AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

AGH UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Analiza zbioru MyAnimeList

Dominik Dziuba
Miłosz Filus

https://www.kaggle.com/azathoth42/myanimelist

27.01.2021

Opis zbioru

Dane o anime: anime_filtered.csv

- 9.85 MB
- 14474 rekordy danych
- 31 atrybutów, w tym: 8 numerycznych, 7 kategorycznych i 15 tekstowych

Dane o użytkownikach: users_filtered.csv

- 34.3 MB
- 116133 rekordy danych
- 16 atrybutów, w tym: 10 numerycznych, 1 kategoryczny, 4 tekstowe

Wybór interesujących atrybutów

Dla zbioru o anime:

- Odrzucenie niepotrzebnych danych tekstowych (jak tytuły w innych językach)
- Odrzucenie danych o pochodzeniu danego anime (producent, wydawcy itp.)
- Odrzucenie danych o piosenkach początkowych i końcowych
- Ostatecznie wzięto pod uwagę dane o: tytuł (do indentyfikacji), liczba odcinków, długość odcinków, informacje o ocenie użytkowników i popularności, typ anime (np. czy serial telewizyjny czy film), gatunki

Dla zbioru o użytkownikach:

- Odrzucenie ciężkich w weryfikacji danych o lokacji użytkowników
- Odrzucenie danych o implementacji serwisu
- Odrzucenie danych o ostatnim loginie
- Ostatecznie wzięto pod uwagę informacje takie jak: nazwa użytkownika (do indentyfikacji), wiek, wiek konta, płeć, średnie oceny, ilość oglądanych anime

Wstępna obróbka danych

Dla obydwu zbiorów była dołączona wstępna wersja obrobiona danych i taka została wykorzystana:

- Rekordy anime nie zawierające poprawnych danych o pochodzeniu czy producencie oraz takie, które nie były jeszcze w momencie zbierania danych transmitowane.
- W zbiorze użytkowników, zostali usunięci tacy którzy nie podali lokacji, daty urodzenia czy płci.

Dla zbioru o użytkownikach dodatkowo zostało wykonane:

- Daty urodzenia i rejestracji w serwisie zamieniono na okresy od podanych dat
- Odfiltrowano użytkowników o nierealnych wiekach poza przedziałem [8, 80]
- Wyróżniliśmy podzbiór użytkowników, którzy nie ocenili ani jednego anime (mały w porównaniu z resztą zbioru)

Dla zbioru o anime:

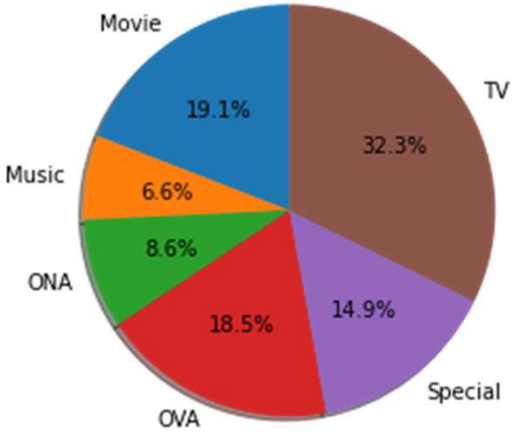
- Konwersja kategorii wiekowych na dane numeryczne
- Konwersja długości odcinków z danych tekstowych na numeryczne

Dla obydwu zbiorów usunięto wszystkie niepełne rekordy (mające puste pola).

Wybrane statystyki

Dla zbioru o anime:

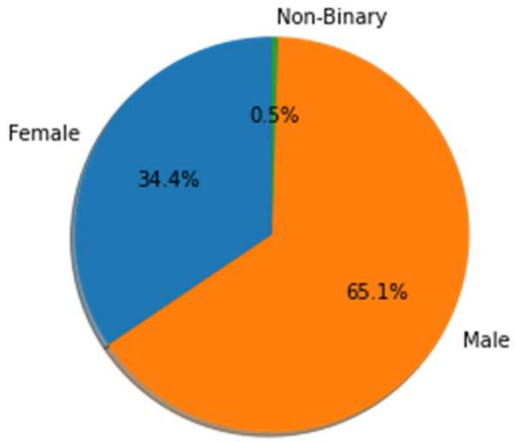
- Średnia liczba odcinków: 12
- Średnia długość odcinka: 16 min (mediana: 20 min)
- Średnia ocena użytkowników: 6.5



Rys. 1. Rozkład typów anime

Dla zbioru o użytkownikach:

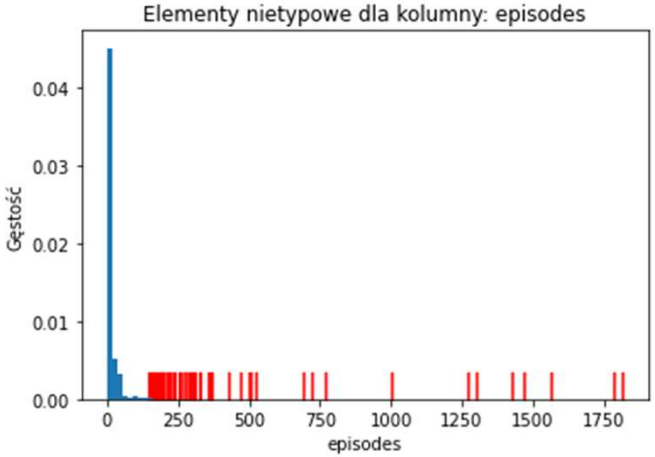
- Średnia oglądanych serii: 11
- Średnia ukończonych serii: 164
- Liczba dni spędzonych na oglądaniu: 58
- Średni wiek: mężczyźni – 28 lat,
  kobiety – 27 lat, non-binary – 26 lat
- Średnie ukończone serie:
  mężczyźni – 145,
  kobiety – 84,
  non-binary – 144

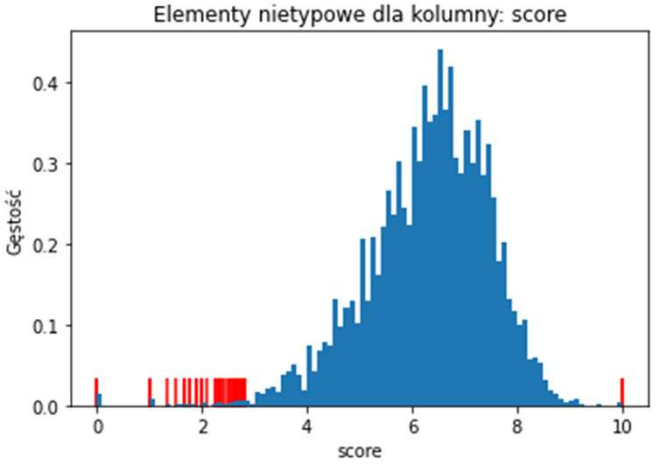


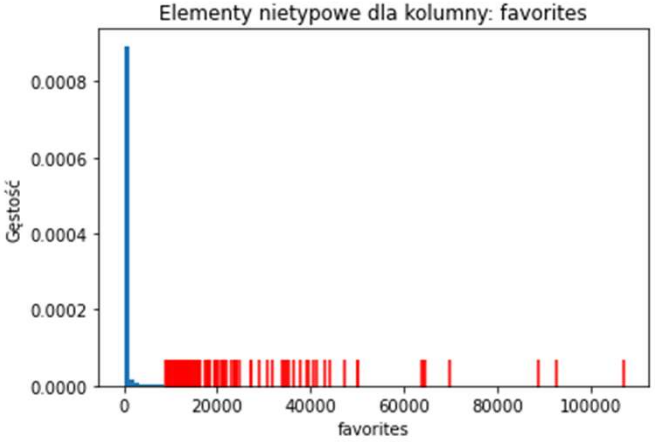
Rys. 2. Rozkład płci użytkowników

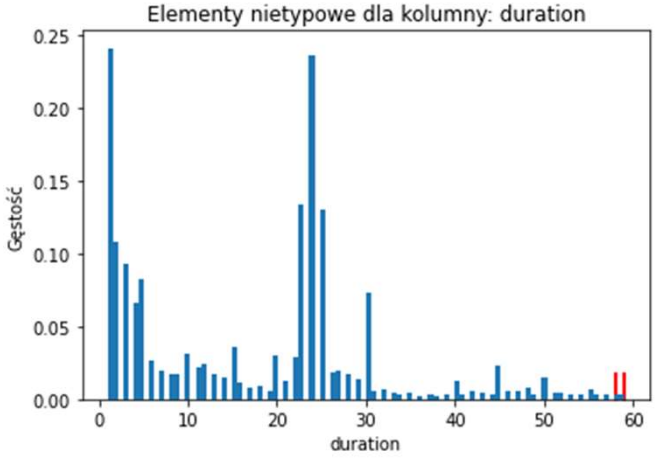
Elementy nietypowe - anime

W przypadku obydwu zbiorów wykorzystano metodę 3 sigm, gdzie w niektórych przypadkach w zbiorze dotyczącym użytkowników ilość sigm została ręcznie dobrana. Kolejno obrazki z rozkładem atrybutów i czerwonymi liniami wskazującymi elementy odchylone.

