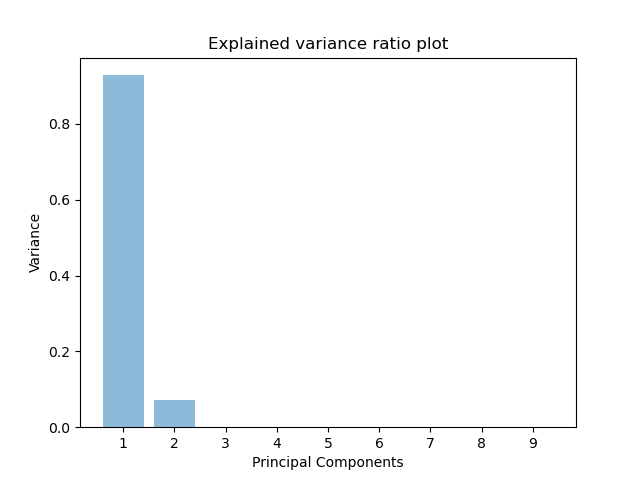
**Forrás**

Az adatbázis, amivel dolgoztam zenei műfajok adatait tartalmazza, amelyek a Spotify-ról lettek gyűjtve.

(Eredeti forrás: <https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks?select=data_by_genres_o.csv>)

Az attribútumok közül néhányat szándékosan nem olvasok be, csakis azokat, hagyom meg amik könnyen értelmezhetőek és beszédesek. (Például a műfajok nevei lényegtelenek.)

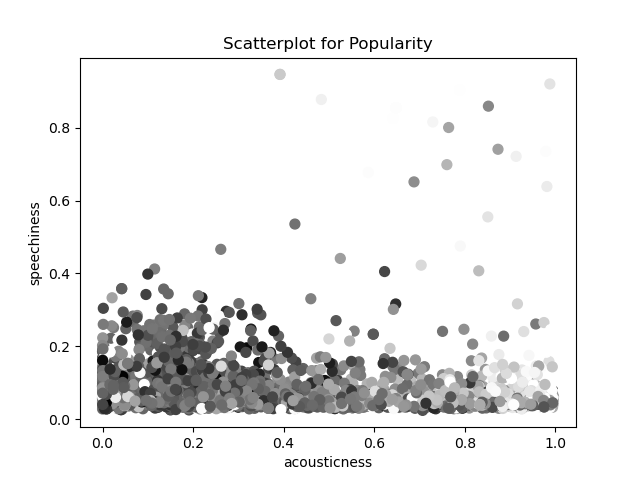
**PCA**

Az első ábra egy PCA illesztés, ez az egyes attribútumok relatív jellentőségét, súlyát ábrázolja.

(Látható, hogy eszerint a becslés szerint 2 adat már nagyrészt leírja az adatállományt.)

**SelectKBest**

A következő ábra 2 automatikusan kiválasztott legjelentősebb adat függvényeként ábrázolja a célértékünket. Célértéknek az adott műfaj népszerűségét választottam. (A sötétebb pontok jelzik a népszerűbb műfajokat.)



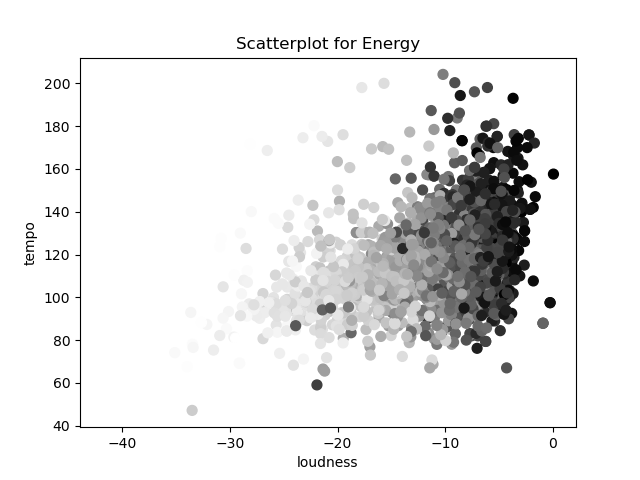
(Látható, hogy a program a beszédesség és az akusztikusság függvényében ábrázolta a népszerűséget.)

Az ábráról nagyjából az olvasható le, hogy a kevésbé akusztikus műfajok népszerűbbek. Illetve a beszédességnél ránézésre 0.2 körül vannak a legnépszerűbb műfajok.

**Egy másik scatterplot érdekességnek**

Mivel az automatikusan választott adatok között nem volt annyira látványos az összefüggés, érdekességképp manuálisan választottam 3 adatot, amely látványosabb ábrát eredményez.

(Az energikusság-ot ábrázoltam a hangerő és a tempó függvényeként.



