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 «Системное программирование»

ОТЧЁТ

по лабораторной работе №4

на тему «Классификация с помощью дерева решений»

Выполнил
Студент группы КЭ-120
_____ Д.А. Снегирева
«____»_____2020 г.
Email: dashasneg@mail.ru

Челябинск 2020

ЗАДАНИЕ

Выполните классификацию набора данных из задания 3 с помощью построения дерева решений, фиксируя критерий выбора атрибута разбиения (information gain, gain ratio, index gini) и варьируя соотношение мощностей обучающей и тестовой выборок (от 60%:40% до 90%:10% с шагом 10%). Выполните визуализацию построенных деревьев решений. Вычислите показатели качества классификации: аккуратность (accuracy), точность (precision), полнота (recall), F-мера. Выполните визуализацию полученных результатов в виде диаграмм.

СОДЕРЖАНИЕ

1 КРАТКИЕ СВЕДЕНИЯ О НАБОРАХ ДАННЫХ И СРЕДСТВАХ РЕАЛИЗАЦИИ

В данной работе был использован набор данных Wine Data Set (https://archive.ics.uci.edu/ml/datasets/wine), содержащий результаты химического анализа вин, выращенных в одном и том же регионе Италии, и состоящий из 13 атрибутов.

В качестве одного из средств реализации была использована библиотека scikit-learn, простое средство для анализа данных.

Репозиторий задания: https://github.com/DasHaSneg/BigDataMiningCourse

Каталог задания: 4 decision_tree

2 ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ

При определении классификатора был задан параметр criterion равным gini.

Для визуализации были вычислены показатели качества классификации: аккуратность (accuracy), точность (precision), полнота (recall) и F-мера для разных значений соотношений мощностей обучающей и тестовой выборок (от 60%:40% до 90%:10% с шагом 5%). Для каждого показателя были построены диаграммы, которые представлены на рис. 1.

