



**Zene Popularitás Predikció egy Nemzetközi Toplistán Túlélés-elemzés és
Gépi-tanuló Modellek Alkalmazásával**

TUDOMÁNYOS DIÁKKÖRI KONFERENCIA

**(Music Popularity Prediction in an International Top List Using Survival
Analysis and Machine-learning Models)**

Palásti László

Pénzügy és Számvitel, Bsc, 2. évfolyam,

Konzulens: Madari Zoltán

Dátum: 2024.04.14.

Absztrakt:

A modern zenei streaming vállalatok által létrehozott Big Data adatbázisok okán leegyszerűsödött egy dal teljesítményének, valamint az egyes felhasználók fogyasztási szokásainak elemzése. Elsődleges célom annak feltérképezése és bemutatása, milyen zenei, szövegi, illetve egyéb jellemzők befolyásolják leginkább egy dal túlélését a nemzetközi toplistákon. Egy dal népszerűségének egyik fő indikátora az ilyen toplistákon való szereplés, így a zeneipari vállalatok zene-portfóliójuk kialakítása során érdekeltek abban, hogy ismerjék az egyes jellemzők hatását a dalok toplistákon való teljesítményére. Erre a kérdésre a dalok érzelmi elemzése, valamint parametrikus és nem-parametrikus túlélés-regressziója által válaszolok. A lemezkiadó cégek szintén érdekeltek abban, hogy a már elkészült termékeik számára felállítsanak olyan benchmarkokat, amelyekhez képest teljesítményük jól mérhető lesz. Második kutatási kérdésem keretein belül olyan statisztikai módszer keresése a célom, amely segítségével a legnagyobb pontosságú becslést lehet adni egy adott dal túlélésére. Ezt mind parametrikus és nem-parametrikus hagyományos túlélés-modellek (Cox-proportionális kockázati modell, Weibull-AFT modell), valamint egyéb gépi-tanulási túlélés-modellek (véletlen túlélés-erdő, DeepHit, Deepsurv) összehasonlításával és kiértékelésével végzem el. Eredményeim szerint a dalok túlélésének becslésére a legnagyobb becslési pontosságot a Random Survival Forest, valamint a DeepSurv modellek képesek túlteljesíteni a hagyományos túlélés modellek becslőképességét. Az egyes változók hatásai közül kiemelendő az explicit dalok és a Covid-19 kitörését követően a listára felkerült dalok rosszabb teljesítménye, valamint az, hogy a fogyasztók jelentősen pozitívan értékelték azokat a dalokat, amelyek esetén a ritmikus, vagy zenei pozitívítási értékek magasak.

Kulcsszavak: Túlélés-Elemzés, Gépi Tanulás, DeepSurv, DeepHit, Random Survival Forest, zeneipar

JEL kódok: L82, C55, D12, C45, C52, C53

Abstract:

The Big Data databases created by modern music streaming companies have simplified the analysis of a song's performance and the consumption habits of individual users. My primary objective is to identify and demonstrate which musical, lyrical and other characteristics have the most influence on the survival of a song in the international charts. One of the main indicators of a song's popularity is its presence in such charts, and music companies are interested in knowing the impact of certain characteristics on the performance of a song in the charts when designing their music portfolio. I answer this question through an emotional analysis of songs and parametric and non-parametric survival regressions. Record labels are also interested in setting up benchmarks for their existing products against which their performance can be well measured. My second research question is to find a statistical method to give the most accurate estimate of the survival of a song. I will do this by comparing and evaluating both parametric and non-parametric traditional survival models (Cox proportional hazards model, Weibull-AFT model) as well as other machine-learning survival models (random survival forest, DeepHit, Deepsurv). My results show that the Random Survival Forest and DeepSurv models are able to outperform the estimation accuracy of traditional survival models in estimating song survival. Among the effects of individual variables, we highlight the poorer performance of explicit songs and songs that were added to the list after the Covid-19 outbreak, and the fact that consumers rated songs with high rhythmicity or musical positivity scores significantly positively.

Keywords: survival analysis, machine learning, DeepSurv, DeepHit, Random Survival Forest, music industry

JEL codes: L82, C55, D12, C45, C52, C53

Tartalom

0. Jegyzékek.....	F
1. Ábrajegyzék.....	F
2. Táblázatjegyzék.....	F
I. Bevezetés.....	F
II. Szakirodalmi áttekintés:.....	2
III. Kutatási kérdések és kiinduló hipotézisek:	4
IV. Alkalmazott statisztikai módszertan	5
1. Szentiment elemzés – VADER Lexikon használata	5
2. Túlélés-elemzés	6
a) Kaplan-Meier modell.....	8
b) Log rank tesztstatisztika.....	8
c) Cox-Arányos Hazárdmodell	9
d) Weibull-AFT (Accelerated Failure Time) modell	9
3. Gépi tanulást alkalmazó túlélésmodellek.....	10
a) Random Survival Forest.....	10
b) DeepHit	11
c) DeepSurv	12
V. Adatbázis bemutatása, adatbeszerzési problémák:.....	14
1. Leíró statisztika az adatbázisról (főleg kimeneti változó).....	15
VI. Szentiment elemzés	19
VII. Kaplan-Meier túlélésgörbe	22
1. Explicit és nem explicit dalok túlélésének vizsgálata	23
2. Év szerinti bontású túlélés egyezőség vizsgálat	24
VIII. Magyarázó változók parciális hatásának elemzése.....	26
1. Cox-Arányos hazárdmodell.....	26
2. Weibull-AFT modell	28
IX. Gépi-tanuló modellek túlélési valószínűségek becslésére.....	31
1. Random Survival Forest	31
2. DeepHit	33
3. DeepSurv.....	34
X. Összegzés.....	36
XI. Limitációk.....	37

XII. Hivatkozásjegyzék	38
------------------------------	----

0. Jegyzékek

1. Ábrajegyzék

1. ábra: DeepHit modell felépítése. Forrás: https://doi.org/10.1609/aaai.v32i1.11842	12
2. ábra: Egész adatsor eredményváltozásának eloszlásdiagramjai.....	16
3. ábra: Szűkített adatsor eloszlásdiagramjai	17
4. ábra: Magyarázó változók eloszlásdiagramjai	17
5. ábra: Korrelációs mátrix a modell változóival	18
6. ábra: Abszolút szószám eloszlása.....	19
7. ábra: Dalokban szereplő eltérő szavak számának eloszlása.....	19
8. ábra: Abszolút szószám eloszlása.....	19
9. ábra: Pozitív, negatív, neutrális szentiment értékek eloszlásfüggvényei éves bontásban	21
10. ábra: Kaplan-Meier túlélésgörbe az egész adatsorról a kimeneti változó kvartilis értékeivel	22
11. ábra: Explicit és nem explicit dalok túlélésfüggvényei.....	23
12. ábra: Egyes években listára került dalok túlélésfüggvényei, a kimeneti változó kvartilis értékeivel	24
13. ábra: Cox-modell alapfeltételt kielégítő magyarázó változók hatása egy dal hazárfüggvényére	27
14. ábra: Egyes magyarázó változók hatása a túlélésfüggvényre a Weibull-AFT modellben .	31
15. ábra: A tesztalmaz 10 elemű mintájára készített becsült túlélésfüggvény az RSF modell alapján	33
16. ábra: DeepHit modell (1 réteg 32 neuron; 2 réteg 64-64 neuron) veszteségfüggvénye.....	34
17. ábra: DeepSurv modell (2 rejtett réteg: 32-32 neuron) veszteségfüggvénye	35

2. Táblázatjegyzék

1. táblázat: Modellben szereplő változók leírása.....	14
2. táblázat: Pozitív és negatív szentimentértékek kvartilis értékei éves bontásban.....	20
3. táblázat: Kaplan-Meier túlélésmodell túlélési valószínűség becslése.....	22
4. táblázat: Egyes évszám párok páronkénti log rank próba empirikus szignifikanciaszintjei	25
5. táblázat: Az egyes évek becsült túlélésértékei a Kaplan-Meier modell alapján.....	25
6. táblázat: Cox-modell koefficiensek, szórásuk, p-értékeik.....	26
7. táblázat: Weibull-AFT modell koefficiensei, szórásuk, p-értékeik	28
8. táblázat: Random Survival Forest bemeneti paraméterei.....	32
9. táblázat: RSF modell átlagos változófontosságai	32
10. táblázat: DeepHit modell implementálásai és modell-értékelési paraméterei.....	33
11. táblázat: DeepSurv modell és modell-értékelési paraméterei	35
12. táblázat: Prediktív modellek teljesítmény a Harrell-féle C-index szerint	37

I. Bevezetés

A modern streaming szolgáltatók megjelenését megelőzően limitált eszköztár állt rendelkezésre a zenefogyasztás elemzésére, számos probléma nehezítette meg a folyamatot. Mivel a fogyasztási adatok ekkor kizárólag eladásokból, valamint rádióadásokból származtak, nem volt lehetőség az egyedi felhasználók szokásainak pontos feltérképezésére. A fizikai eladásokból befolyó információ sok esetben késztett a feldolgozási munkálatok okán, valamint egy album eladása esetén nem lehetett pontos információt nyújtani az egyedi dalok teljesítményéről. Szintén jelentős problémát jelentett az illegális zenefogyasztás, nem volt lehetőség a nem legális úton beszerzett albumok elemzésére.

A modern zenei streaming szolgáltatók megjelenésével, valamint az ezen szolgáltatók által létrehozott Big Data adatbázisok által átalakult a fogyasztás-elemzés, mérhetőbbé váltak zenefogyasztási szokásaink. A változások lehetőséget adnak arra, hogy pontosabban elemezzük az egyéni fogyasztói szokásokat, illetve az albumok és egyes dalok teljesítményét.

Dolgozatomban két kérdésre keresem a választ. Elsődleges célom annak feltérképezése és bemutatása, milyen zenei, szövegi, illetve egyéb jellemzők befolyásolják leginkább egy dal túlélését a nemzetközi toplistákon. A lemezkiadók és egyéb zeneipari szereplők alapvetően termékként tekintenek az általuk forgalmazott albumokra, ezért fontos számukra annak monitorozása, milyen piaci teljesítménye van az egyes termékeknek (ezen termékekben belül az egyes daloknak). Egy dal népszerűségének egyik fő indikátora a nemzetközi toplistákon való szereplés, így a zeneipari vállalatok zene-portfóliójuk kialakítása során érdekeltek abban, hogy ismerjék az egyes jellemzők hatását a dalok toplistákon való teljesítményére. Erre a kérdésre a dalok érzelmi elemzése, valamint parametrikus és nem-parametrikus túlélés-regressziója által válaszolok.

A lemezkiadó cégek szintén érdekeltek abban, hogy a már elkészült termékei számára felállítson olyan benchmarkokat, amelyekhez képest teljesítményük jól mérhető lesz. Második kutatási kérdésem keretein belül olyan statisztikai módszer keresése a célom, amely segítségével a legnagyobb pontosságú becslést lehet adni egy adott dal túlélésére. Ez mind parametrikus és nem-parametrikus hagyományos túlélés-modellek (Cox-Arányos hazard modell, Weibull-AFT modell), valamint egyéb gépi-tanulási túlélés-modellek (véletlen túlélés-erdő, DeepHit, Deepsurv) összehasonlításával végzem el.

II. Szakirodalmi áttekintés:

Ezen fejezet célja bemutatni azt, milyen eredményre jutottak azon korábban megjelent cikkek és tanulmányok, amelyeket kutatási kérdéseim megválaszolásának szempontjából relevánsnak tartok.

Az általam keresett tudományos szakirodalom három csoportra osztható. Egyrészt, mivel az adatbázisomban szereplő dalok jelentős része a Covid-19 megjelenését követően került fel a toplistára, célom volt olyan kutatások feltérképezése, amelyek a felhasználók fogyasztásában bekövetkező változásokat vették szemügyre. A fogyasztáseleméleti szakirodalom mellett a túlélés-elemzés zeneipari alkalmazásáról, valamint a zeneiparban használt egyéb gépi tanulást elemzési eszközül felhasználó népszerűség predikcióról szóló tanulmányok ismertetését tartom fontosnak.

A mély-tanuló algoritmusok túlélés-elemzésben történő felhasználásáról és felhasználási területeik ismertetéséről írt tanulmányt Wiegerebe és tsai. (2023). Tanulmányukban összehasonlították a különböző módszertan szerint (Cox-alapú, Időben diszkrét, Parametrikus, PEM-alapú, ODE-alapú, Rangsor-alapú és egyéb) becslő modelleket becslési képességük és értelmezhetőségük szempontjából. A tanulmány szerint a vizsgált modellek egyszerű túlélés-modellé szűkítése az egyes DL modelleknek, emellett a modellek a klasszikus túlélés-modelleknél flexibilisebbek. A későbbiekben említett, majd alkalmazott DL alapú modellek szintén elemét képezték Wiegerebe és tsai. (2023) tanulmányának.

Katzman és tsai. (2018) bemutatták az általuk megalkotott DeepSurv modellt, amely a Cox-Arányos házardmodell mélytanuló algoritmust felhasználó modifikációja. A megalkotott deep feed-forward neurális hálót felhasználó modellt eredetileg páciensek személyre szabott kezelés-ajánló rendszeréhez használták fel, a modell az $\hat{h}_\theta(x)$ log-valószínűség függvényre végez becslést az említett neurális hálóval. A kutatás eredménye szerint a klasszikus túlélés-modelleknél a DeepSurv pontosabb egyéni túlélés-becslést képes adni.

Egy másik, mélytanuló algoritmust felhasználó túlélés-modell a DeepHit, amelyet Lee és tsai. (2018) mutattak be tanulmányukban. A korábbi túlélés-elemző modellek bizonyos feltételezésekkel éltek a sztochasztikus folyamat alakja kapcsán, amely sok esetben nem teljesül, így egy adott modell gyakran nem alkalmazható bizonyos adatsorra. A DeepHit nem tesz feltevéseket a mögöttes sztochasztikus folyamatra vonatkozóan, és lehetővé teszi azt, hogy

a kapcsolat a kovariánsok és a kockázat(ok) között idővel megváltozzon. A legfontosabb, hogy a DeepHit zökkenőmentesen kezeli az egymással versengő kockázatokat; azaz a beállítások, amelyekben egynél több lehetséges érdekes esemény van. (Lee és tsai. 2018) Elemzésem szempontjából fontos, hogy a DeepHit által nyújtott újdonságok közül, az általam felhasznált adatsor tulajdonságai okán, nem vagyok képes a versengő kockázatok bemutatására.

A szakirodalomban nem található olyan tanulmány, amely valamely hasonló, mély-tanuló algoritmusokat alkalmazó módszer segítségével végez zene-popularitás kapcsán túlélés-bebecslést.

