Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ **ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ**

КАФЕДРА **КОМПЬЮТЕРНЫЕ СИСТЕМЫ И СЕТИ (ИУ6)**

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.03.03 Прикладная информатика**

ОТЧЕТ

по лабораторной работе № 3
Вариант 15

Название:

Дисциплина: Прикладной анализ данных

Студент ИУ6-55Б _____ В.К. Полубояров
(Подпись, дата) (И.О. Фамилия)

Преподаватель _____ М.А. Кулаев
(Подпись, дата) (И.О. Фамилия)

Москва, 2023

1. Не забудьте удалить таргеты из предыдущих лабораторных работ из вашей выборки.
2. Нормирование (масштабирование) исходных данных.Обратите внимание,что данные (коэффициенты, числа) для нормализации (масштабирования) рассчитываются только на основе обучающей выборки. И затем уже применяются к тестовым данным.
3. С помощью библиотеки sklearn сделать fit-predict модели иерархической кластеризации. Произвести кластеризацию 3 раза – с каждым из типов связей, которые мы проходили на занятии (параметр linkage). Построить дендрограмму для каждого типа связи и определить оптимальное число кластеров по ней. Выберите наилучший вариант (по вашему мнению) и обоснуйте ваш выбор. Получите итоговые метки кластера для каждого объекта на основе наилучшего варианта и определенного вами по дендрограмме наилучшего числа кластеров.
4. С помощью библиотеки sklearn сделать fit-predict модели k-средних. Перебрать по сетке различные варианты числа кластеров. Для каждого посчитать метрику Дэвиса-Болдина. Определить оптимальное число кластеров на основе значений этой метрики (выбрать наилучший вариант кластеризации).
5. Посчитайте индекс Рэнда между наилучшей кластеризацией из п.3 и наилучшей кластеризацией из п. 4. Сделать вывод о близости выбранных вами вариантов на основе этого индекса.
6. Для одного из наилучших вариантов для каждого кластера посчитать среднее значение признаков в каждом кластере. Проинтерпретировать кластеры на основе различий между средними значениями признаков в различных кластерах (постараться дать «логичные» названия).

Задание 1. Подготовка данных.

Исходя из варианта были подготовлены исходные данные - отобраны строки, соответствующие регионам: Центральный, Центрально-Черноземный, Северо-Кавказский, Западно-Сибирский, Восточно-Сибирский, Дальневосточный районы.

С помощью библиотеки Pandas были получены данные в виде массивов, где X - целевые признаки.

```
df = pd.read_excel('data3.xlsx')
df = df.drop(df.columns[[0, 1]], axis=1)
X = df
```

Рисунок 1 - разбиение данных на две выборки.

Задание 2. Нормирование данных

Для корректной работы с данными необходимо нормировать данные. Для нормирования была выбрана “Стандартизация”, которую часто называют Z-оценкой. Она рассчитывается с помощью формулы, отдельно для каждого x.

```
mean_X = X_train.mean()
std_X = X_train.std()

#нормализация тренировочных и тестовых данных
for column in X_train.columns:
    X_train[column] = (X_train[column] - mean_X[column]) / std_X[column]
    X_test[column] = (X_test[column] - mean_X[column]) / std_X[column]
```

Рисунок 2 - нормирование данных.

Задание 3. Иерархическая кластеризация.

Построим дендрограмму для метода ближайших соседей.

```
# Построение матрицы связей
Z = linkage(X, method='single')

# Построение дендрограммы
plt.figure(figsize=(10, 5))
plt.title(f'Dendrogram single linkage')
dendrogram(Z)
plt.show()
```

Рисунок 3 - построение дендрограммы.

