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

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” 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)/ ТГУ 2011 г. Стр. 41;
2. Диагностика и лечение артериальной гипертензии (электронный ресурс)/ 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)/

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%

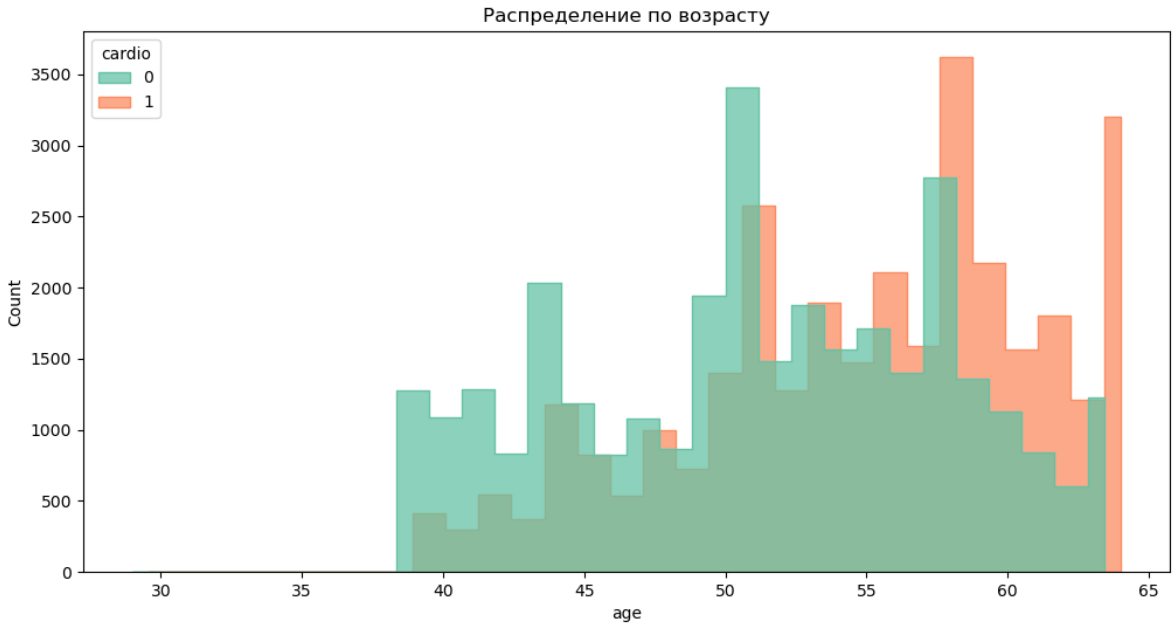
Образ жизни:

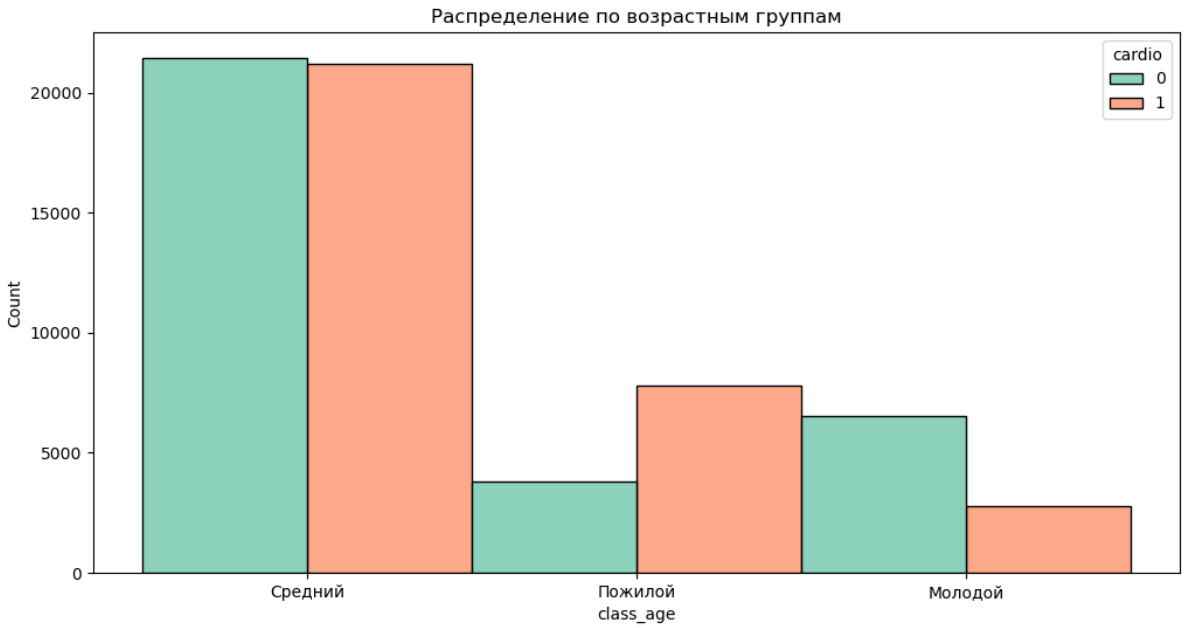
- **Курение:** некурящие — 91%, курящие — 9%
- **Алкоголь:** непьющие — 95%, пьющие — 5%
- **Физическая активность:** активные — 80%, неактивные — 20%

