**МИНОБРНАУКИ РОССИИ**

**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ  
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**Кафедра вычислительной техники**

**ОТЧЁТ  
по дисциплине «Машинное обучение»**

| Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| --- | --- |
| Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Гатауллин Р.И. |

**Лабораторная работа № 3  
Тема: «Кластеризация»**

Санкт-Петербург 2023

**Оглавление**

[Цель работы 3](#_30j0zll)

[Задачи 3](#_1fob9te)

[Ход работы 4](#_2et92p0)

[Теория 4](#_tyjcwt)

[KMeans. Метод K-средних 5](#_17dp8vu)

Иерархическая кластеризация [8](#_26in1rg)

[DBSCAN 11](#_lnxbz9)

[Вывод 13](#_3rdcrjn)

**Цель работы**

Получение и закрепление навыков предобработки данных и применение методов машинного обучения для решения задач кластеризации.

# **Задачи**

1. Обучение моделей и подбор параметров
2. Оценка обученных моделей

**Ход работы**

# **Теория**

Задача кластеризации – это одна из основных задач в машинном обучении, которая группирует отдельные экземпляры данных в кластеры со сходными характеристиками. Кластеризацию можно также использовать для определения в наборе данных связей, которые невозможно логически отследить просмотром или наблюдением данных.

Ниже представлена краткая информация о некоторых алгоритмах кластеризации:

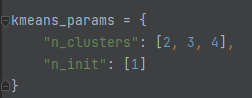
1. K-средних: наиболее популярный метод кластеризации. Смысл алгоритма в том, что он стремится минимизировать квадратичное отклонение точек кластеров от центров этих кластеров.

2. DBScan: плотностной алгоритм пространственной кластеризации с присутствием шума. Алгоритм кластеризации, основанный на плотности данных. Алгоритм группирует вместе точки, которые тесно расположены (точки со многими близкими соседями), помечая как выбросы точки, которые находятся одиноко в областях с малой плотностью (ближайшие соседи которых лежат далеко).

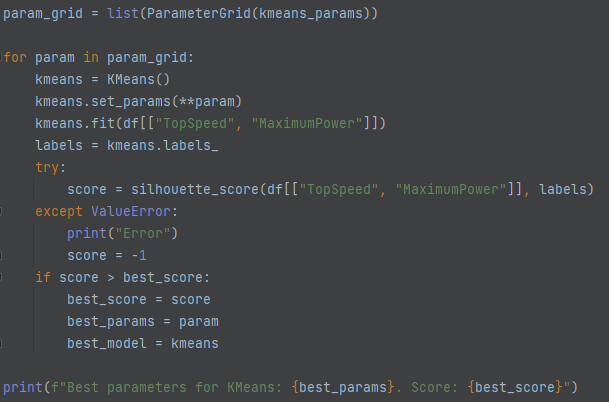
3. Иерархическая кластеризация: совокупность алгоритмов упорядочивания данных, направленных на создание иерархии (дерева) вложенных кластеров. Алгоритмы иерархической кластеризации предполагают, что анализируемое множество объектов характеризуется определенной степенью связности.

