SZAKDOLGOZAT

Sádt Dominik

2025

Budapesti Corvinus Egyetem

Eötvös Lóránd Tudományegyetem

Természettudományi kar

Ok-specifikus halandósági ráták előrejelzése gépi tanulási módszerekkel

SZAKDOLGOZAT

Sádt Dominik

Biztosítási és pénzügyi matematika MsC

Aktuárius specializáció

Témavezető:

Dr. Kovács László

Budapest, 2025

Tartalomjegyzék

Ábrajegyzék

1. Bevezetés

A hosszú élet kockázata, vagyis a longevity kockázat egyre nagyobb kihívást jelent a biztosítási és nyugdíjrendszerek számára világszerte. A halálozási valószínűségek évszázadok óta csökkenő trendet mutatnak, és ez a csökkenés a modern időkben is folytatódik. A javuló egészségügyi ellátás, az életmódbeli változások és az orvostudomány fejlődése mind hozzájárulnak ahhoz, hogy az emberek hosszabb ideig éljenek. Ennek következtében a statikus halandósági táblák alkalmazása a járadék- és nyugdíjtermékek esetében jelentős alulbecsléshez vezethet a jövőbeli kifizetések tekintetében.

A hosszabb élettartam nem csupán az egyéni pénzügyi tervezés szempontjából jelentős kérdés, hanem makrogazdasági és társadalmi kihívásokat is felvet. Az állami nyugdíjrendszerek fenntarthatósága veszélybe kerülhet, ha a várható élettartam növekedése nincs megfelelően figyelembe véve a nyugdíjrendszer finanszírozásában. Az Európai Unióban működő biztosítók számára a Szolvencia II keretrendszer előírja a longevity kockázat megfelelő modellezését. Ez azt jelenti, hogy a biztosítóknak halandósági ráták előrejelzésére alkalmas modelleket kell alkalmazniuk a megfelelő tőkekövetelmények meghatározása érdekében. A halandósági trendek modellezése tehát nem csupán elméleti kérdés, hanem gyakorlati jelentőséggel is bír.

A halandósági táblák előrejelzésére több matematikai és statisztikai modell létezik. A szakirodalomban a Lee-Carter modell a legelterjedtebb, amely a halálozási valószínűségek időbeli trendjeinek modellezésére egy faktoros idősorelemzést alkalmaz. Emellett több más módszer is létezik, például az APC (age-period-cohort) modellek és a CBD (Cairns-Blake-Dowd) modell, amelyek különböző megközelítésekkel próbálják megragadni a halandósági trendeket.

A klasszikus modellek azonban feltételezik, hogy a halandósági mintázatok viszonylag stabilak és lineárisan jól strukturálhatók egy vagy néhány fő komponens mentén. Napjainkban azonban a nagy mennyiségű, heterogén adat és a társadalmi-gazdasági tényezők gyors változása olyan modellezési rugalmasságot igényel, amelyre a hagyományos statisztikai megközelítések nem minden esetben alkalmasak. Itt lépnek be a gépi tanulási (machine learning) módszerek, amelyek képesek megragadni a komplex, nemlineáris mintázatokat is az adatokban.

A dolgozat célja annak vizsgálata, hogy a gépi tanulási algoritmusok milyen mértékben képesek javítani a klasszikus Lee-Carter modell által adott halálozási ráta-előrejelzéseket, különös

tekintettel az egyes halálokok szerinti bontásra. A dolgozatban először bevezetem az alkalmazott modellezési eszközök módszertani alapjait, majd bemutatom a kutatáshoz felhasznált halandósági adatokat. Az elemzés lényegi részében a Lee-Carter modell és különböző gépi tanulási eljárások, például döntési fák, gradient boosting és random forest teljesítményét hasonlítom össze.

Két különböző megközelítést is alkalmazok, melyeket Susanna Levantesi és Dorethe Skovgaard Bjerre munkái inspiráltak: ők eltérő módon integrálták a gépi tanulási módszereket a halandósági előrejelzésekbe. A célom ezen megközelítések teljesítményének részletes összehasonlítása országonként és a leggyakoribb halálokonként, valamint annak vizsgálata, hogy a gépi tanulás képes-e hozzáadott értéket nyújtani a klasszikus halandósági modellezéshez.

2. A halandóság modellezés

A halandóság számszerűsítésének egyik legalapvetőbb mutatója a halandósági ráta, amely egy adott időszakra és populációra vonatkozóan azt fejezi ki, hogy a populáció mekkora hányada hunyt el az adott időszakban. Matematikailag a következőképpen határozható meg:

$$m = \frac{D}{E} \tag{1}$$

ahol m a halandósági ráta, D a vizsgált időszakban elhunytak száma, E pedig a populáció nagyságát jelenti, amelyet valamilyen módon definiálnunk kell. A populáció létszáma többféleképpen is értelmezhető. Az egyik lehetőség a vizsgált időszak kezdetén életben lévő egyének száma, amelyet kezdeti kitettségnek nevezünk és E^0-val jelölünk. Egy másik megközelítés az időszak alatt élő egyének átlagos száma, amelyet központi kitettségnek nevezünk és E^c-vel jelölünk. Ez utóbbit úgy számíthatjuk ki, hogy a vizsgált időszak kezdetén életben lévő személyek által megélt egyéni időmennyiségeket összegezzük. A kitettségtől függően eltérő halandósági ráták léteznek. Ha a kezdeti kitettséget használjuk, akkor kezdeti halandósági rátáról beszélünk, amelyet m^0-val jelölünk, míg, ha a központi kitettséget vesszük figyelembe, akkor központi halandósági rátát kapunk, amelynek jele m^c.

A halandósági modellezésben kulcsfontosságú fogalom a túlélési függvény, amely az élettartam eloszlását írja le. Az L élettartamhoz tartozó túlélési függvényként azt a $G: \mathbb{R}^+ \rightarrow [0,1]$ függvényt értjük, amely az alábbi egyenlettel adható meg:

$$G(y) = P(L \geq y), \qquad y \geq 0. \tag{2}$$

Ez a függvény néhány alapvető tulajdonsággal rendelkezik:

$$G(0) = 1, \tag{3}$$

$$G(y) = 1 - F(y), \qquad y \geq 0, \tag{4}$$

ahol $F: \mathbb{R}^+ \rightarrow [0,1]$ az L élettartam eloszlásfüggvénye.

Amennyiben egy egyén már elért egy adott $x \geq 0$ életkort, akkor a hátralévő élettartam $(L - x)$ eloszlása feltételes eloszlásként értelmezhető az $L \geq x$ feltétel mellett. Ennek megfelelően definiálható a reziduális túlélési függvény (G_x) az alábbi módon:

$$G_x(y) = P(L - x \geq y | L \geq x) = \frac{P(L \geq x+y)}{P(L \geq x)} = \frac{G(x+y)}{G(x)} \tag{5}$$
$$(x, y \geq 0).$$

Egy jelentős mutató az x éves korban várható hátralévő élettartam, amely a következő integrállal számítható ki:

$$e_x = E(L - x | L \geq x) = \int_x^\infty G_x(y)dy \tag{6}$$

Kiemelendő az $x = 0$ eset, amely a születéskori várható élettartamot adja meg:

$$e_x = E(L) = \int_0^\infty G_x(y)dy \tag{7}$$

Az élettartam eloszlásának modellezése során két fő megközelítés lehetséges: egyik esetben a folytonos eloszlásokkal dolgozunk, míg a másik esetben az L értékeit diszkrét valószínűségi változóként kezeljük (például egész számú élettartam esetén). (Vékás, 2016)

2.1 Folytonos modell

Folytonos eloszlások alkalmazása esetén a halandóság mérésére célszerű olyan mutatószámot használni, amely az egyes életkorokhoz tartozó pillanatnyi értéket adja meg, és nem függ a

vizsgált időintervallum hosszától. Természetesen egy adott pillanatban a halálozási valószínűség mindig nulla, mivel folytonos eloszlás esetén bármely $y \geq 0$ értékre

$$P(L = y) = 0. \tag{8}$$

Ugyanakkor határértékes megközelítéssel meghatározható egy úgynevezett pillanatnyi évesített halálozási valószínűség, ha egy rövid időtartamra számított évesített halálozási valószínűséget úgy vizsgálunk, hogy az időtartam hossza a nullához tart. Az így nyert mérőszámot halálozási intenzitásnak nevezzük, amelyet más néven hazárdrátaként vagy kockázati rátaként is emlegetnek:

$$\mu(y) = \lim_{\beta \to 0^+} \frac{P(L < y + \beta | L \geq y)}{\beta} \tag{9}$$

Bár a halálozási intenzitás önmagában nem tekinthető valószínűségnek, mégis egyfajta évesített pillanatnyi halálozási valószínűségként értelmezhető. A fenti meghatározás intuitív és szemléletes, azonban a gyakorlatban a halálozási intenzitásfüggvényt gyakran egyszerűbb az alábbi összefüggés alapján kiszámítani:

$$\mu(y) = \lim_{\beta \to 0^+} \frac{F(y + \beta) - F(y)}{\beta G(y)} = \frac{f(y)}{G(y)}. \tag{10}$$

2.2 Diszkrét modell

Az aktuáriusi gyakorlatban kiemelt jelentőségűek a teljes években mért élettartamokra épülő halandósági táblák. A halandósági tábláknak olyan táblázatokat hívunk, amikben egy adott évre vonatkozó elméleti halálozási (q_x), túlélési valószínűségei (p_x) és a várható élettartam szerepelnek életkoronként. A valószínűségeket azt mutatják meg, hogy mennyi a valószínűsége, hogy egy x éves ember a következő születésnapja előtt meghal. Tehát:

$q_x =$ a valószínűség, hogy valaki, aki megélte x-edik életévét meghal az $(x + 1)$.-edik életéve betöltése előtt.

Mivel ezek az adatok egy adott időpontban végzett felmérésből származnak, minden q_x eltérő években született, de egy időben élő generációk halálozási valószínűségeit reprezentálja, miközben egy pillanatképet ad a népesség egészének halandósági viszonyairól. A túlélési valószínűség a q_x-ből számolható, ugyanis:

$p_x = 1 - q_x =$ a valószínűég, hogy valaki, aki megélte x-edik életévét megéli az $(x + 1)$.-edik életévét is.

Bizonyítható, hogy a p_x-ek szorzata megadja annak a valósíznűségét, hogy aki megélte az x. életévét, életben lesz t év múlva is:

$t|p_x = p_x \cdot p_{x+1} \dots p_{x+t+1} =$ a valószínűség, hogy valaki, aki megélte x-edik életévét megéli az $(x+t)$.-edik életévét is. (Banyár, 2003)

A halandósági tábla tehát nem egy generációnak az adatait tartalmazza, hanem egy pillanatnyi állapotot.

Továbbá a koréves túlélési valószínűségek és a túlélési függvény diszkrét értékei között az alábbi kapcsolat áll fenn:

$$p_x = \frac{G(x+1)}{G(x)} \quad (x \epsilon \mathbb{N},\ G(x) > 0), \tag{11}$$

$$G(x) = \prod_{y=0}^{x-1} p_y \qquad (x \epsilon \mathbb{N})\,. \tag{12}$$

A teljes években történő modellezésnek két fő oka van: egyrészt az elérhető demográfiai adatok jellemzően egész éves bontásban állnak rendelkezésre, másrészt a számítások során az integrálok helyett egyszerűbb véges összegzésekkel lehet dolgozni. Az egészértékű modell matematikai alapjainak tisztázásához célszerű a T élettartamot szétbontani egy egész és egy tört rész összegére:

$$T = T_e + T_t \tag{13}$$

ahol T_e az élettartam egész része, míg T_t annak tört része. A koréves halálozási valószínűségek csak az élettartam egész részének eloszlását határozzák meg, ezért a tört rész viselkedésére külön feltevést szokás alkalmazni. A halandósági táblák általában tartalmaznak egy feltételezett maximális életkort, $\omega \epsilon \mathbb{N}$ amelyre teljesül, hogy

$$P(T > \omega) = 0. \tag{14}$$

Magyarországon a KSH ezt az értéket $\omega = 100$ évnek veszi. A maximális életkorra vonatkozó feltevésből következik, hogy az élettartam egész részének eloszlását teljes mértékben meghatározzák a q_x $(x = 0,1,2, \dots, \omega - 1)$ halálozási valószínűségek, kiegészítve a $q_\omega = 1$ feltétellel, amely biztosítja, hogy az adott modellben az ω évnél magasabb életkor megélése nem következhet be. A halandósági táblákban a p_x és $q_x$értékeken túl szerepel még a

$$l_x = l_0 G(x) \quad (x = 0, 1\,, \dots, \omega), \tag{15}$$

$$d_x = l_{x+1} - l_x \quad (x = 0, 1, \dots, \omega - 1) \tag{16}$$

továbbélési és kihalási rend is, ahol általában $l_0 = 100\,000$ a halandósági tábla kiinduló száma. A továbbélési rend a túlélési függvény egy konstansszorosa, és egy feltételes, 100 000 fővel induló populáció esetén megadja, hogy hányan érik el legalább az x éves kort, feltételezve, hogy a halálozási valószínűségek időben állandók maradnak. A kihalási rend hasonló módon az adott korban elhunytak várható számát mutatja. A halandósági táblák rendszerint tartalmazzák a következő képlet segítségével meghatározható várható hátralévő élettartamokat is:

$$e_x = \frac{1}{l_x} \sum_{i=x+1}^{\omega} l_i + \frac{1}{2} \qquad (x = 0, 1, \dots, \omega - 1)\,. \tag{17}$$

Az életbiztosítási ipar működése szorosan kapcsolódik a halandósági táblákhoz. Az életbiztosítási szerződések árazása, az egyének kockázati profiljának meghatározása, valamint a tartalékok képzése mind a halandósági táblák alapján történik. A biztosítótársaságok számára kritikus fontosságú, hogy pontos információkkal rendelkezzenek arról, hogy a biztosítottak körében milyen a várható halálozás.

A halandósági táblák alkalmazása nem korlátozódik az életbiztosítás területére; azok fontos szerepet játszanak a népesedési elemzésekben is. A demográfusok és a kormányzati szervek ezeket az adatokat használják a népességi folyamatok megértéséhez és a jövőbeni trendek előrejelzéséhez. Ezek az előrejelzések kritikus jelentőségűek az olyan hosszútávú stratégiák kidolgozásában, mint a nyugdíjrendszer fenntarthatósága, az egészségügyi szolgáltatások tervezése, valamint a munkaerőpiaci folyamatok elemzése. A halandósági táblák alapján a kormányzatok és az egyéb érdekeltek adatalapú döntéseket hozhatnak a társadalmi és gazdasági kihívások kezelésére. (Banyár, 2003)

3. A Lee-Carter modell elméleti alapja

3.1 A klasszikus modell

A módszert (Lee & Carter, 1992) mutatta be a "Modeling and Forecasting U.S. Mortality" c. cikkben. A modell a mai napig az egyik legjobban használható statisztikai eszköz halandósági valószínűségek és várható élettartam előrejelzésére. A továbbiakban bemutatom a szakdolgozat központi elemeként szolgáló modell legfontosabb részleteit.

Legyen m_{at} egy ország halandósági rátájának logaritmusa a korcsoportban $(a = 1, 2, \dots, A)$ és t időben $(t = 1, 2, \dots, T)$. Ezeket a rátákat a

$$m_{at} = \alpha_a + \beta_a \gamma_t + \epsilon_{at} \tag{18}$$

egyenlettel tudjuk modellezni, ahol α_a, β_a, és γ_t becsülendő paraméterek, ϵ_{at} véletlen hibatag, amiről általában azt feltételezzük, hogy normális eloszlású, 0 várható értékű, σ^2 szórásnégyzetű. A paraméterezés nem egyértelmű, mivel invariáns a következő transzformációkra:

$$\beta_a \rightarrow c\beta_a \tag{19}$$

$$\alpha_a \rightarrow \alpha_a - \beta_a c \tag{20}$$

$$\gamma_t \rightarrow \frac{1}{c}\gamma_t \tag{21}$$

$$\gamma_t \rightarrow \gamma_t + c \tag{22}$$

minden $c\epsilon\mathbb{R},\ c \neq 0$.

Ezek szerint a felírt egyenlethez tartozó likelihoodnak végtelen ekvivalens maximuma van, amik azonos előrejelzéseket adnak. A következő megkötéseket használjuk a gyakorlatban:

$$\sum_t \gamma_t = 0 \tag{23}$$

$$\sum_a \beta_a = 1. \tag{24}$$

A $\sum_t \gamma_t = 0$ megkötésből következik, hogy α_a az empirikus átlag a korcsoportban, $\alpha_a = m - a$. Eszerint felírható a modell a centrált log-halálozási ráta szerint: $\tilde{m}_{at} = m_{at} - \bar{m}_{at}$. Mivel ϵ_{at} normális eloszlásúak, tehát átírható az eredeti modell egy fixed effect modellé:

$$m_{at} \sim N(\sim\bar{\mu}_{at}, \sigma^2) \tag{25}$$

$$E(\tilde{m}_{at}) = \bar{\mu}_{at} = \beta_a\gamma_t. \tag{26}$$

Itt csak $A + T$ paraméterek használjuk az $A\ x\ T$ mátrix tagjainak becslésére:

$$\tilde{m} = \begin{pmatrix} \tilde{m}_{5,0} & \cdots & \tilde{m}_{5,4} \\ \vdots & \ddots & \vdots \\ \tilde{m}_{80,0} & \cdots & \tilde{m}_{80,4} \end{pmatrix} \tag{27}$$

A Lee-Carter modell felfogható egy kontingencia táblának is, ahol a cellák értékeit a paraméterek becsléseivel közelítjük. Feltételezzük, hogy a sorok (kor) és oszlopok (idő) függetlenek egymástól, illetve a várható értéke a cellának az adott peremértékek szorzata

$E(m_{at}) = \beta_\alpha\gamma_t$. (Lee & Miller, 2001) szerint ez a feltétel a legtöbb adatsorra nem teljesül.

β_a és γ_t becslésére maximum likelood módszert szoktak alkalmazni. Mivel végtelen ekvivalens maximuma van, a standard optimalizációs megoldások nem lesznek hatékonyak. Emiatt Lee és Carter cikkükben a centrált korprofilokat tartalmazó mátrix szinguláris érték felbontását alkalmazták (SVD) elsődlegesen.

Tétel (Szinguláris érték felbontás):

Egy tetszőleges $A \in \mathbb{R}^{m\times n}$ mátrix szinguláris érték felbontásán a

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T = U\hat{S}V^T, \quad \hat{S} \coloneqq \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix} \tag{28}$$

értjük, ahol $U \in R^{m\times m}, \quad V \in R^{n\times n}$ ortogonális mátrixot, és S diagonális mátrix:

$$S = \mathrm{diag}(\sigma_1, \dots, \sigma_r), \tag{29}$$

ahol $\sigma_1 \geq \cdots \geq \sigma_r > 0$ egyértelmű és A szinguláris értékeinek hívjuk.

