|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **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»**

Отчет по лабораторной работе №5

**«Ансамбли моделей машинного обучения. Часть 1»**

по дисциплине «Технологии машинного обучения»

Выполнил:

студент группы ИУ5Ц-84Б

Падалко К.Р.

подпись, дата

Проверил:

к.т.н., доц., Ю.Е. Гапанюк

подпись, дата

2025 г.

**СОДЕРЖАНИЕ ОТЧЕТА**

[1. Цель лабораторной работы 3](#_Toc195640800)

[2. Задание 3](#_Toc195640801)

[3. Основные характеристики датасета 3](#_Toc195640802)

[4. Листинг 4](#_Toc195640803)

[4.1. Анализ датасета 4](#_Toc195640804)

[4.2. Описательная статистика 7](#_Toc195640805)

[4.3. Машинное обучение 8](#_Toc195640806)

[4.3.1. Разделение выборки 8](#_Toc195640807)

[4.3.2. Масштабирование данных 8](#_Toc195640808)

[4.3.3. Обучение модели 8](#_Toc195640809)

[4.3.3.1. Случайный лес 8](#_Toc195640810)

[4.3.3.2. Сверхслучайные деревья 8](#_Toc195640811)

[4.3.3.3. AdaBoost 8](#_Toc195640812)

[4.3.3.4. Градиентный бустинг 8](#_Toc195640813)

[4.4. Оценка качества моделей 9](#_Toc195640814)

[5. Вывод 10](#_Toc195640815)

## Цель лабораторной работы

Изучение ансамблей моделей машинного обучения.

## Задание

1. Выберите набор данных (датасет) для решения задачи классификации или регрессии.
2. В случае необходимости проведите удаление или заполнение пропусков и кодирование категориальных признаков.
3. С использованием метода train\_test\_split разделите выборку на обучающую и тестовую.
4. Обучите следующие ансамблевые модели:

* две модели группы бэггинга (бэггинг или случайный лес или сверхслучайные деревья);
* AdaBoost;
* градиентный бустинг.

1. Оцените качество моделей с помощью одной из подходящих для задачи метрик. Сравните качество полученных моделей.

## Основные характеристики датасета

Название датасета: Набор данных о видах ирисов.

Ссылка: https://www.kaggle.com/datasets/uciml/iris

**О датасетах**

Этот набор данных содержит информацию о различных аспектах ирисов (цветков) из трех видов: Setosa, Versicolor и Virginica. В наборе представлены характеристики, такие как длина и ширина чашелистика и лепестка для 150 образцов ирисов. Данные используются для классификации видов ирисов на основе этих характеристик.

Набор данных включает 150 строк, каждая из которых представляет один ирис, и 5 столбцов.

Этот датасет использован для задач классификации и обучения моделей машинного обучения, таких как k-ближайших соседей, дерева решений, логистической регрессии и других методов классификации.

**Структура данных**

sepal length (длина чашелистика) — измеряется в сантиметрах.

sepal width (ширина чашелистика) — измеряется в сантиметрах.

petal length (длина лепестка) — измеряется в сантиметрах.

petal width (ширина лепестка) — измеряется в сантиметрах.

species (вид) — категориальная переменная, указывающая на вид ириса, который представлен в строке (Setosa, Versicolor или Virginica).

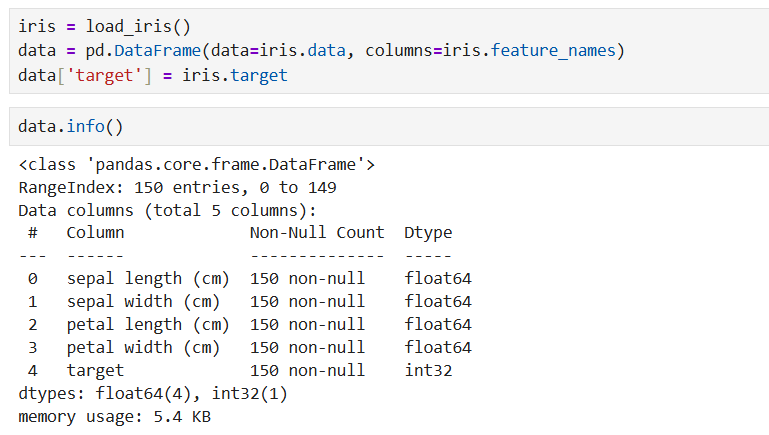
**Выбор признаков для машинного обучения**

