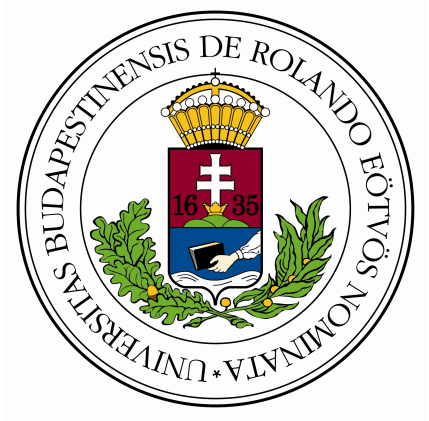
Budapesti Corvinus Egyetem
Eötvös Loránd Tudományegyetem



Adatalapú kockázatelbírálási módszerek az életbiztosításban

MSc szakdolgozat

Huszárik Ádám
Biztosítási és pénzügyi matematika MSc
Aktuárius specializáció
2020

Témavezető: Dr. Kovács Erzsébet

Köszönetnyilvánítás

Ezúton szeretném kifejezni hálámat mindenkinek, aki valamilyen módon hozzájárult a dolgozat elkészítéséhez. Külön szeretném kiemelni Dr. Kovács Erzsébetet, aki hasznos észrevételeivel és szakértelmével segített a dolgozatírás során. A konzultációs lehetőségek és a szakemberekkel egyeztetett interjúk nélkül a dolgozat nem készülhetett volna el jelen formájában.

Tartalomjegyzék

Táblázatok jegyzéke

Ábrák jegyzéke

1. Bevezetés

A kockázatelbírálás központi szerepet játszik egy biztosító életében. Minél több adat áll rendelkezésre egy adott ügyfélről, annál pontosabban lehet felmérni, hogy mennyi kockázatot jelent. Azonban nem feltétlen az jelenti az optimális megoldást, hogy a biztosító minél több információt megkérdez az ügyféltől és az alapján áraz. Ugyanis számolni kell azzal is, hogy egy túlbonyolított folyamat esetén mind az ügyfél, mind a biztosító munkatársainak csökken az elégedettsége, ami kihatással lehet a jövőbeli értékesítésekre. Ezért fontos megtalálni az egyensúlyt, hogy kellő mennyiségű információ álljon rendelkezésre, ugyanakkor ennek megszerzése ne okozzon jelentős többletterhet. Ebben segíthetnek az utóbbi évtizedben elterjedt telematikus eszközök, melyek rengeteg adatot gyűjtenek, és így lehetővé válik a használatot figyelembe vevő árazás. A korábbi – osztályozáson alapuló - kockázat becslés helyett a viselkedés, a tényleges igénybevétel lehet az árazás alapja, ami sokkal tagoltabb ügyfélszegmentációt és dinamikus árazást tesz lehetővé. Ugyanakkor azt is mérlegelni kell, hogy a kockázatközösség túlzott tagolása nem okoz-e a biztosítótársaság számára kockázatnövekedést.

Magyarországi életbiztosítások esetében a jelenlegi árazás egészséges populációból indul ki, amire pótlékot számolnak az ügyféltől kapott információk alapján. A fő szempontok a kor, a káros szenvedélyek jelenléte és az egészségi állapot, amelyet jellemzően egy kérdőív vagy orvosi vizsgálat formájában tud meg a biztosító. Fontos vizsgálni, hogy a jelenlegi folyamatot lehet-e javítani azáltal, hogy csak a legnagyobb hatással bíró és a lehető legmegbízhatóbb adatokat használjuk fel az árazáshoz. Ezen adatok megszerezhetősége is létfontosságú kérdés. Továbbá maga az adatszerzés folyamata is fontos tényező, mert ezzel akár motiválhatjuk is az ügyfelet az egészségesebb életmódra, ami hosszabb távon a biztosítónak is kifizetődő. Már több országban is lehet találkozni hasonló konstrukcióval, ami jutalmazza a biztosítottakat, ha szűréseken vesznek részt, egészségesen táplálkoznak és aktív, sportos életet élnek. Az egyéni adatok, köztük a lépésszámlálás, pulzus, vérnyomás, stb. folyamatos gyűjtése és monitorozása hatékonyabb kockázatelbírálást alapoz meg. Így a mérőeszközök használata a költséges kezdeti/egyszeri orvosi vizsgálat helyett/mellett folyamatos kontrollt biztosít.

A téma elméleti feldolgozása mellett a Survey of Health, Ageing and Retirement in Europe (SHARE) adatbázisban található adatokat használom, ami lehetővé teszi az 50 év feletti európai polgárok adatainak az elemzését. Az egyre növekvő várható élettartam miatt, egyre több ember szeretne idősebb korában életbiztosítást kötni, amit a biztosítók is egyre szívesebben fogadnak a javuló demográfiai tendenciák ismeretében. Ebből következik, hogy érdemes foglalkozni ennek

a korosztálynak a kockázataival, hiszen más szempontok kerülhetnek előtérbe, amiket az európai piacon egyelőre nem vesznek figyelembe. Ráadásul a szokásos életbiztosítási termékek mellett (kockázati, vegyes, befektetéshez kötött), ahol végeznek kockázatelbírálást, az időskori járadékok meghatározása szempontjából is indokolt foglalkozni a témával, hiszen az életpálya során felhalmozott vagyon kifizetési üteme nagy mértékben függhet az egyén egészségi állapotától.

