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-404
А.Ю. Емельянова

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:

1) построенные деревья решений для заданного набора данных;

2) 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/Sun1ess-sea/Technologies-of-analytical-information-processing

**Визуализация**

В данной практической работе были построены и визуализированы некоторые деревья решений с разными соотношениями разбиения данных на

обучающую и тестовую выборки. Они представлены на рисунках 1-4.

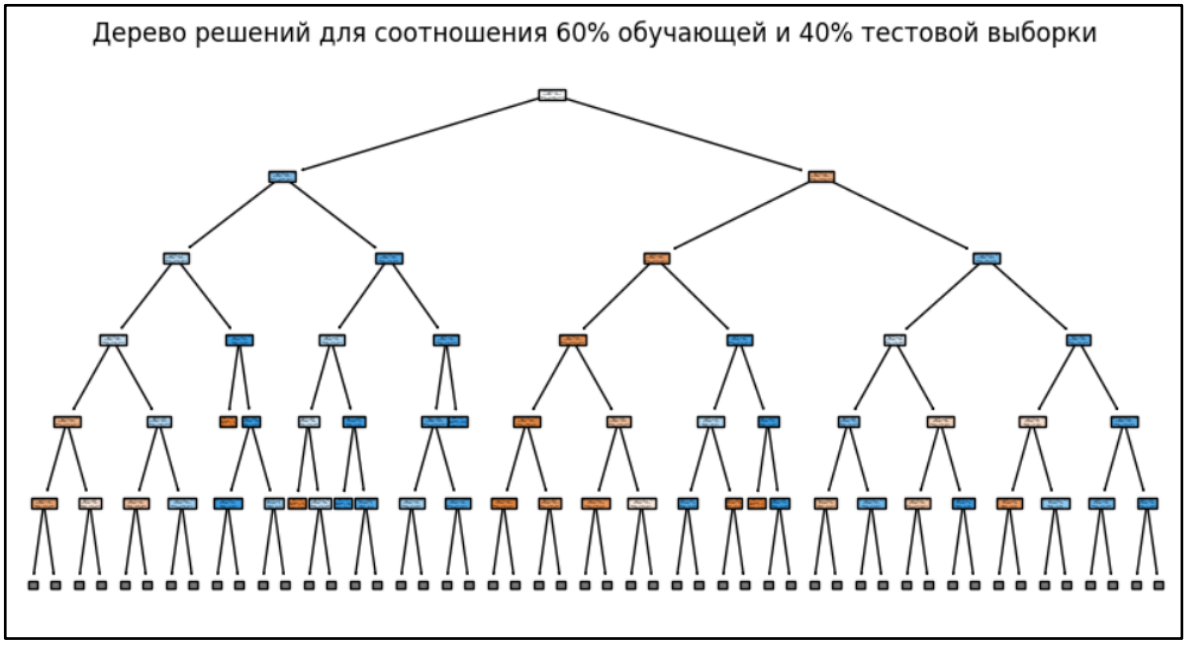


Рисунок 1 – Дерево решений с соотношением 60% на 40% обучающей и тестовой выборки

