PROIECT

PREDICTIA POPULARITATII UNEI PIESE PE SPOTIFY

Studenti: Muresan Vlad Catalin & Nendrean Flavius

Informatica Economica Grupa 4, Anul 3

INTRODUCERE

Una dintre cele mai mari și profitabile industrii din ziua de azi este industria muzicală, evaluată la peste 26 miliarde de euro. Fie că ești artist sau simplu ascultător, cele două cele mai mari platforme pe care poți asculta sau promova muzica sunt YouTube și Spotify. În timp ce Spotify este cea mai utilizată platformă pentru redarea conținutului muzical, YouTube este cea mai mare platformă de streaming în general, cu cele mai multe vizualizări alocate clipurilor muzicale.

În mod evident, unul dintre principalele interese pentru artiști este crearea unor piese care să atragă publicul, să aibă un număr mare de stream-uri și să genereze venituri cât mai mari.

În continuare, vom aborda următoarele întrebări de cercetare la care ne propunem să răspundem:

1. Există o legătură între dansabilitate (danceability), nivelul de vorbire (speechiness), durată (duration), nivelul de audiență live (liveness), tempou (tempo) și energie (energy) și numărul de stream-uri pe care îl are o piesă?

2. În cazul în care există această legătură, cât de puternică este aceasta?

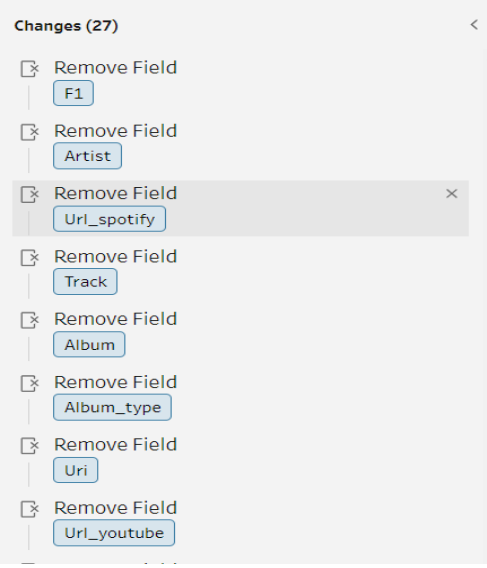
3. Aceste variabile influențează în mod similar sau în proporții diferite numărul de stream-uri?

4. Putem estima numărul de vizualizări/stream-uri pe care o piesă le va avea în funcție de dansabilitate, nivelul de vorbire, durată, nivelul de audiență live, tempou sau energie?

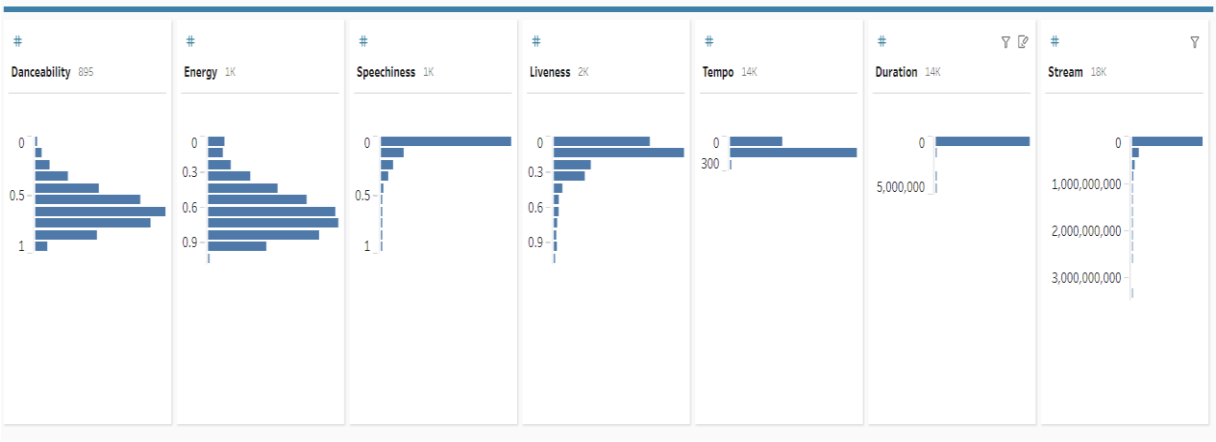
Setul de date:

Sursa setului de date utilizat în cadrul acestui proiect este https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube. Autorul acestui set de date a furnizat detalii referitoare la mai multe aspecte legate de anumite piese preluate de la diverși artiști, inclusiv informații tehnice despre acestea, cum ar fi tempoul, cheia, volumul, valența, acustica etc., precum și numărul de vizualizări și like-uri de pe YouTube și numărul de stream-uri de pe Spotify.

Curățarea bazei de date: Pentru a curăța baza de date, am utilizat platforma specializată pentru modelarea datelor tabelare, Tableau Prep. Deoarece am selectat doar coloanele Dansabilitate, Nivel de vorbire, Durată, Nivel de audiență live, Tempou, Energie și numărul de stream-uri (Stream), setul nostru de date va arăta ca în modelul de mai jos:



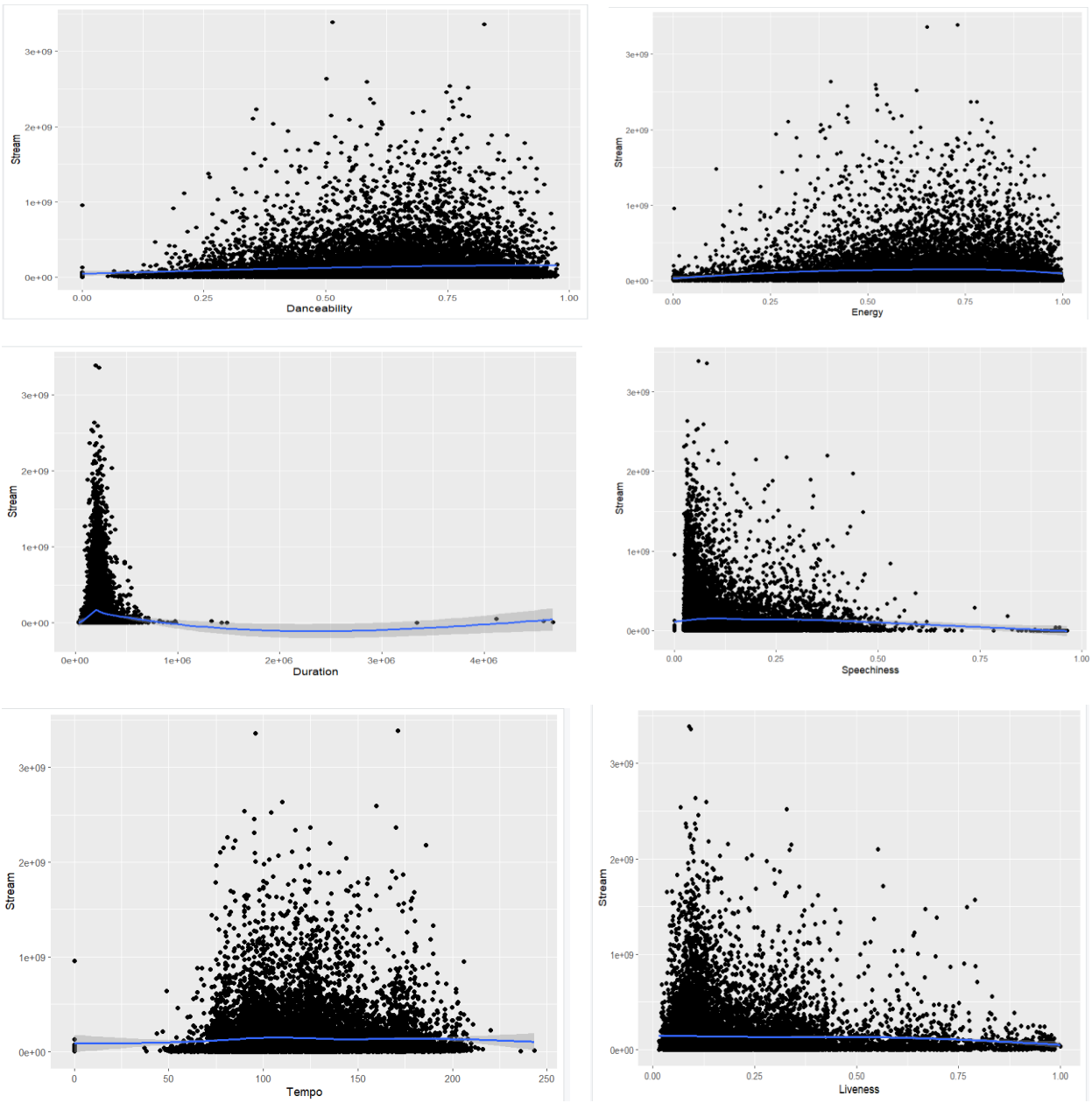
Prin urmarea acestui proces de curățare a setului de date, vom obține urmatorele coloane, mai relevante, care vor facilita analiza și interpretarea datelor în cadrul proiectului nostru:



Rezultate si discutii:

Observăm că setul de date conține atât variabile numerice, cât și variabile categorice, iar în contextul întrebărilor de cercetare, variabila pe care încercăm să o prezicem este tot de natură numerică. În consecință, lucrăm într-o ipoteză în care ne propunem să realizăm o predicție numerică.

Primul pas în atingerea obiectivului nostru este de a genera graficele pentru fiecare variabilă din setul nostru de date în mod individual:



Astfel, observăm că există o posibilă legătură între variabilele energy, duration, liveness speechiness si tempo, pe de-o parte, și variabila stream pe de altă parte. Este posibil ca unele dintre aceste relații să fie liniare, indicând o corelație între nivelul de energie al unei piese, durata acesteia, prezența unei audiențe în înregistrare și nivelul de vorbire în piesă, precum și tempo-ul variat a acesteia și popularitatea piesei în rândul utilizatorilor de pe platforma Spotify.

Pe baza acestei constatări și în scopul de a determina dacă variabilele energy, duration, liveness, speechiness și tempo sunt variabile independente care influențează numărul de stream-uri, precum și modul în care se manifestă această influență și relația dintre acestea, am decis să utilizăm în primul rând metoda celor mai mici pătrate - regresia liniară. Prin aplicarea regresiei

liniare, vom căuta o relație matematică între variabilele menționate și numărul de stream-uri, cu scopul de a identifica coeficienții de regresie care să ne ofere o înțelegere mai profundă a modului în care aceste variabile pot influența popularitatea unei piese. Această alegere a fost determinate de elementele observate cu ocazia vizualizării graficelor realizate din care rezultă că există o relație posibil liniară între variabilele independente menționate și numărul de stream-uri, având în vedere că informațiile obținute sunt destinate atat artiștilor care doresc să lanseze o piesă de succes și cat si publicului lor care nu deține neapărat cunoștințe avansate în domeniul statisticii considerăm că regresia liniară este o metodă preferabilă, deoarece oferă un grad mai ridicat de interpretabilitate.

Astfel,ne propunem să identificăm o relație f între Y (**Stream**) și **X=(X1, X2,..., Xp)** (Danceability, Energy, Speechiness, Duration, Liveness, Tempo) in asa fel încât **Y = f(X)+ε.**

Vom presupune că forma funcției f este liniară, exprimată prin ecuația **Y = β0 + β1 X1 + ... + βp Xp**. Vom efectua ajustări pe datele de antrenament pentru a obține estimări pentru coeficienții **β0, β1, ..., βp**.

