Иванов Артем Евгеньевич, Низовцев Дмитрий Валерьевич

«Прогнозирование сердечно-сосудистых заболеваний»

1. Постановка задачи

Цель данного кейса - разработать модель, способную предсказывать наличие сердечно-сосудистых заболеваний у пациентов на основе различных медицинских показателей. Сердечно-сосудистые заболевания являются одной из ведущих причин смертности в мире, и их своевременное обнаружение может значительно повысить эффективность лечения и профилактики.

2. Описание датасета

Используемый датасет представляет собой набор данных о пациентах с различными медицинскими показателями. Датасет включает следующие поля:

- Общее здоровье
- Мед осмотр
- Рак кожи
- Другие виды рака
- Диабет
- Депрессия
- Возраст
- Пол
- ИМТ
- Упражнения
- Рост
- Вес
- Курение
- Алкоголь
- Употребление фруктов
- Наличие сердечно-сосудистых заболеваний (целевая переменная)

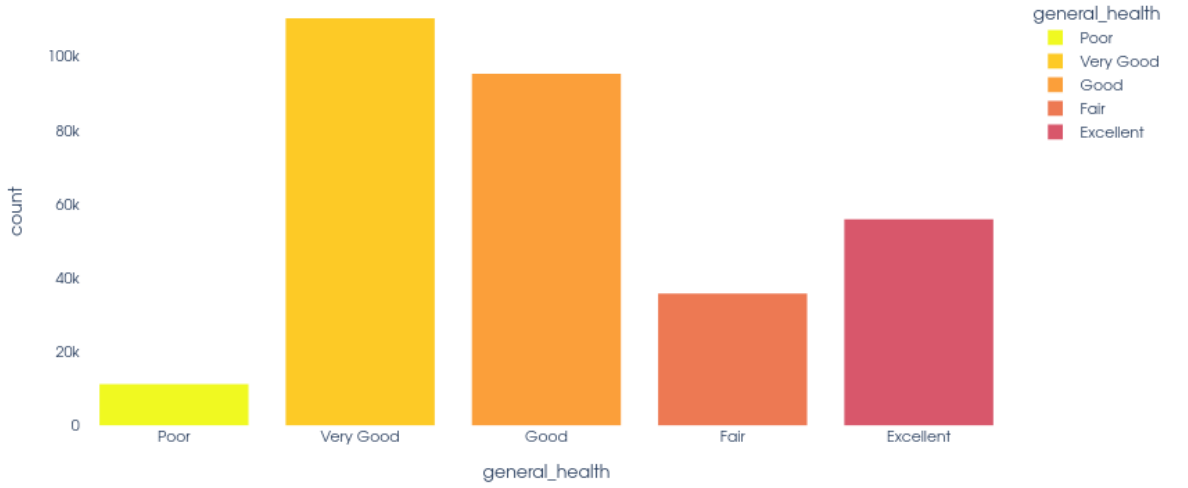
3. Ход работы

3.1. Гипотеза

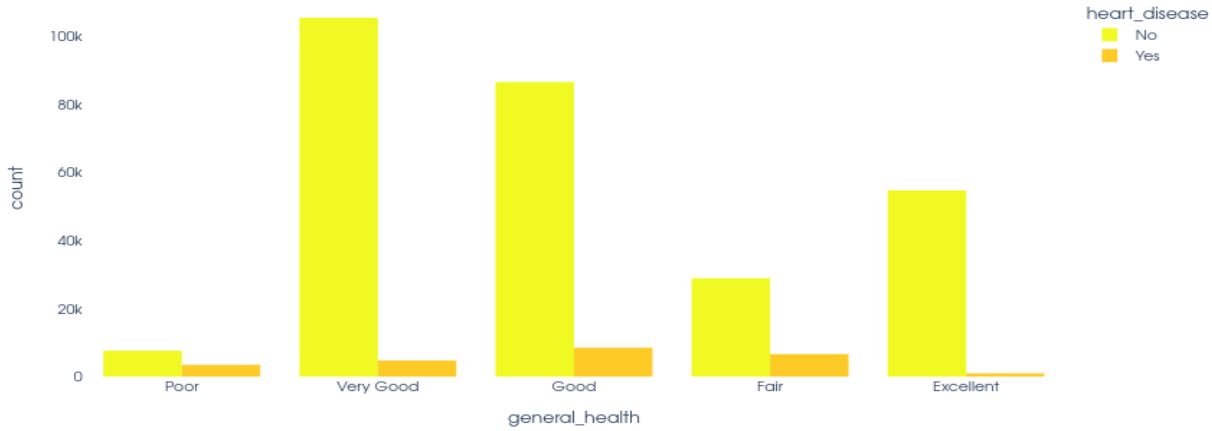
Предполагается, что наибольшее влияние на наличие сердечно-сосудистых заболеваний оказывают возраст, сахарный диабет, а также образ жизни (курение, употребление алкоголя, физическая активность). Основная задача – выделить эти ключевые признаки и построить модель, способную предсказывать наличие заболевания с высокой точностью.