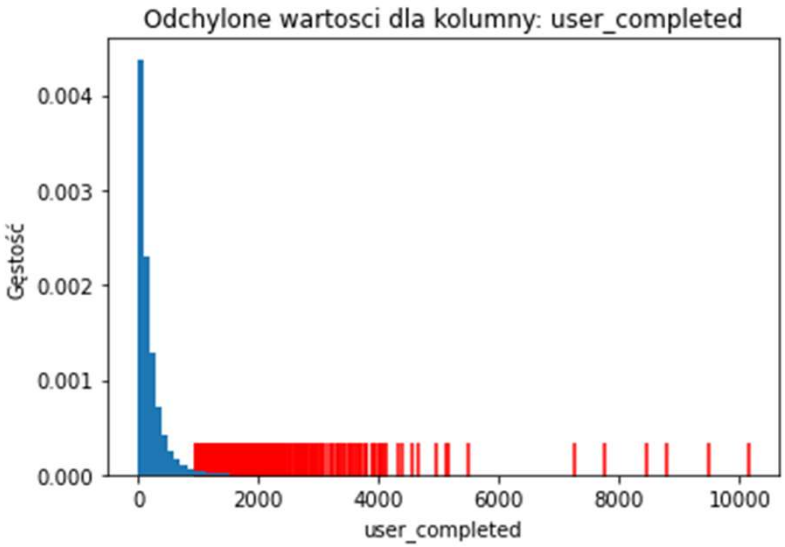


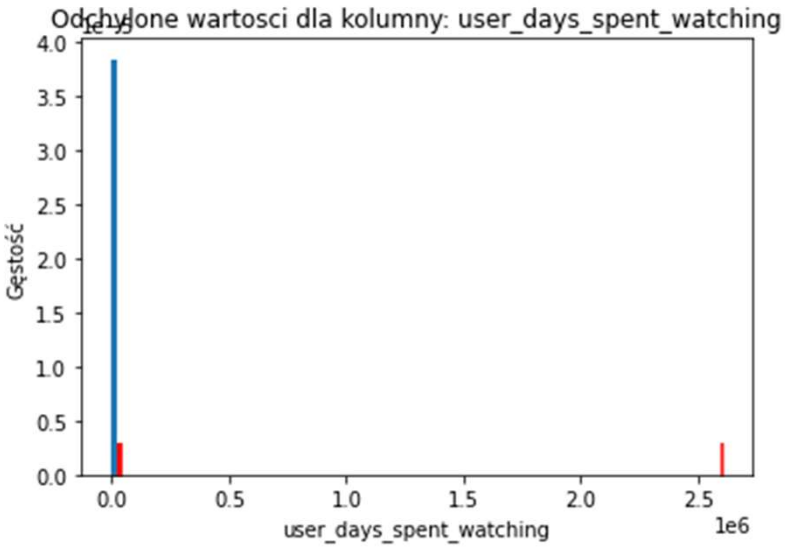


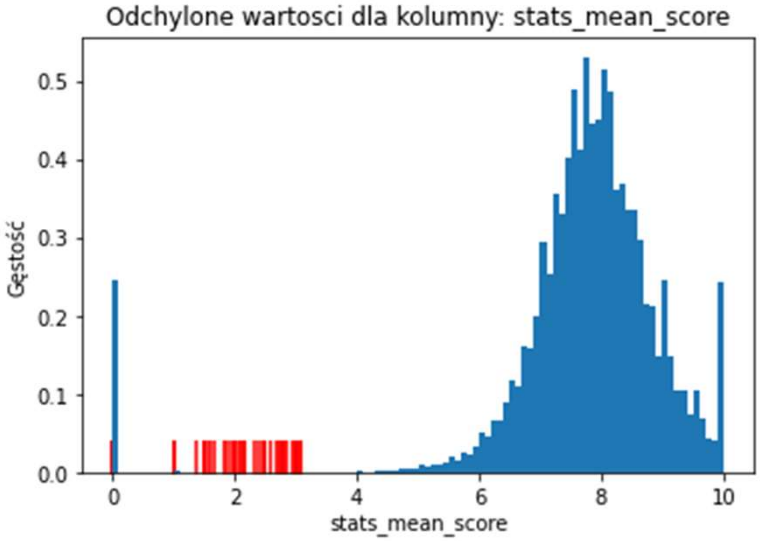


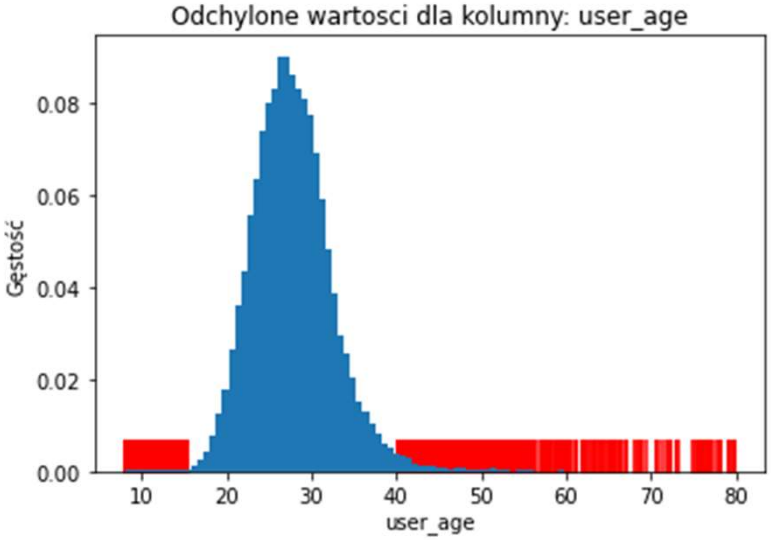


Elementy nietypowe - użytkownicy

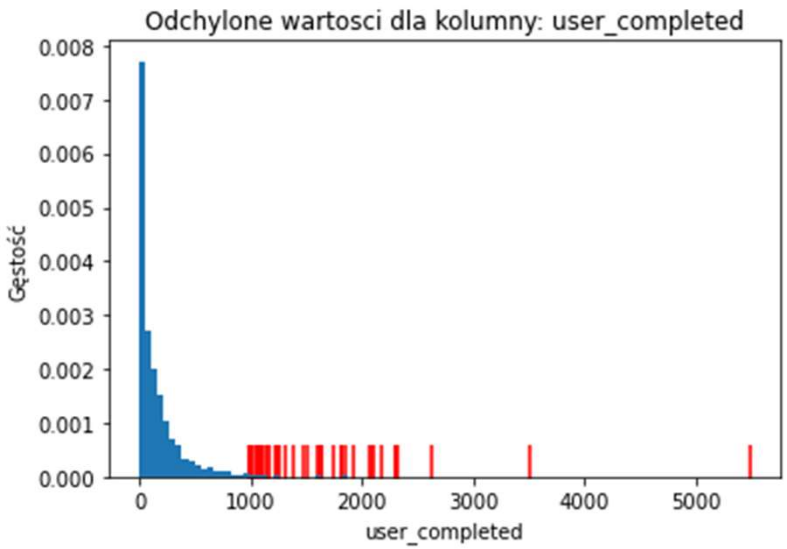


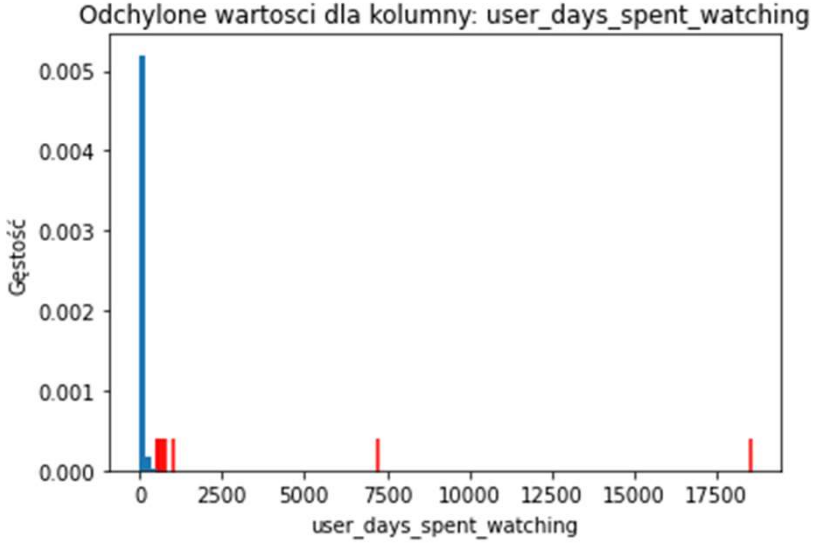


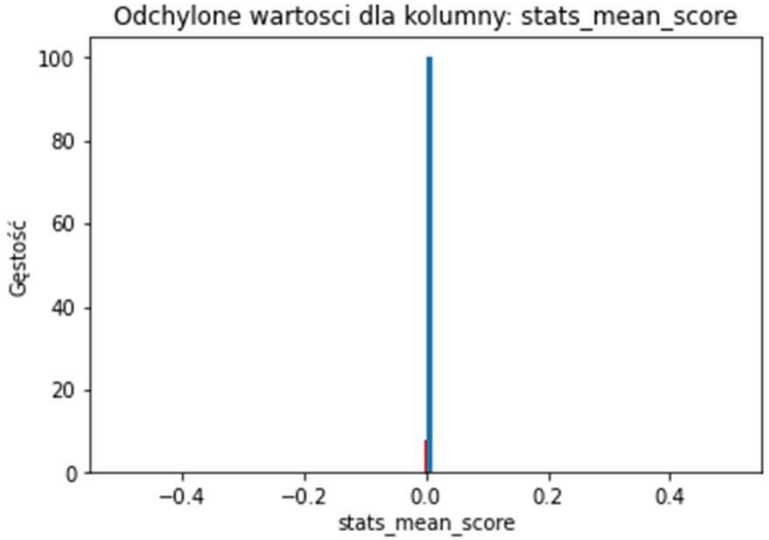


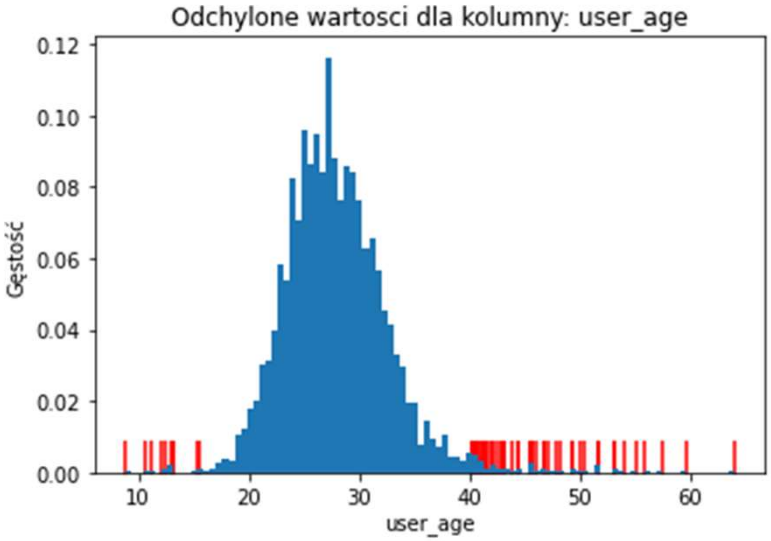


Elementy nietypowe – użytkownicy nie oceniający



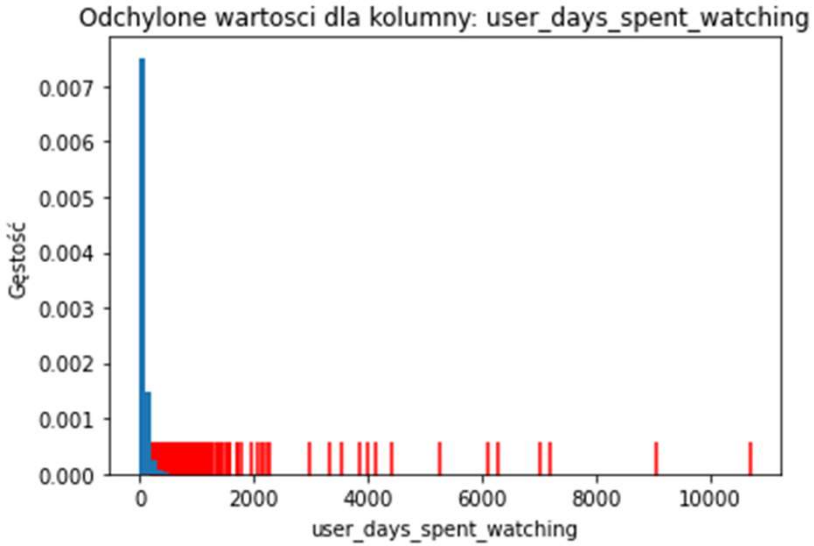


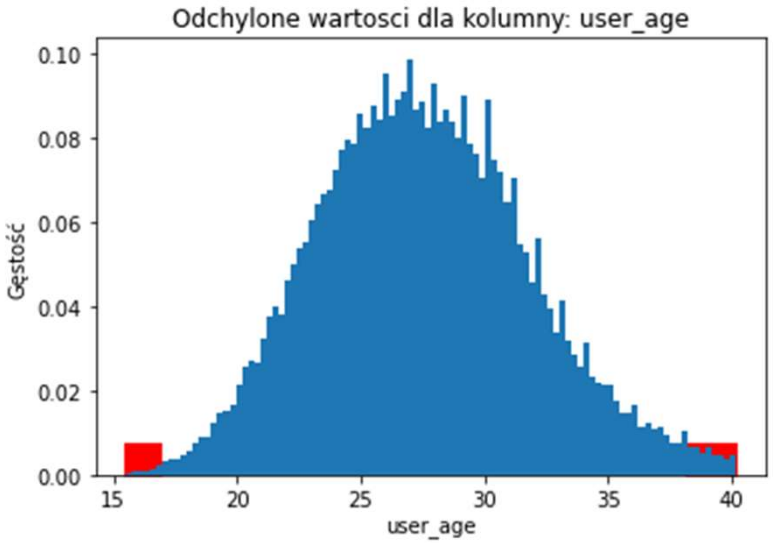




