|  |  |
| --- | --- |
|  | **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:***

***\_\_\_\_\_\_\_\_\_\_\_Предсказание риска развития\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_ болезни сердца \_\_\_\_\_\_\_\_\_\_\_\_\_***

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

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

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

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

Студент \_\_\_ИУ5-65Б\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_Обухов А.А.\_\_\_\_\_**

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

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Гапанюк Ю.Е.\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_ Гапанюк Ю.Е.\_\_\_\_\_**

(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 \_\_\_\_\_\_\_\_\_\_

(Индекс)

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

(И.О.Фамилия)

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

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_\_\_Предсказание риска развития болезни сердца \_\_\_\_\_\_\_\_\_\_\_

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

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

Студент группы \_\_ИУ5-65Б\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Обухов Антон Александрович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Исследовательская\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_\_\_\_\_\_\_\_\_НИР\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

Техническое задание \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_Исследовать методы машинного обучения для решения задачи классификации\_\_\_\_\_\_

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

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

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

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

Расчетно-пояснительная записка на \_\_26\_\_ листах формата А4.

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

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

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

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

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

Дата выдачи задания « 13 » февраля 2024 г.

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

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Обухов А.А.\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

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

**Содержание**

[**Введение** 4](#_Toc133806658)

[**Постановка задачи** 6](#_Toc133806659)

[**Выполнение работы** 7](#_Toc133806660)

[**Заключение** 25](#_Toc133806662)

[**Список использованной литературы** 26](#_Toc133806663)

# **Введение**

В современном мире заболевания сердца остаются серьезной проблемой. Кардиологи уделяют особое внимание сбору и анализу данных о факторах риска, чтобы вовремя предсказать и предотвратить развитие сердечно-сосудистых заболеваний.

В данном исследовании мы используем данные из отчетов о медицинских осмотрах и истории болезни пациентов, чтобы разработать модель машинного обучения для прогнозирования риска сердечно-сосудистых заболеваний. Мы будем использовать алгоритмы классификации для определения факторов риска, включая возраст, пол, тип грудной боли, артериальное давление, уровень холестерина и другие показатели здоровья.

Наша цель состоит в том, чтобы создать эффективную модель, которая может помочь врачам быстро и точно определить пациентов с высоким риском развития сердечно-сосудистых заболеваний и принять своевременные меры для профилактики и лечения. Такая модель может способствовать улучшению здоровья населения и спасению жизней.

Для достижения поставленной цели были определены следующие этапы:

1. Поиск и выбор набора данных для построения моделей машинного обучения для решения задачи регрессии или классификации.
2. Проведение разведочного анализа данных.
3. Выбор признаков, подходящих для построения моделей.
4. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.
5. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения.
6. Выбор метрик для последующей оценки качества моделей.
7. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии.
8. Формирование обучающей и тестовой выборок на основе исходного набора данных.
9. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров и оценка качества моделей на основе тестовой выборки.
10. Подбор гиперпараметров для выбранных моделей. Построение оптимальных моделей.
11. Формирование выводов о качестве построенных моделей на основе выбранных метрик.

# **Постановка задачи**

Данное исследование направлено на решение задачи классификации, а именно, на прогнозирование риска развития заболеваний сердца.

Имеются данные о состоянии здоровья пациентов, которые включают информацию о таких факторах, как возраст, пол, тип грудной боли, артериальное давление, уровень холестерина и другие показатели здоровья. Каждого пациента можно классифицировать как имеющего высокий или низкий риск развития заболеваний сердца.

Целью задачи является создание модели машинного обучения, которая будет использовать имеющиеся данные для предсказания риска развития заболеваний сердца у пациентов. Для этого мы будем использовать различные алгоритмы классификации, такие как K ближайших соседей, метод опорных векторов, дерево решений, случайный лес и градиентный бустинг. Модель должна обучаться на тренировочных данных и проверяться на тестовых данных для оценки ее точности и эффективности.

Результатом работы должна быть модель, которая сможет предсказывать риск развития заболеваний сердца у пациентов, и помочь врачам оптимизировать профилактические и лечебные мероприятия для улучшения здоровья населения.

# **Выполнение работы**

Для решения задачи классификации был выбран набор данных содержащий информацию о доставках.

В наборе данных присутствуют следующие столбцы:

* age – возраст в годах sex - пол (1 = мужчина; 0 = женщина)
* cp - тип грудной боли (1: типичная стенокардия, 2: атипичная стенокардия, 3: неангинозная боль, 4: бессимптомная)
* trestbps - артериальное давление в покое (в мм рт. ст.)
* chol - холестерин в сыворотке в мг/дл
* fbs - уровень сахара в крови натощак > 120 мг/дл (1 = верно, 0 = неверно)
* restecg - результаты электрокардиографии в покое (0: нормально, 1: наличие аномалии
* ST-T (инверсия зубца T и/или элевация или депрессия ST > 0,05 мВ), 2: указание на возможную или определенную гипертрофию левого желудочка по критериям Эстеса)
* thalach - максимальный достигнутый пульс
* exang - стенокардия, вызванная физической нагрузкой (1 = да; 0 = нет)
* oldpeak - депрессия ST, вызванная физической нагрузкой, по сравнению с состоянием покоя
* num - диагноз болезни сердца (0: сужение диаметра сосудов < 50%, 1: сужение диаметра > 50%)

Данный датасет использован для решения задачи классификации - предсказания риска развития болезней сердца.

Загружаем данные, получаем общую информацию о датасете и делаем предположения о влиянии признаков на целевую переменную. В наборе данных содержится 261 строка и 14 столбцов, из которых 13 типа int64 и 1 типа flaot.

Пропусков не было обнаружено.

Строим график pairplot для визуализации распределения данных попарно для множества колонок.

