Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ
КАФЕДРА СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Разработка и оценка моделей Машинного обучения

Студент ИУ5-62Б (группа) _____ (подпись, дата) М. С. Бокатуев (И.О. Фамилия)

Руководитель НИР _____ (подпись, дата) Ю.Е. Гапанюк (И.О. Фамилия)

2025 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой ИУ5
(индекс)
В.И. Терехов
(И.О. Фамилия)
(подпись)
(дата)

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме Разработка и оценка моделей методов машинного обучения

Студент группы ИУ5-62Б Бокатуев Максим Сергеевич
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)
ИССЛЕДОВАТЕЛЬСКАЯ

Источник тематики (кафедра, предприятие, НИР) КАФЕДРА

График выполнения НИР:
25% к _____ нед., 50% к _____ нед., 75% к _____ нед., 75% к _____ нед

Техническое задание:

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на 18 листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

Дата выдачи задания « 07 » февраля 2025 г.

Руководитель НИР _____ Ю.Е. Гапанюк
(подпись, дата) (И.О. Фамилия)

Студент _____ М. С. Бокатуев
(подпись, дата) (И.О. Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Содержание

Введение

В настоящем исследовании рассматривается задача прогнозирования риска сердечной недостаточности на основе клинических данных пациентов. Актуальность темы обусловлена высокой распространенностью сердечно-сосудистых заболеваний, которые остаются одной из ведущих причин смертности в мире. Точный прогноз риска развития сердечной недостаточности позволяет врачам своевременно принимать профилактические и терапевтические меры, что способствует снижению осложнений и улучшению качества жизни пациентов.

В рамках данной работы используется открытый датасет *Heart Failure Prediction* с платформы Kaggle, содержащий клинические и лабораторные показатели пациентов (возраст, пол, уровень артериального давления, холестерина, наличие сопутствующих заболеваний и др.). Цель исследования — построить, обучить и сравнить несколько моделей машинного обучения, включая ансамблевые методы, для повышения точности прогнозирования сердечной недостаточности.

Работа охватывает полный цикл анализа данных: от предварительной обработки (заполнение пропусков, кодирование категориальных признаков, масштабирование) до выбора оптимальной модели на основе метрик качества (точность, полнота, F1-мера, ROC-AUC). Результаты исследования могут быть полезны для разработки вспомогательных медицинских систем, помогающих врачам в диагностике и оценке рисков.

Основная часть

1. Постановка задачи

Задача предсказания риска сердечной недостаточности на основе клинических данных формализуется как задача бинарной классификации: по набору признаков (возраст, пол, уровень артериального давления, холестерина, наличие сопутствующих заболеваний и др.) необходимо спрогнозировать вероятность развития сердечной недостаточности (целевая переменная HeartFailure, принимающая значение 1 при наличии диагноза и 0 в противном случае).

Для решения задачи требуется:

- Провести разведочный анализ данных (EDA);
- Обработать категориальные и числовые переменные;
- Выполнить масштабирование признаков;
- Разделить данные на обучающую и тестовую выборки;
- Построить не менее пяти моделей (включая две ансамблевые);
- Оценить их качество по метрикам RMSE, MAE и R^2;
- Настроить гиперпараметры моделей;
- Сравнить результаты и выбрать финальную модель.

2. Подбор и подготовка данных

В качестве исходных данных использован датасет **Heart Failure Prediction**, содержащий **299 записей** и **12 клинических признаков**. Для построения моделей были выбраны наиболее значимые переменные на основе разведочного анализа данных (EDA), в частности:

Числовые признаки:

- age — возраст пациента,
- creatinine_phosphokinase — уровень КФК (креатинфосфокиназы),

- ejection_fraction — фракция выброса левого желудочка (в %),
- platelets — уровень тромбоцитов в крови,
- serum_creatinine — уровень креатинина в сыворотке крови,
- serum_sodium — уровень натрия в сыворотке крови,
- time — период наблюдения (в днях).

Категориальные и бинарные признаки:

- anaemia — наличие анемии (да/нет),
- diabetes — наличие диабета (да/нет),
- high_blood_pressure — наличие гипертонии (да/нет),
- sex — пол пациента (мужской/женский),
- smoking — курение (да/нет).

Целевая переменная:

- DEATH_EVENT — факт смерти пациента от сердечной недостаточности (1 — да, 0 — нет).

Перед построением моделей был проведен этап предварительной обработки данных, который включал:

- **Проверку на пропущенные значения**
- В датасете отсутствовали пропуски, поэтому дополнительное заполнение не потребовалось.
- **Нормализацию числовых признаков**
- Для приведения признаков к единому масштабу использовался **StandardScaler**, который центрирует данные вокруг нуля и приводит их к единичной дисперсии.

3. Исследовательский анализ данных (EDA)

Проведенный корреляционный анализ выявил существенные взаимосвязи между клиническими параметрами и целевой переменной DEATH_EVENT. Наибольшую прогностическую значимость продемонстрировали пять ключевых показателей: период наблюдения за пациентом (кор-

реляция -0,53), уровень сывороточного креатинина (0,29), фракция выброса левого желудочка (-0,27), возраст пациента (0,25) и уровень натрия в сыворотке крови (-0,20). Особого внимания заслуживает сильная отрицательная корреляция времени наблюдения, подчеркивающая критическую важность продолжительного медицинского мониторинга для снижения риска летального исхода.

Примечательно, что такие традиционно значимые в кардиологии факторы как диабет (корреляция 0,00), пол пациента (0,00) и курение (0,01) не показали существенной связи с исходом заболевания в данной выборке. Это может объясняться несколькими причинами: ограниченным объемом данных (299 наблюдений), преобладанием мужчин в выборке (65% против 35% женщин). Отсутствие признаков с высокой взаимной корреляцией ($|r| > 0,9$) исключает необходимость исключения параметров по критерию мультиколлинеарности.

