МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

Отчет по проекту

**Предсказание эффективности лечения препаратом  
 на основе данных пациента**

по дисциплине «Интеллектуальный анализ данных»

Выполнили: студенты группы 932102

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Осипова А.Д.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Локотецкий М.Е.

Проверил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_Замятин А.В.

Томск – 2023

**ТЕХНИЧЕСКОЕ ЗАДАНИЕ**

**на разработку программного сервиса**

**интеллектуального анализа данных**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1. **Название** | Предсказание эффективности лечения препаратом на основе данных пациента | | | | |
| 1. **Исполнение** | Групповое | | | | |
| 1. **Разработчики** | 1. Осипова А.Д. (К-ближайших соседей)  2. Локотецкий М.Е. (Градиентный бустинг) | | | | |
| 1. **Цель разработки** | Создать инструмент для предсказания эффективности лечения препаратом на основе данных пациента (возраст, пол, группа крови, заболевание, препарат) | | | | |
| 1. **Решаемые задачи** | 1. Отбор данных. 2. Предварительная обработка (чистка данных). 3. Сделать выводы о наличии зависимости группы крови и частоты выявления заболевания. 4. Построение классификационной модели. | | | | |
| 1. **Сроки исполнения** | 6.1 Начало: | 30.10.2023 г. | 6.2 Окончание: | | 22.12.2023 г. |
| 1. **Основные требования к разработке** | | | | | |
| 1. **Данные** | **8.1 Источник данных** | | | <https://www.kaggle.com/datasets/prasad22/healthcare-dataset> | |
| **8.2 Параметры обучающих данных** | | | 3000 векторов; каждый вектор имеет 15 показателей, значимых – 5 и 1 (прогнозируемый) результат тестирования | |
| **8.3 Формат признакового вектора x** | | | **X =** {*x1*,*x2*,…,*xn*},  где  *x1 – возраст*  *x2 – пол*  *x3 – группа-крови (+резус-фактор)*  *x4 – заболевание*  *x5 – используемый препарат*  n = 5 | |
| **8.4 Формат вектора y (опционально)** | | | **Y =** {*y1*,*y2*,…,*ym*},  где  *y1– результат лечения*  *m* = 1 | |
| 1. **Функционал** | **9.1 Функции предварительной обработки (с трансформацией/без трансформации признакового пространства)** | | | Сокращение объема данных (исключение векторов со схожими признаками) | |
| **9.2 Функции содержательной обработки (метод DM, методы DM и принцип интегрирования результата, полученного отдельными методами)** | | | К-ближайших соседей, градиентный бустинг | |
| 1. **Результаты** | **10.1 Представление результатов/что позволят установить (формулировка полученных неочевидных знаний, способствующих принятию на практике конкретных управленческих решений)** | | | Модель позволяет определить наиболее вероятный результат лечения на основе исходных данных о пациенте | |
| 1. **Исследования** | **11.1 Исследовательская компонента (оценка эффективности и пределов применимости используемых методов, рекомендации по использованию методов в решении аналогичных задач)** | | | Проверка гипотезы о наличии зависимости результата лечения от перечисленных признаков.  В результате проведенной работы будут сделаны выводы и даны рекомендации. | |

Ответственные исполнители: \_\_\_\_\_\_\_\_\_\_\_\_\_Осипова А.Д. Заказчик\_\_\_\_\_\_\_\_\_\_\_\_\_А.В. Замятин

\_\_\_\_\_\_\_\_\_\_Локотецкий М.Е.

**Введение.**

Данная работа основана на исследованиях, в которых рассматривается взаимосвязь эффективности лечения препаратом от группы крови пациента. Так, например, иммунитет у людей с первой группой крови лучше справляется с малярией и хуже с гепатитом С и раком печени. Соответственно, для пациентов с первой группой крови и второй для одного и того же, заболевания будут предпочтительны различные препараты. Подобные исследования существуют и для других факторов, которые мы рассматриваем в работе. Появляется задача предсказания эффективности лечения препаратом пациента в зависимости от его данных.

**Постановка задачи.**

В качестве источника данных для работы был выбран сайт <https://www.kaggle.com/>, данные по ссылке <https://www.kaggle.com/datasets/prasad22/healthcare-dataset> содержат набор данных по здравоохранению.

Исходный набор данных содержит 15 признаков (14 входных и 1 целевой). Целевой признак Test Results содержит информацию о результатах лечения пациента и принимает три значения: Normal, Abnormal, Inconclusive.

В рамках данной работы мы создали прототип инструмента ИАД, который позволит создать прогноз лечения пациента (неубедительный, удовлетворительный, неудовлетворительный) на основе информации о нем.

