**Документация**

**Дипломная работа по теме:**

**«Анализ данных о сердечно-сосудистых заболеваниях (поиск инсайтов, составление рекомендаций стейкхолдерам)»**

**Профессия “Аналитик данных”, DA-118**

**Сорокопуд Елизавета Григорьевна**

**г. Москва, 2025**

**Оглавление**

**Введение………………………………………………………………………………3**

**Блок 1. Описание исходного датасета и типов данных……..4**

**Блок 2. Подготовка и преобразование данных…………………..5**

**Блок 3. Очистка Данных………………………………………………………5**

**Блок 4. Анализ данных для стейкхолдеров………………………..8**

**4.1 Взаимосвязь ССЗ с биологическими признаками….……..…..8**

**4.1.1 Взаимосвязь ССЗ с показателем возраста…………..…………8**

**4.1.2 Взаимосвязь ССЗ с показателем пола……………….………….9**

**4.1.3 Взаимосвязь ССЗ с показателем роста…………..…………….10**

**4.1.4 Взаимосвязь ССЗ с показателем веса………………….………11**

**4.1.5 Взаимосвязь ССЗ с показателем ИМТ………………….………12**

**4.2 Взаимосвязь ССЗ с диагностическими признаками……..…13**

**4.2.1 Взаимосвязь ССЗ с показателем давления…………..………13**

**4.2.2 Взаимосвязь ССЗ с показателем холестерина…..………….14**

**4.2.3 Взаимосвязь ССЗ с показателем глюкозы………..………….15**

**4.3 Взаимосвязь ССЗ с образом жизни…………………………..………16**

**4.3.1 Взаимосвязь ССЗ с показателем курения………..…………..16**

**4.3.2 П Взаимосвязь ССЗ с показателем алкоголя………..……….17**

**4.3.3 Взаимосвязь ССЗ с показателем активного образа жизни…………………………………………………………………………………..18**

**Блок 5. Построение модели Логистической Регрессии…...19**

**5.1 Оценка качества прогнозной модели (RMSE)....................19**

**Итоги проекта и заключение………………………………………......21**

**Введение**

**Цели проекта:**

В рамках осуществления научно-исследовательской деятельности в лаборатории, специализирующейся на комплексном анализе медицинских показателей пациентов, необходимо произвести всестороннее исследование имеющихся в распоряжении организации информационных массивов.

Целью данного исследования является идентификация потенциальных рисков развития сердечно-сосудистых заболеваний на основании анализа анамнеза пациентов.

Полученные в результате исследования данные должны быть использованы для разработки научно обоснованных рекомендаций по профилактике сердечно-сосудистых заболеваний для ключевых заинтересованных сторон системы отчетности (стейкхолдеров).

**Задачи:**

1. Создание формы единого отчета на основе предварительно подготовленных и очищенных данных, что позволит организации создать единую иерархию метрик;

2. Изучить взаимосвязь избыточного веса и артериального давления с риском развития сердечно-сосудистых заболеваний для разработки профилактических мер, включающих коррекцию питания и повышение физической активности, что позволит своевременно предотвращать развитие патологий;

3. Создание прогностической модели на основе метода логистической регрессии, способной оценивать вероятность развития сердечно-сосудистых заболеваний у пациентов на основе анализа их медицинских показателей;

**Блок 1. Описание исходного датасета и типов данных (13 столбцов)**

Для исследования был взят датасет “[**Cardiovascular Disease dataset**](https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset?resource=download)” со статистикой сердечно-сосудистых заболеваний.

|  |  |  |  |
| --- | --- | --- | --- |
| **№** | **Имя Столбца** | **Описание** | **Тип данных** |
| ***1*** | ***Id*** | Номер пациента | int |
| ***2*** | ***age*** | Возраст пациента | int |
| ***3*** | ***gender*** | Пол пациента: 1: Женщина, 2: Мужчина. | int |
| ***4*** | ***height*** | Рост пациента | int |
| ***5*** | ***weight*** | Вес пациента | float |
| ***6*** | ***ap\_hi*** | Систолическое артериальное давление | int |
| ***7*** | ***ap\_lo*** | Диастолическое артериальное давление | int |
| ***8*** | ***cholesterol*** | Уровень холестерина: 1: нормальный, 2: повышенный, 3: высокий. | int |
| ***9*** | ***gluc*** | Уровень глюкозы: 1: нормальный, 2: повышенный, 3: высокий. | int |
| ***10*** | ***smoke*** | Наличие потребления никотина | int |
| ***11*** | ***alco*** | Наличие потребления алкоголя | int |
| ***12*** | ***active*** | Наличие физической нагрузки | int |
| ***13*** | ***cardio*** | Наличие или отсутствие сердечно-сосудистых заболеваний | int |

