Decision tree

Мозжевилова Е.Д.
Нгуен Тхи Хань Хуен

5030102/20201

План презентации

Определение

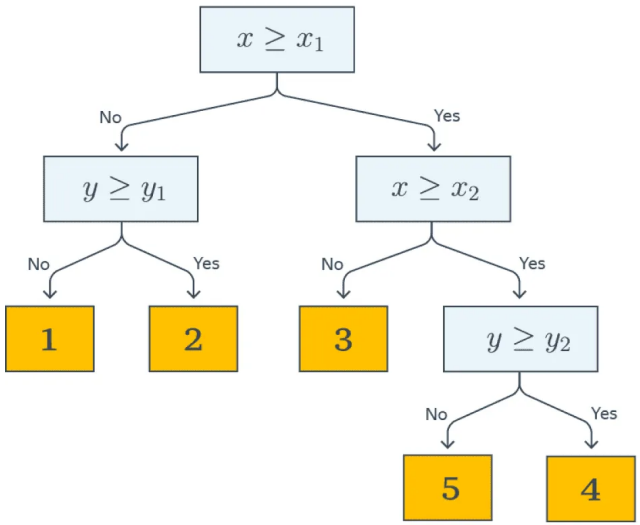
Решающее дерево —

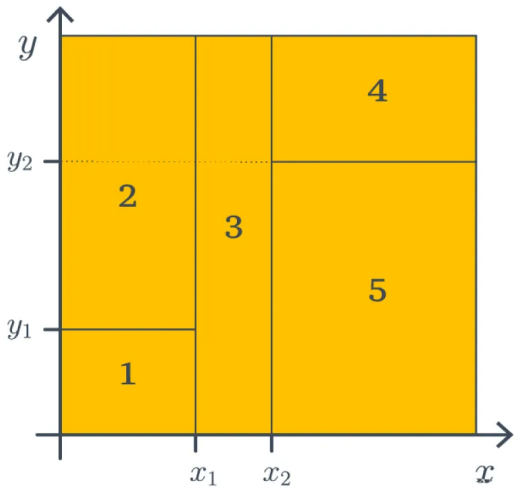
это алгоритм машинного обучения с учителем, используемый для задач классификации и регрессии. Это древовидная структура, которая моделирует процесс принятия решений, рекурсивно разбивая данные на подмножества на основе значений признаков (features).

Иллюстрация

Проще говоря

это набор вопросов "если-то", организованных в виде дерева, который приводит к решению.





Замечание

Дерево можно рассматривать как кусочно-постоянную аппроксимацию.

