ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ «НЕТОЛОГИЯ»

**ДИПЛОМНАЯ РАБОТА**

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

по направлению подготовки DAU-75 Аналитик данных

Подрезова Екатерина Николаевна

Москва, 2025

**Содержание:**

[1. Введение 3](#_Toc198021808)

[2. Постановка бизнес-задачи 3](#_Toc198021809)

[2.1 Проблема: 3](#_Toc198021810)

[2.2 Бизнес-задача. 3](#_Toc198021811)

[3. Описание данных 3](#_Toc198021812)

[Описание датасета. 3](#_Toc198021813)

[4. Разведочный анализ (EDA) 4](#_Toc198021814)

[4.1. Проверка на пропущенные значения. 4](#_Toc198021815)

[4.2. Проверка на дубликаты. 4](#_Toc198021816)

[4.3. Анализ аномальных значений. 4](#_Toc198021817)

[4.4. Очистка данных от аномалий. 5](#_Toc198021818)

[4.5. Анализ зависимостей. 5](#_Toc198021819)

[4.5.1. Анализ зависимости наличия ССЗ от возраста. 5](#_Toc198021820)

[4.5.2. Анализ зависимости наличия ССЗ от веса и индекса массы тела (BMI). 5](#_Toc198021821)

[4.5.3. Анализ зависимости наличия ССЗ от уровня холестерина. 6](#_Toc198021822)

[4.5.4. Анализ зависимости наличия ССЗ от уровня холестерина и глюкозы. 7](#_Toc198021823)

[4.5.4 Матрица корреляции. 8](#_Toc198021824)

[5. Построение модели классификации 9](#_Toc198021825)

[5.1. Логистическая регрессия. 9](#_Toc198021826)

[5.2. Случайный лес. 10](#_Toc198021827)

[5.3. XGBoost. 12](#_Toc198021828)

[5.4. Выводы по результатам обучения моделей и ROC-анализу: 12](#_Toc198021829)

[5.5. Пример внедрения модели. 13](#_Toc198021830)

[Заключение 14](#_Toc198021831)

# **1. Введение**

Сердечно-сосудистые заболевания (ССЗ) — **главная причина смертности** в мире (по данным ВОЗ). Раннее выявление факторов риска позволяет предотвратить развитие болезней и снизить нагрузку на систему здравоохранения.

* 17,9 млн человек умирают ежегодно от ССЗ.
* В России ситуация особенно острая: 47% смертей связаны с ССЗ (Росстат, 2022)
* 80% преждевременных инфарктов и инсультов можно предотвратить.
* Ранняя диагностика снижает риск осложнений на 30–50%.

Цель работы:

* Провести анализ данных о пациентах с ССЗ.
* Выявить ключевые факторы риска.
* Построить модель классификации для прогнозирования заболевания.
* Разработать рекомендации для медучреждений и пациентов.

# **2. Постановка бизнес-задачи**

## **2.1 Проблема:**

* Терапевты принимают решение за 10-15 минут.
* Нет автоматизированных систем оценки риска.
* Пациенты не осознают свой риск, пока не станет слишком поздно.

## **2.2 Бизнес-задача.**

Разработать систему оценки индивидуального риска ССЗ на основе стандартных клинических показателей, которая позволит:

**Для медицинских учреждений:**

* Выявлять пациентов из группы риска для приоритетного наблюдения.
* Оптимизировать нагрузку на врачей.

**Для пациентов:**

* Получать персонализированную оценку риска.
* Получать конкретные рекомендации по снижению риска.
* Мотивировать на изменение образа жизни.

# **3. Описание данных**

## **Описание датасета.**

Использован датасет с [Kaggle: Cardiovascular Disease Dataset](https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset?resource=download).

Признаки включают возраст, рост, вес, пол, давление, уровень холестерина и глюкозы, а также бинарные показатели (курение, алкоголь, активность).

**Признаки:** 12 медицинских и антропометрических показателей.

**Целевая переменная:** cardio (бинарный: 0/1 — наличие ССЗ).

|  |  |
| --- | --- |
| **Признак** | **Описание** |
| age | Возраст в днях |
| height | Рост (см) |
| weight | Вес (кг) |
| gender | Пол (1 – женщина, 2 – мужчина) |
| ap\_hi | Систолическое давление (мм рт. ст.) |
| ap\_lo | Диастолическое давление (мм рт. ст.) |
| cholesterol | Уровень холестерина (1: норма, 2: выше нормы, 3: сильно выше) |
| gluc | Уровень глюкозы (1: норма, 2: выше нормы, 3: сильно выше) |
| smoke | Курение (0: нет, 1: есть) |
| alco | Употребление алкоголя (0: нет, 1: есть) |
| active | Физическая активность (0: нет, 1: есть) |
| cardio | Наличие ССЗ (0: нет, 1: есть) |

# **4. Разведочный анализ (EDA)**

## **4.1. Проверка на пропущенные значения.**

* Пропущенные значения отсутствуют во всех столбцах.
* Дополнительная обработка пропусков не требуется.

