Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

*Дисциплина «Технологии машинного обучения»*

**Отчёт**

по лабораторной работе №6

«Ансамбли моделей машинного обучения»

*Вариант 12*

Студент:

Крюков Г. М.

Группа ИУ5-61Б

Преподаватель:

Гапанюк Ю. Е.

Москва, 2020 г.

**Цель лабораторной работы:**

Изучение ансамблей моделей машинного обучения.

**Задание:**

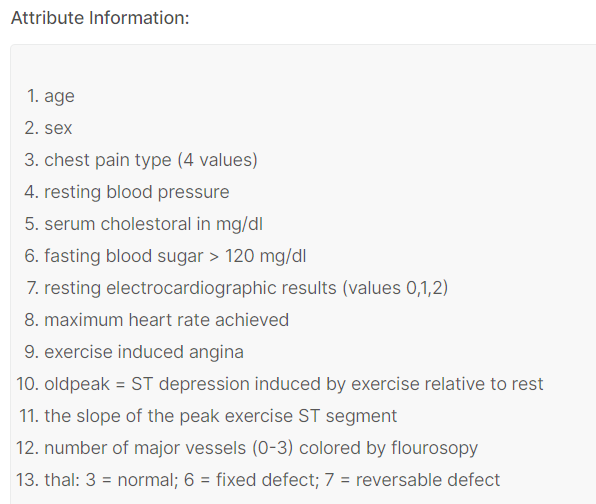
1. Выберите набор данных (датасет) для решения задачи классификации или регресии.
2. В случае необходимости проведите удаление или заполнение пропусков и кодирование категориальных признаков.
3. С использованием метода train\_test\_split разделите выборку на обучающую и тестовую.
4. Обучите две ансамблевые модели. Оцените качество моделей с помощью одной из подходящих для задачи метрик. Сравните качество полученных моделей.

**Выполнение работы:**

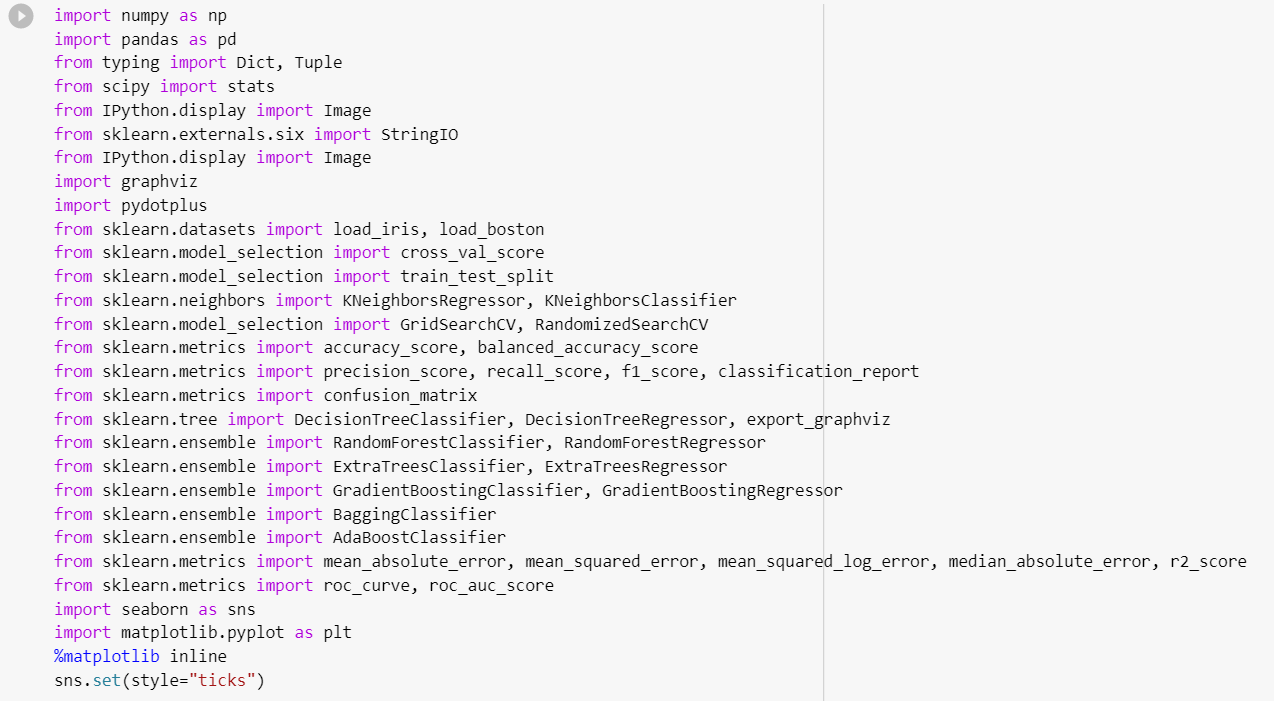
Выбранный датасет:

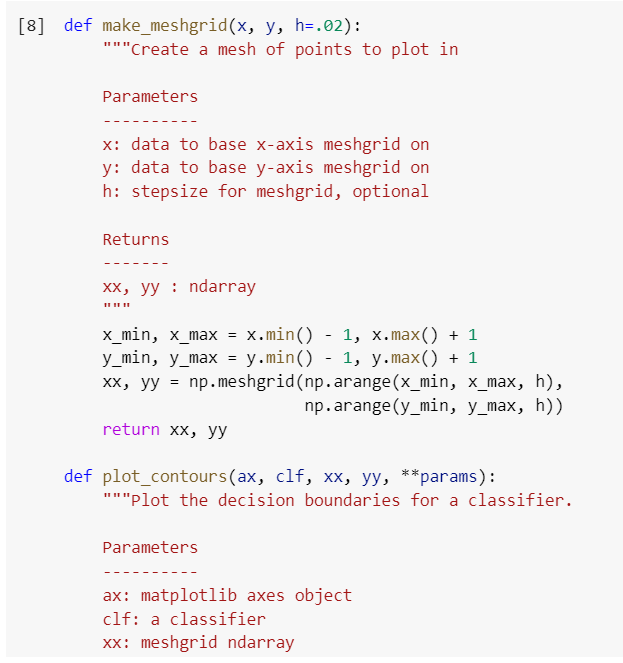
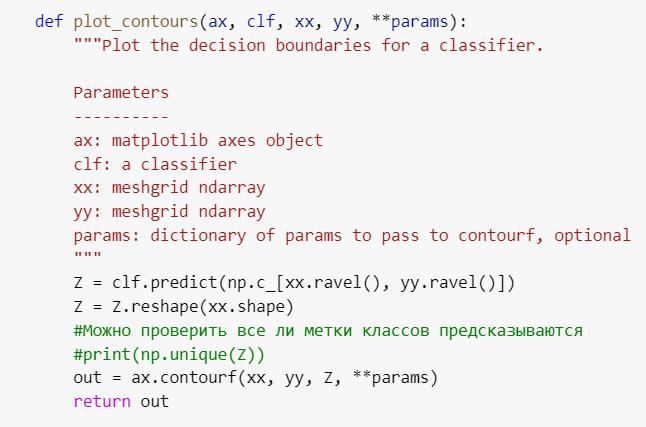
[https://www.kaggle.com/ronitf/heart-disease-uci/data#](https://www.kaggle.com/ronitf/heart-disease-uci/data)

Датасет содержит 76 атрибутов, но все опубликованные эксперименты относятся к использованию подмножества из 14 из них. Поле "цель" относится к наличию у пациента сердечно-сосудистых заболеваний. Это целочисленное значение от 0 (отсутствие присутствия) до 4.

