|  |  |
| --- | --- |
|  | **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 | ИУ5-63Б |  |  |  | **К.А. Буйдина** |
|  | (группа) |  | (подпись, дата) |  | (И.О. Фамилия) |
| Руководитель НИР |  |  |  |  | **Ю.Е. Гапанюк** |
|  |  |  | (подпись, дата) |  | (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) | |

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| по теме | | Разработка и оценка моделей машинного обучения | | | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
| Студент группы | | | | ИУ5-63Б | | | |  | | | | | | | |
|  | | | | Буйдина Кристина Анатольевна | | | | | | | | | | | |
| 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:*** | | | | | *решение задачи машинного обучения на основе материалов* | | | | | | | | | | |
| дисциплины. Выбор датасета, первичный анализ, выбор метрик для оценки качества моделей, | | | | | | | | | | | | | | | |
| построение базового решения, оценка качества, подбор гиперпараметров. | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
| ***Оформление научно-исследовательской работы:*** | | | | | | | | | | |  | | | | |
| Расчетно-пояснительная записка на \_\_\_\_\_ 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** |  |  |  | **К.А. Буйдина** |
|  |  | (подпись, дата) |  | (И.О. Фамилия) |

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

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 4](#_Toc197107953)

[1. ПОСТАНОВКА ЗАДАЧИ 5](#_Toc197107954)

[2. АНАЛИЗ ДАТАСЕТА 6](#_Toc197107955)

[3. ВЫБОР МОДЕЛЕЙ И МЕТРИК ДЛЯ ОЦЕНКИ КАЧЕСТВА 8](#_Toc197107956)

[4. ПОСТРОЕНИЕ БАЗОВОГО РЕШЕНИЯ 9](#_Toc197107957)

[5. ПОДБОР ГИПЕРПАРАМЕТРОВ 10](#_Toc197107958)

[6. ОЦЕНКА КАЧЕСТВА МОДЕЛЕЙ 12](#_Toc197107959)

[7. ВЕБ-ПРИЛОЖЕНИЕ 13](#_Toc197107960)

[ЗАКЛЮЧЕНИЕ 15](#_Toc197107961)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 16](#_Toc197107962)

# **ВВЕДЕНИЕ**

В современном мире очень важно сделить за здоровьем. Каждый день люди читают множество негативных новостей, едут в общественном транспорте, где долго приходится стоять, много часов проводят в сидячем положении и редко занимаются спортом. Все это влияет на ключевые показатели здоровья, из-за чего возникает большой риск сердечных заболеваний. В рамках представленной работы мной было проведено исследование, целью которого был анализ различных моделей машинного обучения, которые могут быть использованы для прогнозирования потенциального риска сердечного заболевания.

Исследование было сосредоточено на анализе набора данных «Heart Failure Prediction Dataset», содержащего информацию о 1000 наборов показателей здоровься пациентов, у которых было положительное или отризательное значение сердечного заболевания. Датасет включает такие характеристики, как возраст, пол, тип боли в груди, артериальное давление в состоянии покоя, холестерин, ЭКГ, ЧСС, стенокардия, депрессия, наклон сегмента пиковой нагрузки и наличие сердечного заболевания. Целью работы стало построение и оценка моделей машинного обучения для прогнозирования наличия сердечного заболевания на основе имеющихся данных.

Модели машинного обучения представляют собой мощный инструмент для анализа сложных взаимосвязей между различными характеристиками пациентов и их влиянием на наличие сердечного заболевания. Представленное исследование включает в себя всесторонний анализ данных, включая предварительное исследовательское изучение, выбор и обработку признаков, корреляционный анализ, а также построение и сравнение различных моделей машинного обучения с подбором оптимальных гиперпараметров.

Результаты работы могут быть использованы в медицине врачами и обычными пациентами для оценки потенциального наличия сердечного заболевания, а также для более глубокого понимания факторов, влияющих на появления сердечных заболеваний.

# **1. ПОСТАНОВКА ЗАДАЧИ**

Целью научно-исследовательской работы является разработка и оценка моделей машинного обучения для прогнозирования наличия сердечного заболевания. Исследование направлено на выявление ключевых факторов, влияющих на появление сердечных заболеваний, и создание моделей, способных с высокой точностью предсказывать потенциалую вероятность наличия заболевания у пациента на основе различных характеристик здоровья.

Датасет включает в себя следующие признаки:

| **Поле** | **Описание** | **Тип данных** |
| --- | --- | --- |
| Age | Возраст | int64 |
| Sex | Пол | string |
| ChestPainType | Тип боли в груди | string |
| RestingBP | Артериальное давление в состоянии покоя | int64 |
| Cholesterol | Сывороточный холестерин | int64 |
| FastingBS | Уровень сахара в крови | int64 |
| RestingECG | Результаты электрокардиограммы | string |
| MaxHR | Максимальная частота сердечных сокрацений | int64 |
| ExerciseAngina | Стенокардия, вызванная физическим напряжением | string |
| Oldpeak | Числовое значение, измененное в депрессии | float64 |
| ST\_Slope | Наклон сегмента ST пиковой нагрузки | string |
| HeartDisease | Сердечное заболевание | int64 |

# **2. АНАЛИЗ ДАТАСЕТА**

Первичный анализ датасета:

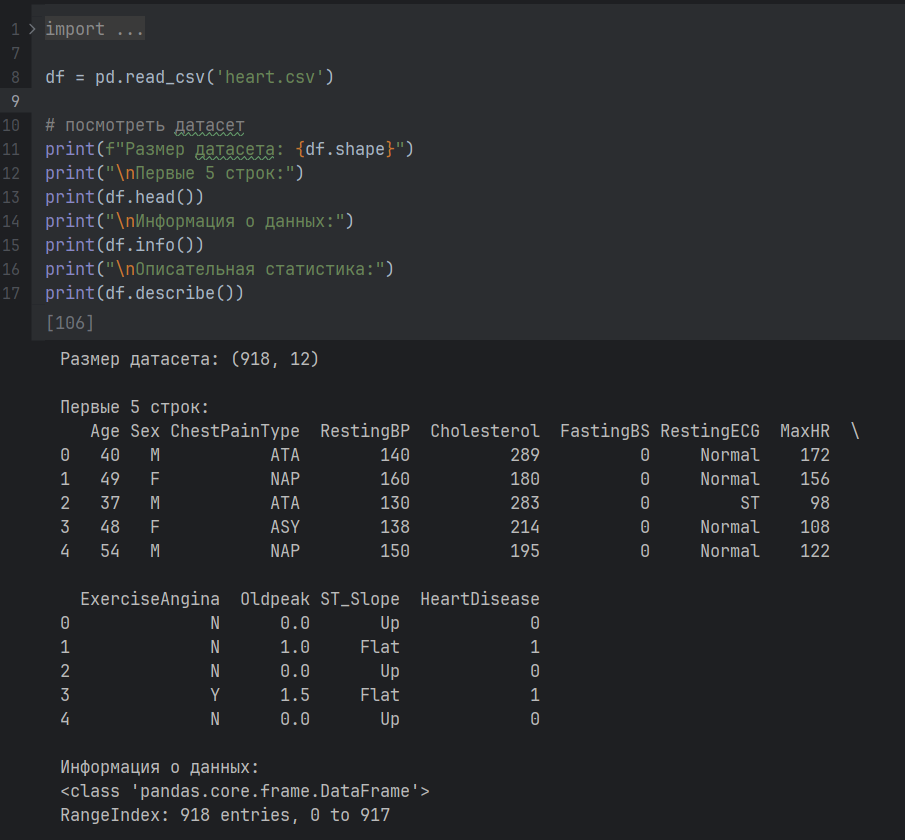


Рисунок 1 – Общий просмотр содержания датасета.

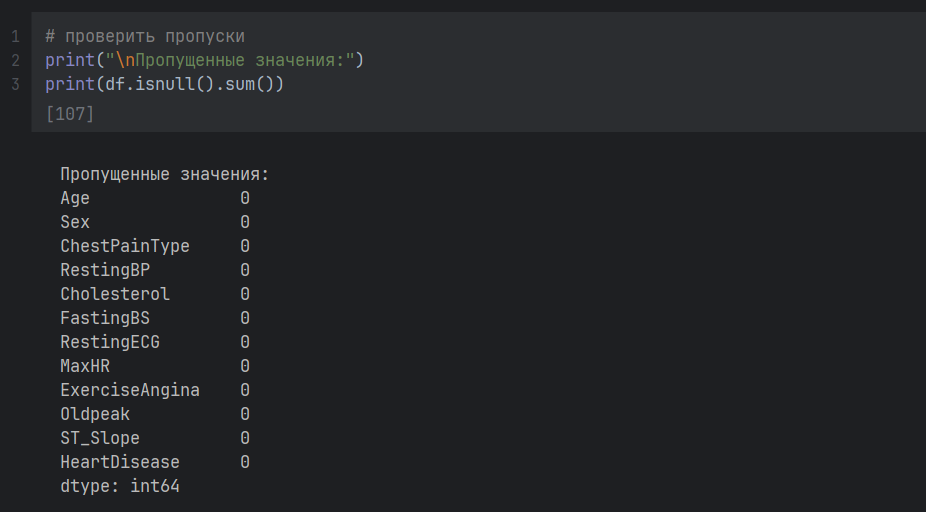


Рисунок 2 – Анализ пропусков.

Визуализация данных:

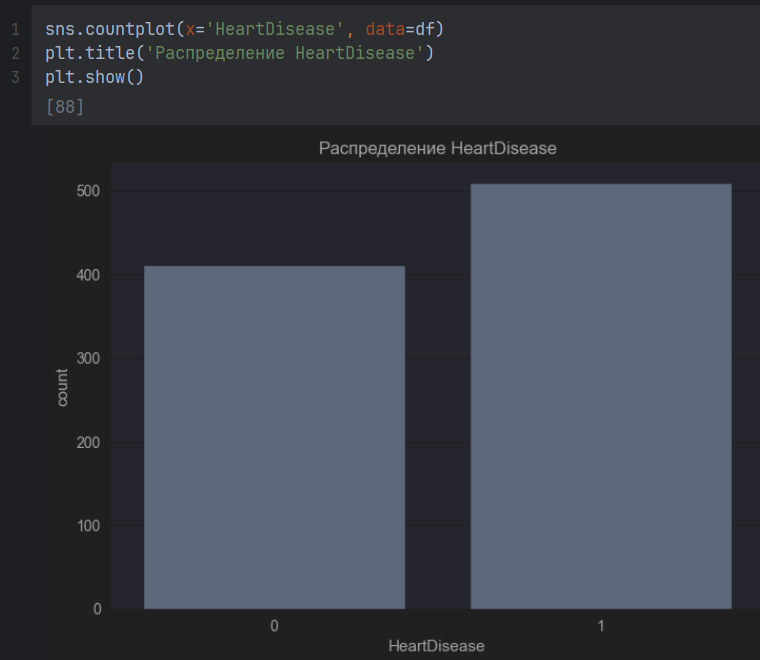


Рисунок 3 – Распределение значений ключевого признака.



