|  |  |
| --- | --- |
|  | **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 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***\_\_\_\_\_\_\_Предсказание риска возникновения\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_сердечной недостаточности\_\_\_\_\_\_\_\_\_\_\_\_\_***

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

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

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

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

Студент \_\_\_ИУ5-65Б\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_Акулова 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)

*2023 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 возникновения сердечной недостаточности\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

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

Студент группы \_\_ИУ5-65Б\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Акулова Анастасия Алексеевна\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Исследоваткльская\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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 \_\_32\_\_ листах формата А4.

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

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

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

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

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

Дата выдачи задания « 13 » февраля 2023 г.

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

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_Акулова А.А.\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

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

**Содержание**

[**Введение** 4](#_Toc136809437)

[**Постановка задачи** 6](#_Toc136809438)

[**Выполнение работы** 7](#_Toc136809439)

[**Создание веб-приложения** 24](#_Toc136809440)

[**Заключение** 26](#_Toc136809441)

[**Список использованной литературы** 27](#_Toc136809442)

# **Введение**

Сердечно-сосудистые заболевания (ССЗ) являются причиной смерти номер 1 в мире, унося, по оценкам, 17,9 миллиона жизней каждый год, что составляет 31% всех смертей в мире. Четыре из 5 смертей от ССЗ происходят из-за сердечных приступов и инсультов, и одна треть этих смертей происходит преждевременно у людей в возрасте до 70 лет. Сердечная недостаточность является распространенным событием, вызванным ССЗ, 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 котором модель машинного обучения может оказать большую помощь.

В данной работе мы будем использовать данные, полученные из медицинских исследований, чтобы построить модель машинного обучения, которая сможет предсказывать возникновение сердечной недостаточности у пациентов. Мы будем использовать алгоритмы классификации для определения факторов риска.

Целью данной работы является разработка эффективной модели, которая может помочь врачам быстро и точно определить вероятность возникновения сердечной недостаточности у пациентов и принять меры для предотвращения данного заболевания.

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.
12. Создание веб-приложение для демонстрации хотя бы одной модели машинного обучения.

# **Постановка задачи**

Данная работа по машинному обучению направлена на решение задачи классификации, а именно, предсказание риска возникновения сердечной недостаточности у пациентов.

Имеются данные о пациентах, которые включают информацию о таких факторах, как возраст, пол, тип рисунка груди, стенокардия, атипичная стенокардия, неангинальная боль, кровяное давление в состоянии покоя, уровень холестерина, уровень сахара в крови натощак, эектрокардиограмма в состоянии покоя, ГЛЖ, достигнутая максимальная частота сердечных сокращений, стенокардия, вызванная физической нагрузкой, наличие депрессии, наклон сегмента ST при максимальной нагрузке. Каждый пациент может быть классифицирован как имеющий риск развития сердечной недостаточности или не имеющий риска.

Целью задачи является создание модели машинного обучения, которая будет использовать имеющиеся данные для предсказания риска возникновения сердечной недостаточности у пациента. Для этого мы будем использовать различные алгоритмы классификации, такие как K ближайших соседей, метод опорных векторов, случайный лес, метод логистической регрессии 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 возникновение сердечной недостаточности у пациентов с высокой точностью и помочь врачам принимать меры для предотвращения данного заболевания.

# **Выполнение работы**

Для решения задачи классификации был выбран набор данных содержащий информацию о пациентах, которые могут быть подвержены сердечной недостаточности.

В наборе данных присутствуют следующие столбцы:

* Возраст: возраст пациента [годы]
* Пол: пол пациента [M: мужчина, F: Женщина]
* Тип рисунка груди: тип боли в груди [TA: типичная стенокардия, ATA: атипичная стенокардия, NAP: неангинальная боль, ASY: бессимптомная]
* Давление в состоянии покоя: кровяное давление в состоянии покоя [мм рт. ст.]
* Уровень холестерина: сывороточный холестерин [мм / дл]
* Уровень сахара в крови натощак [1: если уровень сахара в крови натощак > 120 мг / дл, 0: в противном случае]
* Электрокардиограмма в состоянии покоя: результаты электрокардиограммы в состоянии покоя [Нормальный: нормальный, ST: аномалия зубца ST-T (инверсия зубца T и / или подъем или депрессия ST > 0,05 мВ), ГЛЖ: вероятная или определенная гипертрофия левого желудочка по критериям Эстеса]
* MaxHR: достигнутая максимальная частота сердечных сокращений [Числовое значение от 60 до 202]
* Стенокардия, вызванная физической нагрузкой [Y: Да, N: Нет]
* Oldpeak: oldpeak = ST [Числовое значение, измеряемое при депрессии]
* ST\_Slope: наклон сегмента ST при максимальной нагрузке [Up: подъем, Flat: равнина, Down: спад]
* HeartDisease: выходной класс [1: болезнь сердца, 0: норма]

Данный датасет использован для решения задачи классификации - предсказания наличия или отсутствия сердечной недостаточности у пациентов.

Загружаем данные, получаем общую информацию о датасете и делаем предположения о влиянии признаков на целевую переменную. В наборе данных содержится 918 строк и 12 столбцов, из которых 6 типа int64, 5 типа object и 1 типа float64.

Не обнаруживаем пропуски в колонках.

Не обнаруживаем дубликаты в калонках.

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

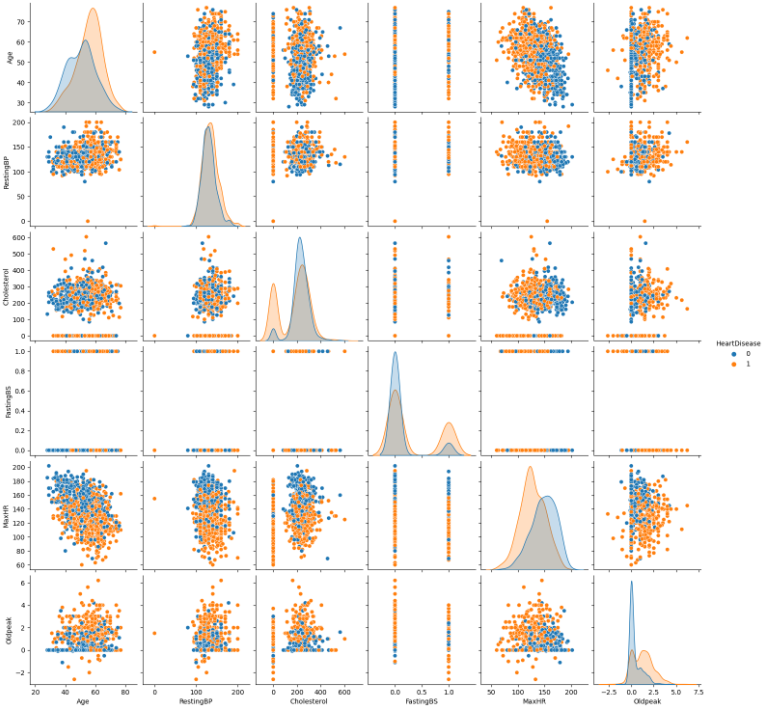


Рисунок 1 - Визуализация распределения данных попарно для множества колонок

Проверяем сбалансированы ли классы в нашем наборе данных. Получаем следующую гистограмму:

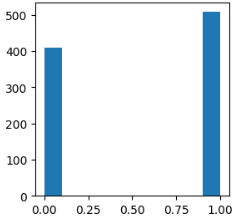
****

Рисунок 2 - Гистограмма классов

Видим, что классы сбалансированы.

