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)

Школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Отделение школы (НОЦ) Отделение информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Алгоритмическое и программное обеспечение выделения значимых предикторов из медицинской документации осмотра пациента

УДК 004.4.021:61:002.1:616-071.2

Студент

Группа	ФИО	Подпись	Дата
8ПМ7И	Демченко Ирина Сергеевна		

Руководитель ВКР

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

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель	Потехина Н.В.			

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Горбенко М.В.	к.т.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

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

Томск – 2019 г.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

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

Школа информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Отделение школы (НОЦ) Отделение информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Губин Е.И.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

магистерской диссертации

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

Студенту:

Группа	ФИО
8ПМ7И	Демченко Ирине Сергеевне

Тема работы:

Алгоритмическое и программное обеспечение выделения значимых предикторов из медицинской документации осмотра пациента	
Утверждена приказом директора (дата, номер)	№1436/с от 25.02.2019г.

Срок сдачи студентом выполненной работы:	06.06.2019

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 данной работы является история болезни пациента. Непосредственным предметом исследования является один из документов истории болезни – «Осмотр в стационаре при поступлении». Для анализа доступны 70 историй болезни пациентов, страдающих рожистыми воспалениями. Деперсонализированные истории болезни 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; извлечение признаков из текстовых данных; оценка значимости признаков; построение классификатора фрагментов документа «Осмотр в стационаре при поступлении»; 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)*	

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Потехина Нина Васильевна
Социальная ответственность	Горбенко Михаил Владимирович
Раздел на иностранном языке	Диденко Анастасия Владимировна

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

Литературный обзор

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	25.02.2019

Задание выдал руководитель / консультант (при наличии):

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

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Демченко Ирина Сергеевна		

Министерство образования и науки Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Уровень образования магистратура
Отделение школы (НОЦ) Отделение информационных технологий
Период выполнения весенний семестр 2018/2019 учебного года

Форма представления работы:

Магистерская диссертация

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	06.06.2019

Дата контроля	**Название раздела (модуля) / вид работы (исследования)**	**Максимальный балл раздела (модуля)**
7.03.2019	Аналитический обзор	15
15.03.2019	Объект и методы исследования	20
15.04.2019	Расчеты и аналитика	20
05.05.2019	Результаты	20
15.05.2019	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	15
20.05.2019	Социальная ответственность	10

Составил преподаватель:

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

СОГЛАСОВАНО:

Руководитель ООП	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ПМ7И	Демченко Ирина Сергеевна

Школа	ИШИТР	Отделение школы (НОЦ)	Информационных технологий
Уровень образования	Магистр	Направление/специальность	09.09.04 Программная инженерия

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного-исследовательского проекта: материально-технических, энергетических, финансовых, информационных и человеческих*	*Оклад руководителя - 33664 руб.* *Оклад инженера- 21760 руб.* *Стоимость материальных ресурсов определялась согласно прейскурантам компаний*
2. *Нормы и нормативы расходования ресурсов*	*Накладные расходы 16%;* *Районный коэффициент 30%.* *Норма амортизации 33,3%*
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	*Коэффициент отчислений на уплату во внебюджетные фонды 30 %.*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала научно-исследовательского проекта*	*Анализ перспективности технических решений посредством Quad-анализа* *Диаграмма Исикавы* *Оценка готовности научно-исследовательского проекта к коммерциализации*
2. *Разработка устава научно-исследовательского проекта*	*Определение цели научно-исследовательского проекта, требований к проекту, описание заинтересованных стороны проекта, рабочей группы.*
3. *Планирование процесса управления научно-исследовательским проектом: 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. *Диаграмма Ганта* 4. *Бюджет затрат* 5. *Реестр рисков*	

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель ОСГН ШБИП	Потехина Нина Васильевна	-		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Демченко Ирина Сергеевна		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8ПМ7И	Демченко Ирине Сергеевне

Инженерная школа	**ИШИТР**	**Отделение**	Информационных технологий
Уровень образования	Магистратура	**Направление/специальность**	Программная инженерия

Исходные данные к разделу «Социальная ответственность»:	
1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	Алгоритм классификации текстовых блоков из истории болезни пациента
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности: – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	Рабочее место должно соответствовать требованиям ГОСТ 12.2.032-78. В соответствии с СН-245-71 в помещении должен быть организован воздухообмен. В соответствии с СН-181-70 рекомендуются следующие цвета окраски помещений: потолок – белый или светлый цветной; стены – сплошные, светло-голубые; пол – темно-серый, темно-красный или коричневый.
2. Производственная безопасность 2.1. Анализ выявленных вредных факторов при разработке и эксплуатации проектируемого решения в следующей последовательности: – физико-химическая природа вредности, её связь с разрабатываемой темой; – действие фактора на организм человека; – приведение допустимых норм с необходимой размерностью (со ссылкой на соответствующий нормативно-технический документ); – предлагаемые средства защиты; – (сначала коллективной защиты, затем – индивидуальные защитные средства). 2.2. Анализ выявленных опасных факторов при разработке и эксплуатации проектируемого решения в следующей последовательности: – механические опасности (источники, средства защиты; – термические опасности (источники, средства	Вредные факторы: - Электромагнитные излучения - Микроклимат - Освещенность рабочей зоны - Шум на рабочем месте Опасные факторы: - Статическое электричество - Короткое замыкание - Пожароопасность

защиты); – электробезопасность (в т.ч. статическое электричество, молниезащита – источники, средства защиты); – пожаровзрывобезопасность (причины, профилактические мероприятия, первичные средства пожаротушения).	
3. Экологическая безопасность: – защита селитебной зоны – анализ воздействия объекта на атмосферу (выбросы); – анализ воздействия объекта на гидросферу (сбросы); – анализ воздействия объекта на литосферу (отходы); – разработать решения по обеспечению экологической безопасности со ссылками на НТД по охране окружающей среды.	- Анализ негативного воздействия на окружающую природную среду: утилизация люминесцентных ламп, компьютеров и другой оргтехники
4. Безопасность в чрезвычайных ситуациях: – перечень возможных ЧС при разработке и эксплуатации проектируемого решения; – выбор наиболее типичной ЧС; – разработка превентивных мер по предупреждению ЧС; – разработка действий в результате возникшей ЧС и мер по ликвидации её последствий.	Наиболее типичная ЧС – пожар. Для повышения устойчивости объекта к пожарам необходимо использовать огнеупорные материалы, а также ознакомить персонал с режимом работы объекта в случае возникновения ЧС и обучить выполнению конкретных работ по ликвидации очагов поражения. Предусмотренные средства пожаротушения (согласно требованиям противопожарной безопасности СНиП 2.01.02-85): огнетушитель ручной углекислотный ОУ-5, пожарный кран с рукавом и ящик с песком (в коридоре). Кроме того, каждое помещение оборудовано системой противопожарной сигнализации.

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Горбенко Михаил Владимирович	к.т.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Демченко Ирина Сергеевна		

Планируемые результаты обучения по направлению

09.04.04 «Программная инженерия»

Код результата	Результат обучения (выпускник должен быть готов)
Общие по направлению подготовки 09.04.04 «Программная инженерия»	
Р1	Проводить научные исследования, связанные с объектами профессиональной деятельности
Р2	Разрабатывать новые и улучшать существующие методы и алгоритмы обработки данных в информационно-вычислительных системах
Р3	Составлять отчеты о проведенной научно-исследовательской работе и публиковать научные результаты
Р4	Проектировать системы с параллельной обработкой данных и высокопроизводительные системы
Р5	Осуществлять программную реализацию информационно-вычислительных систем, в том числе распределенных
Р6	Осуществлять программную реализацию систем с параллельной обработкой данных и высокопроизводительных систем
Р7	Организовывать промышленное тестирование создаваемого программного обеспечения
Профиль «Технологии больших данных» / «Big data solutions»	
Р8	Исследовать и анализировать большие данные, создавать их модели и интерпретировать структуры данных в таких моделях
Р9	Понимать принципы создания, хранения, управления, передачи и анализа больших данных с использованием новейших технологий, инструментов и систем обработки данных в высокопроизводительных сетях
Р10	Применять теорию распределенной системы управления базами данных к традиционным распределенным системам реляционных баз данных, облачным базам данных, крупномасштабным системам машинного обучения и хранилищам данных

РЕФЕРАТ

Выпускная квалификационная работа 101 с., 16 рис., 18 табл., 25 источников, 3 прил.

Ключевые слова: история болезни, текстовые данные, классификация, оценка значимости признаков, метод опорных векторов

Объектом исследования является (ются) истории болезни пациентов, страдающих рожистыми воспалениями, а именно документ «Осмотр в стационаре при поступлении».

Цель работы – разработка алгоритмического и программного обеспечения выделения значимых предикторов из медицинской документации осмотра пациента

В процессе исследования проводились аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; извлечение признаков из текстовых данных; оценка значимости признаков; построение классификатора фрагментов документа «Осмотр в стационаре при поступлении».

В результате исследования выявлены значимые признаки из медицинской документации осмотра пациента, разработана модель классификации фрагментов из истории болезни пациента

Область применения: здравоохранение

Экономическая эффективность/значимость работы применение системы позволит автоматически распределять вводимую пользователем (врачом) информацию по соответствующим блокам.

В будущем планируется работа по внедрению алгоритма в медицинские системы заполнения документации

Оглавление

Введение

Современный мир постоянно меняется и развивается. На данный момент, мобильных телефонов существует больше, чем людей. Люди используют виртуальных ассистентов, автомобили, управляемые автопилотом, а также ищут информацию в интернете о том или ином симптоме заболевания.

Здравоохранение и аналитика являются одними из самых быстрорастущих областей в промышленности и разработке учебных программ [1].

Современный мир называют эрой данных, потому что ежедневно мы собираем огромный объем данных. Данные получают как из социальных сетей, так и от конкретных датчиков. По некоторым оценкам к 2020 году каждый человек будет создавать 1,7 мегабайта данных в секунду [2]. В то же время, имея так много данных и не используя их, возникает вопрос, почему мы все еще собираем и храним так много данных? Очевидно, что мы должны использовать современные технологии не только для сбора и хранения, но и для извлечения знаний из доступных данных.

Система здравоохранения генерирует почти 1/3 мировых данных, и заинтересованные стороны в области здравоохранения надеются на аналитику данных и медицинскую информатику, благодаря которым желают устранить медицинские ошибки, сокращая количество повторных обращений, предоставляя медицинскую помощь на основе фактических данных и демонстрируя качественные результаты. Существует значительная потребность в использовании растущих объемов данных при помощи аналитики для анализа и принятия решений в здравоохранении [3].

Несмотря на то, что медицина была восприимчива к преимуществам больших данных и искусственного интеллекта, она медленно внедряла быстро развивающиеся технологии, особенно по сравнению с такими секторами как финансы, развлечения и транспорт [4].

Все данные можно разделить по их типу на структурированные и неструктурированные данные соответственно. Структурированные данные обладают высокой степенью организованности и упрощают поиск информации. Для этой цели структурированные данные обычно хранятся в реляционной базе данных. Неструктурированные данные не имеют предопределенной модели или схемы. Так как неструктурированные данные не имеют идентифицируемой структуры, это и создает сложности для поиска информации. Электронная почта, текстовые сообщения, публикации в социальных сетях являются хорошими примерами неструктурированных данных. Около 80% мировых данных представлены в неструктурированном виде. Далеко не всегда возможно преобразовать неструктурированные данные в структурированную модель, однако аналитика неструктурированных данных улучшается с использованием науки о данных и таких методов машинного обучения, как обработка естественного языка (NLP).

Большая часть медицинских данных представлена в виде изображений, например, результатов рентгенографии, или текстовых данных, будь то в рукописном или машинописном варианте. Очевидно, эти данные являются неструктурированными и сложнее поддаются анализу. Возможно, это послужило одной из причин медленного внедрения технологий машинного обучения в области медицины.

Целью работы является разработка алгоритмического и программного обеспечения выделения значимых предикторов из медицинской документации осмотра пациента, а также построение классификатора.

В рамках данной работы проводится обработка текстовых медицинских данных. Так как зачастую текстовые данные содержат достаточно объемный набор признаков, в рамках данного исследования выявляются наиболее значимые признаки для дальнейшего анализа. Также предложен подход классификации блоков электронных записей врачебного осмотра для сбора предикторов и формирования эффективной схемы лечения.

1. Обзор литературы

В современном мире к большинству задач применяются методы машинного обучения. Методы машинного обучения и анализа данных помогают прогнозировать будущее, строить модели принятия решений для выбора наилучшей из двух и более альтернатив.

Модели принятия решений принимают разные формы, и одной из них является прогнозное моделирование. Прогнозное моделирование – это использование алгоритма и программного обеспечения для больших наборов данных для прогнозирования потенциальных результатов, где алгоритм представляет собой формулу или расчет, используемый для решения проблемы в модели. В сфере здравоохранения существует огромная возможность использовать большие данные для прогнозного моделирования.

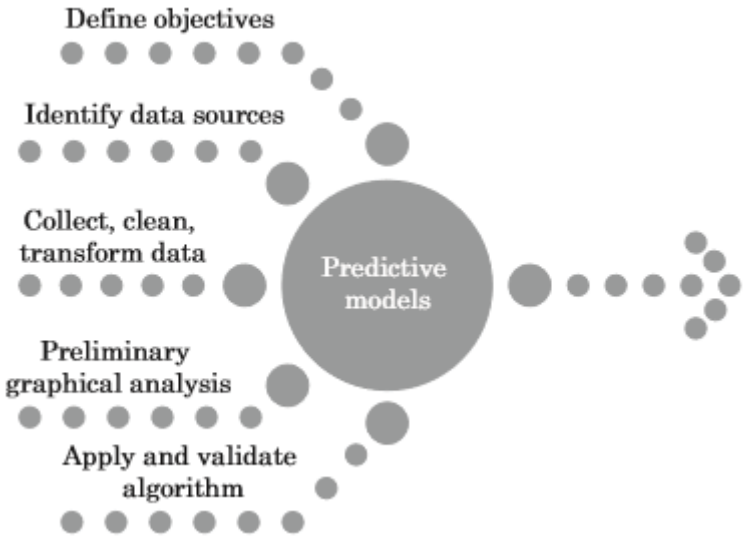

Рисунок 1 – Этапы прогнозного моделирования

На Рисунок 1 показаны этапы прогнозного моделирования. Получив оценку или прогноз будущих результатов, ученые должны включить их в процесс принятия решений. Если у них есть прогнозы, необходимо поделиться этой информацией и убедиться, что медицинские работники правильно интерпретируют результаты работы модели.

Несколько научных исследований были сосредоточены на обработке текстовой информации, доступной в наборах данных здравоохранения.

Анализ текстовых данных является извлечением полезных знаний из текстовых данных.

Общая стратегия построения модели с использованием интеллектуального анализа текста включает следующие шаги:

- Сбор документов (например, медицинских карт),
- Предварительная обработка данных (например, исключить стоп-слова, лемматизировать, преобразовать в случай с одним размером, если необходимо),
- Извлечение признаков из текстовых данных (преобразование текстовых данных в числовые форматы),
- Построение модели машинного обучения (выбор модели, её параметров, оценка точности),
- Интерпретация результатов (например, интерпретация кластеров).

Текст содержит много качественной информации, которую сложно использовать в статистическом моделировании. В области здравоохранения врачи выражают свое мнение словами, которые содержат полезную информацию, не отраженную в других источниках. Эта информация может в дальнейшем использоваться для разработки интеллектуальных моделей и улучшения процесса здравоохранения. Тем не менее, традиционное построение модели требует в качестве входных данных количественной информации. Поэтому при анализе текстовых данных, на этапе извлечения признаков из текстовых данных, текст следует преобразовать в цифровую форму.

В настоящий момент существует три наиболее популярных метода форматирования текстовых данных к виду пригодному для подачи на вход алгоритмам машинного обучения.

1. *Word2Vec*

Слова можно кодировать разными способами. Самый простой способ — это их занумеровать, т.е. составить полный словарь из текста, собрать все

возможные словоформы, использованные в тексте, и пронумеровать все эти слова. Однако такой способ кодирования не несёт никакой смысловой нагрузки, т.е. по коду нельзя сказать, насколько близкими по смыслу являются слова, например, номер 7 и номер 457.

Для построения "осмысленного" пространства для слов был разработан метод word2vec [5], который отображает слова W в векторное пространство $V \subset \mathbb{R}^n$.

$$word2vec: W \rightarrow V$$

При этом, совместно употребляемые в тексте T слова из W отображаются в близкие (в смысле евклидовой метрики) точки пространства V. К тому же, над точками V можно выполнять операции, имеющие смысл в W.

Результатом работы ***word2vec*** является набор векторов (матрица) – кодов слов, которая получается с помощью обучения определённой нейросети на некотором тексте (упорядоченном множестве слов).

2. *TF-IDF*

TF-IDF (от англ. *TF — term frequency*, *IDF — inverse document frequency*) — это статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. При использовании меры *TF-IDF*, вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции. Данная мера часто используется в задачах информационного поиска и анализа текстов.

3. *GloVe*

Модель Global Vectors (*GloVe*) предложена лабораторией компьютерной лингвистики Стенфордского университета, сочетает в себе черты сингулярного разложения и метода *word2vec*. *GloVe* – модель обучения без учителя для представления слов, которая превосходит другие модели в

задачах на аналогии слов, сходство слов и задачам распознавания именованных сущностей.