Jelen esetben tehát

$$m \cong BLU', \tag{30}$$

ahol a becslés β-ra B első oszlopa, γ_t-re pedig $\beta'\bar{m}t$. Előfordulhat, hogy $\tilde{m}$ felbontása nem létezik, ekkor a $C = \tilde{m}\tilde{m}'$ mátrix legnagyobb sajátértékhez tartozó sajátvektoraként kiszámítható. (Girosi & King, 2007)

(Alho, 2000) megállapítása szerint a (18) egyenlettel leírt modell nem alkalmas teljes mértékben a vizsgált jelenség megbízható leírására, mivel a szinguláris értékfelbontáson alapuló legkisebb négyzetes becslés (OLS) egyik fő hátránya, hogy homoszkedasztikus hibákat feltételez, vagyis azt, hogy a hiba szórása minden korcsoportra nézve azonos. Ez azzal a feltételezéssel is összefügg, hogy a hibák normáleloszlásúak, ami a valóságban gyakran nem teljesül.

A gyakorlatban ugyanis a halandósági intenzitás (force of mortality) logaritmusának szórása jelentősen eltérhet az egyes korcsoportok között: az idősebb korcsoportok esetében jellemzően nagyobb szórással számolhatunk, mivel ezekben a csoportokban kevesebb haláleset történik. Ennek következtében a statisztikai ingadozások mértéke is nagyobb, ami megkérdőjelezi az OLS-módszer alapfeltételezéseinek érvényességét.

3.2 Poisson modell

Mivel a halálozások száma diszkrét, eseményszámláló jellegű változó, (Brillinger, 1986) nyomán indokoltabbnak tűnik Poisson-eloszlás feltételezése a halálozási adatokra. A legkisebb négyzetek módszerével kapcsolatban korábban említett problémák elkerülése érdekében a továbbiakban Poisson regressziós keretben megfogalmazott becslési eljárást alkalmazunk.

Legyen $D_{xt} \sim Poisson(E_x \mu_x(t))$, ahol $\mu_x(t) = \exp(\alpha_x + \beta_{k,t})$.

A paramétereket a (18) egyenletben megadott feltételek továbbra is korlátozzák. A halandósági intenzitás (force of mortality) a továbbiakban is egy log-bilineáris formát követ:

$$\ln(\mu_x(t) = \alpha_x + \beta_x * k_t \tag{31}$$

Az α_x, β_x és k_t paraméterek értelmezése a klasszikus Lee-Carter modellhez hasonló:

α_x: az életkorhoz kötődő átlagos halandósági szintet írja le, azaz megmutatja, hogy az adott korcsoportban milyen szinten helyezkedik el a halálozási arány időben átlagolva.

β_x: z adott életkor érzékenységét fejezi ki az időbeli halandósági változásokra, vagyis azt, hogy a halandósági trendek változása milyen mértékben érinti az adott korcsoportot.

k_T: az időben bekövetkező, minden életkorra kiterjedő halandósági szintváltozást reprezentálja; egyfajta globális időbeli trendet jelenít meg, amely az összes korcsoport halandóságára hatással van

A α_x, β_x és k_t paraméterek becslésére a továbbiakban nem a szinguláris értékfelbontáson alapuló módszert alkalmazzuk, hanem a (31) modell alapján felírt log-likelihood függvény maximalizálásával határozzuk meg ezeket a paramétereket. A választott valószínűségi keret, a halálozási események Poisson-eloszlású modellezése lehetővé teszi, hogy figyelembe vegyük az adatok diszkrét természetét és a különböző korcsoportok közötti varianciaeltéréseket is (Brouhns et al., 2002).

Ha $d_{x,t}$, az ténylegesen megfigyelt halálozások száma, akkor a likelihood függvény a következő:

$$L_{x,t}(\theta;\, d_{x,t}) = \frac{\lambda_{x,t}^{d_{x,t}}\, e^{-\lambda_{x,t}}}{d_{x,t}!} \tag{32}$$

a log-likelihood függvény ekkor

$$\begin{aligned}\ln L_{x,t}(\theta; d_{x,t}) &= \ln\left(\frac{\lambda_{x,t}^{d_{x,t}} e^{-\lambda_{x,t}}}{d_{x,t}!}\right) \\ &= \ln\left(\lambda_{x,t}^{d_{x,t}} e^{-\lambda_{x,t}}\right) - \ln(d_{x,t}!) \\ &= d_{x,t}\ln(\lambda_{x,t}) + \ln(e^{-\lambda_{x,t}}) - \ln(d_{x,t}!) \\ &= d_{x,t}\ln(\lambda_{x,t}) - \lambda_{x,t} - \ln(d_{x,t}!)\end{aligned} \tag{33}$$

Mivel a $D_{x,t}$-ről feltettük, hogy függetlenek, az összetett log-likelihood függvény felírhatjuk a következő alakban:

$$\begin{aligned}l(\theta) &= \sum_{x,t} d_{x,t}\ln(\lambda_{x,t}) - \lambda_{x,t} - \ln(d_{x,t}!) \\ &= \sum_{x,t} d_{x,t}\ln(E_{x,t}\mu_{x,t}) - E_{x,t}\mu_{x,t} - \ln(d_{x,t}!) \\ &= \sum_{x,t} d_{x,t}(\ln(E_{x,t}) + \ln(\mu_{x,t})) - E_{x,t}\mu_{x,t} - \ln(d_{x,t}!) \\ &= \sum_{x,t} d_{x,t}\ln(E_{x,t}) + d_{x,t}\ln(\mu_{x,t}) - E_{x,t}\mu_{x,t} - \ln(d_{x,t}!) \\ &= \sum_{x,t} d_{x,t}\ln(\mu_{x,t}) - E_{x,t}\mu_{x,t} - \ln(d_{x,t}!) + d_{x,t}\ln(E_{x,t}) \\ &= \sum_{x,t}\{d_{x,t}\ln(\mu_{x,t}) - E_{x,t}\mu_{x,t}\} + C \\ &= \sum_{x,t}\{d_{x,t}\ln(e^{\alpha_x+\beta_x k_t}) - E_{x,t}e^{\alpha_x+\beta_x k_t}\} + C \\ &= \sum_{x,t}\{d_{x,t}(\alpha_x + \beta_x k_t) - E_{x,t}e^{\alpha_x+\beta_x k_t}\} + C\end{aligned} \tag{34}$$

ahol a C konstans és $\theta = (\alpha_x, \beta_x, k_t)$.

A maximális valószínűségi becsléseket, az $\hat{\alpha}_x, \hat{\beta}_x, \hat{k}_t$ értékeket az $l(\theta)$ parciális deriváltjainak nullára állításával kapjuk meg. Iteratív módszert használunk az $\hat{\alpha}_x, \hat{\beta}_x, \hat{k}_t$ becslésére, mivel az optimalizálási problémát nem lehet zárt formulával megoldani, ezért egy frissítési sémát alkalmazunk, és a kiindulási értékekkel $\widehat{\alpha^0}_x, \widehat{\beta^0}_x, \widehat{k^0}_t$ kezdjük a frissítési folyamatot.

Az $v+1$-edik iterációs lépésben egy paraméterhalmazt frissítünk, miközben a többi paramétert az aktuális becsléseiknél rögzítjük, a következő frissítési séma használatával:

$$\widehat{\theta^{(v+1)}} = \widehat{\theta^{(v)}} - H^{-1}\left(\widehat{\theta^{(v)}}\right)\frac{\partial l}{\partial \theta}\left(\widehat{\theta^{(v)}}\right), \tag{35}$$

ahol $H = \frac{\partial^2 l}{\partial \theta\, \partial \theta'}$ Hesse-mátrix. Ezt az algoritmust hívják Newton-Raphson módszernek (Groot, 2011).

3.3 Előrejelzés

Ahhoz, hogy előrejelzést adjunk, Lee és Carter feltételezték, hogy β konstans, és hogy $\hat{\gamma}$ egy egyváltozós idősor modell. A legjobban illeszkedő idősor az ARIMA(0,1,0) eltolásos (driftes) véletlen bolyongás:

$$\hat{\gamma}_t = \hat{\gamma}_t - 1 + \theta + \xi_t \tag{36}$$

$$\xi \sim N(0, \sigma_{rw}^2),$$

ahol θ a drift paraméter, aminek ML becslése $\theta = \frac{\hat{\gamma}_T - \hat{\gamma}_1}{(T-1)}$, tehát csak az első és utolsó γ-tól függ a paraméter értéke. Ahhoz, hogy két periódusra jelezzünk előre beillesztjük $\hat{\theta}$ a drift paraméter helyére, $\hat{\gamma}_{t-1}$ helyére pedig az egyel visszaléptetett, definíció szerinti értékét helyettesítjük:

$$\hat{\gamma}_t = \hat{\gamma}_{t-1} + \hat{\theta} + \xi_t = \left(\hat{\gamma}_{t-2} + \hat{\theta} + \xi_{t-1}\right) + \hat{\theta} + \xi t. \tag{37}$$

Általánosítva $\hat{\gamma}_t$előrejelzését $T + (\Delta t)$ időpontra a következő írható fel:

$$\hat{\gamma}_{T+(\Delta t)} = \hat{\gamma}_T + (\Delta t)\hat{\theta} + \sum_{l=1}^{(\Delta t)} \xi_{T+l-1} \tag{38}$$

Ebből előrejelezhetők pontbecslések, amik egy egyenest követnek $\hat{\theta}$ meredekséggel.

$$\mu_{T+(\Delta t)} = \hat{\gamma}_T + (\Delta t)\widehat{\theta}. \tag{39}$$

Az elején felírt egyenlethez, ha az eddig tárgyaltakat behelyettesítjük, megkapjuk a log-mortalitás pontbecslését:

$$\mu_{T+(\Delta t)} = \bar{m} + \hat{\beta}\hat{\gamma}_{T+(\Delta t)}\,. \tag{40}$$

3.4 Limitációk

Számos erőssége és előnye ellenére az Lee-Carter modell nem hibátlan. Több erőfeszítés történt a szakirodalomban, hogy ezeket a problémákat a Lee-Carter modell kiterjesztésével a korlátokat

kezelje. Ebben a szakaszban bemutatom (Basellini, 2023) által publikált, a legjelentősebb és legszélesebb körben alkalmazott Lee-Carter kiterjesztéseket.

3.4.1 $\hat{k}_t$ paraméterezése

A Lee-Carter halandósági modell fő előnye, hogy a szinguláris érték-felbontás egyetlen tagjával képes megragadni a halandósági mintázatokat. Ez azonban pontatlanságot eredményezhet a becsült halálozási rátákban és az elhalálozások számában, különösen az idősebb korcsoportoknál. Ennek ellensúlyozására egy második lépésben módosítani lehet a $\hat{k}_t$ halandósági indexet, hogy az illesztett modell jobban megfeleljen a megfigyelt halálozási adatoknak. Basellini három módszert mutat be, egyik (Lee & Miller, 2001) megközelítése, miszerint az életkor-specifikus halálozások helyett a születéskor várható élettartamhoz igazítja a modellt, ami függetleníti a populációs adatoktól. (Booth et el, 2002) ötlete, hogy Poisson regresszióval és az eltérések minimalizálásával illeszti a modellt a megfigyelt életkor-specifikus halálozási számokhoz, csökkentve a legfrissebb időszak előrejelzési hibáját. Említésre méltó még (Rabbi & Mazzuco, 2021) munkája, ahol a születéskor várható élettartam helyett a halálozási életkorok szórását használják, hogy jobban tükrözze a halandósági javulásokat.

3.4.2 Kiindulási hibák

A Lee-Carter modell nem illeszkedik pontosan a megfigyelt halálozási rátákhoz a kiindulási évben, ami szintén előrejelzési hibát okozhat. (Lee & Miller, 2001) rámutattak, hogy az eredeti Lee-Carter előrejelzés torzítást okozott az USA születéskor várható élettartamában, ezért ők a megfigyelt kiindulási rátákat használták, hogy megszüntessék ezt a torzítást. (Stoeldraijer et al., 2018) nyolc európai ország adatait elemezve kimutatták, hogy a pontosság és a stabilitás között kompromisszum van: az utolsó megfigyelt év használata növeli a pontosságot, míg az utóbbi évek átlagának használata nagyobb stabilitást biztosít.

3.4.3 Statisztikai keretrendszer

A Lee-Carter modell statisztikai keretrendszere korlátozott, mert nem épít explicit sztochasztikus folyamatra, és a modell egyszerű szinguláris érték felbontást használ. Ez feltételezi, hogy a hibák normális eloszlásúak és homoszkedasztikusak, ami a valós halálozási adatoknál nem teljesül, mivel az idősebb korosztályokban a halálozási arányok nagyobb szórást mutatnak.

A statisztikai hiányosságok kezelésére több alternatív megközelítés született. Az egyik az ún. Poisson-keret, ami a halálesetek számát Poisson-eloszlással modellezi, ami jobban illeszkedik a halandósági adatok természetes változékonyságához (Brouhns et al., 2002). A túldiszperziós

poisson-modellnél egy negatív binomiális eloszlást vagy kvázi-likelihood módszereket alkalmaznak a variancia pontosabb modellezésére. (Wilmoth, 1993) Az adatok túldiszperziót mutatnak, ha a variancia nagyobb, mint az elméleti modell által feltételezett érték. Általánosított lineáris modellekkel is kiküszöbölhetők a hiányosságok, mivel a súlyozott legkisebb négyzetek vagy logit-transzformációval történő modellezés segít a paraméterek pontosabb becslésében (Renshaw & Haberman, 2003). Végül a Bayes-féle modellezés esetében pedig a Lee-Carter modell Poisson-keretbe ágyazása Bayes-statisztikai módszerekkel javítja az illesztést és az előrejelzést (Czado et al., 2005).

3.4.4 Rögzített b_x

A Lee-Carter modell egyik alapfeltevése, hogy a halandósági mintázat időben rögzített. Azonban (Lee & Miller, 2001) rámutattak, hogy több kutató szerint a modell paraméterei időben változhatnak. A 20. század első felében a fiatalok halandósága gyorsan csökkent, míg az időseké lassabban. A század második felében ez a tendencia megfordult, amit a modell nem tud megfelelően kezelni.

(Tuljapurkar et al., 2000) és (Lee & Miller, 2001) javaslata szerint az illesztési időszakot 1950-től kezdődően érdemes meghatározni, hogy a feltételezés jobban teljesüljön. (Booth et al., 2002) szintén korlátozta az illesztési időszakot a jobb illeszkedés érdekében. A rögzített mintázat problémája különösen hosszú távú előrejelzéseknél jelentős. (Li et al., 2013) ezt úgy oldotta meg, hogy bevezette a „rotáció" módszerét, amely lehetővé teszi a halandóság változó korstruktúráját. Egy időben változó paramétert alkalmaznak, amely fokozatosan módosul az eredeti szinguláris értékfelbontás alapján meghatározott értéktől egy végső mintázatig. Ez a megközelítés az ENSZ halandósági előrejelzéseiben is alkalmazásra került.

3.4.5 $\hat{k}_t$ linearitása

(Lee & Carter, 1992) a driftes véletlen bolyongás idősormodellt találták a legmegfelelőbbnek az Egyesült Államok adataihoz modellükben. Ez a lineáris modell az idősor első és utolsó pontján halad át. Bár ez a módszer széles körben használt, más ARIMA modellek esetenként pontosabbak lehetnek (de Jong & Tickle, 2006).

A Lee-Carter modell hosszú távú előrejelzésekre összpontosít, és hosszú idősorok esetén az esetleges lineáris eltérések kevésbé befolyásolják az eredményt. Azonban ezek az eltérések befolyásolhatják a drift paramétert, ami torzítást eredményezhet az előrejelzés első évében, valamint növelheti az előrejelzési bizonytalanságot, ezért az illesztési időszak helyes megválasztása kiemelten fontos a pontos előrejelzés érdekében.

3.4.6 Merevség

Az Lee-Carter modell csak a szinguláris érték felbontás első tagját használja, míg a többi komponensek alkotják a reziduálist. Ez a megközelítés azonban merevséget eredményez, mivel az első időindex lineáris. A magasabb rendű tagok nemlineáris mintázatai potenciálisan javíthatnák az előrejelzés pontosságát, de ezek jellemzően figyelmen kívül maradnak. Emellett az egyetlen időindex használata azt eredményezi, hogy a halandósági javulási tényezők korreláltak életkor szerint, ami nem valósághű.

3.4.7 Simítás

A Lee-Carter modell előrejelzése az Egyesült Államok halálozására nézve jó eredményeket mutatott, mivel nagy populáció és az ötéves korcsoportok stabilizáló és simító hatást eredményezett. Ennek ellenére a modell által illesztett és előrejelzett halálozási ráták gyakran ingadozóak, különösen, ha az elemzés egyéves korcsoportokat alkalmaz. Ez az életkori mintázatok szabálytalanságából és az időbeli trendek egyenetlenségéből fakad, utóbbi kumulálódó hatással van az előrejelzésre.

(Girosi & King, 2008) megjegyezték, hogy a modell életkor-specifikus halálozási rátái idővel egyre kevésbé lesznek simák, végül irreálissá válnak, és eltérnek bármilyen előre meghatározott trendtől, függetlenül a kezdeti adatoktól. Ennek a problémának a kezelésére különböző megközelítéseket javasoltak, amelyek szinte kivétel nélkül spline-alapú simítást alkalmaznak. Ez történhet az adatok előzetes simításával, paraméter becslés közbeni, vagy utólagos simításával.

3.4.8 Előrejelzési időszak

Az Lee-Carter modell egyik újítása az előrejelzési intervallum bevezetése volt. Azonban az eredeti modell által előrejelzett születéskor várható élettartam intervallumát több cikkben is kritika érte.

(Lee & Carter, 1992) az előrejelzési intervallum számításakor csak az innovációs hibát vették figyelembe, és ennek alapján származtatták az összes halandósági mérőszám előrejelzésének konfidencia intervallumát. Ez azonban nem veszi figyelembe, hogy a logaritmikus halálozási ráták életkor szerinti változásai korreláltak, így az intervallumok hibásak lehetnek.

Mivel az analitikus korrelációs becslés bonyolult, egy szimulációs megközelítés a preferált módszer. Ebben az eljárásban egy nagy mintaszámú véletlenszerűen generált jövőbeli halálozási séma segítségével határozzák meg az előrejelzési intervallumokat. Az egyes mutatók

(pl. halálozási ráták, várható élettartam stb.) előrejelzési intervallumát az önálló szimulált eloszlásaik percentilisei alapján számítják.