Usuwanie elementów nietypowych

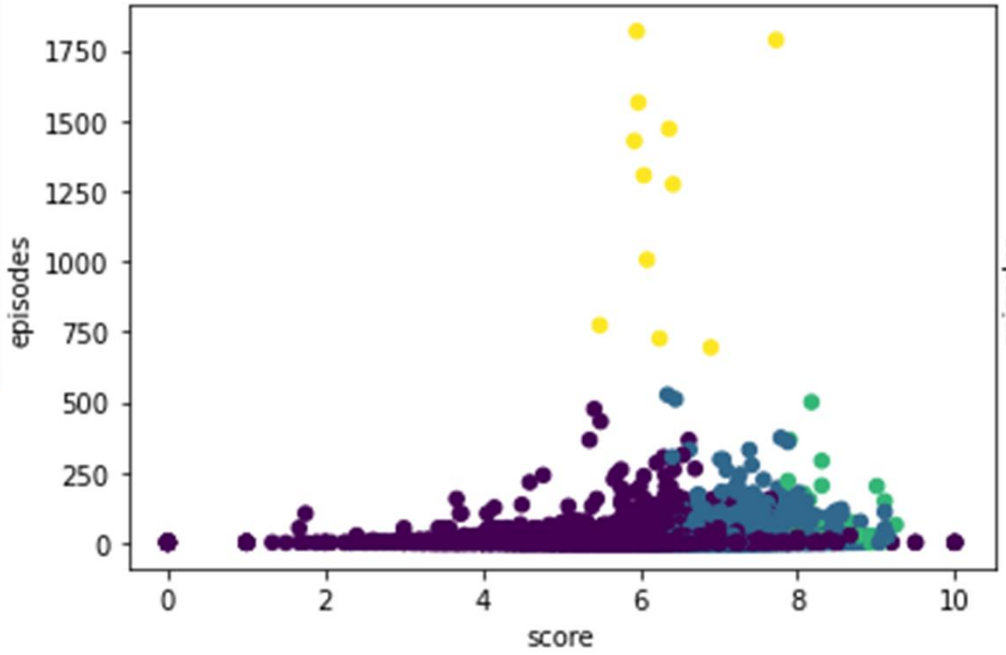
Ze zbioru anime nie były usuwane żadne elementy nietypowe, ponieważ elementy odstające w tym zbiorze wnoszą wiele informacji. Natomiast ze zbioru użytkowników usunięto takich, którzy oglądali (a przynajmniej tak byli oznaczeni) nierealistyczną ilość odcinków, mieli niepoprawne (mało prawdopodobne) ilości lat. Dodatkowo usuwanie ze zbioru użytkowników jest możliwe dzięki faktu, że ilość rekordów jest stosunkowo duża.