Рисунок 4 – Распределение значений остальных признаков.

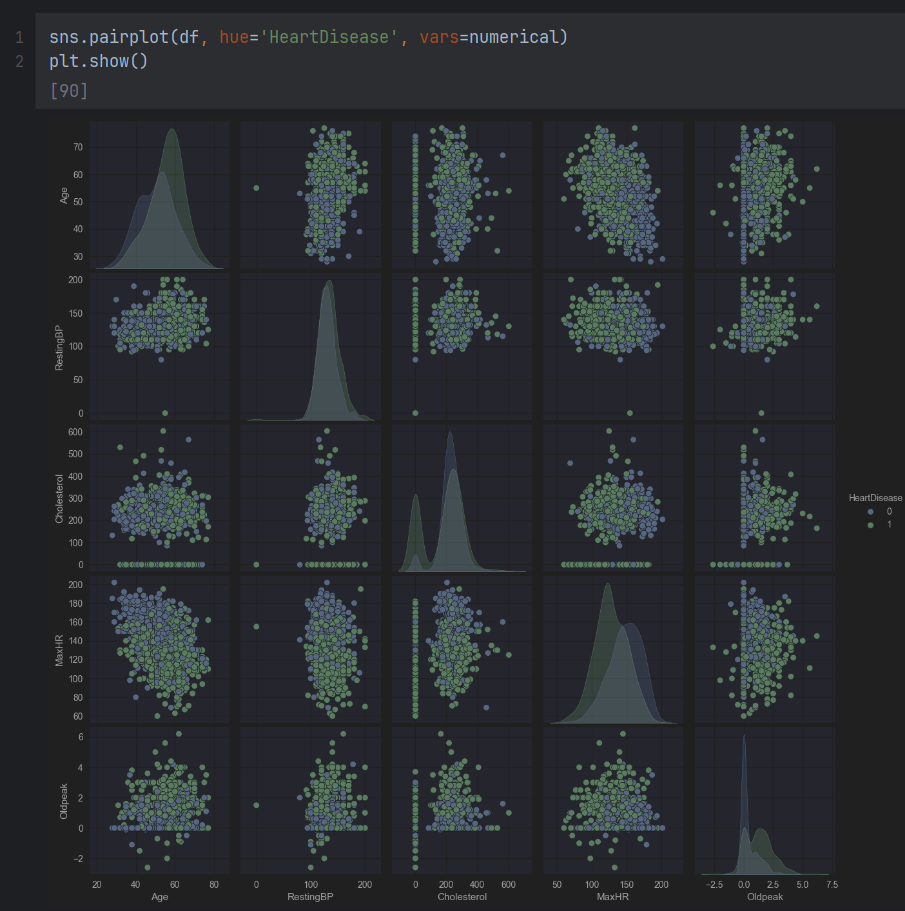


Рисунок 5 – Распределение значений признаков в зависимости от значения ключевого параметра.

Построим корреляционную матрицу признаков:

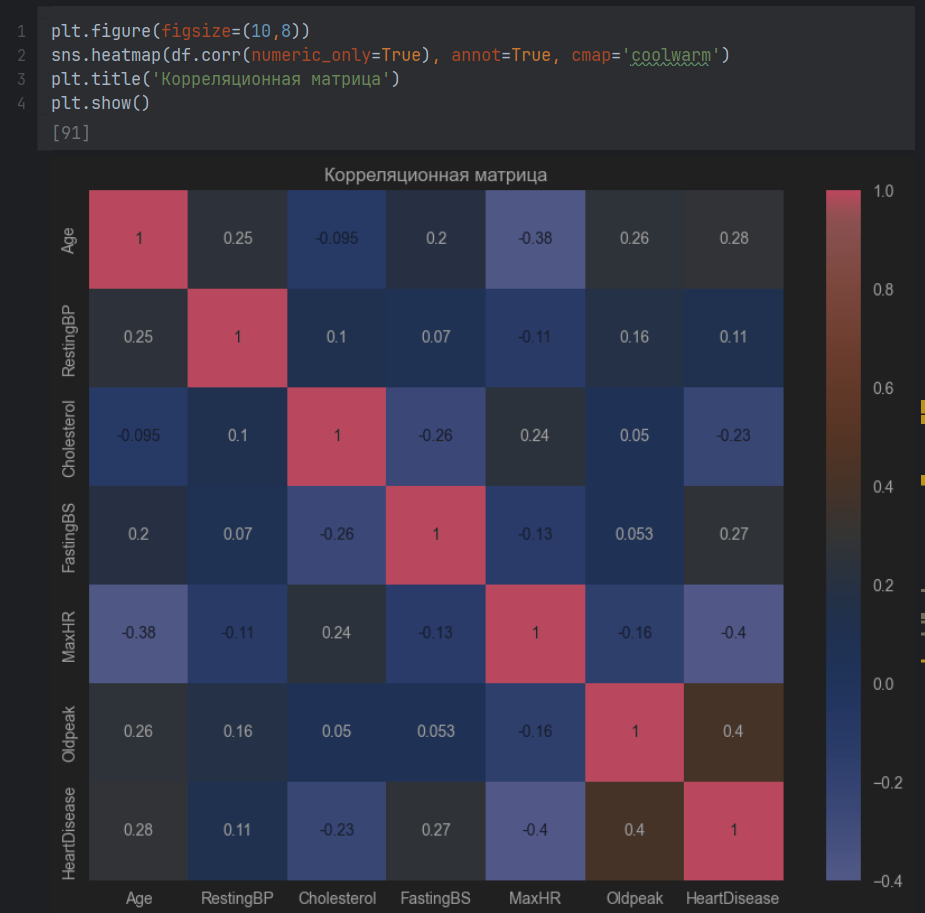


Рисунок 6 – Корреляционная матрица признаков.

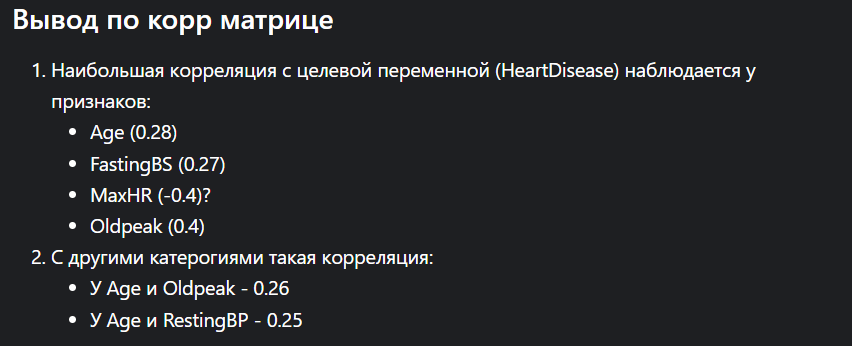


Рисунок 7 – Вывод по корреляционной марице.

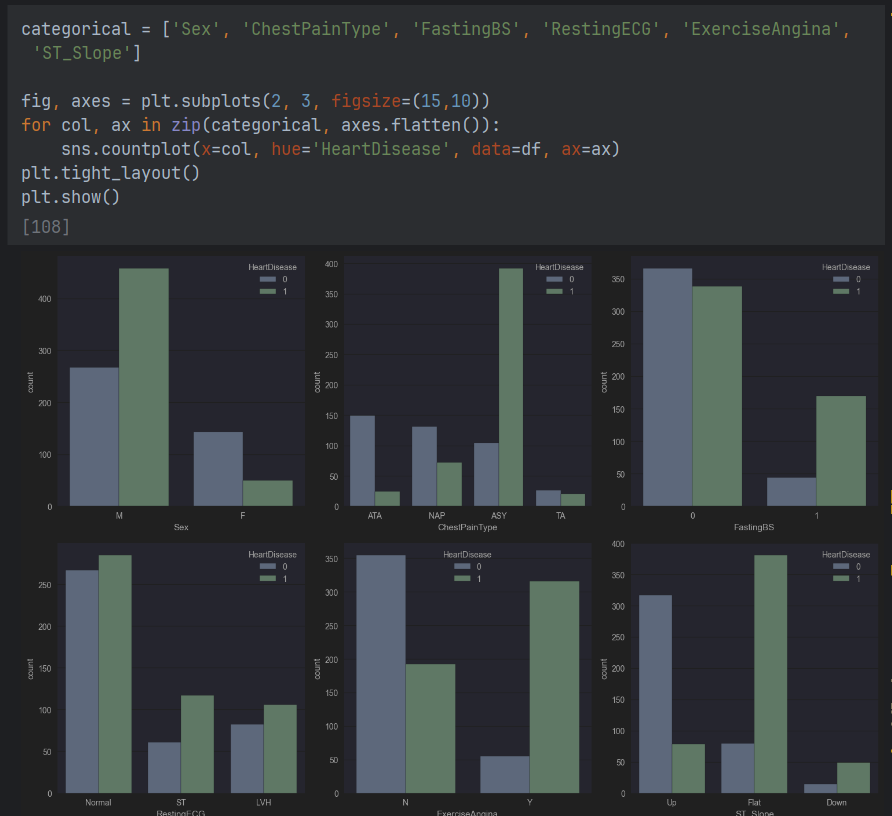


Рисунок 8 – Визуализация распределения признаков, которые больше всего коррелируют с ключевыми параметрами.

