Министерство науки и высшего образования Российской Федерации

ФГБОУ ВО «Ульяновский государственный технический университет»

Кафедра «Вычислительная техника»

**Лабораторная работа №5**

Дисциплина: «Системы искусственного интеллекта»

«Алгоритмы кластеризации данных»

Вариант 14

Выполнил студент

группы ИВТАCбд-31

Клишин К. С.

Проверил:

ассистент кафедры «ВТ»

Хайруллин И. Д.

Ульяновск, 2024

**Цель работы**

1. Произвести масштабирование признаков (scaling).
2. С использованием библиотеки sklearn написать программу с использованием алгоритмов кластеризации данных, позволяющую разделить исходную выборку на классы, соответствующие предложенной вариантом задаче.
3. Провести эксперименты и определить наилучший алгоритм кластеризации, параметры алгоритма. Необходимо использовать не менее 3-х алгоритмов. Данные экспериментов необходимо представить в отчете (графики, ход проведения эксперимента, выводы).

# Ход работы

По варианту задан датасет Dresses\_Attribute\_Sales, который представляет собой данные о продажах женских платьев на Aliexpress.

Загрузим датасет, и сразу произведём масштабирование признаков:

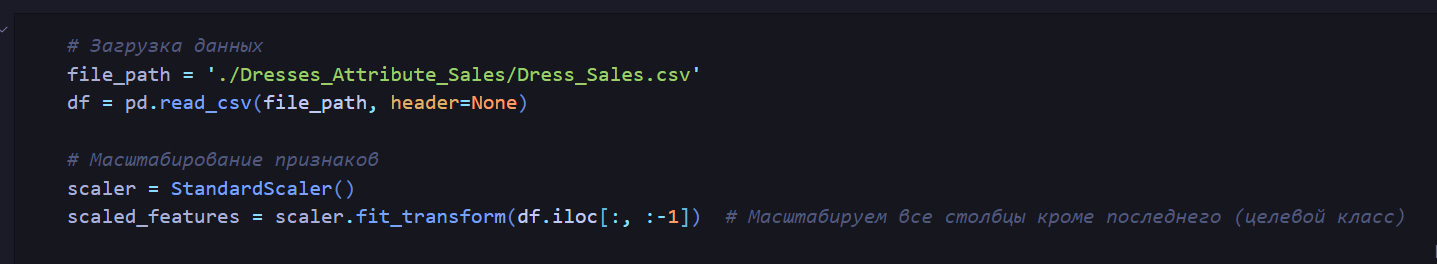


Рис. 1 – Загрузка датасета и масштабирование признаков

Используем метод локтя для определения необходимого количества кластеров для алгоритма кластеризации KMeans:

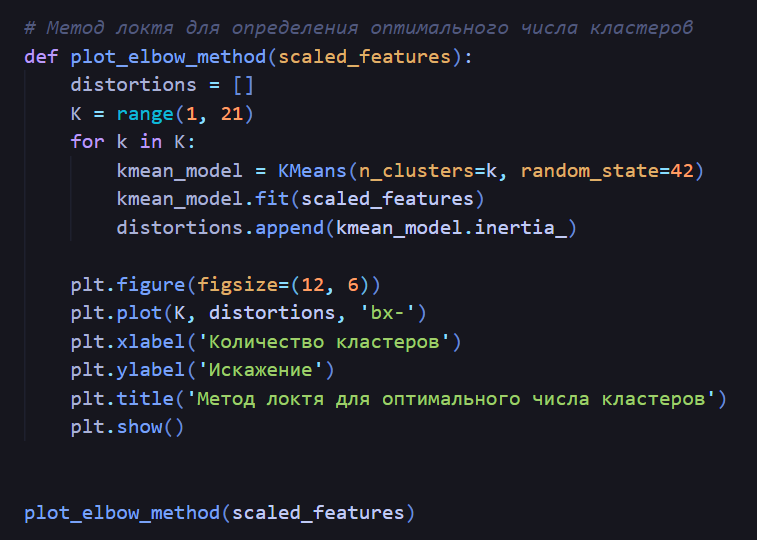


Рис. 2 – Метод локтя для определения необходимого количества кластеров

Получаем следующий график:

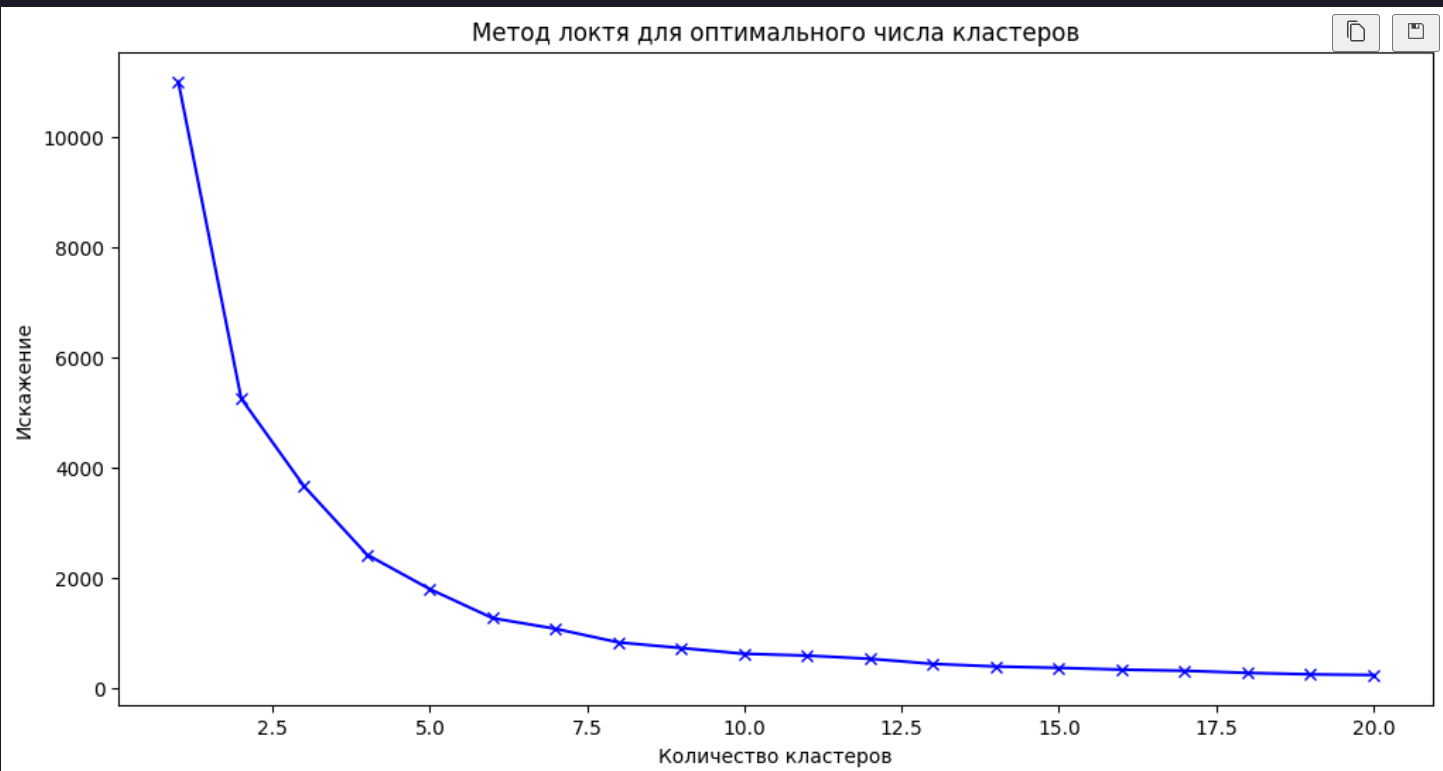


Рис. 3 – График метода локтя

Исходя из графика понимаем, что оптимальное количество кластеров – 6.

