Приложение № 1

к договору №

от

Техническое задание

|  |  |
| --- | --- |
| ИСПОЛНИТЕЛЬ  Самозанятая Саврей Юлия Михайловна  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/Саврей Ю.М./ | ЗАКАЗЧИК  ООО "\_\_\_\_\_\_\_\_"  Директор  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_/\_\_\_\_\_\_\_\_\_\_\_/ |

Оглавление

[1. Общие положения 3](#_Toc167650029)

[1.1 Наименование НИР 3](#_Toc167650030)

[1.2 Формулировка проблемы 3](#_Toc167650031)

[1.3 Нормативная документация 4](#_Toc167650032)

[1.4 Источники разработки 5](#_Toc167650033)

[1.5 Консультанты со стороны Заказчика 5](#_Toc167650034)

[1.6. Требования к патентной чистоте и лицензированию 5](#_Toc167650035)

[2. Список терминов и определений 5](#_Toc167650036)

[3. Список сокращений 6](#_Toc167650037)

[4. Цели и задачи работы 6](#_Toc167650038)

[5. Позиционирование 6](#_Toc167650039)

[6. Входные и выходные данные 7](#_Toc167650040)

[7. Функциональные требования 7](#_Toc167650041)

[8. Ограничения 9](#_Toc167650042)

[9. Допущения и зависимости 10](#_Toc167650043)

[10. Атрибуты качества 10](#_Toc167650044)

[11. Требования к защищенности 11](#_Toc167650045)

[12. Требования к развертыванию 12](#_Toc167650046)

[13. Документирование 12](#_Toc167650047)

[14. Порядок приемки работ 12](#_Toc167650048)

[15. Приложения 13](#_Toc167650049)

# 1. Общие положения

## 1.1 Наименование НИР

«GiggleRX» (шифр проекта “Прогнозирование диагноза болезни по описанию симптомов”)

GiggleRX – "рецепт с улыбкой" или "веселый рецепт". Это название подчеркивает, что бот предоставляет медицинские рекомендации и информацию, но делает это в дружелюбной манере, что может помочь пользователям чувствовать себя более комфортно при использовании сервиса.

## 1.2 Формулировка проблемы

Актуальность задачи:

В современном мире здоровье становится все более приоритетной областью для людей. Однако, доступ к медицинским услугам может быть ограниченным или затрудненным из-за различных факторов, таких как удаленность от медицинских учреждений, высокие затраты на медицинское обслуживание и другие. В этом контексте возникает актуальная задача разработки телеграмм бота, способного прогнозировать возможный диагноз болезни по описанию симптомов.

Портрет целевой аудитории

* Возраст: от 16 до 55 лет
* ﻿﻿Местоположение: Санкт-Петербург, городская область

Целевая аудитория и ее потребность:

1. Люди с ограниченным доступом к медицинским учреждениям: это могут быть жители удаленных районов, где доступ к квалифицированным медицинским специалистам ограничен.

2. Занятые люди: люди, у которых ограничено время для посещения медицинских учреждений из-за занятости на работе или других обстоятельств.

3. Люди с медицинской тревожностью: те, кто испытывает тревогу относительно своего здоровья, могут обратиться к боту для получения предварительной оценки своих симптомов.

4.Молодежь: особенно молодые люди, которые привыкли использовать технологии для решения различных задач, могут обратиться к боту в первую очередь.

5. Люди с ограниченными финансовыми ресурсами: те, кто не может позволить себе регулярные посещения врачей из-за финансовых затрат, могут воспользоваться услугами бота для начальной оценки своего состояния здоровья.

6. Люди, живущих в регионах с недостаточным медицинским обслуживанием: это включает в себя как удаленные области, так и города с недостаточным числом медицинских специалистов.

7. Медицинские специалисты: врачи и медицинский персонал, которые могут использовать инструмент для дополнительной проверки своих предположений и получения дополнительных данных для обсуждения с пациентами.

