely At Random) is a mechanism for generating missing values, in which the probability of missing for each record of the set is the same. For example, if a sociological survey was conducted in which one randomly selected question was not asked to every tenth respondent. Moreover, the respondents answered all the other questions asked, then the MCAR mechanism takes place. In this case, ignoring or excluding records containing missing data does not lead to distortion of the results.

MAR (Missing At Random). In practice, the data is usually omitted not by chance, but due to some patterns. Missing values are classified as MAR if the probability of missing can be determined based on other information available in the dataset (gender, age, position, education, etc.) that does not contain missing values. In this case, deleting or replacing the omissions with the value “Missing value”, as in the case of MCAR, will not significantly distort the results.

MNAR (Missing Not At Random) is a mechanism for generating missing values, in which data is missing depending on unknown factors. MNAR assumes that the probability of missing could be described based on other attributes, but there is no information on these attributes in the dataset. As a consequence, the probability of missing cannot be expressed based on the information contained in the dataset.

There are several ways to eliminate missing values. Below are the most common ways to eliminate missing values:

– Filling in the missing values with zero (0). If the numeric attribute takes both negative and positive values, filling in 0 will indicate the neutrality of the filled case. This method is also applicable to categorical features, when the value 0 characterizes the absence of this feature in an individual observation [7].

– Filling in missing values with statistical indicators. The arithmetic mean, mode, and median are applied. If the attribute does not have outliers, then the average value is used, otherwise the median is used, because this indicator is resistant to outliers.

– Filling in missing values with values found by auxiliary models. Based on the values of other features and known values of a feature that has missing values, a machine learning model is trained and the missing values are predicted.

– Creating indicator variables. The missing values are replaced with zeros and a new attribute is added, which takes the value 1 for observations with missing values and 0 for observations without missing values, or vice versa.

– Filling in the missing values with the value of the neighboring observation. This method is used when filling in time series when subsequent values are strongly related to the previous ones.

– Exclusion from the observation dataset with missing values. If the data set is large enough and there are few missing values, then deleting the observation will not have a significant impact on further analysis. Also, the observation is deleted if the missing values are present in a large number of features.

– Exclusion of a feature with missing values from the dataset. If a feature contains a large number of missing values that cannot be restored, such a feature is removed from the dataset [1].

Dimensionality Reduction

With a real-world dataset, there are usually tons of attributes, and if their number is not reduced, it may affect the performance of the model later when this dataset is passed to it. Reducing the number of features while keeping as much variation in the dataset as possible will have a positive impact in many ways, such as:

– Requiring less computational resources

– Increasing the overall performance of the model

– Preventing overfitting (when the model becomes too complex and the model memorizes the training data, instead of learning, so in the test data the performance decreases a lot)

– Avoiding multicollinearity (high correlation of one or more independent variables). Also, applying this technique will reduce the noise data.

Next, we will consider the main types of dimensionality reduction that can be applied to the data in order to improve them for later use.

Feature selection refers to the process of selecting the most important variables (features) related to the prediction variable, in other words, selecting the attributes which contribute most to the model. Here are some techniques for this approach that can be applied either automatically or manually:

– Correlation Between Features: This is the most common approach, which drops some features that have a high correlation with others.

– Statistical Tests: Another alternative is to use statistical tests to select the features, checking the relationship of each feature individually with the output variable.

– Recursive Feature Elimination (RFE): The Recursive Feature Elimination, also known as Backward Elimination, where the algorithm trains the model with all features in the dataset, calculating the performance of the model, and then drops one feature at a time, stopping when the performance improvement becomes negligible [2].

– Variance Threshold: Another feature selection method is the variance threshold, which detects features with high variability within the column, selecting those that got over the threshold. The premise of this approach is that features with low variability within themselves have little influence on the output variable.

Also, some models automatically apply a feature selection during the training. The decision-tree-based models can provide information about the feature importance, giving a score for each data feature. The higher the value, the more relevant it is for the model.

As the name suggests, the linear methods use linear transformations to reduce the dimensionality of the data.

The most common approach: The Principal Component Analysis, a method that transforms the original features in another dimensional space captures much of the original data variability with far fewer variables. However, the new transformed features lose the interpretability of the original data, and it only works with quantitative variables.

Other types of linear methods are Factor Analysis and Linear Discriminant Analysis.