A SHARE adatfelvételi hullámok során jellemzően ugyanazokon az embereken végzik el a vizsgálatokat, amiből megtudhatjuk, hogy kik haltak meg két tetszőleges hullám között. 2011-ben és 2017-ben magyar adatok is rögzítésre kerültek, illetve további európai országokon is elvégezték az adatfelvételt ebben a két hullámban, amik alapján különböző osztályozó eljárásokkal modellezhetjük, hogy mik lehetnek a halál bekövetkeztének befolyásoló tényezői.

A dolgozat az alábbi módon épül fel:

- A 2. fejezetben ismertetem az adatalapú kockázatelbíráláshoz köthető cikkeket és azok eredményeit. Majd áttekintem az időskori kockázatelbírálás során felmerülő kérdéseket, és a piacon megjelenő legújabb kockázatelbírálási technikákat. Végül az adatszerzés más lehetséges formáira térek ki.
- A 3. fejezet során röviden bemutatom a használt modellek elméletét, illetve az összehasonlíthatóságukra szolgáló lehetséges mutatókat.
- A 4. fejezetben bemutatom az elemzéshez használt adatbázist, és az adat-előkészítés lépéseit. Továbbá feltáró elemzés segítségével vizsgálom a célváltozó és a magyarázó változók kapcsolatát.
- Az 5. fejezetben ismertetem a modellezés lépéseit, majd a változószelekciós eljárásokkal vizsgálom az elemzés szempontjából fontos változókat. Ezután a 3. fejezetben ismertetett algoritmusokat felhasználva készítem el a modelljeimet.
- Végezetül a 6. fejezetben foglalom össze a dolgozatból levonható következtetéseket.

2. Adatokon alapuló kockázatelbírálás

Életbiztosítások esetén különösen fontos szerepet játszik a kockázatelbírálás. Ugyanis ha elbírálás nélkül csak a legjobbnak tűnő ügyfeleket fogadná a biztosító, akkor nagyban csökkenne a profitabilitása, mert rengeteg ügyfelet utasítana el. Éppen ezért az a lényege, hogy az ügyfelek kockázatát alaposan felmérjék, és különböző kockázati csoportokba sorolják be őket, ami alapján meg lehet határozni a megfelelő díjat. Valamint a biztosítási piacot érintő antiszelekciós hatás ellen is védelmet nyújt, ami szerint azok kötnek biztosítást, akiknél nagyobb eséllyel következik be a biztosítási esemény. Jellemzően ez egy költséges és időigényes folyamat eredménye, ugyanis a biztosítottnak rengeteg adatot kell magáról szolgáltatnia, amit többnyire zaklatásként élnek meg, míg a biztosítónak szakembereket kell alkalmaznia, akik a tapasztalataik alapján megfelelően fel tudják mérni a kockázatot. A szokásos gyakorlat szerint (Interjú Dr. Halász Katalinnal) a biztosítónak dolgozó orvos az ügyfél vizsgálati eredményeit feltölti egy nagy viszontbiztosító (Swiss Re, Munich Re) kockázatelbíráló rendszerébe, ami eredményül megadhatja például, hogy hány százalékos díjemelést célszerű alkalmazni, vagy hogy további orvosi véleményre van-e szükség. A rendszerek előnye, hogy sokan használják, aminek köszönhetően megbízható támpontot adnak az orvosoknak, de természetesen még egyéb negatív/pozitív hatású szempontokat is figyelembe vehetnek saját belátásuk szerint. A költségeket tovább növeli, hogy bizonyos biztosítási összeg, tartam vagy kor felett további orvosi vizsgálatokat követel meg a biztosító, amiknek helyes értelmezése szintén a kockázatelbíráló szakembereket terheli. A kérdések csökkentésével vagy automatizálásával enyhíteni lehetne a terheket, amivel több pénzt spórolhat a biztosító, mint amennyit veszít a kissé pontatlanabb kockázatelbírálás miatt.

A MuleSoft által végzett kutatás (Seekings [2018]) is alátámasztja az ügyfelek elégedetlenségét a jelenlegi rendszerrel szemben, így nem meglepő, hogy mind a két fél részéről igény van arra, hogy leegyszerűsítsék, automatizálják ezt a folyamatot. Ennek jelentőségére hívja fel a figyelmet Sallai [2019] is, vagyis hogy túl bonyolultak és nem kellőképpen személyre szabottak a szolgáltatások, aminek következtében más megoldásokat választhatnak az ügyfelek. A Liferay és az International Data Corporation által készített tanulmány (Liferay & IDC [2020]) is megerősíti, hogy a biztosítók tisztában vannak a fogyasztók igényeivel, és ennek fényében egyre inkább ügyfélközpontúvá válnak. Ennek alapja azonban a különböző rendszerek összehangolása és automatizálása, amihez az adatalapú elemzési módszerek integrálására van szükség.

2.1. Prediktív modellek alkalmazása

A prediktív módszerek alkalmazása gyorsabbá, hatékonyabbá, konzisztensebbé teheti a kockázatelbírálást, feltéve hogy rendelkezésre állnak a megfelelő adatok. A Deloitte munkatársai (Batty et al. [2010]) segítségével fejlesztett modellek aránylag nagy pontossággal fel tudták mérni az ügyfelek mortalitási mutatóit, amiket felhasználva jelentős mértékben lehet csökkenteni a kockázatelbírálókra háruló terhet. Viszont a gyakorlatba való átültetés előtt sok szempontot figyelembe kell venni, melyekre a következő alfejezetben térek ki.

