**Документация**

**Дипломная работа по теме:**

**“Анализ данных о сердечно-сосудистых заболеваниях**

**(поиск инсайтов, составление рекомендаций стейкхолдерам,**

**построение модели классификации наличия заболевания)”**

**Профессия “Аналитик данных”, DAU-65**

**Степаненко Ольга Викторовна**

**г. Санкт-Петербург, 2024 г.**

[**Введение 3**](#_3zinw1ek8298)

[**Блок 1. Описание исходного датасета и типов данных 4**](#_s7ed1s1v586g)

[**Блок 2. Подготовка и очистка данных 5**](#_z92ck3mifizc)

[**2.1 Подготовка данных, поиск аномалий 5**](#_4f1gghih5ijt)

[**2.2 Итоговый датасет с изменениями и выводами по очистке данных 5**](#_spv4ltyvj0v6)

[**Блок 3. Анализ данных 6**](#_m21i0kw487x7)

[**3.1 Однофакторный анализ. Влияние на целевую переменную разных факторов риска 6**](#_ifa3us5uijfm)

[**3.1.1 Выводы и рекомендации по однофакторному анализу 11**](#_yh2gv1z78z4h)

[**3.2 Матрица корреляции Спирмена 12**](#_rcfbls8z0flr)

[**Блок 4. Построение модели логистической регрессии 13**](#_8r6db3feg88v)

[**4.1 Матрица ошибок 14**](#_ni3yfuabkqaj)

# 

# **Введение**

**Цели проекта:**

В рамках набора данных о сердечно-сосудистых заболеваниях, собранных по результатам объективных обследований и субъективной информации, предоставленной пациентом, провести исследование данных, найти взаимосвязь между наличием сердечно-сосудистых заболеваний и факторами риска. Дать рекомендации стейкхолдерам, построить модель классификации наличия заболевания.

**Задачи:**

1. Подготовить данные для исследования.
2. Выявить какие факторы риска имеют наибольшее влияние на заболевание сердечно-сосудистой системы для предотвращения заболеваемости.
3. Построить модель классификации наличия заболевания.

**Стейкхолдеры:**

1. Медицинские работники, которые проводят санитарно - просветительную работу.
2. Пациенты.

# **Блок 1. Описание исходного датасета и типов данных**

Для исследования был взят датасет “[Cardiovascular Disease dataset](https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset)” с набором данных о сердечно-сосудистых заболеваниях.

| № | Имя столбца | Описание | Единицы измерения | Тип данных |
| --- | --- | --- | --- | --- |
| 1 | id | Идентификатор |  |  |
| 2 | age | Возраст | Дни | int64 |
| 3 | gender | Пол  1-женщины  2-мужчины |  | int64 |
| 4 | height | Рост | см | int64 |
| 5 | weight | Вес | кг | float64 |
| 6 | ap\_hi | Систолическое давление |  | int64 |
| 7 | ap\_lo | Диастолическое давление |  | int64 |
| 8 | cholesterol | Уровень холестерина  1-в норме  2- выше нормы  3-значительно выше нормы |  | int64 |
| 9 | gluc | Уровень глюкозы  1-в норме  2- выше норм  3-значительно выше нормы |  | int64 |
| 10 | smoke | Курение |  | int64 |
| 11 | alco | Употребление алкоголя |  | int64 |
| 12 | active | Физическая активность |  | int64 |
| 13 | cardio | Наличие или отсутствие сердечно-сосудистого заболевания (целевая переменная) |  | int64 |

# **Блок 2. Подготовка и очистка данных**

## **2.1 Подготовка данных, поиск аномалий**

Столбец “ age” переведен из дней в годы.

По статистическим данным датасета обнаружены аномальные значения в столбцах “ap\_hi”, “ap\_lo”, “height”, “weight”, в том числе и отрицательные.

Произведен анализ выбросов по перечисленным выше столбцам, подсчитана мода и среднее значение для решения о замене аномальных значений.

