Problem biznesowy

Serwis muzyczny "Pozytywka" ma problem z nowo dodawanymi wykonawcami, którzy nie mają przypisanego gatunku muzycznego. To może utrudniać użytkownikom wyszukiwanie i odkrywanie nowej muzyki, a także wpływać na skuteczność algorytmów rekomendacyjnych serwisu, co przekłada się na mniejszą chęć korzystania z niego i zarazem mniejsze zyski. Zadaniem biznesowym jest jak najlepsze dopasowanie gatunku / gatunków muzycznych do każdego artysty, który go nie posiada, bazując na podstawie utworów, które wykonuje.

Zadanie modelowania

Mamy do czynienia z zadaniem klasyfikacji. Polega ono na przypisywaniu gatunków muzycznych do zespołów. Badając liczbę próbek danych (ok. 100 000 rekordów utworów muzycznych), zdecydowaliśmy się zastosować klasyfikator najbliższych sąsiadów.

W modelu będziemy wykorzystywać parametry utworów dotyczące ich cech muzycznych / użytkowych:

- popularity
- release_date (jedynie rok)
- dancebility
- energy
- speechiness
- valence
- duration_ms
- explicit

Dane wejściowe: Utwory muzyczne (w formie liczbowych parametrów opisujących ich cechy) artystów, którzy nie posiadają przypisanego gatunku muzycznego w serwisie.

Dane wyjściowe: 1 gatunek muzyczny najbardziej pasujący do danego artysty w modelu podstawowym oraz kilka gatunków pasujących do stylu muzycznego artysty w modelu rozszerzonym.

Założenia:

- Model będzie klasyfikował artystę do jednej z parunastu grup gatunków muzycznych.
- Ze względu na dużą liczbę pierwotnych klas (125 gatunków) konieczne będzie ręczne pogrupowanie podobnych gatunków muzycznych.
- Nie posiadamy kosztownych zasobów wiedzy eksperckiej, dlatego grupowanie odbędzie się na podstawie własnego researchu.
- Zbiory: treningowy, testowy i walidacyjny zostaną wybrane z danych artystów, którzy posiadają przypisany gatunek muzyczny.
- Zdolność do generalizacji modelu zostanie zbadana z wykorzystaniem metody walidacji krzyżowej.

Kryteria sukcesu

Biznesowe kryteria sukcesu: Każdy wykonawca znajdujący się w serwisie powinien posiadać min. 1 przypisany gatunek muzyczny. Średnia ocena satysfakcji korzystania z serwisu wśród klientów powinna wzrosnąć wraz z dynamiką przybywających nowych użytkowników.

Analityczne kryteria sukcesu: Model powinien osiągnąć czułość na poziomie min. 16%* i precyzję na poziomie min. 4%*.
Obie miary liczone są jako makro-średnie

(a) czułości:

$$\frac{1}{N}\sum_{i=1}^{N}(\frac{TP_i}{TP_i + FN_i})$$

(b) precyzji:

$$\frac{1}{N}\sum_{i=1}^{N}(\frac{TP_i}{TP_i + FP_i})$$

gdzie TP_i - liczba prawdziwych pozytywnych przypadków dla klasy i
FN_i - liczba fałszywych negatywnych przypadków dla klasy i
FP_i - liczba fałszywych pozytywnych przypadków dla klasy i

Model powinien dawać podobne wyniki na różnych zbiorach danych z walidacji krzyżowej, co świadczy o jego dobrej zdolności do generalizacji.

*Są to wartości 2 razy większe w porównaniu z wynikami naiwnego klasyfikatora przypisującego klasę większościową wszystkim artystom, po pogrupowaniu gatunków muzycznych z założeniem, że jeden artysta posiada jeden przypisany gatunek.

	precision	recall	f1-score	support
asian pop	0.00	0.00	0.00	1509
classical	0.00	0.00	0.00	314
country	0.00	0.00	0.00	408
electronic	0.00	0.00	0.00	1140
folk	0.00	0.00	0.00	1422
hip hop	0.00	0.00	0.00	1968
jazz/r&b	0.00	0.00	0.00	1355
latin	0.00	0.00	0.00	2227
metal	0.00	0.00	0.00	437
pop/dance	0.26	1.00	0.41	5089
rock	0.00	0.00	0.00	3216
world music	0.00	0.00	0.00	529
accuracy			0.26	19614
macro avg	0.02	0.08	0.03	19614
weighted avg	0.07	0.26	0.11	19614