În primă fază, am examinat relația dintre numărul de stream-uri (Stream) și gradul de dansabilitate (Danceability), presupunând că Stream ≈ β0 + β1 * **Danceability**, unde β0 reprezintă interceptul și β1 reprezintă panta. Am utilizat R-Studio pentru a calcula valorile interceptului, pantei, erorilor standard, t-statistic și p-value. Rezultatele obținute sunt următoarele:

```
Call:
lm(formula = Stream ~ Danceability, data = Streamuri)

Residuals:
       Min         1Q     Median         3Q        Max
-170393521 -117495828  -81816760    1841150 3260874376

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   70605223    6825913   10.34   <2e-16 ***
Danceability 107083053   10620214   10.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 245700000 on 19547 degrees of freedom
Multiple R-squared:  0.005174,  Adjusted R-squared:  0.005123
F-statistic: 101.7 on 1 and 19547 DF,  p-value: < 2.2e-16
```

Astfel, dacă dacă ținem cont doar de Danceability:

Stream = 70605223 + 107083053 * **Danceability**

Folosind valorile calculate, am determinat intervalele de încredere pentru parametrii βi, cu un nivel de încredere de 95%, utilizând formula **βi [βˆ ∈ i − 2SE(βˆi), βˆi + 2SE(βˆi)]**. Astfel:

CI pentru β0 : [57225852 , 83984595] : numarul de streamuri, cel mai probabil, se va incadra intre aceste intervale, daca nu vom tine cont de dansabilitatea piesei.

CI pentru β1 : [86266526, 127899579] : Acesta sugerează că dacă danceability crește cu o unitate (în acest caz, cu 0.1 deoarece valoarea maximă a danceability este 1), atunci ne așteptăm ca numărul de streamuri să crească cu o valoare între 86,266,526 și 127,899,579, presupunând că toate celelalte variabile rămân constante.

```
                  2.5 %    97.5 %
(Intercept)  57225852  83984595
Danceability 86266526 127899579
```

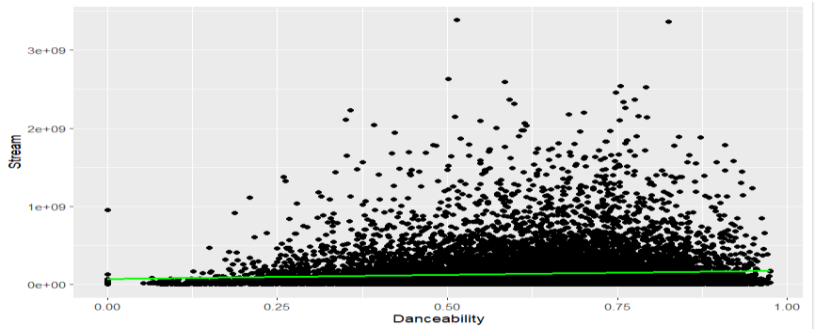
Având în vedere aceste informații, am procedat la testarea ipotezelor formulate:

H0 - Ipoteza nulă: Nu există o relație între variabila X (Danceability) și variabila Y (Stream) - testarea H0 presupune $\beta 1 = 0$.

Ha - Ipoteza alternativă: Există o relație între variabila X (Danceability) și variabila Y (Stream) - testarea H1 presupune $\beta 1 \neq 0$.

Analizând această problemă, observăm că valoarea p-value este foarte mică, ceea ce ne conduce la concluzia că ipoteza nulă este respinsă. Astfel, putem afirma că relația dintre variabila X (Danceability) și variabila Y (Stream) nu se datorează întâmplării, ci există un alt motiv care le leagă.

Pentru a continua analiza, vom genera un nou set de date (grid) din setul inițial de date, luând în considerare 100 de valori ale variabilei Danceability pentru intervalul 0-1 , urmând să calculăm, pentru acest interval, valorile variabilei stream. Dupa ce v-om calcula aceste valori și le v-om aplica pe modelul creat, v-om crea o nouă variabilă, grid_Danceability ce va cuprinde predicțiile si datele nou calculate care vor fi marcate pe un grafic nou, utilizând o dreaptă de culoarea verde, după cum se poate observa în imaginea de mai jos:



Bazându-ne pe valoarea calculată pentru R^2, care este foarte aproape de 0, putem trage o concluzie preliminară că nu există o corelație puternică între variabila danceability și variabila stream, în ceea ce privește modelul luat în considerare. Aceasta sugerează că variația din variabila danceability nu poate explica în mod semnificativ variația în variabila Stream.

Apoi, am examinat relația dintre numărul de stream-uri (Stream) și nivelul de energie(Energy) transmis presupunând că Stream ≈ β0 + β1 * **Energy**, unde β0 reprezintă interceptul și β1 reprezintă panta. Am utilizat R-Studio pentru a calcula valorile interceptului, pantei, erorilor standard, t-statistic și p-value. Rezultatele obținute sunt următoarele:

```
Call:
lm(formula = Stream ~ Energy, data = Streamuri)

Residuals:
       Min         1Q     Median         3Q        Max
-155178971 -117765303  -85178421    1609343 3244521874

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 104368582    5523699  18.895  < 2e-16 ***
Energy       51547716    8243003   6.254  4.1e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 246100000 on 19547 degrees of freedom
Multiple R-squared:  0.001997,  Adjusted R-squared:  0.001946
F-statistic: 39.11 on 1 and 19547 DF,  p-value: 4.096e-10
```

Astfel, dacă dacă ținem cont doar de variabila Energy:

Stream = 104368582 + 51547716 * **Energy**

Folosind valorile calculate, am determinat intervalele de încredere pentru parametrii βi, cu un nivel de încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi), βˆi + 2SE(βˆi)]. Astfel:

CI pentru β0 : [93541661 , 115195503] : numarul de streamuri, cel mai probabil, se va incadra intre aceste intervale, daca nu vom tine cont de energia piesei.