Строим таблицу средних значений с группировкой по целевому признаку и делаем следующие предположения:

* для сердечной недостаточности характерно наличие высокого кровяного давления в состоянии покоя и сахара в крови, также низкого халестерина
* среди людей с сердечной недостаточностью чаще всего встречаются люди более взрослые
* среди людей с зафиксированное сердечной недостаточностью люди могут страдать депрессией
* у больных сердечной недостаточностью сердце частота сердечных сокращений реже

Построим диаграмму зависимости наличия депресии у пациентов с сердечной недостаточностью.

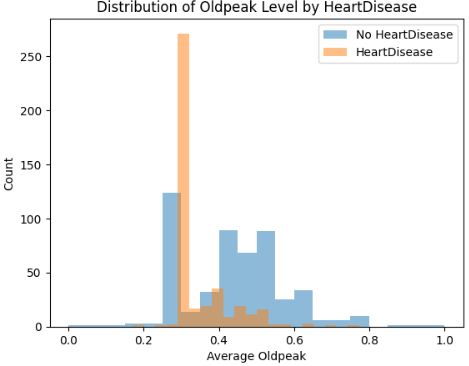


Рисунок 3 - Гистограмма зависимости депресии и вероятности сердечной недостаточности

Люди с сердечной недостаточностью могут впасть в сильную депрессию

Построим диаграмму зависимости наличия холестерина и сердечной недостаточности.

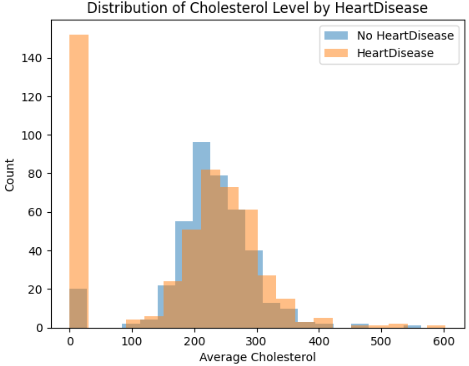


Рисунок 4 - Гистограмма зависимости наличия холестерина и вероятности сердечной недостаточности

Подтверждаем предположение, что низкий холестерин влияет на заболевание сердечной недостаточностью

Смотрим на соотношение наличия высокого кровяного давления и наличия сердечной недостаточностью.

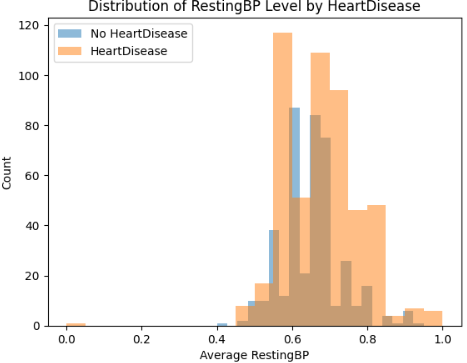


Рисунок 5 - Гистограмма зависимости наличия кровяного давления и вероятности сердечного заболевания

Кровяное давление при сердечной недостаточности чуть выше

Проверяем количество сахара в крови на риск сердечной недостаточности.

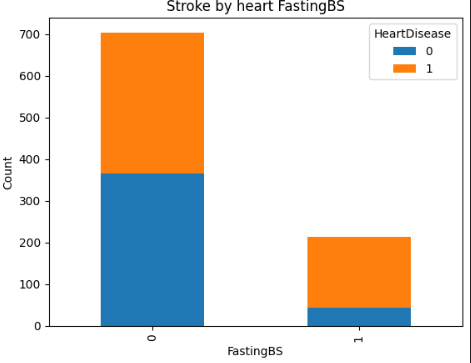


Рисунок 6 - Гистограмма зависимости наличия сахара в крови и вероятности сердечной недостаточности

У людей с сердечной недостаточностью сильно больше уровень сахара в крови.

Проверяем пол человека на риск сердечной недостаточности.

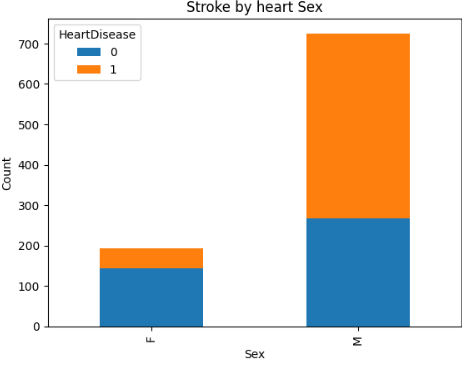


Рисунок 7 - Гистограмма зависимости пола и вероятности сердечной недостаточности

Заметим, что среди женщин довольно мало тех, у кого присутствует сердечная недостаточность.

Далее приведем данные к нужному формату. Сначала масштабируем численные признаки методом MinMaxScaler. Посмотрим на распределения колонок до и после масштабирования.

Таблица 1 - Распределение численных колонок до и после масштабирования

|  |  |
| --- | --- |
| До масштабирования | После масштабирования |
| Рисунок 8 - Распределение возраста до масштабирования | Рисунок 9 - Распределение возраста после масштабирования |
| Рисунок 10 - Распределение кровяного давления в состоянии покоя до масштабирования | Рисунок 11 - Распределение кровяного давления в состоянии покоя после масштабирования |
| Рисунок 12 - Распределение холестерина до масштабирования | Рисунок 13 - Распределение холестерина после масштабирования |
| *Рисунок 14 - Распределение сахара в крови на тощак до масштабирования* | *Рисунок 15 - Распределение сахара в крови на тощак после масштабирования* |
| *Рисунок 16 - Распределение пульса до масштабирования* | *Рисунок 17 - Распределение пульса после масштабирования* |
| *Рисунок 18 - Распределение депрессии до масштабирования* | *Рисунок 19 - Распределение депрессии до масштабирования* |

Распределение не изменилось.

Проводим корреляционный анализ данных. Строим тепловую карту корреляций.

