

DAU-21 Аналитик данных с нуля до middle

**Дипломная работа**

Анализ данных о сердечно-сосудистых заболеваниях (поиск инсайтов, составление рекомендаций стейкхолдерам, построение предиктивной модели классификации наличия заболевания).

Королев Валерий

Москва

2022

**ОГЛАВЛЕНИЕ**

**Введение………………………………………………………………………………...2**

**ГЛАВА 1. ЗАДАЧИ ИССЛЕДОВАНИЯ……………………………………………..3**

**ГЛАВА 2. ОПИСАНИЕ ДАННЫХ…………………………………………………...4**

**ГЛАВА 3. ПОДГОТОВКА ДАННЫХ………………………………………………..5**

**ГЛАВА 4. АНАЛИЗ ДАННЫХ………………………………………………………..7**

**ГЛАВА 5. ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ МЕЖДУ ПРИЗНАКАМИ….8**

5.1. Взаимосвязь целевого признака cardio с другими признаками……………**8**

5.1.1. Диагностические показатели - cardio……………………………….**8**

5.1.2. Объективные показатели (ИМТ, возраст, пол) - cardio……………**10**

5.1.3. Поведенческие факторы - cardio……………………………………**11**

5.2. Взаимосвязь поведенческих и диагностических признаков………………**11**

5.3. Взаимосвязь ИМТ с поведенческими и диагностическими признаками...**11**

**ГЛАВА 6. ПОСТРОЕНИЕ ПРЕДИКТИВНОЙ МОДЕЛИ………………………..12**

**ЗАКЛЮЧЕНИЕ………………………………………………………………………..14**

**ВВЕДЕНИЕ**

По данным Всемирной организации здравоохранения (ВОЗ) сердечно-сосудистые заболевания (ССЗ) являются основной причиной смерти во всем мире, от которой по оценкам каждый год умирает 17,9 млн человек. Сердечно-сосудистые заболевания представляют собой группу болезней сердца и кровеносных сосудов. Более четырех из пяти смертей от ССЗ происходит в результате инфаркта и инсульта, причем треть из этих случаев смерти носит преждевременный характер и отмечается среди людей в возрасте до 70 лет.

Наиболее значительными поведенческими факторами риска болезней сердца и инсульта являются нездоровое питание, низкий уровень физической активности, употребление табака и вредное употребления алкоголя. Последствия поведенческих факторов риска могут проявляться в виде повышенного кровяного давления, повышенного уровня глюкозы или липидов в крови, избыточного веса или ожирения. Эти «промежуточные факторы риска» могут выявляться на уровне учреждений первичной медико-санитарной помощи и указывать на повышенный риск инфаркта, инсульта, сердечной недостаточности и других осложнений.

Доказано, что прекращение употребления табака, уменьшение потребления соли, потребление фруктов и овощей, регулярная физическая активность и отказ от вредного употребления алкоголя снижают риск развития сердечно-сосудистых заболеваний. Для стимулирования здоровых форм поведения важнейшее значение имеет проведение политики в области здравоохранения, обеспечивающей создание благоприятных условий для здорового образа жизни и повышающей его экономическую доступность.

Условием профилактики преждевременной смертности является выявление всех лиц, подверженных наиболее высокому риску ССЗ, и обеспечение предоставления им надлежащего лечения. Важнейшим условием предоставления необходимого лечения и консультирования всем, кому это необходимо, является обеспечение наличия в учреждениях первичного звена соответствующих лекарственных препаратов и базовых медицинских технологий, необходимых для диагностики и лечения неинфекционных заболеваний.

Если обратиться к статистике сердечно-сосудистых заболеваний за рубежом, то в США, Западной Европе, Австралии, Канаде наблюдается тенденция снижения смертности от ИБС благодаря своевременному диагностированию и улучшению методик выявления заболеваний на ранних стадиях. За счет регулярных обращений к врачу и применению эндоваскулярной хирургии удается увеличить продолжительность жизни на 10-15 лет.

