|  |  |
| --- | --- |
|  | **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)** |

ФАКУЛЬТЕТ ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ

КАФЕДРА СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***Прогнозирование сердечной недостаточности \_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент   ИУ5-31М **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ С.В. Гришин\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **Ю.Е. Гапанюк**

(Подпись, дата) (И.О.Фамилия)

Консультант \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2024 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)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме   Прогнозирование сердечной недостаточности \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы   ИУ5-31М

  Гришин Станислав Васильевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

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

  учебная

Источник тематики (кафедра, предприятие, НИР)   кафедра

График выполнения НИР: 25% к \_\_\_ нед., 50% к \_\_\_ нед., 75% к \_\_ нед., 100% к \_\_\_ нед.

***Техническое задание***   *Провести разведочный анализ данных. Провести корреляционный анализ данных. Выбрать наиболее подходящие модели и сделать выводы о качестве построенных моделей.*

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на \_15\_ листах формата А4.

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания «\_\_» \_\_\_\_\_\_\_\_ 2024 г.

**Руководитель НИР**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Ю.Е. Гапанюк**

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ С.В. Гришин\_ \_**

(Подпись, дата) (И.О.Фамилия)

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

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 4](#_Toc154431314)

[1. Описание датасета 5](#_Toc154431314)

[2. Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. 6](#_Toc154431315)

[3. Выбор признаков, подходящих для построения моделей. Кодирование категориальных признаков. Масштабирование данных. 9](#_Toc154431316)

[4. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения. 10](#_Toc154431321)

[5. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии. 11](#_Toc154431324)

[6. Формирование выводов о качестве построенных моделей на основе выбранных метрик. 11](#_Toc154431325)

[ЗАКЛЮЧЕНИЕ 14](#_Toc154431326)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 15](#_Toc154431327)

**ВВЕДЕНИЕ**

В качестве предметной области был выбран датасет с информацией о сердечной недостаточности. В исследовании будет решаться задача бинарной классификации.

Сердечно-сосудистые заболевания (ССЗ) являются причиной смерти номер 1 во всем мире, унося примерно 17,9 миллиона жизней ежегодно, что составляет 31% всех смертей в мире. Четыре из 5 смертей от сердечно-сосудистых заболеваний связаны с сердечными приступами и инсультами, и одна треть этих смертей происходит преждевременно среди людей в возрасте до 70 лет. Сердечная недостаточность является распространенным явлением, вызванным сердечно-сосудистыми заболеваниями, и этот набор данных содержит 11 признаков, которые можно использовать для прогнозирования возможного заболевания сердца.

Люди с сердечно-сосудистыми заболеваниями или с высоким сердечно- сосудистым риском (из-за наличия одного или нескольких факторов риска, таких как гипертония, диабет, гиперлипидемия или уже установленное заболевание) нуждаются в раннем выявлении и лечении, в чем большую помощь может оказать модель машинного обучения.

1. Описание датасета

В качестве набора данных мы будем использовать набор данных прогнозирования инсульта: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Age: возраст пациента [лет]

Sex: пол пациента [M: Мужской, F: Женский]

ChestPainType: тип боли в груди [TA: типичная стенокардия, ATA: атипичная стенокардия, NAP: неангинальная боль, ASY: бессимптомная]

RestingBP: артериальное давление в состоянии покоя [мм рт.ст.]

Cholesterol: холестерин сыворотки [мм/дл]

FastingBS: уровень сахара в крови натощак [1: если FastingBS > 120 мг/дл, 0: иначе]

RestingECG: результаты электрокардиограммы в покое [Normal: нормальная, ST: анома- лия ST-T (инверсия T и/или элевация или депрессия ST > 0,05 мВ), LVH: вероятная или определенная гипертрофия левого желудочка по критериям Эстеса]

MaxHR: максимальная достигнутая частота сердечных сокращений [Числовое значение от 60 до 202]