**Блок 2. Подготовка и преобразование данных**

В ходе исследования качества данных были сделаны следующие

изменения:

* Значения столбца “**age**” переведены из дней в года;
* В столбце “**weight** ” - Изменен тип данных на “**Int**”;

**Блок 3. Очистка аномалий в данных**

В процессе обработки данных выявлены аномальные значения показателей артериального давления: отрицательные значения, а также выходящие за физиологические пределы показатели в столбцах **ap\_hi** (за пределами диапазона 84–180 мм рт. ст.) и **ap\_lo** (за пределами диапазона 45–110 мм рт. ст.), что несовместимо с жизнедеятельностью пациентов или указывает на необходимость экстренной госпитализации. Дополнительно проведена верификация разницы между систолическим и диастолическим давлением (**ap\_hi > ap\_lo** с допустимым диапазоном 30–70 мм рт. ст.). Строки с отклонениями от установленных критериев были исключены из выборки на основании:

1. ВАРИАЦИИ СИСТЕМНОГО АРТЕРИАЛЬНОГО ДАВЛЕНИЯ ЧЕЛОВЕКА ([электронный ресурс](https://docs.yandex.ru/docs/view?tm=1754311555&tld=ru&lang=ru&name=posobie_Borodin.pdf&text=верхнее%20и%20нижнее%20давление%20научное%20исследование&url=https%3A%2F%2Fcosphys.rff.tsu.ru%2Fwp-content%2Fuploads%2F2014%2F11%2Fposobie_Borodin.pdf&lr=65&mime=pdf&l10n=ru&sign=32831bd7449326456a58c0ba6abfff61&keyno=0&nosw=1&serpParams=tm%3D1754311555%26tld%3Dru%26lang%3Dru%26name%3Dposobie_Borodin.pdf%26text%3D%25D0%25B2%25D0%25B5%25D1%2580%25D1%2585%25D0%25BD%25D0%25B5%25D0%25B5%2B%25D0%25B8%2B%25D0%25BD%25D0%25B8%25D0%25B6%25D0%25BD%25D0%25B5%25D0%25B5%2B%25D0%25B4%25D0%25B0%25D0%25B2%25D0%25BB%25D0%25B5%25D0%25BD%25D0%25B8%25D0%25B5%2B%25D0%25BD%25D0%25B0%25D1%2583%25D1%2587%25D0%25BD%25D0%25BE%25D0%25B5%2B%25D0%25B8%25D1%2581%25D1%2581%25D0%25BB%25D0%25B5%25D0%25B4%25D0%25BE%25D0%25B2%25D0%25B0%25D0%25BD%25D0%25B8%25D0%25B5%26url%3Dhttps%253A%2F%2Fcosphys.rff.tsu.ru%2Fwp-content%2Fuploads%2F2014%2F11%2Fposobie_Borodin.pdf%26lr%3D65%26mime%3Dpdf%26l10n%3Dru%26sign%3D32831bd7449326456a58c0ba6abfff61%26keyno%3D0%26nosw%3D1))/ ТГУ 2011 г. Стр. 41;
2. Диагностика и лечение артериальной гипертензии ([электронный ресурс](https://docs.yandex.ru/docs/view?tm=1754311939&tld=ru&lang=ru&name=rec01.pdf&text=верхний%20предел%20ад%20научное%20исследование&url=https%3A%2F%2Fmedinef.ru%2Fwp-content%2Fuploads%2F2011%2F05%2Frec01.pdf&lr=65&mime=pdf&l10n=ru&sign=fb9199164f416f951aa8992a3fbf4e2f&keyno=0&nosw=1&serpParams=tm%3D1754311939%26tld%3Dru%26lang%3Dru%26name%3Drec01.pdf%26text%3D%25D0%25B2%25D0%25B5%25D1%2580%25D1%2585%25D0%25BD%25D0%25B8%25D0%25B9%2B%25D0%25BF%25D1%2580%25D0%25B5%25D0%25B4%25D0%25B5%25D0%25BB%2B%25D0%25B0%25D0%25B4%2B%25D0%25BD%25D0%25B0%25D1%2583%25D1%2587%25D0%25BD%25D0%25BE%25D0%25B5%2B%25D0%25B8%25D1%2581%25D1%2581%25D0%25BB%25D0%25B5%25D0%25B4%25D0%25BE%25D0%25B2%25D0%25B0%25D0%25BD%25D0%25B8%25D0%25B5%26url%3Dhttps%253A%2F%2Fmedinef.ru%2Fwp-content%2Fuploads%2F2011%2F05%2Frec01.pdf%26lr%3D65%26mime%3Dpdf%26l10n%3Dru%26sign%3Dfb9199164f416f951aa8992a3fbf4e2f%26keyno%3D0%26nosw%3D1))/ Москва 2009 г. Стр. 2, таблица 1;