По статистике ВОЗ от сердечно-сосудистых заболеваний в России в ближайшие 15 лет удастся спасти 4 миллиона человек за счет улучшившейся диагностики и применения мирового опыта. Однако в 50% случаев успешность лечения зависит от самого заболевшего. Если же больной будет игнорировать диспансеризацию, лечение и коррекцию вредных привычек, то врачи останутся бессильны, а шанс умереть преждевременно возрастет на 25%.

**ГЛАВА 1. ЗАДАЧИ ИССЛЕДОВАНИЯ**

Исследование проводится на основании данных, полученных по результатам диагностических обследований пациентов с признаками и симптомами сердечно-сосудистых заболеваний при их обращении в медицинские учреждения.

Задачами исследования являются:

* выявление влияния поведенческих факторов на параметры диагностических признаков и риск возникновения ССЗ.
* выявление взаимосвязи объективных физических признаков (пол, возраст, рост, вес) с диагностическими признаками и риском возникновения ССЗ.
* определение признаков, имеющих статистически значимую взаимосвязь с наличием ССЗ, построение модели прогнозирования возникновения ССЗ на основе имеющихся данных, оценка ее точности.

Результаты исследования будут интересны прежде всего в медицинской практике в части возможного влияния различных признаков на риск возникновения ССЗ, степени влияния и использования модели для улучшения ранней диагностики заболевания.

Как было указано ранее, снижение уровня заболеваемости может быть достигнуто за счет отказа от негативных поведенческих факторов, влияющих на возможность возникновения ССЗ. Поэтому результаты исследования могут быть интересны для государственных, общественных организаций, занимающихся пропагандой и развитием здорового образа жизни и физической активности населения.

**ГЛАВА 2. ОПИСАНИЕ ДАННЫХ**

Датасет сформирован по результатам диагностических обследований и содержит 70 000 записей.

Признаки можно разделить на следующие типы:

1. Фактические данные:

* **gender** - пол, категориальный признак, 1 - женский, 2 - мужской [[1]](#footnote-0)
* **age -** возраст, количественный признак, указан в днях[[2]](#footnote-1)
* **height -** рост, количественный признак, указан в см.
* **weight** - вес, количественный признак, указан в кг.

1. Поведенческие факторы:

* **smoke** - курение, категориальный признак, 1 - да, 0 - нет.
* **alco** - употребление алкоголя, категориальный признак, 1 - да, 0 - нет.
* **active** - физическая активность, категориальный признак, 1 - да, 0 - нет.

1. Диагностические данные:

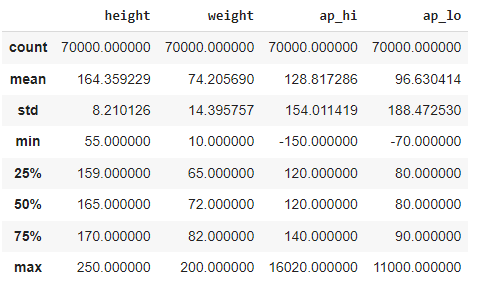
* **ap\_hi -** систолическое кровяное давление, количественный признак.
* **lo\_hi** - диастолическое кровяное давление, количественный признак.
* **cholesterol** - уровень холестерина в крови, категориальный признак, 1 - норма, 2 - выше нормы, 3 - значительно выше нормы.
* **gluc -** уровень сахара в крови, категориальный признак, 1 - норма, 2 - выше нормы, 3 - значительно выше нормы.

Целевая переменная **cardio -** сердечно-сосудистое заболевание, категориальный признак, 1 - да, 0 - нет.

В датасете нет пропусков данных, все признаки заполнены.

**ГЛАВА 3. ПОДГОТОВКА ДАННЫХ**

Данные признаков **height, weight, ap\_hi, ap\_lo**, судя по минимальным и максимальным значениям, требуют корректировки.



Для указанных выше признаков определим границы выбросов и проведем очистку данных. Рассчитаем границы выбросов, исходя из методики межквартильного диапазона с коэффициентом 1.5.

* **height**

Нижняя граница выбросов: женщины - 143.5, мужчины - 150.0.

Верхняя граница выбросов: женщины - 179.5, мужчины - 190.0.

