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: |
| «Задачи кластеризации и 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](#_Toc118469870)

[Подготовка данных для анализа 4](#_Toc118469871)

[Вывод 12](#_Toc118469872)

# Постановка задачи

Получить набор данных из 100 наблюдений с параметрами, согласно номеру варианта. С использованием Python провести кластерный и дискриминантный анализ данных. При этом необходимо:

* получить значения основных показателей описательной статистики и построить диаграммы рассеяния для всех признаков;
* выполнить стандартизацию данных;
* построить и вывести на экран дендрограмму;
* провести кластеризацию данных методом k-средних, вывести на экран результаты кластеризации; показать на этих же диаграммах средние значения для каждого кластера;
* построить диаграмму средних значений для каждого кластера и каждого измерения;
* создать инструмент для графического отображения результатов дискриминантного анализа – решающих поверхностей и данных;
* провести линейный дискриминантный анализ данных, для каждой пары признаков: вывести на экран решающие поверхности и данные, оценить точность;
* провести квадратичный дискриминантный анализ данных, для каждой пары признаков: вывести на экран решающие поверхности и данные, оценить точность; сравнить с точностью линейного дискриминантного анализа.

По варианту №10 необходимо сгенерировать данные с 4 кластерами, 2 признаками и дисперсии равной 1.5.

# Подготовка данных для анализа

Подключаем библиотеки, необходимые для выполнения работы.

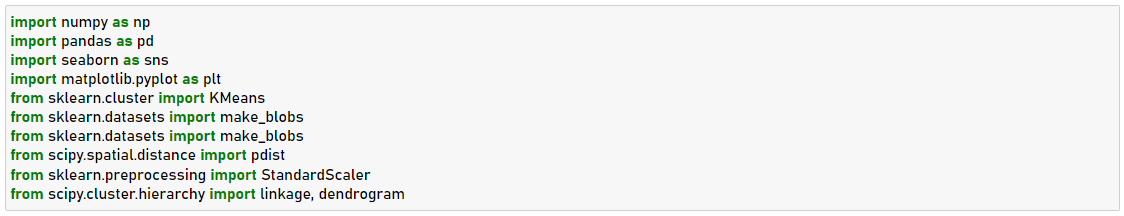


Рисунок 1 – Подключение библиотек в Python

Получаем исходные данные.

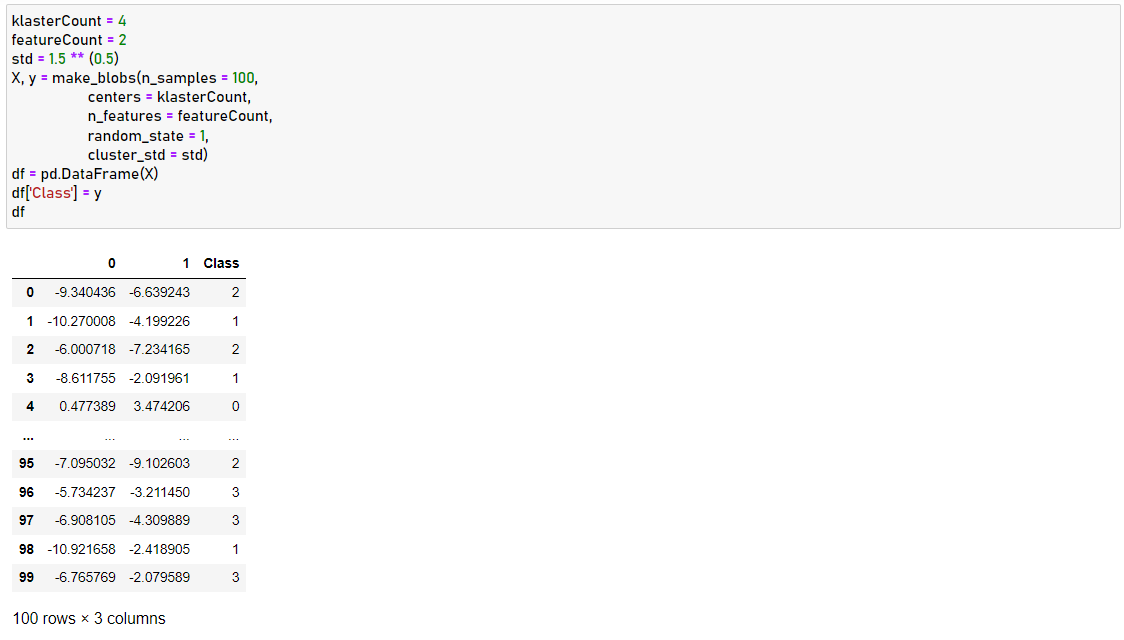


Рисунок 2 – Генерация исходных данных

Построим диаграммы рассеивания.

