Федеральное государственное бюджетное образовательное учреждение

высшего профессионального образования

**Национальный исследовательский университет «МЭИ»**

**Институт информационных и вычислительных технологий**

**Кафедра Прикладной Математики и Искусственного интеллекта**

Лабораторная работа №3

**«Алгоритм CART»**

**Выполнил:** студент Алибеков А. А.

Группа: А-05м-24

**Проверил:** Михайлов И. С.

**Москва 2024**

**Постановка задачи:**

Реализовать алгоритм CART для построения дерева решений. Процесс решения должен отображаться на форме. На каждом этапе работы алгоритма вычисляется подходящий признак для разбиения на основе анализа impurity узлов. Необходимо показать выбранный признак и его значение impurity. В качестве языка реализации можно использовать любой язык программирования.

В программе должно задаваться количество уровней дерева решений. Необходимо выполнить построение дерева решений для различных значений уровней. Оценить точность классификации. Полученные результаты отобразить в таблице и на графике. Сделать вывод об оптимальном значении количества уровней дерева решений.

После построения основного дерева решений необходимо проверить его устойчивость. Для этого следует удалить некоторые записи из обучающего множества и проверить работу (качество классификации) дерева решений на той же тестовой выборке. Выполнить данную процедуру три раза. Сформировать таблицу, на каждой строке которой будут содержаться данные об общем количестве записей в обучающей выборке, количестве объектов в тестовой выборке, количестве правильно классифицированных и количестве неправильно классифицированных объектов. Сравнить результаты.

**Принцип работы дерева решений (CART)**

Алгоритм строится следующим образом:

1) создаётся корневой узел на основе наилучшего разбиения;

2) тренировочный набор разбивается на 2 поднабора: всё что соответствует условию разбиения отправляется в левый узел, остальное — в правый;

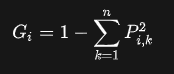
3) далее рекурсивно для каждого тренировочного поднабора повторяются шаги 1-2 пока не будет достигнут один из основных критериев останова: максимальная глубина, максимальное количество листьев, минимальное количество наблюдений в листе или минимальное снижение загрязнения в узле.

**Выбор наилучшего разбиения**

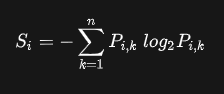
Для быстрого поиска правильного ответа необходимо задавать вопросы, которые исключат наибольшее количество неверных вариантов.

Показатель того, насколько хорошо вопрос в решающем узле позволяет отделить верный ответ от неверных, называется мерой загрязнённости узла. В случае классификации для оценки качества разбиения узла используются следующие критерии:

* Неопределённость (загрязнённость) Джини — мера разнообразия в распределении вероятностей классов. Если все элементы в узле принадлежат к одному классу, то неопределённость Джини равна 0, а в случае равномерного распределения классов в узле неопределённость Джини равна 0.5.



* Энтропия Шеннона — мера неопределённости или беспорядка классов в узле. Она характеризует количество информации, которое необходимо для описания состояния системы: чем выше значение энтропии, тем менее упорядочена система и наоборот.



* Ошибка классификации — величина, отображающая долю неправильно классифицированных элементов в узле: чем меньше данное значение, тем меньше загрязнённость в узле.



