Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

*Дисциплина «Технологии машинного обучения»*

**Отчёт**

по рубежному контролю №2

Тема: «Технологии использования и оценки моделей машинного обучения.»

*Вариант 3*

Студент:

Белкина Е.В.

Группа ИУ5-61Б

Преподаватель:

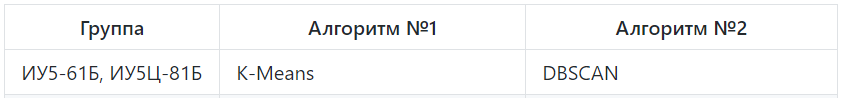
Гапанюк Ю.Е.

Москва, 2020 г.

**Задание**

Задача 2. ***Кластеризация данных*** (по вариантам)

Кластеризуйте данные с помощью двух алгоритмов кластеризации (варианты по группам приведены в таблице).



Сравните качество кластеризации с помощью следующих метрик качества кластеризации (если это возможно для Вашего набора данных):

Adjusted Rand index

Adjusted Mutual Information

Homogeneity, completeness, V-measure

Коэффициент силуэта

Сделайте выводы о том, какой алгоритм осуществляет более качественную кластеризацию на Вашем наборе данных.

Набор данных:

<https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine>

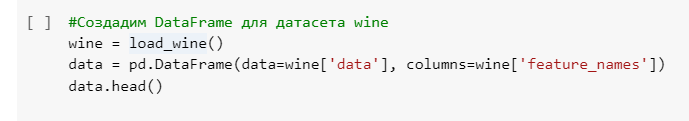
**Выполнение задания**

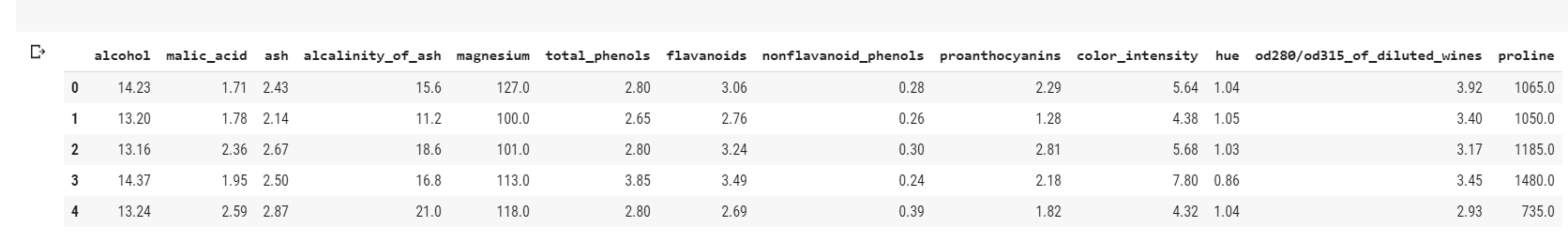
1. Импортируем необходимые библиотеки с помощью команды import.



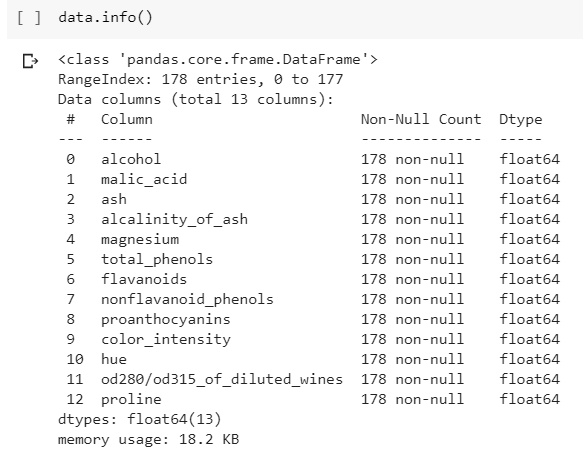
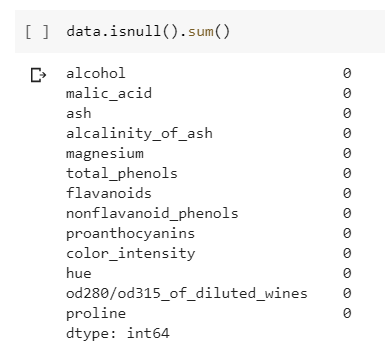
1. Импортируем датасет load\_wine из sklearn в соответствии с заданием варианта