Сделаем масштабирование и кодирование данных:

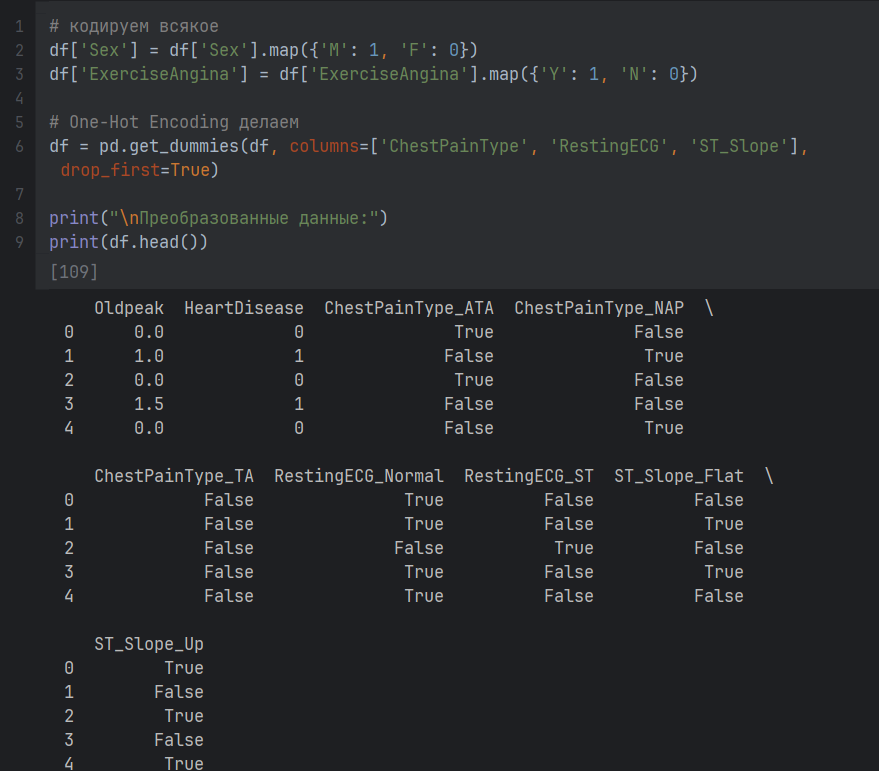


Рисунок 9 – Кодирование признаков.

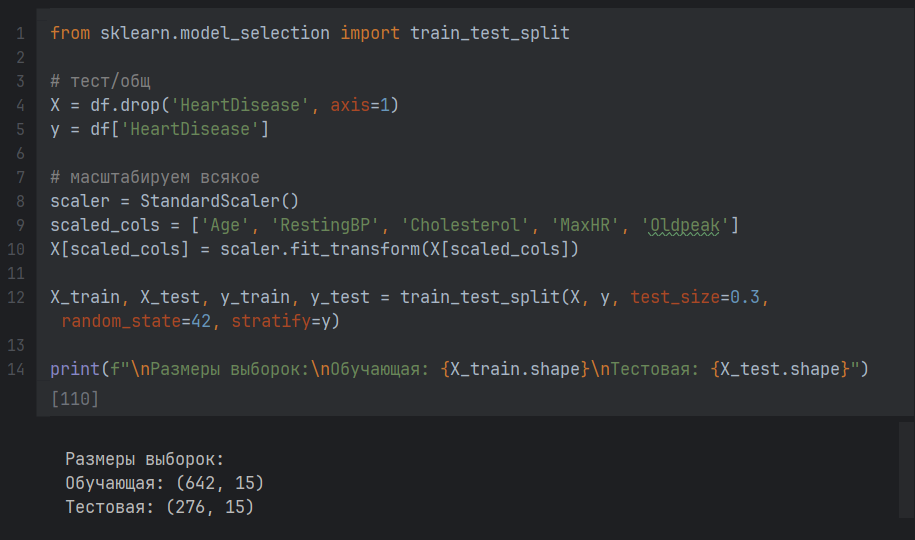


Рисунок 10 – Масштабирование признаков.

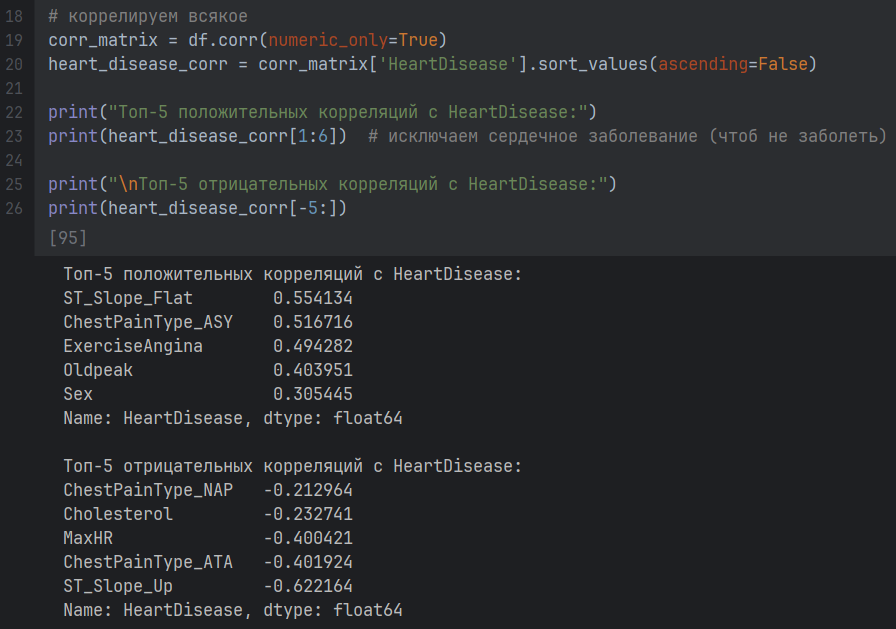


Рисунок 11 – Более точный анализ корреляций признаков.

# **3. ВЫБОР МОДЕЛЕЙ И МЕТРИК ДЛЯ ОЦЕНКИ КАЧЕСТВА**

Для оценки качества моделей выберем следующие метрики:

**• Accuracy (Точность общего предсказания)** - доля правильно предсказанных сердечных заболеваний. Показывает, насколько точно модель определяет наличие заболевания.

• **Precision (Точность)** - доля правильных предсказаний для каждой категории среди всех предсказанных в этой категории. Важна для понимания надежности предсказаний модели для конкретных диапазонов.

**• Weighted F1-score** - среднее значение F1-метрики с учетом частоты каждой категории характеристик. Обеспечивает сбалансированную оценку качества модели для всех диапазонов.

**• Recall (Полнота)** - показывает, насколько хорошо модель определяет наличие сердечного заболевания, на которое влияют другие характеристики. Важна для оценки способности модели не пропускать случаи с высокой потенциальной вероятностью сердечного заболевания.

**• ROC-AUC**  - показывает способность бинарного классификатора различать положительный и отрицательный классы при различных пороговых значениях классификации.

С учетом следующих факторов:

* Требование задания НИРС: не менее 5 моделей, не менее 2 ансамблевых моделей;
* Решение задачи регрессии и классификации (преобразование задачи);
* Датасет содержит числовые признаки.

Выберем следующие модели:

* Линейная регрессия;
* Метод ближайших соседей;
* Метод опорных векторов.
* Дерево решений;
* Случайный лес;
* Градиентный бустинг;

# **4. ПОСТРОЕНИЕ БАЗОВОГО РЕШЕНИЯ**

Построим базовое решение для каждой из моделей:

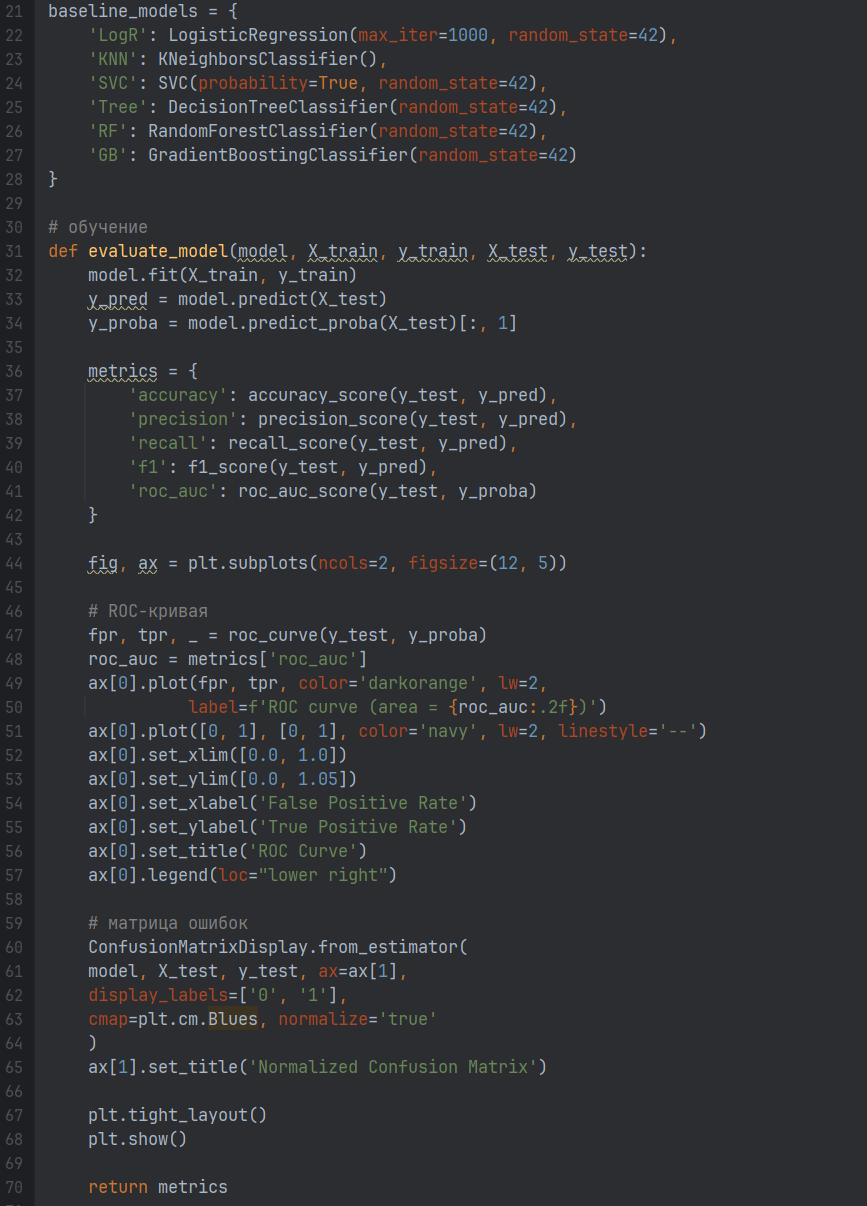
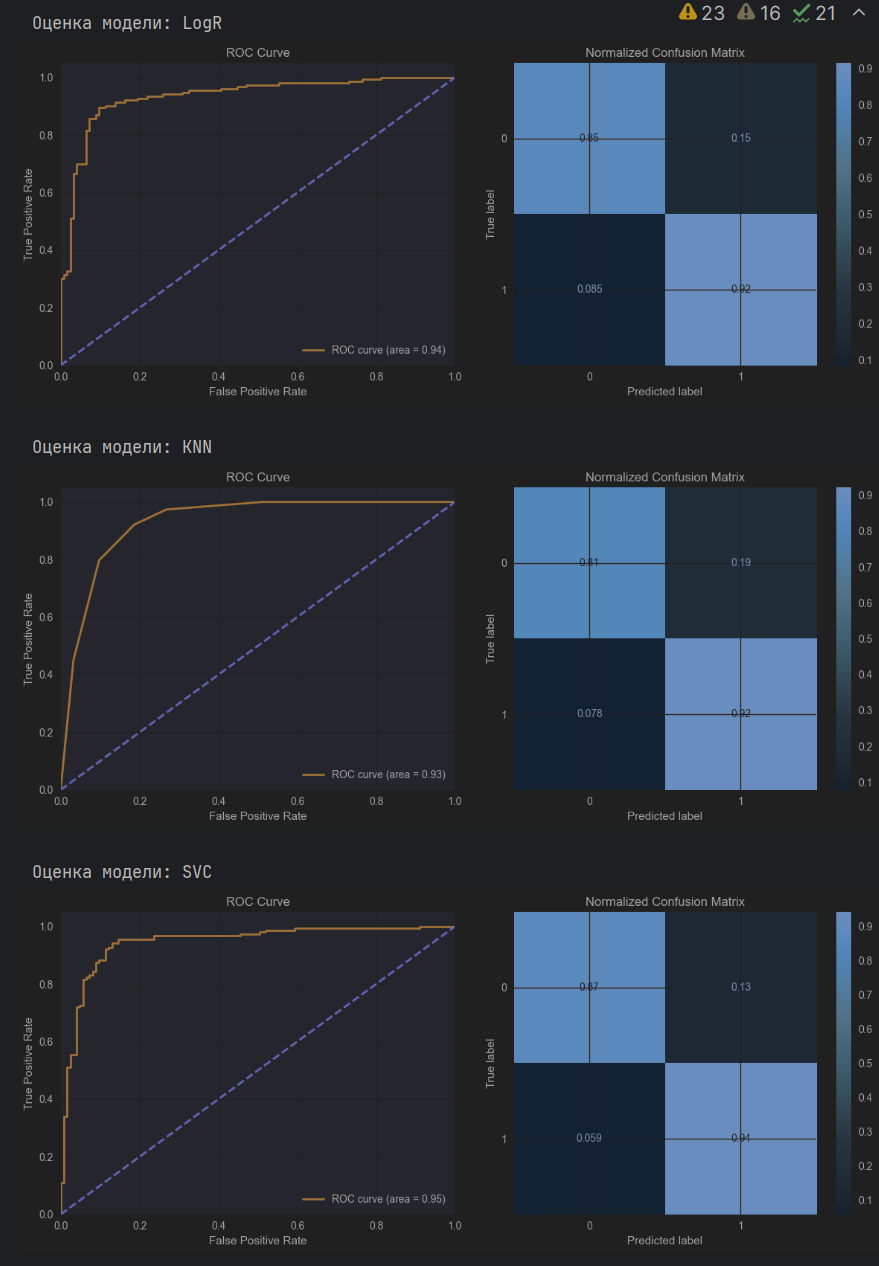


Рисунок 12 – Базовое решение.



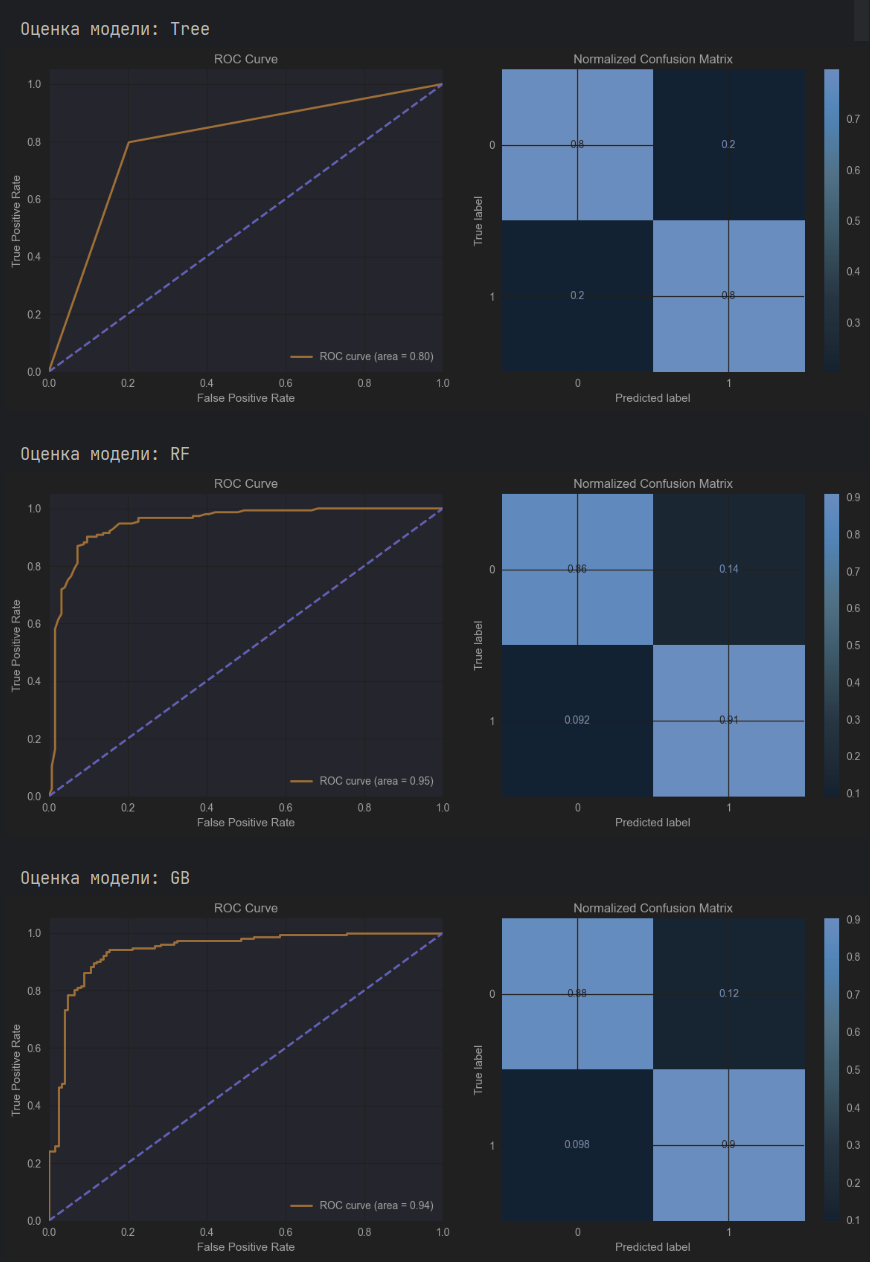


Рисунок 13 – Визуализация точности базового решения.