Egy brit toplista albumainak túlélését vizsgáló tanulmány, Strobl és Tucker (2000) írása, az egyes albumok teljes és átlagos eltöltött idejét, illetve az egyes előadók albumainak számát a toplistán mind erősen jobbra ferdének találta. A ferdeségi jellemzők mellett azt állapították meg, hogy a kezdeti siker, a szezonális kereslet, illetve az album típusa szignifikáns jellemzők egy album toplistán való túlélése szempontjából. A tanulmányban elsők között alkalmaztak túlélés-regressziós módszereket a zeneipari elemzésben, említésre méltó, hogy a periódusban kizárólag eladási, valamint egyéb piaci jellemzők álltak az írók rendelkezésére adatként, így az egyes dalok túlélésének bemutatása nem volt lehetséges.

Hasonló eredményt ért el Im és tsai. (2018) tanulmánya. Munkájuk során egy koreai toplistát vizsgáltak túlélés-elemzési szempontból. A túlélés szempontjából legszignifikánsabb változó az adott dal albumán a dal címdal jellege volt. Szintén pozitívan járul hozzá egy nagy lemezkiadó cég jelenléte az adott dalnál. A szupersztár effektus kizárólag abban az esetben volt szignifikáns, amennyiben a címdalok túlélését vizsgálták. Mind a letöltött, mind az online hallgatott zenék esetén negatív hatása volt egy dal túlélésére az illegálisan meghallgatások számának.

A dalok Spotify API rendszerén keresztül elérhető jellemzőit számos tanulmány használta fel adatként, a legtöbb esetben különféle népszerűség becslési gépi tanulási módszerek alkalmazásának bemutatására. Sharma és tsai.(2022) a Random Forest, K-Nearest Neighbour és Linear Support Vector classifier módszereket hasonlították össze az F1 érték alapján. Kutatási eredményük szerint a zenepopularitás-becslés szempontjából a Random Forest gépi tanuló módszer a leghatékonyabb.

Szintén a Spotify API adatait használja fel Rutger, Nijkamp (2018) munkája, amely többváltozós regressziót alkalmazott a népszerűségre ható tényezők vizsgálatára, az általa

alkalmazott modell szerint egy adott dal lejátszásainak számát a zenei jellemzők kis mértékben befolyásolják.

Kaimann és tsai. (2019) kutatásában 2017. január és 2020. január között vizsgálták a Spotify Global Top 200 lejátszási listán szereplő dalok túlélését. Pozitív irányú kapcsolatot fedeztek fel egy dal túlélési ideje és a dal mögött álló lemezkiadó cég nagysága között. Szintén szignifikáns a kapcsolat a túlélési idő és a verseny mértéke (többi dal teljesítménye a lejátszási listán) között. Hasonló állítható egy dal terjedésének hatásáról is.

Egy a tanulmány írásakor hiánypótló zene adatbázis (SpotGenTrack Popularity Dataset) és egy a korábbi módszertannál pontosabban becsülő mélytanuló rendszer (HitMusicNet) bemutatását olvashatjuk Gutiérrez és tsai. (2020) munkájában. A tanulmány írói egy olyan tanuló algoritmust alkottak meg, amely figyelembe veszi egy adott dal esetén a korábban elérhető eladási adatok mellett mind az egyes zenei, szövegi és meta-adatokat.

Parada-Cabaleiro és társai 2024.03.28-án publikált tanulmányukban az elmúlt 50 év zeneszövegeit vizsgálták. Tanulmányuk angol nyelvű zeneszövegeket vizsgál öt műfajban, szövegi komplexitás, szentiment értékek, népszerűség szempontjából. Eredményeik szerint idővel a pop műfajú dalok zenei komplexitása csökkent mind lexikálisan, mind strukturálisan. Ezen felül a dalok érzelmi tónusa negatívabb lett, személyesebb zeneszövegekkel.

III. Kutatási kérdések és kiinduló hipotézisek:

Dolgozatom célja olyan hipotézisek vizsgálata, valamint olyan kutatási kérdések körüljárása, amelyek segítenek egy dal jellemzőinek a dal toplistán való túlélésére gyakorolt hatásának megértésében. Alapvető kérdésem az, hogy az egyes zenei jellemzők közül mely növeli egy dal túlélési idejét, illetve mely csökkenti. Mivel szövegelemzési módszertant is felhasználok elemzésem során, megvizsgálom, van-e fellelhető szövegi trend a Global Top 200-ra a vizsgált periódusban felkerült dalok között.

Dolgozatom elemzési feladatok mellett modellszelekciós feladatot is végez, céloom meghatározni, mely gépi tanuló algoritmus képes a legpontosabb túlélésbecslésre, valamint melyik képes két adatpontot a legnagyobb pontossággal helyesen sorba rendezni túlélési idejük alapján.

Az általam vizsgált periódusnak része a Covid-19 megjelenését, valamint az első lezárásokat követő időszak. Elemzésem során a felsoroltak mellett választ keresek arra, volt-e különbség a Covid-19 megjelenését megelőző, valamint az ezt követő időszakok között.

A kutatási kérdéseimre az alábbi kiinduló hipotézisek vizsgálatával keresem a választ:

- A Covid időszakban a Global Top 200 lejátszási listára felkerült dalokra jellemzőbb a negatív szentiment érték, mint a nem ebben az időszakban a listára került dalok esetén.
- A Covid időszakban a listára felkerült dalok esetén kisebb volt az egy dal általi átlagos túlélési idő, mint az ezt megelőző időszakban felkerült dalok között.
- A mélytanuló algoritmusokat használó becselő modellek pontosabb túlélés-becslést képesek adni, mint egy hagyományos túlélés-modell.
- A vizsgált időszakban is kijelenthető, hogy csökkent a dalok szövegi komplexitása.

Látható, hogy az általam megválaszolni kívánt kutatási kérdések túlélés-elemzés centrikus megválaszolásával korábban nem foglalkozott a szakirodalom. Jelenleg a Spotify API által elérhetővé tett jellemzők túlélésre való hatásait, valamint az új, prediktív mélytanuló modellek teljesítményét nem tárták még fel, ebből kifolyólag munkám mind a változók hatásának vizsgálata, mind a modell implementáció szempontjából releváns a szakirodalomban.

IV. Alkalmazott statisztikai módszertan

1. Szentiment elemzés – VADER Lexikon használata

A szentiment elemzés a Természetes Nyelvi Feldolgozás (NLP) által használt módszer, amelynek célja a szöveges adatok szentiment vagy érzelmi tónusának meghatározása. Az egyes szövegelemeket megfelelő egységekre lebontva klasszifikálja a pozitív, neutrális, negatív szentiment értékű klaszterek egyikébe. A zeneszövegek szentiment elemzésével feltárhatunk szövegi trendeket egyes előadók, műfajok, toplisták dalaiban, illetve a kapott szentiment értékeket prediktív modelljeinkben magyarázó változóként használhatjuk.

A zeneszövegek szövegi komplexitásának elemzésére a Type-Token rátát (TTR) használom, amely egy egyszerű természetes nyelvfeldolgozási modell, amely azt adja meg, hogy egy szövegben milyen arányát képzik az egyedi szavak az abszolút szószámnak. A TTR egyhez közeli értéke magas szövegi komplexitásra utal, a nullához közeli érték esetén a szövegben sok ismétlés található. (Furner, 2016)

A szentiment értékek meghatározásához szükség van valamilyen szótárra. Minden szótár rendelkezik valamilyen, általában egyedi szentiment rendszerrel, sok esetben egy adott szótár bizonyos szövegekre specializáltan készült el, például a Loughran-lexikon kifejezetten pénzügyi és közgazdasági bejelentések, egyéb szövegek elemzésére használatos. Dolgozatom során a VADER (Valence Aware Dictionary and sEntiment Reasoner) szótárat fogom használni, amely elsődlegesen közösségi-média szövegek elemzésére lett megalkotva. Mivel dolgozatom során toplistán szereplő dalok szentiment értékeit fogom felhasználni, olyan szótárra volt szükség, amely minél több szlengszót tartalmaz.

A VADER klaszterezési módszere során minden szövegelemet (ez például lehet szó, szókapcsolat, vagy sor) ellát egy valence értékkel, azaz a szöveg érzelmi tartalmát leíró értékkel, majd ezt az értéket normalizálja (X). Ez a normalizált, súlyozott összegző módszer abban az esetben hasznos, ha az elemzés során célunk minden mondathoz egy egydimenziós szentiment értéket csatolni.

Ha $X \geq 0,5$, a szövegelem szentiment értéke pozitív.

Ha $X \in [-0,5; 0,5)$, a szövegelem szentiment értéke neutrális.

Ha $X < -0,5$, a szövegelem szentiment értéke negatív. (Hutto és Gilbert, 2014)

Elemzésem során, a VADER alkalmazásához a Python NLTK (Natural Language Toolkit) csomagját használtam fel.

2. Túlélés-elemzés

A túlélés-elemzés, vagy eseményig bekövetkező-idő-elemzés egy statisztikai módszertan, amely azt vizsgálja, hogy mekkora az egy adott eseményig eltelt idő mértéke. A vizsgált populáció minden egyedéhez (ez lehet egy orvosi tanulmány során a beteg, mérnöki tanulmány során a gép, pénzügyi tanulmány során a hitelfelvevő) feltételezzük, hogy köthető egy tényleges túlélési időpont (Y), valamint egy tényleges cenzúra időpont (C). A túlélési idő a kritikus esemény (orvosi tanulmány során a beteg halála vagy betegségbe visszaesése, mérnöki tanulmány során a gép elromlása, pénzügyi tanulmány esetén a hitelfelvevő csődjelentése) bekövetkezéséig eltelt időt jelenti, a cenzúra pedig azt az időpontot, amikor az egyed kiesik a vizsgálatból. (Egyed kiesik a vizsgálati időszak során a kritikus esemény bekövetkezése nélkül, vagy a kritikus esemény az egyed esetén be sem következik a vizsgálati idő során.)

$$\delta = \begin{cases} 1, & \text{if } T \leq C \\ 0, & \text{if } T > C \end{cases}$$

A cenzúra torzítása miatt megalkotunk egy státusz-indikátort (δ), amely értéke 1, amennyiben a vizsgálati időszakban megfigyelhető a kritikus esemény, illetve 0, amennyiben ez nem következik be. A túlélés-elemzés során három, egymásba transzformálható függvényt becsülhetünk: a túlélés függvényt, a kumulált kockázat függvényt, valamint a kumulált kockázat függvényt.

A túlélés függvény vagy túlélés görbe egy csökkenő meredekségű függvény, amely a t időpontnál tovább való túlélést mutatja meg.

$$S(t) = P(T > t)$$

A hazard függvény a pontban t utáni halálozási rátát jelenti, feltéve, hogy az egyed T időpontig nem halt meg.

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}$$

Hazárd függvény kapcsolata a túlélés függvénnyel:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} =$$

$$\lim_{\Delta t \rightarrow 0} \frac{\frac{P((t < T \leq t + \Delta t) \cap (T > t))}{\Delta t}}{\frac{P(T > t)}{P(T > t)}} = \lim_{\Delta t \rightarrow 0} \frac{\frac{P(t < T \leq t + \Delta t)}{\Delta t}}{\frac{S(t)}{S(t)}} = \frac{f(t)}{S(t)},$$

ahol $f(t)$ a sűrűségfüggvény t pontban felvett értéke, azaz annak valószínűsége, hogy az egyed esetén a kritikus esemény rögtön a t időpontot követően következik be.

A kumulált hazárd függvény a hazárd a t időpontig bekövetkező halálozások összegét jelenti:

$$H(t) = \int_0^t h(x) dx$$

Bebizonyítható, hogy a túlélés függvény és a kumulált kockázat függvény közötti kapcsolat a következő:

$$S(t) = e^{-H(t)}$$

a) Kaplan-Meier modell

A Kaplan-Meier modell egy nemparametrikus túlélés-modell, amely egy adott adatsor túlélés becslésére használatos.

- d_k = k időpontban kiesett (meghalt) betegek száma
- r_k = a d_k időpont előtt a modellből még nem kiesett betegek száma (risk set)

A teljes valószínűség tétele miatt,

$$P(T > d_k) = P(T > d_k | T > d_{k-1})P(T > d_{k-1}) + P(T > d_k | T \leq d_{k-1})P(T \leq d_{k-1}),$$

Felhasználva, hogy d_i monoton fogyó,

$$S(d_k) = P(T > d_k) = P(T > d_k | T > d_{k-1})P(T > d_{k-1}), \text{ ahol } P(T > d_{k-1}) = S(d_{k-1}).$$

Eszerint viszont

$$S(d_k) = P(T > d_k | T > d_{k-1}) * \dots * P(T > d_2 | T > d_1)P(T > d_1).$$

A **Kaplan-Meier közelítő** felhasználja, hogy $\hat{P}(T > d_i | T > d_{i-1}) = (r_i - d_i)/r_i$,

egyenlő:

$$\hat{S}(dk) = \prod_{j=1}^k \frac{r_j - d_j}{r_j}$$

A Kaplan-Meier modell kizárólag egész adatsorra vetített túlélés-becslésre képes, azaz az egyedi betegek egyéb magyarázó változóit nem veszi figyelembe. Amennyiben az egyes betegekre vetítve célunk túlélést becsülni, más modelleket kell használnunk. (Kaplan és Meier, 1958)

b) Log rank tesztstatisztika

A Log rank tesztstatisztika egy kétmintás statisztikai próba, amellyel két populáció túlélésfüggvényének egyezőségét vizsgáljuk. Alapvetően jobboldali cenzúrázott adatok esetén használható, nemparametrikus teszt. A teszt során vizsgált hipotézisek:

- H_0 : q_1 és q_2 populáció túlélésfüggvényei megegyeznek.
- H_1 : q_1 és q_2 populáció túlélésfüggvényei nem egyeznek meg.

A modell minden egyes tapasztalt időpont esetén kiszámolja q_1 és q_2 populáció kumulált kumulatív függvényét adott $j \in J$ időpontokban. A teszt statisztika $L = \frac{\bar{X} - \mu}{\sigma(\bar{X})}$ alakú, ahol:

- $x = \sum_{j=1}^J q_{1j}$, amely a vizsgált időszak alatt az összes elhalálozott első populációba tartozó beteg,
- $\mu = \frac{r_1^J}{r_j} q_{1j}$, azaz az első populációba tartozó betegek átlagos halálozási aránya,
- $Var(\sum_{j=1}^J q_{1j}) \approx \sum_{j=1}^J (Var(q_{1j}))$, vagyis az első populáció halálozási számának varianciájának közelítése. (Bland és Altman, 2004)

c) Cox-Arányos Hazárdmodell

A Cox-Arányos Hazárdmodell (továbbiakban Cox-modell) a túlélésregressziót végrehajtó modellek egyik legnépszerűbb változata, amelynek célja egy adatbázison belül az egyes betegek túlélésének modellezése. A Cox-modell célja a kockázatfüggvény modellezése az adatbázisban szereplő magyarázó változókkal. A kockázatfüggvény egy adott t időpontban azt mondja meg, hogy egy adott beteg milyen valószínűséggel esik ki a modellből, azaz hal meg. A Cox-modell kulcsfeltételezése szerint a kockázat ráta logaritmus a magyarázó változók és egy populációs szintű alaphazárd érték lineáris függvénye. Ez matematikailag:

$$h(t | x) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p), \text{ ahol:}$$

- $h(t | x)$ az X magyarázó változókkal rendelkező beteg kockázat függvénye t időpontban ($t \in [1; p]$).
- $h_0(t)$ az alaphazárd függvény értéke, azaz azon kockázatfüggvény értéke, amely esetén minden magyarázó változó zérus.
- $\beta_1, \beta_2, \dots, \beta_n$ együtthatók az $x_1, x_2, \dots, x_n$ magyarázó változóhoz tartozó együtthatók.