Для машинного обучения выберем целевой признак – вид ирисов. Сопоставим с остальными признаками, а именно, характеристики цветов, вывялим примерный вид ириса.

## Листинг

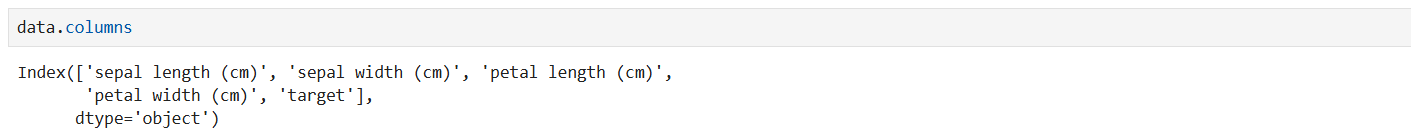
## Анализ датасета



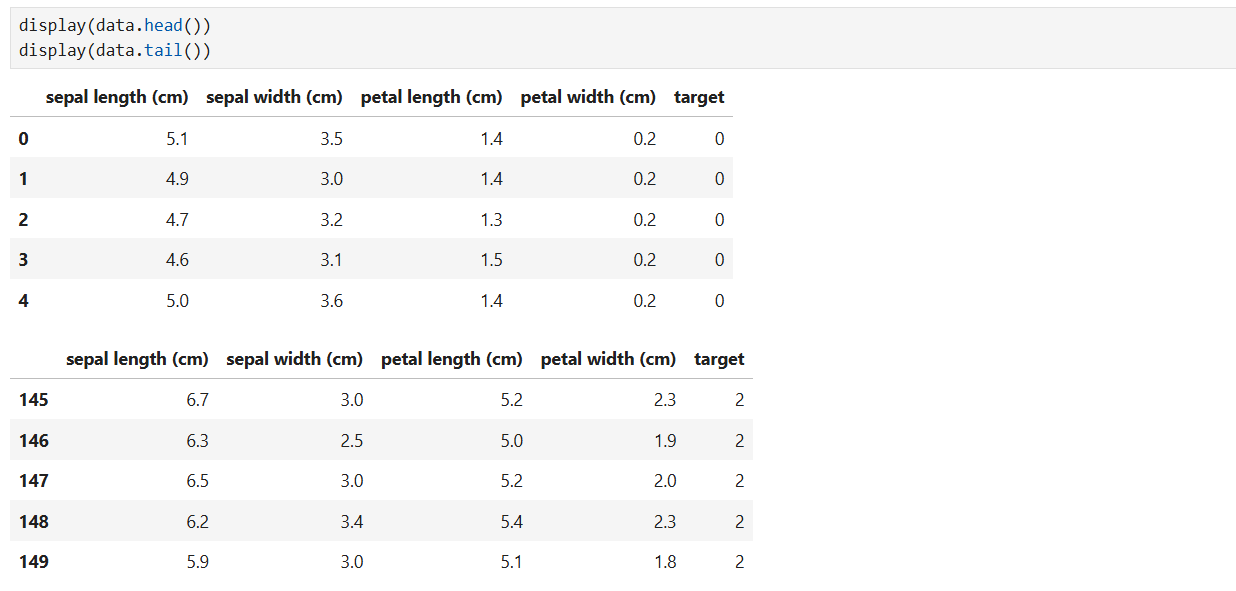