Структура

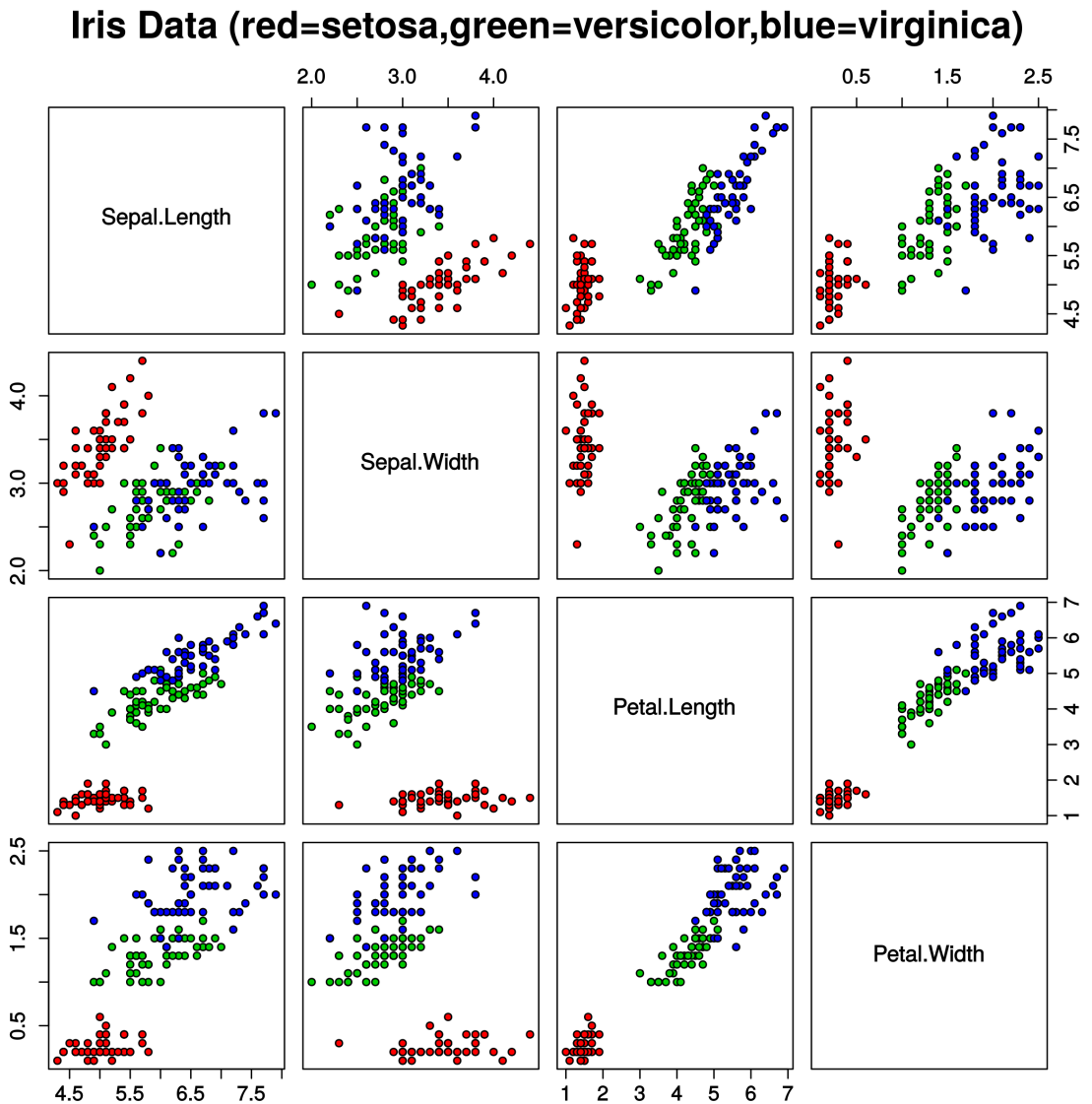
Основные понятия

- **Корневой узел**: Начало дерева, весь набор данных
- **Внутренние узлы (node)**: Решающие правила и подмножества наблюдений, которые им удовлетворяют.
- **Листья (leaf)**: Классифицированные деревом наблюдения, каждый лист ассоциируется с одним из классов.

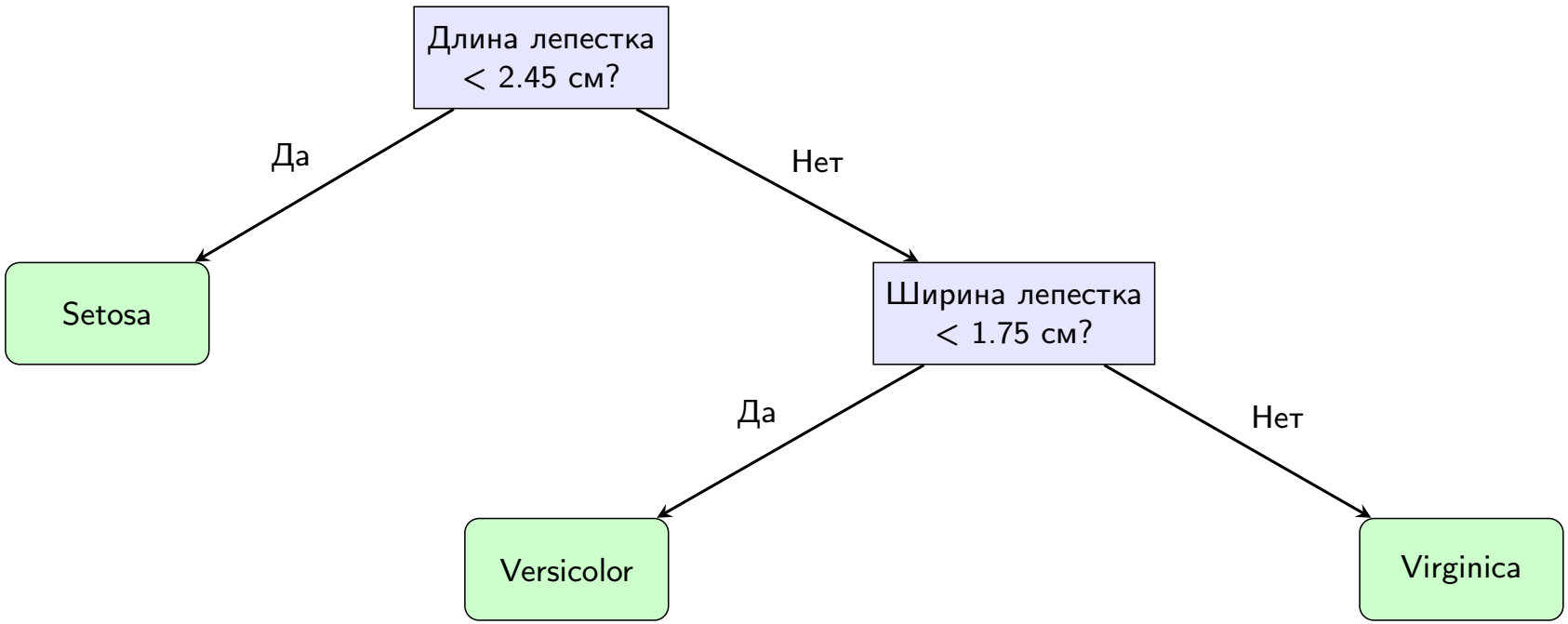
Пример

Смотрим предыдущий слайд:
Узлы отмечены голубым, листья жёлтым.