4. Gépi tanulás a mortalitás modellezésben

A gépi tanulás (ML) alkalmazása a biztosításmatematikában számos területen előnyös lehet. Az ML-alapú modellek hatékonyan alkalmazhatók kárbecslésre, biztosítási díjak meghatározására, csalások detektálására, valamint a kockázati portfólió optimalizálására. Mivel a biztosítási adatok gyakran nagy méretűek és heterogének, a gépi tanulási algoritmusok különösen hasznosak lehetnek az összetett kapcsolatok feltárásában és a pontosabb előrejelzések készítésében. A következő fejezetekben részletesen bemutatom a szakdolgozatban alkalmazott gépi tanulási módszereket. A kutatás során három különböző gépi tanulási algoritmust alkalmazok, két eltérő módszertani megközelítés szerint. Ennek megfelelően a dolgozatban összesen hat előrejelzési modellt elemzek és hasonlítok össze egymással.

4.1 Első megközelítés

(Levantesi & Pizzorusso, 2019) kutatása alapján vegyük a következő kategorikus változókat, amik egy egyedet írnak le: életkor (a), év (t), születés év (c), halálok (z), ország (l). Mindegyik egyedhez hozzárendeljük az $x = (a, t, c, z, l)\ \epsilon\ X$ változók vektorát, ahol $A = \{0, \dots, \mu\}, T = \{t_1, \dots, t_n\}, C = \{c_1, \dots, c_m\}, Z = \{S001, \dots S0016\},\ L = \{ESP, \dots, USA\}$.

Feltételezzük, hogy hogy a halálok száma D_x a következő feltételeknek eleget tesz:

- D_x független $x\ \epsilon\ X$ -től,
- $D_x \sim Poi(m_x \cdot E_x)\ \ \forall\ x \epsilon X$,

ahol m_x a centrált halálozási ráta, E_x a kitettség. Legyen d_x^{mdl} a becsült várható halálozások száma, m_x^{mdl} a hozzá tartozó centrált halálozási ráta. Ekkor továbbra is fennállnak a következő állítások:

- $m_x = m_x^{mdl}$
- $D_x \sim Poi(\vartheta_x \cdot d_x^{mdl})$, ahol $\vartheta_x \equiv 1,\ d_x^{mdl} = m_x^{mdl} E_x$.

A $\vartheta_x \equiv 1$ feltétel azt jelenti, hogy a választott mortalitási modell (jelen esetben Lee-Carter) tökéletesen illeszkedik az adatokra. Ez a valóságban általában nem áll fent, mivel gyakran alul- ($\vartheta_x > 1$) vagy túlilleszkedik ($\vartheta_x < 1$), ezért kalibráljuk ϑ_x-t x alapján, három különböző gépi tanulási módszerrel (döntési fa, véletlen erdő, és gradiens boosting). Legyen ϑ_x a gépi tanulási algoritmus eredménye, amit $\frac{D_x}{d_x^{mdl}}$-re fogok alkalmazni. Így az összefüggés:

$$\frac{D_x}{d_x^{mdl}} \sim a + t + c + z + l. \tag{41}$$

Legyen $\hat{\vartheta}_x$ gépi tanuló algoritmus által becsült paraméter, tehát a megoldása a fenti egyenletnek. $\hat{\vartheta}_x$ -val korrigáljuk a centrált halálozási rátát, hogy jobban illeszkedjen az adatokra:

$$\tilde{m}_x^{m\,dl} = \hat{\vartheta}_x \cdot m_x^{mdl}, \quad \forall x \epsilon X. \tag{41}$$

4.2 Második megközelítés

Ebben a megközelítésben is a klasszikus sztochasztikus halandósági modelleket gépi tanulási módszerekkel kombináljuk, hogy pontosabb előrejelzéseket kapjunk (Bjerre, 2022) cikke alapján. Az eljárás lényege, hogy először egy hagyományos modellt, például a Lee-Carter modellt illesztünk az adatokra, és ez alapján előrejelzést készítünk a halálozási ráták logaritmusára. Tehát először kiszámítjuk az aktuális és jövőbeli időpontra vonatkozó becsült értékeket, tehát m_x^{mdl} és m_{x+t}^{mdl}.

Ezt követően meghatározzuk a modell által vissza nem adott eltéréseket, vagyis a reziduumokat, amelyek a következő módon alakulnak:

$$r_t^{model} = m_t - m_x^{mdl} \tag{42}$$

Ezek a reziduumok tartalmazhatják azokat a mintázatokat, amelyeket a klasszikus modell nem tudott megragadni. Ezeket az eltéréseket egy gépi tanulási algoritmus segítségével modellezzük, ahol a magyarázóváltozók az előbbiekhez hasonlóan életkor (a), év (t), születés év (c), halálok (z), és ország (l):

$$r_t^{model} \sim a + t + c + z + l. \tag{43}$$

Miután az algoritmusokat ráillesztettük a reziduumok mintázataira, előre tudjuk jelezni a jövőbeli időpontra eső reziduumokat:

$$r_{t+h}^{model,DT}, r_{x+h}^{model,RF} \text{ és } r_{x+h}^{model,GB}. \tag{44}$$

Végül ezeket a becsült reziduumokat hozzáadjuk az eredeti klasszikus modell előrejelzéseihez, így egy javított előrejelzést kapunk:

$$m_{x+h}^{model,DT} = m_{x+h}^{model} + r_{x+h}^{model,DT}, \tag{45}$$

$$m_{x+h}^{model,RF} = m_{x+h}^{model} + r_{x+h}^{model,RF}, \tag{46}$$

illetve

$$m_{x+h}^{model,GB} = m_{x+h}^{model} + r_{x+h}^{model,GB}. \tag{47}$$

4.3 Döntési fa

A döntési fák olyan prediktív modellek, amelyek az adatok alapján hoznak létre egy fastruktúrát, és ezen keresztül képesek döntéseket hozni. Az egész modell egy hierarchikus struktúrát alkot, ahol az adathalmaz a gyökérben kezdődik, majd rekurzív osztásokon keresztül kisebb, homogénebb részhalmazokra bontódik, egészen a levelekig, amelyek a végső előrejelzést tartalmazzák.

Az osztási folyamat célja, hogy a csomópontokon belüli elemek minél homogénebbek legyenek, míg az egyes csomópontok közötti különbség a lehető legnagyobb legyen. Ezt különböző mérőszámokkal lehet mérni, amelyek közül az osztályozási problémáknál a leggyakrabban alkalmazottak a Gini-index és az entrópia.

A Gini-index azt méri, hogy egy csomópontban mekkora a valószínűsége annak, hogy két véletlenszerűen kiválasztott elem különböző osztályba tartozik. Ha p_i az i-edik osztály relatív gyakorisága, akkor a Gini-index képlete a következő:

$$\text{Gini} = 1 - \sum_{i=1}^{K} p_i^2 \,, \tag{48}$$

ahol K az osztályok száma. Az optimális osztás az, ahol a Gini-index minimális, azaz a csomópontban a lehető legnagyobb az osztályok közötti homogenitás.

Az entrópia a bizonytalanság mértékét fejezi ki egy csomóponton belül, és az információelmélet alapfogalma. Egy csomópont entrópiája az alábbi képlettel számolható:

$$H = -\sum_{i=1}^{K} p_i \log_2 p_i \,. \tag{49}$$

Az osztási folyamat során gyakran az információs nyereséget (angolul information gain) veszik figyelembe, amely az osztás előtti és utáni entrópiák különbségeként értelmezhető:

$$\text{Information Gain} = H_{\text{szülő}} - \sum_{j=1}^{M} \frac{N_j}{N} H_j \tag{50}$$

ahol $H_{szülő}$ a szülőcsomópont entrópiája, H_j a j-edik részcsomópont entrópiája, N_j az adott részcsomópont elemeinek száma, N a szülőcsomópont elemeinek száma, és M az osztások száma. Az optimális osztás az, amely maximális információs nyereséget eredményezi, így csökkentve a bizonytalanságot a részcsomópontokban (Cryer & Chan, 2008).

Regressziós problémák esetén a döntési fa célja az, hogy a célváltozó varianciáját minimalizálja az egyes csomópontokban. Ha egy csomópontban a célváltozó átlagát $\bar{y}$ jelöljük, akkor a variancia:

$$Var = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2, \tag{51}$$

ahol N az adott csomópontban található elemek száma. Az elágazás során az optimális osztási szabályt az határozza meg, hogy a két részcsomópont varianciájának, az adott csomópontban lévő elemek arányával súlyozva, mekkora csökkenést eredményez az eredeti varianciához képest:

$$\Delta\,\mathrm{Var} \;=\; \mathrm{Var}_{\mathrm{szülő}} \;-\; \left(\frac{\mathrm{N}_{\mathrm{bal}}}{\mathrm{N}}\;\mathrm{Var}_{\mathrm{bal}} \;+\; \frac{\mathrm{N}_{\mathrm{jobb}}}{\mathrm{N}}\;\mathrm{Var}_{\mathrm{jobb}}\right), \tag{52}$$

Az eljárás során az algoritmus rekurzívan választja ki a legjobb osztási szabályt minden csomópontban, amíg a további osztás nem eredményez számottevő javulást, vagy amíg egy előre meghatározott minimális csomópontméretet el nem érünk.

Fontos hangsúlyozni, hogy az algoritmus teljesítményét nagymértékben befolyásolják a hiperparaméterek. Például a fa maximális mélysége korlátozza a modell komplexitását, ezáltal segít megelőzni a túlilleszkedést. A csomópontok felosztásához szükséges minimális mintaszám, valamint a levelekben elvárt mintaszám biztosítja, hogy a fa ne váljon túlzottan specifikussá. Emellett a sztochasztikus elemek, például a tanításhoz használt véletlenszerű részhalmazok alkalmazása növeli a modell rugalmasságát és robusztusságát. E paraméterek megfelelő finomhangolása elengedhetetlen a predikciós pontosság és az általánosíthatóság közötti egyensúly eléréséhez (Hastie et al., 2009).

Mortalitás modellezés esetében legyen $(X_t)_{t\epsilon T}$ X-nek egy tagja, ekkor a döntési fa által becsült paraméter:

$$\hat{\vartheta}(x) = \sum_{t\epsilon T} \bar{\vartheta}_t \mathbb{1}_{\{x\epsilon\ \ T\}}. \tag{53}$$

4.4 Véletlen erdő

A véletlen erdő (random forest) egy tanulási módszer, amely több döntési fából épít fel egy erdőt, és ezen fák eredményeit aggregálja a végső predikciók meghatározásához. Az eljárás két alapvető technikán alapszik: a bagging (bootstrap aggregálás) és a véletlenszerű változó kiválasztás minden egyes fa osztási pontján. A bagging során a tanulási adathalmazból többszörös, véletlenszerű mintavételeket készítünk, így minden egyes döntési fát egy külön mintán tanítunk meg. Ez a módszer csökkenti az egyedi fák által mutatott magas varianciát, ami jelentős előny a modell robusztusságának növelésében. Ezen felül, a véletlenszerű változó kiválasztás biztosítja, hogy az egyes fák eltérő szerkezetűek legyenek, hiszen minden osztási

pontnál nem a teljes változóhalmazt, hanem csak annak egy véletlenszerű részét vesszük figyelembe. Ezt a technikát ensemble módszernek nevezik. Ennek köszönhetően csökken az egyes döntési fák közötti korreláció, ami az ensemble eljárások egyik legjelentősebb előnye.
Regressziós feladatok esetén a random forest végső előrejelzését úgy kapjuk meg, hogy az egyes fák előrejelzéseit átlagoljuk. Matematikailag ezt úgy írhatjuk le, hogy ha $\hat{\vartheta}^t(x)$ jelöli a t-edik fa előrejelzését, akkor a végső predikció

$$\hat{\vartheta}(x) = \frac{1}{T}\sum_{t=1} \hat{\vartheta}^t(x) \tag{54}$$

, ahol T a fák száma. Osztályozási feladatoknál pedig az egyes fák által adott osztálycímkék többségi szavazata határozza meg a végső predikciót:

$$\hat{\vartheta} = \text{mode}\left(\hat{\vartheta}^1(x), \hat{\vartheta}^2(x), \dots, \hat{\vartheta}^T(x)\right) \tag{55}$$

A modell hibájának varianciájára vonatkozó képlet azt mutatja, hogy a random forest a következő képlettel becsülhető:

$$\text{Var}\left(\hat{\vartheta}\right) = \rho\sigma^2 + \frac{1-\rho}{T}\sigma^2, \tag{56}$$

ahol σ^2 az egyes fák hibájának varianciája, ρ pedig az egyes fák közötti átlagos korreláció. Ez a képlet azt szemlélteti, hogy minél kisebb ρ (azaz minél kevésbé korreláltak az egyes fák), annál nagyobb a variancia csökkenés hatása, különösen akkor, ha T , azaz a fák száma növekszik (Breiman, 2001).
A fenti elméleti háttér rávilágít arra, hogy a random forest nem csak az egyedi döntési fák előnyeit ötvözi, hanem a bagging és a véletlenszerű változó kiválasztás révén jelentősen javítja az előrejelzések pontosságát és robusztusságát, miközben csökkenti a túltanulás (overfitting) kockázatát.

4.5 Gradiens boosting

A gradient boosting egy olyan gépi tanulási módszer, amelyet (Friedman, 2001) fejlesztett ki, és amelynek célja gyenge predikciós modellek, jellemzően döntési fák iteratív módon történő kombinálása egy erősebb modellé. A módszer lényege, hogy minden egyes iterációban egy új

döntési fát illesztünk az előző modellek hibáira, így csökkentve a végső előrejelzési hibát. Formálisan, az algoritmus egy adott veszteségfüggvény minimalizálását célozza meg. Legyen F a predikciós modell, amely iteratívan épül fel, és jelölje $\hat{m}_k(F)$a k-adik iterációban meghatározott predikciót. Ekkor a frissítési szabály a következő:

$$\hat{m}_k(F) = \hat{m}_{k-1}(F) + v \sum_{j=1}^{J_k} \gamma_{j,k} I\left(F \in R_{j,k}\right) \tag{57}$$

ahol v egy konstans tanulási ráta, J_k k-adik fa leveleinek száma, $\gamma_{j,k}$ pedig a j-edik leveléhez tartozó predikciós érték, amelyet a negatív gradiens alapján határozunk meg. A negatív gradiensek ($\gamma_{j,k}$) a veszteségfüggvény deriváltjaiból származnak, és az aktuális modell hibáit reprezentálják.

(Friedman, 2001) továbbfejlesztette az algoritmust a sztochasztikus gradient boosting bevezetésével, amely jelentősen javítja mind a predikciós teljesítményt, mind a számítási hatékonyságot. Ez a megközelítés az egyes iterációkban a teljes adathalmaz helyett annak egy véletlenszerű részhalmazát használja fel az új döntési fa illesztéséhez.

5. Kiértékeléshez használt metrikák

A gépi tanulási modellek előrejelző teljesítményének objektív értékeléséhez elengedhetetlen a megfelelő kiértékelő metrikák alkalmazása. A regressziós problémák esetében, ahol a cél a valós számok előrejelzése, a három leggyakrabban használt hiba-alapú mérőszám a gyökér-négyzetes átlagos hiba (RMSE), a medián abszolút hiba (MAE) és a százalékos átlagos abszolút hiba (MAPE). Ezek a metrikák különböző tulajdonságokkal rendelkeznek, így eltérő szempontok alapján adnak visszajelzést a modell teljesítményéről.

Az RMSE a becsült értékek és a tényleges megfigyelések közötti eltérések négyzetének átlaga után vett négyzetgyök. Kiszámítási képlete:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{58}$$

ahol y_i a tényleges, $\hat{y}_i$ pedig a becsült érték, n pedig a megfigyelések száma. Az RMSE erősen érzékeny a nagy hibákra, mivel a négyzetre emelés kiemeli a kiugró eltéréseket. Éppen ezért különösen hasznos akkor, ha a nagy előrejelzési hibák elkerülése elsődleges cél.

A MAE a tényleges és becsült értékek közötti abszolút eltérések számtani átlaga:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{59}$$

Ez a metrika egyenlően kezeli a kisebb és a nagyobb hibákat, és kevésbé érzékeny a kiugró értékekre, mint az RMSE. Előnye, hogy közvetlenül értelmezhető az adatok mértékegységében.

A MAPE a hibákat a tényleges értékhez viszonyítja, így százalékos formában fejezi ki az előrejelzési hiba nagyságát:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{60}$$

Ezáltal lehetővé teszi különböző skálájú problémák összehasonlítását is. Hátránya ugyanakkor, hogy ha a tényleges értékek közel vannak a nullához, akkor a MAPE értelmezhetetlenné válhat vagy extrém nagy hibát eredményezhet. Ennek kiküszöbölésére bizonyos esetekben módosított százalékos hibamértékeket alkalmaznak.

6. Adatok bemutatása

Az elemzés során felhasznált adatok a *Human Mortality Database* (HMD*)* részeként elérhető *The Human Cause-of-Death Data series* (HCD@HMD) adatbázisból származnak. A HCD@HMD sorozat a HMD egy kiegészítő adathalmaza, amely más haláloki adatforrásoktól eltérően egységesített, időben összehasonlítható halálozási okokat tartalmaz. A halálokok besorolása egy állandó, fix halálokozati listán alapul, amely figyelembe veszi a *Betegségek Nemzetközi Osztályozásában* (ICD) történt változásokat is, így biztosítva az időbeli összehasonlíthatóságot. Az adatbázis időszakosan frissül, és az adatok országos szintű bontásban érhetők el.

A jelen dolgozat elemzéseiben kizárólag a halálesetek abszolút számára, valamint a népességre vetített kitettség értékeire volt szükség, életkor és nem szerinti bontásban. Az adatok országos szinten kerültek feldolgozásra.

Ország	Adatok elérhetősége
Fehéroroszország	1955-2019
Franciaország	1979-2020
Japán	1981-2021

Lettország	1956-2019
Litvánia	1956-2019
Moldova	1965-2018
Lengyelország	1959-2019
Románia	1980-2019
Spanyolország	1980-2021
USA	1979-2021

1. táblázat Az elemzésben szereplő országok

Mivel az egyes országokban eltérő időszakokra állnak rendelkezésre halandósági adatok, a tanuló- és teszthalmaz szétválasztása minden esetben az adott ország saját idősorához igazodva, egységesen 70-30%-os arányban történt. Az adatok forrása a Human Mortality Database (HMD), amely részletes dokumentációval ismerteti a halandósági ráták és táblák előállításának módszertanát. A feldolgozási folyamat hat fő lépésből áll, amelyek hat különálló adatcsoporthoz kapcsolódnak, és ezek mindegyike elérhető a HMD adatbázisában.

Az első lépés az éves, nemek szerinti élveszületési adatok összegyűjtése az egyes országokra vonatkozóan, lehetőség szerint hosszabb időtávon. Ezekre az adatokra elsősorban a fiatalabb korcsoportokra vonatkozó népességbecslések pontosítása érdekében van szükség. Ezt követi a halálozások rögzítése, amely során az adatgyűjtés célja a lehető legnagyobb részletezettség biztosítása. Amennyiben az adatok csak aggregált formában érhetők el, a HMD egységes becslési eljárásokat alkalmaz a halálesetek születési év, halálozási év és betöltött életkor szerinti megbontására.

