TALLINNA TEHNIKAÜLIKOOL

Infotehnoloogia teaduskond

Hanna Kätlin Ardel, Kristina Jumankina, Ken Kruuser

SPOTIFY LUGUDE KLASTERDAMINE JA POPULAARSUSE PROGNOOSIMINE

Statistilised meetodid masinõppes

Tallinn 2023

SISUKORD

1. UURIMISPROBLEEM

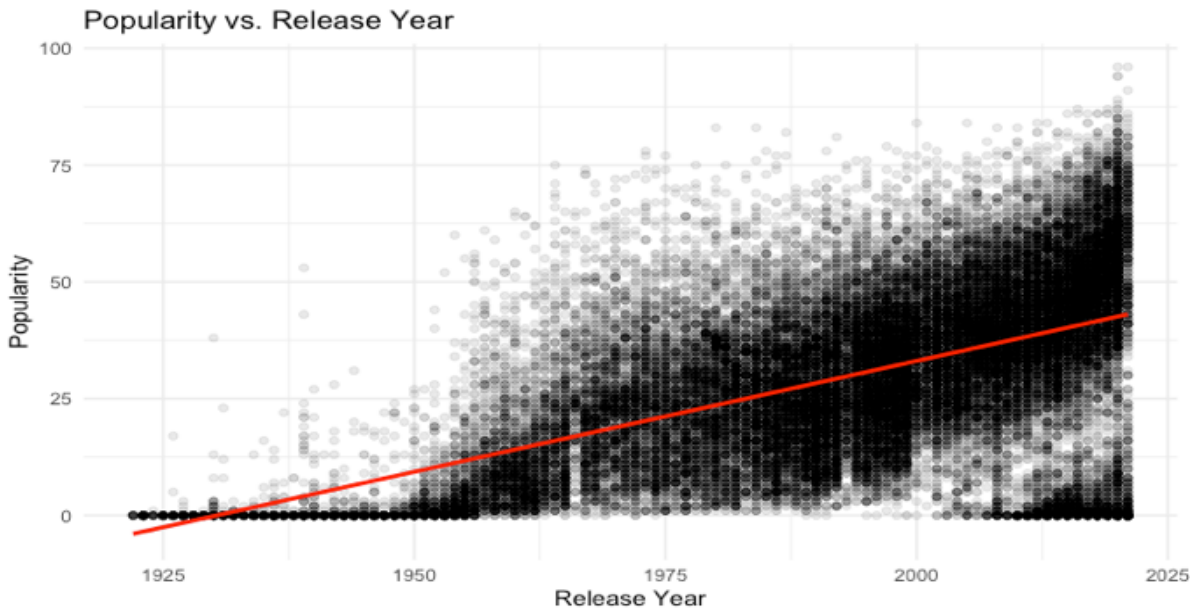
Muusika on kogu inimkonna ajaloo vältel olnud meie kultuuri lahutamatu osa. Küll aga on muusika tarbimine tänu digitaalplatormide arengule muutunud inimestele lihtsaks ja kättesaadavaks, mis on viinud platvormide kasutuse suurtesse massidesse ja toonud muusika kuulamise erinevate igapäevaste tegevuste juurde. Muusikatööstusele kui ärile on seejuures kriitilise tähtsusega oskus liigitada muusikat žanritesse ja mõista mustreid ning tendentse, mis muudavad loo populaarseks ning suurendavad seetõttu kasutajate muusika tarbimist ja ka platvormide kasutamise aega. Selle mõistmine on oluline ka teisele platvormiga seotud olevatele osapooltele, sealhulgas artistidele, produtsentidele, plaadifirmadele ja striimimisteenustele. Lugude populaarsuse hindamine ja analüüs on võimalik tänu platvormide kogutud andmetele muusikalugude ja artistide kohta. Nende andmete põhjal on omakorda võimalik läbi erinevate metoodikate ja trendide uurimise kaudu võimalik jõuda äriliselt vajalike järeldusteni.

Varasemad uuringud on uurinud psühholoogia, kultuuri, identiteedi ja isiklike veendumuste mõju muusikaeelistustele. Indiviiditasandil seostas Greenberg (*et al.* 2015) muusikažanri eelistuse isiku kognitiivsete võimetega, sh isiku empaatia ja süstemaatilisus. Andrews (*et al*. 2020) kinnitasid isiksuse, kultuuri ja uskumuste mõju muusikaeelistustele. Väites, et inimese karakter ei ole ainus ega ka mitte peamine muusikaeelistuse mõjutaja, kuid selle asemel on olulisteks mõjutajateks isiklikud ja kultuurilised väärtused koos teiste teguritega. (Petitbon & Hitchcock 2022, 169) Lisaks andsid North ja Hargreaves tõestust selle kohta, et indiviidi poliitilised ja sotsiaalsed veendumused mõjutavad muusikalisi eelistusi (Andrews *et al.* 2020).

Muusikaeelistuste mõistmine võimaldab plaadifirmadel ja teistel muusika loomise ja müümisega seonduvatel ettevõttetel panustada muusikatööstuse ja muusikaturu arengusse, selliselt, et tagatud on suur turg ehk palju kasutajaid ning kuulajaid. Samuti on üheks eesmärgiks seejuures tagada enda ettevõtte majanduslik kasumlikkus. Muusikaeelistuste mõistmisel on üheks oluliseks teadmiseks ka see, et erinevatele ajastutele on omased erinevad muusikastiilid (Petitbon & Hitchcock 2022, 169).

Striimimisteenuste pakkujatele nagu Spotify võib sellise sisuga uuringu tulemuste kasutamine parandada soovitusalgoritme, sisu omandamise strateegiaid ja kasutajate kaasatust. Artistid ja plaadifirmad saavad ülevaate praegustest trendidest ja eelistustest, mis aitavad juhtida nende

loomingulisi ja turundustegevusi. Akadeemilisest vaatepunktist annab uuring panuse muusikaanalüütika valdkonda, ühendades žanri klassifitseerimise populaarsuse analüüsiga.



Joonis 1. Lugude populaarsuse ja nende väljaandmise aasta vaheline seos. Autorite koostatud programmis RStudio.

Muusikaeelistus on oma loomulikult ebatäpne kontseptsioon ja eeldefineeritud muusikažanrid kalduvad piirama muusikaeelistuse paindlikkust. Žanrite määramine on võimalus muusika ja artistide klassifitseerimiseks subjektiivse mõõdikuga selle järgi, kuidas teatud muusikamustrid kõlavad, eesmärgiga grupeerida sarnaseid helimustreid, instrumente ja rütme omavahel. Olemas on küll üldine konsensus, millised laulud kuuluvad teatud žanritesse, võivad siiski mõned lood kuuluda mitmesse žanrisse ja žanrite endi tajumine võib erineda sõltuvalt kuulajast. Küll aga leidub juba ka masinõppel põhinevaid uuringuid, mis muudaksid žanrite määramise võimalusel objektiivsemaks. Näiteks on tehtud uuring, mis pakkus välja närvivõrgu algoritmi muusikažanrite klassifitseerimiseks heliliste omaduste põhjal. (Petitbon & Hitchcock 2022, 169)

Uurimisküsimused

Uurimisküsimus 1. Muusikažanrite klassifikatsioon: kuidas saab kategoriseerida muusikapalad erinevatesse žanritesse, lähtudes lugude omadustest?

Uurimisküsimus 2. Faktorid, mis mõjutavad loo populaarsust: millised on peamised faktorid, mis mõjutavad loo populaarsuse indeksit?

Uurimisküsimus (1) on avastuslik (*exploratory*) ja seetõttu kasutatakse klasteranalüüsi metoodikat. Püstitatud uurimisküsimus seostub avastuslikkusega (*exploratory*), kuna kasutatavad meetodid pakuvad selgeid ja tõlgendatavad tulemusi, mis aitavad mõista muusikažanrite omadusi ja kuuluvust vastavalt lugude omadustele. Uurimisküsimus (2) on prognoosiv (*prediction*) ja seetõttu kasutatakse juhusliku metsa (*random forest*) metoodikat. Juhusliku metsa meetodi kasutamine võimaldab prognoosida loo populaarsust ning seeläbi hinnata, millised tegurid mõjutavad enim populaarsuse indeksit. Kasutatavad meetodid on juhendamata masinõppe meetodid.