Каждый из методов имеет свои преимущества и недостатки. Для данной работы наиболее актуален метод *TF-IDF*, так как есть необходимость в оценке важности слов в контексте не обусловленной лишь частой употребления слова.

2. Объект и методы исследования

2.1. Описание объекта

Объектом исследования данной работы является история болезни пациента. Непосредственным предметом исследования является один из документов истории болезни – «Осмотр в стационаре при поступлении». Данный медицинский документ содержит подробную информацию о состоянии пациента при поступлении в стационар. Документ «Осмотр в стационаре при поступлении» включает в себя следующие 11 блоков информации:

1. номер пациента, пол и возраст,
2. дата и время осмотра,
3. жалобы (раздел содержит описание беспокоящих пациента факторов, записанных со слов пациента),
4. анамнез болезни (хронология развития симптомов заболевания со слов пациента до поступления под наблюдение врача),
5. анамнез жизни (описание условий жизни и труда пациента, его поведенческих и пищевых привычек, а также ранее перенесенных заболеваний),
6. анамнез врачебно-трудовой экспертизы (ВТЭ, информация о листе нетрудоспособности),
7. объективный статус (указание наличия или отсутствия патологий по каждому из рассматриваемых органов и систем организма),
8. локальный статус (приводятся максимально подробные данные исследования поражённой системы),
9. диагноз при поступлении,
10. обоснование диагноза (в логической последовательности обосновывается диагноз с указанием только тех данных, которые этот диагноз подтверждают.),
11. диагноз.

Истории болезни, используемые при выполнении данной работы предоставлены Сибирским государственным медицинским университетом. Выборка содержит 74 истории болезни пациентов, страдающих рожистыми воспалениями.

Приложение Б содержит пример оцифрованного документа «Осмотр в стационаре при поступлении» из истории болезни пациента с диагнозом рожа. Оцифрованный документ представляет собой чередование следующих строк: название блока документа, содержание блока, пустая строка.

2.2.Методы

2.2.1. Метод извлечения текстовых признаков *TF-IDF*

В качестве исходных данных для данной работы имеется набор предварительно оцифрованных текстовых документов (*.txt*) с информацией из медицинского документа «Осмотр в стационаре при поступлении».

Классификаторы и алгоритмы машинного обучения не могут напрямую обрабатывать текстовые документы в их первоначальном виде, так как большинство моделей ожидают получить на вход числовые векторы признаки. Поэтому на этапе предварительной обработки текстовые данные преобразуются в более управляемое представление.

Один из наиболее распространенных подходов для извлечения признаков из текста заключается в использовании модели «мешка слов» (*bag of words model*): модели, в которой для каждого документа принимается во внимание наличие и частота слов, но порядок, в котором они употреблены, игнорируется.

TF (*term frequency* — частота слова) — отношение числа вхождений некоторого слова к общему числу слов документа. Таким образом, оценивается важность слова в пределах отдельного документа.

$$\mathrm{tf}(t,d) = \frac{n_t}{\sum_k n_k}, \tag{1}$$

где n_t есть число вхождений слова t в документ, а в знаменателе — общее число слов в данном документе.

IDF (*inverse document frequency* — обратная частота документа) — инверсия частоты, с которой некоторое слово встречается в документах коллекции. Основоположником данной концепции является Карен Спарк Джонс[6]. Учёт *IDF* уменьшает вес широкоупотребительных слов. Для каждого уникального слова в пределах конкретной коллекции документов существует только одно значение *IDF*.

$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{ d_i \in D \mid t \in d_i \}|}, \quad (2)$$

где $|D|$ — число документов в коллекции;

$|\{ d_i \in D \mid t \in d_i \}|$ – число документов из коллекции D, в которых встречается t (когда $n_t \neq 0$).

Выбор основания логарифма в формуле не имеет значения, поскольку изменение основания приводит к изменению веса каждого слова на постоянный множитель, что не влияет на соотношение весов.

Таким образом, мера *TF-IDF* является произведением двух сомножителей:

$$\text{Tf-idf}(t,d,D) = \text{tf}(t,d) \times \text{idf}(t,D). \quad (3)$$

Большой вес в *TF-IDF* получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

В частности, для каждого фрагмента ИБ в нашем наборе данных мы рассчитаем меру, называемую Term Frequency, Inverse Document Frequency, сокращенно обозначенной как tf-idf.

2.2.2. Метод опорных векторов

Метод опорных векторов – мощный и широко используемый алгоритм машинного обучения, который можно рассматривать как расширение персептрона. В методе опорных векторов задача оптимизации состоит в том,

чтобы максимизировать расстояние между разделяющей гиперплоскостью (границей решения) и самыми близкими к этой гиперплоскости объектами обучающей выборки, так называемыми опорными векторами [7]. Такие модели, как правило, имеют более низкую ошибку обобщения, тогда как модели с малым зазором более подвержены переобучению.

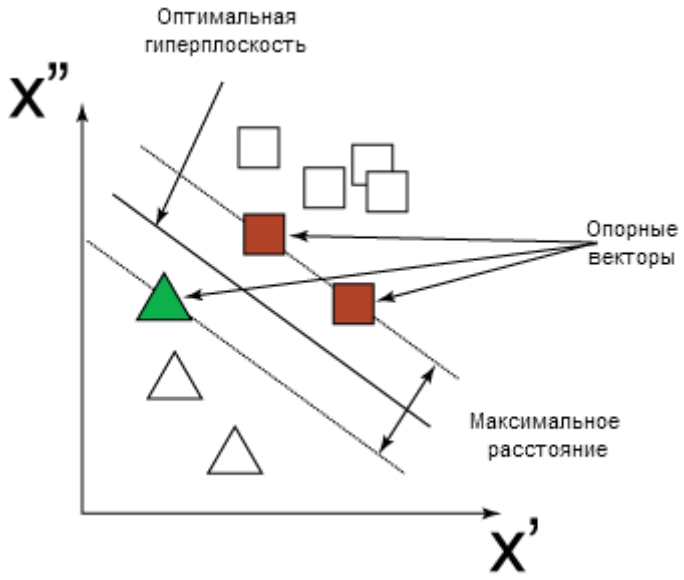

Рисунок 2 – Иллюстрация метода опорных векторов

Рисунок 2 иллюстрирует метод опорных векторов, оптимальную гиперплоскость, опорные векторы и максимальное расстояние.

Преимуществами данного метода являются:

- высокая скорость нахождения решающих функций;
- метод сводится к решению задачи квадратичного программирования в выпуклой области, которая всегда имеет единственное решение;
- метод находит разделяющую полосу максимальной ширины, что позволяет в дальнейшем осуществлять более уверенную классификацию.

Однако, стоит отметить, что метод опорных векторов чувствителен к шумам и стандартизации данных.

2.2.3. Критерий хи-квадрат для отбора признаков

При применении критерия хи-квадрат для выбора признаков, мы подсчитаем хи-квадрат статистику между каждым признаком и целевой функцией и выберем желаемое количество признаков с наибольшим значением критерия. Статистика хи-квадрат измеряет зависимость между признаком и целевой функцией и рассчитывается по следующей формуле:

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}, \quad (4)$$

где t – признак, c – целевой класс (функция), A – количество документов класса c, в которых встречается признак t, B –количество документов, класс которых отличен от c, в которых встречается признак t, C – количество документов класса c, в которых нет признака t, D – количество документов, класс которых отличен от c, в которых нет признака t; N – общее количество документов.

Очевидно, что если признак и целевая класс независимы, то данный признак не имеет значимости для классификации наблюдений и значение статистики будет равно 0. Для каждого класса считают статистику между этим классом и каждым признаком, а затем для признаков вычисляют среднее значение этой статистики по следующей формуле:

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} P(c_i)\, \chi^2(t, c_i), \quad (5)$$

где m – количество классов, $P(c_i)$ – вероятность принадлежности документа классу c_i.

В итоговый набор признаков отбираются либо заданное число признаков с наибольшим значением статистики (5), либо признаки, значение статистики (5) для которых больше некоторого предопределенного порога.

3. Расчеты и аналитика

3.1. Выбор программного обеспечения

В качестве языка программирования для решения поставленной задачи был выбран язык Python.

Python — высокоуровневый язык программирования, построенный на идея императивного, объектно-ориентированного и функционального программирования. Язык создан Гвидо ван Россумом в 1989 году и с тех пор непрерывно совершенствуется [8]. Преимущества Python:

– открытая разработка;

– язык довольно прост в изучении, особенно на начальном этапе;

– особенности синтаксиса стимулируют программиста писать хорошо читаемый код;

– имеет большое сообщество, позитивно настроенное по отношению к новичкам;

– множество полезных библиотек и расширений языка можно легко использовать в своих проектах благодаря предельно унифицированному механизму импорта и программным интерфейсам;

– механизмы модульности хорошо продуманы и могут быть легко использованы;

– абсолютно всё в Python является объектами в смысле ООП, но при этом объектный подход не навязывается программисту;

– Python имеет множество различных библиотек для машинного обучения.

Исходя из преимуществ, язык программирования Python подходит для решения поставленной задачи, т.е. разработки алгоритма классификации, а также извлечения значимых признаков из текстовых данных.

3.2. Используемые Python библиотеки

В ходе реализации проекта были использованы некоторые Python библиотеки для работы с определенными типами данных и использования специальных функций, расширяющих возможности языка.

Библиотека *NumPy* – это библиотека, добавляющая поддержку многомерных массивов и матриц, а также математических операций над этими массивами. В ходе работы были использованы объекты *NumPy* – *numpy.ndarray*, а также такие функции библиотеки как *arrange*(), *linspace*() и др.

Библиотека *pandas* – высокоуровневая Python библиотека, построенная поверх библиотеки Python. Данная библиотека предоставляет возможности использования структур *DataFrame* и *Series*. В ходе работы исходный набор данных хранился и обрабатывался в структуре *DataFrame*, а также была использована функция *pandas.read_csv* для считывания файла с исходными данными в структуру *DataFrame*.

Библиотека *Natural Language Processing* (*NLTK*) – библиотека для работ с естественными языками. Средствами данной библиотеки в работе при предварительной обработке данных были отфильтрованы стоп-слова, а также была проведена лемматизация.

Библиотека *Matplotlib* представляет собой модуль-пакет для Python, с помощью которого можно создавать рисунки и графики различных форматов. В ходе данной работы средства библиотеки были использованы для визуализации этапов работы и полученных результатов.

Seaborn также является Python библиотекой для визуализации данных. Данная библиотека построена на основе *Matplotlib* и предоставляет высокоуровневый интерфейс для отображения информативной статистической графики.

Библиотека *Scikit-learn* предоставляет возможности реализации ряда алгоритмов обучения, сокращения размерности данных, извлечения и отбора

признаков. Несколько алгоритмов обучения, а также функции для извлечения и отбора признаков были использованы в текущей работе.

3.3. Загрузка и предварительный анализ данных

Оцифрованные документы «Осмотр в стационаре при поступлении» из историй болезни пациентов содержатся в файле *medforms.txt*. Пример информации для одного пациента представлен в Приложении Б. Средствами библиотеки *pandas*, исходные данные загружаются в объект типа *DataFrame* – *text*. Из структуры данных в файле очевидно, что в text строки-названия блоков и строки с содержанием блоков документа чередуются, поэтому следующим шагом создается новый объект структуры *DataFrame df* со столбцами ‘*text’* и ‘*name’*, в которых представлены соответственно содержание и название блока документа. Далее для удобства добавляем столбец ‘*category_id’*, в котором закодируем цифрами имена блоков.

На этапе предварительного разведочного анализа данных требуется убедиться в сбалансированности используемого набора данных. В противном случае, необходимо учитывать факт несбалансированности данных при построении модели, а также при оценке точности модели. Используемый набор данных полностью сбалансирован, так как имеется одинаковое количество фрагментов каждого раздела документа из истории болезни.

Все предоставленные для анализа документы из историй болезни пациентов заполнены в полном объеме, что означает сбалансированность выборки данных.

3.4. Предварительная подготовка и выделение признаков из текстовых данных

На данном этапе работы средствами библиотеки *NLTK* была проведена фильтрация стоп-слов. Под стоп-словами в данном случае понимаются часто употребляемые слова (например, «а», «в», «что», «этот», «такой»), не несущие смысловой нагрузки. В процессе работы алгоритма, такие слова лишь занимают время обработки, поэтому логично удалить их из

рассмотрения. В текущей работе для удаления стоп-слов из рассмотрения был использован стандартный список стоп-слов для русского языка *nltk.corpus.stopwords.words('russian')*.

Для извлечения признаков из текстовых данных использован метод TF – IDF в реализации *sklearn.feature_extraction.text.TfidfVectorizer*, со следующими параметрами: минимальное количество упоминаний одного слова (*min_df*) – 2, *ngram_range* = (1,2) – извлечение слов и словосочетаний состоящих из 2 слов в качестве признаков. Извлеченные текстовые признаки содержатся в объекте *features*, метки категорий – *labels*.

Далее, из списка всех выделенных признаков *features*, при помощи метода хи-квадрат *sklearn.feature_selection_chi2*, выделяем значимые слова и словосочетания для каждого раздела документа.

3.5. Разделение данных на обучающее и тестовое подмножества

Наиболее распространенной стратегией выборки множества для обучения является разделение данных на обучающее и тестовое подмножества. Некоторые описывают подмножество тестирования как данные проверки или данные для оценки, но принцип тот же: мы строим модель на обучающем подмножестве, а затем оцениваем качество полученной модели на тестовом подмножестве, которое не было ранее использовано для обучения. Подмножества обучения и тестирования обычно создаются путем случайного выбора записей, так что каждая запись принадлежит одному и только одному подмножеству [9].

В данной работе, для разбиения множества исходных данных на обучающее и тестовое подмножества, будем использовать метод *sklearn.model_selection.train_test_split*. По умолчанию данный метод делит исходную выборку в следующих отношениях: 75% исходных данных будут использоваться для обучения, а оставшиеся 25% записей для тестирования полученной модели.

3.6.Построение классификатора

3.6.1. Выбор модели классификатора

Для построения классификатора в ходе работы были проведены исследования для нескольких моделей библиотеки *scikit-learn*, а также *catboost*. В процессе поиска наилучшей модели классификатора для данной задачи были рассмотрены следующие модели:

- Модель «случайного леса» – *sklearn.ensemble. RandomForestClassifier*,
- Метод опорных векторов – *sklearn.svm.LinearSVC*,
- Наивный байесовский классификатор для многоклассовой задачи – *sklearn.naive_bayes.MultinominalNB*,
- Логистическая регрессия – *sklearn.linear_model. LogisticRegression*,
- Классификатор *CatBoost* – *catboost.CatBoostClassifier*.

На данном этапе работы, классификаторы рассматриваются без специально подобранных макропараметров. Каждая модель оценивается по пять раз на случайном небольшом наборе записей из набора данных. Качество модели оценивается с помощью метрики точности – *accuracy*. Решение о выборе модели принимается на основе среднего значения точности пяти полученных моделей этого типа. Для данной задачи наилучшей моделью стал метод опорных векторов – *LinearSVC*.

3.6.2. Подбор оптимальных параметров классификатора

Теперь, когда модель выбрана, будем подбирать оптимальные макропараметры, ведь эффективность метода зависит от выбранных параметров. Для этого полезно знать, какие параметры были использованы данным методом по умолчанию:

```
sklearn.svm.LinearSVC(penalty='l2', loss='squared_hinge', d
ual=True, tol=0.0001, C=1.0, multi_class='ovr', fit_interce
pt=True, intercept_scaling=1, class_weight=None, verbose=0,
 random_state=None, max_iter=1000).
```

При детальном рассмотрении каждого из параметров модели, в данной задаче имеет смысл зафиксировать значения *random_state* и *dual*. Примем *random_state*=0 и *dual*=True, исходя из рекомендаций документации при случае числа записей в данных меньшем, чем число переменных.

Определим список параметров модели, которые будем варьировать при подборе:

- *'multi_class'* – данный параметр определяет стратегию мультиклассового разделения в случае наличия в данных объектов, относящихся более чем к двух классам.
    - *'multi_class'='ovr'* – реализация стратегии «один против всех». В данном случае обучается количество классификаторов равное количеству классов в данных. Классификатор с самым высоким значение функции выхода присваивает новый объект к определенному классу.
    - *'multi_class'='crammer_singer'* – реализация стратегии «один против одного». Здесь также обучается количество классификаторов равное количеству классов в исходных данных, однако теперь объект присваивается тому классу, к которому его отнесло большинство классификаторов.
- *'tol'* – параметр, используемый для критерия остановки. Рассмотрим три значения: 10^{-3}, 10^{-4}, 10^{-5}.
- *'C'* – штрафной параметр ошибки. Будем рассматривать значения: [0.01,0.1,1,10,100].

Далее будем производить исчерпывающий поиск по указанным параметрам. Параметры оптимизируются путем перекрестной проверки по сетке параметров. Каждая комбинация проверяется с использованием кросс-проверки, и выбирается та, которая проявила себя лучше всех других на кросс-проверке. Финальная модель, которая используется для тестирования и

классификации новых данных, обучается затем на всем множестве с использованием выбранных оптимальных параметров.