Потребность аудитории заключается в том, чтобы быстро и удобно получить информацию о возможном диагнозе заболевания на основе предоставленных симптомов, что помогает им принять более осознанные решения относительно своего здоровья.

Проблематика:

Одними из основных проблем являются сложность и неполнота симптомов, которые пользователь может описать. Также существует необходимость в точности и достоверности предоставляемой информации, чтобы избежать неправильных диагнозов и рекомендаций, которые могут привести к негативным последствиям.

Бизнес-цель заказчика:

Основной бизнес-целью заказчика является предоставление пользователям удобного и надежного инструмента для предварительной оценки их здоровья на основе симптомов. Это может способствовать увеличению доступности медицинской помощи, улучшению здоровья населения и снижению нагрузки на медицинские учреждения за счет предотвращения случаев самодиагностики и самолечения.

Бизнес-требования и бизнес-правила:

1. Точность диагностики: бот должен обладать высокой точностью прогнозирования диагноза на основе предоставленных симптомов.

2. Удобство использования: бот должен быть интуитивно понятным и легким в использовании для широкого круга пользователей.

3. Обновление базы данных: необходимо регулярно обновлять базу данных бота новыми медицинскими данными и симптомами для улучшения точности диагностики.

4. Рекомендации и дальнейшие действия: после предполагаемого диагноза бот должен предоставлять рекомендации пользователю, такие как обращение к врачу для подтверждения диагноза или предоставление дополнительной информации о возможных методах лечения.

5. Интеграция с медицинскими учреждениями: предусмотреть возможность интеграции бота с медицинскими учреждениями для получения дополнительной экспертной консультации при необходимости (на будущую оптимизацию телеграмм бота).

## 1.3 Нормативная документация

Федеральный закон "О персональных данных" от 27.07.2006 № 152-ФЗ (ФЗ-152):

- Этот закон регулирует обработку персональных данных, включая сбор, хранение, использование и распространение таких данных. В рамках проекта необходимо соблюдать требования ФЗ-152 при обработке персональных данных пользователей (если таковые собираются).

Нормативно-техническая документация по защите персональных данных (НТД по ЗПД):

- Включает в себя требования и рекомендации по обеспечению безопасности и конфиденциальности персональных данных, включая методы защиты, организацию работы с данными и процедуры реагирования на инциденты.

Внутренние правила и политики безопасности:

- В рамках проекта могут быть разработаны внутренние правила и политики безопасности, определяющие процессы работы с данными, доступ к ним, обработку и хранение информации.

Регламенты и стандарты Telegram:

- Учитывая, что проект осуществляется в среде Telegram, важно также соблюдать положения регламентов и стандартов этой платформы, связанные с безопасностью и обработкой данных.

Основанием для выполнения/разработки НИР является заданный пет-проект по дисциплине «Инженерия машинного обучения (Bonus Track)» факультета программной инженерии и компьютерных технологий НИУ ИТМО.

## 1.4 Источники разработки

Данный проект запускается впервые

На основе результатов анализа данных разработать модель машинного обучения и искусственного интеллекта, способную прогнозировать диагноз болезни на основе описания симптомов.

Описание процесса встраивания модели:

После разработки модели ее следует интегрировать в процесс работы телеграмм бота:

1. Интеграция модели: взять соответствующие программные интерфейсы (API), позволяющие взаимодействовать между телеграмм ботом и моделью прогнозирования диагноза.

2. Разработать метод прогнозирование диагноза болезни на основе машинного обучения и искусственного интеллекта.

3. Тестирование интеграции: провести тестирование взаимодействия между телеграмм ботом и моделью, убедившись, что все работает корректно и безошибочно.

4. Оптимизация производительности: в случае необходимости оптимизировать производительность модели для обеспечения ее эффективной работы в реальном времени при обработке запросов от пользователей.

5. Запуск в эксплуатацию: после успешного завершения тестирования интеграции модель может быть запущена в эксплуатацию в составе телеграмм бота.

Временные задержки по данным и работе конвейера:

- Разработка модели: от нескольких месяцев до полугода, в зависимости от сложности и объема модели.