The non-linear methods (or manifold learning methods) are used when the data doesn't fit in a linear space. The idea behind this technique is that in a high dimensional space, most of the important features lie in a small number of low dimensional manifolds. Many algorithms make use of this approach.

The Multi-Dimensional Scaling (MDS) is one of those, and it calculates the distance between each pair of objects in a geometric space. This algorithm transforms the data to a lower dimension, and the pairs that are close in the higher dimension remain in the lower dimension as well.

The Isometric Feature Mapping (Isomap) is an extension of MDS, but instead of Euclidean distance, it uses the geodesic distance.

Other examples of non-linear methods are Locally Linear Embedding (LLE), Spectral Embedding, t-distributed Stochastic Neighbor Embedding (t-SNE) [1].

Data Transformation

There are two main types of data: categorical and numeric. Figure 3 shows the classification of data by type.

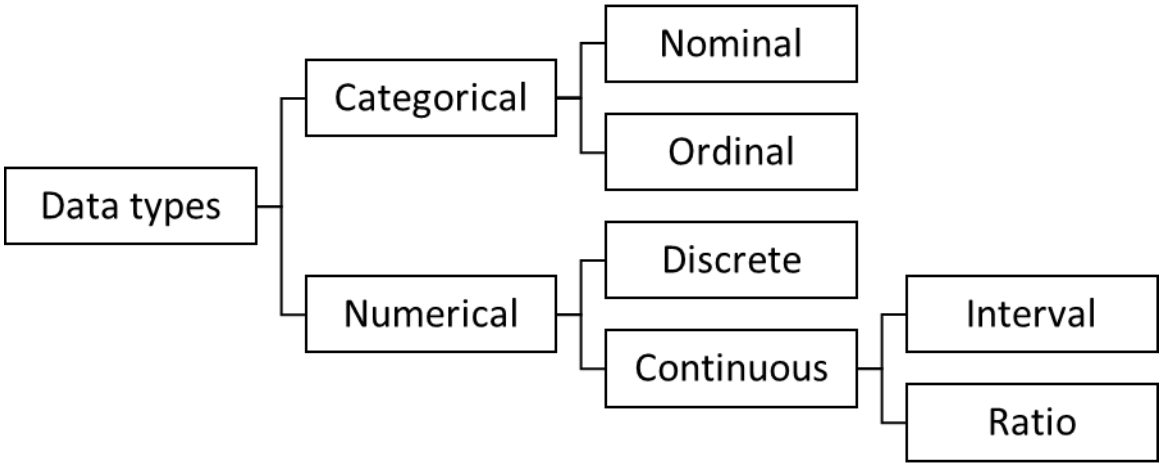


Figure 3. Data classification by type

Nominal values are discrete units and are used to denote variables that have no quantitative value, nominal data have no order. An example of nominal data can be gender (female, male).

Ordinal values are discrete and ordered units. An example of ordinal data is the stages of education (primary, secondary, further and higher).

Discrete data is a type of data that cannot be measured, but can be counted. An example would be the number of heads in 100 coins. To check whether the data is discrete, it is necessary to answer the following two questions: can they be counted and can they be divided into smaller and smaller parts?

Continuous data represents measurements, so their values cannot be counted, but they can be measured. An example would be a person's height.

Interval values represent ordered units having the same difference. An example is the air temperature (-10°C, -5°C, 0°C, +5°C, +10°C).

The ratio values are also ordered units that have the same difference, but the ratios have absolute zero, i.e. the absence of one or another attribute is possible. An example would be height, weight, length, etc.

Most machine learning and optimization algorithms show better results if the features are on the same scales. The process of bringing features to the same scale is called scaling. Scaling is applied to numeric type data. There are two general approaches to bringing different features to the same scale: normalization and standardization.

Normalization means bringing features to the range [0, 1], minimax scaling is applied to each feature column, where the normalized value $x_{norm}^{(i)}$ from the sample $x^{(i)}$ can be calculated as follows [4]:

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}},$$

where $x^{(i)}$ is a single sample, x_{min} is the smallest value in the feature column, x_{max} is the maximum value in the feature column.

With the help of standardization, the feature columns are centered in the zero mean value, i.e. equal to 0, with a single standard deviation, i.e. equal to 1, as a result of which the feature columns take the form of a normal distribution. The standardization procedure can be represented by the following formula:

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x},$$

where μ_x is the empirical mean of a single feature column, σ_x is the standard deviation.

