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 №3 по дисциплине:

|  |
| --- |
| «Методы анализа данных» |
| Задачи кластеризации и 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 100 наблюдений с параметрами, согласно номеру варианта (таблица 3.1). 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.

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

Вариант 5:

Количество кластеров – 3.

Количество признаков – 4.

Дисперсия значений – 1.

# 2 Ход работы

# 2.1 Создаём синтетические данные для кластеризации



Рисунок 2.1.1 – Код для создания.

# 2.2 Построение диаграммы рассеяния и стандартизация данных.

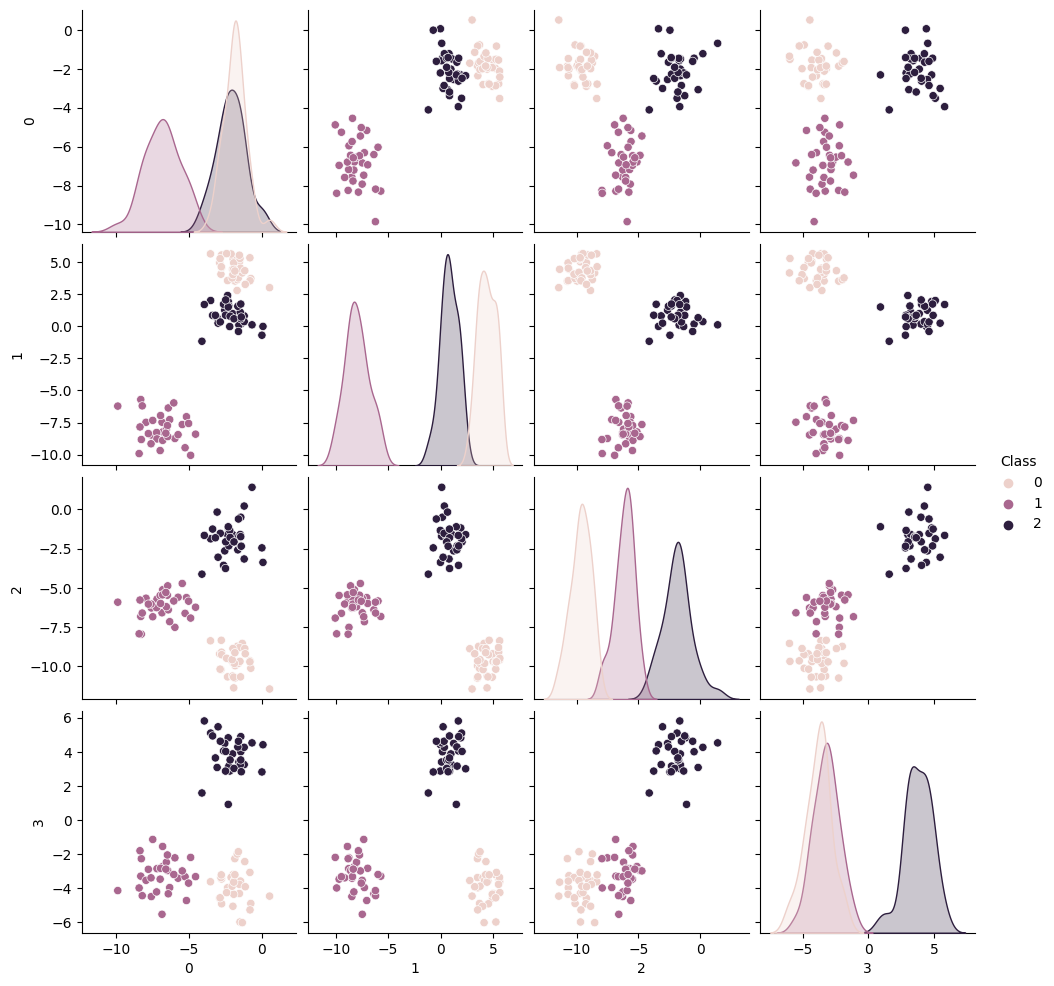


Рисунок 2.2.1 – Диаграмма рассеяния.

