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

ОТЧЕТ

по практической работе 3

«Классификация с помощью дерева решений»

по дисциплине

«Технологии аналитической обработки информации»

Выполнил:_____
студент группы КЭ-403
О.С. Мазжухин

Проверил:_____
преподаватель
А.И. Гоглачев
Дата: _____
Оценка: _____

Челябинск – 2025

Формулировка задания

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 наборе Census Income (данные о результатах переписи населения, в т.ч. о годовом доходе -- ниже или выше $50000: скачать обучающую выборку в формате CSV, тестовую выборку в формате CSV, скачать описание). 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 наборе 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.

Гиперссылка на каталог репозитория с исходными текстами, наборами данных и другими материалами:

https://github.com/LN4rkot1k/informationProcessing

Визуализация

Визуализация построенных деревьев решений для разных соотношений разбиения данных на обучающую и тестовую выборки представлена на

рисунках 1-4:

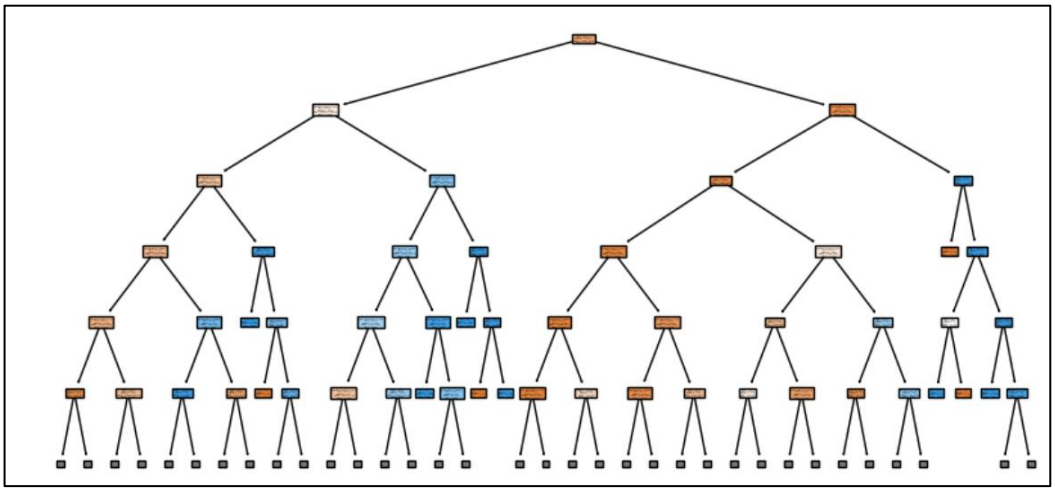

Рисунок 1 – Дерево решений с разбиением 60% на 40%

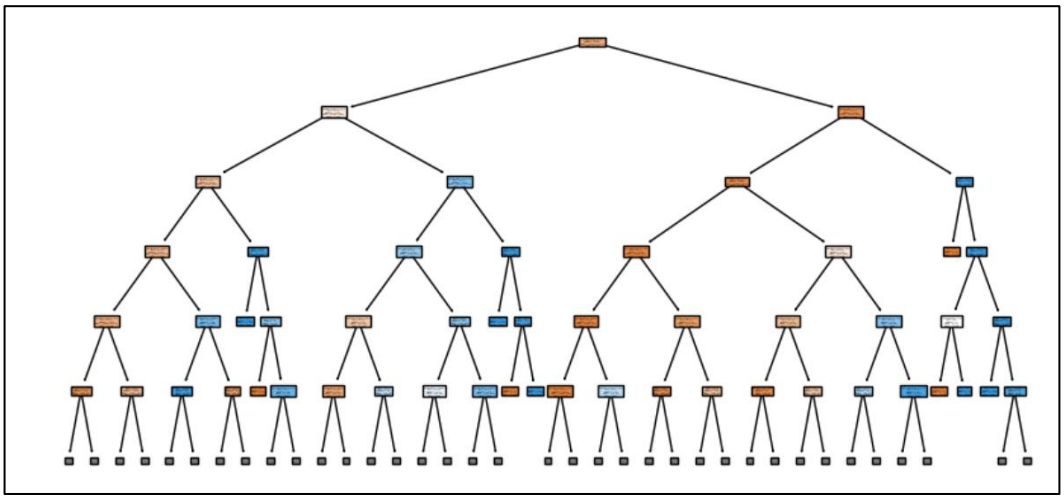

Рисунок 2 – Дерево решений с разбиением 70% на 30%

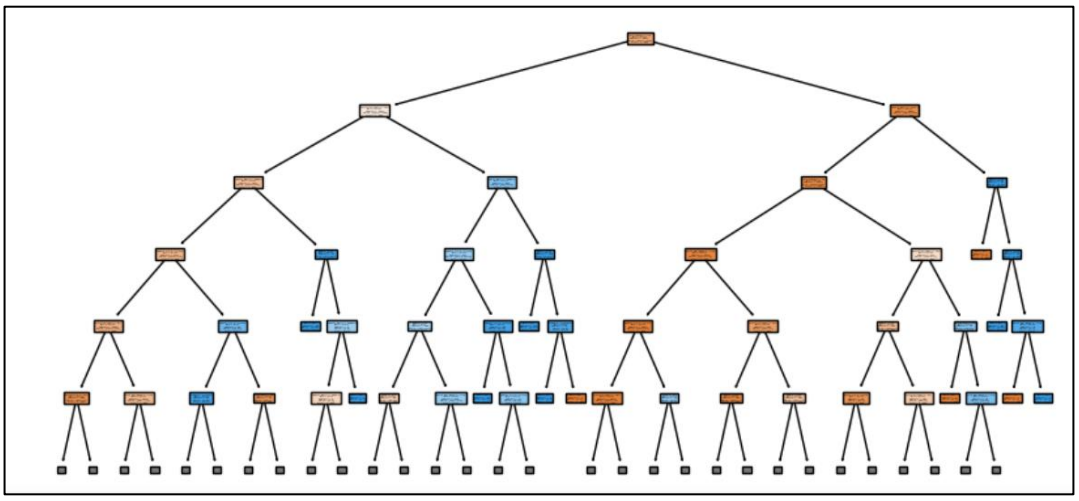

Рисунок 3 – Дерево решений с разбиением 80% на 19%

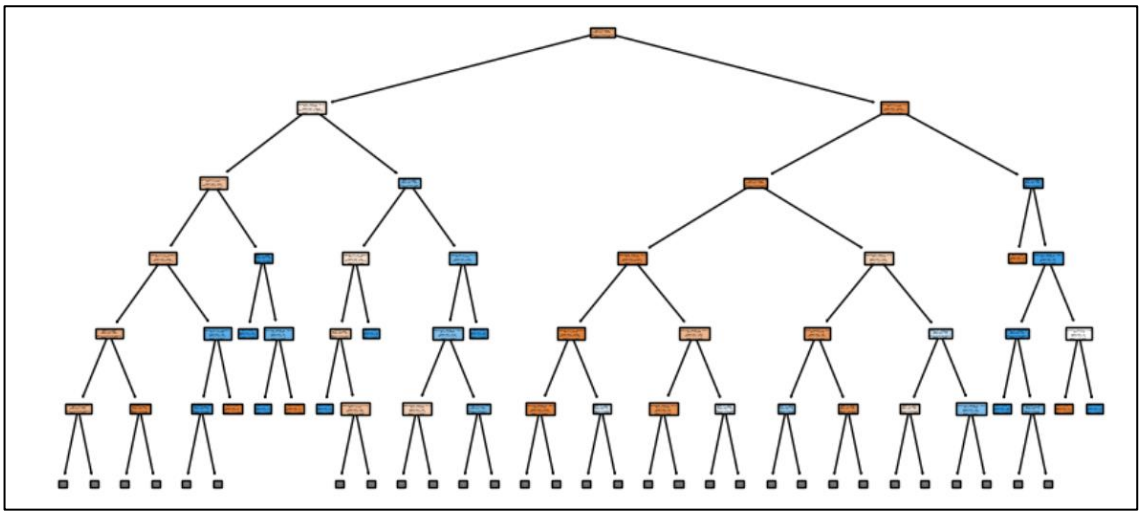

Рисунок 4 – Дерево решений с разбиением 90% на 9%

Чем больше тренинговая выборка, тем больше строк данных анализирует дерево, соответственно, результат обучения модели будет более точным. Однако, при слишком большом количестве тренинговых данных модель может легко переобучиться, а при слишком маленьком количестве данных, наоборот, переобучиться.