2. ANDMED

Antud töö käsitleb Spotify laulude andmestiku, mis on pärit *Kaggle* platvormilt ning need on kogutud *Spotify Web API* abil (Spotify Dataset … 2023). Andmestik võimaldab vastata uurimisküsimustele eelkõige seetõttu, et sisaldab palju erinevaid muutujaid, mis kirjeldavad erinevate tunnuste põhjal laulude omadusi, samuti sisaldab andmestik laulude populaarsuse indeksit.

Andmestik sisaldab ülevaadet lauludest, mis on Spotify platvormil saadaval olnud 2021 aastal. Vanim laul andmestikus on seejuures 1921 aastast. Andmestik koosneb 586 672 reast ja 20 veerust. Iga muutuja kirjeldus:

- ***Id*** - loo unikaalne identifitseerimisnumber;
- ***Name*** - muusikapala nimi;
- ***Popularity*** - kirjeldab muusikapala populaarsust. Lugude populaarsus varieerub andmestikus palju, keskmine populaarsuse skoor on 27 (skaalal 0-100), mõned lood on saavutanud maksimaalse populaarsuse 100;
- ***Duration_ms*** - lugude kestus mõõdetuna millisekundites. Näitab laia vahemikku, keskmiselt on muusikapala kestusega umbes 214 893 millisekundit (või umbes 3.58 minutit). Andmestiku pikim lugu on üle 5.6 miljoni millisekundi (üle 93 minuti);
- ***Explicit*** - näitab, kas loo sisu on eksplitsiitne, enamasti on väärtused 0 (mitte-eksplitsiitne), väike osa lugudest on eksplitsiitsed (väärtusega 1);
- ***Artists*** - iga looga seotud artistide nimed;
- ***Id_artists*** - artistidele vastavad ID-d;
- ***Release Date*** - andmekogumik sisaldab iga loo avaldamise kuupäeva, vanim lugu avaldati aastal 1900, uusim lugu avaldati aastal 2021.

Andmestik sisaldab veel mitmeid omadusi, mis kirjeldavad lugude muusikalisi omadusi:

- ***Danceability, Energy*** **ja** ***Valence*** on numbrilised mõõdikud, mis varieeruvad vahemikus 0 kuni 1 ja mille keskmised väärtused on vastavalt 0.577, 0.549 ja 0.564. Need omadused näitavad sobivust tantsimiseks, intensiivsust ja üldist positiivsust, mida lugu edasi annab;

- ***Key, Loudness* ja *Mode*** - andmekogumik sisaldab muusikalist võtit (vahemikus 0 kuni 11), valjusus (mõõdetuna detsibellides) ja helilaad (mažoor või minoor) – andes teavet muusikalise tooni kohta;
- ***Speechiness, Acousticness* ja *Instrumentalness*** on samuti mõõdetud pakkudes teavet kõnelejate hulga, akustilise olemuse ja lugude instrumentaalsuse kohta;
- ***Liveness*** - näitab elava publiku olemasolu salvestuses;
- ***Tempo* ja *Time Signature*** - andmekogumik sisaldab lugude tempot (lööke minutis) ja nende taktimõõte, mis on loo rütmi põhilised aspektid.

Tabelis 1 on välja toodud andmestiku muutujate kohta kirjeldav statistiline tabel. Tabelis on välja toodud iga tunnuse miinimum ja maksimum väärtus, esimene kvartiil, mediaan, keskmine ja kolmas kvartiil.

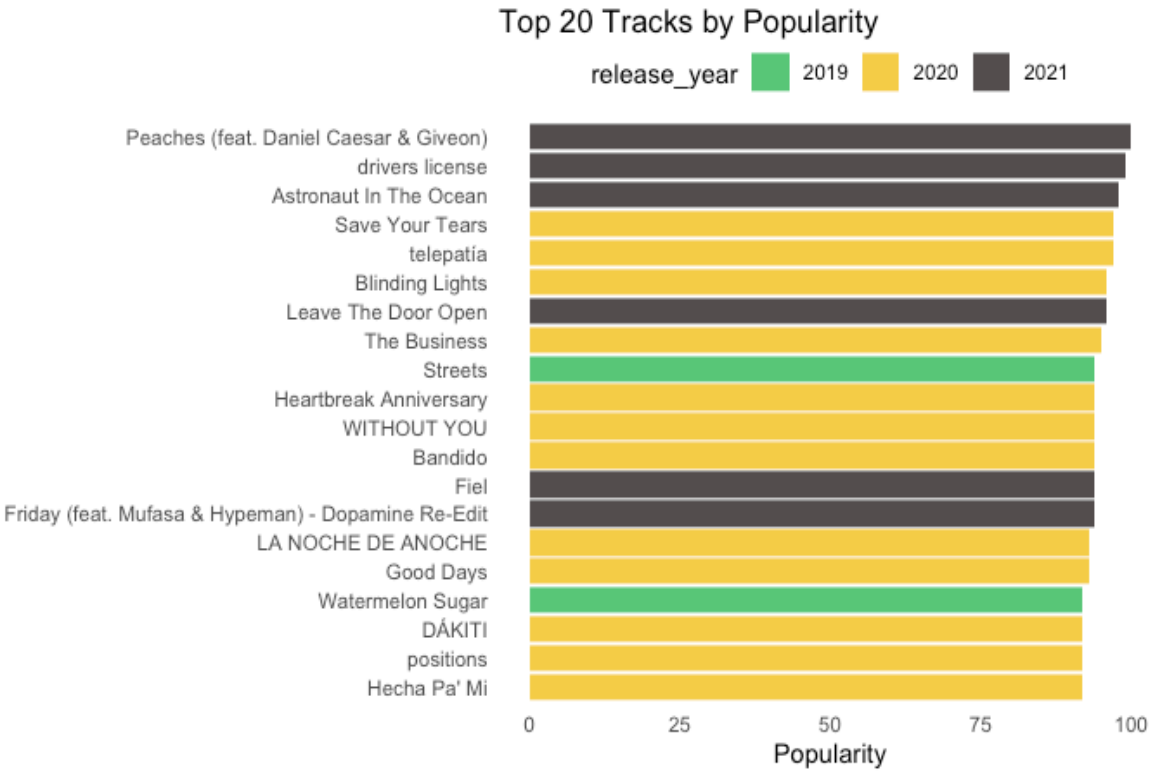
```
     id                 name              popularity      duration_ms
Length:586672      Length:586672      Min.   :  0.00   Min.   :   3344
Class :character   Class :character   1st Qu.: 13.00   1st Qu.: 175093
Mode  :character   Mode  :character   Median : 27.00   Median : 214893
                                      Mean   : 27.57   Mean   : 230051
                                      3rd Qu.: 41.00   3rd Qu.: 263867
                                      Max.   :100.00   Max.   :5621218
   explicit         artists           id_artists        release_date
Min.   :0.00000   Length:586672      Length:586672      Length:586672
1st Qu.:0.00000   Class :character   Class :character   Class :character
Median :0.00000   Mode  :character   Mode  :character   Mode  :character
Mean   :0.04409
3rd Qu.:0.00000
Max.   :1.00000
 danceability        energy            key            loudness
Min.   :0.0000   Min.   :0.000   Min.   : 0.000   Min.   :-60.000
1st Qu.:0.4530   1st Qu.:0.343   1st Qu.: 2.000   1st Qu.:-12.891
Median :0.5770   Median :0.549   Median : 5.000   Median : -9.243
Mean   :0.5636   Mean   :0.542   Mean   : 5.222   Mean   :-10.206
3rd Qu.:0.6860   3rd Qu.:0.748   3rd Qu.: 8.000   3rd Qu.: -6.482
Max.   :0.9910   Max.   :1.000   Max.   :11.000   Max.   :  5.376
     mode         speechiness      acousticness    instrumentalness
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000000
1st Qu.:0.0000   1st Qu.:0.0340   1st Qu.:0.0969   1st Qu.:0.0000000
Median :1.0000   Median :0.0443   Median :0.4220   Median :0.0000245
Mean   :0.6588   Mean   :0.1049   Mean   :0.4499   Mean   :0.1134508
3rd Qu.:1.0000   3rd Qu.:0.0763   3rd Qu.:0.7850   3rd Qu.:0.0095500
Max.   :1.0000   Max.   :0.9710   Max.   :0.9960   Max.   :1.0000000
   liveness         valence           tempo       time_signature
Min.   :0.0000   Min.   :0.0000   Min.   :  0.0   Min.   :0.000
1st Qu.:0.0983   1st Qu.:0.3460   1st Qu.: 95.6   1st Qu.:4.000
Median :0.1390   Median :0.5640   Median :117.4   Median :4.000
Mean   :0.2139   Mean   :0.5523   Mean   :118.5   Mean   :3.873
3rd Qu.:0.2780   3rd Qu.:0.7690   3rd Qu.:136.3   3rd Qu.:4.000
Max.   :1.0000   Max.   :1.0000   Max.   :246.4   Max.   :5.000
```

Tabel 1. Statistiline kokkuvõte andmekogumi tunnuste kohta. Autorite koostatud programmis RStudio.