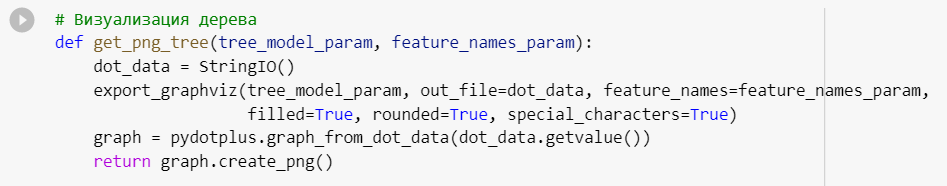


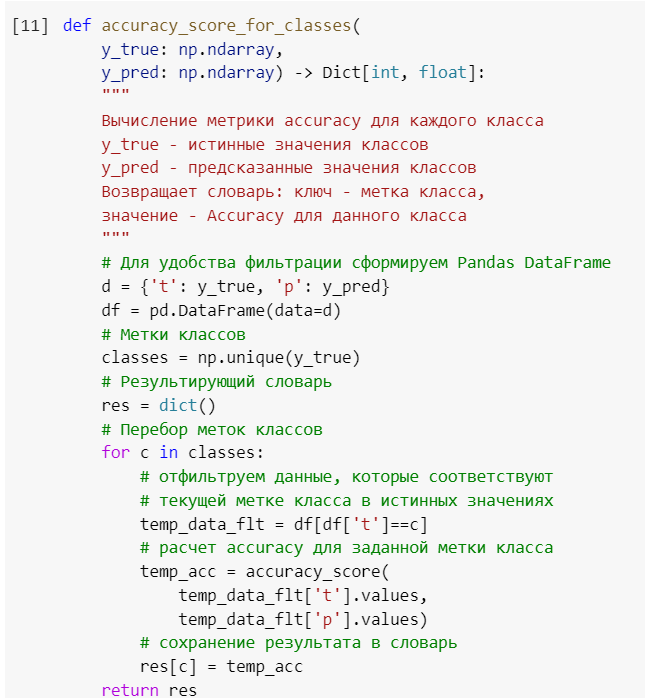
Подготовка данных:

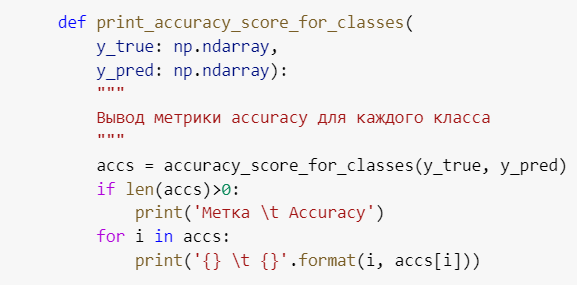


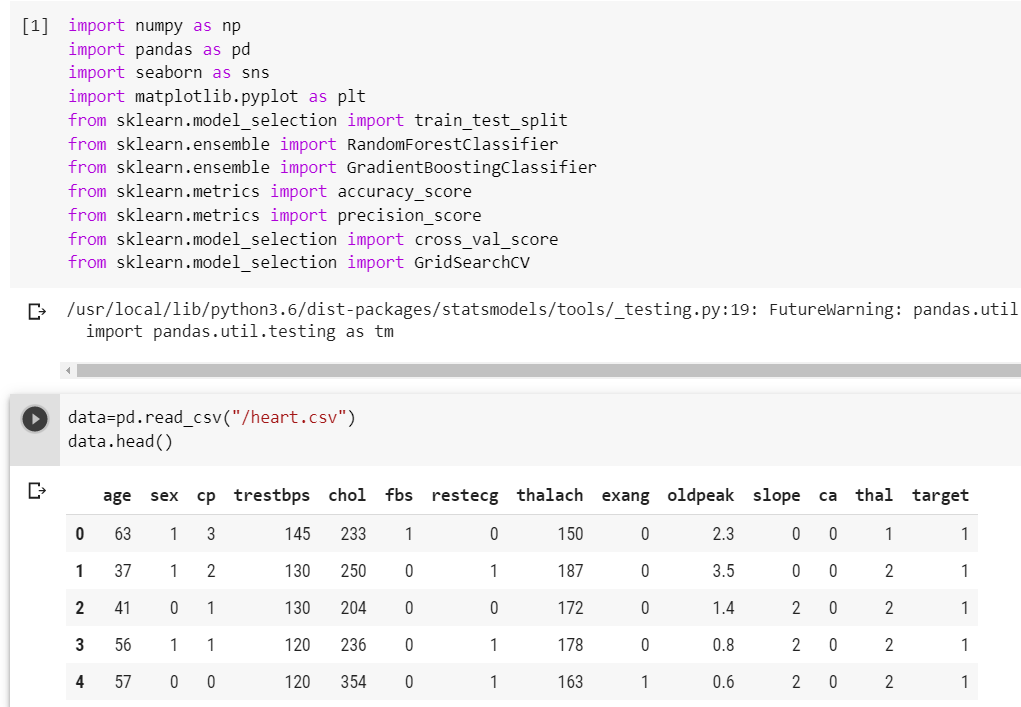




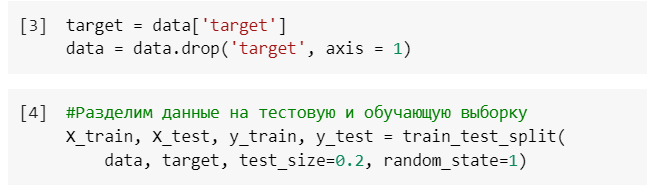




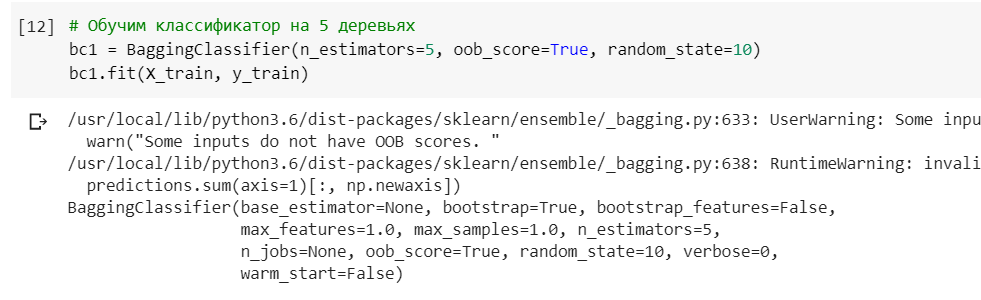