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)



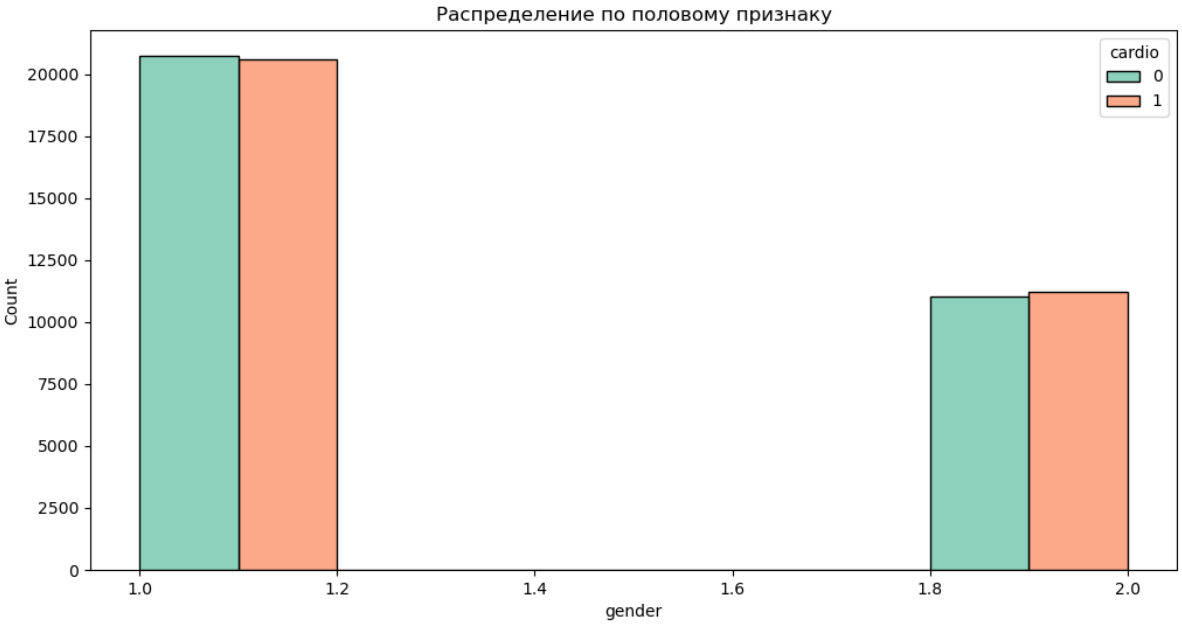


	Молодой	Средний	Пожилой
СС3	9%	67%	24%
Без	20%	68%	12%

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 = 0.17787724191871 24)