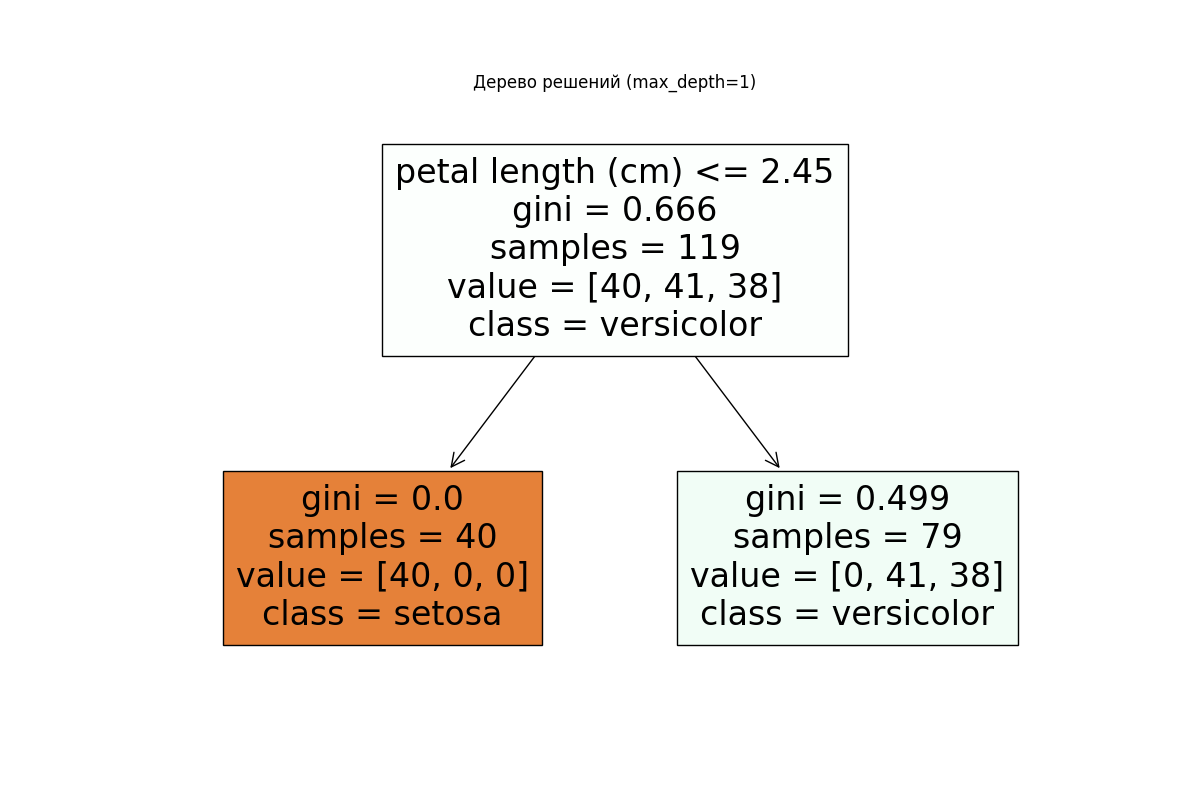
В данном случае Pi,k — это доля k-го класса среди обучающих образцов в i-ом узле.