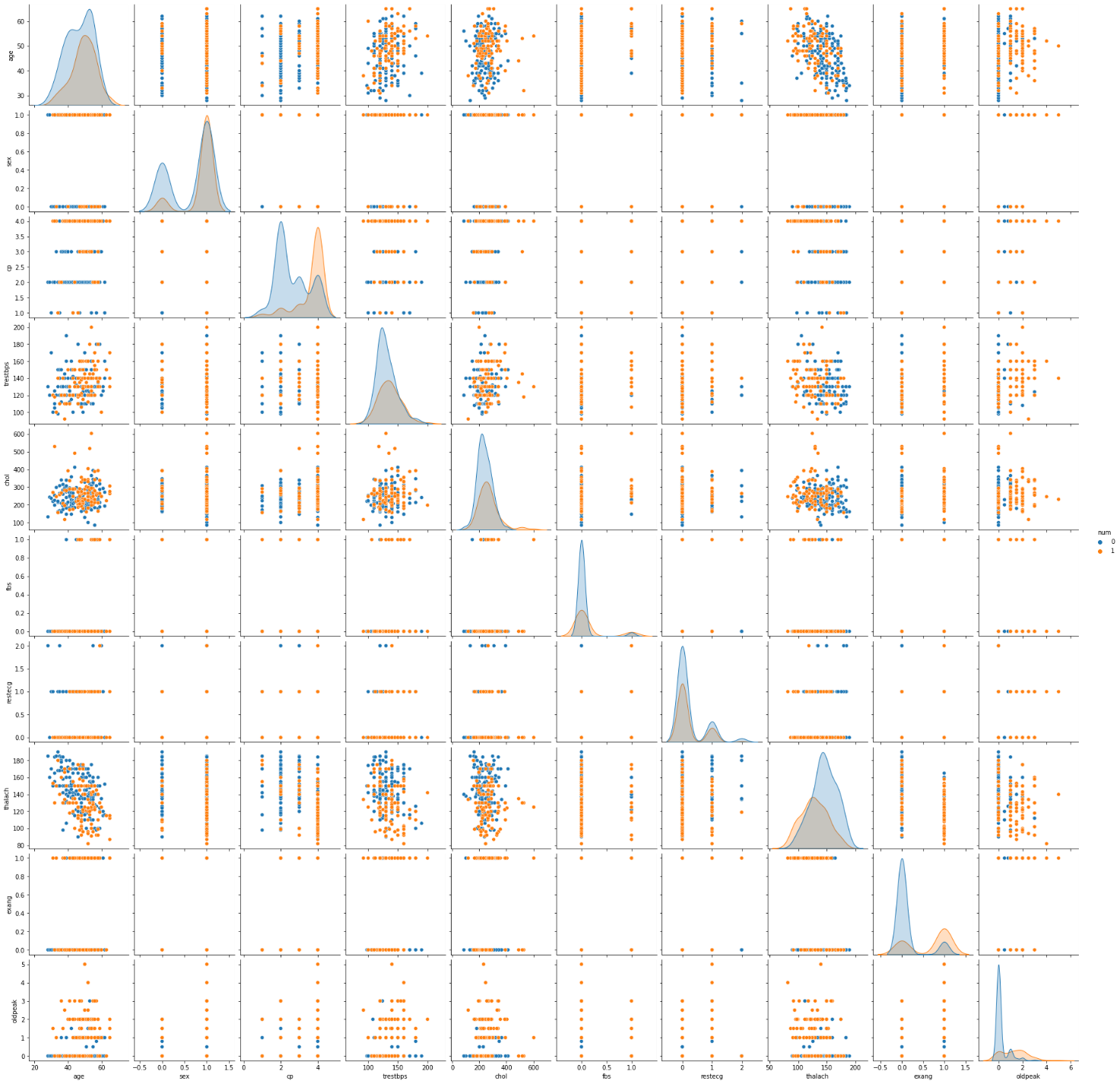


Рисунок 1 - Визуализация распределения данных попарно для множества колонок

Проверяем сбалансированы ли классы в нашем наборе данных. Получаем следующее:

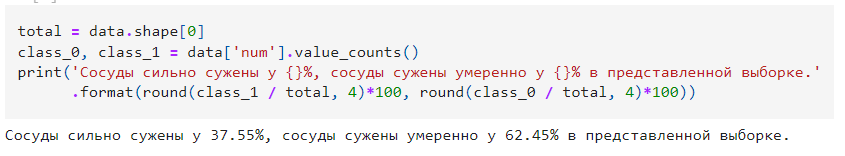


Рисунок 2 – Анализ балансировки

Видим, что классы немножко не сбалансированы.