- Тестирование модели: от одного дня до двух дней, в зависимости от требований к тестированию и доступности данных.

- Интеграция модели: от часа до двух, в зависимости от сложности интеграции и наличия необходимых ресурсов.

- Оптимизация производительности: от нескольких недель до нескольких месяцев, в зависимости от сложности модели и требований к производительности.

- Запуск в эксплуатацию: после успешного завершения предыдущих этапов, процесс запуска в эксплуатацию может занять от нескольких дней до нескольких недель.

## 1.5 Консультанты со стороны Заказчика

Иванов Иван Иванович, главный специалист, email@company.com, +7(999)888-77-66, доступен для консультаций в рабочее время.

## 1.6. Требования к патентной чистоте и лицензированию

Используемые библиотеки и фреймворки должны распространяться под лицензиями, совместимыми с коммерческим использованием.

# 2. Список терминов и определений

Telegram-бот — это мини-программа внутри мессенджера, которая управляется текстовыми командами в чате по принципу «вопрос — ответ».

API — программный интерфейс, то есть описание способов взаимодействия одной компьютерной программы с другими.

Программа-парсер — это программа, сервис или скрипт, который собирает данные с указанных веб-ресурсов, анализирует их и выдает в нужном формате.

Prompt – это инструкция, которая используется при создании команды для нейросети. От правильно составленного промпта зависит релевантность информации, которую искусственный интеллект выдаст вам в ответ.

GigaСhat – русскоязычная нейросеть от Сбера.

Random Forest Tree Classification method - это мощный алгоритм машинного обучения, который используется для прогнозирования на основе множества решений, полученных от различных деревьев решений.

# 3. Список сокращений

API - Application Programming Interface

Tg-bot – Telegram-bot

# 4. Цели и задачи работы

Цель: предоставить пользователям возможность получить предварительную оценку своего здоровья на основе предоставленных симптомов. Т.е. разработать телеграмм-бота, который на основе симптомов пользователей спрогнозирует диагноз болезни. Это позволит пользователям принимать более осознанные решения относительно своего медицинского состояния и обращаться за медицинской помощью вовремя.

Задачи:

1. Сбор данных о симптомах и диагнозах

1.2 Реализация метода машинного обучения по датасету

1.3 Подключение искусственного интеллекта

2. Написание программного кода для телеграмм-бота

3. Запуск бота и сбор обратной связи от пользователей

4. Настройка искусственного интеллекта на основе данных от пользователей

5. Тестирование и оптимизация бота

6. Интеграция с медицинскими учреждениями (на будущее)

# 5. Позиционирование

GiggleRX представляет собой Telegram-бот, который использует метод машинного обучения для прогнозирования болезней, и искусственный интеллект для анализа болезней (их описания), предоставленных пользователями, и определения возможных заболеваний. Бот активно взаимодействует с API GigaChat для прогнозирования болезней и использует API Telegram для общения с пользователями.

Пользовательские ответы временно записываются в базу данных для последующего анализа и использования в модели телеграмм-бота (после ответа бота – удаляются).

Информация передается через API Telegram с использованием защищенного протокола HTTPS, что обеспечивает шифрование данных в процессе передачи.

Условия эксплуатации:

- Стабильное подключение к интернету для взаимодействия с телеграмм-ботом.

Данные из базы экспортируются в формате CSV-файлов, которые содержат структурированную информацию, готовую для анализа машинным алгоритмом. Они подаются на вход модели машинного обучения для обучения или прогнозирования.

Роли и количество пользователей:

- Разработчик: 1 человек, занимающийся разработкой модели машинного обучения и искусственного интеллекта, кода телеграмм-бота и интеграции между ними.

- Пользователи: потенциально неограниченное количество пользователей, которые могут использовать бота для предварительной оценки своего здоровья.