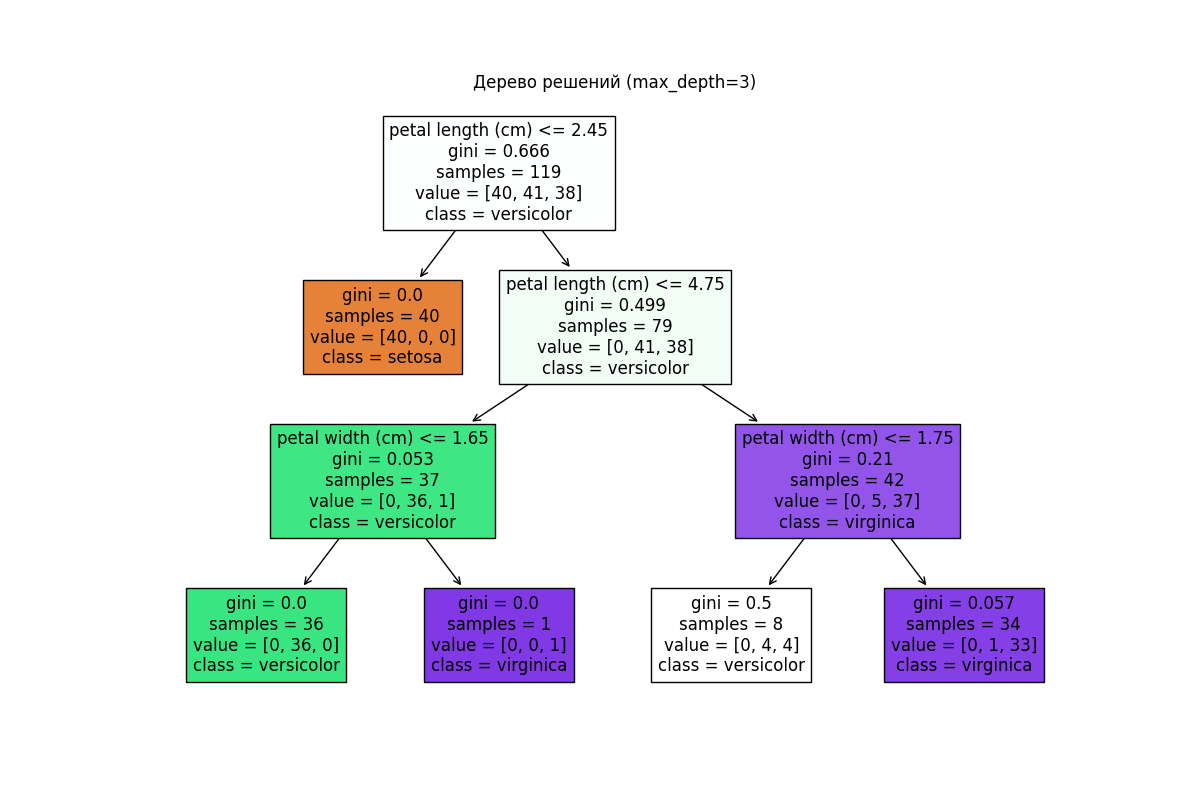
На практике чаще всего используются неопределённость Джини и энтропия Шеннона за счёт большей информативности.