Преобразуем датасет Scikit-learn в Pandas Dataframe



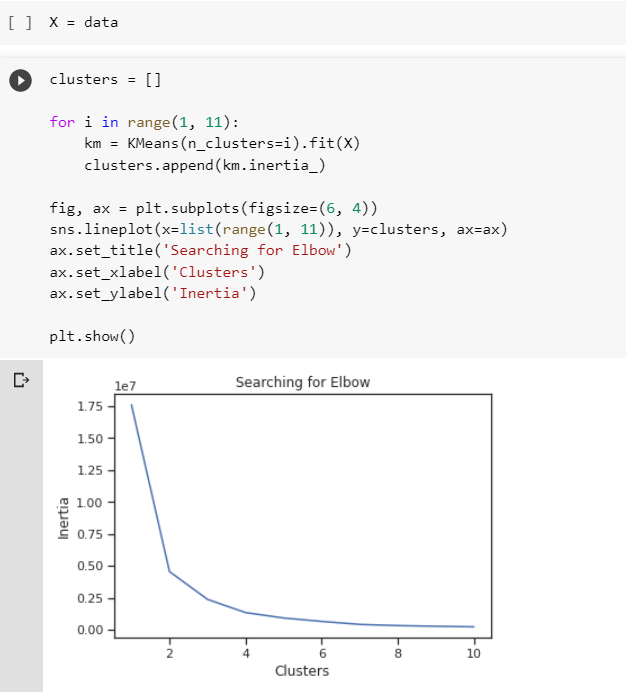


1. Проверим наличие пропусков данных

Можем видеть, что пропуски данных в датасете отсутствуют.

1. Подбор гиперпараметра количества кластеров



Используем "правило локтя".

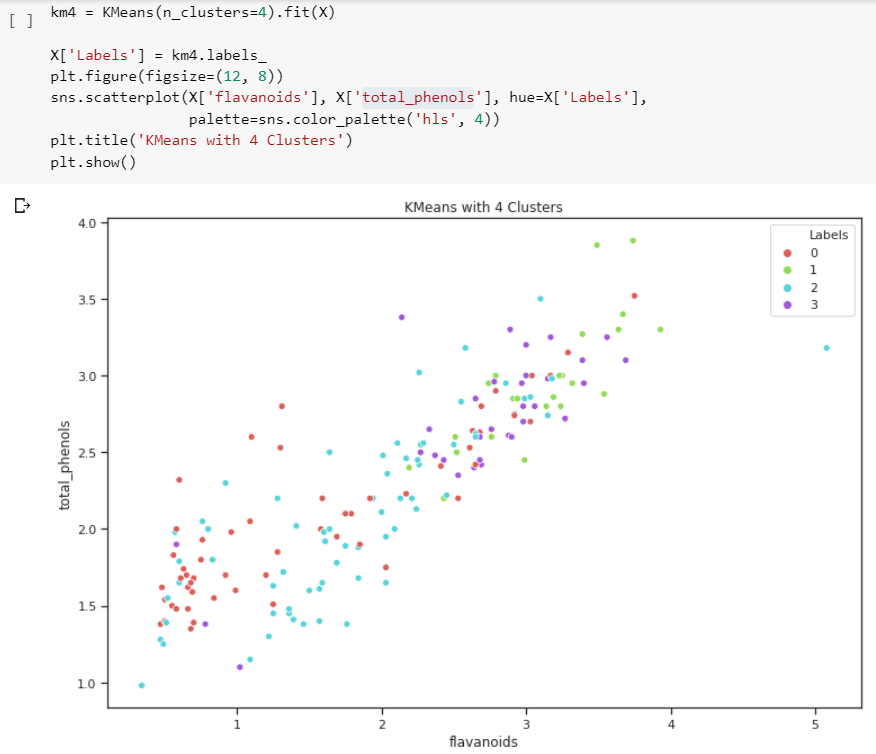
Видим, что после 4 кластеров уменьшение инерции резко замедляется.

