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

Дипломная работа по программе «Аналитик данных»

Сорокопуд Е. Г.
Группа: DA-118

2025 г.

Описание цели и задач

Описание цели и задач

В лаборатории провести масштабное исследование медицинских данных пациентов с целью раннего выявления рисков сердечно-сосудистых заболеваний. На основе полученных результатов сформировать практические рекомендации по профилактике ССЗ для всех участников медицинского процесса.

Задачи:

1. Разработка унифицированной отчётной формы для создания иерархической системы метрик на базе очищенных данных.

2. Изучение связи избыточного веса и давления с риском ССЗ для разработки мер профилактики через коррекцию питания и повышение активности.

3. Создание прогностической модели на основе логистической регрессии для оценки риска развития сердечно-сосудистых заболеваний по медицинским показателям пациентов.

Описание данных и их предобработки

2

Описание данных и их предобработки

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

Датасет - "Cardiovascular Disease dataset"

- Значения столбца "**age**" переведены из дней в года;
- В столбце "**weight** " - Изменен тип данных на "**Int**";
- Добавлен столбец "**IMT**" – индекс массы тела;

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.

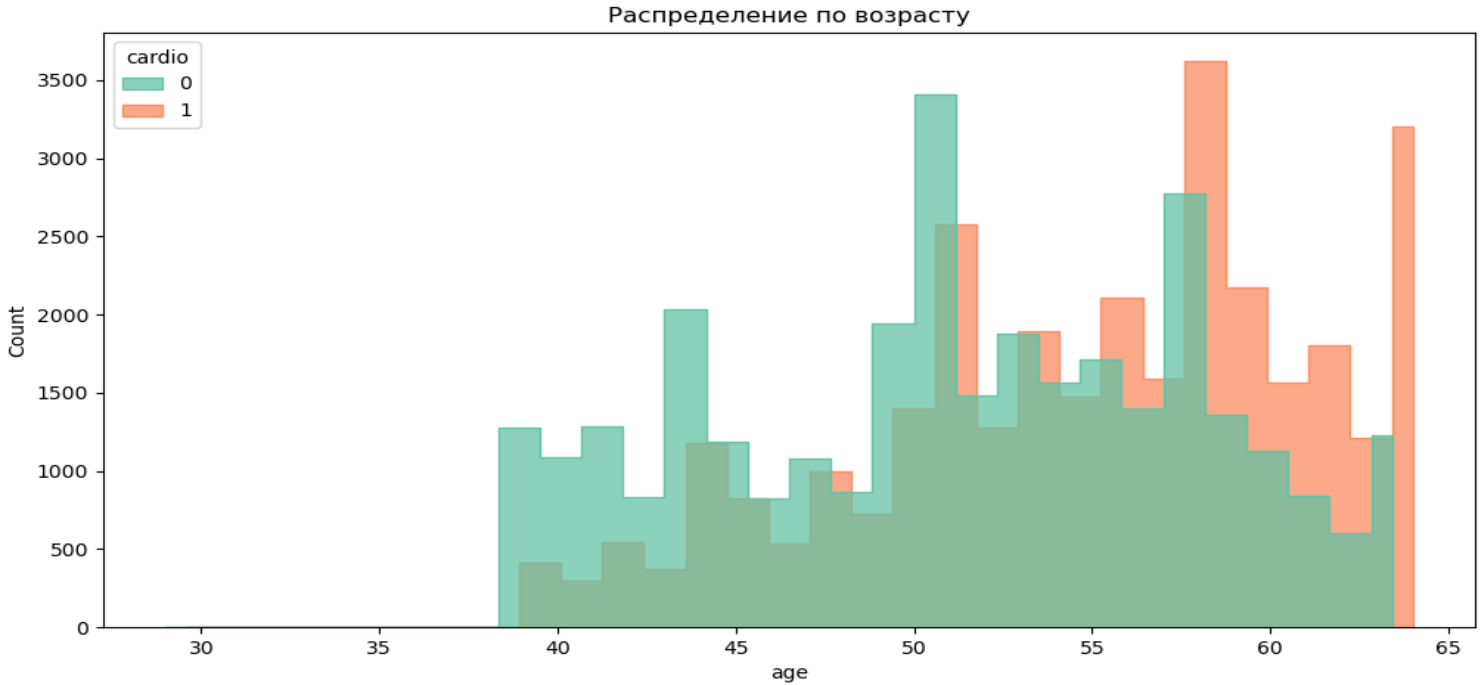
Описание проведенного исследования

3

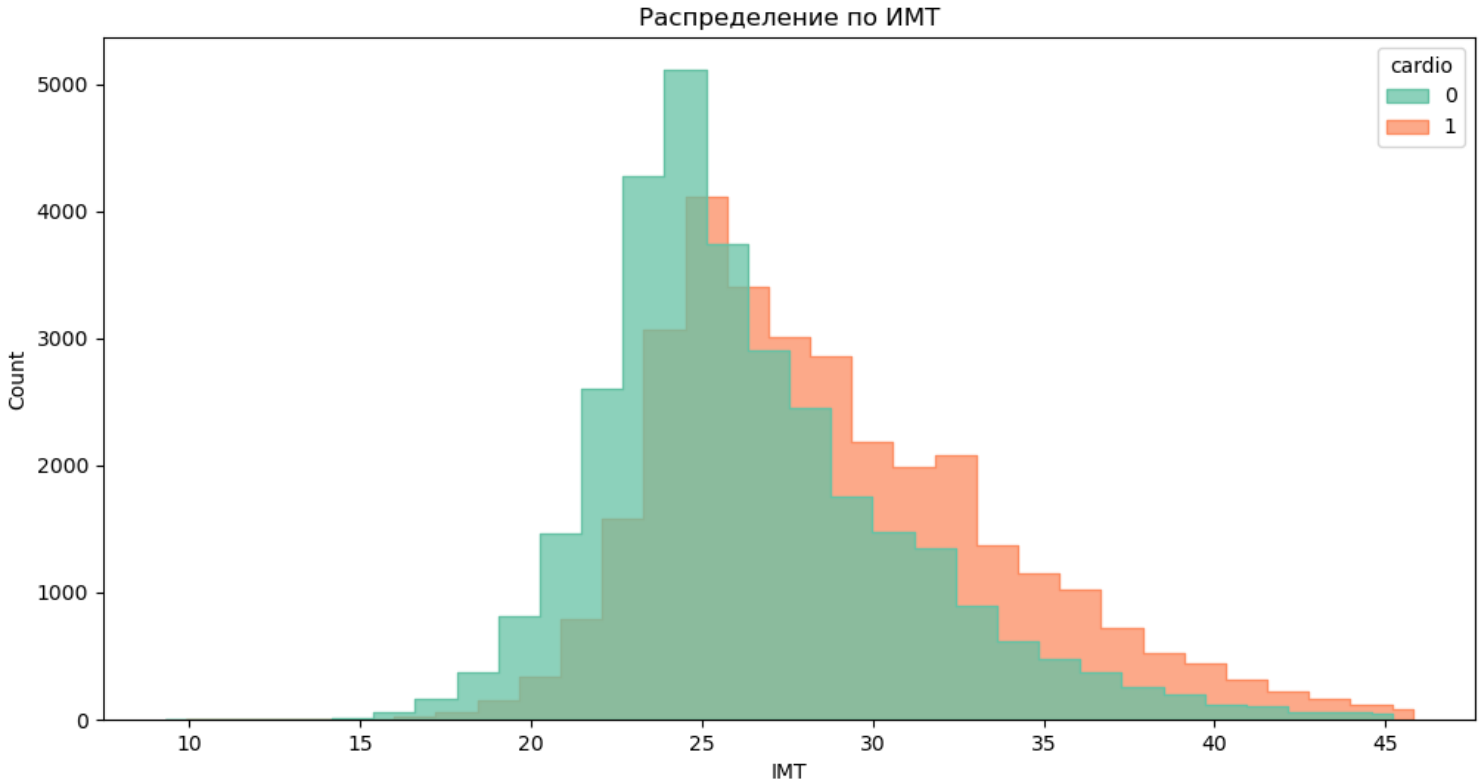
Описание проведенного исследования

По результатам корреляционного анализа выявлены значимые взаимосвязи между следующими параметрами:

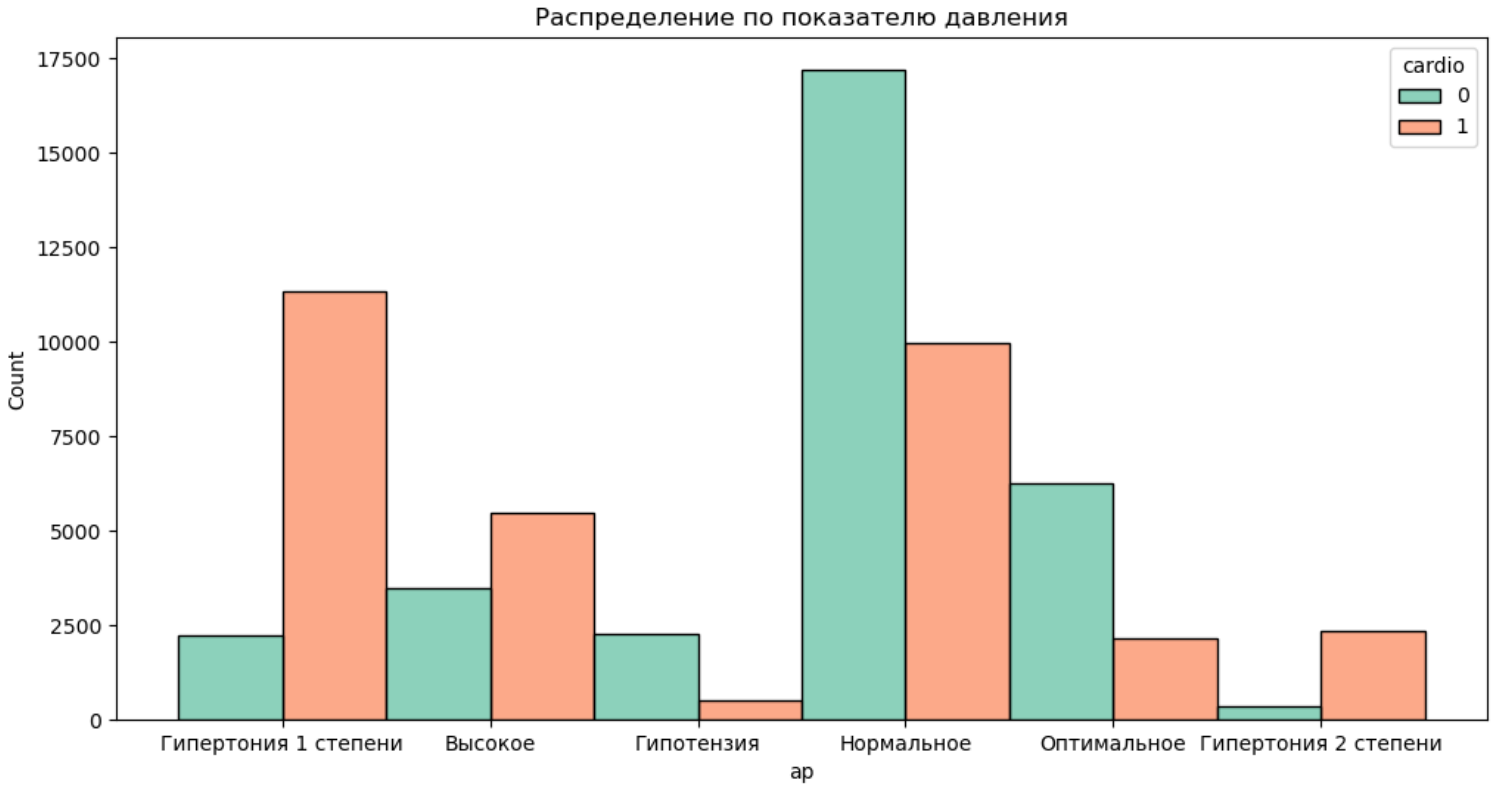
1. Выявлена связь возраста и ССЗ (p-value = 0.0);



