

# MUSIC RECOMENDATION

Nguyễn Thị Minh Ly - 23020399

Đặng Minh Nguyệt - 23020407



# NỘI DUNG TRÌNH BÀY

01

Đặt vấn đề

02

Dữ liệu

03

Phương pháp

04

Kết quả

# ĐẶT VẤN ĐỀ

01

## Nhu cầu khám phá

- Người dùng bị quá tải lựa chọn và khó tìm nghệ sĩ phù hợp sở thích.
- Cần hệ gợi ý để tiết kiệm thời gian và nâng cao trải nghiệm.

02

## Hạn chế của gợi ý hiện có

- Các nền tảng lớn thường phức tạp, không rõ ràng cách gợi ý.
- Người dùng thiếu công cụ minh bạch, đơn giản để tìm nghệ sĩ tương tự theo cách họ muốn.

03

## Tiềm năng hệ thống

- Hệ thống gợi ý nghệ sĩ có thể tích hợp vào ứng dụng âm nhạc, công cụ tìm kiếm hoặc nền tảng giải trí.
- Dễ mở rộng sang gợi ý bài hát, playlist, hoặc phân tích xu hướng âm nhạc cá nhân.



# DỮ LIỆU

## MÔ TẢ

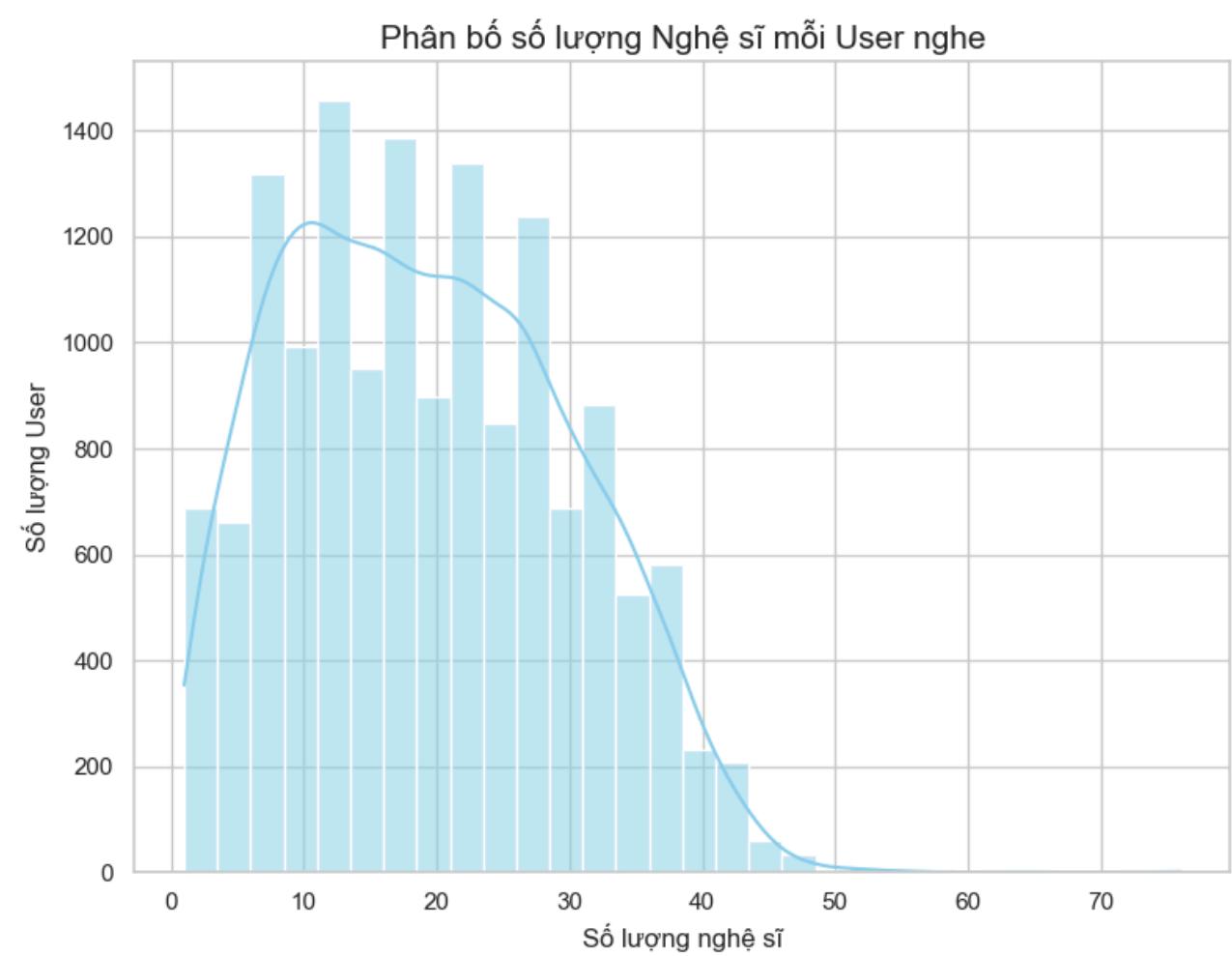