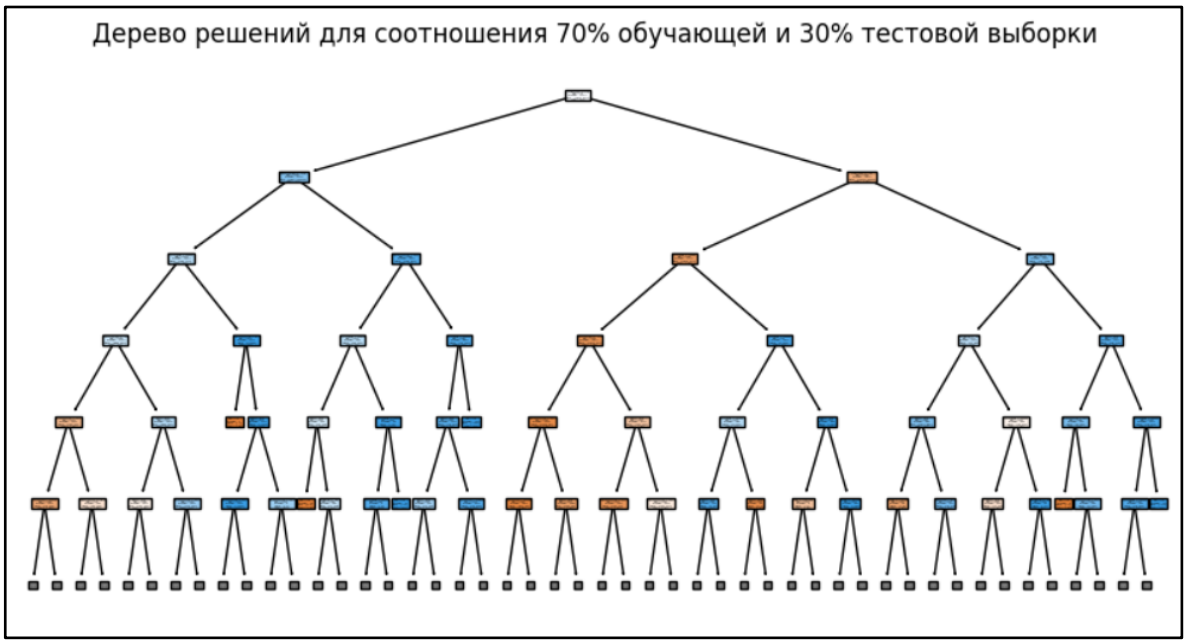


Рисунок 2 – Дерево решений с соотношением 70% на 30% обучающей и тестовой выборки

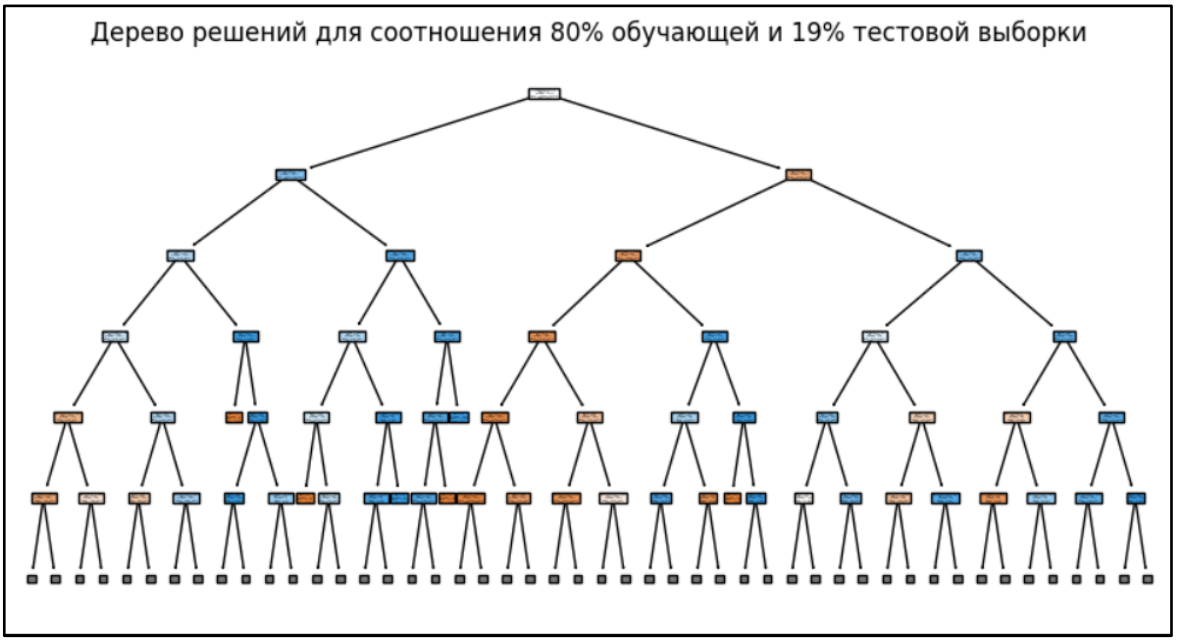


Рисунок 3 – Дерево решений с соотношением 80% на 19% обучающей и тестовой выборки