Analüütiline ülevaade andmetest

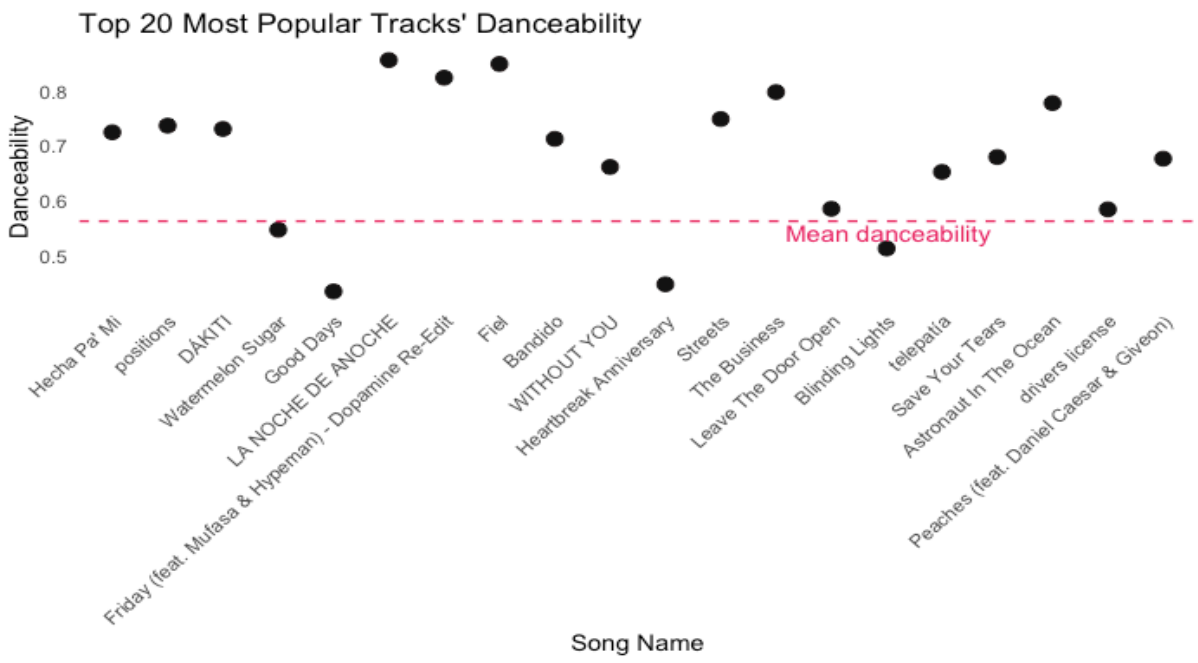
Spotify laulude andmestik on väga suure mahuga, seega anname järgnevalt ülevaate andmetest läbi erinevate visuaalide loomise.

Joonisel 2 olev tulpdiagramm kujutab aastatel 2019 kuni 2021 Spotify kahekümne populaarseima loo võrdlust. Populaarsuse indeksid on skaalal 0 kuni 100. Joonisel horisontaalsed tulbad näitavad iga loo populaarsuse taset. Erinevad värvid tähistavad iga loo väljaandmise aastat, pakkudes kiiret võrdlevat vaadet aastate kaupa.



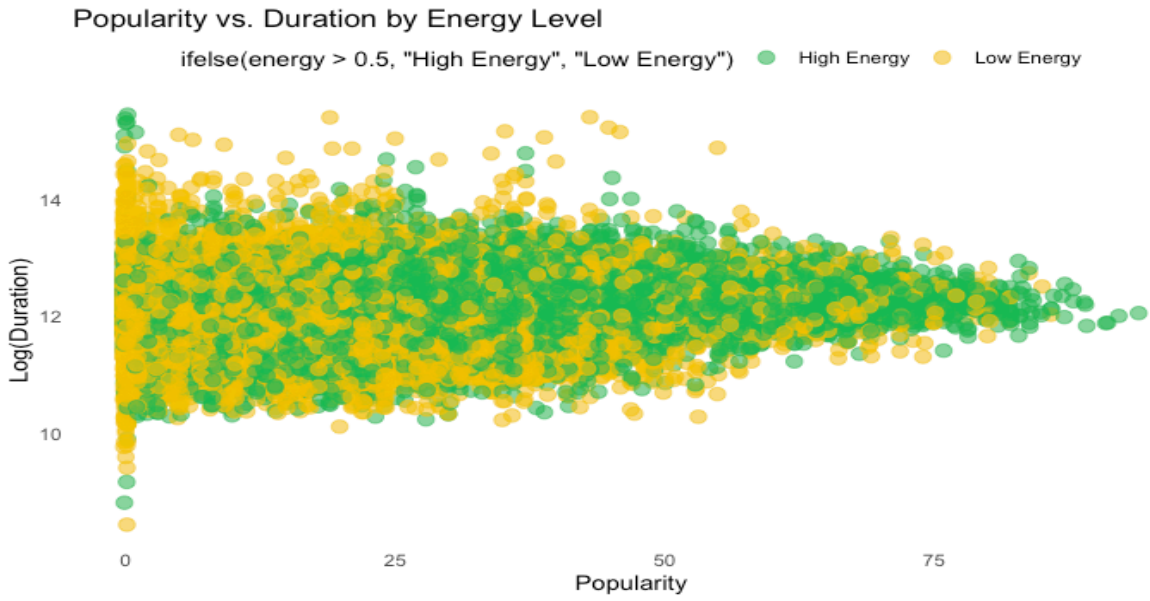
Joonis 2. Kakskümmend Spotify populaarseimat lugu 2019-2021. Autorite koostatud programmis RStudio.

Järgmine hajuvusdiagramm (Joonis 3) kujutab kahekümne kõige populaarsema loo *danceability* skoore. Iga punkt esindab ühte lugu, mille asukoht Y-teljel näitab selle tantsitavust skaalal 0.5 kuni 0.8. Punane katkendlik joon tähistab nende lugude keskmist *danceability* skoori, mille abil on võimalik võrrelda üksikuid lugusid keskmisega. X-telg loetleb laulude nimed, võimaldades tuvastada, milline on iga konkreetse populaarse loo *danceability*.



Joonis 3. Kahekümne kõige populaarsema loo danceability skoorid. Autorite koostatud programmis RStudio.