3.2. Визуализация медицинских показателей

Провели общую оценку здоровья

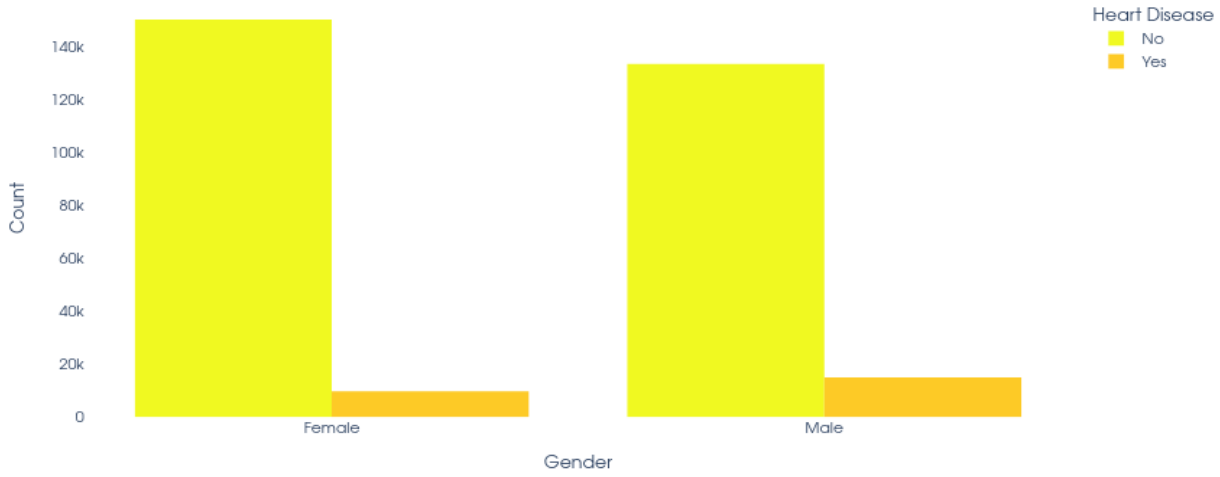


Также анализируем наличие проблем с сердцем в этих группах.

