Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

ИРКУТСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ

ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Институт информационные системы и анализ данных

Центр программной инженерии

**ОТЧЁТ**

к лабораторной работе №5 по дисциплине:

|  |
| --- |
| «Методы анализа данных» |
| Построение деревьев решений для задачи классификации |
| наименование темы |

Выполнилстудент ИСТб-21-1 Д.И. Морозов

номер группы подпись И. О. Фамилия

дата

Проверил Доцент Е.А. Осипова

Должность подпись И. О. Фамилия

дата

Иркутск – 2023 г.

# 1 Постановка задачи

А. Выбрать среду программирования для языка Python.

Б. Получить набор данных из 200 наблюдений с параметрами, согласно номеру варианта (табл. 1.1). Построить деревья решений и выполнить их анализ. При этом необходимо:

* разделить исходную выборку на две равные части: для обучения и для тестирования;
* обучить дерево решений и вывести его на экран;
* используя функцию рисования решающих поверхностей из лабораторной работы № 3, вывести на экран решающие поверхности и данные для построенного дерева решений для каждой пары признаков;
* оценить точность модели на тестирующей выборке;
* в случае, если точность модели менее 0,9, изменяя значение параметров criterion, max\_depth, min\_samples\_split, min\_samples\_leaf попытаться ее повысить.

В. Выполнить анализ полученных на каждом этапе результатов и оформить отчет по лабораторной работе.

Вариант 1.

Количество кластеров – 2.

Количество признаков – 5.

Стандартное отклонение - 3.

# 2 Описание исходных данных.

В качестве набора данных была создан датасет со случайными значениями.

Получим исходные данные для числа классов N=2:

N=2, D=5

X, y = make\_blobs(n\_samples=100, centers=N, n\_features=D, random\_state=None, cluster\_std=3)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.5)

df=pd.DataFrame(X\_train)

df['Class']=y\_train

3 Выполнение работы.

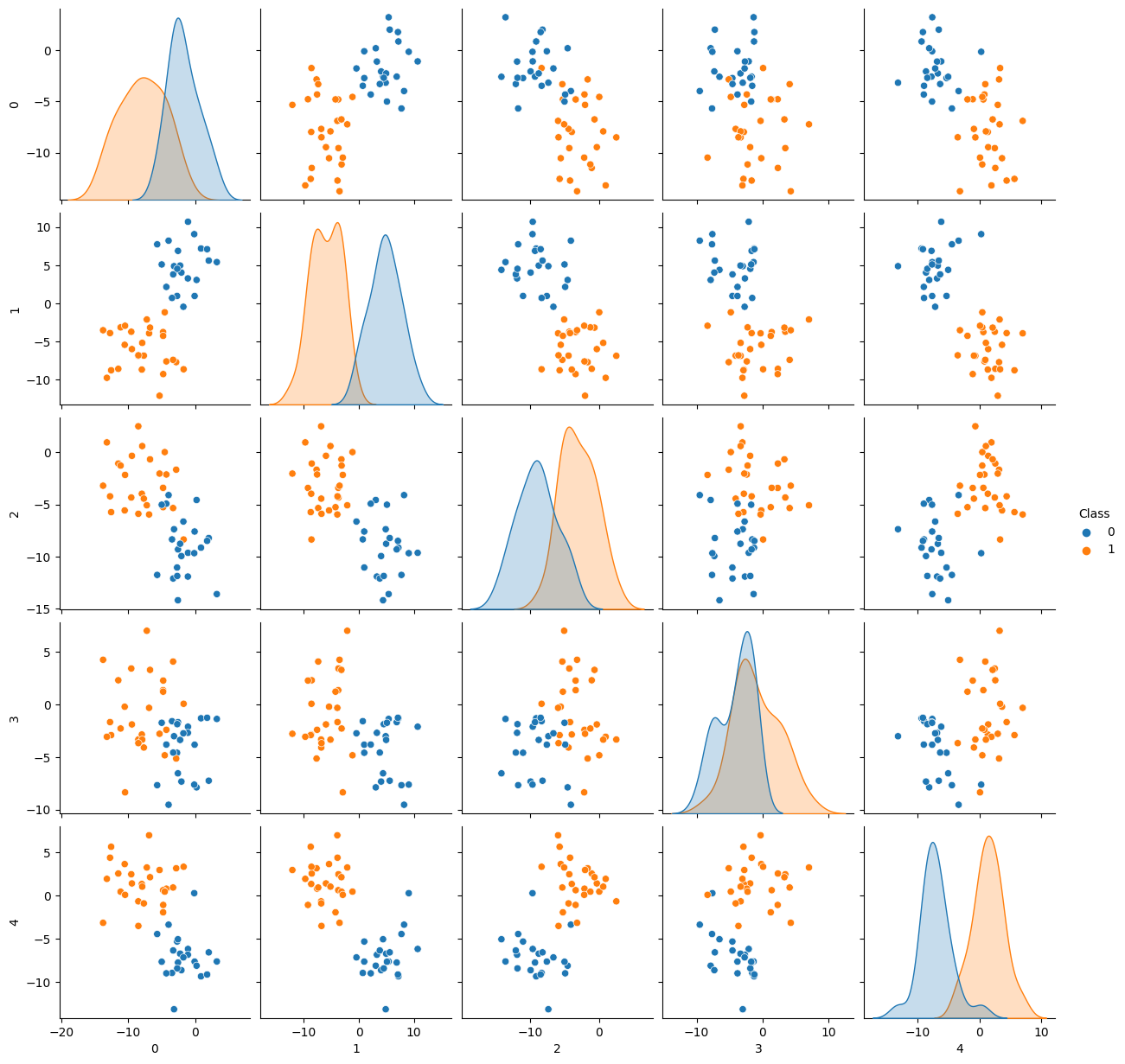
Вывод диаграмм рассеяния.   