Далее – запускаем три алгоритма кластеризации:

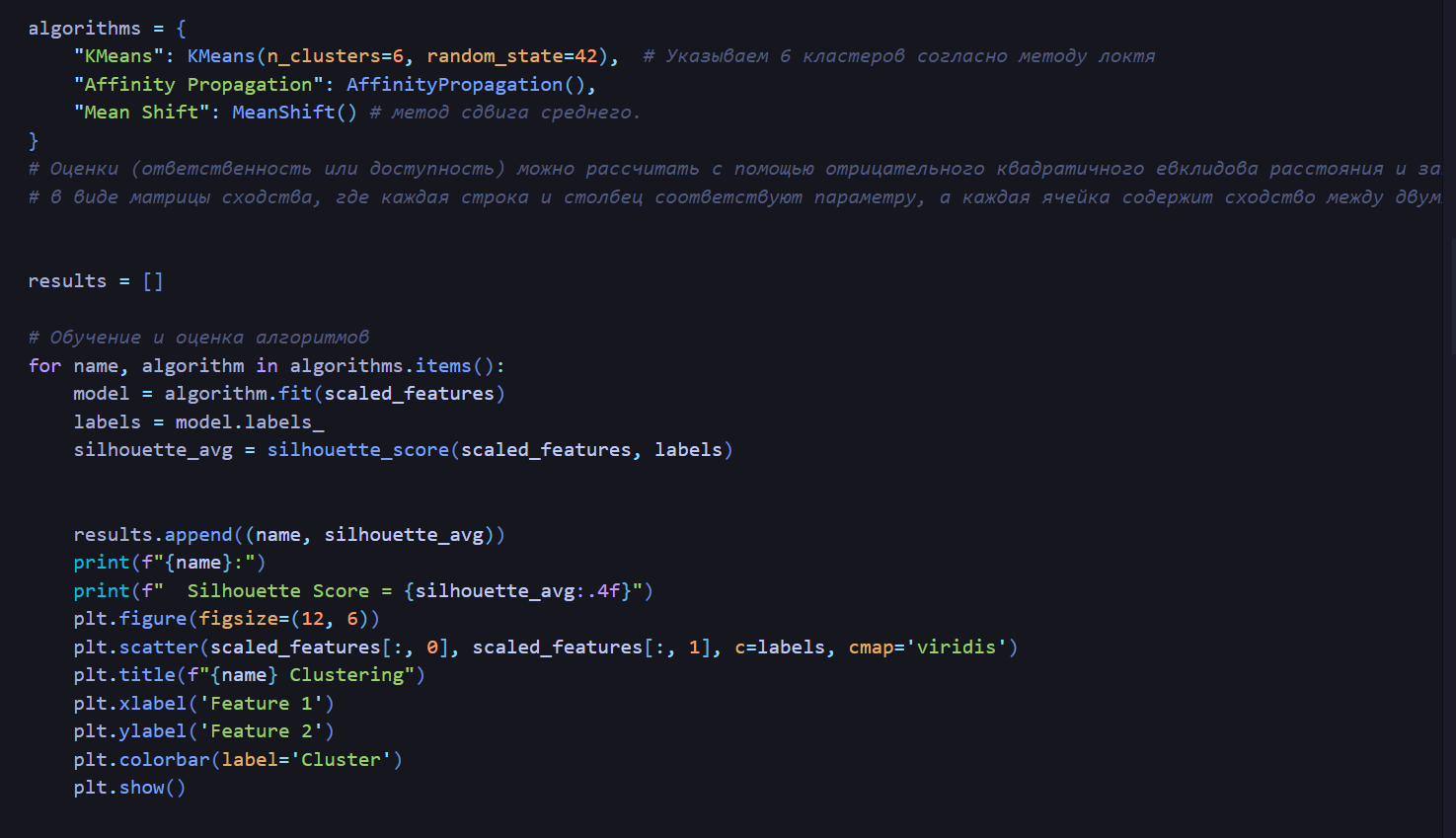


Рис. 4 – Код для тестирования каждого из алгоритмов кластеризации

Получаем следующие результаты Silhouette Score = 0.5949:

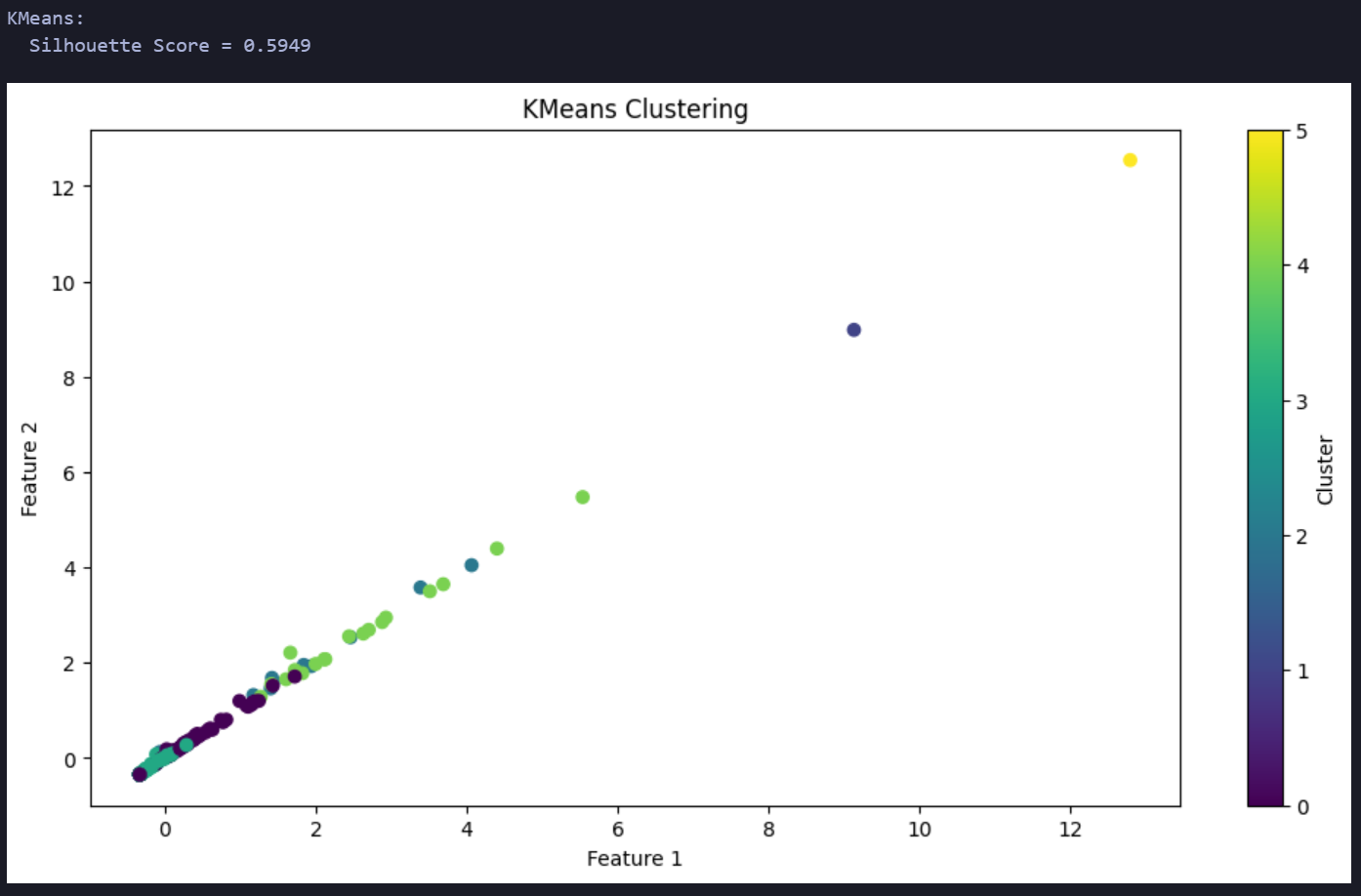


Рис. 5 – Результаты работы алгоритма кластеризации KMeans.

