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

Дисциплина: Прикладной анализ данных

Название: Решение задачи кластеризации

Студент ИУ6-55Б _____ А.Д. Шевченко
(Группа) (Подпись, дата) (И.О. Фамилия)

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

Москва, 2023

Цель: исследовать методы кластеризации

Формулировка:

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).
https://scikitlearn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html

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.

https://scikitlearn.org/stable/modules/generated/sklearn.metrics.rand_score.html#sklearn.metrics.rand_score

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).

Основная часть

1. Первым делом была проведена загрузка excel-файла с исходными данными. Файл был считан в датафрейм с помощью библиотеки pandas (рисунок 1).

```
# Загрузка файла с уже удаленными таргетами

X = pd.read_excel('ADA_LR3_source.xlsx')
X.head()
```

	x1	x2	x3	x4	x5	x6	x7	x8	x9
0	8.5	16.3	6.8	5.6	17.4	163	151	23.6	2344
1	9.3	12.6	7.2	5.5	25.3	194	239	9.2	1809
2	8.7	4.6	6.5	4.2	16.2	152	192	26.9	2406
3	8.6	6.2	6.1	4.0	17.4	190	205	20.1	2023
4	8.1	11.4	7.7	6.4	5.9	183	198	22.0	1419

Рисунок 1 – загрузка исходных данных

Далее была произведена Z-нормализация данных (рисунок 2).

```
# Z-нормализация данных

mean = np.mean(X, axis=0)
deviation = np.std(X, axis=0)
X_norm = (X - mean) / deviation

X_norm.head()
```

	x1	x2	x3	x4	x5	x6	x7	x8	x9
0	-0.286871	0.380753	-0.578697	1.355041	0.039305	-0.089130	-0.151352	-0.585908	1.223650
1	0.047072	-0.747237	0.035964	1.242121	2.792850	0.432408	2.681192	-2.108728	0.152280
2	-0.203386	-3.186134	-1.039693	-0.225840	-0.378955	-0.274192	1.168356	-0.236928	1.347809
3	-0.245128	-2.698355	-1.654355	-0.451680	0.039305	0.365113	1.586800	-0.956038	0.580828
4	-0.453843	-1.113071	0.804291	2.258402	-3.969020	0.247346	1.361484	-0.755110	-0.628719

Рисунок 2 – Z-нормализация данных

2. Вторым шагом стало построение дендрограмм для каждого типа связей. Код изображен на рисунке 3, результат построения – на рисунках 4-6.

```
linkage_types = ['single', 'average', 'complete']

for linkage_type in linkage_types:
  link = linkage(X_norm, method=linkage_type)
  plt.figure(figsize=(15, 5))
  plt.title(f'Иерархическая кластеризация (тип связей - {linkage_type})')
  plt.xlabel('Номер объекта')
  plt.ylabel('Расстояние')
  dendrogram(link)
```