Загрузка данных

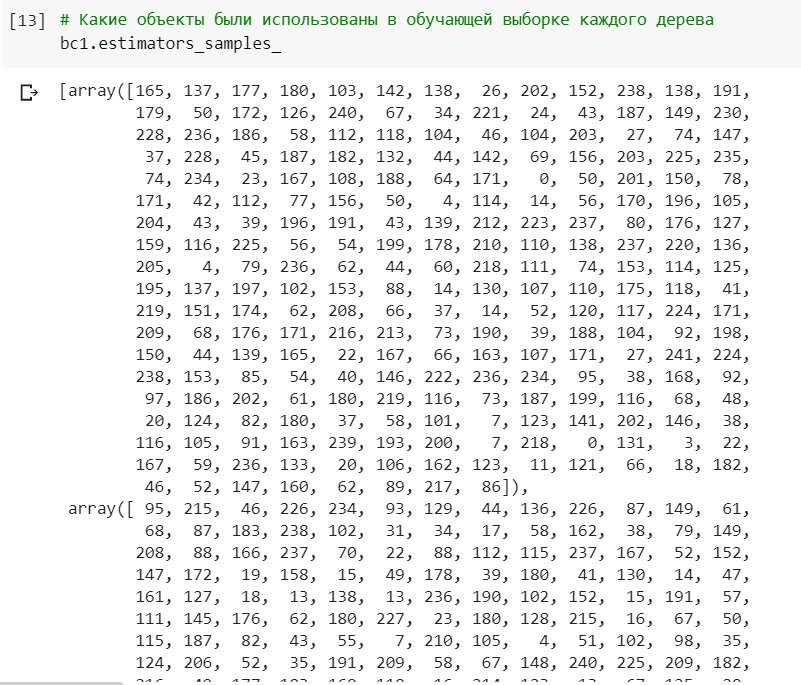


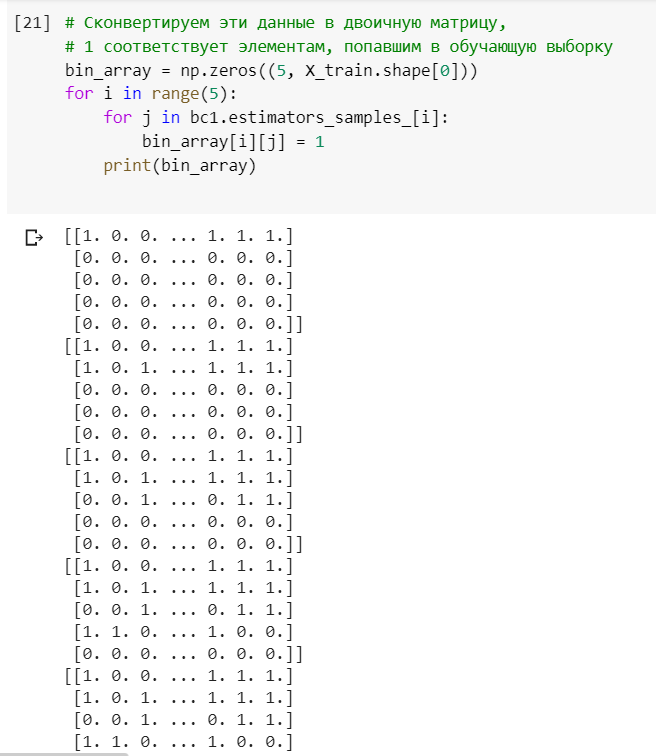
Разделение данных на обучающую и тестовую выборки