2.1.1. Megbízhatóság

A hagyományos aktuáriusi kommutációs számok helyett, a növekvő adatmennyiség és a gépi tanulás jelentheti a jövőt a biztosítási piacon történő árazás során. Fontos azonban látni, hogy az adatok forrása, megbízhatósága és megszerzésének költsége is létfontosságú kérdés. Az ember neme éppen ezért volt nagyon hasznos választóvonal, amíg a gender direktíva el nem törölte használatának lehetőségét az Európai Unióban 2012-ben. Ennél könnyebben megszerezhető és értékesebb információ nemigen áll a biztosító rendelkezésére, hiszen az önbevallásos alapon közölt adatokban nem lehet teljes mértékben megbízni, a hivatalos adatokhoz pedig csak költségek árán juthat a biztosító. Olykor még a hatósági iratokban sem lehet maradéktalanul megbízni, hiszen előfordulhat, hogy valaki nem a lakcímkártya szerinti címen él.

A halandósági adatokra vonatkozó legpontosabb képet maga a biztosító állománya adja. Kockázati és vegyes életbiztosítások esetén a biztosító célja, hogy minél kedvezőbb mortalitási mutatókkal rendelkezzen az állománya, amit egyrészt a kockázatelbírálással tud módosítani, másrészt viszont akaratán kívül is eléri ezt a célját a megtakarítási jellegű biztosításoknál, hiszen jellemzően a jobban szituált emberek engedhetnek meg maguknak ilyen biztosítást, akik alapból is kedvezőbb életkilátással rendelkeznek. Ezenfelül további szűrőként funkcionál a várakozási idő is, ami szerint a biztosító kockázatviselésének kezdete nem azonnali. Azonban az állomány mérete sokszor nem elegendő, hiszen életbiztosítások esetében jóval kevesebb biztosítási esemény történik, mint például gépjármű biztosítások esetében, ami megnehezíti a mortalitás becslését. Erre láthatunk viszont egy példát Gschlössl et al. [2011] cikkében egy német biztosító adatain, ahol Poisson regresszió segítségével becsülik meg az egyes korokhoz tartozó mortalitási rátákat a tartam, biztosítási összeg és kockázatelbírálás függvényében. Ennek ellenére azonban, érdemes fontolóra venni a nyilvánosan hozzáférhető adatbázisok használatának lehetőségét, melyet a biztosító korrigálhat saját adatai alapján.

Minél több adat áll rendelkezésre egy adott csoportról, annál pontosabban lehet felmérni a kockázatukat és előrejelezni halandóságukat különböző módszerekkel, aminek köszönhetően olyanok is köthetnek biztosítást, akiknek korábban esélye sem volt arra. Ilyen például a kanadai Manulife (Matt [2016]) kezdeményezése, melynek köszönhetően már HIV-fertőzött emberek is biztosításhoz juthatnak, amennyiben a modell kedvező életkilátásokat jósol.

2.1.2. Alkalmazott modellek

A kockázatelbírálási folyamat gyorsításához elengedhetetlen a megfelelő modell kiválasztása, így nem meglepő módon a Kaggle oldalon is található ezzel kapcsolatos verseny, melyet a Prudential életbiztosító hirdetett meg a saját anonim és sztenderdizált adataival. Azonban meg kell említeni hátrányként, hogy az adatállományhoz nem adták meg a változók pontos leírását, csak hogy milyen jellegű változóról van szó (például családdal, egészséggel kapcsolatos). Ebből következik, hogy nem lehet konkrét következtetéseket levonni a lényeges változókra vonatkozóan, csak a modellek teljesítményét lehet összehasonlítani. Boodhun és Jayabalan [2018] és Govindarajula [2019] is ezt az adatállományt tanulmányozták, ami közel 60 ezer megfigyelésből és 126 magyarázó változóból állt. Lineáris regresszió, logisztikus regresszió, döntési fa és neurális háló teljesítményét vetették össze (AUC és MSE segítségével), aminek a célja, hogy kategorizálják az ügyfeleket különböző kockázati csoportokba, amit a magyarázó változók szelektálásával, és sűrítésével érnek el. Boodhun és Jayabalan cikkében a szelektálással kapott magyarázó változók pontosabb előrejelzéshez vezettek, mint az új komponensekkel meghatározott modellek. Govindarajula viszont főkomponens elemzéssel redukálja a változók számát, tekintve hogy a konkrét változók hiányában amúgy sem lehetne az eredményekből messzemenő következtetéseket levonni. Az előbbi esetében a döntési fa bizonyult a legmegbízhatóbb algoritmusnak, míg az utóbbinál a neurális hálók vezettek a legpontosabb becsléshez. A Kaggle oldalán lévő legjobb megoldások között azonban lineáris regresszióval is találkozhatunk, ami azt vetíti előre, hogy érdemes minél több módszert kipróbálni a modellezés során.

