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

Дипломная работа по программе «Аналитик данных»

Арсентьев Я. А.
Группа: DA-114

2025 г.

Описание задачи исследования.

1

Описание задачи исследования, стейкхолдеры.

- Предсказать наличие или отсутствие сердечно-сосудистого заболевания (ССЗ) по результатам обследования пациента.
- Определение признаков, имеющих статистически значимую взаимосвязь с наличием ССЗ.
- Связь негативных привычек с возникновением ССЗ.

Круг стейкхолдеров.

- 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.

Описание данных

2

Описание данных

- Набор данных состоит из **70 000** записей с данными о пациентах по **12** признакам

```
cardio.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70000 entries, 0 to 69999
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   id           70000 non-null  int64  
 1   age          70000 non-null  int64  
 2   gender       70000 non-null  int64  
 3   height       70000 non-null  int64  
 4   weight       70000 non-null  float64
 5   ap_hi        70000 non-null  int64  
 6   ap_lo        70000 non-null  int64  
 7   cholesterol  70000 non-null  int64  
 8   gluc         70000 non-null  int64  
 9   smoke        70000 non-null  int64  
 10  alco         70000 non-null  int64  
 11  active       70000 non-null  int64  
 12  cardio       70000 non-null  int64  
dtypes: float64(1), int64(12)
memory usage: 6.9 MB
```

Описание данных

- **Фактические данные:**
- • **gender** - пол, категориальный признак, 1 - женский, 2 - мужской.
- • **age** - возраст, количественный признак, указан в днях.
- • **height** - рост, количественный признак, указан в см.
- • **weight** - вес, количественный признак, указан в кг.
- **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.
- **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.

Предобработка данных

3

Предобработка данных

- Так как возраст в датасете указан в минутах

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

- Переведем в привычный формат, полный возраст в годах. Создаем новый столбец age_years. С помощью команды (cardio['age_years'] = cardio['age']//365.25) переводим минуты в года.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	age_years
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0	50.0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1	55.0
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1	51.0
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1	48.0
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0	47.0