Рисунок 1 – Диаграммы рассеяния.

Построение деревьев решений с использованием критерия Джини, минимальным количеством ля разделения равным 15 и глубиной дерева не больше 5.

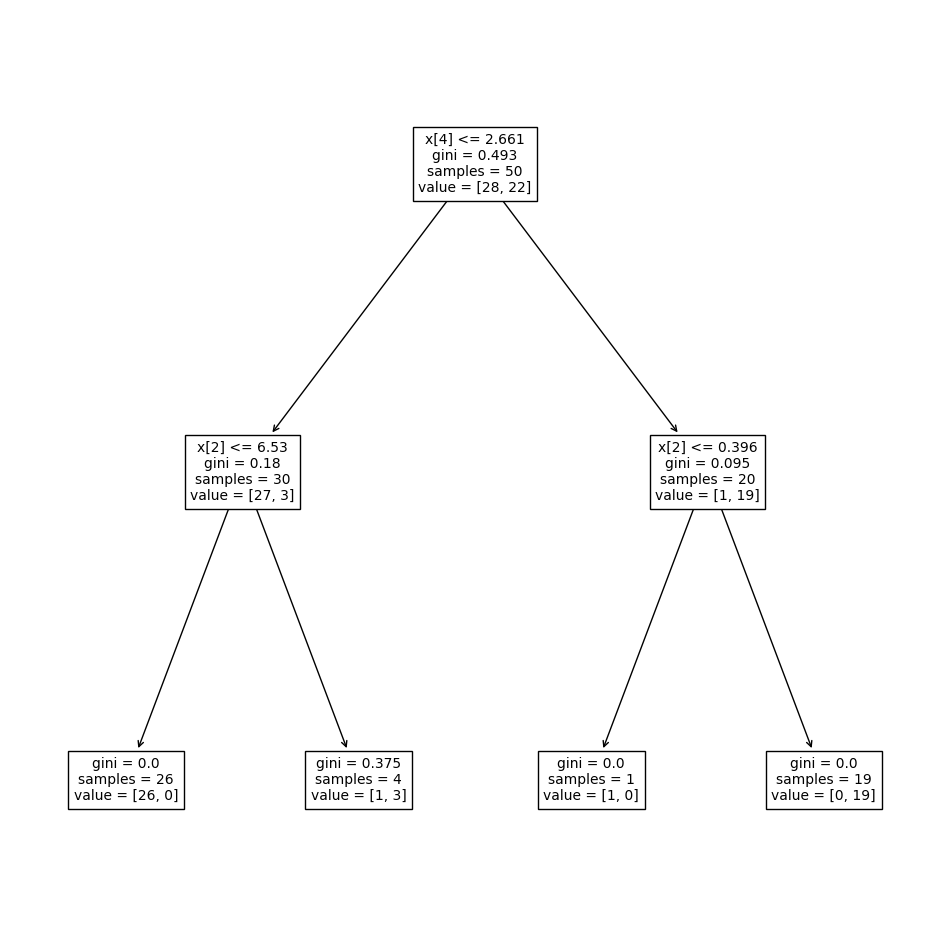


Рисунок 2 –Деревья решений.