Biddle et al. [2018] egy ausztráliai életbiztosító adatain vizsgálták az automatizálás lehetőségét, viszont a kockázati csoportokba sorolás helyett a kizárások detektálása és a kérdőív optimalizálása volt a cél. Logisztikus regresszió és *Gradient Boosting* módszerekkel tettek kísérletet a leggyakoribb kizáró okok meghatározására, melyeket *Recursive Feature Elimination* és *Lasso* változószelekciós eljárásokkal választottak ki. Az előbbi módszer a *Backward* algoritmusra hasonlít, azonban itt a lépésenként elhagyott és a végső modellben szereplő változók

számát határozzuk meg. *Lasso* eljárással pedig a béta esztimátorok értéket csökkentjük egy büntetőparaméter segítségével, aminek köszönhetően a nem lényeges változók súlya nullához fog közelíteni. A szerzők a faktoranalízis lehetőségét elvetik, hiszen akkor nem lehetne az egyedi hatását vizsgálni az egyes változóknak. Az eredményekből arra lehetett következtetni, hogy a *Gradient Boosting* módszer csaknem negyed annyi változóból tudott hasonlóan jó becslést adni az AUC értékek alapján, ami a gyakorlati szempontokat figyelembe véve sokkal kedvezőbbnek tűnik, hiszen ebben az esetben sokkal több olyan kérdést el lehet hagyni a kockázatfelmérésből, amik végső soron semelyik kizáró ok meglétéhez sem járulnak hozzá.

Yi Tan és Guo-Ji Zhang [2005] egy kínai életbiztosító adatait használva kategorizálják be az ügyfeleket három kockázati csoportba:

- alacsony, akiknek jelentkezését elfogadja a biztosító;
- közepes, akiknek jelentkezését csak további vizsgálatokkal fogadják el;
- magas, akik elutasításra kerülnek.

A szerzők azt is figyelembe veszik, hogy az egyes ügyfelekről nem feltétlenül azonos információ áll rendelkezésre, ugyanis a feltett kérdések köre függ az életkortól, biztosítási összegtől és tartamtól. Ezért a hiányzó adatokat szimuláció segítségével pótolják, majd SVM használatával becsülik meg, hogy egy adott ügyfél magas kockázatú-e vagy sem. Mivel élet- és egészségbiztosítási adatok is rendelkezésre álltak, ezért azt tekintették magas kockázatú ügyfélnek, aki meghalt vagy több, mint háromszor igényelt szolgáltatást. Az SVM finomhangolása után közel 90%-os találati arányt tudtak elérni mind a tanuló, mind pedig a tesztelő adatokon, azonban arra nem térnek ki a szerzők, hogy ez milyen szerkezetben áll elő. Ugyanis hiába teljesít látszólag jól a modell, ha például több magas kockázatút sorol be rosszul, mint jól. Hiszen éppen ezekre az esetekre kíváncsi a biztosító.

Arora és Vij [2012] szerint a neurális hálók jelenthetik a megoldást az automatizálásra, kiegészítve az elmosódott halmazok logikájával (*fuzzy logic*), ugyanis a való élet is tele van bizonytalansággal. Ez a kombinált rendszer levehetné a kockázatelbírálókról a terhet, akiknek így több idejük lehetne a bonyolultabb ügyek kivizsgálására. Ehhez hasonlóan Nikolopoulos és Duvendack [1994] is egy hibrid megoldást javasolnak, ami a genetikus algoritmus technikáját ötvözi a *GINESYS* rendszer által létrehozott klasszifikációs szabályokkal. Az egyesített rendszer megbízhatóbb előrejelzést adott arra vonatkozóan, hogy melyik ajánlatokat kell elutasítani. Ezekből is látszik, hogy kombinált megoldások lehetőségét sem szabad kizárni a lehetséges módszerek halmazából.

Joram et al. [2017] empirikus számítások helyett egy tudásalapú rendszert dolgoztak ki, amelyben meghatározták a jelentős magyarázó változókat, és egy szabályrendszert. Az általuk készített program meghatározza a bemeneti paraméterek függvényében a kockázatot, ami hatékonyabbá teheti a folyamatot. Azonban a szerzők csupán 5 ordinális skálán értelmezett szempont figyelembevételével írták meg a programjukat, ami jelentősen csökkenti az egyedi esetek megkülönböztethetőségének lehetőségét.

Ahogy látható már sokan foglalkoztak a kockázatfelmérés automatizálásának vizsgálatával. A költségek csökkentése és az ügyfelek bevonzása/megtartása szempontjából ezek elengedhetetlen változtatások a biztosítási piacon. Azonban, ahogy Batty et al. [2010] és Abrokwah [2015] is írják, nem feltétlenül kell/szabad csak ezekre a modellekre hagyatkoznunk a kockázatelbírálás során, hanem inkább egy előzetes szűrőként kell rájuk tekinteni, amik kategorizálják az ügyfeleket és megmondják, hogy kiknél érdemes további vizsgálatokat elvégezni szerződéskötés előtt. Ezért a dolgozat elemzés részében én is ezt tűzöm ki célul.