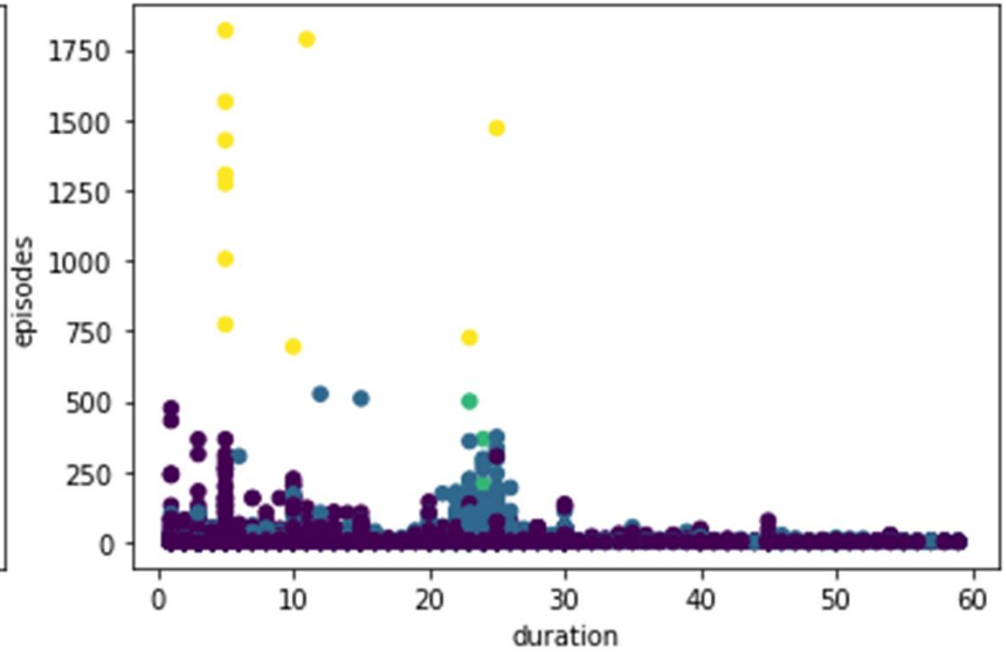




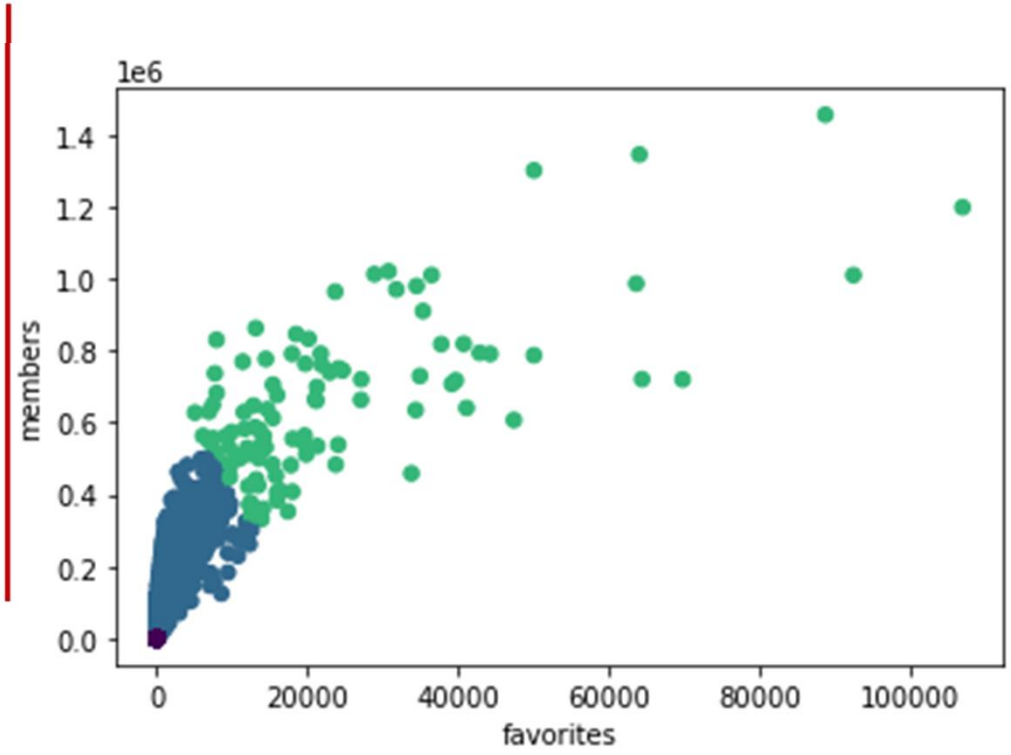
Klasteryzacja - anime

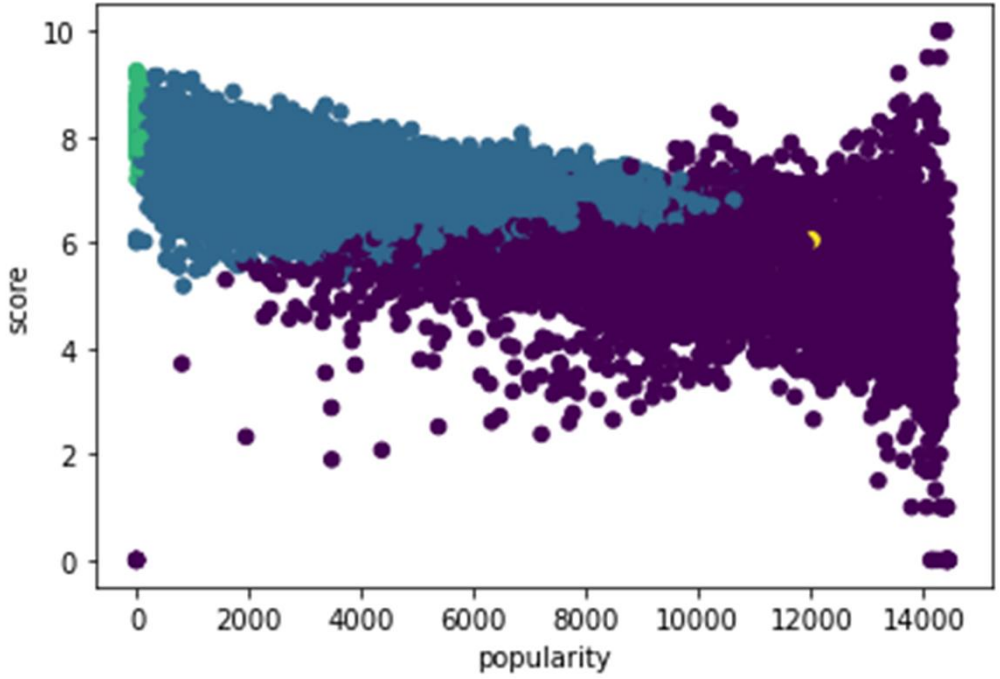
W przypadku obydwu zbiorów wykorzystano metodę KMeans do klasteryzacji. W przypadku tego zbioru, jak i tego o użytkownikach, ustalono liczbę klastrów na 4.



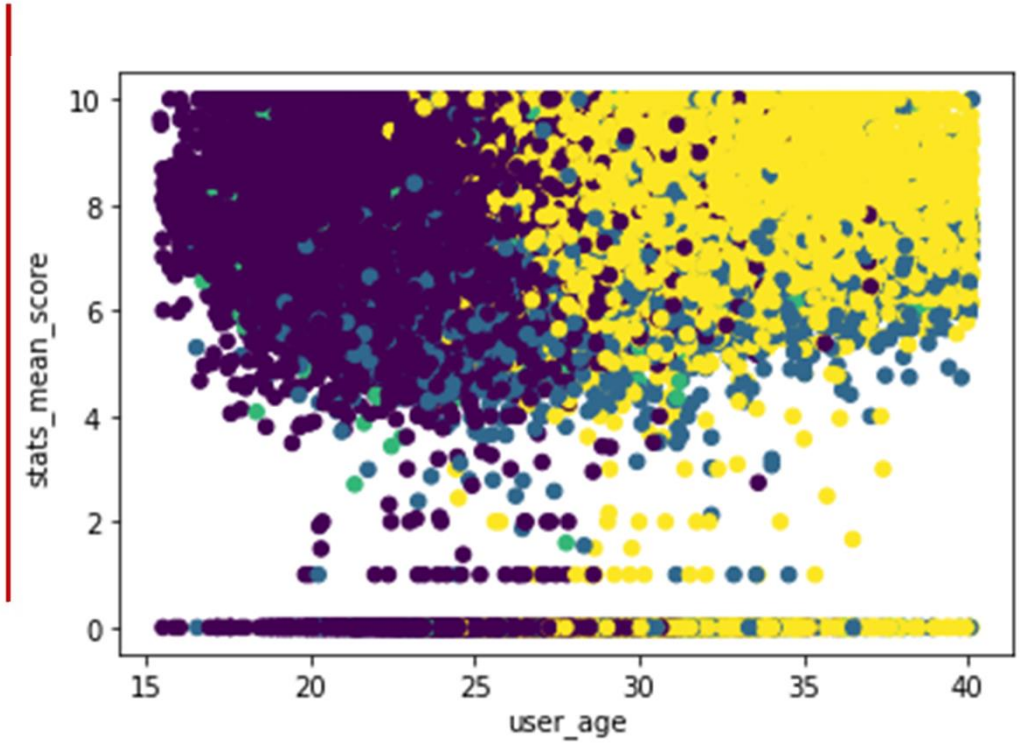


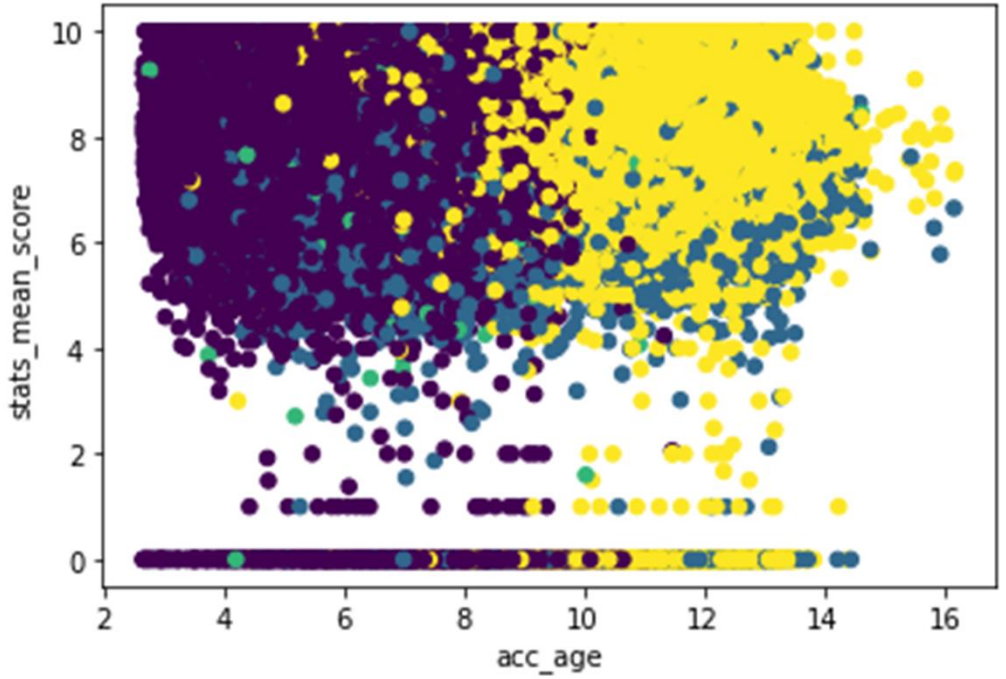
Klasteryzacja – anime c. d.