4 Построение решающих поверхностей.

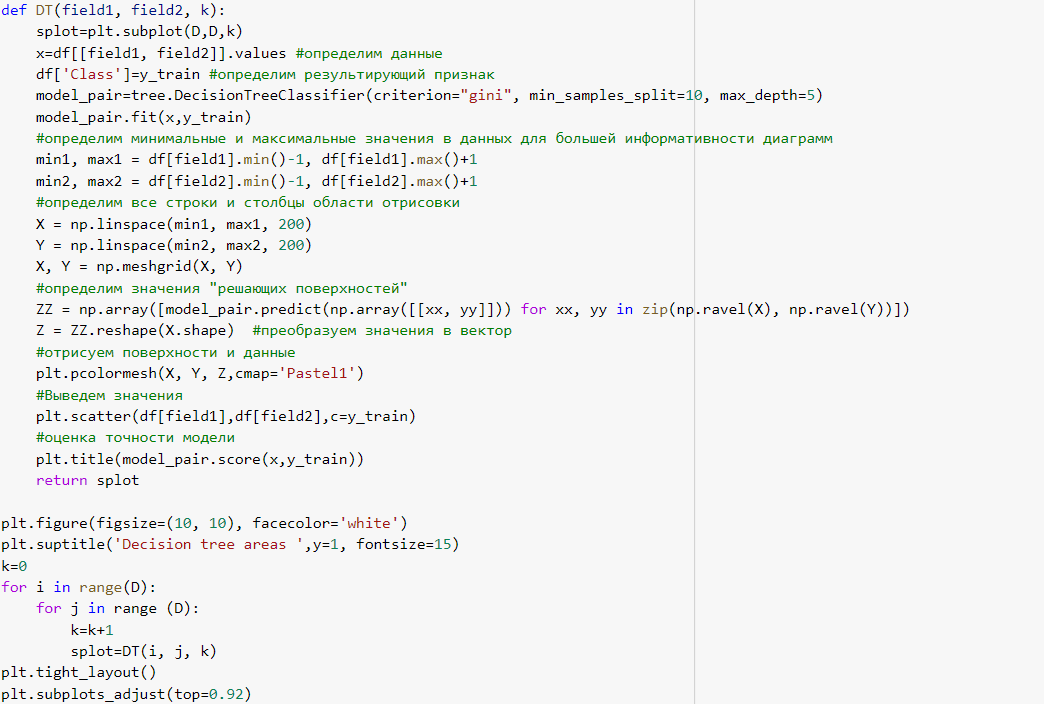
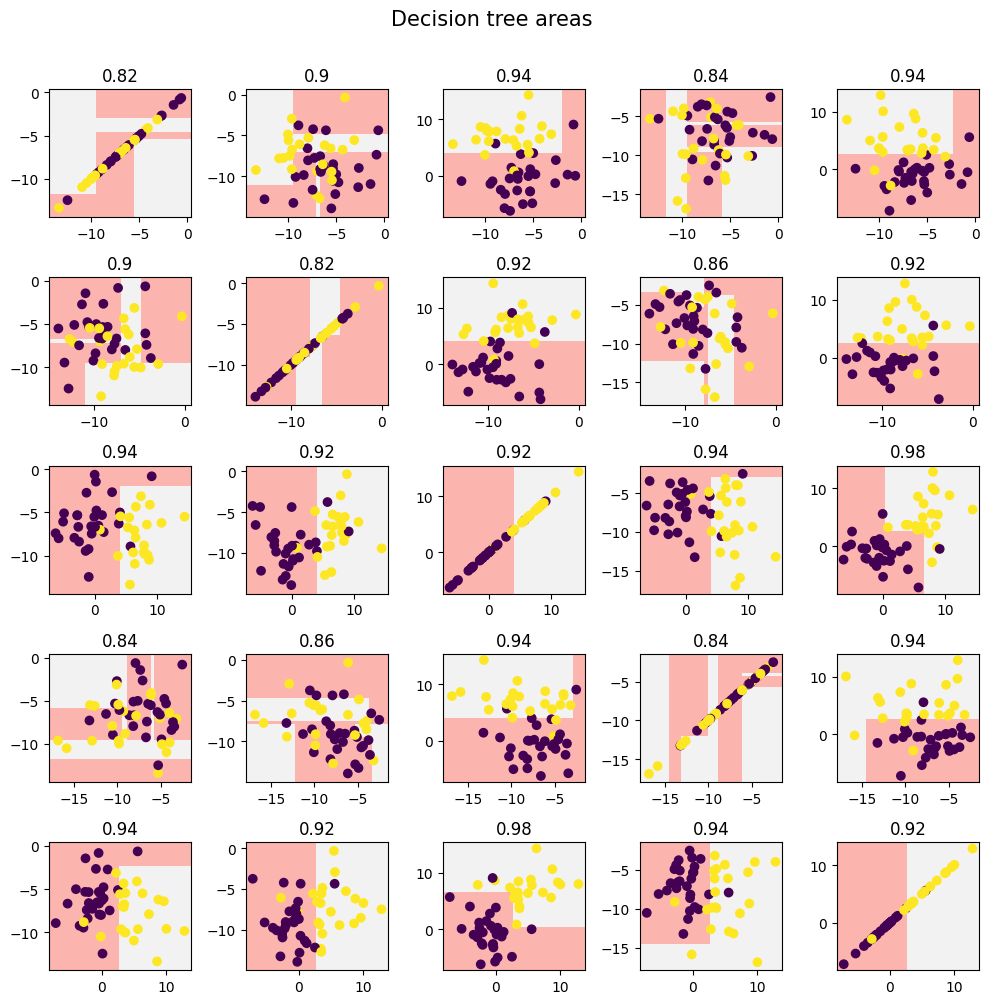


