Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

|  |  |
| --- | --- |
| Институт | математики и компьютерных наук |
| Кафедра | компьютерной безопасности |

ОТЧЕТ

ЛАБОРАТОРНАЯ РАБОТА 9.

КЛАСТЕРИЗАЦИЯ

Выполнил: Окунев Николай Александрович,

студент 2 курса

группы КМБ-с-о-23-1

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025 г.

Проверено с оценкой:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025 г.

Ставрополь, 2025

1. **Цели и задачи**

Цель лабораторной работы: научится производить кластерный анализ данных на основе метода К-средних. Основные задачи: – получение навыков рефакторинга кода в проектах машинного обучения; – изучение принципов определения оптимального количества кластеров в методах кластерного анализа; – изучение возможностей языка Python для реализации кластероного анализа.

1. **Теоретическое обоснование**

Кластеризация – это разбиение множества объектов на подмножества (кластеры) по заданному критерию. Каждый кластер включает максимально схожие между собой объекты. Представим переезд: нужно разложить по коробкам вещи по категориям (кластерам) – например одежда, посуда, декор, канцелярия, книги. Так удобнее перевозить и раскладывать предметы в новом жилье. Процесс сбора вещей по коробкам и будет кластеризацией. Критерии кластеризации определяет человек, а не алгоритм, – этим она отличается от классификации. Этот метод машинного обучения часто применяют в различных неструктурированных данных – например если нужно автоматически разбить коллекцию изображений на мини-группы по цветам. Кластерный анализ применяют в разных сферах:  в маркетинге – для сегментирования клиентов, конкурентов, исследования рынка;  медицине – для кластеризации симптомов, заболеваний, препаратов; биологии – для классификации животных и растений;  социологии – для разбиения респондентов на однородные группы;  компьютерных науках – для группировки результатов при поиске сайтов, файлов и других объектов.

**3.Методика и порядок выполнения работы**

3.1 Учебная задача Задание. На основе предоставленного набора данных Mall\_custumers.csv реализуйте модель кластеризации методом К-средних. Подключаем библиотеки и загружаем имеющийся набор данных:

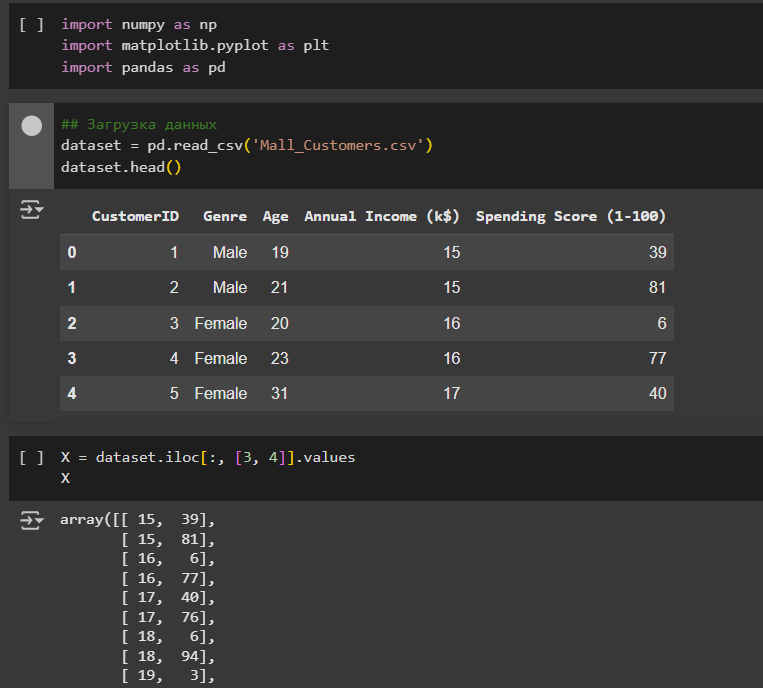
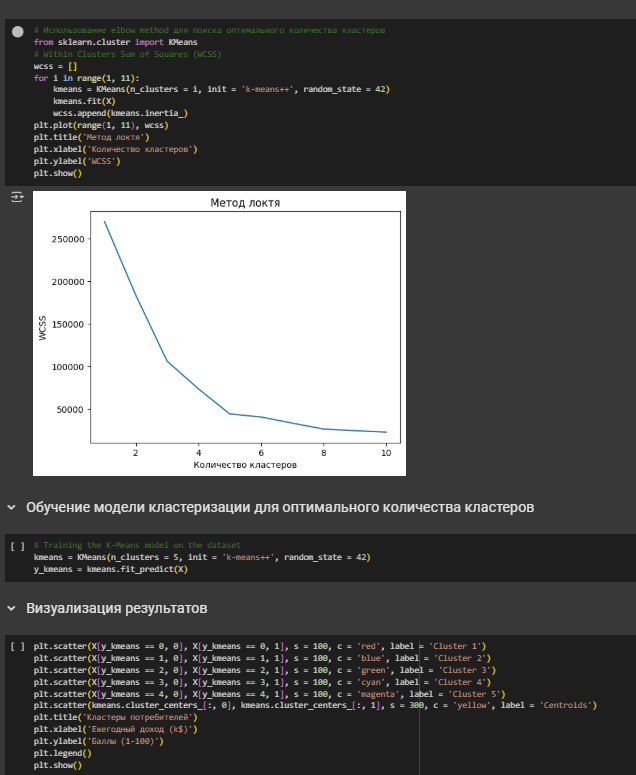