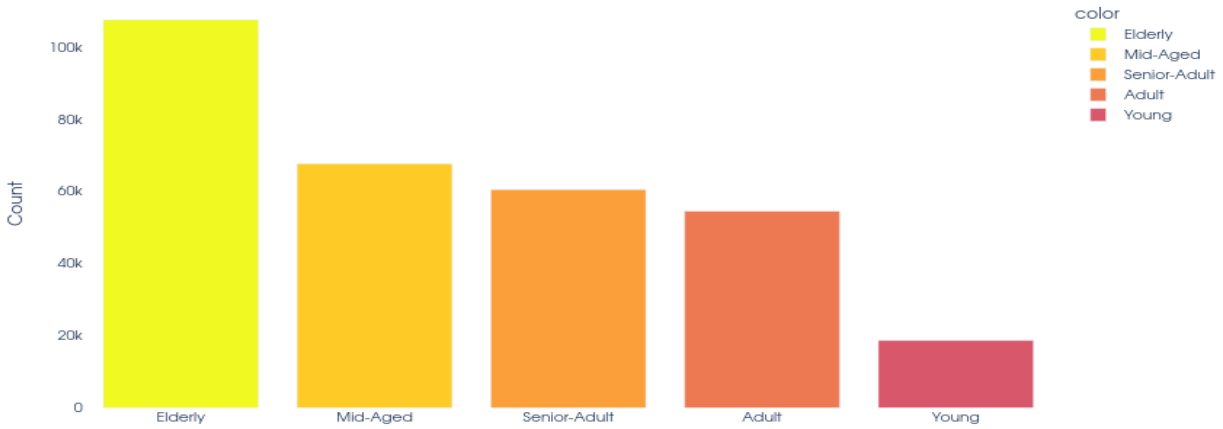


Провели демографический анализ

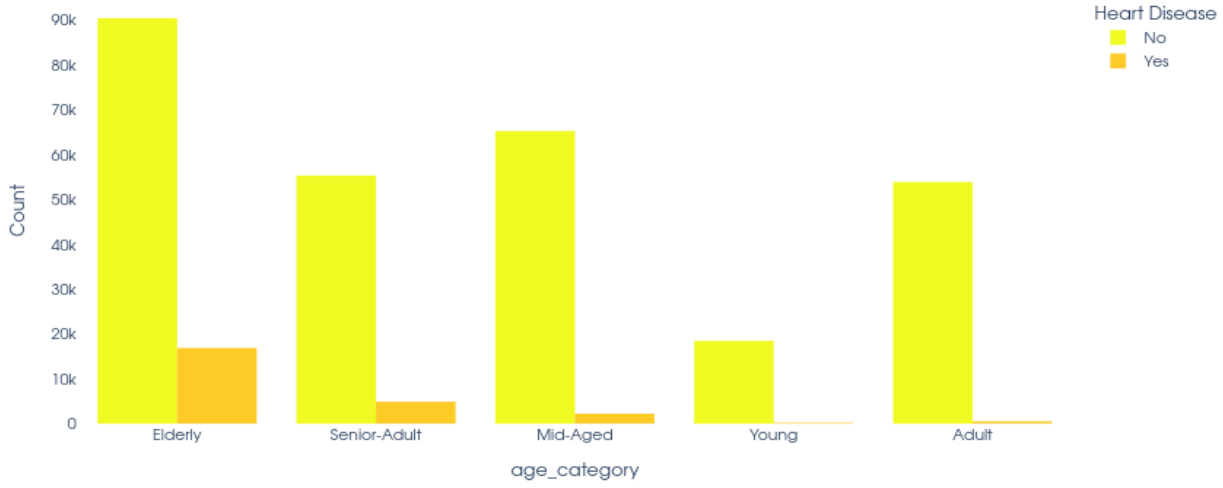
Далее узнаем наличие сердечных заболеваний у каждого пола



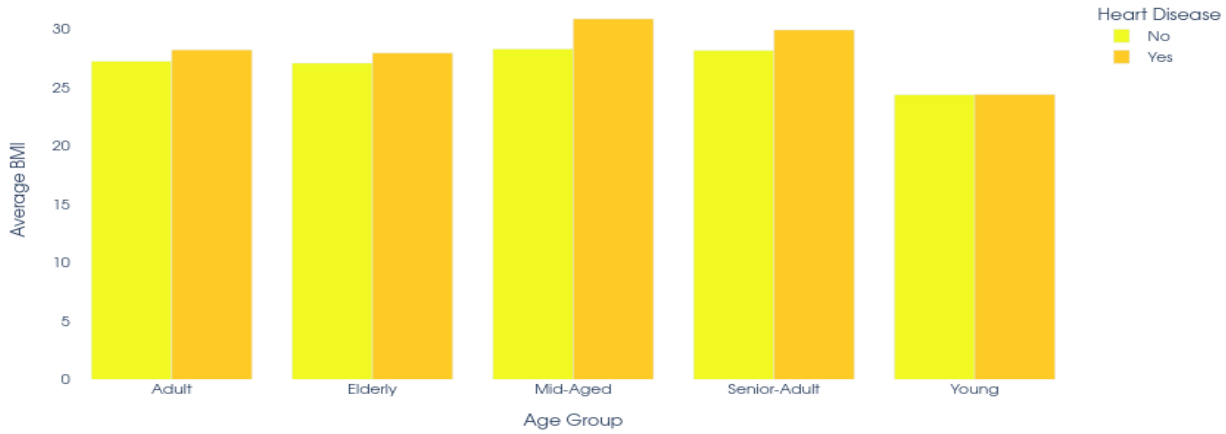
Провели анализ по возрастным данным



Далее узнаем, какая из возрастных категорий наиболее подвержена сердечным заболеваниям

