**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 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Чишко В. О.

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

Санкт-Петербург

2024

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

**Краткое описание датасета**

Данные о продажах видеоигр в разных регионах. Представлена информация продаж Америки, Европы, Японии, остальных и общемировая. Датасет взят с сайта www.kaggle.com, вот ссылка на сам датасет: https://www.kaggle.com/datasets/gregorut/videogamesales/data Данные реальны.

Представлены как числовые, так и строковый тип данных:

* Rank - топ по продажам – (числовой)
* Name - наименование видеоигры – (строковый)
* Platform - платформа, на которой игра выпускалась – (строковый)
* Year - год выхода – (числовой)
* Genre - жанр – (строковый)
* Publisher - издатель – (строковый)
* NA\_Sales - продажи в Америке (в миллионах) – (числовой)
* EU\_Sales - продажи в Европе (в миллионах) – (числовой)
* JP\_Sales - продажи в Японии (в миллионах) – (числовой)
* Other\_Sales - продажи в остальных частях всета (в миллионах) – (числовой)
* Global\_Sales - Общемировые продажи – (числовой)

**Добавление нового атрибута**

Добавим новый атрибут "Возраст игры", для которого бралось значение даты выхода игры и оно вычиталось из даты создания датасета, т.е. 2016 года.

**"Причесать" датасет**

В датасете были найдены пропуски и выбросы. Часть пропусков была заполнена, а часть, которую не возможно заполнить, была удалена. Выбросы же были просто удалены. По итогам “причесывания” было удалено около 6000 записей из 16000 изначальных.

**Построить графики зависимости одной переменной от другой**

Сначала, были рассмотрены разные распределения, и был выбран атрибут «Жанр», как критерий для выявления зависимостей. Была простроена корреляционная матрица и по этим данным построили графики зависимостей одних зависимостей от других. В основном брались корреляции между ценами в разных регионах и их зависимость от жанров. Все графики представлены в работе. По графикам можно понять, что распродажи по жанрам разбросаны преимущественно хаотично, но во всех идёт перевес в сторону жанров “Action” и “Shooter”.

**Применить метод KMeans**

Для дальнейшей кластеризации нормализуем данные. Для этого воспользуемся функцией StandardScaler библиотеки sklearn.preprocessing.

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.

Алгоритм был применен только на самые наглядные пары атрибутов и по ним были построены графики. Сами графики представлены в работе.

**Вывод**

В ходе данной работы был добавлен новый атрибут в датасет, проведена первичная обработка: удалены выбросы по атрибутам продаж, обработаны и удалены пропущенные значения, дубликатов обнаружено не было, нормализированы данные в датасете.

Были построены графики зависимостей атрибутов от различных жанров, а также изучен метод кластеризации KMeans, с помощью которого мы выделили кластеры для данного датасета.