Строим скрипичные диаграммы для числовых колонок

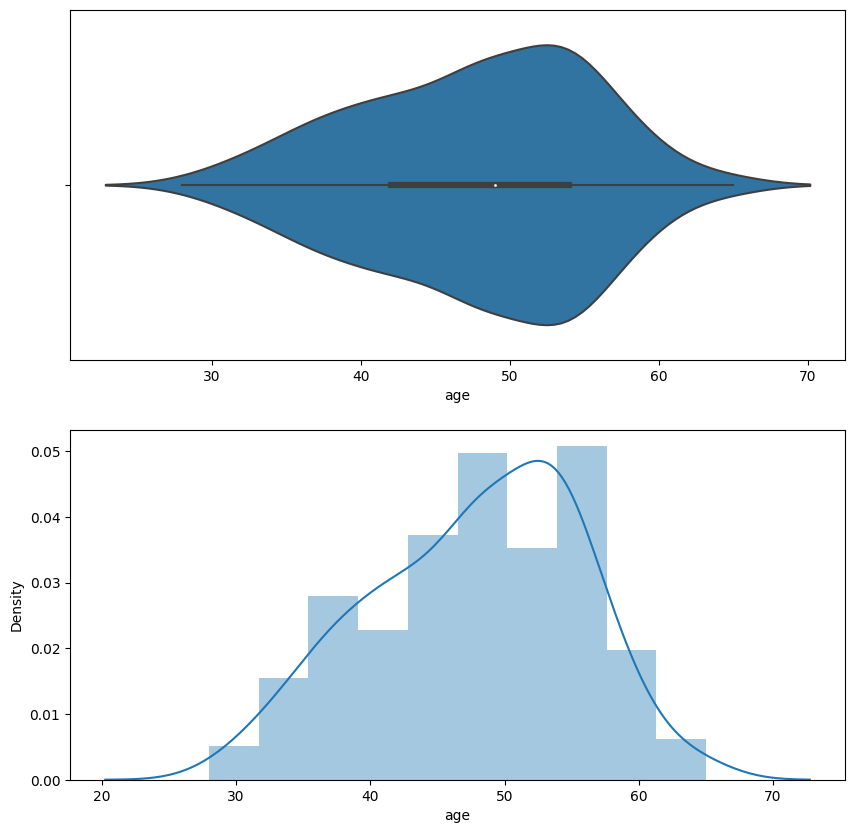


Рисунок 3 - Гистограмма возростов

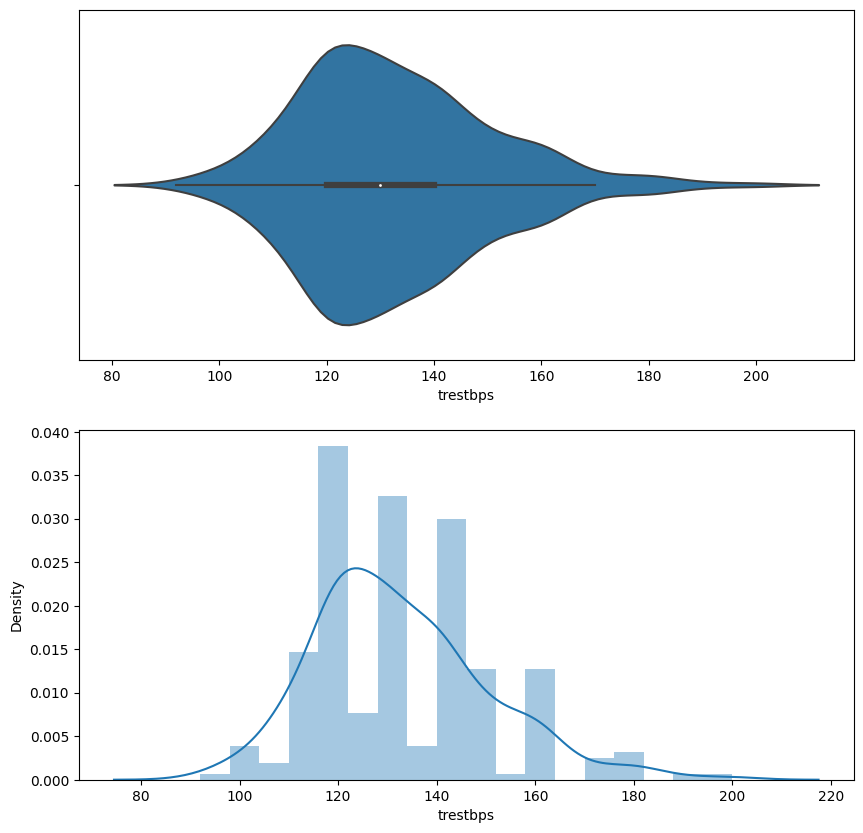


Рисунок 4 - Гистограмма артериального давления в покое

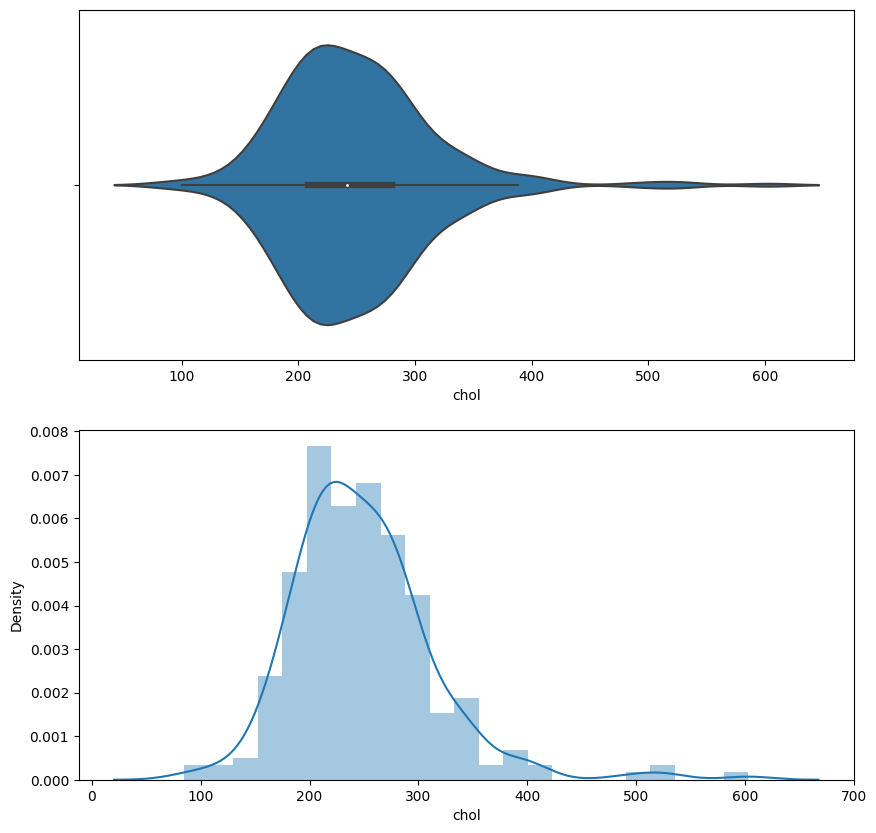


Рисунок 5 - Гистограмма холестерина

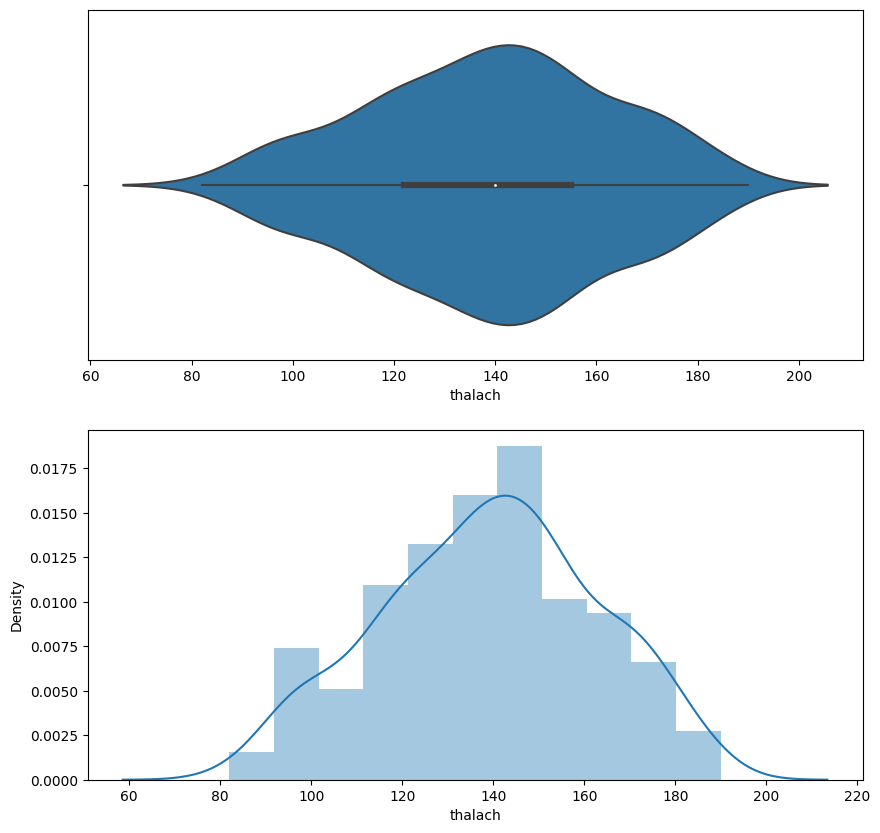


Рисунок 6 - Гистограмма максимального пульса

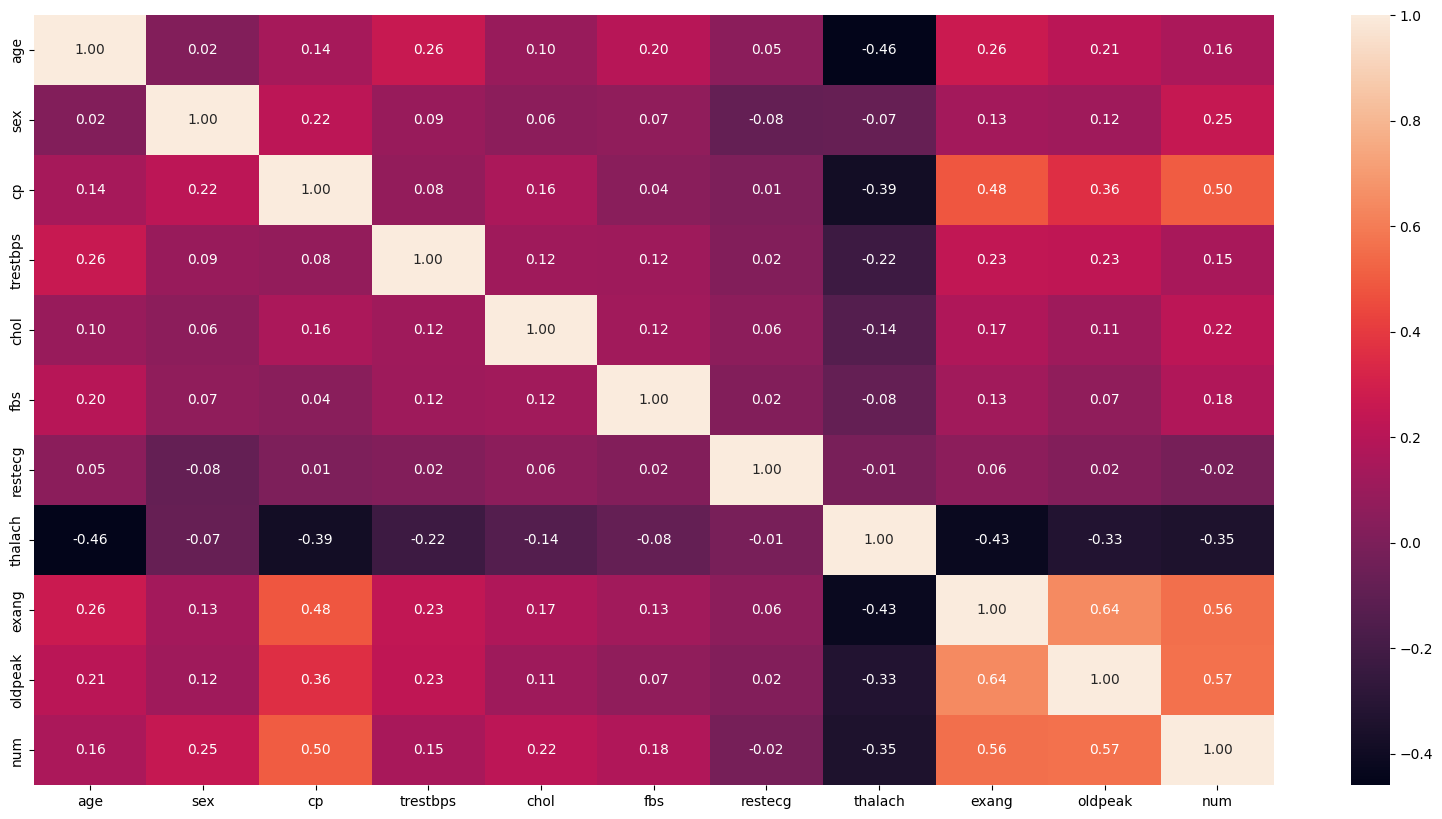


Рисунок 7 - Тепловая карта корреляций

Выводы:

* Целевой признак классификации "num" наиболее сильно коррелирует с грудной болью cp (0,5); со стенокардией (0,56); с депрессией ST-сегмента (0.57). Эти признаки обязательно следует оставить в модели классификации.
* У нас нет признаков которые имеют корреляцию, близкую по модулю к 1, поэтому никакие признаки не следуют исключать из модели по этому признаку.
* Однако у нас существует признак, который имеет корреляцию, близкую по модулю к 0 по отношению к целевому параметру, а именно: restecg (-0.02) - результаты электрогардиографии в покое. Также этот признак имеет корреляцию, близкую по модулю к 0, с остальными параметрами.
* Большие по модулю значения коэффициентов корреляции свидетельствуют о значимой корреляции между исходными признаками и целевым признаком. На основании корреляционной матрицы можно сделать вывод о том, что данные позволяют построить модель машинного обучения.

Выберем метрики для оценки качества модели:

#### Метрика precision:

* Можно переводить как точность, но такой перевод совпадает с переводом метрики "accuracy".
* 𝑝𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛=𝑇𝑃/(𝑇𝑃+𝐹𝑃)
* Доля верно предсказанных классификатором положительных объектов, из всех объектов, которые классификатор верно или неверно определил как положительные.
* Используется функция [precision\_score.](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html#sklearn.metrics.precision_score)
* Метрика recall (полнота):
* 𝑟𝑒𝑐𝑎𝑙𝑙=𝑇𝑃/(𝑇𝑃+𝐹𝑁)
* Доля верно предсказанных классификатором положительных объектов, из всех действительно положительных объектов.