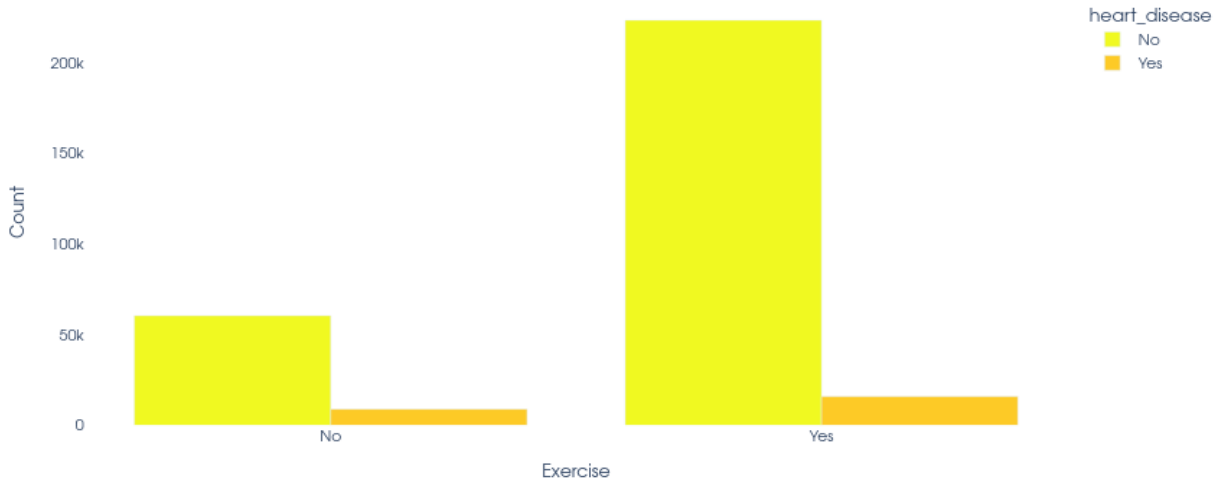


Проверка возрастных групп и их среднего ИМТ в зависимости от сердечно-сосудистых заболеваний

