Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет Информационных технологий и управления
Кафедра Интеллектуальных информационных технологий

Индивидуальная практическая работа №3
по дисциплине «Статистические основы индуктивного вывода»
на тему
Кластерный анализ в пакете STATISTICA

Студент гр. 021703: Е. С. Колосовский

Проверил: А. А. Ефремов

Минск 2023

1 Определение автомобилей на основе дискриминантного анализа

1.1 Задачи

1. Скачать из открытых ресурсов бесплатную демоверсию пакета STATISTICA.
2. Изучить алгоритм выполнения кластерного анализа указанными методами.
3. Скачать в открытых источниках (например, kaggle.com) датасет, включающий не менее 4 переменных и выполнить кластеризацию. Результаты кластеризации обосновать подробно с практической точки зрения исходя из ваших знаний о выбранной предметной области.

1.2 Выполнение

В качестве датасета будет использоваться: `https://www.kaggle.com/datasets/abineshkumark/carsdata`

Известна следующая информация по автомобилям:

1. Страна производства
2. Расход топлива
3. Количество цилиндров
4. Объем двигателя

Часть датасета:

1	14	8	350	US
2	31,9	4	89	Europe
3	17	8	302	US
4	15	8	400	US
5	30,5	4	98	US
6	23	8	350	US
7	13	8	351	US
8	14	8	440	US
9	25,4	5	183	Europe
10	37,7	4	89	Japan
11	34	4	108	Japan
12	34,3	4	97	Europe
13	16	8	302	US
14	11	8	350	US
15	19,1	6	225	US
16	16,9	8	350	US
17	31,8	4	85	Japan
18	16	8	304	US
19	24	4	113	Japan
20	24	4	107	Europe
21	37,2	4	86	Japan

Таблица 1 – Таблица с данными об автомобилях

С помощью метода локтя попробуем определить примерное количество кластеров, которые можно выделить(исследование проведено с помощью python-скрипта):

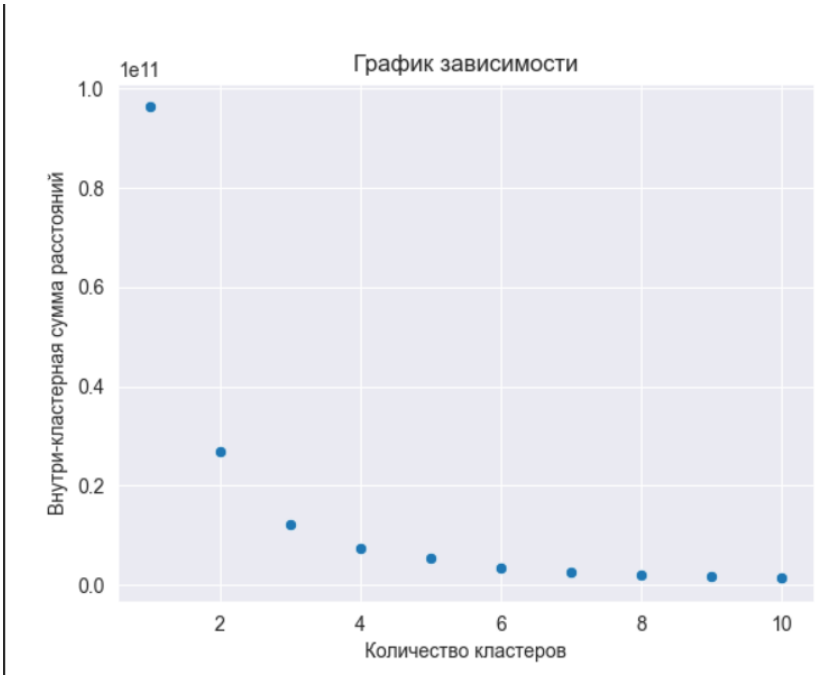

Рис. 1 – Результат исследования датасета с целью определения количества кластеров

С помощью метода локтя мы можем увидеть, что датасет делится на 4 кластера. Выявим наличие естественных кластеров с помощью иерархической классификации(была проведена с помощью python-скрипта).