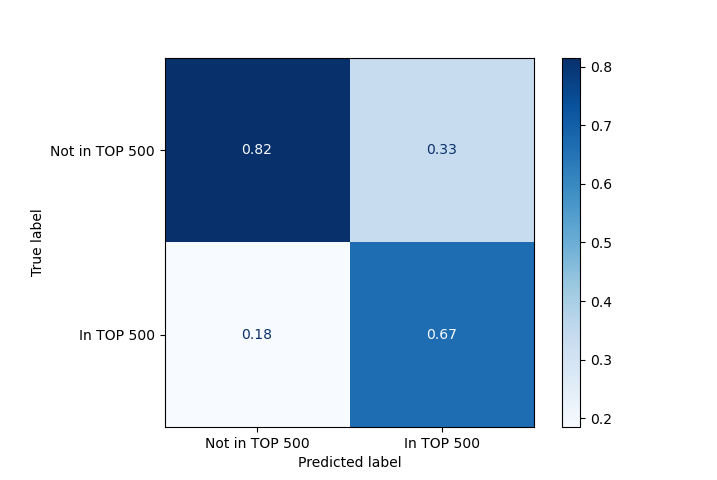
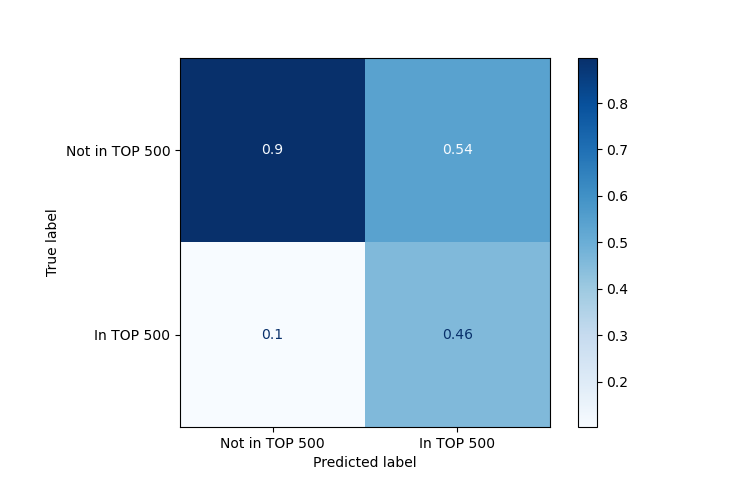
Az eredmény eléggé látványos és beszédes. A legfeltűnőbb jelenség az, hogy az adatok egy háromszöget alkotnak. Ez elsősorban a bal felső sarok üressége miatt érdekes, ami azt jelzi, hogy egy műfaj szinte sosem lehet egyszerre gyors és halk.

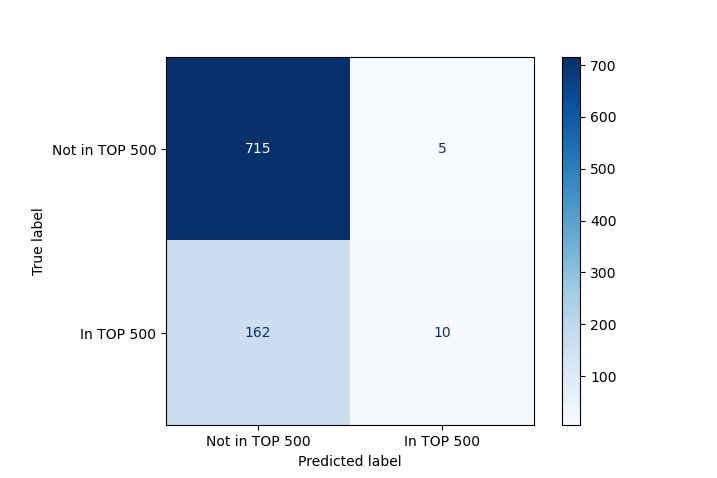
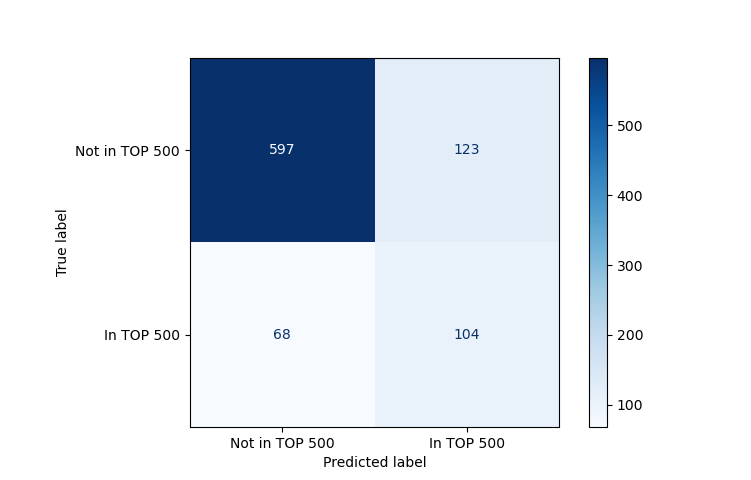
Ha a célértéket, azaz az energikusság-ot nézzük, azt látjuk, hogy minél hangosabb egy műfaj, jellemzően annál energikusabb.