A harmadik lépés a népességméret éves, január 1-jei állapotának megbecslése. Ezt vagy közvetlenül külső forrásokból szerzik be, vagy népszámlálási adatok, valamint születési és halálozási számok kombinációjával számítják ki. Ezt követi a halálozási kockázatnak kitett népesség, az ún. expozíció meghatározása, amely az előző lépésben becsült népességértékeken alapul, kiegészítve egy korrekciós tényezővel. Ez utóbbi figyelembe veszi, hogy a halálesetek nem egyenletesen oszlanak el az év során az egyes korcsoportokon belül.

Az ötödik lépésben kerülnek kiszámításra a halandósági ráták, amelyek az adott kor-idő intervallumra eső halálesetek számának és a megfelelő kitettségi értéknek a hányadosai. Végül a hatodik lépésben ezekből a rátákból halálozási valószínűségek származtathatók, amelyek alapján összeállíthatók a halandósági táblák. Ezek a táblák olyan kulcsfontosságú demográfiai mutatókat tartalmaznak, mint például a születéskor várható élettartam és más élettartamra vonatkozó statisztikák.

Fontos megjegyezni, hogy a HMD által közzétett adatokból a nyilvánvaló, durva hibákat kiszűrték és javították. Ilyen például, ha egy statisztikai táblázatban egy érték tévesen nagyságrendekkel nagyobb (például 3 800 helyett 38 000), ezek az egyértelmű rögzítési hibák automatikusan korrigálásra kerülnek. Ugyanakkor az adatok nem kerültek módosításra a szisztematikus életkormeghamisítás (például téves életkor-megadás) vagy a lefedettségi hibák (az események vagy népesség túl- vagy alulszámlálása) esetében, így ezek a torzulások a vizsgálatban is jelen lehetnek.

Az elemzéshez használt két adatstruktúra (D_x és E_x) szinte megegyező, a kitettségekben értelemszerűen nem szerepelnek halálokok:

Változó	Leírás
country	Országok azonosítója
year	Év
age	Életkor
cohort	Születési év
cause	Halál oka

2. táblázat Az elemzésben használt változók

Mivel a jelen kutatás elsődleges célja annak vizsgálata, hogy a halandósági ráták modellezésének módszertana országonként és halálokonként eltérő eredményekhez vezet-e, az adatok minden ország esetében aggregálásra kerültek nemek szerint. A vizsgálatban szereplő adatbázis korcsoportos bontásban tartalmazza az egyes halálokokhoz kapcsolódó halálesetek számát. Az alábbi változók az adott haláloki kategóriák szerint, korcsoportonként rögzített esetszámokat jelölik:

Változó	Leírás
total	Minden korosztály aggregálva
d0	0
d1	1-4
d5	5-9
d10	10-14
d15	15-19
d20	20-24
d25	25-29
d30	30-34
d35	35-39

d40	40-44
d45	45-49
d50	50-54
d55	55-59
d60	60-64
d65	65-69
d70	70-74
d75	75-79
d80	80-84
d85p	85+

3. táblázat Az elemzésben használt korcsoportok

A kitettségi (expozíciós) adatok feldolgozása során ugyanazt a korcsoportonkénti bontást alkalmaztam, mint a halálesetek esetében, biztosítva ezzel az összehasonlíthatóságot a két adattípus között. A halálozási okok kategorizálása a Betegségek Nemzetközi Osztályozása (International Classification of Diseases, ICD) 10. verziója szerint történt. Az adatbázis többféle aggregálási szinten kínál haláloki bontást, azonban a jelen szakdolgozatban a legáltalánosabb, legmagasabb szintű (legösszesítettebb) haláleseti kategóriákat alkalmaztam az elemzéshez.

Sorszám	**Név**	**IDC10 kód**
S001	Fertőző betegségek	A00-B99
S002	Daganat	C00-D48
S003	A vér és a vérképző szervek betegségei	D50-D89
S004	Endokrin, táplálkozási és anyagcsere-betegségek	E00-E88
S005	Mentális és viselkedési zavarok	F01-F99
S006	Az idegrendszer és az érzékszervek betegségei	G00-G44, G47-H93
S007	Szívbetegségek	I00-I51
S008	Cerebrovaszkuláris betegség	G45, I60-I69
S009	A keringési rendszer egyéb és nem meghatározott rendellenességei	I70-I99, K64
S010	Akut légzőszervi megbetegedések	J00-J22, U04, U07
S011	Egyéb légzőszervi betegségek	J30-J98

S012	Emésztőrendszeri megbetegedések	K00-K63, K65-K92
S013	A bőr és a bőr alatti szövetek, a vázizomzat betegségei és a kötőszövet	L00-M99
S014	A húgy- és nemi szervek betegségei, valamint a terhességgel, szüléssel és gyermekágyi időszakkal kapcsolatos szövődmények.	N00-O99
S015	A perinatális időszakban kialakuló állapotok és veleszületett rendellenességek/anomáliák	P00-Q99, R95
S016	Egyéb külső okok	V01-Y89

4. táblázat A HMD adatbázisban szereplő halálokok nevei és a hozzá tartozó IDC10 kód

Az elemzés során országonként kizárólag a kettő legmagasabb mortalitási rátával rendelkező halálozási ok modellezésére kerül sor.

7. Eredmények értékelése

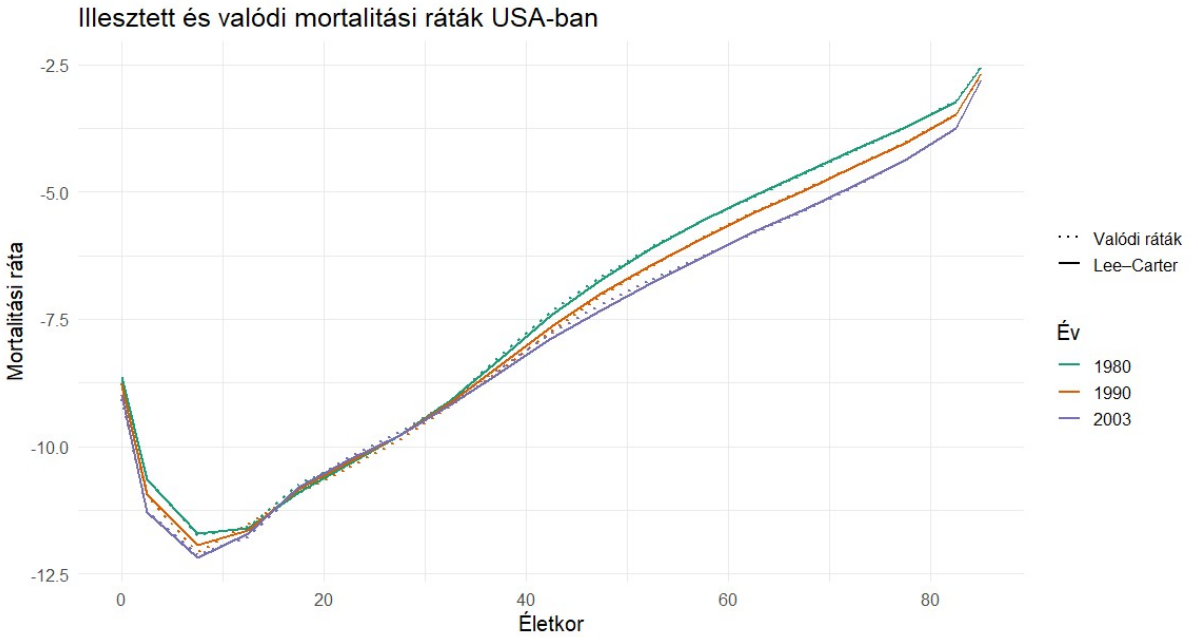
Az elemzés során néhány egyszerűsítő feltétellel éltem, amelyeket az alábbiakban foglalok össze.

- Elsőként feltételeztem, hogy a halálesetek egymástól független események, tekintettel arra, hogy egy személy kizárólag egy adott halálok következtében halhat meg. Ugyanakkor ez a megközelítés figyelmen kívül hagyja a versengő kockázatok jelenségét, amely szerint egy adott halálok bekövetkezése megakadályozza más halálokok realizálódását. Ennek következtében, ha például a jövőben a rákos megbetegedések kezelése javul, és emiatt kevesebben halnak meg rákban 60 évesen, akkor ezek az emberek potenciálisan később más okból, például szív- és érrendszeri betegségben halhatnak meg 65 évesen. Az ilyen típusú eltolódások azt eredményezhetik, hogy bizonyos halálokok előrejelzése alulbecsült lesz, mivel a modell nem képes megfelelően kezelni azt a dinamikát, amikor az egyik halálok csökkenése másik halálok gyakoribbá válását vonja maga után.
- Az elemzés nem különítette el a férfi és női halandósági rátákat, mivel a nemi megkülönböztetés a biztosítási gyakorlatban, az ide vonatkozó uniós irányelvek értelmében nem megengedett, így nem releváns tényező.
- Az egyes országok esetében eltérő mennyiségű adat áll rendelkezésre, ezért az elemzett időszak hossza országonként különböző. A modell tanításához és teszteléséhez használt

adathalmazokat minden ország esetében külön, 70-30 százalékos arányban osztottam fel tanuló és teszt adatkészletre, nem pedig az összevont, teljes adatbázis alapján.

- Mivel halálokokat vizsgáltam, mind a meghaltak számánál, mind a kitettségeknél csak korcsoportok álltak a rendelkezésemre, így az elemzéshez a korcsoportok számtani közepét használtam, mint életkor.
- A vizsgálat középpontjába azokat a halálokokat állítottam, amelyek minden ország esetében a leggyakoribbak közé tartoznak.
- A Lee-Carter modell illesztéséhez minden ország és halálok kombinációjára külön halandósági táblát állítottam össze, és azokat egymástól független populációként kezeltem. Ennek eredményeképpen összesen 10 ország × 2 halálok = 20 darab halandósági tábla képezte az elemzés alapját.

A szükséges adat-előkészítési lépéseket követően ezekre a szeparált halandósági adatsorokra külön-külön illesztettem a poisson eloszlású Lee-Carter modellt az R-ben található StMoMo csomaggal. Így tehát a paraméterbecsléshez maximum likelihood módszert használtam.



1. ábra Lee-Carter modell illesztése amerikai adatokra

A fenti ábra a tanító adatokon végzet Lee-Carter illesztést mutatja három évben, csak a szívbetegségben (S007) elhunytak alapján.

Az illesztést követően mindkét korábban bemutatott módszertani megközelítést elvégeztem. A gépi tanulási algoritmusok alkalmazása során minden esetben öt rétegű keresztvalidációs

eljárást alkalmaztam a hiperparaméterek optimális kiválasztása érdekében, amely biztosította a modellek túlilleszkedésének elkerülését. A (x) táblázat tartalmazza a random forest keresztvalidációjának eredménytábláját. A modell paraméterei a legkisebb RMSE alapján lettek kiválasztva.

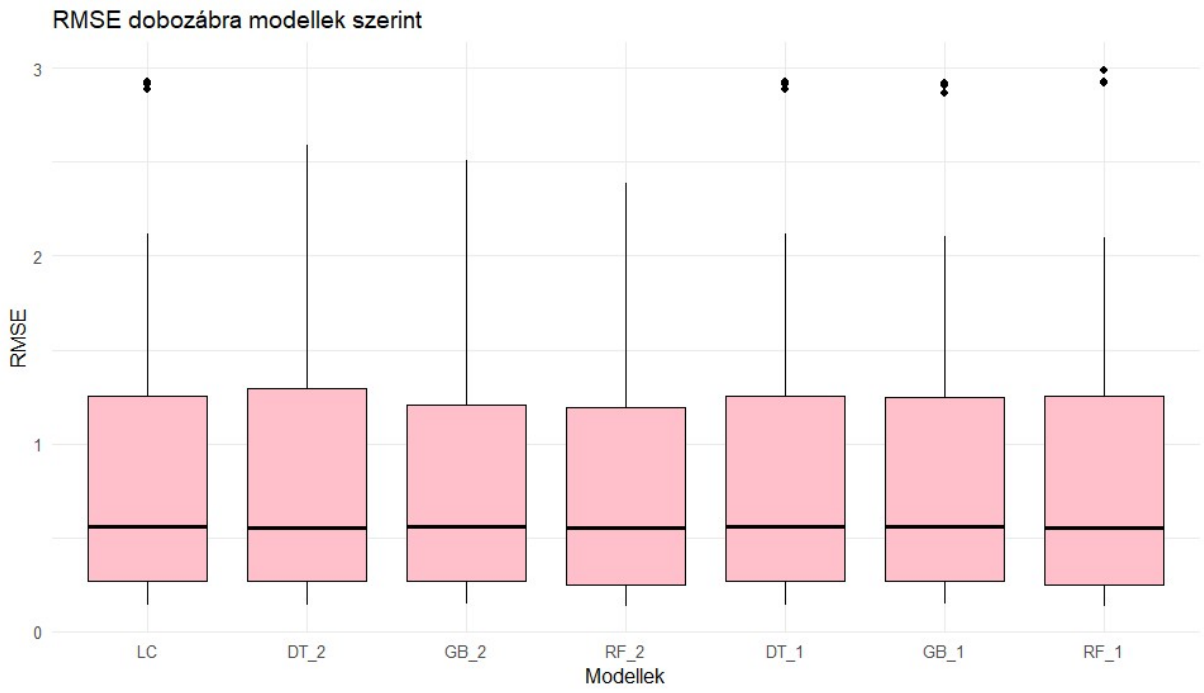
RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0,3910	0,0073	0,1428	0,0168	0,0096	0,0041
0,3897	0,0127	0,1416	0,0165	0,0099	0,0039
0,3886	0,0174	0,1395	0,0163	0,0127	0,0038
0,3875	0,0226	0,1367	0,0161	0,0137	0,0036
0,3869	0,0271	0,1341	0,0169	0,0111	0,0038

5. táblázat Keresztvalidáció eredménye

Az előrejelzéseket a kizárólag tesztadathalmazon hajtottam végre, tehát az értékelés kizárólag a korábban nem látott adatokon történt, ezzel szimulálva a valós alkalmazási környezetet. Az egyes modellek teljesítményének objektív értékeléséhez a logaritmizált mortalitási rátákra a korábban bemutatott három sztenderd hibamértéket alkalmaztam (RMSE, MAE, MAPE). Ezeket az értékeket többféle bontásban számítottam ki, lehetővé téve a modellek teljesítményének részletes és többdimenziós összehasonlítását.

Jelölés	Modell
DT_1	Első megközelítés – döntési fa alapú
RF_1	Első megközelítés – véletlen erdő alapú
GB_1	Első megközelítés – gradient boost alapú
DT_2	Második megközelítés – döntési fa alapú
RF_2	Második megközelítés – véletlen erdő alapú
GB_2	Második megközelítés – gradient boost alapú

6. táblázat Modellekhez használt jelölésrendszer



2. ábra RMSE alakulása modellenként

A 2. ábrán szereplő modellek közé tartozik a klasszikus Lee-Carter (LC) modell, valamint hat gépi tanulási eljárás: döntési fa (DT_1, DT_2), gradient boosting (GB_1, GB_2), illetve random forest (RF_1, RF_2). Mivel modellenként külön becslést alkalmaztam országonként, azon belül halálokonkénti bontásban, az így keletkezett RMSE-k értékeinek eloszlását dobozdiagramon ábrázoltam.

A dobozdiagramról leolvasva nem egyértelmű, de a Lee-Carter modell medián RMSE értéke a legmagasabb, ami arra utal, hogy ez a modell nagyobb tipikus hibával dolgozik, mint a gépi tanulási alternatívák. Emellett a LC modellnél megfigyelhető a nagy szórás is, több kiugró értékkel, ami a predikciók változékonyságára és kevésbé robusztus jellegére utal.

Ezzel szemben a gépi tanulási modellek, különösen a második módszertan szerintiek (DT_2, GB_2, RF_2), lényegesen alacsonyabb medián értékeket mutatnak. Különösen kiemelkedő az RF_2 modell, amely alacsony hibaszintet ér el, miközben az értékek szórása is kismértékű marad. Ez arra enged következtetni, hogy az RF_2 modell nemcsak pontosabb, hanem stabilabb is a predikció során.

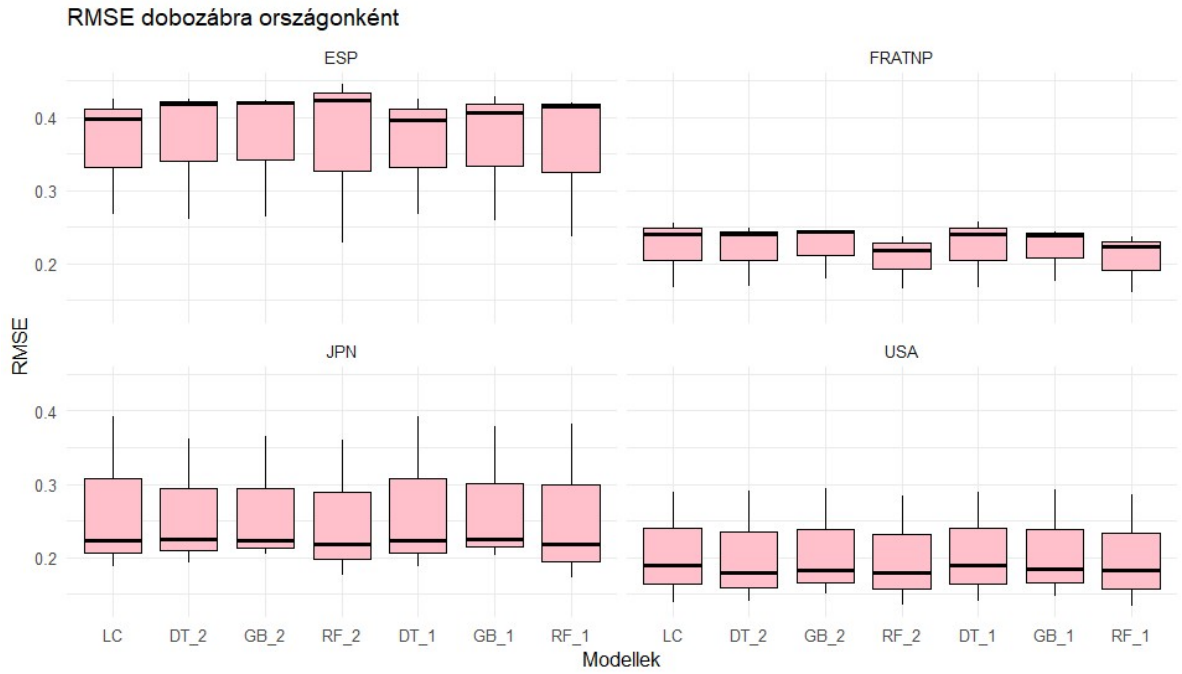
Az első megközelítésből való modellek (DT_1, GB_1, RF_1) medián értéke szintén alacsonyabb a LC modellénél, ugyanakkor ezeknél nagyobb szórás és több kiugró érték figyelhető meg, ami az algoritmusok érzékenységét jelzi bizonyos predikciós esetekre.