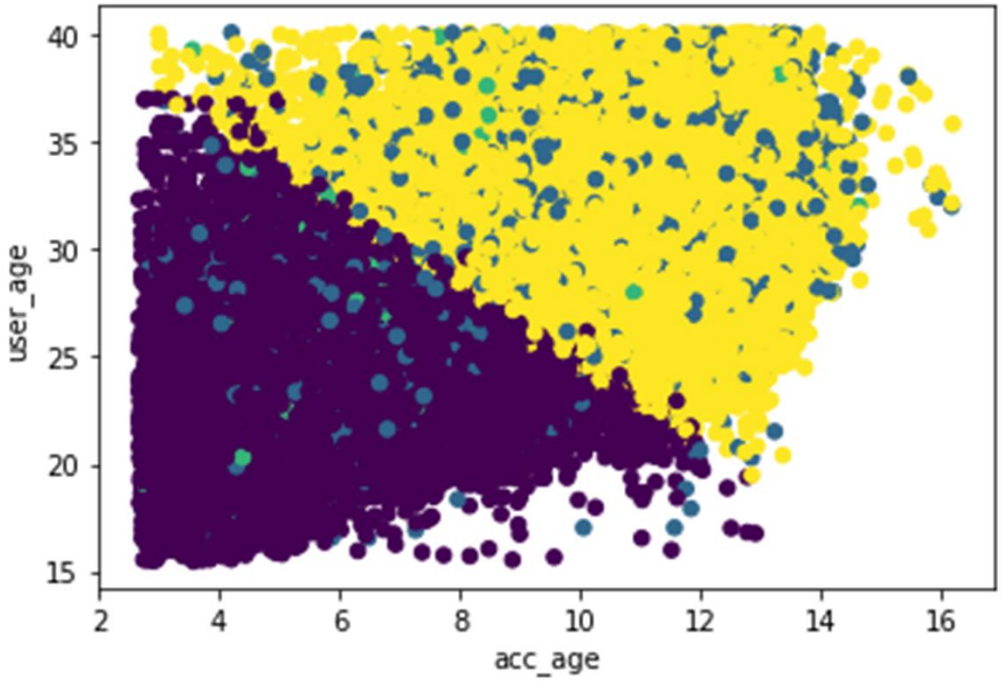


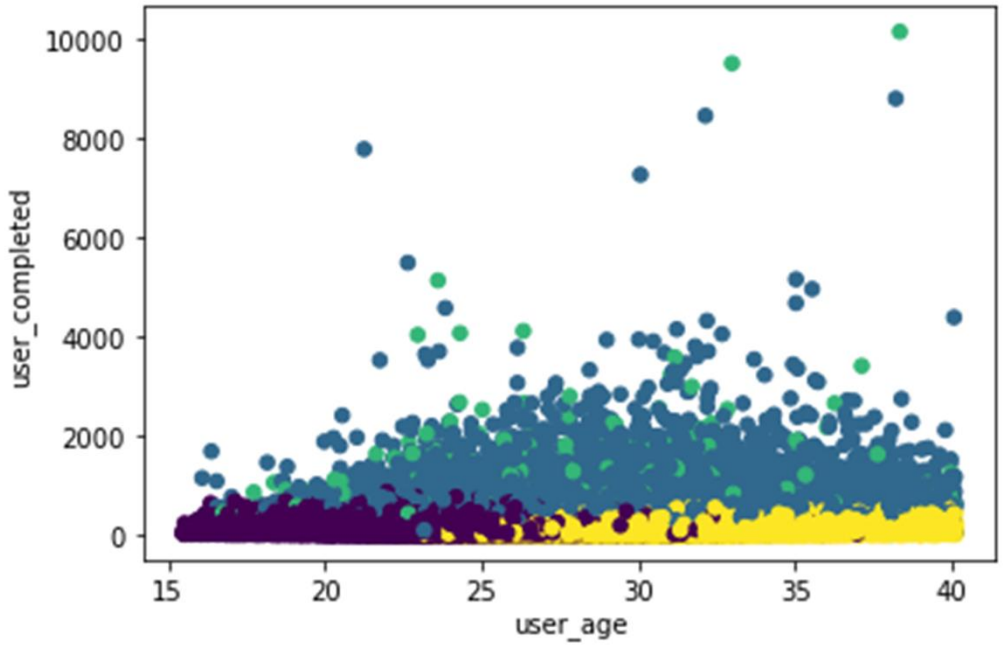
Klasteryzacja - użytkownicy



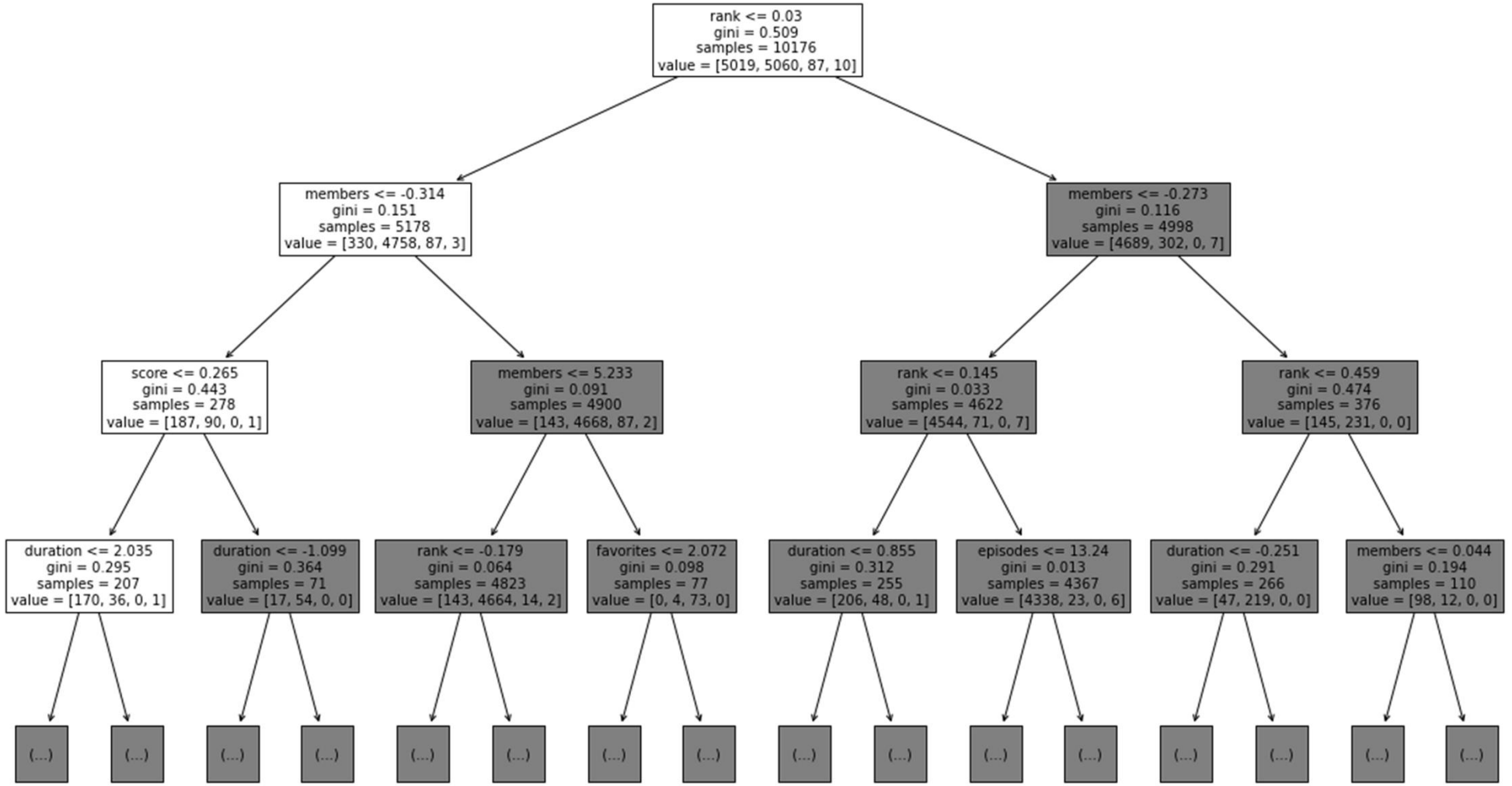


Klasteryzacja – użytkownicy c. d.

