|  |  |
| --- | --- |
|  | **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 | РАДИОТЕХНИЧЕСКИЙ |
| КАФЕДРА | СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ |

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

|  |
| --- |
|  |
| ***Разработка и оценка моделей*** |
| ***машинного обучения*** |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | РТ5-61Б |  |  |  | А.В. Пересыпко |
|  | (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 методов машинного обучения | | | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
| Студент группы | | | | РТ5-61Б | | | |  | | | | | | | |
|  | | | | Пересыпко Александр Владимирович | | | | | | | | | | | |
| 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 \_\_\_\_\_ 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** |  |  |  | А.В. Пересыпко |
|  |  | (подпись, дата) |  | (И.О. Фамилия) |

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

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 4](#_Toc199437512)

[1. ПОСТАНОВКА ЗАДАЧИ 5](#_Toc199437513)

[2. ПОДГОТОВКА ДАННЫХ 6](#_Toc199437514)

[3. ИССЛЕДОВАТЕЛЬСКИЙ АНАЛИЗ ДАННЫХ 7](#_Toc199437515)

[4. ОБРАБОТКА И ПРЕОБРАЗОВАНИЕ ПРИЗНАКОВ 11](#_Toc199437516)

[5. ПОСТРОЕНИЕ И СРАВНЕНИЕ МОДЕЛЕЙ 12](#_Toc199437517)

[6. НАСТРОЙКА ГИПЕРПАРАМЕТРОВ 13](#_Toc199437518)

[7. ВЕБ-ПРИЛОЖЕНИЕ 14](#_Toc199437519)

[ЗАКЛЮЧЕНИЕ 16](#_Toc199437520)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 17](#_Toc199437521)

ВВЕДЕНИЕ

Машинное обучение сегодня является одним из самых динамично развивающихся направлений анализа данных и широко применяется для решения задач классификации, регрессии, кластеризации и других. В данной работе рассматривается задача многоклассовой классификации: по характеристикам участка леса необходимо определить тип лесного покрова (Cover Type).

Цель исследования — построить, обучить и сравнить несколько моделей машинного обучения, включая ансамблевые методы, на примере датасета CoverType. В рамках работы реализован полный цикл подготовки и анализа данных: проведён разведочный анализ, выполнена предобработка и масштабирование признаков, реализовано обучение и подбор гиперпараметров для различных моделей, а также проведена оценка их качества по ряду метрик. Итоговые результаты визуализированы и представлены в виде интерактивного веб-приложения, что демонстрирует практическую применимость полученного решения.

1. ПОСТАНОВКА ЗАДАЧИ

В рамках курсовой работы рассматривается задача многоклассовой классификации: прогнозирование типа лесного покрова участка по его характеристикам. Для решения используется открытый датасет CoverType, включающий 54 признака для каждого участка леса площадью 30×30 м в США. Целевая переменная — Cover\_Type (от 1 до 7).

Для решения задачи требуется:

1. Выполнить разведочный анализ данных;
2. Обработать пропуски и закодировать категориальные переменные;
3. Провести масштабирование признаков;
4. Сформировать обучающую и тестовую выборки;
5. Построить не менее пяти моделей;
6. Оценить их качество по нескольким метрикам;
7. Настроить гиперпараметры моделей;
8. Сравнить результаты и обосновать выбор финальной модели.

# ПОДГОТОВКА ДАННЫХ

В датасете CoverType (fetch\_covtype) все признаки (54 штуки) делятся на три группы:

Непрерывные (10 признаков рельефа и дистанций)

* Elevation (высота)
* Aspect (ориентация склона)
* Slope (уклон)
* Horizontal\_Distance\_To\_Hydrology
* Vertical\_Distance\_To\_Hydrology
* Horizontal\_Distance\_To\_Roadways
* Hillshade\_9am
* Hillshade\_Noon
* Hillshade\_3pm
* Horizontal\_Distance\_To\_Fire\_Points

Бинарные «заповедники» Wilderness Area (4 колонки) – Wilderness\_Area1, Wilderness\_Area2, Wilderness\_Area3, Wilderness\_Area4 (каждая = 1, если участок находится в соответствующем заповеднике)

Бинарные «типы почвы» Soil Type (40 колонок) – Soil\_Type1, Soil\_Type2, …, Soil\_Type40 (каждая = 1, если на участке именно этот тип почвы)

Целевая переменная – Cover\_Type (1…7), задаёт тип лесного покрытия.