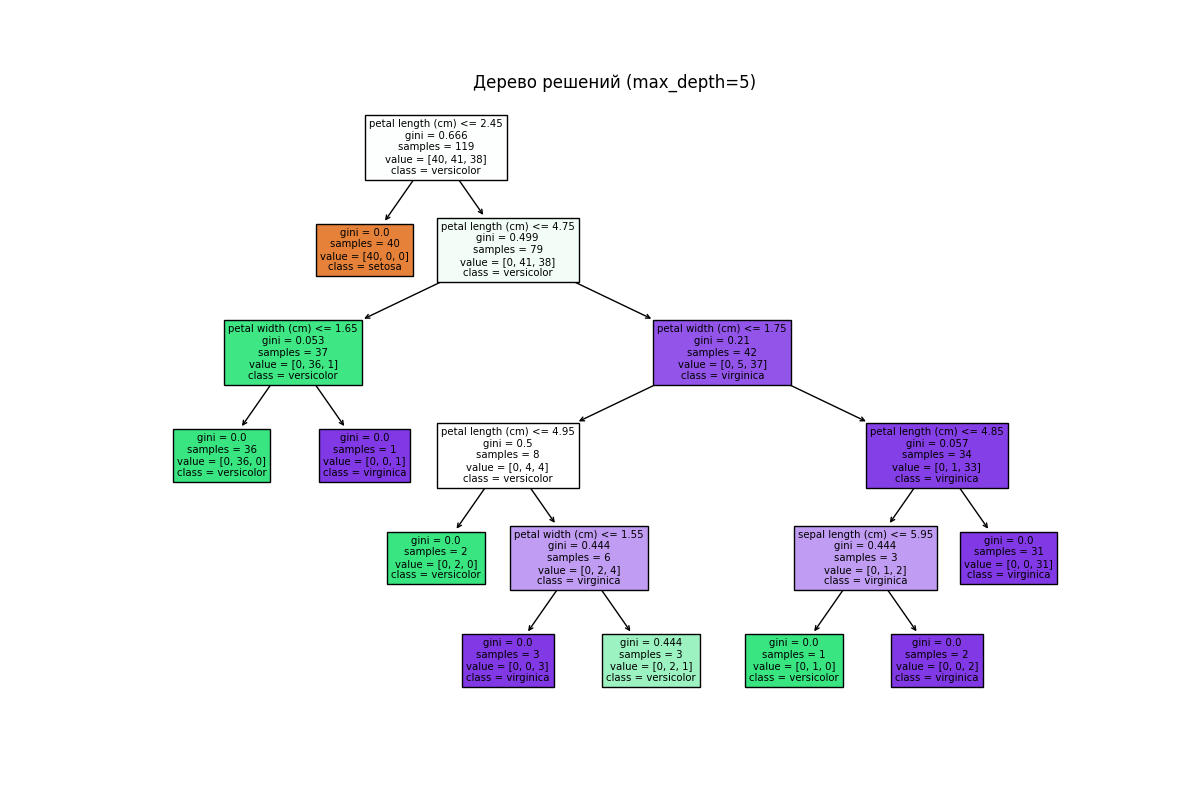
**Примеры работы программы:**

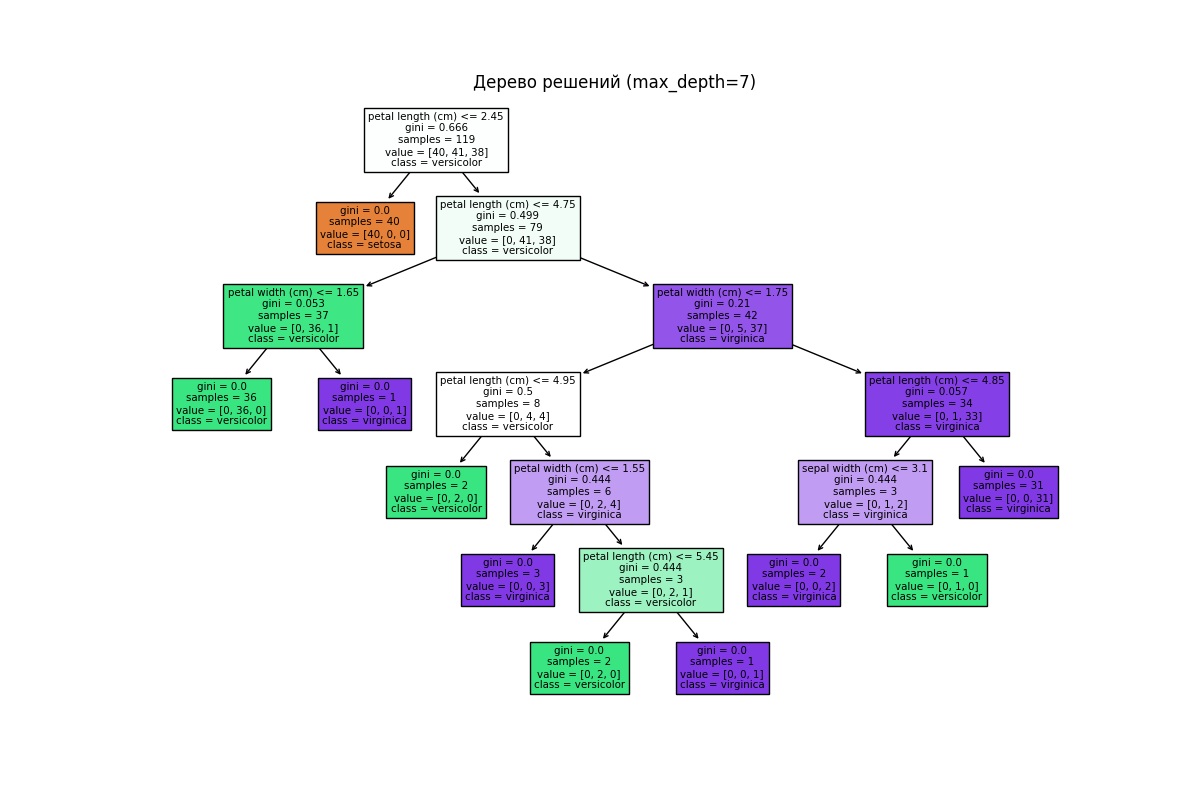
Программа была разработана на основе данных iris модуля sklearn datasets. Для обучающей выборки взято 104 значений (70%), для тестовой выборки – 45 (30%).