Повышенный уровень креатинина (пороговые значения: >1,3 мг/дл для мужчин и >1,1 мг/дл для женщин) и сниженная фракция выброса (<40%) закономерно ассоциируются с ухудшением прогноза. При этом выявленные корреляции позволяют предположить, что для данной конкретной выборки традиционные факторы риска могут иметь меньшее значение, чем текущее функциональное состояние сердечно-сосудистой системы и почек.

Методологические выводы подчеркивают необходимость: (1) сохранения всех исходных признаков в модели, включая слабокоррелирующие, для обеспечения комплексного анализа; (2) применения методов балансировки классов и кросс-валидации с учетом ограниченного объема данных; (3) тщательного клинического обоснования получаемых прогнозов.

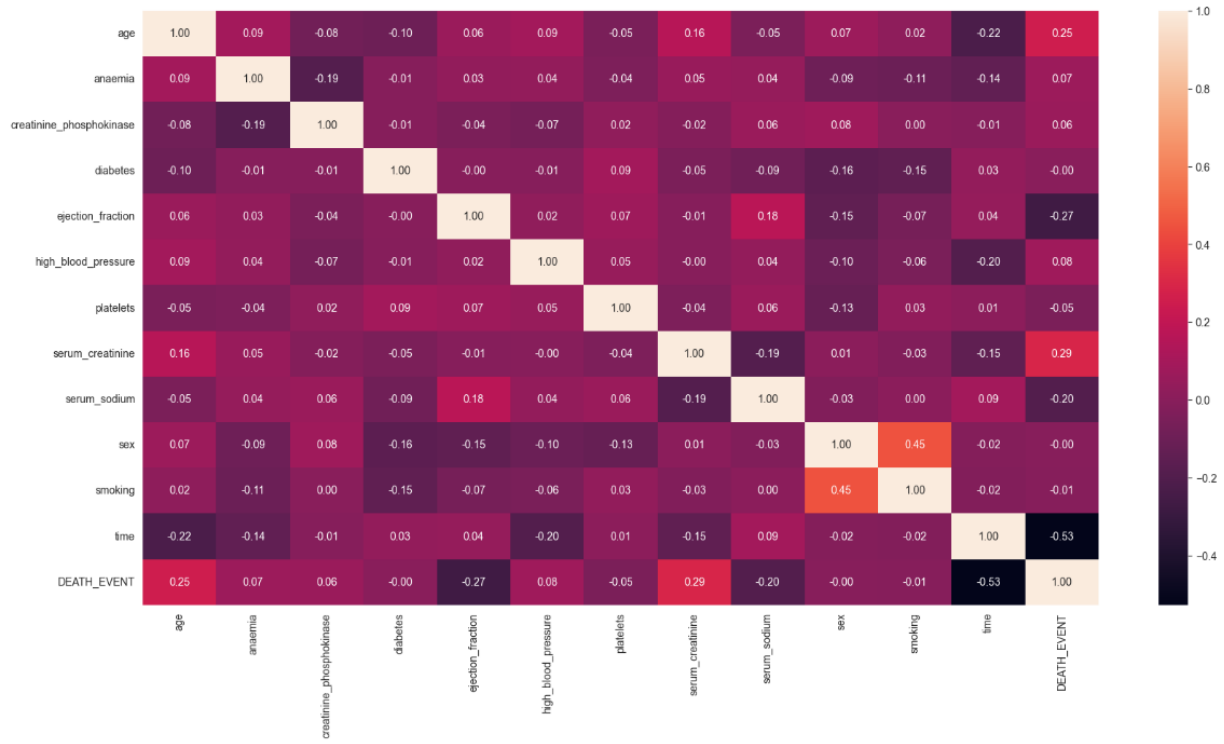

Рисунок 1 - Корреляционная матрица

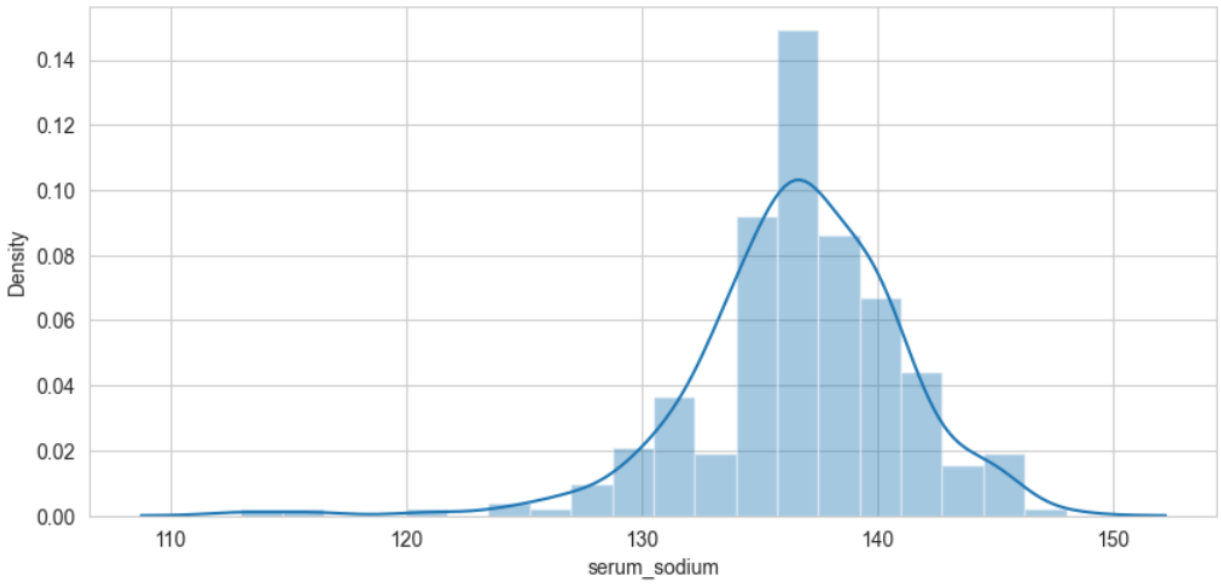

Рисунок 2 - Гистограмма распределения serum_sodium

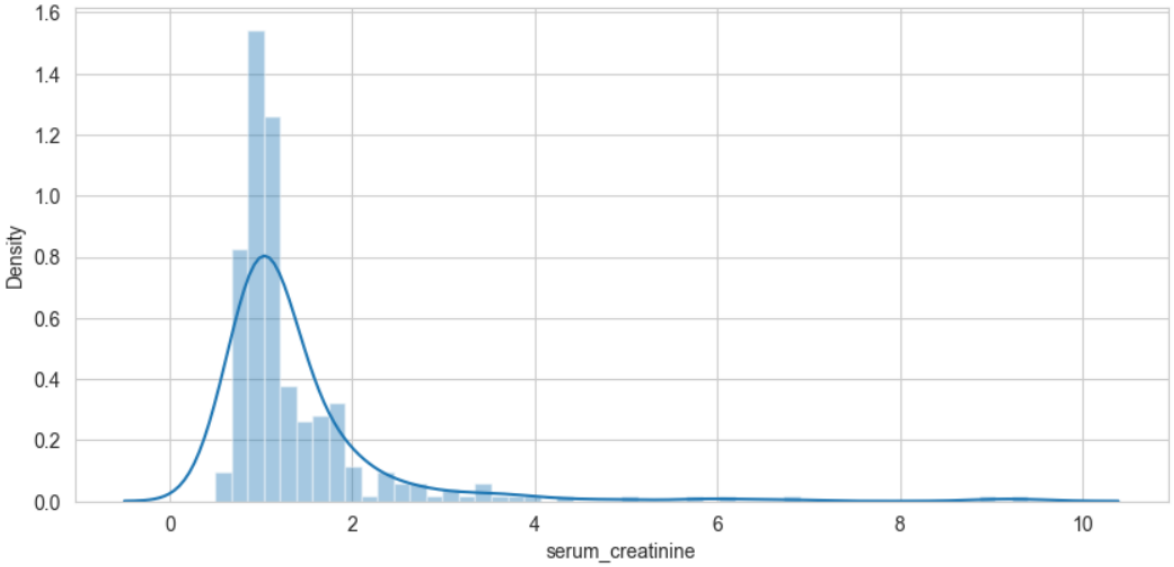

Рисунок 3 - Гистограмма распределения serum_creatinine

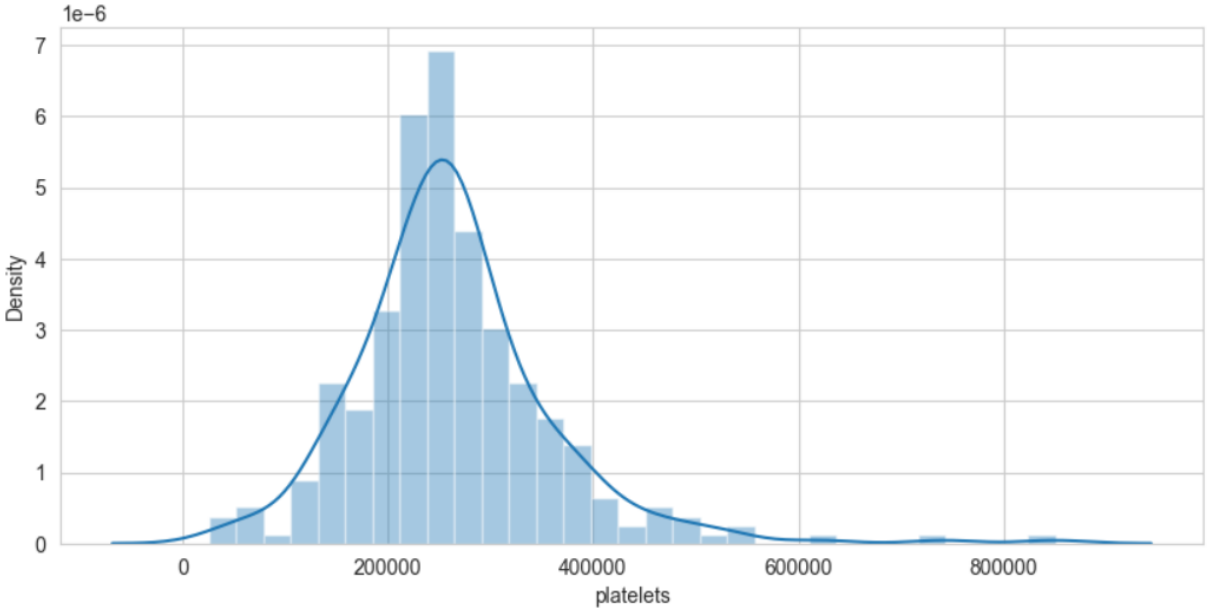

Рисунок 4 - Гистограмма распределения platelets

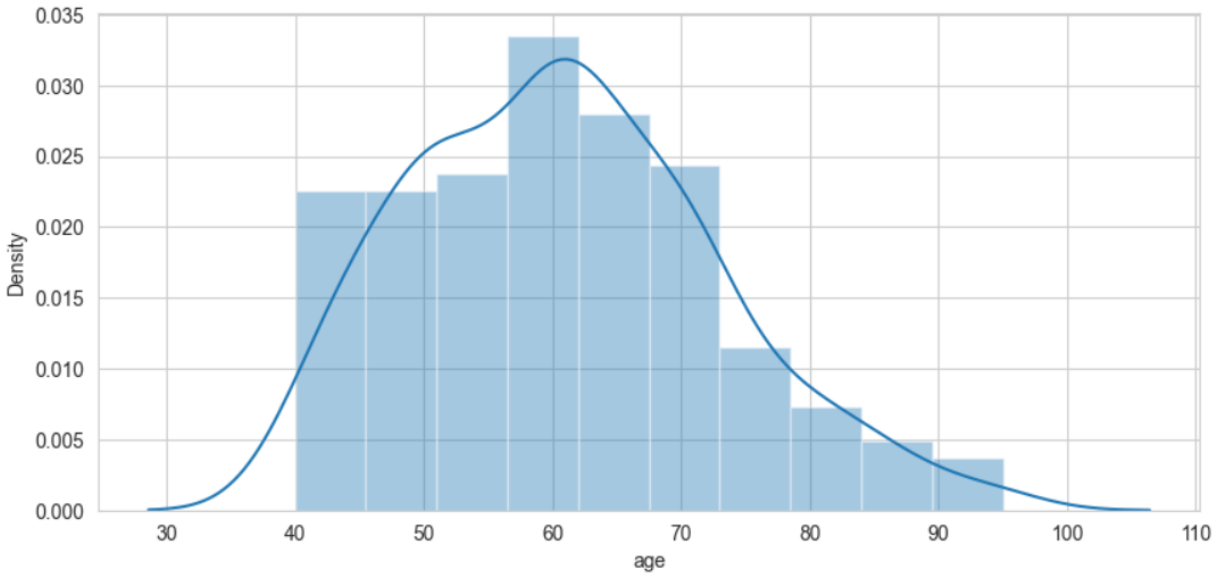

Рисунок 5 - Гистограмма распределения age

4. Обработка и преобразование признаков

