Eksploracja danych - propozycja projektu

1. Cel badania

Celem badania jest budowa optymalnego modelu klasyfikacji wieloklasowej za pomocą odpowiednich algorytmów uczenia maszynowego. Jego zadaniem będzie przyporządkowywanie gatunku muzycznego (blues, muzyka klasyczna, country, folk, jazz, newage, reggae, rock, metal) do utworów muzycznych na podstawie zmiennych określających charakter danego utworu oraz jego dane techniczne.

2. Opis zbioru badawczego

Zbiór danych został przeze mnie stworzony z wykorzystaniem API udostępnionego przez serwis Spotify, które pozwala na ekstrakcję cech muzycznych utworów obecnych na wybranej playliście lub profilu artysty. Do tego celu wykorzystałem paczkę `spotifyr`, która umożliwiła komunikację z API w R. Dla każdego gatunku wybrałem od jednej do trzech playlist stworzonych przez Spotify, tak aby łączna liczba obserwacji dla każdego gatunku mieściła się w przedziale 250-300. Pierwotny zbiór zawiera następujące zmienne:

- `playlist_id` - identyfikator playlisty, z której pochodzi utwór
- `playlist_name` - nazwa playlisty, z której pochodzi utwór
- `playlist_img` - adres zdjęcia playlisty, z której pochodzi utwór
- `playlist_owner_name` - nazwa twórcy playlisty, z której pochodzi utwór
- `playlist_owner_id` - id twórcy playlisty, z której pochodzi utwór
- `danceability` - zmienna ilościowa zawierająca się w przedziale 0-1, mówi o tym, jak bardzo utwór nadaje się do tańczenia na podstawie elementów takich jak tempo, stabilność rytmu i ogólnej regularności
- `energy` - zmienna ilościowa zawierająca się w przedziale 0-1, jest miarą intensywności utworu

- `key` - zmienna kategoryczna określająca tonację utworu, wartość odpowiada danej tonacji używając następującej notacji
- `loudness` - zmienna ilościowa zawierająca się w przedziale -60-0, określa średnią głośność utworu w decybelach
- `mode` - zmienna jakościowa dychotomiczna, określa czy utwór jest ze skali durowej (1) czy mollowej (0)
- `speechiness` - zmienna ilościowa przyjmująca wartości z przedziału 0-1, określa obecność słów recytowanych (nieśpiewanych). Przykładowo wartości bliskie 1 mogą odpowiadać audiobookowi, natomiast poniżej 0.33 najprawdopodobniej odpowiadają utworowi muzycznemu
- `acousticness` - zmienna ilościowa przyjmująca wartości z przedziału 0-1, mówi o tym jak bardzo akustyczny jest utwór. Wartości bliskie 1 wyrażają wysokie prawdopodobieństwo tego, że utwór jest akustyczny
- `instrumentalness` - zmienna ilościowa przyjmująca wartości z przedziału 0-1 określająca czy utwór zawiera wokal. Wartości bliskie 1 wskazują na to, że utwór jest instrumentalny
- `liveness` - zmienna ilościowa z przedziału 0-1, określa obecność żywej widowni w utworze, wartości bliskie jeden wskazują na to, że utwór został nagrany np. na koncercie
- `valence` - zmienna ilościowa z przedziału 0-1 określająca jak pozytywnie brzmi utwór. Przykładowo: wartości bliskie 1 - utwór wesoły, wartości bliskie 0 - utwór smutny, depresyjny
- `tempo` - zmienna ilościowa określająca tempo utworu
- `track_id` - identyfikator utworu
- `analysis_url` - link do analizy przeprowadzonej przez Spotify na utworze
- `time_signature` - oszacowane metrum muzyczne. Przyjmuje wartości z przedziału 3-7, gdzie np. wartość 7 odpowiada metrum 7/8
- `added_at` - data dodania utworu na platformę
- `is_local` - zmienna binarna określająca, czy utwór pochodzi z biblioteki lokalnej
- `track.disc_number` - zmienna określająca numer płyty, z której pochodzi utwór

- `track.duration_ms` - długość utworu w milisekundach
- `track_explicit` - zmienna binarna określająca, czy utwór zawiera treści wulgarne
- `track.name` - nazwa utworu
- `track.popularity` - zmienna ilościowa z przedziału 0-100 określająca popularność utworu, gdzie wartości bliskie 100 oznaczają, że utwór jest hitem
- `track.album.name` - nazwa albumu, z którego pochodzi utwór
- `track.album.release-date` - data wydania albumu, z którego pochodzi utwór
- `track.album.release_date_precision` - określa dokładność podanej daty wydania albumu, z którego pochodzi utwór, np. wartość "year" oznacza, że data została podana z dokładnością jedynie do roku, natomiast "day" oznacza, że data jest dokładna co do dnia
- `track.album.total_tracks` - ilość utworów na albumie, z którego pochodzi utwór
- `key_name` - zmienna typu string określająca tonację utworu
- `genre` - zmienna objaśniana kategoryczna, określająca gatunek muzyki, do którego należy utwór.

Tuż po ekstrakcji zmiennych za pomocą API, poza wymienionymi przeze mnie zmiennymi, w surowym zbiorze występowało również kilka zmiennych takich jak `track.album.href`, które przyjmowały same linki, więc zdecydowałem się je pominąć w opisie.

Chcę również wspomnieć, że powyższy opis zmiennych może ulec zmianie, ponieważ zamierzam w dalszym ciągu eksperymentować z API i spróbować wyodrębnić jeszcze inne zmienne.