

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

**«Дальневосточный федеральный университет»**

**(ДВФУ)**

**ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ**

(ШКОЛА)

**Департамент информационных и компьютерных систем**

**ОТЧЕТ**

по дисциплине «системы искусственного интеллекта»

|  |  |
| --- | --- |
| Выполнил студенты группы Б9122-09.03.03пикд | |
|  | Зверев Р. И. |
|  | |
| Проверил преподаватель | |
|  | Бочарова В. В. |
|  | |
|  | |
| зачтено/не зачтено | |

г. Владивосток

2025 г

**Оглавление**

[**Цель работы** 3](#_Toc212596312)

[**Введение** 4](#_Toc212596313)

[**Описание данных** 5](#_Toc212596314)

[**Нахождение оптимального числа кластеров** 6](#_Toc212596315)

[**Визуализация кластеризации датасета** 9](#_Toc212596316)

[**Kmeans** 9](#_Toc212596317)

[**EM алгоритм** 10](#_Toc212596318)

[**Заключение** 13](#_Toc212596319)

[**Список литературы** 14](#_Toc212596320)

# **Цель работы**

Целью работы является освоение алгоритмов кластеризации, такие как Kmeans и EM алгоритм на наборе данных Quake.

**Постановка задачи**

В данной работе рассматривается задача кластеризации данных.

Необходимо реализовать следующие этапы и функции:

* Построить графики показателей (сумма квадратов расстояний, отношение среднего внутрикластерного расстояния к внекластерному) для различного количества кластеров для набора данных Quake;
* Подобрать оптимальное количество кластеров.

# **Введение**

В этой работе проводится практическое исследование методов кластеризации на реальном наборе данных Quake. Будут рассмотрены два подхода — жёсткий алгоритм K-means и вероятностный EM (Gaussian Mixture) — с реализацией и сравнением их поведения. Для оценки качества кластеризации будут построены графики зависимостей ключевых метрик от числа кластеров (сумма квадратов расстояний внутри кластеров и отношение среднего внутрикластерного к внекластерному расстоянию), после чего на основе этих графиков будет выбрано оптимальное число кластеров и сделаны выводы о структуре данных.

# **Описание данных**

Датасет Quake предоставляет данные о землетрясениях. Всего в нем 2178 объектов (записей) и 4 признака. Признаки:

* **Focal\_depth**: глубина точки гипоцентра землетрясения;
* **Latitude**: Широта точки землетрясения;
* **Longitude**: Долгота точки землетрясения;
* **Richet**: Сила землетрясения по шкале Рихтера.

# **Нахождение оптимального числа кластеров**

Построим график количества кластеров и суммы квадратов внутри кластерных расстояний:

Данный график и метод его построения называется «Метод локтя». Оптимальное количество кластеров находится в точке «локтя». На данном графике это число 3 явно и 6 тоже похоже.

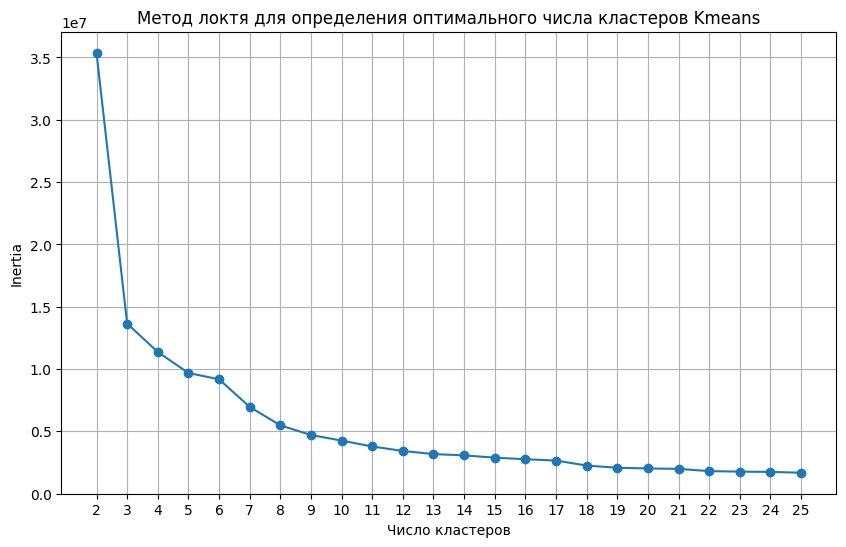


Рисунок 1 Метод локтя для определения оптимального числа кластеров

Воспользуемся методом «Силуэт кластера» (Silhouette Score), который измеряет, насколько объект похож на объекты своего кластера по сравнению с объектами других кластеров. График выглядит так:

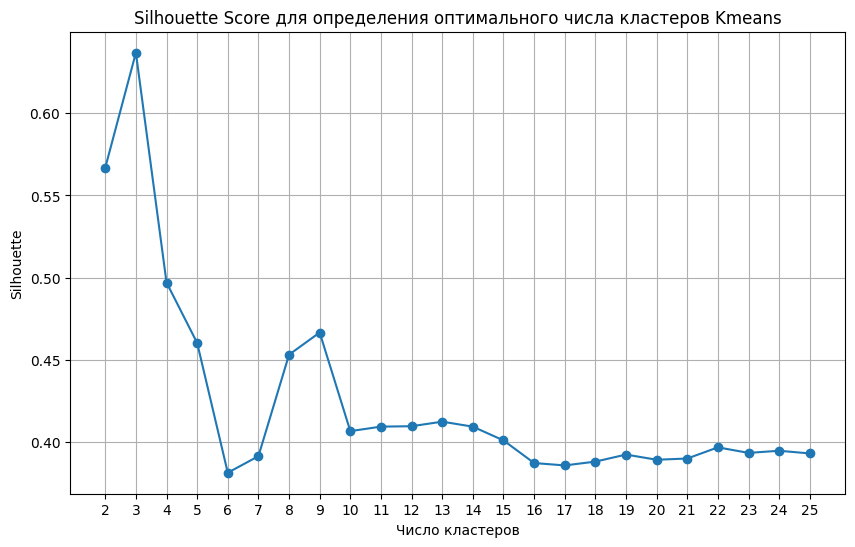
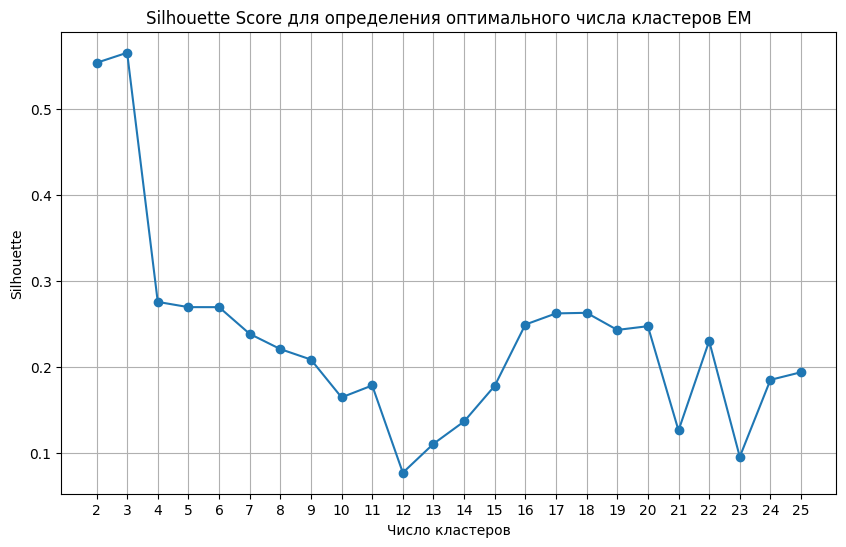


Рисунок 2 График Silhouette Score для определения оптимального числа кластеров

В данном методе оптимальное число — это верхняя точка на графике. В данном случае это 3. Этот метод более точен, поэтому оптимальное число кластеров 3.

Для выявления оптимального числа кластеров для EM алгоритма воспользуемся одним методом - «Silhouette Score». Его график будет таким:

Рисунок 3. Silhouette Score для EM алгоритма

По графику видно, что оптимальное число кластеров для EM алгоритма это 3. Для двух алгоритмов получилось, что оптимальное число кластеров — 3.