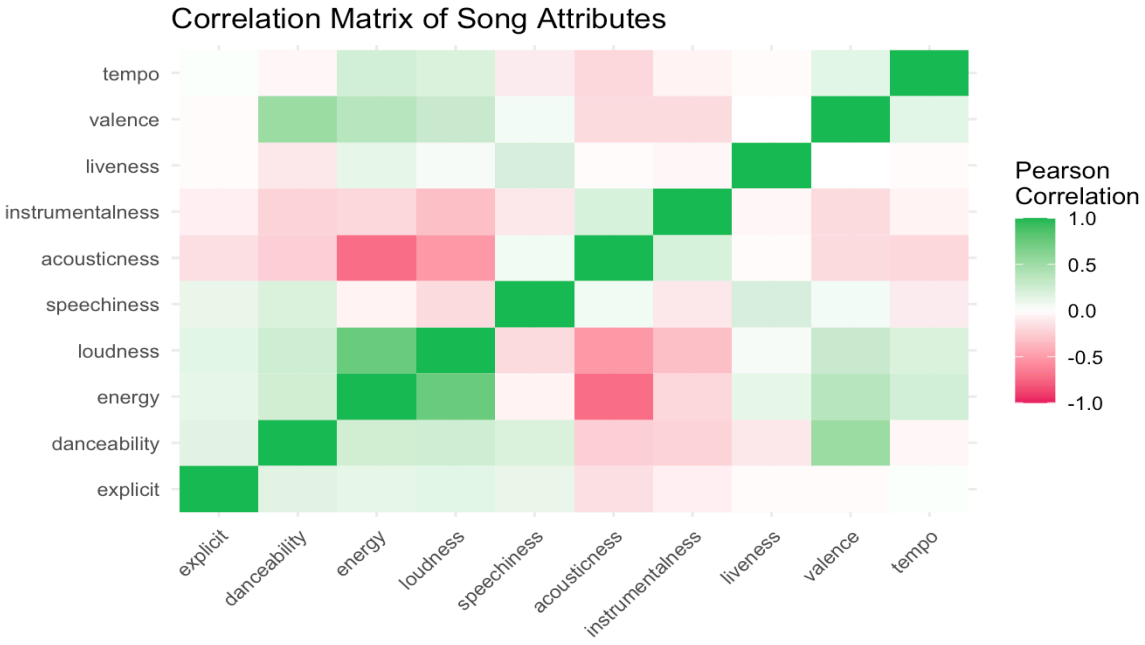
Alljärgnev hajuvusdiagramm (Joonis 4) kujutab populaarsuse ja logaritmiliselt teisendatud loo kestvuse seost, eristades "High Energy" ja "Low Energy" kategooriad. Populaarsus on kuvatud x-teljel ja y-telg mõõdab kestvust logaritmilisel skaalal, et lihtsustada laia andmevahemiku haldamist. Graafik näitab võimalikke trende loo energiataseme, populaarsuse ja kestvuse vahel.



Joonis 4. Populaarsuse ja (log teisendatud) kestvuse seos. Autorite koostatud programmis RStudio.

Korrelatsioonimaatriks (vt Joonis 5) kujutab erinevate lauluomaduste vahelisi korrelatsioone. Roheline tähistab positiivseid ja punane negatiivseid korrelatsioone, tumedamad värvid

näitavad tugevamaid seoseid. Diagonaalil olev rida kõige tumedamaid rohelisi ruute näitab iga omaduse täiuslikku positiivset korrelatsiooni iseendaga. See kompaktne visuaalne abivahend aitab tuvastada, millised laulu omadused kipuvad koos tõusma ja langema.



Joonis 5. Laulude omaduste korrelatsioonimaatriks. Autorite koostatud programmis RStudio.

3. METOODIKA

Analüüsi läbiviimiseks on kasutatud kahte erinevat meetodit, mida autorid pidasid kõige sobivamaks uurimisküsimustele vastamisel. Esimene uurimisküsimus on seotud muusikapalade kategoriseerimisega erinevatesse žanritesse. Sellele uurimisküsimusele vastuse leidmiseks kasutati klasteranalüüsi. Klasteranalüüs on juhendamata õppe metoodika ning antud uuringus võimaldab klasteranalüüs kategoriseerida muusikapalad žanritesse, lähtudes nende sisemistest omadustest nagu tempo, rütm ja meloodia. Selle eesmärk on selgitada muusikapalade andmetest mustreid ja rühmitusi, mis ei pruugi olla kohe nähtavad tavapäraste žanrite kaudu.

Teine uurimisküsimus on seotud loo populaarsust mõjutavate faktorite leidmisega. Sellele uurimisküsimusele leitakse vastus juhusliku metsa meetodiga. Juhuslik mets on samuti juhendamata õppe meetod ning sobib antud uurimisküsimusele vastamiseks, sest juhusliku metsa meetod on hea üldistusvõimega ning suudab efektiivselt tuvastada seoseid ja mustreid suure hulga muutujate vahel, mis on kasulik loo populaarsust mõjutavate faktorite kindlaks tegemisel ja lugude populaarsuse prognoosimisel. Meetodi üldistusvõime on valitud andmestiku puhul oluline ka seetõttu, et andmestikus ei ole populaarsete ja ebapopulaarsete lugude suhe tasakaalus, mis tähendab seda, et väga paljudel teistel meetodite abil on sellisest andemestikust keeruline leida seoseid ja mustreid.

PCA ja klasteranalüüs

Esimesele uurimisküsimusele saadi tulemus kombineerides peakomponentide analüüsi (PCA) ja klasteranalüüsi meetodeid. Nende meetodite koos kasutamine sobib antud konteksti just seetõttu, et PCA meetodi kasutamine esimese sammuna võimaldab andmekogumi mõõtmeid vähendada ja seejärel saab klasteranalüüsi abil leida loodud andmeruumis mustreid.

Peakomponentide analüüs on statistiline meetod andmekogumi mõõtmete vähendamiseks. PCA meetod on efektiivne meetod suurte andmekogude analüüsimisel, mis sisaldavad ühe vaatluse kohta suurt hulka mõõtmeid, samal ajal säilitades maksimaalse teabehulga ja võimaldades mitmemõõtmeliste andmete visualiseerimist. Selle meetodi rakendamisel teisendatakse andmed lineaarse teisendamisega uude koordinaatsüsteemi, kus (enamik) andmete varieerumist saab kirjeldada väiksemate mõõtmetega kui algandmed (*Ibid*.).

Läbiviidud analüüsi esimene etapp kasutab PCA-d, et vähendada laulude omaduste kõrgemõõtmelisust *(high-dimensionality)* hallatavamaks komponentide arvuks, säilitades

samal ajal nii palju varieeruvust kui võimalik. See on oluline, kuna see suurendab järgneva klasteranalüüsi efektiivsust ja tõlgendatavust. PCA-d viiakse läbi R-is, kasutades baaspaketi funktsiooni *prcomp*. Säilitatavate peakomponentide arv määratakse, arvestades komponentide poolt selgitatud kumulatiivset varieeruvuse osakaalu, millele määratakse lävend 80%.

PCA-le järgnevas etapis kasutame klasteranalüüsi, et kategoriseerida laulud nende peamiste komponentide tulemuste põhjal eraldi gruppidesse. Klastrianalüüs on juhendamata õppe tehnika, mis tuvastab meie andmetes klastreid. Klasteranalüüs on teostatud PCA-teisendatud andmetel, selleks et tagada klastri moodustamine andmete kõige olulisemate alusmustrite põhjal, vabaneda üleliigsest mürast, mis võib esineda esialgses kõrgemõõtmelises ruumis.

Kasutame *k-means* klastri algoritmi, mis jaotab laulud *k* eraldiseisvateks ja mittekattuvateks klastriteks. *K* väärtus (klastri arv) määratakse küünarnuki meetodi või *sillhouetti* meetodi abil, mis tuvastab punkti, kus rohkemate klastrite lisamine ei paranda oluliselt klastrite sees olevate ruutude summat. Klastreid teostatakse samuti R-is, kasutades statistikapaketi funktsiooni *kmeans*.

Metoodika viimane samm seisneb klastri tõlgendamises lauluomaduste kontekstis. Me uurime iga klastri tsentroide, et mõista klastri liikmete määratlevaid omadusi. See analüüs aitab kaasa žanrite või stiilide kvalitatiivsele tõlgendamisele, mida iga klaster võib esindada.

Juhuslik mets

Teisele uurimisküsimusele saadi tulemus juhusliku metsa meetodi kasutamisega. Antud meetodit on kasutatud selleks, et leida loo populaarsust mõjutavaid faktoreid. Juhusliku metsa meetod sobib hästi teise uurimisküsimusele tulemuse leidmiseks, sest juhuslik mets suudab efektiivselt käsitleda andmestikku, kus on mitmeid erineva sisuga muutujaid, tabades komplekssseid seoseid muutujate vahel. Tegemist on mitmedimensioonilise andmestikuga, siis juhusliku metsa meetodi kasutamine vähendab ülesobitamise ohtu.