Максимальный рост в данных у женщин 250 см, что можно считать аномальным и исключить. Остальные значения 198 см и менее считаем допустимыми. Максимальный рост у мужчин 207 см также считаем допустимым. Для установления нижней границы выбросов возьмем следующее медицинское определение: Карликовым ростом (карликовостью, нанизмом, дварфизмом) считается состояние, когда высота взрослого мужчины не превышает 130 см, а женщины – 120 см.

* **weight**

Нижняя граница выбросов: женщины - 37.5, мужчины - 42.5.

Верхняя граница выбросов: женщины - 105.5, мужчины - 110.5.

Максимальные значения веса в датасете 200 кг (по одному значению у женщин и мужчин). Количество людей, страдающих ожирением, в мире только растет. Поэтому вес свыше 100 кг нельзя считать аномальным. В контексте данного направления исследования излишний вес может быть одним из факторов, влияющих на возникновение ССЗ. Поэтому удаляем только данные, в которых вес меньше 37.5 кг у женщин и 42.5 кг у мужчин.

* **ap\_hi, ap\_lo**

Систолическое и диастолическое (верхнее и нижнее) кровяное давление являются связанными признаками, поэтому рассматриваем их вместе в поиске аномалий. Здесь аномалии ярко выражены: отрицательные значения и недопустимо высокие (превышение в 10-100 раз нормальные значения). Расчетные границы следующие:

Систолическое давление: нижняя - 90, верхняя - 170.

Диастолическое давление: нижняя - 65, верхняя - 105.

В медицинской практике критическим считается давление ниже 80 для систолического и ниже 50 для диастолического. Эти показатели и примем в качестве нижней границы допустимости данных.

Максимально зафиксированное высокое давление 310/220. В качестве верхней границы допустимости данным возьмем этот показатель, уменьшенный на 20%, т.е. 250 для систолического и 180 для диастолического давления. Кроме того, удалим записи, в которых **ap\_hi < ap\_lo**, что является недопустимым.

В итоге удалили 1 463 строк.

В данных содержится признак вес **weight,** играющий существенную роль в причинах возникновения ССЗ. Однако, абсолютное значение не всегда объективно, т.к. вес связан с ростом человека. Поэтому в медицине чаще используется относительный показатель Индекс массы тела (ИМТ) (англ. body mass index (BMI)). ИМТ позволяет оценить степень соответствия массы человека и его роста и тем самым, косвенно, оценить, является ли масса недостаточной, нормальной, избыточной (ожирение). Есть несколько методик расчета, возьмем самую распространенную (индекс Кетле):

**ИМТ = вес (кг) / (рост(м)) \*\* 2**

Введем признак **BMI**, рассчитанный по данной формуле.

**ГЛАВА 4. АНАЛИЗ ДАННЫХ**

После очистки данных итоговый датасет содержит информацию о 68 537 пациентах в возрасте от 39 до 64 лет. Половой состав: женщины - 65%, мужчины - 35%. Доля пациентов с диагностированным ССЗ составляет 49%. Заболеваемость среди мужчин и женщин практически одинакова: женщины - 49%, мужчины - 50%.

Рассмотрим заболеваемость в различных группах пациентов:



**Поведенческие признаки.** Большинство пациентов не курят, не употребляют алкоголь, физически активны. При этом заболеваемость среди них практически такая же, как в противоположной группе и в среднем по всем пациентам около 49%. Можно предположить, что в имеющихся данных влияние этих признаков на возникновение ССЗ будет небольшим.

**Диагностические признаки и ИМТ.** По отдельным признакам (холестерин, сахар) большинство пациентов имеют нормальные показатели, но при этом высокие уровни заболеваемости. Границы нормального артериального давления довольно условны, взяты исходя из возрастного состава пациентов в диапазоне 120 - 145. Высокий уровень заболеваемости в этой группе может быть связан с неблагоприятными показателями по другим признакам. Хотя доля каждого отдельного признака достаточно высокая, доля пациентов, имеющих нормальные показатели по всем этим признакам, составляет 4.9%. А высокий уровень заболеваемости может быть связан с наличием других факторов, не присутствующих в имеющихся данных. Большинство пациентов (95.1%) имеют хотя бы один признак, не соответствующий норме.