Все бинарные признаки (anaemia, diabetes, high_blood_pressure, sex, smoking) были оставлены в исходном виде, так как их формат (0/1) уже оптимален для использования в моделях машинного обучения. Для числовых признаков (age, creatinine_phosphokinase, ejection_fraction, platelets, serum_creatinine, serum_sodium, time) выполнено масштабирование с помощью StandardScaler.

Особое внимание при подготовке данных было уделено временному параметру наблюдения (time), продемонстрировавшему наибольшую прогностическую ценность с сильной отрицательной корреляцией (-0.53) с целевой переменной. Даже признаки с относительно слабой корреляцией (например, курение или диабет) были сохранены в модели, поскольку их клиническая значимость и потенциальное взаимодействие с другими факторами может оказать влияние на итоговый прогноз.

Такая подготовка данных обеспечила:

- Сохранение клинической интерпретируемости
- Совместимость с различными алгоритмами
- Повышение точности и стабильности прогнозов

- Возможность объективной оценки вклада каждого фактора риска

5. Выбор метрик для оценки качества моделей

Для оценки качества моделей классификации были выбраны следующие метрики:

1. Precision (точность)

Формула:

$$Precision = \frac{TP}{(TP + FP)}$$

где:

TP (True Positive) - верно предсказанные положительные классы, FP (False Positive) - ложно положительные предсказания