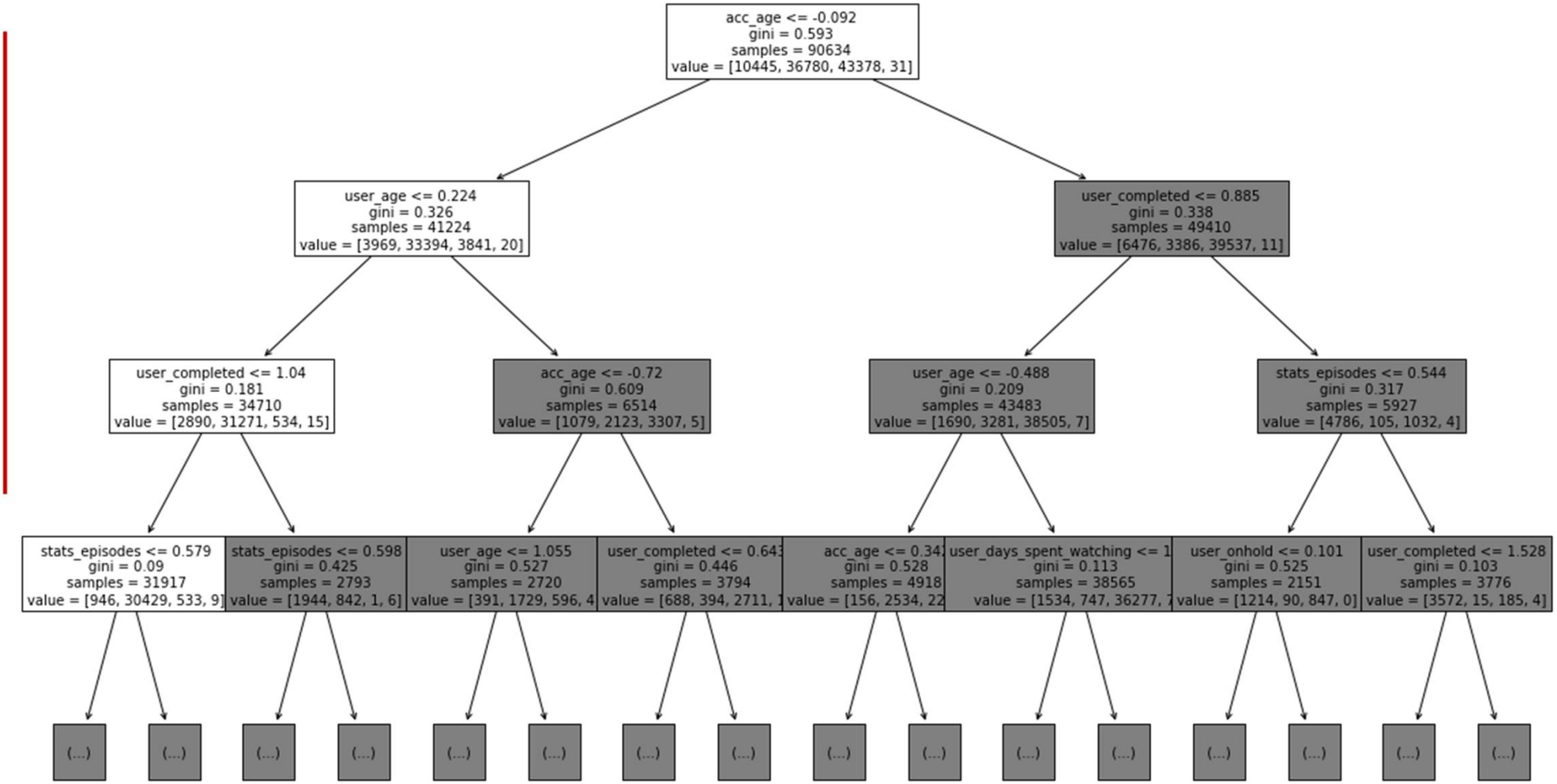




Klasteryzacja – wpływ atrybutów na podział anime

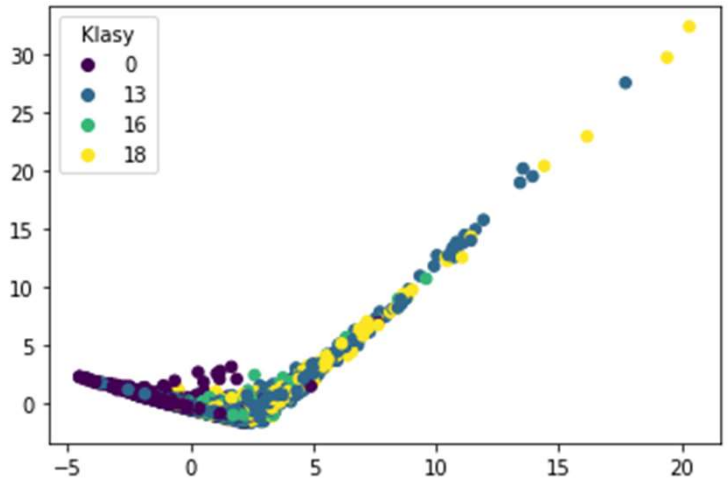


Klasteryzacja – wpływ atrybutów na podział użytkowników

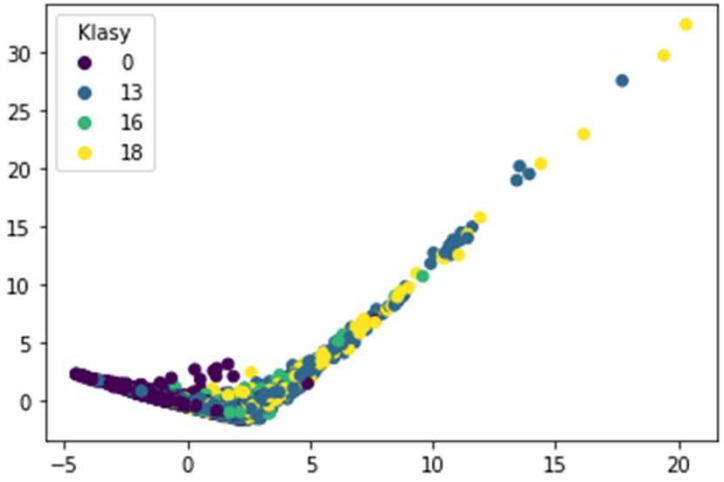


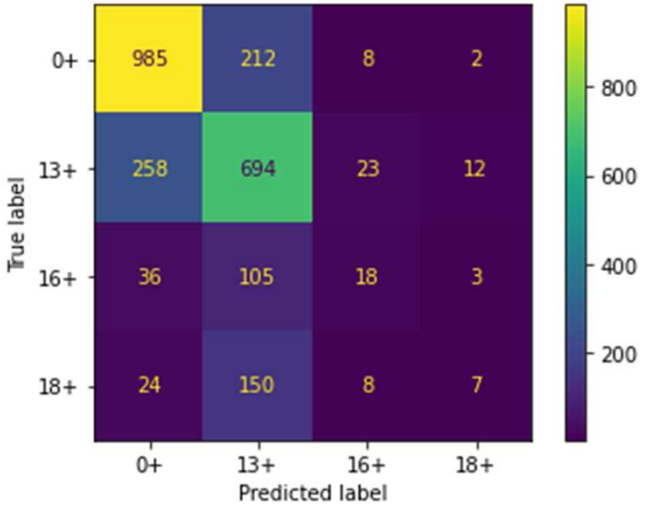
Klasyfikacja – anime ograniczenie wiekowe

