МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

Учреждение образования

«Гомельский государственный технический университет

имени П.О.Сухого»

Факультет автоматизированных и информационных систем

Кафедра «Информационные технологии»

**ЛАБОРАТОРНАЯ РАБОТА №2**

по дисциплине: «Введение в кластерный анализ и классификацию»

на тему: **«**Кластеризация данных**»**

Выполнил: студент гр. ИТИ-22

Василенко Р.В.

Принял: ассистент

Карась О. В.

Гомель 2024

**Цель работы:** разработать программное обеспечение для решения задачи кластеризации (обучение без учителя)

Ход выполнения работы:

1. Загрузить исходные данные согласно варианта (таблица 2) из csvфайлов

2. На основании считанных данных сформировать DataFrame

3. Провести анализ на полноту данных

4. Построить матрицу диаграмм рассеивания, тепловую карту и матрицу корреляции (scatterplotmatrix, heatmap, plotcorr)

5. Провести анализ на основе построенных графиков

6. Согласно индивидуального задания, выполнить кластеризацию данных. Оценить качество классификации для каждого метода с помощью функционала качества

7. Построить дендрограммы

8. Выполнить отбор информативных признаков и уменьшить размерность пространства признаков

9. Построить матрицу диаграмм рассеивания, тепловую карту и матрицу корреляции (scatterplotmatrix, heatmap, plotcorr) для изменённого пространства признаков

10.Повторно выполнить кластеризацию данных. Оценить качество классификации для каждого метода с помощью функционала качества

11.Проверить, к какому классу принадлежит произвольный объект

12.Сделать выводы по результатам проделанной работы

Вариант 2 - Минимальное оставное дерево. K средних k = 5.

**Ход работы и результаты выполнения**

Чтение данных из .csv файла происходит с помощью библиотеки pandas, команды .read\_csv(). На рисунке 1 изображен результат чтения csv файла.

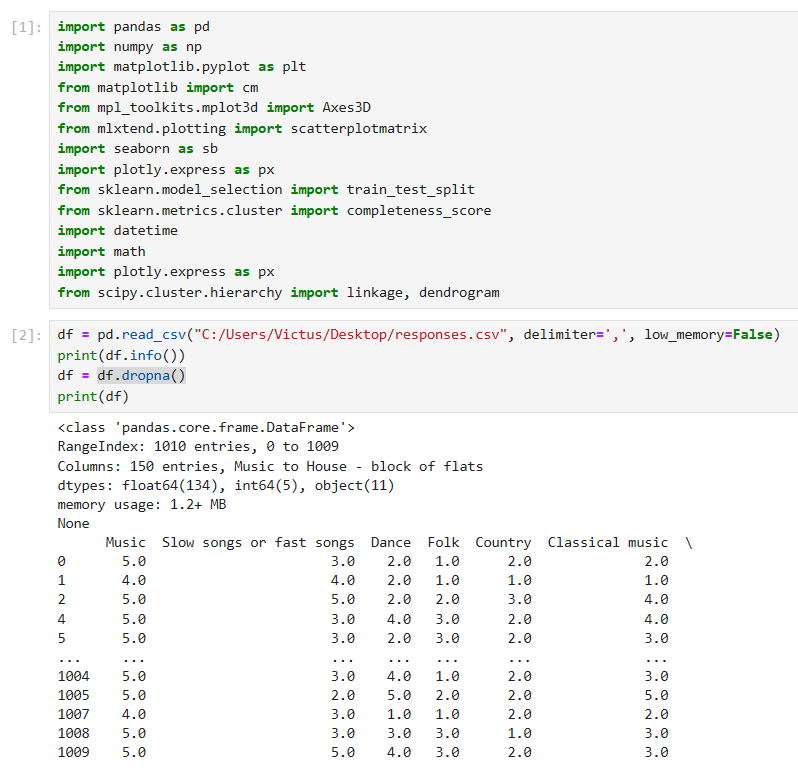
**

Рисунок 1 – Результат выполнения задания.