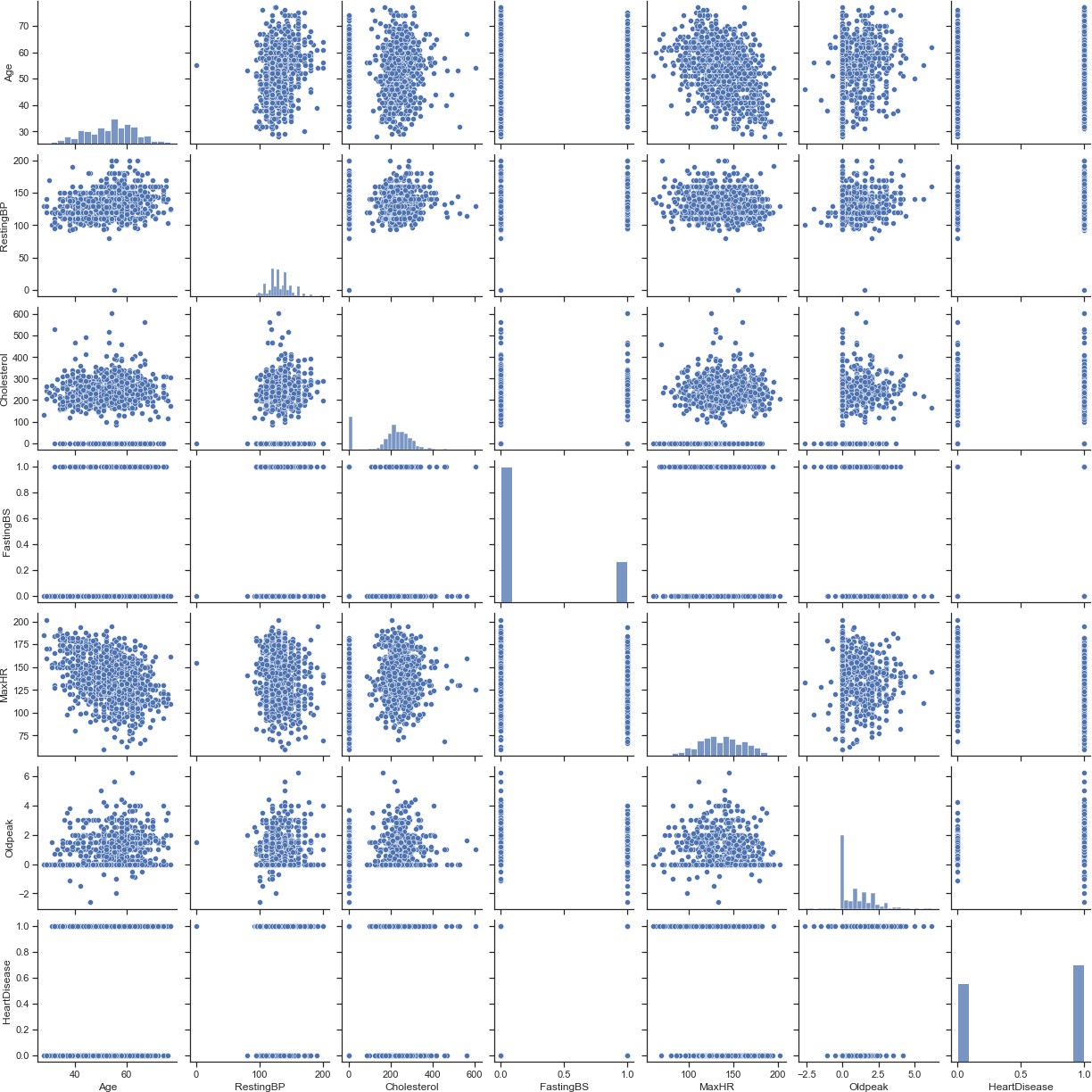
ExerciseAngina: стенокардия, вызванная физической нагрузкой [Y: Да, N: Нет]