Mivel egy kockázat, és nem egy túlélés modell a Cox-modell, a függvényértékek annak valószínűségét adják meg, hogy egy adott beteg a t időpontot követően közvetlenül meghal. (Cox, 1972)

d) Weibull-AFT (Accelerated Failure Time) modell

Az AFT modellek két populáció túlélését hasonlítják össze, amelyek valamilyen, α paraméterben megegyeznek.

$$S_a(t) = S_b\left(\frac{t}{\alpha(x)}\right).$$

Például amennyiben $\alpha=5$, a b populációban lévő betegek ötször olyan lassan öregszenek, mint az a populációban lévők. Természetesen lehet az α nem konstans függvény, a legáltalánosabb formája a

$$\alpha(x) = \exp(b_0) + \sum_{i=1}^n b_i x_i.$$

Az AFT modellek alkalmazásához kiválasztunk valamilyen parametrikus formát $S(t)$ -re. Dolgozatomban a Weibull-eloszlást használom fel, ebben az esetben a kockázati függvény:

$$H(t, x) = \left(\frac{t}{\alpha(x)} \right)^{\rho}$$

A Weibull-eloszlásnak két paramétere van:

- α skálázási paraméter,
- ρ a változás sebességét írja le:
 - $\rho < 1$ esetén a kiesési ráta csökken az idő előrehaladtával,
 - $\rho = 1$ esetén konstans a kiesési ráta,
 - $\rho > 1$ esetén a kiesési ráta növekszik az idő előrehaladtával.

A Weibull-AFT modell segítségével a túlélést becsüljük, így az egyes koefficiensek azt mutatják a modellben, hogy az adott magyarázó változó ceteris paribus változásával hányszorosára nő a túlélési idő. (Liu és tsai. 2023)

3. Gépi tanulást alkalmazó túlélésmodellek

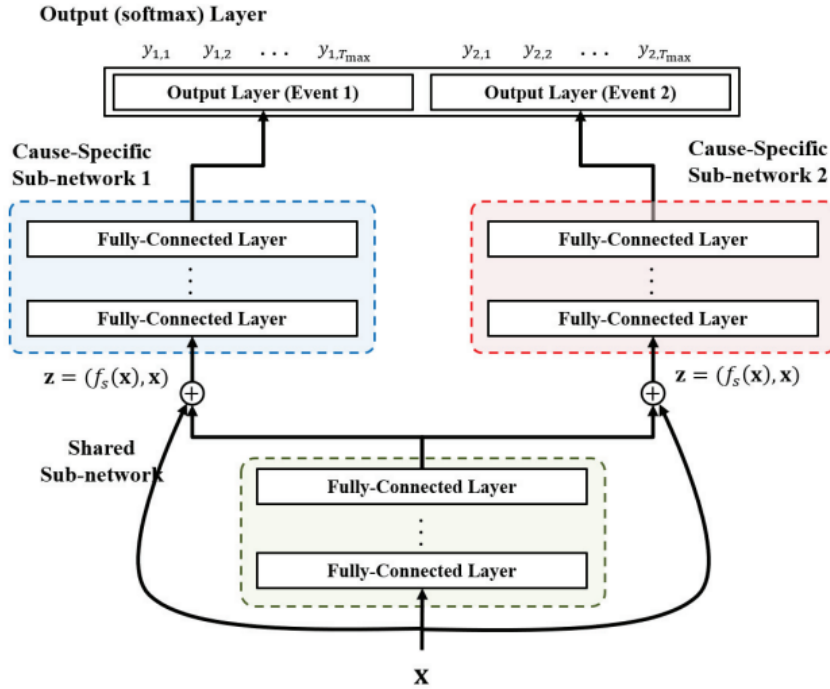
a) Random Survival Forest

Alapelveiben a Random Survival Forest a klasszifikációs és regressziós problémákat megoldó Random Forest módszer kiterjesztése, amely túlélési valószínűségek, valamint túlélésgörbék becslését végzi, kifejezetten jobboldali cenzúrárt tartalmazó túlélésadatok elemzésére fejlesztve. Ahogy a szimpla RF modellek, a Random Survival Forest modellek is döntési fák együttesét építik fel, azonban klasszifikációs vagy folytonos jellegű kimenetek előrejelzése helyett a RSF modellek minden egyes döntési fája a túlélési valószínűségeket jelzi előre különböző időpontokban.

A modell implementálása során az adatból B számú bootstrap mintát, azaz az adatbázissal megegyező elemszámú, azonban visszatevéses mintát húzunk. A szerzők munkája során az eredeti adatpontok 37%-a nem szerepeltek a mintákban, ezek lettek az Out-of-Bag adatpontok (OOB adatpontok).

A modell kiértékelésére alapvetően a Harrell féle C-indexet használjuk, amely az ROC-görbe alatti terület származtatott értéke. A modell kiértékelését a tesztalmazon, jelen esetben az OOB adatpontokon hajtjuk végre. A kiértékelés során vesszük az összes lehetséges párt az OOB adatpontok közül. Legyen $c=1$, ha egy pár esetén a becsült túlélési idő a kisebb tapasztalt túlélési idővel rendelkező adatpont esetén kisebb, ellenkező esetben $c=0.5$. amennyiben összesen p db pár van, $C-index = \frac{\sum_{i=1}^p c_{-p}}{p}$, azaz a modellel azt vizsgáljuk, hogy adott két adatpont esetén mekkora eséllyel képes a modell pontosan meghatározni, hogy a két adatpont közül melyiknek nagyobb a túlélési ideje. Minimumértéke 0.5, maximumértéke 1. A C-index 0.5-0.65 intervallumon gyengén, 0.65-0.85 intervallumon jó teljesítményt, 0.85-1 között kiváló teljesítményt jelent. (Harrell és tsai., 1982)

A DeepHit egy tensor flow környezetben létrehozott többfeladatú mélytanuló algoritmus túléléselemzési feladatok elvégzésére. A modell nem él a sztochasztikus folyamatról előzetes feltételezésekkel, a tanulási folyamat során magát a túlélési idők eloszlását tanulja meg, ebből számolja a túlélési valószínűségeket. A DeepHit modell képes egyszerre több veszélyhelyzetet kezelni, azaz olyan szituációkban is működik a modell, amikor egyszerre több célesemény jelen van. (Lee és tsai., 2018)



1. ábra: DeepHit modell felépítése. Forrás: <https://doi.org/10.1609/aaai.v32i1.11842>

Az ábrán látható a modell felépítése: egy közös rész-hálóból, a K egymással versengő célesemény saját rész-hálójából, majd egy végső kimeneti rétegből áll. A közös rész-háló minden egyes $k \in K$ célesemény rész-hálójának átadja a bemeneti halmaz leképezését, majd minden k esemény rész-hálója a közös rész-háló kimeneti értékét, valamint az eredeti magyarázó változó értékeket felhasználva kimeneti értéként vektorformában annak valószínűségét adja meg, hogy először k esemény fog bekövetkezni. A végső, kimeneti softmax függvényt felhasználó réteg vektorformában adja meg azt, hogy az egyes időpontokban mekkora az egyes k események bekövetkezésének valószínűsége.

A modell által felhasznált veszteségfüggvény cenzúrázott adatok használatára alkalmas. Mind az egész adatsor log-valószínűségét (L_1), mind az egyes k célesemény egyedi veszteségfüggvényét (L_2) felhasználja. Az L_1 segítségével a DeepHit képes az egyes adatpontok túlélési idejéről általános képet adni, az L_2 pedig a Harrell féle C-index-ből transzformáltan kapott veszteségfüggvény segítségével az egyes k célesemények egyedi veszteségfüggvényével korrigálja L_1 -et. A kettő összegeként adja a modell a teljes veszteségfüggvényt (Lee és tsai., 2018)

c) DeepSurv

A DeepSurv egy többrétegű feedforward neurális háló, amelynek célja a Cox-modell log-kockázat függvényének becslése. A modell x bemeneti értéke számos egymáshoz kapcsolódó rejtett rétegen keresztül halad, majd kimeneti értéként a $\hat{h}_\theta(x)$, vagyis a Cox-modell ismeretlen függvénye (θ maga a neurális háló jelölése). A modell tanítása során a célfüggvényként az átlagos negatív log-parciális-valószínűség függvényét vesszük a Cox-parciális-valószínűség függvénynek, $L_c(\beta) = \prod_{i=1}^n \frac{\exp(\hat{h}\beta(x_i))}{\sum_{j=1}^{r_i} \exp(\hat{h}\beta(x_j))}$ -nek, ahol n az adatpontok száma, r_i az adott t időpontban mért még meg nem halt betegek száma:

$$l(\theta) := -\frac{1}{N_{E=1}} \sum_{i:E_i=1} (\hat{h}\theta(x_i) - \log \sum_{j=1}^{r_i} e^{\hat{h}\theta(x_j)}) + \alpha \|\theta\|_2^2, \text{ ahol}$$

$N_{E=1}$ azon adatpontok száma, ahol bekövetkezett a célesemény, α pedig egy regularizációs paraméter, amely a túlillesztés elkerülése végett szerepel. A modell aktivációs függvénye a SELU, optimalizálási folyamata során az Adam optimalizálót használja. (Katzman és tsai., 2018)

A modell a negatív log-parciális-valószínűség minimalizálása segítségével képes pontos becslést adni a Cox-féle házárdfüggvényre. Mind a DeepSurv, mind a DeepHit modellek esetén a modellek teljesítményének értékelésére a C-index, valamint Brier-érték az általánosan elfogadott módszer. A Brier-érték valószínűségi becslések kiértékelésére használatos mérték, amellyel így a modell abszolút teljesítményére szolgálunk becsléssel. A Brier-érték $(B) = \frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2$, ahol N a mintaelemszám, p_i modelljeink esetén az egyes becsült túlélési valószínűség, a DeepHit esetén a direkt valószínűség, a DeepSurv esetén a Cox-modellből transzformált valószínűség. Az o_i pedig az adott i mintaelem bináris kimenetele ([nem meghalt; halt meg] = [0;1]).

A közönséges és gépi-tanuló modellek implementálására felhasznált, valamint egyéb elemzéshez felhasznált Python csomagok: *numpy*, *seaborn*, *matplotlib.pyplot*, *pandas*, *scipy.stats*, *lifelines*, *sklearn*, *sksurv*, *torch*, *pycox*.

V. Adatbázis bemutatása, adatbeszerzési problémák:

Az általam használt adatbázis az összes olyan Spotify-on elérhető dalt és alább részletezett jellemzőjét tartalmazza, amely 2017.01.01., illetve 2020.11.05. között felkerült a Global Top 200, azaz a nemzetközi 200 leghallgatottabb dal listájára. Az adatbázis saját adatbányászatom eredménye, amelyhez segítségül felhasználtam a Spotify API, valamint a Musixmatch API eszköztárát. Az adatbázis számos változójáról részletes leírás található a <https://developer.spotify.com/documentation/web-api/reference/get-audio-features> címen. Az adatbázisban megtalálható kiinduló változók bemutatására szolgál az alábbi táblázat.

1. táblázat: Modellben szereplő változók leírása

Változó	Leírás
track_url	Dal url-je (pl.: https://open.spotify.com/track/1Mzm1CNqVxtwbCYhOo843s?si=7565dd4b3834498f)
artist_URL	Előadó url-je (pl.: https://open.spotify.com/artist/0k17h0D3J5VfsdmQ1iZtE9?si=46a8aa2277d9430d)
track_name	Dal címe (pl.: Echoes – 2011 Remastered Version)
artist_name	Előadó neve (pl.: Pink Floyd)
lyrics	Dalszöveg
first_date	Első nap, amikor az adott dal szerepelt a toplistán a vizsgált periódusban.
last_date	Utolsó nap, amikor az adott dal szerepelt a toplistán a vizsgált periódusban.
position_vector	Egy vektor, amely minden az adott dal által elért pozíciót tartalmaz
mean_position	A position_vector átlagolásával kapott átlagosan elért helyezés.
artist_followers	Előadó követőinek száma a platformon.
artist_popularity	Az előadó 0-100 skálán terjedő népszerűségi mutatója, amely az előadó összes dalának népszerűségi értékéből származtatott szám.
artist_genres	Azon zenei műfajok, amelyben egy adott előadó dalokat gyárt.
explicit	Amennyiben explicit egy adott dal, értéke 1, ellenkező esetben 0.
during_covid	Azon dalok esetén, amelyek a nemzetközi Covid-19 okozta lezárásokat követően szerepeltek, értéke 1, egyébként 0.
track_pop	Egy dal 0-100 skálán terjedő népszerűségi mutatója.
danceability	Egy dal [0;1] intervallumon értelmezett, egyéb zenei jellemzők által számolt mutatója, amely azt írja le, mennyire alkalmas egy adott dal táncolásra.
energy	Egy dal [0;1] intervallumon értelmezett, ritmikai jellemzőkből származtatott mutató, amely megmutatja, milyen az adott dal gyorsasága.
instrumentalness	Egy dal [0;1] intervallumon értelmezett mutatója, amely abban az esetben nő, ha egyre kevesebb az adott dalban elhangzó szavak száma.
loudness	A dal decibelben mért átlagos hangossága.
speechiness	Egy dal [0;1] intervallumon értelmezett mutatója, amely arra ad választ, hogy a dal milyen arányban épül fel zeneszövegből.

acousticness	Annak megbízhatósági értéke, hogy egy adott dal akusztikus.
liveness	Annak megbízhatósági értéke, hogy az adott dal élő felvétel.
valence	Olyan [0;1] intervallumon értelmezett mutató, amely arra ad választ, hogy az adott dal milyen mértékben mutat zenei pozitivitást.
tempo	Egy ütések per perc (BPM) mutató.
duration_ms	A dal hossza milliszekundumokban.
key	[-1;11] skálázású mutató, amely megmutatja, milyen kulcs jellemzi az adott dalt (Amennyiben nem ismert, az értéke -1)
mode	A dal modális jellegét ismertető mutató. A dūr jelleg 1, a moll jelleg 0.
duration	Egy adott dalnak a lejátszási listában eltöltött napjainak száma.
status	Stáuszust (életben van-e az egyed) jelző mutató, mivel jelen adattábla nem tartalmaz cenzúrát, minden értéke 1.