* Используется функция [recall\_score.](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html#sklearn.metrics.recall_score)
* Метрика 𝐹1-мера
* Для того, чтобы объединить precision и recall в единую метрику используется 𝐹𝛽-мера, которая вычисляется как среднее гармоническое от precision и recall:
* 𝐹𝛽=(1+𝛽^2)⋅(𝑝𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛⋅𝑟𝑒𝑐𝑎𝑙𝑙)/(𝑝𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛+𝑟𝑒𝑐𝑎𝑙𝑙)
* где 𝛽 определяет вес точности в метрике.
* На практике чаще всего используют вариант F1-меры (которую часто называют F-мерой) при 𝛽=1:
* 𝐹1=2⋅(𝑝𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛⋅𝑟𝑒𝑐𝑎𝑙𝑙)/(𝑝𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛+𝑟𝑒𝑐𝑎𝑙𝑙)
* Для вычисления используется функция [f1\_score.](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score)
* Метрика ROC AUC
* Основана на вычислении следующих характеристик:
* 𝑇𝑃𝑅=𝑇𝑃/(𝑇𝑃+𝐹𝑁) - True Positive Rate, откладывается по оси ординат. Совпадает с recall.
* 𝐹𝑃𝑅=𝐹𝑃/(𝐹𝑃+𝑇𝑁) - False Positive Rate, откладывается по оси абсцисс. Показывает какую долю из объектов отрицательного класса алгоритм предсказал неверно.
* Идеальная ROC-кривая проходит через точки (0,0)-(0,1)-(1,1), то есть через верхний левый угол графика.
* Чем сильнее отклоняется кривая от верхнего левого угла графика, тем хуже качество классификации.
* В качестве количественной метрики используется площадь под кривой - ROC AUC (Area Under the Receiver Operating Characteristic Curve). Чем ниже проходит кривая тем меньше ее площадь и тем хуже качество классификатора.
* Для получения ROC AUC используется функция [roc\_auc\_score.](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html#sklearn.metrics.roc_auc_score)

Выберем модели для решения задачи классификации:

* Логистическая регрессия
* Метод ближайших соседей
* Решающее дерево
* Случайный лес
* Градиентный бустинг

Формируем обучающую и тестовую выборку в соотношении 8:2. Оставляем колонки все колонки кроме «num» т.к. они влияют на целевой признак.