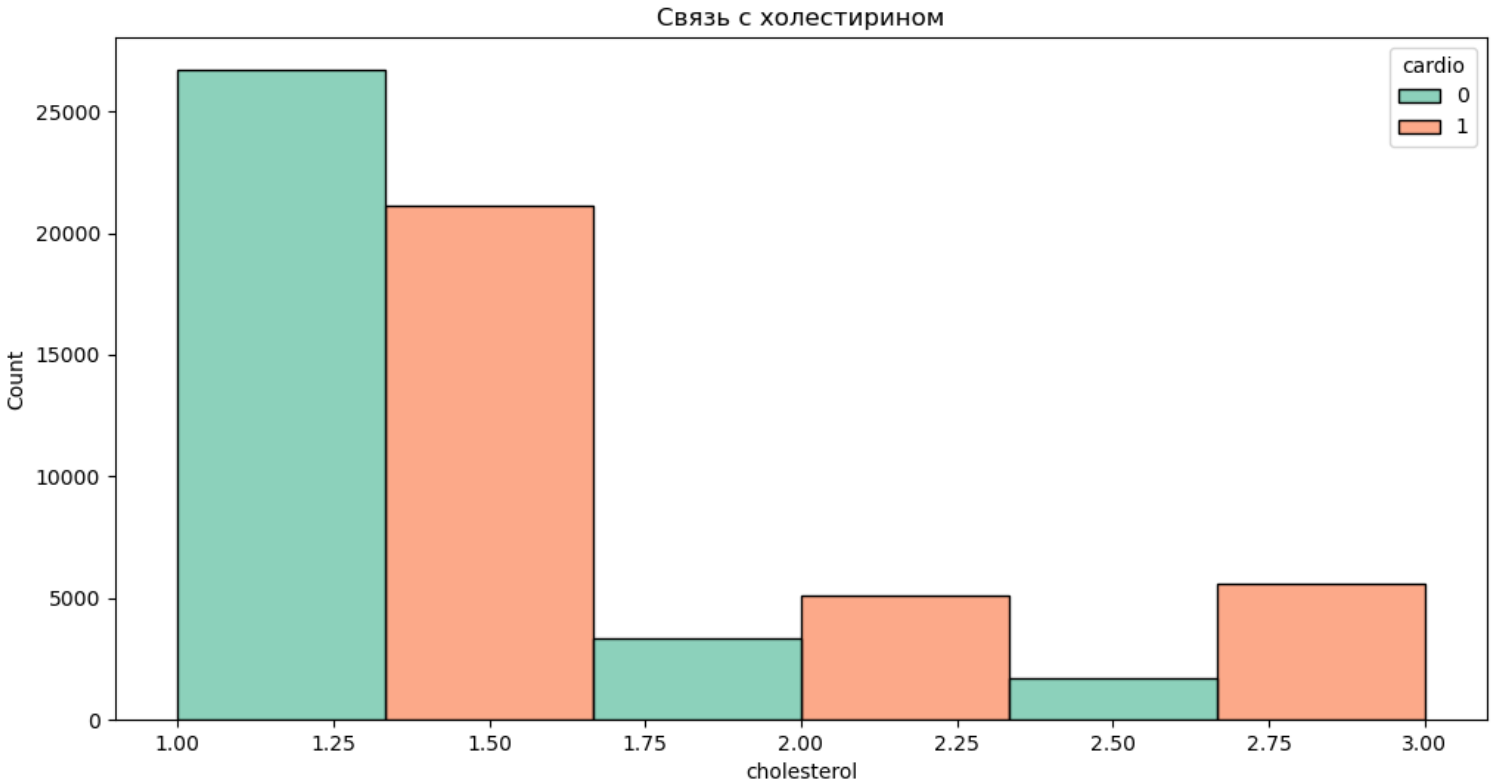
2. Выявлена связь ИМТ и ССЗ (p-value = 0.0);