CI pentru β1 : [35390726, 67704705] : Acesta sugerează că dacă danceability crește cu o unitate (în acest caz, cu 0.1 deoarece valoarea maximă a variabilei energy este 1), atunci ne așteptăm ca numărul de streamuri să crească cu o valoare între 35,390,726 si 67,704,705, presupunând că toate celelalte variabile rămân constante:

```
> confint(streamuri_energy)
                2.5 %    97.5 %
(Intercept) 93541661 115195503
Energy      35390726  67704705
>
```

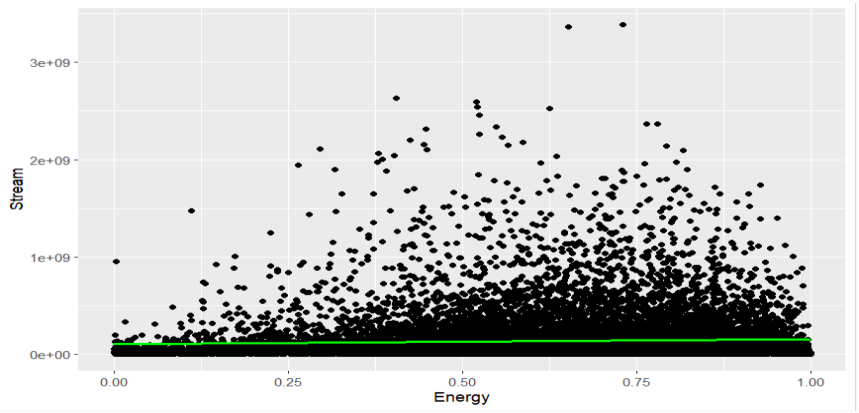
Având în vedere aceste informații, am procedat la testarea ipotezelor formulate:

H0 - Ipoteza nulă: Nu există o relație între variabila X (Energy) și variabila Y (Stream) - testarea H0 presupune β1 = 0.

Ha - Ipoteza alternativă: Există o relație între variabila X (Energy) și variabila Y (Stream) - testarea H1 presupune β1 ≠ 0.

Analizând această problemă, observăm că valoarea p-value este foarte mică, ceea ce ne conduce la concluzia că ipoteza nulă este respinsă. Astfel, putem afirma că relația dintre variabila X (Energy) și variabila Y (Stream) nu se datorează întâmplării, ci există un alt motiv care le leagă.

Pentru a continua analiza, vom genera un nou set de date (grid) din setul inițial de date, luând în considerare 100 de valori ale variabilei Energy pentru intervalul 0-1 , urmând să calculăm, pentru acest interval, valorile variabilei Stream. Dupa ce v-om calcula aceste valori și le v-om aplica pe modelul creat, v-om crea o nouă variabilă, grid_Energy ce va cuprinde predicțiile si datele nou calculate care vor fi marcate pe un grafic nou, utilizând o dreaptă de culoarea verde, după cum se poate observa în imaginea de mai jos:



Bazându-ne pe valoarea calculată pentru R^2, care este foarte aproape de 0, putem trage o concluzie preliminară că nu există o corelație puternică între variabila Energy și variabila Stream, în ceea ce privește modelul luat în considerare. Aceasta sugerează că variația din variabila Energy nu poate explica în mod semnificativ variația în variabila Stream.

In continuare, am examinat relația dintre numărul de stream-uri (Stream) și nivelul de vorbire in piesa (Speechiness) transmis presupunând că Stream ≈ β0 + β1 * **Speechiness**, unde β0 reprezintă interceptul și β1 reprezintă panta. Am utilizat R-Studio pentru a calcula valorileinterceptului, pantei, erorilor standard, t-statistic și p-value. Rezultatele obținute sunt următoarele:

```
lm(formula = Stream ~ Speechiness, data = Streamuri)

Residuals:
       Min         1Q     Median         3Q        Max
-139703795 -119008457  -87099996    2005597 3248069298

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 140703803    2367790  59.424   <2e-16 ***
Speechiness -37672454   16583240  -2.272   0.0231 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 246300000 on 19547 degrees of freedom
Multiple R-squared:  0.0002639, Adjusted R-squared:  0.0002128
F-statistic: 5.161 on 1 and 19547 DF,  p-value: 0.02311
```

Astfel, dacă dacă ținem cont doar de variabila Speechiness:

Stream = 140703803 + (-37672454) * **Speechiness**

Folosind valorile calculate, am determinat intervalele de încredere pentru parametrii βi, cu un nivel de încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi), βˆi + 2SE(βˆi)]. Astfel:

CI pentru β0 : [136062733 , 145344874] : numarul de streamuri, cel mai probabil, se va incadra intre aceste intervale, daca nu vom tine cont de gradul de vorbire in piesa.

CI pentru β1 : [-70177019, -5167889] : Acesta sugerează că dacă variabila Speechiness crește cu o unitate (în acest caz, cu 0.1 deoarece valoarea maximă a variabiliei speechiness este 1), atunci ne așteptăm ca numărul de streamuri să scada cu valori cuprinse intre -70,177,019 si -5,157,889 presupunând că toate celelalte variabile rămân constante:

```
                2.5 %    97.5 %
(Intercept) 136062733 145344874
Speechiness -70177019  -5167889
```

Având în vedere aceste informații, am procedat la testarea ipotezelor formulate:

H0 - Ipoteza nulă: Nu există o relație între variabila X (Speechiness) și variabila Y (Stream) - testarea H0 presupune β1 = 0.

Ha - Ipoteza alternativă: Există o relație între variabila X (Speechiness) și variabila Y (Stream) - testarea H1 presupune β1 ≠ 0.

Analizând această problemă, observăm că valoarea p-value nu este una foarte mica, ceea ce poate duce la concluzia ca ipoteza nula s-ar putea sa fie corecta, iar relatia dintre Y(Stream) si X (Speechiness) sa se datoreze sansei. De asemenea, luam in conisderare si valoarea foarte mică, apropiată de 0, a lui R^2, care intareste ipoteza valorii nule existente.

In urmatoaraea parte, am examinat relația dintre numărul de stream-uri (Stream) și durata piesei (Duration) transmis presupunând că Stream ≈ β0 + β1 * **Duration**, unde β0 reprezintă

interceptul și β1 reprezintă panta. Am utilizat R-Studio pentru a calcula valorile interceptului, pantei, erorilor standard, t-statistic și p-value. Rezultatele obținute sunt următoarele:

```
Call:
lm(formula = Stream ~ Duration, data = Streamuri)

Residuals:
       Min         1Q     Median         3Q        Max
-140698178 -119318204  -86986787    1906173 3248857549

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.422e+08  3.582e+06  39.688   <2e-16 ***
Duration    -2.247e+01  1.388e+01  -1.619    0.106
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 246300000 on 19547 degrees of freedom
Multiple R-squared:  0.000134,  Adjusted R-squared:  8.289e-05
F-statistic:  2.62 on 1 and 19547 DF,  p-value: 0.1055
```

Astfel, dacă dacă ținem cont doar de variabila Duration:

Stream = 1.422e+08 + (-2.247e+01) *

Duration

Folosind valorile calculate, am determinat intervalele de încredere pentru parametrii βi, cu un nivel de încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi), βˆi + 2SE(βˆi)]. Astfel:

CI pentru β0 : [1.351376e+08 , 1.1491793e+08] : numarul de streamuri, cel mai probabil, se va incadra intre aceste intervale, daca nu vom tine cont de durata piesei.