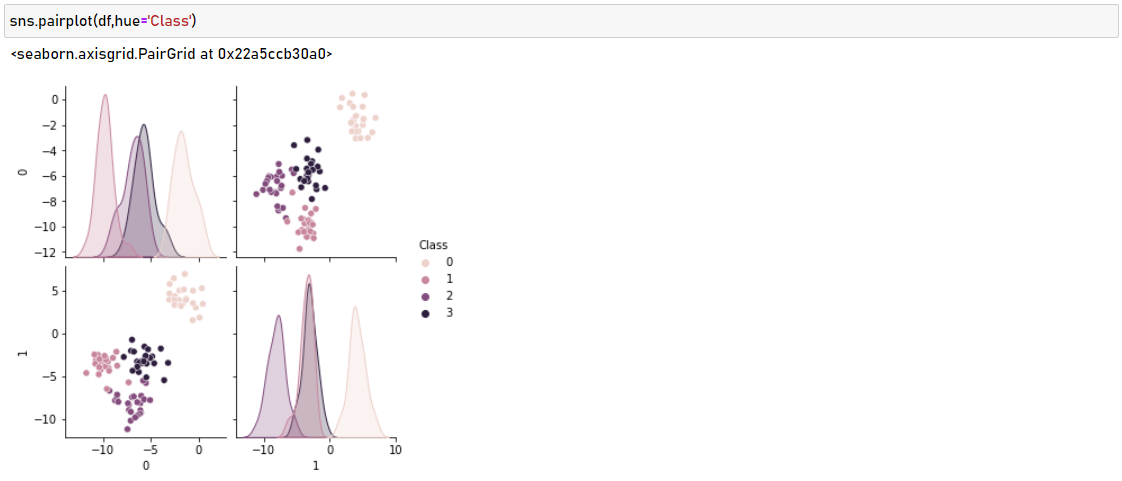


Рисунок 3 – Диаграммы рассеивания

Произведем нормализацию данных для возможности получать расстояния между объектами. Нормализация позволит избавиться от разности единиц измерений и приведет все признаки к безразмерным величинам.

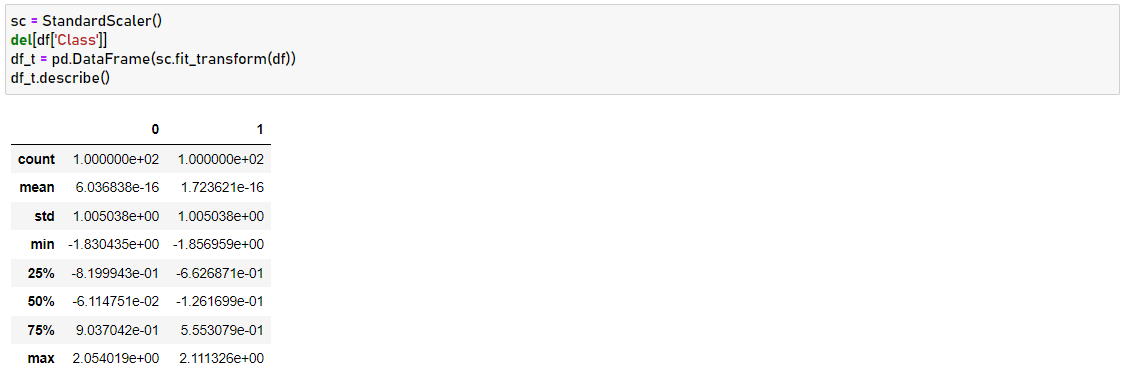


Рисунок 4 – Нормализация данных

Построим дендограмму. При этом будем использовать евклидовы расстояния и метод дальнего соседа.