Ирисы Фишера



Ирисы Фишера



Алгоритмы обучения

Существует несколько алгоритмов построения деревьев (ID3, C4.5, CART). Наиболее распространенным является **CART (Classification and Regression Trees)**. Его процесс можно описать как «жадный» рекурсивный алгоритм разделения данных.

Алгоритм CART

1. Начинаем с корневого узла, содержащего все данные.
2. Жадно ищем лучшее разбиение для каждого узла по каждому признаку и каждому возможному порогу. Цель — найти разбиение, которое максимально "очистит" дочерние узлы.
3. Разделяем выбранный узел, по выбранному правилу.
4. **Критерии остановки:**
   - Все узлы стали чистыми (или почти чистыми).
   - Достигнута максимальная глубина дерева.
   - Количество объектов в узле стало меньше минимально допустимого.

Критерии разделения

Критерии

Чтобы выбрать лучший признак для разделения в узле, алгоритм использует метрики, которые измеряют «нечистоту» узлов. Цель — сделать дочерние узлы как можно более «чистыми», то есть содержащими объекты преимущественно одного класса.

Для категориальных признаков (классификация):

- Индекс Джини
- Энтропия

Для непрерывных признаков (регрессия):

- Дисперсия: минимизация MSE (Mean Squared Error)
- Отклонение Пуассона
- средняя абсолютная ошибка

Критерии разделения для классификации

Индекс Джини (Gini Index)

$$I_{Gini}(node) = 1 - \sum_{k=1}^{K} (p_k)^2$$

где p_k — доля объектов класса k в узле. Максимален, когда классы распределены равномерно. Стремится к нулю, когда узел содержит объекты только одного класса.

Энтропия (Entropy)

$$I_{Entropy}(node) = -\sum_{k=1}^{K} p_k \cdot \log_2(p_k)$$

Также максимальна при равномерном распределении классов и равна нулю для чистового узла.

Критерии разделения для регрессии

Среднеквадратичная ошибка (MSE)

$$I_{MSE}(node) = \frac{1}{N_{node}} \sum_{i \in node} (y_i - \bar{y}_{node})^2$$

где $\bar{y}_{node}$ — среднее значение целевой переменной в узле.

Средняя абсолютная ошибка (MAE)

$$I_{MAE}(node) = \frac{1}{N_{node}} \sum_{i \in node} |y_i - \tilde{y}_{node}|$$

$\tilde{y}_{node}$ — медиана. Более устойчива к выбросам, чем MSE.

Отклонение Пуассона (Poisson Deviance)

Используется для данных, распределенных согласно закону Пуассона.

$$I_{Poisson}(node) = \frac{1}{N_{node}} \sum_{i \in node} \left(y_i \cdot \log \left(\frac{y_i}{\bar{y}node} \right) - (y_i - \bar{y}node) \right)$$

Пример работы критериев для классификации(1/8)

Исходные данные

№	Возраст	Доход (тыс. руб.)	Покупает
1	25	40	0
2	28	35	0
3	30	75	1
4	40	80	1
5	45	50	1

Распределение классов

Всего объектов: $N = 5$

Класс «0»: $N_0 = 2$

Класс «1»: $N_1 = 3$

Доли классов:

$p_0 = \frac{2}{5} = 0.4$

$p_1 = \frac{3}{5} = 0.6$

Расчёт неоднородности

$$I_{Gini} = 1 - \sum_{k=1}^{2} (p_k)^2 = 0.48$$

$$I_{Entropy} = -\sum_{k=1}^{2} p_k \cdot \log_2(p_k) = 0.971$$

Пример (2/8)

Выбираем возможные признаки разделения

- Сортируем значения признака по возрастанию
- Берём середины между соседними различными значениями
- Каждая середина — потенциальный порог разделения

Признак «Возраст»

Упорядоченные значения:
25, 28, 30, 40, 45

Кандидаты-пороги:
$\frac{25+28}{2} = 26.5$
$\frac{28+30}{2} = 29$
$\frac{30+40}{2} = 35$
$\frac{40+45}{2} = 42.5$

Признак «Доход»

Упорядоченные значения:
35, 40, 50, 75, 80

Кандидаты-пороги:
$\frac{35+40}{2} = 37.5$
$\frac{40+50}{2} = 45$
$\frac{50+75}{2} = 62.5$
$\frac{75+80}{2} = 77.5$

Пример (3/8)

Порог: $age \leq 26.5$

Левое поддерево ($age \leq 26.5$): №1 (25 лет, класс 0)