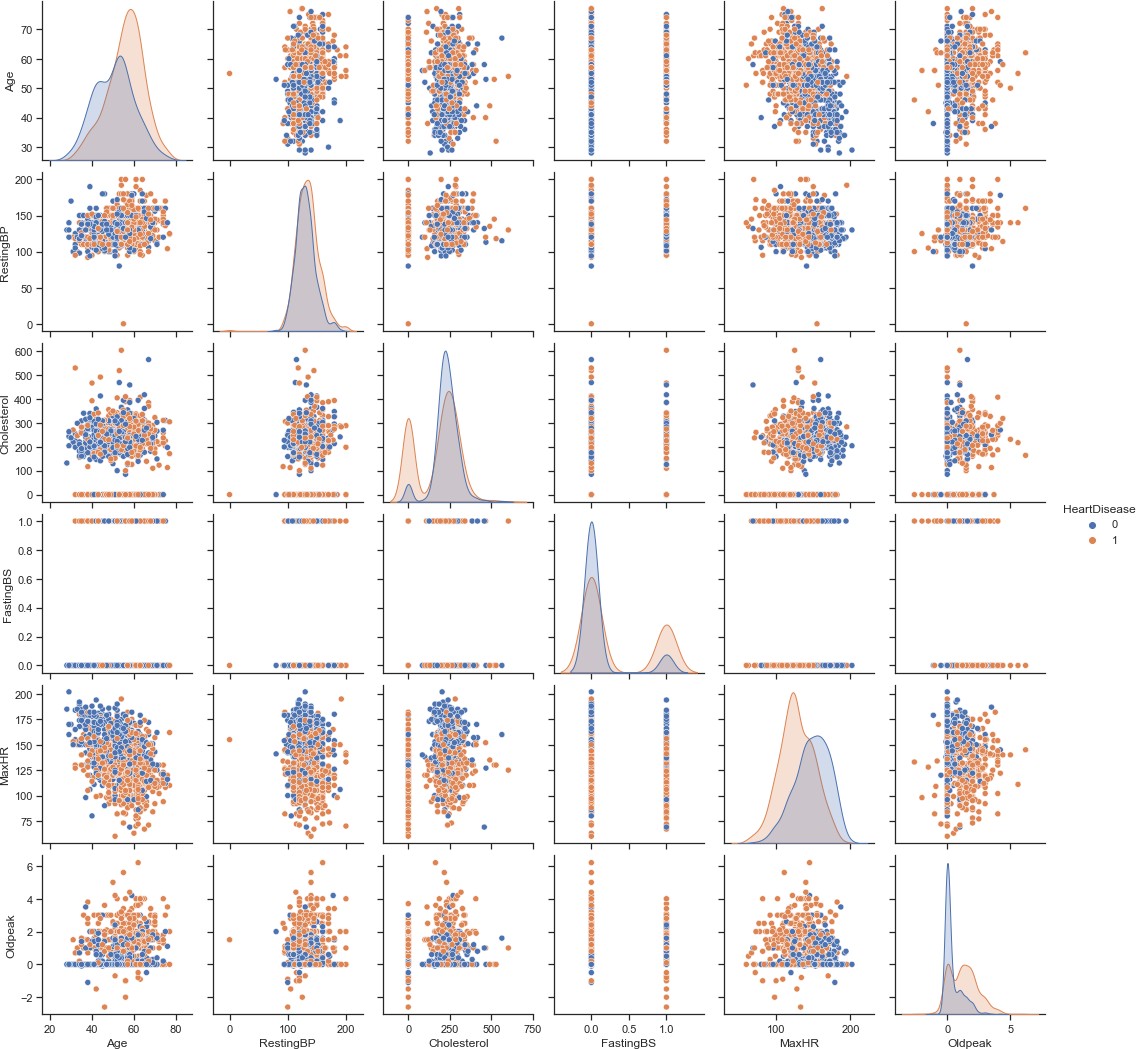
Oldpeak: oldpeak: ST [Числовое значение, измеренное в депрессии]

ST\_Slope: наклон сегмента ST пикового упражнения [Up: восходящий, Flat: плоский, Down: нисходящий]

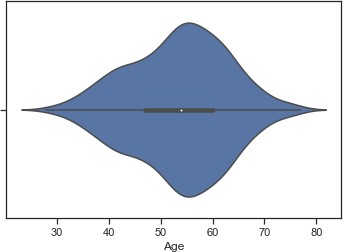
HeartDisease: выходной класс [1: болезнь сердца, 0: нормальный]

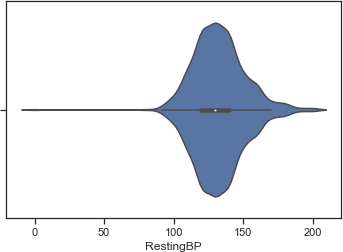
1. Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных.