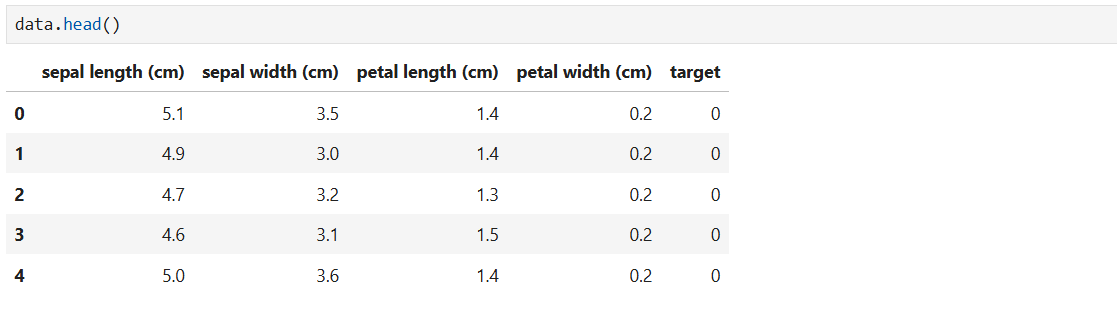
В датасете содержатся 150 строки, имеются 2 различные типы: int32 и float64.

Просмотр названий столбцов.

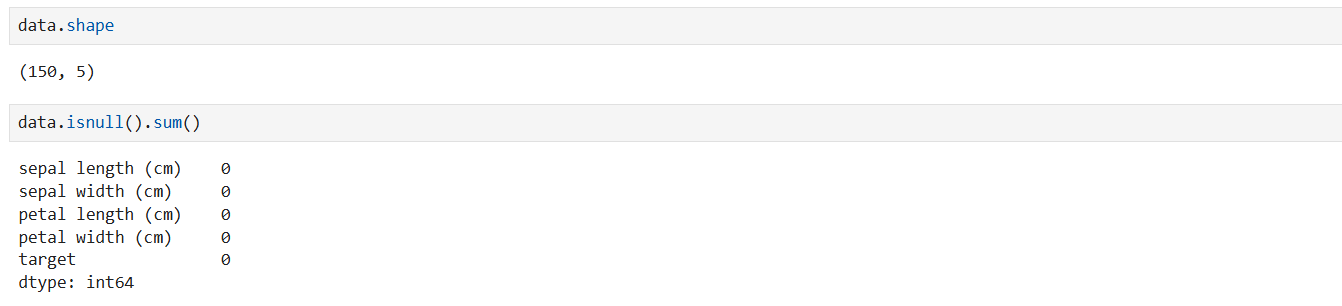


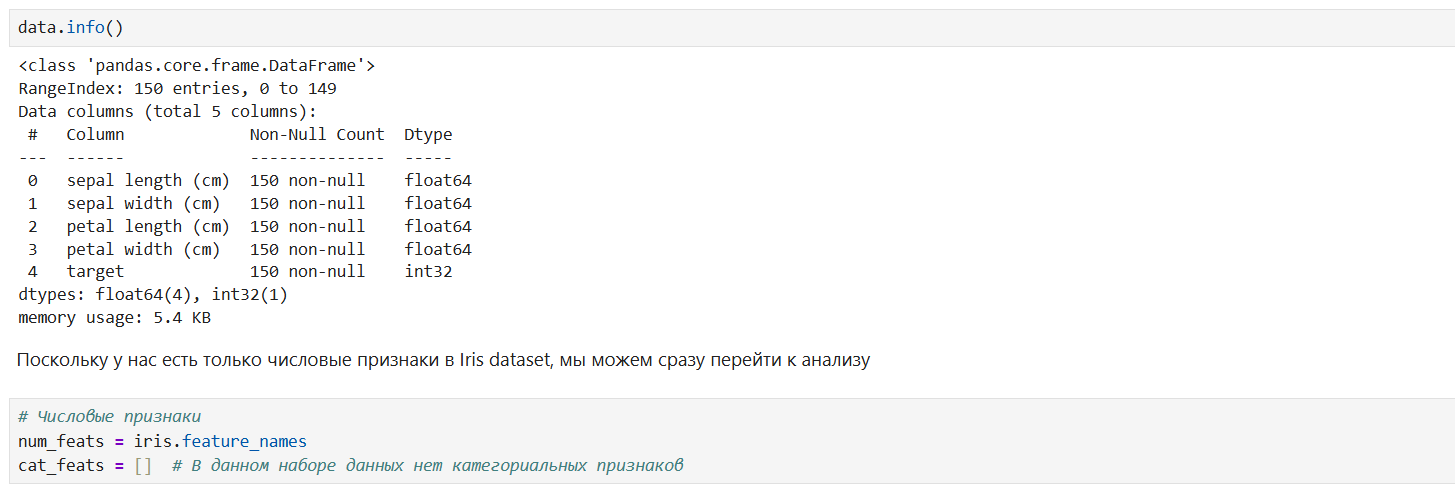
Первые и последние пять строк датасета.