## **4.2. Проверка на дубликаты.**

* Полные дубликаты не обнаружены.
* Все 70,000 записей уникальны.

## **4.3. Анализ аномальных значений.**

* Давление: обнаружены физиологически невозможные значения давления:
* Верхнее давление > 250 или < 70 мм рт.ст.
* Нижнее давление > 150 или < 40 мм рт.ст.
* Случаи, когда нижнее давление превышает верхнее.

Всего найдено 1332 аномальных записей (1.9% данных).

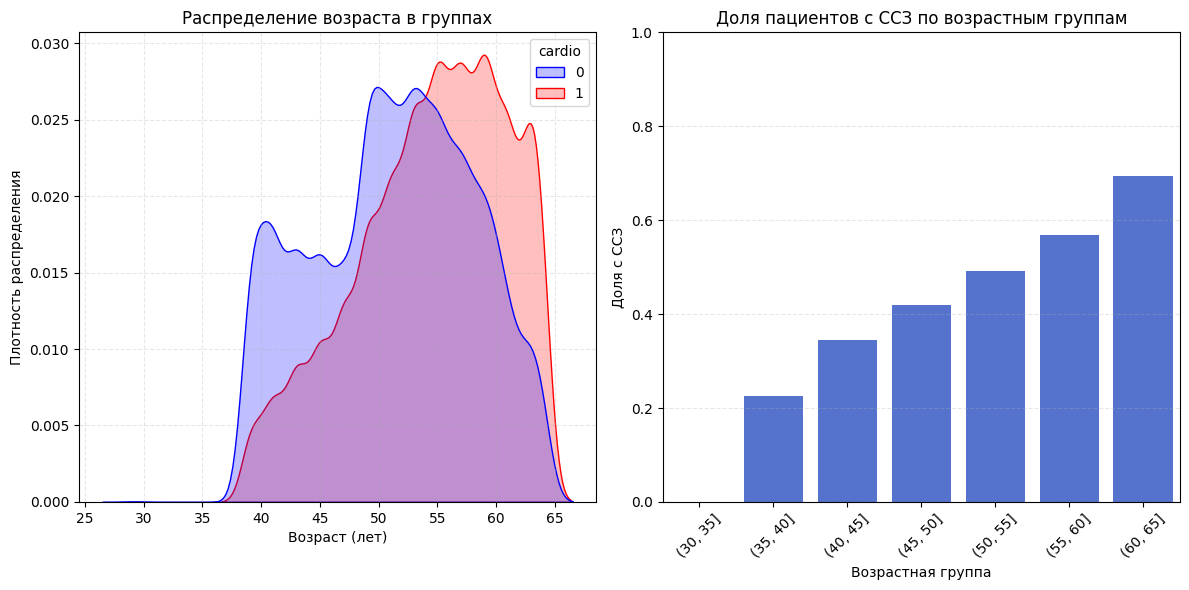
* Рост/вес:
* Аномалии роста: 29 записей (0.04% данных).
* Аномалии веса: 7 записей (0.01% данных).
* Возраст:
* Аномалии возраста не обнаружены (0.00% данных).
* Все пациенты в возрасте от 30 до 65 лет.
* Данные соответствуют ожидаемому диапазону.
* Холестерин и глюкоза:
* Аномалии холестерина: 0 записей (0.00% данных).
* Аномалии глюкозы: 0 записей (0.00% данных).
* Аномалии глюкозы: 0 записей (0.00% данных).
* Пол, курение, алкоголь, активность, ССЗ:
* Аномалии по полу: 0 записей (0.00% данных).
* Аномалии по курящим: 0 записей (0.00% данных).
* Аномалии по употребляющим алкоголь: 0 записей (0.00% данных).
* Аномалии по физически активным: 0 записей (0.00% данных).
* Аномалии по ССЗ: 0 записей (0.00% данных).
* Все значения соответствуют ожидаемым категориям 0, 1).
* Индекс массы тела:
* Средний BMI составляет 27.6 (избыточный вес).
* 25% пациентов имеют BMI > 30.2 (ожирение).
* Минимальное значение: 3.5 (аномально низкое).
* Максимальное значение: 298.7 (аномально высокое).
* Аномальные значения BMI: 271 записей (0.39% данных).