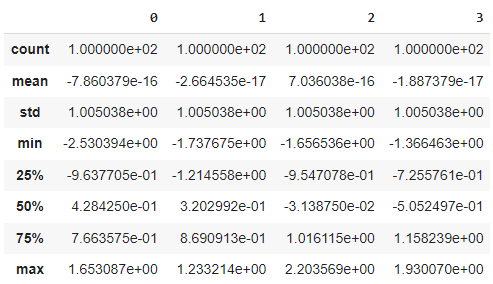


Рисунок 2.2.2 – Показатели после стандартизации данных.

# 1.3 Иерархическая кластеризация и построение дендрограммы.

Код построения иерархической кластеризации методом complete.

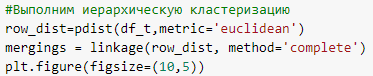


Рисунок 2.3.1 – Код.

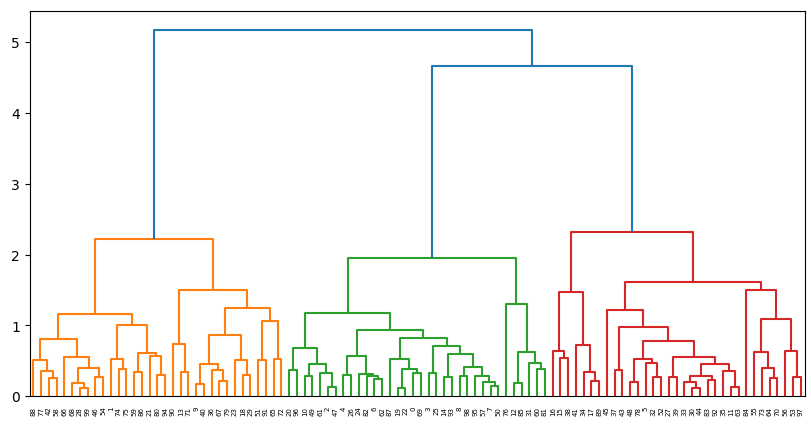


Рисунок 2.3.2 – Дендрограмма для кластеризации.

# 2.4 Кластеризация методом к-средних.

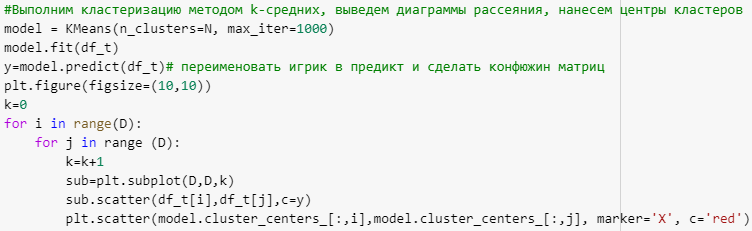
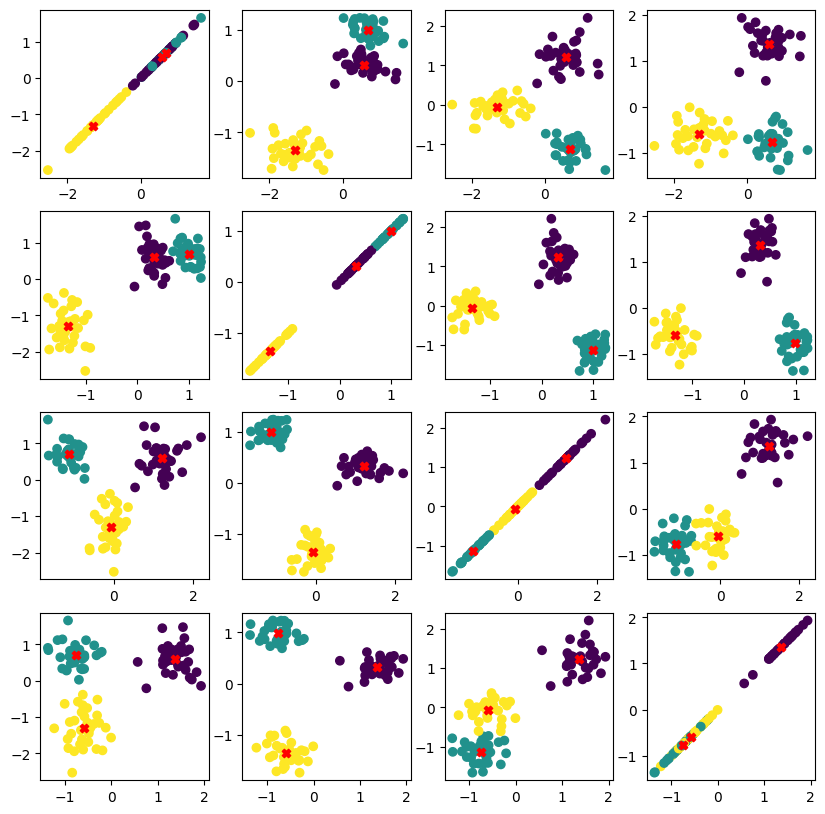