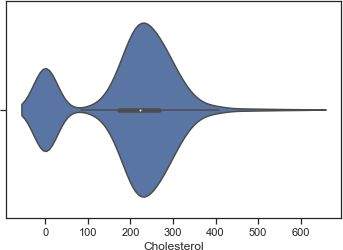
Парные диаграммы:

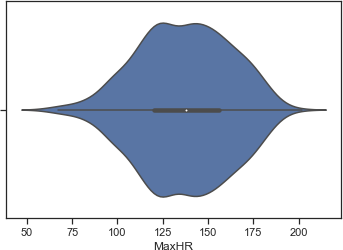


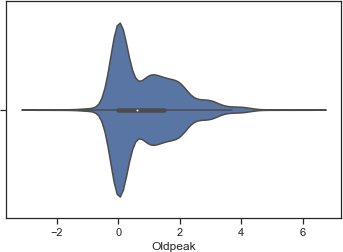
Скрипичные диаграммы для числовых колонок:

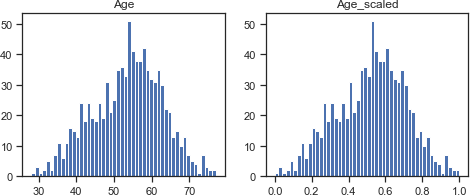


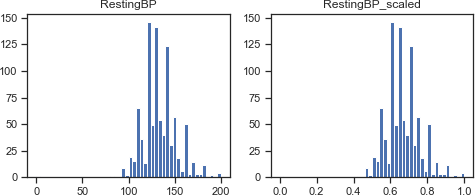


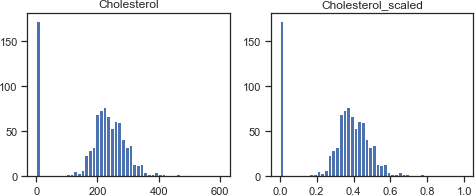


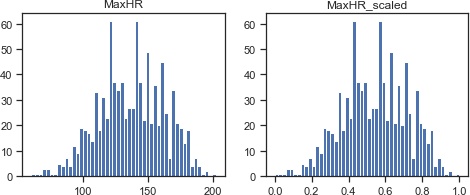


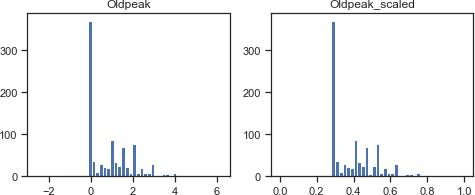


1. Выбор признаков, подходящих для построения моделей. Кодирование категориальных признаков. Масштабирование данных.

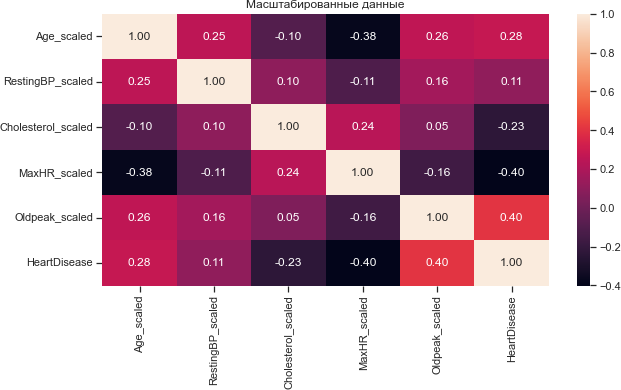
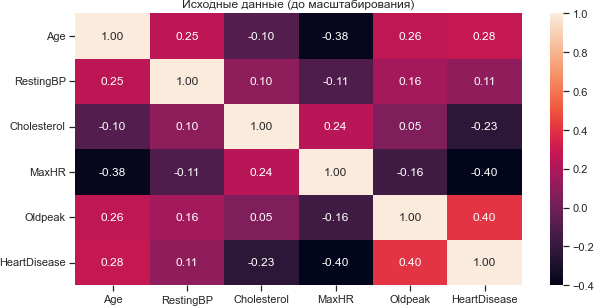








1. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения.