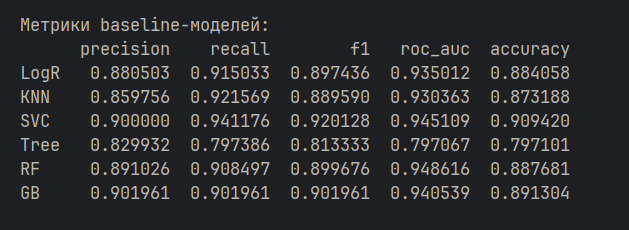
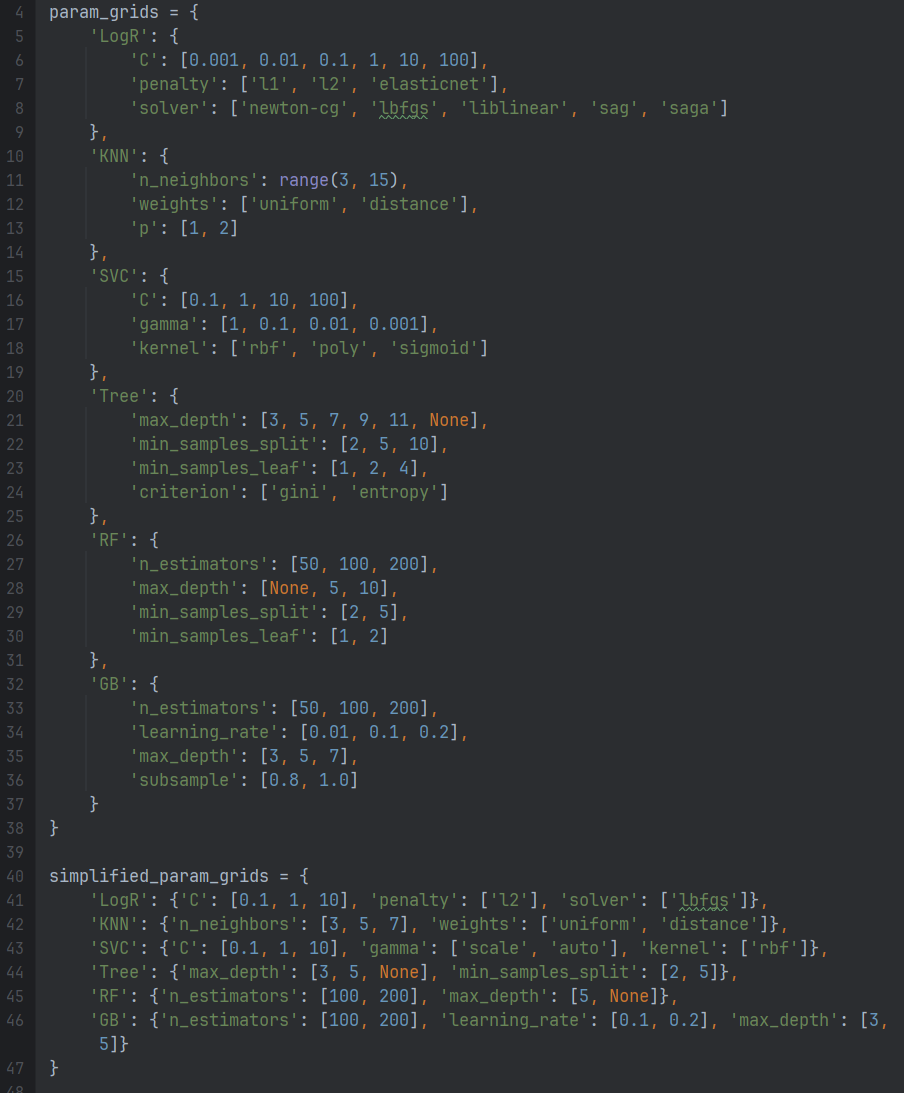
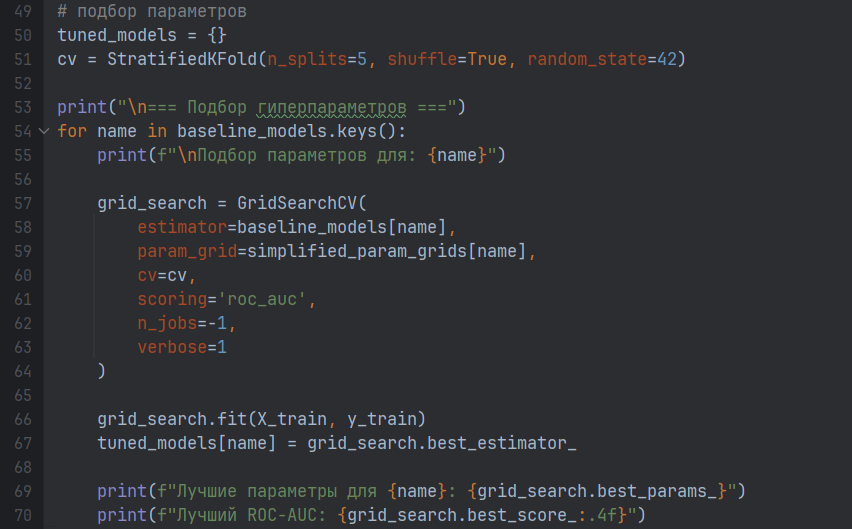


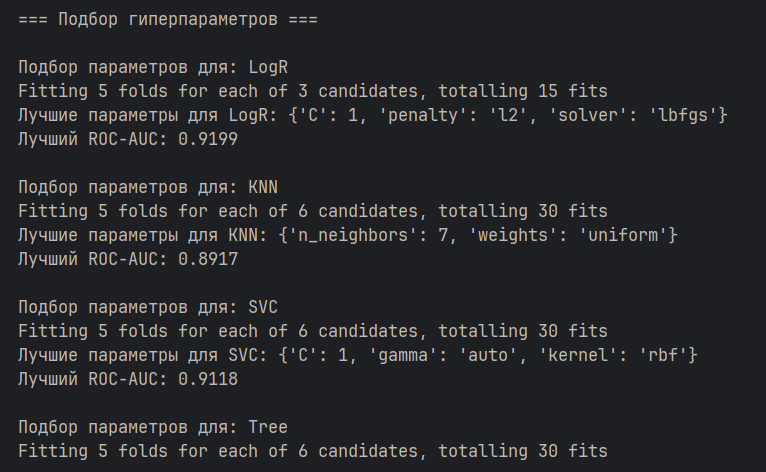
Рисунок 14 – Оценка точности базового решения.

# **5. ПОДБОР ГИПЕРПАРАМЕТРОВ**

Для повышения точности моделей произведём подбор гиперпараметров моделей с использованием функции **GridSearchCV** и оптимизацию.







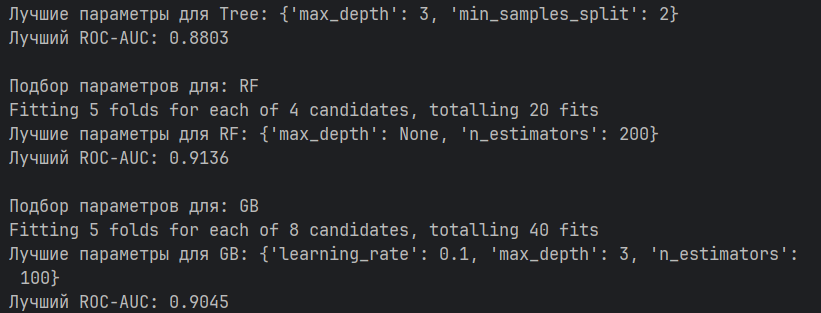
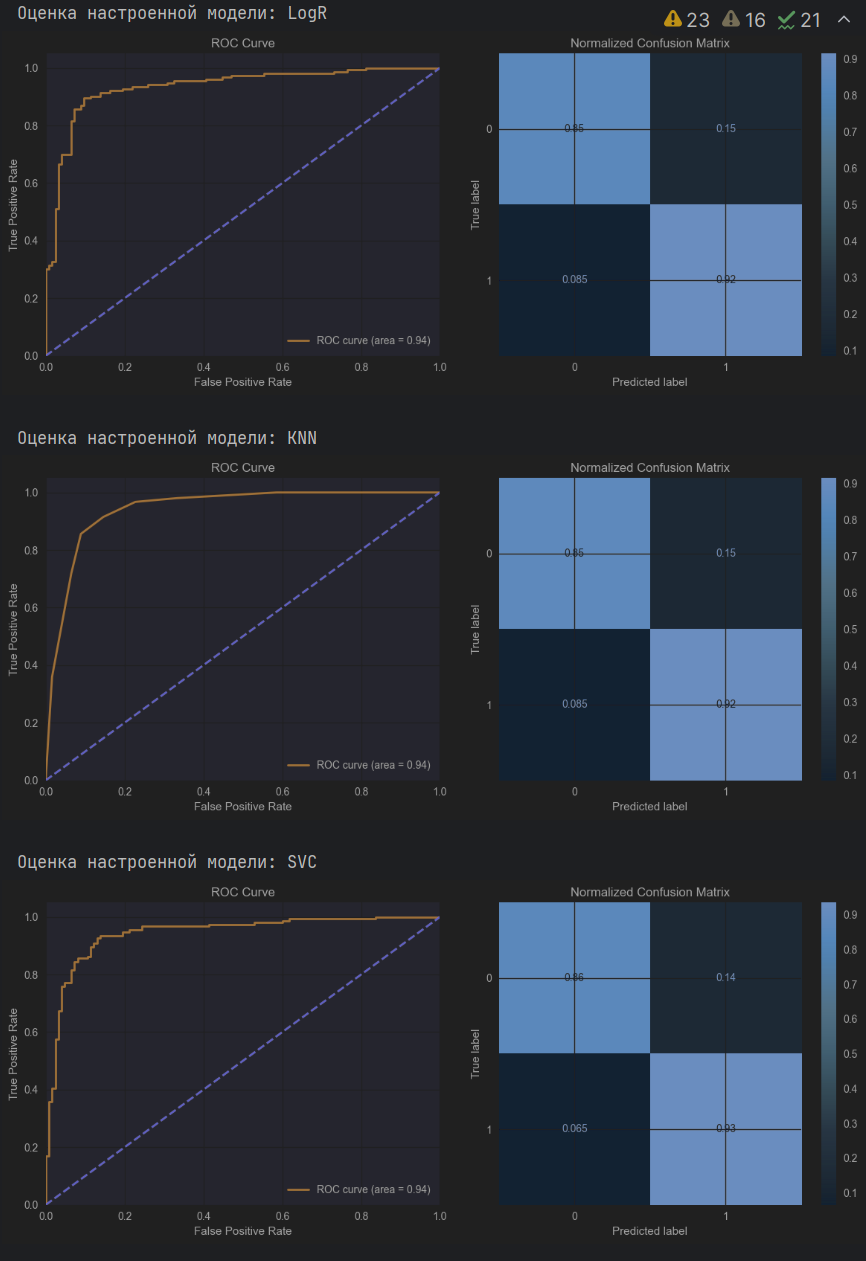


Рисунок 15 – Подбор гиперпараметров.

# **6. ОЦЕНКА КАЧЕСТВА МОДЕЛЕЙ**

Построим диаграммы, отображающие оценки качества исследуемых моделей после подбора гиперпараметров:



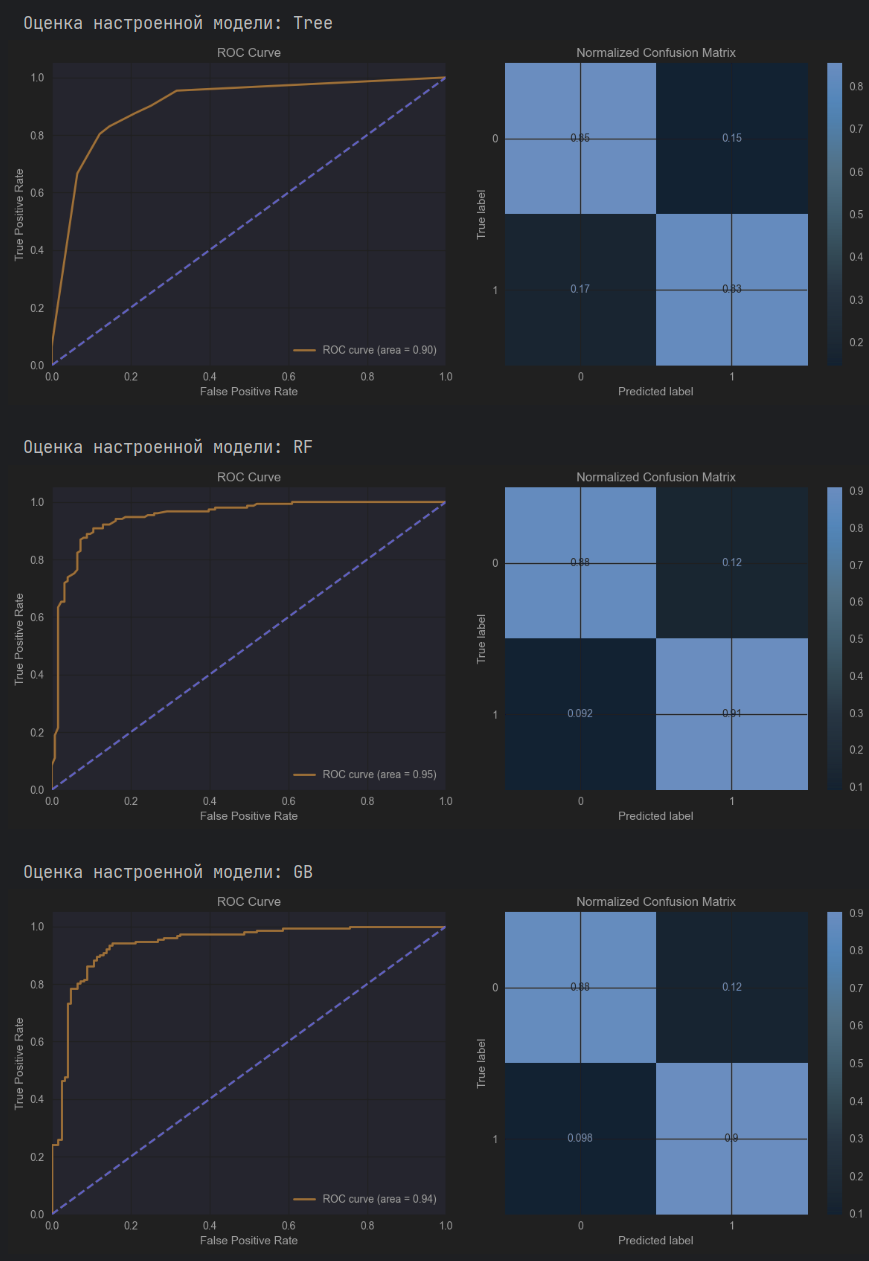


Рисунок 16 – Оценка моделей после подбора гиперпараметров.

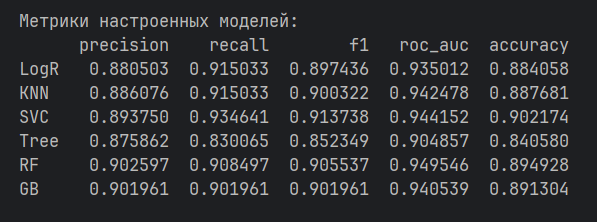
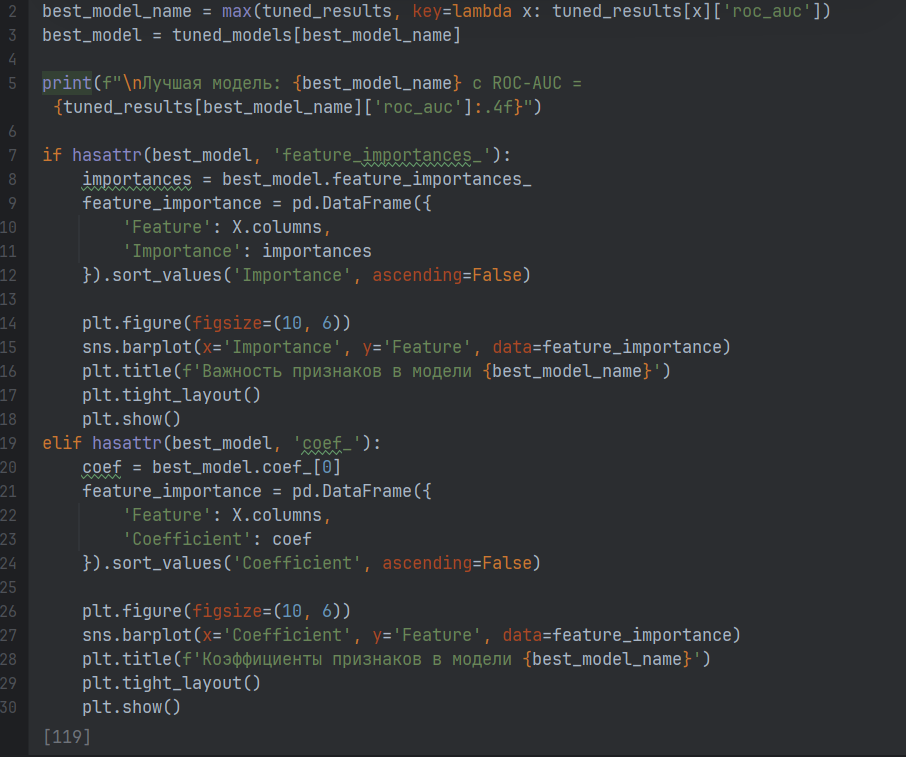


Рисунок 17 – Оценка моделей.



Рисунок 18 – Выводы по оценке моделей.

В ходе исследования наибольшую точность показала модель **Случайный лес (Random Forest)**.



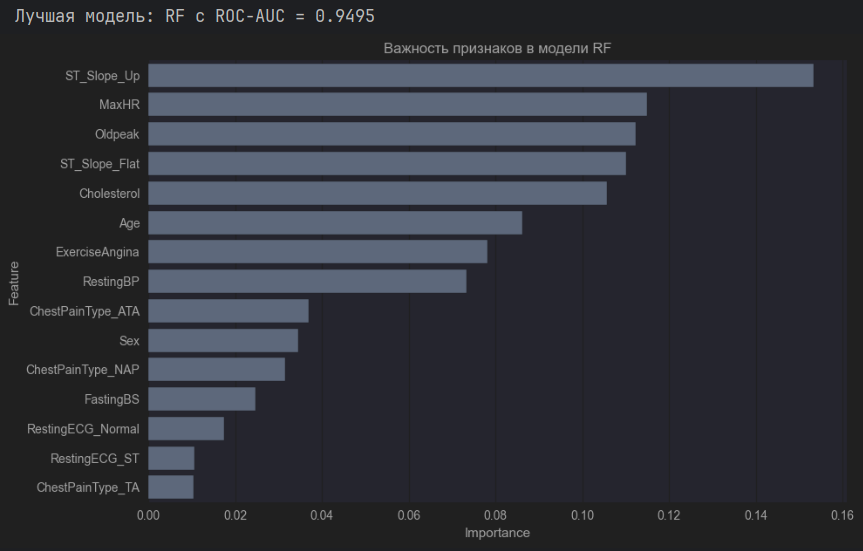
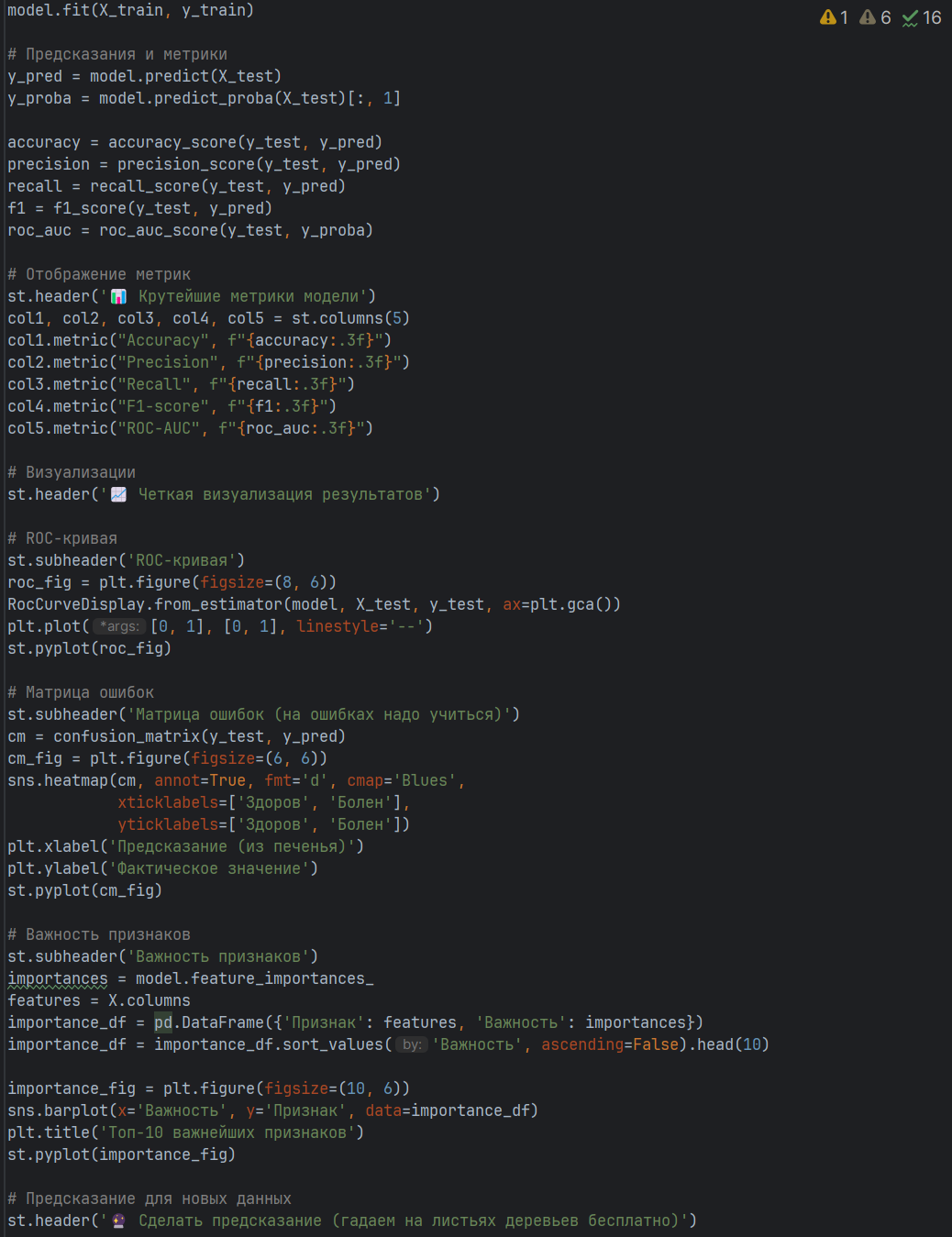


Рисунок 19 – Визуализация параметров, взияющих на модель.