# **KMeans. Метод K-средних.**

Используемые параметры



Код метода

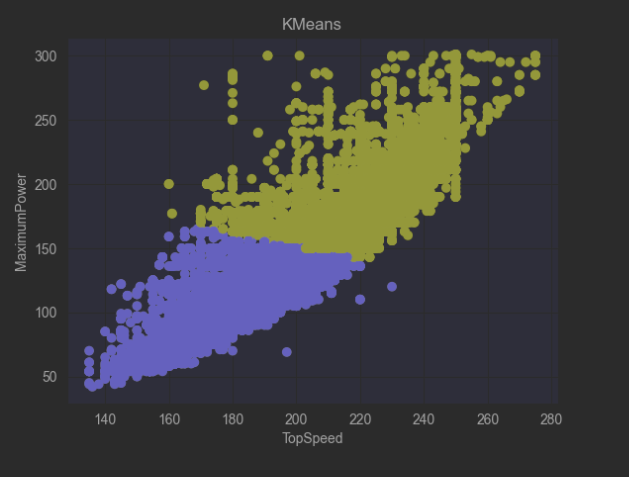


Оценка достоверности 52.96%



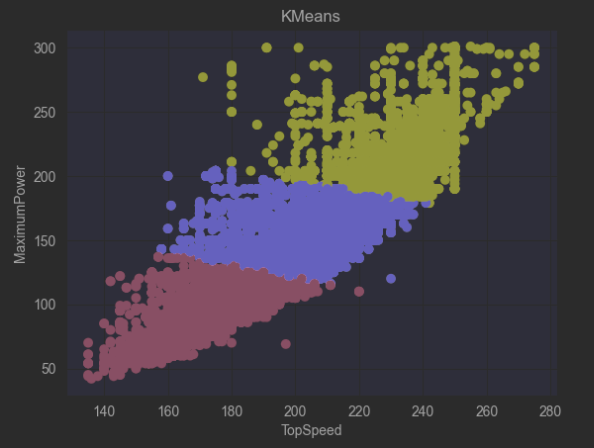
Визуализация полученных данных

Для 2 кластеров



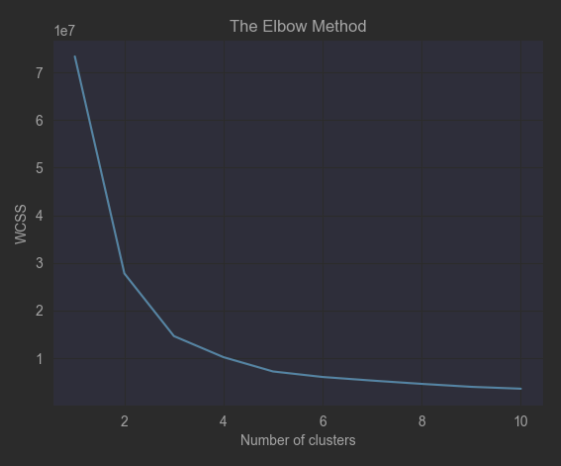


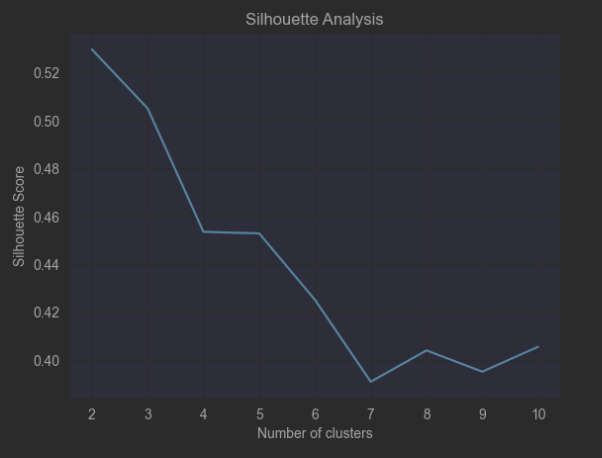
Для 3 кластеров

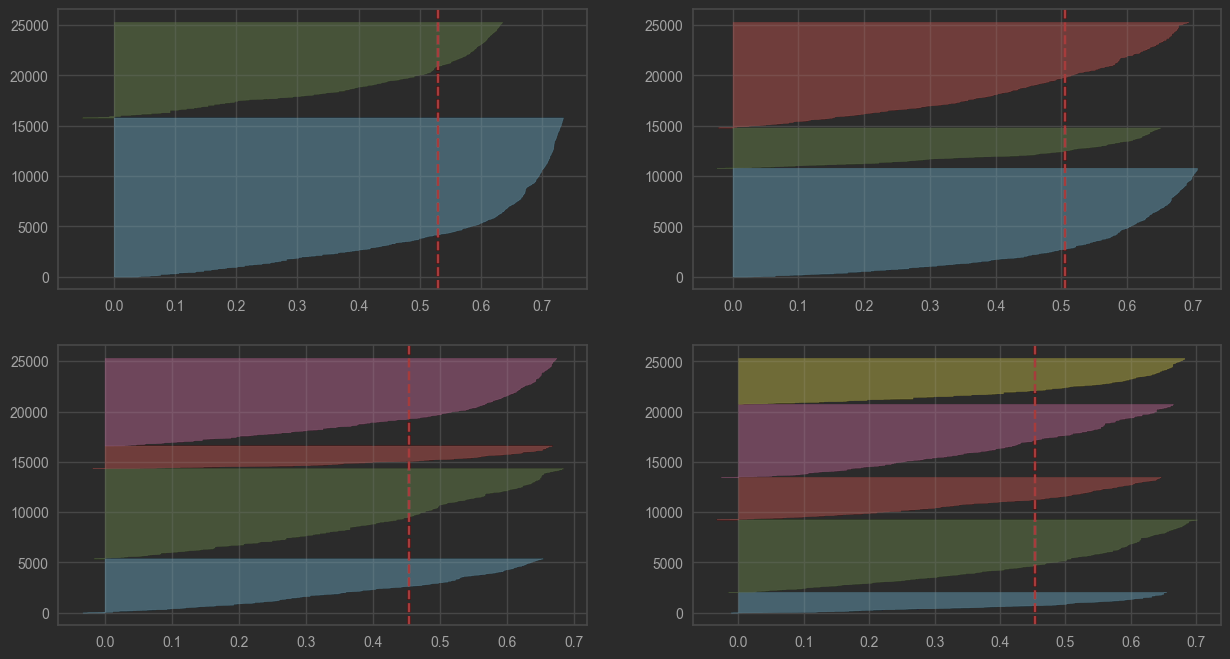




Метод локтя

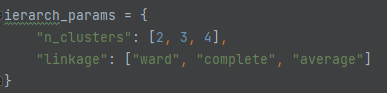