Осуществлено дополнение выборки новым параметром — **ИМТ** (индекс массы тела), который был рассчитан для нормализации анализируемой совокупности. Установлены граничные значения показателя с учётом трёхсигмового отклонения: минимальное значение определено как - **-3σ** (9,3), максимальное - **+3σ** (45,8). Данные корректировки позволили обеспечить репрезентативность выборки и повысить достоверность последующего статистического анализа. С основанием на источник:

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ ИНДЕКСА МАССЫ ТЕЛА СТУДЕНТОВ СПЕЦИАЛЬНОГО УЧЕБНОГО ОТДЕЛЕНИЯ В ПРОЦЕССЕ ИХ ФИЗИЧЕСКОГО ВОСПИТАНИЯ ([электронный ресурс](https://docs.yandex.ru/docs/view?tm=1754473747&tld=ru&lang=ru&name=Попкович.pdf&text=имт%20исследования&url=https%3A%2F%2Felib.bsu.by%2Fbitstream%2F123456789%2F55333%2F1%2F%25D0%259F%25D0%25BE%25D0%25BF%25D0%25BA%25D0%25BE%25D0%25B2%25D0%25B8%25D1%2587.pdf&lr=65&mime=pdf&l10n=ru&sign=c76b38824d9129d06e394f57a796f664&keyno=0&nosw=1&serpParams=tm%3D1754473747%26tld%3Dru%26lang%3Dru%26name%3D%25D0%259F%25D0%25BE%25D0%25BF%25D0%25BA%25D0%25BE%25D0%25B2%25D0%25B8%25D1%2587.pdf%26text%3D%25D0%25B8%25D0%25BC%25D1%2582%2B%25D0%25B8%25D1%2581%25D1%2581%25D0%25BB%25D0%25B5%25D0%25B4%25D0%25BE%25D0%25B2%25D0%25B0%25D0%25BD%25D0%25B8%25D1%258F%26url%3Dhttps%253A%2F%2Felib.bsu.by%2Fbitstream%2F123456789%2F55333%2F1%2F%2525D0%25259F%2525D0%2525BE%2525D0%2525BF%2525D0%2525BA%2525D0%2525BE%2525D0%2525B2%2525D0%2525B8%2525D1%252587.pdf%26lr%3D65%26mime%3Dpdf%26l10n%3Dru%26sign%3Dc76b38824d9129d06e394f57a796f664%26keyno%3D0%26nosw%3D1))/ Белорусский государственный университет, г .Минск, стр. 1, таблица 1;

Результат итогового вида обработанного датасета, а также выводы по причинам очистки данных приведены в таблице:

|  |  |  |  |
| --- | --- | --- | --- |
| **№** | **Новое имя**  **Столбца** | **Преобразование**  **данных** | **Очистка**  **данных** |
| ***1*** | ***Id*** | int | При предобработке данных из датасета было удалено 6454 строки с некорректными значениями показателей артериального давления (**ap\_hi**, **ap\_lo**) и **ИМТ**, а также избыточные данные, способные исказить результаты анализа. |
| ***2*** | ***age*** | int |
| ***3*** | ***gender*** | Int |
| ***4*** | ***height*** | int |
| ***5*** | ***weight*** | int |
| ***6*** | ***ap\_hi*** | int |
| ***7*** | ***ap\_lo*** | int |
| ***8*** | ***cholesterol*** | int |
| ***9*** | ***gluc*** | int |
| ***10*** | ***smoke*** | int |
| ***11*** | ***alco*** | int |
| ***12*** | ***active*** | int |
| ***13*** | ***cardio*** | int |
| ***14*** | ***IMT*** | float |

**Описание данных:**

**Основные показатели:**

* **Гендерный состав:** женщины — 65%, мужчины — 35%
* **Возрастная структура:** молодой — 15%, средний — 67%, пожилой — 18%
* **Сердечно-сосудистые заболевания:** присутствуют — 49%, отсутствуют — 51%