# **7. ВЕБ-ПРИЛОЖЕНИЕ**

Реализуем веб-приложение для демонстрации влияния гиперпараметров на точность модели **Линейной регрессии,** **Случайный лес, Дерево решений**. Используем фреймворк **Streamlit**.







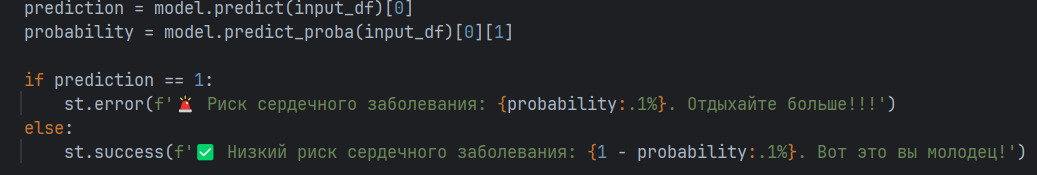
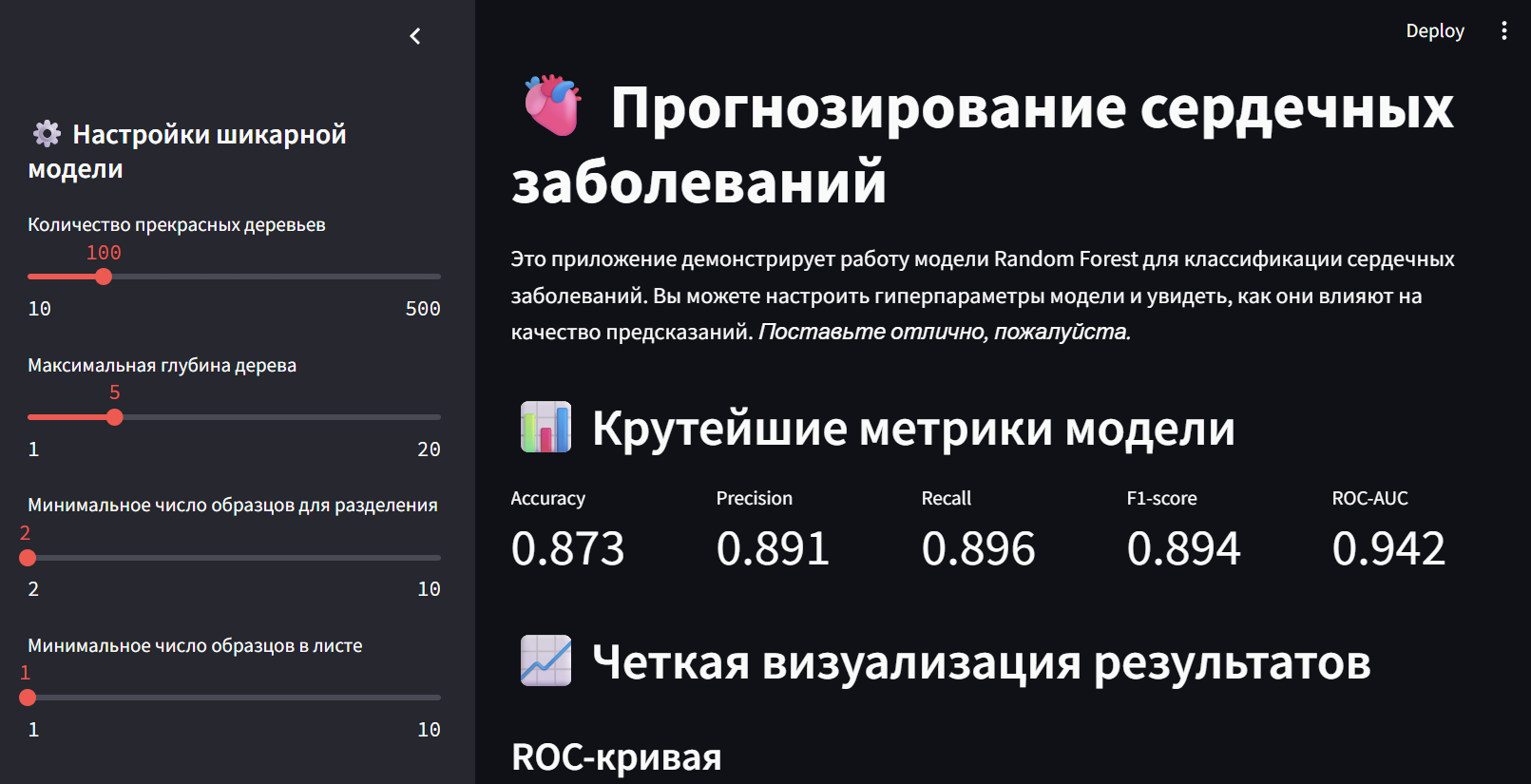
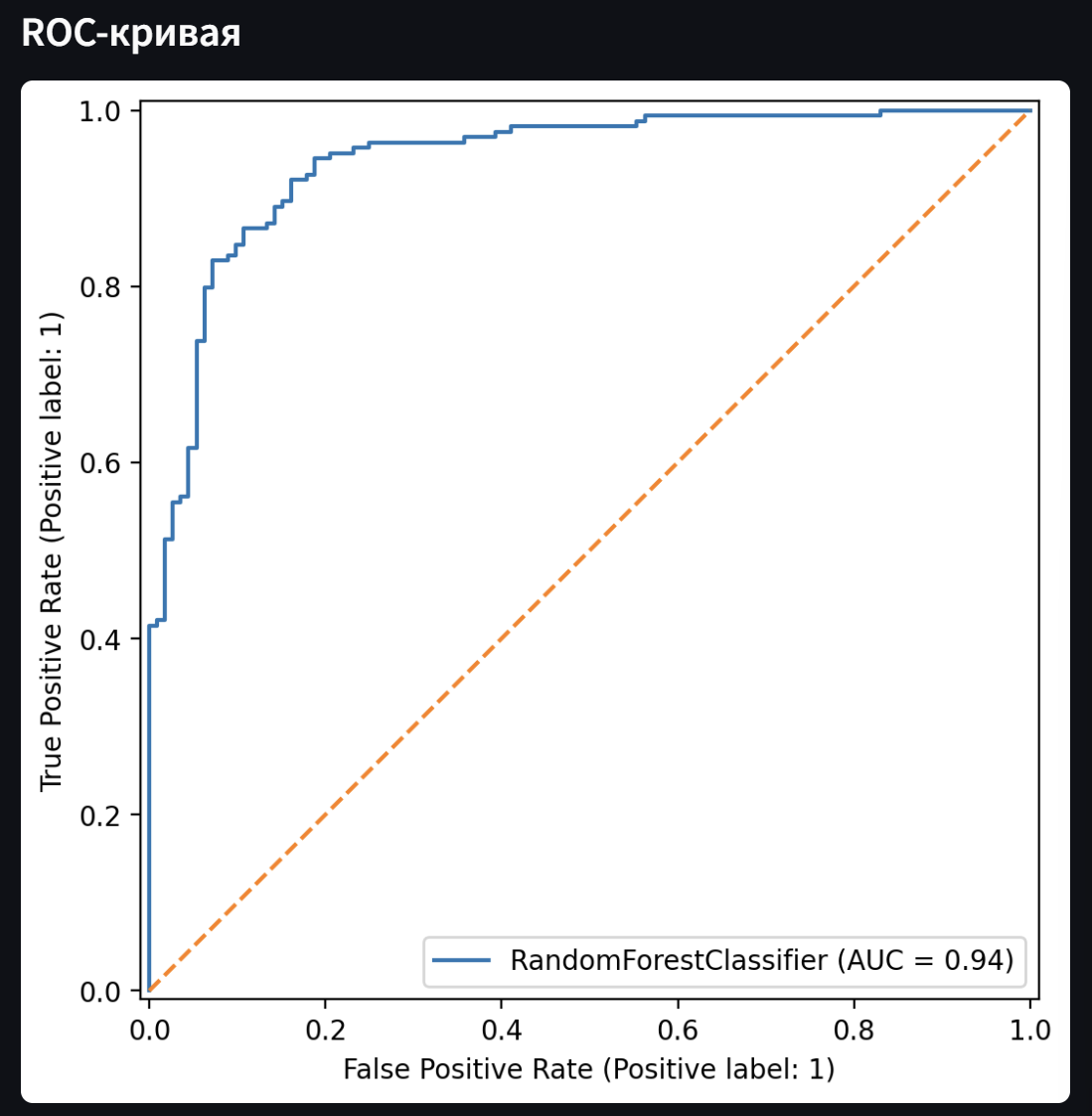
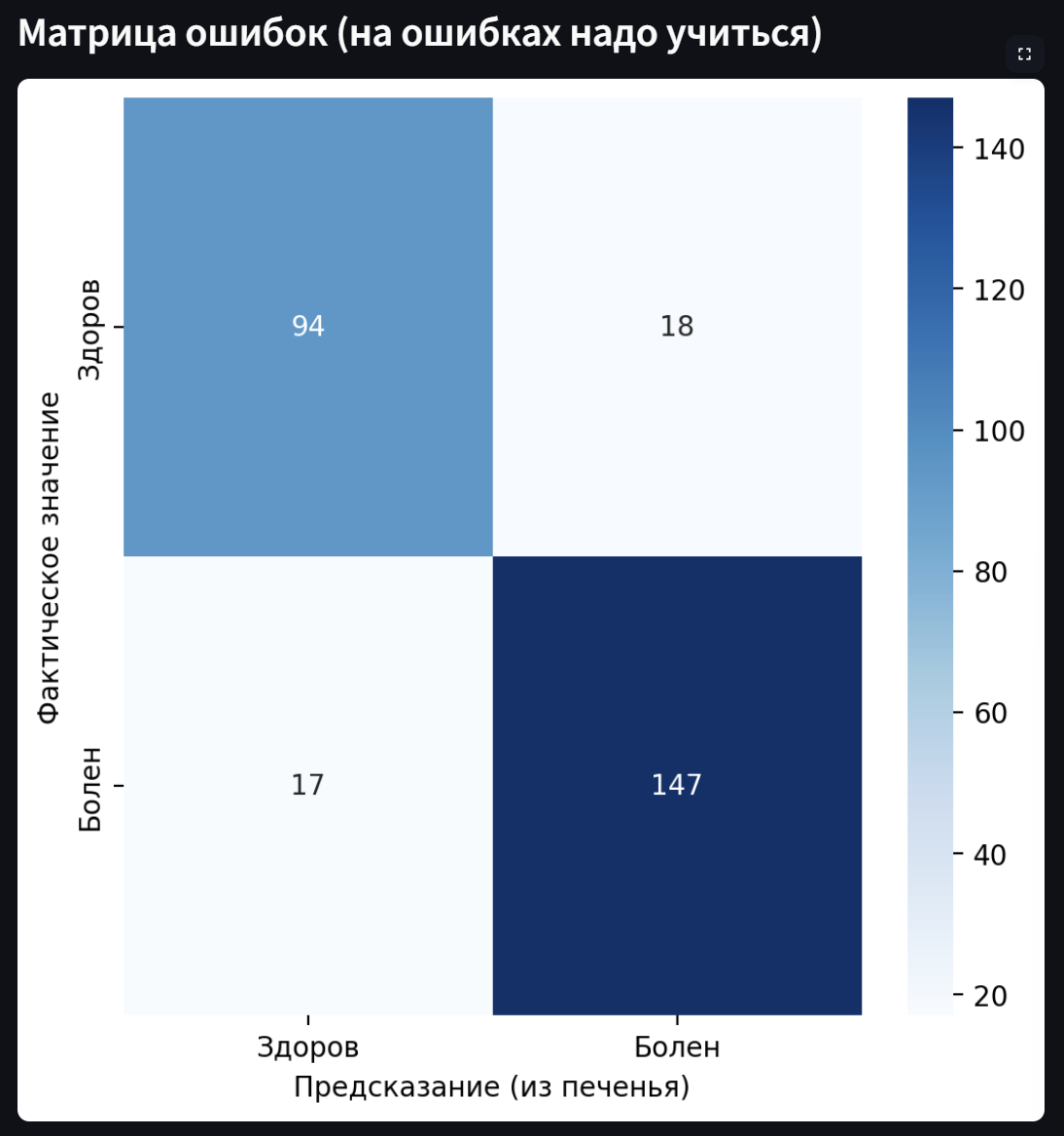
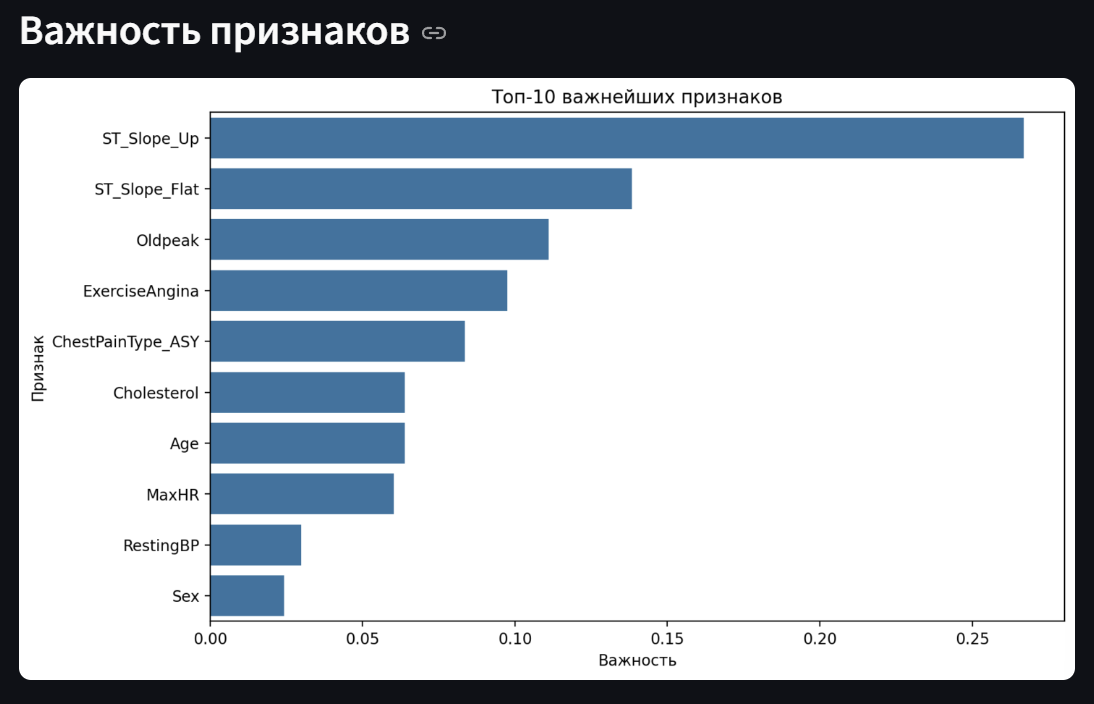