**ГЛАВА 5. ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ МЕЖДУ ПРИЗНАКАМИ**

Исследуем возможные зависимости между признаками:

* Взаимосвязь целевого признака **cardio** с другими признаками.
* Взаимосвязь поведенческих и диагностических признаков.
* Взаимосвязь веса с поведенческими и диагностическими признаками

Для этого рассчитаем корреляцию между признаками и оценим статистическую значимость этой зависимости. Для расчета корреляции воспользуемся следующими методами:

* Матрица корреляции Пирсона - для числовых признаков.
* Хи-квадрат статистика, тест Фишера - для категориальных признаков.
* Бисериальная корреляция - для числовых и категориальных признаков.

При оценке статистической значимости показателей исходим из уровня значимости alfa = 0.05.

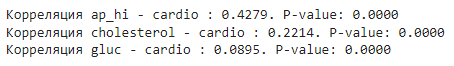
**5.1. Взаимосвязь целевого признака cardio с другими признаками.**

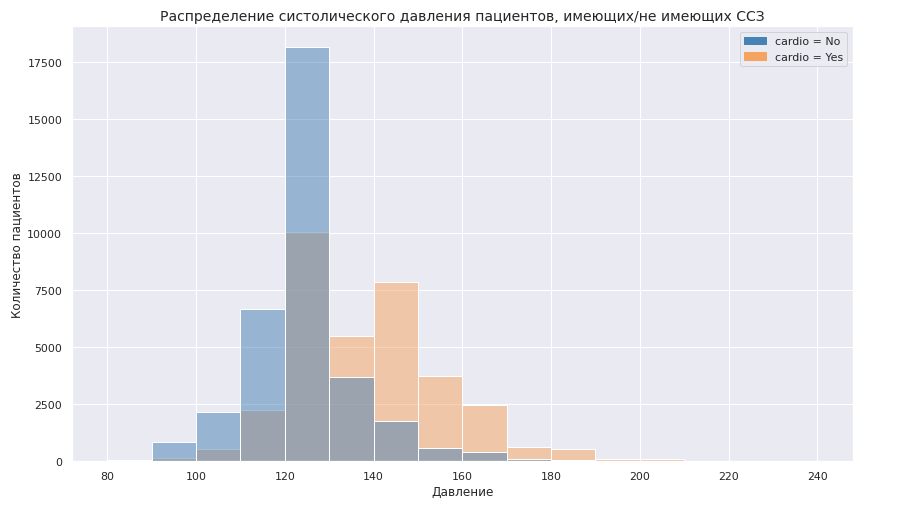
Определим корреляционную связь и статистическую значимость для дальнейшего использования при построении предиктивной модели.

**5.1.1. Диагностические показатели - cardio.**

Поскольку систолическое и диастолическое давление - два связанных признака и возникновение ССЗ связано, в первую очередь, с уровнем систолического давления, то в дальнейшем анализе будем рассматривать только этот признак.

Анализ показывает корреляционную зависимость возникновения ССЗ от значений диагностических параметров, подтверждается их статистическая значимость, т.е. P-value < alfa.:



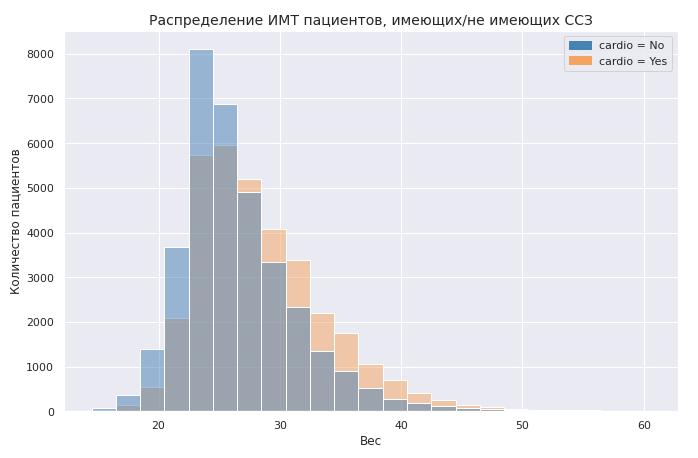


|  |  |
| --- | --- |

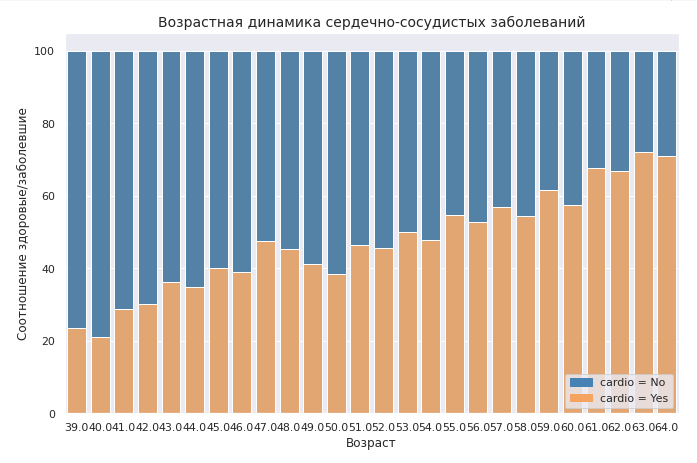
**5.1.2. Объективные показатели (ИМТ, возраст, пол) - cardio.**

Подтверждается статистическая значимость коэффициентов корреляции между ИМТ, возрастом и целевым признаком.

****



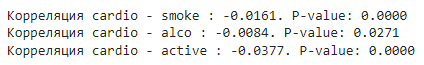




Влияние пола не является статистически значимым (p-value = 0.06).

**5.1.3. Поведенческие факторы - cardio.**

Корреляция присутствует, хотя значения невысокие. Статистическая значимость подтверждается. Отрицательные коэффициенты корреляции признаков **smoke** и **alco** c целевым признаком свидетельствуют о том, что в исследуемом наборе данных у пациентов, курящих и употребляющих алкоголь, меньшая склонность к ССЗ.



**5.2. Взаимосвязь поведенческих и диагностических признаков.**

Поведенческие признаки **smoke** и **alco** имеют положительные коэффициенты корреляции с диагностическими показателями, т.е. у пациентов, курящих и употребляющих алкоголь, более высокие уровни давления и холестерина (исключение: **smoke - gluc**), что, в свою очередь, способствует возникновению ССЗ.





**5.3. Взаимосвязь ИМТ с поведенческими и диагностическими признаками.**

Из объективных признаков рассмотрим отдельно взаимосвязь веса с поведенческими и диагностическими признаками. Вес во многом зависит от образа жизни, питания, а излишний вес (ожирение) является одной из причин возникновения различных заболеваний, в т.ч. сердечно-сосудистых и приводит к изменению различных диагностических признаков.

Подтверждается положительная корреляция ИМТ с диагностическими признаками:



Коэффициенты корреляции ИМТ с поведенческими признаками имеют небольшие значения, но связь статистически значима.



В части физической активности можно однозначно сказать, что большая активность способствует снижению веса. Отрицательный коэффициент корреляции с признаком smoke подтверждает факт увеличения веса у бросивших курить. В части алкоголя сложно определить причинную связь: алкоголь способствуют увеличению веса или просто среди полных людей больше употребляющих алкоголь.

**ГЛАВА 6. ПОСТРОЕНИЕ ПРЕДИКТИВНОЙ МОДЕЛИ**

Для решения задачи прогнозирования возникновения ССЗ (целевой признак **cardio**) воспользуемся несколькими моделями классификации и сравним их качество, используя функцию roc\_auc\_score.

Также построим для каждой модели Матрицу путаницы, рассчитаем показатели. Для данной тематики важным показателем является Recall (чувствительность), который показывает сколько заболевших определила модель из общего количества, имеющих заболевание

1. Логистическая регрессия.

Изменение набора зависимых признаков незначительно влияет на точность модели. При наборе признаков: BMI, ap\_hi, cholesterol, gluc, smoke, alco, active, age\_years.





1. Варьирование гиперпараметров модели.