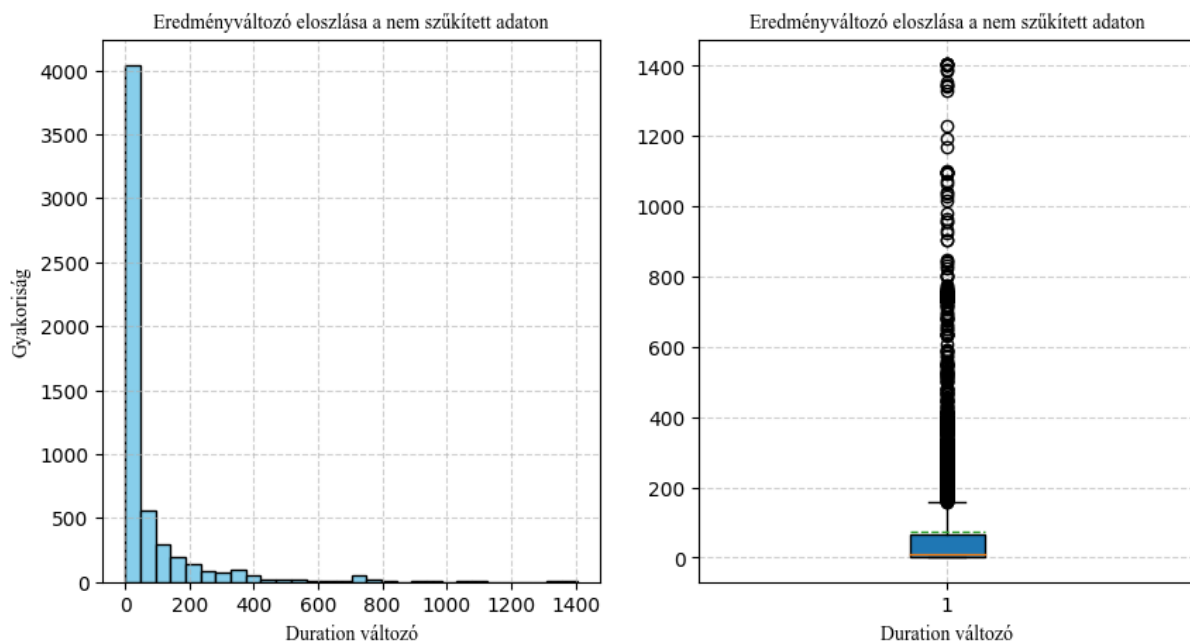
A Spotify API-segítségével elérhető dalokat a *spotipy* Python csomag segítségével töltöttem le, a letöltési szakasz ezen részén nem merültek fel problémák. A zeneszövegek letöltéséhez szükség volt egy tematikus API felhasználására, én a Musixmatch API mellett döntöttem. A letöltési folyamat során a *beautifulsoup* Python csomagot használtam fel.

A zeneszöveg adatbányászata során keletkeztek letöltési hibák, névlegesen kettő. Bizonyos dalokat nem talált meg a program, egyes, nagyobb szószámmal rendelkező dalok esetén pedig rossz adatot töltött le. A hibásan letöltött dalok adatait nem fogom vizsgálni, így a kiinduló 5917 elemű adatbázis 5736 eleművé redukálódott.

1. Leíró statisztika az adatbázisról

Az adatbázisban szereplő dalok közül a vizsgált periódusban egy adott szám átlagosan 72,90 napot tölt el a Global Top 200 lejátszási listán. Ettől az átlagos értéktől átlagosan 158,05-tel tér el a lejátszási listán való szereplési napok száma (Az eredményváltozó értelmezési tartománya a $[1;\infty)$). Az értékek terjedelme 1405. Az Q_1 értéke 2, a Q_3 értéke 65. A medián értéke 10. Mivel a medián és a Q_1 közötti távolság (8) sokkal kisebb a medián és a Q_3 közötti távolságnál (55), az eredményváltozó eloszlása balra ferde, jobbra elnyúló. Ezt bizonyítja az aszimmetria mutató 4.023 értéke is.

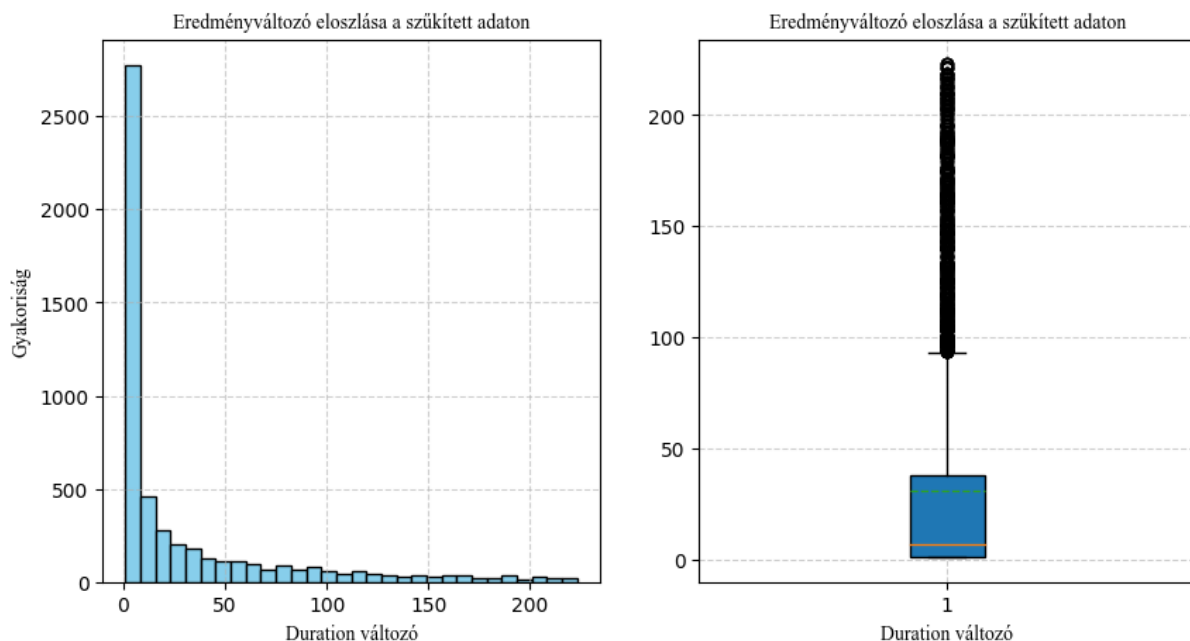
A hisztogramot és doboz-ábrát szemügyre véve látható, hogy a felső outlier értékek okozta torzítás teljes mértékben értékelhetetlenné teszi ezt a két ábrát.



2. ábra: Egész adatsor eredményváltozásának eloszlásdiagramjai

Mivel hatalmas a Q_3 és a Q_4 közötti különbség, a későbbi modellépítés pontosságának növelése céljából szűkítést végzek az adatsoron. Azokat a sorokat távolítom el az adatbázisból, amelyek esetén az eredményváltozó értéke nagyobb, mint a $Q_3 + \sigma(duration)$.

A szűkített adat 5187 sorból áll, egy dal átlagosan 30,95 napot tölt a Global Top 200 listán. Az Q_1 érték 1, a medián 7, a Q_3 érték 38. A szűkített adat esetén is elmondható, hogy az eredményváltozó mediánja és a Q_3 közötti távolság kisebb, mint a medián és az Q_1 közötti távolság. Ez, valamint a ferdeségi mutató 2,03 értéke balra ferde, jobbra elnyúló eloszlásra utalnak. Az eloszlás ferdesége jól látható a szűkített adatsor eredményváltozójának hisztogramján és dobozábráján.

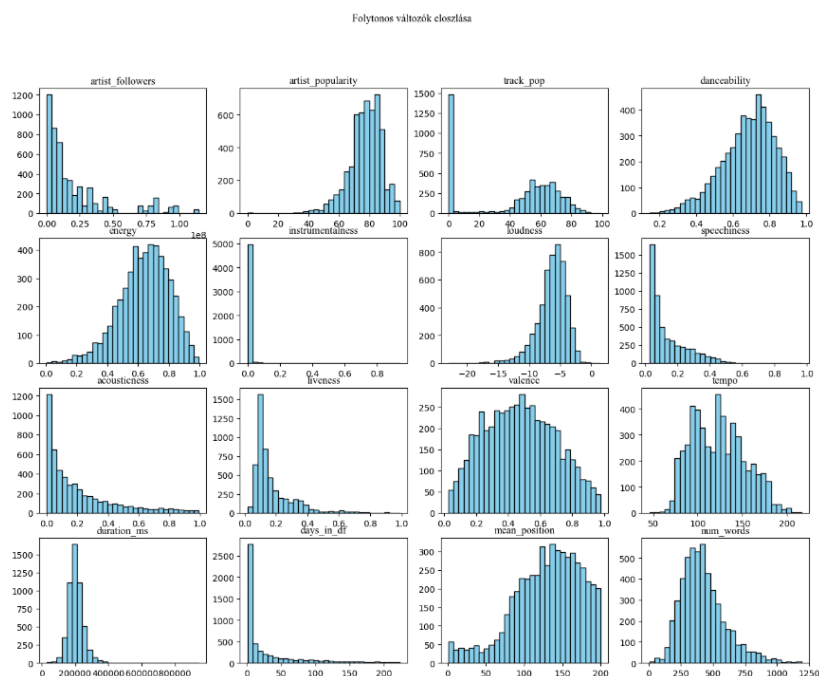


3. ábra: Szűkített adatsor eloszlásdiagramjai

Az adatbázis szűkítését követően az eredményváltozó maximumértéke 1405-ről 223-ra csökkent, így elemzésem során nem fogom vizsgálni az éveket a listán töltő dalokat, célom az tipikus teljesítményhez közelebb teljesítő dalok elemzése.

Az eredményváltozó tulajdonságainak ismertetése mellett az elemzés szempontjából fontos, hogy fogalmunk legyen az egyes magyarázó változók viselkedéséről. Az egyes hisztogramokon a felhasznált magyarázó változók hisztogramjai láthatóak.

A *track_pop* változó eloszlása érdekes képet mutat a változóról. Az eloszlás láthatóan kétmódusú, abszolút maximumát egyértelműen a nulla körül veszi fel, azonban 50-75 között szintén magasabb a gyakoriságok



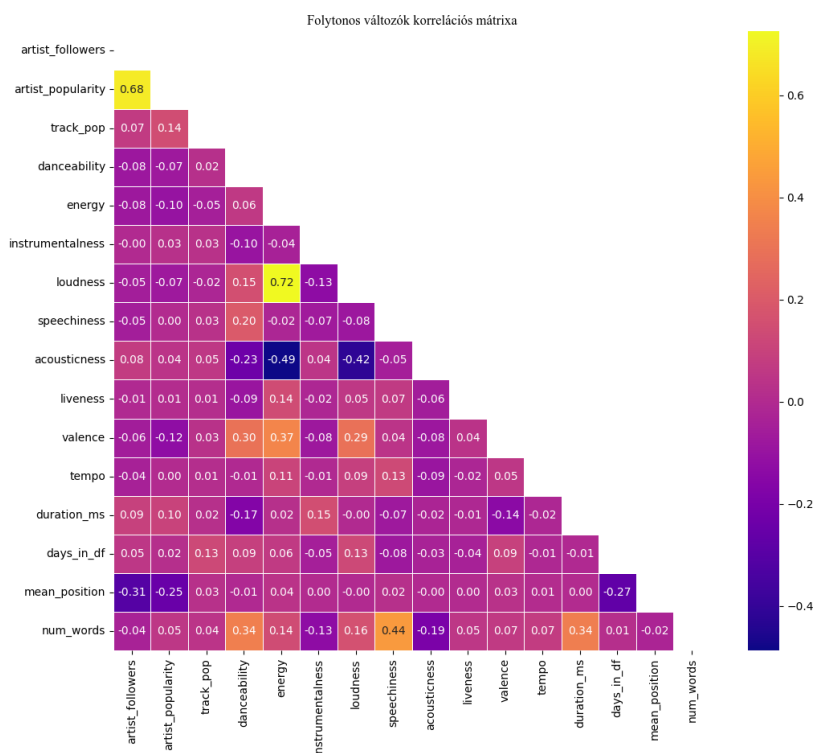
4. ábra: Magyarázó változók eloszlásdiagramjai

értéke. A változót az alapján számítják, hogy adott időpontban mennyiszer játszották a számot, így a nulla érték biztosan nem hiányzó adat reprezentációként fordul elő.

Az *acousticness* és a *danceability* változók eloszlása alapján azt állíthatjuk, hogy a vizsgált időszakban a Global Top 200 lejátszási listákon szereplő számok között nem jellemző, hogy nagy az akusztikus elemek száma a zenében. A *danceability* eloszlása alapján azt feltételezem, hogy a “táncolhatóság” olyan változó, amely jellemző a Global Top 200-on jelen lévő számokra.

A folytonos változók korrelációs mátrixában a páronkénti kétváltozós korrelációs együtthatókat találjuk, amelyek az egyes változók

között mért lineáris kapcsolat irányát, valamint erejét mutatják. Alapvetően a legtöbb érték gyenge pozitív, vagy negatív kapcsolatot mutat. Erős pozitív irányú lineáris kapcsolat van, 0,68-as értékkel, az *artist_popularity* és az *artist_followers* között. Mivel amellet, hogy a két változó közötti pozitív kapcsolat erős, hasonló magyarázó képességük van, a túlélésmodellek során nem fogom az *artist_followers* változót felhasználni. Szintén erős pozitív irányú kapcsolat van a *loudness* és az *energy* változók között, 0,72 értékkel. Gyenge lineáris



5. ábra: Korrelációs mátrix a modell változóival

kapcsolatok között nincs olyan erősségű, mint amelyet a pozitív irányú kapcsolat között lehet tapasztalni. Az *acousticness* és *energy* változók között -0,49, valamint az *acousticness* és *loudness* változók között -0,42 értékek a legjelentősebb negatív irányú kétváltozós korrelációs együtthetők.

VI. Szentiment elemzés

A kvantitatív szövegelemzés egy olyan módszer, amely számok és statisztikai módszerek felhasználásával elemzi és értelmezi a szöveges adatokat. Ennek során a szavak előfordulása, gyakorisága, összefüggései mind vizsgálati témái az elemzésnek.