Рисунок 2.4.1 – код построения.

  
Рисунок 2.4.2 – диаграммы рассеяния с центрами кластеров.

**2.5 Добавление AgglomerativeClustering и графики средних для каждого кластера.**

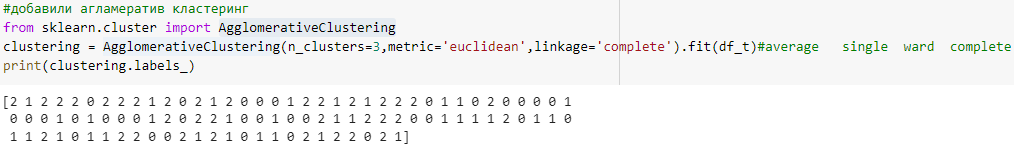
****

Рисунок 2.5.1 – Код для создания AgglomerativeClustering и вывода.

На основе графика средних для каждого кластера можно сделать вывод о том, какие признаки наиболее характерны для каждого кластера. Если графики средних для разных кластеров различаются значительно, то можно сделать вывод о том, что кластеризация проведена эффективно и объекты были правильно распределены по кластерам. Если же графики средних похожи друг на друга, то это может указывать на то, что кластеризация проведена неоптимально и нужно использовать другие методы для анализа данных.

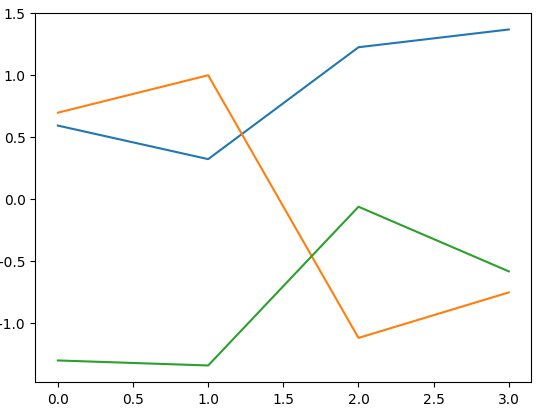


Рисунок 2.5.2 –Графики средних для каждого кластера.

**2.6 Дискриминантный анализ и отображение решающих поверхностей.**

Создадим функцию, реализующую инструмент для графического отображения результатов дискриминантного анализа – решающих поверхностей и отображения данных

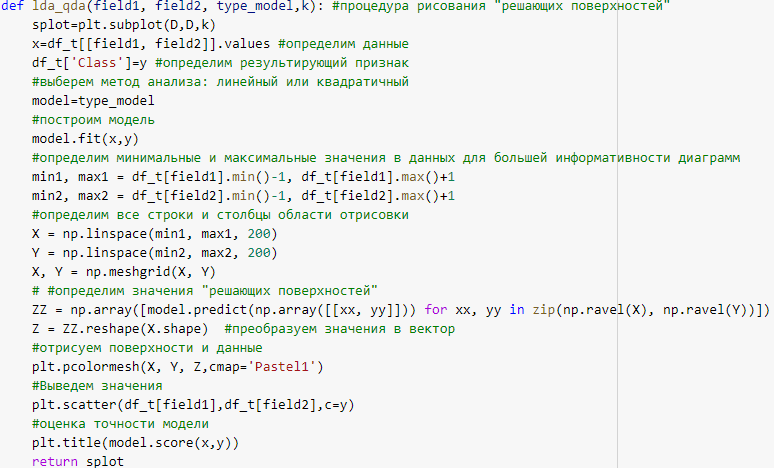


Рисунок 2.6.1 – Код создания функции.