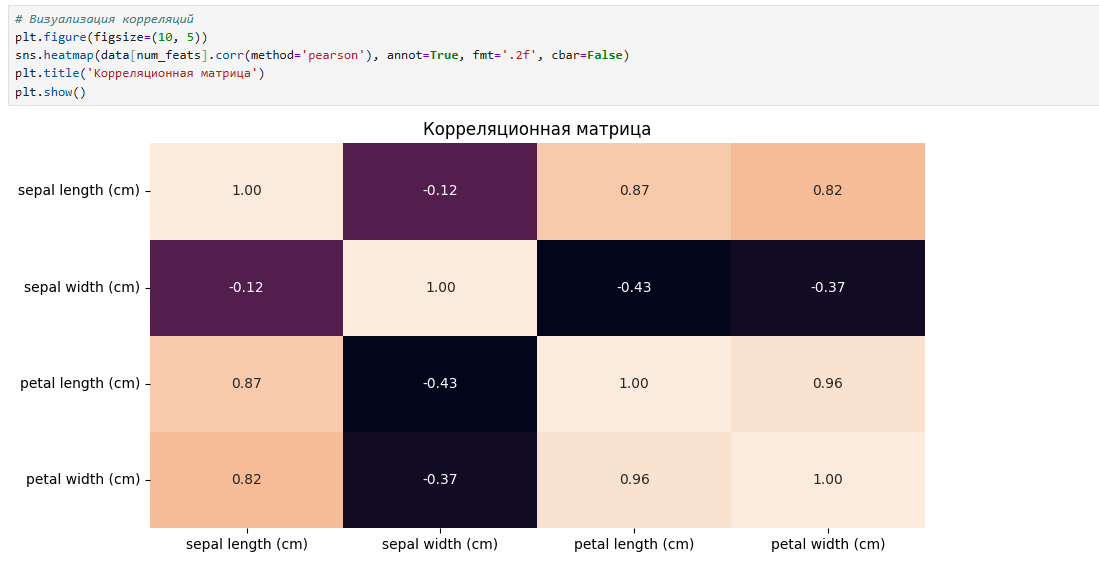




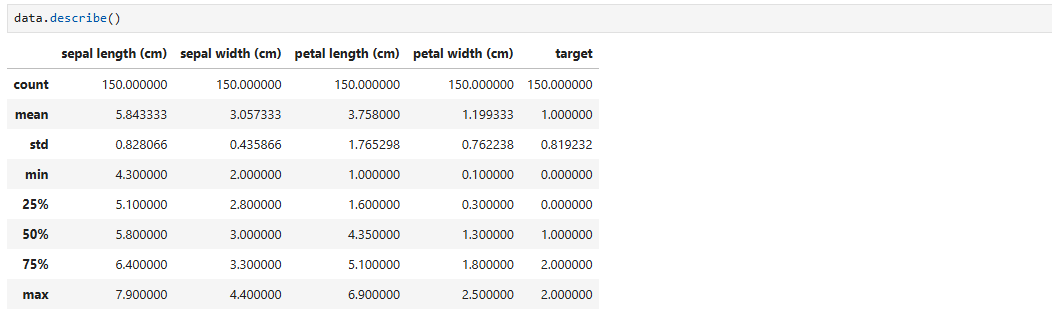
Количество строк и столбцов.