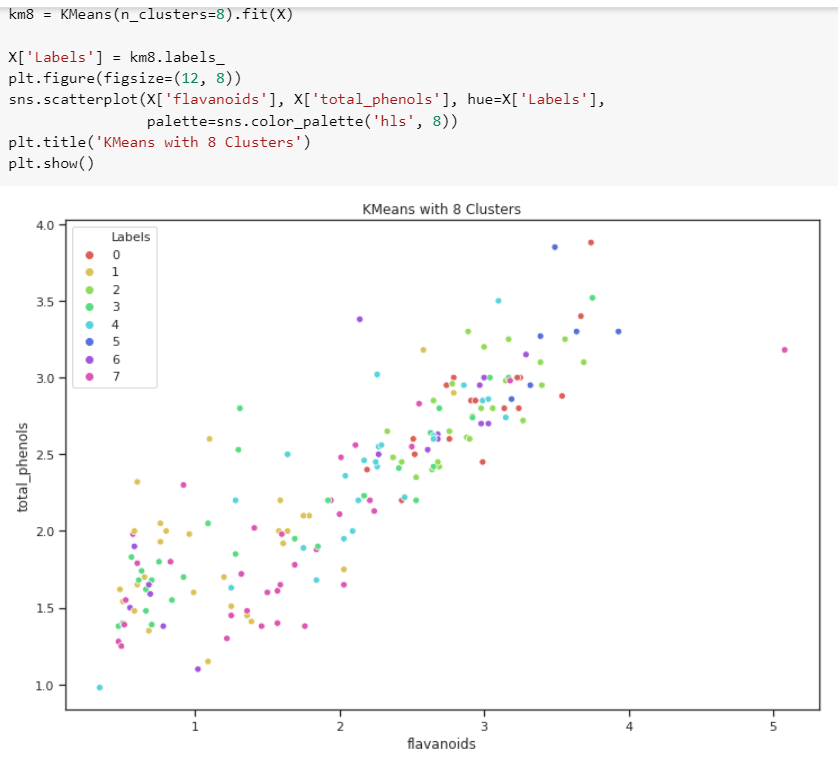
1. **Кластеризация K-Means**

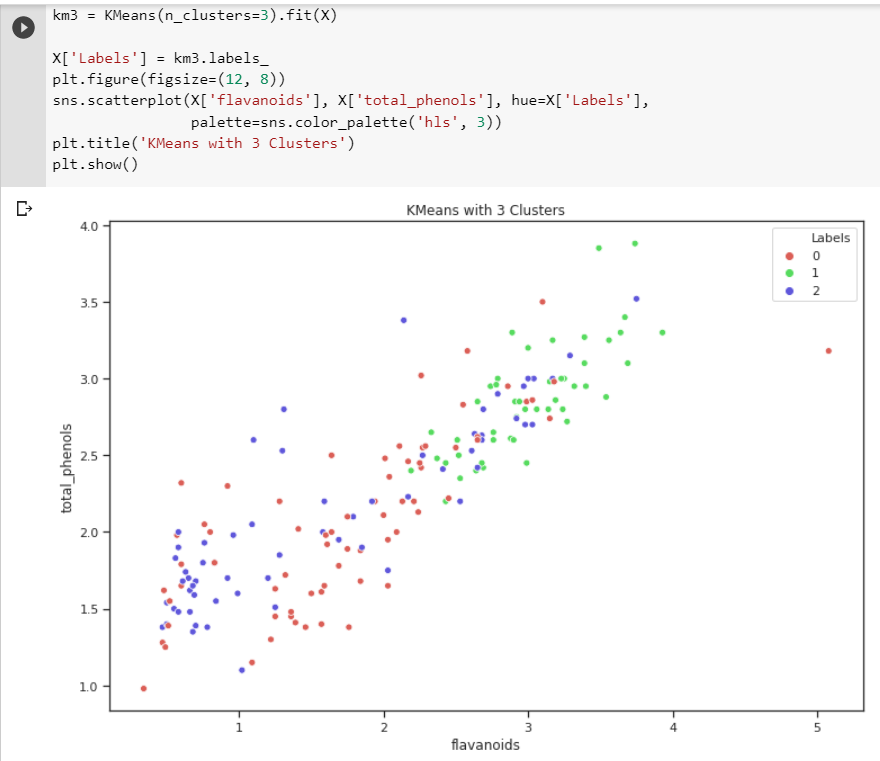
The KMeans algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares (see below). This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields.

n\_clusters*int, default=8*

The number of clusters to form as well as the number of centroids to generate.







1. **Кластеризация DBSCAN**

Perform DBSCAN clustering from vector array or distance matrix.

