Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

**Иркутский национальный исследовательский   
технический университет**

|  |
| --- |
| Институт информационных технологий и анализа данных |
| наименование института |

|  |
| --- |
| Отчет по дисциплине |
| «Методы анализа данных» |
| по теме: |
| «Построение деревьев решений для задачи классификации» |
|  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Выполнил студент группы |  | АСУб-20-1 |  |  |  | Устюжанин В. М. |
|  |  | Шифр группы |  | Подпись |  | И.О. Фамилия |
| Проверил преподаватель |  |  |  |  |  | Осипова Е. А. |
|  |  |  |  | Подпись |  | И.О. Фамилия |

Иркутск 2022 г.

**Содержание**

[Постановка задачи 3](#_Toc122681166)

[Выполнение лабораторной работы 4](#_Toc122681167)

[Вывод 11](#_Toc122681168)

# 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.

Варианту 7 соответствуют следующие параметры:

Количество кластеров N – 3.

Количество признаков D – 5.

Дисперсия значений cluster\_std – 4.

# Выполнение лабораторной работы

Подключим библиотеки, необходимые для выполнения задания.

  
Рисунок 1 – Подключение библиотек

Получим исходные данные для числа кластеров N=3:

  
Рисунок 2 –

Построим диаграммы рассеяния:

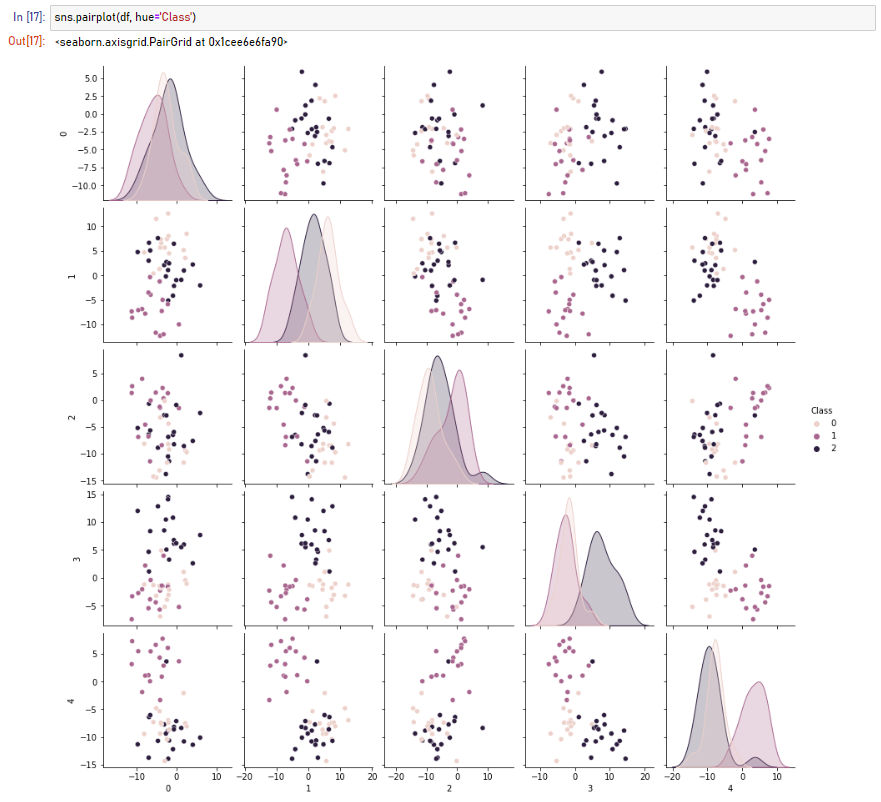


Рисунок 3 – Диаграммы рассеяния для 5 признаков

Обучим дерево решений:

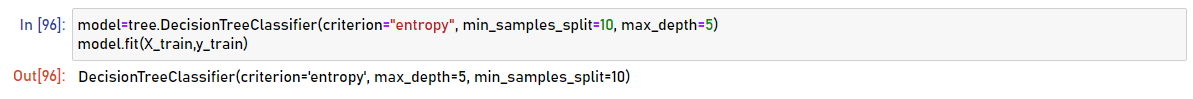


Рисунок 4 – Обучение диаграммы рассеивания

Выведем дерево решений:

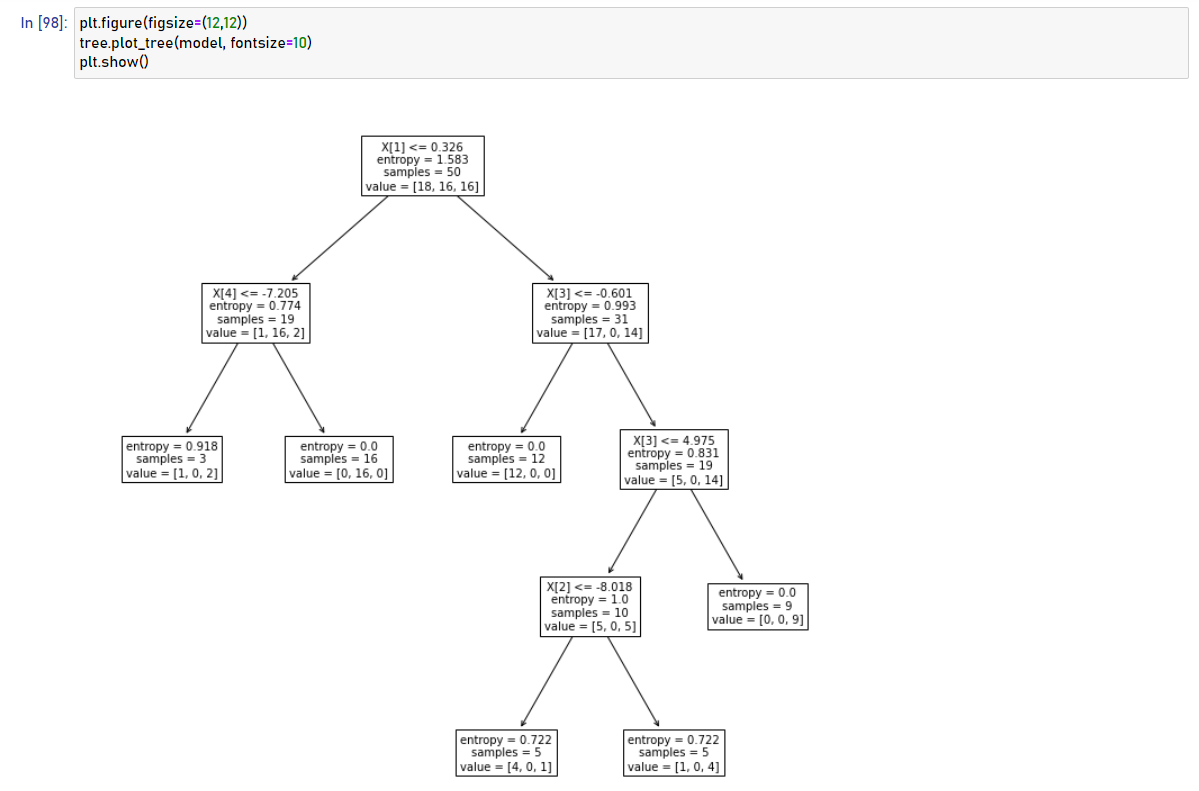


Рисунок 6 – Дерево решений

Немного модифицируем и используем функцию отрисовки решающих поверхностей из лабораторной работы №3.