As for categorical features, they are most often presented in the form of text values that require transcoding into numbers.

Given the fact that the values of ordinal features are not equivalent to each other, i.e. the order matters, it is optimal to match each value with its ordinal number in a sorted list of all possible values. As an example, we can present the pre–formation of European sizes as follows: M – 1, L - 2, XL – 3.

When working with nominal variables, this approach is not applicable, each individual value is equivalent to the others and it is incorrect to put ordinal numbers in accordance with them. For example, the following matches cannot be put in the “color” attribute: blue is 0, green is 1, red is 2. Blue is not less than

green, and red is not more than blue. Therefore, a method called One Hot Encoding is used for nominal features. The essence of this method is that a dummy attribute is created for each unique value in the nominal attribute column. This attribute takes the value 1 if the presence of this attribute is characteristic of a particular observation, otherwise it is 0. Table 1 shows the initial values and the result of One Hot Encoding of the nominal attribute “color” [3].

Table 32. Example of a One Hot Encoding

Color	Color_green	Color_red	Color_blue
green	1	0	0
red	0	1	0
blue	0	0	1

Conclusion

In this part of the work, the methods of data preprocessing were considered.

The most well-known methods for detecting outliers were described. With the help of extreme data analysis, approximation method, and projection methods, such anomalies in the data as outliers can be detected.

Methods to eliminate missing values in the data were considered. There are the following three mechanisms for generating missing values: Missing Completely At Random, Missing At Random, Missing Not At Random. The missing values can be imputed by statistical indicators, a zero value, the value of a neighboring observation, values predicted by machine learning algorithms.

To reduce the use of computing resources, increase the overall performance of the model, prevent overfitting, and avoid multicollinearity, dimensionality reduction methods are used.

There are different types of data. Depending on the data type, various transformations can be performed on the data. If the numerical data are in different scales, then scaling is necessary. There are two types of scaling: standardization and normalization. If categorical features are presented in a non-numeric format, then it is necessary to encode the values of the features.

The process of preprocessing the initial dataset is a very important step in data analysis. The quality of the analysis is directly related to the quality of the data.

References

1. Data Preprocessing: 6 Techniques to Clean Data – Scalable path [electronic resource]. – URL: https://www.scalablepath.com/data-science/data-preprocessing-phase/. Accessed 5.06.2020.

2. 6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples). - Towards Data Science – URL: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779. – Accessed 27.05.2020.

3. Machine learning algorithms in simple language. Part 1 [electronic resource]. – URL: https://medium.com/nuances-of-programming/алгоритмы-м.. – Accessed 27.05.2020.

4. Sebastian Raschka. Python Machine Learning. Birmingham: Packt Publishing, 2015. – 454 p.

5. Understanding Boxplots. - Towards Data Science [electronic resource]. – URL: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51. – Accessed 27.05.2020.

6. What are outliers in the data? – Engineering Statistics Handbook [electronic resource]. – URL: https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm. Accessed 5.06.2020.

7. What Is Data Preprocessing in ML? – Serokell [electronic resource]. – URL: https://serokell.io/blog/data-preprocessing/. Accessed 5.06.2020.

Приложение Б – Признаки исходного набора данных

№	Название признака	Тип данных	Оставлен для анализа + Исключен -	Название листа рабочей книги
1.	№ медицинской карты	Текст	-	Лист «База данных» Общая информация
2.	Дата поступления (госпитализации)	Дата	-	
3.	Дата выписки	Дата	-	
4.	Возраст	Целое число	+	
5.	Рост	Целое число	Индекст масс тела	
6.	Вес	Целое число		
7.	Пол	Текст	+	
8.	Диагноз заключительный	Текст	+	
9.	Форма КЭ	Текст	-	
10.	Форма ИКБ	Текст	-	
11.	Поражение органов при ИКБ	Текст	-	
12.	2-х волновое течение КЭ	Текст	-	
13.	Осложнения КЭ ОНГМ / судороги	Текст	-	
14.	Осложнения КЭ Парезы	Текст	-	
15.	Сопутствующие болезни	Текст	+	
16.	Дата присасывания клеща	Дата	-	Лист «База данных» Извещение
17.	Дата начала заболевания	Дата	-	
18.	Присасывание клеща	Текст	+	
19.	Локализация присасывания на теле	Текст	+	
20.	Единичные или множественные укусы клеща	Текст	+	
21.	Наличие эритемы на месте укуса клеща	Текст	+	
22.	Факт присасывания клеща в анамнезе (в	Текст	+	

