**Отчёт по тестовому заданию**

**Выполнила София Бердус**

**ML**

Имеется набор данных (см. файл **DataSet.ods**). Необходимо:

1. Убрать лишние данные. К ним относятся все строки с **CVE-ID**=«NVD-CWE-Other» и **CVSS-V3**=«None»;
2. Разбить записи на кластеры (по ключевым словам в **DESCRIPTION**), выбрав оптимальное количество. Проанализировать кластеры по **CWE-ID** и **SEVERITY**;
3. Вывести ключевые слова-паттерны для каждого кластера;

**Описание**

В процессе работы появилась ошибка, связанная с отсутствием библиотеки odfpy. Поскольку эта библиотека уже была установлена в моей системе, но не распознавалась, создание нового окружения позволило избежать проблем и обеспечить стабильную работу.

Таким образом, первоначально я провела работу в терминале, перейдя в рабочую директорию и выполнив следующий блок команд:

conda create -n new\_env python=3.9

conda activate new\_env

conda install -c conda-forge odfpy

Затем вручную установила все необходимые библиотеки в новое окружение.

Задание 1.

Код для очистке данных приведен в файле clean.py. После его запуска очищенные данные сохраняются в файл cleaned\_data.csv (рис. 1).

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 1 – Очищенные данные

Задание 2, 3.

Сначала произведена загрузка и предварительная обработка данных. Затем выполнена кластеризация с помощью алгоритма KMeans. Оптимальное количество кластеров, равное 5, определено на основе классического “метода локтя”, график представлен на рисунке 2.

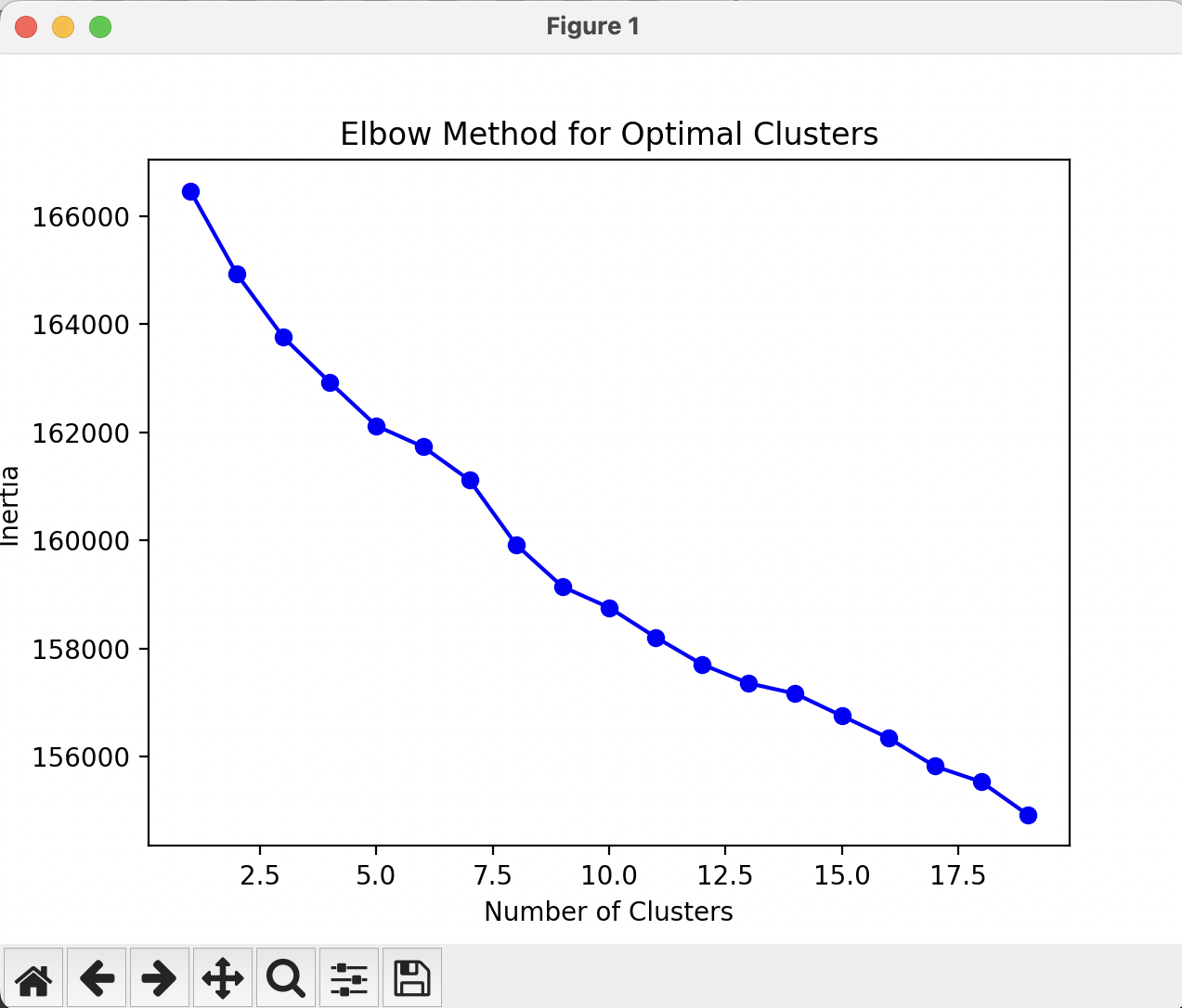


Рисунок 2 – График

На графике инерция существенно уменьшается на интервале от ~2 до 5 кластеров. После точки с 5 кластерами степень уменьшения инерции становится менее выраженной, следовательно после 5 кластеров добавление новых кластеров не приносит значительного улучшения.

По итогам выполнения кластеризации выводятся ключевые слова, присущие каждому кластеру (рис. 3).

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

Рисунок 3 – Результат выполнения программы

Итоговые данные сохраняются в файл clustered\_data.csv (рис. 4).

Изображение выглядит как текст, снимок экрана, чек, число

Автоматически созданное описание

Рисунок 4 – Итоговые данные