## **2.2 Итоговый датасет с изменениями и выводами по очистке данных**

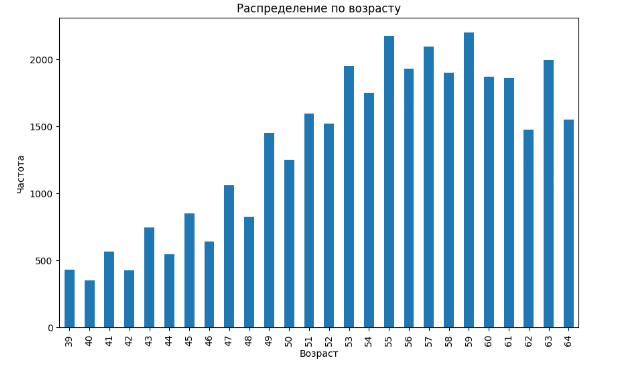
| № | Имя столбца | Очистка и преобразование данных |
| --- | --- | --- |
| 1 | id | Без изменений |
| 2 | age | Возраст пересчитан на количество лет без учета високосного года |
| 3 | gender | Без изменений |
| 4 | height | Рассчитаны выбросы и заменены на моду |
| 5 | weight | Рассчитаны выбросы и заменены на моду |
| 6 | ap\_hi | Отрицательные значения заменены по модулю, рассчитаны выбросы и заменены на моду |
| 7 | ap\_lo | Отрицательные значения заменены по модулю, рассчитаны выбросы и заменены на моду |
| 8 | cholesterol | Без изменения |
| 9 | gluc | Без изменения |
| 10 | smoke | Без изменения |
| 11 | alco | Без изменения |
| 12 | active | Без изменения |
| 13 | cardio | Без изменения |

# **Блок 3. Анализ данных**

В данном блоке будет проведен анализ данных и даны рекомендации стейкхолдерам.

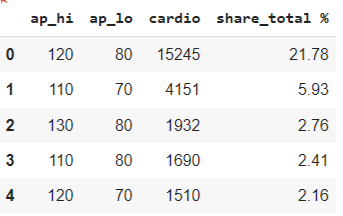
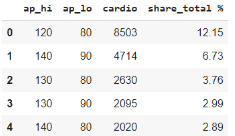
Стейкхолдерами являются врачи, медицинский персонал и конечное звено пациенты.

## **3.1 Однофакторный анализ. Влияние на целевую переменную разных факторов риска**



Из графика видно, что с возрастом увеличивается количество пациентов с заболеванием сердечно- сосудистой системы.

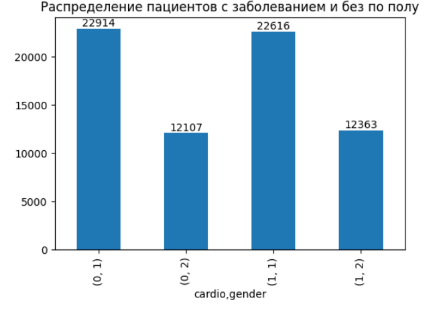
С заболеванием ССС Без заболевания ССС



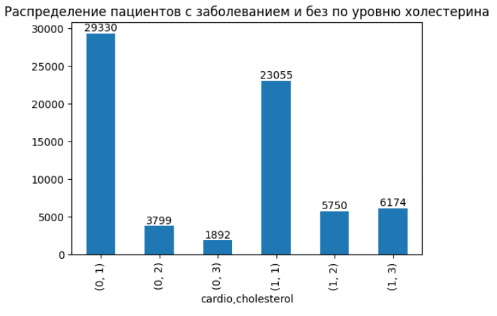
Таблицы отсортированы в порядке убывания по столбцу “cardio”.

При сравнении двух таблиц видно, что АД в норме (120/80) у пациентов без заболевания ССС почти в два раза встречается чаще, чем у пациентов имеющих то или иное заболевание ССС.