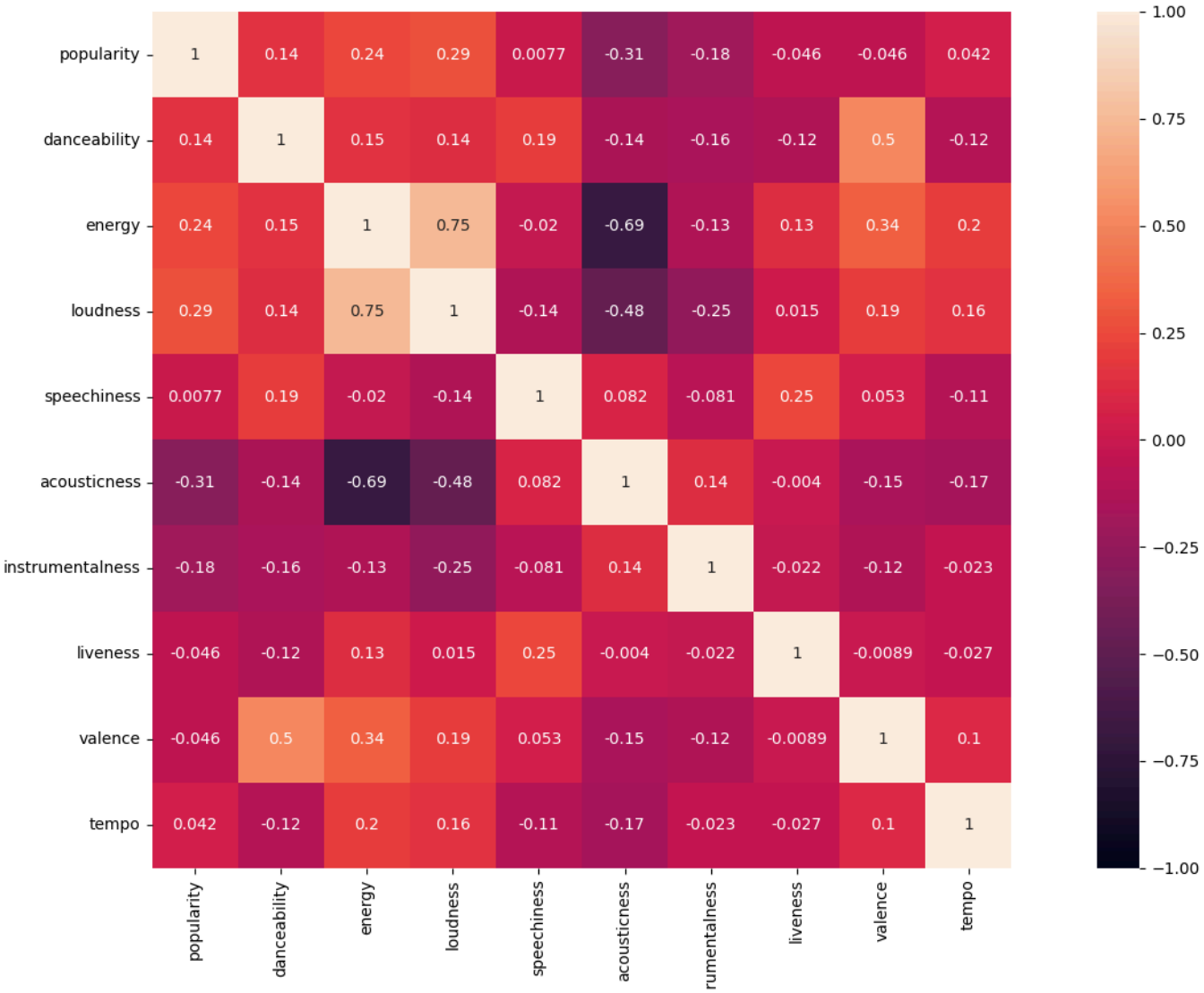
Analiza Danych

Braki i błędy w danych

- Atrybut ‘mode’ wskazujący modalność ścieżki dźwiękowej (dur =1 lub moll = 0) nie będzie przez nas uwzględniony ze względu na duże braki jego występowania. Ponad 103 000 piosenek nie ma określonej modalności, podczas gdy tylko 26 000 rekordów posiada sprecyzowaną wartość.
- 1262 piosenki posiadają sygnaturę czasową (parametr ‘time_signature’) poniżej dopuszczalnego dolnego limitu. Jednak i tak nie będziemy uwzględniać tego atrybutu ze względu na niską wartość współczynnika wzajemnej informacji.
- 53 piosenki posiadają wartość parametru ‘loudness’, która nie mieści się w typowej skali między -60 a 0 decybeli. Uznaliśmy te wartości za szumy pomiarowe, które należy zignorować.
- Część piosenek w polu “release_date” posiada jedynie rok wydania piosenki bez dokładnej daty. W naszym modelu i tak będziemy uwzględniać sam rok.
- Plik artists.jsonl posiada 27650 rekordów artystów o poprawnie zdefiniowanym id, imieniu i przypisanych gatunkach.
- Dane z plików sessions.jsonl i users.jsonl tymczasowo nie są planowane do wykorzystania w modelu, lecz po wstępnej analizie nie wykazują żadnych nieprawidłowości. W pliku z sesjami znajdują się puste wartości track_id, jednak wynikają one z sesji odtwarzania reklamy.
- Dane z pliku track_storage.jsonl nie są użyteczne dla realizowanego modelu.

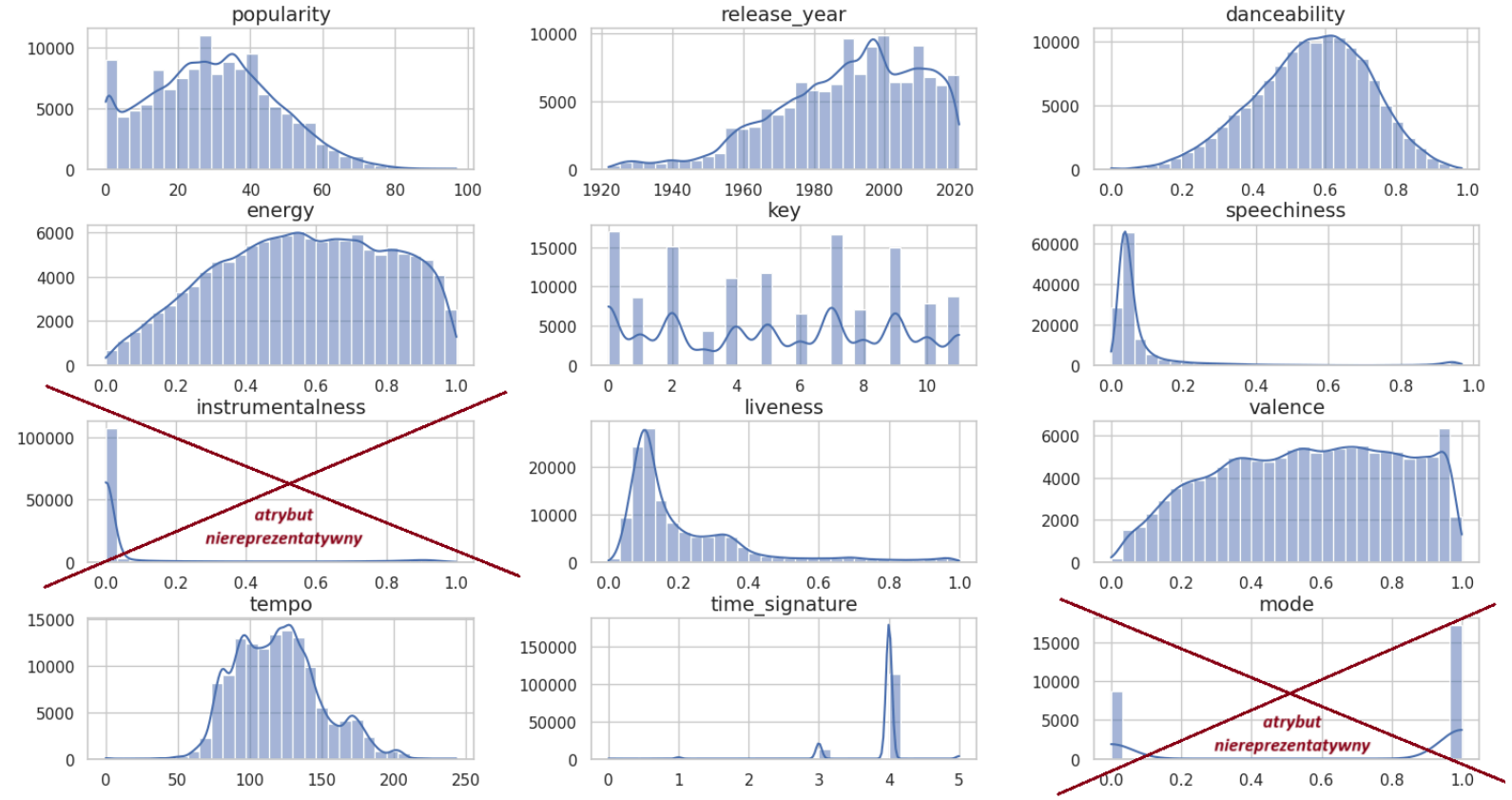
Korelacje między ciągłymi atrybutami

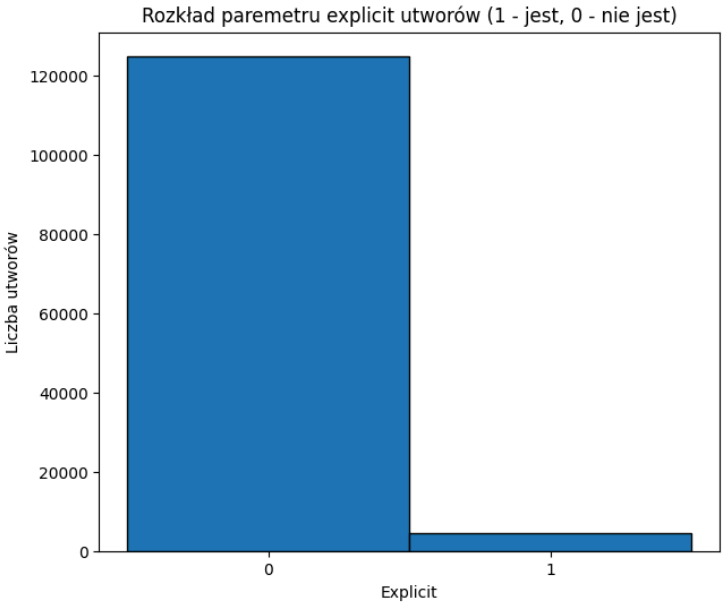
Korelacje między niektórymi zmiennymi o wartościach ciągłych:

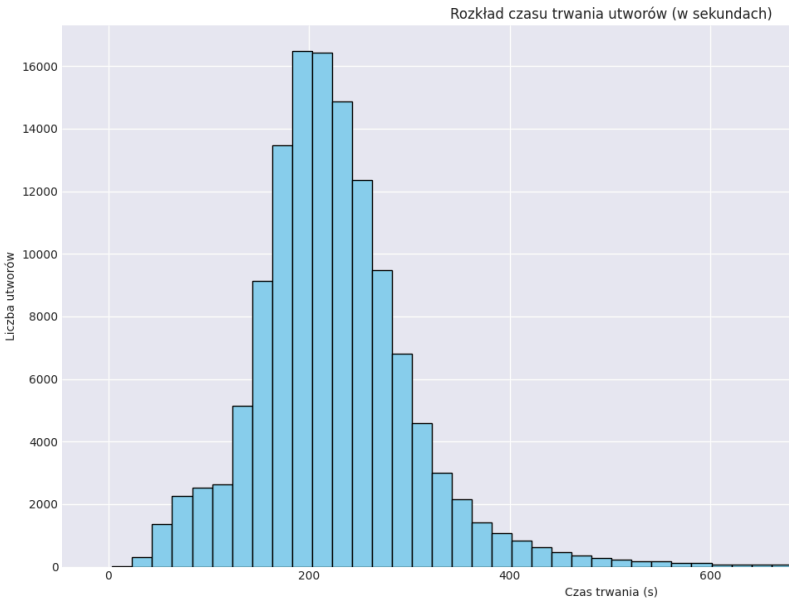


Współczynniki energy, loudness, acousticness są ze sobą mocno skorelowane. Inną zauważalnie mocniej skorelowaną parą współczynników jest danceability i valence.

Rozkłady atrybutów

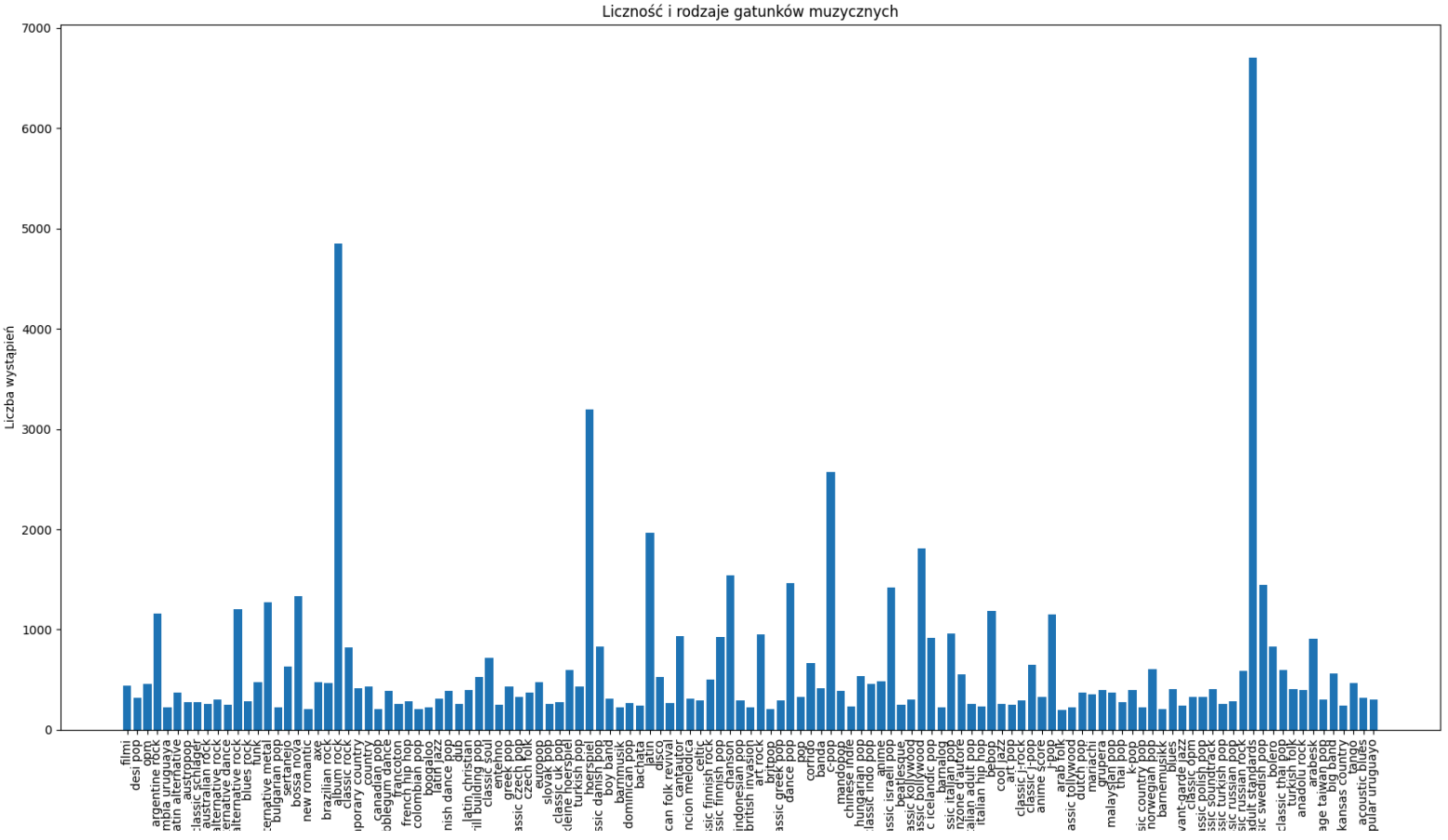






Ograniczenie liczby gatunków

W oryginalnych danych liczebność gatunków muzycznych to 3953, którym odpowiada 129648 utworów. Po usunięciu z danych gatunków muzycznych reprezentowanych przez mniej niż 200 utworów, zostało 125 gatunków muzycznych i 86429 utworów. Poniżej rozkład liczby utworów przypadających na gatunek muzyczny:



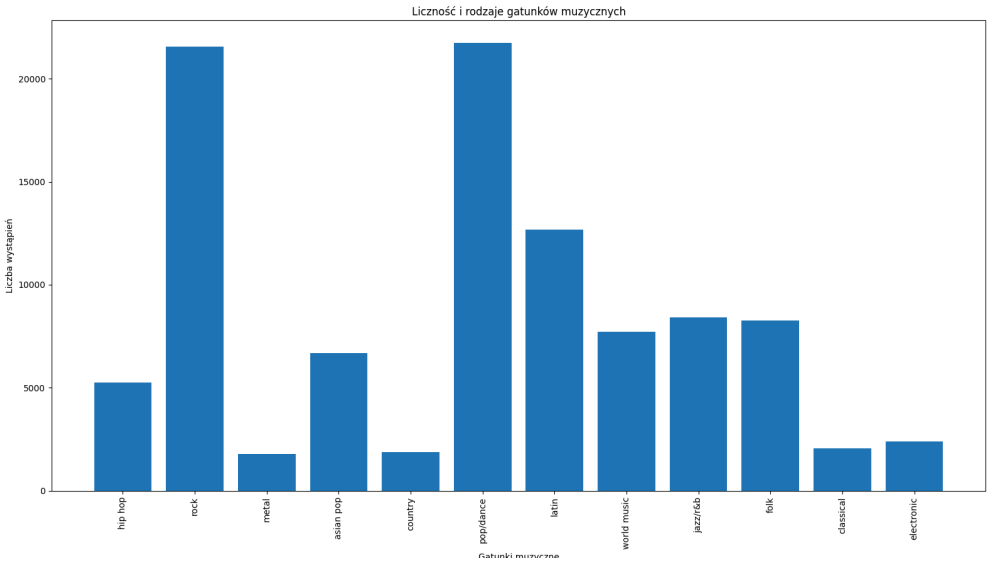
Wszystkie ze 125 gatunków przydzieliliśmy ręcznie do nowych, utworzonych grup. Ponadto dla każdej grupy stworzyliśmy listę słów kluczowych, która pozwoli klasyfikować do niej pozostałe z kilku tysięcy gatunków. Przydzielenia do danych grup gatunków odbywają się zgodnie z priorytetami: najpierw sprawdzane jest czy gatunek nie znajduje się na liście 125 najpopularniejszych i nie jest ręcznie przydzielony, a następnie sprawdzane jest dopasowanie słów kluczowych w odpowiedniej kolejności np. słowo ‘metal’ ma większy priorytet niż dopasowanie po słowie ‘classical’.

Nowa grupa	Stare gatunki
Rock	'argentine rock', 'austropop', 'australian rock', 'australian alternative rock', 'alternative dance', 'alternative rock', 'blues rock', 'alternative metal', 'new romantic', 'brazilian rock', 'album rock', 'classic rock', 'british invasion', 'art rock', 'britpop' oraz wszystkie posiadające słowo ‘rock’ w nazwie
Pop / Dance	'desi pop', 'opm', 'bulgarian pop', 'canadian pop', 'bubblegum dance', 'europop', 'slovak pop', 'classic uk pop', 'boy band', 'dance pop', 'pop', 'c-pop', 'classic italian pop', 'italian adult pop', 'art pop', 'dutch pop', 'classic norwegian pop', 'classic opm', 'classic polish pop', 'classic turkish pop', 'classic russian pop', 'classic swedish pop', 'finnish dance pop', 'disco' oraz wszystkie posiadające w nazwie słowa: ‘pop’, ‘dance’, ‘opm’, ‘disco’
Asian Pop (opcjonalnie)	'anime', 'anime score', 'classic j-rock', 'classic j-pop', 'j-pop', 'c-pop', 'mandopop', 'chinese indie', 'classic indo pop', 'k-pop', 'thai pop', 'classic thai pop', 'vintage taiwan pop', 'classic malaysian pop' oraz wszystkie posiadające w nazwie słowo ‘anime’ lub wyrażenie ‘pop’ + słowo z listy asian_nationalities = ['afghan', 'armenian', 'azerbaijani', 'bahraini', 'bangladeshi', 'bhutanese', 'bruneian', 'burmese', 'cambodian', 'chinese', 'cypriot', 'emirati', 'filipino', 'georgian', 'indian', 'indonesian', 'iranian', 'iraqi', 'israeli', 'japanese', 'jordanian', 'kazakhstani', 'kuwaiti', 'kyrgyzstani', 'laotian', 'lebanese', 'malaysian', 'maldivian', 'mongolian', 'nepalese', 'north korean', 'omani', 'pakistani', 'palestinian', 'saudi', 'singaporean', 'south korean', 'sri lankan', 'syrian', 'taiwanese', 'thai', 'timorese', 'turkish', 'turkmen', 'uzbekistani', 'vietnamese', 'yemeni']
Latin	'filmi', 'cumbia uruguaya', 'latin alternative', 'sertanejo', 'bossa nova', 'axe', 'latin jazz', 'latin christian', 'bachata', 'latin', 'corrido', 'banda', 'mariachi', 'grupera', 'bolero', 'tango' oraz wszystkie zawierające jedno ze słów z listy: latin_music_genres = [‘latin’, ‘latino’, 'salsa', 'merengue', 'reggae', 'reggaeton', 'cumbia', 'bolero', 'flamenco', 'tango', 'ranchera', 'mariachi', 'norteña', 'samba', 'bossa nova', 'trova', 'son', 'rumba', 'mambo', 'cha-cha-cha', 'fado', 'vallenato', 'pop latino', 'rock en español', 'jazz latino’, ‘mexico']

Country	'contemporary country', 'country', 'arkansas country', 'classic country pop’ oraz wszystkie zawierające słowo ‘country’
World Music	'kleine hoerspiel', 'hoerspiel', 'cantautor', 'cancion melodica', 'celtic', 'chanson', 'arab folk', 'classic tollywood', 'canzone d'autore', 'anadolu rock', 'arabesk', 'canto popular uruguayo'
Jazz / R&B	'bebop', 'cool jazz', 'avant-garde jazz', 'classic soul' oraz wszystkie zawierające słowo ‘jazz’, ‘r&b’, ‘funk’, ‘soul’, ‘swing’, ‘saxophone’, ‘bop’
Electronic	'dub' oraz zawierające któreś ze słów electronic_music_words = ['techno', 'rave', 'electronica', 'dubstep', 'dark', 'house', 'industrial', 'dub', 'synth', 'new-age', 'deep', 'nu-disco', 'garage', 'EDM', 'breakbeat', 'trance', 'fusion', 'gabber', 'future', 'beats', 'effects', 'electronic', 'bass', 'FX', 'hardstyle', 'hard', 'kick']
Hip Hop	'italian hip hop', 'french hip hop' oraz wszystkie zawierające wyrażenie ‘hip hop’
Classical	'classic soundtrack', 'adult standards' oraz wszystkie zawierające któreś ze słów ‘classical’, ‘piano’, ‘soundtrack’, ‘instrumental’
Folk	'american folk revival', 'classic finnish rock', 'classic finnish pop', 'classic greek pop', 'classic hungarian pop', 'classic israeli pop', 'beatlesque', 'classic kollywood', 'classic bollywood', 'classic icelandic pop', 'barnalog', 'classic italian pop', 'classic russian rock', 'big band', 'acoustic blues', 'czech folk’, 'turkish folk' oraz wszystkie zawierające słowo ‘folk’
Metal	Wszystkie zawierające słowo ‘metal’

Wyniki po grupowaniu gatunków

Rozkład liczności utworów na gatunek po przypisaniu do utworu gatunku, który występuje najczęściej dla danego wykonawcy.



W przypadku gatunków metal, country, classical i electronic mamy mało wystąpień, konieczne będzie zbalansowanie danych.