A túlélésmodellek alkalmazását megelőzően célszerűnek tartom az adatbázisban szereplő dalok egy olyan jellemzője szerinti bemutatását, amelyre vonatkozóan az adatbázis alapból nem tartalmaz mérhető értéket. A zeneszövegeket hosszuk alapján, valamint szentiment értékeik alapján mutatom be, az alábbi sorrendben:

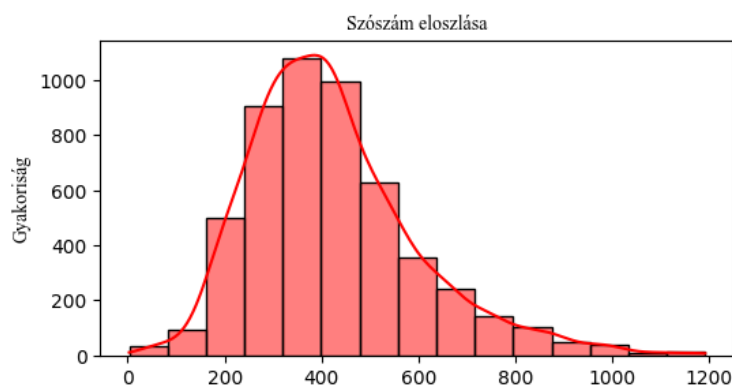
- abszolút szószám,
- abszolút különböző szavak száma,
- szövegi komplexitás,
- pozitív és negatív szentiment érték.

A vizsgált 5736 dal átlagosan 417,74 szóból áll, a Q_1 értéke 298, a Q_3 értéke 505, a medián (Q_2) 394. Mivel a medián értéke közelebb áll Q_1 -hez, mint Q_3 -hoz, az abszolút szószám balra ferde, jobbra elnyúló. Ezt mutatja be az alábbi ábra.

Az abszolút különböző szavak esetén szintén balra ferde, jobbra elnyúló eloszlást tapasztalunk. A Q_1 értéke 101, a Q_2 140, a Q_3 pedig 199.

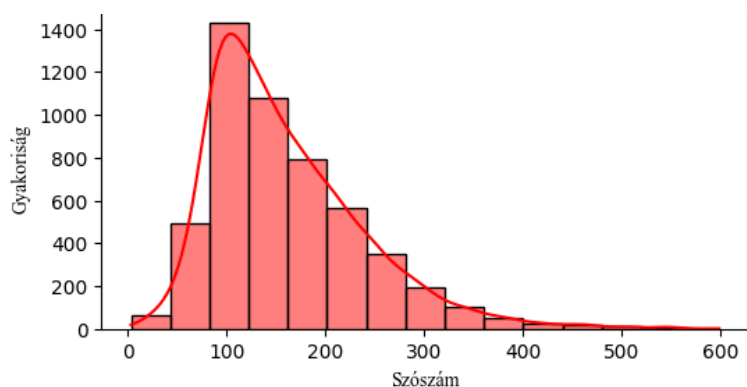
A dalok zenei komplexitását a Type-Token rátával vizsgáltam meg, amelynek átlagos értéke az adatbázisban 31,85%, azaz átlagosan 100 dalszövegben szereplő szóra 31,85 új szó jut. Egyes tanulmányok szerint hosszú távon csökkent a dalok zenei komplexitása, mivel az én adatbázisom négyéves terjedelmet fed

le, azt vizsgálom meg, ez alatt a négy éves terjedelem alatt változott-e a zenei komplexitás a Global Top 200-ra felkerült dalok esetén. A 2017-ben listára kerülő dalok esetén a TTR értéke átlagosan 29,79%, a 2020-ban a listára kerülő dalok esetén a TTR értéke 32,97%. Látszólag tehát rövid távon a 2017-ben listára került dalok átlagos zenei komplexitása alacsonyabb, mint



8. ábra: Abszolút szószám eloszlása

6. ábra: Abszolút szószám eloszlása



7. ábra: Dalokban szereplő eltérő szavak számának eloszlása

a 2020-ban a listára került dalok esetén. Annak érdekében, hogy megbizonyosodjunk arról, hogy szignifikánsan nem egyenlő a két átlagos érték, elvégzünk egy kétmintás t-próbát:

- H_0 : A 2017-ben a listára került dalok átlagos TTR értéke megegyezik a 2020-ban a listára került dalok átlagos TTR értékével.
- H_1 : A 2017-ben a listára került dalok átlagos TTR értéke nem egyezik meg a 2020-ban a listára került dalok átlagos TTR értékével.

A próbafüggvény értéke $-8,87 \cdot 10^{-18}$ empirikus szignifikanciaszinttel. Bármely szokványos szignifikanciaszinten elutasítom a nullhipotézist, így azt állíthatom, hogy szignifikáns különbség van a két átlagos TTR érték között. Rövid távon tehát Parada-Cabaleiro és tsai. 2024-es munkájával ellentmondó eredményre jutott kutatásom.

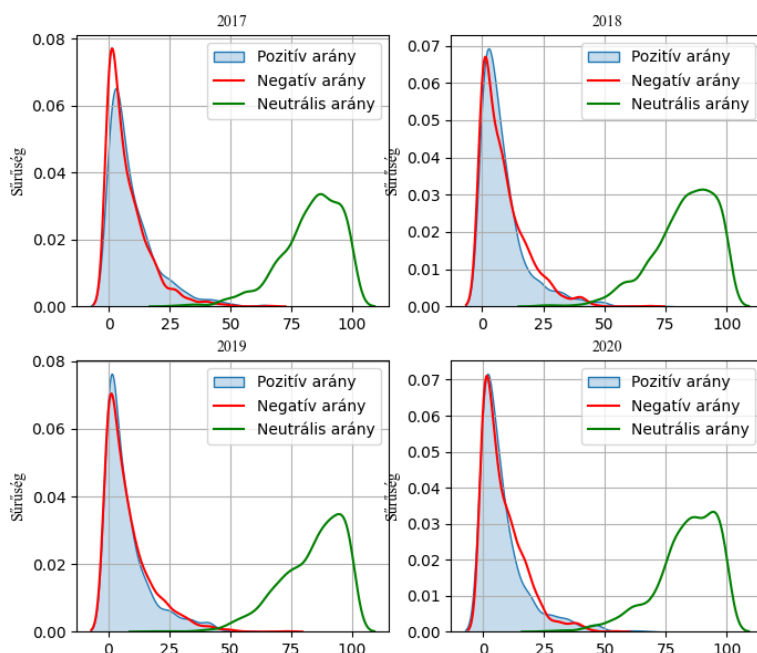
Az egyes dalok szöveghosszának bemutatását követően a dalok szentiment értékeit vizsgáltam meg. Elemzésem során a VADER (Valence Aware Dictionary and sEntiment Reasoner) lexikont alkalmaztam, amelyet a szakirodalomban általánosan a közösségi média szövegadatok elemzésére használnak.

2. táblázat: Pozitív és negatív szentimentértékek kvartilis értékei éves bontásban

	pos							
	count	mean	std	min	25%	50%	75%	max
2017	1209	9,410%	10,415%	0,000%	2,273%	6,000%	13,208%	65,625%
2018	1393	8,468%	9,900%	0,000%	1,887%	5,405%	11,111%	75,000%
2019	1290	7,875%	9,956%	0,000%	0,844%	4,324%	10,526%	66,667%
2020	1295	8,151%	10,386%	0,000%	1,156%	4,762%	10,827%	68,750%
	neg							
	count	mean	std	min	25%	50%	75%	max
2017	1209	7,502%	8,742%	0,000%	0,000%	4,598%	11,111%	66,129%
2018	1393	8,666%	9,275%	0,000%	1,515%	6,061%	12,903%	68,000%
2019	1290	8,264%	9,623%	0,000%	0,000%	5,128%	11,864%	72,881%
2020	1295	8,198%	8,609%	0,000%	1,695%	5,455%	12,445%	54,167%

Minden dal esetén soronként meghatároztam a szentiment értéket, majd ezt a [negatív, neutrális, pozitív] klaszterek egyikébe soroltam. Az általam bemutatott számértékek a dalok klasztereinek elemszám megoszlását mutatják (például: [pozitív, neutrális, negatív] = [0.3, 0.4, 0.3]). A vizsgált dalok esetén a pozitív szentiment értéke átlagosan 8,72%, a neutrális szentiment értéke átlagosan 83,28%, míg a negatív szentiment értéke átlagosan 8,001% volt.

Az ábra az egyes szentimenttípusok sűrűségfüggvényeit tartalmazza éves bontásban. Minden listára kerülési évben egyértelműen látható a pozitív és negatív szentimentértékek balra ferdesége, valamint a neutrális szentimentérték balra elnyúlása. A táblázat szerint a 2019 és 2020 az átlagos pozitív szentimentérték 0,276%-ponttal, míg a medián pozitív szentimentérték 0,483%-ponttal nőtt. Arra, hogy ez a különbség szignifikáns-e, egy kétmintás t-próbát végeztem el, ahol:



9. ábra: Pozitív, negatív, neutrális szentiment értékek eloszlásfüggvényei éves bontásban

- H0: A 2019-ben listára került és a 2020-ban listára került dalok átlagos pozitív szentimentértéke egyenlő.
- H1: A 2019-ben listára került és 2020-ban listára került dalok átlagos pozitív szentimentértéke nem egyenlő.

A próbafüggvény értéke -0,69 49,02%-os empirikus szignifikanciaszint mellett. Nem tudom ennyi információval birtokomban elutasítani a nullhipotézist, így azt állíthatom, hogy a két évben megegyezett az átlagos pozitív szentimentérték.

A negatív szentimentértékek esetén látható, hogy az átlagos szentimentérték 2019-ről 2020-ra 0,066%-ponttal csökken, a medián szentimentérték pedig 0,327%-ponttal nő. Hasonló módszerrel, kétmintás t-próbával vizsgálom a két szentimentvektor egyezőségét.

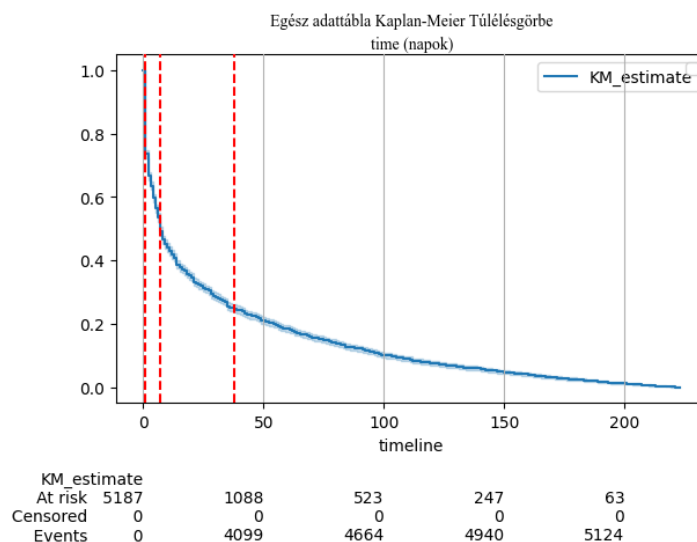
- H0: A 2019-ben és 2020-ban a listára került dalok átlagos negatív szentimentértéke egyenlő.
- H1: A 2019-ben és 2020-ban a listára került dalok átlagos negatív szentimentértéke nem egyenlő.

A próbafüggvény értéke 0,1853 85,30%-os p-érték mellett, így egyik szokásosan használt szignifikanciaszint mellett sem vagyok képes a nullhipotézis elutasítására. A rendelkezésemre álló információ alapján azt állíthatom, hogy a 2019-es évben mind a pozitív, mind a negatív szentimentértékek átlaga nem tért el szignifikánsan. Ez cáfolja egyik kiinduló hipotézisemet,

mely szerint a Covid-időszakban az emberek fogyasztási szokásai változása keretein belül magasabb negatív szentimentértékekkel bíró dalokat hallgattak.

A képzett szentiment vektorok az adatsor minden elemének szentiment értékét tartalmazzák, így a pozitív és negatív szentimenteket felhasználom, mint a magyarázó változó a túlélés modellekben.

VII. Kaplan-Meier túlélésgörbe



10. ábra: Kaplan-Meier túlélésgörbe az egész adatsorról a kimeneti változó kvartilis értékeivel

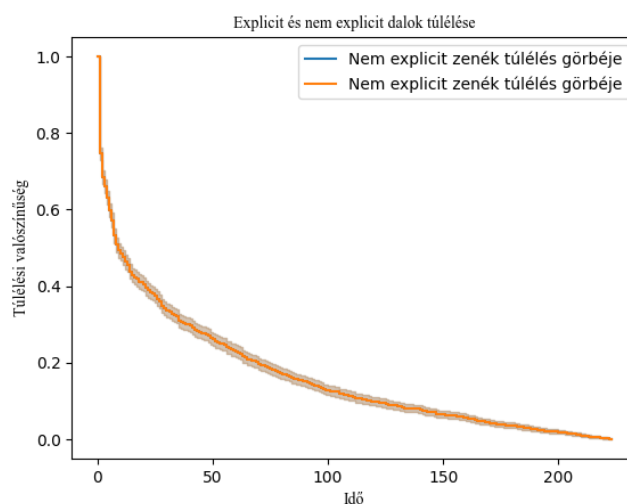
A Kaplan-Meier túlélésgörbén az egész adatsorra vonatkoztatott túlélés-bebecslést találjuk. Az első 50 nap után már 4099 dal kiesett a modellből, 100 nap után további 565, 150 nap után 276, végül a legalább 200 napig a Global Top 200 listán fennmaradó dalok száma 63.

3. táblázat: Kaplan-Meier túlélésmodell túlélési valószínűség becslése

Eltelt napok	Túlélési valószínűség
1	73,53%
2	66,71%
3	63,56%
4	59,69%
5	56,33%
10	44,11%
50	20,98%
100	10,08%
150	4,76%
200	1,21%
250	0,00%

A táblázatban a Kaplan-Meier túlélésgörbe általi becsült túlélési valószínűségek találhatóak az eltelt napok függvényeként. A modell szerint annak valószínűsége, hogy egy dal már az első napon kiesik, 26,47%. Az ötvenedik napon egy dal túlélési valószínűsége 20,98%, kétszáz nap esetén pedig 1,21%.