Рисунок 9.1 – Набор данных для задачи кластеризации

Исходный набор данных содержит сведения о посетителях торгового центра. В наборе присутствуют признаки, представленные в таблице 9.1. Признак набора данных Описание CustomerID Идентификатор клиента Genre Пол Age Возраст Annual Income Годовой доход Spending Score Баллы, присвевыемые клиенту специалистами по анализу данных торгового центра (от 1 до 100). Чем больше клиент тратит – тем больше баллов ему присваивается. Решение. Для разработки модели необходимо реализовать следующий код:



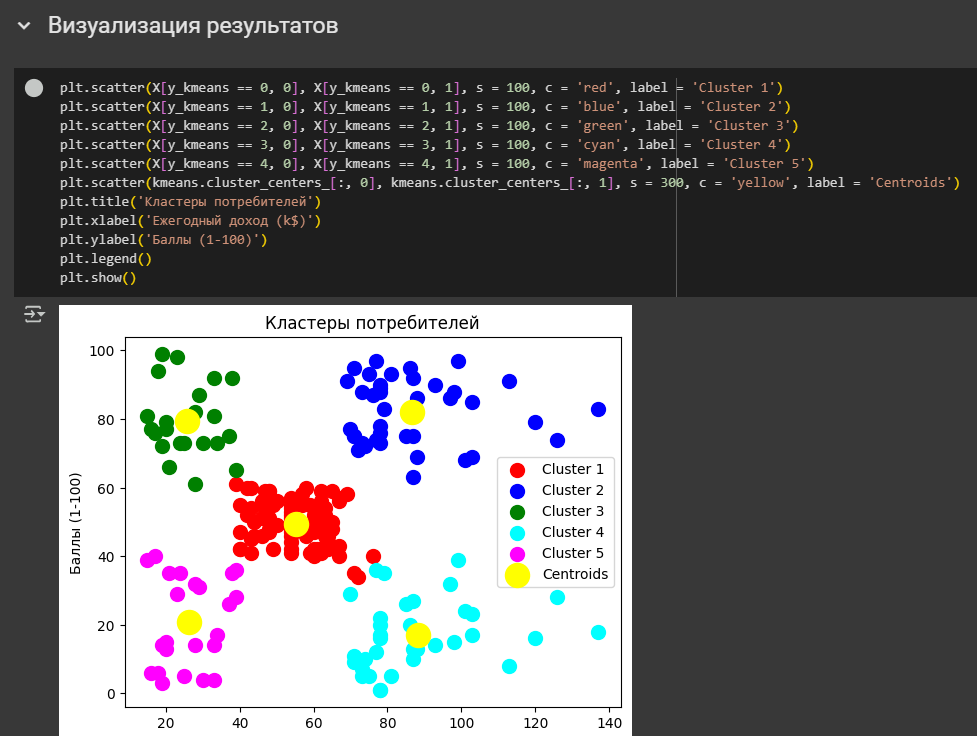


Рисунок 9.2 – Реализация метода кластеризации методом К-средних

**5. Контрольные вопросы**

1)Что такое кластерный анализ?