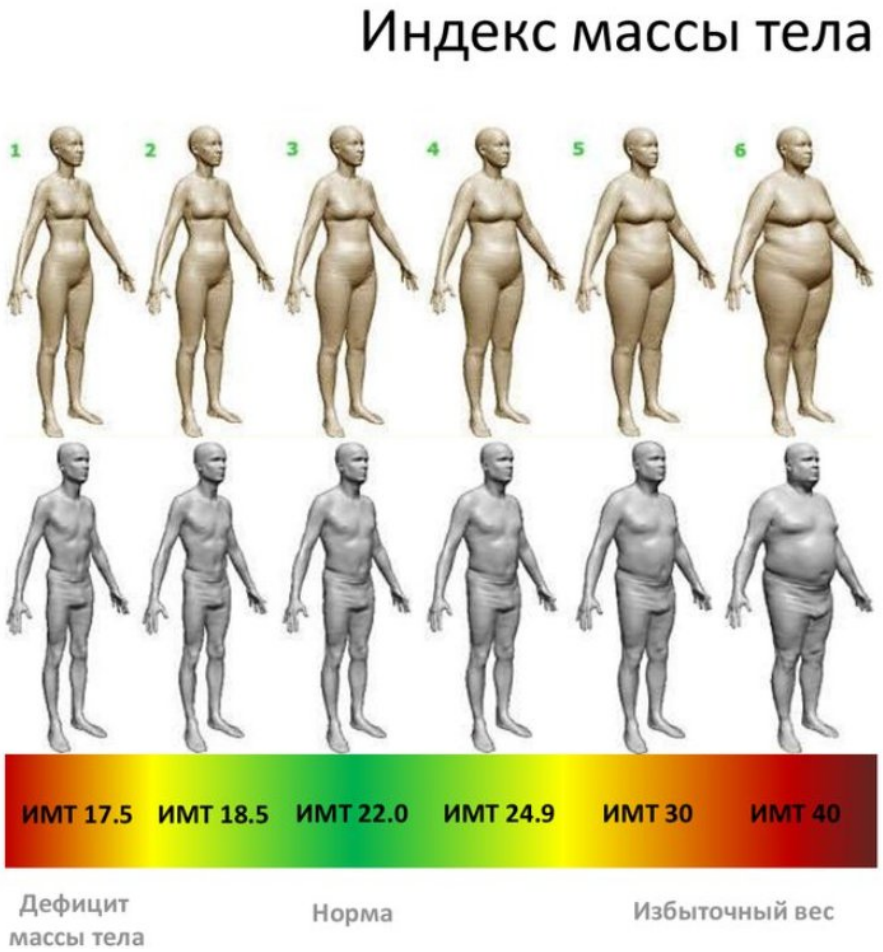
Предобработка данных

- Судя по данным нужна чистика количественных признаков(рост, вес, систолическое кровяное давление**,** диастолическое кровяное давление) в таблице от выбросов.

	height	weight	ap_hi	ap_lo
count	70000.000000	70000.000000	70000.000000	70000.000000
mean	164.359229	74.205690	128.817286	96.630414
std	8.210126	14.395757	154.011419	188.472530
min	55.000000	10.000000	-150.000000	-70.000000
25%	159.000000	65.000000	120.000000	80.000000
50%	165.000000	72.000000	120.000000	80.000000
75%	170.000000	82.000000	140.000000	90.000000
max	250.000000	200.000000	16020.000000	11000.000000

Предобработка данных

- В данных содержится признак вес **weight** и рост **height** и на основании этих данных мы можем создать колонку **Индекс массы тела** (ИМТ) который часто используется в медицине. Введем признак **BMI**, рассчитанный по формуле **ИМТ = вес (кг) / (рост(м)) ** 2**

