Отчет

**Описания задач:**

|  |
| --- |
| 1. Взять неразмеченный набор данных  2. Написать программу кластеризации данных из этого набора. Использовать 2 разных метрики схожести и 2 метода разных видов  3. Подсчитать 2 внутренние и 2 внешние меры качества работы методов  4. Написать отчет, в котором:  - описать выбранный набор данных  - описать выбранные метрики, методы кластеризации (не рассказанные – подробно), меры качества  - сравнить качество работы методов между собой  - сделать выводы |

**Выбранный набор данных:**

Ирисы Фишера

|  |
| --- |
| Ирисы Фишера состоят из данных о 150 экземплярах ириса, по 50 экземпляров из трёх видов — Ирис щетинистый (Iris setosa), Ирис виргинский (Iris virginica) и Ирис разноцветный (Iris versicolor). Для каждого экземпляра измерялись четыре характеристики (в сантиметрах):  1. Длина наружной доли околоцветника (англ. sepal length);  2. Ширина наружной доли околоцветника (англ. sepal width);  3. Длина внутренней доли околоцветника (англ. petal length);  4. Ширина внутренней доли околоцветника (англ. petal width).  https://machinelearningmastery.ru/standard-machine-learning-datasets/ |

**Методы класстеризации:**

1. K-means cluster in sk-learn library
2. Dendrogram cluster in scipy library

|  |
| --- |
| from sklearn.cluster import KMeans  from scipy.cluster import hierarchy as sch |

**Методы оценки:**

Внутренние меры оценки качества

|  |
| --- |
|  |

|  |
| --- |
|  |

Внешние меры оценки качества

|  |
| --- |
|  |

|  |
| --- |
|  |

**Сравнение качество работы методов**:

Значения меры:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Внутренние меры** | | **Внешние меры** | |
| Silhouette | Calinski-Harabasz | Jaccard | Fowlkes-Mallows |
| K-means | 0.420658 | 177.828798 | 0.037037 | 0.559645 |
| Dendrogram | 0.312965 | 66.965441 | 0.013333 | 0.650002 |

График кластеризации

|  |
| --- |
|  |

**Вывод:**

-1 < Silhouette < 1

0 < Calinski-Harabasz

0 < Jaccard < 1

0 < Fowlkes-Mallows < 1

1. Схожесть кластеров довольно высокие у обоих методов
2. Значения Jaccard и Calinski-Harabasz значительно ниже наблюдалось у Dendrogram. То есть схожесть с target лучше выделяется в K-means методе.
3. Силуэтноть у обойх методах положителен.