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

**Кафедра информационных систем**

**ОТЧЕТ**

**по практической работе №2**

**по дисциплине «Машинное обучение»**

**Тема: «Метод кластеризации K-mean»**

Студент гр. 2373 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Красножён Е. Д.

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Татчина Я. А.

Санкт-Петербург

2024

**Цель работы:** познакомиться с методом кластеризации K-mean с помощью пакета sklearn.

**Краткое описание датасета**

Датасет о 1215 покемонах и их характеристиках.  
В датасете представлены следующие атрибуты:

* images: Путь к изображению покемона (строковой тип. В данной работе не используется)
* index: Номер покемона в Покедексе (числовой тип. Был удалён из датасета, так как используется стандартная индексация).
* Name: Название Покемона (строковой тип).
* Type 1: Первый из типов Покемона (строковой тип).
* Type 2: Второй из типов Покемона, если присутствует (строковой тип).
* Total: Сумма всех характеристик Покемона (числовой тип).
* HP: Очки здоровья Покемона (числовой тип).
* Attack: Показатель базовой атаки Покемона для обычных атак (числовой тип).
* Defense: Базовый показатель защиты Покемона от обычных атак (числовой тип).
* SP. Atk.: Показатель базовой атаки Покемона для особых атак (числовой тип).
* SP. Def.: Базовый показатель защиты Покемона от особых атак (числовой тип).
* Speed: Определяет, какой Покемон атакует первым в раунде (числовой тип).

**Добавление нового атрибута**

Был добавлен новый атрибут "Avg", который яляется средним значением всех характеристик Покемона.

**Обработка данных**

Дубликатов в датасете не было найдено.

Пропуски были найдены только в столбце Type 2 и были изменены на ‘None’

Выбросы были обнаружены с помощью графиков seaborn.boxplot() и строчки с выбросом были удалены из датасета

**Построить графики зависимости одной переменной от другой**

С помощью seaborn.regplot() были построены графики зависимостей одной переменной от другой. По графикам было выявлено, что графики зависимости с переменной Total и графики зависимости с переменной Avg идентичны друг другу при одинаковом втором атрибуте.

Ярковыраженных кластеров замечено не было и для кластеризации были выбраны следующие пары атрибутов, где кластеры «примерно» заметны: 'Total' и 'HP'; 'Defense' и 'Total'; 'HP' и 'SP. Atk.'; 'Attack' и 'Speed'; 'Defense' и 'SP. Def'; 'SP. Atk.' и 'SP. Def'.

**Применить метод KMeans**

Для кластеризации данных был применён метод KMeans.

Алгоритм KMeans группирует данные, пытаясь разделить выборки на n групп с одинаковой дисперсией, минимизируя критерий, известный как инерция или сумма квадратов внутри кластера. Этот алгоритм требует указания количества кластеров. Он хорошо масштабируется для большого количества данных и используется в самых разных областях применения.

Для определения количества кластеров воспользуемся методом локтя. Для этого посчитаем значения инерции алгоритма при различных количествах кластеров и выберем значение инерции, после которого инерция перестаёт существенно уменьшаться.

**Вывод**

В ходе данной работы был добавлен новый атрибут

Была проведена первичная обработка: обработаны пропущенные значения, удалены выбросы, дубликатов обнаружено не было.

Были построены графики зависимостей атрибутов от различных жанров, а затем кластеризованы некоторые группы данных пар атрибутов с помощью метода KMeans.

Все графики представлены в .ipynb файле.