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 №4**

по дисциплине «Интеллектуальный анализ больших данных»

Тема: Классификация с помощью дерева решений

Выполнил

студент группы КЭ-120

Глизница Максим Николаевич

E-mail: letadllo@mail.ru

Челябинск 2021

# 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.

# Краткие сведения о наборах данных

Использованный набор данных:

Mushroom Classification (<https://www.kaggle.com/uciml/mushroom-classification>). Содержит информацию о типичных признаках определённых видов грибов, таких как форма шляпки, запах, и других. Также включает информацию о том, является ли данный вид гриба ядовитым. Расшифровка используемых в файле сокращений находится в файле attr.txt. Набор содержит 3916 ядовитых и 4208 неядовитых грибов

# 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: 4. DecTree.

# Визуализация показателей качества

Для визуализации были использованы соотношения мощностей обучающей и тестовой выборки от 60%:40% до 90%:10% с шагом 10%. Был использован классификатор DecisionTreeClassifier, реализующий алгоритм классификации с помощью дерева решений. В качестве критерия выбора атрибута разбиения была выбрана энтропия. Для того, чтобы дерево было легче визуализировать, его глубина была ограничена 3 уровнями.

Была выполнена визуализация всех построенных деревьев. Результаты визуализации приведены на рис. 1.