Строим базовое решения, выводим значениями метрик и ROC-кривую.

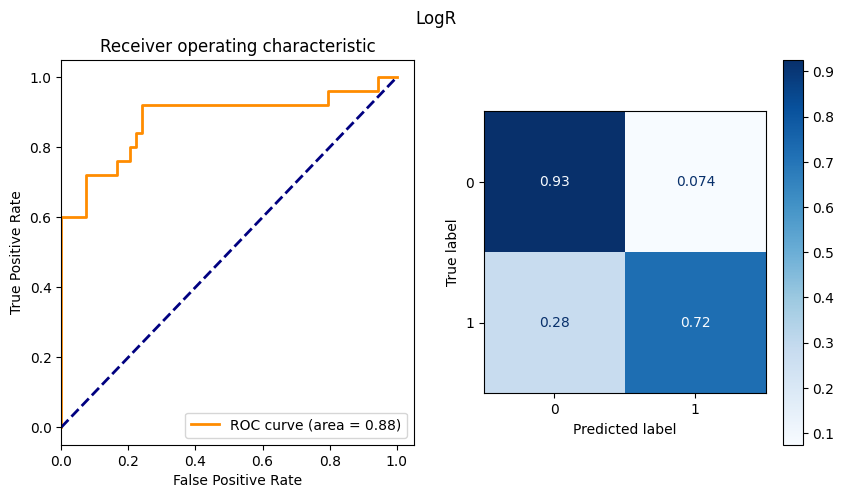


Рисунок 8 - ROC-кривая базовой модели LogR

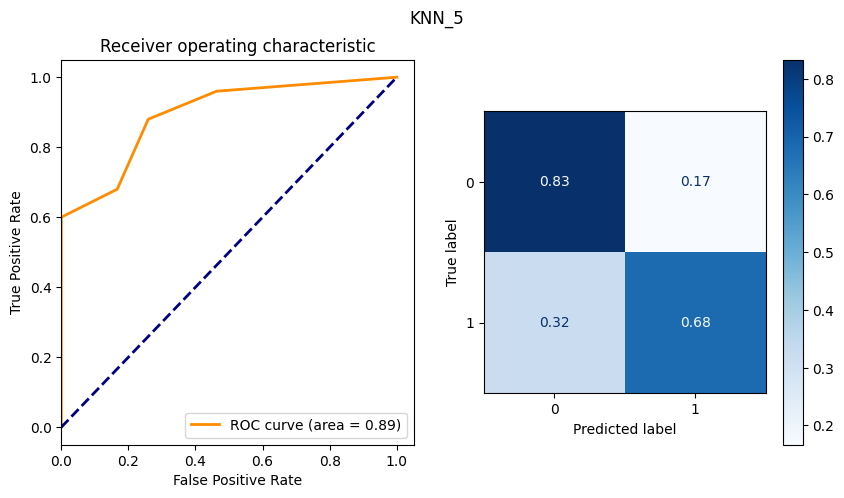
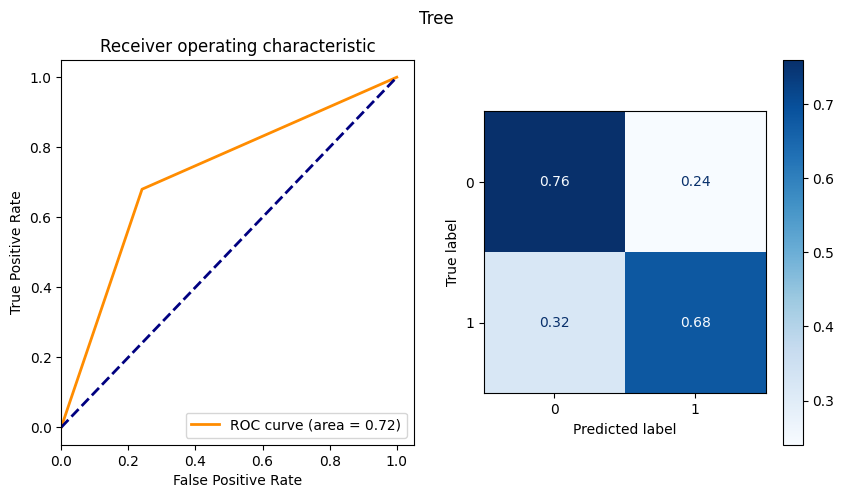