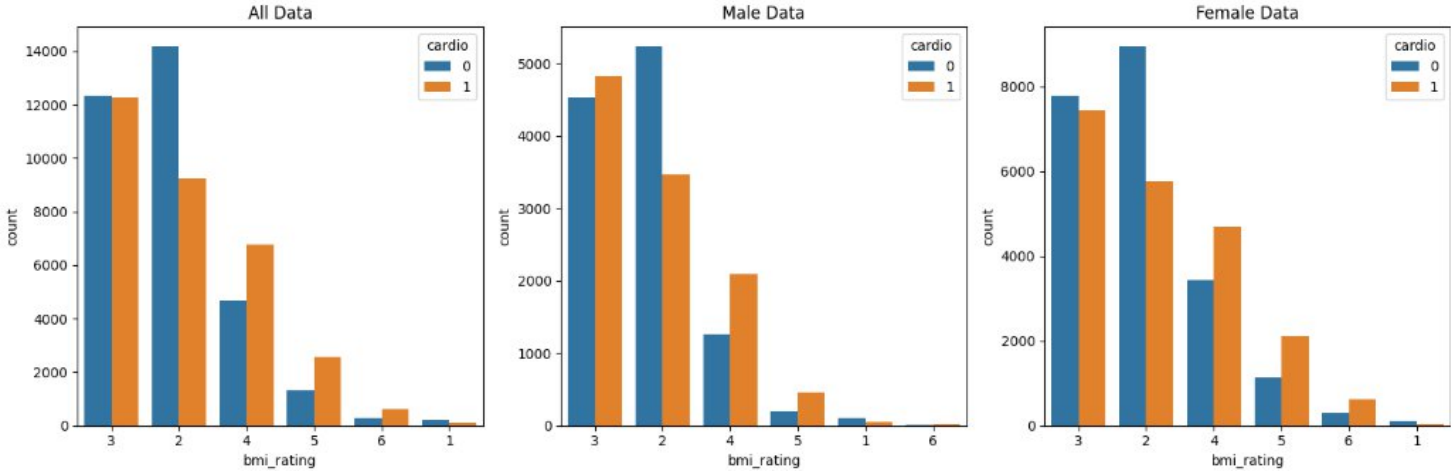


Описание проведенного исследования

4

Описание проведенного исследования

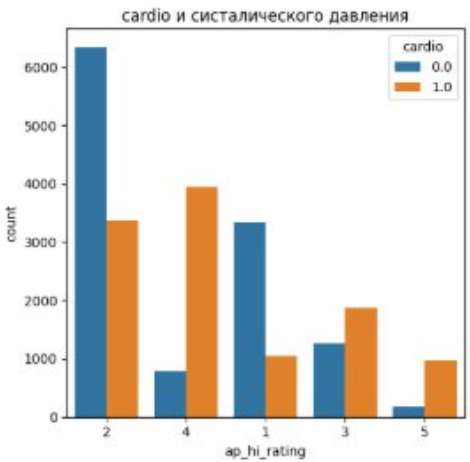
- На основе корреляционной матрицы создаем визуализацию корреляции между целевой переменной cardio и индекса массы тела (bmi), для всех данных, отдельно для мужчин и отдельно для женщин.

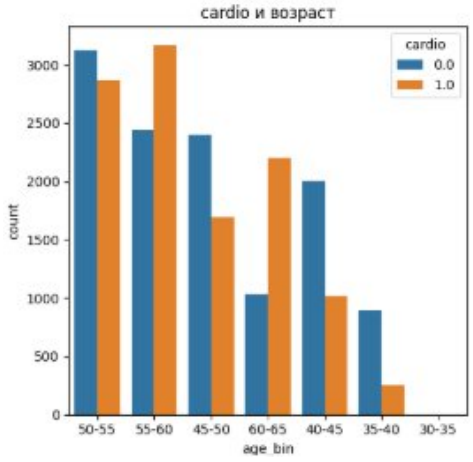


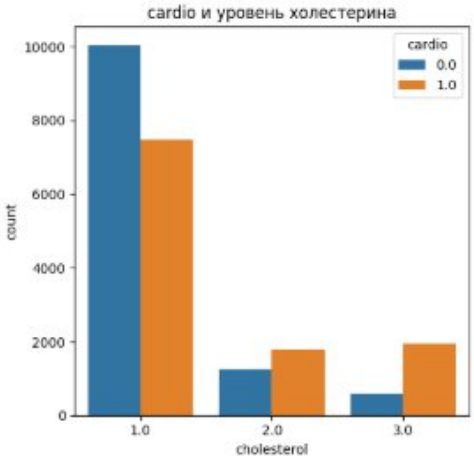
- Вывод: Наибольший риск ССЗ у мужчин в 3, 4, 5 категориях - это индексы массы тела с предожирением и ожирением.
- У женщин наибольший риск ССЗ в 4, 5, 6 категориях - это индексы массы тела с ожирением 1,2,3 степени. Во 2 категории с нормальным ИМТ и 3 категории с избыточной массой тела (предожирение) женщины менее всего подвержены ССЗ.

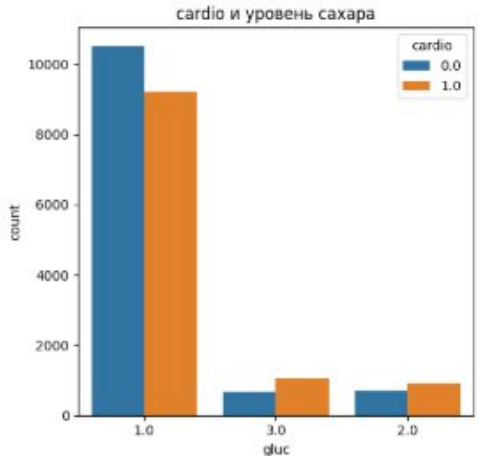
Описание проведенного исследования

- На основе корреляционной матрицы создаем визуализацию корреляции между целевой переменной cardio и давления, возраста, сахара и холестерина в крови.