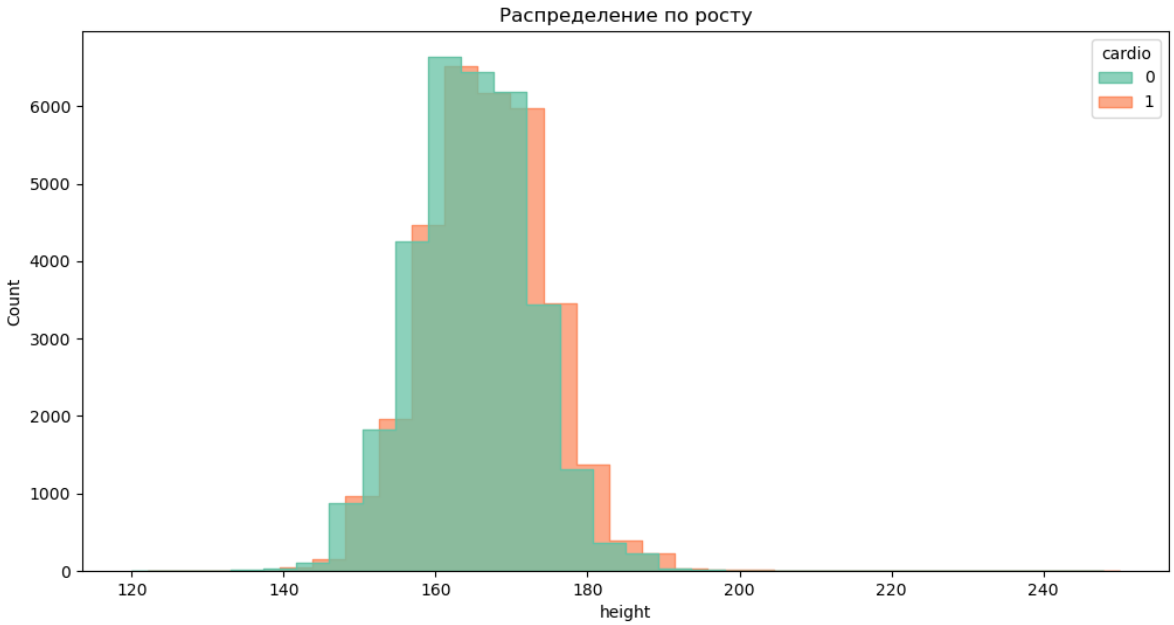
	Женщины	Мужчины
ССЗ	65%	35%
Без	65%	35%

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.

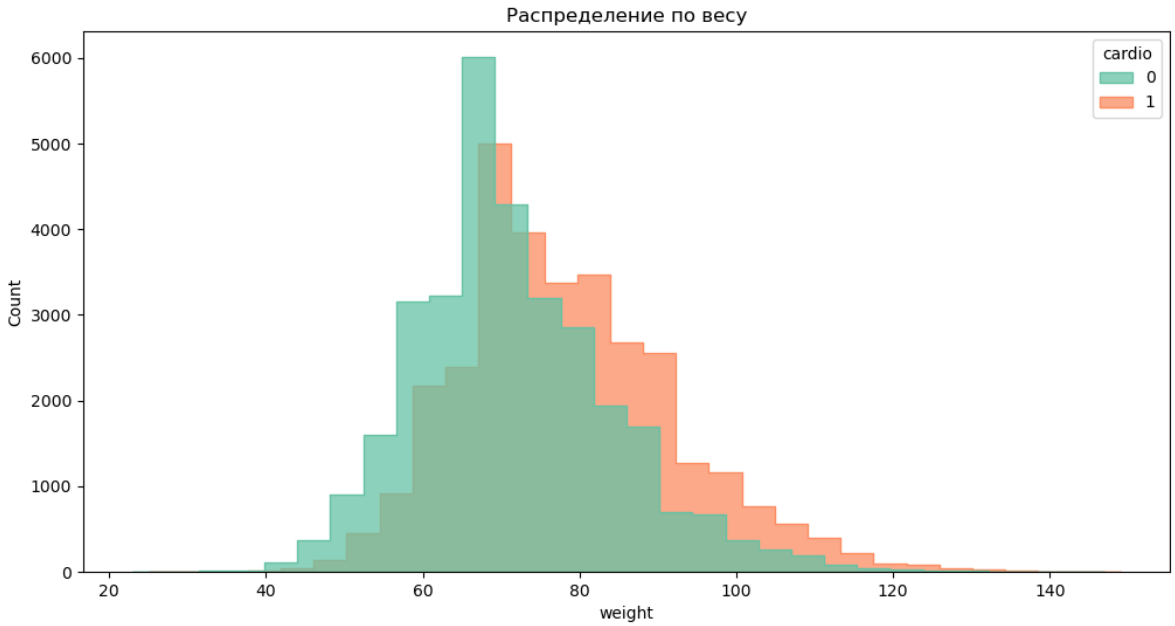
4.1.3 Взаимосвязь С-СЗ с показателем роста (p-value = 0.00590021921825544)



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)



Вывод:

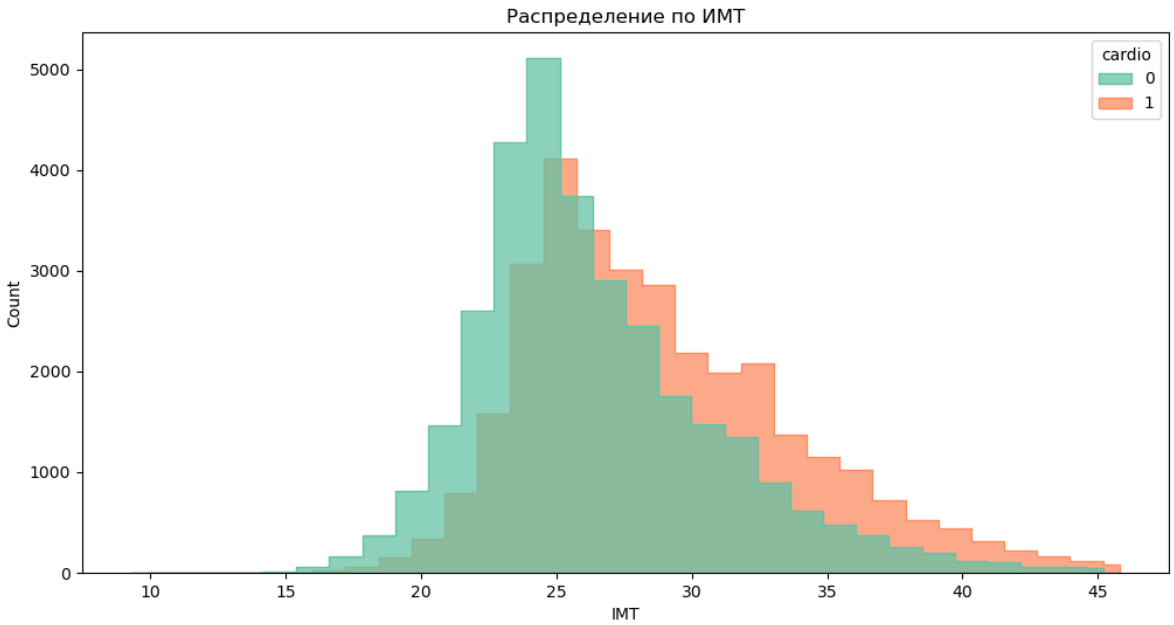
Статистический анализ данных (включая графический анализ и Т-критерий) выявил закономерность: среди пациентов с массой тела 80 кг и выше отмечается преобладание лиц с диагностированными сердечно-сосудистыми заболеваниями над здоровыми.

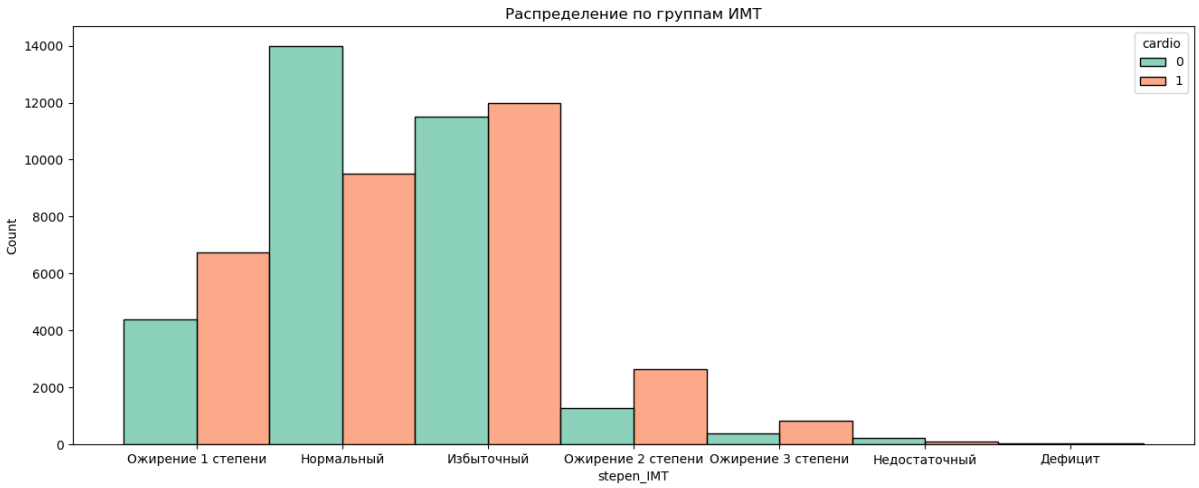
Тем не менее для корректной оценки влияния весового фактора на развитие ССЗ требуется расчёт индекса массы тела (ИМТ). Данный этап исследования планируется реализовать в дальнейшем, что позволит получить более точные и объективные результаты.

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.