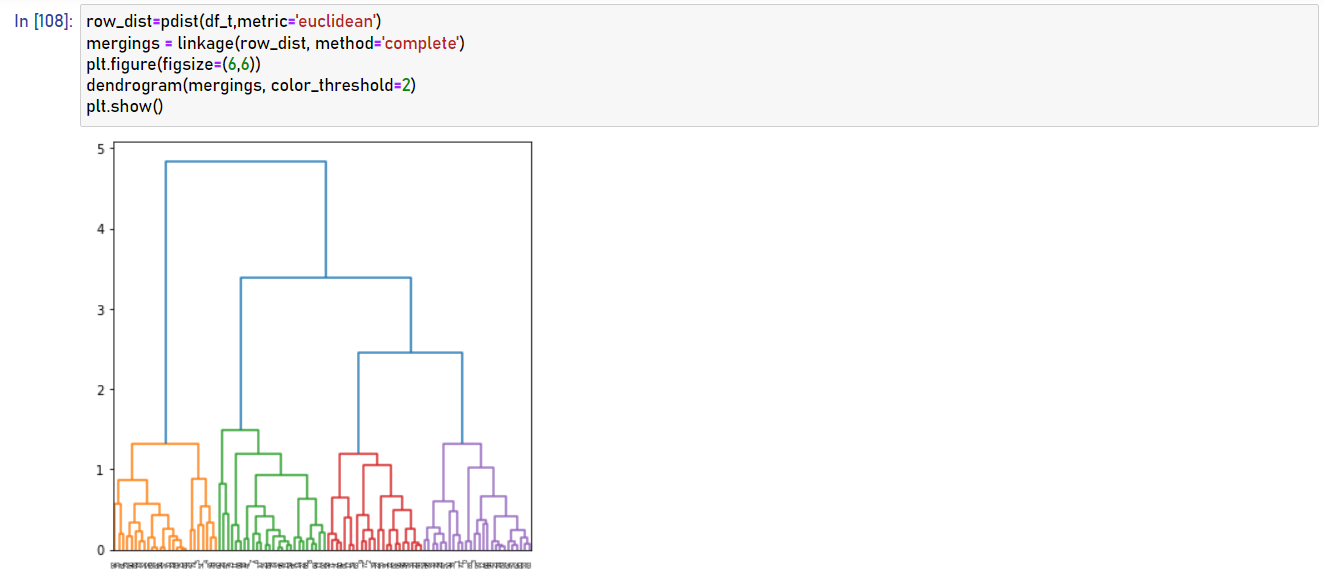


Рисунок 5 – Дендограмма

Проведем кластеризацию данных методом k-средних, выведем результаты и отметим на этих же диаграммах средние значения для каждого кластера.

Первоначально выбирается желаемое количество кластеров. Поскольку нам известно, что в нашем наборе данных есть 4 кластера, установим параметр модели **n\_clusters** равный четырем.

Теперь случайным образом из входных данных выбираются четыре элемента выборки, в соответствие которым ставятся четыре кластера, в каждый из которых теперь включено по одной точке, каждая при этом является центроидом этого кластера.

Далее ищем ближайшего соседа текущего центроида. Добавляем точку к соответствующему кластеру и пересчитываем положение центроида с учетом координат новых точек.  Алгоритм заканчивает работу, когда координаты каждого центроида перестают меняться. Центроид каждого кластера в результате представляет собой набор значений признаков, описывающих усредненные параметры выделенных классов.