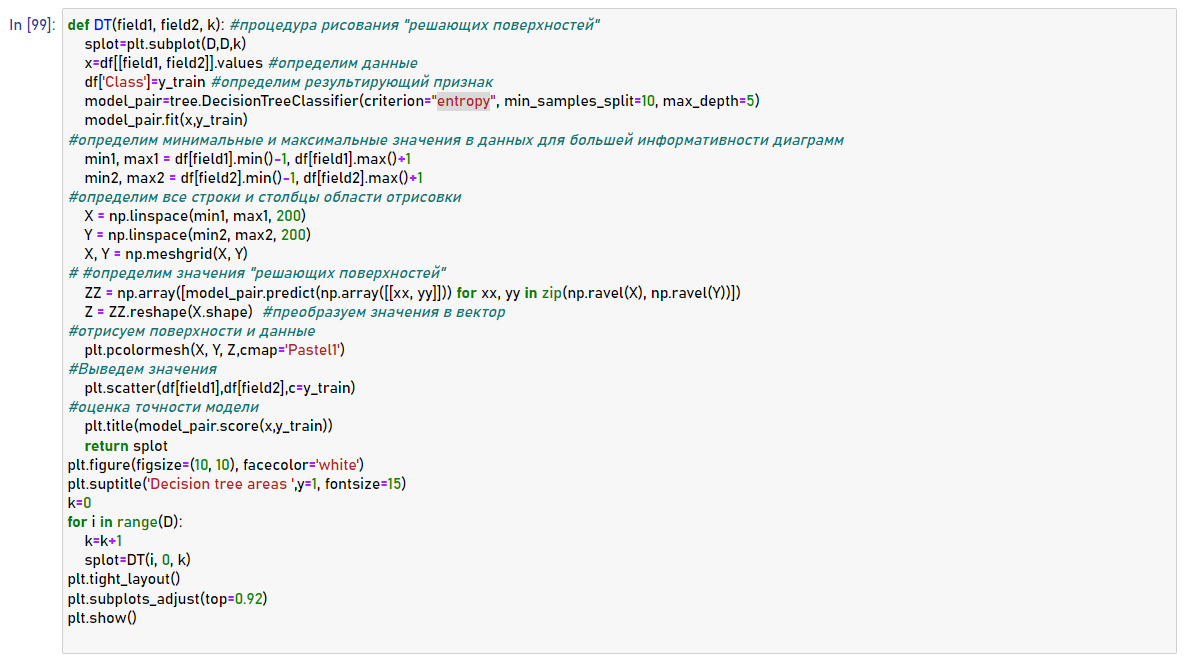


Рисунок 7 – Функция для отрисовки решающих поверхностей

Построим и выведем на экран деревья решений и решающие поверхности для каждой пары признаков:

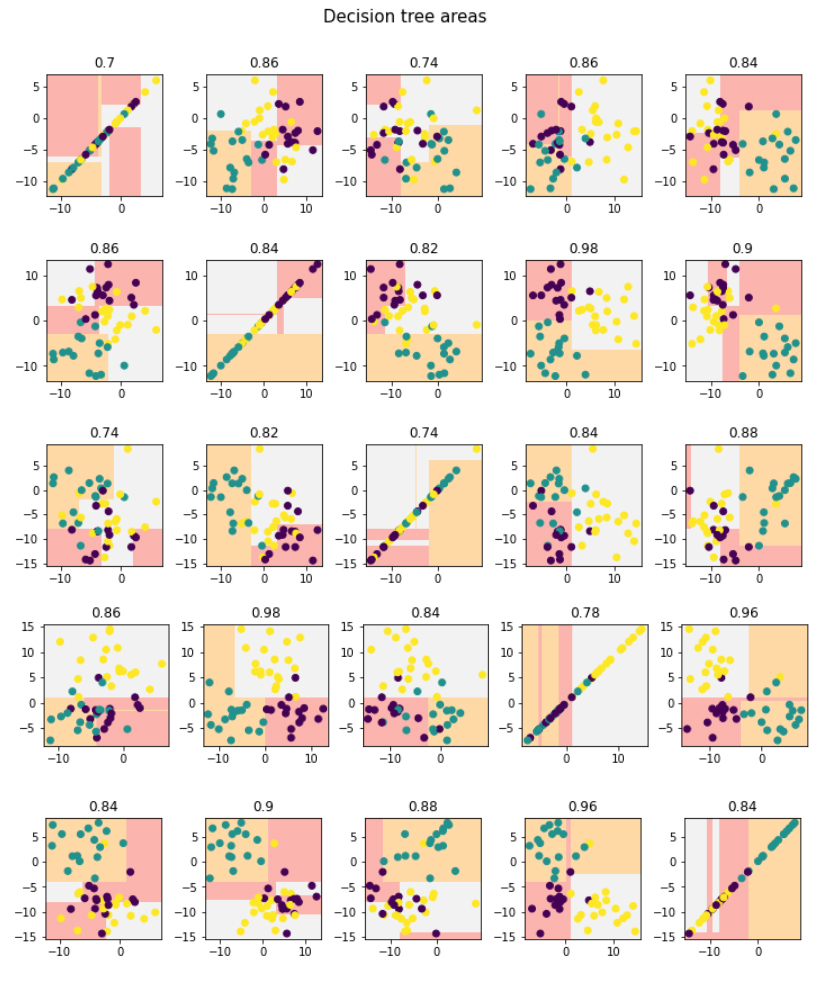


Рисунок 8 – Решающие поверхности

Оценим точность модели на тестирующей выборке:

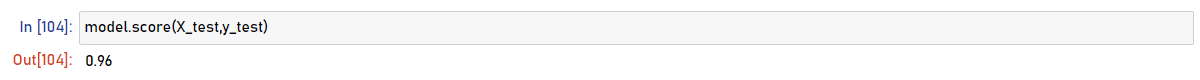


Рисунок 9 – Оценка точности

Точность модели можно считать оптимальной.

# Вывод

В процессе выполнения лабораторной работы №5 были ознакомлены основные понятия деревьев решений, функции для Python. Была построена модель дерева решений и определена и повышена его точность.