Для исчерпывающего поиска по сетке параметров создадим объект *grid_cv* как экземпляр класса *sklearn.model_selection.GridSearchCV*. В качестве параметров передаем данному методу модель, сетку параметров, метрику точности, а также параметр для определения стратегии перекрестной проверки. Будем использовать в качестве метрики качества модели *scoring='accuracy'*, и определим *cv*=5. В результате исчерпывающего поиска по сетке, в поле *best_estimator_* объекта *grid_cv* находится классификатор с наилучшими параметрами из представленных в сетке.

3.6.3. Построение классификатора с оптимальными параметрами

Далее будем рассматривать классификатор с оптимальными параметрами, выбранными при помощи поиска по сетке. Построим модель на всем обучающем подмножестве данных *X_train*, с помощью получившейся модели, предскажем значения на тестовом наборе данных *X_test*.

Оценим полученную модель при помощи известных значений меток для тестового подмножества данных *y_test* и предсказанного при помощи модели *y_pred*. Построим матрицу ошибок *sklearn.metrics.confusion_matrix*, для наглядности визуализируем ее средствами библиотеки *seaborn* – *seaborn.heatmap*. С помощью данного инструмента мы получаем легко интерпретируемую матрицу ошибок: элементы матрицы, находящиеся на главной диагонали, свидетельствуют о количестве верно предсказанных меток классов, в то время как недиагональные элементы матрицы говорят о количестве элементов тестового подмножества для которых метка класса была определена моделью ошибочно.

Также для оценки качества модели используется *sklearn.metrics.classification_report*, в результате которого имеем точность, чувствительность и гармоническое среднее точности и чувствительности, подсчитанное для каждого класса.

4. Результаты

4.1. Оценка значимости признаков

Из файла с доступным набором данных загружено 352 записи. При помощи метода *TF-IDF* из текстовых данных было выделено 2510 признаков.

С использованием критерия хи-квадрат были выявлены наиболее значимые признаки для каждого раздела документа «Осмотр при поступлении в стационар». Ниже представлен список слов и словосочетаний, наиболее значимых в каждом блоке документа:

```
# 'Анамнез ВТЭ':
  . Most correlated unigrams:
. нетрудоспособности
. нуждается
  . Most correlated bigrams:
. лист нетрудоспособности
. амбулаторно выдавался
# 'Анамнез болезни':
  . Most correlated unigrams:
. заболела
. клинику
  . Most correlated bigrams:
. заболела остро
. инфекционную клинику
# 'Анамнез жизни':
  . Most correlated unigrams:
. возрасту
. отрицает
  . Most correlated bigrams:
. бытовые условия
. семеиныи анамнез
# 'Дата и время осмотра':
  . Most correlated unigrams:
. время
. дата
  . Most correlated bigrams:
. 10 2016
. 2016 время
# 'Диагноз':
  . Most correlated unigrams:
. рожа
. a46
  . Most correlated bigrams:
. тяжести a46
. a46 рожа
# 'Диагноз при поступлении':
  . Most correlated unigrams:
. a46
. рожа
  . Most correlated bigrams:
. локализованная среднеи
. a46 рожа
# 'Жалобы':
  . Most correlated unigrams:
. повышение
. слабость
  . Most correlated bigrams:
. отек правои
. повышение температуры
# 'Локальный статус':
  . Most correlated unigrams:
. ощупь
. горячая
  . Most correlated bigrams:
. четкими границами
. горячая ощупь
# 'Номер пациента, пол и возраст':
  . Most correlated unigrams:
. пол
. женскии
  . Most correlated bigrams:
. пол женскии
. женскии возраст
# 'Обоснование диагноза':
  . Most correlated unigrams:
. начало
. интоксикации
  . Most correlated bigrams:
. учитывая острое
. острое начало
# 'Объективный статус':
  . Most correlated unigrams:
. сознание
. безболезненныи
  . Most correlated bigrams:
. выражение лица
. сознание полное
```

Сравним результаты полученные при помощи критерия хи-квадрат с результатами частотного анализа употребления слова в блоке на примере блока «Локальный статус». Гистограмма распределения часто употребляемых слов для блока «Локальный статус» представлена на Рисунок 3.

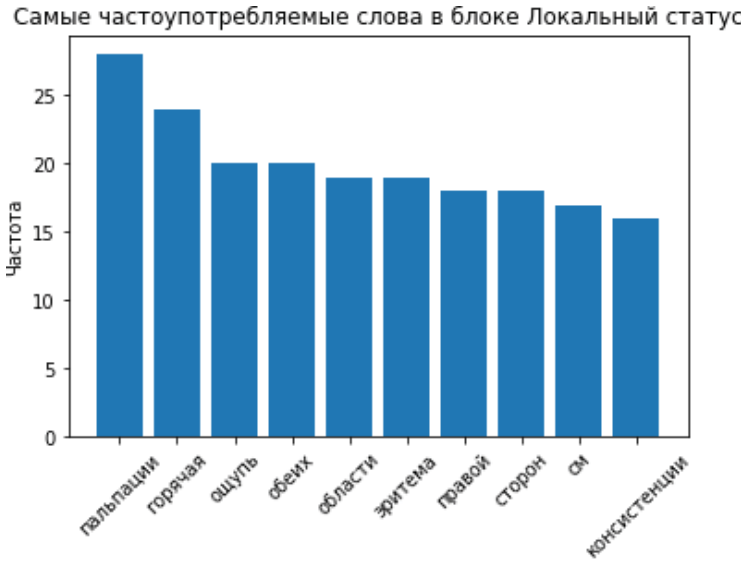

Рисунок 3 – Гистограмма распределения часто употребляемых слов в блоке «Локальный статус»

При изучении представленной выше гистограммы и результатов работы критерия хи-квадрат, можно сделать вывод о том, что самое часто употребляемое слово для этого блока «пальпации» не является значимым. Исходя из логики построения отображения текстовых данных методом *TF-IDF*, данное слово также часто встречается во фрагментах, относящихся к другим блокам, а потому частота употребления данного слова не может считаться значимой. Сравнивая результаты частотного анализа слов тех же блоков с оценкой критерия хи-квадрат, можно сделать вывод о том, что частота употребления конкретного слова в блоке не пропорциональна его значимости.

4.2. Классификатор фрагментов документа из истории болезни

Для построения классификатора в ходе работы были проведены исследования для нескольких моделей библиотеки *scikit-learn*, а также

catboost. Результаты оценок рассматриваемых моделей представлены на Рисунок 4.

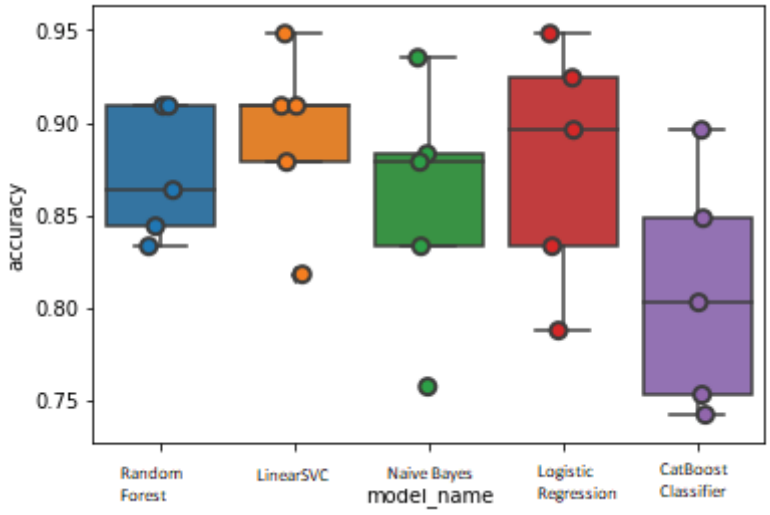

Рисунок 4 – Оценка точности различных моделей

В Таблица 1 представлена средняя точность полученных классификаторов.

Таблица 1 Средняя точность оцениваемых моделей классификаторов

Название классификатора	**Среднее значение точности**
CatBoostClassifier	0.808658
LinearSVC	0.892641
LogisticRegression	0.877922
MultinomialNB	0.857576
RandomForestClassifier	0.871861

Наибольшей средней точностью обладает модель *LinearSVC*. Данная модель была выбрана для дальнейшей настройки параметров.

В результате исчерпывающего поиска по сетке выбранных значений параметров, наилучшей оказалась модель со следующими параметрами:

```
LinearSVC(C=100, class_weight=None, dual=True, fit_intercept=True,
     intercept_scaling=1, loss='squared_hinge', max_iter=1000,
     multi_class='ovr', penalty='l2', random_state=0, tol=1e-05,verbose=0)
```

Для оценки качества построенной модели, воспользуемся матрицей ошибок, представленной на Рисунок 5.

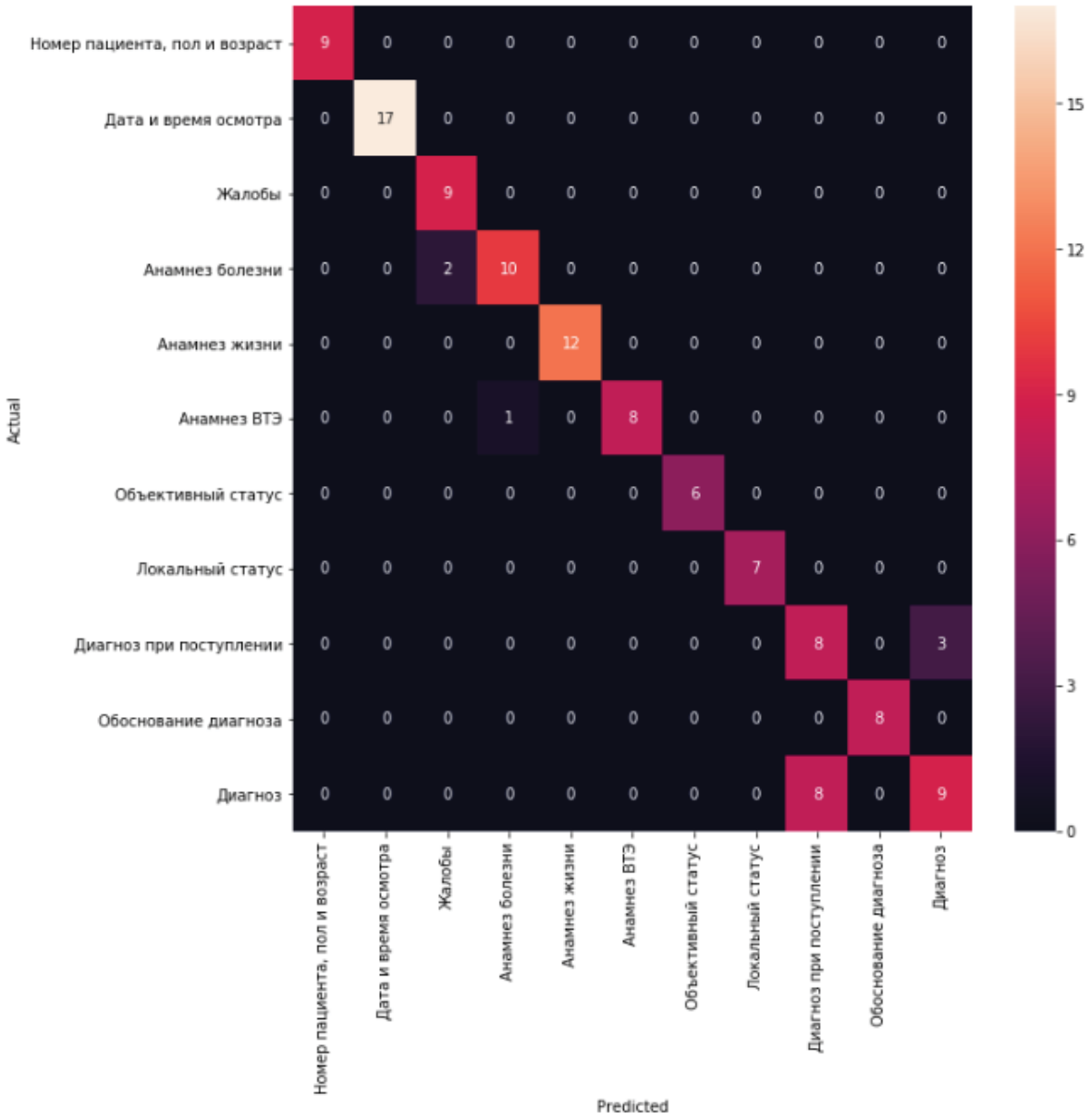

Рисунок 5 – Матрица ошибок

Как видно из матрицы ошибок, в модель правильно классифицирует большую часть примеров из тестовой выборки, однако всё же ошибается в классификации некоторых. Так, например, два фрагмента, являющийся анамнезом болезни, были классифицированы моделью как жалобы.

Для получения более точного представления о точности классификатора, обратимся к отчёту классификатора по тестированию модели, представленному наРисунок 6.

	precision	recall	f1-score	support
Номер пациента, пол и возраст	1.00	1.00	1.00	9
Дата и время осмотра	1.00	1.00	1.00	17
Жалобы	0.82	1.00	0.90	9
Анамнез болезни	0.91	0.83	0.87	12
Анамнез жизни	1.00	1.00	1.00	12
Анамнез ВТЭ	1.00	0.89	0.94	9
Объективный статус	1.00	1.00	1.00	6
Локальный статус	1.00	1.00	1.00	7
Диагноз при поступлении	0.50	0.73	0.59	11
Обоснование диагноза	1.00	1.00	1.00	8
Диагноз	0.75	0.53	0.62	17
micro avg	0.88	0.88	0.88	117
macro avg	0.91	0.91	0.90	117
weighted avg	0.89	0.88	0.88	117

Рисунок 6 – Отчет классификатора о тестировании

Для того чтобы иметь представление о причинах ошибок в классификации, рассмотрим более подробный отчёт об ошибках модели, представленный на Рисунок 7.

'Анамнез ВТЭ' predicted as 'Анамнез болезни' : 1 examples.

	name	text
346	Анамнез ВТЭ	Лист трудоспособности выдан амбулаторно № 265...

'Диагноз' predicted as 'Диагноз при поступлении' : 8 examples.

	name	text
208	Диагноз	A46 рожа левого плеча и предплечья, эритемпто...
263	Диагноз	A46 Рожа правого плеча и предплечья, распрост...
230	Диагноз	A46 Рожа правой голени, эритематозная форма, ...
65	Диагноз	A46 Рожа лица и ушной раковины слева, эритема...
120	Диагноз	A46 Рожа лица, эритематозная форма, локализир...
142	Диагноз	A46 Рожа лица, эритематозная форма, первичная...
186	Диагноз	A46 Рожа левой голени, эритематозная форма, л...
274	Диагноз	A46 Рожа правой стопы эритематозно-геморрагич...

'Диагноз при поступлении' predicted as 'Диагноз' : 3 examples.

	name	text
294	Диагноз при поступлении	A46 Рожа лица и волостистой части головы, эри...
272	Диагноз при поступлении	A46 Рожа
173	Диагноз при поступлении	A46 Рожа, рожа носа, эритематозная, первичная...

Рисунок 7 – Ошибки классификатора

Рассмотрим ошибки модели более подробно. Действительно, модель классифицировала два фрагмент категории «Анамнез болезни» как «Жалобы», ведь в этих категориях происходит описание конкретной проблемы пациента. Восемь фрагментов категории «Диагноз» были классифицированы как «Диагноз при поступлении», и три наоборот. Эти ошибки объясняются тем, что зачастую при более тщательном осмотре пациента диагноз при поступлении подтверждается и врач копирует запись из одной категории в другую.

4.2.1. Выявление признаков, значимых для классификатора

Благодаря построенному классификатору имеется возможность оценки значимости признаков для конкретного классификатора. Рассмотрим эти признаки на основе значений поля *coef_* объекта модели и выведем по два значимых слова и словосочетания для каждого блока.

```
# 'Анамнез ВТЭ':
  . Top unigrams:
       . инвалидность
       . нуждается
  . Top bigrams:
       . поступления стационар
       . момент поступления
# 'Анамнез болезни':
  . Top unigrams:
       . клинику
       . заболела
  . Top bigrams:
       . инфекционную клинику
       . 04 17
# 'Анамнез жизни':
  . Top unigrams:
       . отрицает
       . детстве
  . Top bigrams:
       . алкоголь употребляет
       . курит алкоголь
# 'Дата и время осмотра':
  . Top unigrams:
       . 2016
       . 2017
  . Top bigrams:
       . 10 2016
       . 2016 время
# 'Диагноз':
  . Top unigrams:
       . a46
       . рожа
  . Top bigrams:
       . рожа распространенная
       . первичная тяжелои
# 'Диагноз при поступлении':
  . Top unigrams:
       . локализованная
       . распространенная
  . Top bigrams:
       . a46 рожа
       . локализованная первичная
# 'Жалобы':
  . Top unigrams:
       . слабость
       . боль
  . Top bigrams:
       . повышение температуры
       . гиперемия кожи
# 'Локальный статус':
  . Top unigrams:
       . эритема
       . горячая
  . Top bigrams:
       . четкими границами
       . горячая ощупь
# 'Номер пациента, пол и возраст':
  . Top unigrams:
       . женскии
       . мужскои
  . Top bigrams:
       . рожа мужскои
       . рожа женскии
```

```
# 'Обоснование диагноза':
  . Top unigrams:
       . интоксикации
       . начало
  . Top bigrams:
       . острое начало
       . диагноз рожа
```

```
# 'Объективный статус':
  . Top unigrams:
       . состояние
       . удовлетворительное
  . Top bigrams:
       . состояние удовлетворитель
ное
       . сознание полное
```

Как видно из сравнения списка значимых признаков согласно критерию хи-квадрат и значимых признаков при построении модели, данные списки признаков практически идентичны, что дает основания считать выбор значимых признаков по критерию хи-квадрат уместным.

5. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

5.1. Предпроектный анализ

В рамках данной работы разрабатывается алгоритм классификации фрагментов истории болезни по существующим блокам документа «Осмотр в стационаре при поступлении» из истории болезни. Данный алгоритм может быть использован в системах поддержки принятия решения врача. Потенциальными потребителями таких систем являются предприятия здравоохранения. Такие системы помогут врачу быстрее и, иногда, более корректно принять решение о способе лечения данного пациента на основе схожих случаев.

Потенциальными потребителями результатов научно-исследовательского проекта помимо медицинского персонала можно считать также представителей научного сообщества, заинтересованных в продолжении исследования или в использовании его технических результатов.

Для эффективного использования научного потенциала научно-исследовательского проекта необходимо прилагать усилия не только к непосредственно её разработке, но и к проведению её анализа с точки зрения экономических требований.

Задачами раздела являются:

- Определить перспективности с помощью технологии Quad;
- Проанализировать факторы и определить связанные с ними проблемы, используя диаграмму Исикавы;
- Оценить готовности научно-исследовательского проекта к коммерциализации;
- Определить цель и результат научно-исследовательского проекта, составить рабочую группу научно-исследовательского проекта;

– Спланировать работу, распределить задачи между участниками научно-исследовательского проекта и определить трудоемкость работ для каждого исполнителя;

– Сформировать бюджет научно-исследовательского проекта;

– Провести анализ рисков.

5.1.1. Технология Quad

В рамках данной работы, анализ конкурентных технических решений невозможен ввиду отсутствия открытых данных о наличии и свойствах подобных решений. Обычно, подобные программные решения разрабатываются для нужд конкретного медицинского учреждения и не распространены за его рамками.

Технология QuaD представляет собой гибкий инструмент измерения характеристик, описывающих качество новой разработки и её перспективность на рынке и позволяющие принимать решение целесообразности вложения денежных средств в научно-исследовательский проект. В основе технологии лежит нахождение средневзвешенной величины показателей, как представлено в Таблица 2.

Таблица 2 – Оценочная карта технологии QuaD

Критерий оценки	**Вес критерия**	**Баллы**	**Максимальный балл**	**Относительное значение**	**Средневзвешенное значение**
Энергоэффективность	0,15	100	100	1	0,15
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	0,05	0	100	0	0
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

Критерий оценки	Вес критерия	Баллы	Максимальный балл	Относительное значение	Средневзвешенное значение
Прозрачность кода	0,05	100	100	1	0,05
Итого	**1**				**0,8725**

Значение качества по технологии QuaD равное 0,8725 говорит о том, что такая разработка считается перспективной, так как у рассматриваемой разработки высокие показатели по всем наиболее важным критериям, таким как надежность, функциональная мощность и т.д.

5.1.2. Диаграмма Исикавы

Диаграмма причины-следствия Исикавы – это графический метод анализа и формирования причинно-следственных связей, инструментальное средство для систематического определения причин проблемы и последующего графического представления.

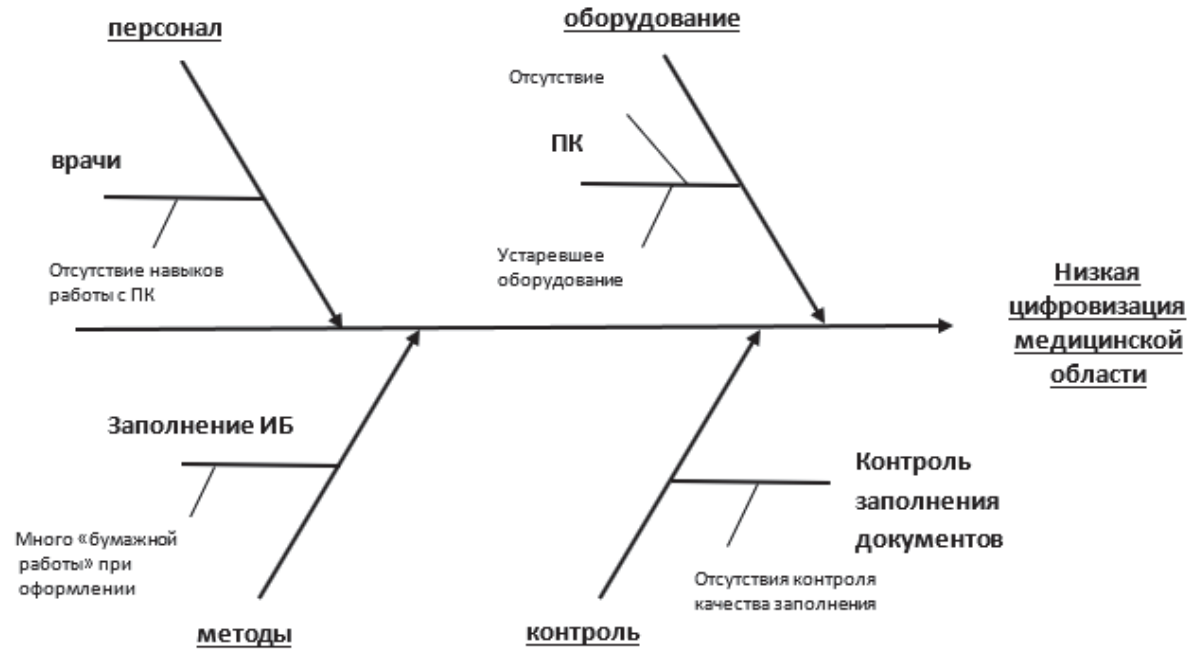

Рисунок 8 – Диаграмма Исикавы

Из диаграммы видны возможные причины низкой цифровизации медицинской области, начиная от устаревшего или полностью отсутствующего оборудования, заканчивая человеческими факторами такими как отсутствие навыков работы с ПК.

5.1.3. Оценка готовности научно-исследовательского проекта к коммерциализации

На любой стадии научной разработки полезно оценить степень её готовности к коммерциализации, а также оценить уровень собственных знаний для её проведения. По результатам такого анализа можно сделать вывод о готовности научно-исследовательского проекта к коммерциализации, а также о возможности привлечения иных специалистов в команду проекта.

Таблица 3 – Бланк оценки степени готовности научно-исследовательского проекта к коммерциализации

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

№ п/п	Наименование	Степень проработанности научно-исследовательского проекта	Уровень имеющихся знаний у разработчика
11	Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок	1	1
12	Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот	1	1
13	Проработаны вопросы финансирования коммерциализации научной разработки	1	1
14	Имеется команда для коммерциализации научной разработки	1	1
15	Проработан механизм реализации научно-исследовательского проекта	3	3
	ИТОГО БАЛЛОВ	35	35

Итоговый балл $Б_{сум}$ по каждому направлению равен 35 баллам. Если значение $Б_{сум}$ лежит в интервале от 30 до 44 баллов, то перспективность такой разработки считается средней. Отсюда следует вывод о том, что для дальнейшей коммерциализации научно-исследовательского проекта необходимо вовлечение сторонних специалистов в этой области.

Успех продвижения товара на рынок во многом зависит от правильности выбора метода коммерциализации. Для данного научно-исследовательского проекта наиболее приемлемым методом коммерциализации является инжиниринг, предполагающий предоставление заказчику на основе договора комплекса или отдельных видов инженерно-технических услуг. Так как возможно в процессе использования программного продукта потребуется его доработка и адаптация к конкретной

среде использования вплоть до внедрения в производство, а данный метод коммерциализации позволяет это.

5.2. Инициация научно-исследовательского проекта

Группа процессов инициации состоит из процессов, направленных на определение нового научно-исследовательского проекта. В рамках процессов инициации определяются цели, а также заинтересованные стороны проекта.

5.2.1. Цели и результат научно-исследовательского проекта

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

Таблица 4 содержит информацию о заинтересованных сторонах научно-исследовательского проекта.

Таблица 4 – Заинтересованные стороны научно-исследовательского проекта

Заинтересованные стороны научно-исследовательского проекта	**Ожидания заинтересованных сторон**
Отделение информационных технологий ТПУ	Научные публикации Защита магистерской диссертации
Медицинский персонал	Уменьшение времени работы с медицинской документацией, благодаря автоматизации процесса классификации фрагментов истории болезни
Пациенты	Получение более качественных консультаций врачей, благодаря уменьшению работы врача с документацией
Научное сообщество	Алгоритм извлечения важных текстовых признаков

В Таблица 5 представлена информация о иерархии целей научно-исследовательского проекта и критериев их достижения.

Таблица 5 – Иерархия целей научно-исследовательского проекта и критерии их достижения

Цели проекта:	Разработать алгоритмическое и программное обеспечение выделения значимых предикторов из медицинской документации осмотра пациента
Ожидаемые результаты проекта:	Алгоритмическое и программное обеспечение выделения значимых предикторов из медицинской документации осмотра пациента
Требования к результату проекта:	Выбраны значимые признаки
	Построен классификатор блоков текста, принадлежащих медицинской документации
	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 Таблица 6 определены участники научно-исследовательского проекта и их роли.

Таблица 6 – Рабочая группа научно-исследовательского проекта

№ п/п	ФИО, должность	Роль в проекте	Функции
1	Аксёнов Сергей Владимирович	Научный руководитель	Составление научных задач, контроль выполнения проекта, проверка разработки, проверка документации
2	Демченко Ирина Сергеевна	Инженер	Проектирование, реализация

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.

5.3. Планирование управления научно-исследовательским проектом

Планирование комплекса предполагаемых работ осуществляется в следующем порядке:

- определение структуры работ в рамках научно-исследовательского проекта;
- определение участников каждой работы;
- установление продолжительности работ;
- построение графика проведения научно-исследовательского проекта.

Для выполнения научно-исследовательского проекта формируется рабочая группа, в состав которой могут входить научные сотрудники и преподаватели, инженеры, техники и лаборанты, численность групп может варьироваться. По каждому виду запланированных работ устанавливается соответствующая должность исполнителей.

В данном разделе составлен перечень этапов и работ, распределение исполнителей по данным видам работ.

5.3.1. План научно-исследовательского проекта

В рамках планирования научно-исследовательского проекта построим календарный и линейный графики. Календарный план проекта представлен в Таблица 7.

Таблица 7 – Календарный план проекта

Код работы (из ИСР)	**Название**	**Длительность, дни**	**Дата начала работ**	**Дата окончания работ**	**Состав участников (ФИО ответственных исполнителей)**
1	Выбор научного руководителя	7	15.02.19	21.02.19	Демченко И.С.
2	Составление и утверждение темы	4	22.02.19	25.02.19	Аксёнов С.В.
3	Составление календарного плана-графика выполнения работы	3	26.02.19	28.02.19	Аксёнов С.В., Демченко И.С.
4	Подбор и изучение литературы по теме магистерской работы	14	01.03.19	14.03.19	Демченко И.С.
5	Анализ предметной области	7	15.03.19	21.03.19	Демченко И.С.
6	Предварительная обработка и оцифровка исходных данных	17	22.03.19	07.04.19	Демченко И.С.
7	Выявление значимых признаков	10	08.04.19	17.04.19	Демченко И.С.
8	Построение классификатора	18	18.04.19	05.05.19	Демченко И.С.
9	Согласование выполненной работы с научным руководителем	7	06.05.19	12.05.19	Аксёнов С.В., Демченко И.С.
10	Выполнение других частей работы (финансовый менеджмент, социальная ответственность)	14	13.05.19	26.05.19	Демченко И.С.
11	Подведение итогов, оформление работы	5	27.05.19	31.05.19	Аксёнов С.В., Демченко И.С.
ИТОГО:		**106**	**15.02.19**	**31.05.19**	

5.3.1.1. Разработка графика проведения научно-исследовательского проекта

Наиболее удобным и наглядным способом отслеживания выполнения проектной работы является диаграмма Ганта.

Диаграмма Ганта – горизонтальный ленточный график, на котором работы по теме представляются протяженными во времени отрезками, характеризующимися датами начала и окончания выполнения данных работ.

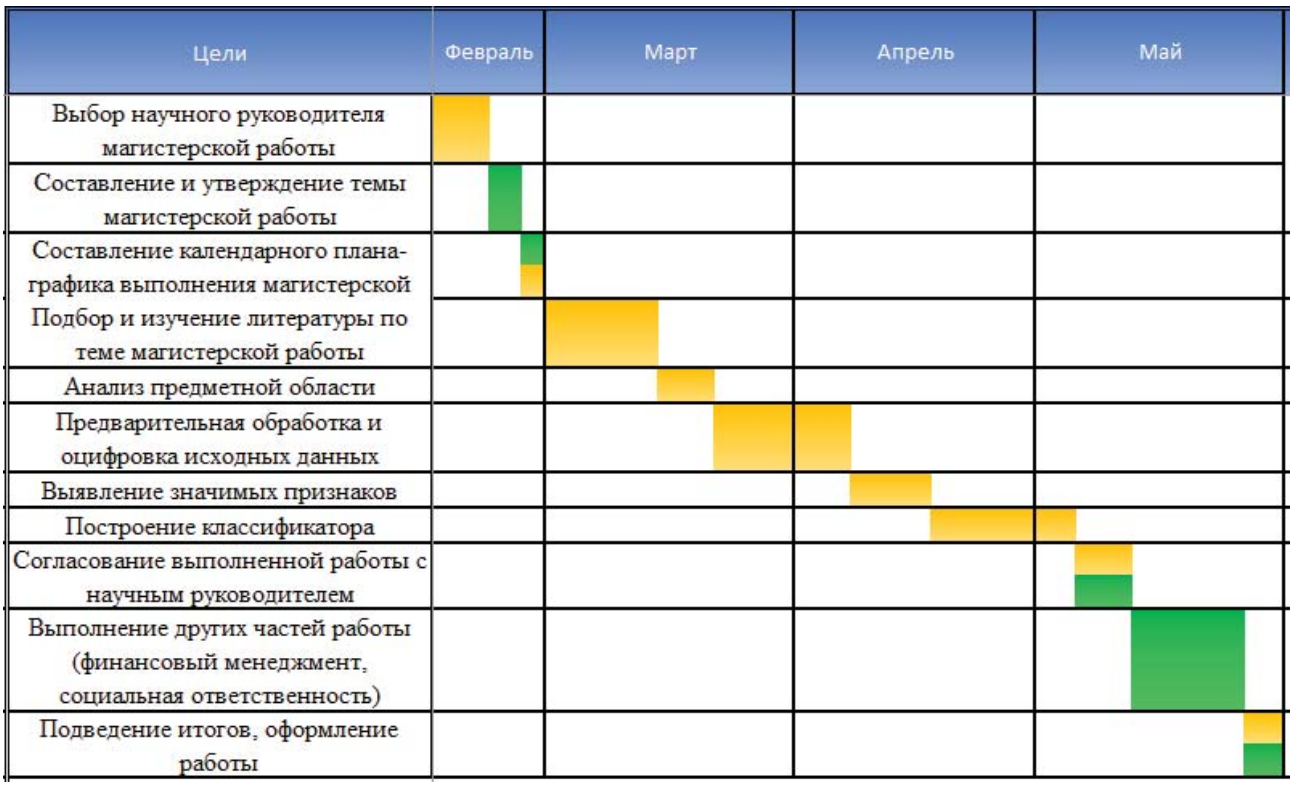

Рисунок 9 – Диаграмма Ганта

Из диаграммы Ганта наглядно видны границы этапов научно-исследовательского проекта. Длительность его выполнения 3,5 месяца.

5.3.2. Бюджет научно-исследовательского проекта

При планировании бюджета научно-исследовательского проекта должно быть обеспечено полное и достоверное отражение всех видов расходов, связанных с его выполнением. В процессе формирования бюджета научно-исследовательского проекта используется следующая группировка затрат по статьям:

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

5.3.2.1. Расчет материальных затрат

Материальные затраты на проведение данного научно-исследовательского проекта состоят только из затрат на оформление документации, в том числе канцелярских принадлежностей, на общую сумму 2000 руб.

5.3.2.2. Амортизационные затраты

В ходе работы использовался ПК, первоначальная стоимость которого составляет 47000 руб. Срок полезного использования ПК составляет 3 года. Будем использовать ПК в течение 4 месяцев для написания ВКР, согласно плану работ. Тогда:

- норма амортизации:

$$\text{Ан} = \frac{1}{n} * 100\% = \frac{1}{3} \times 100\% = 33{,}33\%$$

- годовые амортизационные отчисления:

$$\text{Аг} = 47000 \times 0{,}33 = 15510 \text{ рублей}$$

- ежемесячные амортизационные отчисления:

$$\text{Ам} = \frac{15510}{12} = 1292{,}5 \text{ рублей}$$

- итоговая сумма амортизации основных средств:

$$\text{А} = 1292{,}5 \times 4 = 5170 \text{ рублей}$$

5.3.2.3. Заработная плата исполнителей

Рассчитаем заработную плату исполнителей научно-исследовательского проекта. Примем, что оклад инженера составляет 21760 руб., а оклад научного руководителя составляет 33664 руб. Таблица 8 содержит расчет основной заработной платы исполнителей.