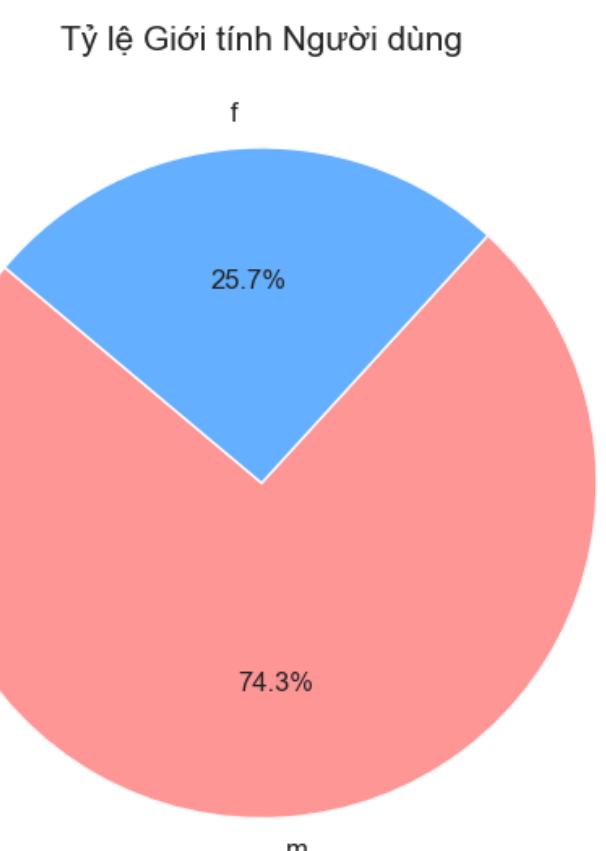
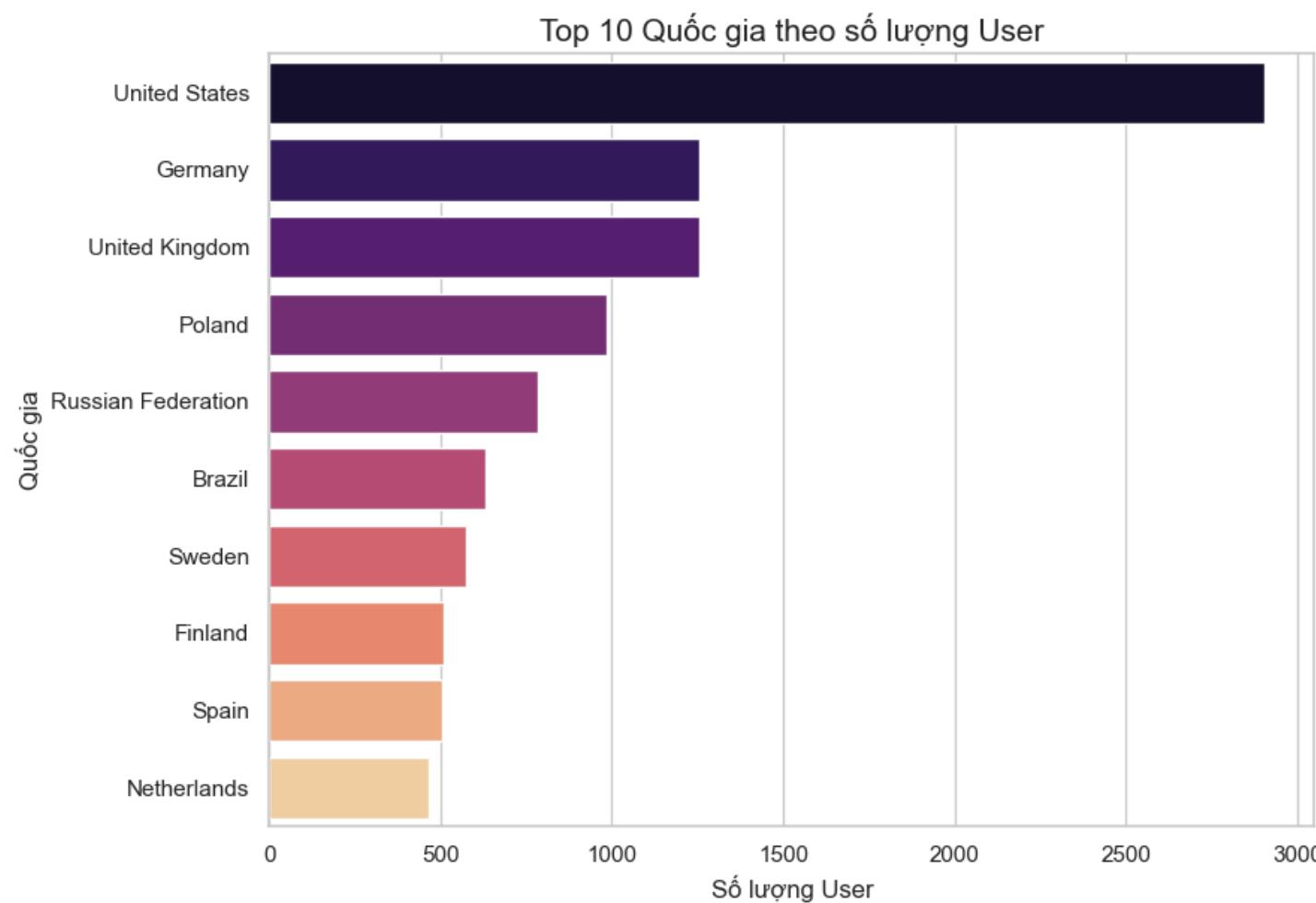
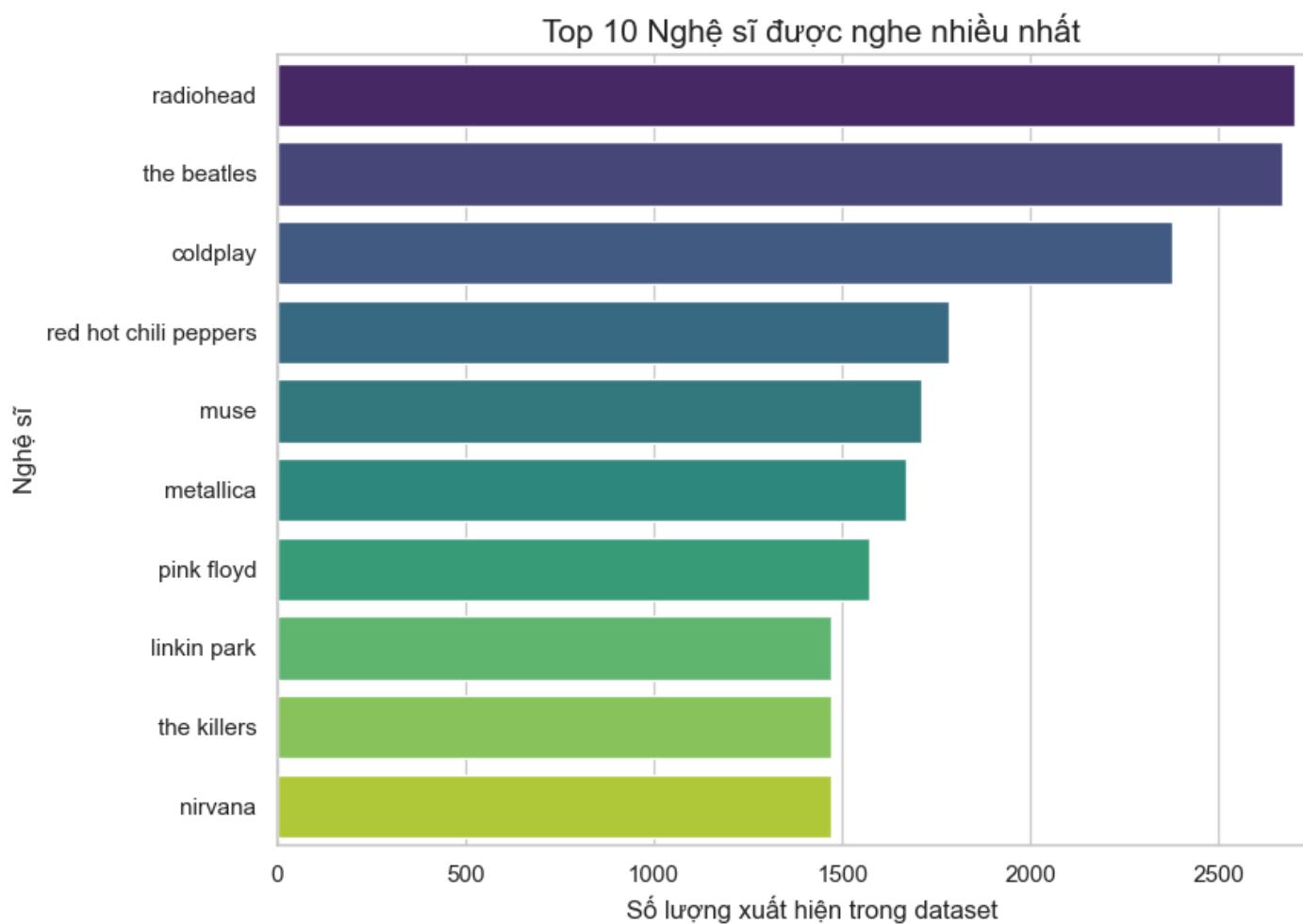
- **Miền dữ liệu:** Hành vi người dùng trong lĩnh vực âm nhạc.
- **Kích thước:** ~300.000 dòng
- **Thuộc tính chính:**
  1. *user\_id*: Mã định danh người dùng
  2. *sex*: Giới tính
  3. *music\_artist*: Nghệ sĩ mà người dùng nghe
  4. *country*: Quốc gia của người dùng