Далее – Affinity Propagation с Silhouette Score = 0.4003:

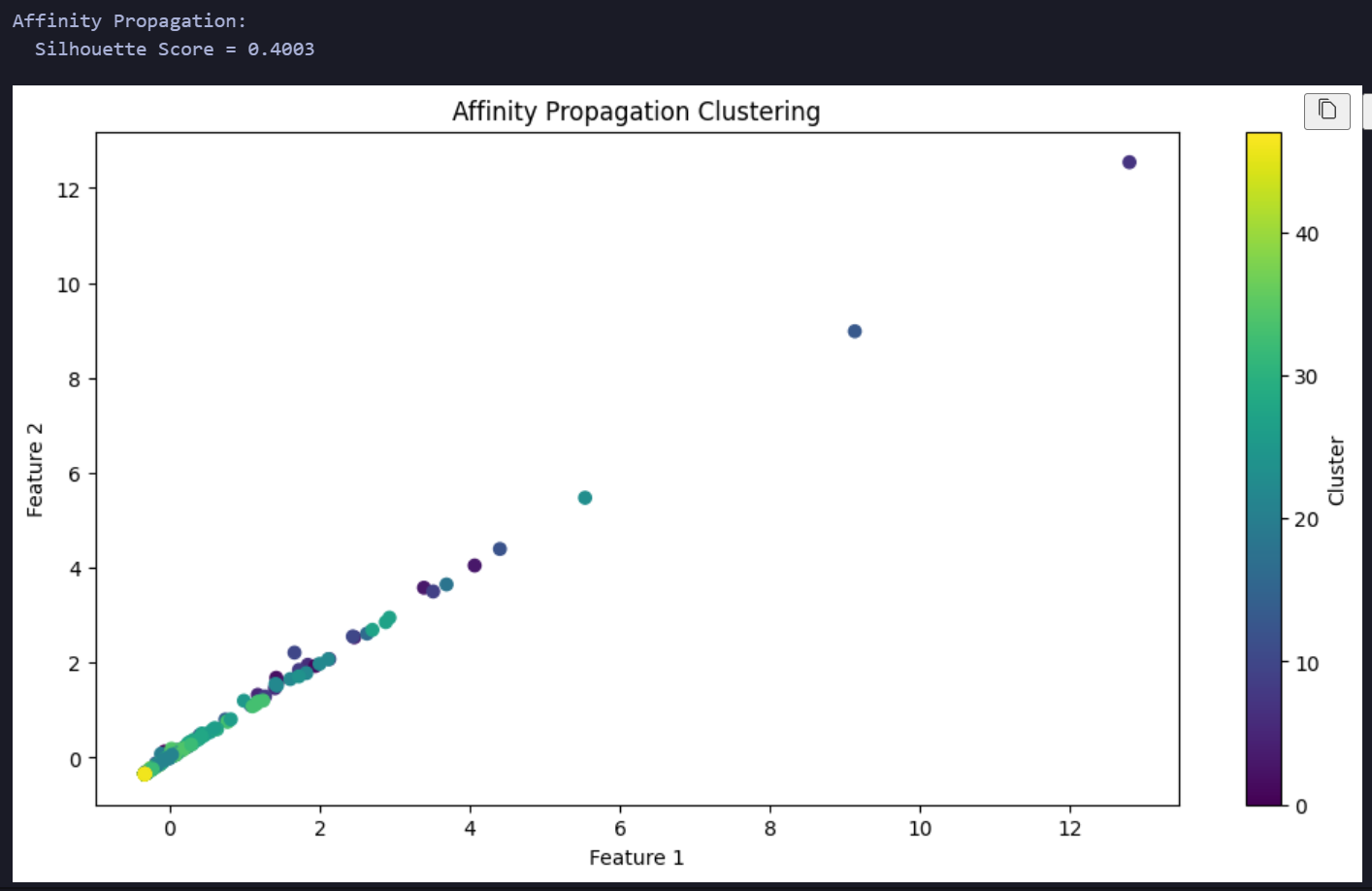
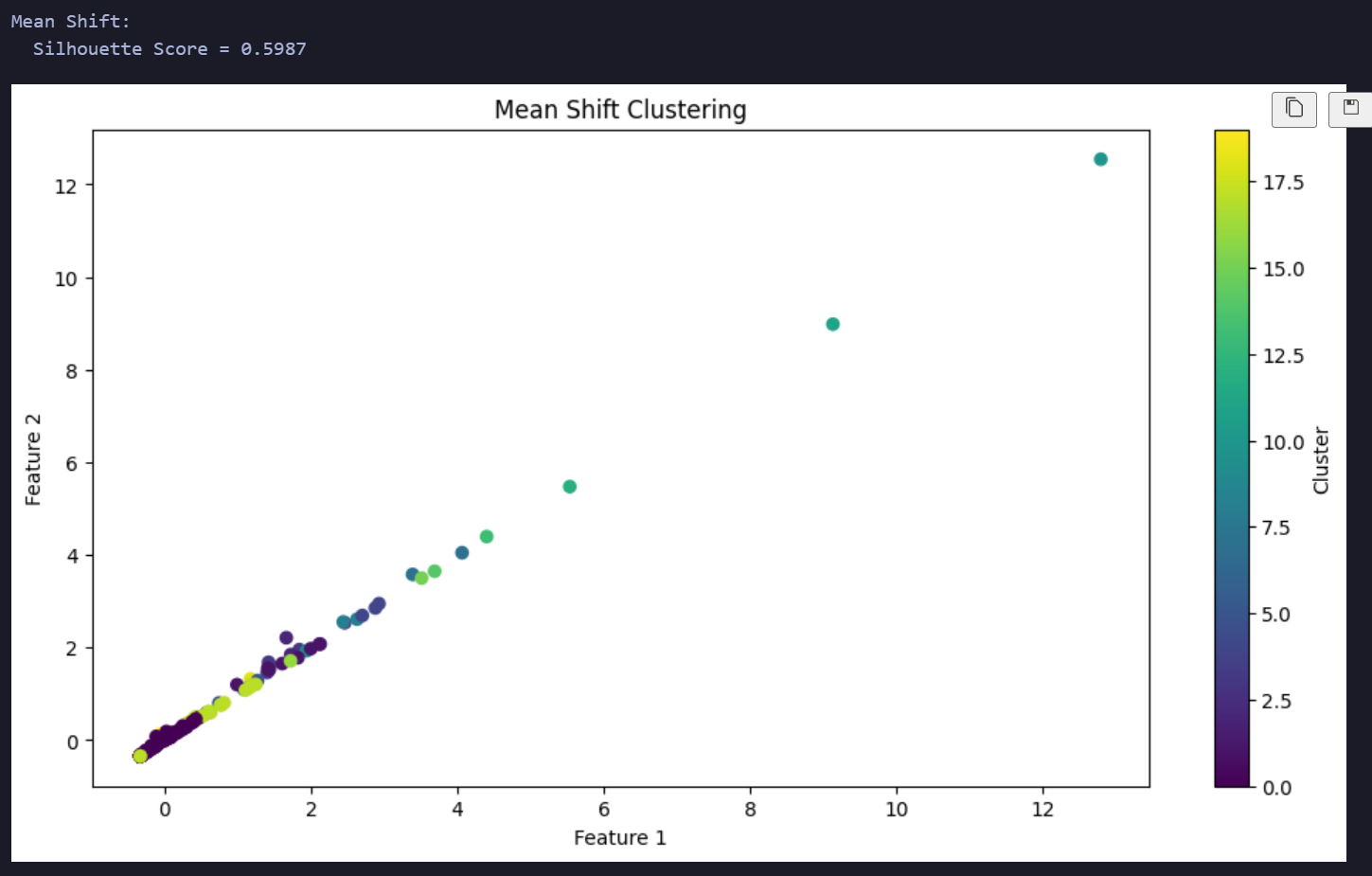


Рис. 6 – Результаты работы алгоритма кластеризации Affinity Propagation

И третий алгоритм – Mean Shift с Silhouette Score = 0.5987:



**Вывод**

В рамках данной лабораторной работы была проведена кластеризация текстовых данных с применением нескольких алгоритмов. В ходе работы были осуществлены различные этапы и методы анализа. В результате экспериментов было выявлено, что лучшие результаты показали алгоритмы KMeans и Mean Shift, оба показали схожие показатели. В то же время, Affinity Propagation оказался менее эффективным по сравнению с ними.