# Классификация с помощью дерева решений

Кузьмичева Ольга, КЭ-403

# Задание

1. Разработайте программу, которая выполняет классификацию заданного набора данных с помощью дерева решений. Параметрами программы являются набор данных, критерий выбора атрибута разбиения (Information gain, Gain ratio, Gini index).
2. Проведите эксперименты на двух наборах из различных предметных областей. Наборы данных должны существенно отличаться друг от друга по количеству атрибутов и/или классов. Например, наборы grades (сведения об оценках школьников за письменную контрольную работу) и Census Income (данные о результатах переписи населения, в т.ч. о годовом доходе -- ниже или выше $50000). В качестве обучающей выборки для построения дерева используйте 100% исходных данных.
3. Выполните визуализацию построенных деревьев решений.
4. Доработайте программу, добавив в список ее параметров долю, которую занимает обучающая выборка от общего размера набора данных, и обеспечив вычисление и выдачу в качестве результатов следующих показателей качества классификации: аккуратность (accuracy), точность (precision), полнота (recall), F-мера.
5. Проведите эксперименты на ранее выбранных наборах данных, фиксируя критерий выбора атрибута разбиения и варьируя соотношение мощностей обучающей и тестовой выборок от 60%:40% до 90%:10% с шагом 10%.
6. Выполните визуализацию полученных результатов в виде следующих диаграмм:
    - построенные деревья решений для заданного набора данных;
    - показатели качества классификации в зависимости от соотношения мощностей обучающей и тестовой выборок для заданного набора данных.
7. Подготовьте отчет о выполнении задания и загрузите отчет в формате PDF в систему. Отчет должен представлять собой связный и структурированный документ со следующими разделами:
    - формулировка задания;
    - гиперссылка на каталог репозитория с исходными текстами, наборами данных и др. сопутствующими материалами;
    - рисунки с результатами визуализации;
    - пояснения, раскрывающие смысл полученных результатов.

В соответствии с заданием была разработана программа, которая выполняет классификацию заданного набора данных с помощью дерева решений. Для задачи классификации была использована библиотека sklearn. В репозитории размещено четыре файла: 3lab_tree1.ipynb, 3lab_tree2.ipynb – реализация классификации с помощью дерева решений на разных датасетах, adult.csv – датасет, используемый в 3lab_tree2.ipynb, grades.csv – датасет, используемый в 3lab_tree1.ipynb.

## Эксперимент 1

В результате классификации первого датасета (grades - сведения об оценках школьников за письменную контрольную работу), были получены деревья решений, а также показатели качества классификации для размеров тестовой выборки: 40%, 30%, 20%, 10%.

На рисунке 1 изображено наиболее оптимально классифицированное дерево решений (величина тестовой выборки равна 10%, критерий выбора атрибута – information gain).

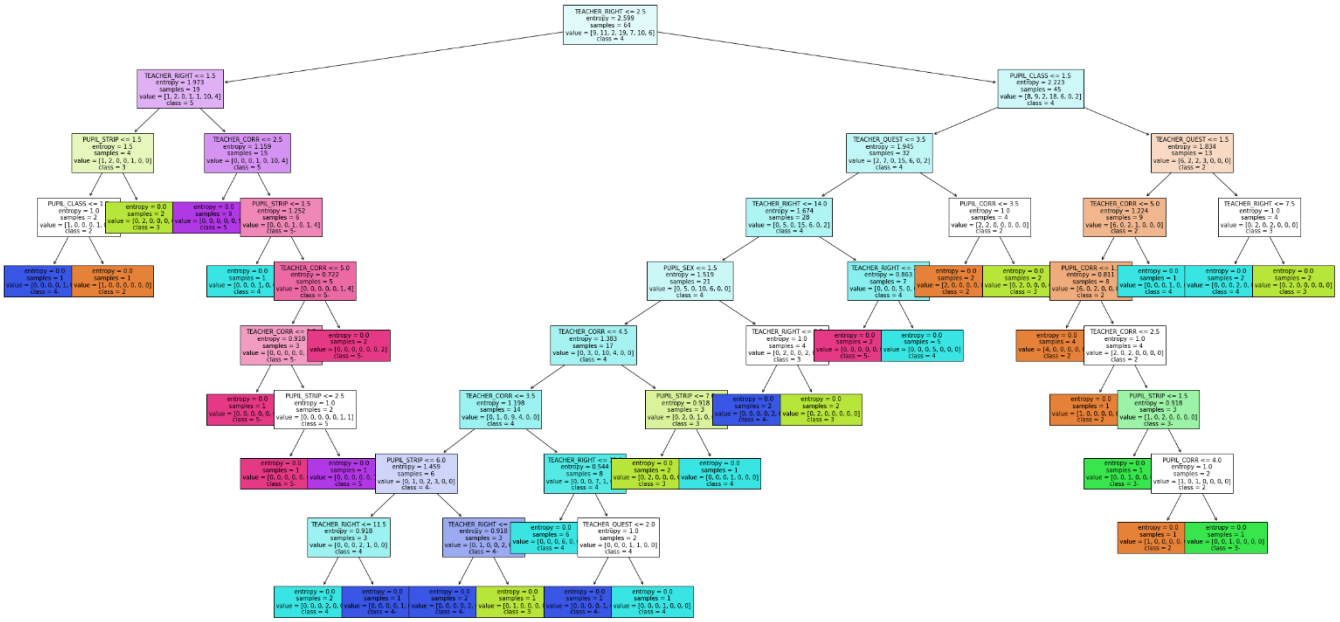

Рисунок 1 – Дерево решений

В узлах дерева находится следующая информация:

1. Условие разветвления (значения условий представлены в закодированном виде).
2. Энтропия определения классов на шаге – чем выше энтропия, тем более неоднозначно определены классы
3. Количество рассматриваемых данных за шаг
4. Значения определения классов – чем выше значение, чем вероятнее соответствующий класс

5. Класс – текстовое определение класса для наглядности. Для данного датасета классы равны «2», «3», «3-», «4», «4-», «5», «5-».