	Дефицит	Недостаточный	Нормальный	Избыточный	Ожирение 1 степени	Ожирение 2 степени	Ожирение 3 степени
ССЗ	<1%	<1%	30%	38%	21%	8%	3%
Без	<1%	1%	44%	36%	14%	4%	1%

Вывод:

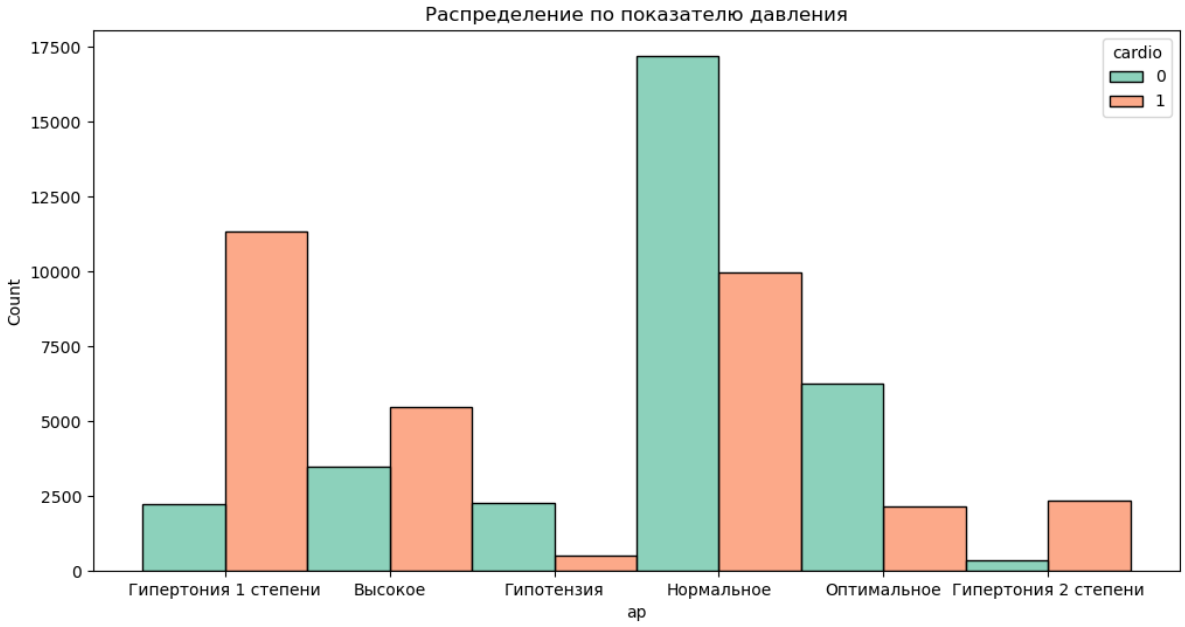
Анализ графических данных и т-критерия выявил прямую корреляцию между индексом массы тела (ИМТ) и вероятностью диагностики сердечно-сосудистых заболеваний: при

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.



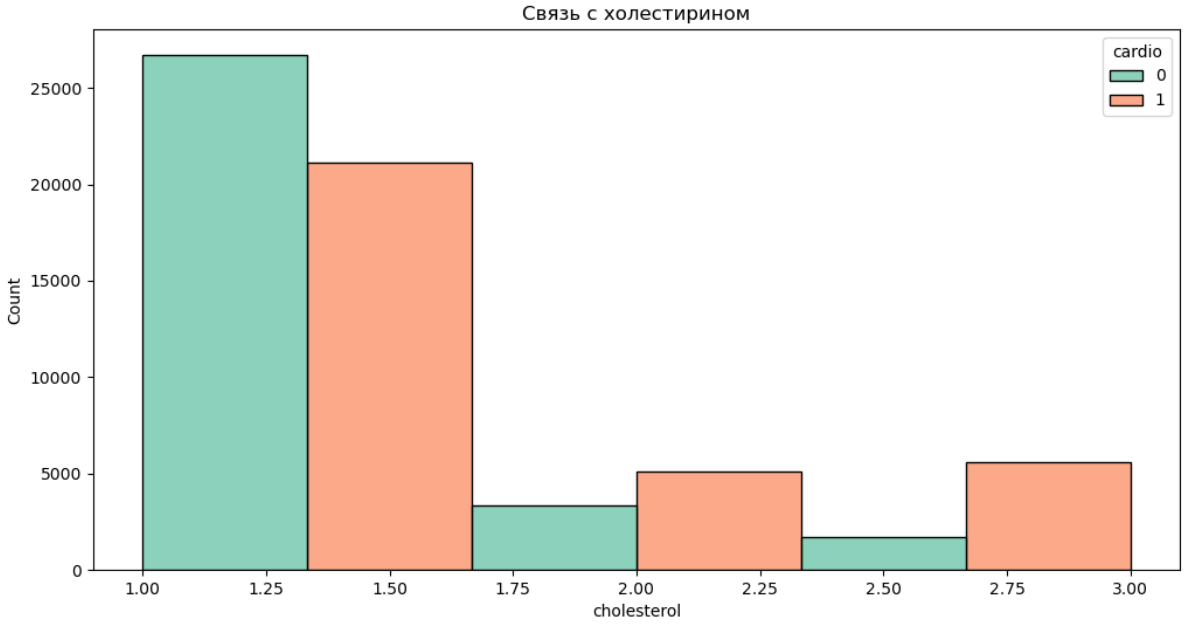
	Гипотензия	Оптимальное	Нормальное	Высокое	Гипертония 1 степени	Гипертония 2 степени
ССЗ	2%	7%	31%	17%	36%	7%
Без	7%	20%	54%	11%	7%	1%