**Подготовка данных.**

На первом этапе было сделано профилирование данных с помощью библиотеки pandas\_profiling. Всего в датасете 10000 записей, 12 категориальных признака (object), 1 непрерывный (float64), 2 признака принимают только целочисленные значения (int64). Среди признаков есть как информация о самом пациенте (возраст, пол, заболевание и т.д.), так и неинформативные для нашей работы признаки (номер палаты, название больничного учреждения и т.д.). Все признаки, которые не оказывают влияние на результат были удалены. В итоге было отобрано 6 признаков, которые пригодны для построения модели (таблица 1).

|  |  |  |  |
| --- | --- | --- | --- |
| Номер | Входной/Целевой | Тип | Признак |
| 1. | Входной | Числовой | Age |
| 2. | Категориальный | Gender |
| 3. | Blood Type |
| 4. | Medical Condition |
| 5. | Medication |
| 6. | Целевой | Test Results |

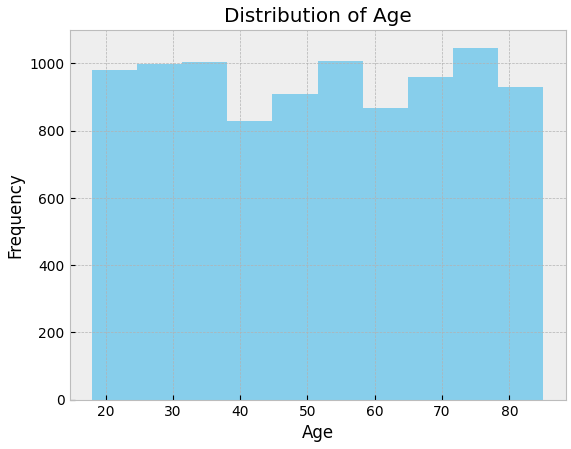
Таблица 1

Целевой признак Test Results (результаты теста) принимает три значения: Normal, Abnormal, Inconclusive.

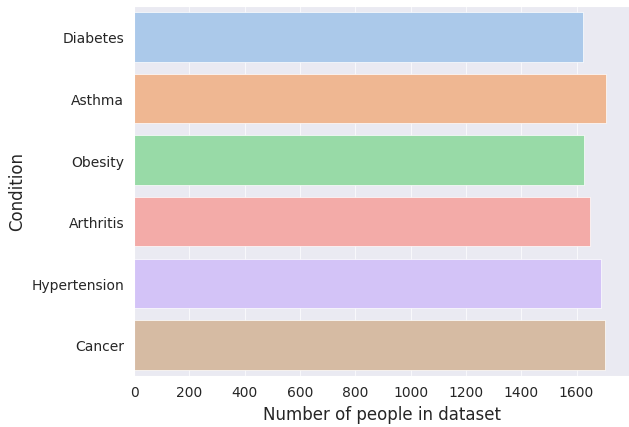
Пропусков обнаружено не было. Все дубликаты были исключены из датасета, в результате чего количество записей уменьшилось до 9530. Все категориальные признаки были закодированы с помощью ordinal encoding.

Было произведено деление исходного набора данных на тренировочную и тестовую выборки. Целевой признак сбалансирован. Построим графики для нормализованного пространства:

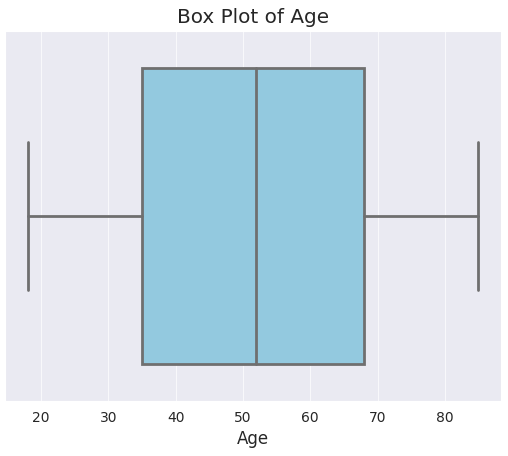
* Гистограмма с количеством людей и соответствующим возрастом



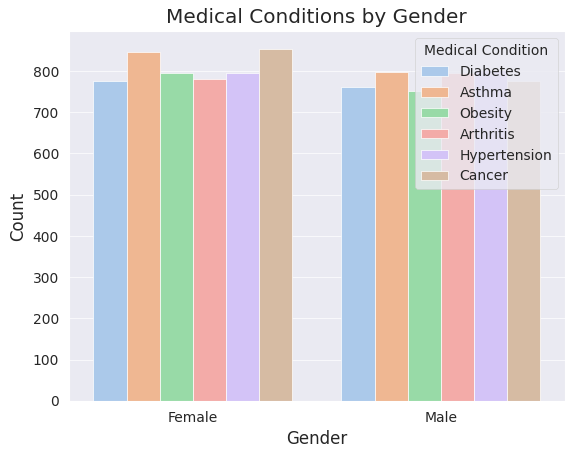
* Соотношение количества людей в наборе данных с заболеваниями.