Проверка на полноту данных происходит с помощью метода .info(), в котором мы можем проверить, есть ли в столбцах нулевые ячейки. С помощью .dropna() мы избавляемся от всех записей, в которых есть нулевые значения

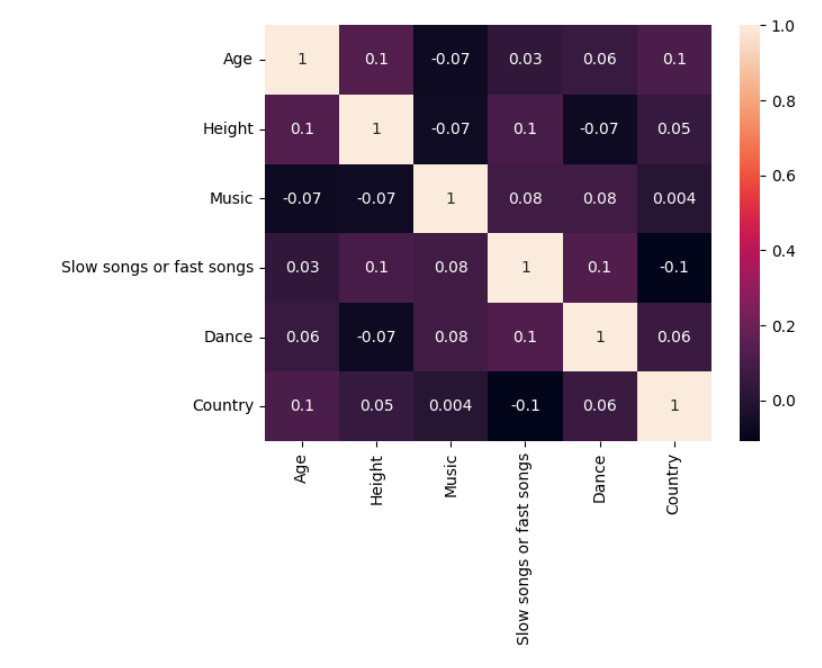
**

Рисунок 2 – Цветовая карта матрицы корреляции.

Для визуального изображения собранных данных используется библиотека matplotlib. На рисунке 3 изображена матрица диаграмм рассеивания

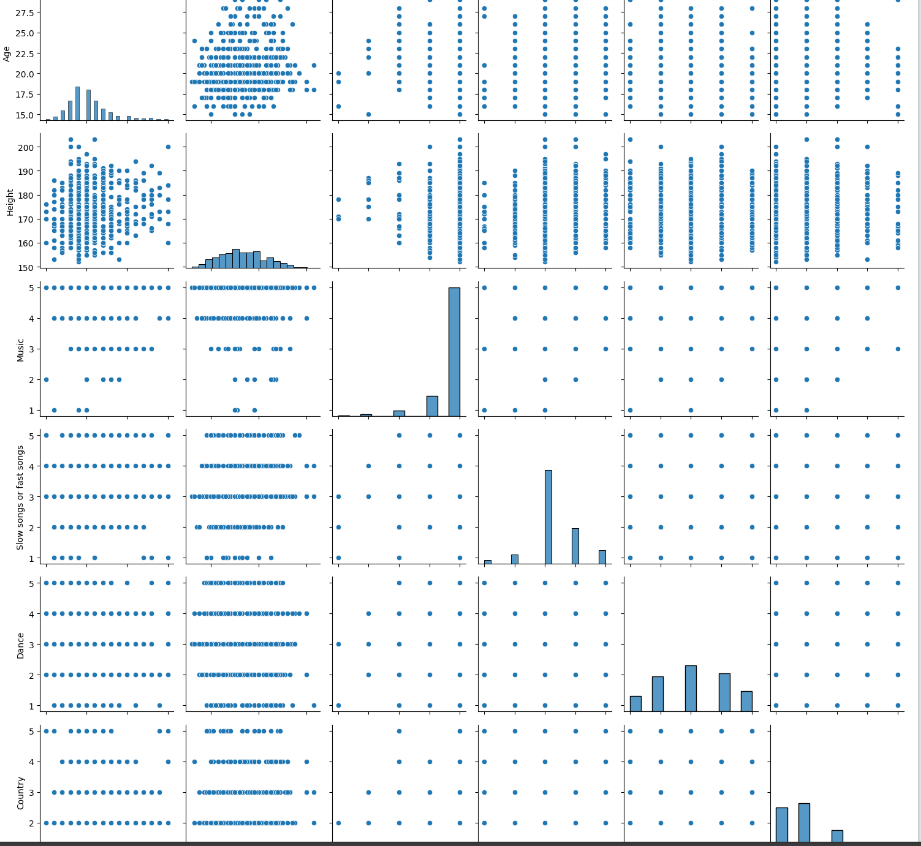


Рисунок 3 – Матрица диаграмм рассеивания

Матрица диаграмм рассеивания изображается благодаря библиотеке seaborn и matplotlib.pyplot. На рисунке 4 изображено соотношение мужчин и женщин в записях