Рисунок 3 –Код решающих поверхностей.



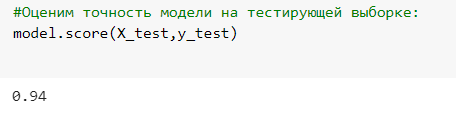


Рисунок 6 – Оценка точности.

5 Вычисление матриц ошибок и оценка точности.

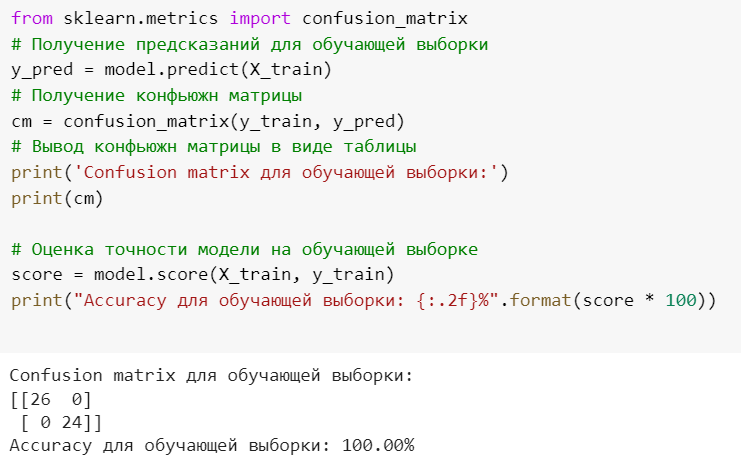


Рисунок 7 – Матрица ошибок и оценка точности для обучающей выборки

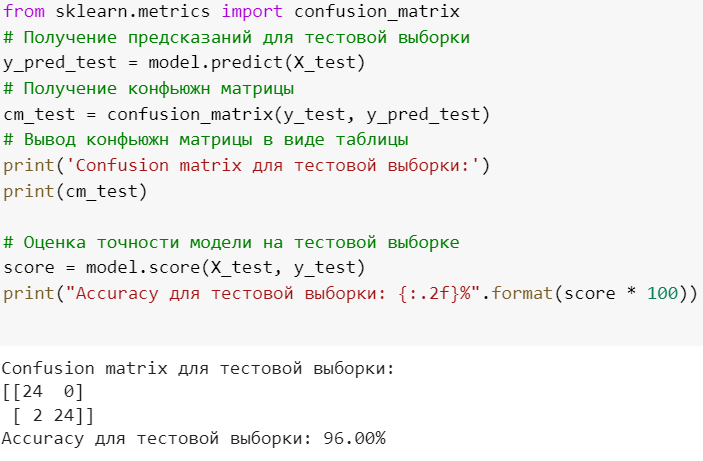


Рисунок 8 – Матрица ошибок и оценка точности для тестовой выборки

5 Лес решений.