1. Explicit és nem explicit dalok túlélésének vizsgálata



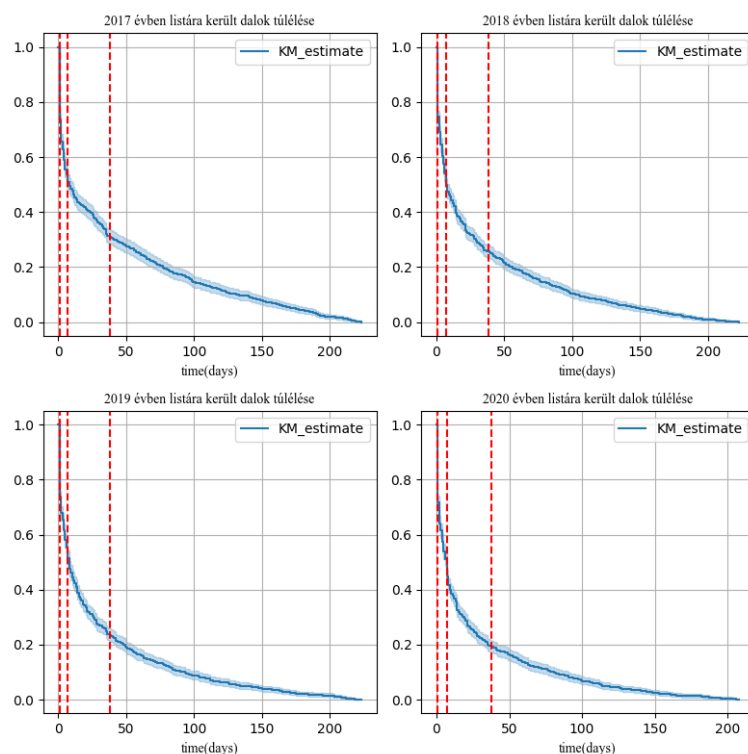
11. ábra: Explicit és nem explicit dalok túlélésfüggvényei

A Kaplan-Meier modell felhasználásával össze lehet hasonlítani bizonyos jellemzői alapján az adatsort. Először is megvizsgáljuk, szignifikánsan eltér-e az explicit dalok túlélése a nem explicit dalok túlélésétől. Az ábrán látható a két populáció túlélésgörbéje, amelyek vizuálisan nem térnek el egymástól radikálisan. Azt feltételezem, nem különbözik szignifikánsan a két populáció túlélésfüggvénye.

- H0: Az explicit és a nem explicit dalok túlélésfüggvénye egyenlő.
- H1: A két populáció túlélésfüggvénye különbözik.

A hipotézis tesztelésére a log rank tesztet végzem el, a próbafüggvény értéke 86.6, $p < 0.005$. Ez alapján bármely szokványos szignifikanciaszint mellett elutasítjuk a nullhipotézist, azaz azt állítom, hogy az explicit és a nem explicit dalok túlélése nem egyezik meg.

2. Év szerinti bontású túlélés egyezőség vizsgálat



12. ábra: Egyes években listára került dalok túlélésfüggvényei, a kimeneti változó kvartilis értékeivel

Hasonlóképpen járok el az egyes években a műsorlistára került dalok esetén. Az adatbázisban négy naptári év során a listára kerülő dalok szerepelnek, ez alapján osztottam fel a populációt négy részpulációra. Az ábrák az egyes években a listára felkerült dalok túlélésfüggvényeit, valamint függőleges vonalakkal a kvartilis értékeit mutatják. Ebben az esetben is a log rank tesztet végzem el a populációk összes két elemű kombinációja esetén.

- H_0 : i és j ($i \neq j$) populációk túlélésfüggvénye megegyezik.
- H_0 : i és j ($i \neq j$) populációk túlélésfüggvénye nem egyezik meg.

4. táblázat: Egyes évszám párok páronkénti log rank próba empirikus szignifikanciaszintjei

i	j	p érték
2017	2018	<0.005
2017	2019	<0.005
2017	2020	<0.005
2018	2019	0.31
2018	2020	<0.005
2019	2020	<0.005

A vizsgált populáció párok közül egyedül a 2018 és 2019 között a műsorlistára felkerült dalok túlélésgörbéit tekinthetjük szignifikánsan megegyezőnek, a többi populáció pár túlélésfüggvényeit bármely szokásosan használt szignifikanciaszint mellett különbözőnek tekintem.

5. táblázat: Az egyes évek becsült túlélésértékei a Kaplan-Meier modell alapján

Eltelt idő	P(2017)	P(2018)	P(2019)	P(2020)
1	73,78%	74,73%	73,72%	71,81%
2	65,67%	68,99%	67,75%	64,17%
3	62,86%	64,47%	65,27%	61,54%
4	59,14%	60,45%	61,55%	57,53%
5	55,58%	57,57%	58,06%	53,98%
10	47,73%	45,73%	44,50%	38,61%
50	27,63%	21,46%	18,99%	16,22%
100	14,47%	10,48%	8,76%	6,87%
150	7,86%	4,95%	3,95%	2,47%
200	1,90%	0,93%	1,55%	0,54%
250	0,00%	0,00%	0,00%	0,00%
300	0,00%	0,00%	0,00%	0,00%

A túlélésgörbék összehasonlítása nem ad arra vonatkozóan képet, hogy melyik évben felkerült dalok esetén kisebb, melyikben nagyobb egy adott időpontban egy dal túlélési valószínűsége, a táblázat ezt mutatja be.

Az első 5 napot követően jelentős különbségek vannak az egyes túlélési valószínűségek között, a 2019-ben a listára került dalok esetén 58,06% a túlélési valószínűség, a 2020-ban listára került dalok esetén ez az érték 53,98%.

Szintén kiemelő, hogy bármely időpont esetén tekintjük meg a valószínűségeket, a 2020-ban felkerült dalok esetén tapasztaljuk a legkisebb túlélési valószínűséget. Ez igazolja azon kiinduló hipotézisemet, miszerint a Covid-időszakban a lejártsági listára felkerült dalok túlélési ideje kisebb, mint az ezt az időszakot megelőzően a listára felkerült dalok.

Megjegyzendő azonban, hogy az általam vizsgált időintervallum 2020. év közben lezárult, így az adat biztosan tartalmaz olyan dalokat, amelyek a valóságban hosszabb időt töltöttek a lejátszási listán.

VIII. Magyarázó változók parciális hatásának elemzése

Az egész adatsorra vonatkozó becsléseket követően vizsgáljuk meg, az egyes dalok egyedi jellemzői hogyan járulnak hozzá egy adott dal túléléséhez. Ezt mind a hazard modellezésével, Cox-moddal, mind pedig a túlélési idő modellezésével, Weibull-AFT modellel vizsgálom.

1. Cox-Arányos hazardmodell

A Cox-modell a Kaplan-Meier, vagy az AFT modellekkel ellentétben nem a túlélési valószínűséget, hanem a hazardrát modellezi (annak a valószínűsége, hogy az adott beteg pont t időpont után hal meg). A koefficiensek alapvetően az egyes változók hatásának természetes alapú logaritmusai. Az egyes koefficienseket úgy értelmezhetjük, hogy amennyiben az adott változó egységnyi mértékben változik, az adott dal hazard rátája $e^{\text{koefficiens}}$ -re változik.

A Cox-modell azzal az alapfeltételezéssel él, hogy a hazard ráta időben arányos bármely két dal között. Az alapfeltételezés tesztelését követően azt kaptam eredményül, hogy számos változó esetén nem teljesül. A modell szignifikáns paramétereit így kizárólag azon változók esetén értelmezem, amelyek esetén érvényesül a feltétel.

6. táblázat: Cox-modell koefficiensek, szórásuk, p-értékeik

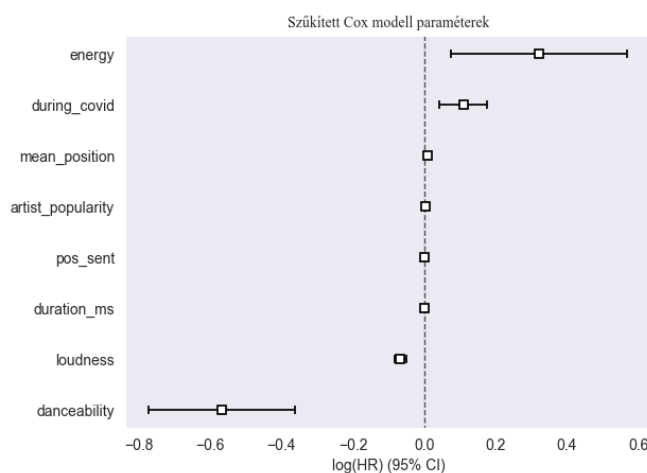
covariate	β_i	$\exp(\beta_i)$	se(coef)	p
<i>artist_popularity</i>	0,2485%	100,2488%	0,1326%	6,0977%
<i>danceability</i>	-			
	56,6622%	56,7439%	10,4502%	0,0000%
<i>energy</i>	32,1718%	137,9496%	12,6478%	1,0969%
<i>loudness</i>	-6,6344%	93,5809%	0,8265%	0,0000%
<i>duration_ms</i>	0,0000%	100,0000%	0,0000%	76,7930%
<i>during_covid</i>	10,9231%	111,5420%	3,3843%	0,1249%
<i>mean_position</i>	0,8148%	100,8181%	0,0364%	0,0000%
<i>pos_sent</i>	0,2047%	100,2049%	0,1373%	13,6079%

A szignifikáns, folytonos magyarázó változók közül több változó magyarázóhatása kifejezetten marginális. Az *artist_popularity* változó minden más változatlansága melletti egységnyi növekményével a hazardráta, azaz annak valószínűsége, hogy az adott dal egy t időpontot

követően kiesik a modellből, 0,21%-ponttal nő. Hasonló állítható *pos_sent* változóról, azaz a dal pozitív szentiment értékéről, a változó egységnyi növekménye, minden más változatlansága mellett a hazardráta 0,2%-os növekedésével jár. Az átlagos elért pozíció (*mean_position*) változó egységnyi növekménye esetén 0,82%-ponttal nő a kiesési valószínűség minden más változatlansága mellett.

A *danceability*, *loudness* változók mind szignifikánsan csökkentik a kiesési valószínűséget. A *danceability* változó egységnyi növekménye esetén, minden más változatlanságát feltételezve, a hazard ráta a 56,74%-ára csökken. A másik változó kapcsán is hasonló állítható, a *loudness* 93,58%-ára csökkenti a hazardrát a *danceability* változónál leírt feltételek mellett.

A hazard ráta ceteris paribus növekményéhez az alábbi változók járulnak hozzá nagymértékben: *energy*, *during_covid*. Az *energy* változó a Spotify API dokumentációja szerint olyan dalok esetén vesz fel magas értéket, amelyek heavy-metal műfajúak, míg egy Bach szimfónia esetén alacsony a változó értéke. Az *energy* változó egységnyi, ceteris paribus növekedésével 1,3795-szörösére nő annak valószínűsége, hogy egy dal kiesik a műsorlistából adott t időpontot követően. A *liveness* mellett a *during_covid* változó jár magas negatív magyarázó erővel. Amennyiben egy dal a Covid-19 okozta Európai lezárásokat követően került fel a toplistára, minden más változatlansága mellett 11,54%-ponttal nő annak valószínűsége, hogy a dal adott t időpontot követően kiesik a műsorlistából.



13. ábra: Cox-modell alapfeltételt kielégítő magyarázó változók hatása egy dal hazardfüggvényére

Az egyes magyarázó változók hatásának erősségét mutatja az ábra, amelyen az egyes magyarázó változók koefficiensei, valamint ezek 95%-on meghatározott konfidencia intervallum határai szerepelnek. Mind a *danceability*, mind az *energy* változó esetén szélesebb konfidencia intervallumokat tapasztalunk, mint a *during_covid*, vagy a *loudness* szignifikáns

magyarázó változók esetén, ami arra utal, a *during_covid* és a *loudness* változók esetén pontosabb a becsült parciális hatás.

2. Weibull-AFT modell

Céлом az, hogy az összes, szignifikánsnak tekinthető, modellemben szereplő magyarázó változó hatását vizsgáljam a dalok túlélésére vetítve. Mivel az AFT modellek során nem szükséges olyan feltételeknek megfelelni, mint amelyek a Cox-modell esetén alapfeltételezéseknek minősülnek, célszerű ilyen modellek alkalmazásával is vizsgálni az adatbázist. A Weibull-AFT modell nem a hazardrát, hanem a túlélést modellezi, ezért ezen modell esetén a koefficiensek a Cox-modellével ellentétes jelentéstartalommal bírnak. Amellett, hogy a modell alkalmazásával a Cox-modell alapfeltételezését megsértő változók hatásait is képes vagyok vizsgálni, a két modell szimultán alkalmazása fontos keresztvalidálási eszköz.

7. táblázat: Weibull-AFT modell koefficiensei, szórásuk, p-értékeik

param	Változó	β_i	$\exp(\beta_i)$	se(coef)	p
	<i>acousticness</i>	- 0,0144	0,9857	10,9520%	89,5231%
	<i>artist_popularity</i>	- 0,0062	0,9938	0,2029%	0,2160%
	<i>danceability</i>	1,3206	3,7458	17,5087%	0,0000%
	<i>duration_ms</i>	0,0000	1,0000	0,0000%	92,8818%
	<i>during_covid</i>	- 0,2564	0,7738	5,1818%	0,0001%
	<i>energy</i>	- 0,3992	0,6709	21,2046%	5,9774%
	<i>explicit</i>	- 0,5054	0,6032	5,1945%	0,0000%
	<i>instrumentalness</i>	- 1,7321	0,1769	30,4450%	0,0000%
	<i>ispop</i>	0,0086	1,0086	4,6188%	85,3111%
	<i>key</i>	0,0078	1,0078	0,5796%	17,8360%
	<i>liveness</i>	- 0,2873	0,7503	15,3841%	6,1817%
	<i>loudness</i>	0,0758	1,0788	1,2905%	0,0000%
	<i>mean_position</i>	- 0,0136	0,9865	0,0535%	0,0000%
	<i>mode</i>	- 0,0034	0,9966	4,3226%	93,7359%
	<i>neg_sent</i>	- 0,0027	0,9973	0,2590%	30,1482%
	<i>pos_sent</i>	- 0,0046	0,9954	0,2087%	2,7325%
	<i>speechiness</i>	- 0,6269	0,5342	19,3731%	0,1212%
	<i>tempo</i>	0,0000	1,0000	0,0701%	97,8902%
	<i>track_pop</i>	0,0078	1,0079	0,0664%	0,0000%
	<i>valence</i>	0,2204	1,2466	11,6480%	5,8462%
lambda_	Intercept	5,0232	151,8964	34,4312%	0,0000%
rho_	Intercept	- 0,4035	0,6680	1,0382%	0,0000%

A Cox-modellel ellentétben a Weibull-modell a névadó eloszlásából kifolyólag tartalmaz értelmezendő paramétereket. Az modell λ paramétere a skálázást jelöli, (a két vizsgált populáció túlélésgörbéi közötti „gyorsítást”), a ρ paraméter pedig a modell változási

gyorsaságát. Mivel értéke 0.668, az idő előrehaladtával egyre csökken annak valószínűsége, hogy az adott dal túlél. A következőkben megvizsgálom az egyes szignifikáns (5%-os szignifikanciaszinten vizsgálva) változók hatását az egyes túlélésre.

Vannak olyan magyarázó változók a Weibull-modellben, amelyek kizárólag marginális pozitív hatással bírnak, névlegesen egy dal kulcs (*key*) értéke egységnyi növekménye esetén 0,78%-ponttal növeli minden más változatlansága mellett egy dal túlélési valószínűségét. Az, hogy egy dal milyen népszerű (*track_pop*), kizárólag 0,79%-os túlélési valószínűséggel jár minden más változó változatlanságát feltételezve.