7.1 Országonkénti eredmények

Az országonkénti bontás során indokoltnak láttam az országok két csoportra osztását, mivel a becslési hibák nagysága alapján egyértelmű elkülönülés figyelhető meg. Ez az eltérés összefüggésbe hozható a rendelkezésre álló adatok mennyiségével, amely nagyvonalakban a népességszám szerinti kategóriákra is ráilleszthető.

Ország	Mintanagyság (meghaltak száma)
USA	206 277 306
Japán	83 869 238
Franciaország	45 812 542
Lengyelország	41 435 869
Spanyolország	30 394 352
Románia	20 481 448
Lettország	4 630 439
Moldova	4 271 163
Litvánia	3 809 270
Fehéroroszország	2 165 409

7. táblázat Országonkénti mintanagyság



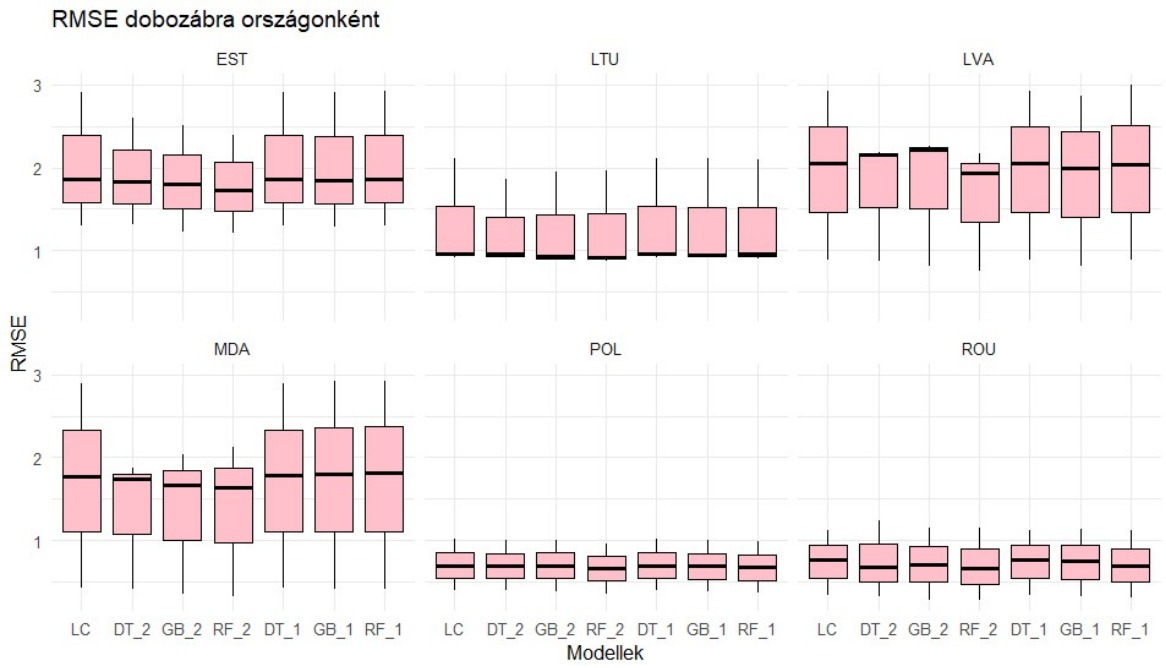
3. ábra RMSE alakulása országonként (ESP, FRATNP, JPN, USA)

Az országokat két csoportba soroltam: az első csoportba Spanyolország, Franciaország, Japán és az Egyesült Államok tartozik, míg a második csoportot Lengyelország, Románia, Lettország, Moldova, Litvánia és Fehéroroszország alkotja.

Az országonként nézve több halálokra készült előrejelzés, így a kapott RMSE-ket itt is érdemes egy dobozdiagramon ábrázolni. A nyugat-európai és tengerentúli országok modellillesztéseinek RMSE-értékei alapján megfigyelhető a 3.ábrán, hogy a különböző gépi tanulási megközelítések rendszerint pontosabb előrejelzéseket nyújtanak, mint a klasszikus Lee-Carter modell. A négy vizsgált ország (Spanyolország, Franciaország, Japán és az Egyesült Államok) eltérő demográfiai mintázatai ellenére a javulás tendenciája általánosan érvényesül.

Franciaország és az USA esetében a különbségek különösen szembetűnők: a gépi tanulási modellek alacsonyabb medián RMSE értékkel és szűkebb eloszlással rendelkeznek, ami erősebb illeszkedést és kisebb hibaszórású predikciókat jelent. Ez arra utal, hogy ezekben az országokban a klasszikus modellek nem tudják megfelelően megragadni a halandósági trendek komplexitását, míg a gépi tanuló modellek képesek alkalmazkodni az összetettebb szerkezetekhez.

Spanyolországban a különbségek mérsékeltebbek, de a gépi tanulási modellek itt is előnyt mutatnak. A legkisebb eltéréseket Japánban figyelhetjük meg, ahol a Lee-Carter modell teljesítménye közel azonos szinten van a gépi tanulási modellekével. Ez arra utalhat, hogy Japán halálozási mintázatai időben és kor szerint annyira egyenletesek és kiszámíthatók, hogy egy klasszikus lineáris modell is jól működik.

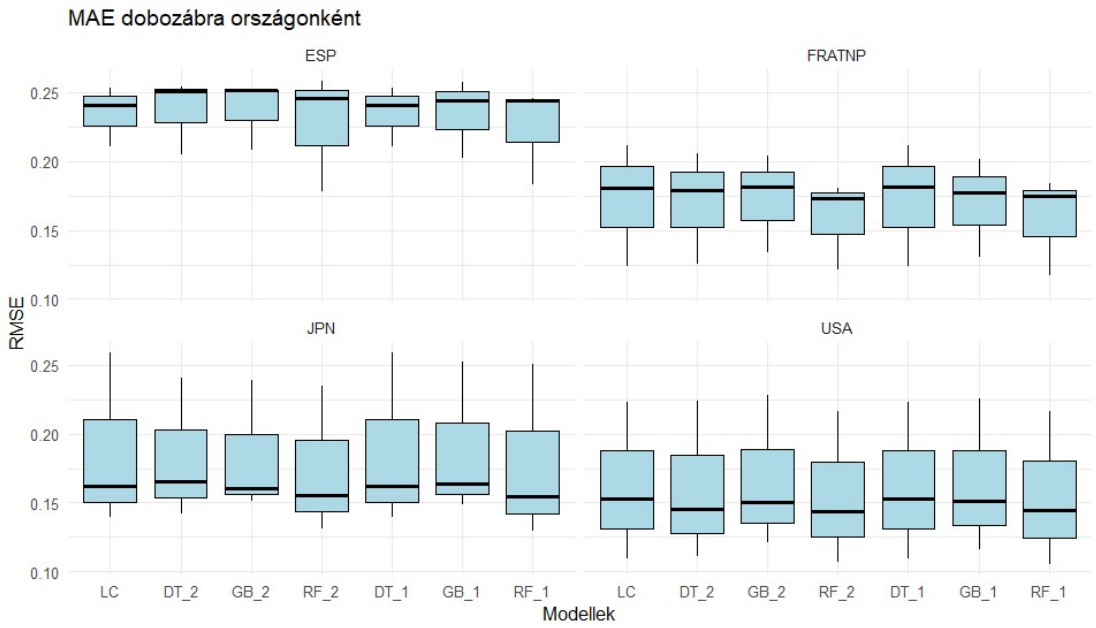


4. ábra RMSE alakulása országonként (EST, LTU, LVA, MDA, POL, ROU)

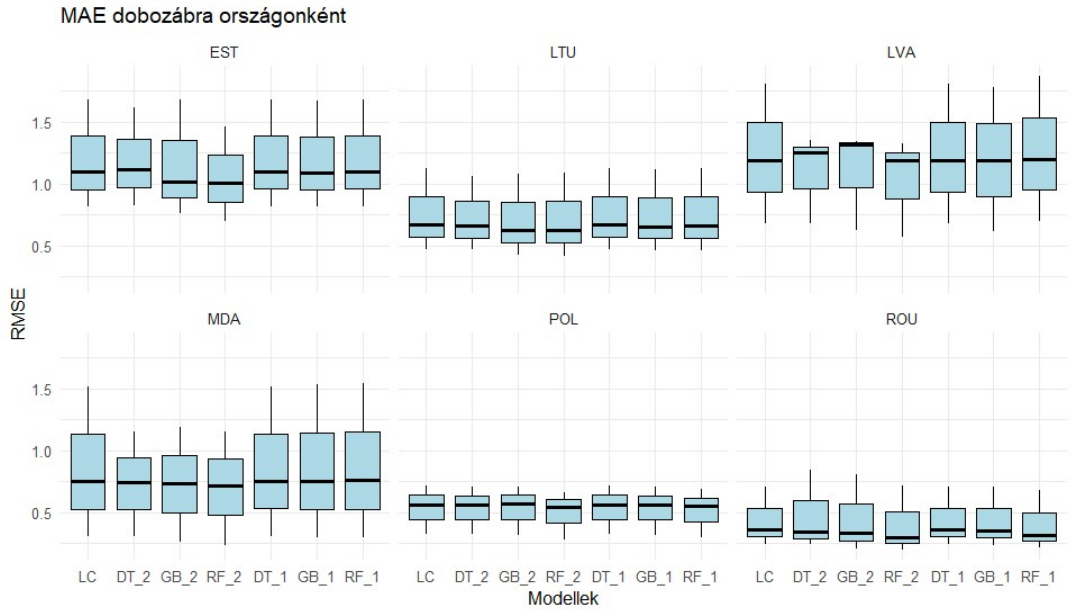
A kelet-közép-európai országokban, ahol gyakran tapasztalható nagyobb halandósági ingadozás, a gépi tanulási modellek különösen fontos szerepet töltenek be a predikciós hibák csökkentésében. A vizsgált országokban (Észtország, Litvánia, Lettország, Moldova, Lengyelország, Románia) eltérő mértékben, de szinte minden esetben megfigyelhető a gépi tanulási modellek javulása a Lee-Carter modellhez képest.

A legnagyobb javulás Litvániában, Moldovában és Lengyelországban jelentkezik, ahol a RF_2 és GB_2 modellek alacsonyabb RMSE értéket és kisebb szórást mutatnak. Ezekben az országokban a klasszikus modell hibája nemcsak magasabb, de eloszlása is szélesebb, ami a predikciók megbízhatatlanságát jelzi. A gépi tanulási modellek viszont jobban tudják kezelni a nemlineáris mintázatokat, valamint a heterogén adatokból is képesek tanulni.

Ugyanakkor olyan országokban, mint Lettország és Észtország, a javulás mérsékeltebb. Ezekben az esetekben bár a gépi tanulási modellek mediánja jellemzően alacsonyabb, a szórásuk hasonló vagy nagyobb is lehet, mint a klasszikus modellé. Ez felveti annak lehetőségét, hogy bizonyos országok esetében a modellteljesítmény javítása nemcsak a módszertani választáson, hanem a bemeneti adatok minőségén és stabilitásán is múlik.



5. ábra MAE alakulása országonként (ESP, FRATNP, JPN, USA)



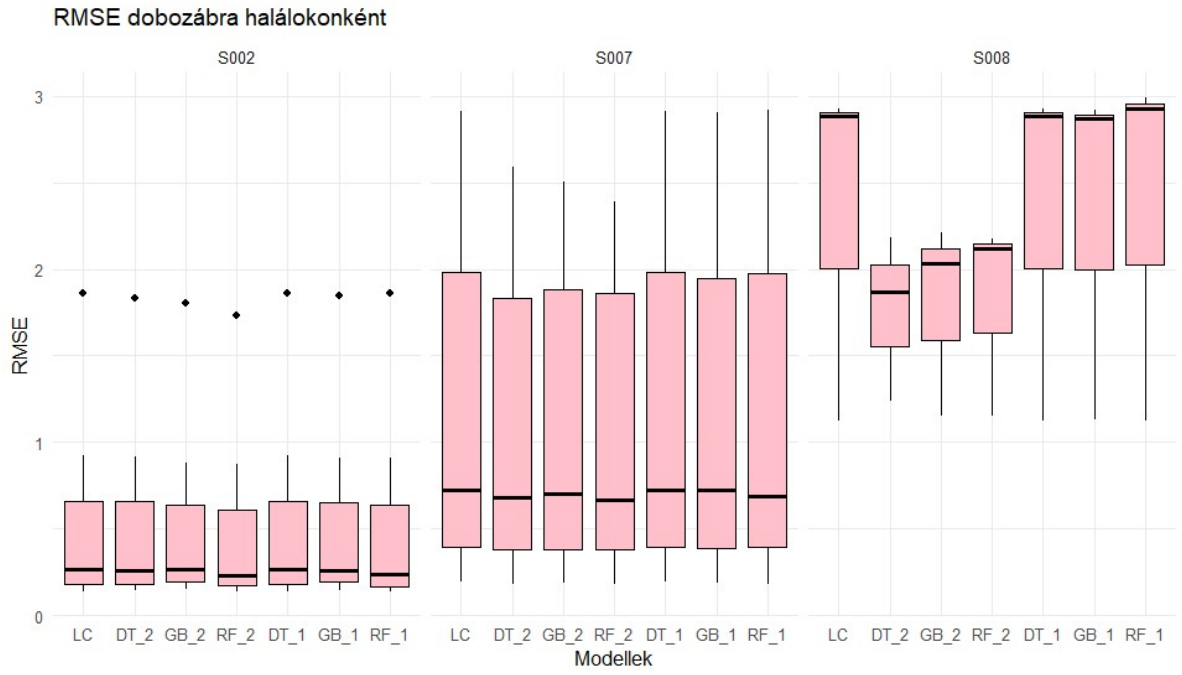
6. ábra MAE alakulása országonként (EST, LTU, LVA, MDA, POL, ROU)

A modellek országonkénti teljesítményének összehasonlítását az MAE értékek elemzésével is elvégeztem. Ez látható az 5-6. ábrán. Míg az RMSE a nagyobb hibákra erősebben reagál, az MAE kiegyensúlyozottabb képet ad az átlagos predikciós eltérés mértékéről, ezért alkalmas a modellek stabilitásának, valamint a szélsőérték-érzékenység értékelésére.

Több országban (pl. ESP, JPN, EST, LVA) az MAE értékek kisebb szórást mutatnak, mint az RMSE, ami arra utal, hogy bár lehetnek szélsőséges predikciós hibák (amit az RMSE jobban kiemel), a modellhibák többsége mérsékelt. A gépi tanulási modellek itt is jellemzően jobb

teljesítményt mutatnak, de a különbségek kevésbé látványosak, mint az RMSE esetében. Ez arra utal, hogy a LC modell több esetben elsősorban a nagy hibák miatt teljesít rosszul, míg a kisebb eltérések gyakoriságában kevésbé marad el a gépi tanulástól.

7.2 Halálokonkénti eredmények

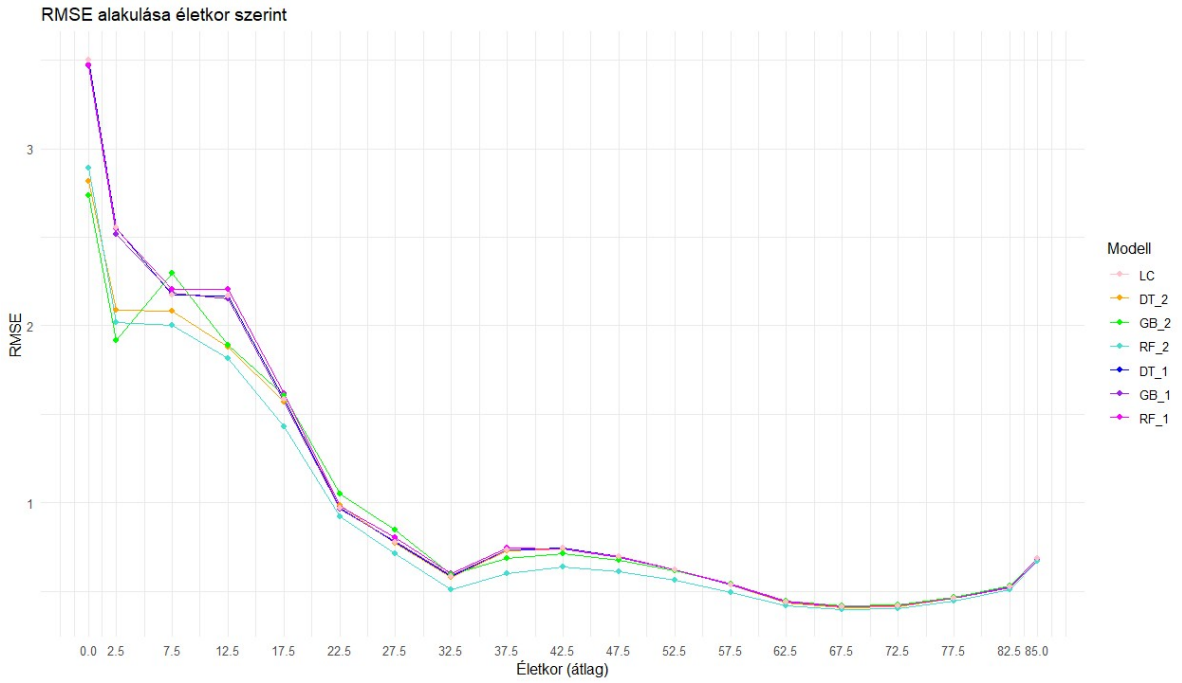


7. ábra RMSE alakulása halálokonként

Halálozási okonként csoportosítva az eredményeket, a modellek összehasonlítása alapján a véletlenerdő-alapú megközelítések mutatják a legstabilabb teljesítményt a daganat (S002) és szívbetegség (S007) általi halál kategóriákban, amelyet alacsonyabb medián RMSE-értékek, illetve daganat (S002) okozta halál esetében szűkebb szórás is alátámaszt. A daganatos halálozás (S002) kategóriában megfigyelhető kiugró értékek elsősorban Fehéroroszországnak tulajdoníthatók, mivel ebből az országból áll rendelkezésre a legkevesebb adat erre a halálokra vonatkozóan. Ezzel szemben a cerebrovaszkuláris betegségben elhunytak (S008) kategória esetében az RMSE értékek jelentősen volatilisabbak. Itt a második módszerrel illesztett modellek számottevő javulást eredményeznek, míg az első módszer gyakran nem nyújt érdemi előrelépést a Lee-Carter modellhez képest, sőt, bizonyos esetekben rontja is a becslés pontosságát. A következtetések alapján a daganat általi halál (S002) a legjobban modellezhető, alacsony hibaszinttel és stabil eloszlással, míg a cerebrovaszkuláris betegségben elhunytak (S008) a legnehezebb, jelentős hibavariabilitással.