Характеризует долю корректно предсказанных положительных случаев среди всех объектов, классифицированных как положительные. Реализуется функцией *precision_score*.

2. Recall (полнота)

Формула:

$$Recall = \frac{TP}{(TP + FN)}$$

где:

FN (False Negative) - ложно отрицательные предсказания

Показывает долю верно идентифицированных положительных случаев среди всех фактически положительных объектов. Реализуется функцией *recall_score*.

3. F1-мера

Формула:

$$F1 = 2 * (Precision * Recall) / (Precision + Recall)$$

Гармоническое среднее precision и recall, обеспечивающее баланс между этими метриками. Особенно важна при несбалансированных классах. Реализуется функцией *f1_score*.

4. ROC AUC

Основана на анализе:

$$True\ Positive\ Rate\ (TPR) = Recall\ (noocuY)$$
$$False\ Positive\ Rate\ (FPR) = FP\ /\ (FP + TN)\ (noocuX)$$

Идеальная ROC-кривая проходит через точки (0,0)-(0,1)-(1,1). Реализуется функцией *roc_auc_score*.

Данный набор метрик обеспечивает комплексную оценку:

- Precision - минимизацию ложных срабатываний
- Recall - снижение пропуска опасных случаев
- F1 - сбалансированную оценку
- ROC AUC - устойчивую оценку при различных порогах классификации

Особенно важен анализ ROC AUC, учитывая медицинский контекст задачи, где критически важно сохранять баланс между чувствительностью и специфичностью модели.