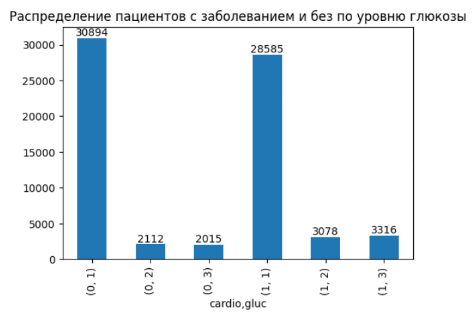
Также в таблице с заболеванием ССС видна динамика к повышенному АД (140/90, 140/80) . Это уже гипертоническая болезнь 1 ст.



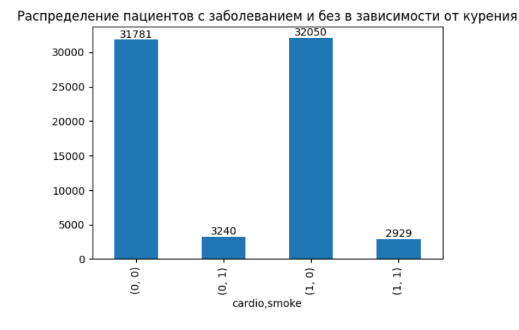
По гендерному типу различий в данном датасете нет.



По графику видно, что у пациентов с повышенным уровнем холестерина риск СС заболеваний выше, чем у пациентов с нормальным уровнем холестерина.

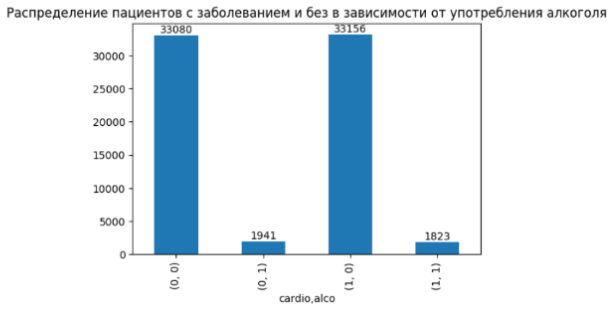


По графику видна зависимость наличия заболевания от повышенного уровня глюкозы. При сахарном диабете разрушается стенка сосудов, что приводит к серьезным патологиям сердечно -сосудистой системы.



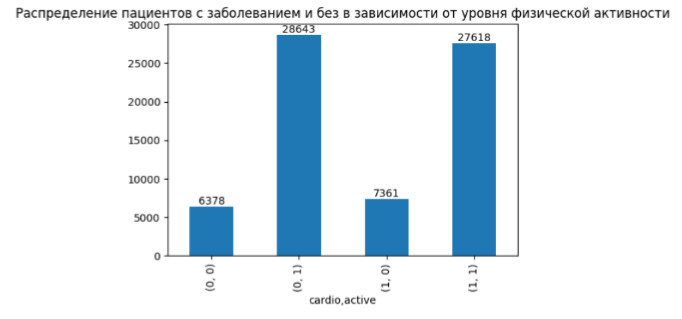
Курение как монофактор почти не оказывает влияния на заболевания сердечно-сосудистой системы, к тому же это субъективная характеристика.

Для полноценной оценки влияние данного фактора на заболеваемость сердечно-сосудистой системы нужно оценивать объективно стаж курения и количество выкуренных сигарет.



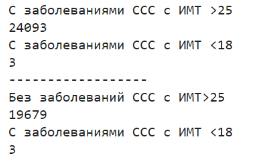
Употребление алкоголя как монофактор почти не оказывает влияния на заболевания сердечно-сосудистой системы, это также субъективная характеристика.

Для полноценной оценки влияние данного фактора на заболеваемость сердечно-сосудистой системы нужно оценивать объективно дозы употребляемого алкоголя и количество лет употребления.



Пациенты, у которых присутствует физическая активность менее подвержены заболеваниям сердечно- сосудистой системы, хотя данный фактор тоже является субъективным и оценка физической активности полноценно и объективно может изменить оценку данных.

Далее был подсчитан индекс массы тела (далее по тексту ИМТ) , так как проанализировать отдельно влияние роста и веса не объективно.