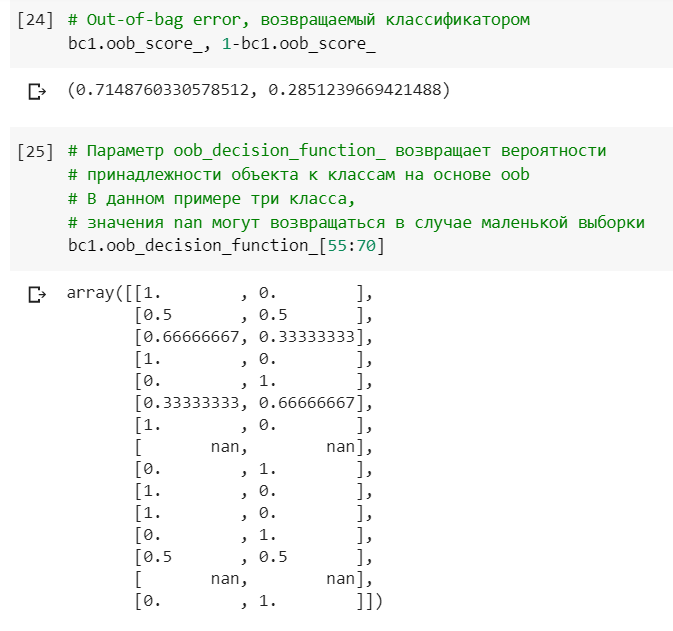
1. **Бэггинг**



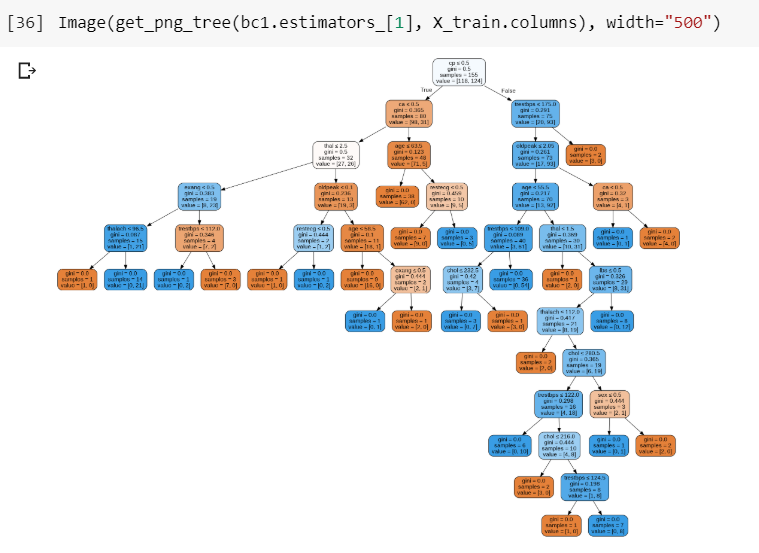


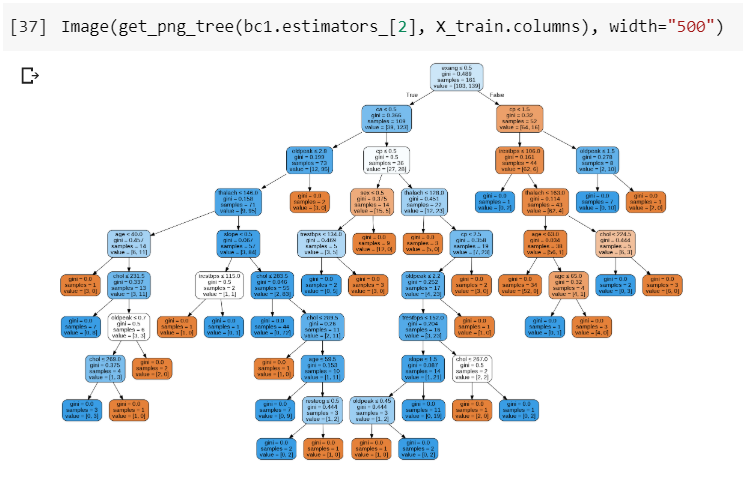


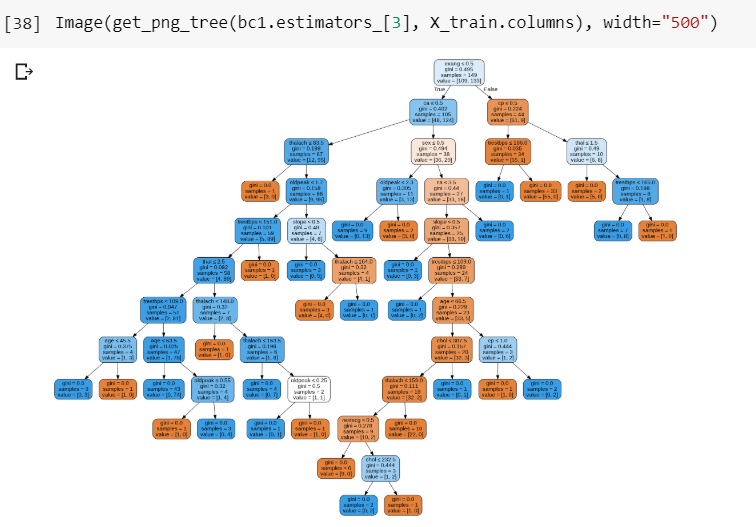


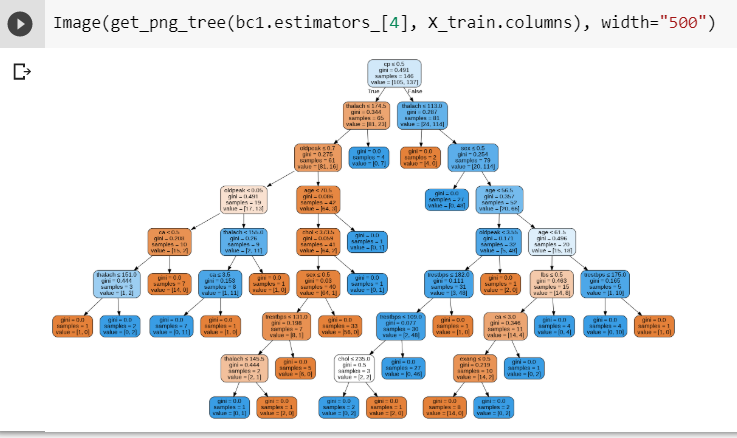