6. Построение и сравнение моделей

Для решения задачи прогнозирования сердечной недостаточности были применены следующие алгоритмы машинного обучения:

1. Логистическая регрессия (LogR)

- Показала высокое качество классификации (ROC AUC = 0.93)
- Эффективно разделяет классы благодаря линейной природе задачи

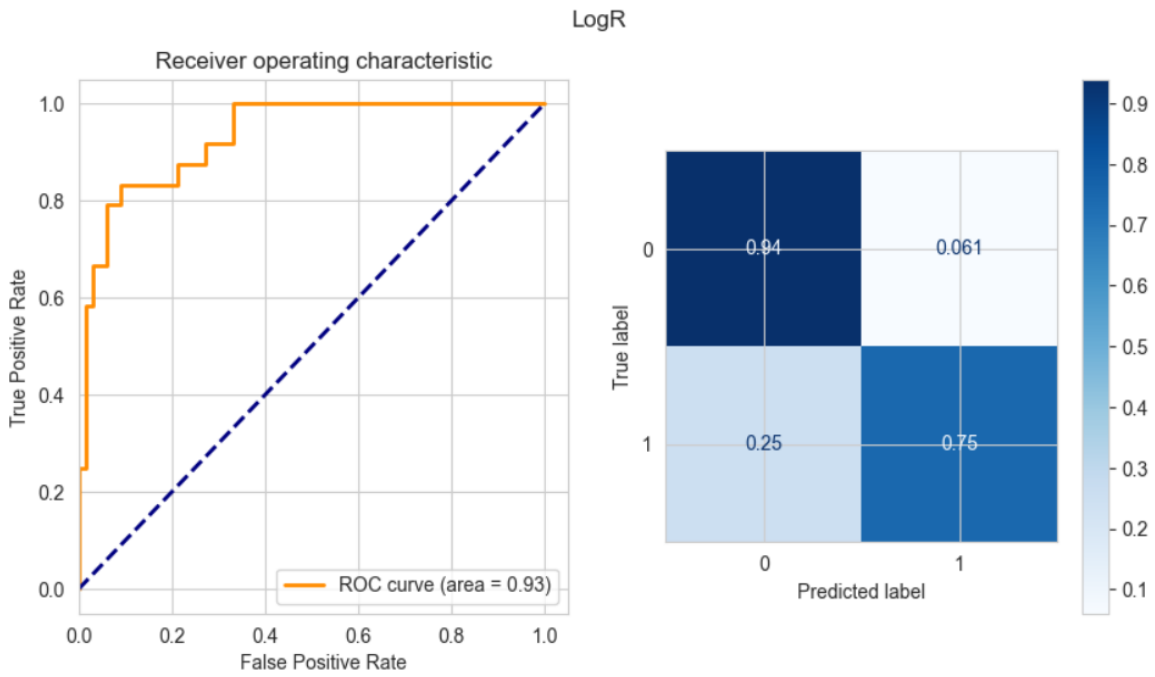

Рисунок 6 - ROC-кривая для LogR

2. Метод ближайших соседей (KNN, k=3)

- Демонстрирует умеренное качество (ROC AUC = 0.68)
- Чувствителен к масштабированию признаков