Рисунок 9- ROC-кривая базовой модели KNN



*Рисунок 10 - ROC-кривая базовой модели Decision Tree*

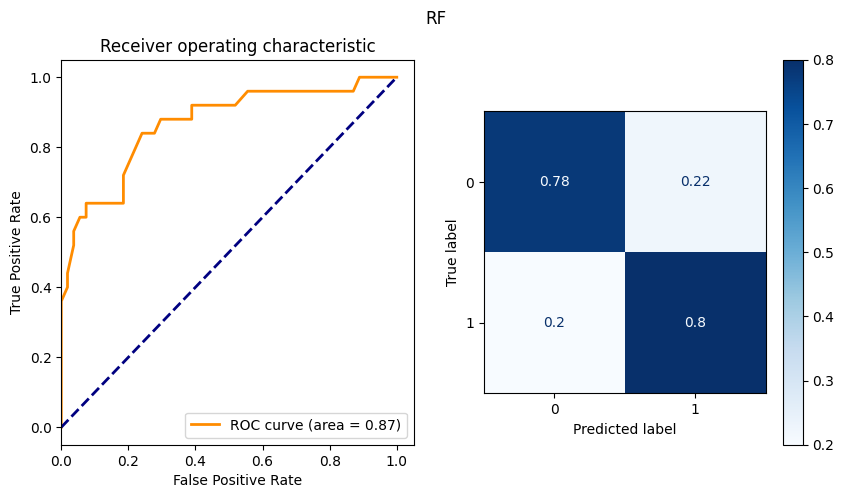


Рисунок 11 - ROC-кривая базовой модели Random Forest

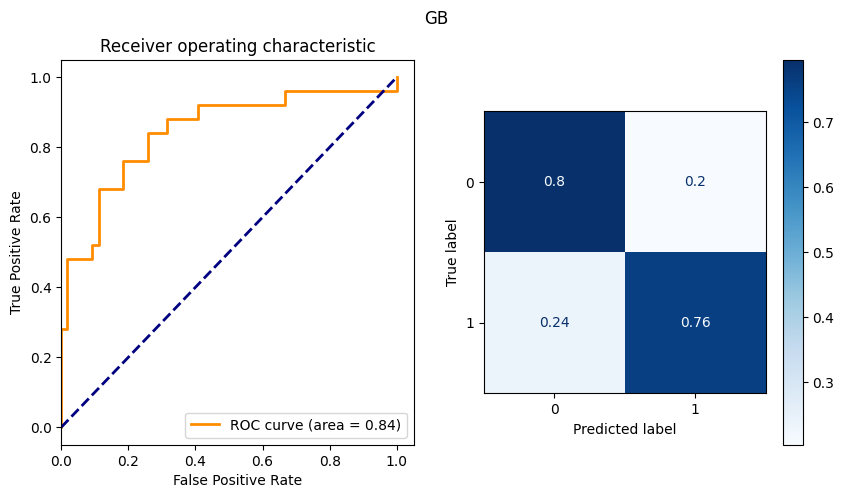


Рисунок 12 - ROC-кривая базовой модели Gradient Boosting

Используем GridSearch для поиска оптимальных гиперпараметров для каждой модели.

KNeighboursClassifier:

Best hyperparameters: { 'n\_neighbors': 29,}

LogisticRegression:

Best hyperparameters: {'C': 0.25, 'random\_state': 0}

DecisionTreeClassifier:

Best hyperparameters: {'criterion': 'entropy', 'random\_state': 0}

RandomForestClassifier:

Best hyperparameters: {'criterion': 'entropy', 'n\_estimators': 119, 'random\_state': 0}

GradientBoostingClassifier:

Best hyperparameters: {'criterion': 'friedman\_mse', 'learning\_rate': 0.5, 'loss': 'exponential', 'n\_estimators': 11, 'random\_state': 0}