## **4.4. Очистка данных от аномалий.**

После обработки данных осталось 68 406 записей из исходных 70 000. Было исключено 1 594 наблюдений (2.28% от общего объема).

## **4.5. Анализ зависимостей.**

## **4.5.1. Анализ зависимости наличия ССЗ от возраста.**



Ключевые наблюдения:

1. Распределение возраста:

* В группе с ССЗ пик плотности смещен вправо (к старшим возрастам)
* Наибольшая разница между группами наблюдается в диапазоне 50-70 лет

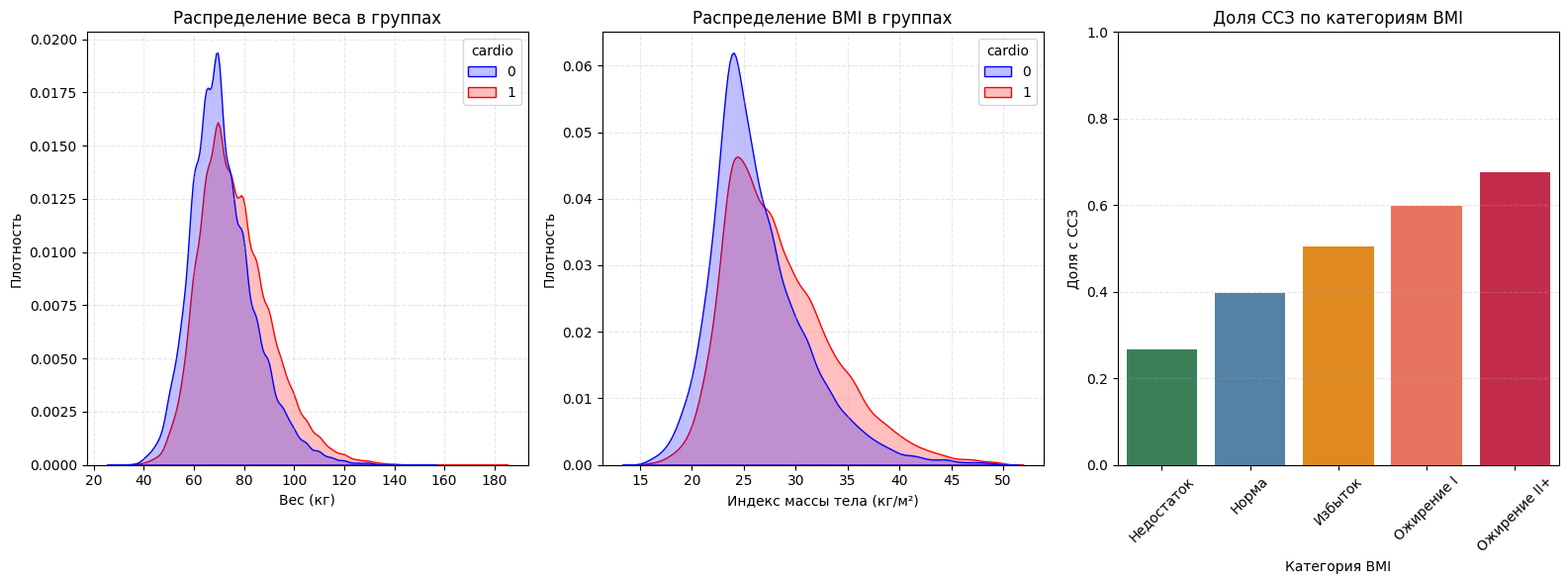
2. Возрастные группы:

* Четко прослеживается рост доли ССЗ с увеличением возраста
* Наиболее резкий скачок происходит после 45 лет

3. Практические выводы:

* Возраст является значимым фактором риска ССЗ
* Особое внимание следует уделять пациентам старше 45 лет

## **4.5.2. Анализ зависимости наличия ССЗ от веса и индекса массы тела (BMI).**



Ключевые наблюдения:

1. Распределение веса:

* У пациентов с ССЗ распределение смещено вправо (более высокие значения).
* Наибольшая разница наблюдается в диапазоне 70-100 кг.

