**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ «ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**Кафедра ИС**

**ОТЧЕТ**

**по лабораторной работе №2**

**по дисциплине «Машинное обучение»**

**Тема: Метод кластеризации KMeans**

Студентка гр. 2373 Никитина М.Ю.

Преподаватель Татчина Я.А.

Санкт-Петербург

2024

Задание:

1. К текущему датасету необходимо добавить новый аттрибут, например, если есть дата рождения, то посчитать возраст или определить популярность фильма по количеству просмотров и т.п.

2. "Причесать" датасет: удалить выбросы и дубли, обработать пропущенные значения, найти кривые данные и т.п.

3. Построить графики зависимости одной переменной от другой (plt.scatter), описать какие кластеры (группы) вы видите. Построить несколько разных графиков, найти ярковыраженные группы.

4. Изучить перечисленные ноутбуки, по примеру, попробовать применить метод KMeans к вашему датасету.

Для выполнения данной лабораторной работы был взята база данных Abba Music Characteristics (Spotify), скаченная с сайта <https://www.kaggle.com/datasets/thedevastator/abba-music-characteristics-spotify?resource=download>.

В данном датасете представлена информация насчет самых популярных песен (хитов) шведской группы ABBA на известной платформе Spotify. Все данные реальны.

Датасет включает в се6я следующие атрибуты:

album - название альбома (String)

track\_number - трек-номер песни в альбоме (Integer)

uri - уникальный идентификатор песни (String)

acousticness - мера акустики песни (Float)

danceability - насколько хорошо песня хороша для танцев (Float)

energy - энергичность песни (Float)

instrumentalness - инструментальность песни (Float)

liveness - оригинальность песни (Float)

loudness - громкость песни (Float)

speechiness - мера кол-ва текста в песни (Float)

tempo - темп песни (Float)

valence - позитивность песни (Float)

popularity - популярность песни (Integer)

artist - исполнитель песни (String)

По результатам 1-й лабораторной работы дубликатов в датасете нет, как и пропущенных значений, и выбросов.

Для определения кластеризации были использованы следующие зависимости:

1. танцевальность песен от их энергичности
2. акустика песен от инструментальности
3. популярность песен от оригинальности
4. позитивность песен от темпа
5. громкость песен от кол-ва текста
6. мелодичность песен от инструментальности

Полученные графики зависимости:

* Изображение выглядит как снимок экрана, диаграмма

  Автоматически созданное описаниетанцевальность песен от их энергичности:
* Изображение выглядит как текст, снимок экрана, линия, График

  Автоматически созданное описаниеакустика песен от инструментальности:
* Изображение выглядит как снимок экрана, текст, диаграмма, линия

  Автоматически созданное описаниепопулярность песен от оригинальности:
* позитивность песен от темпа:

Изображение выглядит как снимок экрана, диаграмма, текст, линия

Автоматически созданное описание

* Изображение выглядит как текст, снимок экрана, линия, График

  Автоматически созданное описаниегромкость песен от кол-ва текста
* мелодичность песен от инструментальности:

Изображение выглядит как текст, диаграмма, линия, График

Автоматически созданное описание

Кластеризация:

Для графиков кластеризации был выбран атрибут n\_clusters со значением 2, так как посмотрев на графики, каждый из них можно условно разделить на 2 части (т.к. данные не слишком разбросаны). Например, на графике ниже можно выделить группу песен с энергичностью менее 1.0 и группу с энергичностью более 1.0.

* Изображение выглядит как текст, снимок экрана, Мультимедийное программное обеспечение, программное обеспечение

  Автоматически созданное описаниеэнергичность и танцевальность:
* акустичность и инструментальность:

Изображение выглядит как текст, снимок экрана, диаграмма, Графическое программное обеспечение

Автоматически созданное описание

* популярность и оригинальность:

Изображение выглядит как текст, снимок экрана, диаграмма, Графика

Автоматически созданное описание

* Изображение выглядит как текст, снимок экрана, Графика

  Автоматически созданное описаниепозитивность и темп:
* громкость и кол-во текста:

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

* мелодичность и инструментальность:

Изображение выглядит как текст, снимок экрана, диаграмма

Автоматически созданное описание

Вывод:

В данной лабораторной работе мы удалили все выбросы, дубликаты, а также пропущенные значения в каждом атрибуте. Также были построены графики зависимостей одной переменной от другой. Был изучен метод кластеризации KMeans, с помощью которого на графиках мы смогли выделить основные кластеры и описать их.