# **Визуализация кластеризации датасета**

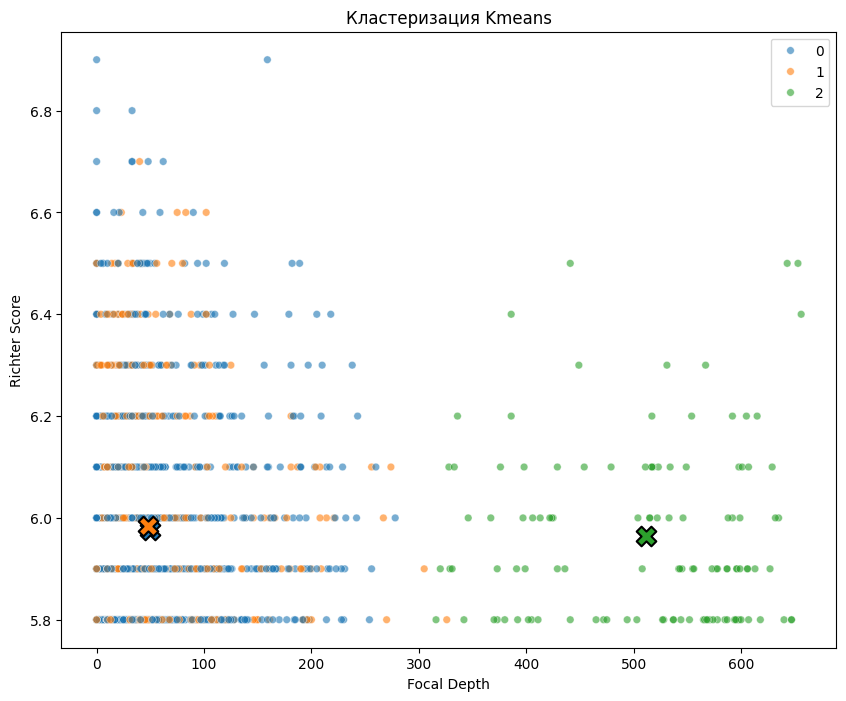
## **Kmeans**

В результате кластеризации датасета по оптимальному числу класетров получилась такая картина:

* **1ый кластер:** 1437 записей;
* **2ой кластер:** 623 записи;
* **3ий кластер:** 118 записей.

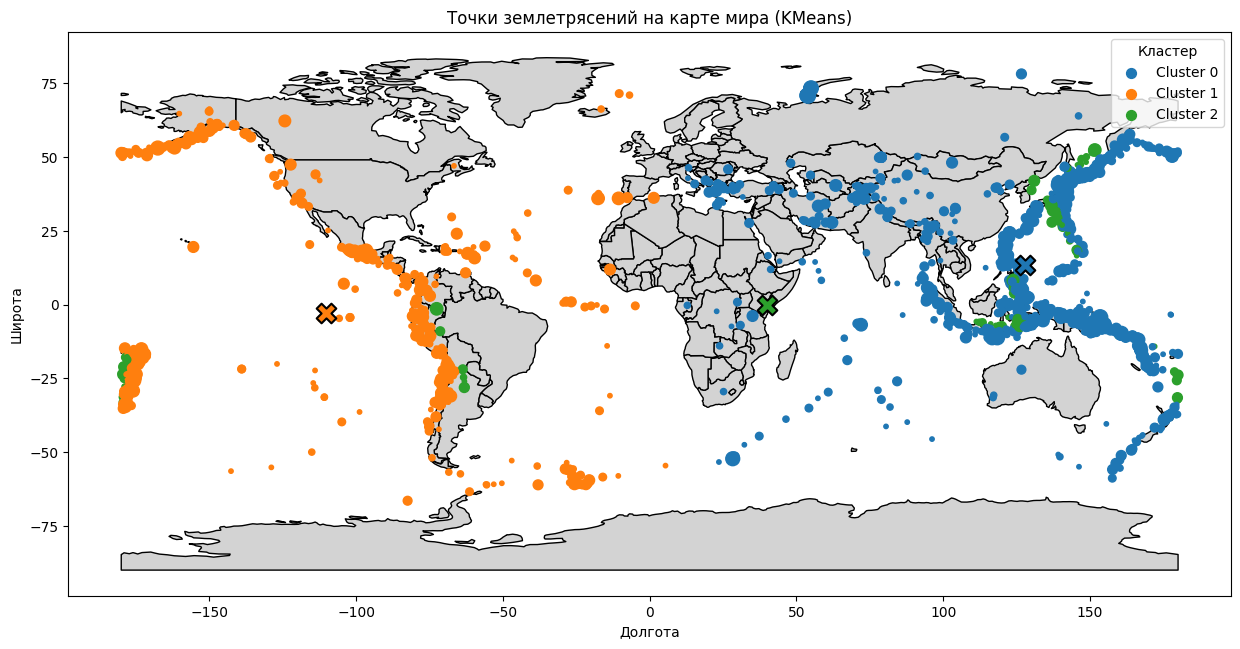
Посмотрим на распределение кластеров с точки зрения глубины землетрясения и шкалы Рихтера:

Как видно, чем ниже глубина гипоцентра землетрясения, тем сильнее балл по шкале Рихтера. Основная часть землетрясений достигает 6,5 баллов по данной шкале.

Рисунок 4. Визуализация кластеров по глубине землетрясения и шкале Рихтера

Рассмотрим кластеры на мировой карте:

Можно заметить, благодаря такой визуализации, что два самых больших кластера разделяют восточную и западную части света, а еще, что более интересно, разделяют между собой Тихоокеанское огненное кольцо. Третий кластер при кластеризации алгоритмом Kmeans отвечает за некие выбросы по всему свету и не имеет четкой принадлежности к части света.

Рисунок 5. Кластеры на карте мира

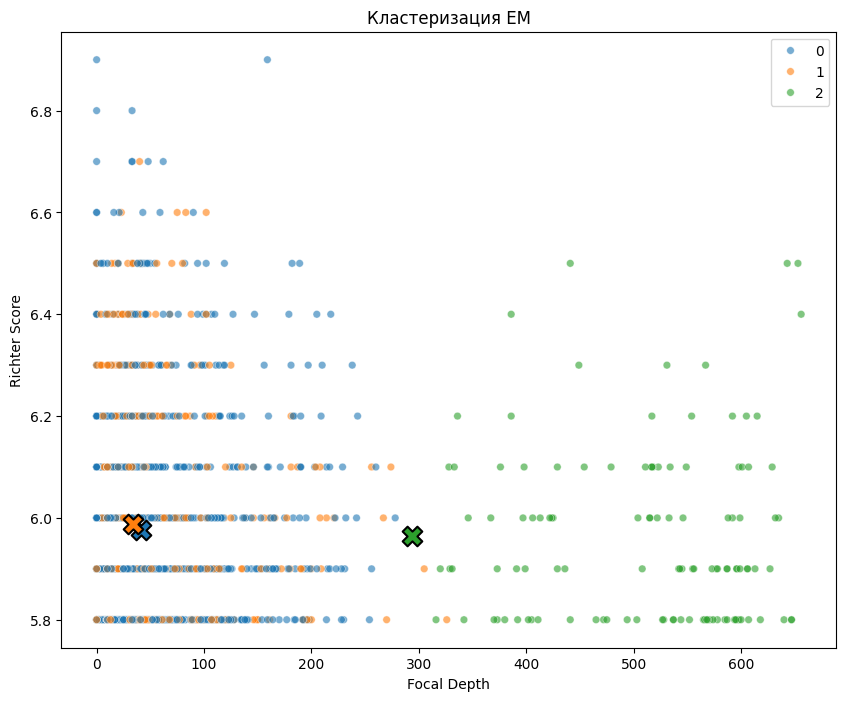
## **EM алгоритм**

В результате кластеризации датасета по оптимальному числу кластеров получилась такая картина:

* **1ый кластер:** 1293 записи;
* **2ой кластер:** 604 записи;
* **3ий кластер:** 281 запись.