CI pentru β1 : [-4.968656e+01, 4.738569e+00] : Acest interval sugereaza ca daca durata va creste cu o unitate stream-urile piesei vor varia cu o valoare aflata in intervalul -4.968656e+01 si 4.738569e+00.

```
                   2.5 %       97.5 %
(Intercept)  1.351376e+08 1.491793e+08
Duration    -4.968656e+01 4.738569e+00
```

Având în vedere aceste informații, am procedat la testarea ipotezelor formulate:

H0 - Ipoteza nulă: Nu există o relație între variabila X (Duration) și variabila Y (Stream) - testarea H0 presupune β1 = 0.

Ha - Ipoteza alternativă: Există o relație între variabila X (Duration) și variabila Y (Stream) - testarea H1 presupune β1 ≠ 0.

Analizând această problemă, observăm că valoarea p-value este relative crescuta (0.10555) iar R^2 este la un nivel foarte mic, acest lucru determinand o posibilitate ridicata ca ipoteza nula sa fie reala iar relatia dintre cele doua variabile sa fie bazata pe sansa.

Bazându-ne pe valoarea calculată pentru R^2, care este foarte aproape de 0, putem trage o concluzie preliminară că nu există o corelație puternică între variabila Duration și variabila Stream, în ceea ce privește modelul luat în considerare. Aceasta sugerează că durata nu poate explica în mod semnificativ variația în variabila Stream.

Apoi, am examinat relația dintre numărul de stream-uri (Stream) și nivelul de influenta a modului in care este cantata piesa (Liveness) transmis presupunând că Stream ≈ β0 + β1 * **Liveness**, unde β0 reprezintă interceptul și β1 reprezintă panta. Am utilizat R-Studio pentru a calcula valorile interceptului, pantei, erorilor standard, t-statistic și p-value. Rezultatele obținute sunt următoarele:

```
Call:
lm(formula = Stream ~ Liveness, data = Streamuri)

Residuals:
       Min         1Q     Median         3Q        Max
-145657875 -119226386  -85752588    1959752 3243748225

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 147774442    2693541  54.863  < 2e-16 ***
Liveness    -55767888   10659133  -5.232 1.69e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 246200000 on 19547 degrees of freedom
Multiple R-squared:  0.001398,  Adjusted R-squared:  0.001347
F-statistic: 27.37 on 1 and 19547 DF,  p-value: 1.695e-07
```

Astfel, dacă dacă ținem cont doar de variabila Liveness:

1477744442 + (-55767888) * **Liveness**

Folosind valorile calculate, am determinat intervalele de încredere pentru parametrii βi, cu un nivel de încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi), βˆi + 2SE(βˆi)]. Astfel:

CI pentru β0 : [142494872,153054013] : numarul de streamuri, cel mai probabil, se va incadra intre aceste intervale, daca nu vom tine cont de gradul de audienta live a piesei.

CI pentru β1 : [-76660698, -34875078]: Acesta sugerează că dacă variabila liveness crește cu o unitate (în acest caz, cu 0.1 deoarece valoarea maximă a variabilei liveness este 1), atunci ne așteptăm ca numărul de streamuri să ia o valoare cuprinsa intre intervalul -76660698 si -34875078 presupunând că toate celelalte variabile rămân constante:

```
                2.5 %    97.5 %
(Intercept) 142494872 153054013
Liveness    -76660698 -34875078
```

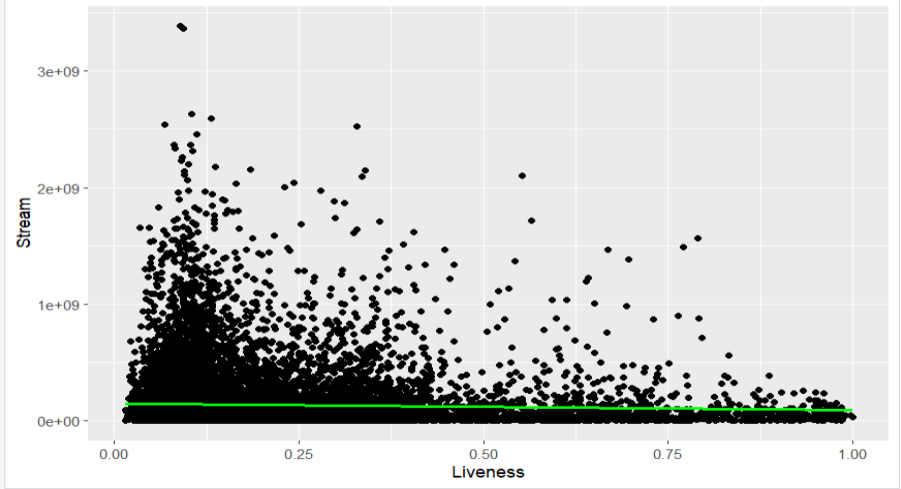
Având în vedere aceste informații, am procedat la testarea ipotezelor formulate:

H0 - Ipoteza nulă: Nu există o relație între variabila X (Liveness) și variabila Y (Stream) - testarea H0 presupune β1 = 0.

Ha - Ipoteza alternativă: Există o relație între variabila X (Liveness) și variabila Y (Stream) - testarea H1 presupune $\beta 1 \neq 0$.

Analizând această problemă, observăm că valoarea p-value este foarte mică, ceea ce ne conduce la concluzia că ipoteza nulă este respinsă. Astfel, putem afirma că relația dintre variabila X (Liveness) și variabila Y (Stream) nu se datorează întâmplării, ci există un alt motiv care le leagă.