Из данных видно, что при повышенном ИМТ риск заболевания возрастает. 240093/19679\*100-100= 22,4%

### **3.1.1 Выводы и рекомендации по однофакторному анализу**

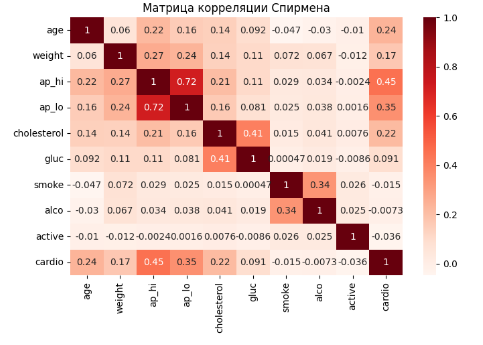
Исходя из выше представленного анализа данных, можно сделать вывод, что пол, употребление алкоголя и курение как отдельные факторы не оказывают прямого влияния на заболевания ССС. Но в данном датасете это достаточно субъективная оценка.

Такие факторы как глюкоза и холестерин выше нормы и значительно выше нормы, гиподинамия (отсутствие или сниженная физическая активность), возраст а также повышенный ИМТ оказывают значительное влияние на заболевание сердечно-сосудистой системы..

Рекомендации стейкхолдерам: медработникам, которые проводят санитарно- профилактические работы доводить до пациентов информацию о рисках заболевания ССС, проводить беседы о снижении массы тела, своевременном ежегодном обследовании на наличие сахарного диабета, повышенного уровня холестерина, рекомендовать занятия спортом, согласно возраста и наличия заболеваний.

## **3.2 Матрица корреляции Спирмена**

Построим матрицу корреляции Спирмена. Посмотрим есть ли взаимосвязь между всеми показателями относительно друг друга и в т.ч. заболеваемости.



1. Видно достаточно существенную взаимосвязь между систолическим и диастолическим АД , что вполне логично, т.к. эти два показателя непосредственно связаны.
2. Существует связь между целевой переменной (заболевания ССС) и АД (как систолическим так и диастолическим), что тоже вполне логично, т.к. эти показатели также непосредственно связаны.
3. Небольшая связь между холестерином и заболеваниями ССС
4. Также есть небольшая связь между возрастом и ССС заболеваниями.

## 

## 

## **Блок 4. Построение модели логистической регрессии**

1. Формируем данные

Из данных удаляем неинформативный столбец ID -идентификатор пациента

1. Делим данные на признаки и целевую переменную

Признаки (возраст, пол, рост, вес, систолическое давление, диастолическое давление, уровень глюкозы, уровень холестерина, уровень глюкозы, курение, употребление алкоголя, физическая активность)

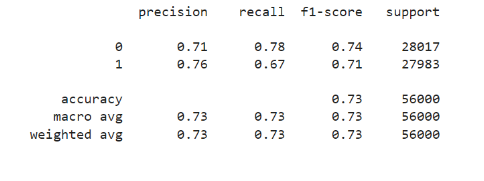
Целевая переменная (заболевание ССС)

1. Используем алгоритм логистической регрессии и обучаем модель

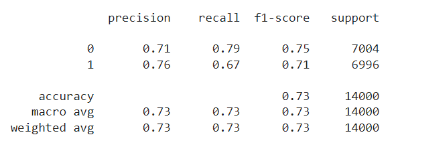
Использован алгоритм логистической регрессии так как предсказывается зависимая переменная на основании значений набора переменных- предикторов и зависимая переменная имеет два значения.