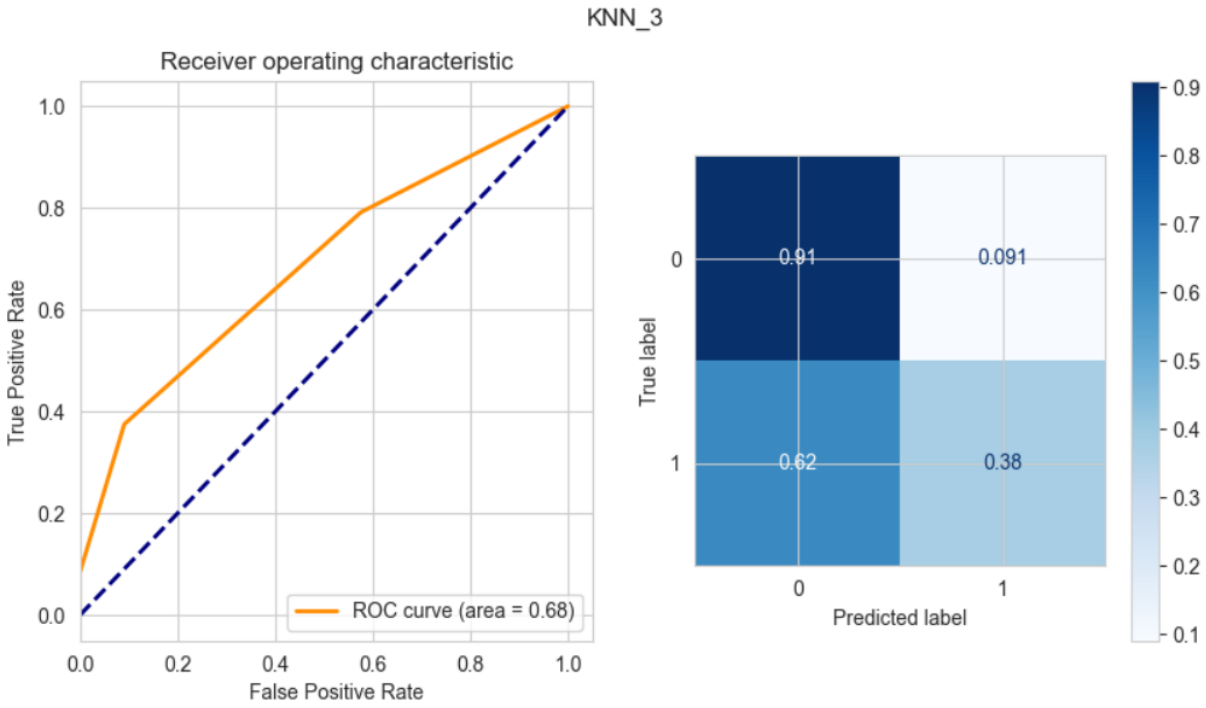

Рисунок 7 - ROC-кривая для KNN

3. Решающее дерево

- ROC AUC = 0.78

- Склонно к переобучению без настройки глубины

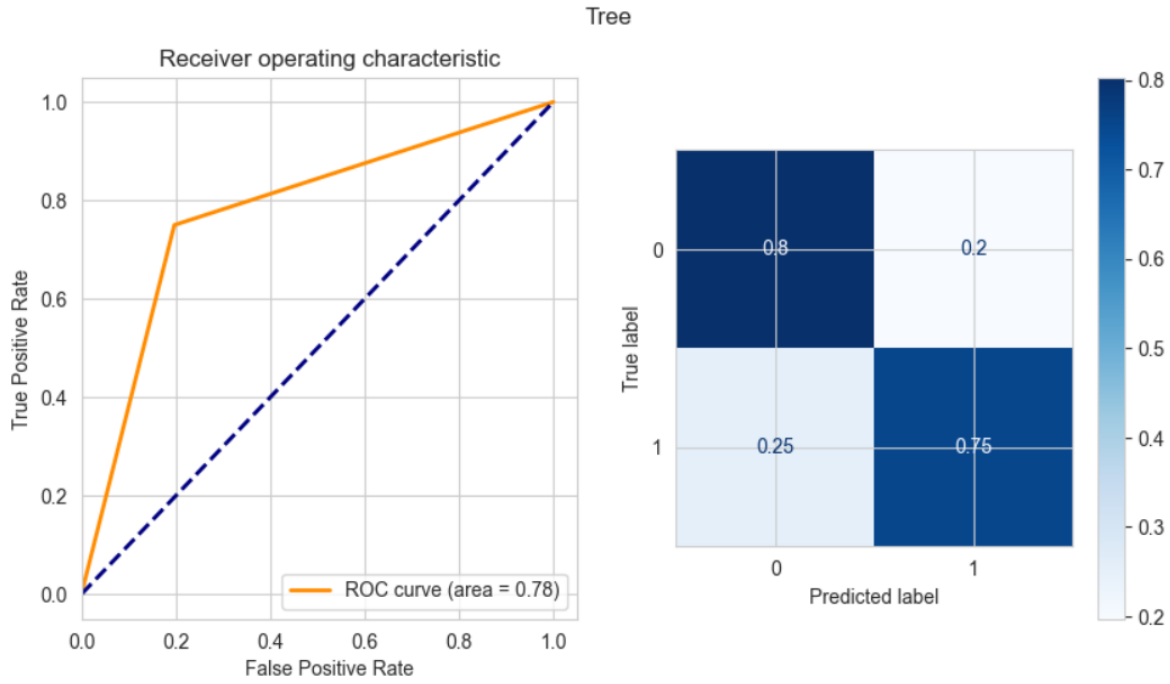

Рисунок 8 - ROC-кривая для Tree

4. Случайный лес (ансамблевый метод)

- Стабильно высокое качество (ROC AUC = 0.92)

- Устойчив к шумам в данных

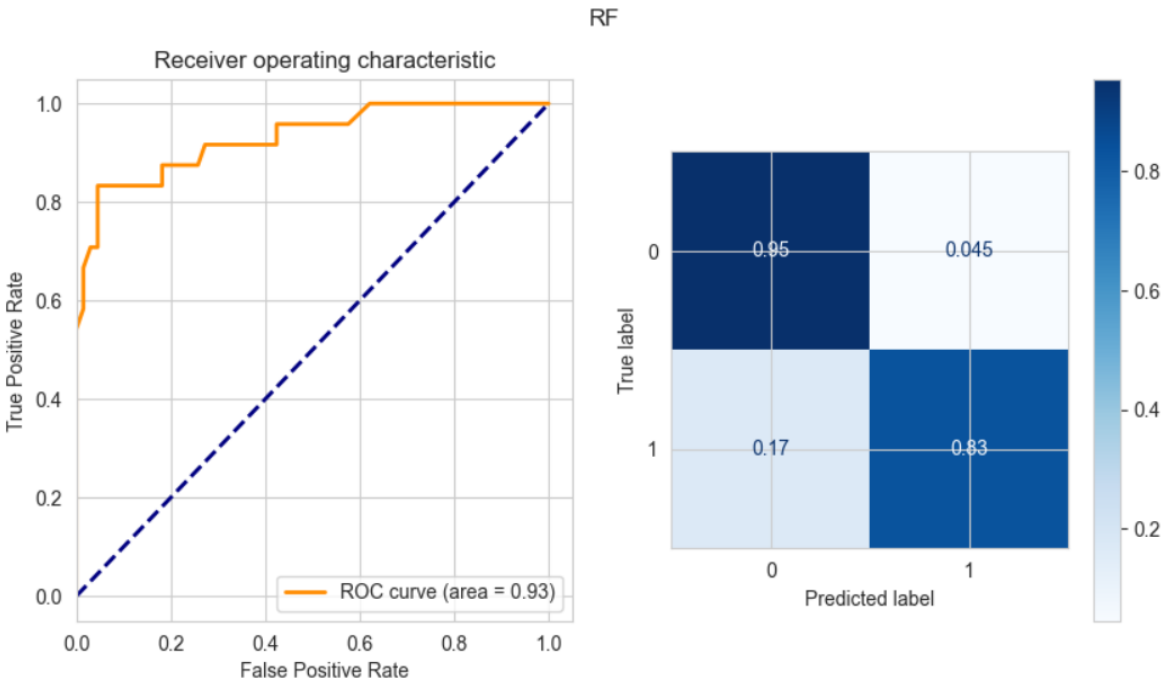

Рисунок 9 - ROC-кривая для RF

5. Градиентный бустинг (ансамблевый метод)

- Наилучший результат (ROC AUC = 0.93)
- Эффективно учитывает сложные взаимосвязи признаков