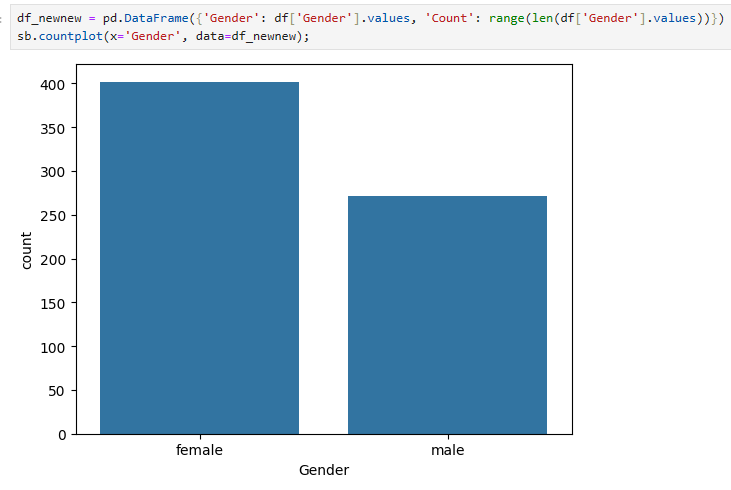


Рисунок 4 – Соотношение мужчин и женщин.

На рисунке 5 на графике в виде точек показаны все записи в соответствии роста и веса

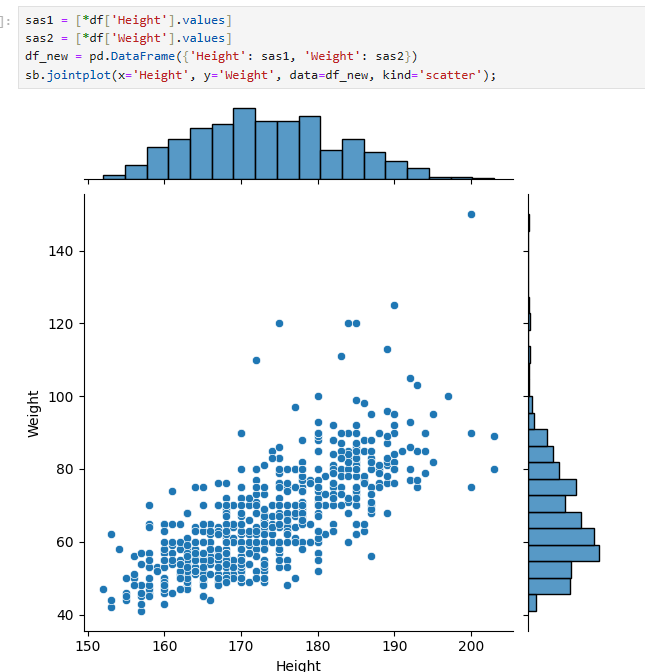


Рисунок 5 – Диаграмма роста от веса

На рисунке 6 изображена цветовая карта матрицы корреляции роста от веса.

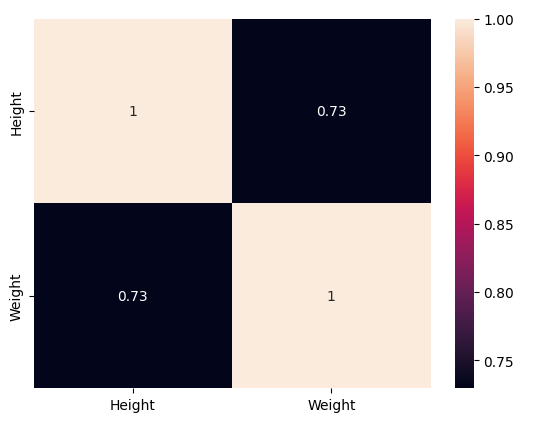


Рисунок 6 – Цветовая карта роста от веса

На рисунке 7 изображен алгоритм Прима, реализованный на python. Сначала создается список всех ребер графа, затем они сортируются по возрастанию и из наименьших из них составляется одна компонента связи.