	прошлом)			
23.	Дата введения антибиотика	Дата	-	
24.	Дата введения иммуноглобулина	Дата	-	
25.	Была ли плановая вакцинация	Текст	+	
26.	Температура при поступлении	Дробное число	-	Лист «База данных» Осмотр Оценка витальных функций
27.	Пульс в мин / частота сердечных сокращений	Целое число	-	
28.	Систолическое давление (СД)	Целое число	-	
29.	Диастолическое давление (ДД)	Целое число	-	
30.	Употребление речной рыбы семейства карповых	Текст	+	Лист «База данных» Осмотр Анамнез жизни
31.	Злоупотребление алкоголем	Текст	+	
32.	Курение	Текст	+	
33.	Употребление наркотических препаратов	Текст	+	
34.	Головная боль	Текст	+	Лист «База данных» Осмотр Жалобы
35.	Головокружение/мушки	Текст	+	
36.	Светобоязнь /гиперэстезия	Текст	+	
37.	Болезн. глазн. яблок при движении	Текст	+	
38.	Боль в горле при глотании	Текст	+	
39.	Боль по ходу позвоночника	Текст	+	
40.	Тошнота	Текст	+	
41.	Рвота	Текст	+	
42.	Онемение / парестезии пальцев / лица	Текст	+	
43.	Боли в мышцах / Миалгия	Текст	+	
44.	Боли /Ломота в суставах	Текст	+	
45.	Озноб	Текст	+	
46.	Сонливость	Текст	+	
47.	Заторможенность	Текст	+	

48.	Слабость / Утомляемость	Текст	+	
49.	Бессоница	Текст	+	
50.	Возбуждение	Текст	+	
51.	Сыпь / Экзантема на коже	Текст	+	
52.	Боли в сердце / грудной клетке	Текст	+	
53.	Одышка	Текст	+	
54.	ДРУГИЕ жалобы	Текст	+	
55.	Покраснение / увеличение миндалин / задней стенки глотки	Текст	+	Лист «База данных» Осмотр Объективный статус
56.	л/узлы лимфоузлы	Текст	+	
57.	Печень не выступает / выступает из-под края реберной дуги	Текст	+	
58.	Изменения суставов: покраснение/ отечность / болезненность при движении	Текст	+	
59.	Сознание	Текст	+	Лист «База данных» Невролог
60.	Походка	Текст	+	
61.	Поза Ромберга	Текст	+	
62.	ПНП пальце-носовая проба	Текст	+	
63.	Речь	Текст	+	
64.	ЧН черепные нервы / глазные щели, зрачки, движение глазных яблок	Текст	+	
65.	Конвергения	Текст	+	
66.	Нистагм	Текст	+	
67.	Болезненность тригеминальных точек (точек тройничного нерва)	Текст	+	
68.	Носогубная складка	Текст	+	
69.	Лицевой нерв	Текст	+	
70.	Глоточный рефлекс	Текст	+	

71.	Язык	Текст	+	
72.	объем движений конечностей	Текст	+	
73.	Сила мышц	Текст	+	
74.	Тонус мышц	Текст	+	
75.	Рефлексы (периостальные рефлексы, сухожильные рефлексы)	Текст	+	
76.	Менингиальные знаки: ригидность затылочных мыщц	Текст	+	
77.	Менингиальные знаки: симптом Кернига	Текст	+	
78.	Менингиальные знаки: симптомы Брудзинского	Текст	+	
79.	Симптом Лассега	Текст	+	
80.	Симптом Вассермана	Текст	+	
81.	Симптом Бабинского	Текст	+	
82.	План лечения (Этиотропный препарат)	Текст	-	
83.	Температура	Дробное число	-	Лист «Температурный лист»
84.	Систолическое давление (СД)	Целое число	-	
85.	Диастолическое давление (ДД)	Целое число	-	
86.	Пульс в мин. /ЧСС	Целое число	-	
87.	Эритроциты / RBC	Дробное число	+	Лист «Общий анализ крови»
88.	Гемоглобин / HGB	Целое число	+	
89.	Лейкоциты / L / WBC	Дробное число	+	
90.	Эозинофилы / EOS%	Дробное число	+	
91.	Нейтрофилы / NEU%	Дробное число	+	
92.	Лимфоциты / LYM%	Дробное число	+	
93.	Моноциты / MON%	Дробное число	+	
94.	Базофилы / BAS%	Дробное число	+	
95.	Скорость оседания эритроцитов / СОЭ	Целое число	-	