**Показатели здоровья:**

* **Холестерин:** нормальный — 76%, повышенный — 13%, высокий — 11%
* **Глюкоза:** нормальная — 85%, повышенная — 7%, высокая — 8%
* **Артериальное давление:** гипотензия — 4%, оптимальное — 13%, нормальное — 43%, высокое — 14%, гипертензия 1 ст. — 21%, гипертензия 2 ст. — 4%
* **ИМТ:** дефицит — <1%, недостаточный — <1%, нормальный — 37%, избыточный — 37%, ожирение 1 ст. — 17%, ожирение 2 ст. — 6%, ожирение 3 ст. — 2%

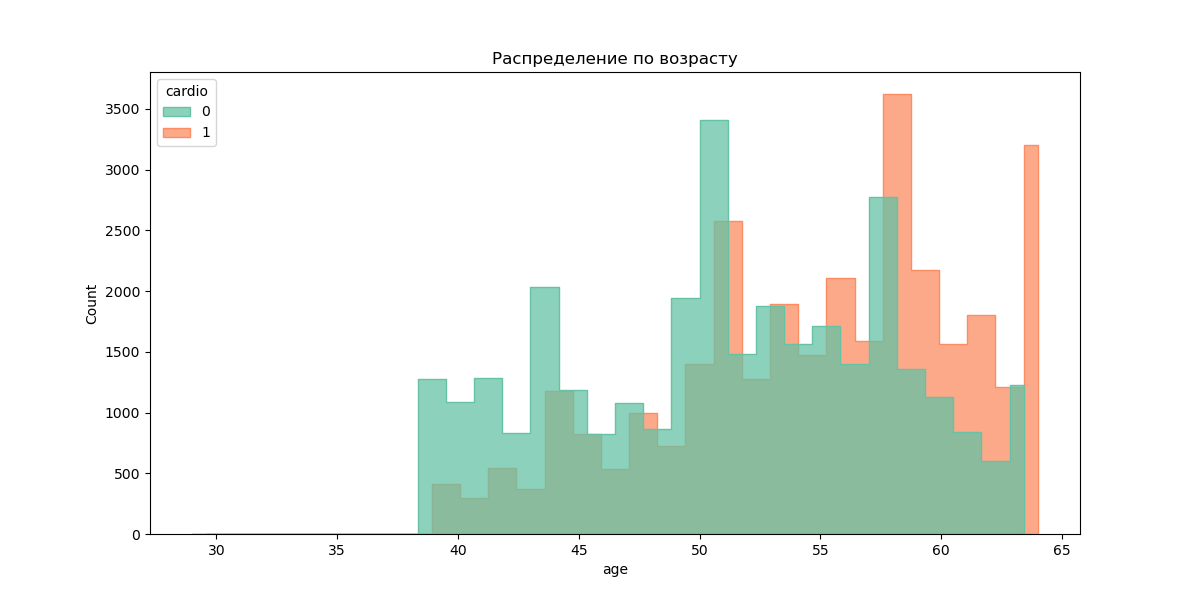
**Образ жизни:**

* **Курение:** некурящие — 91%, курящие — 9%
* **Алкоголь:** непьющие — 95%, пьющие — 5%
* **Физическая активность:** активные — 80%, неактивные — 20%

**Блок 4. Анализ данных для стейкхолдеров**

**4.1 Взаимосвязь С-СЗ с биологическими признаками**

**4.1.1 Взаимосвязь С-СЗ с показателем возраста (p-value = 0.0)**

****

**Изображение выглядит как текст, Шрифт, снимок экрана, число

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, диаграмма, Прямоугольник

Автоматически созданное описание**

**Вывод:**

Пациенты были стратифицированы по возрастным группам согласно классификации ВОЗ: молодые (18–44 года), средние (45–59 лет) и пожилые (60–75 лет). Анализ данных показал, что с возрастом повышается риск развития сердечно-сосудистых заболеваний, при этом переломный момент наблюдается около 55 лет, когда количество пациентов с ССЗ начинает превышать число здоровых, а в группе пожилых отмечается значительное преобладание пациентов с диагностированными сердечно-сосудистыми заболеваниями.

**4.1.2 Взаимосвязь С-СЗ с показателем пола**

**(p-value = 0.1778772419187124)**

**Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание**

**Вывод:**

Результаты демонстрируют, что как среди мужской, так и среди женской выборки пациентов распределение случаев ССЗ происходит без существенной зависимости от пола. Это означает, что гендерный фактор сам по себе не может рассматриваться как значимый предиктор риска развития сердечно-сосудистых заболеваний.