# 6. Входные и выходные данные

Входные данные:

Входные данные представляют собой структурированные данные о симптомах и соответствующих им возможных диагнозах. Эти данные включают в себя описания симптомов, их характеристики, а также соответствующие им возможные заболевания или состояния. Примеры входных данных могут включать в себя:

- Описание симптомов (например, боль в груди, кашель, высокая температура и т. д.).

- Характеристики симптомов (например, интенсивность, частота, длительность и т. д.).

- Возраст и пол пациента.

- Личные медицинские истории (предыдущие заболевания, хронические состояния и т. д.).

Выходные данные:

Результаты проекта представлены в форме предполагаемого диагноза или набора возможных диагнозов, соответствующих описанным симптомам. Это может быть список заболеваний или состояний, вероятностные оценки вероятности каждого диагноза или рекомендации о дальнейших действиях (например, обратиться к врачу для дополнительного обследования и т. д.).

Также выходные данные представляют картинки котов, чтобы поднять настроение пользователям.

Взаимодействие между компонентами:

Взаимодействие между компонентами проекта будет осуществляться посредством API, что обеспечит передачу данных между системами в структурированном формате и обеспечит их совместимость и согласованность.

Требования к обезличиванию данных:

Личные данные пользователей не будут сохраняться в базе данных и не будут использоваться в процессе прогнозирования диагноза болезни. Только необходимые медицинские данные будут использоваться для обработки запросов и предоставления рекомендаций.

# 7. Функциональные требования

*Методы, которые должны реализовываться разрабатываемыми технологиями*

Пользователь предоставляет данные через чат ТГ-бота, вводя информацию о своем возрасте, симптомах болезни, по желанию – личных медицинских историй.

Данные пользователя считываются нейросетью и алгоритмом модели машинного обучения для предоставления рекомендаций.

Вывод рекомендаций пользователю в понятном формате через чат ТГ-бота.

Машинное обучение основано на Random Forest Tree Classification method. Это мощный алгоритм машинного обучения, который используется для прогнозирования на основе множества решений, полученных от различных деревьев решений.

*Способы доступа к выходным данным, их отображения (визуализации);*

Для получения доступа к данным, пользователю необходимо ввести в ТГ-бот нужные данные (те описание симптомов), на основе которых и строится конечный ответ программы. Ответ программы пользователь увидит в этом же чате с ТГ-ботом.

*Предобработка входных данных и работа МО системы*

Загрузка и предобработка данных:

* Загружается датасет с информацией о симптомах и заболеваниях.
* Данные предобрабатываются: выделяются признаки (симптомы) и целевая переменная (прогнозируемая болезнь).
* Строки, представляющие симптомы, инициализируются значением 0.
* Данные также начинает обрабатывать нейросеть Сбера – GigaChat

Обучение модели:

* Данные разбиваются на обучающую и тестовую выборки.
* Создается и обучается классификатор RandomForestClassifier.
* Оценивается точность модели.
* Создается prompt для нейросети, на основе которого будут строяться ответы GigaChata

Выбор симптомов:

* Функция SimiliarSymptoms выбирает похожие симптомы, которые могут быть связаны с заданным первым симптомом.

Прогнозирование болезни:

* Функция GetDisease принимает список симптомов, подготавливает входные данные и использует обученную модель для предсказания болезни.
* Также нейросеть принимает список симптомов и подготавливает входные данные

Выдача информации о болезни:

* Функция DiseaseInfo и нейросеть GigaChat возвращает информацию о заболевании.
* Также с помощью API телеграмм-бот возвращает картинку котиков после сообщения с информацией о заболевании

*Контроль характеристик данных*

1. Проверка наличия необходимых симптомов для прогнозирования диагноза у модели машинного обучения (датасет модели ограничен). Если таковых нет – нейросеть берет на себя ответ пользователю.

2. Проверка корректности формата и структуры входных данных перед передачей их в нейронные сети.

*Архивирование данных*

Архивирование данных не требуется, так как процесс обучения и прогнозирования будет происходить в реальном времени с использованием актуальных данных.

# 8. Ограничения

Не использовать запатентованные алгоритмы

Языки программирования: Python

Для разработки и отладки программного обеспечения используется современная интегрированная среда разработки (IDE) – PyCharm.