Analüüs viidi läbi R-is, kasutades selleks *randomForest* funktsiooni. Enne *randomForest* funktsiooni kasutamist eemaldati andmestikust *NA* (*Not Available*) väärtused ning eemaldati ebavajalikud dimensioonid nagu id, loo nimi, artisti nimi, artisti id, loo väljastamise kuupäev. Analüüsiks võeti juhuslikul viisil 954 populaarset ja 954 mitte populaarset lugu. Loo populaarsus määrati *popularity* veeru järgi. Kui populaarsuse indeks oli suurem või võrdne 80-ga, siis lugu määrati populaarseks, alla 80 populaarsuse väärtusega lugu määrati mitte populaarseks. Lugusid, kus populaarsuse indeks on üle või võrdne 80-ga, on andmestikus 0.16%. Pool antud valimist määrati treeningandmeteks ning teine pool jäi testandmestikuks.

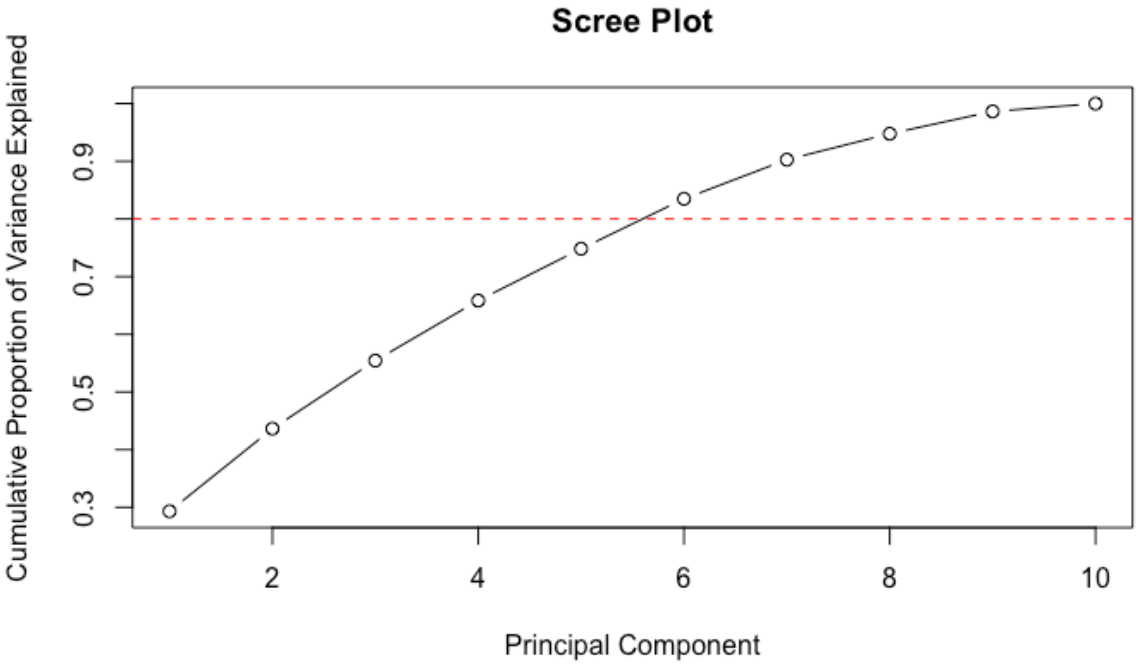
Peale andmetöötlust kasutati *randomForest* funktsiooni, et luua juhusliku metsa mudel, mis kalkuleerib iga dimensiooni olulisust populaarsuse määramisel. Ennustamiseks kasutati *predict* funktsiooni, mis kasutas eelnevalt loodud mudelit ning testandmestikku, et prognoosida lugude populaarsuse indeksit.

Mudeli soorituse hindamiseks kasutati segadusmaatriksit ning arvutati erinevaid sooritusnäitajaid, kasutades selleks *confusionMatrix* funktsiooni. Antud näitajad aitavad mõista mudeli efektiivsust ning annavad ülevaate mudeli tundlikkuse ja spetsiifilisuse tasakaalust.

4. TULEMUSED JA JÄRELDUSED

4.1 PCA ja klasteranalüüs

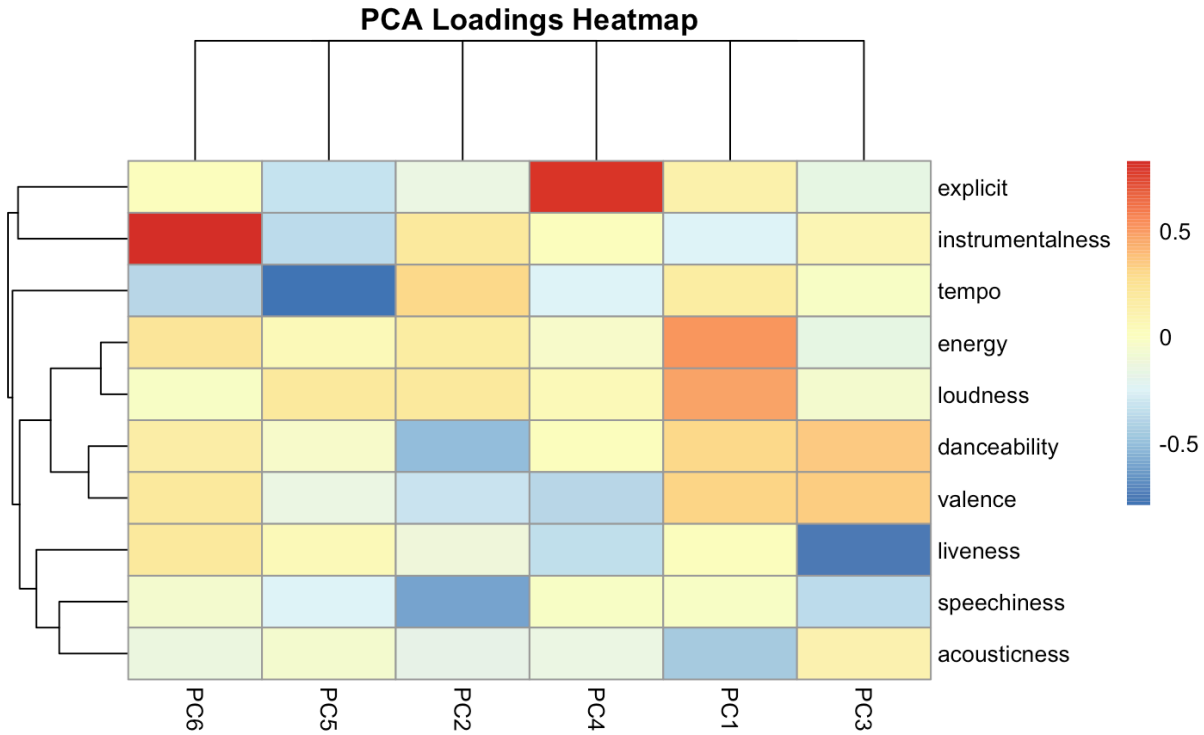
PCA meetod aitas Spotify andmestiku laule selgelt mõtestada, tuues esile kuus põhikomponenti, mis annavad ülevaate loodud muusikapalade mitmekesisusest. Need esimesed kuus peakomponenti selgitasid kumulatiivselt üle 80% andmekogumi varieeruvusest (vt Joonis 6). Seega pidasid autorid piisavaks kuue komponenti kasutamist, millega saada kätte enamik andmestiku teave ilma ülemäärase kohandamiseta.



Joonis 6. PCA: Tähtsusega komponentide tuvastamine. Autorite koostatud programmis RStudio.