Bizonyos változók hasonló mértékű, viszont negatív módon befolyásolják a dalok túlélését. Meglepően az *artist_popularity*, azaz egy előadó népszerűségi értéke egységnyi növekménye esetén 99,38%-ára csökkenti egy adott dal túlélési valószínűségét minden más változatlansága mellett. Egy dal átlagos elért pozíciója a modell szerint szintén minimális, de negatív hatással bír a túlélésre. Az *artist_popularity* változó feltételeivel megegyező feltételek között a *mean_position* változó 98,65%-ára csökkenti a túlélési valószínűséget. A modell az egyes szentiment értékek, pozitív és negatív egyaránt, esetén negatív túlélési hatást jelez előre, a negatív *ceteris paribus* növekedése esetén 99,73%-ára, pozitív *ceteris paribus* növekedése esetén 99,54%-ára csökken egy adott dal túlélési valószínűsége. Az erős érzelmi töltetű zeneszövegeket tehát a fogyasztói közeg kis mértékben negatívan értékeli.

Egyes változók jelentősebb pozitív, vagy negatív hatással bírnak a túlélésre. A *danceability* változó egységnyi, *ceteris paribus* növekménye esetén egy dal túlélése 274,58%-ponttal nő, a *loudness* változó hasonló feltételek melletti növekményével 7,88%-ponttal nő a túlélési valószínűség, a *valence* változó esetén (szintén adott feltételek mellett) a túlélési valószínűség az 124,55%-ára növekszik. Látható, hogy a gyengén negatív zeneszöveg szentiment értékekkel szemben a *valence* változó esetén, amely a zenei elemek pozitivitását is magába foglalja, erős pozitív hatást tapasztalunk a túlélési valószínűségekre. A pozitív hatással bíró változókkal szemben a *during_covid*, *energy*, *explicit*, *instrumentalness*, *liveness*, *speechiness* változók mind jelentős negatív hatással bírnak az egyes dalok túlélésére.

A *during_covid* változót vizsgálva azt állíthatjuk, hogy az egyéb magyarázó változók változatlansága mellett a Covid-19 okozta lezárásokat követően listára felkerült dalok esetén a túlélési valószínűség 77,28%-a lesz a nem ebben az időszakban felkerült dalok túlélésének. Az *energy* változó olyan műfajok esetén veszi fel a Spotify API dokumentációja szerint

maximumértékét, mint a metál, vagy a magasabb ritmusértékű rock dalok, míg például egy Bach szimfónia esetén lesz minimális. A változó *ceteris paribus* egységnyi növekedése mellett a túlélési valószínűség 67,09%-ára csökken. Az explicit változóval ellátott dalok esetén a túlélési valószínűség, az egyéb változók változatlanlansága mellett, a nem explicit dalok túlélési valószínűségének 60,32%-a. A korábbiakhoz hasonló állítható az *instrumentalness* változó hatásáról. A változó abban az esetben maximális, ha az adott dal nem tartalmaz zeneszöveget, így főként elektronikus jellegű, valamint klasszikus és jazz stílusú dalok esetén magas az értéke. A változó egységnyi növekményével, minden más változatlanlansága mellett, 82,31%-ponttal csökkenti egy dal túlélési valószínűségét. A toplistákra való teljesítményt így ez a változó növekedése képes legradikálisabban rontani. Mind a *liveness*, azaz egy adott dal élőzeneiségének számszerűsített értéke, vagy a *speechiness*, avagy annak mértéke, hogy egy dalban szereplő beszélt szöveg (versfelolvasás, beszélgetés) mérésére szolgáló változó mind szignifikánsan csökkenti egy dal túlélési valószínűségét. A *liveness* változó *ceteris paribus* egységnyi növekménye 75,13%-ára, a *speechiness* változó *ceteris paribus* egységnyi növekménye 53,42%-ára csökkenti a túlélési valószínűséget.

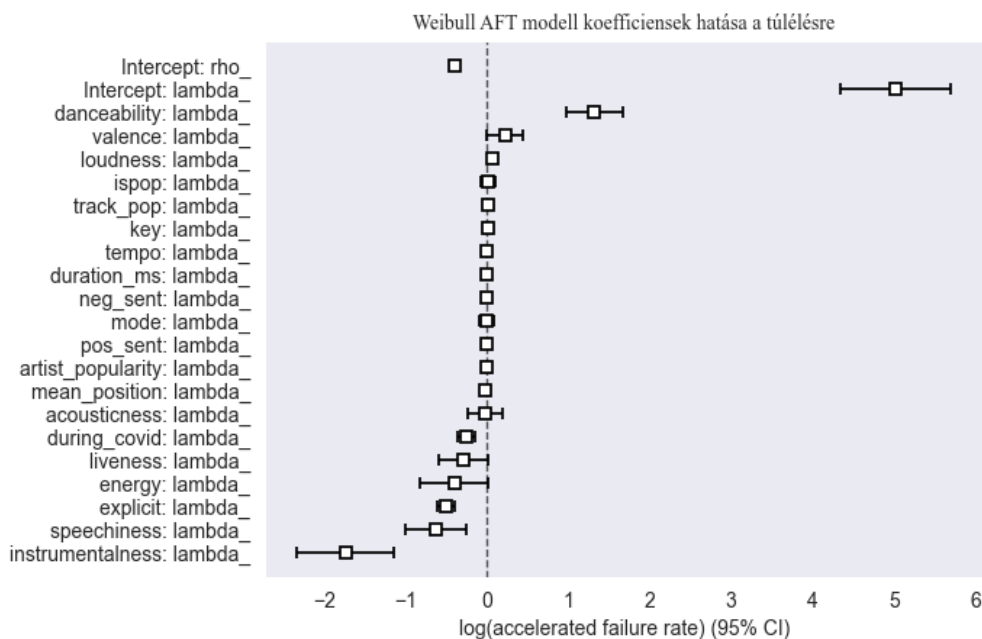
A Cox-modell és a Weibull-AFT modell paramétereit összehasonlítva teljesül a kereszt-validációs feltétel, miszerint a túlélési valószínűségeket befolyásoló magyarázó változók hatásának irányát mindkét modell azonosan becsülte meg.

A modellek által az alábbi állításokat tehetjük a dalok Global Top 200 listán való túlélését illetően:

A Covid-19 okozta lezárásokban szignifikánsabban alacsonyabb adott egyéb magyarázó változó értékeket feltételezve egy dal túlélési valószínűsége. Ebből az eredményből az fakad, hogy nem vagyok képesek elutasítani kiinduló hipotézisemet, így a Covid-időszakban a listára került dalok esetén csökken a túlélési valószínűség, így a túlélési idő. Ez azzal a jelentéssel bír, hogy a lezárások alatt a Spotify felhasználói nagyobb mértékben válogattak az általuk hallgatott dalok között, így a toplisták nagyobb rendszerességgel cserélődtek ki.

A szignifikánsan negatív túlélés-befolyásoló hatással bíró magyarázó változók közül több, mint a *speechiness*, *energy*, illetve az *instrumentalness* műfajszűkítő erővel bírnak. Mivel a *speechiness* változó olyan „dalok” esetén maximális, amelyek esetén magas a dalban található szöveget illetően a beszélt szöveg aránya, és negatívan befolyásolja egy dal túlélését, azt lehet állítani, hogy a rap és egyéb magas beszélt szöveg aránnyal rendelkező dalok túlélése adott egyéb magyarázó változók mellett kisebb, mint az egyéb műfaj dalai esetén, ahol a változó

értéke kisebb. Ahogy korábban ezt részleteztem, a dokumentáció szerint az *energy* változó a rock és a heavy-metal dalok esetén veszi fel maximumát. Így az ilyen műfajú dalok túlélési valószínűsége, minden más változatlanágát feltételezve, kisebb, mint az egyéb, kisebb *energy* értékkel bíró dalok esetén. Ezzel ellentétes eredményt közöl az *instrumentalness* változó, amely minimumértékét pont a magas szóhasználatú dalok esetén veszi fel, így például egy rapszám esetén az alacsony szóhasználatú dalokhoz mérve magasabb túlélési valószínűségeket tapasztalunk.



14. ábra: Egyes magyarázó változók hatása a túlélésfüggvényre a Weibull-AFT modellben

A szignifikáns magyarázó változók közül az *instrumentalness* és az *energy* bírnak a legszélesebb konfidencia intervallummal, így ezen két változó esetén mondható az, hogy a legpontatlanabb a koefficiens becslése a szignifikáns változók közül.

IX. Gépi-tanuló modellek túlélési valószínűségek becslésére

A következő fejezetben bemutatom, milyen pontossággal képesek a Random Survival Forest, DeepHit és DeepSurv modellek az általam vizsgált adatbázis esetén a túlélési valószínűség becslésére.

1. Random Survival Forest

Az RSF modell implementálása során 25%-os teszhalmazmérettel dolgoztam. A modellt lefuttattam a bemeneti paraméterek összes lehetséges, összesen 7200 kombinációján, azt az összehasonlítási mértékegység a Harrell-féle C-index volt.

8. táblázat: Random Survival Forest bemeneti paraméterei

Paraméter	Érték
n_estimators	150
min_samples_split	5
min_samples_leaf	25
max_features	sqrt
n_jobs	-1
random_state	124
verbose	1

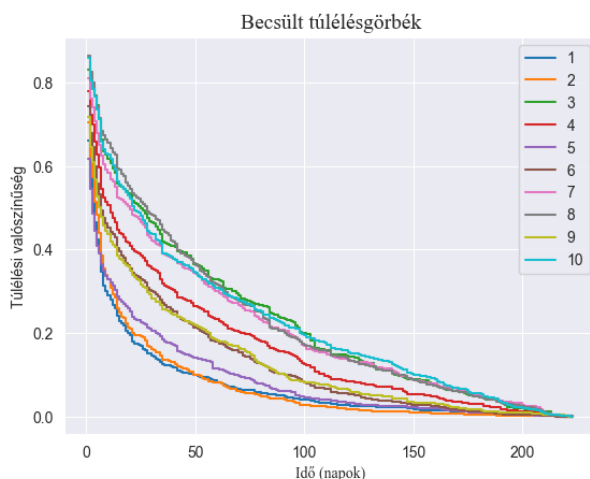
A táblázat a legjobban teljesítő modell bemeneti paramétereit tartalmazza. A modellben 150 döntési fa található, minden belső csomópont szétválasztásához legalább 5 elemű, a végső csomópontokban legalább 25 elemű minta szükséges. Az ideális szétválasztás esetén a kiválasztott magyarázó változók közül legfeljebb $\sqrt{\text{magyarázó változók száma}}$ számosságú változó kerülhet kiválasztásra. Az egyéb bemeneti értékek nem a modell struktúráját, hanem a futás ellenőrizhetőségét biztosítják.

9. táblázat: RSF modell átlagos változófontosságai

Változó	Átlagos változófontosság	Változófontosság szórása
<i>track_pop</i>	7,9282%	0,6184%
<i>mean_position</i>	6,6494%	0,4446%
<i>danceability</i>	1,0729%	0,1691%
<i>explicit</i>	0,8324%	0,1921%
<i>loudness</i>	0,4538%	0,2438%
<i>instrumentalness</i>	0,3183%	0,1674%
<i>duration_ms</i>	0,1826%	0,1040%
<i>artist_popularity</i>	0,1603%	0,0485%
<i>valence</i>	0,1405%	0,0924%
<i>pos_sent</i>	0,1178%	0,0434%
<i>acousticness</i>	0,1135%	0,0463%
<i>speechiness</i>	0,1065%	0,1073%
<i>tempo</i>	0,0988%	0,0748%
<i>liveness</i>	0,0836%	0,0771%
<i>ispop</i>	0,0508%	0,0331%
<i>key</i>	0,0392%	0,0295%
<i>neg_sent</i>	0,0290%	0,1667%
<i>during_covid</i>	0,0200%	0,0211%
<i>mode</i>	-0,0107%	0,0175%
<i>energy</i>	-0,0204%	0,0467%

A táblázat az egyes magyarázó változók változófontosságát tartalmazza. Az átlagos változófontosság azt mutatja meg, hogy az egyes magyarázó változók egymáshoz mérve milyen

arányban járultak hozzá a modell prediktív képességéhez, a magas értékkel bíró magyarázó változók, mint a *danceability*, *track_pop*, vagy *mean_position* nagy befolyásoló erővel bírtak az egyes túlélésgörbék becslésében. A modell a hagyományos túlélésmodellekkel ellentétben nem bizonyos zenei jellemzők esetén, hanem meta-adatok esetén adja a legjelentősebb magyarázó erővel bíró változókat.



15. ábra: A teszhalmaz 10 elemű mintájára készített becslott túlélésfüggvény az RSF modell alapján

A teljesítményt a Harrell-féle C-index segítségével mérem. A C-index értéke 69,82%, azaz a modell 69,82%-os valószínűséggel helyes sorrendben túlélési időt becsülni két adatpont esetén.

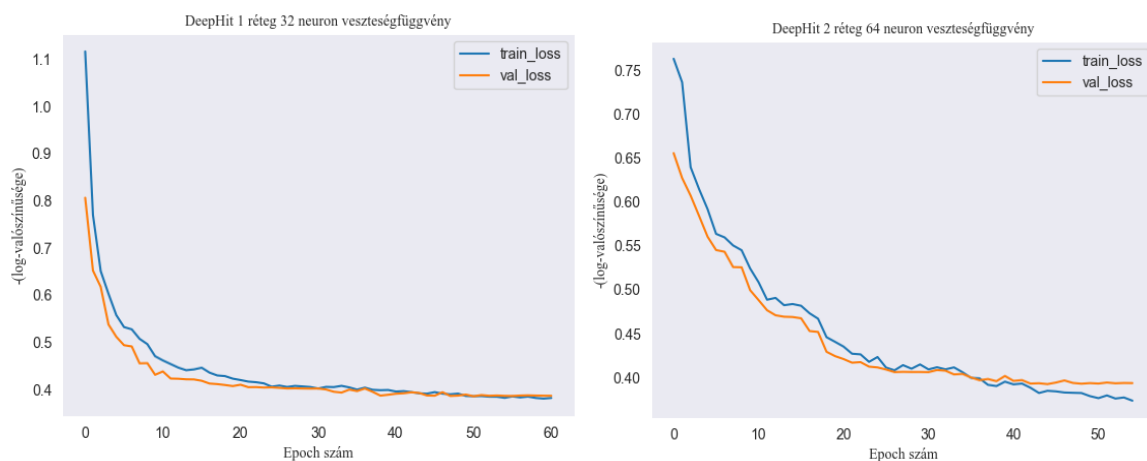
2. DeepHit

A modell futtatása során esetemben nincsenek versengő célesemények, így a modell második rész-hálózata kizárólag egy hálóból áll. A következő táblázat az egyes általam vizsgált modelleket mutatja be (a modellek kizárólag a rejtett rétegek méretében különböztek), a megfelelő abszolút és relatív becslési képességgel együtt. Az összes modell futtatása esetén 20%-os dropout értéket határoztam meg.