Контейнер, который позволяет упаковать проект телеграмм-бота, включая все его зависимости, в стандартный единый блок – Docker

Не использовать иностранные библиотеки, которые могут подпадать под санкции или экспортные ограничения

Ограничения на использование сторонних библиотек и сервисов:

Ограничение запросов API GigaChat: для бесплатной версии Gigachat PRO лимит токенов генерации текста составляет 50000. Одновременные запросы (доступное количество одновременных обращений к API) – 1 поток

Также используется платформа Telegram и его API. Существуют некоторые лимиты для бота внутри мессенджера:

* Длина @username для бота: 5–32 символа.
* Длина информации о боте: до 120 символов.
* Длина описания бота: до 512 символов.
* Количество кнопок: до 100 штук.
* Отправление ботом сообщений нескольким разным пользователям / в чаты, которые подключены к боту: до 30 сообщений с интервалом от 1 секунды.
* Количество inline-результатов: до 50 элементов на страницу.

Ограничения по импортозамещению:

Все компоненты отечественного производства, кроме Telegram API.

Ограничения, вытекающие из нормативной документации:

Проект должен соответствовать нормативной документации, касающейся обработки медицинских данных и защиты персональной информации. Это включает в себя соблюдение законодательства о защите данных и конфиденциальности, а также соответствие стандартам безопасности в области здравоохранения.

Аппаратные ограничения на вычислительные мощности, параллельность обработки, доступ к интернет и локальной сети:

Существует ограничение на получение данных от Telegram-бота. Пользователи всегда должны иметь подключение к интернету, чтобы получать информацию, однако история взаимодействия может быть доступна и в офлайн-режиме.

Ограничения на используемые языки программирования и средства разработки, хранилища данных:

Ограничений нет

# 9. Допущения и зависимости

Используемые библиотеки:

* telebot (или pyTelegramBotAPI) – для создания и управления Телеграм-ботом
* requests – для выполнения HTTP-запросов
* gigachat – для взаимодействия с API GigaChat
* logging – для ведения логов
* pandas - библиотека для работы с данными, предоставляющая высокопроизводительные структуры данных и инструменты для анализа данных.
* sklearn (scikit-learn) - библиотека для машинного обучения, включающая инструменты для предварительной обработки данных, моделирования и оценки.
* collections - встроенная библиотека, предоставляющая специализированные контейнерные типы данных.
* numpy - библиотека для работы с многомерными массивами и матрицами, а также для выполнения математических операций высокого уровня.
* warnings - встроенная библиотека для управления предупреждениями в коде.
* random - встроенная библиотека для генерации случайных чисел и выполнения различных случайных операций.
* re - встроенная библиотека для работы с регулярными выражениями.
* json - встроенная библиотека для работы с JSON (JavaScript Object Notation) данными.
* os - встроенная библиотека для взаимодействия с операционной системой.

Большая часть данных о болезнях берется из API GigaChat PRO, поэтому проект сильно зависит от его корректности данных.

Если предметная область разрастется, то будет необходима смена модели МО.

# 10. Атрибуты качества

Все данные проходят предварительную обработку.

Что будет делать система если упал канал связи между пользователем и сервером?

Если потеряна связь с Telegram: ожидание связи, чтобы продолжить взаимодействие.

Качество данных: важно обеспечить качество и достоверность входных данных, включая описания симптомов и соответствующие им диагнозы. Это включает в себя проверку данных на отсутствие ошибок, аномалий и недостаточную информацию.

Надежность:

1. Вероятность работы ПО без сбоев: надежность программного обеспечения может измеряться через процент успешно завершенных операций, среднее время наработки на отказ (MTBF) и максимальную приемлемую вероятность отказа за определенный период времени.

Устойчивость:

Система должна быть устойчива к неверному вводу данных, все введенные данные проверяются, так что в случае некорректно полученной информации система не перестаёт работать и обрабатывать запросы.