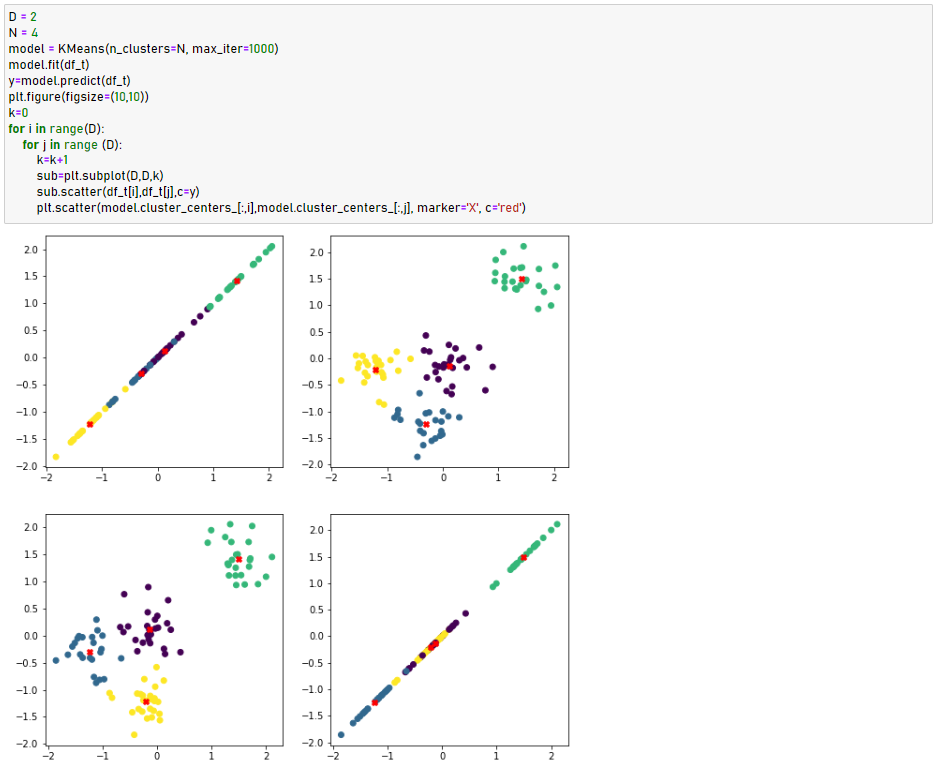
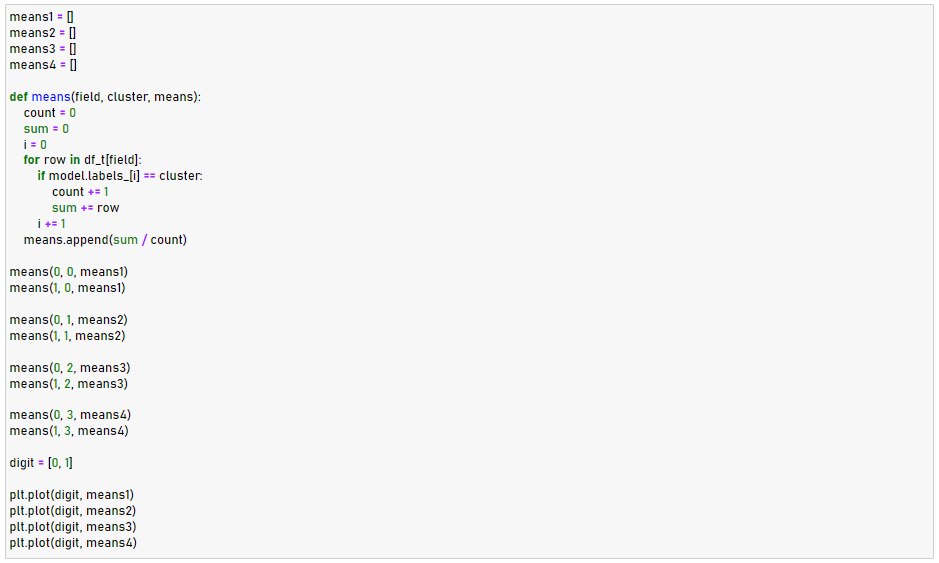


Рисунок 6 – Результаты кластеризации

Теперь построим диаграммы средних значений для каждого кластера по каждому признаку. График средних, где линии показывают масштабированные средние по каждому из измерений для каждого кластера.



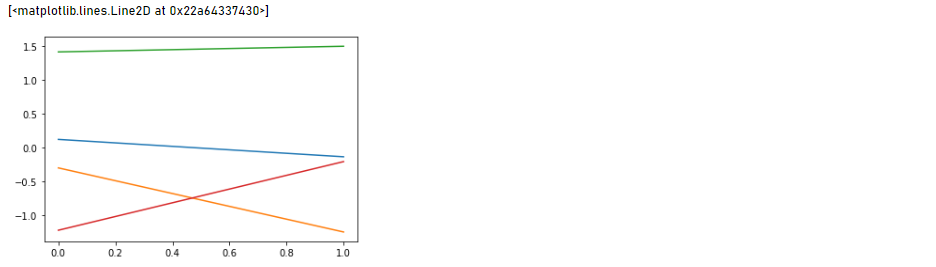


Рисунок 7 – Диаграмма средних

Дискриминантный анализ используется для принятия решения о том, какие переменные различают (дискриминируют) две или более возникающие совокупности (группы)

Используя следующую вспомогательную функцию, будем производить линейный и квадратичный дискриминантный анализ данных, для каждой пары признаков.