- $N_{left} = 1,\ N_0 = 1,\ N_1 = 0,\ p_0 = 1.0,\ p_1 = 0.0,\ I_{Gini}^{left} = 0.0,\ I_{Entropy}^{left} = 0.0$

Правое поддерево ($age > 26.5$): №2-5 (классы: 0,1,1,1)

- $N_{right} = 4,\ N_0 = 1,\ N_1 = 3,\ p_0 = 0.25,\ p_1 = 0.75$
- $I_{Gini}^{right} = 1 - (0.25^2 + 0.75^2) = 0.375$
- $I_{Entropy}^{right} = -(0.25 \cdot \log_2(0.25) + 0.75 \cdot \log_2(0.75)) = 0.811$

Общий индекс Джини: $I_{Gini} = \frac{1}{5} \cdot 0.0 + \frac{4}{5} \cdot 0.375 = 0.3$
Общая энтропия: $I_{Entropy} = \frac{1}{5} \cdot 0.0 + \frac{4}{5} \cdot 0.811 = 0.649$

Порог: $age \leq 29$

Левое поддерево ($age \leq 29$): №1-2 (25,28 лет, классы: 0,0)

- $N_{left} = 2,\ N_0 = 2,\ N_1 = 0,\ p_0 = 1.0,\ p_1 = 0.0,\ I_{Gini}^{left} = 0.0,\ I_{Entropy}^{left} = 0.0$

Правое поддерево ($age > 29$): №3-5 (классы: 1,1,1)

- $N_{right} = 3,\ N_0 = 0,\ N_1 = 3,\ p_0 = 0.0,\ p_1 = 1.0,\ I_{Gini}^{right} = 0.0,\ I_{Entropy}^{right} = 0.0$

Общий индекс Джини: 0.0, **Общая энтропия:** 0.0

Пример (4/8)

Порог: $age \leq 35$

Левое поддерево ($age \leq 35$): №1-3 (классы: 0,0,1)

- $N_{left} = 3,\ N_0 = 2,\ N_1 = 1,\ p_0 = 0.667,\ p_1 = 0.333$
- $I_{Gini}^{left} = 1 - (0.667^2 + 0.333^2) = 0.444$
- $I_{Entropy}^{left} = -(0.667 \cdot \log_2(0.667) + 0.333 \cdot \log_2(0.333)) = 0.918$

Правое поддерево ($age > 35$): №4-5 (классы: 1,1)

- $N_{right} = 2,\ N_0 = 0,\ N_1 = 2,\ I_{Gini}^{right} = 0.0,\ I_{Entropy}^{right} = 0.0$

Общий индекс Джини: $\frac{3}{5} \cdot 0.444 + \frac{2}{5} \cdot 0.0 = 0.267$
Общая энтропия: $\frac{3}{5} \cdot 0.918 + \frac{2}{5} \cdot 0.0 = 0.551$

Пример (5/8)

Порог: $age \leq 42.5$

Левое поддерево ($age \leq 42.5$): №1-4 (классы: 0,0,1,1)

- $N_{left} = 4,\ N_0 = 2,\ N_1 = 2,\ p_0 = 0.5,\ p_1 = 0.5$
- $I_{Gini}^{left} = 1 - (0.5^2 + 0.5^2) = 0.5$
- $I_{Entropy}^{left} = -(0.5 \cdot \log_2(0.5) + 0.5 \cdot \log_2(0.5)) = 1.0$

Правое поддерево ($age > 42.5$): №5 (класс: 1)

- $N_{right} = 1,\ N_0 = 0,\ N_1 = 1$
- $I_{Gini}^{right} = 0.0,\ I_{Entropy}^{right} = 0.0$

Общий индекс Джини: $\frac{4}{5} \cdot 0.5 + \frac{1}{5} \cdot 0.0 = 0.4$
Общая энтропия: $\frac{4}{5} \cdot 1.0 + \frac{1}{5} \cdot 0.0 = 0.8$

Пример (6/8)

Порог: $income \leq 37.5$

Левое поддерево ($income \leq 37.5$): №1-2 (классы: 0,0)

- $N_{left} = 2,\ N_0 = 2,\ N_1 = 0,\ I_{Gini}^{left} = 0.0,\ I_{Entropy}^{left} = 0.0$

Правое поддерево ($income > 37.5$): №3-5 (классы: 1,1,1)

- $N_{right} = 3,\ N_0 = 0,\ N_1 = 3,\ I_{Gini}^{right} = 0.0,\ I_{Entropy}^{right} = 0.0$

Общий индекс Джини: 0.0 **Общая энтропия:** 0.0

Порог: $income \leq 45$

Левое поддерево ($income \leq 45$): №1-2,5 (классы: 0,0,1)

- $N_{left} = 3,\ N_0 = 2,\ N_1 = 1,\ p_0 = 0.667,\ p_1 = 0.333$
- $I_{Gini}^{left} = 0.444,\ I_{Entropy}^{left} = 0.918$