Рисунок 3 – код построения дендрограмм

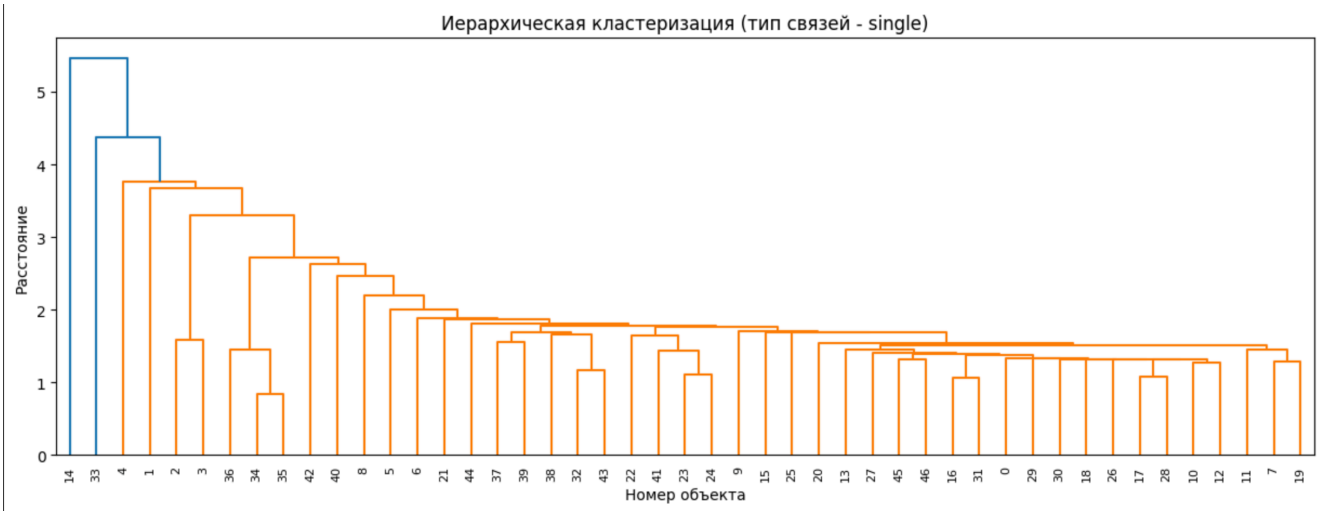

Рисунок 4 – иерархическая кластеризация (тип связей – ближайший сосед)

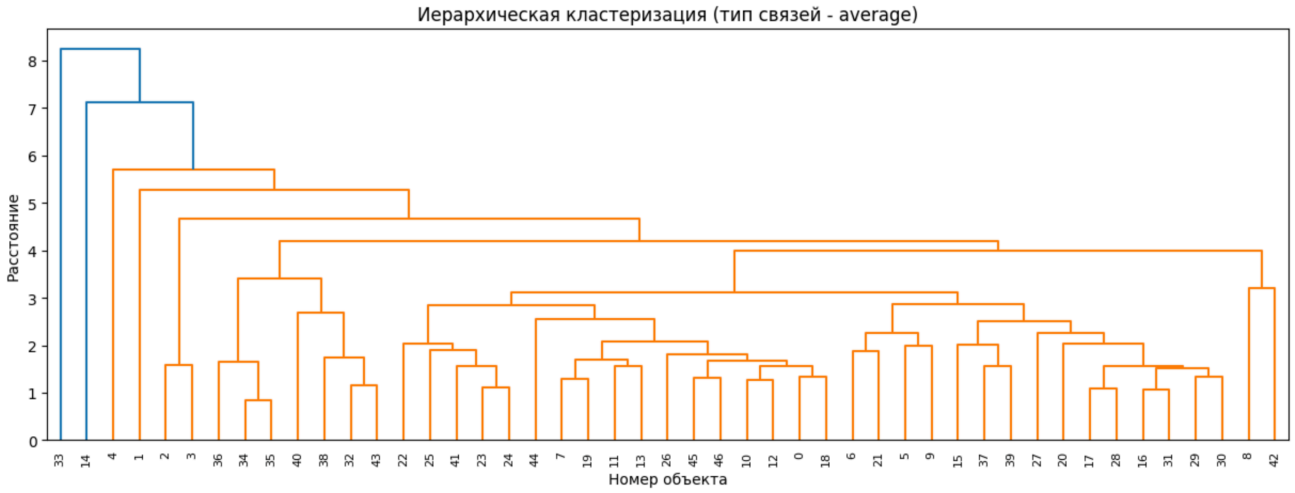

Рисунок 5 – иерархическая кластеризация (тип связей – между центрами кластеров).