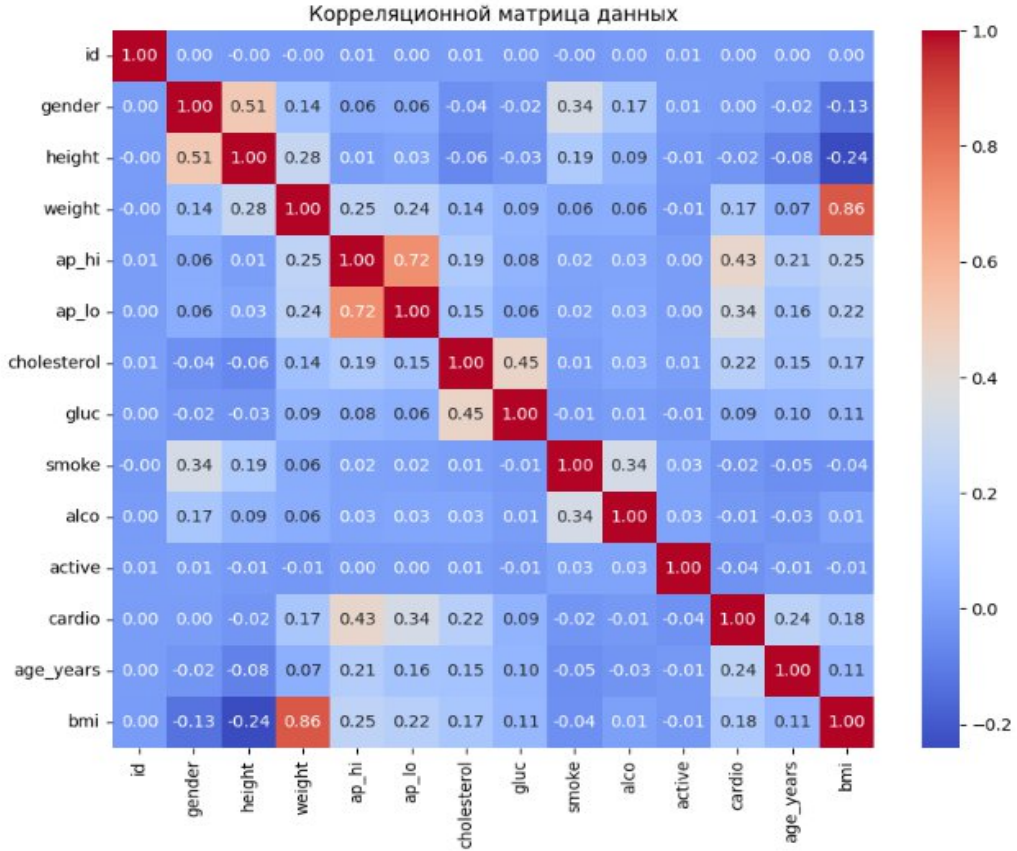




- Систолическое давление: 1 – Оптимальное артериальное давление(АД), 2 - Нормальное АД, 3 – Высокое нормальное АД, 4 - Артериальная гипертензия 1-й степени, 5 - Артериальная гипертензия 2-й степени, 6 - Артериальная гипертензия 3-й степени
- Уровень сахара и холестерина в крови: 1 - норма, 2 - выше нормы, 3 - значительно выше нормы.