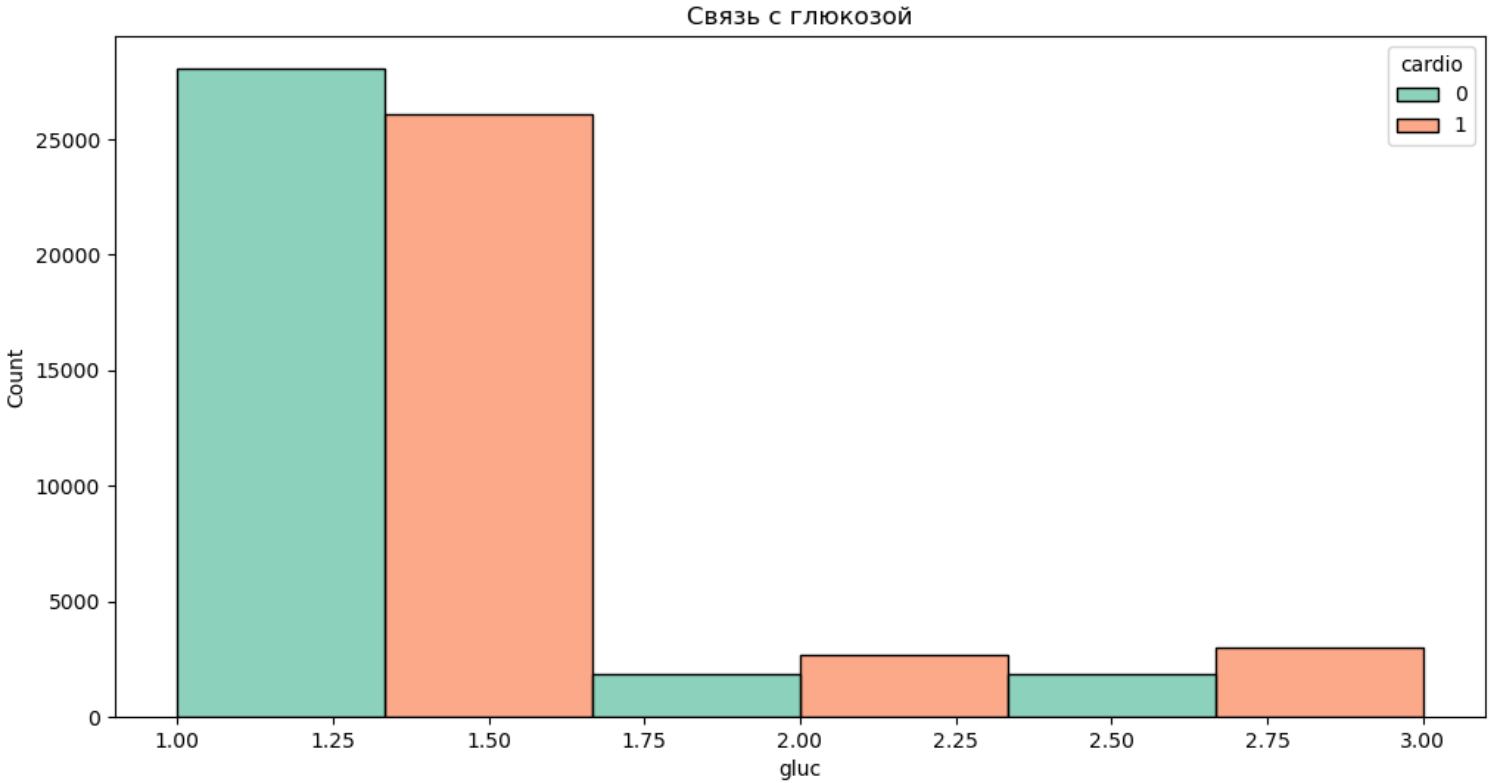
3. Выявлена связь давления и ССЗ (p-value = 0.0);