Левая ветвь дерева обозначает «да», а правая «нет». В конечных узлах дерева энтропия равна нулю, а класс определен однозначно.

Из полученного дерева, например, можно сделать следующий вывод: если оценка учителя >= 72 баллам, то поставленная оценка скорее всего будет «5», «-5», в другом случае, поставленная оценка скорее всего будет «4» или «3».

На рисунке 2 изображена таблица показателей качества классификации. Аккуратность (accuracy) обозначает долю данных, по которым классификатор принял правильное значение (в данном случае, удалось достичь значения 0.62). Точность (precision) обозначает долю данных, действительно принадлежащих данному классу относительно всех данных, которые система отнесла к этому классу. Полнота (recall) – это доля найденных классификатором данных, принадлежащих классу относительно всех данных этого класса. F-мера (f1-score) представляет собой гармоническое среднее между точностью и полнотой.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 2 | 1.00 | 0.50 | 0.67 | 2 |
| 3 | 1.00 | 0.50 | 0.67 | 2 |
| 4 | 0.50 | 1.00 | 0.67 | 1 |
| 4- | 0.50 | 1.00 | 0.67 | 1 |
| 5 | 0.00 | 0.00 | 0.00 | 1 |
| 5- | 0.50 | 1.00 | 0.67 | 1 |
| accuracy | | | 0.62 | 8 |
| macro avg | 0.58 | 0.67 | 0.56 | 8 |
| weighted avg | 0.69 | 0.62 | 0.58 | 8 |

Рисунок 2 – Показатели качества классификации

**Эксперимент 2**

В результате классификации второго датасета (Census Income - данные о результатах переписи населения, в т.ч. о годовом доходе - ниже или выше $50000), были получены деревья решений, а также показатели качества классификации для размеров тестовой выборки: 40%, 30%, 20%, 10%.

На рисунке 3 изображено наиболее оптимально классифицированное дерево решений (величина тестовой выборки равна 20%, критерий выбора атрибута – gini index). По полученной классификации можно сделать вывод, что женатые люди или никогда не бывавшие в браке имеют доход выше 50К в 70% случаев, в то время как люди, имеющие семейное положение «разведен», «вдова», «в разлуке» почти никогда не имеют доход выше 50К.

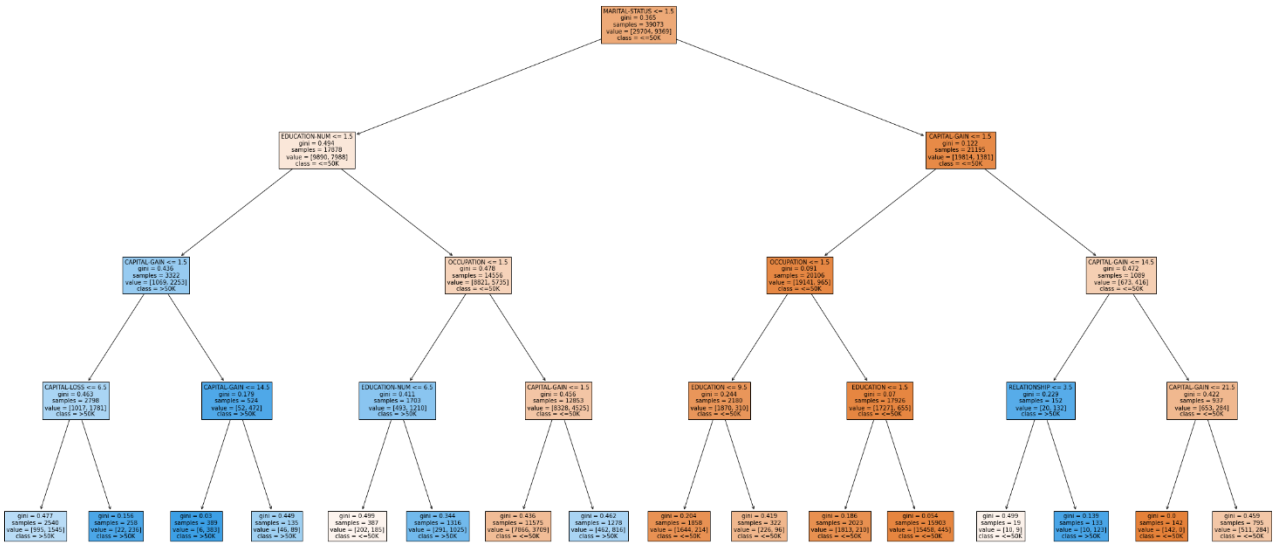

Рисунок 3 – Дерево решений

На рисунке 4 изображена таблица показателей качества классификации. Можно заметить, что удалось достичь аккуратности в 0.82 процента. Более высокие показатели точности с тестовых выборками любого размера по сравнению с первым экспериментом можно объяснить большим количеством данных в датасете.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| <=50K | 0.85 | 0.93 | 0.89 | 7451 |
| >50K | 0.68 | 0.46 | 0.55 | 2318 |
| accuracy | | | 0.82 | 9769 |
| macro avg | 0.76 | 0.69 | 0.72 | 9769 |
| weighted avg | 0.81 | 0.82 | 0.81 | 9769 |

Рисунок 4 – Показатели качества классификации