1. ИССЛЕДОВАТЕЛЬСКИЙ АНАЛИЗ ДАННЫХ

Разведочный анализ данных позволил выявить особенности распределения признаков и их связь с целевой переменной. Это важный этап, позволяющий сформировать гипотезы и принять решения по обработке и отбору признаков.

Построим гистограмму, демонстрирующую распределение типов лесного покрова.

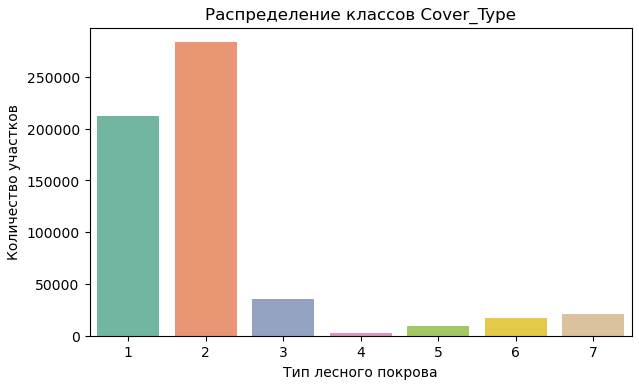


Рисунок 1 – Распределение типов лесного покрова

Распределение классов целевой переменной несбалансировано (некоторые типы встречаются чаще).

Проанализируем, какое распределение числовых признаков:

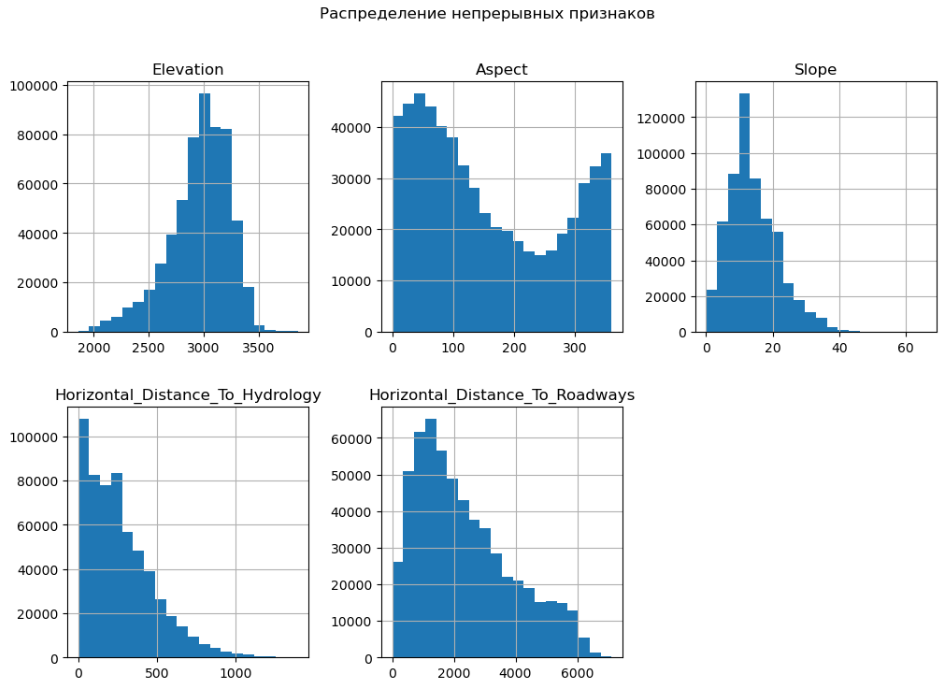


Рисунок 2 – Распределение числовых признаков

Распределения непрерывных признаков в датасете CoverType существенно различаются. Признак Elevation (высота) имеет выраженную асимметрию — большинство участков расположено на определённых высотах, что отражает особенности рельефа местности. Признаки Aspect (ориентация склона) и Slope (уклон) распределены более равномерно, однако встречаются и экстремальные значения. Расстояния до водоёмов и дорог (Horizontal\_Distance\_To\_Hydrology, Horizontal\_Distance\_To\_Roadways) характеризуются длинными "хвостами": большинство значений сконцентрировано вблизи нуля, но присутствуют и значительные выбросы. Эти особенности подтверждают необходимость масштабирования непрерывных признаков перед обучением моделей, а также указывают на возможное влияние выбросов на качество классификации.