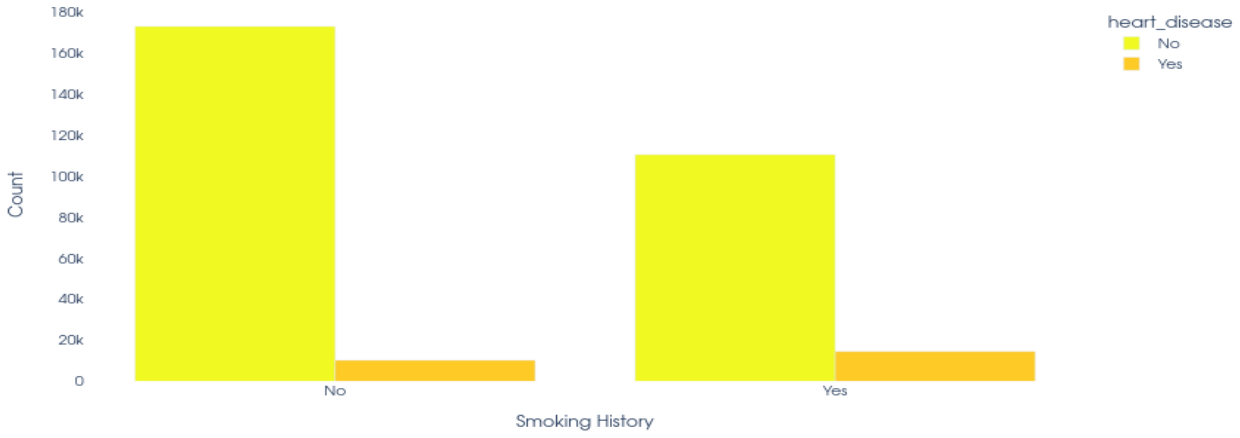


Провели анализ влияния образа жизни

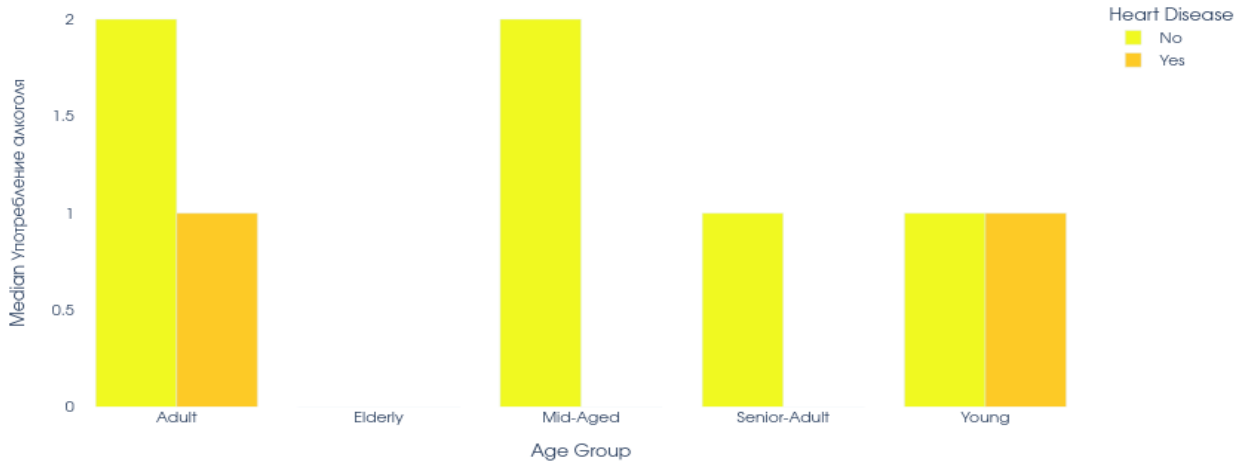
Влияния физических упражнений



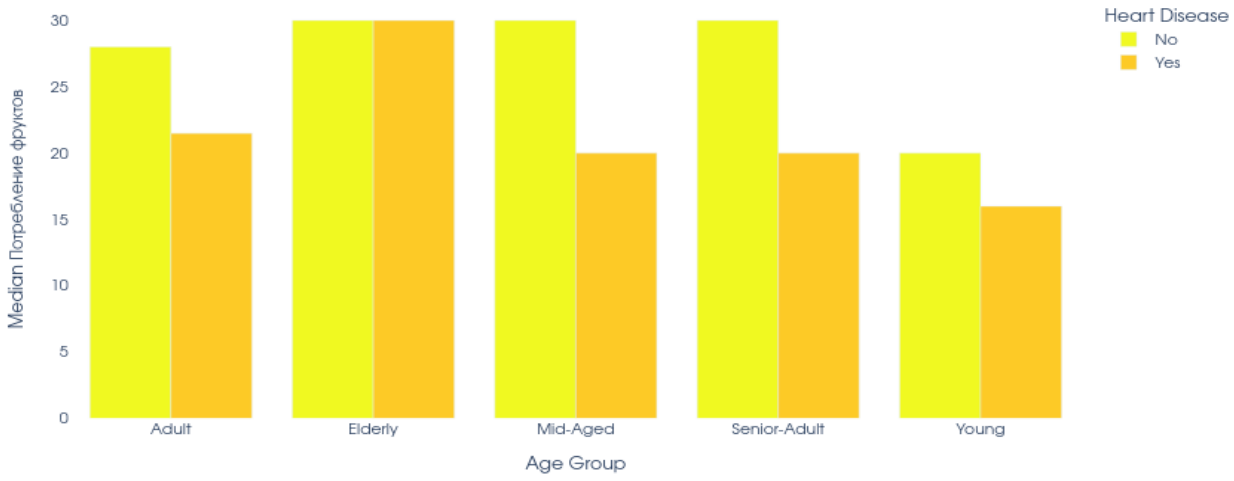
Влияние курения



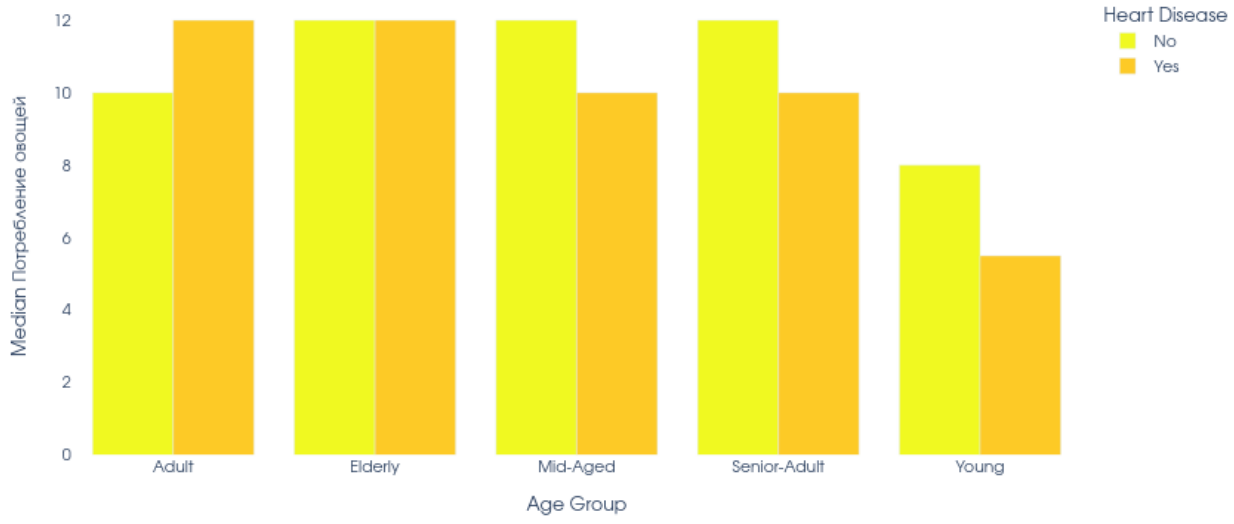
Влияние алкоголя



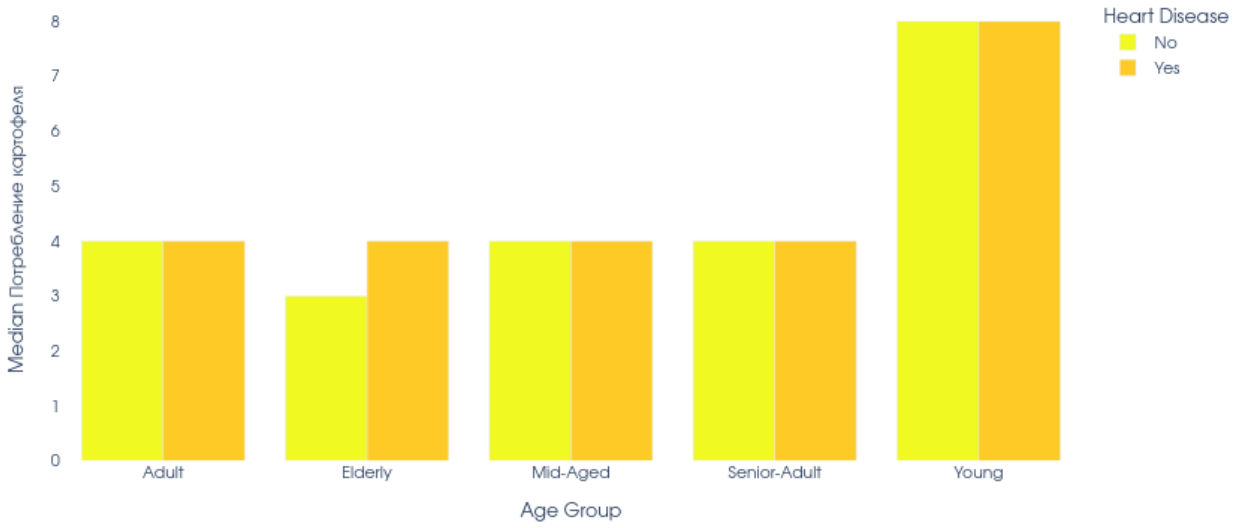
Употребление фруктов



Употребление овощей



Употребление картофеля



Выводы:

Анализ факторов риска показал, что люди со слабым здоровьем имеют больше шансов заболеть сердечно-сосудистыми заболеваниями.

Что касается пола, то данные показывают высокую долю мужчин с диагнозом сердечно-сосудистые заболевания, и у них немного более высокий имт по сравнению с женщинами