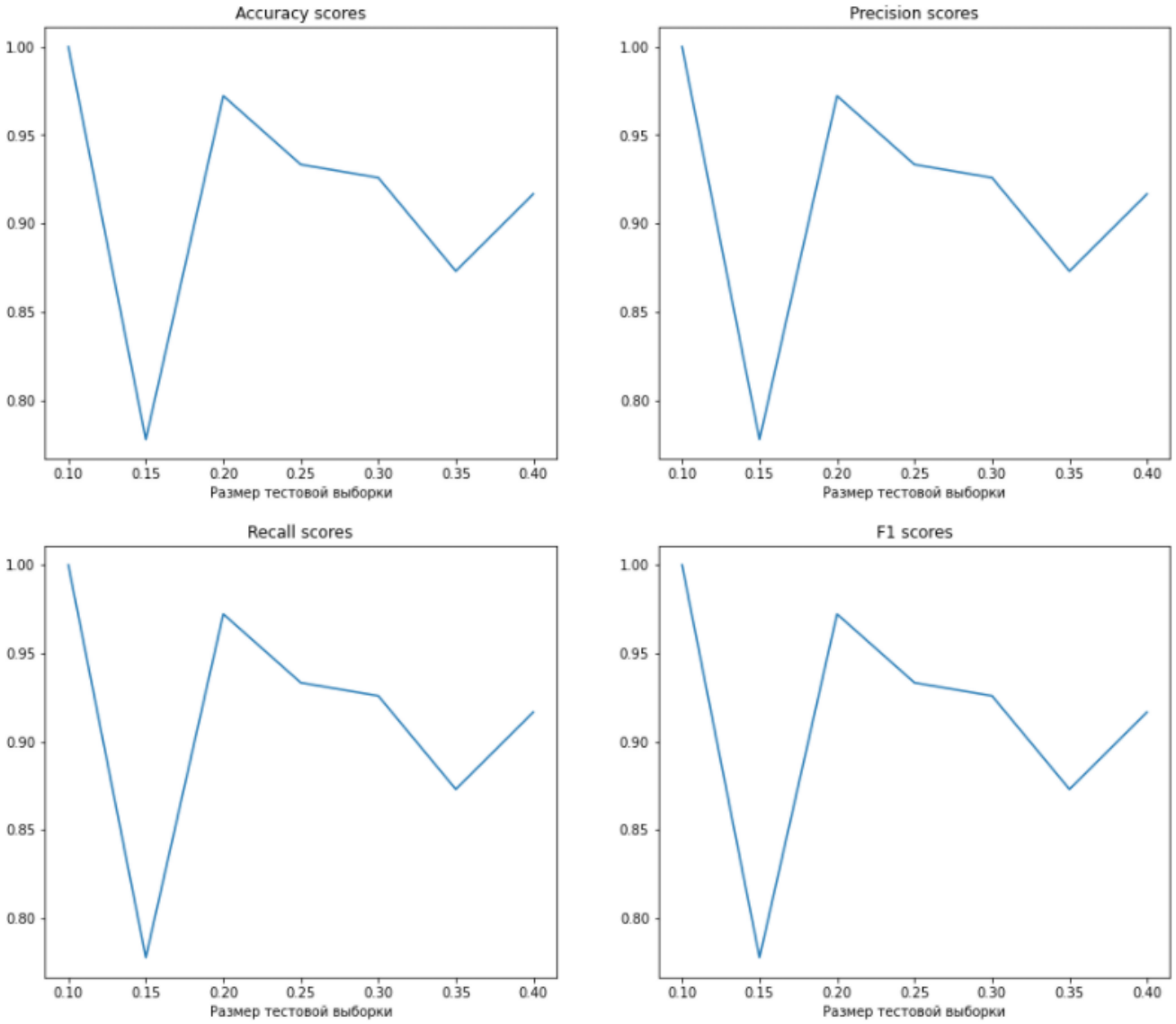

Рис. 1. Результаты

Результаты для каждой метрики идентичны. Исходя из значений, можно сделать вывод, что лучшее достигается при соотношении мощностей обучающей и тестовой выборок 90%:10%.

Кроме того, после получения результатов были сразу построены деревья решения, которые представлены на рис. 2 и 3.