# EDA



1. Nghệ sĩ
2. Quốc gia
3. Giới tính
4. Hành vi





# PHƯƠNG PHÁP

---

# FP-growth

**Khái niệm:** Là thuật toán Khai phá luật kết hợp (Association Rule Mining), tìm ra các mối quan hệ ẩn giữa các đối tượng trong tập dữ liệu lớn.

## Ý tưởng

- **Cấu trúc FP-Tree:** Thay vì quét dữ liệu nhiều lần (như Apriori), thuật toán nén dữ liệu vào một cây tần suất (FP-Tree).
- **Chiến lược "Chia để trị":** Phân rã bài toán lớn thành các đoạn mẫu phổ biến nhỏ hơn, giúp tăng tốc độ xử lý trên tập dữ liệu lớn.

# FP-growth

## Cơ chế gợi ý

1. **Matching (Đối sánh):** Khi người dùng nghe một nghệ sĩ (Input), hệ thống quét tập luật đã sinh ra để tìm các dòng có chứa nghệ sĩ đó trong cột Antecedents (Về trái/Nguyên nhân).
2. **Ranking (Xếp hạng):** Các luật tìm được sẽ được sắp xếp dựa trên chỉ số Lift (Độ nâng) từ cao xuống thấp.
  - Lift > 1: Biểu thị mối liên hệ tương quan dương mạnh mẽ.
3. **Recommendation (Đề xuất):** Trả về danh sách nghệ sĩ nằm ở cột Consequents (Về phải/Kết quả) tương ứng với các luật có Lift cao nhất.