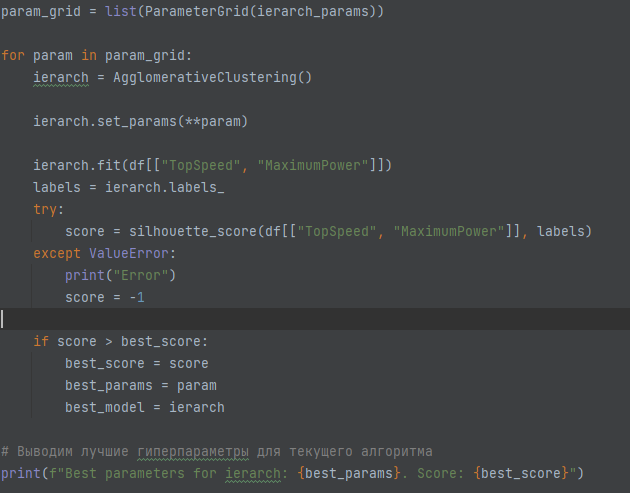


# **Иерархическая кластеризация**

Используемые параметры



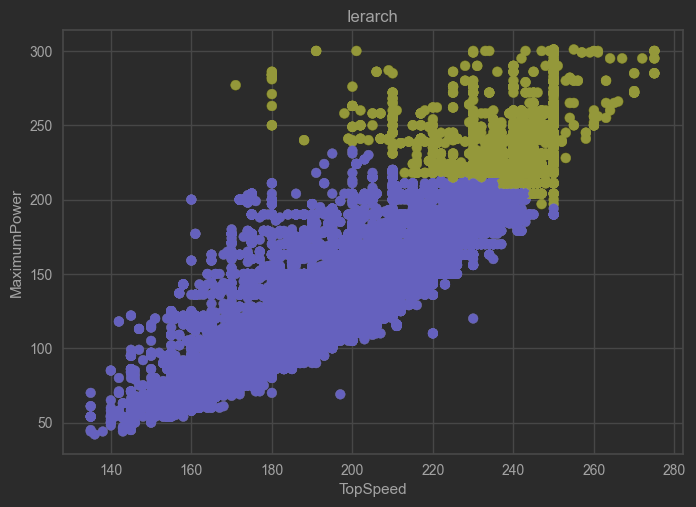
Код метода



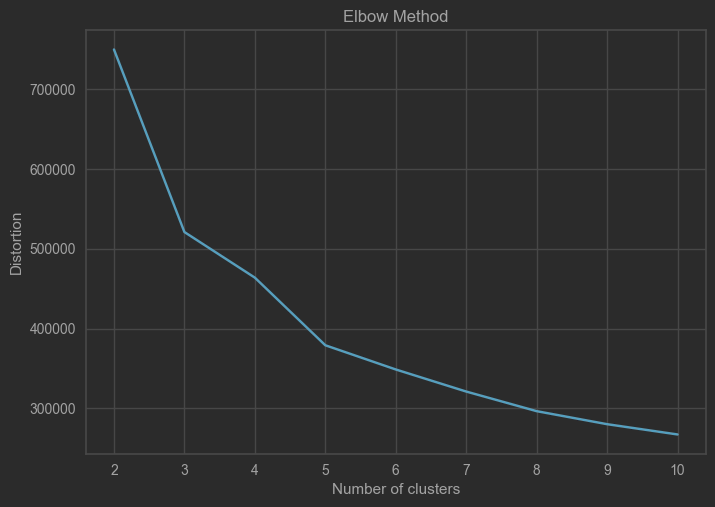
Оценка достоверности 54.32%



Визуализация полученных данных

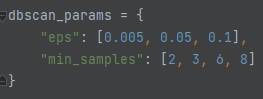


Метод локтя

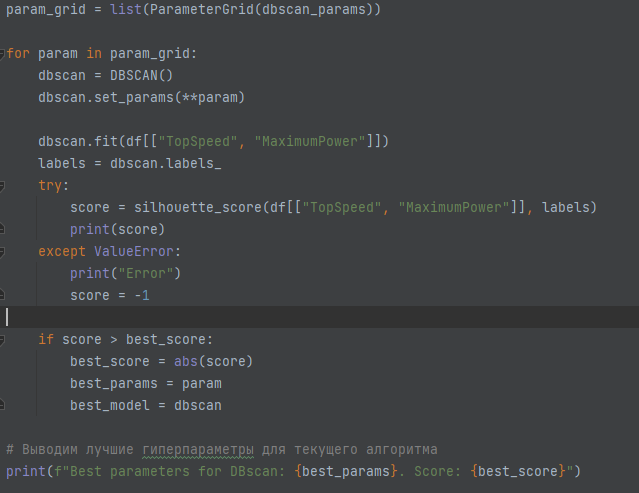


# **DBSCAN**

Используемые параметры



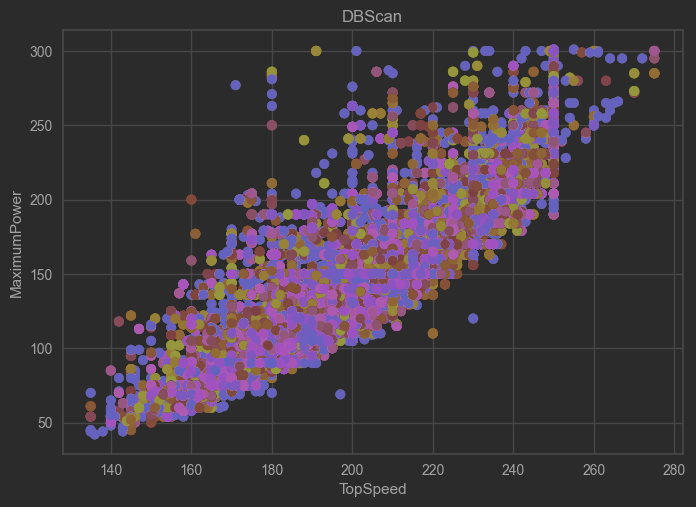
Код метода



Оценка достоверности 93.6%



Визуализация полученных данных



# **Вывод**

В ходе лабораторной работы нами были применены три различных метода кластеризации: K-средних, DBSCAN, Иерархический. KMEANS и LERARCH показали примерно одинаковый результат в 51% при использовании silhouette\_score, а DBSCAN – 93.6%. Данные и умения полученные в ходе работы могли бы упростить работу над прошлой лабораторной и улучшить результаты полученные в ней.