* Визуализация возраста по заболеваниям.



* Анализ связи между полом и заболеванием.



* Распределение результатов теста с количеством людей в наборе данных



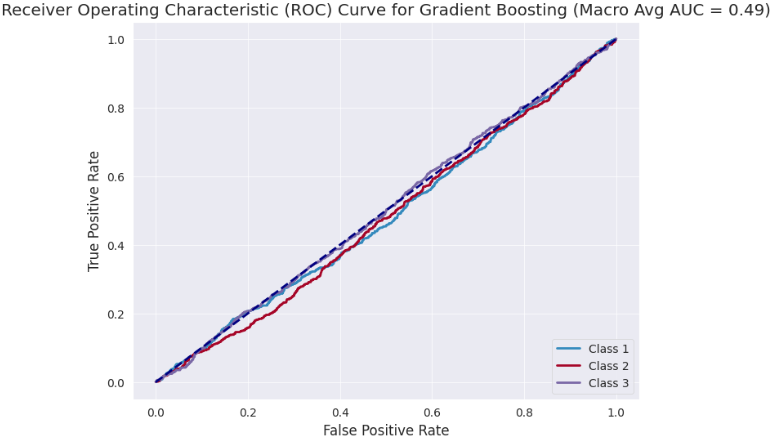
**Моделирование.**

При решении задачи было использовано две модели: Gradient boosting (Градиентный бустинг) и k Nearest Neighbor (k ближайших соседей), рассматривать мы их будем в порядке упоминания.

**Gradient boosting:**

* Проведем биннинг возраста, разделив его на 10 интервалов, для улучшения эффективности.
* Разделим признаки на X и y, где X - матрица признаков, y - предсказываемый вектор целевых значений.
* Разделим данные на обучающую и тестовую выборку.
* Стандартизация признаков с помощью StandardScaler.
* Создание и обучение модели GB, прогнозирование значений для тестового набора данных, вычисление и вывод метрик качества.
* Вычисление макро-усредненного ROC-AUC, вычисление ROC-кривых, построение кривой ROC для каждого класса.

Используем макро-усреднение (Macro-Averaging), который вычисляет ROC-AUC независимо для каждого класса, после чего усредняет их.

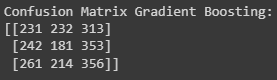


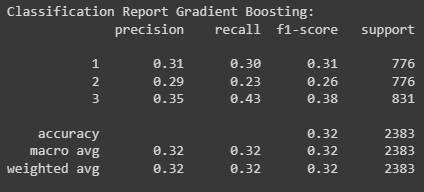
**KNN:**

* Используем GridSearchCV для настройки параметров прогнозирования на тестовых данных.
* Проведем оценку качества модели.

Вычислим матрицу ошибок и составим отчет о классификации:

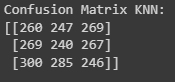
Gradient Boosting

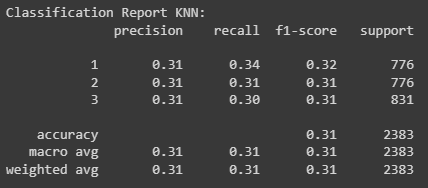




Точность около 31%, f1-score около 31%

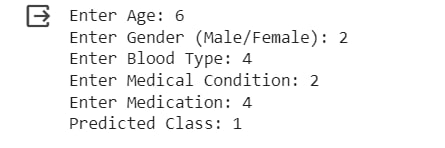
KNN

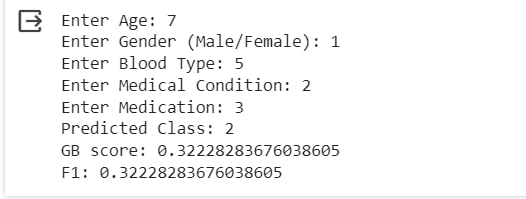




Точность около 32%, f1-score около 32%

Создали GUI:





**Заключение**

В ходе проекта было обучено 2 модели, решающих задачу предсказания эффективности лечения препаратом основе данных пациента. В результате, обе модели имеют схожую производительность. Точность и f1-score находятся на уровне примерно примерно 31-32%. Градиентный бустинг показывает немного лучшие результаты по recall для класса 3, но немного хуже для класса 2 по сравнению с K-ближайшими соседями. Исходя из этих данных, нельзя однозначно определить, какая модель лучше. Выбор между ними может зависеть от конкретных требований задачи, склонности к определенным типам ошибок и других факторов. Может быть полезным провести дополнительные эксперименты или оптимизировать параметры моделей.