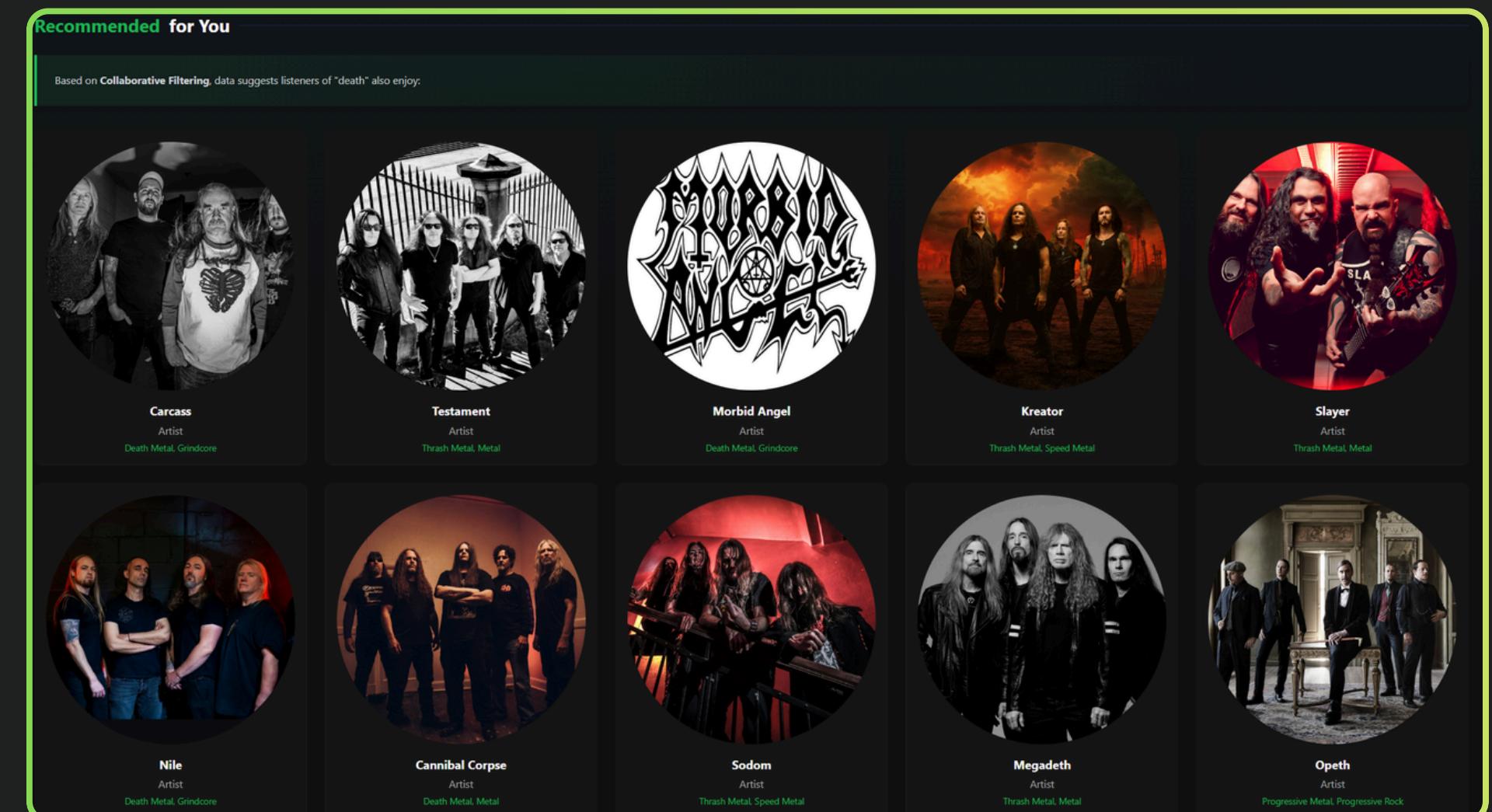
# Item-based Collaborative Filtering

- **Ý tưởng:** Tìm những artist giống nhau, gợi ý các artist tương tự artist user đã thích
- Chỉ giữ lại các nghệ sĩ có tối thiểu 50 người nghe để loại bỏ nhiễu (long-tail noise)
- Tạo **ma trận thưa**:
  - Cấu trúc: Hàng (Rows) = Nghệ sĩ, Cột (Cols) = Người dùng
  - Giá trị: 1 (biểu thị người dùng có nghe nghệ sĩ này)