10. táblázat: DeepHit modell implementálásai és modell-értékelési paraméterei

Rétegek száma	Neuronok száma	Tanulási ráta	Brier-érték	C-index
1	32	5,09%	9,44%	50,80%
1	64	1,39%	9,45%	49,63%
2	32	6,14%	9,30%	48,24%
2	64	5,09%	9,09%	46,01%
2	128	0,55%	9,45%	44,02%
3	32	2,42%	9,41%	46,46%

A vizsgált modellek közül egyedül az egy rétegű, 32 neuronból álló képes az 50,8%-os C-indexével meghaladni az 50%-ot, azaz a vizsgált DeepHit modellek közül egyedül ez a modell képes a véletlennél nagyobb hatékonysággal jól megbecsülni azt, amennyiben a valóságban két adatpont közül az első adatpont esetén hosszabb a túlélési periódus, akkor a modell szerint is az első adatpont esetén lesz minden t időpontban nagyobb a túlélési valószínűség. A Brier-értéket tekintve ennek a modellnek $1-9,44\%=90,56\%$ -os becslési pontossága van, amelynél jobb teljesítményt nyújt a 90,91%-os becslési pontosságú, 2 64 neuronból álló réteget tartalmazó DeepHit modell.



16. ábra: DeepHit modell (1 réteg 32 neuron; 2 réteg 64-64 neuron) veszteségfüggvénye

Az ábrák a két vizsgált DeepHit modell veszteségfüggvényét, azaz a negatív log-valószínűséget mutatják be az Epoch szám függvényében, azaz annak függvényében, hogy az egész adatsor hányszor halad át a modellen a tanulási folyamat során. Látható, hogy mind az egyrétegű, mind a kétrétegű DeepHit modell esetén közel azonos a tanítási és a teszhalmazokon elért veszteség, ami azt jelenti, hogy a modell tanulási folyamatában nem lépett fel a túlillesztés problémája.

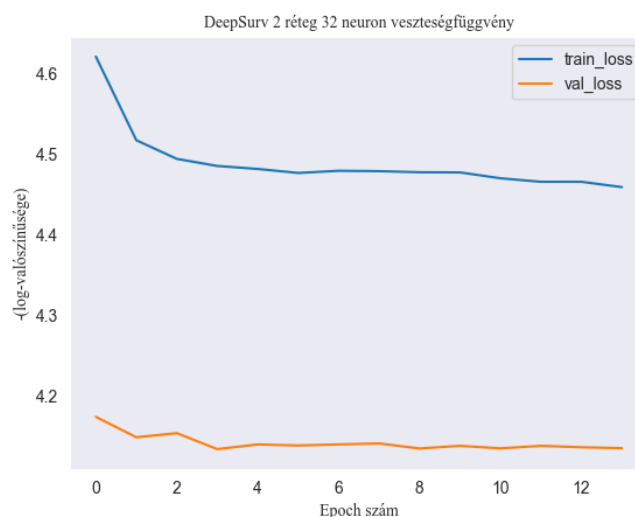
3. DeepSurv

A DeepSurv modellek a korábban ismertetett DeepHit modelleknél minden esetben jobb teljesítményt nyújtanak a túlélési valószínűségek becslése terén. A szelekciós során kettő, illetve három rejtett réteget tartalmazó modelleket teszteltem, az alábbi táblázat mutatja ezek teljesítményét.

11. táblázat: DeepSurv modell és modell-értékelési paraméterei

Rétegek száma	Neuronok száma	Tanulási ráta	Brier-érték	C-index
2	32	8,90%	7,61%	68,91%
2	64	5,09%	7,70%	68,33%
2	128	0,01%	8,45%	67,82%
3	32	5,09%	7,76%	68,59%

A két 32 neuronból álló rejtett rétegű DeepSurv modell képes a vizsgáltak közül a legpontosabb túlélésbecslésre. A Brier-érték a modell esetén 7,61%, azaz a modell 92,39%-os pontossággal képes az egyes adatpontok túlélésfüggvényének becslésére. A modell 68,91%-os valószínűséggel képes arra, hogy két adatpont túlélésének sorrendjét helyesen becsülje meg. Látható, hogy a legrosszabbul teljesítő DeepSurv modell is túlteljesíti a legjobban teljesítő DeepHit modellt.



17. ábra: DeepSurv modell (2 rejtett réteg: 32-32 neuron) veszteségfüggvénye

A veszteségfüggvény szemügyre vételével azonban azt látjuk, hogy jóval alacsonyabb bármely epoch szám esetén a negatív log-valószínűség a teszhalmaz esetén, mint a tanítási halmaz esetén. Ez egy alulillesztett modellre hívja fel a figyelmet, azaz a modell nem volt képes elég részletesen tanulni a tanulási halmaz alapján.

Azzal a feltételezéssel élek, a probléma oka az, hogy számos magyarázó változó esetén nem teljesül a Cox-modell alapfeltételezése. Lefuttattam a modellt kizárólag azokra a magyarázó változókra, amelyek esetén a feltételezés teljesül, ezek esetén mind a Brier-érték, mind a C-index romlott. A Brier-érték a feltételt kielégítő magyarázó változókkal 9,45%, a C-index pedig 62,61%.

X. Összegzés

A dolgozat szakirodalmi áttekintésében ismertettem, hogy korábban milyen túléléselemzési eszközöket alkalmaztak a dalok toplistákon való túlélésének elemzésére, látható volt, hogy korábban főként albumok teljesítményét, egyéb metaadatokat használtak fel magyarázó változóként a túlélésmodellekben. (Strobl és Tucker, 2000) Dolgozatomban megvizsgáltam az egyes zenei jellemzők hatását a dalok túlélésére. Eredményül azt kaptam, hogy a fogyasztói közeg az olyan dalokat, amelyek magas ritmusértékkel bírnak, magas decibel értékkel bírnak, valamint erős zenei pozitívítási értékűek, pozitívan értékeli. Ezzel szemben a zenei jellemzők közül olyan változók járultak hozzá szignifikánsan negatívan egy dal túléléséhez, amelyek vagy a zene szövegének előadási módjára, vagy a zenében található szöveg arányára vonatkoztak. A fogyasztói közeg negatívan értékeli azt, ha egy dal szövege versfelmondás jellegű, nem pedig éneklés, szintén negatív hatással van egy dal túlélésére az, ha a dal explicit jelzővel ellátott.

Mind a vizsgált Cox-modellben, mind a Weibull-AFT modellben szignifikánsan negatív hatása volt a *during_covid* változónak egy dal túlélésére. Egyik kezdeti hipotézisem szerint a Covid időszakban a listára került dalok túlélése szignifikánsan eltér a nem ebben az időszakban listára került dalok túlélésétől. Nem tudtam elutasítani ezt a kiinduló hipotézisemet, így a Covid-19 megjelenését követően a hallgatóközönség fogyasztási szokásai változtak, rövidebb időt töltöttek egy dal hallgatásával, vagy esetleg változatosabbá váltak fogyasztásukban. Ezt erősíti a tény, hogy az egyes évekre futtatott Kaplan-Meier modellek által becsült túlélési-értékeket a log rank próbával összehasonlítva azt kapjuk, szignifikánsan különbözik a 2020-as túlélés a 2019-es túléléstől.

Másik kiinduló hipotézisem szerint a Covid-19 megjelenését követően magasabb lett a negatív szentiment érték az egyes dalokban. Mivel sem a pozitív, sem a negatív szentimentértékeket vizsgálva nem tudtam elutasítani azon nullhipotéziseimet, melyek szerint az adott szentiment értéke megegyezik 2020-2019 (azonos eredmény 2020-2017-re) között, kiinduló hipotézisemet el kell utasítanom, azaz a birtokomban lévő információ mellett nem állíthatom, hogy nagyobb lenne a Covid-19 megjelenését követően listára került dalok negatív szentiment értéke.

Szintén megvizsgáltam a dalok szövegének komplexitását, azzal a kiinduló hipotézissel, hogy az átlagos zenei komplexitás idővel csökken. A Type-Token ráta segítségével számolt komplexitásértékeket összehasonlítva azt az eredményt kaptam, hogy szignifikánsan nagyobb a 2020-as átlagos TTR érték a 2017-es átlagos TTR értéknél, így Parada-Cabaleiro és tsai. 2024 kutatásával ellentétes eredményt kaptam.

Harmadik kiinduló hipotézisem szerint a mélytanuló modellek pontosabb túlélési valószínűség becslésre képesek, mint a hagyományos túlélés modellek, vagyis a Cox-modell és a Weibull-AFT modell.

12. táblázat: Prediktív modellek teljesítmény a Harrell-féle C-index szerint

Modell	Harrell-féle C-index
DeepHit	50,80%
Weibull AFT modell	66,59%
Cox-Arányos Hazárdmodell	66,64%
DeepSurv	68,91%
Random Survival Forest	69,82%

A Harrell-féle C-index alapján mérve a legrosszabb teljesítményt a DeepHit modell nyújtotta. A modell alapvetően azért lett megalkotva, hogy több egymással versengő célesemény esetén nyújtson pontos túlélési valószínűség becslést, futása során a különböző célesemények saját rész-hálózatukon egymás információit is felhasználják tanulási folyamatuk során, így valószínűsíthetően, mivel modellemben nem voltak versengő célfüggvények, a modell nem tudta kellő hatékonysággal megtanulni a túlélési valószínűségeket. A hagyományos Weibull-AFT és Cox-Arányos Hazárdmodellek már elfogadható erősségű C-indexet nyújtottak, a mélytanuló algoritmusok közül a DeepSurv, de leginkább a Random Survival Forest képesek voltak felülmúlni a C-index alapján mérve a hagyományos modellek prediktív képességét.

Kiinduló hipotézisem szerint a gépi-tanuló modellek felülmúlják a hagyományos túlélés modelleket prediktív képességük szerint. Mivel a DeepHit modell nem nyújtott jobb C-indexet a hagyományos modelleknél, sőt éppen jobb eredményt tanúsított, mintha véletlenszerűen határozná meg két adatpont közül, hogy melyik esetén magasabb a túlélés valószínűsége, nem állíthatjuk azt, hogy a gépi-tanuló modellek egyértelműen jobb eredményt mutatnak a predikció terén.

XI. Limitációk

Dolgozatom írása során két hátráltató tényezőt tapasztaltam. Mind a szimpla Cox-modell, mint a Cox-modell ismeretlen függvényét mélytanuló módszerekkel becsülő DeepSurv modell esetén problémát okozott az, hogy számos változó esetén nem teljesül a Cox-modell alapfeltételezése. Mind a hagyományos Cox, mind a DeepSurv modellek esetén így csökkent a C-index értéke, azaz a modellek nem voltak képesek a helyes sorba rendezést ugyanolyan

arányban végrehajtani, mint az alapfeltevést megszegő magyarázó változókat is tartalmazó ilyen modellek.

A másik javítási lehetőséget a vizsgált időintervallum jelentette. Mivel az adatbázis 2020.11.05-ig tartalmazza a Global Top 200-ra felkerült dalokat, majdnem kizárólag a 2020-ban a listára került dalok voltak cenzúrázott adatok. Célszerű lehetne egy olyan adatbázis megalkotása, amelyben a vizsgált időintervallum addig tart, amíg a legutolsó 2020-ban listára felkerült dal esetén be nem következik a célesemény, így pontosabb értéket lehetne adni a szentimentek egyezőségére.

XII. Hivatkozásjegyzék

- Bland, J. M., & Altman, D. (2004). The logrank test. *BMJ. British Medical Journal*, 328(7447), 1073. <https://doi.org/10.1136/bmj.328.7447.1073>
- Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), 187–220. <http://www.jstor.org/stable/2985181>
- Furner, J. (2016). Type–Token Theory and Bibliometrics. In *De Gruyter eBooks* (pp. 119–147). <https://doi.org/10.1515/9783110308464-009>
- Harrell, F., Califf, R., Pryor, D., Lee, K. and Rosati, R. (1982). Evaluating the yield of medical tests. *J. Amer. Med. Assoc.* 247 2543–2546.
- Hutto, C. J., & Gilbert, É. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the . . . International AAAI Conference on Weblogs and Social Media/Proceedings of the International AAAI Conference on Weblogs and Social Media*, 8(1), 216–225. <https://doi.org/10.1609/icwsm.v8i1.14550>
- Im, H., Song, H., & Jung, J. (2018). A survival analysis of songs on digital music platform. *Telematics and Informatics*, 35(6), 1675–1686. <https://doi.org/10.1016/j.tele.2018.04.013>
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics/the Annals of Applied Statistics*, 2(3). <https://doi.org/10.1214/08-aos169>

- Kaimann, D., Tanneberg, I., & Cox, J. (2020). “I will survive”: Online streaming and the chart survival of music tracks. *MDE. Managerial And Decision Economics/Managerial And Decision Economics*, 42(1), 3–20. <https://doi.org/10.1002/mde.3226>
- Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Quarterly Publications of the American Statistical Association/Quarterly Publication of the American Statistical Association*, 53(282), 457–481. <https://doi.org/10.1080/01621459.1958.10501452>
- Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1). <https://doi.org/10.1186/s12874-018-0482-1>
- Lee, C., Zame, W. R., Yoon, J., & Van Der Schaar, M. (2018). DeepHIT: A deep learning approach to survival analysis with competing risks. *Proceedings of the . . . AAAI Conference on Artificial Intelligence*, 32(1). <https://doi.org/10.1609/aaai.v32i1.11842>
- Liu, E., Liu, R., & Lim, K. S. (2023). Using the WEIBULL Accelerated Failure Time Regression model to predict time to health events. *Applied Sciences*, 13(24), 13041. <https://doi.org/10.3390/app132413041>
- Parada-Cabaleiro, E., Mayerl, M., Brandl, S., Skowron, M., Schedl, M., Lex, E., & Zangerle, E. (2024). Song lyrics have become simpler and more repetitive over the last five decades. *Scientific Reports*, 14(1). <https://doi.org/10.1038/s41598-024-55742-x>
- *Prediction of product success: explaining song popularity by audio features from Spotify data - University of Twente Student Theses.* (n.d.). <https://purl.utwente.nl/essays/75422>
- Sharma, N., Pareek, D.P., Pathak, M.P., & Sakariya, M.N. (2022). Journal of Development Economics and Management Research Studies (JDMS), A Peer Reviewed Open Access International Journal, ISSN 2582 5119 (Online), 09 (11), 10 -19, January-March, 2022.

- STROBL, E. A., & TUCKER, C. (2000). The Dynamics of Chart Success in the U.K. Pre-Recorded Popular Music Industry. *Journal of Cultural Economics*, 24(2), 113–134. <http://www.jstor.org/stable/41810719>
- Wiegrebe, S., Kopper, P., Sonabend, R., & Bender, A. (2023). Deep Learning for Survival Analysis: A review. *arXiv* (Cornell University). <https://doi.org/10.48550/arxiv.2305.14961>