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)



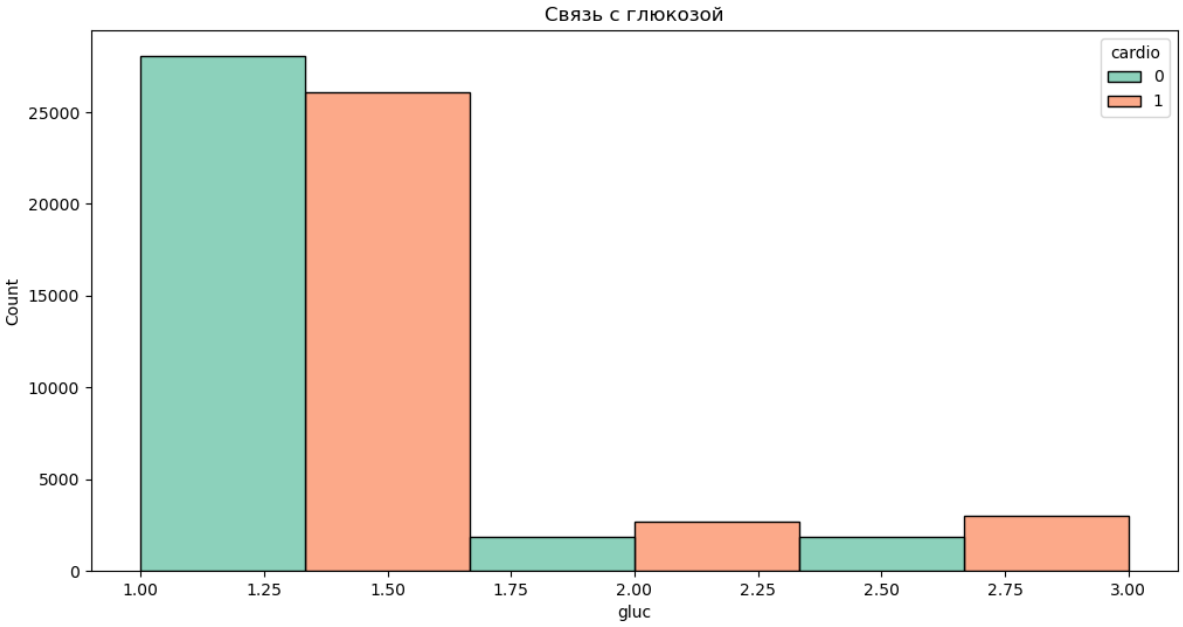
	Нормальный	Повышеный	Высокий
ССЗ	66%	16%	18%
Без	84%	11%	5%

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 = 8.1947086384336 35e-106)



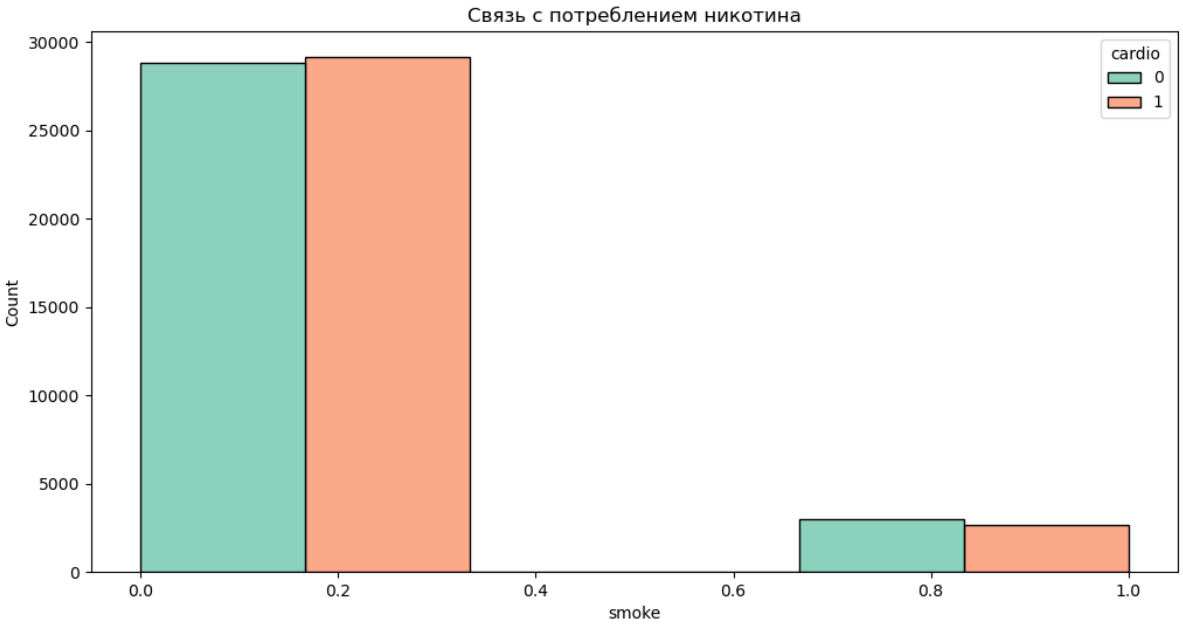
	Нормальная	Повышеная	Высокая
ССЗ	82%	8%	9%
Без	88%	6%	6%