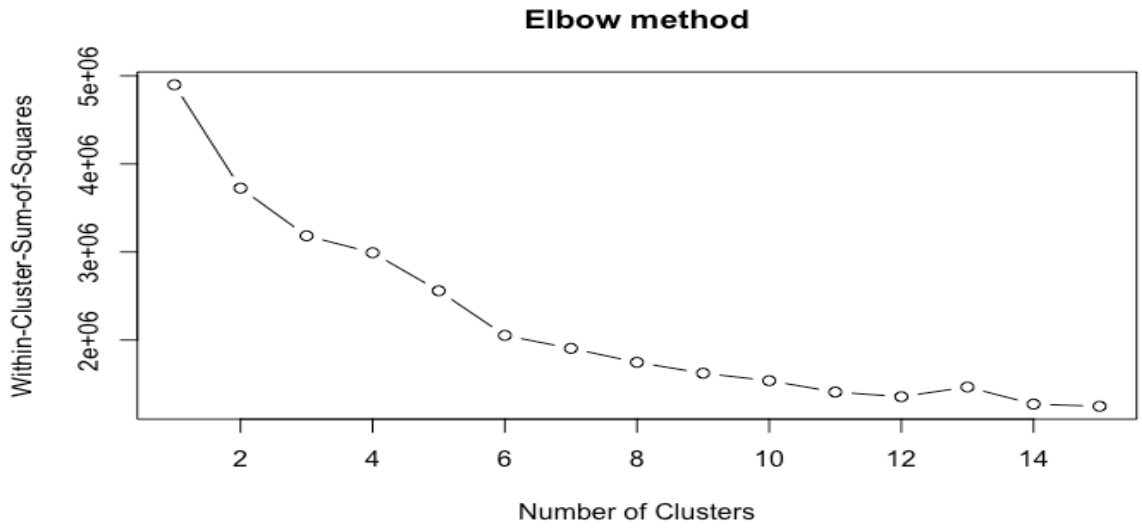
PCA koormuste soojuskaart (vt Joonis 7) andis ülevaade sellest, kuidas iga esialgne muusikapala omadus aitas kaasa kuue valitud peakomponendi kujunemisele. Soojuskaardi legend sinisest punaseni näitas koormuste tugevust ja suunda. Näiteks omadused, millel on tugev positiivne koormus PC-l (punane), andsid positiivset panust vastava komponendi varieeruvusele. Samas, kui tugevate negatiivsete koormustega omadused (sinine), andsid vastupidist panust. Omadused, mille panuse mustrid PC-le olid sarnased, olid hierarhiliselt koondatud kokku, seda näitab soojuskaardi vasakul küljel olev dendrogramm. Mõned suhted on reaalse maailma kontekstis mõistlikud, näiteks muutujad *Energy* ja *Loudness*, mis on kokku

koondatud. Siiski võivad mõned neist suhetest olla esialgu mitte ilmsed ja seda võivad mõjutada spetsiifilised lood andmestikus, näiteks *Speechiness* ja *Acousticness* on koondatud kokku dendrogrammi põhjal, kuid reaalse maailma kontekstis ei pruugi see olla mõistlik.



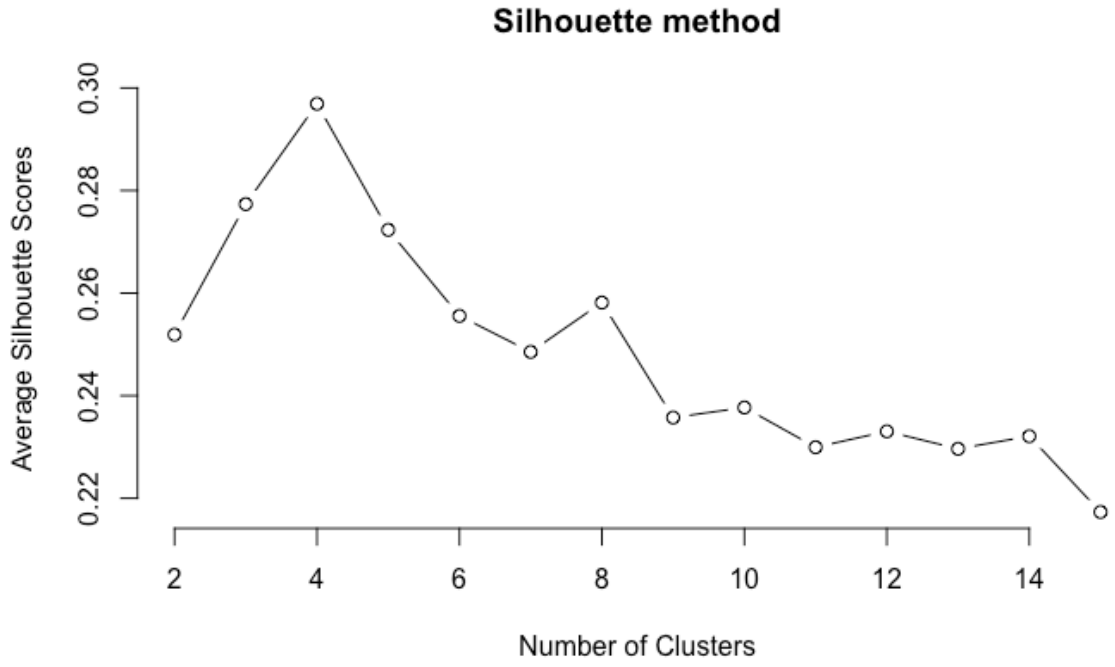
Joonis 7. PCA koormuste soojuskaart. Autorite koostatud programmis RStudio.

PCA on loonud aluse lauluomaduste põhjal žanrite eristavaks jaotuseks. Optimaalse klastrite arvu määramiseks kasutasime esialgu küünarnuki meetodit, kuid selle meetodi rakendamine ei andnud kindlat klastri arvu (vt Joonis 8), mis nõudis edasist uurimist.



Joonis 8. Küünarnuki meetod. Autorite koostatud programmis RStudio.

Seetõttu jätkasime silueti meetodiga, mida rakendasime andmete juhuslikule valimile. See lähenemine andis selgema näidustuse klastrite arvu kohta (vt Joonis 9), mille tulemus soovitass jätkata klasteranalüüsis nelja klastri (k=4) kasutamist.



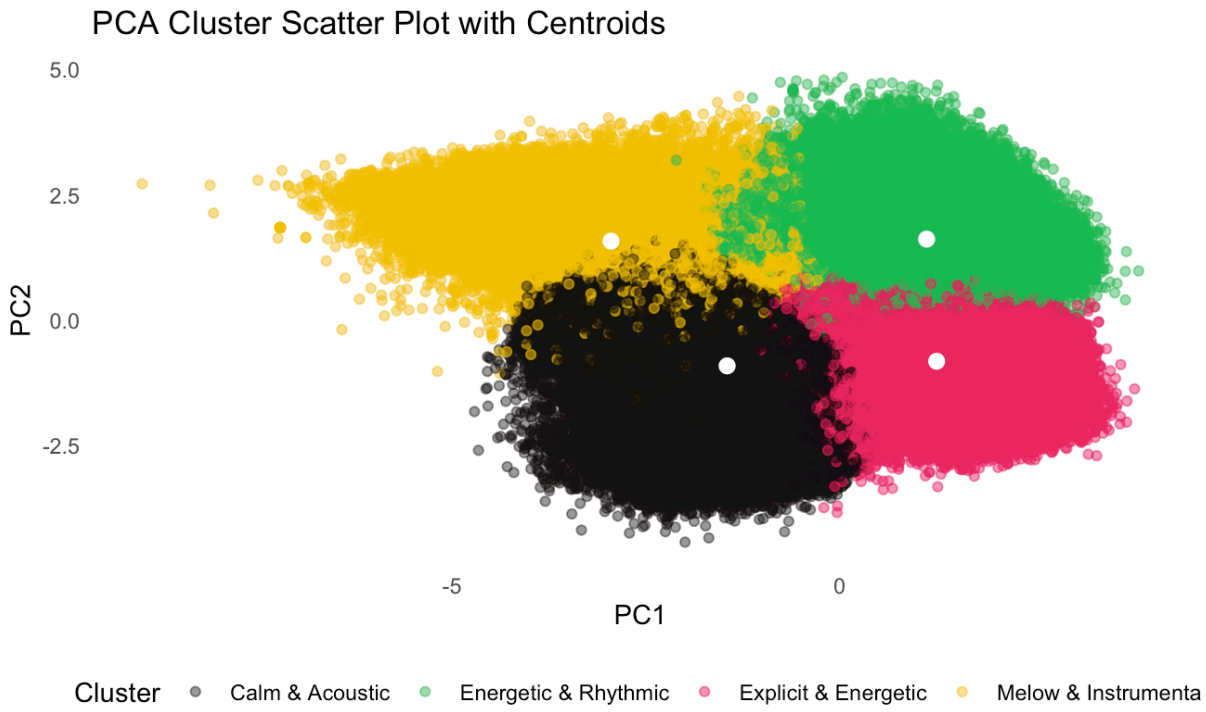
Joonis 9. Silhouette meetod. Autorite koostatud programmis RStudio.