Współczynnik informacji wspólnej

Gatunki są nieuporządkowanymi zmiennymi dyskretnymi, więc konieczne było wyliczenie współczynnika informacji wspólnej.
Do obliczenia współczynnika informacji wspólnej, wartości atrybutów zostały zgrupowane w przedziały:

- popularity - 10 przedziałów, co 10
- dancebility - 10 przedziałów co 0,1
- energy - 10 przedziałów co 0,1
- liveness - przedział od 0,5 do 1 i 25 przedziałów co 0,02
- speechiness - przedział od 0,2 do 1, 8 przedziałów co 0,05
- tempo - 50 przedziałów co 5
- valence - 10 przedziałów co 0,1
- duration_ms - przedziały co 10 sekund

	genre
popularity	0.095230
release_year	0.220934
danceability	0.077078
energy	0.100214
key	0.023027
speechiness	0.125954
liveness	0.026085
valence	0.043403
tempo	0.035863
time_signature	0.028686
duration_ms	0.133490
explicit	0.065591

Atrybutem, który ma największy współczynnik informacji wspólnej jest release_year, z kolei atrybutami, które mają mały wpływ na gatunek są: tempo, liveness, key i time_signature.

Wyniki analizy i selekcja kluczowych zmiennych wejściowych

Wszystkie zmienne wejściowe uwzględniane w modelu będą cechami utworów przypisanych do poszczególnych artystów za pomocą id. Imię/pseudonim artysty nie będzie uwzględniane jako wejście modelu, tak samo jak nazwa utworu, gdyż mogłyby nieść one informację o gatunku tylko w szczególnych przypadkach, a uwzględnienie ich jako wartość liczbową jest trudne.

Z powodu mocnego skorelowania z atrybutów energy, loudness, acousticness w modelu użyte będzie tylko energy.
Z powodu braków w danych nie zostanie użyty atrybut mode.
Z powodu niereprezentatywności nie zostanie użyte instrumentalness
Z powodu małego współczynnika informacji wspólnej, w podstawowym modelu nie zostaną użyte atrybuty tempo, liveness, key i time_signature.

Mamy mało przypadków utworów, które są explicit, ale sam atrybut ma jeden z większych współczynników informacji wspólnej z gatunkiem, więc być może po zbalansowaniu gatunków zostanie zniwelowana nierównomierność atrybutu.

Lista atrybutów użytych w modelu:

- popularity
- release_date (jedynie rok)
- dancebility
- energy
- speechiness
- valence
- duration_ms
- explicit

Autorzy: Michał Kowalczyk, Jakub Kowalczyk

Etap 2 - implementacja

Model podstawowy

Model podstawowy uwzględnia wszystkie wskazane przez nas parametry wejściowe i na ich podstawie przypisuje każdemu utworowi jeden najbardziej dopasowany gatunek. Etykiety gatunków do utworów zostały przez nas przypisane następująco:

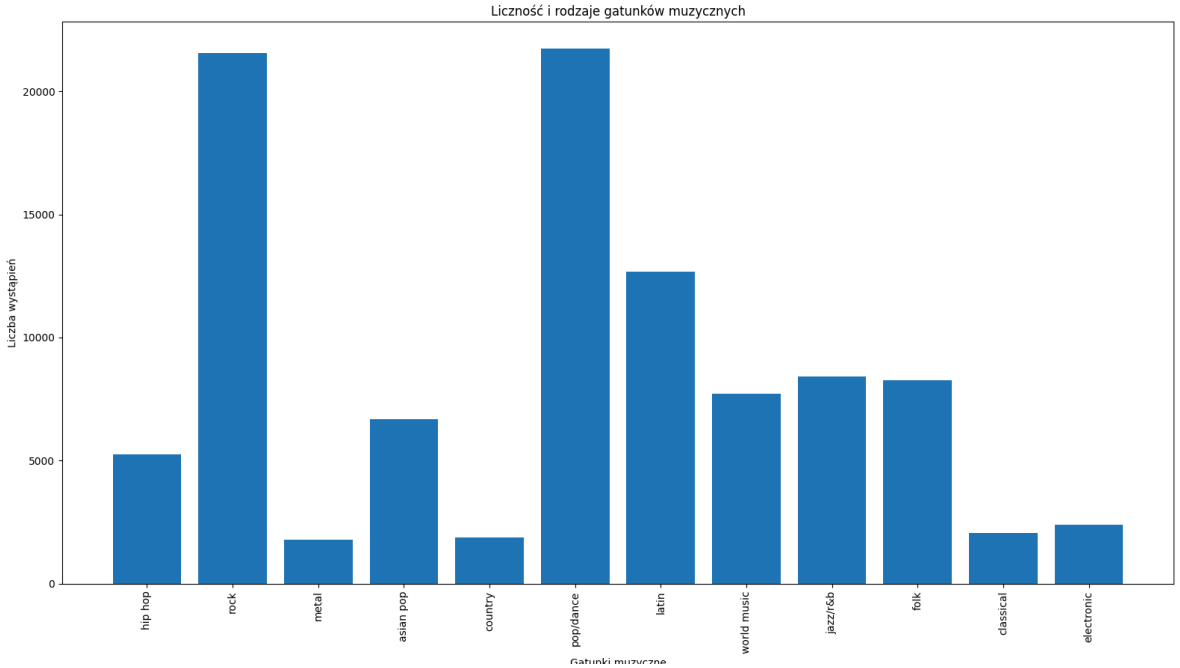
1. Wszystkie aktualne gatunki artysty są mapowane na nowe gatunki np. ["opm", “classic opm”, "pinoy reggae"] staje się [“pop”, “pop”, “latin”].
2. Z nowej listy wybierany jest gatunek występujący najczęściej. Dla [“pop”, “pop”, “latin”] będzie to pop.
3. Nowy gatunek jest przypisywany wszystkim utworom artysty. W podanym przykładzie wszystkie piosenki artysty będą popowe.
4. Jeżeli wszystkie gatunki arytsty są zbyt mało popularne, aby je zaklasyfikować do jakiejś podgrupy kilkunastu nowych gatunków, to wówczas taki rekord jest pomijany.

Następnie dla każdego artysty mającego zdefiniowane gatunki z nowych podgrup, sprawdzamy wszystkie jego piosenki i liczymy średnie prawdopodobieństwa na przypisanie podejrzewanych gatunków. Ten, którego owe prawdopodobieństwo jest największe jest przypisywany artyście.
Istnieje także opcja predykcji gatunku dla konkretnych parametrów utworu lub id artysty znajdującego się w bazie.