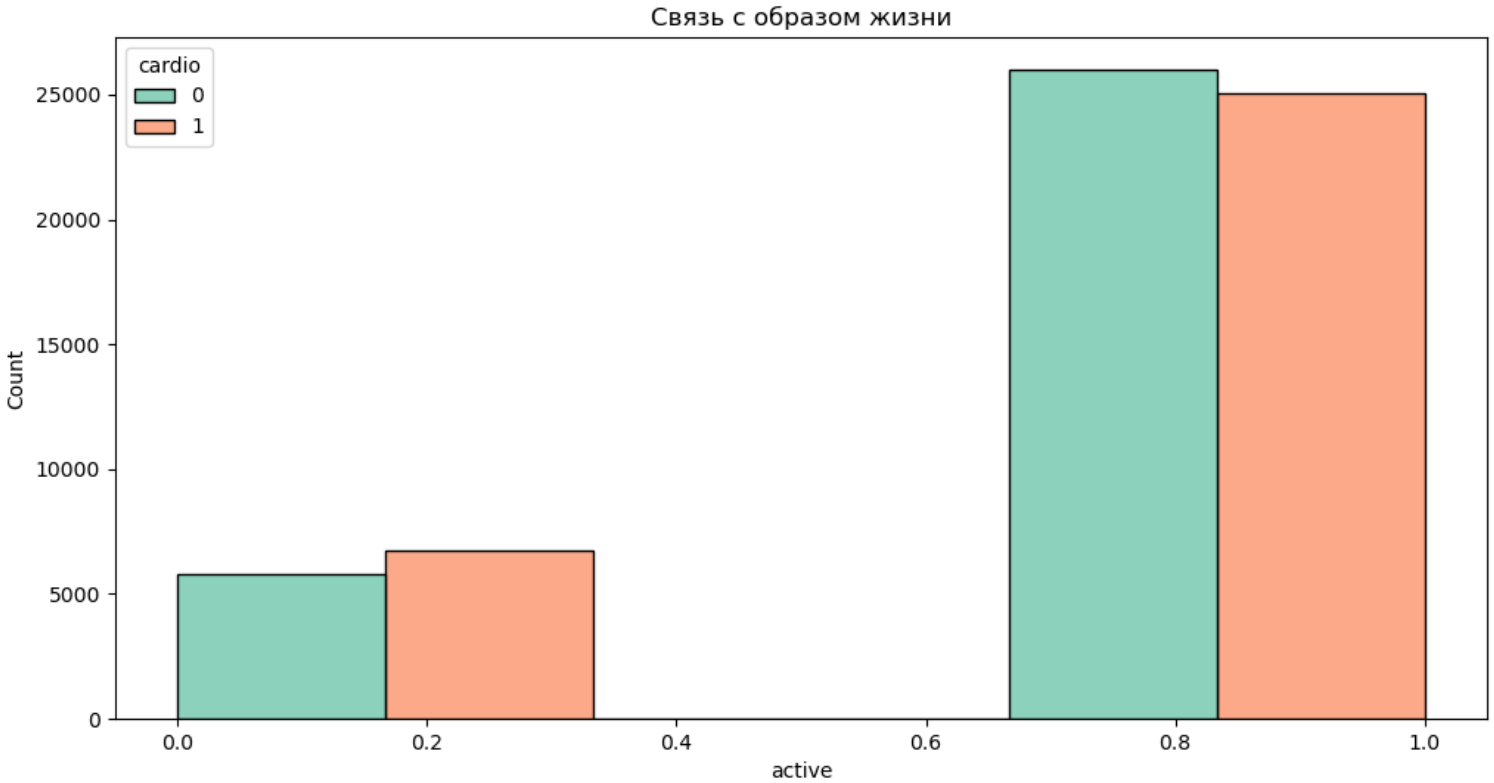
4. Выявлена связь холестерина и ССЗ (p-value = 0.0);



5. Выявлена связь глюкозы и ССЗ (p-value = 8.194708638433635e-106);



6. Выявлена связь активного образа жизни и ССЗ (p-value = 5.293096588842994e-21);



Анализ других показателей датасета:

1. Пол;
2. Курение;
3. Употребление алкоголя;

Не выявил статистически значимой корреляции с риском развития ССЗ согласно t-критерию и графической оценке.

Рост и вес, хотя и показали определённую связь с ССЗ, не рассматриваются как независимые факторы риска, поскольку их влияние реализуется через индекс массы тела (ИМТ).

	Женщины	Мужчины
ССЗ	65%	35%
Без	65%	35%

	Некурящие	Курящие
ССЗ	92%	8%
Без	91%	9%

	Непьющие	Пьющие
ССЗ	95%	5%
Без	95%	5%

Построение модели

4

Построение модели

Цель исследования — создание эффективного инструмента прогнозирования на основе анализа медицинских и демографических показателей пациентов.