Klastrid (vt Joonis 10) tunduvad olevat eristatavad väheste kattuvustega, mis näitab head eraldumist tunnusruumis pärast dimensionaalsuse vähendamist. Punktide tihedus iga klastri sees varieerub ja mõned klastrid tunduvad olevat tihedamad. See võib viidata sellele, et mõned klastrid on tihedamalt grupeeritud lauludega, võimalikult sarnaste tunnustega, samal ajal kui teised klastrid sisaldavad laule, mis on tunnusruumis rohkem hajutatud.

Diagramm (vt Joonis 10) näitab, et esimene peakomponent (PC1) selgitab rohkem varieeruvust kui teine peakomponent (PC2), arvestades andmete laiemat levikut x-teljel võrrelduna y-teljega. See on PCA puhul tüüpiline.

PCA klastrite diagramm näitab nelja eristuvat klastrit. Iga klaster on detailsemalt lahti seletatud *Lisade* peatükis:

- *Calm & Acoustic*
- *Explicit & Energetic*
- *Mellow & Instrumental*
- *Energetic & Rhythmic*



Joonis 10. PCA klaster hajuvusdiagramm keskpunktidega. Autorite koostatud programmis RStudio.

Klastrite keskpunktid on esitatud suuremate punktidena diagrammil (vt Joonis 10) ja tähistavad iga klastri keskmist asukohta PCA tunnusruumis. Keskpunktide paiknemine aitab mõista iga klastri keskset tendentsi – millised on nende peamised tunnused ja kui sarnased või erinevad on laulud klastri sees. Kui keskpunktid on üksteisele lähedamal, viitab see sellele, et vastavad klastrid on omadustelt sarnased või kattuvad. Vastupidiselt, kui keskpunktid paiknevad üksteisest kaugemal, tähendab see, et klastrite laulud on oma omadustelt väga erinevad.

Klaster	1	2	3	4
1	0	2.798908	3.166720	2.860887
2	2.798908	0	4.644771	2.041659
3	3.166720	4.644771	0	4.437775
4	2.860887	2.041659	4.437775	0

Tabel 2. Keskpunktide kaugus. Autorite koostatud programmis RStudio.

Keskpunktide vahelised kaugused on esitatud Tabelis 2. Kõige lähemal asuvad keskpunktid klastrite 1 ja 2 vahel ning 1 ja 4 vahel, mis võib viidata stiililisele sarnasusele või ühisele muusikalisele meeleolule. Suurim kaugus keskpunktide vahel on klastrite 2 ja 3 ning 3 ja 4 vahel, mis näitab, et nende klastrite laulud on kõige selgemini eristatavad.

Käesoleva uuringu klasteranalüüs on edukalt tuvastanud selgelt eristatavad muusikagrupid (kokku neli klastrit), mis põhinevad erinevate lugude omadustel Spotify andmestikus. Iga klaster võimaldab sügavamalt mõista ja kategoriseerida lugusid. Klastrite keskpunktid annavad ülevaate grupi keskmistest omadustest, kuid need ei pruugi täielikult peegeldada iga üksiku loo eripära klastri sees. Keskpunktide vahelised kaugused annavad ülevaade klastrite omavahelisest sarnasusest või erinevusest. Tulemused annavad tõuke edasiseks uurimiseks ja arendamiseks, et paremini mõista, kuidas näiteks muusikalised omadused mõjutavad kuulamiskogemust ja eelistusi.

4.2 Juhuslik mets

Juhusliku metsa analüüsi tulemusena selgus, et olulisimad tunnused lugude populaarsuse indeksi prognoosimisel on helitugevus (*Loudness*), eksplitsiitsus (*ExplicitFactor*) ja laulu positiivsus (*Valence*). See tähendab seda, et mida suuremad on nende muutujate väärtused, seda suurema tõenäosusega on laul populaarne, kuna ExplicitFactor on fiktiivne sõltumatu muutuja (*dummy muutuja*), siis tähendab see seda, et laul on populaarsem, kui ExcplicitFactor'i väärtus on 1 ehk lugu on eksplitsiitne.

Joonisel 11 on välja toodud kaks joonist: *MeanDecreaseGini* ja *MeanDecreaseAccuracy*.

Esimese joonise *MeanDecreaseGini* on mõõdik sellest, kuidas iga muutuja aitab kaasa nn puu sõlmede ja lehtede homogeensusele juhusliku metsa mudelis. Mida kõrgem on Gini skoor, seda olulisema mõjuga on muutuja mudelis. Seega tunnused nagu helitugevus (*Loudness*), eksplitsiitsus (*ExplicitFactor*) ja laulu positiivsus (*Valence*) on selle mõõdiku järgi ühed olulisimad populaarsuse mõjutajad. Helitugevus on seejuures suurima mõjuga muutuja.

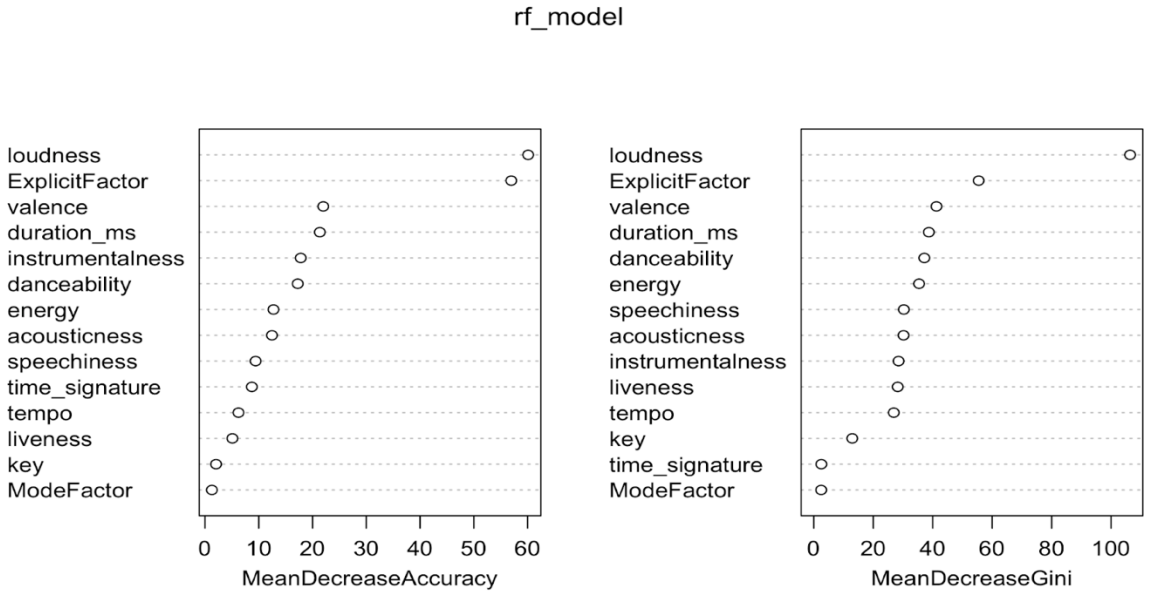
Teise joonise *MeanDecreaseAccuracy* näitab, kui palju kaotab mudel oma täpsust iga muutuja eemaldamisel mudelist. Mida enam mudel oma täpsust kaotab, seda olulisem on muutuja mõju mudelile. Seejuures muutujate olulisus on joonisel kahanevas järjekorras. Jooniselt paistavad taaskord samad kolm tunnust olevat enimtähtsad tunnused lugude populaarsuse hindamisel – helitugevus (*Loudness*) ja eksplitsiitsus (*ExcplicitFactor*) kõige enam ning ka laulu positiivsus (*Valence*).

Helitugevuse (*Loudness*) väärtus on *MeanDecreaseGini* joonisel üle 100, mis tähendab seda, et see muutuja on saadud mudelis väga olulise väärtusega võrreldes teiste muutujatega. Siinkohal tasub märkida, et joonistel välja toodud skaala näitab suhtelist olulisust.

