Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(национальный исследовательский университет)
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

ОТЧЁТ ПО ЗАДАНИЮ №7

по дисциплине «Интеллектуальный анализ больших данных»

Тема: Ансамблевая классификация с помощью бустинга

Выполнил
студент группы КЭ-120
Глизница Максим Николаевич
E-mail: letadllo@mail.ru

Челябинск 2021

1. Задание

Выполните классификацию набора данных из задания 3 с помощью бустинга, варьируя количество участников ансамбля (от 50 до 100 с шагом 10).

Вычислите показатели качества классификации: аккуратность (accuracy), точность (precision), полнота (recall), F-мера. Выполните визуализацию полученных результатов в виде диаграмм. Нанесите на диаграммы соответствующие значения, полученные в заданиях 3, 4, 5, 6.

2. Краткие сведения о наборах данных

Использованный набор данных:

Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009). Содержит информацию о произведённых партиях красного вина "Vinho Verde" (такую как pH и содержание хлоридов), а также оценку качества от 1 до 10 (реально в наборе данных присутствуют значения от 3 до 8). Как предлагается в пояснении к набору данных, высококачественным можно считать вино с оценкой 7 и выше. Таким образом, в наборе данных присутствуют 217 партий высокого качества и 1382 партий низкого качества.

3. Краткие сведения о средствах реализации

Для реализации методов была использована библиотека scikit-learn, включающая в себя множество алгоритмов для анализа данных.

Репозиторий по дисциплине: https://github.com/Airpllane/DAAlgorithms.

Каталог для задания: 7. Boosting.

4. Визуализация показателей качества

Для визуализации были использованы количества участников ансамбля от 50 до 100 с шагом 10. Был использован классификатор GradientBoostingClassifier, реализующий алгоритм классификации с помощью градиентного бустинга.

Были вычислены метрики и выполнена их визуализация на диаграмме. На диаграмму были также нанесены метрики, полученные в результате выполнения предыдущих заданий. Результат приведён на рис. 1.

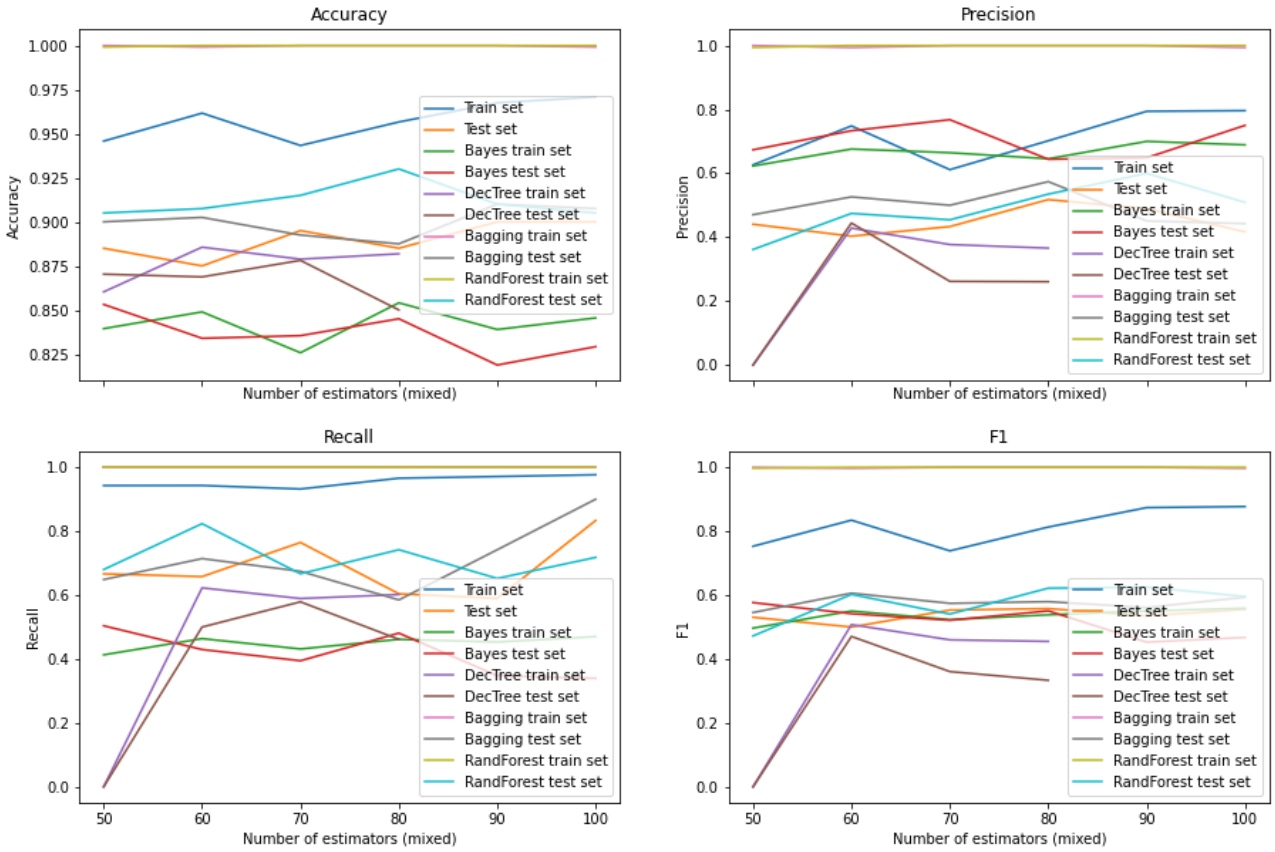

Рис. 1. Результаты визуализации деревьев

Из рисунка можно увидеть, что классификация с помощью градиентного бустинга, в отличие от двух предыдущих заданий, не показывает 100% результатов на тренировочном наборе данных, но при этом показывает очень близкие результаты на тестовых данных. Если сравнить градиентный бустинг с двумя предыдущими алгоритмами, то можно увидеть, что его метрика recall имеет почти такое же значение, как у случайного леса, но при этом случайный лес имеет несколько более высокую точность, из чего можно заключить, что случайный лес подходит для данной задачи лучше, чем бустинг.