- **TF-IDF Transformation**: Giảm trọng số của các nghệ sĩ quá phổ biến (xuất hiện ở hầu hết playlist) để làm nổi bật sở thích đặc trưng của người dùng.

# Item-based Collaborative Filtering

## Cơ chế gợi ý

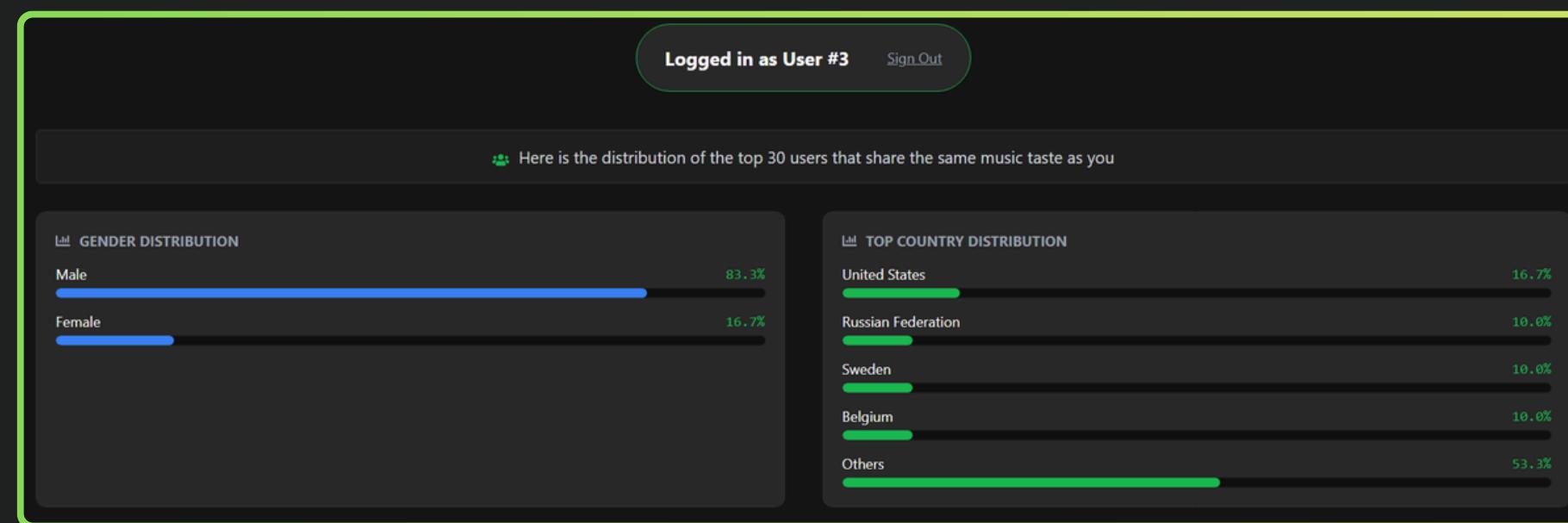
1. Nhận vào tên một nghệ sĩ
2. Truy xuất vector đặc trưng của nghệ sĩ đó từ ma trận
3. Sử dụng knn (dựa trên cosine similarity) để tìm k+1 nghệ sĩ gần nhất trong không gian vector
4. Trả về danh sách Top K nghệ sĩ