Для комплексного понимания взаимосвязей была построена тепловая карта корреляций числовых переменных с целевой переменной.

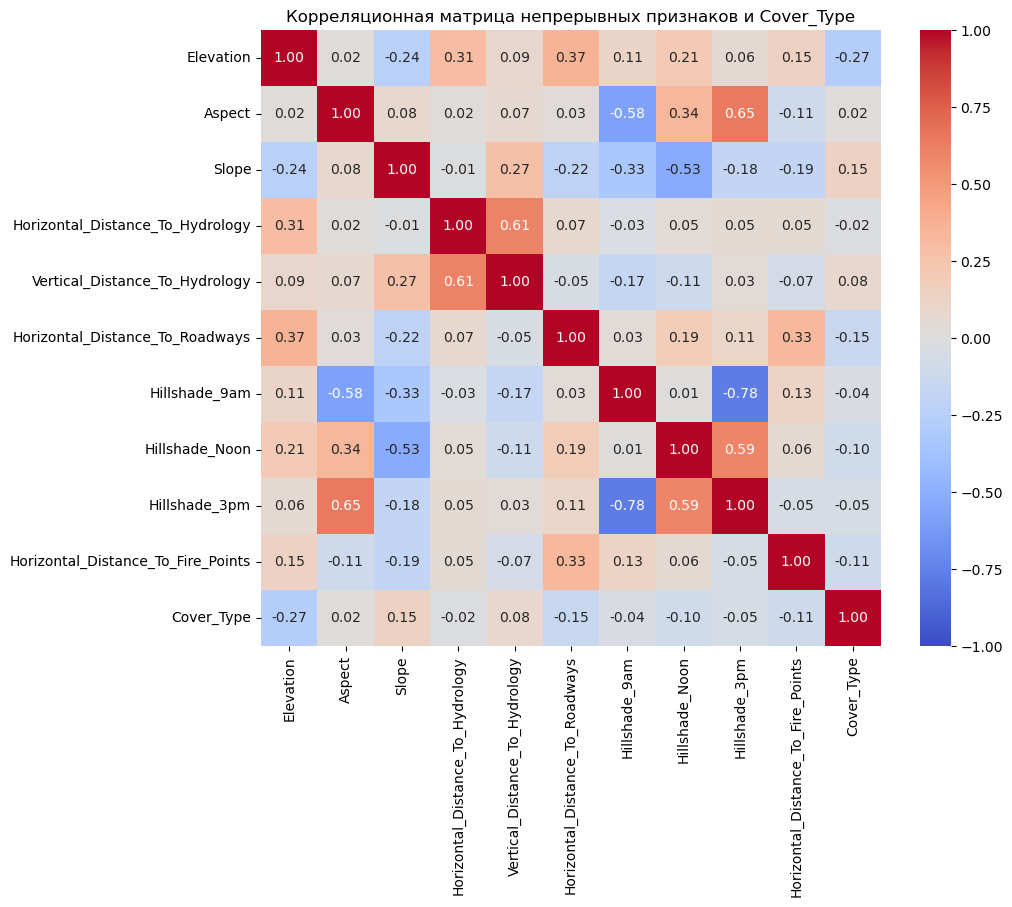


Рисунок 4 – Корреляционная матрица признаков

Наиболее сильную корреляцию с целевой переменной имеют признаки Elevation, Horizontal\_Distance\_To\_Roadways, Horizontal\_Distance\_To\_Fire\_Points. Большинство признаков имеют слабую корреляцию с Cover\_Type, что типично для реальных задач классификации. Между самими признаками есть как сильные, так и слабые взаимосвязи, но мультиколлинеарность выражена неярко.

Данные подходят для построения моделей машинного обучения, так как присутствуют информативные признаки и нет явных проблем с мультиколлинеарностью.

1. ОБРАБОТКА И ПРЕОБРАЗОВАНИЕ ПРИЗНАКОВ

Перед построением моделей обработаем датасет:

1. Проверим пропуски в данных – их нет
2. Масштабирование численных признаков.