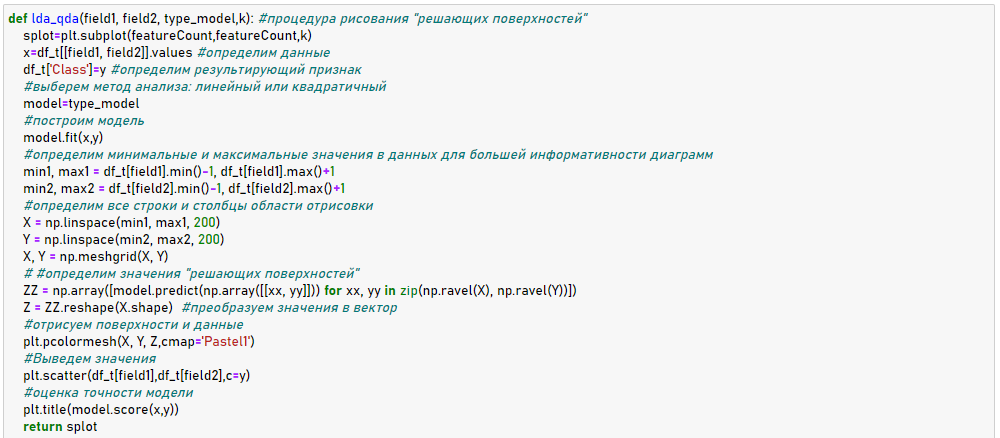


Рисунок 8 – Вспомогательная функция для проведения дискриминантного анализа данных