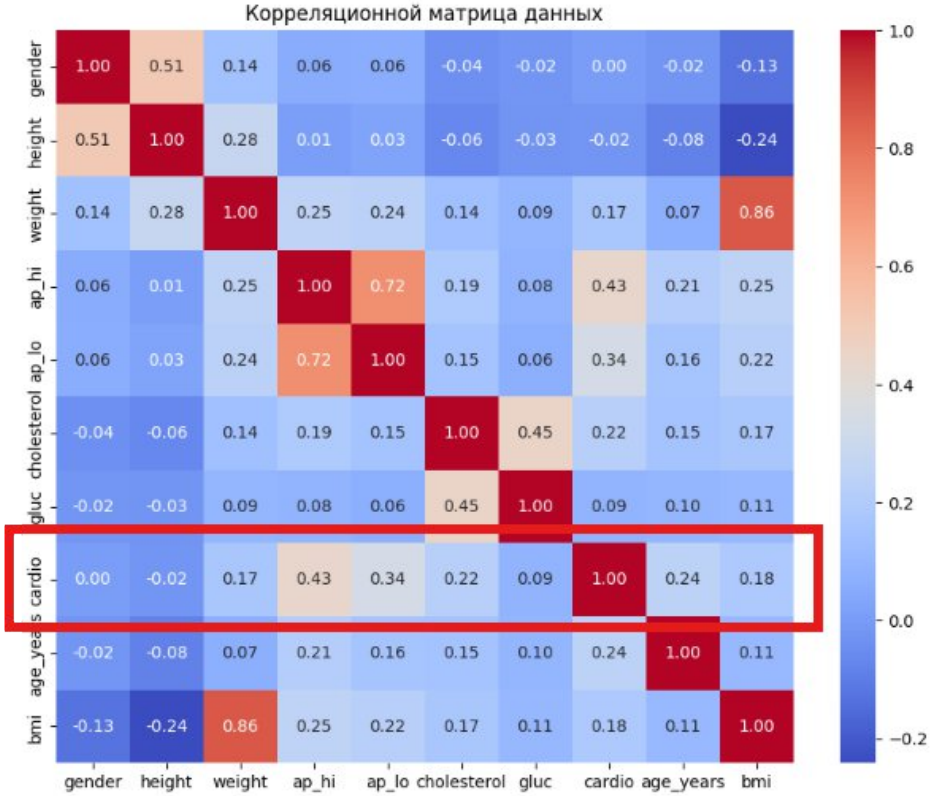
Описание проведенного исследования

- На основе датафрейма создание корреляционной матрицы для определения зависимостей.



Описание проведенного исследования

- Определение зависимостей после удаления столбцов с низкой корреляцией(`'smoke'`, `'alco'`, `'active'`, `'id'`).
- Наблюдается корреляция целевой переменной cardio с весом, ИМТ, давлением, сахаром в крови, холестерином и возрастом.
- давлением: ap_hi - 0.43, ap_lo - 0.34
- уровень холестерина: cholesterol - 0.22
- уровень сахара: gluc - 0.09
- возраст: age_years - 0.24ИМТ: bmi - 0.19.



Выводы и рекомендации

5

Выводы и рекомендации

- Наибольшая корреляция наблюдается между целевой переменной cardio с весом, ИМТ, давлением, сахаром в крови, холестерином и возрастом.
- Диагностические признаки (давление, уровень холестерина и сахара в крови) являются одними из ключевых показателей для контроля риска возникновения ССЗ.
- Максимальный риск ССЗ мужчин у которых ИМТ начиная с предожирения(3), повышенное систолическое давление > 130, с возраста 55лет, также когда уровень холестерина и сахара в крови от 2 - выше нормы.
- У женщин максимальный риск ССЗ когда ИМТ начиная с ожирение 1 степени(4), повышенное систолическое давление > 140, с возраста 55лет, также когда уровень холестерина и сахара в крови от 2 - выше нормы.
- Рекомендации: Контролировать ИМТ так как оно имеет достаточно высокую корреляцию с диагностическими признаки (давление, уровень холестерина и сахара в крови) высокий показатель которых значительно повышают шанс заболеваемости ССЗ. С возрастом от 50лет требуется более частые медицинские обследования сердечно сосудистой системы.
- Не выявлено негативного влияния поведенческих признаков курения и алкоголя на риск возникновения ССЗ, что не сходится с официальными данными Всемирной Организации Здравоохранения(ВОЗ).

Спасибо за внимание!