Pełny zbiór

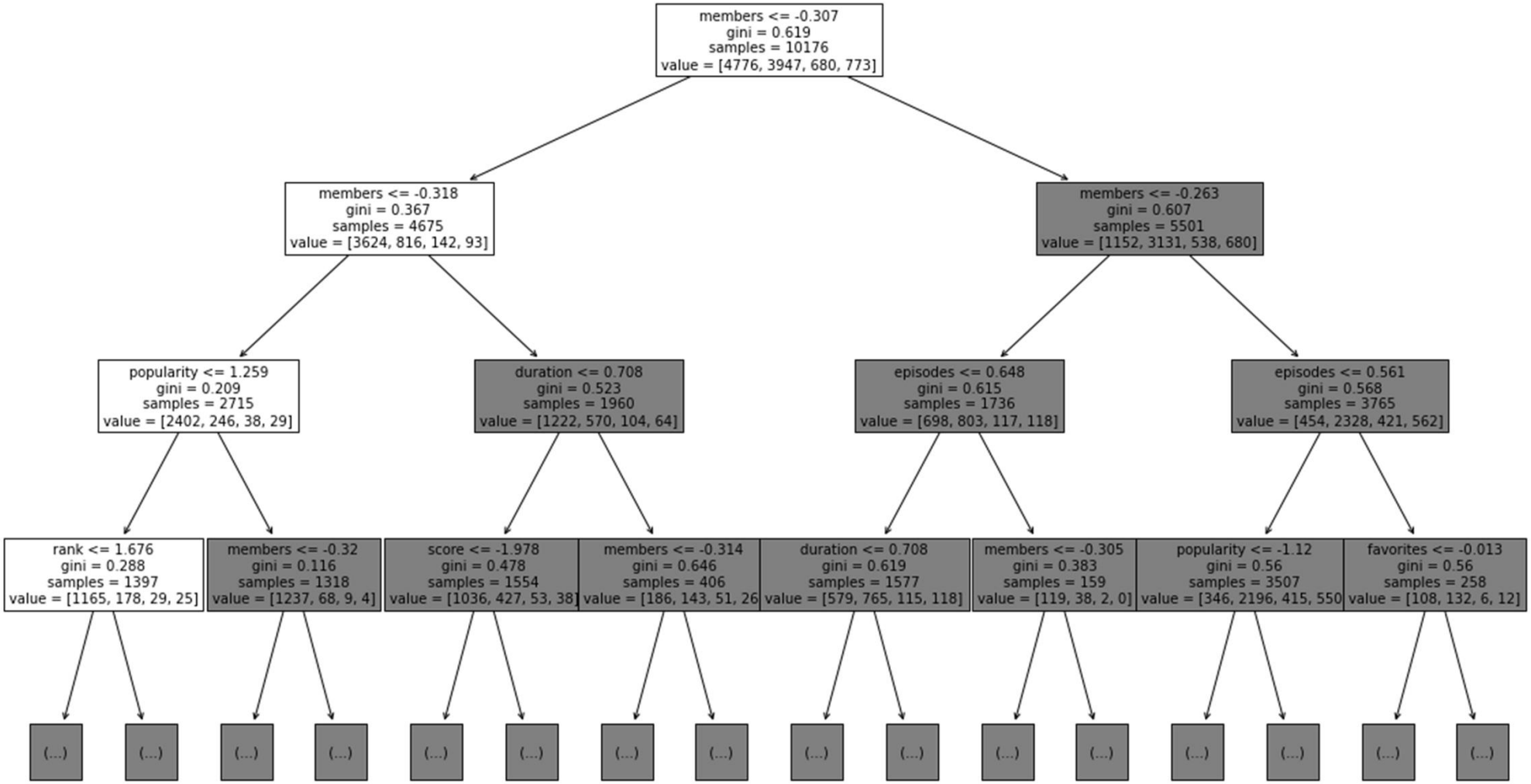


Drzewo decyzyjne - ~70%
dokładność (zbiór testowy ~67%)



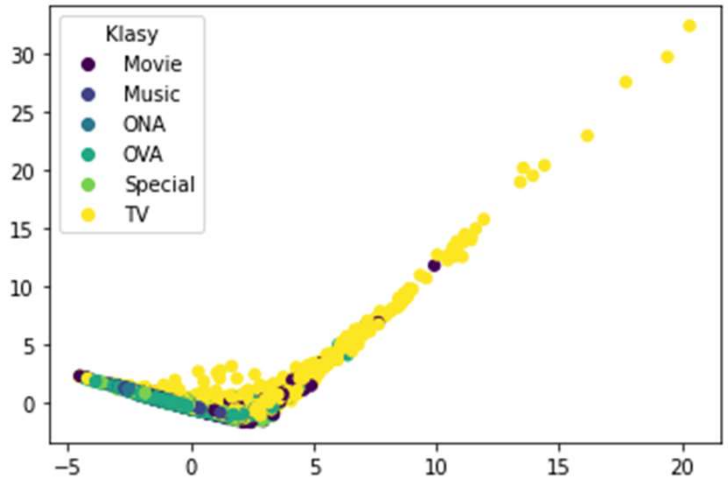


Klasyfikacja – anime ograniczenie wiekowe

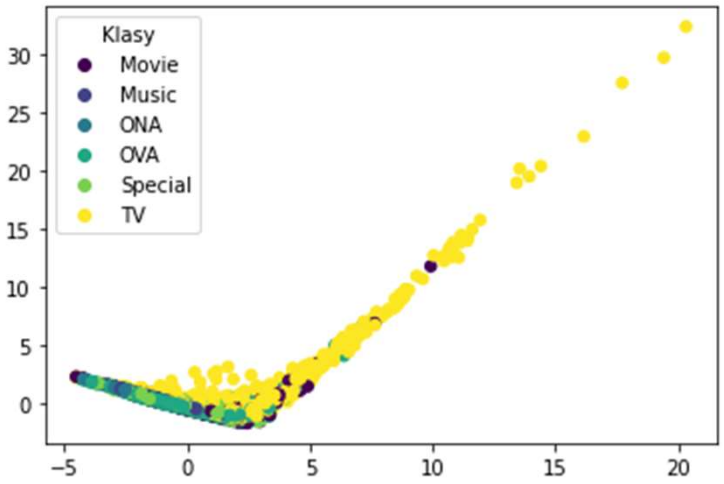


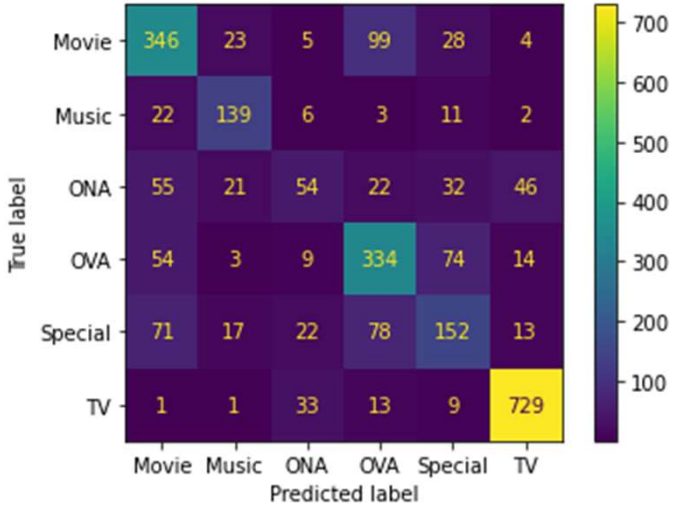
Klasyfikacja – anime typ

Pełny zbiór

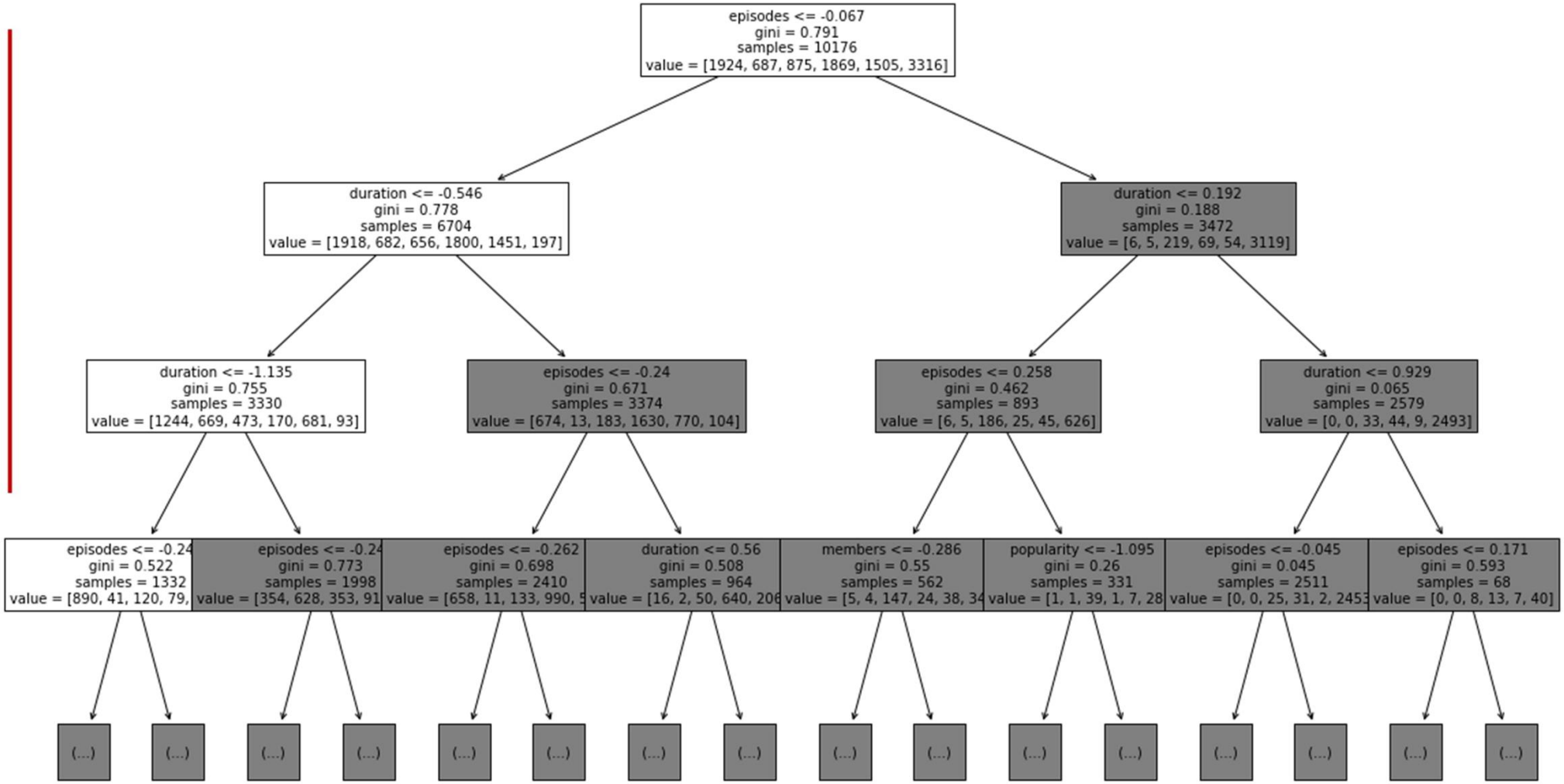


Drzewo decyzyjne - ~75%
dokładność (zbiór testowy ~69%)