Rozwój modelu podstawowego

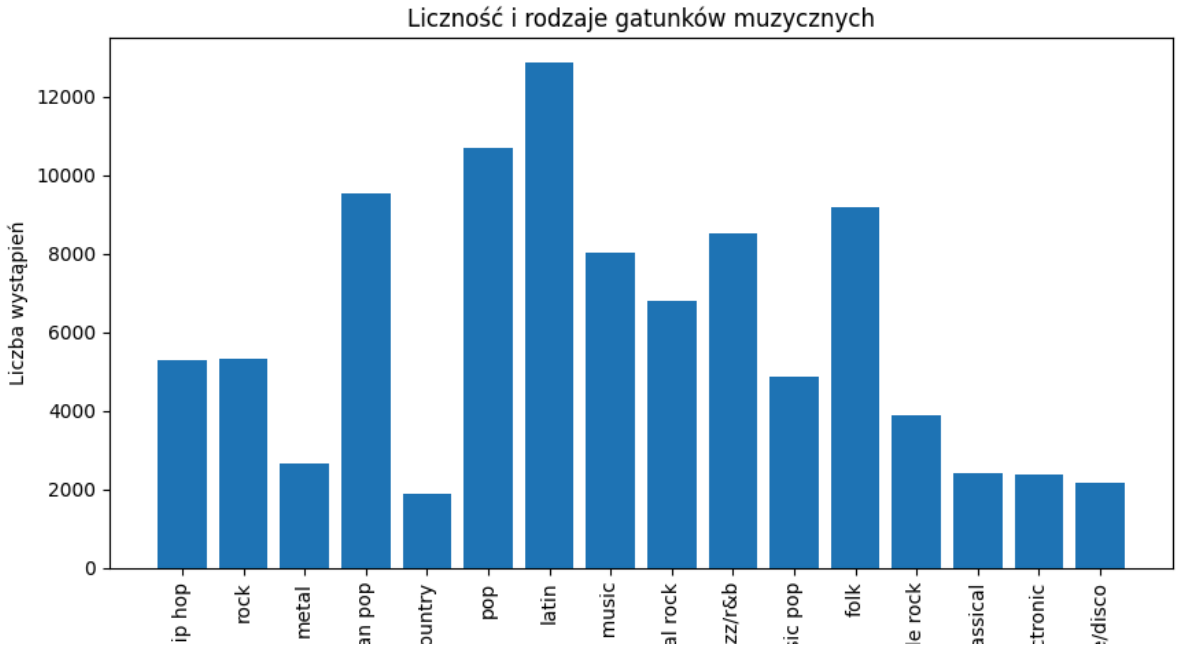
1. Początkowo wykorzystaliśmy klasyfikator najbliższych sąsiadów, zgodnie z planem, dla niezbalansowanych danych. Udało się uzyskać precyzję na poziomie ok. 19% i czułość na poziomie ok. 18%. Przedstawiona tabela dotyczy wyników klasyfikacji dla utworów. Przyjęta została liczba sąsiadów k=3.

	precision	recall	f1-score	support
asian pop	0.10	0.24	0.15	1357
classical	0.05	0.10	0.06	391
country	0.03	0.05	0.04	382
electronic	0.10	0.15	0.12	479
folk	0.14	0.21	0.17	1653
hip hop	0.16	0.20	0.18	1068
jazz/r&b	0.20	0.19	0.19	1726
latin	0.16	0.15	0.16	2540
metal	0.03	0.02	0.03	328
pop/dance	0.28	0.20	0.23	4252
rock	0.31	0.19	0.23	4364
world music	0.66	0.45	0.53	1544
accuracy			0.20	20084
macro avg	0.19	0.18	0.17	20084
weighted avg	0.25	0.20	0.22	20084



2. Po ustaleniu hiperparametru liczby sąsiadów na pierwiastek 4 stopnia z liczby próbek, czyli k = 18, precyzja i czułość wzrosły o około 2 punkty procentowe. Nadal nie był to spektakularny wynik.
3. Po zbalansowaniu danych - rozbiciu rocka i popu na pomniejsze grupy - precyzja spadła do 15%, a czułość do 13%. Klasyfikator nie poradził sobie dobrze ze zwiększeniem liczby przewidywanych klas.

	precision	recall	f1-score	support
asian pop	0.15	0.32	0.20	1882
classic pop	0.06	0.05	0.05	957
classical	0.09	0.04	0.06	493
country	0.06	0.02	0.02	392
dance/disco	0.06	0.01	0.02	453
electronic	0.17	0.07	0.10	496
folk	0.15	0.20	0.17	1763
geographical rock	0.09	0.07	0.08	1346
hip hop	0.14	0.08	0.11	1052
jazz/r&b	0.22	0.22	0.22	1725
latin	0.17	0.26	0.20	2556
metal	0.04	0.01	0.01	558
pop	0.19	0.19	0.19	2204
rock	0.11	0.04	0.06	1030
style rock	0.11	0.02	0.03	818
world music	0.57	0.44	0.50	1610
accuracy			0.18	19335
macro avg	0.15	0.13	0.13	19335
weighted avg	0.18	0.18	0.17	19335



4. Zdecydowaliśmy się na zmianę modelu na las losowy z liczbą drzew 100. Wyniki były znacznie lepsze, jednak wzrósł także czas obliczeń. Dla klasyfikacji utworów udało się uzyskać precyzję ok. 43% i czułość 37%. Dla klasyfikacji artystów były to wyniki na poziomie 95%.

	precision	recall
asian pop	0.34	0.46
classic pop	0.18	0.07
classical	0.41	0.28
country	0.30	0.11
dance/disco	0.47	0.11
electronic	0.72	0.42
folk	0.37	0.41
geographical rock	0.28	0.22
hip hop	0.64	0.74
jazz/r&b	0.47	0.53
latin	0.38	0.53
metal	0.49	0.50
pop	0.33	0.34
rock	0.39	0.43
style rock	0.30	0.24
world music	0.78	0.53
accuracy		
macro avg	0.43	0.37
weighted avg	0.42	0.41

	precision	recall
asian pop	0.93	0.95
classic pop	1.00	0.92
classical	1.00	1.00
country	1.00	0.86
dance/disco	1.00	0.95
electronic	1.00	0.92
folk	0.91	0.92
geographical rock	0.94	0.94
hip hop	0.95	0.99
jazz/r&b	0.97	0.97
latin	0.90	0.97
metal	1.00	0.96
pop	0.92	0.90
rock	0.86	1.00
style rock	1.00	0.96
world music	0.97	1.00
accuracy		
macro avg	0.96	0.95
weighted avg	0.95	0.94

5. W ramach strojenia hiperparametrów wykonaliśmy losowy test na 50 próbkach par hiperparametrów z walidacją krzyżową dzielącą zbiór na 5 podzbiorów. Losowana liczba drzew pochodziła ze zbioru {50, 51, …, 200}, a maksymalna głębokość ze zbioru {5, 6, …, 20} Najlepsze parametry to 173 drzewa o maksymalnej głębokości 19.

```
Model 1 best: {'max_depth': 19, 'n_estimators': 173}
```

6. Po zmianie hiperparametrów na uzyskane w ramach strojenia precyzja i czułość nieznacznie się poprawiły, a w przypadku klasyfikacji gatunków artystów nastąpił nieznaczny spadek.

Base model evaluation:

	precision	recall
asian pop	0.35	0.50
classic pop	0.23	0.04
classical	0.53	0.30
country	0.35	0.09
dance/disco	0.49	0.07
electronic	0.74	0.41
folk	0.39	0.44
geographical rock	0.30	0.19
jazz/r&b	0.49	0.52
latin	0.35	0.55
metal	0.53	0.48
pop	0.33	0.37
rock	0.38	0.44
style rock	0.31	0.24
world music	0.84	0.53
accuracy		
macro avg	0.45	0.37
weighted avg	0.44	0.42

	precision	recall
asian pop	0.67	0.91
classic pop	1.00	0.73
classical	1.00	0.89
country	1.00	0.60
dance/disco	0.96	0.73
electronic	0.96	0.85
folk	0.87	0.82
geographical rock	0.92	0.79
hip hop	0.90	0.94
jazz/r&b	0.96	0.84
latin	0.69	0.83
metal	0.91	0.67
pop	0.76	0.86
rock	0.81	0.77
style rock	1.00	0.81
world music	0.81	0.81
accuracy		
macro avg	0.89	0.80
weighted avg	0.85	0.83

Model rozszerzony

Model rozszerzony opera się na klasyfikatorze k najbliższych sąsiadów.