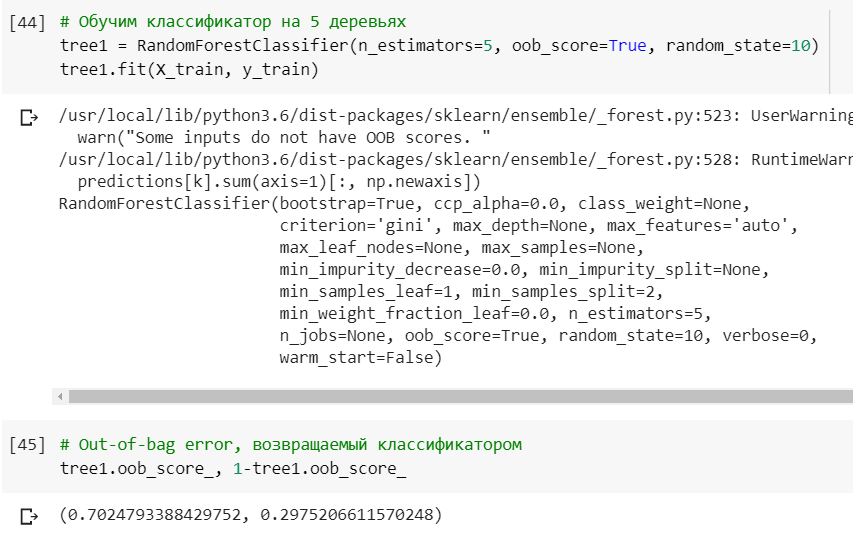


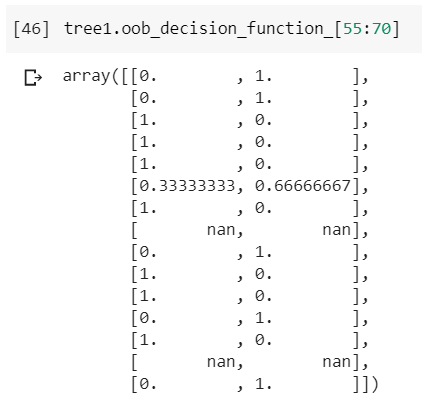


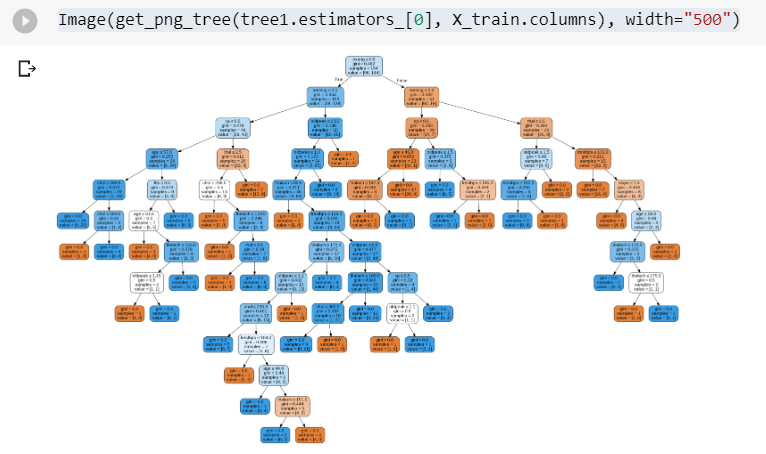


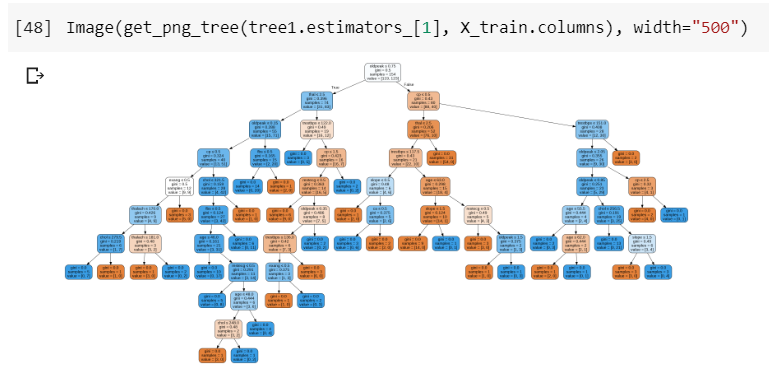
Деревья получаются различными. Таким образом, каждое дерево работает как "слабая модель".