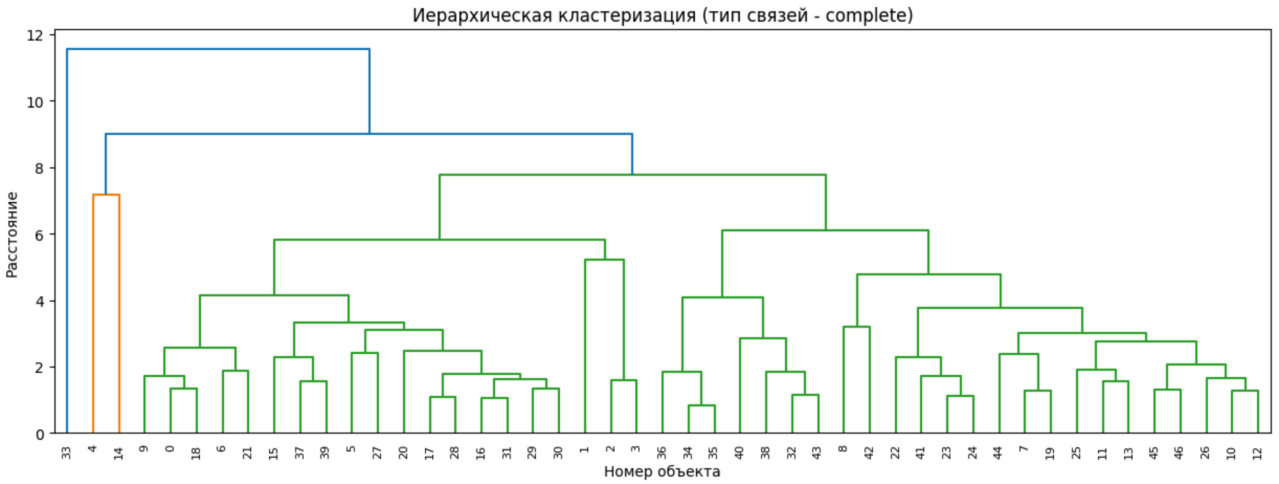

Рисунок 6 – иерархическая кластеризация (тип связей – дальний сосед)

При рассмотрении построенных дендрограмм можно увидеть следующее: кластеризация с типами связей «ближайший сосед» и «между центрами кластеров» оказалась не очень удачной, поскольку объекты оказались разбиты на 4-6 единичных кластеров, а все остальные вошли в один большой кластер. Другое дело – кластеризация с типом связей «дальний сосед». Здесь в результате разбиения получились один единичный кластер, один состоящий из двух объектов и два одинаково больших. Исходя из этого, можно сделать выбор лучшего варианта в пользу типа связей «дальний сосед».

Сделаем fit_predict полученной модели. Получим итоговые метки кластера, задав количество кластеров равным 4 и тип связей «дальний сосед» (complete) (рисунок 7).

```
from sklearn.cluster import AgglomerativeClustering

# Аггломеративная кластеризация с помощью библиотеки sklearn
model = AgglomerativeClustering(n_clusters=4, linkage='complete')
agg_labels = model.fit_predict(X_norm)
agg_labels
```

```
array([1, 1, 1, 1, 0, 1, 1, 2, 2, 1, 2, 2, 2, 2, 0, 1, 1, 1, 1, 2, 1, 1,
       2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 2, 3, 2, 2, 2, 1, 2, 1, 2, 2, 2, 2,
       2, 2, 2])
```

Рисунок 7 – метки кластера с типом связей «дальний сосед»

Полученные метки соответствуют тому, что изображено на дендрограмме (объекты с номерами 4 и 14 были включены в кластер 0, с номером 33 – в кластер 3, левый большой кластер – 1, правый большой кластер – 2).

3. В качестве третьего шага был выполнен fit-predict модели k-средних. По сетке были перебраны различные варианты числа кластеров. Вывод о лучшем числе кластеров был сделан на основе метрики Дэвиса-Болдина. Данная метрика вычисляется следующим образом:

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{S_i+S_j}{d(c_i,c_j)}\right),$$

где k – число кластеров, S – внутрикластерное сходство (например, среднее расстояние точек до центра), d – расстояние между центрами кластеров. Если значение данной метрики низкое, значит, кластеры хорошо разделены (между ними большое расстояние) и каждый кластер хорошо описан своим средним значением (т.е. имеет небольшой разброс). Код и результат изображены на рисунке 8.

```
from sklearn.cluster import KMeans
from sklearn.metrics import davies_bouldin_score, get_scorer_names
from sklearn.model_selection import GridSearchCV

def davies_bouldin_scoring(estimator, X):
  estimator.fit(X)
  return davies_bouldin_score(X, estimator.labels_)

kmeans = KMeans(n_init=10)
param_grid = {'n_clusters': range(2, 9)}

# Создание объекта GridSearchCV
grid_search = GridSearchCV(kmeans, param_grid, scoring=davies_bouldin_scoring)

# Обучение модели и подбор идеальных параметров
grid_search.fit(X_norm)

# Получение лучшего числа кластеров
best_n_clusters = grid_search.best_params_['n_clusters']

# Новая модель уже с указанным лучшим числом кластеров
best_k_means = KMeans(n_init=10, n_clusters = best_n_clusters)
best_k_means.fit_predict(X_norm)
kmeans_labels = best_k_means.labels_

print('Лучшее число кластеров:', grid_search.best_params_['n_clusters'])
```

```
Лучшее число кластеров: 3
```

Рисунок 8 – лучшее число кластеров

Как нам сообщает GridSearchCV, лучшее число кластеров – 3.