Кластерный анализ — это метод машинного обучения без учителя, который используется для разделения набора данных на группы (кластеры) на основе их схожести.

Цель: Найти структуру в данных, разделив объекты на группы так, чтобы объекты внутри каждой группы были более похожи друг на друга, чем объекты из разных групп.

Применение: Кластерный анализ используется в разных областях, таких как маркетинг (сегментация клиентов), биология (классификация видов), анализ социальных сетей (выявление сообществ) и т.д.

2)Перечислите известные методы кластерного анализа.

Существует много разных методов кластерного анализа, вот некоторые из наиболее известных:

Методы на основе центроидов:

K-means

K-medoids

Mini-Batch K-Means

Иерархические методы:

Агломеративная кластеризация

Дивизивная кластеризация

Методы на основе плотности:

DBSCAN

OPTICS

Методы на основе распределений:

Гауссовские смеси (Gaussian Mixture Models)

Спектральная кластеризация

Аффинное распространение (Affinity Propagation)

3)Перечислите классы и функции Python, которые задействованы при реализации кластерного анализа.

При реализации кластерного анализа в Python обычно используются следующие классы и функции из библиотеки scikit-learn:

sklearn.cluster.KMeans: Класс для реализации алгоритма K-means.

sklearn.cluster.AgglomerativeClustering: Класс для реализации агломеративной кластеризации.

sklearn.cluster.DBSCAN: Класс для реализации алгоритма DBSCAN.

sklearn.mixture.GaussianMixture: Класс для реализации гауссовских смесей.

sklearn.metrics.silhouette\_score: Функция для оценки качества кластеризации с помощью коэффициента силуэта.

sklearn.preprocessing.StandardScaler: Класс для масштабирования признаков.

4)Опишите принцип определения оптимального количества кластеров.

Определение оптимального количества кластеров — это важная задача в кластерном анализе. Существует несколько методов для определения оптимального количества кластеров:

Метод локтя (Elbow Method): Строится график зависимости внутрикластерной суммы квадратов (WCSS) от количества кластеров. Оптимальное количество кластеров соответствует "точке локтя" на графике, где WCSS начинает уменьшаться медленно.

Коэффициент силуэта (Silhouette Score): Вычисляется коэффициент силуэта для разных значений количества кластеров. Оптимальное количество кластеров соответствует максимальному значению коэффициента силуэта.

Индекс Калински-Харабаша (Calinski-Harabasz Index): Вычисляется индекс Калински-Харабаша для разных значений количества кластеров. Оптимальное количество кластеров соответствует максимальному значению индекса.

Gap Statistic: Сравнивает внутрикластерную дисперсию с ожидаемой дисперсией для случайного распределения точек.

5)Опишите принципиальные отличия методов регрессии, кластеризации и классификации.

Регрессия, кластеризация и классификация — это три разных типа задач машинного обучения:

Регрессия: Задача предсказания непрерывной числовой переменной.

Обучение: С учителем (есть размеченные данные).

Выход: Непрерывное число.

Пример: Предсказание цены дома на основе его характеристик.

Кластеризация: Задача разделения набора данных на группы (кластеры) на основе схожести объектов.

Обучение: Без учителя (нет размеченных данных).

Выход: Метки кластеров для каждого объекта.

Пример: Сегментация клиентов на основе их покупательского поведения.

Классификация: Задача отнесения объекта к одному из заранее определенных классов.

Обучение: С учителем (есть размеченные данные).

Выход: Класс, к которому относится объект.

**Вывод:** научились производить кластерный анализ данных на основе метода К-средних.