**4.1.3 Взаимосвязь С-СЗ с показателем роста**

**(p-value = 0.00590021921825544)**

**Изображение выглядит как текст, диаграмма, снимок экрана, карта

Автоматически созданное описание**

**Вывод:**

Проведённый графический анализ в сочетании с применением Т-критерия не выявил существенной корреляционной связи между антропометрическим показателем роста и вероятностью развития сердечно-сосудистых заболеваний. Наблюдаемая в ходе статистической обработки слабая взаимосвязь между ростом и ССЗ может быть обусловлена влиянием индекса массы тела на оба показателя, то есть носит косвенный характер, а не является прямым фактором риска.

**4.1.4 Взаимосвязь С-СЗ с показателем веса (p-value = 0.0)**

**Изображение выглядит как диаграмма, карта, снимок экрана, График

Автоматически созданное описание**

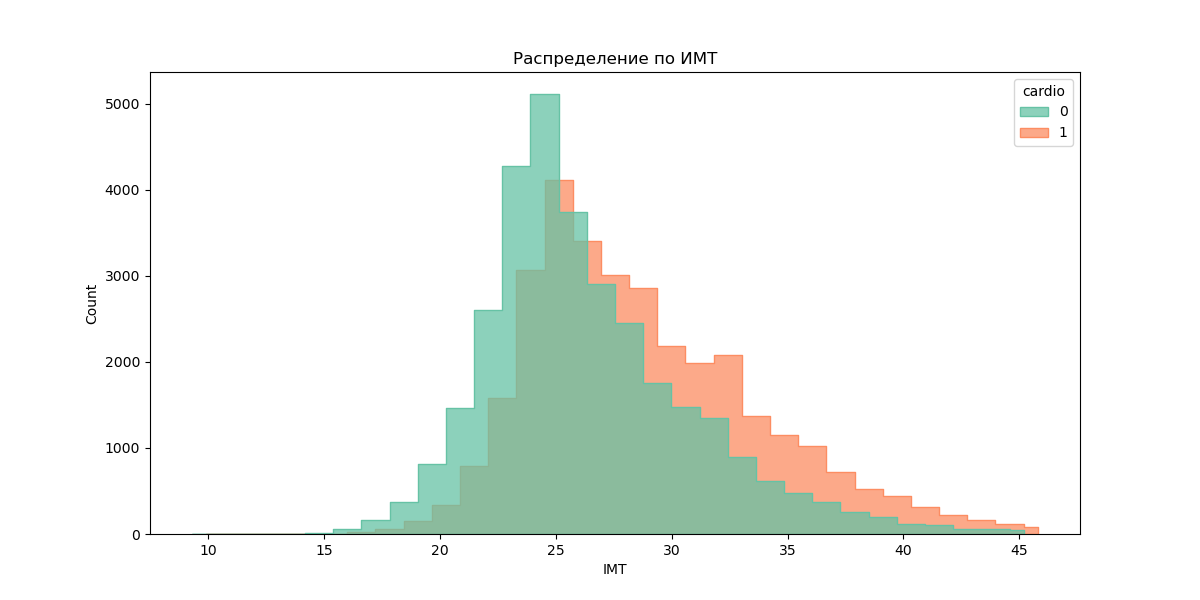
**Вывод:**

Статистический анализ данных (включая графический анализ и Т-критерий) выявил закономерность: среди пациентов с массой тела 80 кг и выше отмечается преобладание лиц с диагностированными сердечно-сосудистыми заболеваниями над здоровыми.

Тем не менее для корректной оценки влияния весового фактора на развитие ССЗ требуется расчёт индекса массы тела (ИМТ). Данный этап исследования планируется реализовать в дальнейшем, что позволит получить более точные и объективные результаты.

**4.1.5 Взаимосвязь С-СЗ с показателем ИМТ (p-value = 0.0)**

**H0:** Распределение риска сердечно-сосудистых заболеваний не отличается при различных значениях индекса массы тела.

**H1:** Распределение риска сердечно-сосудистых заболеваний отличается при различных значениях индекса массы тела.

Изображение выглядит как диаграмма, снимок экрана, График, линия

Автоматически созданное описание

Изображение выглядит как снимок экрана, текст, линия

Автоматически созданное описание

**Вывод:**

Анализ графических данных и т-критерия выявил прямую корреляцию между индексом массы тела (ИМТ) и вероятностью диагностики сердечно-сосудистых заболеваний: при достижении показателя в 27 единиц количество пациентов с ССЗ начинает превышать число здоровых. Для упрощения диагностики ИМТ был распределён по следующим категориям: дефицит массы (≤16), недостаточный вес (16,1–17,9), нормальный вес (18,0–24,9), избыточный вес (25,0–29,9), ожирение 1 степени (30,0–34,9), ожирение 2 степени (35,0–39,9) и ожирение 3 степени (≥40,0). Согласно графику №2, уже при достижении избыточного веса (от 25 единиц ИМТ) наблюдается значительное преобладание пациентов с диагностированными сердечно-сосудистыми заболеваниями.