Pentru a continua analiza, vom genera un nou set de date (grid) din setul inițial de date, luând în considerare 100 de valori ale variabilei Liveness pentru intervalul 0-1 , urmând să calculăm, pentru acest interval, valorile variabilei Stream. Dupa ce v-om calcula aceste valori și le v-om aplica pe modelul creat, v-om crea o nouă variabilă, grid_Liveness ce va cuprinde predicțiile si datele nou calculate care vor fi marcate pe un grafic nou, utilizând o dreaptă de culoarea verde, după cum se poate observa în imaginea de mai jos:



Bazându-ne pe valoarea calculată pentru R^2, care este foarte aproape de 0, putem trage o concluzie preliminară că nu există o corelație puternică între variabila Liveness și variabila Stream, în ceea ce privește modelul luat în considerare. Aceasta sugerează că variația din variabila Liveness nu poate explica în mod semnificativ variația în variabila Stream.

Apoi, am examinat relația dintre numărul de stream-uri (Stream) și tempo-ul in care este cantata piesa. presupunând că Stream $\approx \beta 0 + \beta 1$ * **Tempo**, unde $\beta 0$ reprezintă interceptul și $\beta 1$ reprezintă panta. Am utilizat R-Studio pentru a calcula valorile interceptului, pantei, erorilor standard, t-statistic și p-value. Rezultatele obținute sunt următoarele:

```
Call:
lm(formula = Stream ~ Tempo, data = Streamuri)

Residuals:
       Min         1Q     Median         3Q        Max
-138581473 -119324822  -87167886    2047622 3248325387

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 134514320    7388151  18.207   <2e-16 ***
Tempo           21523      59491   0.362    0.718
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 246400000 on 19547 degrees of freedom
Multiple R-squared:  6.696e-06, Adjusted R-squared:  -4.446e-05
F-statistic: 0.1309 on 1 and 19547 DF,  p-value: 0.7175
```

Astfel, dacă dacă ținem cont doar de variabila Tempo:

134514320 + 21523 * **Tempo**

Folosind valorile calculate, am determinat intervalele de încredere pentru parametrii βi, cu un nivel de încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi), βˆi + 2SE(βˆi)]. Astfel:

CI pentru β0 : [120032913.71 , 148995726.8] : numarul de streamuri, cel mai probabil, se va incadra intre aceste intervale, daca nu vom tine cont de gradul de audienta live a piesei.

CI pentru β1 : [-95084.24 , 138130.7]: Acesta sugerează că dacă variabila Tempo crește cu o unitate atunci ne așteptăm ca numărul de streamuri să ia o valoare cuprinsa intre intervalul -95084.24 si 138130.7 presupunând că toate celelalte variabile rămân constante:

```
                  2.5 %      97.5 %
(Intercept) 120032913.71 148995726.8
Tempo          -95084.24    138130.7
```

Având în vedere aceste informații, am procedat la testarea ipotezelor formulate:

H0 - Ipoteza nulă: Nu există o relație între variabila X (Tempo) și variabila Y (Stream) - testarea H0 presupune β1 = 0.

Ha - Ipoteza alternativă: Există o relație între variabila X (Tempo) și variabila Y (Stream) - testarea H1 presupune β1 ≠ 0.

Analizând această problemă, observăm că valoarea p-value este foarte crescută (0.7175) iar cu un R^2 foarte mic, putem presupune că ipoteza nulă este valabilă în contextul actual.

Pe baza informațiilor obținute prin realizarea modelelor de mai sus, am ajuns la concluzia că există o corelație între anumite variabile independente și numărul de stream-uri al unei piese. Astfel, există posibilitatea ca unele dintre variabilele studiate să influențeze numărul de stream-uri.

Având în vedere această concluzie, am decis că ar fi utilă utilizarea unei metode de regresie liniară cu mai multe variabile. Astfel, am ajuns la următoarea formulă:

Stream ≈ β0 + β1 * Dansabilitate + β2 * Tempou + β3 * Nivel de vorbire + β4 * Durată + β5 * Nivel de audiență live + β6 * Energie.

După introducerea formulei în RStudio, am obținut următoarele valori pentru intercept (β0), pante (β1, β2, β3, β4, β5, β6), erori standard, statisticile t și valorile p.

```
Call:
lm(formula = Stream ~ Danceability + Tempo + Speechiness + Duration +
    Liveness + Energy, data = Streamuri)

Residuals:
       Min         1Q     Median         3Q        Max
-182037072 -117749018  -79307418    2937622 3246159258

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.601e+07  1.136e+07   5.812 6.25e-09 ***
Danceability  9.946e+07  1.142e+07   8.708  < 2e-16 ***
Tempo         2.325e+04  6.048e+04   0.384    0.701
Speechiness  -7.624e+07  1.707e+07  -4.467 7.98e-06 ***
Duration     -1.579e+01  1.391e+01  -1.135    0.256
Liveness     -5.409e+07  1.093e+07  -4.950 7.47e-07 ***
Energy        4.359e+07  8.790e+06   4.959 7.13e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 245300000 on 19542 degrees of freedom
Multiple R-squared:  0.008486,  Adjusted R-squared:  0.008182
F-statistic: 27.88 on 6 and 19542 DF,  p-value: < 2.2e-16
```

Următorul pas al analizei constă în testarea ipotezelor formulate, respectiv:

H0 - Ipoteza nulă: Nu se identifică nicio relație între predictori și variabila dependentă - β1 = β2 = β3 = β4 = β5 = β6

Ha - Ipoteza alternativă: Există cel puțin un coeficient βi care este diferit de zero (βi ≠ 0)

Având în vedere că valoarea statistică F este mult mai mare decât 1 (de exemplu, 27.88) și că p-value are o valoare foarte mică, putem concluziona că există suficiente dovezi statistice pentru a respinge ipoteza nulă și a accepta ipoteza alternativă.