(Engem viszont például meglepett, hogy a tempó növelése ránézésre nem teszi jelentősen energikusabbá a műfajokat.)

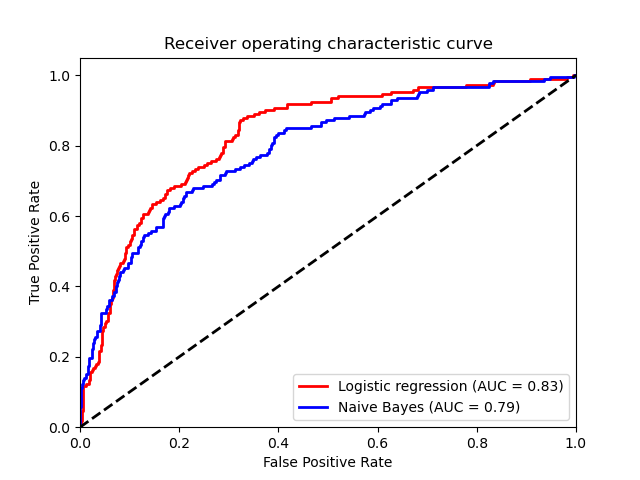
**Predikciók**

A következő lépésben megpróbáltam kitalálni, hogy egy műfaj benne van-e az 500 legnépszerűbben. A bal oldalon látható a logisztikus regresszió tévesztési mátrixa. A jobboldalon pedig a naiv Bayes tévesztési mátrixa. (A felső ábrákon számszerűsítve, az alsókon prediktált érték szerint százalékolva.)

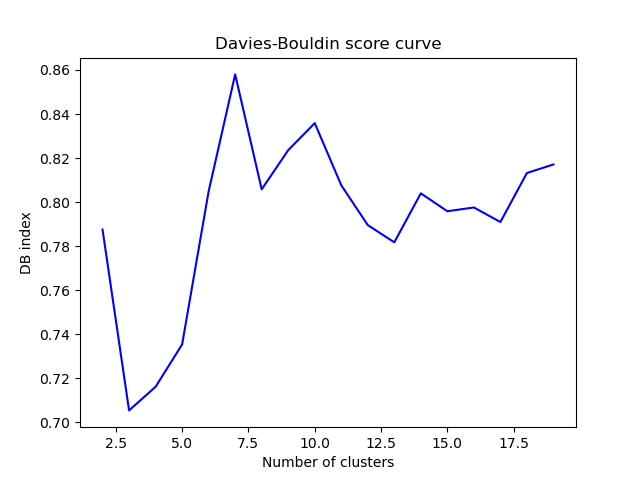
**Logisztikus Naiv Bayes**

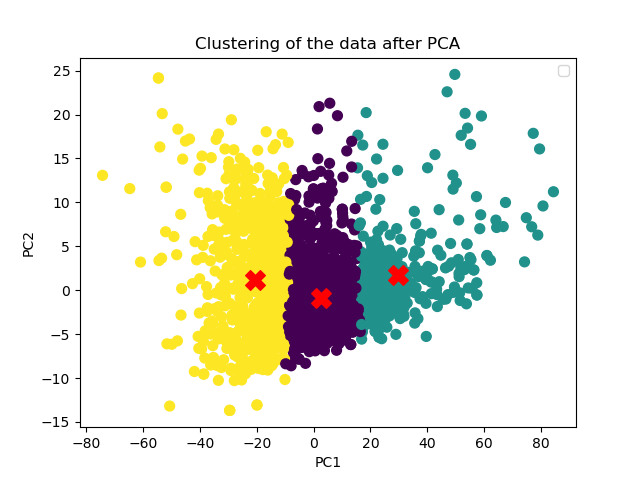
Nagyjából az olvasható le, hogy a logisztikus regresszió szigorúbb, azaz csak nagyon ritkán mondja, hogy egy műfaj benne lesz a TOP 500-ban, de ilyenkor közel 70%-os eséllyel igaza van. Ehhez képest a Bayes gyakrabban mondja, hogy benne lesz, de még ilyenkor is csak 45%-os eséllyel lesz tényleg benne.

Azt viszont, hogy nem lesz benne, jobban jósolja a Bayes 90%-os pontossággal (míg a logisztikus csak 80%-ossal).

A 2 predikciós eljárást ROC görbék alapján is összehasonlítom, ahol a logisztikus regresszió tűnik megbízhatóbbnak. Ez nem meglepő, mivel már a mátrixokból leolvasható, hogy a pozitív értéket, azaz, hogy benne van a TOP 500-ban, a logisztikus regresszió megbízhatóbban jósolta. 

**Klaszterezés**

Az adatokat csoportokba (klaszterekbe) sorolom, ehhez először is keresek egy ideális klaszterszámot Davies Bouldin pontozás segítségével.

Az ábra szerint az ideális klaszterszám a 3 lesz. Tehát 3 klaszterbe sorolom a műfajokat. Itt megint automatikusan választott paraméterek alapján fogunk dolgozni.

Az adatokat jól láthatóan probléma nélkül be tudta sorolni 3 osztályba.