**4.2 Взаимосвязь С-СЗ с диагностическими признаками**

**4.2.1 Взаимосвязь С-СЗ с показателем давления (p-value = 0.0)**

**H0:** Распределение риска сердечно-сосудистых заболеваний не отличается при различных показателях артериального давления.

**H1:** Распределение риска сердечно-сосудистых заболеваний отличается при различных показателях артериального давления.

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание **Изображение выглядит как текст, диаграмма, снимок экрана, График

Автоматически созданное описание**

**Вывод:**

Артериальное давление было классифицировано согласно критериям из источника (пункт 3) по следующим группам: гипотензия (до 100 мм рт. ст. верхнее давление), оптимальное (101–119 мм рт. ст.), нормальное (120–129 мм рт. ст.), высокое (130–139 мм рт. ст.), гипертония 1 степени (140–159 мм рт. ст.) и гипертония 2 степени (160–180 мм рт. ст.).

Графическая визуализация данных и т-критерий показывают, что уже в группе с высоким давлением количество пациентов с сердечно-сосудистыми заболеваниями значительно превышает число здоровых. Это подтверждает прямую взаимосвязь между показателями артериального давления и риском развития сердечно-сосудистых заболеваний.

**4.2.2 Взаимосвязь С-СЗ с показателем холестерина**

**(p-value = 0.0)**

**Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описаниеИзображение выглядит как текст, диаграмма, снимок экрана, Прямоугольник

Автоматически созданное описание**

**Вывод:**

Исследование, основанное на графическом анализе и Т-критерии, показало выраженную корреляцию между уровнем холестерина и частотой выявления сердечно-сосудистых заболеваний.

Анализ данных свидетельствует: чем выше показатели холестерина, тем чаще диагностируются ССЗ, что подтверждает значимость повышенного холестерина как фактора риска развития сердечно-сосудистых патологий.

**4.2.3 Взаимосвязь С-СЗ с показателем глюкозы**

**(p-value = 8.194708638433635e-106)**

**Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, диаграмма, Прямоугольник

Автоматически созданное описание**

**Вывод:**

Анализ показал, что при достижении уровня повышенной глюкозы в крови отмечается небольшое преобладание количества пациентов с диагностированными сердечно-сосудистыми заболеваниями над здоровыми, что указывает на наличие связи между гипергликемией и риском развития ССЗ.

**4.3 Взаимосвязь С-СЗ с образом жизни**

**4.3.1 Взаимосвязь С-СЗ с показателем курения**

**(p-value = 8.689123876667961e-06)**

**Изображение выглядит как текст, снимок экрана, Шрифт, белый

Автоматически созданное описание Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание**

**Вывод:**

Исследование с использованием графиков и Т-критерия показало наличие некой зависимости между курением и вероятностью развития ССЗ, но установить существенную корреляцию не удалось.

**4.3.2 П Взаимосвязь С-СЗ с показателем алкоголя**

**(p-value = 0.013566955101029248)**

**Изображение выглядит как текст, снимок экрана, Шрифт, белый

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, диаграмма, Прямоугольник

Автоматически созданное описание**

**Вывод:**

Статистический анализ (включая графический и Т-критерий) не выявил существенной взаимосвязи между употреблением алкоголя и риском развития сердечно-сосудистых заболеваний.

**4.3.3 Взаимосвязь С-СЗ с показателем активного образа жизни(p-value = 5.293096588842994e-21)**

**Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, диаграмма, Прямоугольник

Автоматически созданное описание**

**Вывод:**

Статистический анализ данных с использованием Т-критерия подтверждает наличие связи между уровнем физической активности и вероятностью развития сердечно-сосудистых заболеваний. Однако детальный анализ, выполненный с помощью графиков и таблиц, показывает, что различия между группами с разной физической активностью выражены незначительно.

Таким образом, хотя статистическая значимость связи подтверждена, практическая разница в риске развития ССЗ между группами с различным уровнем физической активности оказывается небольшой. Это указывает на то, что физическая активность является одним из факторов риска, но её влияние не является определяющим в развитии сердечно-сосудистых заболеваний

**Блок 5. Построение модели Логистической Регрессии**

В рамках исследования была разработана прогностическая модель для оценки риска развития сердечно-сосудистых заболеваний. Основной целью исследования стало создание эффективного инструмента прогнозирования на основе комплексного анализа медицинских и демографических показателей пациентов.