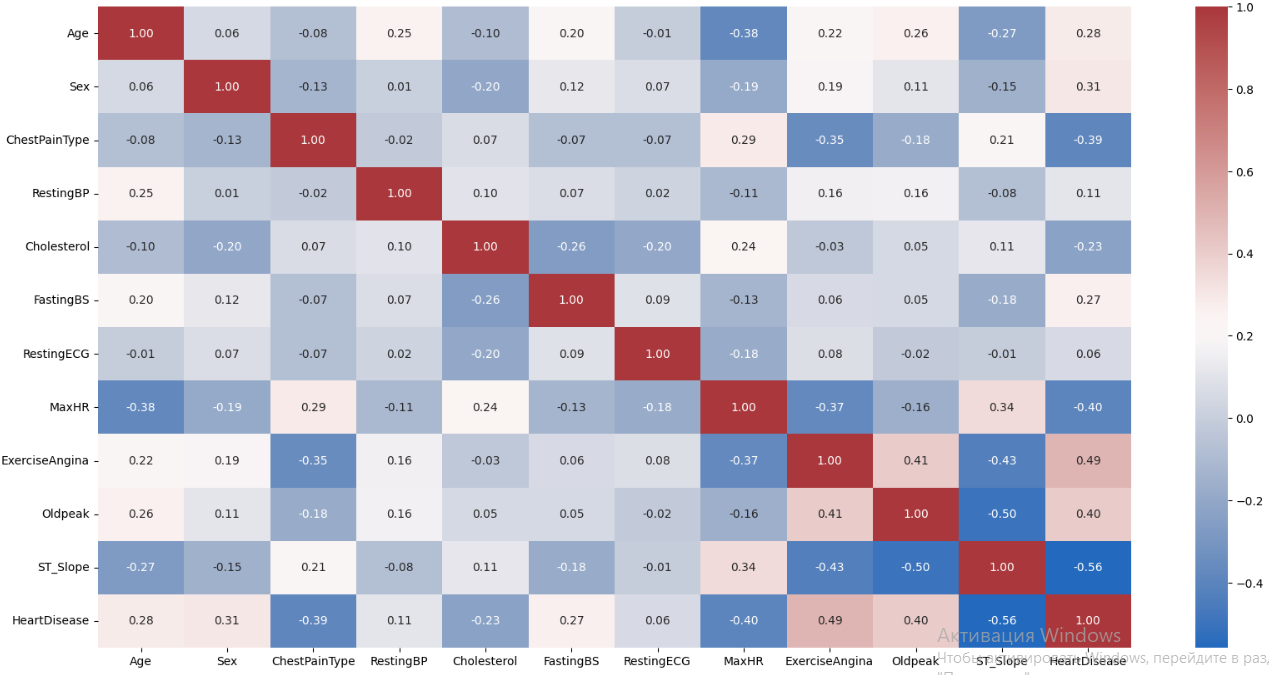


Рисунок 20 - Тепловая карта корреляций

Выводы:

* целевой признак HeartDisease больше всего коррелирует с наличием наклона сегента ST (0.56), стенохардии (0.49), депрессии (0.40).
* Нецелевых признаков, имеющих между собой особо сильную корреляцию, в наборе данных не наблюдается.
* столбец с электрокардиограммой имеет слишком низкую кореляцию с целевым признаком, удалим его
* столбец с кровяным давлением в состоянии покоя оставим, так как ранее мы выявили, что зависимости между ним и сердечой недостаточностью может существовать

Выберем метрики для оценки качества модели:

* - определяется как среднее количество отзывов, полученных по каждому классу. Наилучшее значение равно 1, а наихудшее значение равно 0
* - среднее гармоническое precision и recall. Другими словами, это средневзвешенное значение точности и отзыва.
* - oснована на вычислении следующих характеристик: - True Positive Rate, откладывается по оси ординат. Совпадает с recall. - 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.

Выберем модели для решения задачи классификации:

* KNN;
* Логистическая регрессия;
* Опорных векторов;
* Случайный лес;
* Градиентный бустинг.

Формируем обучающую и тестовую выборку. Удаляем колонку ‘RestingECG, т.к. она почти не влияют на целевой признак. Кодируем категориальные признаки с помощью LabelEncoder