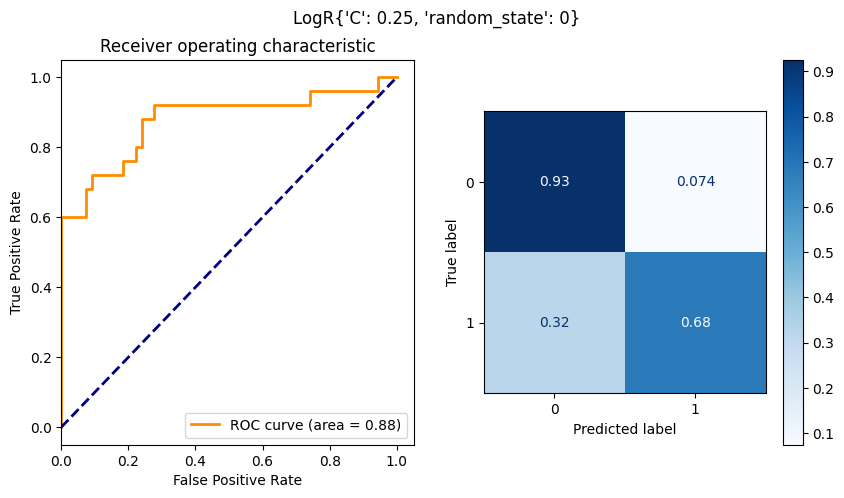


Рисунок 13 - ROC-кривая модели LogR после поиска гиперпараметров

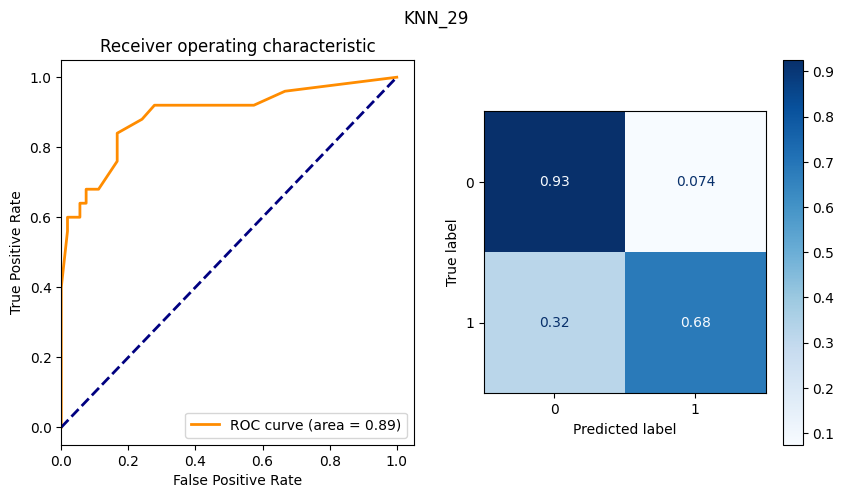


Рисунок 14 - ROC-кривая модели KNN после поиска гиперпараметров

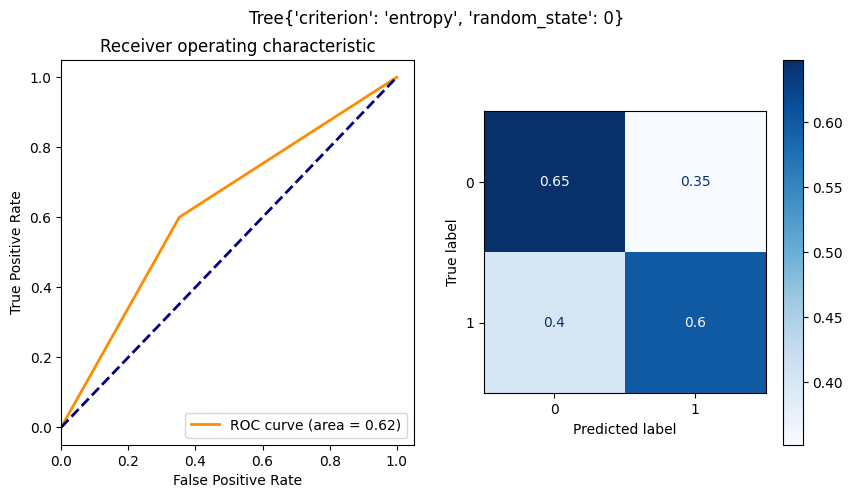


Рисунок 15 - ROC-кривая модели Decision Tree после поиска гиперпараметров

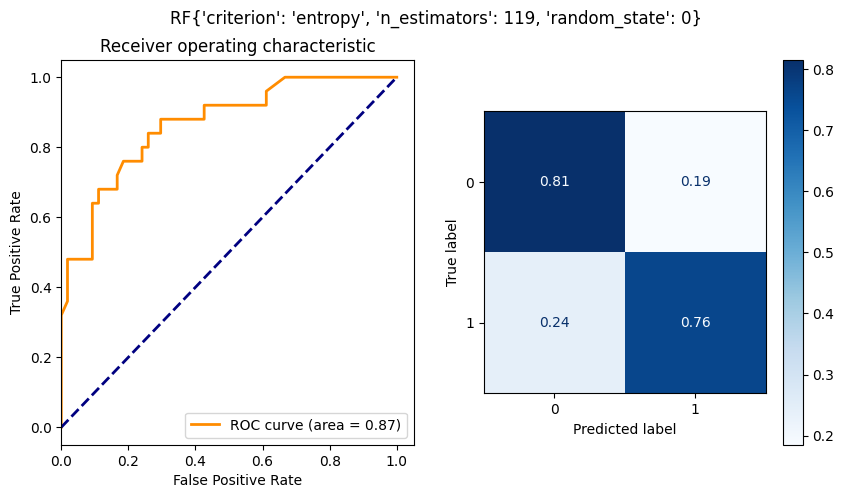


Рисунок 16 - ROC-кривая модели Random Forest после поиска гиперпараметров

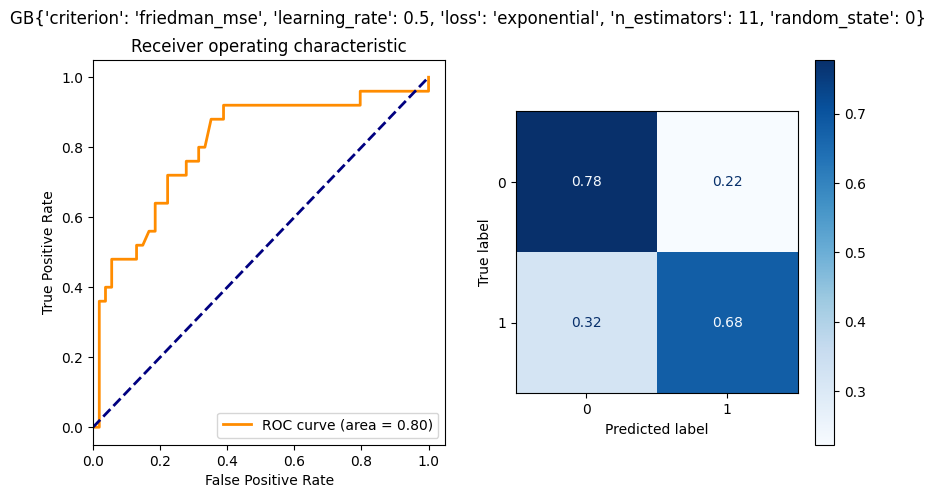
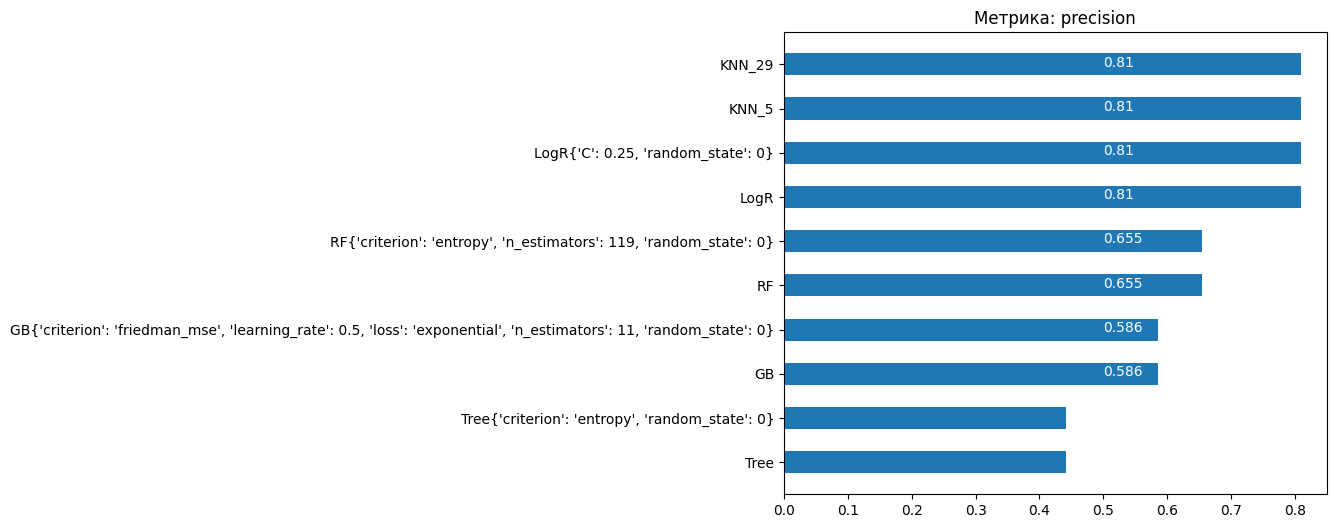
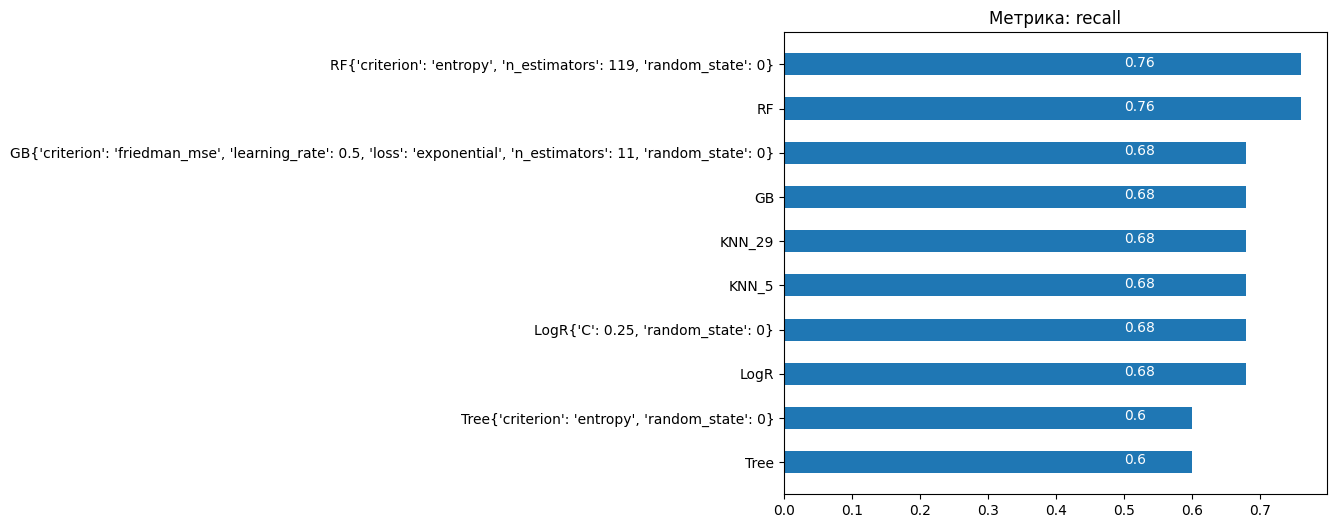
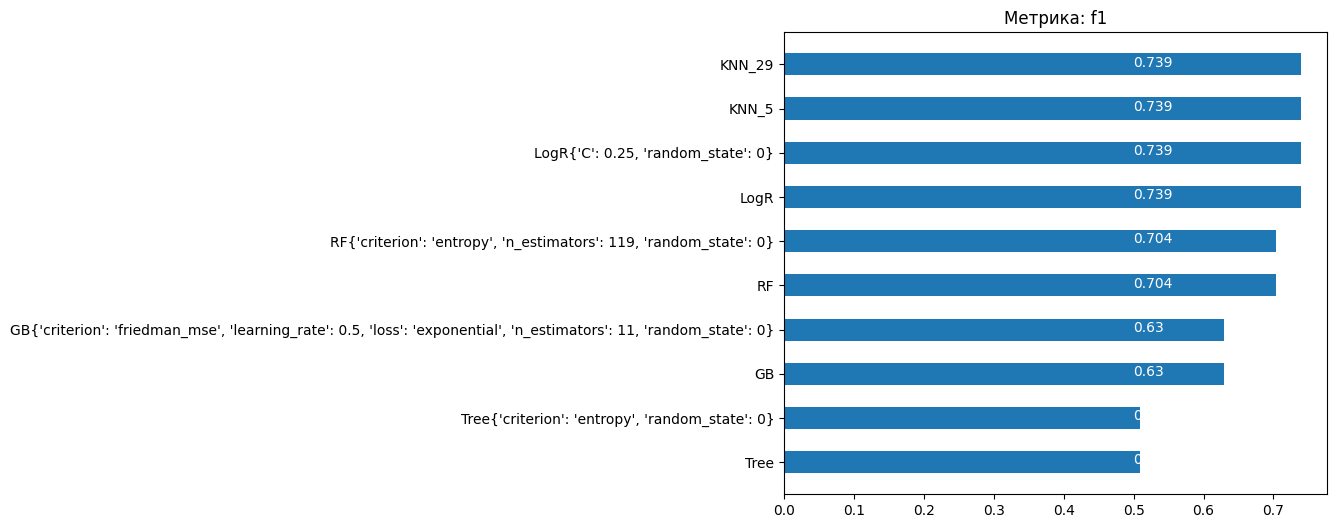