## Описательная статистика



**id:** Это уникальные идентификаторы записей в данных. Каждая строка представляет отдельный образец ириса, и этот столбец не несет дополнительной информации о характеристиках цветов. Он служит лишь для идентификации строки в наборе данных.

**sepallengthcm:** Это длина чашелистика цветка ириса, измеренная в сантиметрах. Среднее значение длины чашелистика составляет 5.84 см. Диапазон значений от 4.3 см до 7.9 см, что указывает на разнообразие в длине чашелистика среди разных видов ирисов.

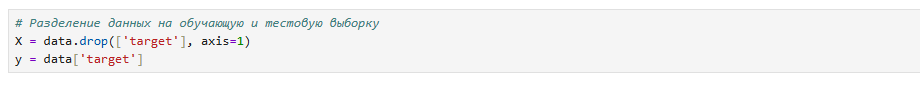
**sepalwidthcm:** Это ширина чашелистика цветка ириса, измеренная в сантиметрах. Средняя ширина чашелистика составляет 3.05 см. Значения варьируются от 2.0 см до 4.4 см, что показывает, что ширина чашелистика также имеет значительные колебания среди ирисов.

**petallengthcm:** Это длина лепестка цветка ириса, измеренная в сантиметрах. Средняя длина лепестка составляет 3.76 см. Длина лепестков варьируется от 1.0 см до 6.9 см, с большими различиями между образцами, что может указывать на разнообразие форм лепестков в зависимости от вида ириса.

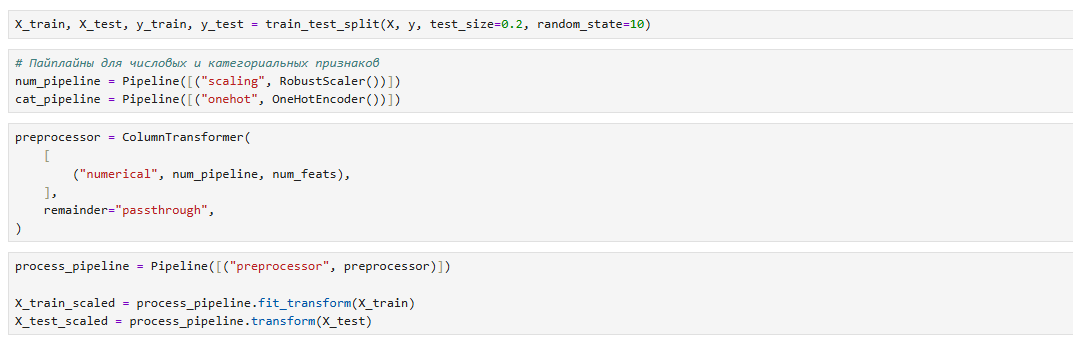
**petalwidthcm:** Это ширина лепестка цветка ириса, измеренная в сантиметрах. Средняя ширина лепестка составляет 1.20 см. Значения варьируются от 0.1 см до 2.5 см, что также указывает на значительный разброс в характеристиках лепестков среди разных видов ирисов.