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

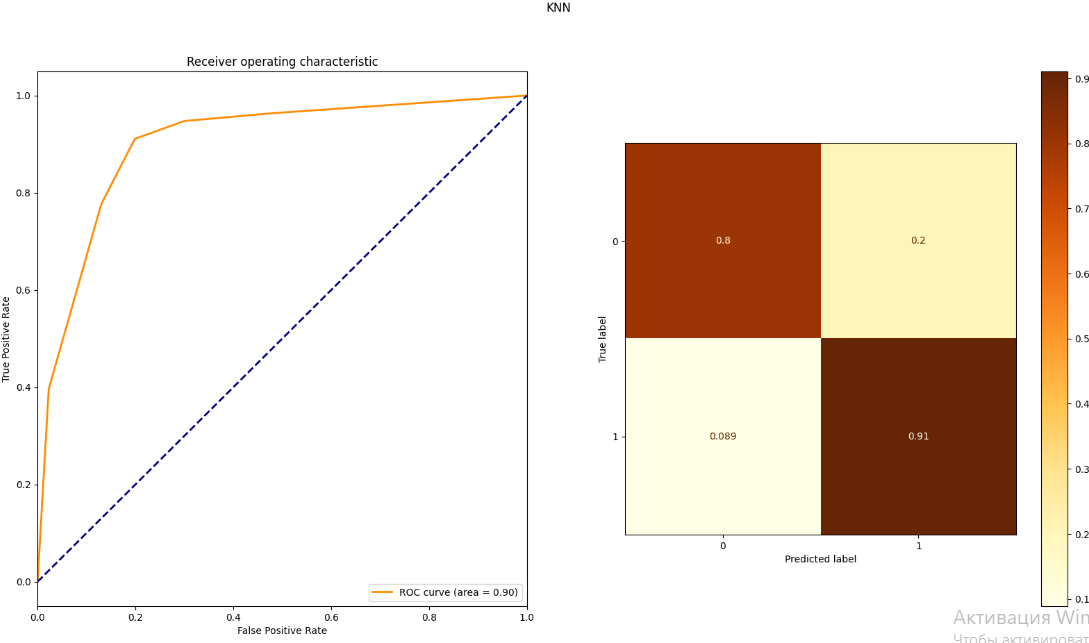


Рисунок 21 - ROC-кривая и матрица ошибок базовой модели KNN

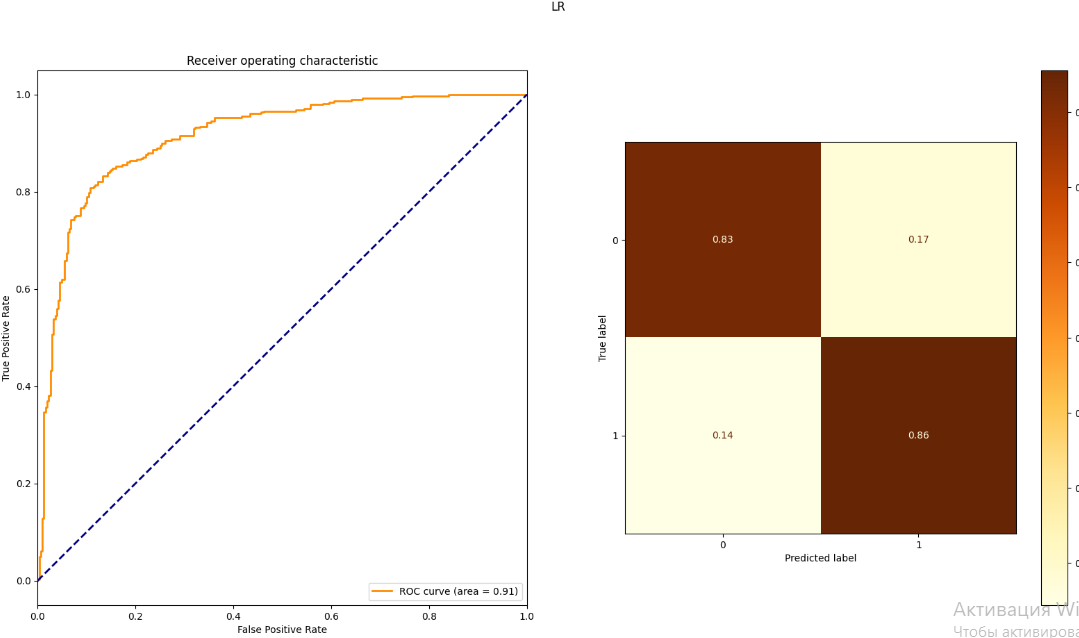


Рисунок 22- ROC-кривая и матрица ошибок базовой модели логистической регрессии

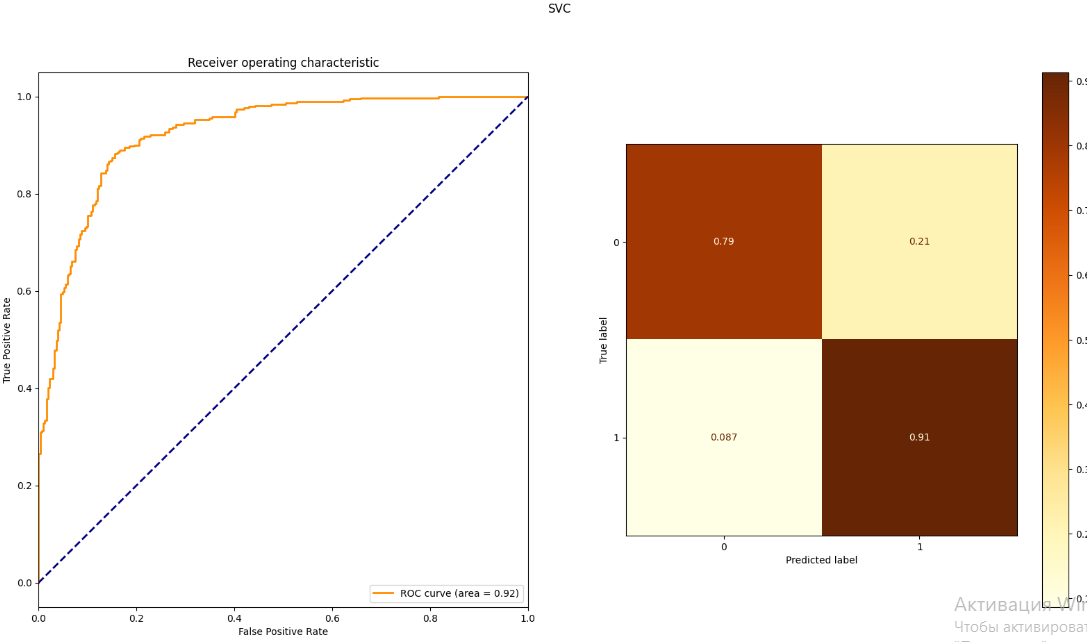


Рисунок 23 - ROC-кривая и матрица ошибок базовой модели опорных векторов

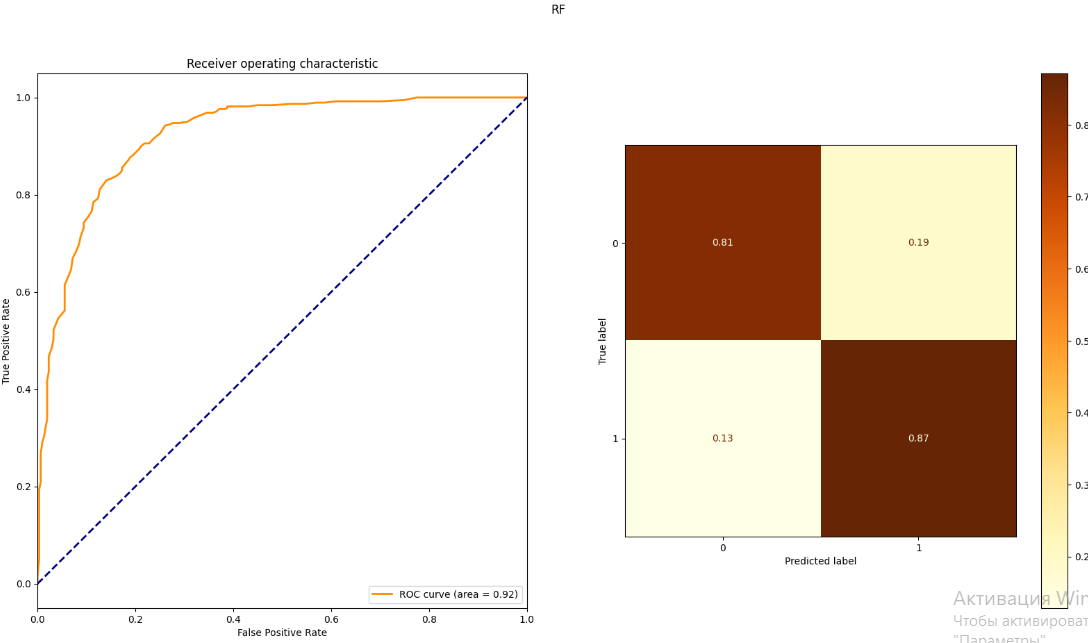


Рисунок 24 - ROC-кривая и матрица ошибок базовой модели случайный лес

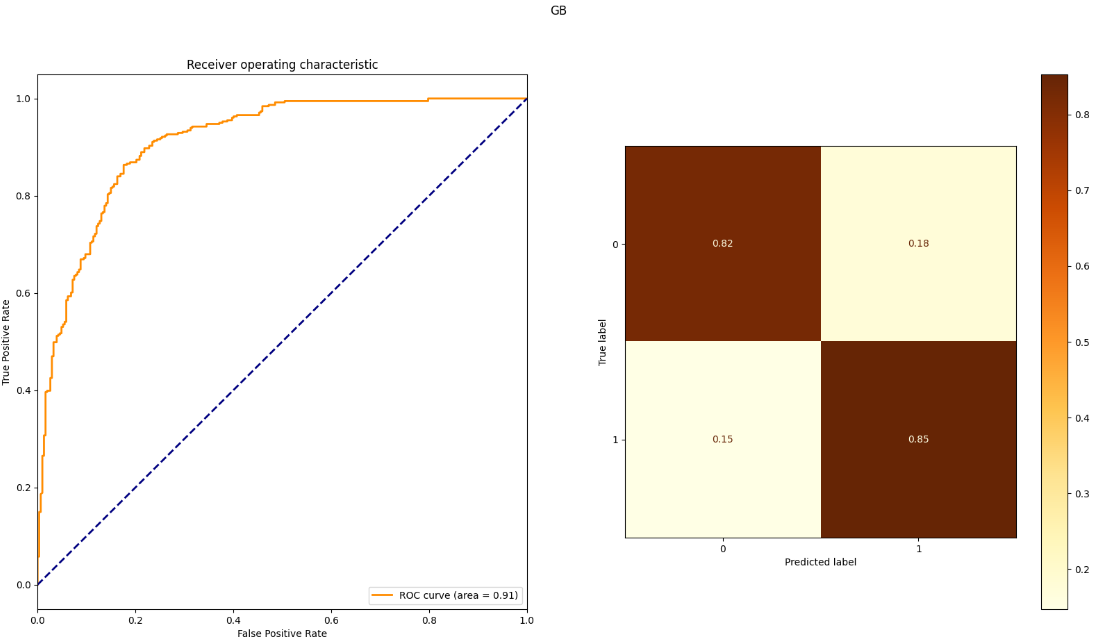


Рисунок 25 - ROC-кривая и матрица ошибок базовой модели градиентный бустинг

Подбор гиперпараметров для выбранных моделей с использованием RandomizedSearchCV.