2. Распределение BMI:

* В группе с ССЗ выражен «хвост» в области высоких значений (ожирение).
* Пик плотности для группы с ССЗ находится в зоне избыточного веса (25-30 кг/м²).

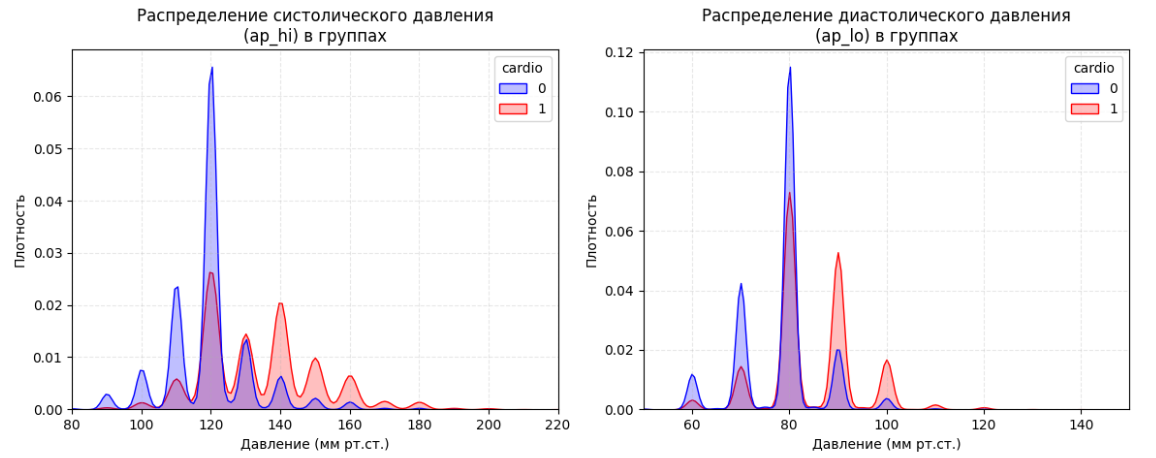
3. Категории BMI:

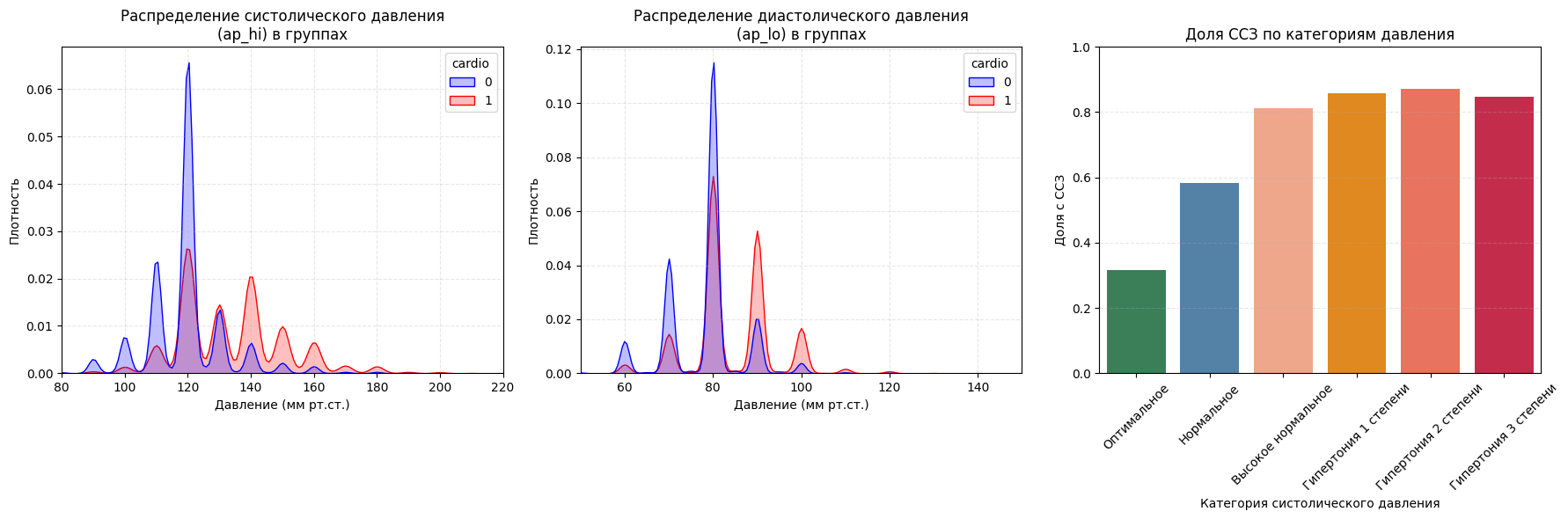
* Четкий рост доли ССЗ с увеличением категории BMI:
* Норма: 39.8%.
* Избыток: 50.6%.
* Ожирение II+: 67.7%.