Визуализация зависимости показателей качества классификации от соотношения мощностей обучающей и тестовой выборок для заданного набора данных представлена на рисунке 5.

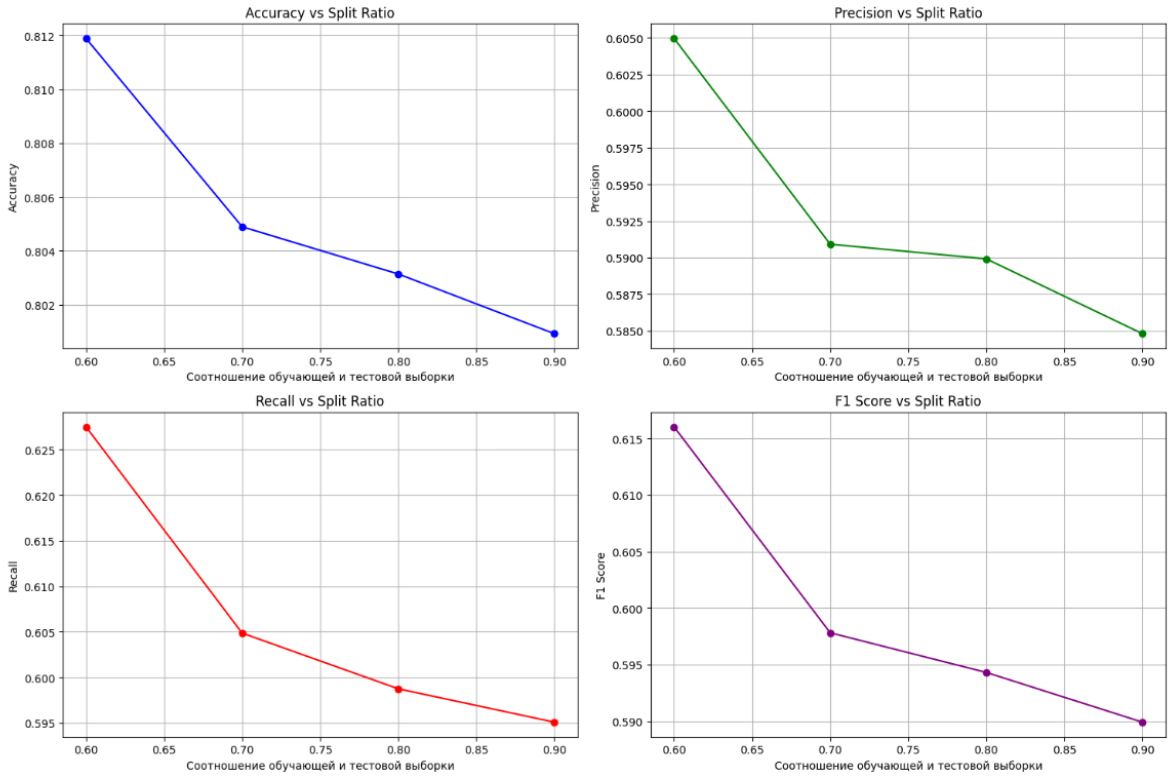

Рисунок 5 – Зависимость метрик от соотношений выборок

Глядя на графики, можем заметить, что значение accuracy остается высоким, что говорит о том, что модель хорошо классифицирует данные, несмотря на увеличение данных. Однако показатели остальных метрик снижаются. Это может говорить о том, что модель начинает хуже классифицировать людей с доходом больше 50К. Также при увеличении обучающей выборки модель стремится к переобучению.