2.2. Időskori kockázatelbírálás

Magyarország vezető életbiztosítóinak (Aegon, Allianz, Generali, Groupama, Posta, NN, Union) termékeit áttekintve azt találtam, hogy jellemzően 60-70 éves korig lehet hagyományos módon kockázati életbiztosítást kötni, azzal a kiegészítéssel, hogy a biztosítás tartama nem tarthat például a biztosított 75 éves korán túl. Ennél idősebb korban csak csekély biztosítási összegért lehet biztosításhoz jutni, többnyire kockázatelbírálás nélkül, várakozási idővel. Viszont ezek a határok változhatnak a jövőben köszönhetően a longevity jelenségnek, vagyis hogy egyre nő az emberek várható élettartama.

Ösztönösen azt éreznénk, hogy minél idősebben akar valaki biztosítást kötni, annál nagyobb súllyal szerepel kockázatként a kora. Azonban a kor nem feltétlenül tekinthető az öregedés jelének, ugyanis egy mai x éves ember jellemzően jobb egészségi állapotban van, mint egy x éves ember volt pár évtizeddel ezelőtt. A változó élethelyzeteknek köszönhetően pedig más szempontok merülnek fel, amik meghatározzák az egyén egészségügyi állapotát. Klein [2013] amerikai biztosítási piacról írt összefoglalásából kiderül, hogy Amerikában már a 2000-es években felismerték az ebben rejlő lehetőségeket, és már a biztosítók fele rendelkezik idősekre szabott kockázatelbíráló programokkal. Ezek a programok jellemzően további kiegészítő kérdésekből állnak (élethelyzet, napi aktivitás, szociális tevékenység, fizikai aktivitás, mentális egészség), illetve az elfogadási határértékek is változhatnak a fiatalabb kori kockázatelbíráláshoz

képest, hiszen például a magasabb vérnyomás vagy súly nem feltétlenül akkora probléma idősebb korban. Továbbá kognitív és funkcionális tesztek végrehajtását is követelheti a biztosító, amelyekből akár a demencia előjeleit is ki lehet szűrni, valamint az öregkori elgyengülés esélyét is fel lehet mérni.

Az utóbbi fontosságára hívja fel a figyelmet Bennett [2004] is, különös tekintettel a társadalmi elszigeteltségre, kiszolgáltatottságra, általános fizikai romlásra, és krónikus betegségekre. Idős korban ugyanis akár már egy esés is végzetes lehet, azonban ennek bekövetkezési valószínűsége és súlyosságának mértéke nagyban függ a körülményektől: mennyire idős, egyedül él-e, esett-e már el korábban, tud-e segítséget hívni, egyéb betegségek gyengítik-e, stb. Ezenkívül még számos tényező befolyásolja az időskori egészségügyi állapotot, amiknek fiatalabb korban nincs akkora jelentősége. Éppen ezért a szerzők határértékeket is megadnak, amelyek normálisnak mondhatók az ügyfelek korától függően. Továbbá a költséghatékonyságra is felhívják a figyelmet, mivel idősebb korban nagyobb valószínűséggel derül fény valamilyen betegségre, így kevés az olyan vizsgálatok száma, ami nem fed fel valamilyen betegséget.

Wells [2017] prezentációjában azt láthatjuk, hogy az amerikai piacon kívül Franciaországban, Dél-Afrikában, és Koreában alkalmaznak külön idősekre szabott kockázatfelmérést. A világ többi részén egyelőre ez egy potenciális terjeszkedési mód, amire nagy valószínűséggel lenne kereslet. A dolgozat 3. fejezetétől kezdődően ennek a célcsoportnak a tulajdonságait elemzem különböző adatelemző módszerekkel. Előtte azonban még ismertetem a jelenleg igénybe vehető szolgáltatásokat, illetve hogy milyen új rendszerek segíthetnek az említett lehetőségeknek a kiaknázásában, amiknek figyelembe vételével személyre szabottabb biztosítási termékek jelenhetnek meg a piacon. A biztosítók megújulásának mozgatórugója nem feltétlen a gyors-, koralapú kockázatelbírálás vagy a szabályozásoknak való megfelelés, hanem a versenyelőny megszerzése. Ennek köszönhetőek a különböző innovatív megoldások.

2.3. Online kockázatfelmérés

Amerikában a hagyományos online kockázati életbiztosítások esetében azonnal lehet ajánlatot kapni néhány adat (kor, nem, saját egészség értékelése, magasság, súly, és dohányzási szokások) megadása, illetve a tartam (jellemzően 2-40 év között) és a biztosítási összeg (jellemzően 50.000$ - 10.000.000$ között) meghatározása után. Ezután azonban további adatokat kell megadni, hogy korrigálni tudják az ajánlatot a kockázat mértékével, vagy akár el is utasíthatják a kérelmet, ha nem tartják biztosíthatónak. Miután megadtuk személyes adatainkat

és kórtörténetünket, egy szakértő elbírálja a jelentkezésünket, ami akár 4-6 hétbe is beletelhet.