Для построения модели был сформирован датасет, включающий различные показатели пациентов. Данные прошли тщательную предварительную обработку, включающую очистку и нормализацию.

В качестве входных параметров модели использовались следующие показатели:

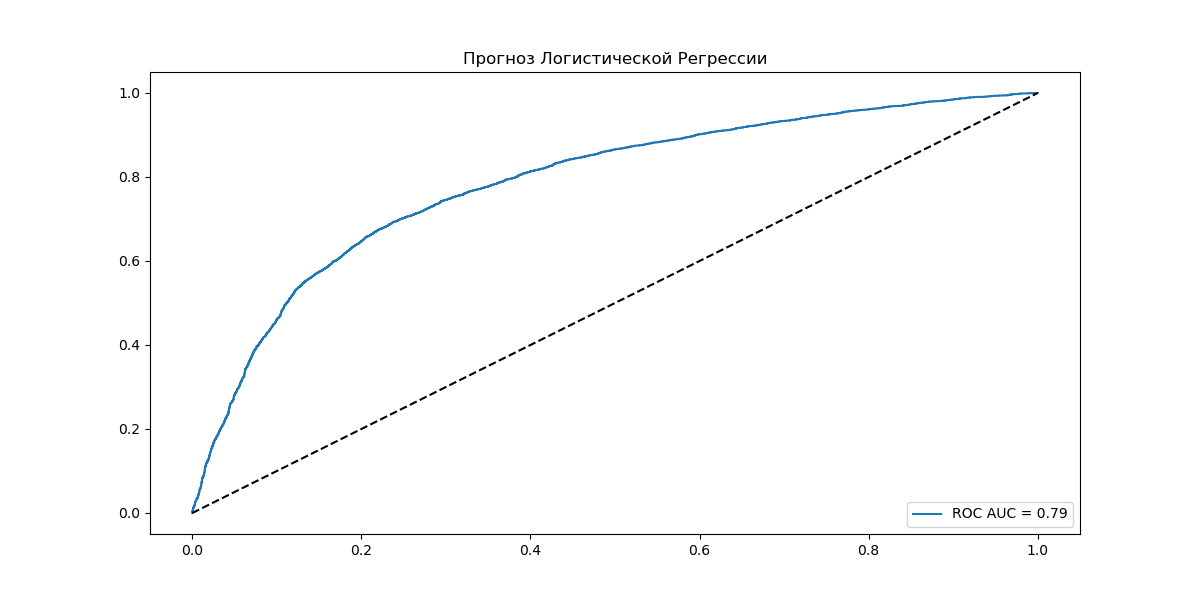
1. Биологические характеристики: пол (gender), возраст (age), рост (height), вес (weight), индекс массы тела (IMT);
2. Диагностические параметры: артериальное давление (ap\_hi и ap\_lo), уровень холестерина (cholesterol), уровень глюкозы (gluc);
3. Показатели образа жизни: курение (smoke), употребление алкоголя (alco), уровень физической активности (active);

Целевой переменной модели являлось наличие или отсутствие риска развития сердечно-сосудистых заболеваний (cardio).

Для построения прогностической модели был выбран метод логистической регрессии, который позволил эффективно оценить вероятность развития заболевания и проанализировать влияние каждого фактора на конечный результат.

**5.1 Оценка регрессионной модели прогнозирования ССЗ**

Оценка эффективности модели показала следующие результаты: точность на обучающей выборке составила 72% (модель не переобучалась), на тестовой выборке — 73%, а итоговая оценка значимости достигла 79%.



**Вывод:**

Результат моделирования позволяют выдвигать обоснованные предположения о возможном наличии сердечно-сосудистых патологий, что существенно облегчает процесс диагностики и помогает врачам принимать взвешенные решения о необходимости проведения дополнительных специализированных обследований.

Таким образом, созданная модель может стать важным инструментом в практике медицинских специалистов, способствуя более раннему выявлению сердечно-сосудистых заболеваний и своевременному назначению необходимого обследования пациентов.

**Итоги проекта и заключение**

Сердечно-сосудистые заболевания (ССЗ) занимают лидирующие позиции среди диагностируемых патологий в современной медицине, оставаясь одной из главных причин смертности во всем мире. Высокая распространенность данных заболеваний определяет постоянную актуальность их изучения и разработки методов ранней диагностики.

**По бизнес-задачам:**

1. В ходе подготовки аналитического отчета была осуществлена комплексная обработка исходных данных компании. Процедуры включали трансформацию и очистку датасета, в результате которой было исключено 6454 строки с некорректными значениями.