Рисунок 7 – Реализация алгоритма Прима

На рисунке 8 изображена созданная по собранным данным дендограмма.

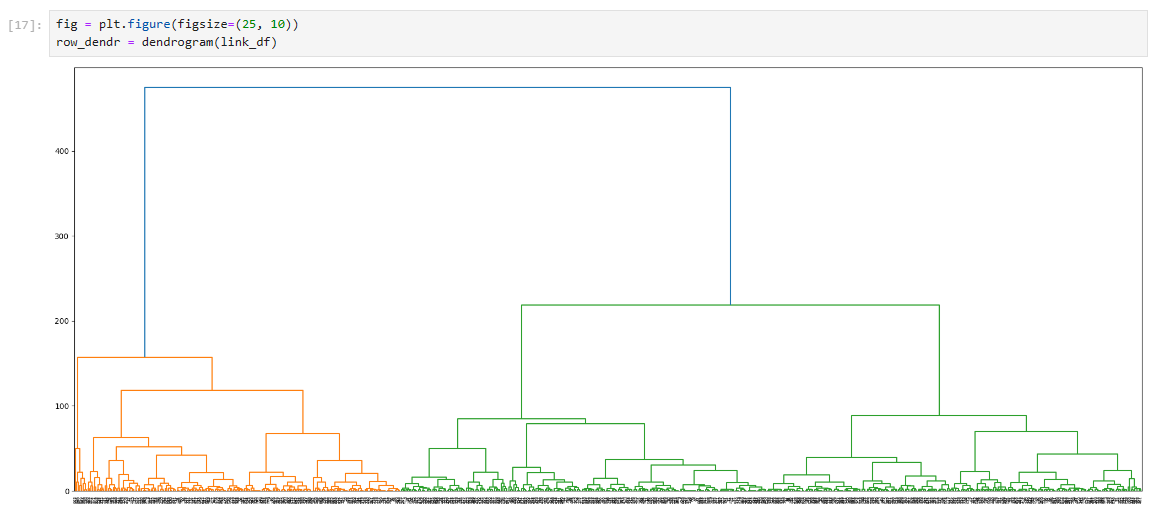
**

Рисунок 8 – Дендограмма

На рисунке 9 изображен результат выполнения алгоритма Прима и последующей кластеризации:

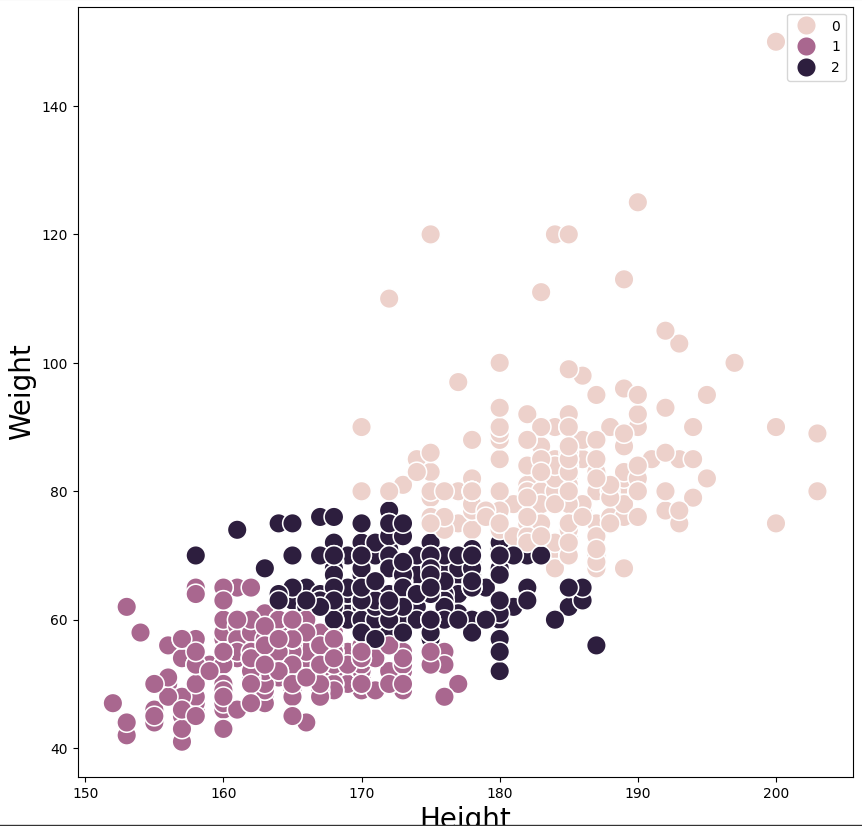


Рисунок 9 – Результат выполнения задания согласно варианту

На рисунке 10 изображена кластеризация по k-средним, k=5.

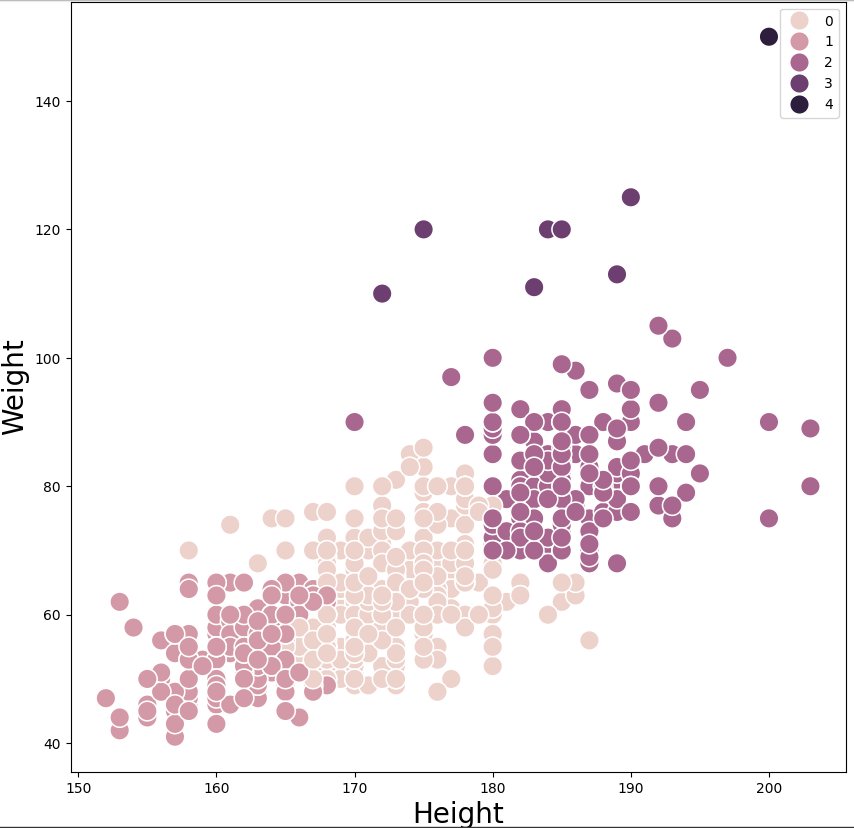


Рисунок 10 – Результат выполнения задания согласно варианту

На рисунке 11 указан алгоритм нахождения принадлежности определенного элемента к кластеру.

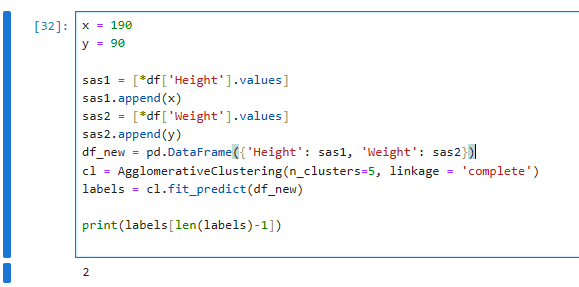


Рисунок 11 – Результат выполнения алгоритма

**Вывод:** в результате выполнения лабораторной работы выполнен анализ данных в csv файле, изображены графики зависимости данных друг от друга, из собранных данных составлены кластеры и произведен анализ функциональности. Были получены три кластера путем реализации метода Прима и пять кластеров методом к-средних, а так же проверена принадлежность некого объекта к кластеру.