2.3.1. Lapetus Solutions - Chronos

Az online életbiztosítások piacán egyedinek számít a Lapetus Solutions algoritmusa: egy kép és néhány adat (kor, nem, súly, magasság, oktatás hossza, dohányzás, aktivitás, okoseszköz párosítása, alvás, idős kort megélő családtagok száma) alapján szinte azonnal képes felmérni az ügyfél kockázatát. A feltöltött kép alapján a program megbecsüli, hogy hány éves, milyen nemű, mi a BMI indexe és hogy dohányzik-e az illető. Ezt összevetve a felhasználó által megadott adatokkal egy becslést készít a várható hátralévő élettartamra, illetve arra, hogy mekkora valószínűséggel éli meg a felhasználó 65 és 85 éves korokat. A rendszer képes felismerni például, ha BMI index alapján tévesen túlsúlyosnak vélik az ügyfelet, vagy ha dohányzó létére az embernek mégis magas a várható hátralévő élettartama. Ez a technológia különösen népszerű lehet az Y generáció körében, mivel ők szívesebben osztanak meg magukról adatot kedvezőbb díjért cserébe (Seekings [2018]). De nem csak ők profitálhatnak ebből a megoldásból, hiszen a mesterséges intelligencia bevezetésével személyre szabottabb ajánlatokat kaphat mindenki, a biztosító költséghatékonyabban működhet, és nem utolsó sorban az ügyfélélmény is jelentősen növekedhet a mostani hagyományos rendszerhez képest. Ezenfelül pénzügyi tervezést, életmód javaslatot és nyugdíjmegtakarítási stratégia kialakítását is lehetővé tesz a Lapetus Solutions.

Azonban azt is fontos kiemelni, hogy a rendszer még nem működik tökéletesen. Saját képekkel kísérletezve azt találtam, hogy akár 5 évet is lehet faragni a becsült életkorból különböző szűrők és képjavító eszközök használatával. De eltekintve az utólagos módosításoktól, a kép készítésének körülményei is befolyásolhatják, hogy mit becsült a program, hiszen fontos tényező a megvilágítás, a szög, de akár az alany kedélyállapota is. Ez pedig azt eredményezi, hogy egy körültekintő ügyfél rengeteg időt fog eltölteni a tökéletes fotó megtalálásával, ami megkérdőjelezhetővé teszi, hogy valóban 10 percet venne csak igénybe a biztosítás megkötése.

2.3.2. Amerikai szabadalom

Bernico és Myers [2019] által bejegyzett szabadalom is azt vetíti előre, hogy van jövője azoknak a megoldásoknak, ahol az ügyfél kép-, videó, hanganyagot készít magáról, hogy többet tudjanak meg az egészséggel kapcsolatos tulajdonságairól. Ehhez vizsgálnák a személy szemmozgását, fogak és ínyek állapotát, illetve a nyaki ütőeret is. Ezekből a felvételekből lehetne a jelentkező egészségügyi jellemzőit megbecsülni, mint például a kort, súlyt, nemet,

edzettségi szintet, pulzust, gyógyszer-, cigaretta-, alkoholfogyasztást, de akár a koleszterin- és vércukorszintet is.

Összességében tehát ez az új irány szélesebb vevőréteget érhet el az innovatív megoldásaik révén. Ráadásul vonzó lehet az idősebb ügyfelek számára is, akiknek már nehezére esne a sok orvosi vizsgálat, és kérdőívek kitöltése. Azonban ahhoz, hogy a teljes kockázatelbírálási folyamat otthonról történhessen meg, még sokat kell javítani a rendszerek megbízhatóságán és a vizsgált tulajdonságok körét is fel kell térképezni.

2.4. Növekvő adatmennyiség lehetőségei és veszélyei

A biztosítás a kockázatkezelés egyik lehetősége, amelyben a hasonló kockázatokat viselő emberek, veszélyközösséget alkotva porlasztják kockázataikat. Vagyis a biztosítási díj fizetésének ellenében a kárt elszenvedő személy kompenzálást kap a közösen felhalmozott vagyonból. Ilyen veszélyközösségek megszervezésében játszanak kulcsszerepet a biztosítók, azonban nem lehet mindenféle kockázattól megszabadulni ilyen módon. Annak érdekében, hogy biztosítani lehessen egy kockázatot, több kritériumnak kell megfelelni: nagyszámú megfigyelés, homogén és véletlenszerű kockázat, egyértelmű és előre felbecsülhető kár, valamint gazdaságosnak is kell lennie. Az életbiztosítások is erre az elvre épülnek, ahol a veszélyközösségeket az azonos kockázati besorolású személyek alkotják.

Ahogy Kivisaari [2018] is írta, minél több paraméter alapján tudjuk besorolni az ügyfeleket, annál inkább csökken a keresztfinanszírozás a csoportok között, ami logikus célkitűzésnek tűnhet, hiszen így mindenki a kockázatának megfelelő biztosítási díjat fizeti, viszont számolni kell a veszélyeivel és korlátaival is. A túl nagy kockázatúak így nehezebben juthatnak biztosításhoz, mivelhogy megszűnik az alacsonyabb kockázatúaktól származó keresztfinanszírozás. Illetve annak a lehetőségét is figyelembe kell venni, hogy azok is kilépnek a biztosítási piacról, akik ráeszmélnek, hogy mennyire alacsony a kockázatuk. Ez összességében a biztosító profitabilitását is ronthatja (Keating [2017]).

Ugyanakkor a biztosíthatóság koncepcióját is szem előtt kell tartani. Erről Eling és Kraft [2017] értekeztek, és azt találták, hogy lehetnek erkölcsi problémák az életbiztosítás terén. Ilyen problémákat vethet fel az árazáshoz figyelembe vett változók használata. Ugyanis az emberek kevésbé fogadják el, ha olyan változók alapján kell magasabb díjat fizetniük, amire nincsen ráhatásuk (Schmeiser et al. [2016]), ami életbiztosítások esetén öröklődő betegségek formájában sokszor előfordulhat. Ehhez persze szükséges az is, hogy az ügyfél tisztában legyen

a kockázatelbírálási folyamattal, amit viszont a biztosító nem szívesen hoz az ügyfelek és így a versenytársak tudomására.