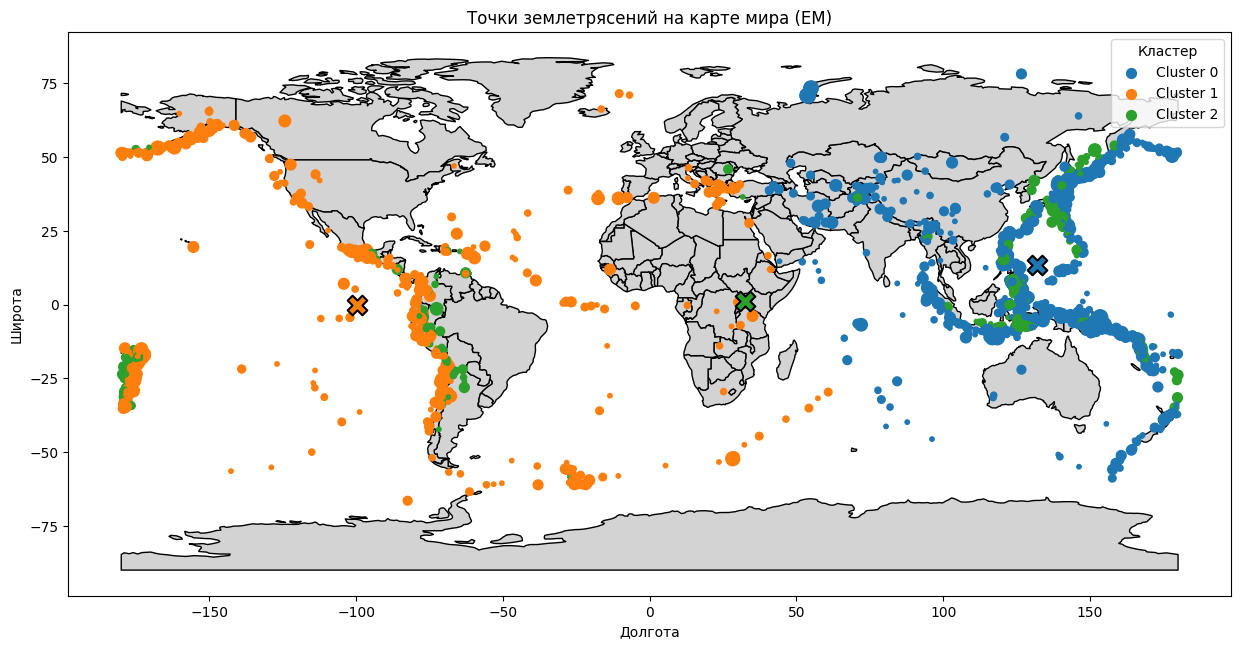
Посмотрим на распределение кластеров с точки зрения глубины землетрясения и шкалы Рихтера:

Как видно, данный алгоритм, так же, отображает закономерность: чем ниже точка гипоцентра землетрясения, тем выше балл по шкале Рихтера.

Рисунок 6. Визуализация кластеров по глубине землетрясения и шкале Рихтера

Рассмотрим, как данный алгоритм распределил кластеры по мировой карте:

Можно заметить, что данный алгоритм не отражает конкретных закономерностей кластеризации по географическому положению, в отличии от Kmeans.

Рисунок 7. Кластеры на карте мира

# **Заключение**

В ходе работы изучен набор данных Quake (2178 записей, 4 признака: глубина гипоцентра, широта, долгота, шкала Рихтера). Проведен анализ кластерной структуры с помощью двух подходов: жёсткого K-means и вероятностного EM. Для выбора числа кластеров были использованы несколько критериев.

Метод локтя (график числа кластеров vs сумма квадратов внутри кластерных расстояний) показал выраженные перегибы в точках вокруг k=3 и k=6 что указывает на наличие нескольких возможных уровней детализации кластеризации. Метод, основанный на отношении среднего внутри кластерного расстояния к вне кластерному был рассчитан и использован как дополнительная контрольная метрика при построении графиков метрик против числа кластеров.

Метод силуэта (Silhouette Score), дающий оценку однородности и отделенности кластеров, выявил максимум при k=3 для K-means — поэтому для K-means был выбран k=3. Для EM по графику силуэта оптимальным оказалось так же k=3.

В ходе визуализации кластеризации разных алгоритмов было замечено, что EM алгоритм хуже выявил географические особенности для кластеров, нежели Kmeans. В результате таких наблюдений можно сделать вывод, что алгоритм Kmeans **лучше**, чем EM.

# **Список литературы**

1. GitHub: исходный код лабораторной работы. – URL: [Лабораторная работа №3.8](https://github.com/FREDY129053/AI_Systems_LABS/blob/main/Lab_3_8.ipynb) (дата обращения: [09.10.2025]). – Текст: электронный.