5. ПОСТРОЕНИЕ И СРАВНЕНИЕ МОДЕЛЕЙ

Для решения задачи многоклассовой классификации типа лесного покрова были построены и обучены пять различных моделей::

1. Логистическая регрессия (LogisticRegression)
2. Дерево решений (DecisionTreeClassifier)
3. Случайный лес (RandomForestClassifier)
4. Градиентный бустинг (GradientBoostingClassifier)
5. Метод опорных векторов (SVM)

Все модели были обучены на одной и той же выборке (80% обучающая, 20% тестовая), и поначалу использовались базовые параметры без настройки.

По результатам базового обучения была получена следующая статистика:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Precision | F1-score |
| Логистическая регрессия | 0.72 | 0.71 | 0.71 |
| Дерево решений | 0.78 | 0.77 | 0.77 |
| Случайный лес | 0.72 | 0.81 | 0.71 |
| Градиентный бустинг | 0.73 | 0.75 | 0.73 |
| Метод опорных векторов | 0.403 | 0.33 | 0.354 |

1. НАСТРОЙКА ГИПЕРПАРАМЕТРОВ

После проведения GridSearchCV и повторной оценки моделей наблюдается улучшение производительности у всех алгоритмов:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Precision | F1-score |
| Логистическая регрессия | 0.723 | 0.71 | 0.71 |
| Дерево решений | 0.875 | 0.88 | 0.87 |
| Случайный лес | 0.852 | 0.86 | 0.85 |
| Градиентный бустинг | 0.860 | 0.86 | 0.86 |
| Метод опорных векторов | 0.706 | 0.70 | 0.69 |

Построим сравнительные гистограммы:

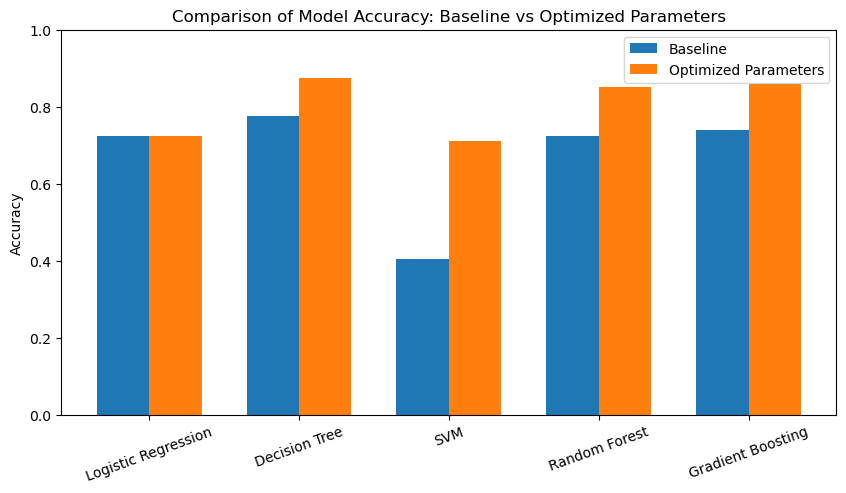
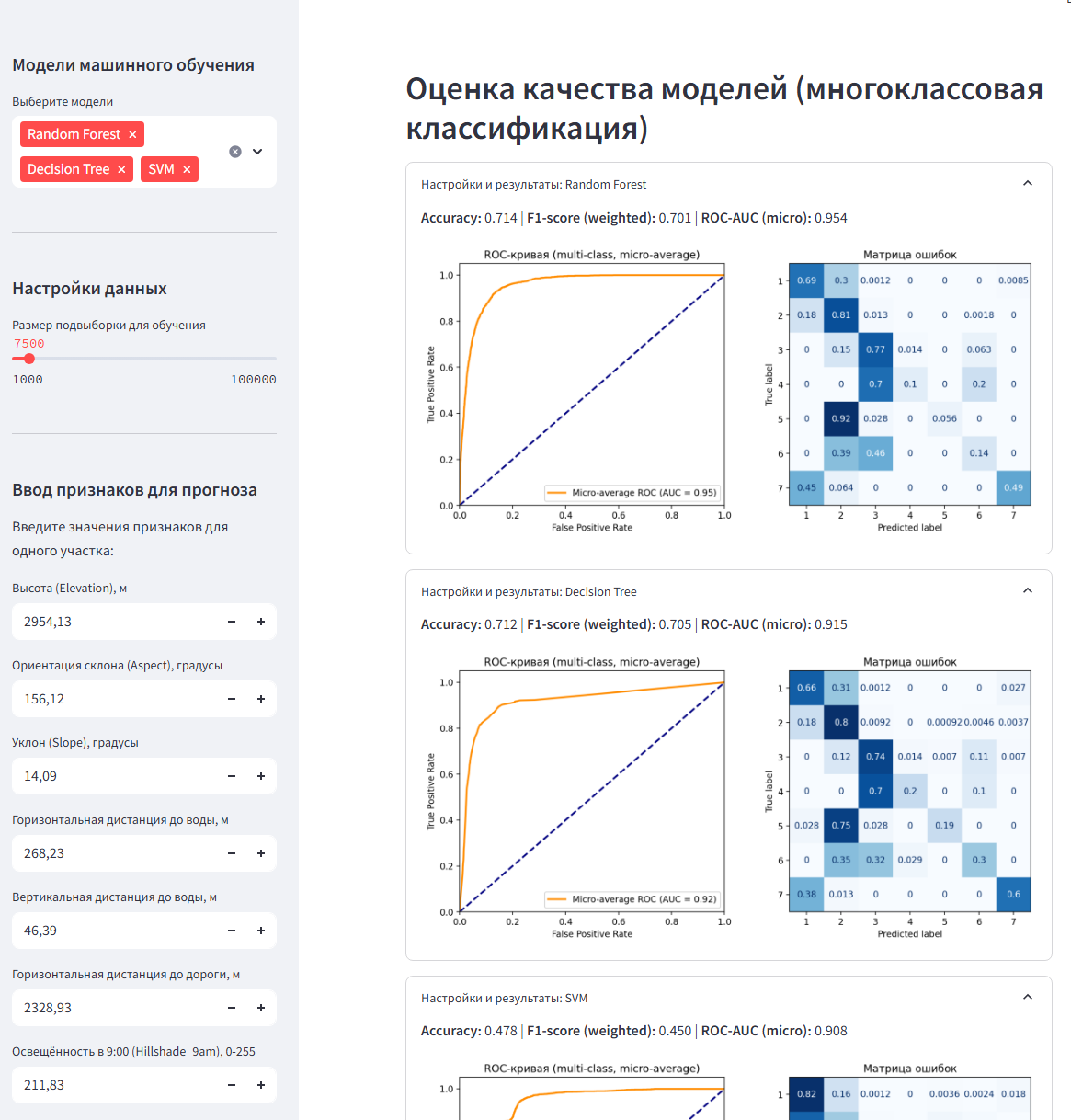