Рисунок 17 - ROC-кривая модели Gradient Boosting после поиска гиперпараметров







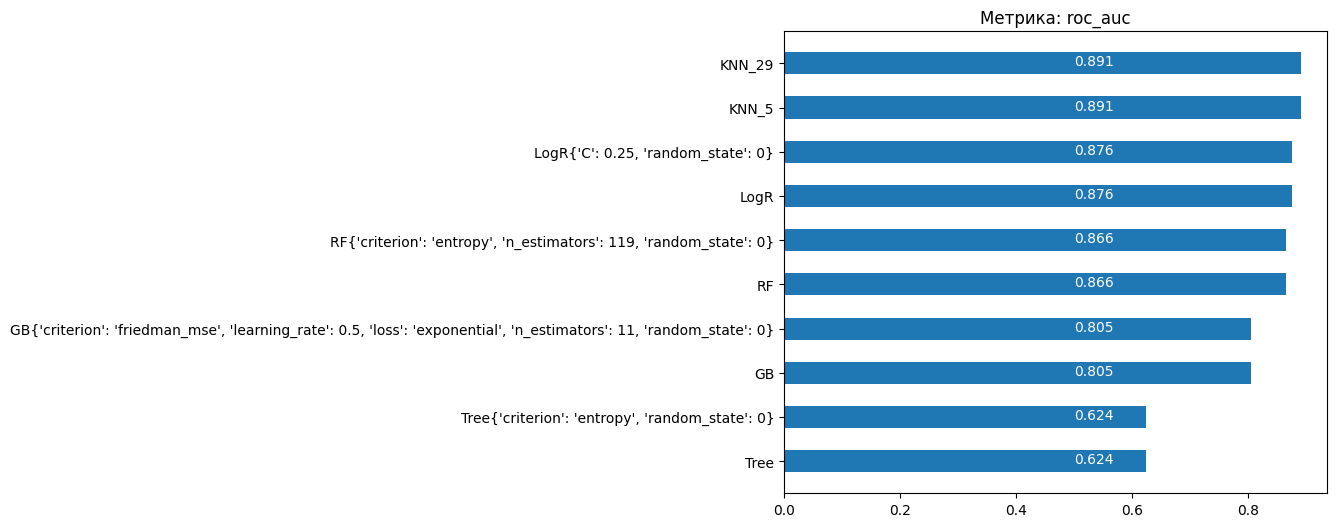


рисунок 27 - Сравнение базовых моделей с моделями после подбора гиперпараметров по 4 метрикам

На основании трех метрик из четырех лучшей для решения данной задачи классификации оказалась модель ближайших соседий после подбора параметров.

# **Заключение**

Классификация параметра, отвечающего за показатель наличия/отсутствия риска развития заболеваний сердца, с помощью методов машинного обучения является актуальной и перспективной задачей в области здравоохранения. Анализ и обработка данных с помощью алгоритмов машинного обучения могут помочь врачам своевременно предсказать пациентов с высоким риском развития сердечно-сосудистых заболеваний и принять меры для профилактики и лечения.

Данные были проанализированы, визуализированы и подготовлены к обучению. Были применены различные алгоритмы, такие как метод ближайших соседей, метод опорных векторов, дерево решений, случайный лес и градиентный бустинг.

В результате исследования было показано, что большинство использованных методов могут достичь хороших результатов, но самыми точными на основании трех метрик из четырех оказались модели градиентного бустинга и метод случайного леса.

Таким образом, разработанная модель машинного обучения может стать ценным инструментом для врачей в борьбе с заболеваниями сердца, помогая им выявлять пациентов с высоким риском и принимать своевременные меры для улучшения здоровья населения.

В результате исследования было показано, что большинство использованных методов могут достичь хороших результатов, но самой точной на основании трех метрик из четырех оказалась модель ближайших соседей.

# **Список использованной литературы**

1. T-test на Python для проверки и получения t-статистики // Помощник Python URL: https://pythonpip.ru/osnovy/t-test-na-python
2. Machine Learning Metrics in simple terms // Medium URL: https://medium.com/analytics-vidhya/machine-learning-metrics-in-simple-terms-d58a9c85f9f6
3. Опорный пример для выполнения проекта по анализу данных. // Jupyter nbviewer URL: https://nbviewer.org/github/ugapanyuk/courses\_current/blob/main/notebooks/ml\_project\_example/project\_classification\_regression.ipynb
4. Репозиторий курса "Технологии машинного обучения", бакалавриат, 6 семестр. // GitHub URL: https://github.com/ugapanyuk/courses\_current/wiki/COURSE\_TMO\_SPRING\_2024/