## Машинное обучение

## Разделение выборки



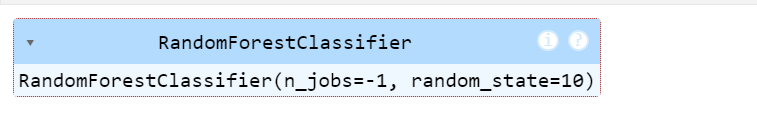
## Масштабирование данных



## Обучение модели

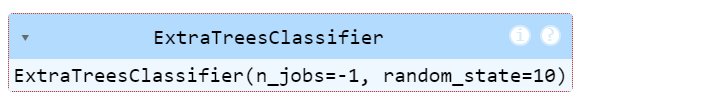
### 4.3.3.1. Случайный лес





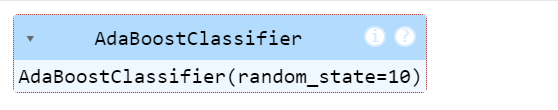
### 4.3.3.2. Сверхслучайные деревья





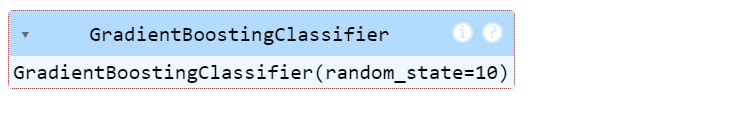