Используем класс **GridSearchCV** для варьирования гиперпараметров “С” и “penalty” модели Логистической регрессии. Последующее обучение модели с параметрами С = 2, penalty = l2 не дали улучшения точности предсказания. Чувствительность ухудшилась.





1. Бэггинг для логистической регрессии.

Точность предсказания, чувствительность не меняются





1. Случайный лес.

Увеличение точности на обучающей выборке, на тестовой практически не изменилась. Чувствительность увеличилась.





1. Градиентный бустинг.

Увеличение точности на обучающей выборке, на тестовой не изменилась. Чувствительность увеличилась.





Разница между моделями не столь существенна в части точности предсказания на тестовой выборке 0.726 - 0.729, на обучающей более значимая 0.726 - 0.742.

Чувствительность 0.664 - 0.705, т.е. модель в лучшем случае определяет только 70% пациентов фактически имеющих заболевание. В то же время показатель правильности определения среди пациентов, не имеющих заболевание Специфичность : specificity имеет более высокие значения и изменяется в диапазоне 0.782 - 0.789.

Разница в значениях данных показателей связана с величиной порогового значения функций активации моделей. В задаче прогнозирования заболеваний лучше ошибиться, спрогнозировав заболевание у здорового человека (в дальнейшем отменить при медицинских обследованиях), чем не определить у действительно болеющего.

**ЗАКЛЮЧЕНИЕ**

В ходе проведенного исследования выполнена следующая работа:

* произведена очистка и подготовка данных.
* проведен анализ данных, описаны основные параметры имеющихся признаков, обозначены некоторые особенности данных.
* на основе расчета коэффициентов корреляции и оценки статистической значимости полученных коэффициентов определены взаимосвязи как между разными группами признаков, так и между признаками и целевой переменной **cardio.**
* построено несколько моделей прогнозирования возникновения ССЗ, оценена их точность, чувствительность (recall), специфичность (specificity).

На основании анализа имеющихся данных можно сделать следующие выводы:

* зафиксирована наибольшая взаимосвязь целевой переменной **cardio** с диагностическим признаками (артериальное давление, уровень холестерина, уровень сахара), объективными признаками (ИМТ, возраст).
* ИМТ имеет положительную корреляцию с диагностическими признаками, что подтверждает негативное влияние излишнего веса на кровяное давление, уровень холестерина и сахара.
* не выявлено негативного влияния поведенческих признаков **smoke** и **alco** на риск возникновения ССЗ. В то же время, существует небольшая положительная корреляция этих признаков с диагностическими признаками, что указывает на косвенное влияние этих признаков на возникновение ССЗ.
* доля пациентов, имеющих нормальные показатели по всем ключевым признакам, составляет 4.9%, но при этом высокий уровень заболеваемости 63.2%, что может быть связано с наличием других факторов, не присутствующих в имеющихся данных.
* наилучшие показатели прогнозирования возникновения ССЗ получены на модели градиентного бустинга: точность roc\_auc\_score 0.742 и 0.729, чувствительность recall 0.705, что в данной задаче вряд ли можно считать достаточным.

В качестве рекомендаций практического использования результатов и задач дальнейшего исследования можно определить следующее:

* диагностические признаки (артериальное давление, уровень холестерина и сахара в крови) являются одними из ключевых показателей для контроля риска возникновения ССЗ.
* необходимо поддерживать нормальное значение ИМТ, т.к. его превышение увеличивает риск возникновения ССЗ.
* увеличение доли заболевших ССЗ с увеличением возраста требует более частых регулярных обследований людей с целью своевременного выявления признаков возникновения ССЗ.
* какие диагностические признаки, не отраженные в данных, могут быть использованы для контроля и прогнозирования возникновения ССЗ.

1. В источнике данных не указана кодировка признака. Средние значения роста и веса по категориям: 1 - рост 161, вес 72.5, 2 - рост 170, вес 77. На основании этого считаем пол: 1 - женский, 2 - мужской. [↑](#footnote-ref-0)
2. Как правило, в медицинских документах фигурирует полный возраст в годах. Поэтому введем признак **age\_years -** полное количество лет, количественный признак. **age\_years = age / 365.25.** [↑](#footnote-ref-1)