Рисунок 20 – Реализация веб-приложения для демонстрации модели.









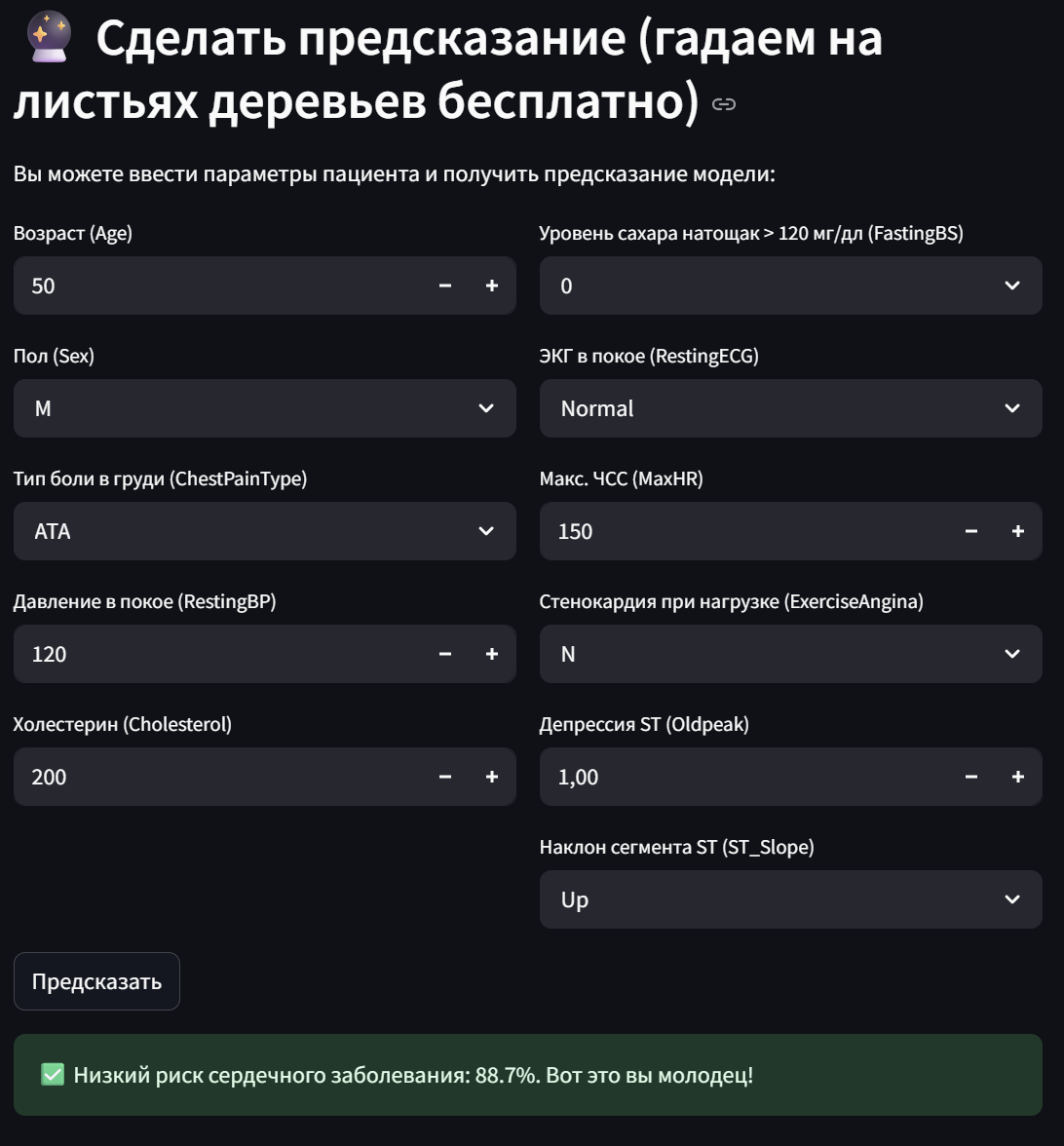


Рисунок 21 – Веб-приложение.

# **ЗАКЛЮЧЕНИЕ**

В ходе выполнения научно-исследовательской было проведено исследование задачи определения сердечного заболевания. Были проанализированы различные модели машинного обучения, в том числе ансамблиевые, а также их оптимизация путём подбора гиперпараметров.

Наиболее высокую точность по исследуемым метрикам показала модель **Случайный лес** после подбора гиперпараметров.

Для демонстрации влияния гиперпараметров на точность модели было разработано веб-приложение.

Таким образом, исследуемые модели машинного обучения и проведенный анализ позволяют не только прогнозировать вероятность наличияя сердечного заболевания, но и выявлять ключевые факторы, влияющие этот параметр. Результаты исследования могут служить основой для дальнейших научных и прикладных работ в области медицины.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Документация Streamlit [Электронный ресурс] // streamlit.io URL: https://streamlit.io/ (дата обращения: 01.05.2025);
2. «Python Data Science Handbook» Джейк Вандер-Плас [Электронный ресурс] // jakevdp.github.io. URL: https://jakevdp.github.io/PythonDataScienceHandbook/ (дата обращения: 02.05.2025);
3. Документация по Python [Электронный ресурс] // Python. URL: https://docs.python.org/3/index.html/ (дата обращения: 01.05.2025);
4. Методические указания НИРС по дисциплине «Технологии машинного обучения».