Атрибуты качества:

Доступность: запланированное время доступности системы для использования и полностью работоспособности. Формально доступность определяется как среднее время наработки на отказ (MTTF) системы, деленное на сумму среднего времени наработки на отказ и среднего времени восстановления после отказа (MTTR). Плановое техническое обслуживание также влияет на доступность системы.

Система доступна круглосуточно при условии подключения к интернету.

Скорость оперирования данными:

Система должна оперировать данными с высокой скоростью для обеспечения оперативного прогнозирования диагноза болезни. Это включает в себя выбор соответствующих аппаратных ресурсов.

# 11. Требования к защищенности

В контексте проекта "Прогнозирование диагноза болезни по описанию симптомов", учитывая, что Телеграмм-бот не будет хранить персональные данные пользователей и защита данных уже предоставляется платформой Telegram, определенные аспекты безопасности не будут применяться или будут минимизированы. Меры безопасности, которые все же могут быть релевантны:

Уровни привилегий пользователей:

Все пользователи имеют одинаковые привилегии в рамках использования функционала бота.

Идентификация и проверка подлинности пользователей:

Так как бот не хранит персональные данные пользователей, нет необходимости в процессах идентификации и аутентификации пользователей.

Конфиденциальность данных:

Бот не будет хранить персональные данные, но все равно важно обеспечить конфиденциальность общения пользователей с ботом. Это может быть достигнуто путем шифрования сообщений через протокол HTTPS.

Уничтожение, повреждение и кража данных:

Поскольку бот не хранит персональные данные пользователей, опасности уничтожения, повреждения или кражи данных минимизируются.

Защита от вредоносного ПО:

Приложение Telegram обеспечивает защиту от вредоносного ПО.

Сетевая безопасность и шифрование данных:

Telegram обеспечивает защищенное соединение через протокол HTTPS, что обеспечивает шифрование данных во время передачи.

Доступ к функциям администратора (управление telegram-bot) должен быть строго ограничен

# 12. Требования к развертыванию

Развертывание должно осуществляться через Docker-контейнеры

Предусмотреть возможность масштабирования для горизонтального распределения нагрузки

Подготовить развертываемые артефакты и документацию по установке

Обеспечить возможность обновления системы до новых версий с минимальным простоем

# 13. Документирование

Для отчетности требуется составить следующие документы:

1. Отчет о НИР (научно-исследовательская работа):

* Описывает цель, задачи, методологию, результаты и выводы проведенного исследования.
* Включает в себя описание архитектуры системы, использованные модели искусственного интеллекта и машинного обучения, данные, анализ результатов и прочие научные аспекты проекта.

1. Инструкция пользователя:

* Обеспечивает пользователя информацией о том, как пользоваться системой.
* Описывает основные функции и возможности системы, шаги по вводу данных, получению результатов и интерфейс взаимодействия с системой.

1. Инструкция развертывания:

* Предоставляет подробные инструкции по установке и настройке системы на целевом окружении.
* Включает в себя необходимые шаги по установке всех зависимостей, запуска сервера и другие действия, необходимые для успешного развертывания системы.

# 14. Порядок приемки работ

Процедура проверки:

**ТГ-бот**

1. Отправить сообщение боту в телеграм “/start”
2. Подробно описать симптомы болезни (раздел 6)
3. Отправить текстовое сообщение телеграмм-боту
4. Получить спрогнозированную болезнь и совет

Выполнение всех функциональных требований (раздел 7)

Производительность в соответствии с пунктом 10

Отсутствие критичных уязвимостей, влияющих на безопасность

Приемка оформляется актом выполненных работ, подписанным обеими сторонами.

# 15. Приложения

1.      чертежи, схемы, описания, расчеты, обоснования, изобретения и другие документы, которые должны быть использованы при выполнении работ (оказании услуг), поставок;

2.      перечень лиц, с которыми должны быть согласованы конкретные технические и организационные решения и вопросы в процессе выполнения работ (оказания услуг), поставок;

3.      справочные и другие материалы.