교과명 : 머신러닝

담당 교수 : 윤직혁 교수님

# 음악의 음향적 특징을 활용한 인기도 예측

## - Spotify 데이터 분석 -

학과 : AI융합학부

학번 : 20241992

이름 : 최희우

## 문제 정의

이번 프로젝트의 최종 목표는 음악의 음향적 특성과 메타데이터를 분석하여 노래의 인기도를 사전에 예측할 수 있는 모델을 구축하는 것이다.

즉, 과거에 존재하는 수많은 노래들의 특성과 인기도 데이터를 학습하여 새로운 노래가 발매되었을 때 그 노래가 대중적으로 성공할 가능성을 수치적으로 예측하고자 한다.

이를 통해 음악 제작자나 프로듀서는 청취자들에게 인기가 높을 것으로 예상되는 음악적 특성을 사전에 파악하고 음악 기획이나 마케팅 전략을 보다 데이터 기반으로 수립할 수 있을 것이다.

연구 가설은 다음 세가지를 바탕으로 진행하려 한다.

첫째, 노래의 핵심 음향적 특징인 'danceability', 'energy', 'valence', 'loudness', 'tempo' 등은 노래의 인기도에 유의미한 영향을 미칠 것이다. 곡의 리듬감, 강도, 밝기 등의 요인은 청취자들의 감정적 반응과 반복 재생 여부에 직접적인 영향을 줄 가능성이 높기 때문이다.

둘째, 노래의 장르와 발매 시기는 음향적 특징과 결합하여 인기도 예측의 정확도를 높이는 보조적 요인으로 작용할 것이다. 이는 음악의 시대적 트렌드와 장르별 특징이 대중적 선호도에 영향을 줄 수 있기 때문이다.

셋째, 인기가 높은 곡들의 음향적 특징 분포는 장르에 따라 다를 것이며 각 장르별로 대표적인 음향적 조합이 존재할 것이다. 예를 들어 팝 장르는 일반적으로 valence와 danceability가 높고 재즈는 acousticness와 instrumentalness가 높은 경향을 보일 것이다.

위의 목표와 연구 가설을 바탕으로 살펴본 구체적인 문제들은 아래와 같다.

곡의 인기도에 가장 큰 영향을 미치는 음향적 특징은 무엇인지, 장르별 인기 곡의 음향적 특징 분포는 어떻게 다르며 음향적 특징과 장르, 발매 연도 정보를 함께 사용할 경우 인기도 예측 정확도는 얼마나 향상되는지 회귀 접근법과 분류 기반 접근 법 중 어느 방식이 더 높은 성능을 보이는 지이다. 이들을 알아보려고 한다.

머신러닝적으로 task는 크게 회귀와 분류를 사용하려 한다.

회귀 접근은 노래의 인기도를 0~100사이의 연속적인 값으로 예측한다. 이를 위해 선형 회귀, 랜덤 포레스트 회귀, XGBoost회귀 등의 모델을 적용할 수 있을 것이다.

분류접근은 인기도를 일정 기준을 기준으로 인기곡과 비인기곡으로 구분하여 예측하는 방식이다. 이 경우 로지스틱 회귀, 랜덤 포레스트, 그래디언트 부스팅 등의 모델을 활용할 수 있을 것이다.

따라서 이번 프로젝트에서는 회귀, 분류 두 가지 접근법을 병행하여 인기도 예측의 정량적, 정성적 분석을 모두 수행하고 각 모델의 예측 성능을 비교함으로써 종합적인 결론을 도출하고자 한다.

## 데이터 정의

이 모델을 구축할 때 사용할 데이터는 kaggle에서 제공한 데이터로 "30000 Spotify Songs"(출처 : https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs?select=spotify_songs.csv) 데이터셋을 활용하려고 한다. 이 데이터는 Spotify의 오디오 특성 API를 통해 수집된 대규모 데이터로 총 약 30000개의 노래에 대한 정보가 포함되어 있다.

각 곡은 장르, 발매일 등의 메타데이터 뿐만 아니라 음향적 속성을 수치화한 변수도 함께 포함하고 있어 내적 특성, 외적 맥락이 인기도에 미치는 영향을 동시에 분석하기에 적합하다 판단하였다.

이 데이터의 변수들은 아래와 같다.

| 변수명 | 자료형 | 설명 |
|---|---|---|
| track_id | character | 노래의 고유 ID |
| track_name | character | 노래 제목 |
| track_artist | character | 노래 아티스트 이름 |
| track_popularity | double | 노래의 인기도를 0~100 사이의 값으로 나타냄. 값이 높을수록 대중적 인기가 높음을 의미함. |
| track_album_id | character | 앨범의 고유 ID |
| track_album_name | character | 노래가 수록된 앨범 이름 |
| track_album_release_date | character | 앨범의 발매일 |
| playlist_name | character | 재생목록(플레이리스트) 이름 |
| playlist_id | character | 재생목록의 고유 ID |
| playlist_genre | character | 재생목록의 장르 |
| playlist_subgenre | character | 재생목록의 하위 장르 |
| danceability | double | 곡이 춤추기에 얼마나 적합한지를 나타내는 |

| | | |
|---|---|---|
| | | 지표. 템포, 리듬 안정성, 비트 강도 등 여러 음악적 요소를 종합하여 산출하며, 0.0은 가장 덜 춤추기 적합하고 1.0은 가장 적합함. |
| energy | double | 곡의 에너지(활동성, 강도)를 0.0~1.0 범위로 표현. 빠르고, 시끄럽고, 강한 곡일수록 값이 높음. |
| key | double | 곡의 전체적인 조성을 숫자로 표현. 예: 0=C, 1=C♯/D♭, 2=D 등. 조성이 감지되지 않으면 -1로 표시. |
| loudness | double | 곡 전체의 평균 음량(dB). 일반적으로 -60~0 dB 범위이며, 값이 높을수록 음량이 큰 곡임. |
| mode | double | 곡의 조성 모드. 1은 장조(major), 0은 단조(minor)를 의미함. |
| speechiness | double | 트랙 내 말소리의 비율을 측정. 값이 1에 가까울수록 말 위주의 녹음(예: 랩, 오디오북), 0.33 미만은 일반적인 음악을 의미함. |
| acousticness | double | 곡이 어쿠스틱(자연음) 기반일 확률을 0.0~0.1으로 표현, 1.0은 매우 어쿠스틱한 곡을 의미함. |
| instrumentalness | double | 곡에 보컬이 없는 정도를 나타냄. 값이 1.0에 가까울수록 악기 연주만으로 구성된 곡일 가능성이 높음. 0.5 이상이면 일반적으로 기악곡으로 간주됨. |
| liveness | double | 청중(라이브 공연)의 존재 가능성을 나타내는 지표. 값이 0.8 이상이면 실제 공연에서 녹음된 곡일 확률이 높음. |
| valence | double | 곡이 전달하는 감정적 긍정성. 0.0~1.0 사이의 값으로, 높을수록 밝고 행복한 분위기를, 낮을수록 슬프고 어두운 분위기를 나타냄. |
| tempo | double | 곡의 전체적인 템포(BPM, 분당 박자 수). 음악의 빠르기나 리듬의 속도를 나타냄. |
| duration_ms | double | 곡의 길이(재생 시간)를 밀리초 단위로 표현. |

이 중 track_popularity는 예측 대상변수로 사용하며 나머지 변수들은 입력 변수로 활용될 것이다. 모델 구축 시 필요없는 ID는 분석에서 제외하며 다른 변수들도 전처리 과정을 통해 정리될 것이다.