# МИНОБРНАУКИ РОССИИ

**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**

# «ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)

**Кафедра ИС**

# ОТЧЕТ

**по лабораторной работе №2**

# по дисциплине «Машинное обучение» Тема: Кластеризация

|  |  |
| --- | --- |
| Студент гр. 2373 | Гермаш А. В. |
| Преподаватель | Татчина Я.А. |

Санкт-Петербург 2024

Цель работы: познакомиться с методом кластеризации K-mean при помощи пакета sklearn.

# Краткое описание набора данных

Для работы был взят отличный от предыдущего набор данных. Датасет является характеристикой использования смартфонов пользователями. Датасет создан искусственно, представлены следующие атрибуты:

* + User ID - идентификатор пользователя
  + Operating System - Android или iOS
  + App Usage Time (min/day) - использование приложений (мин в день)
  + Screen On Time (hours/day) - время включенного экрана смартфона (часов в день)
  + Battery Drain (mAh/day) - разряд батареи в сутки
  + Number of Apps Installed - количество установленных приложений
  + Data Usage (MB/day) - использование сети в мб в день
  + Age - возраст
  + Gender - Пол
  + User Behavior Class - классификация поведения пользователя

# Добавление нового атрибута

Мною был добавлен атрибут Screen Time B. Dr. – отношение разряда батареи смартфона к его времени работы экрана. Атрибут помогает оценить зависимость разряда батареи от активного использования смартфона. Измеряется в [мА].

# Первичная обработка данных

* + Дубликатов не оказалось
  + Пропущенных значений не оказалось
  + Методом межквартального размаха найдено и удалено 29 выбросов в атрибуте Screen Time B. Dr.
  + Удалены атрибуты Gender, User Behavior Class, User ID, Operating System
  + App Usage Time приведено к часам в день
  + Переименованы для удобства некоторые атрибуты

# Графики зависимостей

# Изображение выглядит как текст, диаграмма, линия, снимок экрана Автоматически созданное описание

Рисунок 1 – Матрица графиков рассеивания

На графиках четко видно 5 ярко выраженных групп.

# Кластеризация при помощи KMeans

Проверим наше предположение о 5 кластерах при помощи “метода локтя”. Построим график зависимости инерции KMeans от количества кластеров.

В качестве атрибутов, при помощи которых будем разбивать на группы возьмем Data Usage и Screen Time B. Dr.

Изображение выглядит как текст, линия, снимок экрана, диаграмма

Автоматически созданное описание  
Рисунок 2 – График инерции

По графику можем увидеть, что инерция перестает существенно уменьшаться после N = 3 кластеров.

Посмотрим, как поведёт себя KMeans для N = 3 и N = 5 кластеров.

Изображение выглядит как снимок экрана, Красочность

Автоматически созданное описание

Рисунок 3 – KMeans для N = 3 кластеров

Изображение выглядит как снимок экрана, Красочность

Автоматически созданное описание

Рисунок 4 – KMeans для N=5 графиков

# Выводы

В ходе этой работы была проведена первичная обработка набора данных, был изучен метод кластеризации KMeans, с помощью которого мы попробовали разделить набор данных на кластеры. Для определения оптимального количества кластеров был использован “метод локтя”. Можно сказать, что оптимальным выбором для взятых атрибутов является 3-5 кластеров.