KNN: 'n\_neighbors': 15, 'weights': distance, 'P': 2

Логистическая регрессия: 'C': 0.831763771102671, 'solver': lbfgs, 'class\_weight': {0: 580, 1: 420}

Опорные вектора: 'C': 0.0199526231496888, 'kernel': rbf, 'class\_weight': {0: 366, 1: 634}

Случайный лес: 'n\_estimators': 501, 'max\_features': sqrt, 'criterion': entropy, 'class\_weight': {0: 815, 1: 185}

Градиентный бустинг: 'loss': exponential, 'learning\_rate': 0.013803842646028845, 'n\_estimators': 25, 'criterion': friedman\_mse

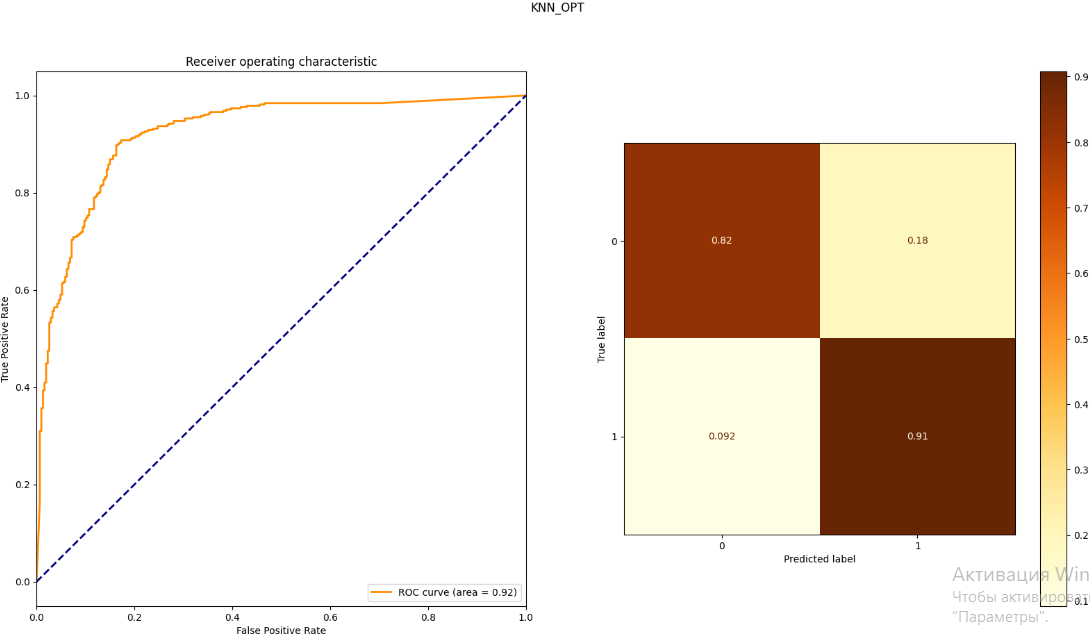


Рисунок 26 - ROC-кривая и матрица ошибок модели KNN после поиска гиперпараметров

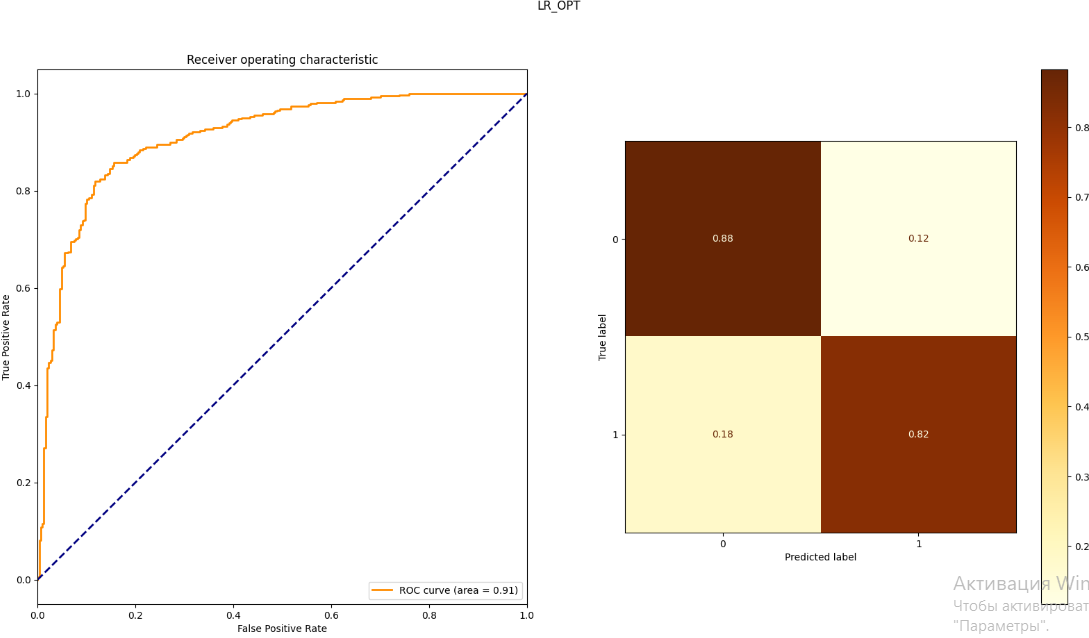


Рисунок 27 - ROC-кривая и матрица ошибок базовой модели логистической регрессии после поиска гиперпараметров

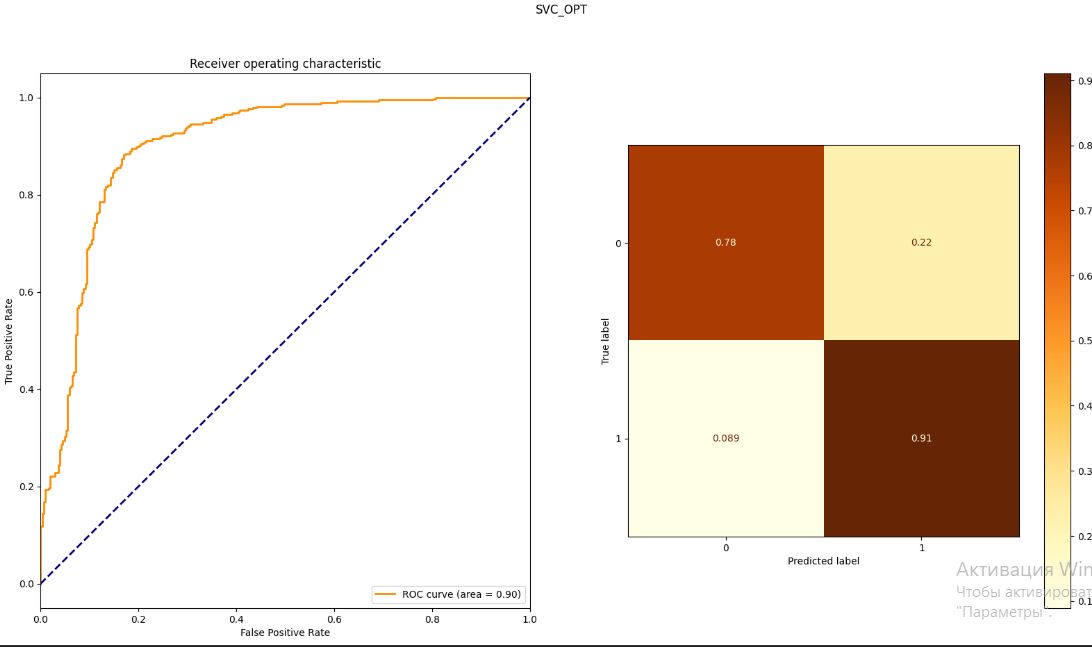


Рисунок 28 - ROC-кривая и матрица ошибок базовой модели опорных векторов после поиска гиперпараметров