Таблица 8 – Расчет заработной платы

Исполнители	**Здн, руб**	**Кпр**	**Кд**	**Кр**	**Тр**	**Зосн**	**Здоп**	**Зп**
Инженер	931,29	0,3	0,3	1,3	81	156903,74	20397,49	177301,23
Научный руководитель	1440,76	0,3	0,3	1,3	9	26971,03	3506,23	30477,26

Полная заработная плата складывается из основной и дополнительной. Дополнительная заработная плата исполнителей рассчитана как 13% от основной заработной платы.

5.3.2.4. Отчисления во внебюджетные фонды (страховые отчисления)

Размер страховых отчислений составляет 30% от суммы основной и дополнительной заработной платы исполнителя. Таким образом, рассчитаем сумму страховых отчислений в Таблица 9.

Таблица 9 – Расчет страховых отчислений

Исполнитель	**Зп**	**Страховые отчисления**
Инженер	177301,23	53190,37
Научный руководитель	30477,26	9143,18

5.3.2.5. Накладные расходы

Величина накладных расходов рассчитывается как 16% от суммы материальных затрат, затрат на специальное оборудование, основной заработной платы, дополнительной заработной платы и страховых взносов.

Таким образом, сумма накладных расходов равна

$$N = (2000 + 5170 + 177301{,}23 + 53190{,}37 + 30477{,}26 + 9143{,}18) \times 0{,}16 = 44365{,}13 \text{ руб}$$

5.3.2.6. Формирование бюджета затрат научно-исследовательского проекта

Таблица 10 содержит информацию о бюджете научно-исследовательского проекта.

Таблица 10 – Бюджет затрат

Наименование	**Сумма, руб.**	**%**
Материальные затраты	2000	0,62
Затраты на специальное оборудование	5170	1,61
Затраты на основную заработную плату	183874,77	57,17
Затраты на дополнительную	23903,72	7,43

Наименование	**Сумма, руб.**	**%**
заработную плату		
Страховые взносы	62333,55	19,38
Накладные расходы	44365,13	13,79
Общий бюджет	**321647,17**	**100%**

Таким образом, большая часть расходов по проекту относится к категории затрат на основную заработную плату исполнителей.

5.3.3. Риски научно-исследовательского проекта

При разработке научно-исследовательского проекта следует понимать и учитывать возможные риски. Таблица 11 содержит результаты анализа возможных рисков.

Таблица 11 – 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. Для коммерциализации научно-исследовательского проекта необходимо вовлечение сторонних специалистов в этой области. Основные работы в рамках данного научно-исследовательского проекта проводились в период с 15 февраля по 31 мая 2019 года. Команда проекта состоит из инженера и научного руководителя. Общий бюджет научно-исследовательского проекта составил 321647,47 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

Аннотация

Представление понятия «Социальная ответственность» сформулировано в международном стандарте (МС) IC CSR-08260008000: 2011 «Социальная ответственность организации».

В соответствии с МС - Социальная ответственность - ответственность организации за воздействие ее решений и деятельности на общество и окружающую среду через прозрачное и этичное поведение, которое:

- содействует устойчивому развитию, включая здоровье и благосостояние общества;
- учитывает ожидания заинтересованных сторон;
- соответствует применяемому законодательству и согласуется с международными нормами поведения (включая промышленную безопасность и условия труда, экологическую безопасность);
- интегрировано в деятельность всей организации и применяется во всех ее взаимоотношениях (включая промышленную безопасность и условия труда, экологическую безопасность).

Введение

Объект исследования – алгоритм классификации текстовых блоков из истории болезни.

Научно-исследовательская работа заключалась в разработке алгоритма классификации текстовых блоков из истории болезни. Работа выполнялась с использованием ЭВМ.

В разделе будут рассмотрены опасные и вредные факторы, оказывающие влияние на производственную деятельность технологического персонала, работающего с автоматизированной системой управления технологическим процессом, рассмотрены воздействия разрабатываемой системы на окружающую среду, правовые и организационные вопросы, а также мероприятия в чрезвычайных ситуациях.

6.1.Правовые и организационные вопросы обеспечения безопасности

6.1.1. Специальные правовые нормы трудового законодательства

Нормы трудового права – это правила трудовых отношений, установленные или санкционированные государством посредством законодательных актов.

Нормы трудового права регулируют любые отношения, связанные с использованием личного труда.

Формы их реализации разнообразны:

- собственно, трудовые отношения;
- организация труда и управление им;
- трудоустройство работников;
- социальное партнерство, коллективные отношения;
- содействие занятости безработных лиц;
- организация профессиональной подготовки и повышения квалификации;
- обеспечение мер по охране труда граждан;
- осуществление контроля и надзора за соблюдением законодательства;
- социальная и правовая защита работников, решение трудовых споров;
- деятельность профессиональных союзов;
- отношения взаимной материальной ответственности работника и работодателя;
- защита прав и интересов работодателей.

Рассмотрим регулирование коллективных отношений.

Настоящий коллективный договор является правовым актом, регулирующим социально-трудовые отношения работников АО «ЕВРАЗ ЗСМК» с работодателем.

Основной задачей коллективного договора является создание необходимых организационно-правовых условий для достижения оптимального согласования интересов сторон трудовых отношений.

По заключенному коллективному договору работодатель обязан:

- соблюдать трудовое законодательство и иные нормативные правовые акты, содержащие нормы трудового права, локальные нормативные акты, условия коллективного договора, соглашений и трудовых договоров;
- предоставлять работникам работу, обусловленную трудовым договором;
- обеспечивать безопасность и условия труда, соответствующие государственным нормативным требованиям охраны труда;
- обеспечивать работников оборудованием, инструментами, технической документацией и иными средствами, необходимыми для исполнения ими трудовых обязанностей;
- обеспечивать работникам равную оплату за труд равной ценности, постоянно совершенствовать организацию оплаты и стимулирования труда, обеспечить материальную заинтересованность работников в результатах их труда;
- выплачивать в полном размере причитающуюся работникам заработную плату в сроки, установленные в соответствии с ТК РФ, коллективным договором, настоящими Правилами, трудовыми договорами;
- вести коллективные переговоры, а также заключать коллективный договор в порядке, установленном ТК РФ;

- знакомить работников под роспись с принимаемыми локальными нормативными актами, непосредственно связанными с их трудовой деятельностью;
- создавать условия, обеспечивающие участие работников в управлении организацией в предусмотренных ТК РФ, иными федеральными законами и коллективным договором формах;
- осуществлять обязательное социальное страхование работников в порядке, установленном федеральными законами;
- возмещать вред, причиненный работникам в связи с исполнением ими трудовых обязанностей, а также компенсировать моральный вред в порядке и на условиях, которые установлены ТК РФ, федеральными законами и иными нормативными правовыми актами РФ;
- принимать необходимые меры по профилактике производственного травматизма, профессиональных или других заболеваний работников, своевременно предоставлять льготы и компенсации в связи с вредными (опасными, тяжелыми) условиями труда (сокращенный рабочий день, дополнительные отпуска и др.), обеспечивать в соответствии с действующими нормами и положениями специальной одеждой и обувью, другими средствами индивидуальной защиты;
- постоянно контролировать знание и соблюдение работниками всех требований инструкций по охране труда, производственной санитарии и гигиене труда, противопожарной безопасности;

Работодатель обязуется проводить аттестацию и сертификацию рабочих мест один раз в пять лет с участием представителя профкома.

Если по результатам аттестации рабочее место не соответствует санитарно-гигиеническим требованиям и признано условно аттестованным, разрабатывать совместно с профкомом план мероприятий по улучшению и

оздоровлению условий труда на данном рабочем месте и обеспечивать их выполнение.

Ежегодно издавать приказ о мероприятиях по охране труда и промышленной безопасности, считать эти мероприятия соглашением по охране труда на год.

Обеспечивать за счет средств работодателя:

- Проведение инструктажей по охране труда, обучение лиц, поступающих на работу с вредными и (или) опасными условиями труда, безопасным методам и приемам выполнения работ со стажировкой на рабочем месте и сдачей экзаменов, проведение периодического обучения по охране труда и проверку знаний требований охраны труда в период работы.
- Проведение обязательных периодических медицинских осмотров (обследований) работников, в том числе женщин в женской консультации, в рабочее время по графику медицинских осмотров, с сохранением за ними места работы (должности) и среднего заработка на время прохождения указанных медицинских осмотров.
- Наличие на производственных участках аптечек для оказания первой помощи пострадавшим и обработки микротравм; наличие в аптечках рекомендованного МЛПУ «Городская клиническая больница №1» перечня средств и медикаментов, их ежегодную замену.
- Выдачу молока работникам Общества в дни фактического выполнения работ, в том числе при выполнении работ временными ремонтными бригадами на местах с наличием вредных факторов в соответствии с медицинскими показаниями в количестве:
  - при длительности смены до 8 часов – 0,5 л (1 талон);

- при длительности смены 11,5 часов – 0,75 л (3 талона на две смены).
- На горячих участках и участках с вредными условиями труда обеспечивать работников сухим чаем из расчета 8 грамм на одного человека в смену. Списки работников, которым необходимо выдавать чай, утверждаются совместным постановлением работодателя и профкома.
- На работах, связанных с загрязнением, выдавать бесплатно банное мыло по норме 400 грамм на одного человека в месяц.
- Выдачу работникам защитных паст в дни работы на основании перечня, утвержденного совместным постановлением работодателя и профкома.
- Бесплатную выдачу витаминных препаратов работникам, подвергающимся воздействию высокой температуры окружающей среды и интенсивному теплооблучению при выполнении работ с особо вредными условиями труда в соответствии со списками, утвержденными совместным постановлением работодателя и профкома.
- Дополнительное страхование работников от несчастных случаев на производстве.

Порядок обеспечения работников спецодеждой, спецобувью и средствами индивидуальной защиты, стирки и дезинфекции устанавливается локальными нормативными актами работодателя, принимаемыми по согласованию с профкомом.

Перечень изменений и дополнений к нормативам, утвержденным законодательством РФ выдачи спецодежды, спецобуви и средств индивидуальной защиты определяется приложением к коллективному договору.

6.1.2. Организационные мероприятия при компоновке рабочей зоны

6.1.2.1. Эргономические требования к рабочему месту оператора ПЭВМ

Проектирование рабочих мест, снабженных видеотерминалами, относится к числу важных проблем эргономического проектирования в области вычислительной техники.

Организация рабочего места программиста или оператора регламентируется следующими нормативными документами:

ГОСТ 12.2.032-78 ССБТ, ГОСТ 12.2.033-78 ССБТ, СанПиН 2.2.2/2.4.1340-03 и рядом других.

Эргономическими аспектами проектирования видеотерминальных рабочих мест, в частности, являются: высота рабочей поверхности, размеры пространства для ног, требования к расположению документов на рабочем месте (наличие и размеры подставки для документов, возможность различного размещения документов, расстояние от глаз пользователя до экрана, документа, клавиатуры и т.д.), характеристики рабочего кресла, требования к поверхности рабочего стола, регулируемость элементов рабочего места.

Главными элементами рабочего места программиста или оператора являются стол и кресло. Основным рабочим положением является положение сидя.

Рациональная планировка рабочего места предусматривает четкий порядок и постоянство размещения предметов, средств труда и документации. То, что требуется для выполнения работ чаще, расположено в зоне легкой досягаемости рабочего пространства.

Моторное поле - пространство рабочего места, в котором могут осуществляться двигательные действия человека.

Максимальная зона досягаемости рук – это часть моторного поля рабочего места, ограниченного дугами, описываемыми максимально вытянутыми руками при движении их в плечевом суставе.

Оптимальная зона – часть моторного поля рабочего места, ограниченного дугами, описываемыми предплечьями при движении в локтевых суставах с опорой в точке локтя и с относительно неподвижным плечом.

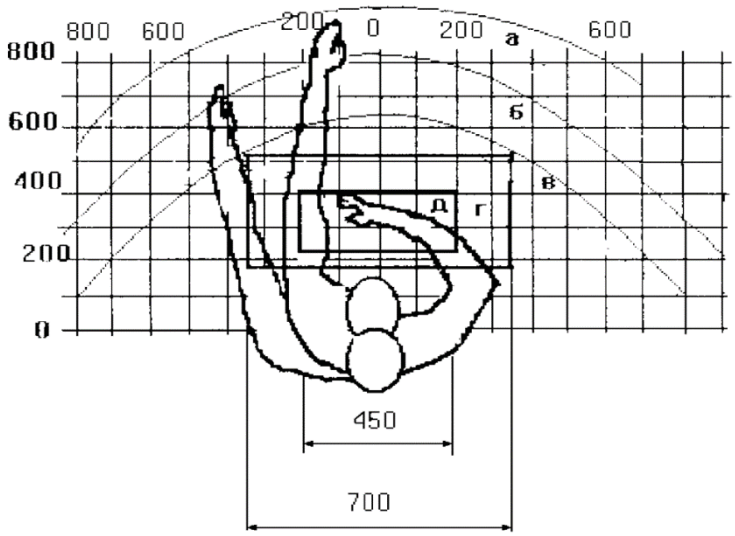

Рисунок 10- Зоны досягаемости рук в горизонтальной плоскости

а - зона максимальной досягаемости;

б - зона досягаемости пальцев при вытянутой руке;

в - зона легкой досягаемости ладони;

г - оптимальное пространство для грубой ручной работы;

д - оптимальное пространство для тонкой ручной работы.

Оптимальное размещение предметов труда и документации в зонах досягаемости:

- дисплей размещается в зоне "а" (в центре);
- системный блок размещается в предусмотренной нише стола;
- клавиатура - в зоне "г"/"д";
- манипулятор "мышь" - в зоне "в" справа;

- документация: необходимая при работе - в зоне легкой досягаемости ладони – "в", а в выдвижных ящиках стола - литература, неиспользуемая постоянно.

Для комфортной работы стол должен удовлетворять следующим условиям:

- высота стола должна быть выбрана с учетом возможности сидеть свободно, в удобной позе, при необходимости опираясь на подлокотники;
- нижняя часть стола должна быть сконструирована так, чтобы программист мог удобно сидеть, не был вынужден поджимать ноги;
- поверхность стола должна обладать свойствами, исключающими появление бликов в поле зрения программиста;
- конструкция стола должна предусматривать наличие выдвижных ящиков (не менее 3 для хранения документации, листингов, канцелярских принадлежностей).
- высота рабочей поверхности рекомендуется в пределах 680-760 мм. Высота поверхности, на которую устанавливается клавиатура, должна быть около 650 мм.

Большое значение придается характеристикам рабочего стула (кресла). Рабочий стул (кресло) должен быть подъемно-поворотным и регулируемым по высоте и углам наклона сиденья и спинки, а также регулируемым по расстоянию спинки от переднего края сиденья. Конструкция стула должна обеспечивать:

- ширину и глубину поверхности сиденья не менее 400 мм;
- поверхность сиденья с закругленным передним краем;
- регулировку высоты поверхности сиденья в пределах 400 – 550 мм и углов наклона вперед до 15° и назад до 5°;

- высоту опорной поверхности спинки 300 ± 20 мм, ширину - не менее 380 мм и радиус кривизны горизонтальной плоскости - 400 мм;
- угол наклона спинки в вертикальной плоскости в пределах 0 ± 30°;
- регулировку расстояния спинки от переднего края сиденья в пределах 260-400 мм;
- стационарные или съемные подлокотники длиной не менее 250 мм и шириной - 50-70 мм;
- регулировку подлокотников по высоте над сиденьем в пределах 230 ± 30 мм и внутреннего расстояния между подлокотниками в пределах 350 - 500 мм.

Поверхность сиденья, спинки и других элементов стула (кресла) должна быть полумягкой с нескользящим, неэлектризующимся и воздухопроницаемым покрытием, обеспечивающим легкую очистку от загрязнения.

Кресло следует устанавливать на такой высоте, чтобы не чувствовалось давления на копчик (это может быть при низком расположении кресла) или на бедра (при слишком высоком).

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

Рабочее место должно быть оборудовано подставкой для ног, имеющей ширину не менее 300 мм, глубину не менее 400 мм, регулировку по высоте в пределах до 150 мм и по углу наклона опорной поверхности подставки до 20°. Поверхность подставки должна быть рифленой и иметь по переднему краю бортик высотой 10 мм.

Необходимо предусматривать при проектировании возможность различного размещения документов: сбоку от видеотерминала, между монитором и клавиатурой и т.п. Кроме того, в случаях, когда видеотерминал имеет низкое качество изображения, например, заметны мелькания, расстояние от глаз до экрана делают больше (около 700 мм), чем расстояние от глаза до документа (300 - 450 мм). Вообще при высоком качестве изображения на видеотерминале расстояние от глаз пользователя до экрана, документа и клавиатуры может быть равным.

Положение экрана определяется:

- расстоянием считывания (0,6–0,7 м);
- углом считывания, направлением взгляда на 20° ниже горизонтали к центру экрана, причем экран перпендикулярен этому направлению.