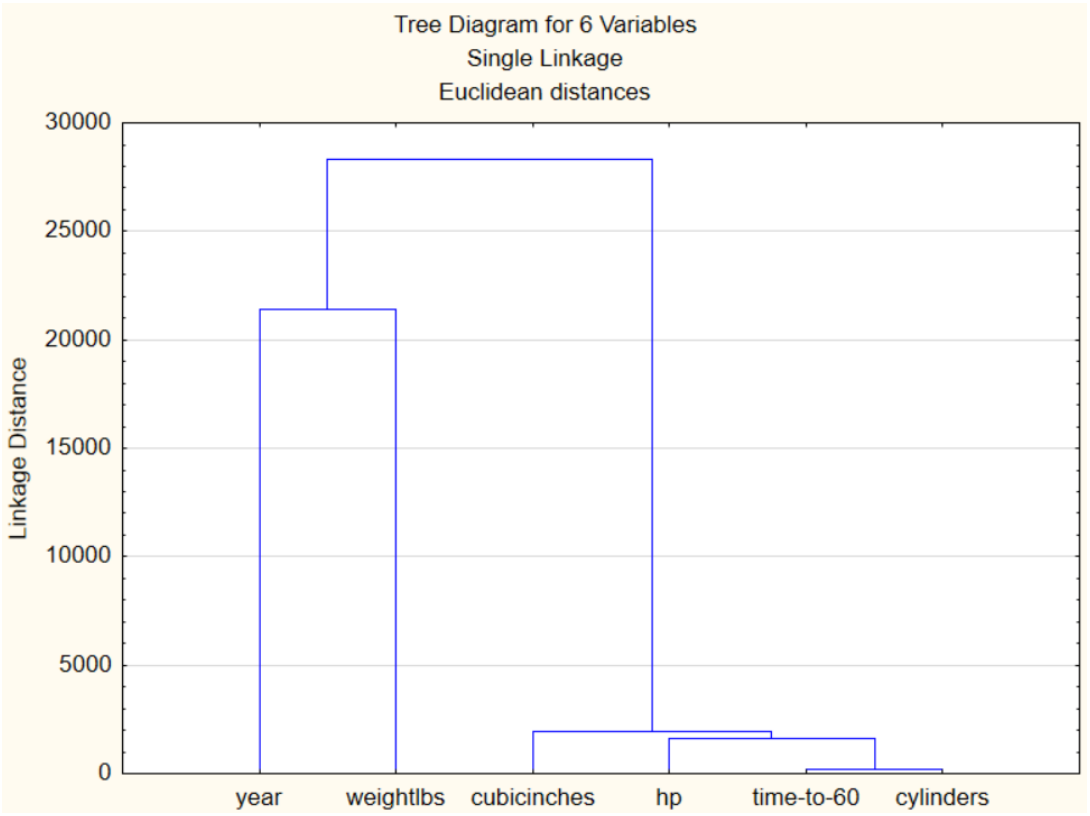

Рис. 2 – Иерархическая классификация

Исходя из результатов иерархической классификации можно сделать вывод, что образуется 4 естественных кластера. Проверим данное предположение, разбив исходные данные методом К средних на 4 кластера, и проверим значимость различия между полученными группами.

	Cluster Means (cars)			
Case ID	**Cluster No. 1**	Cluster No. 2	Cluster No. 3	Cluster No. 4
C_1	3090,500	350,0000	165,0000	10,00000
C_2	1952,500	89,0000	71,0000	9,00000
C_3	2710,000	302,0000	140,0000	9,50000
C_4	2866,000	400,0000	150,0000	9,00000
C_5	2014,500	98,0000	63,0000	10,50000
C_6	2940,000	350,0000	125,0000	12,50000
C_7	3168,500	351,0000	158,0000	10,50000
C_8	3141,500	440,0000	215,0000	8,50000
C_9	2755,000	183,0000	77,0000	12,50000
C_10	2016,000	89,0000	62,0000	10,50000
C_11	2114,000	108,0000	70,0000	10,50000
C_12	2084,500	97,0000	78,0000	10,00000
C_13	3058,000	302,0000	140,0000	11,00000
C_14	2819,000	350,0000	180,0000	9,50000
C_16	3170,000	350,0000	155,0000	11,50000
C_17	2000,000	85,0000	65,0000	11,50000
C_18	2702,000	304,0000	150,0000	10,00000
C_19	2125,500	113,0000	95,0000	10,00000
C_20	2200,500	107,0000	90,0000	9,50000

Рис. 3 – Результат кластеризации методом К-средних

Значение $p < 12$, что говорит о значимом различии.