Правое поддерево ($income > 45$): №3-4 (классы: 1,1)

- $N_{right} = 2,\ N_0 = 0,\ N_1 = 2,\ I_{Gini}^{right} = 0.0,\ I_{Entropy}^{right} = 0.0$

Общий индекс Джини: $\frac{3}{5} \cdot 0.444 + \frac{2}{5} \cdot 0.0 = 0.267$
Общая энтропия: $\frac{3}{5} \cdot 0.918 + \frac{2}{5} \cdot 0.0 = 0.551$

Пример (7/8)

Порог: $income \leq 62.5$

Левое поддерево ($income \leq 62.5$):№1-2,5 (классы: 0,0,1)

- $N_{left} = 3,\ N_0 = 2,\ N_1 = 1$
- $I^{left}_{Gini} = 0.444,\ I^{left}_{Entropy} = 0.918$

Правое поддерево ($income > 62.5$): №3-4 (классы: 1,1)

- $N_{right} = 2,\ N_0 = 0,\ N_1 = 2,\ I^{right}_{Gini} = 0.0,\ I^{right}_{Entropy} = 0.0$

Общий индекс Джини: 0.267 **Общая энтропия:** 0.551

Порог: $income \leq 77.5$

Левое поддерево ($income \leq 77.5$): №1-2,3,5 (классы: 0,0,1,1)

- $N_{left} = 4,\ N_0 = 2,\ N_1 = 2$
- $I^{left}_{Gini} = 0.5,\ I^{left}_{Entropy} = 1.0$

Правое поддерево ($income > 77.5$): №4 (класс: 1)

- $N_{right} = 1,\ N_0 = 0,\ N_1 = 1,\ I^{right}_{Gini} = 0.0,\ I^{right}_{Entropy} = 0.0$

Общий индекс Джини: 0.4 **Общая энтропия:** 0.8

Пример (8/8)

Признак и порог	Индекс Джини	Энтропия	Выигрыш
Возраст 26.5	0.300	0.649	
Возраст 29	**0.000**	**0.000**	**Лучший!**
Возраст 35	0.267	0.551	
Возраст 42.5	0.400	0.800	
Доход 37.5	**0.000**	**0.000**	**Лучший!**
Доход 45	0.267	0.551	
Доход 62.5	0.267	0.551	
Доход 77.5	0.400	0.800	
Исходный узел	0.480	0.971	

Вывод

Наилучшие пороги разделения (дают чистые узлы):

- **Возраст 29** (разделяет на [0,0] и [1,1,1])
- **Доход 37.5** (разделяет на [0,0] и [1,1,1])

Оба порога дают максимальное уменьшение неоднородности.

Переобучение

Переобучение

Переобучение происходит, когда дерево становится слишком глубоким и начинает запоминать обучающие данные, а не изучать общие закономерности. Это приводит к снижению эффективности на новых, ранее не исследованных данных.

Отсечение

Отсечение — важный метод, используемый для предотвращения переобучения в деревьях решений.

- Метод снижает сложность дерева, удаляя ветви с низкой предсказательной силой.
- Повышает производительность модели, помогая дереву лучше обобщать новые данные.
- Упрощает и ускоряет развертывание модели.

Метрики оценки классификации

Основные метрики для классификации

- **Accuracy:** Доля правильных ответов модели
- **Precision:** Доля достоверно положитель из найденых полож
- **Recall:** Доля положительных образцов, которые были правильно классифицированы
- **F1-score:** Гармоническое среднее между точностью и полнотой

Метрики для бинарной классификации

Метрика	**Формула**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-score	$2 \times \frac{Precision \times Recall}{Precision+Recall}$

Обозначения

TP — True Positive (верно предсказанный класс 1),
TN — True Negative (верно предсказанный класс 0),
FP — False Positive (ошибочно предсказанный класс 1),
FN — False Negative (ошибочно предсказанный класс 0)

Метрики для многоклассовой классификации

Подходы к оценке многоклассовой классификации:

- **Micro-average:** Учет всех классов одновременно
- **Macro-average:** Усреднение метрик по всем классам
- **Weighted average:** Усреднение с учетом количества объектов в каждом классе

Например для Precision:

Метрика	**Формула**
Micro Precision	$\frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + \sum_{i=1}^{C} FP_i}$
Macro Precision	$\frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i}$
Weighted Precision	$\sum_{i=1}^{C} w_i \cdot \frac{TP_i}{TP_i + FP_i}$

Обозначения

C — количество классов,

w_i — вес класса i

Метрики для регрессии

Основные метрики оценки регрессионных моделей:

Метрика	**Формула**
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
R^2	$1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$
MAPE	$\frac{100\%}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i-\hat{y}_i}{y_i}\right\rvert$

Обозначения

y_i — истинное значение,

$\hat{y}_i$ — предсказанное значение,

$\bar{y}$ — среднее значение целевой переменной,

n — количество наблюдений

Подготовка

Цель исследования: Сравнить различные алгоритмы и метрики построения решающих деревьев на датасете Iris.