Должна также предусматриваться возможность регулирования экрана:

- по высоте +3 см;
- по наклону от -10° до +20° относительно вертикали;
- в левом и правом направлениях.

Большое значение также придается правильной рабочей позе пользователя. При неудобной рабочей позе могут появиться боли в мышцах, суставах и сухожилиях.

Требования к рабочей позе пользователя видеотерминала следующие:

- голова не должна быть наклонена более чем на 20°;
- плечи должны быть расслаблены;
- локти - под углом 80°–100°;
- предплечья и кисти рук - в горизонтальном положении.

Причина неправильной позы пользователей обусловлена следующими факторами:

- нет хорошей подставки для документов;

- клавиатура находится слишком высоко, а документы – низко;
- некуда положить руки и кисти;
- недостаточно пространство для ног.

Создание благоприятных условий труда и правильное эстетическое оформление рабочих мест на производстве имеет большое значение как для облегчения труда, так и для повышения его привлекательности, положительно влияющей на производительность труда.

6.2.Производственная безопасность

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

Согласно номенклатуре, опасные и вредные факторы по ГОСТ 12.0.003-74 [10] делятся на следующие группы:

- физические;
- химические;
- психофизиологические;
- биологические.

Перечень опасных и вредных факторов, влияющих на персонал в заданных условиях деятельности, представлен в Таблица 12.

Таблица 12 – Перечень опасных и вредных факторов технологии производства

Источник фактора, наименование видов работ	**Факторы**		**Нормативные документы**
	Вредные	**Опасные**	
• Работа с ПЭВМ; • Система отопления; • Система вентиляции; • Источник	Температура и влажность воздуха; Напряженность зрения; Напряженность труда в течение	Электрический ток.	Гигиенические требования к микроклимату производственных помещений СанПиН 2.2.4-548-96 [11]; Нормы естественного и

Источник фактора, наименование видов работ	Факторы		Нормативные документы
	Вредные	Опасные	
освещения.	смены; Естественное и искусственное освещение; Электромагнитные излучения; Повышенная или пониженная влажность воздуха; Повышенный уровень шума.		искусственного освещения предприятий, СНиП 23-05-95 [12]; Допустимые уровни шумов в производственных помещениях. ГОСТ 12.1.003-83. ССБТ [13]; Гигиенические требования к персональным электронно-вычислительным машинам и организации работы, СанПиН 2.2.2/2.4.1340-03 [14]; Защитное заземление, зануление, ГОСТ 12.1.030–81 ССБТ [15].

Эти факторы могут влиять на состояние здоровья, привести к травмоопасной или аварийной ситуации, поэтому следует установить эффективный контроль за соблюдением норм и требований, предъявленных к их параметрам.

6.2.2. Анализ вредных и опасных факторов, которые могут возникнуть на производстве при внедрении объекта исследования

В условиях современного интенсивного использования ЭВМ важное значение имеет изучение психофизиологических особенностей и возможностей человека с целью создания вычислительной техники, обеспечивающей максимальную производительность труда и сохранение

здоровья людей. Игнорирование эргономики может привести к довольно серьезным последствиям.

При внедрении усовершенствованной системы управления технологическим процессом важную роль играет планировка рабочего места. Она должна соответствовать правилам охраны труда и удовлетворять требованиям удобства выполнения работы, экономии энергии и времени оператора.

Основным документом, определяющим условия труда на персональных ЭВМ, являются «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы». Санитарные нормы и правила СанПиН 2.2.2/2.4.1340-03 [14], которые были введены 30 июня 2003 года.

В Правилах указаны основные требования к помещениям, микроклимату, шуму и вибрации, освещению помещений и рабочих мест, организации и оборудованию рабочих мест.

Основным опасным фактором является опасность поражения электрическим током. Исходя из анализа состояния помещения, учебных аудиторий кафедры Программной Инженерии НИ ТПУ, по степени опасности поражения электрическим током можно отнести к классу помещений без повышенной опасности (согласно ПУЭ).

Основным опасным производственным фактором на рабочем месте является высокое напряжение в сети, от которой запитана система управления.

6.2.3. Обоснование мероприятий по защите персонала предприятия от действия опасных и вредных факторов (техника безопасности и производственная санитария)

6.2.3.1. Требования к помещениям для работы с ПЭВМ

В соответствии с основными требованиями к помещениям для эксплуатации ПЭВМ (СанПиН 2.2.2/2.4.1340-03) эти помещения должны иметь естественное и искусственное освещение. Площадь на одно рабочее место пользователей ПЭВМ с ВДТ на базе электронно-лучевой трубки (ЭЛТ) должна составлять не менее 6 $м^2$ и с ВДТ на базе плоских дискретных экранов (жидкокристаллические, плазменные) 4,5 $м^2$.

Для внутренней отделки интерьера помещений с ПЭВМ должны использоваться диффузионно-отражающие материалы с коэффициентом отражения от потолка – 0,7 – 0,8; для стен – 0,5 – 0,6; для пола – 0,3 – 0,5.

6.2.3.2. Микроклимат

Значимым физическим фактором является микроклимат рабочей зоны (температура, влажность и скорость движения воздуха).

Температура, относительная влажность и скорость движения воздуха влияют на теплообмен и необходимо учитывать их комплексное воздействие. Нарушение теплообмена вызывает тепловую гипертермию, или перегрев.

Оптимальные нормы температуры, относительной влажности и скорости движения воздуха производственных помещений для работ, производимых сидя и не требующих систематического физического напряжения (категория Iа), приведены в

Таблица 13, в соответствии с СанПиН 2.2.2/2.4.1340-03 и СанПиН 2.2.4.548-96.

Таблица 13 – Нормы температуры, относительной влажности и скорости движения воздуха

Период года	Категория работы	Температура, С	Относительная влаж. воздуха, %	Скорость движения воздуха, не более м/с
Холодный	Ia	22-24	40-60	0,1
Теплый	Ia	23-25	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, техническим и экономически обоснованным причинам не могут быть обеспечены оптимальные величины.

Допустимые величины показателей микроклимата на рабочих местах представлены в Таблица 14.

Таблица 14 – Допустимые величины показателей микроклимата

Период года	Категория работы	Температура воздуха, °С	Относительная влаж. воздуха, %	Скорость движения воздуха, не более м/с
Холодный	Ia	20-25	15-75	0,1
Теплый	Ia	21-28	15-75	0,1-0,2

В рабочем помещении для выполнения данного научно-исследовательского проекта температурные замеры в холодный период года– февраль 2019 – колебались от 20 до 23 градусов, температура в теплое время года – апрель 2019 – от 23 до 25.

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 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. При недостаточной эффективности центрального отопления должны быть использованы масляные электрические нагреватели.

Радиаторы должны устанавливаться в нишах, прикрытых деревянными или металлическими решетками. Применение таких решеток способствует также повышению электробезопасности в помещениях. При этом температура на поверхности нагревательных приборов не должна превышать 95°С, чтобы исключить пригорание пыли.

6.2.3.3. Освещение

Освещение рабочего места – важнейший фактор создания нормальных условий труда. Освещению следует уделять особое внимание, так как при работе наибольшее напряжение получают глаза.

Освещение делится на естественное, искусственное и совмещенное. Совмещенное сочетает оба вида освещения.

На посту управления, где расположено рабочее место оператора, используется совмещенное освещение.

Для определения приемлемого уровня освещенности в помещении необходимо:

- определить требуемый для операторов уровень освещенности внешними источниками света;
- если требуемый уровень освещенности не приемлем для других операторов, работающих в данном помещении, надо найти способ сохранения требуемого контраста изображения другими средствами.

Рекомендуемые соотношения яркостей в поле зрения следующие:

- между рабочими поверхностями не должно превышать 1:3 – 1:5;

- между рабочими поверхностями и поверхностями стен и оборудования – 1:10.

Освещённость на рабочем месте должна соответствовать характеру зрительной работы, который определяется наименьшим размером объекта различения, контрастом объекта с фоном и характеристикой фона.

Рабочие столы следует размещать таким образом, чтобы видеодисплейные терминалы были ориентированы боковой стороной к световым проемам, чтобы естественный свет падал преимущественно слева.

Освещенность на поверхности стола в зоне размещения рабочего документа должна быть 300 - 500 лк (СНиП 23-05-95, СанПиН 2.2.2/2.4.1340-03). Освещение не должно создавать бликов на поверхности экрана. Освещенность поверхности экрана не должна быть более 300 лк. Следует ограничивать прямую блесткость от источников освещения, при этом яркость светящихся поверхностей (окна, светильники и др.), находящихся в поле зрения, должна быть не более 200 кд/м^2. Показатель ослепленности для источников общего искусственного освещения в производственных помещениях должен быть не более 20.

Согласно СниП 23-05-95 нормы на освещение для оператора поста управления берутся для производственных помещений. Эти нормы представлены в таблице 4.

Таблица 15 – Нормы на освещение для оператора

Характер зрительной работы	**Разряд зрительной работы**	**Подразряд зрительной работы**	**Искусственное освещение**		**Естественное освещение КЕО $e_н$, % при боковом**
			Освещенность при системе общего освещения, лк	**Коэффициент пульсации, $К_п$, %**	
Различение объектов высокой точности	Б	1	300	15	1,0

Расчет системы искусственного освещения на рабочем месте оператора поста управления

Расчет системы искусственного освещения проводится для прямоугольного помещения, размерами: длина A = 5 (м), ширина B = 7 (м), высота H = 4 (м), количество ламп N = 12 (шт).

Вычисления будут, производится по методу светового потока, предназначенного для расчета освещенности общего равномерного освещения горизонтальных поверхностей. Согласно отраслевым нормам освещенности уровень рабочей поверхности над полом составляет 0,8 (м) и установлена минимальная норма освещенности E = 300 (Лк).

Световой поток лампы накаливания или группы люминесцентных ламп светильника определяется по формуле:

$$Ф = E_н \cdot S \cdot K_з \cdot Z \cdot 100 / (n \cdot \eta), \quad (6)$$

Где: $E_н$ – нормируемая минимальная освещённость по СНиП 23-05-95, (Лк);

S – площадь освещаемого помещения, ($м^2$);

$K_з$ – коэффициент запаса, учитывающий загрязнение светильника (источника света, светотехнической арматуры, стен и пр., т.е. отражающих поверхностей), (наличие в атмосфере цеха дыма), пыли;

Z – коэффициент неравномерности освещения. Для люминесцентных ламп при расчётах берётся равным Z = 1,1;

n – число светильников;

η - коэффициент использования светового потока, (%);

Ф– световой поток, излучаемый светильником.

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 ($\rho_{ст}$) и потолка (ρ_n).

Индекс помещения определяется по формуле

$$i = \frac{S}{h \cdot (A + B)} \quad (7)$$

Коэффициенты отражения оцениваются субъективно.

Произведем расчет:

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

где h – расчетная высота подвеса светильников над рабочей поверхностью.

Экономичность осветительной установки зависит от отношения, представленного в формуле:

$$l = \frac{L}{h}, \quad (9)$$

где L – расстояние между рядами светильников, м.

Рекомендуется размещать люминесцентные лампы параллельными рядами, принимая l = 1,4, отсюда расстояние между рядами светильников:

$$L = l \cdot h = 1{,}4 \cdot 3{,}2 = 4{,}48 \text{ (м)} \quad (10)$$

Два ряда светильников будут расположены вдоль длинной стены помещения. Расстояние между двумя рядами светильников и стенами вычисляется по формуле:

$$Л = \frac{B - L}{4} = \frac{7 - 4{,}48}{4} = 0{,}63 \text{ (м)} \quad (11)$$

Определим индекс помещения вычисляя по формуле (7) получаем:

$$i = \frac{35}{3{,}2 \cdot 12} = 0{,}91.$$

Найдем коэффициенты отражения поверхностей стен, пола и потолка.

Так как поверхность стен окрашена в серый цвет, свежепобеленные с окнами без штор, то коэффициент отражения поверхности стен $Р_{ст}$ = 50%. Так как поверхность потолка светлый окрашенный, то коэффициент отражения поверхности потолка $Р_{п}$ = 30%.

Учитывая коэффициенты отражения поверхностей стен, потолка и индекс помещения i, определяем значение коэффициента η = 41%.

Подставив все значения в формулу (6), по которой рассчитывается световой поток одного источника света, получаем:

$$Ф = \frac{300 \cdot 35 \cdot 1{,}5 \cdot 1{,}1}{12 \cdot 0{,}41} = 3521 \text{ (лм)}$$

По полученному световому потоку подбираем лампу, наиболее подходящей является лампа LUNA 250 со световым потоком 3520 (лм).

Выразим E:

$$E = \frac{(F \cdot N \cdot \eta)}{(k)} = \frac{(3520 \cdot 12 \cdot 0{,}41)}{(1{,}5 \cdot 35 \cdot 1{,}1)} = 299.5 \text{ (лм)} \qquad (12)$$

Как видно из расчета, минимальная освещенность в пределах нормы.

Для того чтобы доказать, что использование люминесцентной лампы LUNA 250 является наиболее рациональным, рассчитаем необходимое количество светильников по формуле:

$$N = \frac{(E \cdot k \cdot S \cdot Z)}{(n \cdot \eta \cdot F)}, \qquad (13)$$

где E – норма освещенности E = 300 (Лк);

k – коэффициент запаса учитывающий старение ламп и загрязнение светильников, k = 1,5;

S – площадь помещения;

Z – коэффициент неравномерности освещения, Z = 1,1;

n – число рядов светильников, n = 4;

η – коэффициент использования светового потока, η = 0,41;

F – световой поток, излучаемый светильником.

Подставим численные значения в формулу (13), получим количество светильников в одном ряду:

$$N = \frac{(E \cdot k \cdot S \cdot Z)}{(n \cdot \eta \cdot F)} = \frac{300 \cdot 1{,}5 \cdot 35 \cdot 1{,}1}{0{,}41 \cdot 3520} = 3 \text{ (шт)}$$

Длина одного светильника равна 0,5 (м), в одном светильнике 4 лампы LUNA 250.

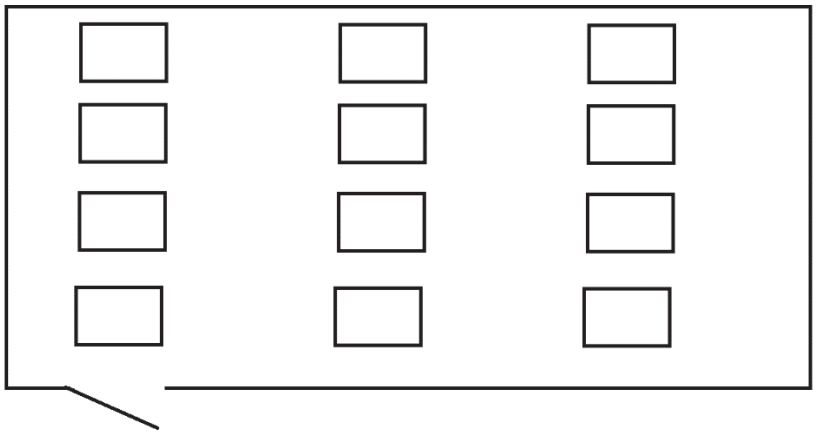

Рисунок 11 – Схема расположения ламп в аудитории КЦ-105 НИ ТПУ

Так как в рассматриваемом помещении количество ламп 12 (шт), по три светильника в четырех рядах, следовательно, нормы безопасности по искусственному освещению в данном случае соблюдены.

6.2.3.4. Шум

В производственных условиях имеют место шумы различной интенсивности и частотного спектра, которые генерируются источниками шумов.

Для исследуемого объекта (производство и пункт управления) основными источниками шумов являются производственное оборудование (внешние источники) и оборудование поста управления (внутренние источники).

ПДУ шума для объектов типа поста управления нормируются ГОСТ 12.1.003-83 [13] и СН 2.2.4/2.1.8.562–96 [16]. Значения ПДУ согласно этим документам представлены в Таблица 16. (для постоянных шумов)

Таблица 16 – Значения ПДУ для постоянных шумов

Рабочие места	Уровни звукового давления (ДБ) в октавных полосах со среднегеометрическими частотами, Гц								Уровни звука и эквивалентные уровни звука, дБА
	63	125	250	500	1000	2000	4000	8000	
ПУ	83	74	68	63	60	78	55	54	65

Для оценки соблюдения ПДУ шума необходим производственный контроль (измерения и оценка). В случае превышения уровней необходимы организационно- технические мероприятия по защите от действия шума (защита временем, расстоянием, экранирование источника, либо рабочей зоны, замена оборудования, использование СИЗ).

6.2.3.5. Электромагнитные излучения

Электромагнитным излучением называется излучение, прямо или косвенно вызывающее ионизацию среды. Контакт с электромагнитными излучениями представляет серьезную опасность для человека, по сравнению с другими вредными производственными факторами (повышенное зрительное напряжение, психологическая перегрузка, сохранение длительное время неизменной рабочей позы).