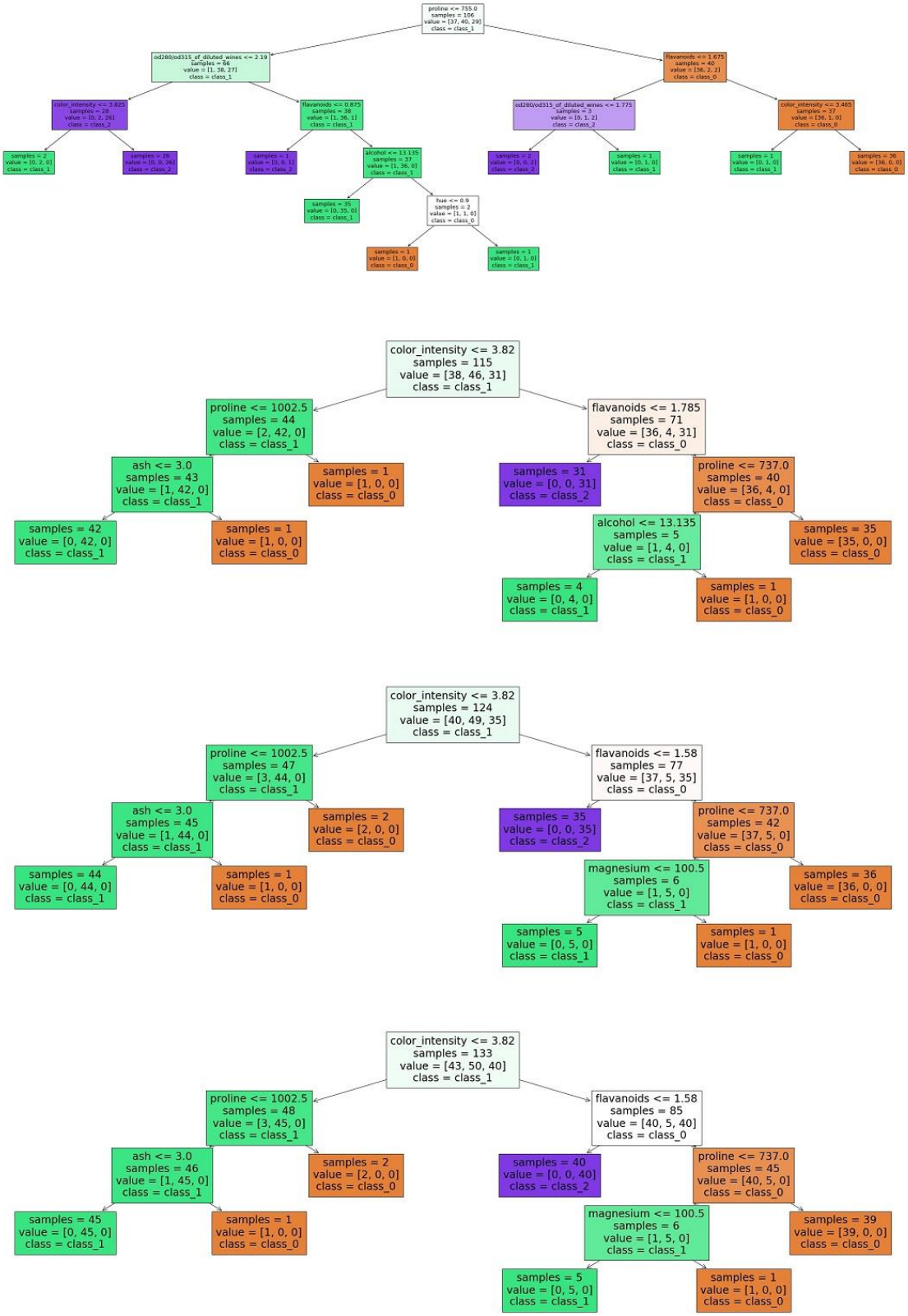

Рис. 2. Первые 4 дерева

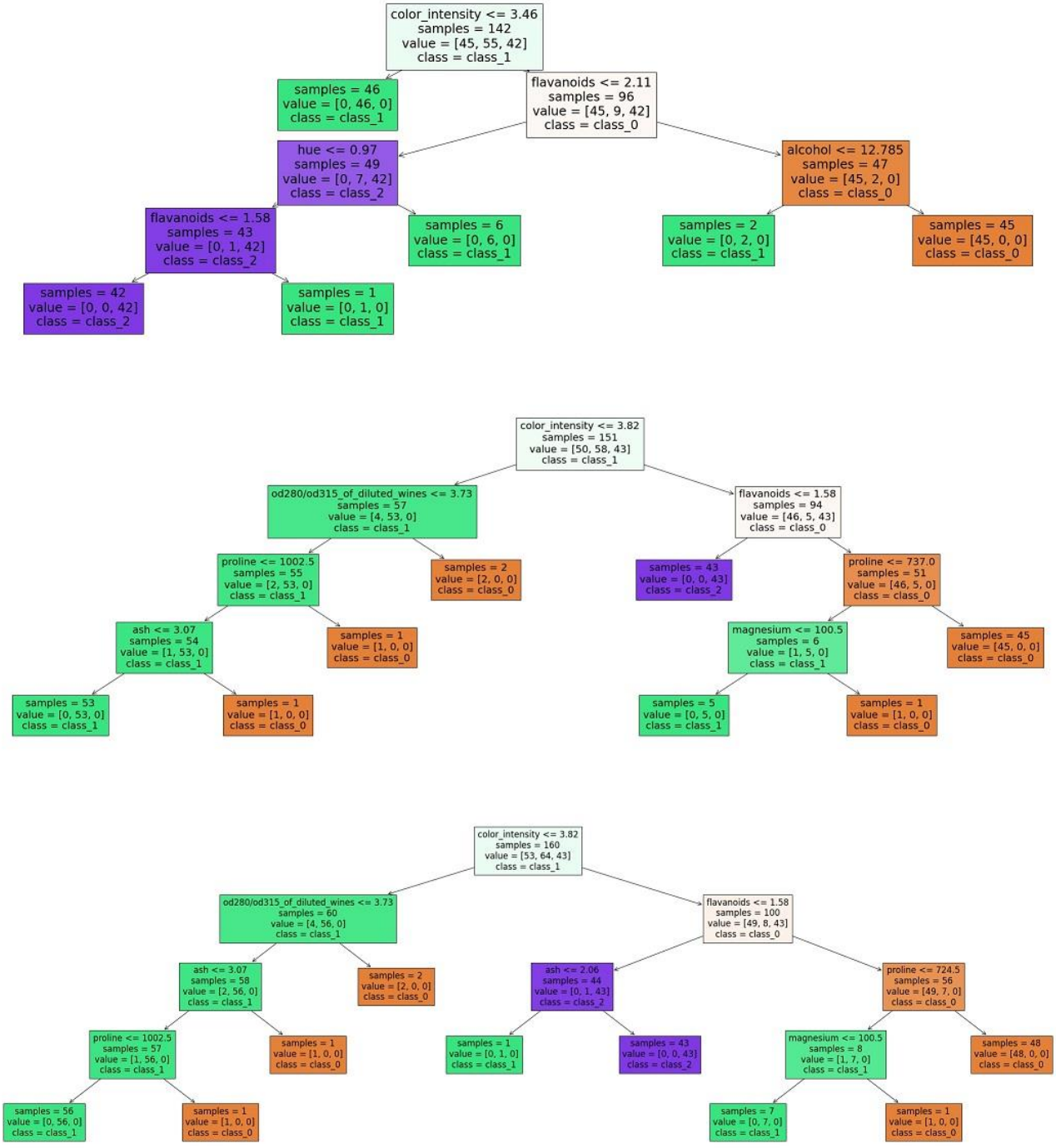

Рис. 2. Оставшиеся 3 дерева

Исходя из рисунков можно сделать вывод, что деревья используют разные признаки при построении.