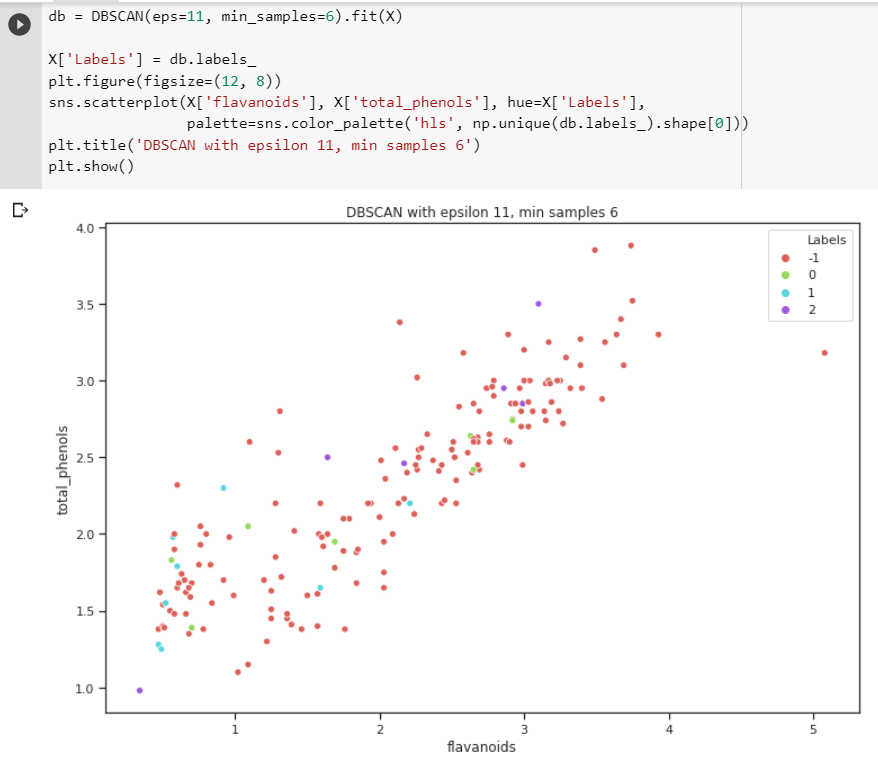
DBSCAN - Density-Based Spatial Clustering of Applications with Noise. Finds core samples of high density and expands clusters from them. Good for data which contains clusters of similar density.