Выполненная предобработка данных позволила создать унифицированную систему метрик, полностью готовую к практическому применению в лабораторных условиях.

2. Исследование показало прямую корреляцию между избыточным весом, артериальным давлением и риском развития сердечно-сосудистых заболеваний (ССЗ). Анализ выявил статистически значимую связь (p-value = 0,00) как между показателями давления и ССЗ, так и между индексом массы тела (ИМТ) и риском развития патологий.

Полученные результаты подтверждают, что мониторинг артериального давления и ИМТ может эффективно использоваться врачами для раннего прогнозирования ССЗ и планирования профилактических мероприятий. Это позволяет рекомендовать включение этих показателей в стандартный скрининг пациентов.

3. Разработана прогностическая модель на базе метода логистической регрессии, продемонстрировавшая высокий уровень прогностической значимости (79%). Полученные результаты свидетельствуют о том, что созданная модель обладает достаточной точностью для оценки вероятности развития сердечно-сосудистых заболеваний у пациентов при анализе их медицинских показателей. Это позволяет использовать её как эффективный инструмент в клинической практике для раннего выявления рисков развития ССЗ.

**Рекомендации:**

В ходе анализа корреляций была выявлена недостаточность информативности данных для предоставления четких выводов, поэтому для улучшения качества данных необходимо доработать отчет и дополнить колонками со следующей информацией:

**Измерение уровня стресса:**

Позволяет провести комплексную оценку работы сердечно-сосудистой системы и своевременно выявить возможные нарушения в работе сердца, что крайне важно для ранней диагностики и профилактики сердечно-сосудистых заболеваний.

**Данные о хронических заболеваниях пациента:**

Помогают составить целостную картину его здоровья: они позволяют оценить предрасположенность к набору избыточного веса, выявить влияние имеющихся патологий на работу различных систем организма и объяснить возможные несоответствия в показателях, которые могут казаться нелогичными.

**Электрокардиограмма:**

Позволит выявить как явные, так и скрытые нарушения в работе сердечно-сосудистой системы. Исследование помогает обнаружить различные патологии сердца на ранних стадиях, включая нарушения ритма, ишемию и другие отклонения в работе сердечной мышцы.

**Генетическое тестирование:**

Позволит вывить наследственную предрасположенность к сердечно-сосудистым заболеваниям. Благодаря ему можно своевременно выявить склонность человека к ССЗ и отнести его к группе риска, что позволит начать профилактические меры до появления первых симптомов заболевания.

**Общий анализ крови**:

Позволит оценить ключевые показатели: ширину распределения эритроцитов, средний объём тромбоцитов и общее количество лейкоцитов. Эти параметры имеют существенное значение для раннего выявления и диагностики сердечно-сосудистых заболеваний. Использование данных показателей значительно ускоряет процесс обследования пациентов и повышает эффективность работы медицинского персонала.

**Общий анализ мочи:**

Позволит своевременно обнаружить маркеры нарушения работы почек при хронической сердечной недостаточности. Исследование помогает выявить патологические изменения в работе почечных сосудов и нефронов, что особенно важно при сердечно-сосудистых заболеваниях. Благодаря быстрому получению результатов врачи могут оперативнее ставить диагноз и назначать лечение, а медицинский персонал — эффективнее выполнять свою работу.

**Вывод:**

В результате проведённого исследования была достигнута главная цель — разработана прогностическая модель на основе метода логистической регрессии с высоким уровнем точности (79%), что подтверждает её эффективность в оценке вероятности развития сердечно-сосудистых заболеваний.

Статистический анализ убедительно продемонстрировал наличие прямой корреляционной зависимости между ключевыми факторами риска — артериальным давлением и индексом массы тела — и вероятностью развития сердечно-сосудистых заболеваний. Подтверждением достоверности полученных результатов служит значение p-value = 0,00, что указывает на наличие статистически значимой связи между исследуемыми параметрами.

Созданная модель позволяет эффективно оценивать риск развития ССЗ на основе анализа медицинских показателей пациентов, выявлять ключевые факторы риска и прогнозировать вероятность возникновения заболеваний. Практическая ценность исследования заключается в возможности использования разработанной модели для раннего выявления пациентов группы риска, планирования профилактических мероприятий и оптимизации стратегии ведения пациентов.

Полученные результаты свидетельствуют об эффективности выбранного подхода к прогнозированию ССЗ и создают основу для дальнейшего совершенствования прогностических моделей в кардиологии. Внедрение разработанной модели в клиническую практику позволит существенно повысить качество медицинской помощи и улучшить профилактику сердечно-сосудистых заболеваний.