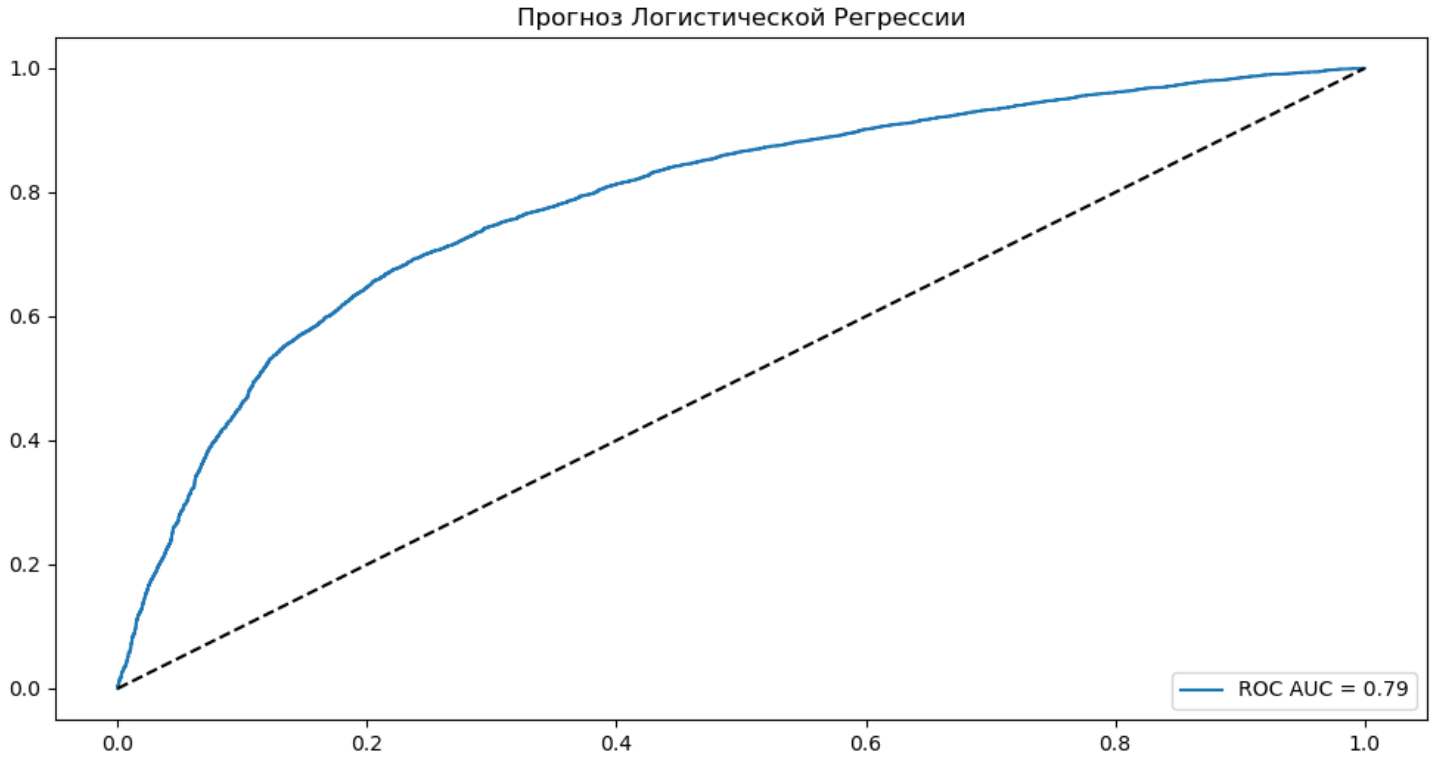
Модель построена на обработанном датасете с показателями пациентов, который прошёл очистку и нормализацию данных.

Параметры:

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

Модель прогнозирует риск сердечно-сосудистых заболеваний (cardio).

Для анализа использована логистическая регрессия, оценивающая вероятность заболевания и влияние факторов.


Прогноз Логистической Регрессии
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC AUC = 0.79

Оценка качества прогнозной модели:

Модель показала высокую эффективность:

- Точность на обучающей выборке: 72%
- Точность на тестовой выборке: 73%
- Итоговая оценка значимости: 79%

Результаты демонстрируют стабильность модели и отсутствие переобучения.

Модель эффективно выявляет риски сердечно-сосудистых заболеваний и помогает врачам своевременно направлять пациентов на дополнительное обследование. Это повышает шансы на успешное лечение.

Выводы и рекомендации

5

Выводы и рекомендации

ССЗ лидируют среди диагностируемых патологий и остаются главной причиной смертности в мире, что определяет актуальность их изучения и разработки методов ранней диагностики.

Вывод по задачам:

1. В результате обработки данных (трансформация, очистка датасета с удалением 6454 некорректных строк) создана готовая к применению в лаборатории система метрик.

2. Выявлена статистически значимая связь (p-value = 0,00) между избыточным весом, артериальным давлением и риском ССЗ, что обосновывает необходимость мониторинга этих показателей при скрининге пациентов.

3.Прогностическая модель на основе логистической регрессии показала высокую точность (79%) в оценке риска развития сердечно-сосудистых заболеваний и может эффективно применяться в клинической практике для раннего выявления рисков.

Рекомендации:

Доработать данные следующими колонками:

- Измерение уровня стресса
- Данные о хронических заболеваниях пациента
- Электрокардиограмма
- Генетическое тестирование
- Общий анализ крови
- Общий анализ мочи

Спасибо за внимание!