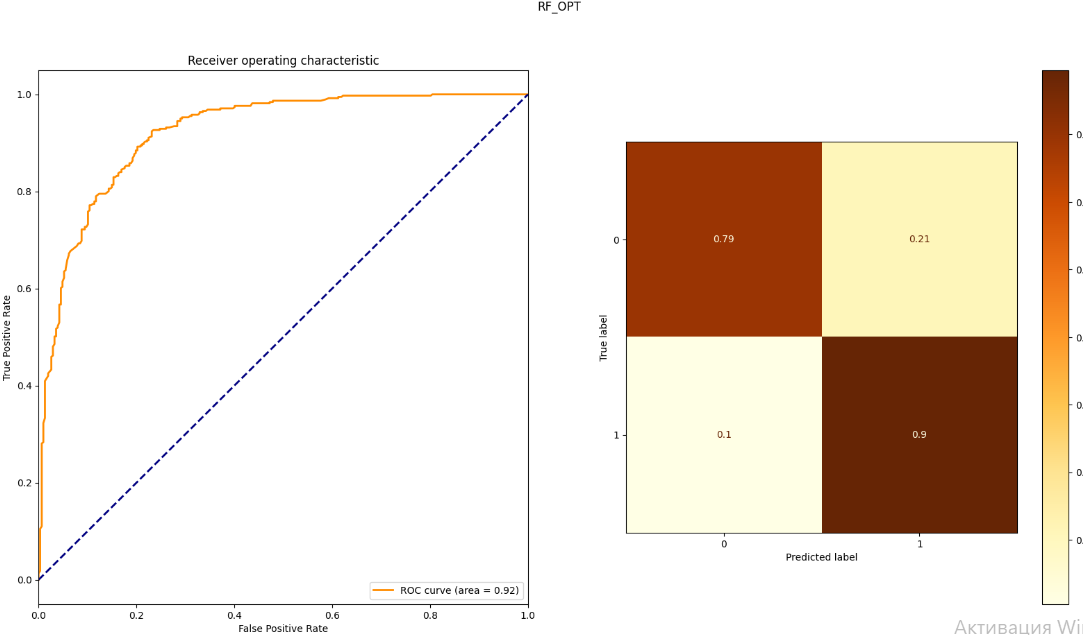


Рисунок 29 - ROC-кривая и матрица ошибок базовой модели случайный лес после поиска гиперпараметров

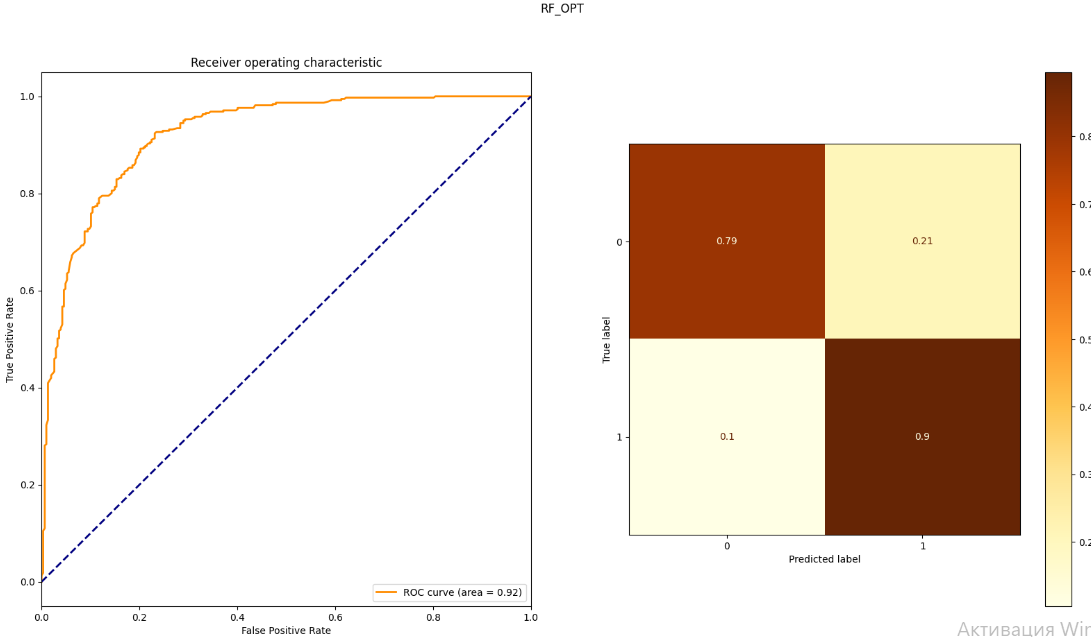
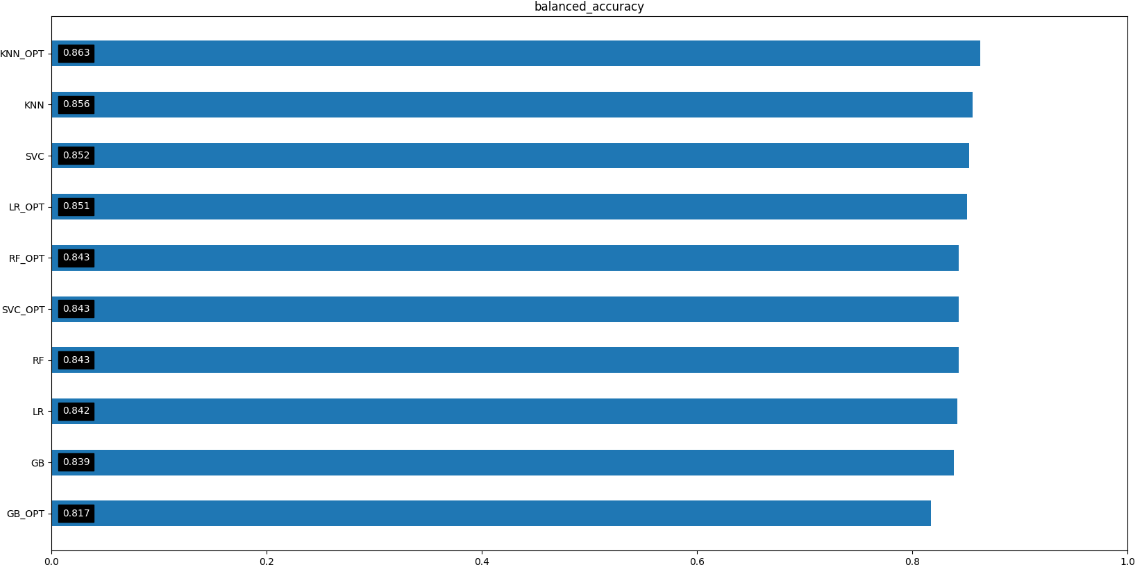
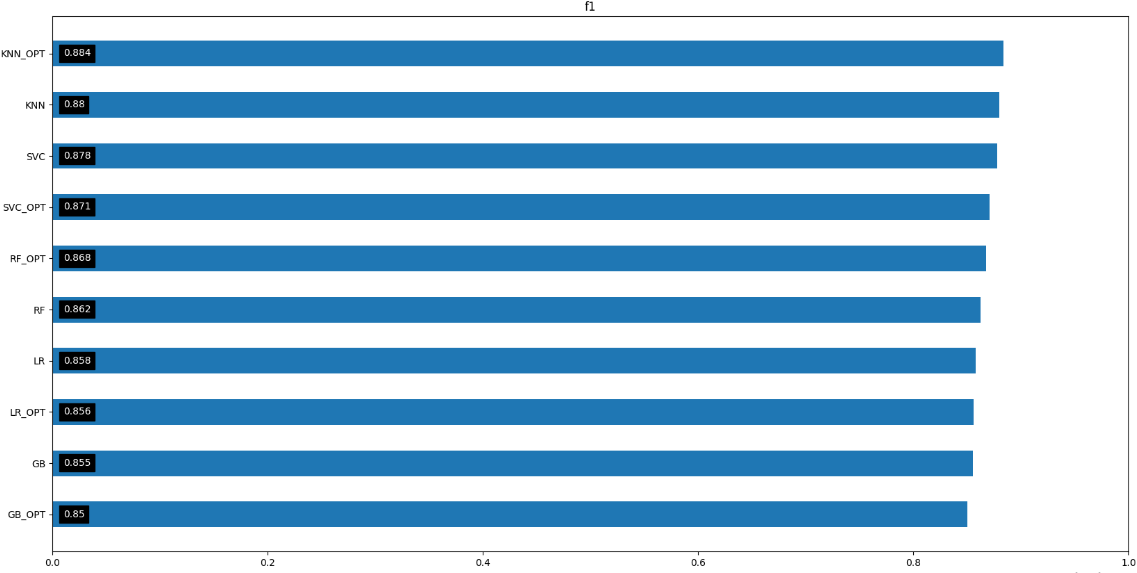