Model będzie trenowany na innym zbiorze, w którym będziemy uwzględniać też mniej . W tym przypadku przypisywanie etykiet gatunków do utworów wygląda następująco:

1. Wszystkie aktualne gatunki artysty są mapowane na nowe gatunki np. ["opm", “classic opm”, "pinoy reggae"] staje się [“pop”, “pop”, “latin”].
2. Gatunki, które są zbyt mało popularne nie pojawią się na nowej liście
3. Dla każdego gatunku znajdującego się na nowej liście na liście utworów zapisujemy kopię rekordu utworu z tym gatunkiem. Jeśli gatunek pojawi się częściej niż raz, utwór zostanie zapisany z tym gatunkiem w wielu kopiach
4. **** dane przesłane w archiwum nie są do końca zgodne z tym opisem, w momencie przesyłania w danych do modelu rozszerzonego został użyty plik, w którym nie były zapisywane kopie utworów przy powtarzaniu się gatunków.

W ramach testowania modelu przeprowadziliśmy test na wszystkich możliwych trójkach parametrów ze zbiorów

```
param_grid = {'n_neighbors' : np.arange(5, 30, 1), 'weights' : ['uniform', 'distance'], 'p' : [1, 2]}
```

z walidacją krzyżową dzielącą zbiór na 5 podzbiorów

```
Model 2 best: {'n_neighbors': 29, 'p': 1, 'weights': 'distance'}
```

Parametr dotyczący ilości sąsiadów znajdował się na górnej granicy, ale ze względu na charakterystykę algorytmu n sąsiadów liczba sąsiadów została przyjęta na 29, ale być może istnieje konieczność dodatkowego sprawdzenia parametrów modelu dla większej ilości sąsiadów.

Complex model evaluation:

	precision	recall
asian pop	0.36	0.42
classic pop	0.05	0.04
classical	0.18	0.18
country	0.57	0.51
dance/disco	0.16	0.17
electronic	0.41	0.34
folk	0.15	0.15
geographical rock	0.27	0.29
hip hop	0.60	0.64
jazz/r&b	0.57	0.71
latin	0.52	0.49
metal	0.52	0.66
pop	0.25	0.21
rock	0.29	0.31
style rock	0.23	0.20
world music	0.58	0.37
accuracy		
macro avg	0.36	0.36
weighted avg	0.36	0.36

	precision	recall
asian pop	0.89	0.89
classic pop	0.93	0.78
classical	0.84	0.89
country	0.96	0.92
dance/disco	0.85	0.73
electronic	0.89	0.79
folk	0.84	0.86
geographical rock	0.68	0.90
hip hop	0.90	0.82
jazz/r&b	0.88	0.90
latin	0.83	0.84
metal	0.88	0.93
pop	0.77	0.84
rock	0.85	0.77
style rock	0.72	0.81
world music	0.73	0.30
accuracy		
macro avg	0.84	0.81
weighted avg	0.84	0.83

Precyzja i czułość zgadywania artystów są na podobnym poziomie co w modelu podstawowym, ale w przypadku przewidywania gatunków utworów jest mniejsza.

Porównanie modeli

Porównujemy modele, sprawdzając, czy poprawny główny gatunek został odgadnięty przez model rozszerzony i co w takiej sytuacji przewidział model podstawowy:

{'true_complex':
    {'true_simple': 662,
    'false_simple': 121},
'false_complex':
    {'true_simple': 137,
    'false_simple': 36}
}

Widzimy, że wyniki są podobne, ilość przypadków, w których jeden z modeli zgadł niepoprawnie, a drugi poprawnie jest podobna, 121 i 137.

Spełnienie kryteriów sukcesu

> Analityczne kryteria sukcesu: Model powinien osiągnąć czułość na poziomie min. 16%* i precyzję na poziomie min. 4%*.

Oba modele spełniły analityczne kryterium sukcesu.

W logach zapisywane są wartości predykcje gatunku wykonane przez dany model

```
2024-01-19 23:09:59,303 - [ModelA_artist] Input: 1Bl6wpkWCQ4KVgnASpvzzA, Prediction: hip hop
```

```
2024-01-19 23:08:20,451 - [ModelB_artist] Input: 1Bl6wpkWCQ4KVgnASpvzzA, Prediction: hip hop
```

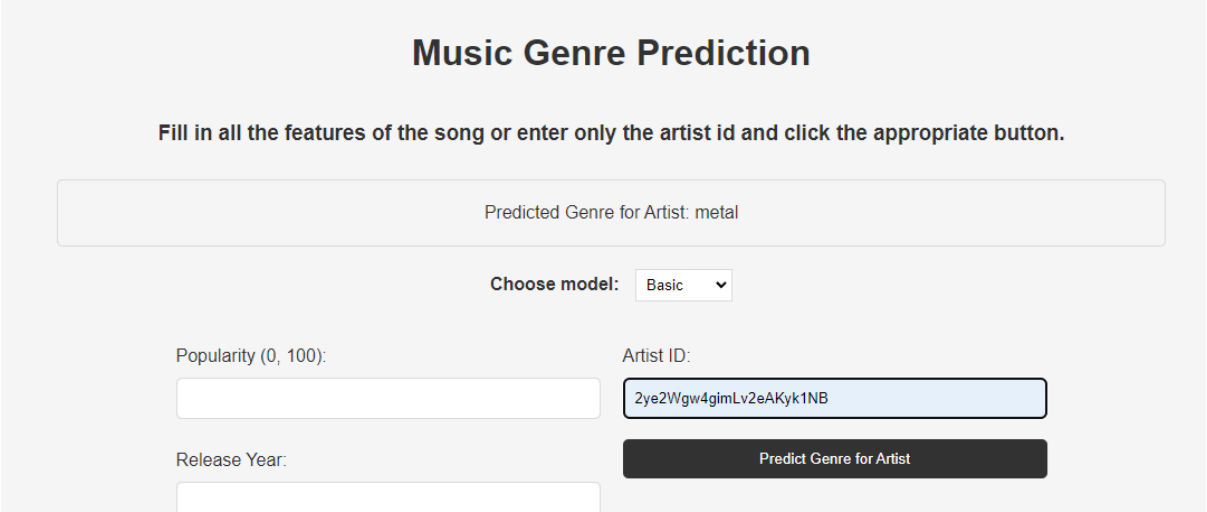
Wraz z działaniem modelu będzie można sprawdzić poprawność przypisanych etykiet przez oba modele.