Когда все устройства персонального компьютера включены, в районе рабочего места программиста, формируется сложное по структуре электромагнитное поле. Реальную угрозу для пользователя компьютера представляют электромагнитные поля. Влияние их на организм человека не обходится без последствий. Исследования показали, что в организме человека под влиянием электромагнитного излучения монитора происходят значительные изменения гормонального состояния, специфические изменения биотоков головного мозга, изменение обмена веществ. Пыль, притягиваемая электростатическим полем монитора, иногда становится причиной дерматитов лица, обострения астматических симптомов, раздражения слизистых оболочек.

Для снижения воздействия электромагнитного излучения следует применять мониторы с пониженным уровнем излучения, также устанавливать защитные экраны, придерживаться регламентированного режима труда и отдыха, а также проводить регулярную гигиеническую уборку помещения.

Нормы электромагнитных полей, создаваемых ПЭВМ приведены в Таблица 17 и Таблица 18, в соответствии с СанПиН 2.2.2/2.4.1340-03 [14].

Таблица 17 – Временные допустимые ЭМП, создаваемых ПЭВМ

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

Таблица 18 – Временные допустимые уровни ЭМП, создаваемых ПЭВМ на рабочих местах

Наименование параметров		**ВДУ**
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
Напряженность электростатического поля		

Для оценки соблюдения уровней необходим производственный контроль (измерения). В случае превышения уровней необходимы организационно- технические мероприятия (защита временем, расстоянием, экранирование источника, либо рабочей зоны, замена оборудования, использование СИЗ).

6.2.3.6. Психофизиологические факторы

Наиболее эффективные средства предупреждения утомления при работе на производстве – это средства, нормализующие активную трудовую деятельность человека. На фоне нормального протекания производственных процессов одним из важных физиологических мероприятий против утомления является правильный режим труда и отдыха (СанПиН 2.2.2/2.4.1340-03 [14]).

Существуют следующие меры по снижению влияния монотонности:

- необходимо применять оптимальные режимы труда и отдыха в течение рабочего дня;
- соблюдать эстетичность производства.

Для уменьшения физических нагрузок организма во время работы рекомендуется использовать специальную мебель с возможностью регулировки под конкретные антропометрические данные, например, эргономичное кресло.

6.2.3.7. Электрический ток

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 (ГОСТ 12.1.030–81 ССБТ [15]);
- 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 (ГОСТ Р 12.1.019-2009 ССБТ [17]):

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

6.3. Экологическая безопасность

Рассмотрим загрязнения литосферы в результате исследовательской деятельности бытовым мусором, на примере люминесцентных ламп. Их эксплуатация требует осторожности и четкого выполнения инструкции по обращению с данным отходом (код отхода 35330100 13 01 1, класс опасности – 1[18]). В данной лампе содержится опасное вещество ртуть в газообразном состоянии. При не правильной утилизации, лампа может разбиться и пары ртути могут попасть в окружающую среду. Вдыхание паров ртути может привести к тяжелому повреждению здоровья.

При перегорании ртутьсодержащей лампы (выходе из строя) её замену осуществляет лицо, ответственное за сбор и хранение ламп (обученное по электробезопасности и правилам обращения с отходом). Отработанные люминесцентные лампы сдаются только на полигон токсичных отходов для захоронения. Запрещается сваливать отработанные люминесцентные лампы с мусором [18].

Бытовой мусор помещений организаций несортированный, образованный в результате деятельности работников предприятия (код отхода 91200400 01 00 4). Агрегатное состояние отхода твердое; основные компоненты: бумага и древесина, металлы, пластмассы и др [18]. Для сбора мусора рабочее место оснащается урной. При заполнении урны, мусор выносится в контейнер бытовых отходов. Предприятие заключает договор с коммунальным хозяйством по вывозу и размещению мусора на организованных свалках.

6.4. Безопасность в чрезвычайных ситуациях

6.4.1. Анализ вероятных ЧС, которые может инициировать объект исследований

Перечень возможных ЧС на объекте исследования может быть достаточно широк. Ограничиваясь местоположением объекта и условиями его эксплуатации, его можно представить следующим (ориентировочным) вариантом:

- наводнение;
- удар молнии;
- пожар на объекте;
- взрыв.

В этом разделе наиболее актуальным будет рассмотрение вида ЧС – пожар, определение категории помещения по пожаровзрывобезопасности в котором происходит управление технологическим процессом, то есть аудитория КЦ–105 НИ ТПУ и регламентирование мер противопожарной безопасности.

Рабочее место оператора поста управления, должно соответствовать требованиям ФЗ Технический регламент по ПБ и норм пожарной безопасности (НПБ 105-03) и удовлетворять требованиям по предотвращению и тушению пожара по ГОСТ 12.1.004-91 [19] и СНиП 21-01-97 [20].

По пожарной, взрывной, взрывопожарной опасности помещение относится к категории Д, т.е. к помещению, в котором находятся негорючие вещества и материалы в холодном состоянии.

Основным поражающим фактором пожара для помещений данной категории является наличие открытого огня и отравление ядовитыми продуктами сгорания оборудования.

6.4.2. Анализ причин, которые могут вызвать ЧС на производстве при внедрении объекта исследований

Пожар в помещении оператора может возникнуть вследствие причин неэлектрического и электрического характера.

К причинам неэлектрического характера относятся халатное и неосторожное обращение с огнем (курение, оставление без присмотра нагревательных приборов).

К причинам электрического характера относятся:

- короткое замыкание;
- перегрузка проводов;
- большое переходное сопротивление;
- искрение;
- статическое электричество.

Режим короткого замыкания – появление в результате резкого возрастания силы тока, электрических искр, частиц расплавленного металла, электрической дуги, открытого огня, воспламенившейся изоляции.

Причины возникновения короткого замыкания:

- ошибки при проектировании.
- старение изоляции.
- увлажнение изоляции.
- механические перегрузки.

Пожарная опасность при перегрузках – чрезмерное нагревание отдельных элементов, которое может происходить при ошибках проектирования в случае длительного прохождения тока, превышающего номинальное значение.

Пожарная опасность переходных сопротивлений – возможность воспламенения изоляции или других близлежащих горючих материалов от

тепла, возникающего в месте аварийного сопротивления (в переходных клеммах, переключателях и др.).

6.4.3. Обоснование мероприятий по предотвращению ЧС и разработка порядка действия в случае возникновения ЧС

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 НИ 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 по п. 3.14 ГОСТ Р 12.2.143-2002 [21].

Согласно Правилам пожарной безопасности, в Российской Федерации ППБ 01-2003 (п. 16) в зданиях и сооружениях (кроме жилых домов) при единовременном нахождении на этаже более 10 человек должны быть разработаны и на видных местах вывешены планы (схемы) эвакуации людей в случае пожара.

План эвакуации людей при пожаре из помещения, где расположена аудитория КЦ–105 НИ ТПУ, представлен на Рисунок 12.

Рисунок 12 - План эвакуации при пожаре

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

Заключение

В рамках данной работы решены поставленные задачи: выделены значимые предикторы для классификации, построен классификатор для распознавания фрагментов документа из истории болезни пациента.

Оценка значимости предикторов проведена по средствам критерия хи-квадрат. Сравнительный анализ предикторов, выбранных по хи-квадрат критерию, с выбранными по частотной оценке, показал недостаточную информативность частотного критерия. Поэтому представление текстовой информации в виде доступном для подачи на вход алгоритмам машинного обучения проведено при помощи метода *TF-IDF*, так как данный метод наделяет объекты свойствами полезными для дальнейшего построения классификатора. Метрика *TF-IDF* позволила учесть частоту значимых слов, при этом уменьшая вес широкоупотребительных слов.

Построен ряд моделей для распознавания фрагмента текста как объекта одного из разделов документа истории болезни. Из полученных моделей выявлена лучшая – метод опорных векторов. Для данной модели осуществлен исчерпывающий поиск по сетке значений параметров с целью определения оптимальных параметров классификатора для данной задачи. Проведена оценка эффективности построенного классификатора, по результатам которой на тестовой выборке имеем следующие результаты: точность классификатора (*precision*) равна 0,89, полнота (*recall*) – 0,88, f-мера (*f1-score*) – 0,88.

В результате проделанной работы для классификации фрагментов документов из истории болезни предлагается использование классификатора *LinearSVC* библиотеки *scikit-learn*. Использование разработанной модели поможет в построении эффективных медицинский информационных систем.

Разработаны разделы «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение», «Социальная ответственность», а также раздел на иностранном языке (английский) – «Data Analytics in Healthcare», размещенный в Приложении А.

Список публикаций и научных достижений

Участие в конференциях:

1. Диплом 1 степени, Международная научно-практическая конференция «Электронные средства и системы управления», г. Томск, 2018 г.;
2. Диплом за участие, Международная научно-техническая конференция студентов, аспирантов и молодых ученых "Научная сессия ТУСУР - 2019", г. Томск;
3. Диплом за участие, Международная научно-практическая конференция «Новая наука: история становления, современное состояние, перспективы развития», г. Казань, 2018 г.
4. Сертификат участника, Международная научно-практическая конференция студентов, аспирантов и молодых ученых «Молодёжь и современные информационные технологии», г. Томск, 2017 г.
5. Сертификат участника, Международная школа «Научный компьютинг, аналитика больших данных и технологии машинного обучения для мегасайнс проектов», г. Дубна, 2018 г.

Участие в конкурсах:

1. Диплом за 1 место, Хакатон «Digital Hack», г. Томск, 2017 г.
2. Диплом за участие, 2 тур конкурсного отбора Стипендиальной программы В. Потанина, г. Томск, 2018 г.
3. Диплом за участие, 2 тур конкурсного отбора Стипендиальной программы В. Потанина, г. Томск, 2019 г.

Премии, звания, стипендии:

1. Именная стипендия ПАО «Транснефть» (г. Москва) студентам ТПУ (с 1 июля 2018 г. по 30 июня 2019 г.);
2. Стипендия Правительства РФ (с 1 сентября 2018г. по 31 августа 2019г.)
3. Повышенная стипендия ТПУ (с 1 февраля 2019 г. по 30 июня 2019 г.)
4. Сертификат владения английским языком ТПУ 3 (C1 – Effectiveness), 2017 г.

Публикации:

1. Демченко, И. С. Modern big data preprocessing techniques [Электронный ресурс] / И. С. Демченко, науч. рук. Е. И. Губин // Новая наука: история становления, современное состояние, перспективы развития: сборник статей по итогам Международной научно-практической конференции. – 2018. – Ч. 1. – [С. 4-7]. – Заглавие с экрана. – Доступ по договору с организацией-держателем ресурса. Режим доступа: https://elibrary.ru/item.asp?id=32852211
2. Д.Д. Богданов, И.С. Демченко. Разработка программы голосового ввода в виде web-приложения для эффективного заполнения медицинских карточек пациентов. – Сборник научных трудов Международной конференции студентов, аспирантов и молодых ученых, 23-26 апреля 2019 г.
3. И.С. Демченко. Построение классификатора для распознавания фрагмента истории болезни. – Сборник научных трудов Международной научно-технической конференции студентов, аспирантов и молодых ученых «Научная сессия ТУСУР – 2019», г. Томск, 22-24 мая 2019 г.
4. Demchenko I.S., Inkhireeva T.A. «Gender recognition by voice»: Материалы XIV Международной научно-практической конференции «Электронные средства и системы управления» (28-30 ноября 2018 г.): в 2 ч. – Ч.2. –Томск: В-Спектр, 2018 – 314с.
5. Казакявичюс И.С., Гергет О.М. «Разработка системы поддержки принятия решения врача, реализующей помощь в выборе управляющего воздействия» // Молодежь и современные информационные технологии: сборник трудов XV Международной научно-практической конференции студентов, аспирантов и молодых ученых, Томск, 4-7 Декабря 2017. - Томск: ТПУ, 2018 - С. 400-401;

Список используемых источников

1. Кобринский Б.А. Системы поддержки принятия решений в здравоохранении и обучении (ФГУ «Московский НИИ педиатрии и детской хирургии Росмедтехнологий», ГОУ ВПО «Российский государственный медицинский университет Росздрава»)
2. Ervin Sejdic, Tiago H. Falk. Signal Processing and Machine Learning for Biomedical Big Data. CRC Press, – 2018.
3. Applied Health Analytics and Informatics Using SAS – Joseph M. Woodside
4. Machine Learning and AI for Healthcare. Big Data for Improved Health Outcomes
5. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality
6. Jones K. S. A statistical interpretation of term specificity and its application in retrieval (англ.) // Journal of Documentation : журнал. — MCB University: MCB University Press, 2004. — Vol. 60, no. 5. — P. 493-502. — ISSN 0022-0418.
7. Рашка С. Python и машинное обучение / пер. с англ. А.В. Логунова. – М.: ДМК Пресс, 2017. – 418 с.: ил.
8. Дронов, В.А. Программирование. — СПб.: БХВ-Петербург, 2006. — 706 с.: ил.
9. Dean Abbott. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. Wiley
10. ГОСТ 12.0.003-74. ССБТ. Опасные и вредные производственные факторы. Классификация.
11. СанПиН 2.2.4-548-96. Гигиенические требования к микроклимату производственных помещений.
12. СНиП 23-05-95. Естественное и искусственное освещение.
13. ГОСТ 12.1.003-83 ССБТ. Шум. Общие требования безопасности.

14. СанПиН 2.2.2/2.4.1340-03. Санитарно–эпидемиологические правила и нормативы «Гигиенические требования к персональным электронновычислительным машинам и организации работы».
15. ГОСТ 12.1.030–81 ССБТ. Защитное заземление, зануление.
16. СН 2.2.4/2.1.8.562–96. Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории застройки.
17. ГОСТ Р 12.1.019-2009 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты.
18. Федеральный классификационный каталог отходов [Электронный ресурс]. – 2013. – Режим доступа: http://www.ecoguild.ru/faq/fedwastecatalog.htm, свободный.
19. ГОСТ 12.1.004-91 ССБТ. Пожарная безопасность. Общие требования.
20. СНиП 21-01-97. Пожарная безопасность зданий и сооружений.
21. ГОСТ Р 12.2.143-2002 ССБТ. Системы фотолюминесцентные эвакуационные. Элементы систем. Классификация. Общие технические требования. Методы контроля.

Приложение А

(справочное)

Data Analytics in Healthcare

Студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Демченко Ирина Сергеевна		

Консультант ОИТ ИШИТР:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Соколова Вероника Валерьевна	к.т.н.		

Консультант – лингвист ОИЯ ШБИП:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Диденко Анастасия Владимировна	к.ф.н.		

Nowadays is called a data-rich era because massive amounts of data are being collected. One can have data both from social media and from specific sensors. It has been predicted that by 2020, each human will create 1.7 megabytes of data per second. At the same time, having so much data and not using it leads to the question, why do we still collecting and storing so much data? Obviously, we should use modern technologies not only for collecting and storing, but for extracting knowledge from data [1].

Medical Data Types

The field of health care is estimated to generate about 1/3 of world data today. There are two general types of data: structured and unstructured.

Structured data is similar to a machine language. Highly organized in its format, structured data facilitates simple, straightforward search and information retrieval operations. Structured data would typically be stored in a relational database for this purpose.

Unstructured data refers to everything else. Unstructured data does not have a predefined model or schema. Data that is unstructured has no identifiable structure within it, and this presents problems for querying and information retrieval. E-mails, text messages, Facebook posts, Twitter tweets, and other social media posts are good examples of unstructured data. Unstructured data examples in healthcare field are MRT images and clinical records. The Gartner report has indicated that 80% of data is in the form of unstructured data. It may not always be possible to transform unstructured data into a structured model, but analytics of unstructured data is improving with the use of data science and machine learning methods such as Natural Language Processing (NLP) assisting in the understanding and classification of sentiment.

There are eight type of sources of health care data:

- Electronic Health Records,
- Health Insurance Claims Data,
- Publically Supported Databases,

- Patient-Reported Outcomes Measures,
- Clinical Registries,
- Genetic Data,
- Sensor data,
- Social Media.

All the sources of health care data meet the four V's of big data: high volume, variety, velocity and veracity. There are some examples that illustrate potential benefits of big data science in health care today.

Potential benefits of Big Data in Healthcare

Precision medicine (PM) is a new approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person. The broad application of PM has recently been enabled by the creation of large-scale biologic databases, powerful methods of characterizing patients, and computational tools.

Pediction Disease Risk. The use of large historic data sets allows researchers to find out the underlying reasons of disease. Moreover, knowing these reasons and applying machine learning algorithms to clinical data one can build a model that will predict the probability of specific person having the disease. This kind of data science application provides an alternative to traditional research methods such as clinical trials.

Health insurance claims repositories are a robust source of data that can be used in retrospective observational research studies. Health insurance data is usually used for validation, but in the «Big data, little data, and care coordination for Medical beneficiaries with Medigap coverage» article authors add survey data and personal interviews and succeed in identifying patients that have a high propensity for successful care coordination.

The Affordable Care Act financially penalizes institutions exceeding set emergency department revisit rate. Therefore, it is important to prevent revisits, it can be done by understanding underlying factors of revisits. The study

«Understanding emergency department 72-hour revisits among Medicaid patients using electronic healthcare records» demonstrates significant differences in the patterns of Medicaid patient encounters that result in a 72-hour revisit to the emergency department versus those that do not. To reduce the revisits rate of those patients, hospitals may use the prediction model and schedule appointments to ensure patients health status.

