«Сравнение алгоритмов построения деревьев решений (ID3, C4.5, CART, CHAID) на примере задачи выбора поставщика»

Цель работы:

Изучить основные алгоритмы построения деревьев решений (ID3, C4.5, CART, CHAID) и применить их к задаче классификации в предметной области *выбора поставщика*. Разработать программную реализацию алгоритмов, подготовить обучающую выборку, выполнить визуализацию построенных деревьев решений и провести сравнительный анализ полученных моделей.

Задачи:

1. *Теоретические задачи:*

- Изучить принципы построения деревьев решений и различия между алгоритмами ID3, C4.5, CART, CHAID.
- Проанализировать используемые критерии разбиения:
- ID3 — информационный выигрыш (Information Gain)
- C4.5 — коэффициент прироста информации (Gain Ratio)
- CART — критерий Джини (Gini Index)
- CHAID — χ^2-критерий (Chi-square)

2. *Практические задачи:*

- Разработать предметную область «Выбор поставщика».
- Сформировать обучающую выборку объёмом 14 примеров и 4 атрибутами: цена, качество, срок поставки, надёжность
- Реализовать алгоритм ID3 на языке C++.
- Сохранить выборку в CSV-файл для последующего анализа.
- Реализовать визуализацию деревьев решений ID3, C4.5, CART, CHAID с помощью Python.
- Сохранить изображения деревьев для дальнейшего сравнения.

3. *Аналитические задачи:*

- Выполнить сравнительный анализ деревьев по следующим метрикам: глубина дерева; количество листьев; корневой признак; структура разбиений.
- Сопоставить различия в поведении алгоритмов и устойчивость моделей к переобучению.
- Определить наиболее значимые факторы выбора поставщика.

В работе решается задача классификации в области выбора поставщика. Каждый поставщик описывается четырьмя атрибутами: *Цена, Качество, Срок поставки* и *Надёжность*; целевой атрибут — *Решение* («Да»/«Нет»). На

основе выборки из 14 примеров построены деревья решений четырьмя алгоритмами: **ID3**, **C4.5**, **CART** и **CHAID**, отличающимися используемыми критериями разбиения.

Выборка сохранена в *supplier_dataset.csv*, а результаты визуализации деревьев — в файлах *id3_tree.png*, *c45_tree.png*, *cart_tree.png*, *chaid_tree.png*. Они используются для анализа структуры деревьев и сравнения алгоритмов.

Для построения дерева решений была сформирована обучающая выборка, включающая 14 примеров, описывающих характеристики потенциальных поставщиков. Каждый объект определяется четырьмя атрибутами: «Цена», «Качество», «Срок поставки» и «Надёжность». Целевым атрибутом является признак «Решение», отражающий итоговый выбор поставщика («Да»/«Нет»). Структура выборки представлена в таблице ниже.

dataset.cpp M | supplier_dataset.csv

data > supplier_dataset.csv

Цена	Качество	Срок поставки	Надёжность	Решение
низкая	высокое	быстрая	высокая	Да
средняя	высокое	нормальная	высокая	Да
высокая	высокое	быстрая	высокая	Нет
низкая	среднее	медленная	средняя	Да
низкая	низкое	медленная	низкая	Нет
средняя	среднее	нормальная	средняя	Да
высокая	среднее	медленная	средняя	Нет
средняя	низкое	быстрая	высокая	Нет
низкая	высокое	медленная	средняя	Да
средняя	высокое	медленная	низкая	Нет
высокая	высокое	нормальная	средняя	Нет
низкая	среднее	нормальная	высокая	Да
средняя	среднее	быстрая	низкая	Нет
низкая	низкое	нормальная	средняя	Нет

Рисунок 1 — Выборка для построения дерева решений

Далее представлено визуализаций по проекту.

На рисунке 2 представлена визуализация графа алгоритма ID3.