1. Оценка модели

тренировочная

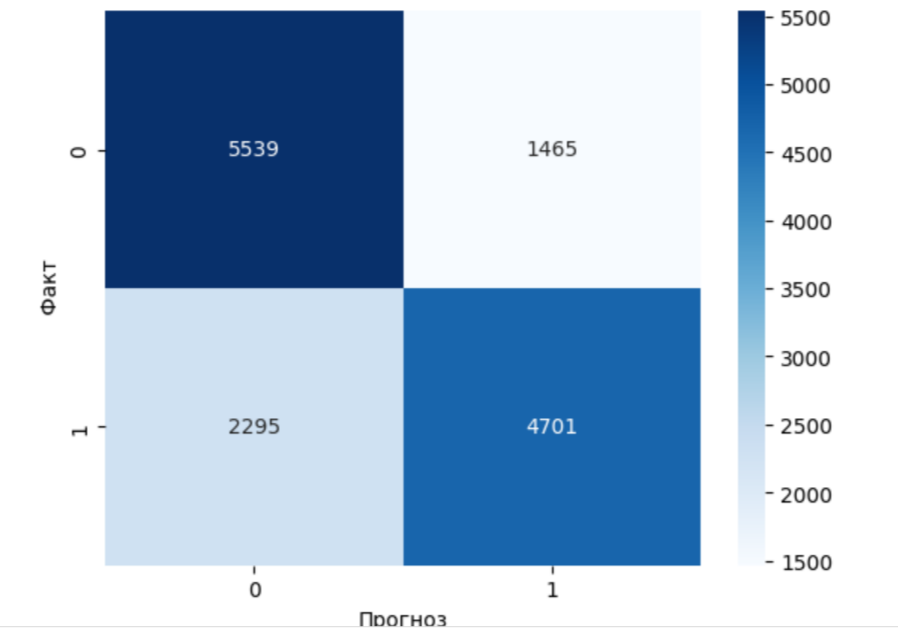


тестовая



Метрика accuracy составляет 0,73. Прогноз по заболеваемости на основании факторов риска достаточный,чтобы с уверенностью сказать, что при наличии факторов риска пациенты будут иметь СС заболевание. Модель можно считать работоспособной.

### **4.1 Матрица ошибок**



1. 4701 классифицированы как больные при наличие факторов риска и они такими являются
2. 5539 классифицированы как здоровые при наличии факторов риска и они такими являются
3. 1465 классифицированы как больные, но на самом деле таковыми не являются
4. 2295 классифицированы как здоровые, но на самом деле таковыми не являются.

**Итоги и выводы**

Сердечно-сосудистые заболевания (ССЗ) являются в настоящий момент одной из самых часто встречающихся причин смерти людей во многих странах мира.

В анализируемом датасете представлены основные факторы риска развития ССЗ, такие как: повышенный уровень глюкозы и холестерина, курение, употребление алкоголя, повышенный уровень АД и гиподинамия (отсутствие или низкая физическая активность).

Набор данных состоял из 70000 записей с субъективными и объективными факторами риска и целевой переменной (заболевания ССС).

В ходе анализа было выявлено, что провоцирующими факторами заболевания являются:

1. Возраст
2. Повышенный уровень холестерина
3. Повышенный уровень глюкозы
4. Гиподинамия
5. Повышенный ИМТ (ожирение)
6. Сочетание нескольких факторов риска с курением и алкоголем

Рекомендации стейкхолдерам:

Медицинским работникам, проводящим санитарно просветительскую работу:

1. Рассказывать пациентам с наличием факторов риска о негативных последствиях вредных привычек
2. Убеждать пациентов регулярно обследоваться для раннего выявления заболевания или его предотвращения.

Пациентам:

1. Ежегодные обследования , контроль анализа крови
2. Отказ от вредных привычек, особенно пациентам имеющим в анамнезе другие факторы риска такие как повышенный холестерин и повышенный уровень глюкозы
3. Активный образ жизни, предотвращение гиподинамии
4. При наличии факторов риска, таких как повышенный уровень глюкозы и холестерина: контроль и соблюдение рекомендаций врача(диета, терапия).

Сбор и анализ таких данных в медицине очень важен, профилактика заболеваний это самая главная цель. Опираясь на данное исследование, есть понимание на что нужно обратить внимание как медицинским работникам, так и пациентам.

Контроль веса, уровня глюкозы, холестерина и добавление в свой распорядок физической активности, может не только снизить риск заболевания, а при уже имеющемся заболевании не допустить его прогрессирования, но и вовсе предотвратить его.