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

Выполнила
Студентка гр. 121701: Д. С. Мулярчик

Проверил: А. А. Ефремрв

Минск 2023

1 Определение видов домов 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/fedesoriano/the-boston-houseprice-data`

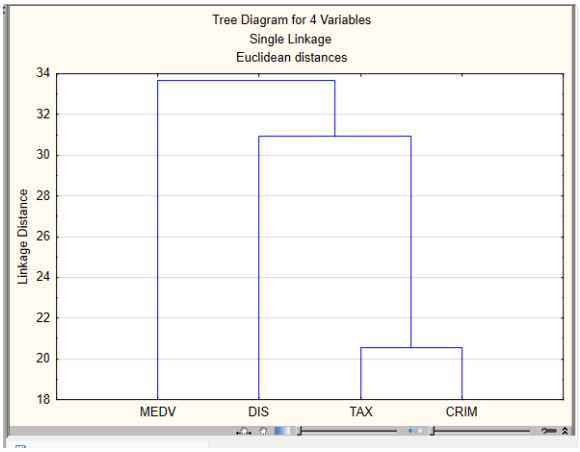
Известна следующая информация об жилье:

1. Уровень преступности
2. Расстоние до бизнес центра
3. Налоги
4. Цена

Часть датасета:

1	0.03237	3	222.0	33.40
2	0.06905	3	222.0	36.20
3	0.02985	3	222.0	28.70
4	0.08829	5	311.0	22.90
5	0.14455	5	311.0	27.10
6	0.21124	5	311.0	16.50
7	0.17004	5	311.0	18.90
8	0.22489	5	311.0	15.00
9	0.11747	5	311.0	18.90
10	0.09378	5	311.0	21.70
11	0.62976	4	307.0	20.40
12	0.63796	4	307.0	18.20
13	0.62739	4	307.0	19.90
14	1.05393	4	307.0	23.10
15	0.78420	4	307.0	17.50
16	0.80271	4	307.0	20.20
17	0.72580	4	307.0	18.20
18	1.25179	4	307.0	13.60
19	0.85204	4	307.0	19.60
20	1.23247	4	307.0	15.20
21	0.98843	4	307.0	14.50
22	0.75026	4	307.0	15.60
23	0.84054	4	307.0	13.90
24	0.67191	4	307.0	16.60
25	0.95577	4	307.0	14.80
26	0.77299	4	307.0	18.40
27	1.00245	4	307.0	21.00
28	1.13081	4	307.0	12.70
29	1.35472	4	307.0	14.50
30	1.38799	4	307.0	13.20

Стандартаризуем выборку и выявим наличие естественных кластеров с помощью иерархической классификации.



Исходя из визуального представления можно сделать вывод, что у нас образуется 4 кла-

стера.Проверим данное предположение, разбив исходные данные методом K средних на 4 кластера, и проверим значимость различия между полученными группами.

	Analysis of Variance (boston)					
Variable	**Between SS**	df	Within SS	df	F	signif. p
CRIM	190,9527	3	314,0473	502	101,7450	0,00
DIS	241,7184	3	263,2816	502	153,6285	0,00
TAX	423,5909	3	81,4092	502	870,6744	0,00
MEDV	339,6293	3	165,3707	502	343,6602	0,00

Рис. 1 – Разбиение на четрые кластера

Значение $p<0.05$, что говорит о значимом различии.

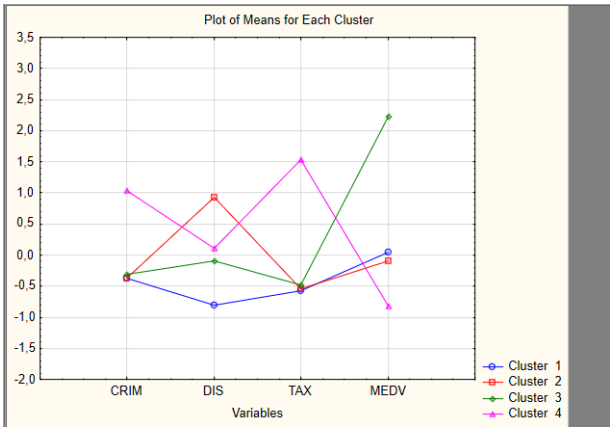


Рис. 2 – Разбиение на четрые кластера

Четыре кластера разбивают жильё на следующие группы:

1. Жильё с низким уровнем преступности, низким расстоянием до центра, низкой ценой аренды и средней ценой покупки
2. Жильё с низким уровнем преступности, высоким расстоянием до центра, низкой ценой аренды и средней ценой покупки
3. Жильё с низким уровнем преступности, средним расстоянием до центра, низкой ценой аренды и средней ценой покупки
4. Жильё с высоким уровнем преступности, средним расстоянием до центра, высокой ценой аренды и низкой ценой покупки

Итог:

Было получено разделение на предполагаемые кластеры.