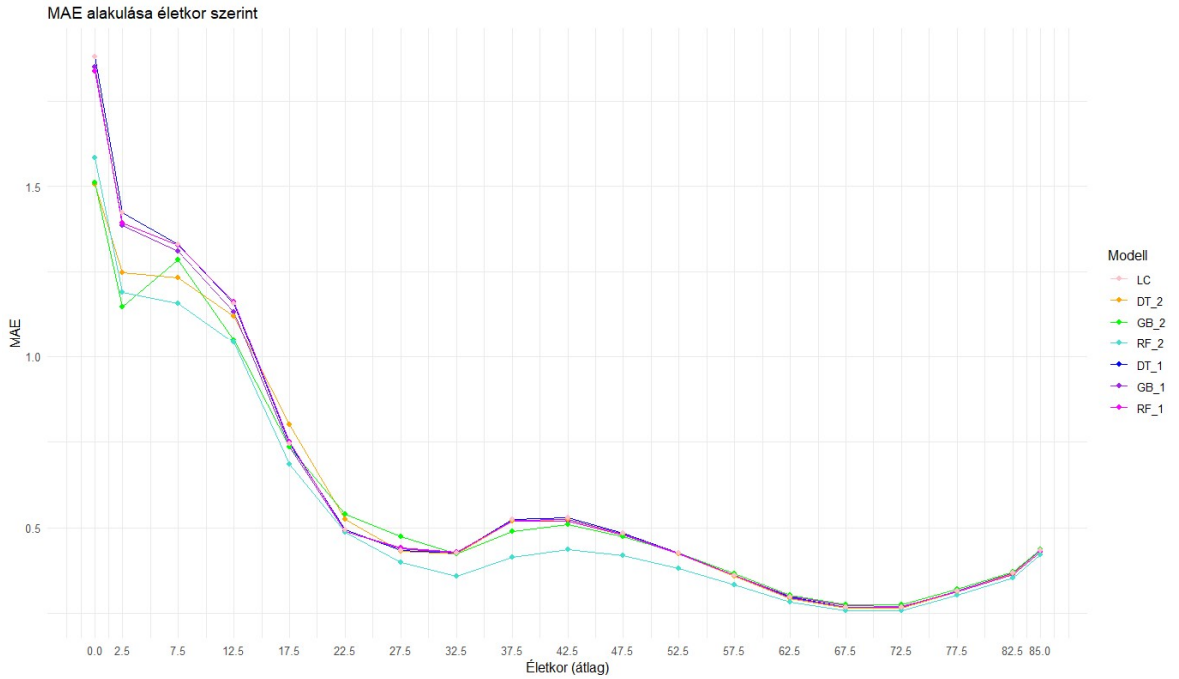
7.3 Korcsoportonkénti eredmények

A következőkben életkoronkénti csoportosítással számoltam ki az eredményeket.



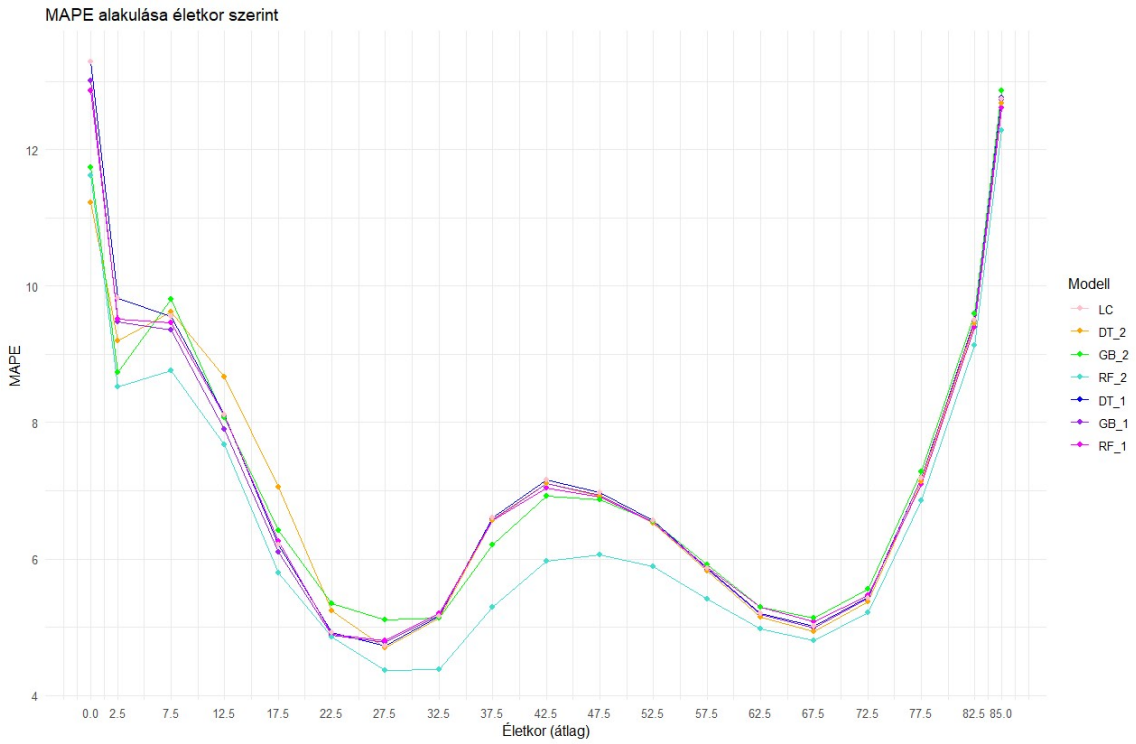
8. ábra RMSE alakulása életkor szerint

Az 8.ábra alapján megfigyelhető, hogy a legmagasabb hibák jellemzően a legfiatalabb (0-5 év) és a legidősebb (80+ év) korcsoportokban jelentkeznek. Ez a tendencia mind a klasszikus LC modell, mind pedig a gépi tanulási korrekciók esetében fennáll, ugyanakkor a különbségek hangsúlyossá válnak a középső életkorokban. A random forest alapú RF_2 modell következetesen alacsonyabb RMSE-értékeket produkál szinte minden életkori csoportban, különösen a 30-70 éves tartományban, amely a népesség többségét is lefedi. A DT_2 és GB_2 modellek szintén javulást mutatnak a LC modellhez képest, de kevésbé konzisztensen. Ezzel szemben az RF_1 modell, amely az első megközelítés szerint korrigált *mx* változót használja, a fiatalabb korcsoportokban kifejezetten rosszabbul teljesít.



9. ábra MAE alakulása életkor szerint

A MAE eredmények hasonló trendeket mutatnak az RMSE-hez képest, de kisebb mértékű szórás jellemzi az egyes modellek közötti különbségeket. A Lee-Carter modell hibája különösen magas a szélső korcsoportokban, míg a gépi tanulási korrekciók minden esetben csökkentik ezt a hibát, különösen az RF_2 modell esetében.



10. ábra MAPE alakulása életkor szerint

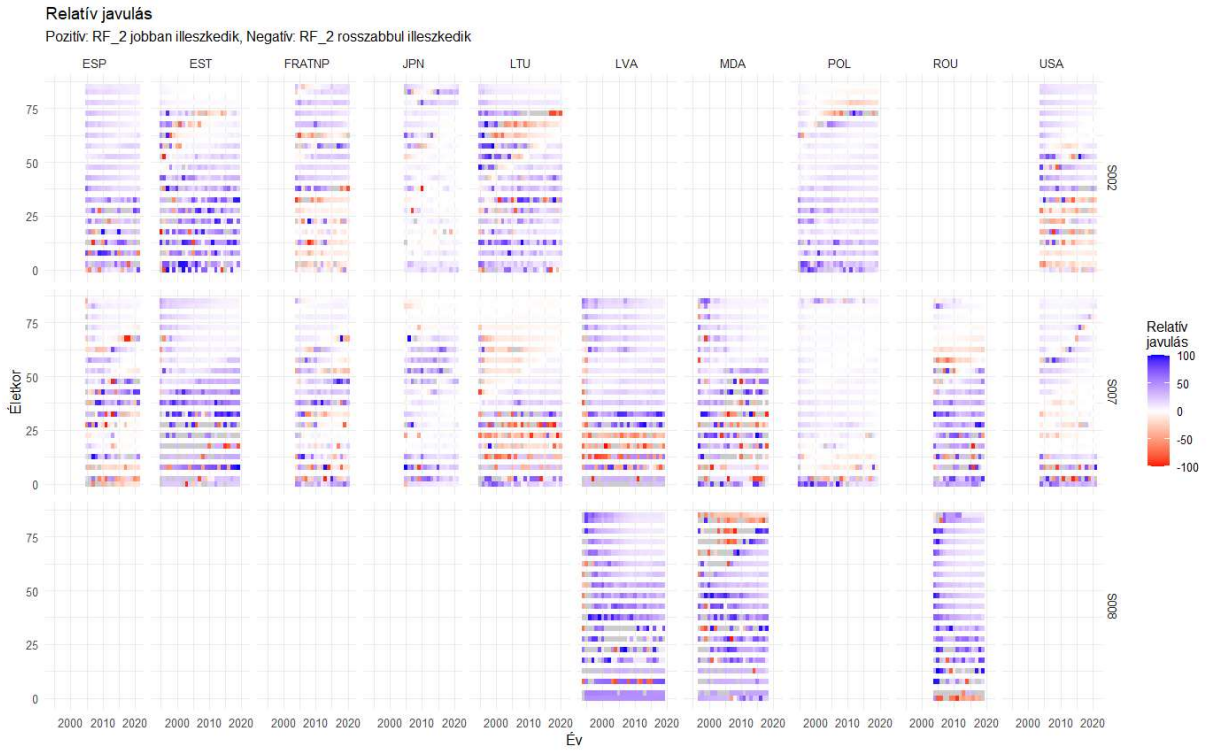
A 10. ábra alapján egyértelműen kirajzolódik, hogy a legnagyobb relatív hibák a 0-5 éves és 80 év feletti korcsoportokban jelentkeznek, ami részben a kis népességszám és az ebből fakadó nagy arányeltérések következménye. Érdekesség, hogy a LC modell torzításai ezekben a korcsoportokban extrém módon megemelkednek, ami jelentős bizonytalanságot jelez az alapmodell előrejelzéseiben. A gépi tanulási korrekciók itt is javulást eredményeznek, de nem egyforma mértékben: míg a GB_2 és DT_2 modellek csak mérsékelt csökkenést érnek el a relatív hibákban, addig az RF_2 modell érezhetően jobban kontrollálja a hibák kilengéseit.

7.4 Legjobban teljesítő modell

A következőkben legjobban teljesítő modellt fogom részletesebben bemutatni. A többi modell eredményét a melléklet tartalmazza, ahol jól láthatóak a különbségek korcsoportonként és országonként. Az alábbiakban bemutatott hőtérkép az RF_2 modell teljesítményének relatív javulását ábrázolja az LC modellhez képest, a halálozási arányok logaritmikus transzformáltjának pontossága alapján. Az elemzés célja annak feltárása, hogy az RF_2 modell mikor és hol teljesít jobban, illetve rosszabbul az LC modellhez képest különböző országok, életkori csoportok és évek vonatkozásában. A 11. ábra alapja a (61) képlettel meghatározott százalékos relatív javulás:

$$Relatív\ javulás\ (\%) = 100 * \frac{|\log(m_x^{LC}) - \log(m_x)| - \left|\log\left(mx_x^{RF_2}\right) - \log(m_x)\right|}{|\log(m_x^{LC}) - \log(m_x)|} \quad (61)$$

ahol m_x^{LC} a Lee Carter modell, $m_x^{RF\_2}$ az RF_2 által előrejelzett mortalitási ráták, m_x pedig a megfigyelt ráták. A pozitív százalékérték (kék) a gépi tanulási modell jobb illeszkedését jelenti a valósághoz, míg a negatív érték (piros) arra utal, hogy az LC modell pontosabb volt az adott korcsoport-időpont-ország hármas metszetében.



11. ábra RF_2 modell relatív javulása a Lee-Carter modellhez képest

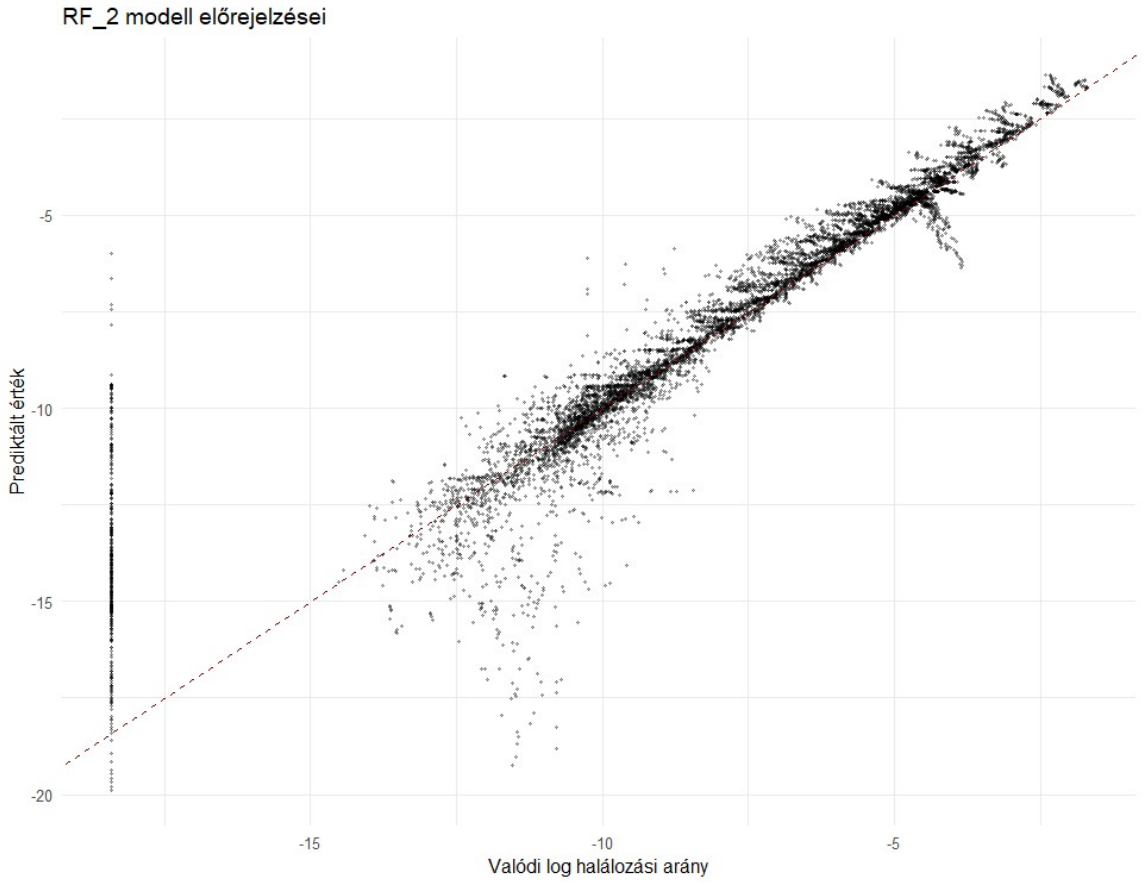
Országonként vizsgálva megállapítható, hogy Spanyolország (ESP) és Lengyelország (POL) esetében az RF_2 modell következetesen jobb predikciókat adott, különösen a 0-40 éves korosztályban, a 2005 utáni időszakban. A kék színárnyalatok itt egyértelmű teljesítményjavulásra utalnak. Egyesült Államok (USA) adatai heterogénebb képet mutatnak: míg több életkor-idő kombinációban jelentős javulás figyelhető meg, főleg az idősebb (65+) korcsoportokban, ugyanakkor számos negatív érték is megjelenik a fiatalabb csoportokban, különösen 2015 után. Japán (JPN) esetében a színskála visszafogott, és kevés erőteljes eltérés látható, ami arra utalhat, hogy az LC modell eredetileg is jól illeszkedett az adott populációhoz, vagy a gépi tanulási modell nem tudott számottevően javítani rajta. Kelet-európai országoknál a javulás itt különösen a fiatal felnőttek (20-40 év) körében hangsúlyos, viszont idősebb korcsoportokban gyakoribb a teljesítményromlás, ami összefügghet a halálozási adatok nagyobb volatilitásával és kisebb elemszámával ezekben a kohorszokban.

Életkorok mintázatát nézve a 0–20 éves tartományban sok országban javulás figyelhető meg. Ez különösen fontos eredmény, mivel a gyermek- és ifjúkori halálozási adatok gyakran torzak vagy hiányosak, így az LC modell itt hajlamosabb hibázni. Ugyanakkor érdemes megjegyezni, hogy ebben a korcsoportban a halandóság további csökkenése már erősen korlátozott, mivel a csecsemőhalandóságot számos országban már rendkívül alacsony szintre sikerült szorítani. Ezt a jelenséget a szakirodalom *rotációként* ismeri, amely szerint a halandóság javulása idővel

inkább az idősebb korosztályok felé tolódik el (Li et al., 2013). A 30-70 éves csoportban szinte minden országban jelentkezik valamilyen mértékű javulás, ami azt jelzi, hogy az RF_2 modell képes jobban megragadni a halálozási mintázatok társadalmi és időbeli trendjeit ebben az aktív korosztályban. A 70 év feletti korosztályban már szórtabb a kép, mivel egyes országokban látványos javulás (pl. USA, POL), míg máshol visszaesés tapasztalható (pl. ROU, LVA).

Érdemes lehet még időbeli trendeket is megfigyelni, ugyanis a 2000-es évek elején több országban kevesebb javulás és több negatív eltérés látható. Ez a modell tanításához használt adatok hiányosságaiból eredhet. A 2010 utáni években sokkal erősebb és kiterjedtebb kékes színfoltok jelennek meg, ami arra utal, hogy a modern gépi tanulási módszerek előnyei leginkább a közelmúltbeli halálozási minták megragadásában nyilvánulnak meg.

A 12. ábrán a vízszintes tengelyen a valódi logaritmizált halálozási arány, míg a függőleges tengelyen az RF_2 modell által becsült értékek szerepelnek. Az ideális esetben az összes pont a vörös szaggatott, 45°-os diagonálisra ($y = x$) illeszkedne, amely a tökéletes predikciót jelenti.