4. Практические выводы:

* Вес и BMI являются значимыми факторами риска ССЗ.
* Особое внимание следует уделять пациентам с:
* Избыточным весом (BMI 25-30 кг/м²).
* Ожирением (BMI > 30 кг/м²).

## **4.5.3. Анализ зависимости наличия ССЗ от давления.**





Ключевые наблюдения:

1. Распределение систолического давления (ap\_hi):

* У пациентов с ССЗ явный сдвиг распределения вправо.
* Основная масса пациентов с ССЗ имеет значения >140 мм рт.ст.

2. Распределение диастолического давления (ap\_lo):

* Аналогичный сдвиг вправо для группы с ССЗ.
* Критический порог начинается около 90 мм рт.ст.

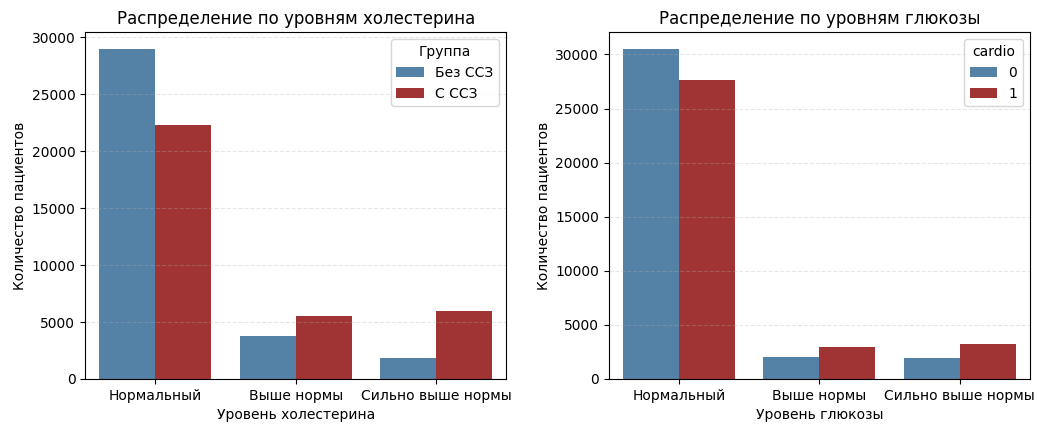
3. Категории давления:

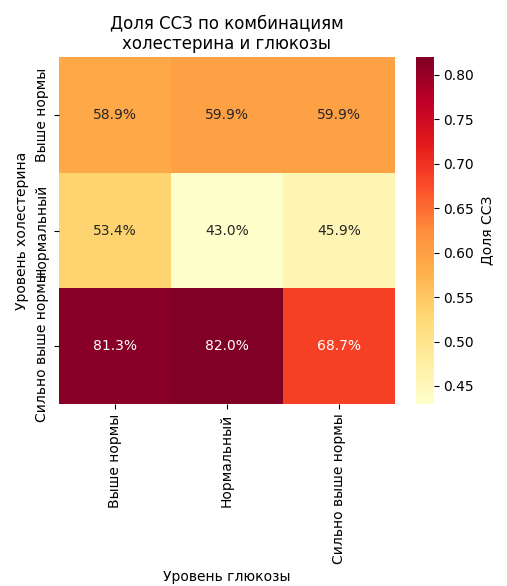
* Четкий рост доли ССЗ с повышением категории:
* Оптимальное (<120): 31.6%.
* Гипертония 1 ст. (140-160): 85.7%.
* Гипертония 3 ст. (>180): 84.6%.

4. Практические выводы:

* Артериальное давление - ключевой фактор риска ССЗ.
* Критические пороги:
* 140/90 мм рт.ст. - точка значительного роста риска.
* 160/100 мм рт.ст. - зона высокого риска.

