**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра информационных систем**

отчет

**по лабораторной работе №2**

**по дисциплине «Машинное обучение»**

Тема: Метод Кластеризации K-Mean

|  |  |  |
| --- | --- | --- |
| Студент гр. 2372 |  | Юрин А.А. |
| Преподаватель |  | Татчина Я.А. |

Санкт-Петербург

2024

**Цель работы.**

Знакомство с методом кластеризации K-mean с помощью пакета sklearn.

**Постановка задачи.**

1. Добавить к набору данных новый атрибут, вычислив его из имеющихся;
2. Очистить данные, удалив выбросы и дубли; обработать пропущенные значения; найти кривые данные и т.п.;
3. Построить графики зависимости одной переменной от другой, описать какие кластеры (группы) можно заметить. Построить несколько разных графиков, найти ярко выраженные группы.
4. Применить метод кластеризации K-mean к набору данных; проанализировать полученные результаты.

**Выполнение работы.**

Для данной лабораторной работы был выбран новый набор реальных данных, содержащий в себе 4600 записей о треках и их прослушиваниях за промежуток с 1987 по 2024 года.

Практически все поля в файле были записаны как вещественные, что не несёт смысла для полей “Release Date”, “Spotify Streams”, “Spotify Playlist Count” и так далее, тип данных этих столбцов был заменён на целочисленный.

Были удалены лишние атрибуты. Так как они не предоставляли нужной информации для выполнения работы.

Был обновлен атрибут “Release Date”. Вместо полной даты выпуска, был выделен только год.

Были созданы новые атрибуты: общее количество прослушиваний, плейлистов и радиовещаний. Они обобщили всю информацию по каждому связанному с ними старым атрибутом. Например, новый атрибут “Total streams” был получен из таких атрибутов как "Spotify Streams", "Pandora Streams", "Soundcloud Streams", "Spotify Playlist Reach", "YouTube Playlist Reach" и "Deezer Playlist Reach". Также при объединении были исключены пустые поля.

После предобработки данных был проведён анализ атрибутов и их визуализация. Так, общее распределение Explict (Контент для взрослых) и Clean (Контент для любого возраста) треков за каждый год. можно увидеть на рисунке 1.

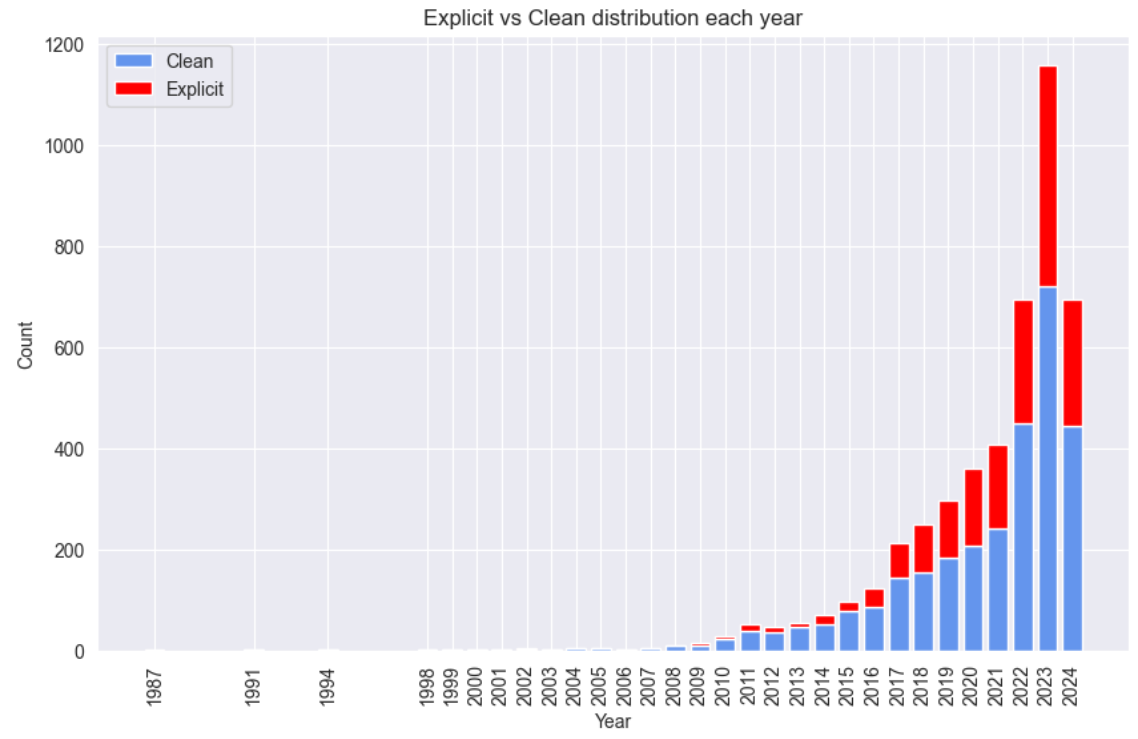


Рисунок 1. Explict и Clean распределение треков за каждый год.

Самые прослушиваемые треки в данном наборе (трек: количество прослушиваний):

1. As It Was: 10906837613,
2. Easy On Me: 8376997154,
3. Heat Waves: 8142020607,
4. STAY (with Justin Bieber): 7807861279,
5. Blinding Lights: 7234759017.

Самые популярные треки по количеству в плейлистах (трек: количество плейлистов с ним):

1. Blinding Lights: 591999,
2. Heat Waves: 487002,
3. STAY (with Justin Bieber): 452539,
4. Sweater Weather: 415962,
5. Smells Like Teen Spirit: 410694.

Самые популярные треки по количеству радиовещаний (трек: количество радиовещаний):

1. Those Oldies but Goodies: 3780513,
2. All I Want for Christmas Is You: 3688803,
3. Sex Music: 3482267,
4. Sexy And I Know It: 3400823,
5. Let It Go - From "Frozen"/Soundtrack Version: 3308111.

Были обнаружены две сильные зависимости. “Total streams” коррелирует с “Total playlists” на 67% и на 54% с “Total broadcasts”, что ожидаемо. Чем сильнее растёт количество прослушиваний, тем больше количество плейлистов и радиовещаний с этими треками.

Для кластеризации методом «локтя» было выбрано число k = 4. После применения метода было выявлено вполне объяснимое распределение по группам, а именно:

1. Самый большой кластер - треки в среднем 2022 года с самым низким рейтингом и низкой популярностью;
2. Наименьший кластер - треки в среднем 2017 года с высоким рейтингом и высокой популярностью, также с самым низким количеством Explicit треков;
3. Небольшой кластер – треки в среднем 2017 года со средним рейтингом и средней популярностью;
4. Большой кластер - треки в среднем 2019 года с небольшим рейтингом и с самой высокой популярностью по количеству прослушиваний.

**Выводы.**

В ходе работы были изучены и обработаны данные музыкальной индустрии за 1987-2024 года. Были выявлены зависимости и проведена кластеризация. В результате были получены навыки кластеризации с помощью метода K-means.