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)



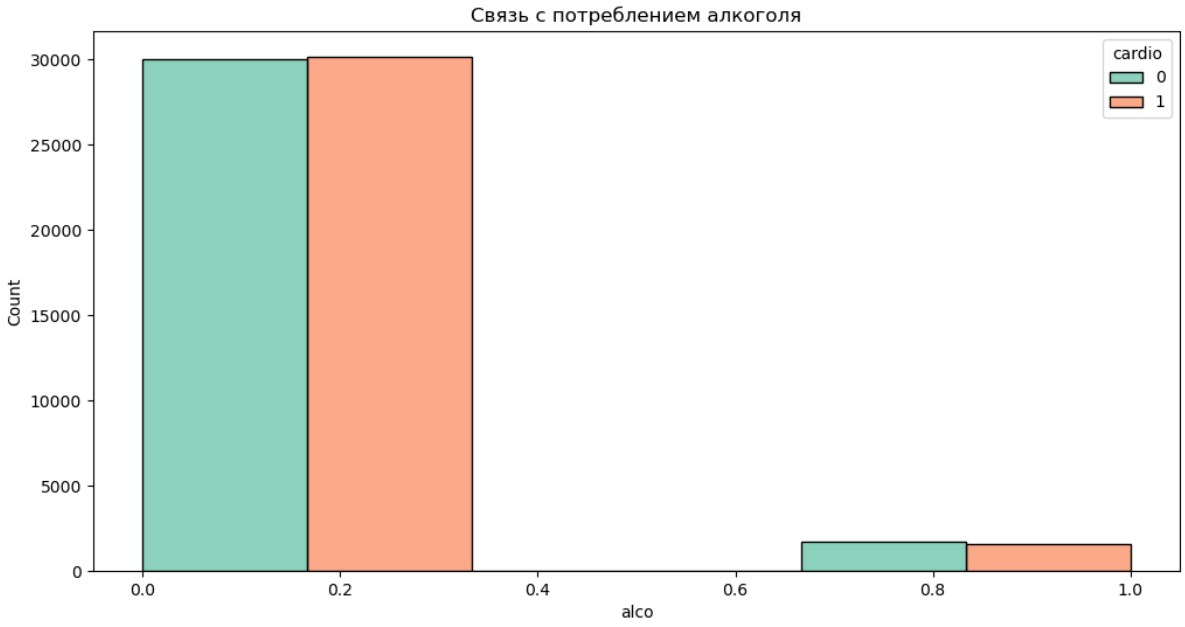
	Некурящие	Курящие
ССЗ	92%	8%
Без	91%	9%

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)