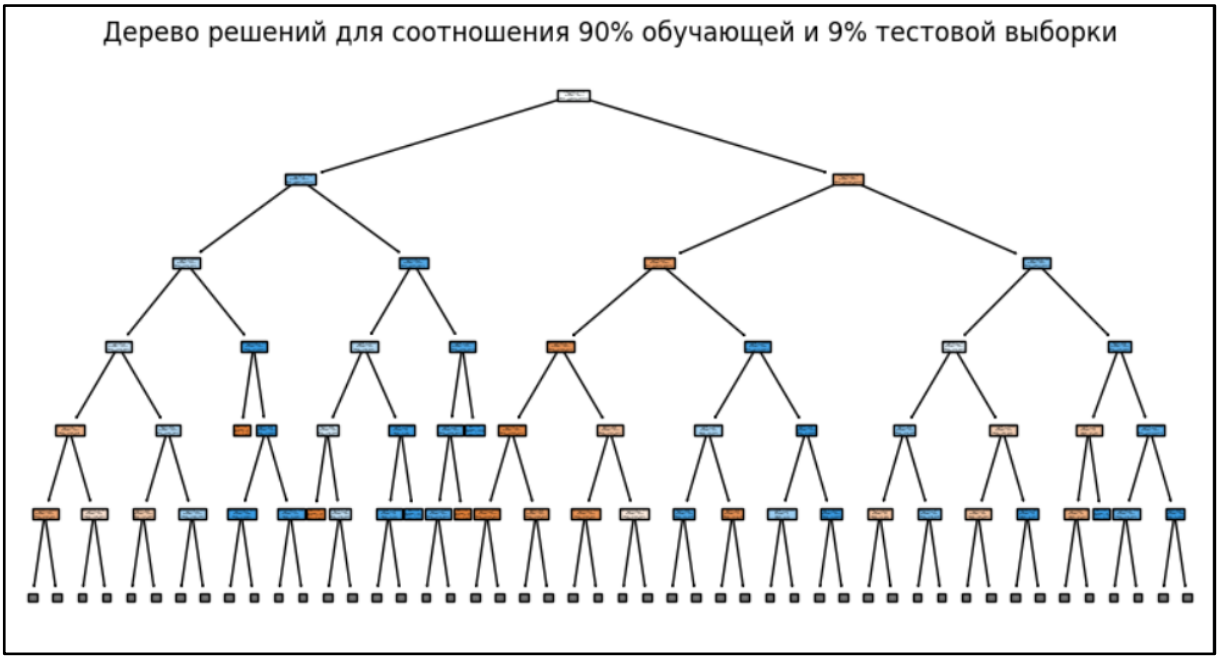


Рисунок 4 - Дерево решений с соотношением 90% на 9% обучающей и тестовой выборки

Для анализа полученных данных сравним показатели метрик при выборке 60% на 40% и при выборке 90% на 9%.

Выборка 60% на 40% даст результаты, которые можно увидеть в списке ниже:

1) Accuracy: 0.8759
2) Precision: 0.8418
3) Recall: 0.9259
4) F1 score: 0.8819

Выборка 90% на 9% даст результаты, которые можно увидеть ниже:

1) Accuracy: 0.9159
2) Precision: 0.8794
3) Recall: 0.9602
4) F1 score: 0.9181

По визуализированным данным и их числовым показателям можно сделать вывод, что чем больше было данных в обучающей выборке, тем выше значения метрик, так как при увеличении обучающей выборки модель лучше обучается, а значит выдает более точные результаты. Модель видит больше примеров, на которых можно учиться, и точнее распознаёт закономерности в данных. Однако, при их слишком большом количестве есть шанс переобучения модели, а при маленьком, наоборот, недообучения.