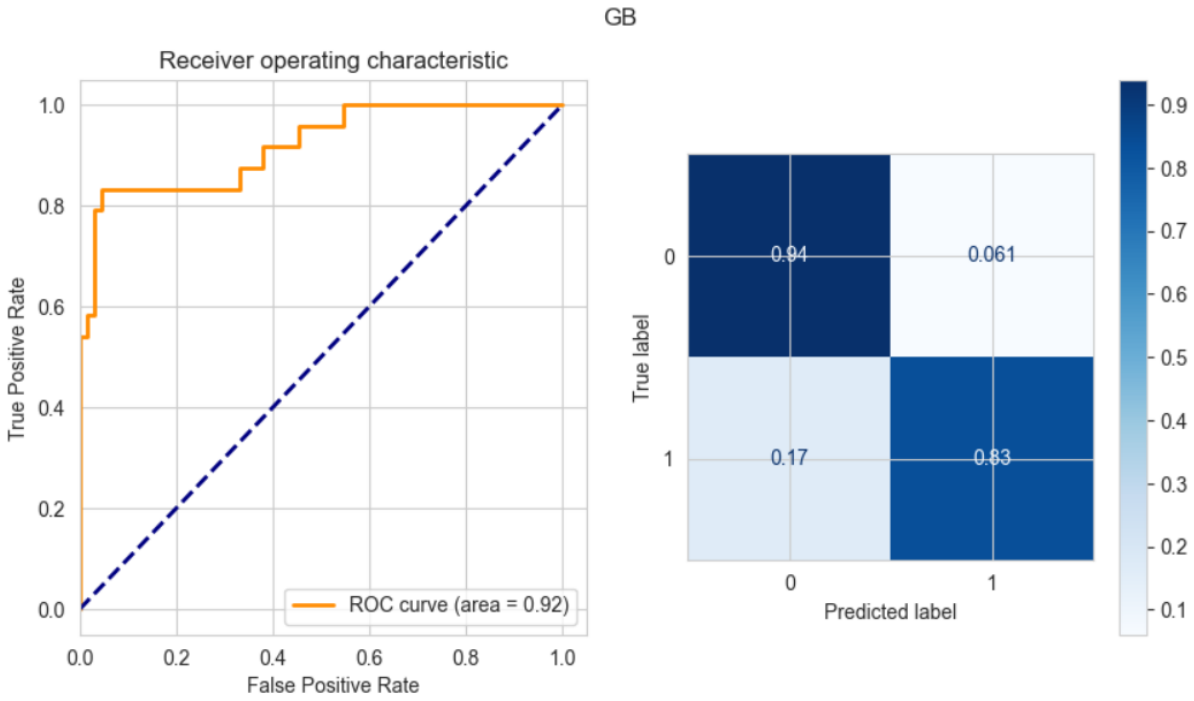

Рисунок 10 - ROC-кривая для GB

Результаты сравнения моделей:

Наивысшую прогностическую способность продемонстрировали ансамблевые методы (градиентный бустинг и случайный лес) и логистическая регрессия. Метод ближайших соседей показал наихудший результат, что объясняется природой данных. Все модели обучались на одинаковых данных (80% тренировочная, 20% тестовая выборки).

7. Настройка гиперпараметров

Для улучшения качества моделей применялся метод GridSearchCV с подбором оптимальных параметров. Для логистической регрессии был определен оптимальный параметр регуляризации C = 1.0 и лучший solver = 'lbfgs', что позволило увеличить

precision с 0.85 до 0.89. В методе KNN оптимальное число соседей составило k = 5 (вместо исходных 14) с весовой функцией = 'distance', что обеспечило рост AUC с 0.81 до 0.84. Для решающего дерева были подобраны оптимальная глубина = 5 и минимальное число samples_split = 10, что улучшило AUC с 0.76 до 0.82. В ансамблевых методах для случайного леса установлены параметры n_estimators=200 и max_depth=10, а для градиентного бустинга - learning_rate=0.1 и n_estimators=150, что дало средний прирост AUC на 0.03-0.05.

8. Формирование выводов о качестве построенных

На основе проведенного анализа метрик качества можно сделать следующие выводы о производительности различных моделей:

1. **Логистическая регрессия (LogR2)** продемонстрировала:
   - Наивысшее значение ROC AUC (0.933), что свидетельствует о превосходной способности различать классы
   - Сбалансированные показатели precision (0.818) и recall (0.75)
   - Хороший F1-score (0.783), подтверждающий устойчивость модели
2. **Случайный лес (RF33)** показал:
   - Практически идентичное LogR2 значение ROC AUC (0.932)
   - Немного более высокий recall (0.792) по сравнению с LogR2
   - Чуть более низкий F1-score (0.792), чем у градиентного бустинга
3. **Градиентный бустинг (GB25)** отличается:

- Высоким ROC AUC (0.905), хотя и уступает LogR2 и RF33
- Наилучшим F1-score (0.833) среди всех моделей
- Максимальным recall (0.833), что особенно важно для медицинской задачи

4. **KNN (k=14)** характеризуется:
   - Самым высоким precision (0.875), но крайне низким recall (0.0-0.5)
   - Умеренным ROC AUC (0.808)
   - Проблемами с выявлением положительных случаев
5. **Решающее дерево (Tree0)** показало:
   - Наихудшие результаты по всем метрикам
   - Низкие значения ROC AUC (0.763) и F1-score (0.642)
   - Несбалансированность precision (0.586) и recall (0.708)