Рисунок 30 ROC-кривая и матрица ошибок базовой модели градиентный

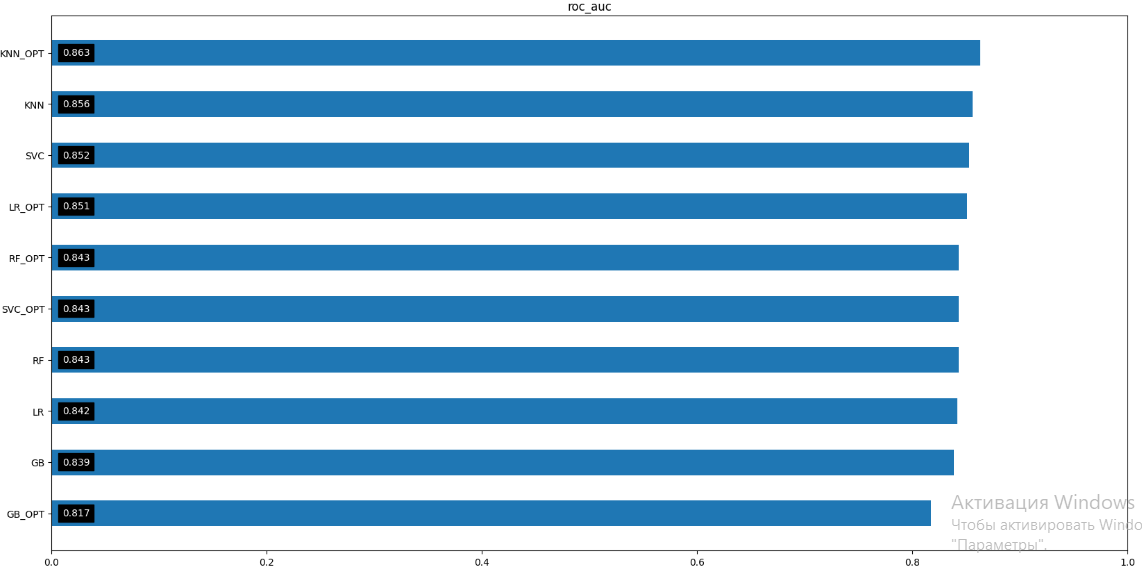
Формирование выводов о качестве построенных моделей на основе выбранных метрик.



*Рисунок 31 Метрика balanced accuracy*



*Рисунок 32 Метрика F1-мера*



*Рисунок 33 Метрика ROC AUC*

Выводы:

* Оптимизация метода ближайших соседей привела к несколько лучшим результатам по всем метрикам
* Оптимизация логистической регрессии привела к несколько лучшим результатам по большинству метрикам
* Оптимизация метода опорных векторов привела к намного худшим результатам по всем метрикам
* Оптимизация случайного леса привела к несколько лучшим результатам по всем метрикам
* Оптимизация градиентного бустинга привела к несколько худшим результатам по всем метрикам
* Лучшие результаты были показаны методом ближайших соседей

# **Создание веб-приложения**

В качестве модели была выбрана логистическая регрессия.

Сначала импортируем необходимые библиотеки, в функции load\_data загружаем набор данных, обрабатываем данные: удаляем лишние столбцы, кодируем категориальные признаки с помощью LabelEncoder и масштабируем числовые с помощью MinMaxScaler.

Затем в фукнции preprocess\_data разделяем данные на тренировочную и тестовую выборку с помощью функции train\_test\_split.

Создаем функцию class\_accuracy\_score, которая будет вычислять значение accuracy для предсказаний модели для каждого класса по отдельности.

Устанавливаем с помощью st.sidebar.header заголовок боковой панели. С помощью st.sidebar.slider создаем ползунки для регулирования значений C и class\_weight, с помощью st.sidebar.radio создаем перключатель для значения solver.

С помощью st.checkbox создаем чекбокс для отображения матрицы корреляции набора данных.

Используя значения, полученные из переключателей и ползунков, создадим и натренируем модель.

С помощью st.header отобразим заголовок сайта.

С помощью st.subheader отобразим подзаголовок для таблицы точностей.

С помощью st.dataframe отобразим результаты использования функции class\_accuracy\_score.

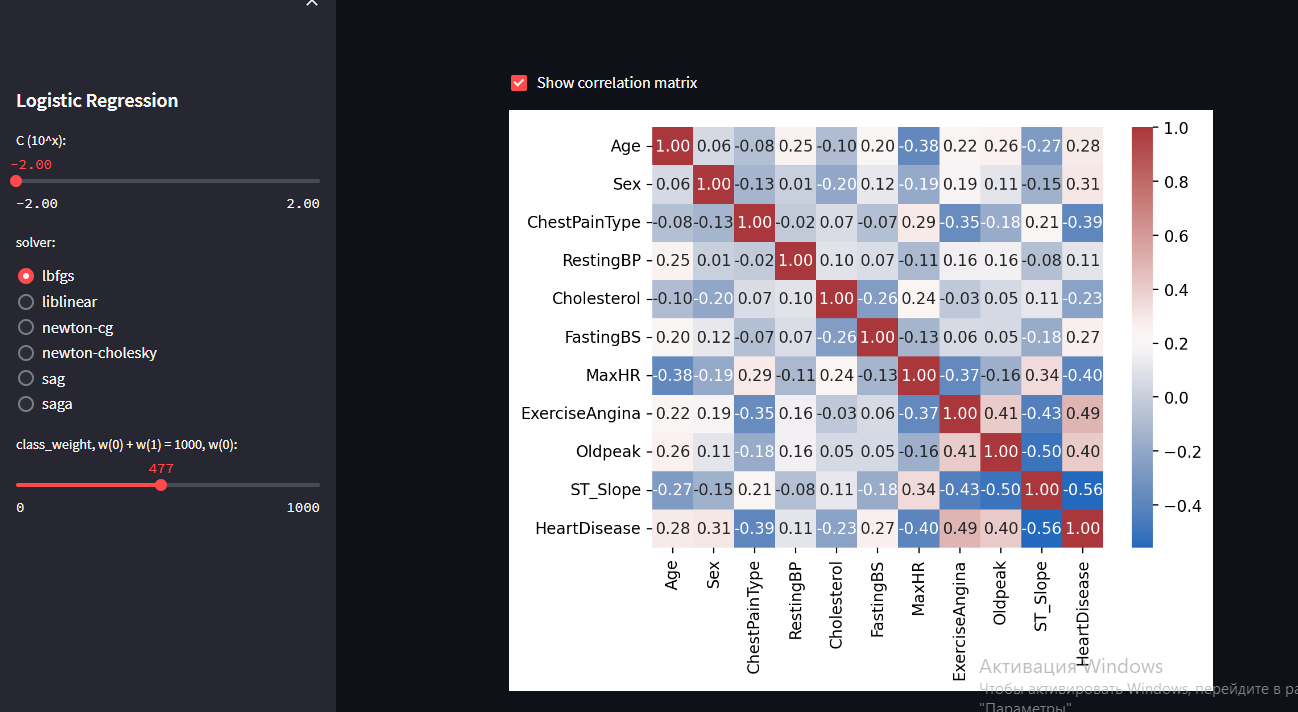
С помощью st.subheader отобразим подзаголовок для матрицы ошибок.