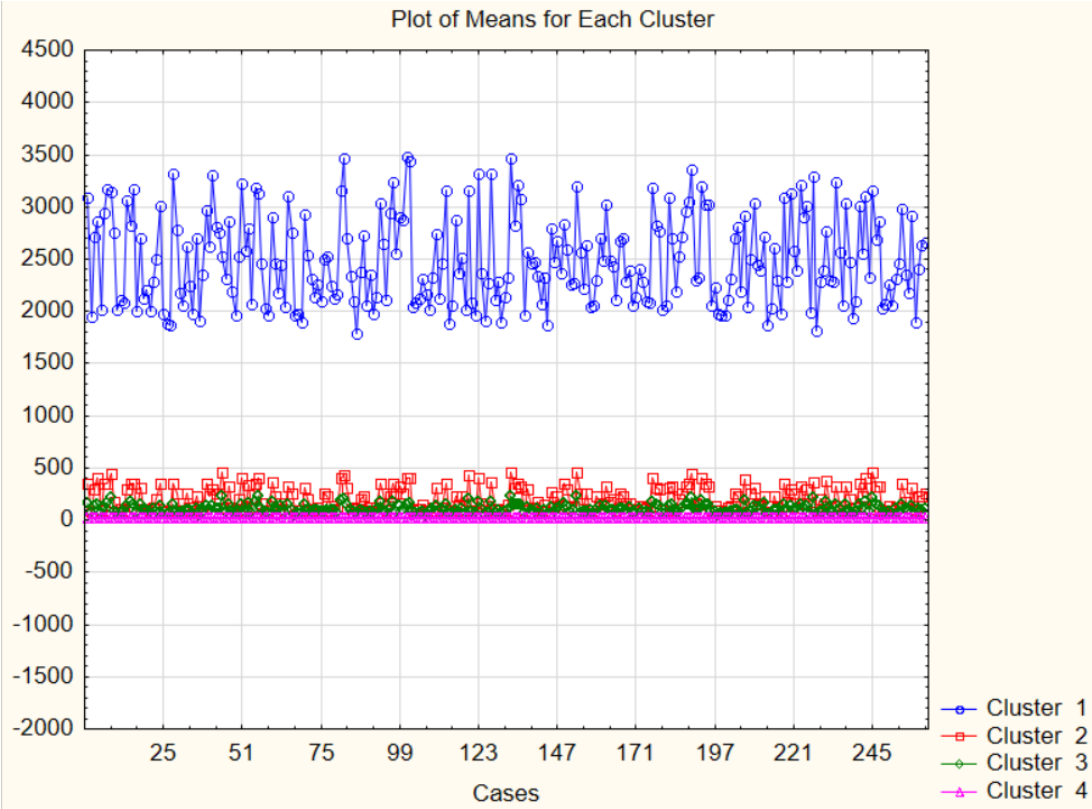

Рис. 4 – Результат разбиения на кластеры

Исходя из полученного графика средних значений переменных по кластерам, можно сделать вывод, что четыре полученных кластера разбивают автомобили на следующие группы:

а Автомобили с большим объемом двигателя, малой мощностью и большим расходом топлива были произведены в США.

б Автомобили с малым объемом двигателя, малой мощностью и небольшим расходом топлива были произведены в Японии.

в Автомобили с большим объемом двигателя, большой мощностью и большим расходом топлива были произведены в Европе.

г Автомобили с большим объемом двигателя, малой мощностью и большим расходом топлива были произведены в Японии.

Итог:

Было получено разделение на предполагаемые кластеры.