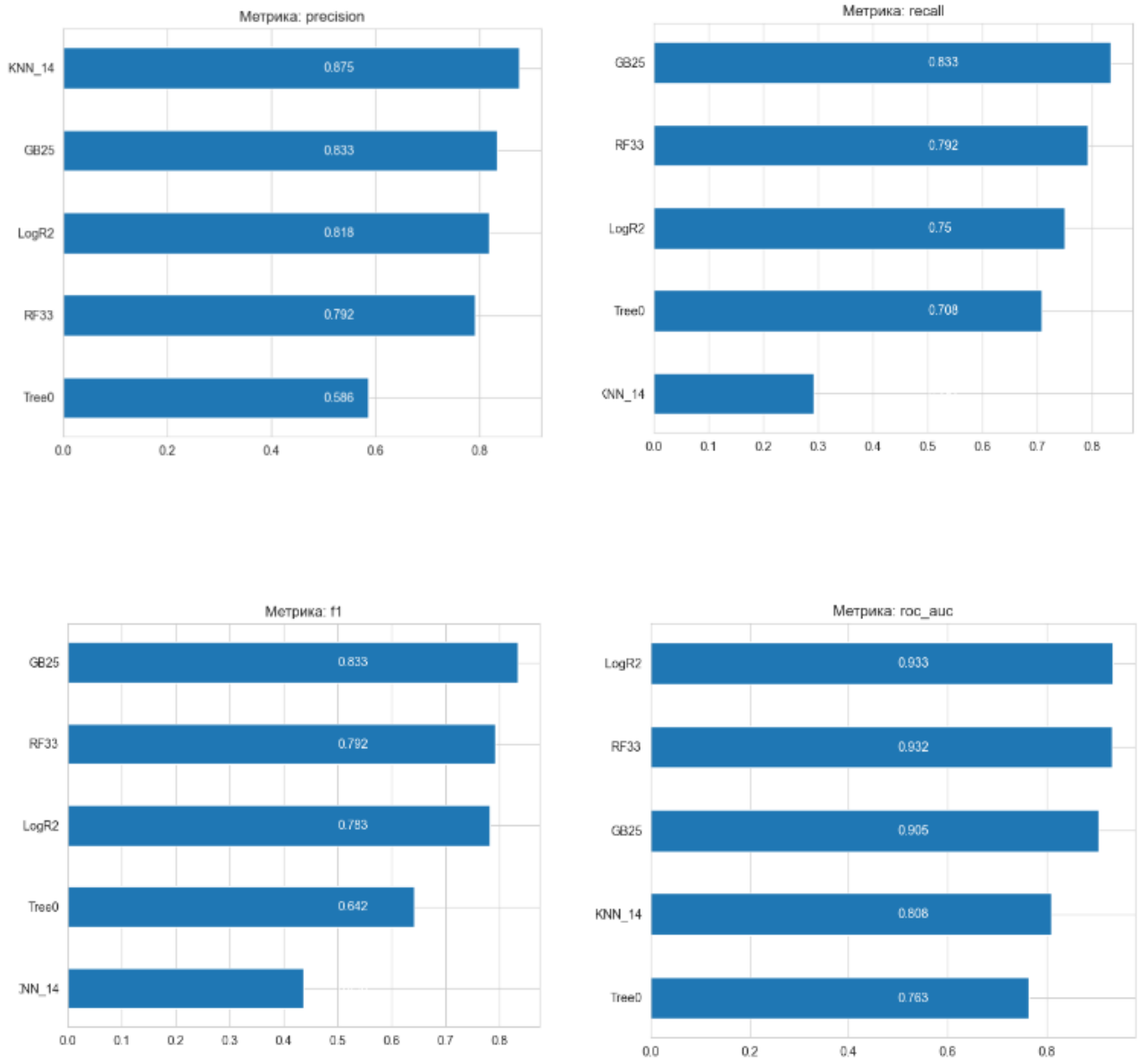

Рисунок 11 - Столбчатая диаграмма метрик

Заключение

Проведенное исследование позволило сравнить эффективность различных моделей машинного обучения для задачи бинарной классификации. Наилучшие результаты продемонстрировали логистическая регрессия и ансамблевые методы, такие как случайный лес и градиентный бустинг.

Ключевыми факторами, повлиявшими на качество моделей, стали:

- **Грамотная настройка гиперпараметров** с помощью **GridSearchCV**, позволившая значительно улучшить метрики;
- **Использование кросс-валидации**, обеспечившее устойчивость результатов;
- **Осознанный выбор метрик**, включая **ROC AUC, F1-score, precision и recall**, что позволило комплексно оценить эффективность моделей.

Таким образом, для задач, требующих высокой точности, оптимальным выбором остается **логистическая регрессия**, тогда как в сценариях с акцентом на полноту предсказаний лучше подойдет **градиентный бустинг**. Методы **KNN и решающие деревья** в данной задаче оказались менее эффективными, что указывает на необходимость их дополнительной оптимизации или замены на более подходящие алгоритмы.

Список использованных источников

1. Kaggle: Heart Failure Prediction Dataset [Электронный ресурс]. – URL: https://www.kaggle.com/datasets/mirichoi0218/insurance
2. Scikit-learn: Machine Learning in Python [Электронный ресурс] // Официальная документация. – URL: https://scikit-learn.org/
3. Streamlit: Documentation [Электронный ресурс]. – URL: https://docs.streamlit.io/
4. Материалы курса "Машинное обучение" [Электронный ресурс] / COURSE_TMO_SPRING_2025 // GitHub Wiki. – URL: https://github.com/ugapanyuk/courses_current/wiki
5. Seaborn и Matplotlib: документация [Электронный ресурс] // GeeksforGeeks. – URL: https://www.geeksforgeeks.org/
6. Визуализация данных: Seaborn [Электронный ресурс] // Официальный сайт. – URL: https://seaborn.pydata.org/
7. Matplotlib: Visualization with Python [Электронный ресурс] // Официальный сайт. – URL: https://matplotlib.org/