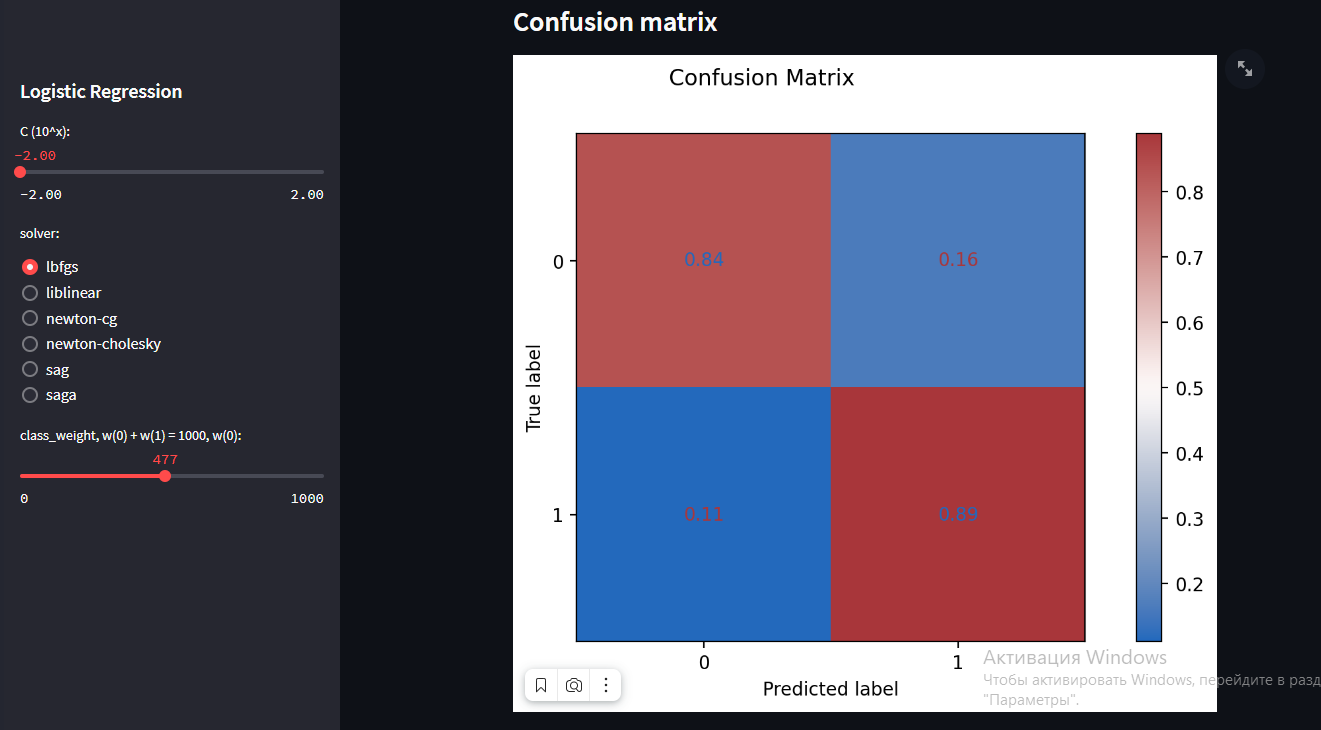
С помощью st.pyplot и ConfusionMatrixDisplay.from\_predictions отобразим матрицу ошибок.

Создадим функцию draw\_roc\_curve для отрисовки ROC кривой по предсказаниям модели.

С помощью st.subheader отобразим подзаголовок для ROC кривой.

С помощью st.pyplot и draw\_roc\_curve отобразим ROC кривую.





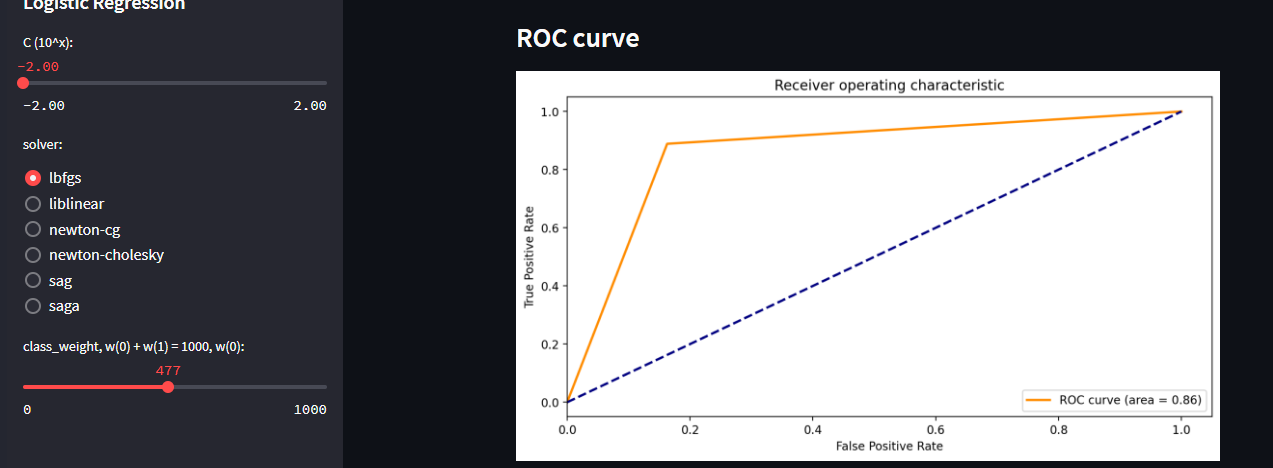


Рисунок 34 - Скриншот веб-приложения

# **Заключение**

Классификация сердечной недостаточности с помощью методов машинного обучения является актуальной и перспективной задачей в медицинской области. Анализ и обработка медицинских данных с помощью алгоритмов машинного обучения могут помочь в определении вероятности возникновения сердечной недостаточности у пациентов и принятии своевременных мер для предотвращения ее возникновения.

В рамках НИР была рассмотрена задача классификации сердечной недостаточности с помощью методов машинного обучения. Данные были проанализированы, визуализированы и подготовлены к обучению. Были применены различные алгоритмы, такие как метод ближайших соседей, логистическая регрессия, опорных векторов, случайный лес, градиентный бустинг.

В результате исследования было показано, что большинство использованных методов могут достичь хороших результатов, но самым точным оказался метод 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 (дата обращения: 30.04.2023).
2. Machine Learning Metrics in simple terms // Medium URL: https://medium.com/analytics-vidhya/machine-learning-metrics-in-simple-terms-d58a9c85f9f6 (дата обращения: 01.05.2023).
3. Опорный пример для выполнения проекта по анализу данных. // Jupyter nbviewer URL: https://scikit-learn.org/stable/tutorial/statistical\_inference/supervised\_learning.html (дата обращения: 25.04.2023).
4. Репозиторий курса "Технологии машинного обучения", бакалавриат, 6 семестр. // GitHub URL: https://github.com/ugapanyuk/courses\_current/wiki/COURSE\_TMO\_SPRING\_2023/ (дата обращения: 25.04.2023).