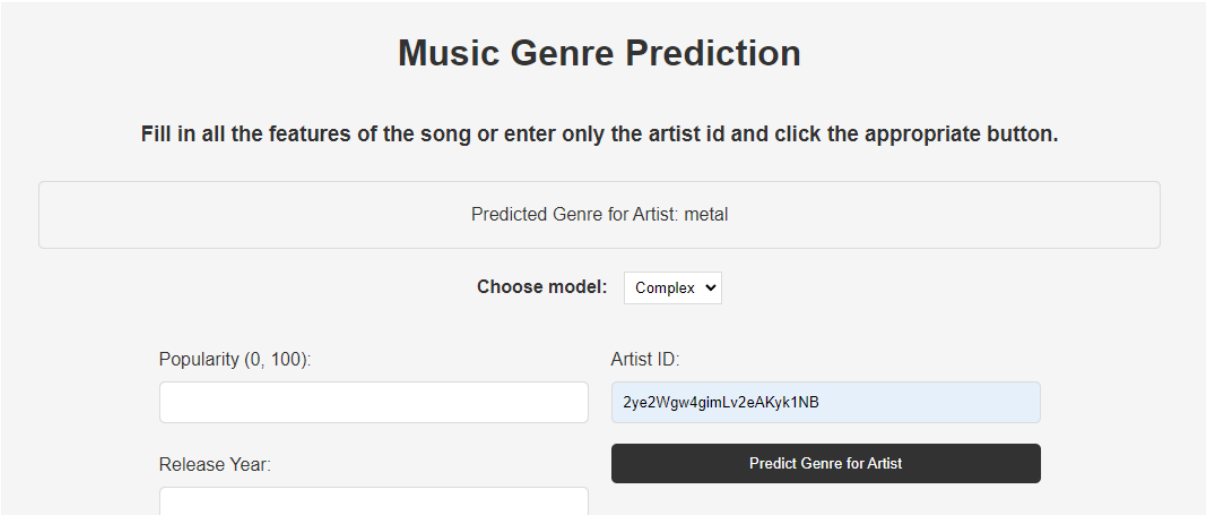
Działanie

Po uruchomieniu lokalnie serwera (app.py) należy przejść pod adres 127.0.0.1:5000/ i odpowiednio wypełnić pola formularza, aby uzyskać predykcję. Pełen raport klasyfikacji dla artystów modelu podstawowego jest tworzony pod adresem 127.0.0.1:5000/artist/report i wyświetlany w konsoli.
Dla modelu rozszerzonego wykorzystujemy 127.0.0.1:5000/artist/report-complex, a do porównania modeli 127.0.0.1:5000/artist/report-compare
Mikroserwis jest gotowy do wdrożenia w projekcie, co można sprawdzić wysyłając przykładowe żądania za pomocą komend:

1. dla utworu: *curl -X POST -H "Content-Type: application/json" -d '{"popularity": 58, "release_year": 1987, "danceability": 0.883, "energy": 0.631, "speechiness": 0.42, "valence": 0.782, "duration_ms": 250387, "explicit": 0}' http://127.0.0.1:5000/predict*
2. dla artysty o danym id:
   curl http://localhost:5000/predict/artist/6n6ot5JVa8YO9z82eNbvd8

Gatunki przewidziane dla artysty o id 2ye2Wgw4gimLv2eAKyk1NB za pomocą modelu podstawowego i rozszerzonego:

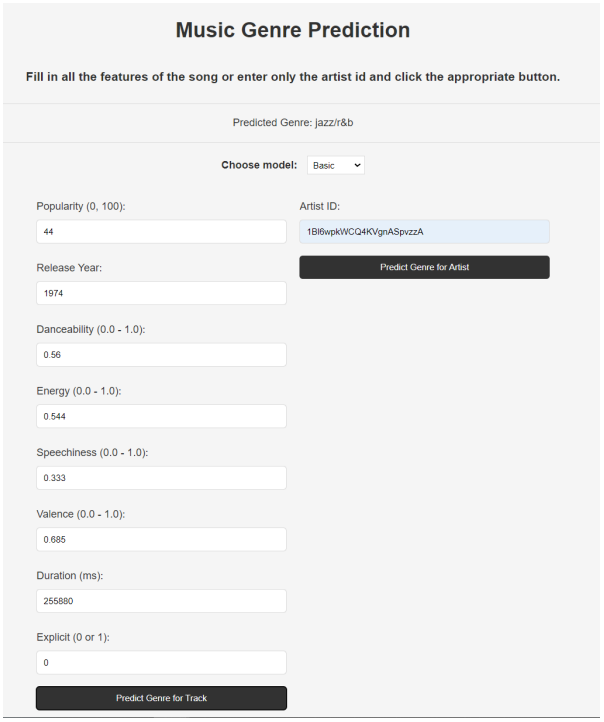


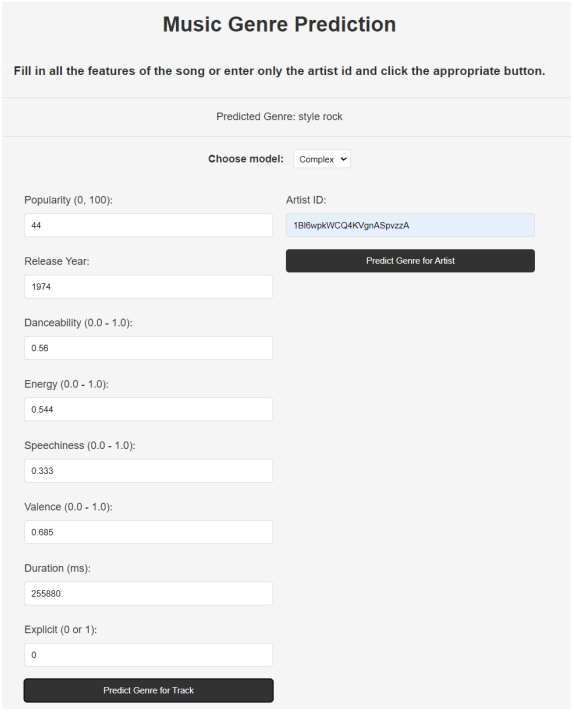


{"id": "2ye2Wgw4gimLv2eAKyk1NB", "name": "Metallica", "genres": ["hard rock", "metal", "old school thrash", "rock", "thrash metal"]}

Gatunki przewidziane dla danych utworu skopiowanych z

{"id":"6pqWrRF9K2PHpBmmRSIte4","name":"Cosmik Debris","popularity":44,"duration_ms":255880,"explicit":0,"id_artist":"6ra4GIOgCZQZMOaUECftGN","danceability":0.56,"energy":0.544,"speechiness":0.333,"valence":0.685,"genre":"rock","release_year":1974}
:


Music Genre Prediction
Fill in all the features of the song or enter only the artist id and click the appropriate button.
Predicted Genre: jazz/r&b
Choose model:
Basic
Popularity (0, 100):
44
Release Year:
1974
Danceability (0.0 - 1.0):
0.56
Energy (0.0 - 1.0):
0.544
Speechiness (0.0 - 1.0):
0.333
Valence (0.0 - 1.0):
0.685
Duration (ms):
255880
Explicit (0 or 1):
0
Predict Genre for Track
Artist ID:
1Bl6wpkWCQ4KVgnASpvzzA
Predict Genre for Artist


Music Genre Prediction
Fill in all the features of the song or enter only the artist id and click the appropriate button.
Predicted Genre: style rock
Choose model:
Complex
Popularity (0, 100):
44
Release Year:
1974
Danceability (0.0 - 1.0):
0.56
Energy (0.0 - 1.0):
0.544
Speechiness (0.0 - 1.0):
0.333
Valence (0.0 - 1.0):
0.685
Duration (ms):
255880
Explicit (0 or 1):
0
Predict Genre for Track
Artist ID:
1Bl6wpkWCQ4KVgnASpvzzA
Predict Genre for Artist