Проведем линейный дискриминантный анализ данных.Для каждой пары признаков выведем на экран решающие поверхности и данные,для каждой пары признаков оценим точность модели дискриминантного анализа.

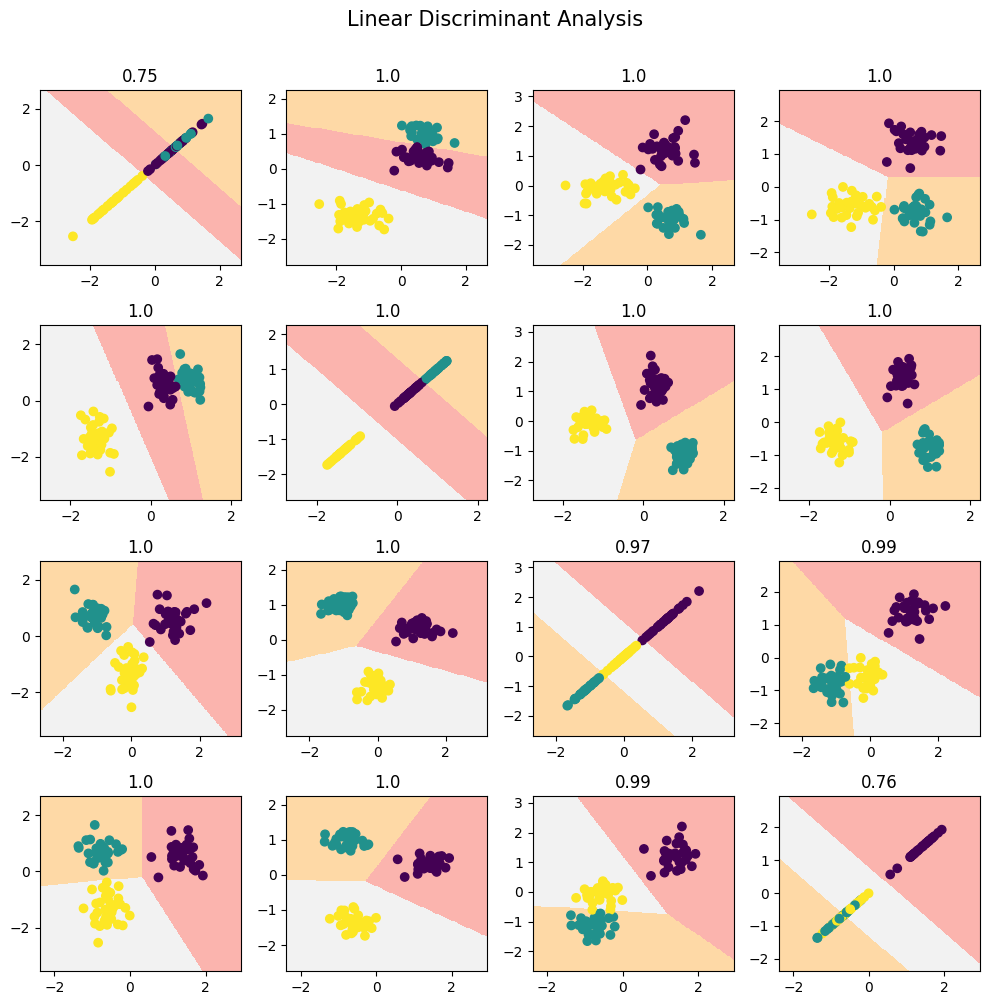


Рисунок 2.6.2 – Решающие поверхности для каждой пары признаков при линейном дискриминантном анализе.