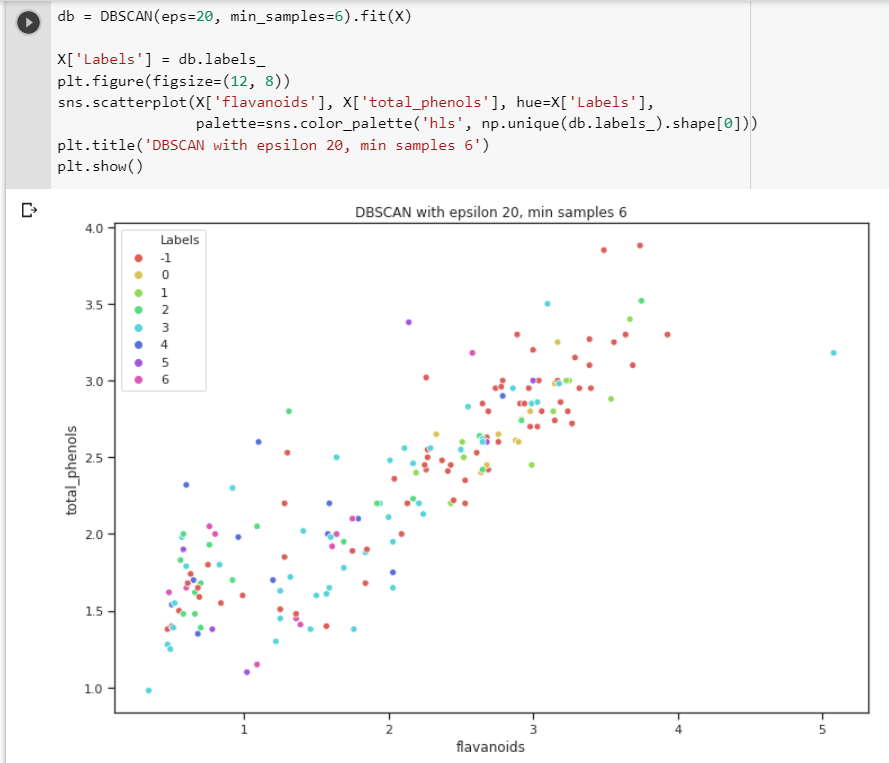
psfloat, default=0.5

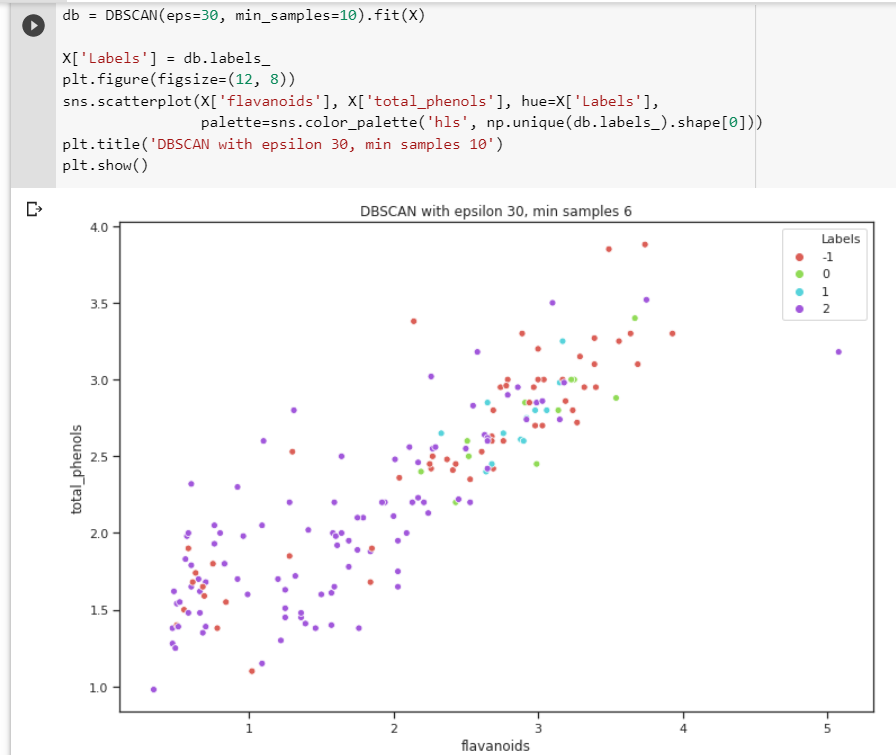
The maximum distance between two samples for one to be considered as in the neighborhood of the other. This is not a maximum bound on the distances of points within a cluster. This is the most important DBSCAN parameter to choose appropriately for your data set and distance function.

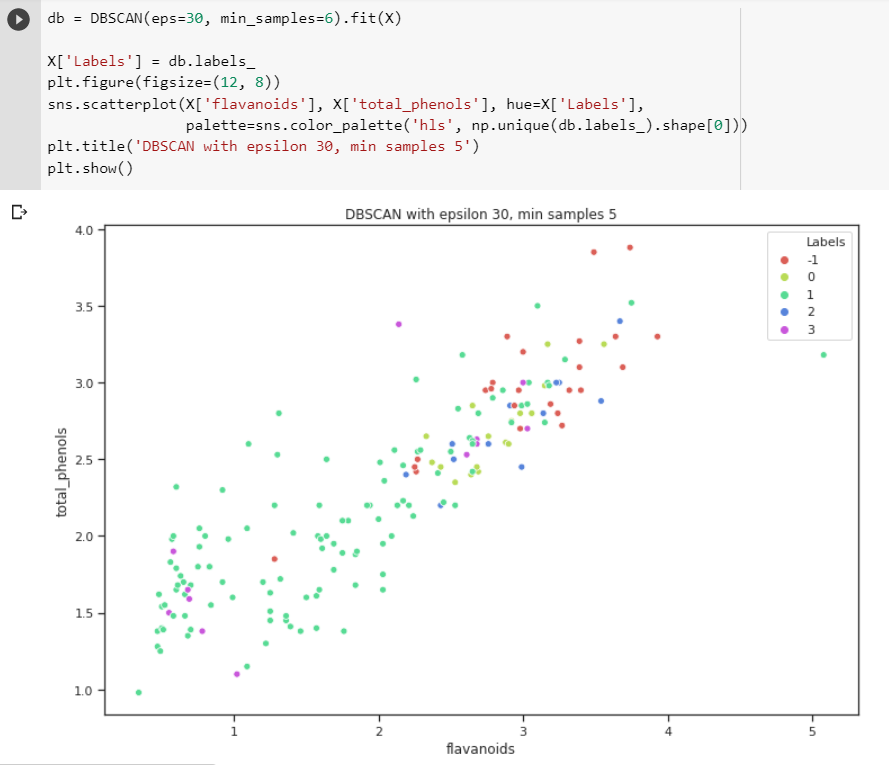
min\_samplesint, default=5

The number of samples (or total weight) in a neighborhood for a point to be considered as a core point. This includes the point itself.