A személyes adatok védelme is fontos kérdés, tekintve hogy olyan bizalmas adatokhoz engedhet hozzáférést a biztosított, amiből következtetni lehet a magánéletére. Abban az esetben pedig, ha nem teszi ezeket elérhetővé, magasabb díjjal szembesülhet, illetve kevesebb engedményt kaphat, mert a biztosító feltételezheti, hogy rejteget előle valamit az ügyfél. Azonban a fogyasztó tudta nélkül is értékes információkhoz juthat a biztosító egy harmadik fél által. A napi internetezés közben generált adatok szintén releváns információkkal szolgálhatnak a biztosítók számára, hiszen keresési előzményekből vagy akár feltöltött képekből is rengeteg értékes információt meg lehet tudni. Bár a jogi és erkölcsi normáknak meg kell felelnie a biztosítóknak, hiszen ha romlik a megítélésük, akkor végső soron még rosszabbul is járhatnak, mint ha eleve figyelembe se vették volna a kérdéses változókat. Ennek a jelenségnek a fontosságára hívja fel a figyelmet GDPR (*General Data Protection Regulation*) rendelet is, amely az Európai Gazdasági Térség országaiban élő természetes személyek adatait védi.

Berthelé [2018] is a túlzott személyre szabás veszélyeiről ír, amelyben azzal érvel, hogy minimális kockázatmegosztás mindenképpen szükséges lesz, hogy egy termék életképességét garantálni tudják. Valamint a kockázatok tökéletes felmérése soha nem lesz lehetséges, mivel külső hatások is befolyásolják azokat, amiket nem lehet mérni.[1] Ugyanakkor az előnyei is megvannak, hiszen egy alapból magas kockázati besorolású ügyfél ezáltal jelezhet, hogy mégsem olyan nagy a kockázata. Vagyis a biztosítási piacot érintő információs aszimmetria mérséklődhet.

Az egyre több adat miatt ezek a problémák egyre többször kerülhetnek előtérbe, azonban ahogy azt már említettem a 2.1.1. alfejezetben, a növekvő adatmennyiség csak akkor lehet releváns a biztosító számára, ha megbízható, pontos és felhasználható. Az újdonságot nem is feltétlenül a biztosítás megkötésénél rendelkezésre álló több adat jelenti, hanem hogy ezt követően is új adatok állnak rendelkezésre az ügyfélről. A jelenlegi hagyományos biztosításoknál ugyanis nem veszik figyelembe, hogy megváltozott-e az ügyfél kockázata a későbbiekben, illetve nem tudnak róla.[2] Ezáltal nem tudják kiszűrni, ha az ügyfél olyan dimenzió mentén növeli a kockázatát, amit egyértelműen befolyásolni tud: például egy nem dohányzó személy a biztosítás megkötése után dohányozni kezd. Ezen kockázatok kezelhetőségének lehetőségét fejtem ki a továbbiakban.

[1] Például ilyen külső hatásnak tekinthető a dolgozat írása közben elterjedő koronavírus is.

[2] Eltekintve azoktól az eseményektől, amelyek bekövetkezése esetén az ügyfelet közlési kötelezettség terheli.

2.5. Használatalapú árazás

A használatalapú árazás azt a dinamikus árazási technikát jelenti, ami egy eszköz segítségével a biztosított viselkedését is figyelembe veszi a díjszabáshoz. Legelterjedtebb formájában a gépjármű-biztosításoknál találkozhatunk vele, ahol az autókba szerelt műszerekkel szereznek információt az autókról, illetve a vezetési szokásokról, hogy szükség esetén korrigálni tudják az árat a tényleges kockázathoz igazítva (Boobier [2016]). A telematikus eszközök elterjedéséről és jelenlegi helyzetéről Hauer et al. [2017] cikkében olvashatunk. Hazai tekintetben még használatuk nagyon kezdetleges, csupán az Aegon, a Posta biztosító és a Generali próbálkozott ilyen koncepciók bevezetésével. Viszont Amerikában és fejlett nyugat-európai országokban már régóta használják ezeket a rendszereket. A pontosabb, kedvezőbb díjszabáson felül további előnyöket is kiemelnek a szerzők: vezetéstechnika javaslatok, értesítés lehetséges veszélyekről, automatikus kárbejelentés, közvetlen kapcsolat autójavító műhelyekkel, stb., melyekbe a fogyasztót interaktív módon a program játékosításával még inkább be lehet vonni. Ezek a gondolatok részben átültethetőek az életbiztosításba is.

2.5.1. Okoseszközök használata

Fontos leszögezni az alfejezet elején, hogy a biztosítót elsősorban nem az érdekli, hogy egészségesebb életmódra ösztönözze ügyfeleit, hanem hogy többletinformáció és új termékei által nagyobb profitot tudjon elérni. Ennek csupán következménye, hogy az emberek tudatosabb életmódot folytatnak, ha olyan termékeket fejlesztenek a biztosítók, amik megkívánják az egészségügyi értékek folytonos monitorozását. Az autókat figyelő rendszerek analógiájára az életbiztosítás területén az okoskarkötők, okosórák, okostelefonok gyűjthetik az adatokat az ügyfelekről. Ezek az eszközök monitorozzák a pulzust, lépésszámot, alvást, sporttevékenységeket, de akár a bevitt kalóriát, folyadékot és testsúlyt is nyomon lehet követni.