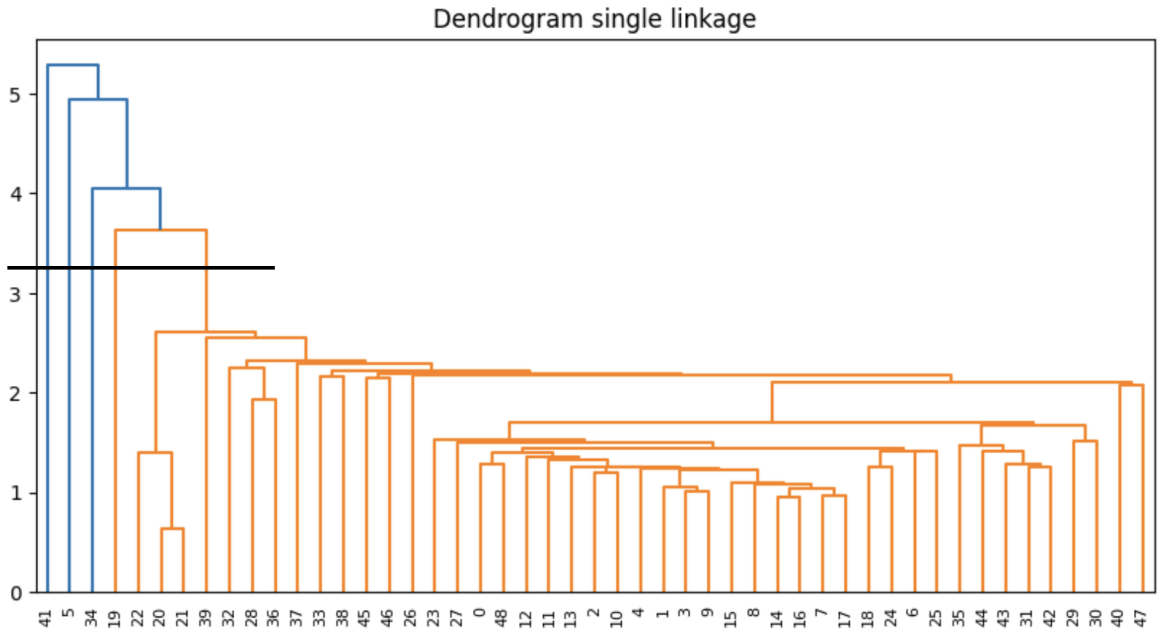

Рисунок 4 - дендрограмма для метода ближайших соседей.

Получившееся число классов (по черной линии на рисунке 4) равно 5. Однако 4 класса из 5 состоят из одного объекта, что говорит о плохом результате кластеризации данным методом.

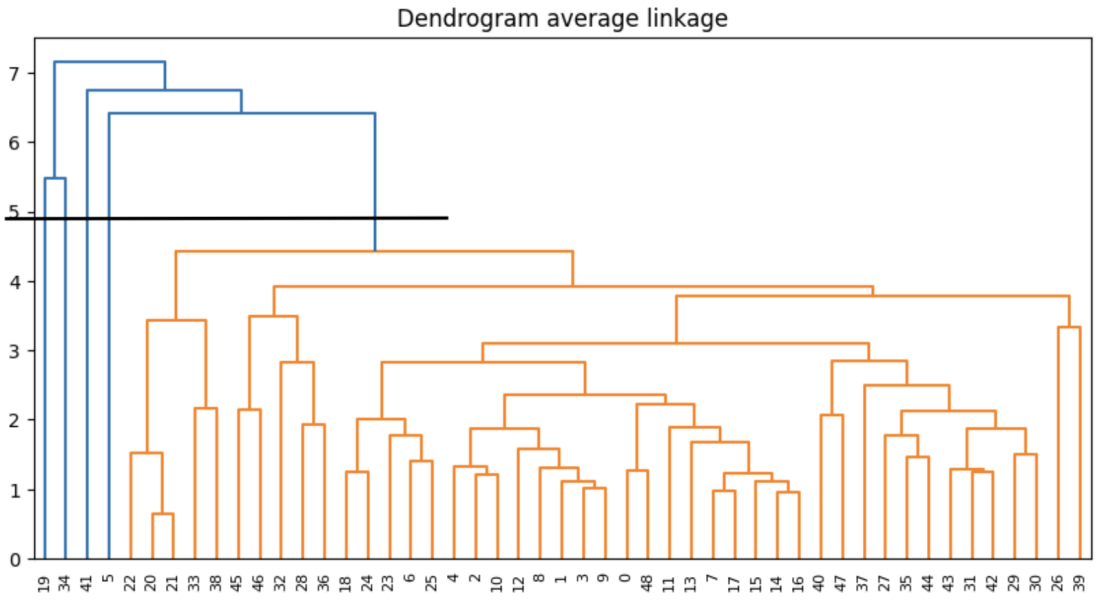

Рисунок 5 - дендрограмма для метода центра кластеров.

Результат очень похож на предыдущий, полученный для метода ближайших соседей.

Количество кластеров равно 5. 4 из 5 единичные, что является плохим результатом.