### 4.3.3.3. AdaBoost



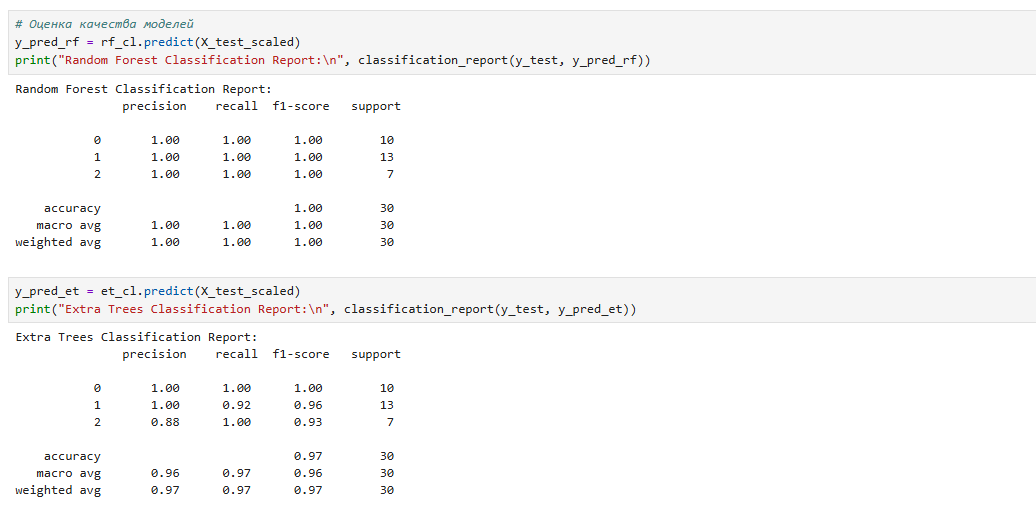
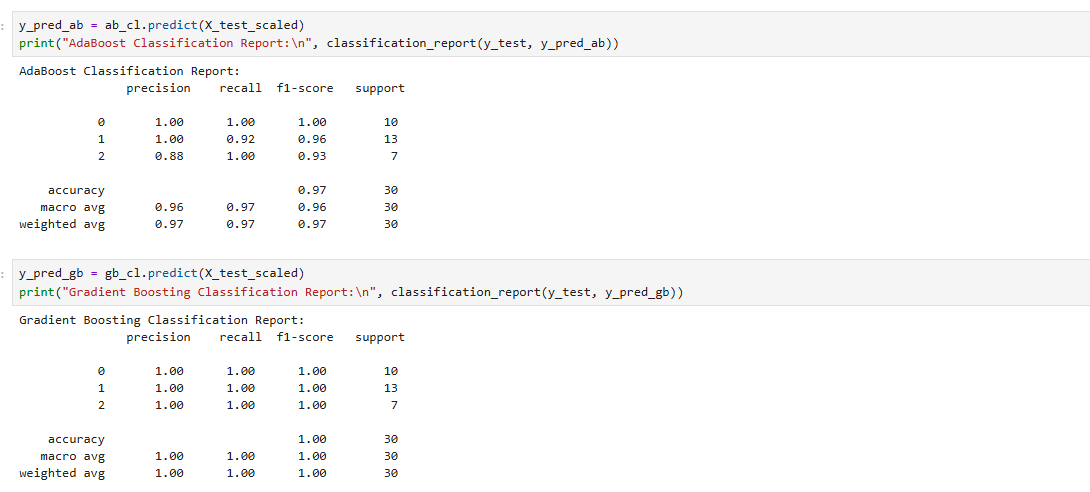
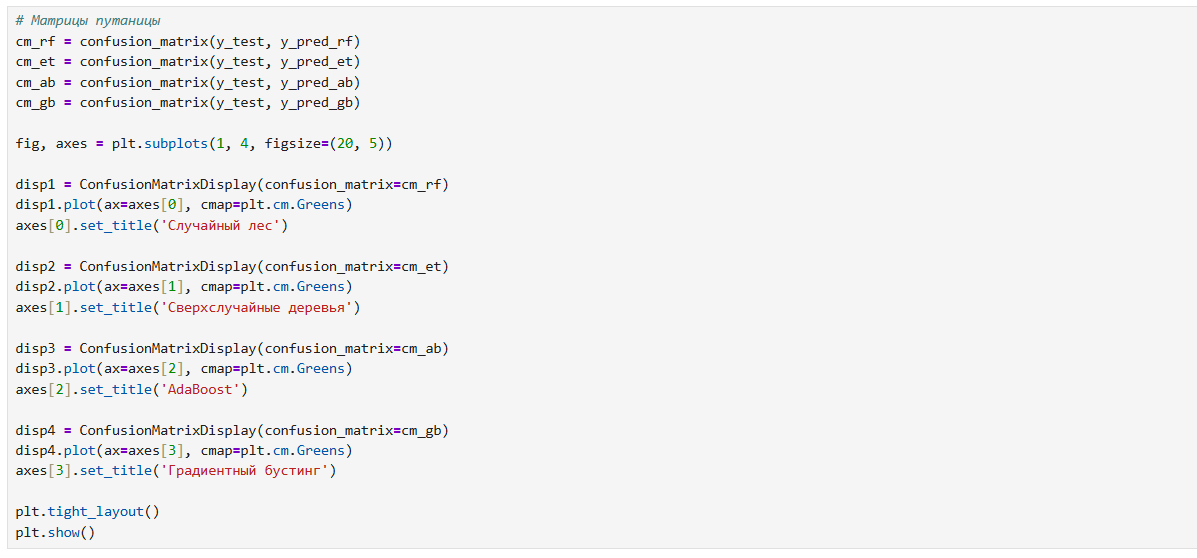
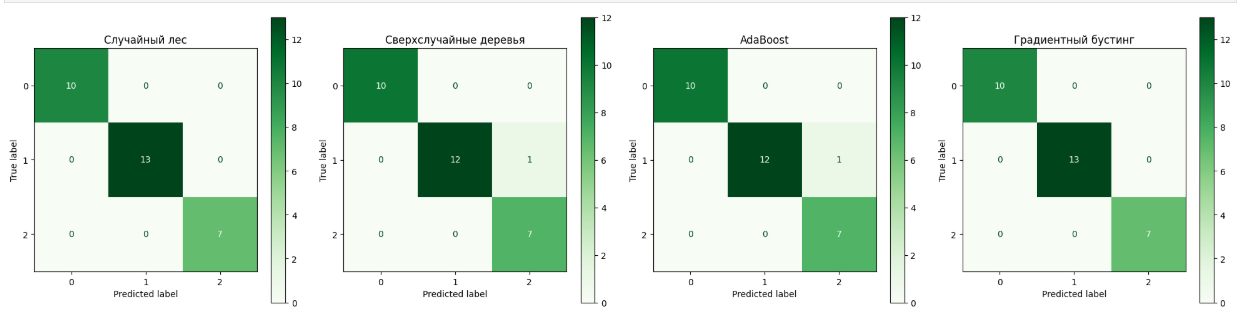


### 4.3.3.4. Градиентный бустинг





## Оценка качества моделей

## Вывод

В ходе лабораторной работы изучили ансамбли моделей машинного обучения.