## **4.5.4. Анализ зависимости наличия ССЗ от уровня холестерина и глюкозы.**





Ключевые наблюдения:

1. Холестерин:

* Четкий рост числа случаев ССЗ с повышением уровня:
* Нормальный: 43.5%
* Сильно выше нормы: 76.2%

2. Глюкоза:

* Менее выраженная, но заметная зависимость:
* Нормальный: 47.5%
* Сильно выше нормы: 61.7%

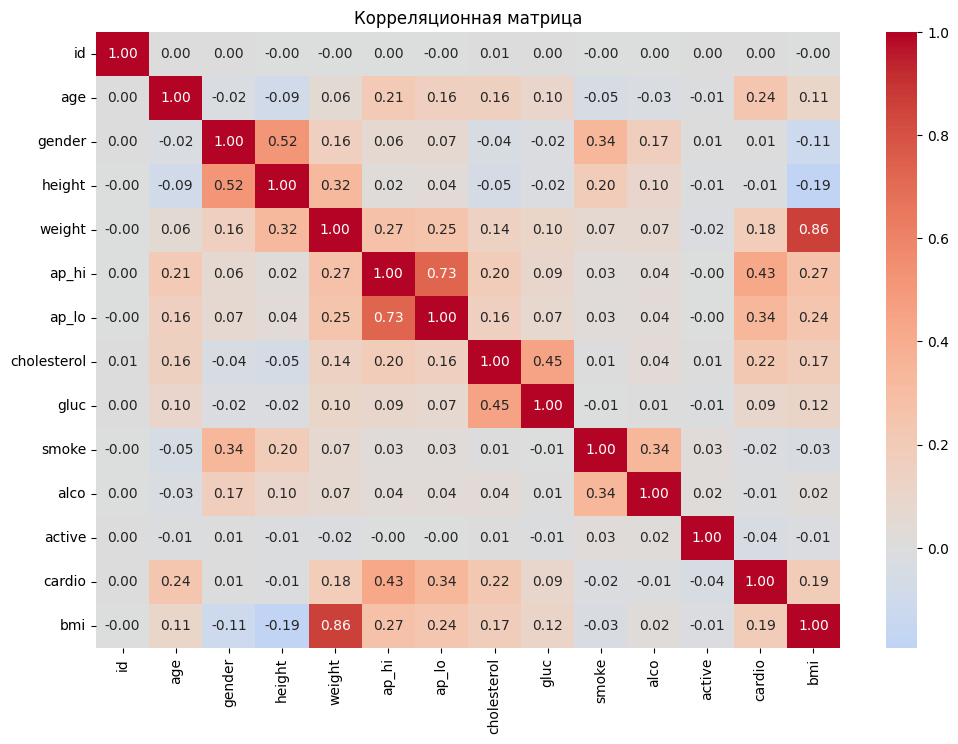
3. Комбинация факторов:

* Наибольший риск при одновременном повышении обоих показателей:
* Холестерин и глюкоза «Сильно выше нормы»: 68.7%
* Минимальный риск при нормальных значениях обоих показателей:
* Оба показателя «Нормальный»: 43.0%

4. Практические выводы:

* Контроль холестерина - приоритетная мера профилактики ССЗ.

## **4.5.4 Матрица корреляции.**



1. Наиболее значимые корреляции с **cardio**:

* **ap\_hi (0.43), ap\_lo (0.34):** Умеренная корреляция.
* **age (0.24):** Умеренная положительная корреляция, что согласуется с медицинскими данными — риск сердечно-сосудистых заболеваний увеличивается с возрастом.
* **weight (0.18)** и **bmi (0.19)**: Слабая положительная связь, указывающая на влияние избыточного веса.
* **cholesterol (0.22):** Умеренная связь, так как высокий уровень холестерина является известным фактором риска.

2. Незначимые корреляции:

* **gender (0.01), height (-0.01), gluc (0.09), smoke (-0.02), alco (-0.01), active (-0.04):** Эти признаки имеют крайне слабую связь с cardio, что означает их минимальное влияние в данной dataset.

3. Другие взаимосвязи:

* **bmi** и **weight (0.86)**: Сильная корреляция, ожидаемо, так как BMI рассчитывается на основе веса и роста. При анализе можно выбрать один из этих признаков, чтобы избежать мультиколлинеарности.
* **cholesterol** и **gluc (0.45)**: Умеренная связь, что может отражать общие метаболические нарушения.

4. Неочевидные наблюдения:

* **height (-0.01)**: Рост практически не влияет на cardio, но имеет слабую отрицательную корреляцию с BMI (-0.29), что логично (более высокие люди могут иметь меньший BMI при том же весе).
* **smoke и alco (0.34):** Умеренная связь между курением и употреблением алкоголя, но оба признака слабо связаны с cardio.