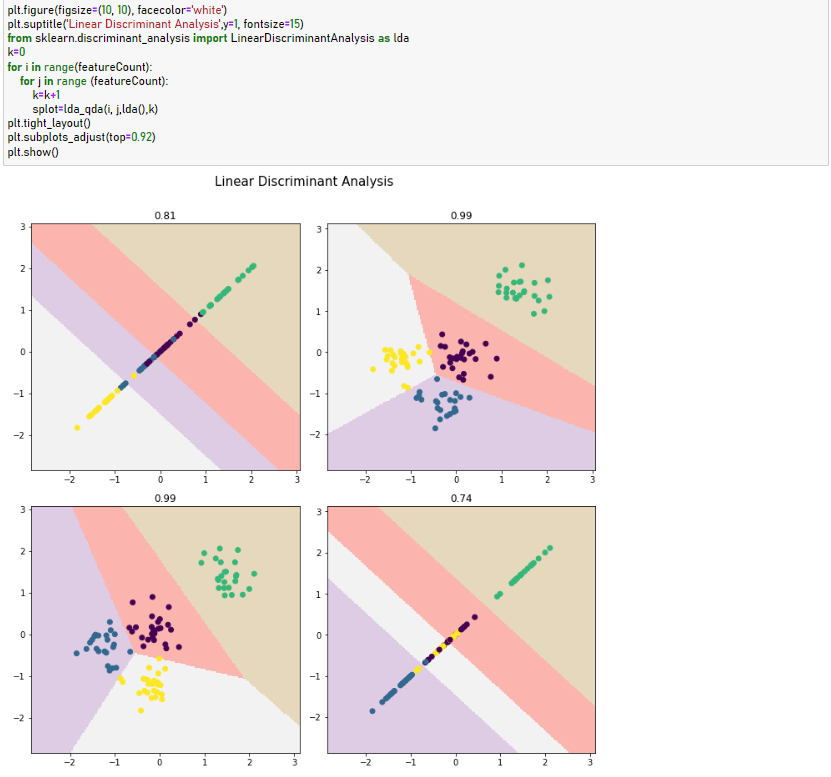


Рисунок 9 – Код для линейного дискриминантного анализа

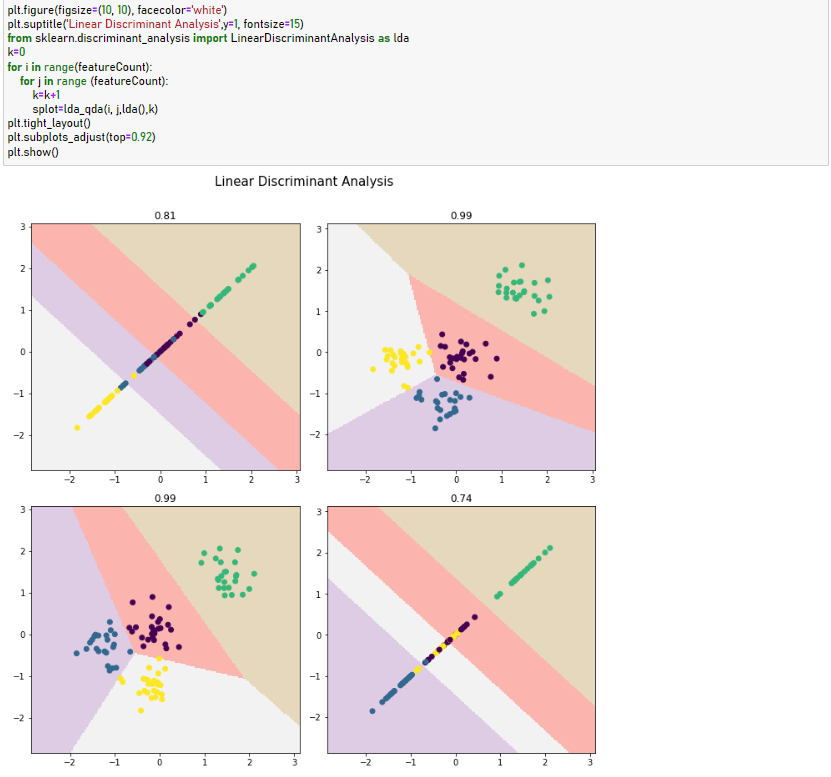


Рисунок 9 – Линейный дискриминантный анализ

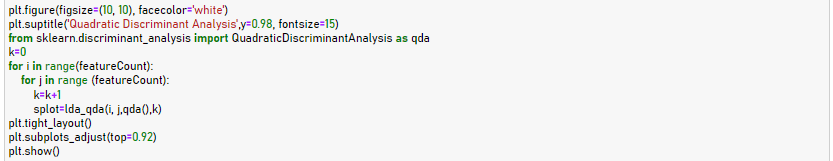


Рисунок 10 – Код для квадратичного дискриминантного анализа

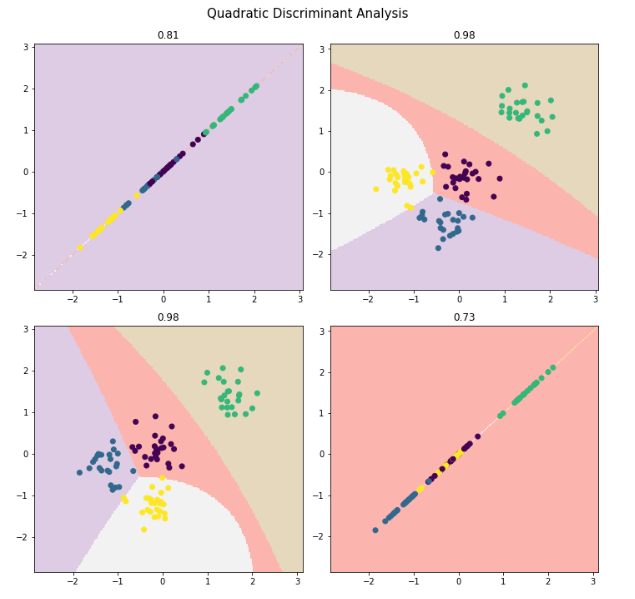


Рисунок 11 – Квадратичный дискриминантный анализ

# Вывод

В ходе данной работы были сгенерированы данные с заданными параметрами, получены навыки в применении иерархических методом классификации и их визуализации. Произведена кластеризация методом k-средних, и дополнительно построена диаграмма средних. И наконец выполнение дискриминантного анализа как линейного, так и квадратичного.