Используемые данные:

- Датасет Iris (150 samples, 4 features)
- 3 класса: Setosa, Versicolor, Virginica
- Разделение: 50% train / 50% test

Метрики оценки:

- Accuracy, Precision, Recall, F1-score
- Визуальный анализ структуры деревьев

Исследуемые модели

Модель	Алгоритм	Критерий разбиения	Дополнительные опции
ID3 (Entropy)	ID3	Entropy	
CART (Gini)	CART	Gini	
Random Split	CART	Gini	Случайный выбор признаков
Deep Tree	CART	Gini	Без ограничения глубины
Pruned Tree	CART	Gini	Ограничение глубины 3

Ключевые параметры

- average = 'macro' для метрик

Реализация на Python

```
X, y = load_iris(return_X_y=True)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5, random_state=42)

models = {
    "ID3 (Entropy)": DecisionTreeClassifier(criterion='entropy'),
    "CART (Gini)": DecisionTreeClassifier(criterion='gini'),
    "Random Split": DecisionTreeClassifier(criterion='gini', splitter='random'),
    "Deep Tree": DecisionTreeClassifier(max_depth=None),
    "Pruned Tree (max_depth=3)": DecisionTreeClassifier(max_depth=3)
}

results = []
for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    results.append({
        "Model": name,
        "Accuracy": accuracy_score(y_test, y_pred),
        "Precision": precision_score(y_test, y_pred,average='macro'),
        "Recall": recall_score(y_test, y_pred, average='macro'),
        "F1-score": f1_score(y_test, y_pred, average='macro'),
    })
```

Результаты оценки моделей

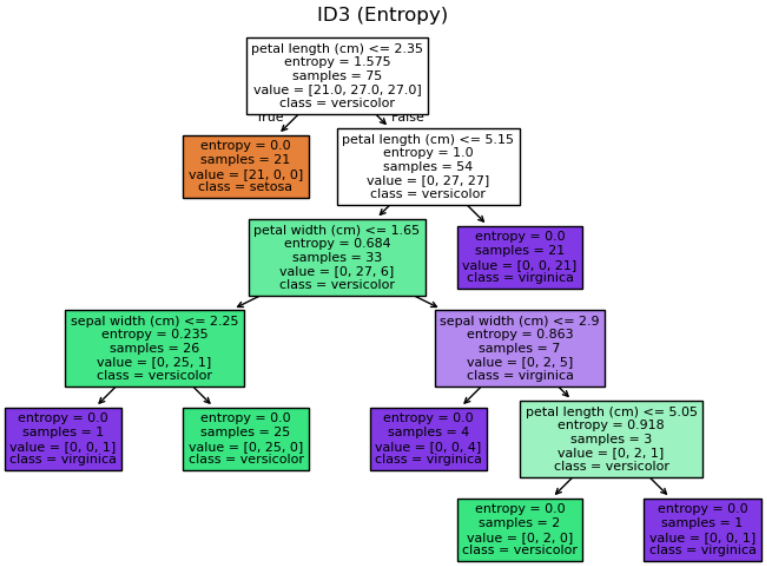
Модель	Accuracy	Precision	Recall	F1-score
ID3 (Entropy)	0.920	0.913	0.913	0.913
CART (Gini)	0.907	0.899	0.899	0.899
Random Split	0.920	0.921	0.913	0.912
Deep Tree	0.907	0.899	0.899	0.899
Pruned Tree	**1.000**	**1.000**	**1.000**	**1.000**

Важное наблюдение

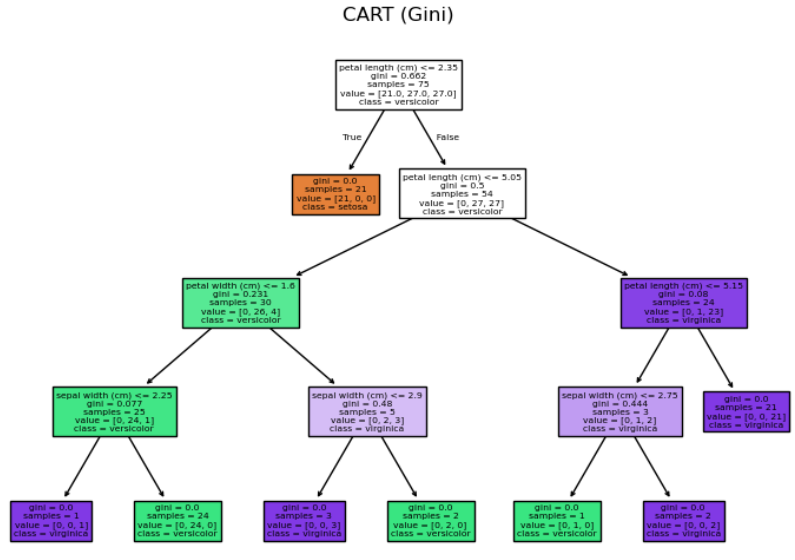
Ограничение глубины улучшило обобщающую способность