4. Следующим шагом было вычисление индекса Рэнда, который позволяет сравнить сходство результатов между двумя разными методами кластеризации (рисунок 9). Значение 1 означает, что результаты полностью совпадают, 0 – результаты абсолютно не совпадают.

```
from sklearn.metrics import rand_score

# Вычисление индекса Рэнда
rand_index = rand_score(agg_labels, kmeans_labels)
rand_index
```

```
0.6512488436632747
```

Рисунок 9 – индекс Рэнда

В нашем случае значение индекса Рэнда получилось приблизительно равным 0.65, что говорит о том, что результаты алгоритмов из пункта 2 и пункта 3 имеют некую схожесть.

5. Пятым шагом стало вычисление среднего значения признаков в каждом кластере (рисунок 10).

```
X_labeled = X_norm.copy()

# Добавление в датафрейм колонки "кластер"
X_labeled['Cluster'] = kmeans_labels

# Вычисление среднего значения признаков в каждом кластере
cluster_means = X_labeled.groupby('Cluster').mean()
cluster_means
```

Cluster	x1	x2	x3	x4	x5	x6	x7	x8	x9
0	-0.366382	0.353171	0.672578	0.564601	-0.281029	0.478072	0.579775	-0.644323	-0.190826
1	1.489520	-0.919992	0.172556	-1.204481	0.198088	-0.657402	-0.970358	1.235366	-1.161846
2	-0.335981	0.050785	-0.922184	-0.059781	0.242284	-0.242524	-0.202474	0.141911	0.850821

Рисунок 10 – среднее значение признаков

Кластер 0: "Экономически-позитивный регион с проблемами здравоохранения"

- Продолжительность жизни (x1): ниже среднего;
- Смертность (x2): выше среднего;
- Число браков (x3): выше среднего;
- Число разводов (x4): выше среднего;
- Младенческая смертность (x5): ниже среднего;
- Отношение дохода к прожиточному минимуму (x6): выше среднего;
- Отношение оплаты труда к прожиточному минимуму (x7): выше среднего;
- Количество населения с доходами ниже прожиточного минимума (x8): ниже среднего;
- Преступность (x9): ниже среднего.

Интерпретация: кластер 0 имеет все положительные показатели выше среднего, а все отрицательные – ниже среднего, за одним исключением: продолжительность жизни ниже среднего, а смертность выше среднего. Это может говорить о том, что сфера здравоохранения в регионе развита не так

хорошо, как остальные.

Кластер 1: “Долгая жизнь, крепкие отношения и экономические трудности”

- Продолжительность жизни (x1): выше среднего;
- Смертность (x2): ниже среднего.
- Число браков (x3): выше среднего;
- Число разводов (x4): ниже среднего;
- Младенческая смертность (x5): выше среднего;
- Отношение дохода к прожиточному минимуму (x6): ниже среднего;
- Отношение оплаты труда к прожиточному минимуму (x7): ниже среднего;
- Количество населения с доходами ниже прожиточного минимума (x8): выше среднего;
- Преступность (x9): ниже среднего.

Интерпретация: кластер 1 можно охарактеризовать высокой продолжительностью жизни и низкой смертностью, что говорит о крепком здоровье населения, высоким числом браков и низком числом разводов, что говорит о хороших социальных отношениях, и низкими отношением дохода к прожиточному минимуму и отношением оплаты труда к прожиточному минимуму и высоким количеством населения с доходами ниже прожиточного минимума, что говорит об экономических трудностях в регионе.

Кластер 2: “Регион в процессе стабилизации”

- Продолжительность жизни (x1): ниже среднего;
- Смертность (x2): чуть выше среднего;
- Число браков (x3): ниже среднего;

– Число разводов (x4): ниже среднего;

– Младенческая смертность (x5): выше среднего;

– Отношение дохода к прожиточному минимум (x6): ниже среднего;

– Отношение оплаты труда к прожиточному минимуму (x7): ниже среднего;

– Количество населения с доходами ниже прожиточного минимума (x8): выше среднего;

– Преступность (x9): выше среднего.

Интерпретация: по этому кластеру можно сказать, что в рассматриваемом регионе дела совсем не очень. Об этом говорят низкие продолжительность жизни, число браков, отношение дохода к прожиточному минимуму, отношение оплаты труда к прожиточному минимуму и высокие преступность и младенческая смертность. Пока что такой регион стабильным назвать нельзя, поэтому дадим ему название «Регион в процессе стабилизации».

Вывод.

В ходе выполнения лабораторной работы было получено представление об иерархической кластеризации и методе k-средних. Говоря об иерархической кластеризации, в частности были изучены различные типы связей. Также было получено понимание, как можно сравнить схожесть результатов двух разных алгоритмов кластеризации.