В зависимости от возрастных групп, у пожилых людей больше всего сердечно-сосудистых заболеваний. Однако у людей среднего возраста, страдающих сердечно-сосудистыми заболеваниями, как правило, ИМТ выше

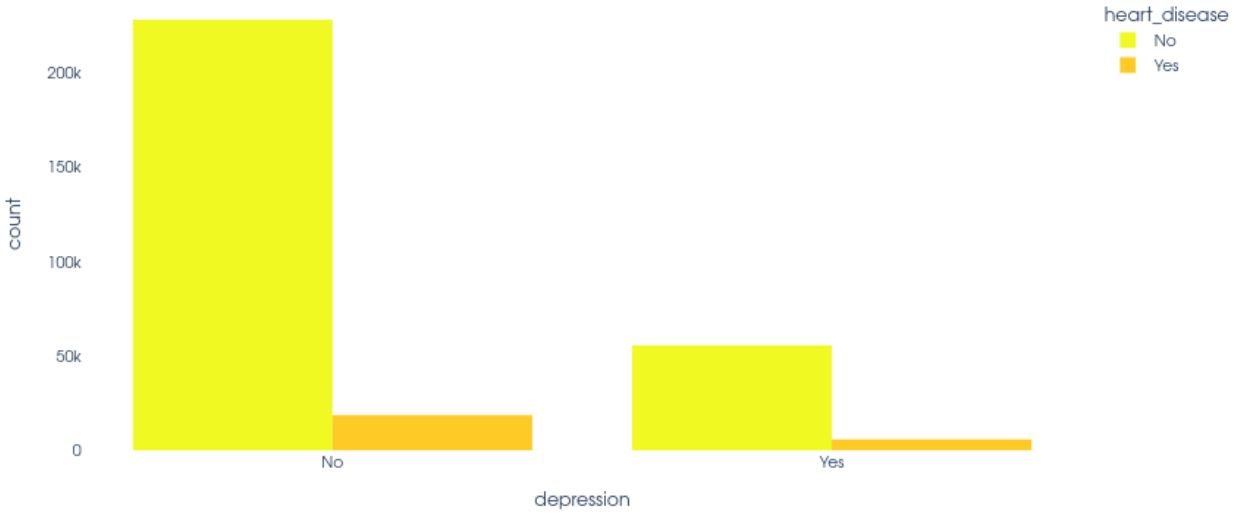
Что касается физических нагрузок, то это не играет существенной роли. Однако курение является важным фактором развития сердечно-сосудистых заболеваний.

Употребление алкоголя оказывает значительное влияние на сердечно-сосудистые заболевания, однако в молодом возрасте это не оказывает существенного влияния. Кроме того, анализ показал, что потребление фруктов не оказывает существенного влияния на людей с сердечными заболеваниями.