Визуализация деревьев: ID3 vs CART

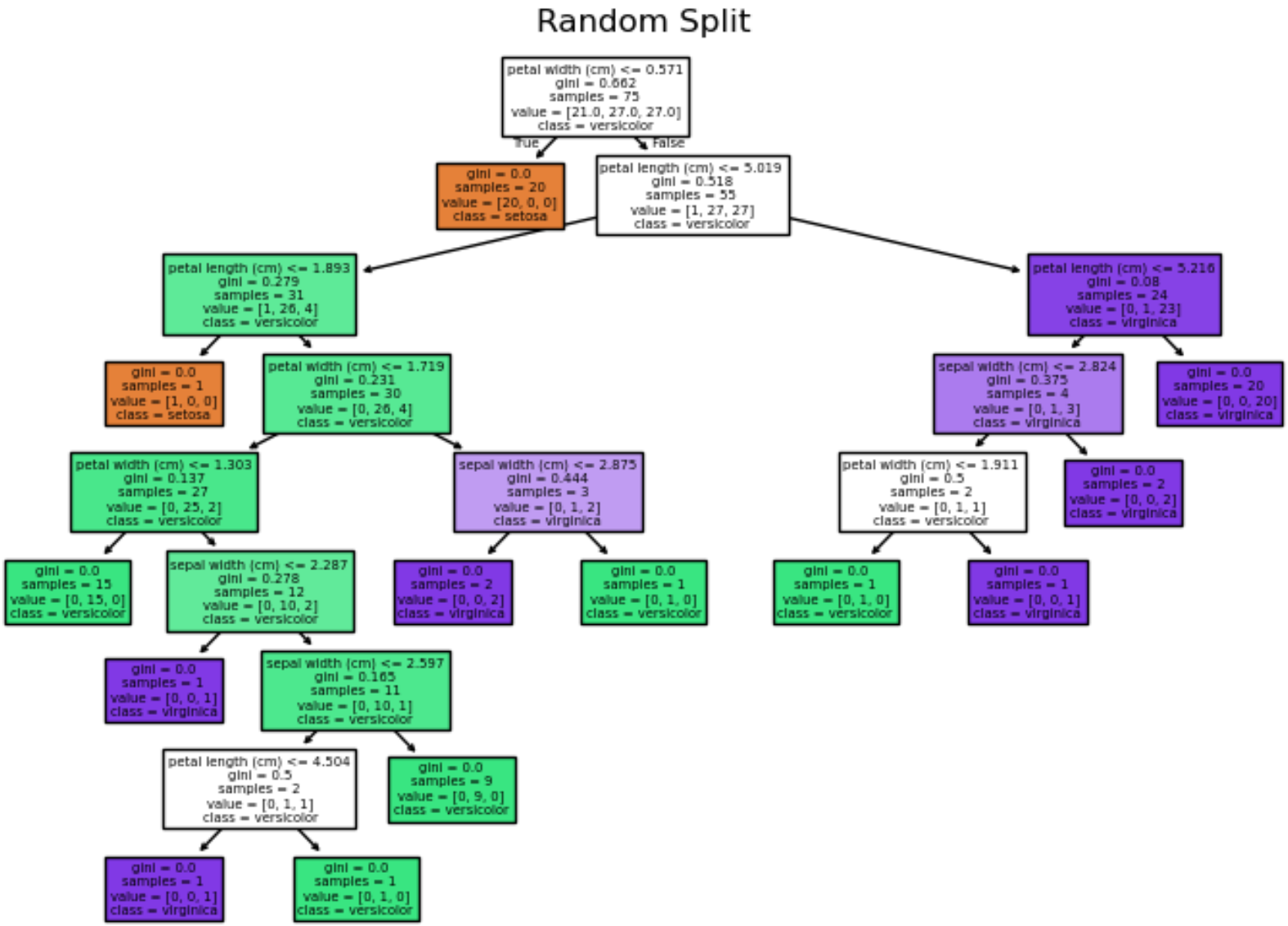
ID3 (Entropy)



CART (Gini)