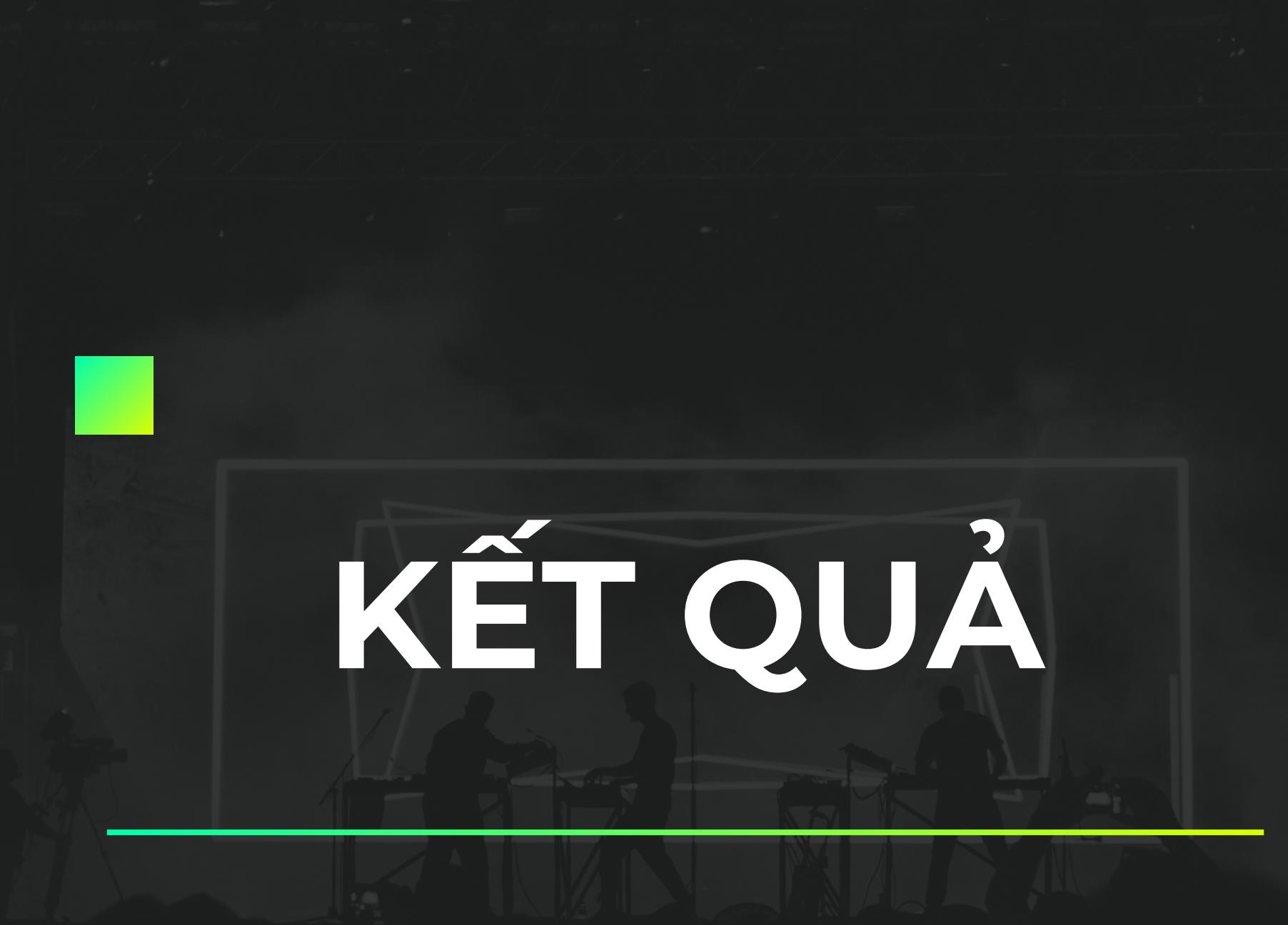
Tính năng thêm

# TÌM NHỮNG NGƯỜI DÙNG CÓ CÙNG GU ÂM NHẠC

- Tương đồng với User-based CF, ta tính toán độ tương đồng giữa các user (dựa trên danh sách các artists mà mỗi user nghe)
- Từ đó lấy ra demographic của các user gần đó: country + sex

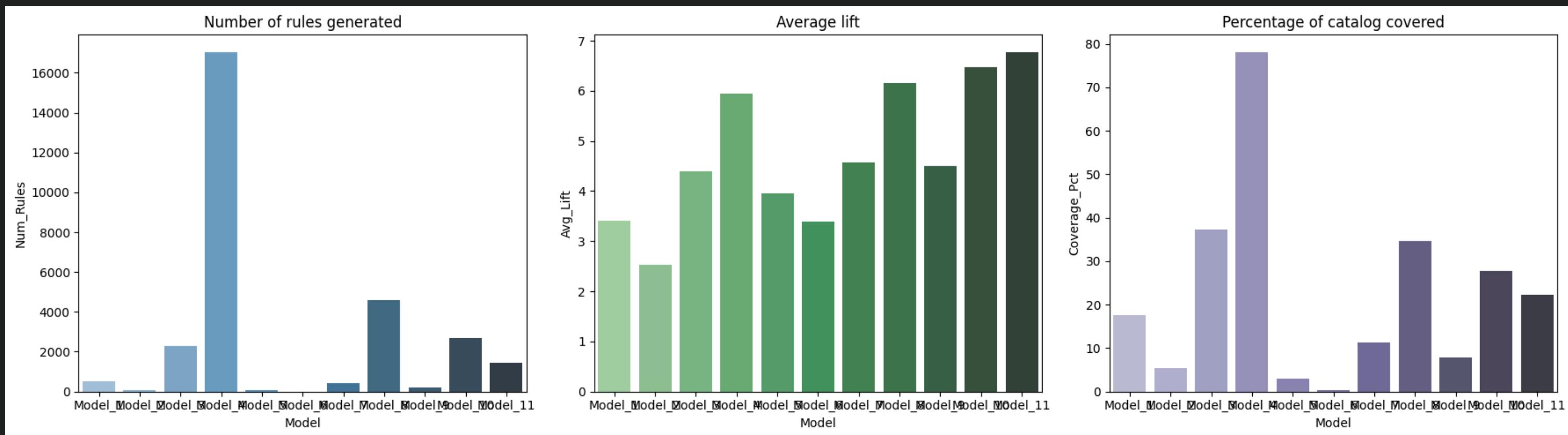


# KẾT QUẢ



# SO SÁNH FP-GROWTH TRÊN 11 BỘ THAM SỐ

*min\_support (0.003-0.02) & min\_confidence (0.3-0.6)*



## Nhận xét:

- Model 4 (0.003-0.3) sinh được lượng luật & độ bao phủ lớn nhất tuy nhiên lift lại thấp hơn → overfitting
- Model 11 và 10 có lift cao nhưng độ bao phủ và lượng luật sinh ra thấp
- Model 8 (MS 0.003-MC 0.5) có sự cân bằng giữa ba yếu tố ổn nhất



**DEMO**

---

WebLink



**CẢM ƠN CÔ VÀ CÁC BẠN  
ĐÃ LẮNG NGHE**

---