12. ábra RF_2 modell előrejelzései a valódi halálozási rátákhoz viszonyítva

A pontok nagy többsége szorosan a diagonális vonal mentén helyezkedik el, különösen a $\log(m_x)$ középtartományában (kb. -10 és -6 között). Ez arra utal, hogy az RF_2 modell a legtöbb esetben pontosan becsüli a halálozási arányokat. A sűrű, koncentrált pontfelhő az RF_2 modell megbízhatóságát és stabilitását támasztja alá. A bal alsó sarokban (-15 alatt) megfigyelhető egy szórtabb, függőleges mintázat, ami azt jelzi, hogy a nagyon alacsony halálozási arányok esetén a predikciók bizonytalanabbak. Ez a tendencia összhangban van a korábbi relatív javulást bemutató hőtérképpel, ahol a 0-5 éves korosztályban (amely tipikusan alacsony halandóságú) az RF_2 modell több országban is gyengébben teljesített. Ez a prediktív bizonytalanság vélhetően az adatok ritkaságával magyarázható. A felsőbb $\log(m_x)$ értékeknél (pl. -6 és afelett) a pontok többsége kissé a diagonális alatt helyezkedik el. Ebből az következik, hogy az RF_2 modell enyhén alábecsül a magas halálozási arányok esetén, például idős korcsoportokban, ahol a halandóság jellemzően nagyobb. Ezt a megállapítás szintén támogatja a hőtérkép, ahol az időskorúaknál több országban vegyes eredményeket láttunk, helyenként romló predikciós teljesítménnyel. A $log\ (m_x)$ -10 és -6 közötti értékeihez tartozó pontok eloszlása különösen szoros illeszkedést mutat, ami azt jelenti, hogy az RF_2 modell ebben a tartományban, azaz a középkorú populáció körében rendkívül megbízható. Ez a megfigyelés ismét alátámasztja az életkor szerinti elemzések eredményeit, ahol szintén ebben az életkori sávban volt a legkisebb hiba.

8. Összefoglalás

A szakdolgozat célja az volt, hogy a klasszikus halandósági előrejelzéshez leggyakrabban használt benchmark Lee-Carter modellt összehasonlítsa korszerű gépi tanulási algoritmusokkal. A vizsgálat során különös figyelmet fordítottam arra, hogy a modelleket többdimenziós bontásban értékeljem, így a predikciós teljesítményt életkoronként, országonként és halálokok szerint is részletesen elemeztem. A predikciók értékeléséhez három hibamértéket (RMSE, MAE és MAPE) használtam, amelyek együttesen összetett képet adtak az algoritmusok pontosságáról és stabilitásáról.

A bemutatott eredmények alapján egyértelműen megállapítható, hogy a gépi tanulási módszerek, különösen a második megközelítés szerint alkalmazott véletlen erdő (RF_2) jelentős javulást eredményeztek a predikciós hibák csökkentésében. Ez a javulás nemcsak aggregált szinten jelentkezett, hanem életkoronkénti és országonkénti bontásban is konzisztensnek bizonyult. A RF_2 modell alacsonyabb medián hibákat, kisebb szórást és

kiegyensúlyozottabb teljesítményt mutatott a legtöbb esetben, ami megerősíti alkalmazhatóságát a halálozási ráta előrejelzésekben.

A halálok szerinti vizsgálatokból az derült ki, hogy az S002 (daganatos betegségek) kategória a legjobban modellezhető, alacsony hibaszintekkel és robusztus predikcióval. Ezzel szemben az S008 kategória (Cerebrovaszkuláris betegség) jelentős predikciós kihívást jelentett, különösen az egyszeres korrekciót alkalmazó gépi tanulási modellek számára. Az S007 kategóriában (szívbetegségek) a modellek teljesítménye közepesnek mondható: itt a gépi tanulás átlagosan jobb illeszkedést biztosított, de a javulás mértéke mérsékeltebb volt.

Az MAE és RMSE együttes elemzése arra is rávilágított, hogy míg a Lee-Carter modell hibái gyakran a szélsőséges esetekből adódnak, a gépi tanulási eljárások nemcsak ezeket csökkentik, hanem az átlagos értéktartományon vett predikciós eltérés mértékét is számottevően mérséklik.

Összességében a szakdolgozat eredményei azt mutatják, hogy a klasszikus demográfiai modellek és a modern gépi tanulási eljárások kombinációja hatékonyan alkalmazható a halandósági mintázatok előrejelzésében. A Lee-Carter modell által szolgáltatott szerkezeti alapot kiegészítve, a gépi tanulás képes korrigálni a rendszeres torzításokat és megragadni azokat a nemlineáris mintázatokat, amelyek a klasszikus megközelítések számára nehezen kezelhetők. E megközelítés hozzájárulhat a demográfiai előrejelzések pontosságának javításához, és megalapozhatja a jövőbeni kutatásokat, amelyek még komplexebb modellek, például mély tanulás irányába mutathatnak.

9. Irodalomjegyzék

Alho, J.M., 2000. Discussion of Lee (2000). *North American Actuarial Journal* 4, 91–93. https://doi.org/10.1080/10920277.2000.10595883

Banyár, J. (2003) *Életbiztosítás*. Budapesti Corvinus Egyetem.

Basellini, U., Camarda, C.G. & Booth, H. (2023). Thirty years on: A review of the Lee–Carter method for forecasting mortality. *International Journal of Forecasting*. 39, 1033-1049. https://doi.org/10.1016/j.ijforecast.2022.11.002

Bjerre, D. S. (2022). Tree-based machine learning methods for modeling and forecasting mortality. *ASTIN Bulletin: The Journal of the IAA* , 52(3), 765-787.

Booth, H., Maindonald, J., & Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, 56, 325-336. https://www.jstor.org/stable/3092985

Breiman, L. (2001). Random Forests. *Machine Learning,* 45, 5-32. https://doi.org/10.1023/A:1010933404324

Brillinger, D.R., (1986). The natural variability of vital rates and associated statistics. *Biometrics* 42, 693–734. https://www.stat.berkeley.edu/pub/users/brill/Papers/biometrics.pdf

Brouhns, N., Denuit, M. & Vermunt, K.J. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics* 31, 373–393

Cryer, J.D., Chan, K. (2008). *Time Series Analysis*. Springer Texts in Statistics.

Czado, C., Delwarde, A., & Denuit, M. (005). Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics & Economics*, 36(3), 260-284.

de Jong, P., & Tickle, L. (2006). Extending Lee-Carter mortality forecasting. *Mathematical Population Studies*, 13, 1-18.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist*, 29(5), 1189-1232. 10.1214/aos/1013203451

Girosi, F., & King, G. (2008). *Demographic forecasting*. Princetion University Press.

Girosi, F., King, G., 2007. *Understanding the Lee-Carter Mortality Forecasting Method.* https://gking.harvard.edu/files/lc.pdf

Groot, R. (2011). *Maximum-Likelihood estimation of the Lee-Carter model* [Szakdolgozat, Risksuniversiteit Groningen]. https://fse.studenttheses.ub.rug.nl/11118/1/Thesis.pdf

Hastie, T., Tibshirani,. R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Texts in Statistics.

Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87(419), 659–671. https://doi.org/10.2307/2290201

Lee, R. D., & Miller, T. (2001). Evaluating the Performance of the Lee-Carter Approach to Modeling and Forecasting Mortality. *Demography*, 38(4), 537– 549. https://doi.org/10.1353/dem.2001.0036

Levantesi, S. & Pizzorusso, V. (2019). Application of Machine Learning to Mortality Modeling and Forecasting. *Risks, 7*(1), 26. **https://doi.org/10.3390/risks7010026**

Li, N., Lee, R.D., & Gerland P. (2013). Extending the Lee-Carter method to model the rotation of age patterns of mortality-decline for long-term projection. *Demography*, 50, 2037-2051.

Rabbi, A. M. F., & Mazzuco, S. (2021). Mortality forecasting with the Lee-Carter method: Adjusting for smoothing and lifespan disparity. *European Journal of Population*, 37, 97-120. https://link.springer.com/article/10.1007/s10680-020-09559-9

Renshaw, A., & Haberman, S. (2003). On the forecasting of mortality reduction factors. *Insurance: Mathematics & Economics*, 32, 379-401.

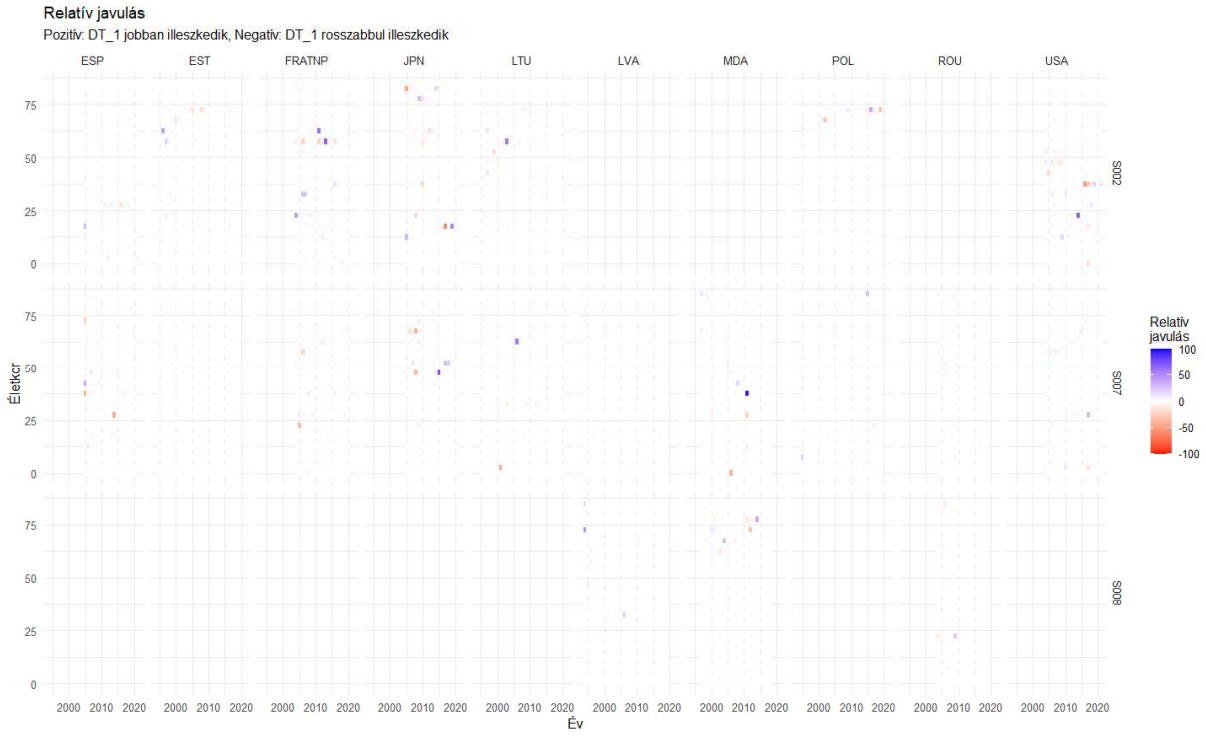
Stoeldraijer, L., van Duin, C., van Wissen, L., & Janssen, F. (2018). Comparing strategies for matching mortality forecasts tot he most recently observed data: exploring the trade-off between accuracy and robustness. *Henus*, 74(1), 1-20. https://genus.springeropen.com/articles/10.1186/s41118-018-0041-y

Tuljapurkar, S., Li, N., & Boe, C. (2000). A universal pattern of mortality decline in the G7 countries. *Nature*, 405(6788), 789-792.

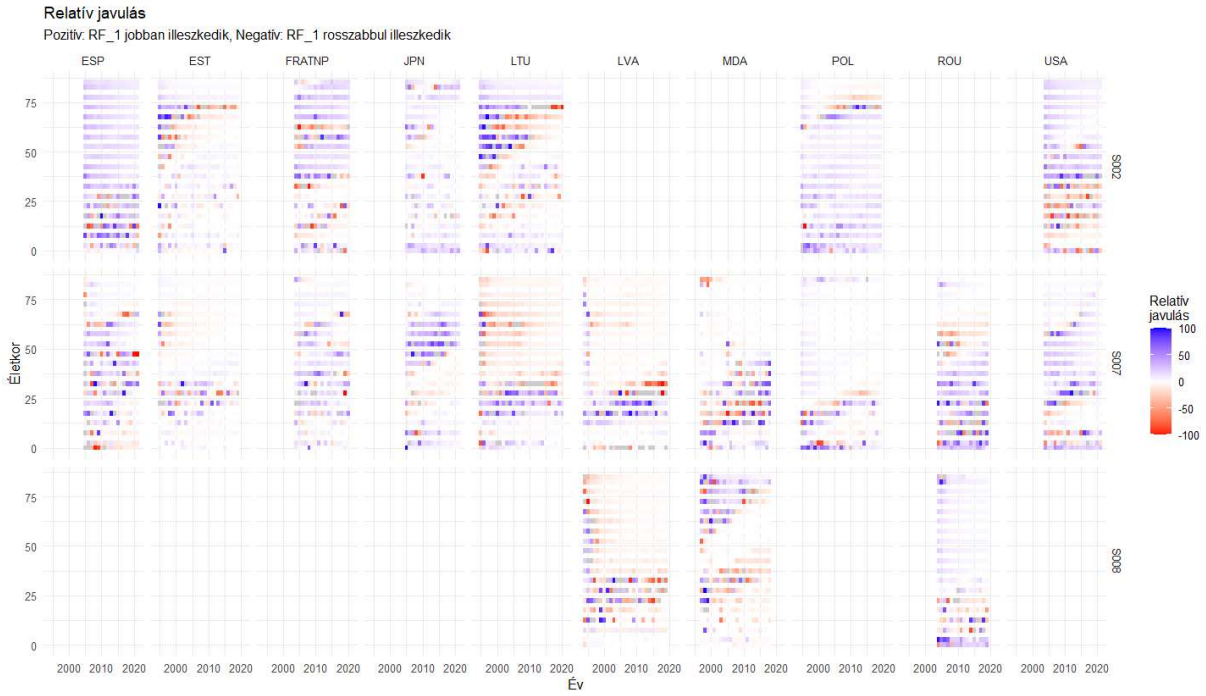
Vékás, P. (2016). *Az élettartam-kockázat modellezése* [Phd, Budapesti Corvinus Egyetem]. https://unipub.lib.uni-corvinus.hu/2398/1/longevity.pdf

Wilmoth, J.R. (1993). *Computational methods for fitting and extrapolating the Lee-Carter model of mortality change*. Technical report, Department of Demography, University of Californi, Berkeley.