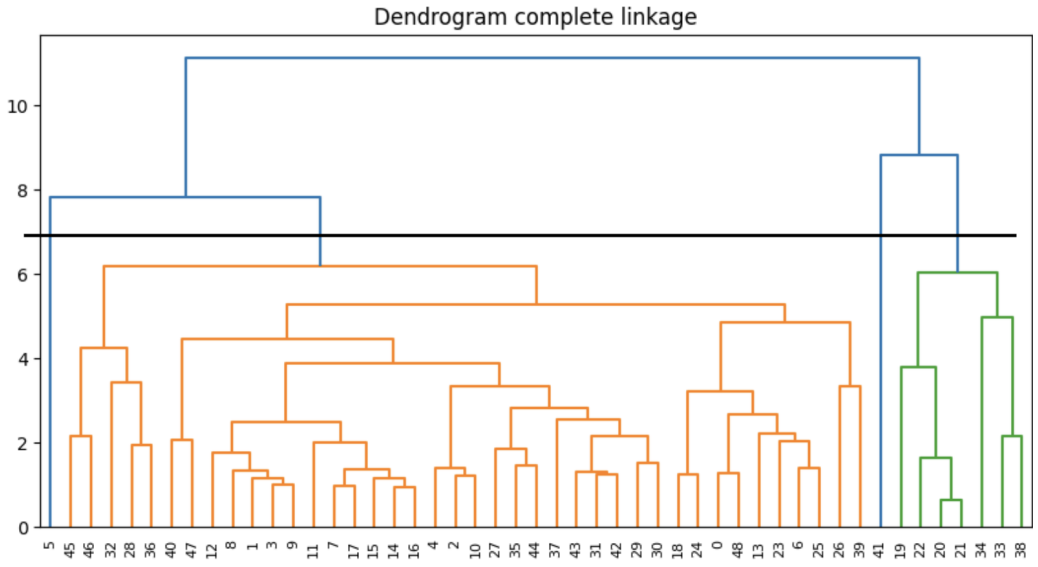

Рисунок 6 - дендрограмма для метода дальнего соседа.

Методом дальнего соседа получилось 4 класса. Два из них являются единичными (Объекты с индексами 5 и 41). Третий класс состоит из 7 объектов. Остальные объекты образуют 4 класс.

По результатам трех различных моделей самым оптимальным для данной выборки является метод дальнего соседа, так как обладает минимальным количеством единичных классов. Также он разделил данные немного “подробнее”: в двух других методах все классы, кроме одного были единичными, а все остальные объекты попали в один большой класс. В случае с методом дальнего соседа разделение оказалось чуть более удачным. Два класса единичные (предположительно - выбросы в данных), и два не единичных класса.
Теперь обучим модель по linkage = complete и n_clusters = 4.

```
best_model = AgglomerativeClustering(linkage='complete', n_clusters=4)

best_model.fit(X)

# Предсказание кластеров
predicted_labels = best_model.fit_predict(X)
predicted_labels
```

Рисунок 7 - обучение модели по выбранному методу.

Теперь определим метки кластеров для наших данных. (см. ниже)

	0	1
0	0	
25	0	
26	0	
27	0	
28	0	
29	0	
30	0	
31	0	
32	0	
47	0	
35	0	
37	0	
39	0	
40	0	
42	0	
43	0	
44	0	
45	0	
46	0	
36	0	
23	0	
24	0	
10	0	
1	0	
2	0	
3	0	
4	0	
6	0	
7	0	
8	0	
9	0	
11	0	
48	0	
13	0	
14	0	
15	0	
16	0	
17	0	
18	0	
12	0	
33	1	
34	1	
21	1	
38	1	
19	1	
20	1	
22	1	
41	2	
5	3	

Рисунок 8 - метки кластеров.

Результат аналогичен тому, что был получен при построении дендрограммы. Следовательно, полученный результат корректный.

Задание 4. Метод K-средних

Сделаем fit-predict модели k-средних. Необходимо подобрать параметр количества кластеров. Подбор будем осуществлять перебором по сетке с помощью метрики Дэвиса-Болдина. Это среднее значение максимального отношения между расстоянием точки от центра ее группы и расстоянием между двумя центрами групп. Ноль — это наименьший возможный результат. Значения, близкие к нулю, указывают на лучшее разделение.

```
def davies_bouldin_score(estimator, X):
    estimator.fit(X)
    labels = estimator.labels_
    score = metrics.davies_bouldin_score(X, labels)
    return score

kmeans = KMeans()
grid_space={'n_clusters': range(2,50)}
grid = GridSearchCV(kmeans, param_grid=grid_space, cv=5, scoring=davies_bouldin_score)
grid.fit(X)
print(grid.best_params_)
```

```
{'n_clusters': 4}
```

Рисунок 9 - подбор параметра количества кластеров по сетке.

В результате наилучшее количество кластеров равно 4.

Задание 5. Индекс Рэнда.

Необходимо посчитать индекс Рэнда между двумя лучшими кластеризациями из пунктов 3 и 4.

Индекс Рэнда — это способ сравнить сходство результатов между двумя разными методами кластеризации. Он оценивает, насколько

много из тех пар элементов, которые находились в одном классе, и тех пар элементов, которые находились в разных классах, сохранили это состояние после кластеризации алгоритмом.

Часто обозначаемый R , индекс Рэнда рассчитывается как:

R = (a+b) / (n C 2)

где:

- a: количество раз, когда пара элементов принадлежит одному и тому же кластеру при использовании двух методов кластеризации.
- b: количество раз, когда пара элементов принадлежит разностным кластерам по двум методам кластеризации.
- n C 2 : количество неупорядоченных пар в наборе из n элементов.