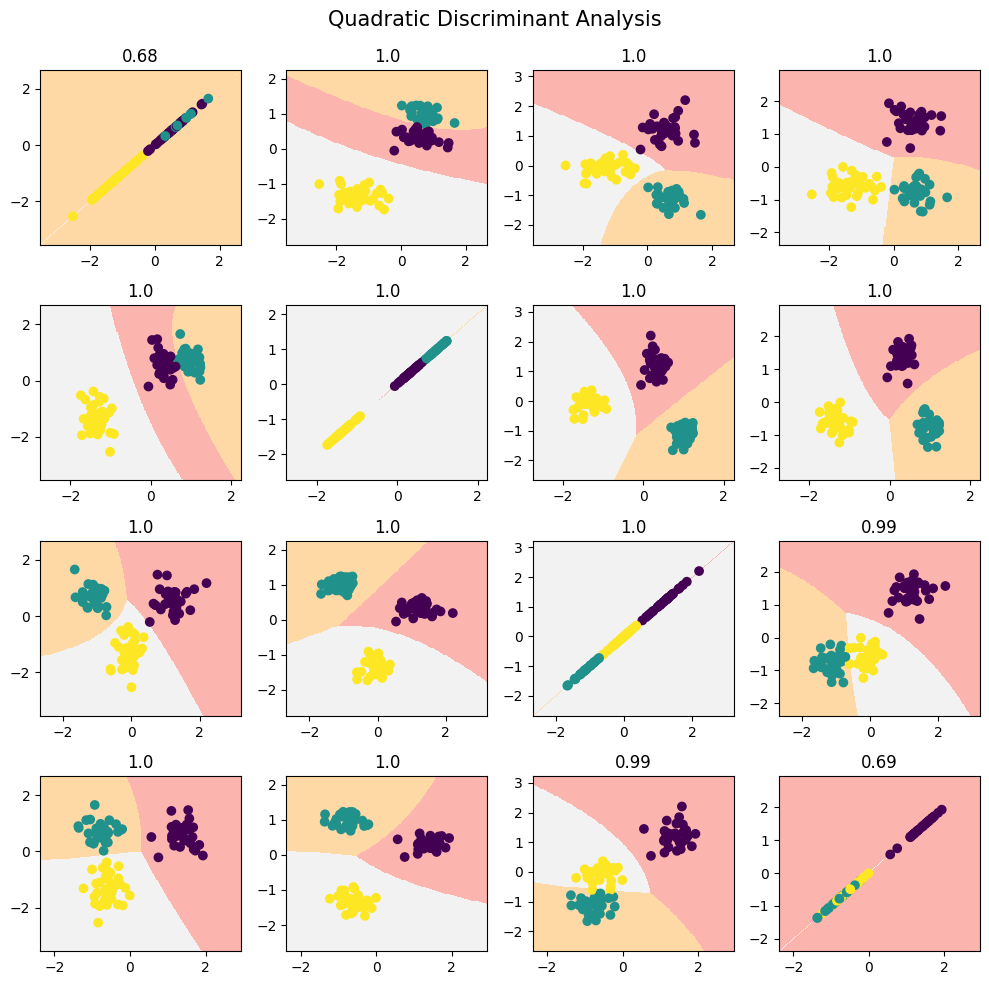


Рисунок 2.6.3 – Решающие поверхности для каждой пары признаков при квадратичном дискриминантном анализе.

**2.7 Нахождение матрицы ошибок и вывод точности классификации.**

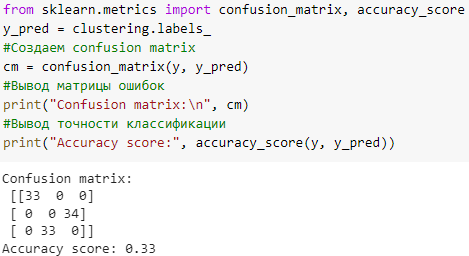


Рисунок 2.7.1 – Матрица ошибок и точность классификации для AgglomerativeClustering.

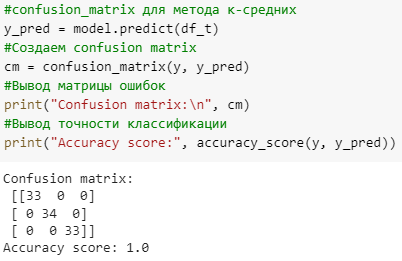


Рисунок 2.7.1 – Матрица ошибок и точность классификации для метода к-средних.