Рисунок 5 – Сравнение моделей по Accuracy

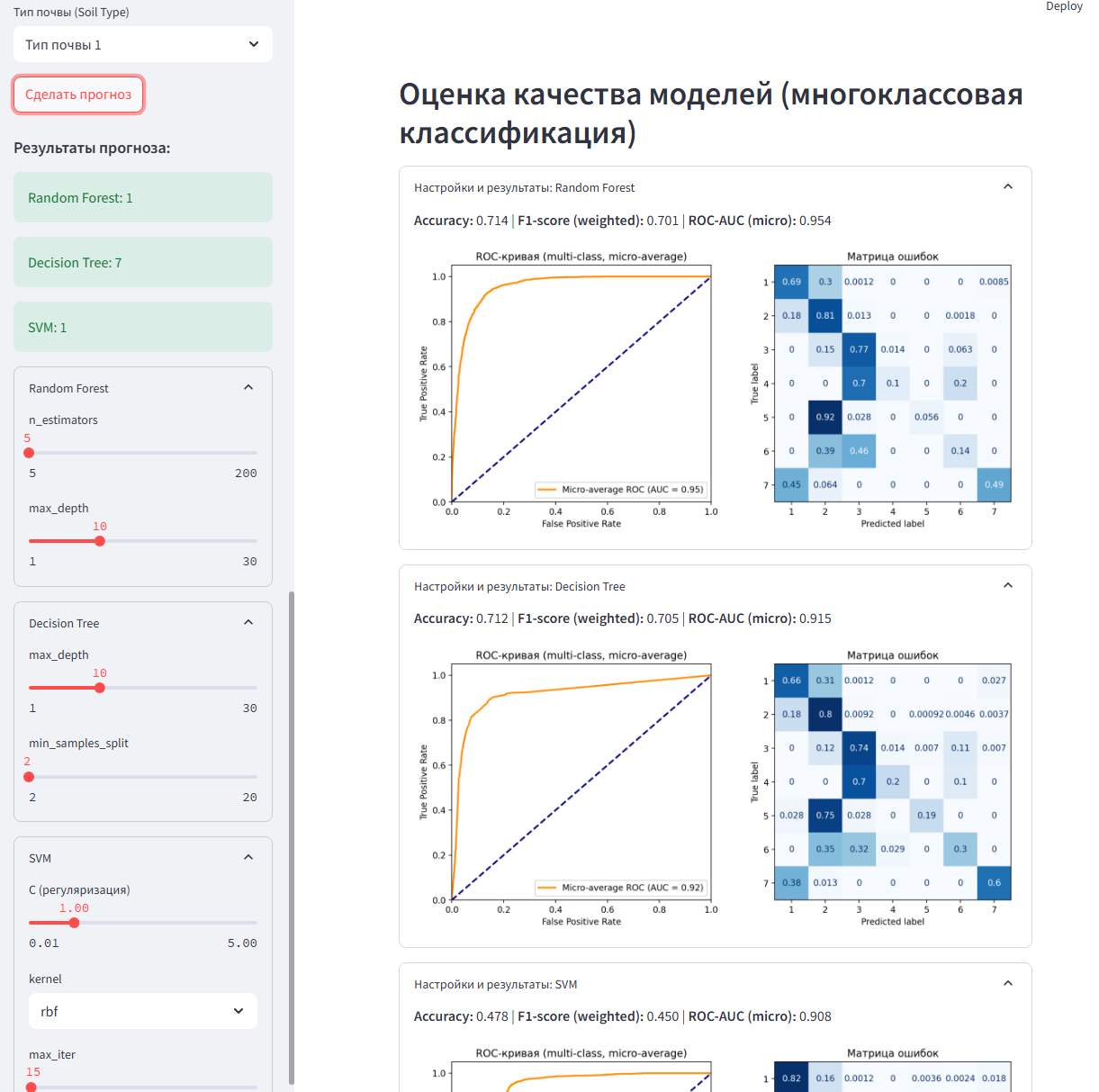
Наилучшие результаты по F1-мере, точности и полноте показала модель градиентного бустинга после настройки гиперпараметров. Будем использовать его для следующего этапа.

1. ВЕБ-ПРИЛОЖЕНИЕ

Реализуем веб-приложение для демонстрации влияния гиперпараметров на точность моделей. Используем фреймворк Streamlit.

****

**Рисунок 8 – Веб-приложение**

****

**Рисунок 9 – Веб-приложение. Предсказание и настройка гиперпараметров**

Изменяя положение параметров, будем получать разные предсказания моделей.

ЗАКЛЮЧЕНИЕ

Результаты проведённого исследования подтверждают, что системный и поэтапный подход к задаче многоклассовой классификации позволяет существенно повысить качество предсказаний. В рамках работы были реализованы все ключевые этапы современного машинного обучения: тщательная обработка и масштабирование признаков, анализ структуры данных, подбор и сравнение различных моделей, а также оптимизация их гиперпараметров.

Особое внимание было уделено анализу метрик качества (accuracy, precision, F1-score), что позволило объективно оценить эффективность каждого алгоритма. Наилучшие результаты показала модель дерева решений, достигнув F1-score выше 0,87 и продемонстрировав хорошую сбалансированность между точностью и полнотой. Это делает её оптимальным выбором для решения задачи классификации типа лесного покрова.

Разработанный код и полученные модели универсальны и могут быть легко адаптированы для других задач классификации. Кроме того, реализована возможность сохранения обученной модели и масштабировщика для дальнейшего использования в продуктивной среде, а также интеграция в веб-интерфейс с помощью Streamlit, что расширяет практическую применимость результатов работы.

**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. fetch\_covtype dataset (https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch\_covtype.html)
2. Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow — O'Reilly, 2019.
3. Документация Scikit-learn — <https://scikit-learn.org/>
4. Материалы курса "Машинное обучение", OpenAI, Stepik, Coursera
5. Python Software Foundation — <https://www.python.org/>
6. Визуализация и EDA: <https://seaborn.pydata.org/>, <https://matplotlib.org/>