# **5. Построение модели классификации**

Основная цель для модели – оценить индивидуальный риск ССЗ на основе стандартных клинических показателей.

Рассмотрим три модели машинного обучения — логистическая регрессия, случайный лес и XGBoost. Для оценки их работы используем стандартные метрики: точность (precision), полнота (recall), F1-мера, матрица ошибок и ROC-AUC.

Для проведения анализа моделей были выбраны: возраст (age), вес (weight), показатели давления (ap\_hi, ap\_lo), холестерин (cholesterol), ИМТ (bmi).

Целевая переменная cardio.

## **5.1. Логистическая регрессия.**

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

Анализ результатов:

Точность модели:

* На тренировочных данных: 0.7237.
* На тестовых данных: 0.7287.
* Разница: 0.0050.

ROC-AUC показатель:

* На тренировочных данных: 0.7885.
* На тестовых данных: 0.7942.
* Разница: 0.0057.

Модель показывает сходные результаты на train и test, признаки переобучения отсутствуют.

Анализ ошибок на тестовых данных:

* Ложноположительные случаи (FP): 1415 (10.34%)
* Ложноотрицательные случаи (FN): 2297 (16.79%)

## **5.2. Случайный лес.**

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

Анализ результатов

Точность модели:

* На тренировочных данных: 0.9563
* На тестовых данных: 0.6972
* Разница: 0.2591

ROC-AUC показатель:

* На тренировочных данных: 0.9931
* На тестовых данных: 0.7522
* Разница: 0.2409

Внимание: Модель демонстрирует признаки переобучения!

Рекомендации: уменьшить глубину деревьев (max\_depth), увеличить min\_samples\_leaf

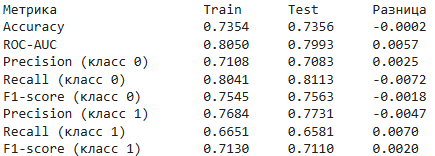
Т.к. модель демонстрирует признаки переобучения подберем оптимальные параметры max\_depth и min\_samples\_leaf.

После оптимизации:

* Точность на тестах выросла с 69.7% до 73.6%.
* Разница уменьшилась с 0.2591 до -0.00.
* Оптимальные параметры ограничивают глубину деревьев и минимальный размер листа.

Анализ финальной модели.

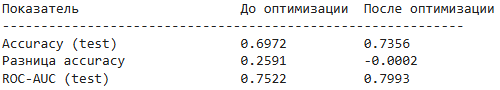
1. Сравнение основных метрик:



2. Диагностика переобучения:

* Переобучение отсутствует (макс. разница 0.0072).

3. Сравнение с исходной моделью:



4. Анализ ошибок классификации:

Всего ошибок: 3618 (26.4%).

* Ложноположительные (FP): 1306 (9.5%).
* Ложноотрицательные (FN): 2312 (16.9%).

Преобладают ложноотрицательные ошибки.

5. Топ-3 наиболее важных признака:

* ap\_hi - 0.30
* ap\_lo - 0.20
* age - 0.15



Наибольший вклад в предсказание вносят:

* Давление (ap\_hi и ap\_lo) совокупно дают 50% важности модели, что указывает на их критическую роль в прогнозировании.
* Возраст и антропометрические данные (bmi и weight) имеют умеренное влияние (по 0.15 и 0.10 соответственно).
* Холестерин (0.10) наименее значим среди анализируемых признаков.

## **5.3. XGBoost.**

Анализ результатов:

Точность модели:

* На тренировочных данных: 0.7565
* На тестовых данных: 0.7342
* Разница: 0.0223

ROC-AUC показатель:

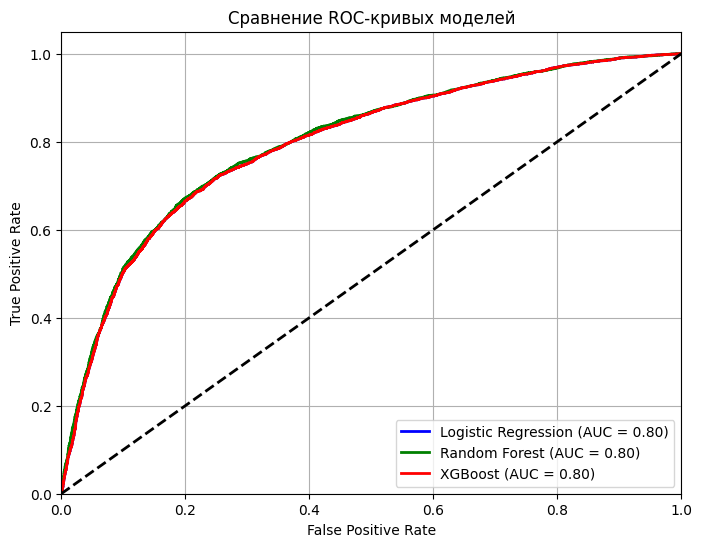
* На тренировочных данных: 0.8357
* На тестовых данных: 0.7957
* Разница: 0.0400