96.	Гематокрит / HCT	Дробное число	-	
97.	MCV	Дробное число	-	
98.	MCH	Дробное число	-	
99.	MCHC	Целое число	-	
100	RDW-CV	Дробное число	-	
101	RDV-SD	Дробное число	-	
102	MPV	Дробное число	-	
103	Тромбоциты / PLT	Целое число	-	
104	PDW	Дробное число	-	
105	PCT	Дробное число	-	
106	P-LCC	Дробное число	-	
107	P-LCR	Дробное число	-	
108	Билирубин общий / BiliT	Дробное число	-	Лист «Биохимический анализ крови»
109	Билирубин прямой / BiliD	Дробное число	-	
110	Общий белок / TP	Целое число	-	
111	Глюкоза / GluC	Дробное число	-	
112	Аспартатаминотрансфераза / AST	Целое число	-	
113	Аланинаминотрансфераза / ALT	Целое число	-	
114	Амилаза общая / Amy	Целое число	-	
115	Щелочная фосфатаза / AlkP	Целое число	-	
116	Мочевина / Urea	Дробное число	-	
117	Креатинин / Crea	Дробное число	-	
118	Натрий / Na-C	Дробное число	-	
119	Калий / K-C	Дробное число	-	
120	Хлор / Cl-C	Дробное число	-	
121	Кальций общий / Ca-C	Дробное число	-	
122	Фосфор / Phos	Дробное число	-	
123	Железо / Fe-Pl	Дробное число	-	

124	С-реактивный белок / CRP32	Дробное число	-	
125	Мочевая кислота / UA	Дробное число	-	
126	Холестерин общий / Chol	Дробное число	-	
127	Триглицериды / Trig	Дробное число	-	
128	Креатинфосфокиназа / CK	Дробное число	-	
129	Гамма-глутамин трансфераза / CGT	Дробное число	-	
130	Лактатдедидрогиназа / LDH	Дробное число	-	
131	Альбумин / AlbG	Дробное число	-	
132	IgM к ИКБ	Текст	-	Лист «Иммуноферментный анализ / ИФА»
133	IgG к ИКБ	Текст	-	
134	IgM к ВКЭ	Текст	-	
135	IgG к ВКЭ	Текст	-	
136	ликвор на ВКЭ	Текст	-	
137	ПЦР к ИКБ Кровь	Текст	-	
138	АГ ВКЭ в крови	Текст	-	
139	IgM ГЭЧ	Текст	-	
140	IgG ГЭЧ	Текст	-	
141	IgM МЭЧ	Текст	-	
142	IgG МЭЧ	Текст	-	
143	Цвет	Текст	-	Лист «Анализ спинномозговой жидкости / Ликвор»
144	Прозрачность	Текст	-	
145	Цитоз	Текст	-	
146	Лимфоциты	Целое число	-	
147	Нейтрофилы	Целое число	-	
148	Белок	Дробное число	-	
149	Глюкоза	Дробное число	-	
150	Хлориды	Целое число	-	
151	Положение оси	Текст	-	Лист «ЭКГ»

152	Ритм	Текст	-	
153	ЧСС	Целое число	-	
154	диффузные изменения миокарда	Текст	-	
155	Гипертрофия миокарда	Текст	-	
156	Другие возможные изменения	Текст	-	
157	Отек диска зрительного нерва	Текст	-	Лист «Окулист»
158	Ангиоретинопатия по гипертензивному типу	Текст	-	
159	Ангиоретинопатия по гипотоническому типу	Текст	-	
160	артерии сужены, вены расширены	Текст	-	
161	Другие возможные изменения	Текст	-	
162	Печень	Текст	-	Лист «УЗИ»
163	Холецистит	Текст	-	
164	ЖКБ	Текст	-	
165	диффузные изменения поджелудочной железы	Текст	-	
166	Увеличение поджелудочной железы	Текст	-	
167	Увеличение селезенки	Текст	-	
168	Другие возможные изменения	Текст	-	