**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра информационных систем**

отчет

**по лабораторной работе №2**

**по дисциплине «Машинное обучение»**

Тема: Метод Кластеризации K-Mean

|  |  |  |
| --- | --- | --- |
| Студент гр. 2372 |  | Михеева Э. Д. |
| Преподаватель |  | Татчина Я. А. |

Санкт-Петербург

2024

**Цель работы**

Знакомство с методом кластеризации K-mean с помощью пакета sklearn.

**Постановка задачи**

1. Добавить новый атрибут к текущему датасету;
2. Удалить выбросы и дубликаты, обработать пропущенные значения;
3. Построить графики зависимости одной переменной от другой, описать, какие кластеры видно;
4. Применить метод KMeans к датасету;
5. Проанализировать полученные результаты.

**Выполнение работы**

Для дальнейшей работы и создания новых атрибутов, данные необходимо очистить. Было замечено, что при создании набора появились невозможные примеры: некоторые записи содержат, например, нулевую ширину экрана. Выделив атрибуты для проверки с помощью исключения логических параметров и атрибутов, которые могут быть нулевыми, очищены данные. В результате очистки был удален 181 пример из 2000.

Решено заменить столбцы высоты и ширины экрана на разрешение и соотношение сторон – из них можно получить оба признака, но они являются более содержательными и простыми для интерпретации.

Далее, проведена подготовка данных к кластеризации. Первоначально, данные были стандартизированы с помощью инструмента StandardScaler из библиотеки sklearn. Это необходимо для уравновешивания атрибутов – некоторые могут принимать значения из большего диапазона и будут сильнее влиять при вычислении расстояний.

С помощью метода локтя был подобран оптимальный параметр k = 4 (рис. 1). Применив алгоритм кластеризации, было получено следующее распределение кластеров: 913, 436, 308 и 162 примера на кластер соответственно (рис. 2).

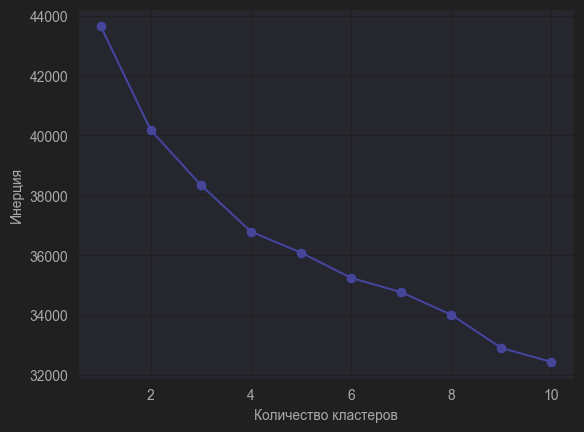


Рисунок . Метод локтя определения параметра k

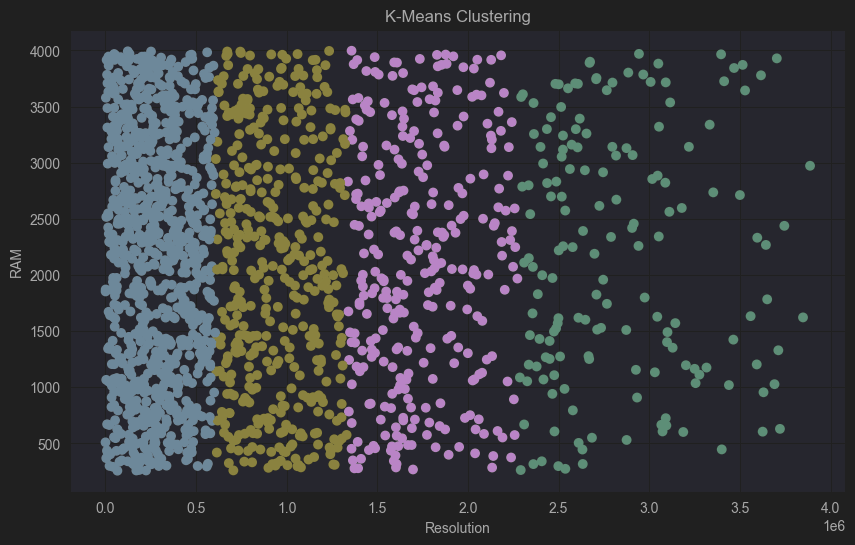


Рисунок . Рассеивание кластеров

Далее, построены столбчатые диаграммы распределения признаков по кластерам. В основном, характеристики распределены равномерно, но снова выделяются разрешение экрана, разрешение камеры и объем оперативной памяти (рис. 3, 4, 5)

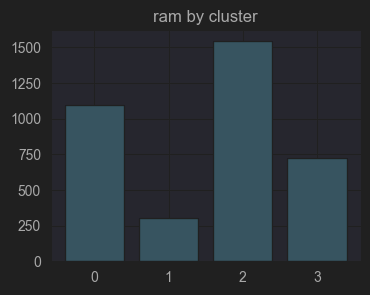


Рисунок . Распределение ОЗУ



Рисунок . Распределение Разрешения экрана

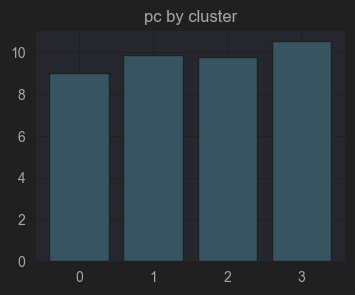


Рисунок . Распределение Разрешения задней камеры

В первую категорию попали дешевые телефоны с низкими показателями, в 0 и 3 – средние бюджетные модели, а второй кластер показывает лучшие характеристики.

**Выполнение работы**

В ходе работы были обработаны пропуски данных, созданы новые атрибуты на основе существующих. Изучен и применен метод кластеризации KMeans, а также методы предобработки данных. В результате получено распределение устройств по четырем категориям.