Индекс Рэнда всегда принимает значение от 0 до 1, где:

- 0: Указывает, что два метода кластеризации не согласуются с кластеризацией любой пары элементов.
- 1: Указывает, что два метода кластеризации полностью согласуются в отношении кластеризации каждой пары элементов.

Рассчитаем индекс Рэнда в программе.

```
# Вычисление индекса Рэнда между предсказанными метками из пункта 3 и пункта 4
rand_index = rand_score(predicted_labels, knn.predict(X))
print("Индекс Рэнда между кластеризациями из пунктов 3 и 4:", rand_index)
```

```
Индекс Рэнда между кластеризациями из пунктов 3 и 4: 0.6003401360544217
```

Рисунок 10 - индекс Рэнда.

Полученное значение - 0.6.

Значение индекса Рэнда равное 0.6 свидетельствует о том, что есть некоторое сходство между кластеризациями, но они также имеют значительные различия.

Задание 6. Интерпретация полученных результатов.

Для интерпретации результатов воспользуемся моделью, полученной с помощью k-средних. Посчитаем среднее значение каждого признака в кластерах.

```
labels = knn.labels_

# Добавление меток кластеров к исходным данным
X_labeled = X.copy()
X_labeled['Cluster'] = labels

# Расчет средних значений для каждого кластера
cluster_means = X_labeled.groupby('Cluster').mean()
cluster_means
```

4 rows | 4 rows × 9 columns | CSV

Cluster	x1	x2	x3	x4	x5	x6	x7	x8	x9
0	0.060393	-0.541650	-0.214980	0.500239	0.553504	-0.227262	0.072397	-0.103088	0.821573
1	-0.529136	0.774356	0.301442	0.009262	-0.306430	-0.011414	-0.113674	-0.314172	-0.428128
2	1.714902	-1.102478	-0.931954	-1.510305	-0.101507	-0.937973	-1.238630	1.842982	-0.408651
3	-0.253844	-0.238748	0.623329	0.572957	-0.170111	2.045171	2.017056	-0.888555	-0.008847

Рисунок 11 - расчет средних значений признаков по кластерам

Рассмотрим кластеры:

- Кластер 0 - “Средние” регионы с высокой преступностью
    - Средняя рождаемость (x1)
    - Низкая смертность (x2)
    - Немного ниже среднего число браков (x3)
    - Выше среднего число разводов (x4)
    - Выше среднего коэффициент младенческой смертности (x5)
    - Умеренное соотношение денежного дохода и прожиточного минимума (x6)

    - Умеренное соотношение средней оплаты труда и прожиточного минимума трудоспособного населения (x7)
    - Ниже среднего количество населения с денежными доходами ниже прожиточного минимума (x8)
    - Высокое число зарегистрированных преступлений (x9)
- Кластер 1 - Регионы с плохой демографией
    - Низкая рождаемость (x1)
    - Высокая смертность (x2)
    - Немного выше среднего число браков (x3)
    - Среднее число разводов (x4)
    - Низкий коэффициент младенческой смертности (x5)
    - Умеренное соотношение денежного дохода и прожиточного минимума (x6)
    - Умеренное соотношение средней оплаты труда и прожиточного минимума трудоспособного населения (x7)
    - Низкая численность населения с денежными доходами ниже прожиточного минимума (x8)
    - Ниже среднего число зарегистрированных преступлений (x9)
- Кластер 2 - “Бедные” регионы
    - Высокая рождаемость (x1)
    - Низкая смертность (x2)
    - Низкое число браков (x3)
    - Среднее число разводов (x4)
    - Ниже среднего младенческой смертности (x5)
    - Низкое соотношение денежного дохода и прожиточного минимума (x6)
    - Низкое соотношение средней оплаты труда и прожиточного минимума трудоспособного населения (x7)

    - Высокая численность населения с денежными доходами ниже прожиточного минимума (x8)
    - Ниже среднего число зарегистрированных преступлений (x9)
- Кластер 3 - Экономически развитые регионы
    - Ниже среднего рождаемость (x1)
    - Ниже среднего смертность (x2)
    - Выше среднего число браков (x3)
    - Выше среднего число разводов (x4)
    - Немного ниже среднего коэффициент младенческой смертности (x5)
    - Высокое соотношение денежного дохода и прожиточного минимума (x6)
    - Высокое соотношение средней оплаты труда и прожиточного минимума трудоспособного населения (x7)
    - Низкая численность населения с денежными доходами ниже прожиточного минимума (x8)
    - Среднего число зарегистрированных преступлений (x9)

Вывод: в ходе выполнения лабораторной работы были изучены методы обучения без учителя: агломеративная кластеризация с различными типами связями и метод k-средних. Также были изучены метрика Дэвиса-Болдина и индекс Рэнда.