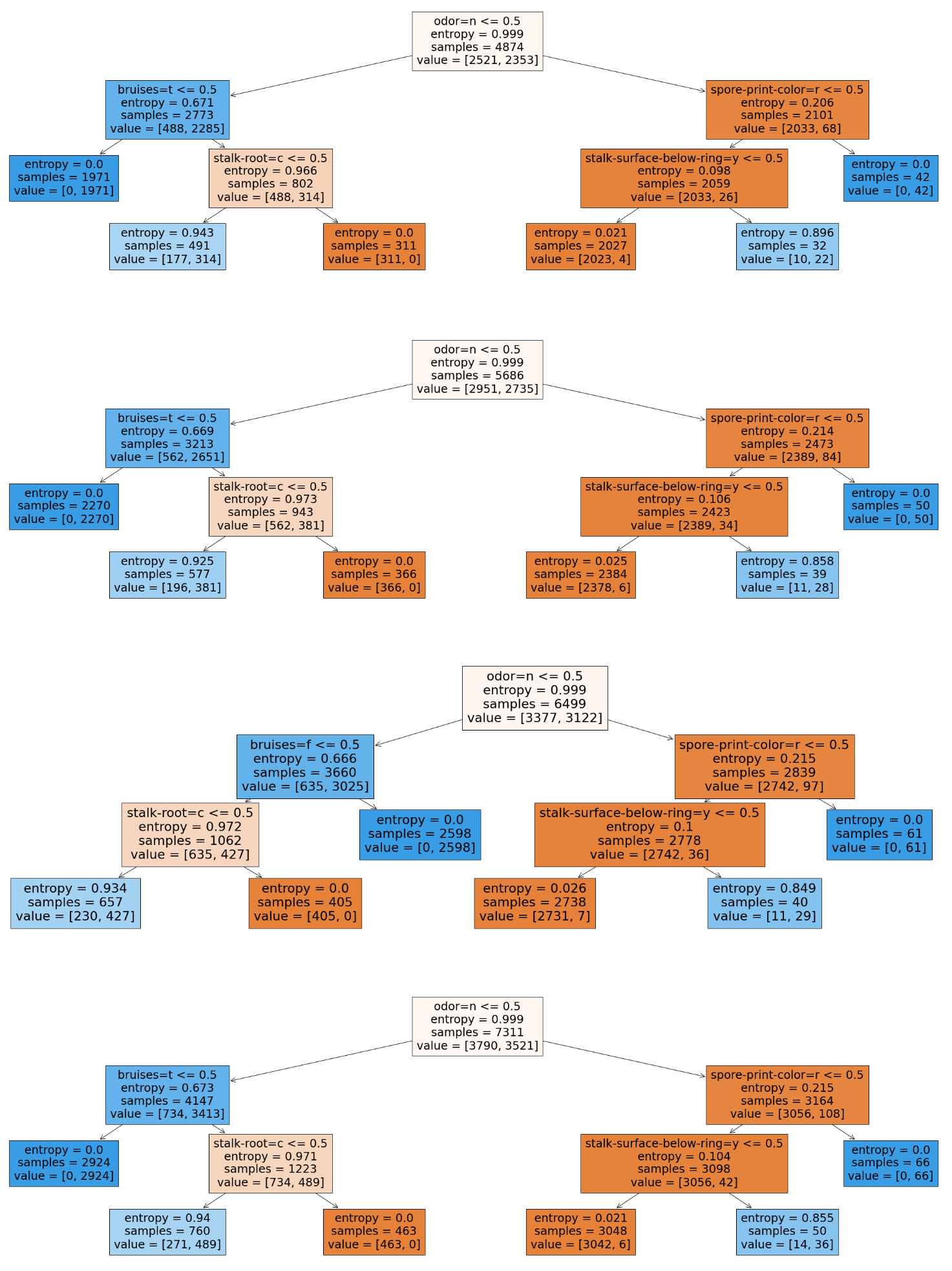


Рис. 1. Результаты визуализации деревьев

Из рисунка можно увидеть, что все построенные деревья практически одинаковы (все используемые признаки оказались одинаковыми, отличие заключается лишь в соотношении экземпляров в нечистых листьях). Первый использованный признак во всех деревьях – отсутствие запаха, из чего можно предположить, что этот признак является одним из самых важных при определении ядовитости гриба.

Далее была выполнена визуализация четырёх метрик качества на выбранных соотношениях. Результат визуализации приведён на рис. 2.

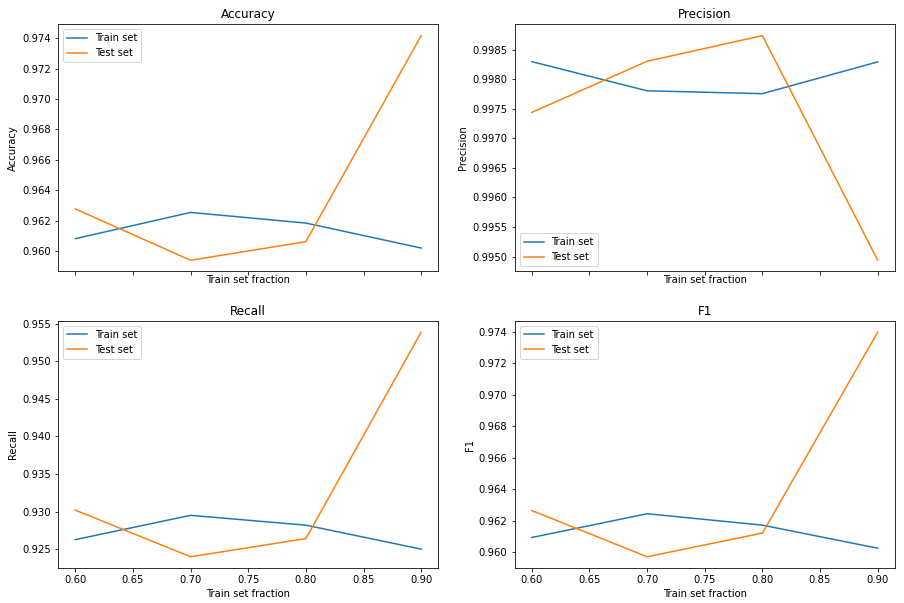


Рис. . Результаты визуализации метрик качества

Можно увидеть, что в соответствии с большинством метрик, наилучшим классификатором оказался построенный на разбиении 90%:10%. Как было замечено ранее, используемые признаки в дереве не отличаются, поэтому вероятная причина этого – в более близком к реальному соотношении долей ядовитых и неядовитых грибов в нечистых листьях.

Как и в прошлом задании, классификация должна стремиться уменьшить количество ложноотрицательных результатов (то есть ядовитые грибы не должны быть классифицированы как неядовитые). Таким образом, наиболее важной метрикой является recall (полнота), которая зависит от количества ложноотрицательных результатов. Эта метрика имеет наиболее низкие значения среди четырёх исследованных, что говорит о том, что качество классификации несколько ниже, чем может показаться изначально.