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

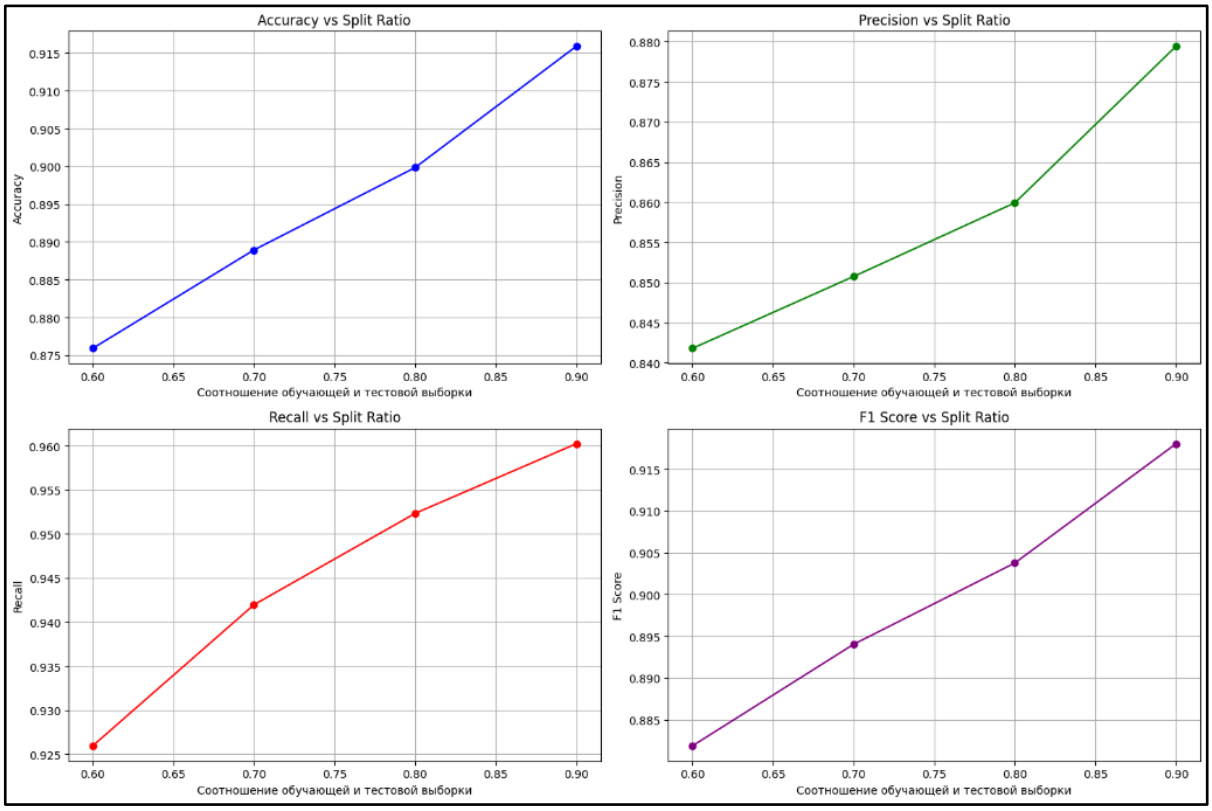


Рисунок 5 – График зависимости метрик от соотношений выборок

Визуализация значений метрик показывает, что чем больше обучающих данных получает модель, тем больше ее точность в дальнейшем, а значит и показатели увеличатся. И наоборот, при небольшом количестве данных для обучающей выборки модель будет иметь худшие показатели, чем могла бы.

Большее количество обучающих данных выдает более высокие показатели метрик, однако необходимо не достигать переобучения.