Модель показывает сходные результаты на train и test, признаки переобучения отсутствуют.

Анализ ошибок на тестовых данных:

* Ложноположительные случаи (FP): 1496 (10.93%)
* Ложноотрицательные случаи (FN): 2141 (15.65%)

## **5.4. Выводы по результатам обучения моделей и ROC-анализу:**



Анализ результатов:

1. По качеству моделей (ROC-AUC):

* Random Forest показал наилучший результат (но после оптимизации): 0.7993
* XGBoost: 0.7957
* Logistic Regression: 0.7957

2. По стабильности:

* Logistic Regression и XGBoost не показали переобучения
* Random Forest требовал оптимизации для устранения переобучения

3. По балансу ошибок:

* Все модели чаще ошибаются в сторону ложноотрицательных прогнозов
* Наименьший FN Rate у XGBoost: 15.65%

Рекомендация:

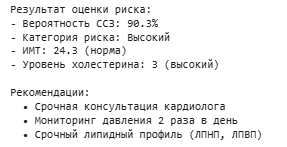
Для дальнейшего использования рекомендуется XGBoost, так как он:

* Демонстрирует наилучшее качество (наибольший AUC).
* Показывает стабильные результаты без переобучения.
* Имеет наиболее сбалансированные ошибки классификации.
* Требует меньше настроек по сравнению с Random Forest.

## **5.5. Пример внедрения модели.**

Входные данные:

* age: 58,
* height: 172,
* weight: 72,
* ap\_hi: 158,
* ap\_lo: 72,
* cholesterol: 3



Сильные стороны:

* Учет комплексного взаимодействия факторов (давление + холестерин + ИМТ).
* Модель демонстрирует высокую уверенность в прогнозе (90.3% вероятность ССЗ для данного пациента), что соответствует категории «Высокий риск».

Основные факторы риска:

* Высокое давление (158/72) — превышает норму (≥140), что автоматически добавляет рекомендацию по мониторингу.
* Возраст (58 лет) — значимый фактор для ССЗ.
* Нормальный ИМТ (24.3) не вносит дополнительный риск, но модель все равно выдала высокую вероятность из-за других факторов.

Потенциальные проблемы: не учитывается другие медицинские признаки, в дальнейшем, можно добавить дополнительные признаки для улучшения прогнозирования вероятности ССЗ.

# **Заключение**

Проведенный анализ включал:

* Предварительную обработку и очистку данных.
* Детальное изучение характеристик всех показателей.
* Выявление взаимосвязей между различными параметрами и заболеваемостью.
* Сравнение эффективности различных прогностических моделей.

Ключевые закономерности:

* Наибольшее влияние на риск ССЗ оказывают: артериальное давление, уровень холестерина, показатели ИМТ и возраст пациентов.
* Вредные привычки (курение, алкоголь) не демонстрируют прямой связи с заболеваемостью, но могут косвенно влиять через другие факторы.

Результаты моделирования:

* Наилучшие показатели продемонстрировала модель градиентного бустинга:
* Демонстрирует наилучшее качество (наибольший AUC)
* Показывает стабильные результаты без переобучения
* Имеет наиболее сбалансированные ошибки классификации
* Требует меньше настроек по сравнению с Random Forest

Рекомендации:

Основные направления профилактики:

* Регулярный контроль давления, холестерина.
* Поддержание нормального веса.
* Усиленное наблюдение за возрастными пациентами.

Перспективы дальнейших исследований:

* Включение дополнительных параметров.
* Совершенствование алгоритмов прогнозирования.
* Изучение скрытых факторов, влияющих на заболеваемость.

Выявленные закономерности подтверждают известные медицинские данные, но также указывают на необходимость более глубокого изучения механизмов развития сердечно-сосудистых заболеваний.