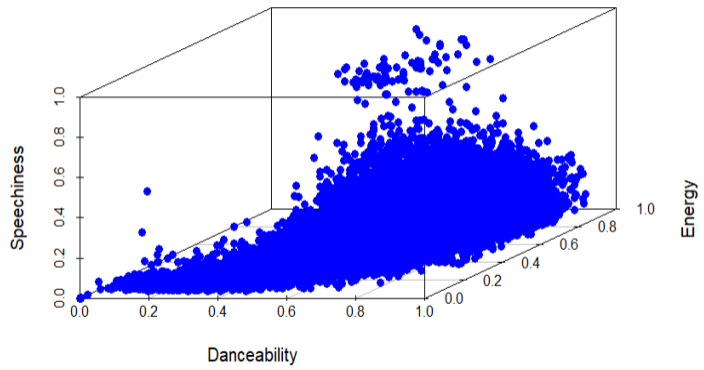
După analizarea datelor, am observat că variabilele Durată și Tempou au o valoare mai mare a statisticilor t și o influență mai redusă în comparație cu celelalte variabile. Ca rezultat, am creat un nou model care include doar variabilele Dansabilitate, Nivelul de vorbire, Nivelul de audiență live și Energie, considerate a fi mai relevante și mai precise în explicarea variației variabilei dependente.Modelul care utilizează doar acești patru predictori (Dansabilitate, Nivelul de vorbire, Nivelul de audiență live și Energie) prezintă cea mai mare acuratețe în comparație cu modelul care utilizează toți cei șase predictori. Cu toate acestea, diferențele în performanță între cele două modele nu sunt majore.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   64696414    8008731   8.078 6.95e-16 ***
Danceability 100102616   11291109   8.866  < 2e-16 ***
Speechiness  -75235439   17025968  -4.419 9.98e-06 ***
Liveness     -54057342   10917243  -4.952 7.42e-07 ***
Energy        43701911    8637269   5.060 4.24e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 245300000 on 19544 degrees of freedom
Multiple R-squared:  0.008412,  Adjusted R-squared:  0.008209
F-statistic: 41.45 on 4 and 19544 DF,  p-value: < 2.2e-16
```

Deci, funcția va fi de forma:

Ystream = 64696414 + 100102616 * Xdanceability + (-75235439) x Xspeechiness + (-54057342) * Xliveness + 43701911 * Xenergy



Încercăm să efectuăm o predicție cu privire la numărul de stream-uri pe care o piesă îl va avea în cazul în care are un nivel de dansabilitate de 0.4, nivelul de audiență live de 0.3, nivelul de energie de 0.9 și nivelul de vorbire de 0.5.

```
> predict(reg_final, newdata = pred_stream, interval="confidence")
       fit      lwr       upr
1 90234258 73843553 106624963
> predict(reg_final, newdata = pred_stream, interval="prediction")
       fit        lwr       upr
1 90234258 -390943324 571411840
>
```

Numărul estimat al stream-urilor este de 90,234,258, situându-se în intervalul de încredere [73,843,553 - 106,624,963] și în intervalul de predicție [-390,943,324 - 571,411,840].

Această constatare, însoțită de o valoare relativ mare a RSE (rădăcina pătrată a erorii medii pătratice) și o valoare relativ mică a R^2 (coeficientul de determinare), indică faptul că există o serie de factori suplimentari care influențează în mod semnificativ numărul de stream-uri al unei piese, în afara celor considerați în modelul de regresie liniară. Acești factori neglijați pot juca un rol important în determinarea succesului unei piese și pot explica variația mare a numărului de stream-uri.

Având în vedere toate informațiile și analizele prezentate, putem formula următoarele răspunsuri la întrebările formulate:

1. Care este relația dintre variabilele Danceability, Energy, Speechiness, Duration, Liveness și Tempo și numărul de stream-uri al unei piese?

- Analiza noastră indică existența unei legături între aceste variabile și numărul de stream-uri al unei piese. Cu toate acestea, influența lor pare să fie mai puțin semnificativă decât s-ar fi anticipat. Alți factori necunoscuți sau neglignați pot juca un rol mai important în determinarea numărului de stream-uri.

2. Cum pot fi aceste variabile utilizate pentru a prezice numărul de stream-uri al unei piese?
- Prin utilizarea unui model de regresie liniară cu multiple variabile putem obține estimări și un interval de încredere pentru numărul de stream-uri al unei piese, acesta având în vedere valorile acestor variabile. Cu toate acestea trebuie sa fim atenti că modelul nostru prezintă o precizie limitată și există alți factori care pot influența rezultatele.

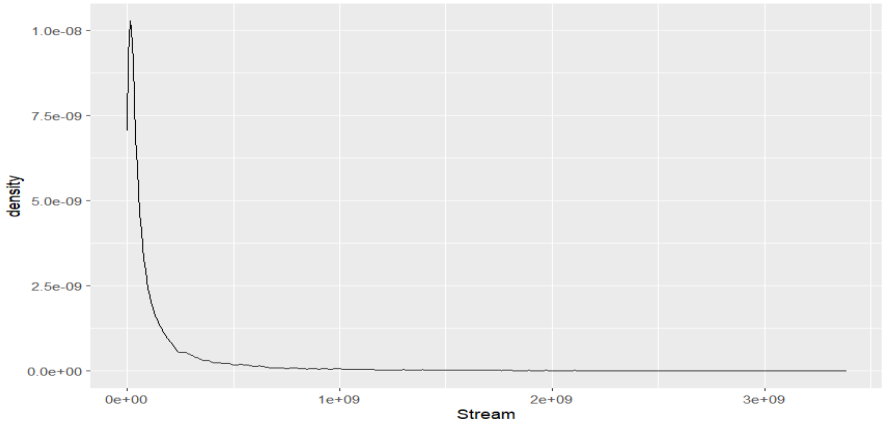
3. Care este gradul de acuratețe al modelului nostru în prezicerea numărului de stream-uri?
- Modelul nostru prezintă o acuratețe moderată în prezicerea numărului de stream-uri al unei piese. Precizia acestuia este influențată de mai mulțe variabile inclusiv variabilitatea datelor și posibilitatea de a exista alti factori neconsiderați în model.

4. Ce informații suplimentare ar trebui luate în considerare pentru a înțelege mai bine succesul fluxului titlului tău?
- Pentru a obține o înțelegere mai largă a succesului unei piese în ceea ce privește fluxurile, popularitatea artistului, preferințele consumatorilor și alțe variabile importante precum tendințele din industria muzicală. Integrarea acestor informații suplimentare va oferi o înțelegere mai buna a impactului asupra numărului de fluxuri de piese.

Arborii de decizie:

Cea de a doua metoda pe care am ales-o pentru a raspunde la intrebariile propuse in proiectul nostru este metoda arborilor de decizie. Am ales aceasta metoda deoarece are un grad ridicat de interpretabilitate al rezultatelor, iar avand in vedere faptul ca persoanele interesate in mod special de rezultatele cercetarii noastre sunt din domeniul muzical sau pur si simplu, simplii ascultatori de muzica, care nu au un grad mare de cunostinte in domeniul statisticii am decis ca interpretabilitatea joaca un rol mai important decat exactitatea datelor.

Primul pas în analiza noastră bazată pe arbori decizionali este generarea unui grafic de densitate pentru a observa distribuția stream-urilor folosind un grafic de densitate:



Din grafic putem observa că majoritatea numărului de stream-uri al pieselor se încadrează în intervalul [0, 1e+09], sau mai concret, între 0 și 1,000,000.

Următorul pas în cercetarea noastră constă în împărțirea setului de date în setul de antrenament (90%) și setul de test (10%). Apoi, vom crea efectiv arborele de antrenament utilizând algoritmul specific. Acest algoritm generează mai mulți arbori prin intermediul cross-validation și selectează arborele și valoarea parametrului cost-complexitate (α) care minimizează suma pătratelor erorilor (SSE) plus α multiplicat cu numărul de noduri terminale dinarbore (|T).

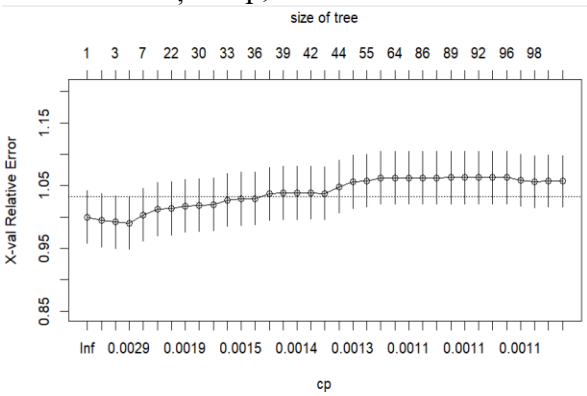
```
node), split, n, deviance, yval
      * denotes terminal node