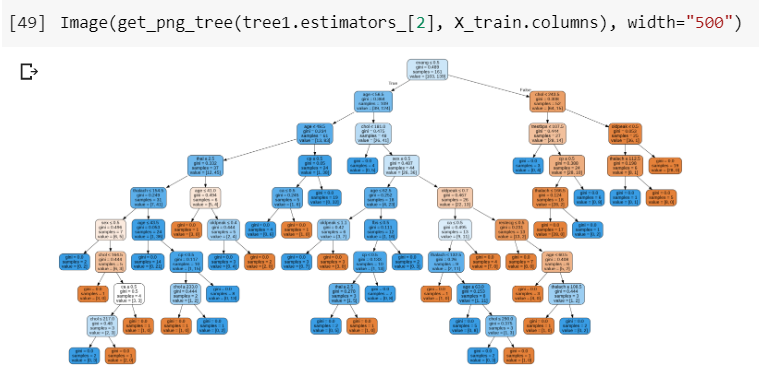
1. **Случайный лес**

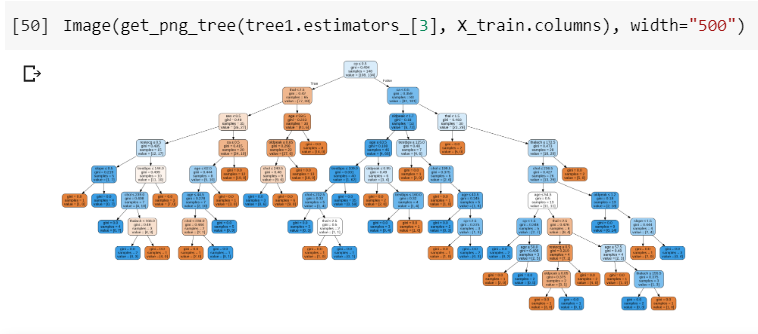


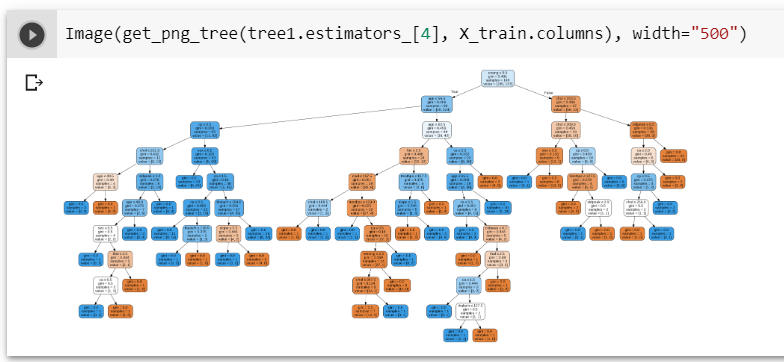












В случае случайного леса деревья получаются более разнообразными, чем в случае бэггинга.

**Сравнение моделей**

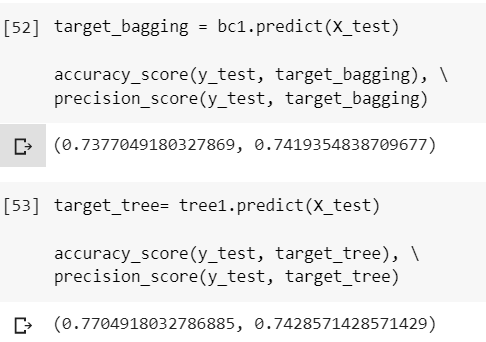
Метрики качества классификации

1) [Accuracy](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html#sklearn.metrics.accuracy_score)

Метрика вычисляет процент (долю в диапазоне от 0 до 1) правильно определенных классов.

#### 2) Метрика precision:

Доля верно предсказанных классификатором положительных объектов, из всех объектов, которые классификатор верно или неверно определил как положительные.



Вывод:

Можем видеть, что обе модели по рассматриваемым метрикам показали приемлемый результат, но модель случайного леса оказалась результативнее, получив меньший процент ошибок.