Helitugevuse muutuja olulisust kinnitab ka *MeanDecreaseAccuracy* joonis, kus selgub, et mudeli selgitusvõime väheneb enim just selle muutuja eemaldamisel mudelist.

Teine oluline muutuja mudelis on eskplitsiitsus (ExcplicitFactor), mis *MeanDecreaseAccuracy* joonisel on sarnase väärtusega nagu helitugevus, kuid *MeanDecreaseGini* joonisel omab see muutuja väiksemat olulisust. Eksplitsiitsus (*ExcplicitFactor*) on huvitava väärtusega muutuja andmestikus, kuna andmestikus olevad laulud on enamuses just väärtusega 0, mis tähendab, et laulud on märgitud mitte eksplitsiitseteks, kuid eksplitsiitsete laulude osakaal andmestikus on vaid 4.4 %.

Kolmas oluline muutuja mõlemal joonisel on laulu positiivsuse muutuja (*Valence*), küll aga on selle muutuja väärtused on väga sarnased ka laulu pikkuse muutuja väärtustele (*Duration_ms*), mistõttu võib tegelikult väita, et need kaks muutujat on üpriski võrdse mõjuga loo populaarsuse indeksi prognoosimisel. Jooniselt on ka näha, et kõige väiksema tähtsusega loo populaarsuse ennustamise mõjutamisel on loo helilaad (*ModeFactor*), mis viitab sellele, et loo populaarsusel ei mängi olulist rolli see, kas loo helilaad on minoor või mažoor.



Joonis 11. Juhusliku metsa analüüsi tulemused. Autorite koostatud programmis RStudio.

Segadusmaatriksi tulemused

Analüüsis kasutatud mudeli täpsus on 78.93%, mis tähendab, et see ennustas loo populaarsust õigesti umbes 79 laulu puhul 100-st. Mitte populaarsed lood ennustati täpsusega 79%. Mudeli tundlikkus oli 78.41%, mis viitab sellele, et mudel tuvastas õigesti 78.41% tegelikult mitte

populaarsetest lugudest. F1 skoor on 0.7882, mis viitab tasakaalule täpsuse ja tundlikkuse vahel. Kappa skoor viitab märkimisväärsele ühtivusele väärtusega 0.5786. Andmestiku levimus ehk *prevalence* viitab, et 50% andmestikust on mitte populaarsed lood ehk kinnitab tasakaalus andmestikku.

Segadusmaatriks näitab ennustuste jaotust, kus tuli välja 374 tõelist negatiivset, 379 tõelist positiivset, 103 valenegatiivset ja 98 valepositiivset. Mudeli tundlikkus ja spetsiifilisus on üle 78%, mis on positiivne, sest mudel suudab hästi tuvastada populaarseid ja mitte populaarseid lugusid. Mudel toimib võrdselt hästi mõlema klassi jaoks.

Mitte populaarsete lugude õigesti ennustamise tõenäosus on 79.24% ja negatiivne ennustusväärtus on 78.63% ehk populaarne lugu ennustati õigesti 78.63% tõenäosusega. Mudel suutis õigesti tuvastada 39.2% mitte populaarseid lugusid ning 49.48% ennustas mudel õigesti mitte populaarseks. McNemari testi p-väärtus 0.7778 viitab sellele, et mudeli soorituses kahe klassis vahel pole suurt erinevust.

Juhusliku metsa metoodika rakendamise tulemusel selgus, et populaarsuse prognoosimisel on olulisimad muutujad seotud loo helitugevuse, eskplitsiitsuse ja loo positiivse meelestatusega (ning ka loo pikkus). Seejuures saab välja tuua, et loo helitugevus oli populaarsuse indeksi prognoosimisel olulisimatest mõjutajatest. Segadusmaatriksi tulemused kinnitasid, et juhusliku metsa mudel toimib positiivselt ja saadud tulemused on oluliselt paremad kui nn juhuslik loo populaarsuse ennustamine. Antud mudel andis küll valitud andmestiku põhjal efektiivse tulemuse, küll aga võime eeldada, et reaalses elus mõjutavad lugude populaarsust veel mitmed teised aspektid, näiteks publitseeritava loo autor - aspekt, mida praeguse analüüsi raames ei saadud hinnata.

KASUTATUD KIRJANDUS

Andrews, C. Gardiner, K. Jain, T. K. Olomi, Y. North, A. C. (2020). *Culture, personal values, personality, uses of music, and musical taste.* Kättesaadav: https://psycnet.apa.org/record/2020-43913-001, 27. detsember 2023.

Greenberg, D.M. Baron-Cohen, S. Stillwell, D.J. Kosinski, M. Rentfrow, P.J. (2015). *Musical preferences are linked to cognitive styles.* Kättesaadav: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0131151, 27. detsember 2023.

Jolliffe, I. T. Cadima, J. (2016). *Principal component analysis: a review and recent developments.* Kättesaadav: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792409/, 27. detsember 2023.

Petitbon, A. M. Hitchcock, D. B. (2022). *What Kind of Music Do You Like? A Statistical Analysis of Music Genre Popularity Over Time.* Kättesaadav: https://people.stat.sc.edu/Hitchcock/jds1040.pdf, 27. detsember 2023.

Spotify Dataset 1921 – 2020, 600+ tracks. (2023). Kättesaadav: https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-19212020-600k-tracks, 27. detsember 2023.

LISAD

Lisa 1. Klastrite nimetamine

Põhinedes klastri statistikal (vt Tabel 3), mis on iga muusikalise atribuudi keskmised väärtused, tõlgendasime igat klastrit järgnevalt:

Esimene klaster:

- Explicit - madal
- Danceability - mõõdukas
- Energy - pigem madal
- Loudness - madal
- Speachiness - kõrge
- Acousticness - kõrge
- Instrumentalness - madal
- Liveness - madal
- Valence - mõõdukas
- Tempo - mõõdukas

Pakutud nimetus: Calm & Acoustic

Näide: Bessie Smith - Nobody Knows You When You're Down and Out https://www.youtube.com/watch?v=kxTyV_cBz7o

Teine klaster:

- Explicit - kõrge
- Danceability - kõrge
- Energy - kõrge
- Loudness - kõrge
- Speachiness - mõõdukas
- Acousticness - madal
- Instrumentalness - madal
- Liveness - madal
- Valence - kõrge
- Tempo - mõõdukas

Pakutud nimetus: Explicit & Energetic

Näide: Rare Earth - Train To Nowhere https://www.youtube.com/watch?v=O5meTliC8-8

Kolmas klaster:

- Explicit - madal
- Danceability - madal
- Energy - madal
- Loudness - madal
- Speachiness - madal
- Acousticness - kõrge
- Instrumentalness - kõrge
- Liveness - mõõdukas
- Valence - madal
- Tempo - madal

Pakutud nimetus: Mellow & Instrumental

Näide: Shuggie's Old Time dee-di-lee-di-leet-deet Slide Boogie https://www.youtube.com/watch?v=j7CNR77GDhI

Neljas klaster:

- Explicit - madal
- Danceability - mõdukas
- Energy - kõrge
- Loudness - kõrge
- Speachiness - madal
- Acousticness - madal
- Instrumentalness - madal
- Liveness - kõrge
- Valence - mõdukas
- Tempo - kõrge

Pakutud nimetus: Energetic & Rhythmic

Näide: The Flying Burrito Brothers - High Fashion Queen https://www.youtube.com/watch?v=JAjH6oD-Hcg

cluster	explicit	danceability	energy	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo
1	0.004074792	0.5205004	0.3190881	-13.287145	0.16418608	0.7311628	0.01936072	0.2145776	0.4412198	108.6051
2	0.100182072	0.6935758	0.6786454	-7.770675	0.08562461	0.2731987	0.04152447	0.1531465	0.7284866	116.2877
3	0.003033958	0.4214245	0.2601213	-17.262149	0.05731136	0.8263173	0.79767913	0.1805846	0.3805871	107.7503
4	0.026468400	0.4753746	0.7343599	-7.177547	0.07439586	0.2058484	0.08570714	0.3191819	0.4940940	138.9145

Tabel 3. Klastrite statistika tabel. Autorite koostatud programmis RStudio.