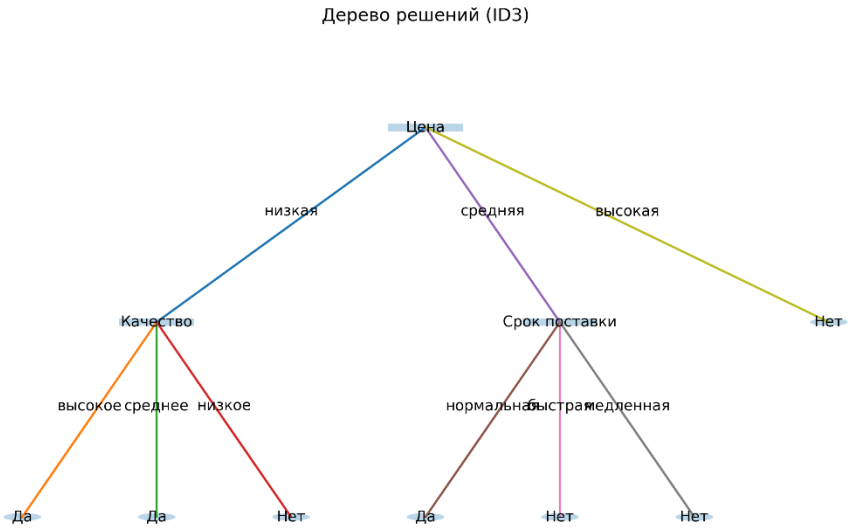

Рисунок 2 — Дерево решений, построенное алгоритмом ID3

На рисунке 3 представлена визуализация графа алгоритма C4.5

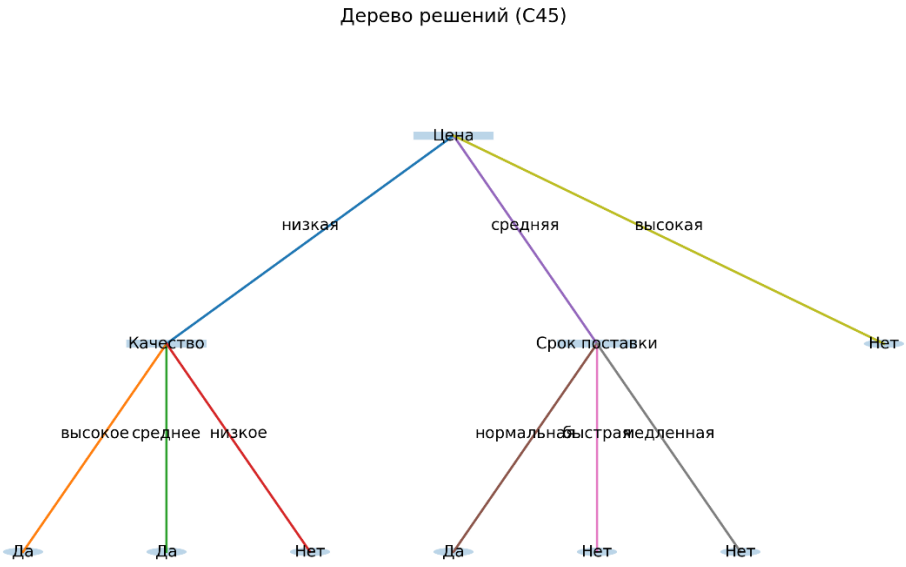

Рисунок 3 — Дерево решений, построенное алгоритмом C4.5

На рисунке 4 представлена визуализация графа алгоритма CART

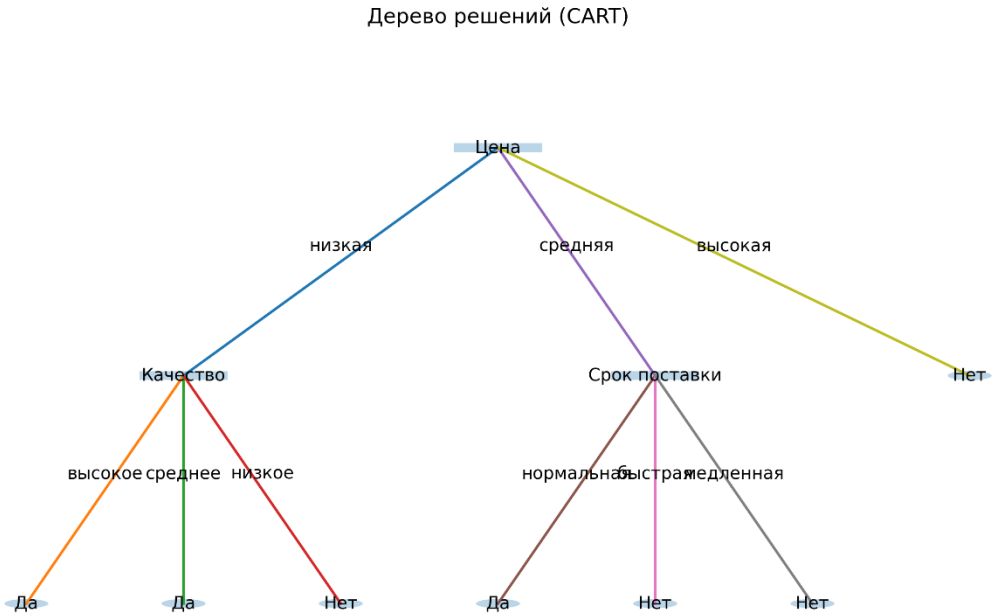

Рисунок 4 — Дерево решений, построенное алгоритмом CART

На рисунке 5 представлена визуализация графа алгоритма CHAID

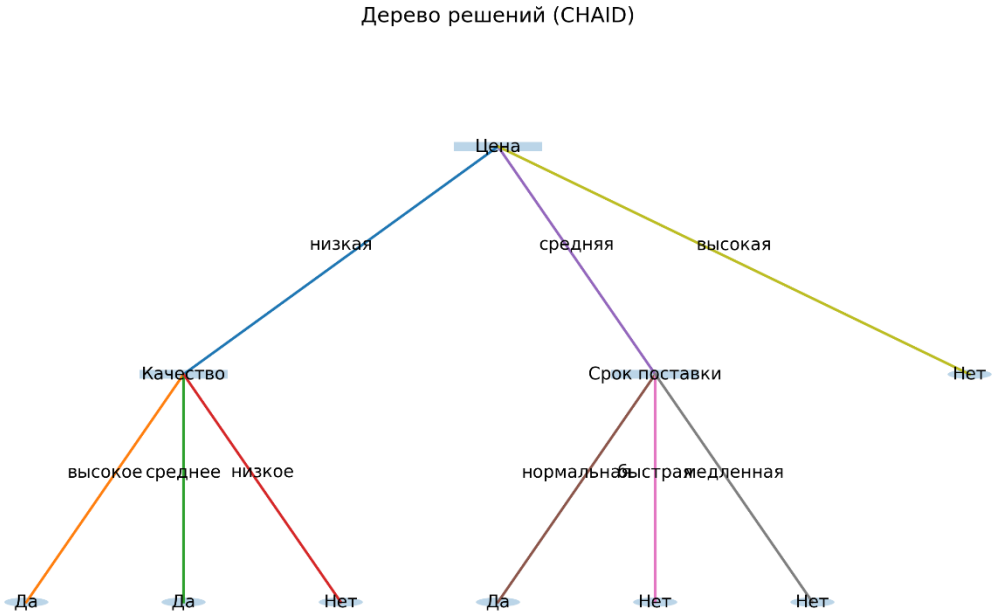

Рисунок 5 — Дерево решений, построенное алгоритмом CHAID

В рамках эксперимента были построены четыре дерева решений на одной и той же обучающей выборке. Модели ID3, C4.5, CART и CHAID различаются используемыми критериями разбиения, что влияет на структуру итоговых деревьев. Во всех алгоритмах корневым признаком стала **Цена**, что подтверждает её наибольшую информативность в задаче выбора поставщика.

ID3 и C4.5 дают похожую структуру дерева, однако C4.5 формирует более устойчивые разбиения.

Вывод: В ходе работы были построены деревья решений четырьмя алгоритмами — ID3, C4.5, CART и CHAID — на обучающей выборке по задаче выбора поставщика. Все алгоритмы корректно обработали данные и выделили одинаково значимый признак в корне дерева, что подтверждает устойчивость модели на небольшой выборке. Полученные результаты и сравнение структур деревьев показали, что методы отличаются критериями разбиения, но дают близкие решения. Это подтверждает применимость деревьев решений для задач классификации и принятия решений.