1. **Метрики качества кластеризации**

1) Adjusted Rand index

Метрика возвращает результат в диапазоне [-1;+1]. Значение близкое к +1 говорит об очень хорошем качестве кластеризации. Значение близкое к 0 соответствует случайным разбиениям. Отрицательные значения говорят о плохом качестве кластеризации.

2) Adjusted Mutual Information

Значение близкое к +1 говорит об очень хорошем качестве кластеризации. Значение близкое к 0 соответствует случайным разбиениям.

3) Homogeneity, completeness, V-measure

Homogeneity - каждый кластер содержит только представителей единственного класса (под классом понимается истинное значение метки кластера). Значение в диапазоне [0;1], 1 говорит об очень хорошем качестве кластеризации.

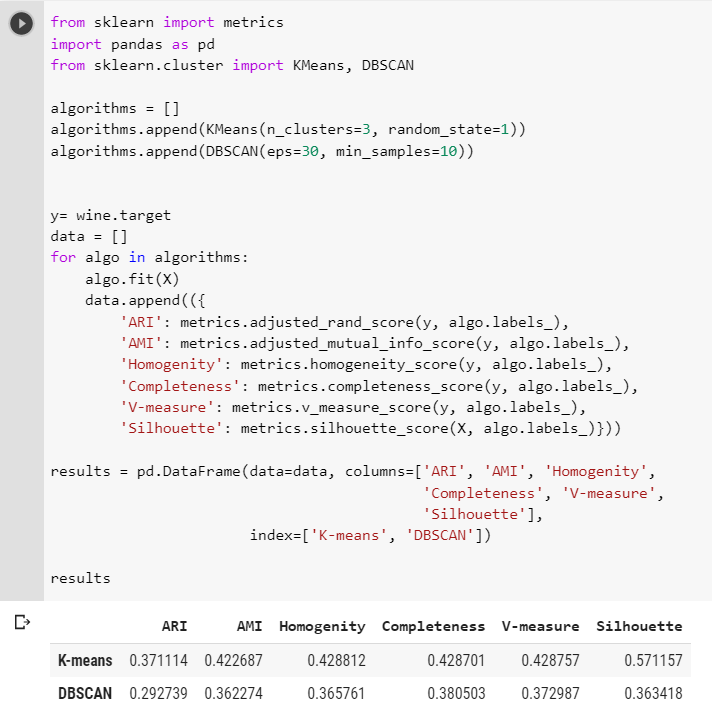
Completeness - все элементы одного класса помещены в один и тот же кластер. Значение в диапазоне [0;1], 1 говорит об очень хорошем качестве кластеризации.

V-measure - среднее гармоническое от Homogeneity и Completeness.

4) Коэффициент силуэта

Данный метод не требует знания истинных значений меток кластеров,ииспользуется подбора оптимального числа кластеров— выбирается число кластеров, максимизирующее значение силуэта.

Силуэтом выборки показывает, насколько среднее расстояние до объектов своего кластера отличается от среднего расстояния до объектов других кластеров. Данная величина лежит в диапазоне [-1;1]. Значения, близкие к -1, соответствуют плохим (разрозненным) кластеризациям, значения, близкие к нулю, говорят о том, что кластеры пересекаются и накладываются друг на друга, значения, близкие к 1, соответствуют "плотным" четко выделенным кластерам. Таким образом, чем больше силуэт, тем более четко выделены кластеры, и они представляют собой компактные, плотно сгруппированные облака точек.



1. Выводы

По данным полученным данным метрик (ARI, AMI, Homogenity, Completeness, V-measure, Silhouette) можем сделать следующие выводы:

* Алгоритм K-Means осуществляет более качественную кластеризацию на используемом в данной работе наборе данных, чем алгоритм DBSCAN, так как значения метрик, полученные для K-Means более близки к 1, что свидетельствует о более высоком качестве кластеризации.