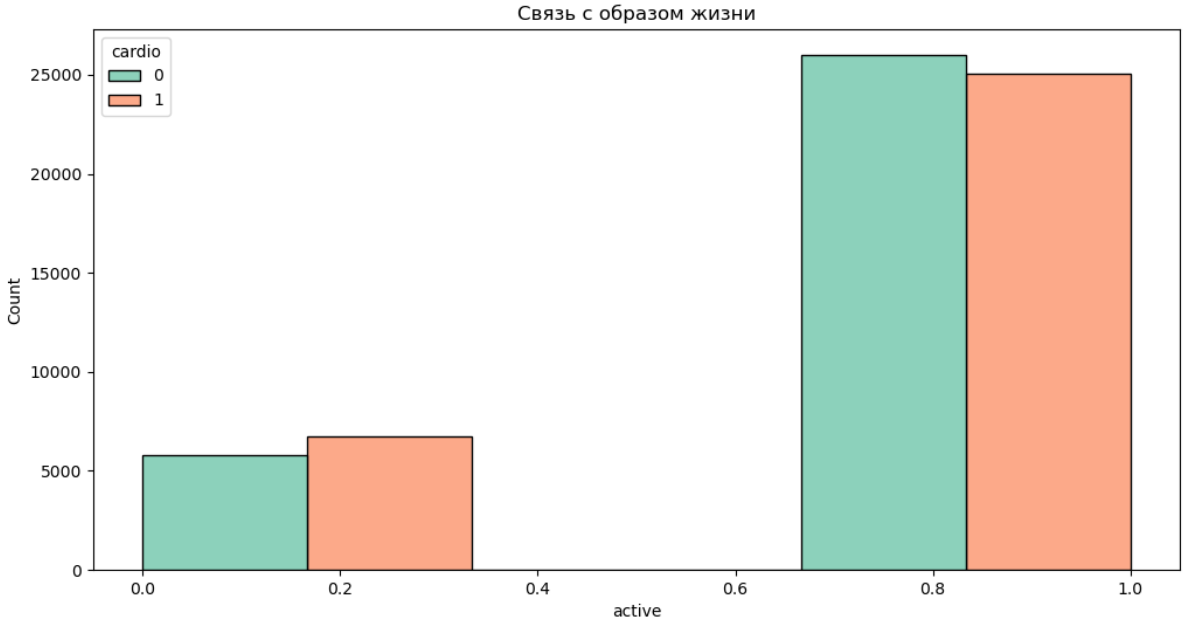


	Непьющие	Пьющие
ССЗ	95%	5%
Без	95%	5%

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 = 5.2930965888842994e-21)



	Ведущие	Неведущие
ССЗ	79%	21%
Без	82%	18%

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:

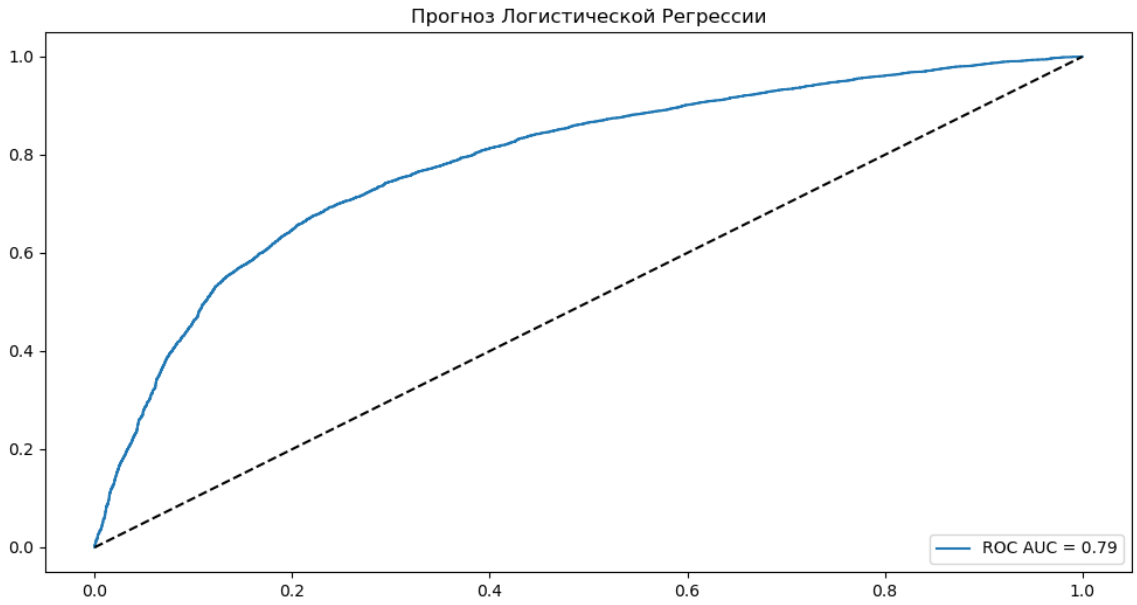
1. Биологические характеристики: пол (gender), возраст (age), рост (height), вес (weight), индекс массы тела (IMT);
2. Диагностические параметры: артериальное давление (ap_hi и ap_lo), уровень холестерина (cholesterol), уровень глюкозы (gluc);
3. Показатели образа жизни: курение (smoke), употребление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