На основе корреляционной матрицы можно сделать следующие выводы:

Корреляционные матрицы для исходных и масштабированных данных совпадают. Целевой признак классификации “HeartDisease” наиболее сильно коррелирует с Oldpeak (0.4) и MaxHR (-0.4). Эти признаки обязательно следует оставить в модели классификации.

1. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии.

Для задачи классификации будем использовать следующие модели:

• Логистическая регрессия

• Метод ближайших соседей

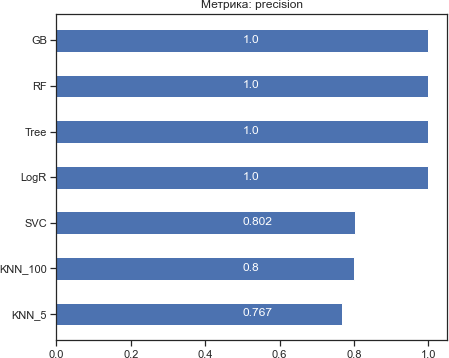
• Машина опорных векторов

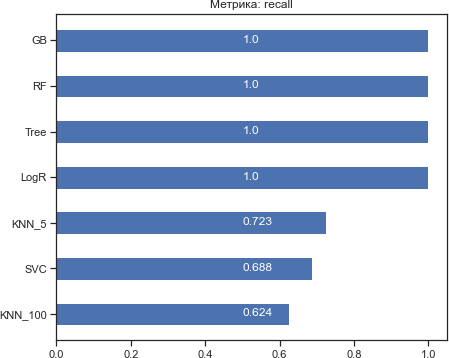
• Решающее дерево

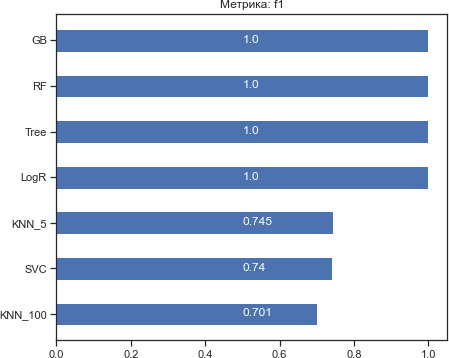
• Случайный лес

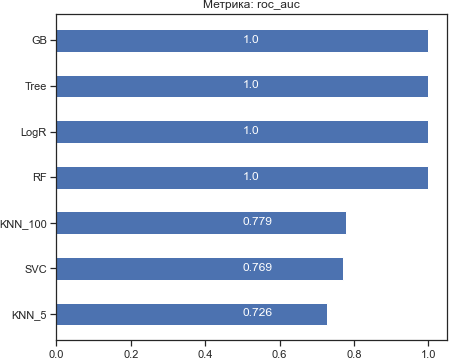
• Градиентный бустинг

1. Формирование выводов о качестве построенных моделей на основе выбранных метрик.









Таким образом, 4 модели – градиентный бустинг, дерево, логистическая регрессия и случайный лес показали одинаково высокий результат.

ЗАКЛЮЧЕНИЕ

Таким образом, было проведено исследование датасета для прогноза сердечной недостаточности. Для задачи классификации использовалось несколько моделей, из которых градиентный бустинг, дерево, логистическая регрессия и случайный лес показали одинаково высокий результат**.**

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Методические указания по программной библиотеке Pandas на языке Python. URL: https://slemeshevsky.github.io/python-course/pandas/pdf/pandas.pdf (дата обращения: 23.12.2024).
2. scikit-learn. URL: https://scikit-learn.org/stable/index.html (дата обращения: 23.12.2024).
3. matplotlib. URL: https://matplotlib.org/stable/api/\_as\_gen/ matplotlib.pyplot.html (дата обращения: 23.12.2024).