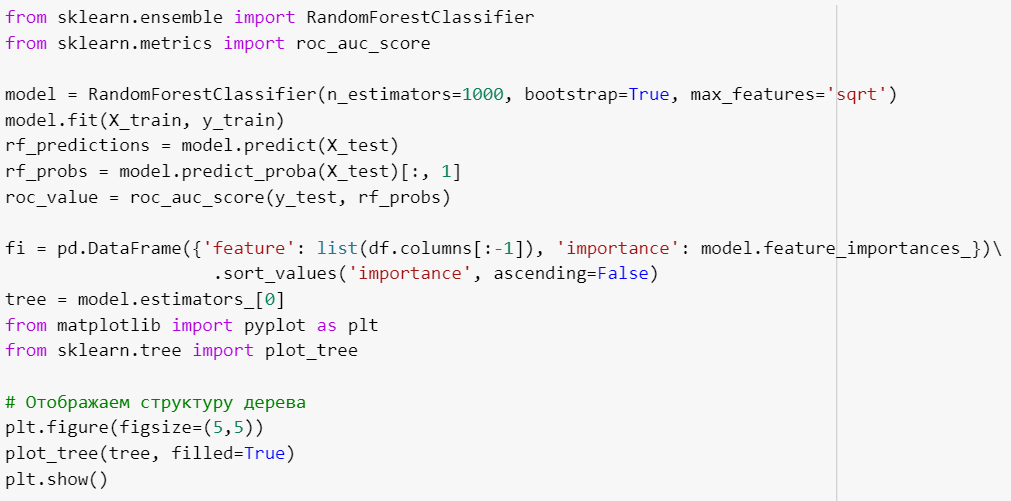


Рисунок 9 – Код леса решений

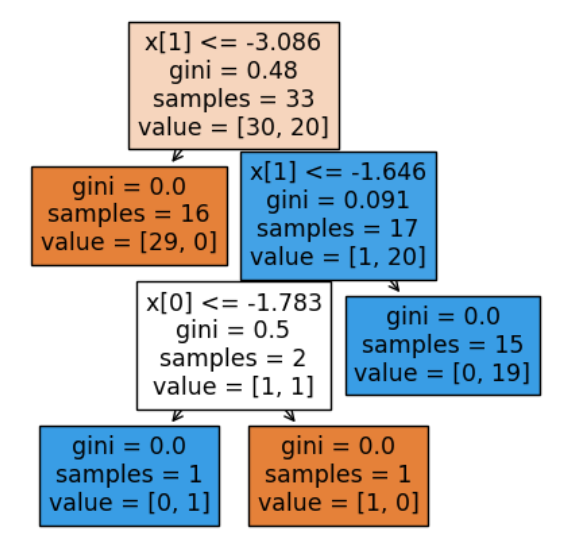


Рисунок 10 – Вид леса решений

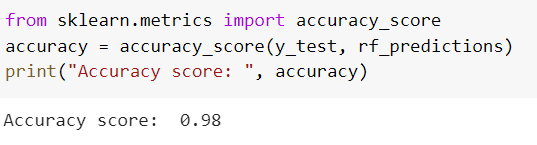


Рисунок 10 – Точность модели

Метод леса решений является более точным по сравнению с простыми деревьями решений. Он позволяет увеличить точность модели за счет создания нескольких деревьев и объединения их результатов. Кроме того, метод леса решений обладает лучшей устойчивостью к переобучению, что позволяет достигать более стабильных результатов при работе с различными наборами данных.

10 Результаты работы

При изучении задачи классификации и построении деревьев решений были получены навыки, которые позволяют узнавать закономерности и определять классы данных на основе их признаков. Для этого использовались обучающие данные, которые описывают каждый класс, а также тестовые данные для проверки точности модели. Одним из главных компонентов исследования была возможность создания оптимальной модели дерева решений с помощью выбора наилучших параметров, таких как глубина дерева или критерий разделения узлов. Также было проанализировано, как классификационная модель может быть использована в различных областях для выявления закономерностей и определения классов данных. В целом, проведенное исследование имеет значение, так как может быть применено в многих отраслях, включая финансы, маркетинг, медицину и другие, для классификации данных и выявления потенциальных зависимостей между признаками.

# Список использованных источников

1 Петров А.В., Бучнев О.С. Лабораторный практикум — Иркутск: ФГБОУ ВО ИРНИТУ, 2022. — 114 с.