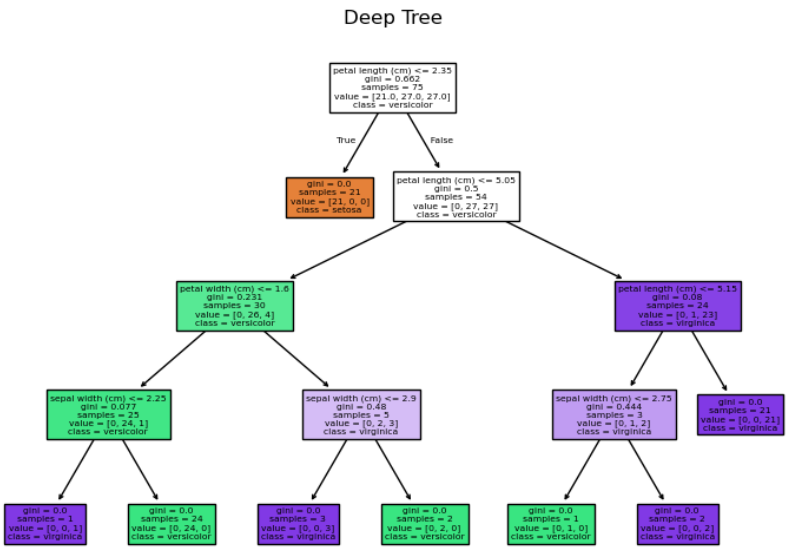


Визуализация деревьев: Random Tree

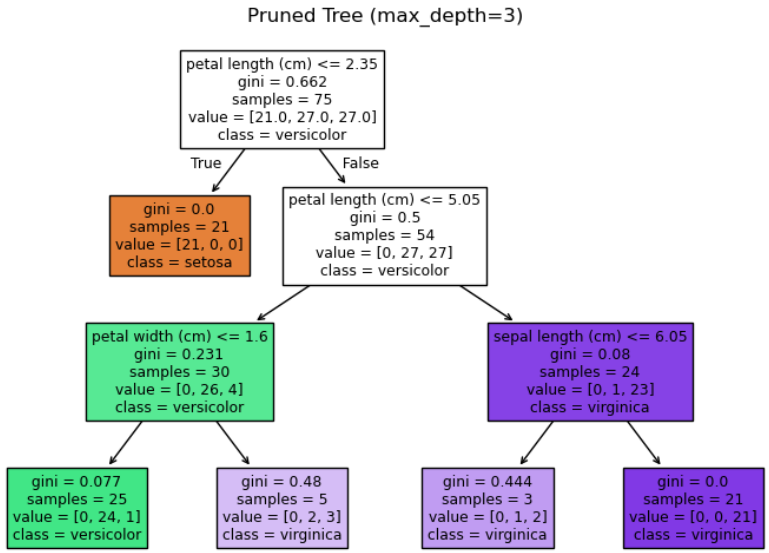


Визуализация деревьев: Deep vs Pruned

Deep Tree



Pruned Tree



Преимущества

Широкая популярность деревьев решений обусловлена следующими их преимуществами:

- Правила формируются практически на естественном языке, что делает объясняющую способность деревьев решений очень высокой.
- Работают как с числовыми, так и с категориальными данными.
- Требуют относительно небольшой предобработки данных, в частности, не требуют нормализации, создания фиктивных переменных, могут работать с пропусками.
- Могут работать с большими объемами данных .

Ограничения

Деревьям решений присущ ряд ограничений:

- Неустойчивость — даже небольшие изменения в данных могут привести к значительным изменениям результатов классификации.
- Поскольку алгоритмы построения деревьев решений являются жадными, они не гарантируют построения оптимального дерева.
- Склонность к переобучению.

Применение

1. Банковское дело и финансы
2. Медицина
3. Маркетинг и бизнес-аналитика
4. Промышленность
5. Интернет-технологии и рекомендательные системы
6. Наука и исследования

Применение

Банковское дело и финансы

- Кредитный скролинг — банк решает, выдавать ли кредит клиенту.
- Определение уровня риска по признакам: возраст, доход, кредитная история.
- Обнаружение мошеннических транзакций (fraud detection).

Медицина

- Диагностика заболеваний: по симптомам и результатам анализов дерево помогает вынести прогноз.

Применение

Маркетинг и бизнес-аналитика

- Сегментация клиентов по поведению (кто скорее купит товар).
- Прогнозирование отклика на рекламные кампании.
- Определение стратегии скидок.

Промышленность

- Контроль качества продукции: «годен» или «брак».
- Оптимизация производственных процессов.

Применение

Интернет-технологии и рекомендательные системы

- Рекомендации фильмов или товаров (например, «купит/не купит»).
- Классификация текстов или спам-фильтры в электронной почте.

Наука и исследования

- Биология: классификация растений или видов животных по характеристикам.
- Социология: прогнозирование поведения людей на основе опросов.
- И многое другое...

Список источников


Вики-портал loginom
https://wiki.loginom.ru/articles/decision-trees.html


scikit-learn
https://scikit-learn.org/stable/modules/tree.html#mathematical-formulation


Яндекс Учебник
https://education.yandex.ru/handbook/ml/article/reshayushchiye-derevya


Статья на geeksforgeeks
https://www.geeksforgeeks.org/machine-learning/decision-tree/

Спасибо за внимание!