Grupa 11:15 pon.	Temat: **Czy z wyekstrachowanych cech utworu da się przewidzieć czy będzie popularny.**	WIMiR Inżynieria Akustyczna
8.01.22	Filip Bober	

1. Wstęp.

Celem tego projektu jest napisanie klasyfikatorów, które będą w stanie przewidzieć na podstawie cechy utworu czy będzie popularny.

Aby tego dokonać modele są uczone na bazie utworów z zestawień z serwisu Spotify Daily Top 200 (codzienne 200 najczęściej słuchanych utworów) dla 35 krajów i świata. Dane były zbierane na przestrzeni 3 lat. Cała baza ma wymiar 170633 x 151, z czego cechy do nauki modelu zajmują 10 kolumn.

2. Przygotowanie i weryfikacja danych

Do projektu zawęziłem dane do Polski. Dzięki temu otrzymałem nową bazę o wymiarach 5273x151. Baza zawiera kolumny(przetłumaczone z języka angielskiego):

-Title: tytuł utworu

-URI: unikalny identyfikator utworu stworzony przez Spotify

-Country: Świat i kraje w których działa Spotify

-Popularity: Popularność liczona na zasadzie 1. miejsce 200pkt., 2. miejsce 199pkt. I tak dalej. Do tego punkty były mnożone:

3 razy za 1. miejsce

2.2 razy za 2. miejsce

1.7 razy za 3. miejsce

1.3 razy za 4-10 miejsce

1 razy za 11-50 miejsce

0.85 razy za51-100 miejsce

0.8 razy za 101-200 miejsce

-Artist: Nazwa wykonawcy

-Album/Single: Czy utwór był wydany w albumie czy jako singiel

-Genre: gatunek wykonawcy wg. Spotify

-Artist_followers: liczba obserwujących wykonawcę na Spotify

-Album: Nazwa albumu

-Release: Data wydania utworu

-Track_Number: numer w albumie

-Track_Album: ilość utworów w albumie

-Danceability: wielkość opisująca jak utwór jest dobry do tańczenia na podstawie tempa, stabilności rytmu, siły bitu i ogólnej regularności. 0.0 najmniej taneczny, a 1.0 najbardziej taneczny.

-Energy: wielkość odczuwalna intensywności i aktywności utworu. Od 0.0 do 1.0

-Key: Klucz w jakim napisany jest utwór. Np. 0=C, 1=C#, 2=D itd. Jeżeli klucz nie został wykryty to wartość wynosi -1

-Loudness: Uśredniona głośność utworu wyrażona w dB

-Mode: Czy utwór jest w tonacji moll (0) czy dur (1)

-Speechiness: wielkość opisująca ile jest słów w utworze. Jeżeli utwór to np. podcast to wartość jest bliżej 1.0. Muzyka mieści się w przedziale 0,33-0,66.

-Acousticness: wielkości pewności czy utwór jest akustyczny. 1.0 reprezentuje wysoką pewność, że utwór jest akustyczny

-Instrumentalness: przewiduje czy utwór nie zawiera wokalu. 1.0 reprezentuje duże prawdopodobieństwo braku wokalu.

-Liveness: wykrywa publiczność na nagraniu. Im bliżej 1.0 tym bardziej prawdopodobne że publiczność występuje w utworze.

-Valence: wielkość od 0.0 do 1.0 opisująca muzyczną pozytywność. Im większa tym weselszy, pogodny utwór.

-Tempo: Tempo utworu wyrażone w BPM

-Duration: Czas trwania w ms.

-Time_signature: Ilość taktów

-Genre of the artist:

-Released:

-Anger: Ilość wyrazów związana ze złością

- Anticipation: Ilość wyrazów związana z oczekiwaniem

-Disgust: Ilość wyrazów związana z obrzydzeniem

-Fear: Ilość wyrazów związana ze strachem

-Joy: Ilość wyrazów związana z radością

-Sadness: Ilość wyrazów związana ze smutkiem

-Surprise: Ilość wyrazów związana z zaskoczeniem

-Trust: Ilość wyrazów związana z zaufaniem

-Bayes:

-LDA: Ilość wyrazów związana z miłością, zbirami, nostalgią, odkrywaniem, zabawą, pożądaniem, nadzieją i celebracją.

- Popumax: Najwyższa pozycja w danych

-Top50_dummy: Wielkość, która równa się 1.0 gdy utwór uzyskał co najmniej 50 miejsce na liście. Jeżeli nie to równa się 0.0.

Cechy wykorzystane do uczenia modeli to Danceability, Energy, Key, Loudness, Mode, Speechiness, Acousticness, Instrumentalness, Liveness, Valence.

3. Klasyfikacja

Do przeprowadzenia klasyfikacji korzystam z klasyfikatorów K-najbliższych sąsiadów oraz z Drzewa Losowego. Optymalizując liczbę sąsiadów wychodzi że dla 10 otrzymano najkorzystniejsze wyniki.

Za „labele" przyjęto kolumnę Top50_dummy, a za „X" cechy utworów wymienione w poprzednim punkcie.

4. Analiza wyników

KNeighbors:

Otrzymane wyniki:

accuracy = 0.6417061611374407

recall = 0.11898016997167139

precision = 0.3853211009174312

F1 = 0.18181818181818185

RandomForest:

Otrzymane: wyniki:

accuracy = 0.7184834123222749

recall = 0.26912181303116145

precision = 0.7089552238805971

F1 = 0.39014373716632444

Wartość precyzji w klasyfikatorze RandomForest jest prawie dwukrotnie większa od klasyfikatora k-sąsiadów. Z racji tego że precyzja mówi nam ile przewidzianych jako popularne utwory faktycznie jest popularnych, wybrano RandomForest jako lepszy klasyfikator w tym badaniu.

5. Podsumowanie

Otrzymane wyniki są satysfakcjonujące, co nie stanowi, że nie można ich poprawić. Zwiększenie liczby rekordów, jak i cechy do uczenia modelu może znacząco wpłynąć na wyniki.

6. Bibliografia

-baza danych: https://www.kaggle.com/pepepython/spotify-huge-database-daily-charts-over-3-years plik: Final database.csv
-dokumentacja biblioteki sci-kit learn: https://scikit-learn.org/stable/