      1) root 13684 8.438238e+20  138147900
        2) Duration_ms< 156776.5 1590 5.033035e+19   82715900
          4) Danceability< 0.5905 589 5.711328e+18   55484850 *
          5) Danceability>=0.5905 1001 4.392526e+19   98738970
           10) Danceability>=0.5935 992 3.265560e+19   93482670 *
           11) Danceability< 0.5935 9 8.221326e+18  678099000 *
        3) Duration_ms>=156776.5 12094 7.879655e+20  145435500
          6) Duration_ms>=272581.5 2306 7.829088e+19  108592000
           12) Danceability< 0.9035 2282 7.241333e+19  106169300 *
           13) Danceability>=0.9035 24 4.590561e+18  338953500
             26) Energy>=0.685 10 1.500302e+17  113343800 *
             27) Energy< 0.685 14 3.567964e+18  500103200 *
          7) Duration_ms< 272581.5 9788 7.058069e+20  154115700
           14) Danceability< 0.5715 3155 1.626523e+20  130698900
             28) Duration_ms< 172643 274 2.849334e+18   71625030 *
             29) Duration_ms>=172643 2881 1.587559e+20  136317200
```

Iată hyperlink-ul către imaginea arborelui de decizie: Arbore de decizie.

Analizând datele generate despre arbore (m1), putem observa că principalul criteriu după care se generează numărul de stream-uri al unei piese este durata mai mică de 156776.5 milisecunde, echivalentul a 2.61 minute, iar apoi urmează nivelul de dansabilitate al piesei.

Conform graficului de mai jos, se observă că valoarea cea mai mică a SSE (suma pătratelor erorilor) este obținută atunci când α (parametrul cost-complexitate) are valoarea de 1.00. În același timp, dimensiunea arborelui asociată acestei valori de α este de 0.0014.



În continuare vom încerca să găsim cel mai bun arbore pentru setul nostru de date de antrenament, utilizând parametrii minsplit (cu un număr minim de instanțe de 20) și maxdepth (cu un număr maxim de noduri interne, între rădăcină și ultima frunză, totalizând maxim 30).

În urma căutării în funcție de parametrii menționați anterior, prin utilizarea metodei hyper_grid, am obținut rezultatul cel mai optim, adică arborele cu următoarele valori: minsplit = 5, maxdepth = 8, cp = 0.006551022, iar în urma predicției avem ca rezultat eroarea de 248,313,968 pentru cele 5865 de instanțe din setul nostru de antrenament.

Observăm totuși că valoarea obținută în urma calculului hyper_grid este identică cu valoarea calculată inițial, deci putem constata că arborele inițial și cel optim sunt egale.

Din moment ce predicția noastră indică o valoare foarte mare a parametrului RMSE (248,313,968), putem concluziona că modelul de arbore de decizie construit nu oferă un nivel înalt de acuratețe în ceea ce privește predicțiile.

Concluzii:

În urma aplicării atât a regresiei liniare, cât și a arborelui de decizie, s-a ajuns la o concluzie similară: modelele rezultate, deși valide, nu sunt foarte fiabile în ceea ce privește exactitatea predicțiilor. Cu toate acestea, modelele pot fi utile pentru a explora și a explica datele din setul de date.

Astfel, utilizarea acestor modele poate oferi o înțelegere mai bună a relațiilor dintre variabilele de intrare (danceability, energy, speechiness, liveness, tempo, durată) și numărul de stream-uri al unei piese. Cu toate acestea, este important de subliniat că nu există un singur criteriu distinct care să influențeze în mod decisiv numărul de stream-uri. Există o multitudine de alte criterii care pot influența acest rezultat.

Prin urmare, aceste modele pot fi folosite ca instrumente utile pentru a explora și a explica relațiile complexe între variabile și numărul de stream-uri, dar trebuie utilizate cu prudență în ceea ce privește realizarea de predicții precise. Este recomandat să se evalueze și să se interpreteze rezultatele modelelor în contextul specific al domeniului și datelor disponibile.