Artificial Intelligence

Most of the examples mentioned above are in fact machine learning models, but are called an artificial intelligence (AI) in media. Many people are surprised to learn that AI is nothing new. These technologies have existed for decades and now are getting spread thanks to cheaper computing and data availability. AI is a subset of computer science that has origins in mathematics, logic, cognitive science (Figure 1).

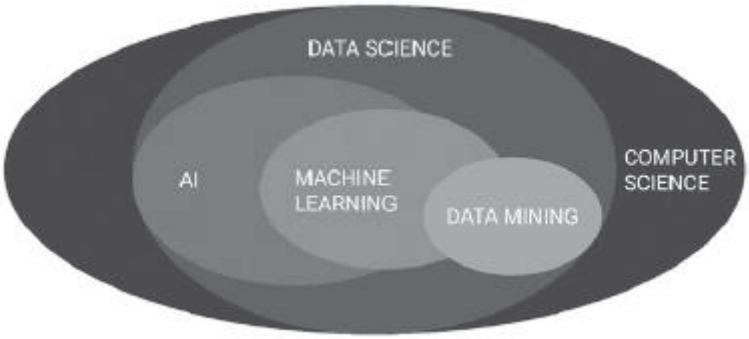

Figure 1 – AI, machine learning, and their place in computer science

The truth is that AI, as its core, is merely programming. The use and application of AI and machine learning in heath care enterprise systems are still relatively new. In the current ecosystem, with healthcare cost increasing, AI and machine learning algorithms can help to save a lot of money by using, for instance, patients' sensor data.

Both patients and healthcare professionals generate a huge amount of data. Blood pressure, geolocation, steps walked and other unstructured data is collected by smartphones today. Data could be extracted from paper documentation, or images scanned.

The ethics of AI are currently without guidelines, regulations. Many assume that AI has an objectivity that puts it above questions of morality, but in healthcare it is not the case. AI algorithms are only as fair and unbiased as the learnings, which come from the environmental data.

Applications of AI in Healthcare

Machine learning and AI are transforming the healthcare industry and improving outcomes, on the other hand, it is unlikely that AI agents will ever completely replace doctors. Machine learning is improving diagnostics, predicting outcomes, and beginning to scratch the surface of personalized care.

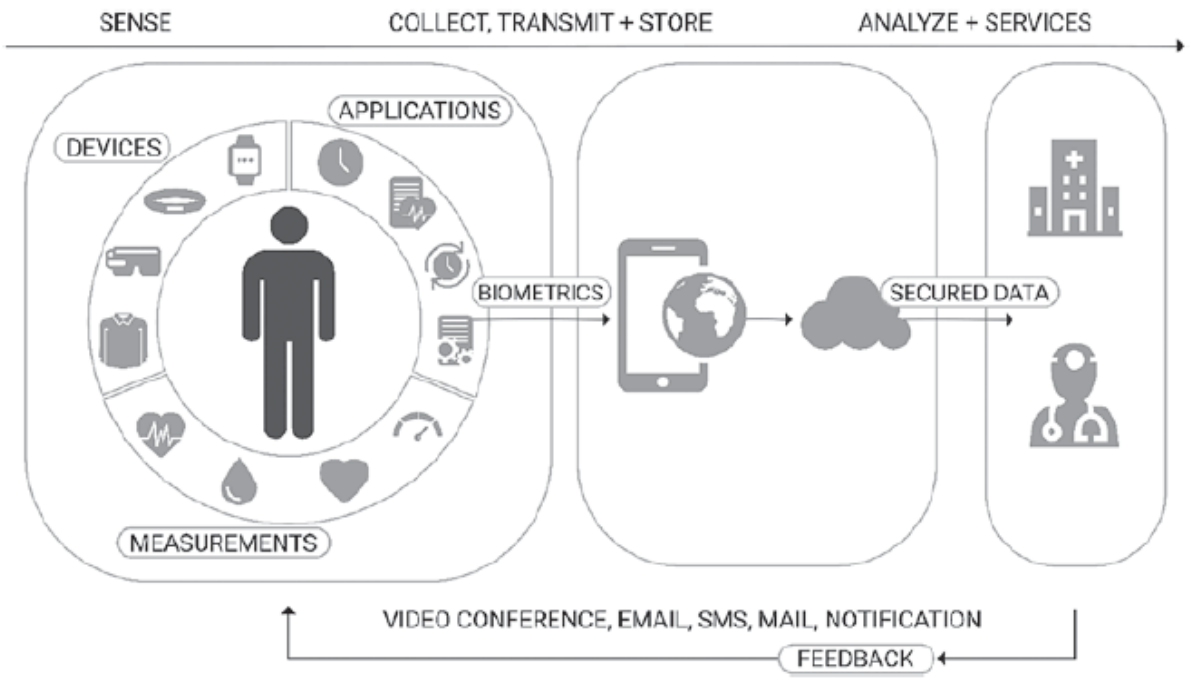

Figure 2 – A data-driven patient – healthcare professional relationship [2]

The potential applications of machine learning in healthcare are vast. Predicting disease, detecting the risk of cancers, suggesting meditation courses – are potential applications. Doctors make challenging decisions daily. It is vital that these decisions are as informed as possible. The use of AI, data mining, and predictive analytics enables clinicians to rationalize decisions and develop evidence-based treatment options. AI does not have to make the decision but can present the most reasonable opportunities to proceed.

Decision Models

Decision models help to forecast the future so that better decisions can be made. A decision is a choice between two or more alternatives. Predictive models support management and clinical decision making, and follow a decision analysis approach. Decision analysis is the process of separating or decomposing a complex decision, and incorporating uncertainty and dynamic assumptions into algorithms to generate alternatives. Alternatively, it can also be described as the use of analytic methods to make better decisions. The process of decision modeling involves framing the decision to be made, and then structuring a quantitative approach to evaluate choices. This decision model has several key steps, shown in Figure 3.

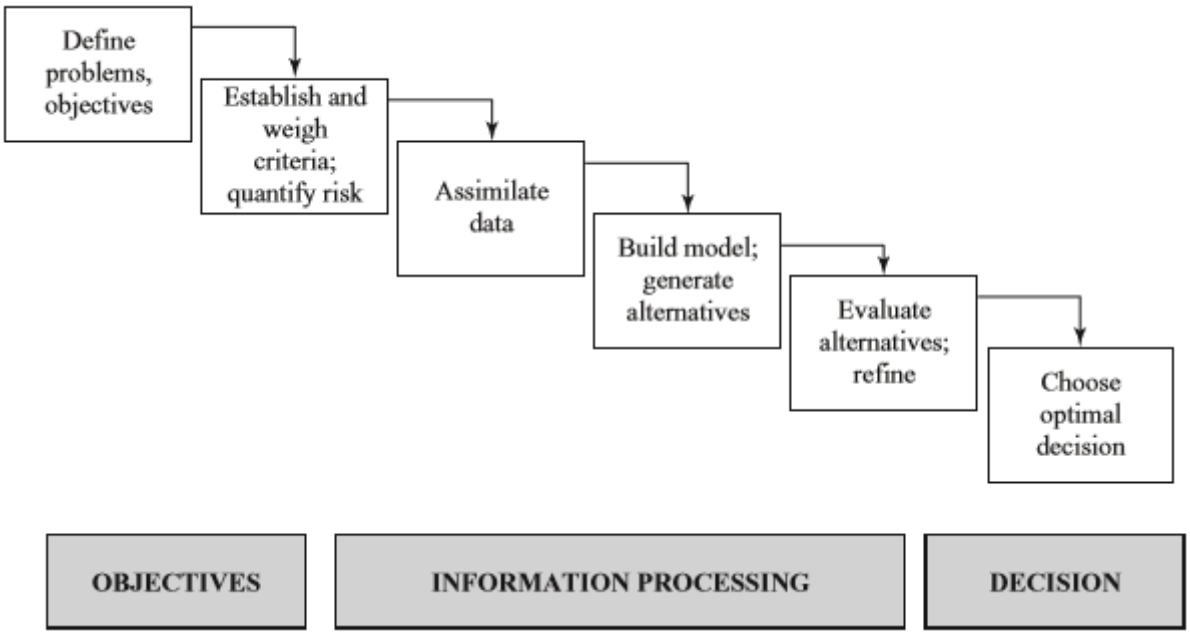

Figure 3 – Decision analysis

Decision models take many forms and one of those is predictive modeling. Predictive modeling is the use of an algorithm and software on large data sets to forecast potential outcomes, where an algorithm is a formula or calculation used to solve a problem in the model. In healthcare, there is a tremendous opportunity to use big data for predictive modeling, for instance, we could use data to identify how to improve patient safety and eliminate medical errors.

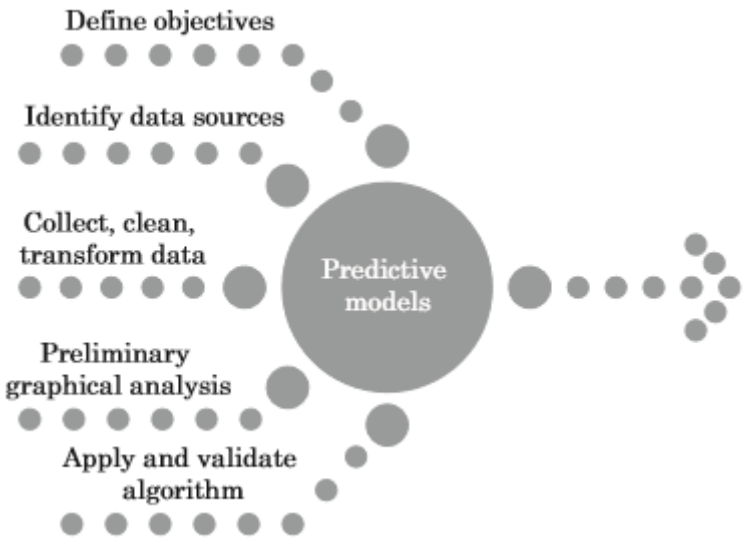

Figure 4 –Predictive modeling steps [3]

Figure 4 shows predictive modeling steps. Once data scientists have an estimation or prediction of future outcomes, they needs to incorporate them into the decision-making process. If they have predictions, it is necessary to share that information and to make sure healthcare professionals understand predictions the right way.

Text Mining in Healthcare

Several research studies have focused on the processing of textual information available in healthcare datasets. Text Mining refers to the discovery of knowledge from textual data. Text contains abundant qualitative information that is difficult to use in statistical modeling. In fields of healthcare, physicians express opinions in terms of words that contain useful information, not captured elsewhere. This information can be further used to develop intelligent models and improve healthcare process. However, traditional model building requires quantifiable information. Text mining converts text into numeric form that allows its use for analysis.

A general strategy for model building using text mining include next steps:

- Documents Collection (for instance, medical records),
- Data Preprocessing (e.g., exclude stop words, lemmatize, transform to one-size case if appropriate),

- Feature Extraction (transform textual data to numeric formats),
- Building a machine learning model (model selection, parameters, accuracy estimation),
- Results interpretation (e.g., interpret clusters).

Text mining is already used in healthcare projects. A brief overview of studies that highlight the significance of textual data and its suitability in research settings is presented below.

One notable research initiative has been performed at the Venderbilt Clinic, New York. The objective is to determine if a natural language processing program (NLP) could automatically code functional status information in accordance with the International Classification of Functioning, Disability, and Health (ICF) requirements.

Event detection is another significant area of research. Hazlehurst et al perform a study to identify vaccine reactions for the Vaccine Safety Datalink Project (VSD). They attempt to do this by analyzing medical care databases and patient medical records. Compared with methods that are used by clinicians this system significantly improves the positive predictive value [4].

Recently, text mining tools have been utilized in healthcare research, e.g. Cerrito and Cerrito analyze the electronic medical records from the emergency department of a hospital over a six month period, using text mining. They have found that similar complaints are treated differently depending on the physician on call. Such differences can affect care quality and costs. Therefore, text mining of prior expert treatment can provide physicians on call with an optimized treatment plan. It can also lead to development of protocols to alleviate disparity in treatment.

Conclusion

Text mining can be an effective tool in healthcare datasets analysis. There are still many challenges of using text mining on healthcare data. Healthcare studies have the added complexity of protection of confidential information. With

the availability of modern tools and techniques to mask confidential data, there is hope that most of these challenges would be overcome.

It is estimated that between 44,000 and 98,000 people die every year due to medical errors, making this a hot topic for research. Using predictive models on patient records will significantly reduce that. While technology can never replace doctors it can serve as a very capable double-check system that could greatly reduce medical error related deaths or complications.

References

1. Ervin Sejdic, Tiago H. Falk. Signal Processing and Machine Learning for Biomedical Big Data. CRC Press, – 2018.
2. Arjun Panesar. Machine Learning and AI for Healthcare: Big Data for Improved Healthcare Outcomes. Apress, – 2019.
3. James R. Langabeer. Performance Improvement in Hospitals and Health Systems: Managing Analytics and Quality in Healthcare, 2nd Edition, – 2018.
4. Uzma Raja, Tara Mitchell, Timothy Day, J. Michael Hardin. Text Mining in Healthcare: Applications and Opportunities / Journal of healthcare Information management: JHIM, – February 2008.

Приложение Б – Пример документа «Осмотр в стационаре при поступлении»

Номер пациента, пол и возраст
Номер 11 пол женский возраст 46

Дата и время осмотра
Дата 22 11 2016 время 23:50

Жалобы
Повышенная температура, головная боль, боли распирающего характера правой нижней конечности, усиливающиеся при ходьбе, болезненность правой паховой области

Анамнез болезни
заболела вечером 21.11.16г - общее недомогание, головные боли. При измерении температура 37,0. 22.11.16 температура повысилась до 38,0, появилась болезненность в правой паховой области, отёчность правой голени и стопы, обратилась в поликлинику, направлена в дежурный хирургический стационар с подозрением на тромбофлебит. Осмотрена дежурным хирургом, проведено УЗИ правой нижней конечности. При измерении Т-39,0, появилась гипермия кожи правой голени. Введена в/м литическая смесь и больная отправлена в инфекционное отделение с Д-зом Рожа правой голени, эритематозная форма.

Анамнез жизни
Заболеваний детсва не помнит. Росла и развивалась соответственно полу и возрасту. Вредных привычек нет, употребление ПАВ - отрицает. ЭПИДАНАМНЕЗ: квартира благоустроенная, контакт с инфеционными больными - отрицает. Питание домашнее, речную рыбу с-ва карповых употребляет редко. Географический анамнез б/о

Анамнез ВТЭ
В листе нетрудоспособности нуждается с 22.11.2016, амбулаторно не выдавался.

Объективный статус
Вес 95кг, рост 165 см, гиперстенического телосложения, повышенного питания. положение активное, доставлена в стационар на личном автотранспорте. сознание полное, выражение лица осмысленное. кожные покровы, видимые слизистые - физиологической окраски, иктеричности склер нет. Губы физиологической окраски, слегка влажные. Высыпаний, трещин - не обнаружено. Слизистая полости рта равномерно розовая, язык обычной величины и формы, умеренно влажный, обложен белым налетом. Мягкое и твёрдое небо розовые, без энатемы, без налета. Неприятного запаха изо рта не ощущается. Миндалины не увеличены, налета на миндалинах нет. Пальпируемые л/узлы не увеличены, не спаяны, при пальпации - безболезненные, паховые справа - до 2,5 см., болезненные при пальпации. Бедренные л/узлы справа также увеличины до 2,5 - 3,0 см, при пальпации чувствительные, кожа над ними не изменена. Дыхание везикулярное, хрипов н выслушиваю. Тоны сердца ясные, ритм правильный, ЧСС 116 в мин. Живот округлой формы, увеличен в размере, мягкий, участвует в акте дыхание, уровней жидкости при перкуссии не определяется. Сосудистых звездочек и грыжевых выпячиваний не наблюдается. Усиление венозного рисунка на передней брюшной стенке нет. При пальпации - живот безболезненный, уплотненный и опухолевидных образований не обнаружено. Резистентность мышц пресса выражена умеренно. Нижнй край печени не выступает из-под края правой рёберной дуги по СКЛ, при пальпации безболезненный. Проекция Ж.П. - безболезненная, пузырные симптомы - отрицательные. Селезенка в положении на правом боку не пальпируется, перкуторно не увеличена. Кишечник - эластичен, безболезненный, урч

аниея, крипитации нет. Почки в положении лёжа не пальпируются, поколачиван ие - безболезненное с обеих сторон. Диурез адекватен, стул ежедневно.

Локальный статус
на правой голени и стопе отёк тестоватой консистенции. края гиперемии чет кие по типу языков пламени. участков размягчения и флюктуации не определяе тся. в паховой области пальпируется болезненный, увеличенный до 3,0 см л/у зел. В области бедренного треугольника так же пальпируется увеличенный до 3,0 см и болезненный л/узел.

Диагноз при поступлении
А46 рожа правой голени и стопы, эритематозная, распространенная, средней степени тяжести, первичная.

Обоснование диагноза
Учитывая острое начало заболевания с повышения температуры, с симптомов о бщей интоксикации, а затем присоединений местных проявлений в виде отека, покраснения кожи на правой голени в стопе, учитывая что подобная клиничес кая картина развивается впервые, можно выставить диагноз: Рожа правой голе ни и стопы, эритематозная, распространенная, средней степени тяжести, перв ичная.

Диагноз
А46 Рожа