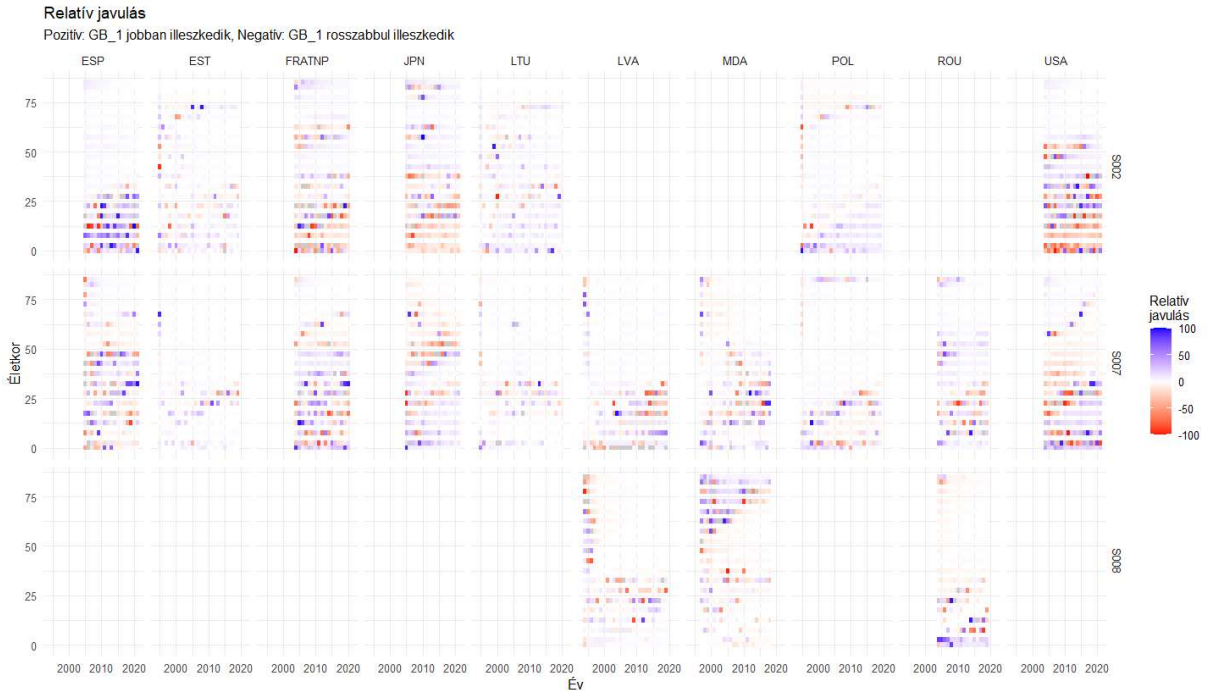
10. Mellékletek



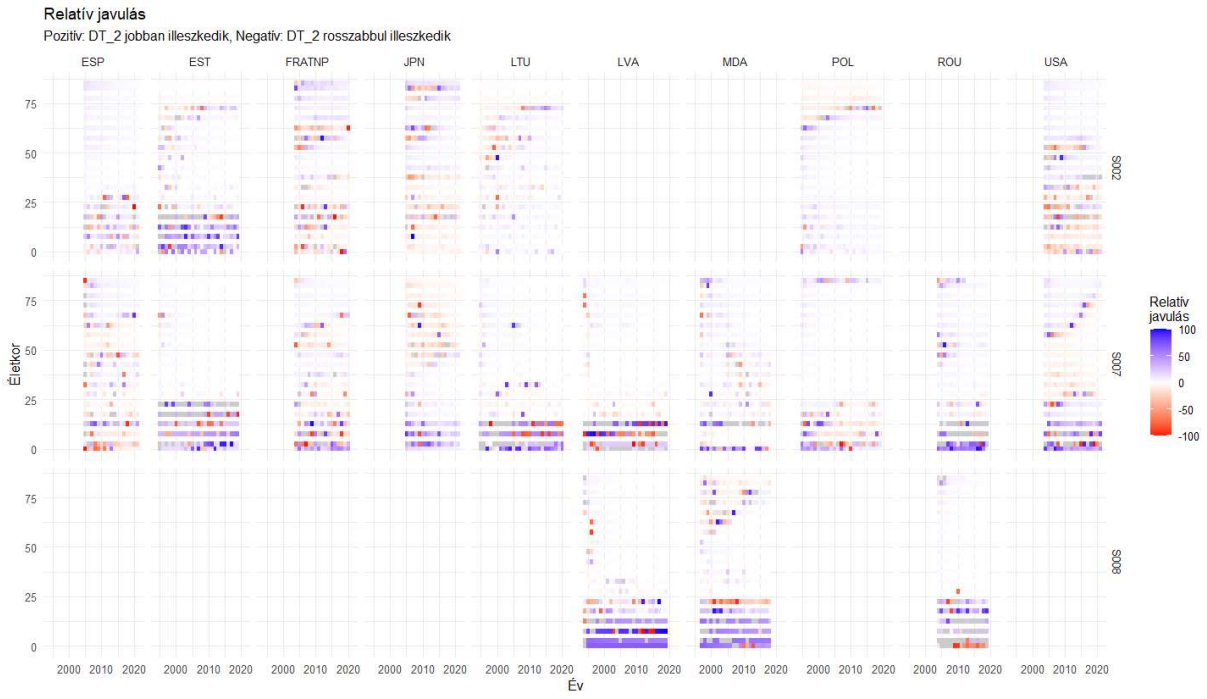
13. ábra Melléklet - DT_1 hőtérkép



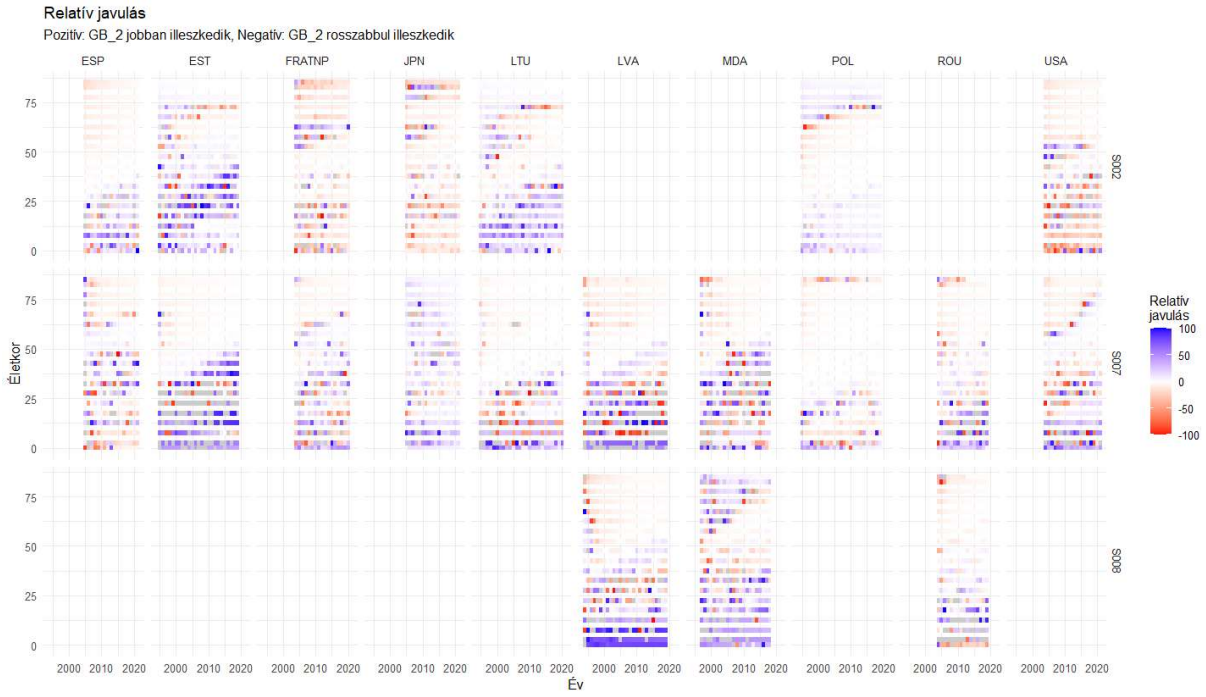
14. ábra Melléklet - RF_1 hőtérkép



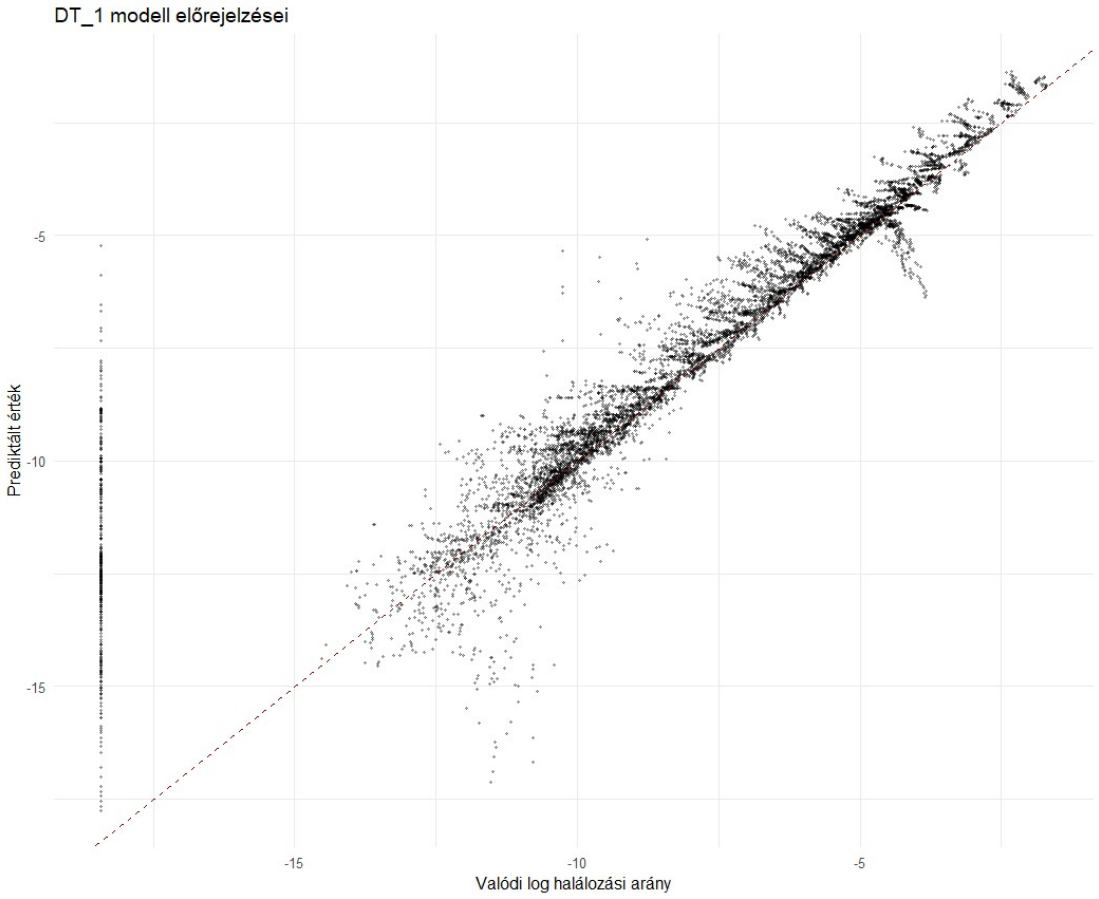
15. ábra Melléklet - GB_1 hőtérkép



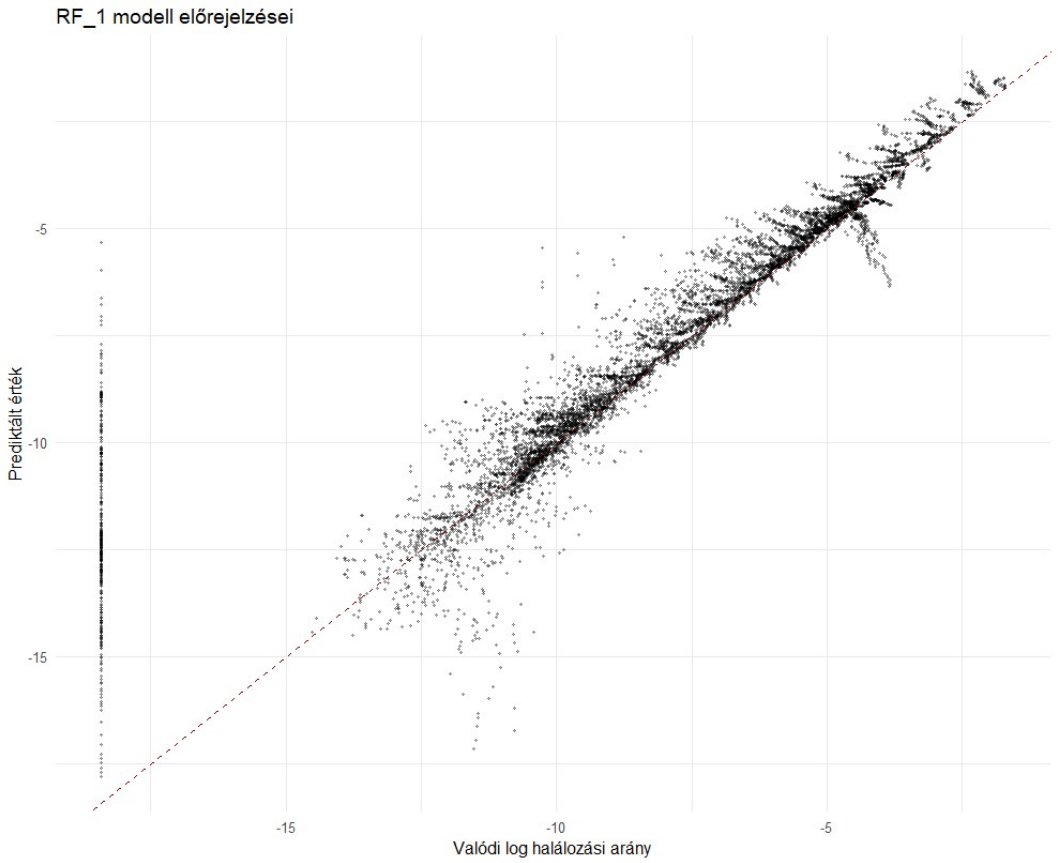
16. ábra Melléklet - DT_2 hőtérkép



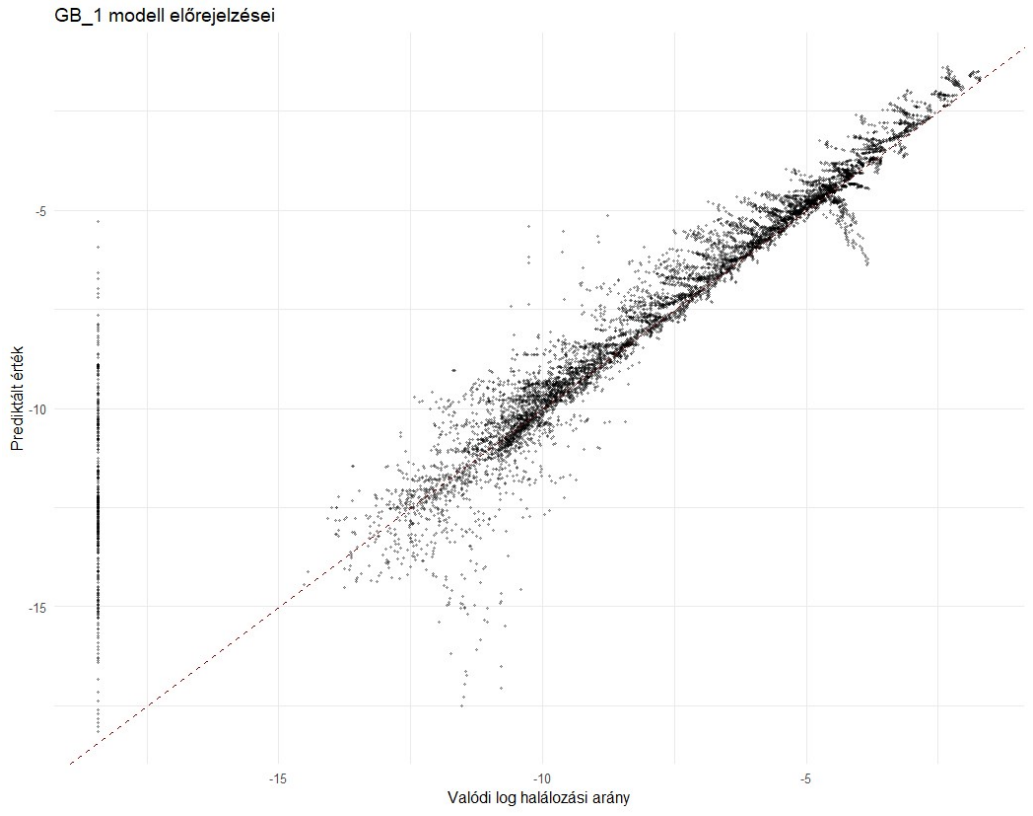
17. ábra Melléklet - GB_2 hőtérkép



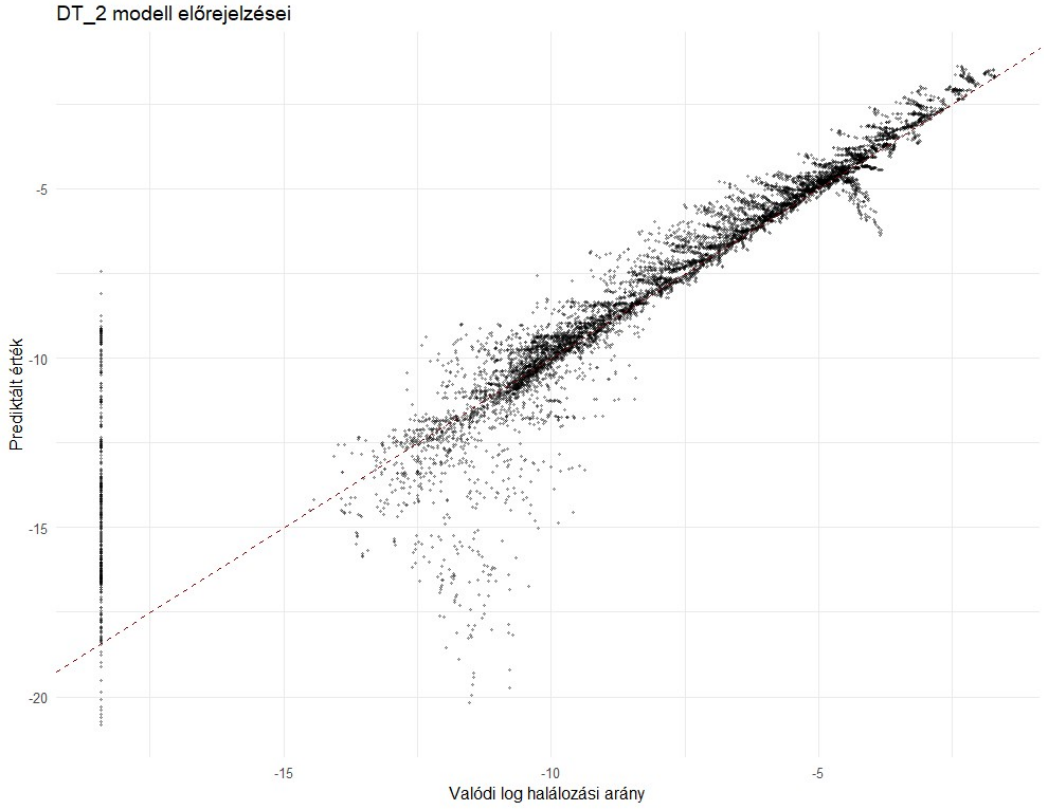
18. ábra Melléklet - DT_1 pontdiagram



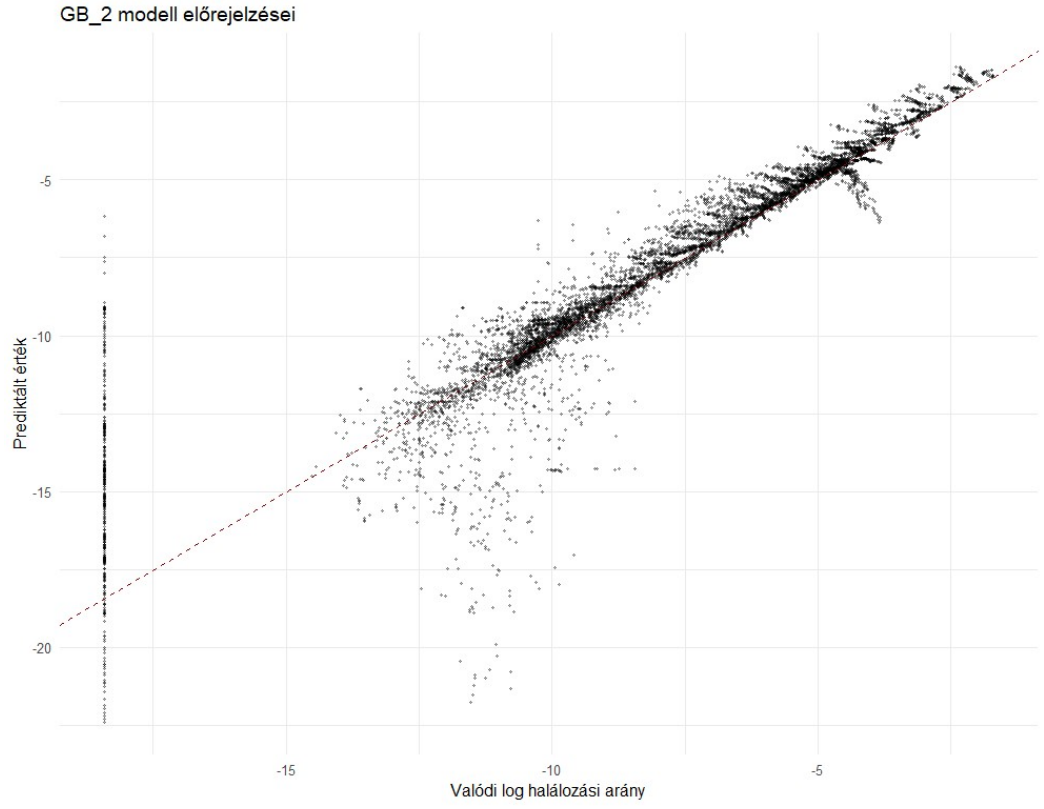
19. ábra Melléklet - RF_1 pontdiagram



20. ábra Melléklet - GB_1 pontdiagram



21. ábra Melléklet - DT_2 pontdiagram



22. ábra Melléklet - GB_2 pontdiagram

MI-eszközhasználati nyilatkozat

Alulírott, **Sádt Dominik** nyilatkozom, hogy szakdolgozatom elkészítése során az alább felsorolt feladatok elvégzésére a megadott MI-alapú eszközöket alkalmaztam:

Feladat	**Felhasznált eszköz**	**Felhasználás helye**	**Megjegyzés**
Nyelvhelyesség javítása	ChatGPT	Teljes dolgozat	Helyesírás, billentyűzet elütés vizsgálata

A felsoroltakon túl más MI-alapú eszközt nem használtam.