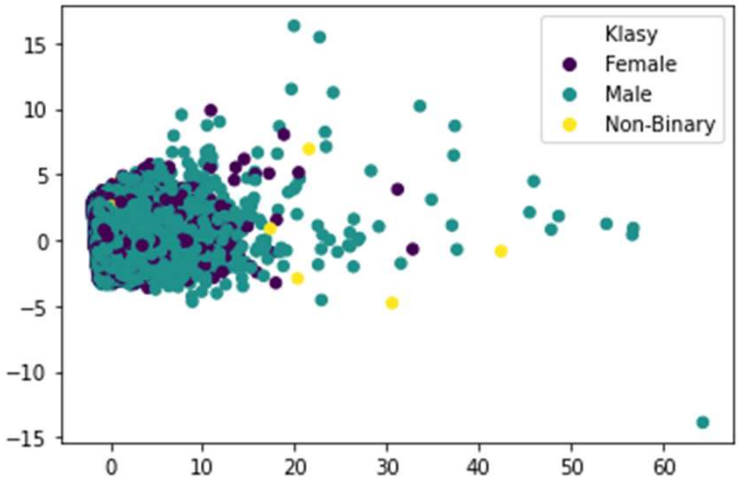


Klasyfikacja – anime typ

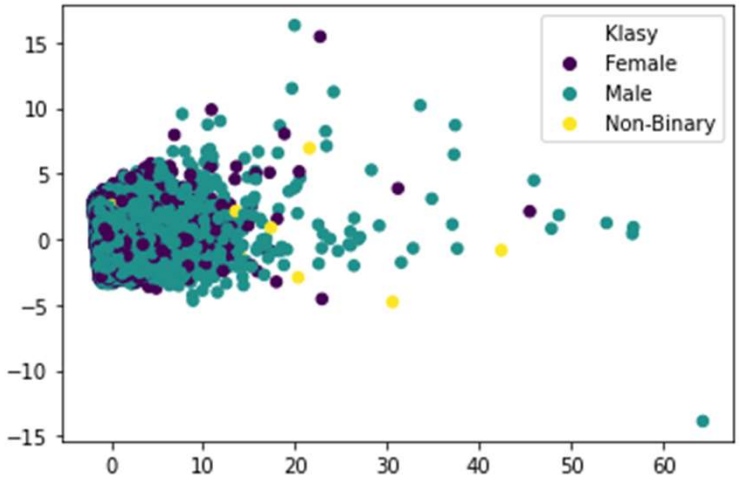


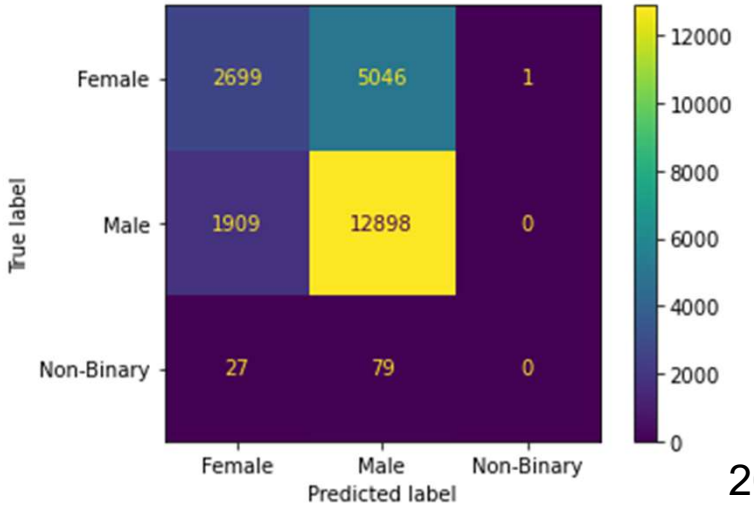
Klasyfikacja – użytkownicy płeć

Pełny zbiór

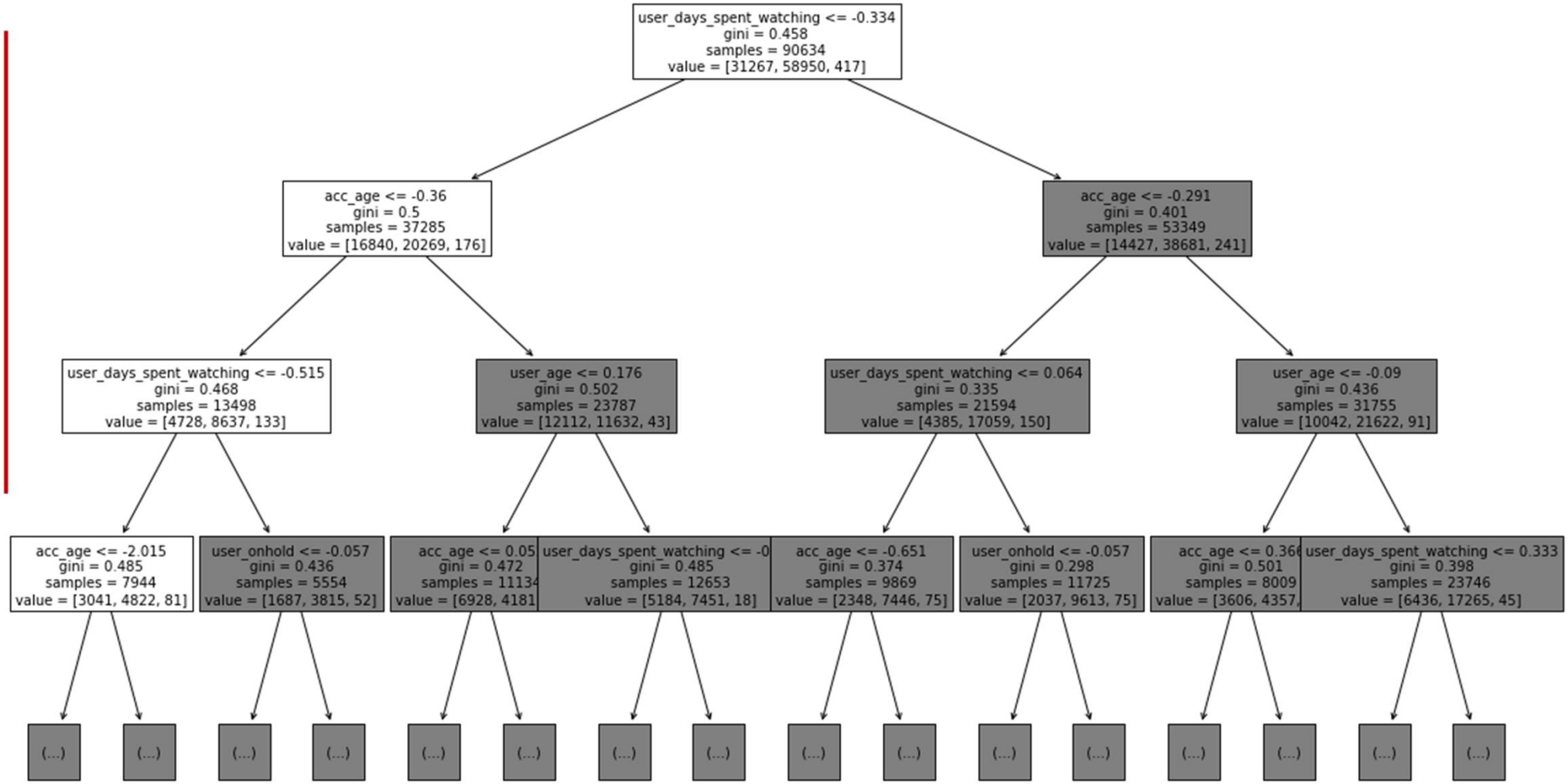


Drzewo decyzyjne - ~70%
dokładność (zbiór testowy ~68%)





Klasyfikacja – użytkownicy płeć



Podsumowanie

- Zbiór potwierdził nasze oczekiwania. Znane anime jest odpowiednio odwzorowane, a elementy danych użytkowników, które są wprowadzane bezpośrednio przez nich są czasami nieprawdziwe.
- Rozkład gęstości niektórych zmiennych, tj. wiek okazał się zaskakujący co do swojego kształtu i wartości jakie go opisują.
- Po usunięciu wartości odstających można było przyjąć opisywane zbiory jako dobrze oddające rzeczywistość.
- Przy pomocy klasteryzacji odkryto wyraźnie różne grupy zarówno anime jaki i użytkowników, definiowane przez mały podzbiór wszystkich atrybutów.
- Przewidywanie, nawet wielu etykiet, przy pomocy zbioru ma wysoką dozę bycia poprawnym, a niektóre metody są lepsze od innych.
- Można przeprowadzić dodatkowe analizy dla innych ilości klastrów, jako że w obydwu zbiorach widać niewielką liczność wybranych klastrów.

Dziękujemy za uwagę ☺