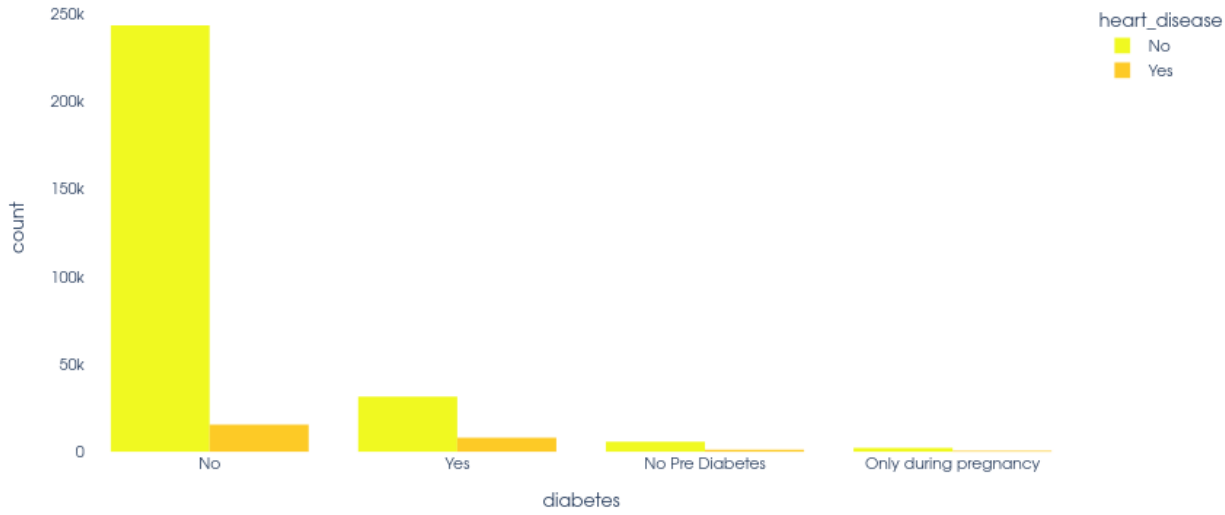
Аналогичным образом, потребление зеленых овощей оказывает значительное влияние на сердечно-сосудистые заболевания, особенно в зрелом возрасте. Кроме того, анализ показал, что употребление жареного картофеля никак не влияет на людей с сердечными заболеваниями.

3.3. Корреляционный анализ

Корреляция между депрессией и сердечными заболеваниями



Корреляция между диабетом и сердечными заболеваниями



Выводы:

Корреляционный анализ показал, что люди, страдающие депрессией, имеют больше шансов заболеть сердечно-сосудистыми заболеваниями.

Аналогичным образом, диабет оказывает значительное влияние на сердечно-сосудистые заболевания.

3.4. Модель прогнозирования

Для прогнозирования наличия сердечно-сосудистых заболеваний была использована логистическая регрессия, а также модель машинного обучения Random Forest .

Основные шаги включали:

1. Предобработку данных: удаление пропусков, нормализация признаков.
2. Разделение данных на обучающую и тестовую выборки.
3. Обучение моделей на обучающей выборке.
4. Оценка точности моделей на тестовой выборке.

Для оценки качества работы алгоритма введем метрики precision (точность) и recall (полнота).
Precision можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющимися положительными, а **recall** показывает, какую долю объектов положительного класса из всех объектов положительного класса нашел алгоритм.

Так как метрики друг от друга не зависят, вводим F-меру, которая поможет найти оптимальный баланс между метриками.

F-мера – среднее гармоническое precision и recall.

Результаты моделей:

- Логистическая регрессия: AUC = 0.72

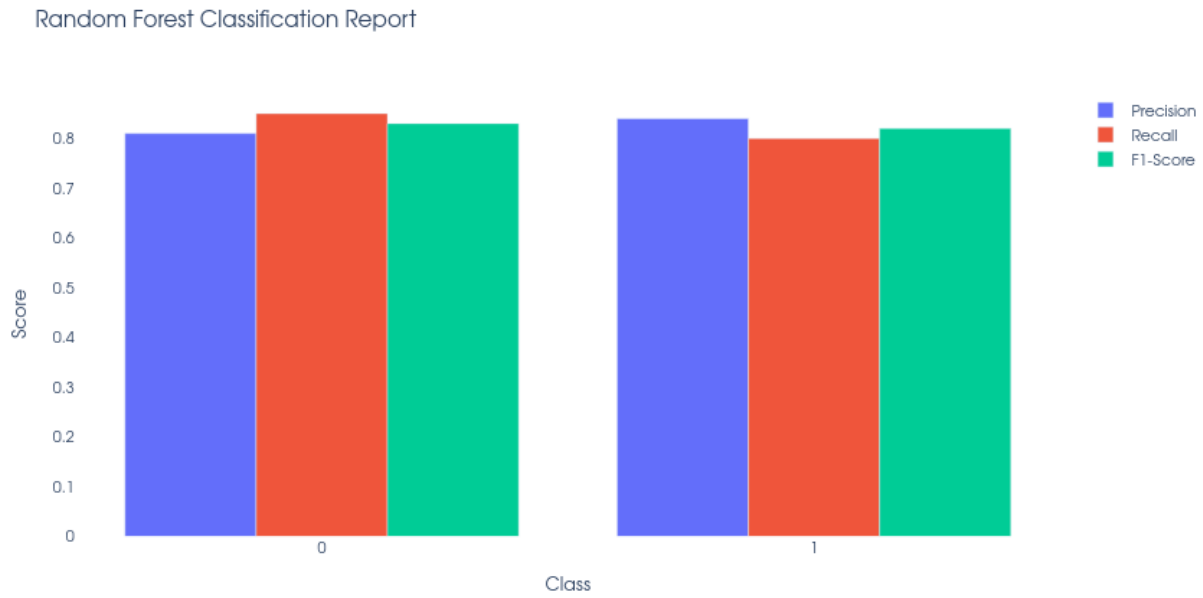
  Отчет о работе модели:

```
======================================== Logistic regression report: ========================================

              precision    recall  f1-score   support

           0       0.72      0.72      0.72     85071
           1       0.72      0.72      0.72     85259

    accuracy                           0.72    170330
   macro avg       0.72      0.72      0.72    170330
weighted avg       0.72      0.72      0.72    170330
```

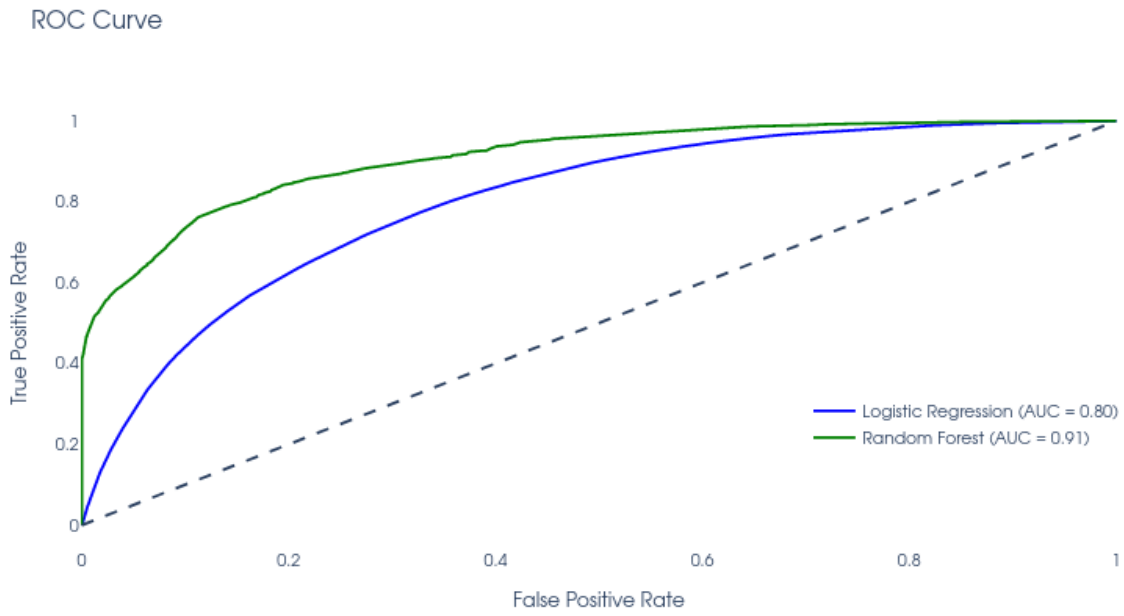
- Random Forest: AUC = 0.83

  Отчет о работе модели:

```
====================================== Random forest report: ========================================

              precision    recall  f1-score   support

           0       0.81      0.85      0.83     85071
           1       0.84      0.80      0.82     85259

    accuracy                           0.82    170330
   macro avg       0.82      0.82      0.82    170330
weighted avg       0.82      0.82      0.82    170330
```



Сравнение результатов с помощью ROC кривых:



Чем больше площадь под кривой, тем лучше работает модель. Легко заметить, какая из моделей оказалась лучше.

4. Выводы

В ходе работы были выполнены следующие шаги:

- Проведен анализ и предобработка данных.
- Построены визуализации распределения медицинских показателей.
- Обучены модели для прогнозирования сердечно-сосудистых заболеваний.
- Модель RandomForest показала наилучший результат с AUC = 0.83.

Данные результаты могут быть использованы для раннего выявления сердечно-сосудистых заболеваний и проведения профилактических мероприятий.