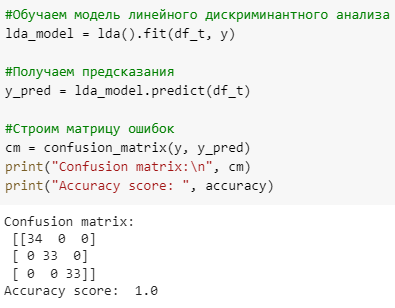


Рисунок 2.7.1 – Матрица ошибок и точность классификации для линейного дискриминантного анализа.

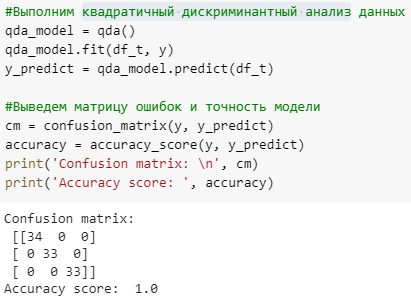


Рисунок 2.7.1 – Матрица ошибок и точность классификации для квадратичный дискриминантный анализа.

**2.8 Оценка качества кластеризации.**

Используя сумму квадратов растояний до центров оценить качество кластеризации.

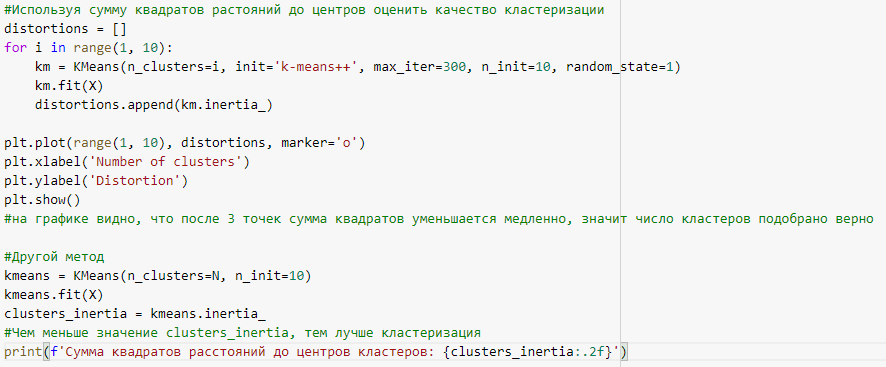


Рисунок 2.8.1 – Код для оценки качества кластеризации.

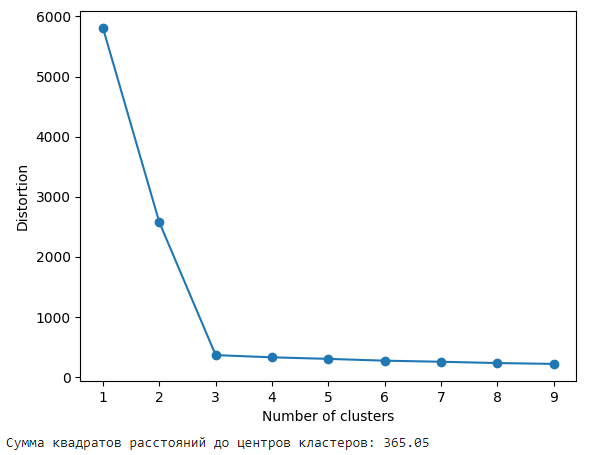


Рисунок 2.8.2 – График оценки и сумма квадратов до центров кластера.

3 Результаты работы

При исследовании задач классификации и кластеризации были разработаны навыки построения моделей, которые позволяют выявлять закономерности и группировать данные в соответствии с их признаками. Для решения задачи кластеризации были использованы методы, которые позволяют сгруппировать данные в N кластеров на основе их похожих значений признаков. При решении задачи классификации были использованы сгенерированные данные как обучающая выборка, где каждый кластер представляет отдельный класс. Были разработаны алгоритмы классификации, которые позволяют предсказывать классы для новых данных на основе обученной модели. Таким образом, проведенное исследование позволяет выявлять закономерности и группировать данные в соответствии с их признаками, а также предсказывать классы для новых данных на основе обученной модели. Это может быть полезно в различных областях, таких как маркетинг, медицина, банковское дело и другие.

# Список использованных источников

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