**Библиотечная реализация(Python).**



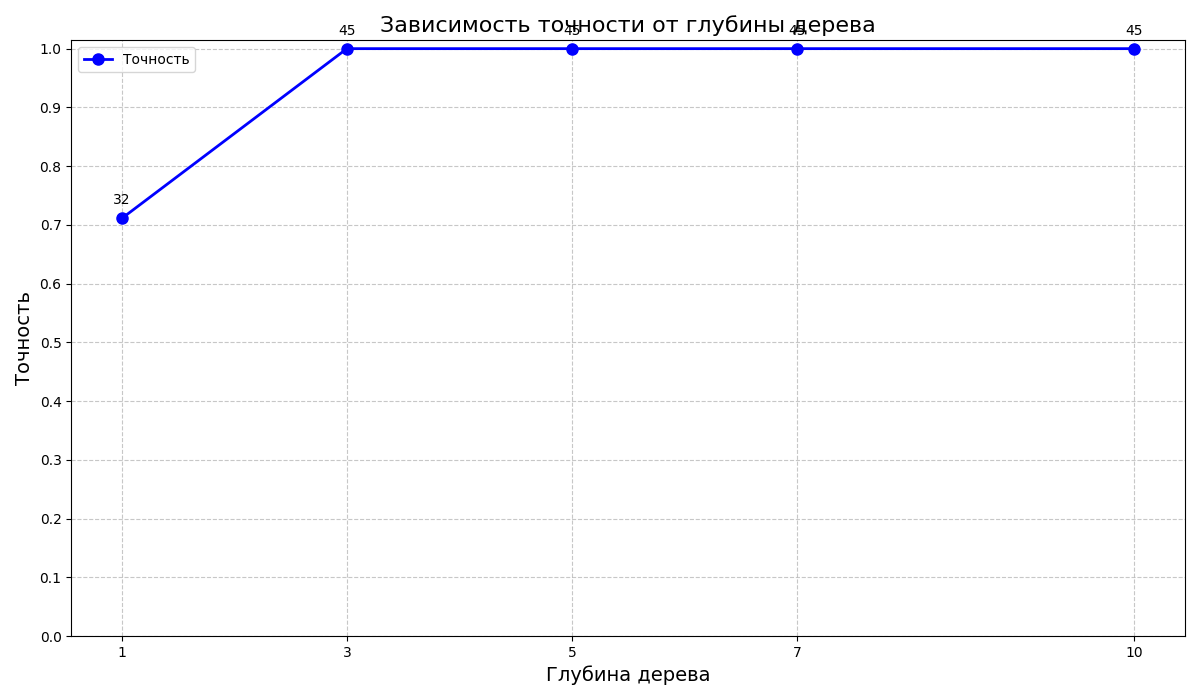






Оценим зависимость точности и кол-ва правильно классифицированных объектов от глубины дерева и отобразим результат в таблице:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Глубина | 1 | 3 | 5 | 7 | 10 |
| Точность | 0,71 | 1 | 1 | 1 | 1 |
| Кол-во правильно  классифиц. | 32 | 45 | 45 | 45 | 45 |



При глубине дерева 3 и более точность остается неизменной и не растет. Поэтому глубина дерева = 3 – оптимальна для выбранной выборки.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Iteration | Train Size | Test Size | Correct | Incorrect |
| 1 | 93 | 45 | 45 | 0 |
| 2 | 93 | 45 | 44 | 1 |
| 3 | 93 | 45 | 45 | 0 |

Количество правильно классифицированных объектов не изменяется значительно между итерациями, что подтверждает, что алгоритм является устойчивым. Небольшое отклонение в одной из итераций незначительно и не влияет на общее качество классификации.

**Собственная реализация(Julia).**

**Вывод:**

В ходе лабораторной работы была поставлена задача классификации, описан метод решения задачи, а именно алгоритм построения дерева классификации CART.

По итогам реализации и экспериментов можно сделать следующие выводы, что метод:

* Алгоритм CART показал достаточно хорошее качество классификации;
* На вход алгоритму необходимо подать глубину дерева;
* В виду подхода к разбиению данных данный алгоритм устойчив к выбросам;
* При увеличении глубины дерева, модель переобучается.