Ezekkel az adatokkal napi szinten való szembesülés eredményezhet egészségtudatosabb életmódot (Neaţu [2015]), bár ennek közvetlen egészségnövelő hatását Finkelstein et al. [2016] vitatják. Kutatásukból, melyben 800 főt vizsgáltak egy éven keresztül, az is kiderült, hogy a monetáris ösztönzőknek csak rövid távon van hatásuk, ezért a kompenzálásnak folyamatosnak kell lennie, hogy hosszabb távon is fent lehessen tartani az ügyfelek érdeklődését. Az eszközök ösztönző hatását Jakicic et al. [2016] is vizsgálták. 470 BMI érték alapján túlsúlyos embernek állítottak össze egy programot, amiben meghatározták az étrendjüket, és a hetente mozgással eltöltött percek irányadó mennyiségét. A vizsgálat két évig tartott, melynek keretein belül havonta

visszajelzést kaptak a haladásukról, illetve hetente emlékeztető üzeneteket a feladatokról. A résztvevők felének továbbá okoskarkötők használatát is biztosították, azonban nem volt megfigyelhető, hogy ez bármilyen pozitív hatással lett volna az eredményre.

Vagyis elmondható, hogy az okoseszközök csak akkor működnek hatásosan, ha van valami további ösztönző, ami elkötelezetté teszi az ügyfelet a használatában, mivel a viseléséből fakadó kezdeti lelkesedés hamar alábbhagy. Ezt a célt szolgálja például a Fitpuli vállalati egészségprogram, ami összekapcsolja egy cég munkatársait, és motiválja őket a tudatosabb életmódra. Azonban egy másik megoldást jelenthet a monitorozásra, ha olyan eszközöket alkalmazunk melyek nem okoznak többletterhet a viselője számára. Ezek az eszközök is megtalálhatóak már a piacon, mint például cipő, talpbetét, alsónemű, öv és számos egyéb ruházati cikk formájában. Használatuk pedig nem igényel extra erőfeszítést az ügyféltől, mivel az adatok automatikusan generálódnak (Spender et al. [2019]). Ezek elterjedése előtt, azonban az okostelefonok szolgálhatnak elsődleges adatforrásként, hiszen rengeteg adatot tárol tulajdonosáról, és kellően sok ember rendelkezik vele, vagyis a biztosítónak sem okozna jelentős többletkiadást a használata. Az eszközök mérésének pontatlansága pedig nem mindenképpen okoz problémát, mert a pontos eredmények helyett inkább a változás detektálása a fontos szempont. Effajta eszközök az idős ügyfelek esetén lényegesen csökkenthetik a kockázatot, ugyanis sok problémát azonnal detektálni tudnak, és értesíteni tudják a megfelelő embert, vagyis használatuk jelentősen csökkenthetik a díjukat.

Ezeknek az eszközöknek a használata már elkezdődött a biztosítási piacon is, ahol applikációk és ösztönzők segítségével (vásárlási utalványok, kedvezmények, kedvezőbb díjak) próbálják rávenni az ügyfeleket, hogy egészségesebb életet éljenek. Az első ilyen egészségprogram a *Discovery Limited* pénzügyi szolgáltató csoporthoz köthető, amelynek *Vitality* terméke már több, mint 17 országban van jelen. Az ügyfelek pontokat gyűjthetnek szűrővizsgálatokon való részvétellel, testmozgással és egészséges táplálkozással, melyeknek megfelelően különböző szinteket érhetnek el és egyre nagyobb kedvezményekben részesülhetnek. Más biztosítók is figyelembe veszik, ha aktív és egészséges ügyfél jelentkezik hozzájuk, mint például a 2.3.1. alfejezetben említett rendszer is, ahova mozgással kapcsolatos adatokat lehet feltölteni. A rendszer megbízhatóságát tovább javíthatja a *Health IQ* által alkalmazott módszer, ami a megfelelő sport(ok) és életmód kiválasztása után ellenőrző kérdéseket tesz fel, hogy kiszűrje a kedvezményre nem jogosult ügyfeleket. Ezen termékek példájára, az okoseszközök kifejezetten idősekre való alkalmazása sem tűnik valószerűtlennek a közeljövőben.

3. Algoritmusok ismertetése

Ahhoz, hogy meg tudjuk becsülni ki hal meg egy adott időszakon belül, olyan módszerre van szükségünk, amellyel lehetséges a bináris klasszifikáció. A lehetséges algoritmusokat a témában fellelhető elemzések során használt gyakoriságuk és a magyarázó változókra tett megkötésük alapján választottam ki. Így a logisztikus regresszió, fa alapú modellek, és a neurális hálók elméletét ismertetem röviden ebben a fejezetben.

3.1. Logisztikus regresszió

A klasszifikációs problémák modellezésére is alkalmasak Nelder és Wedderburn [1972] által javasolt általánosított lineáris modellek. Ezekben a modellekben exponenciális eloszláscsaládbeli kimeneteket lehet modellezni, melyekben a prediktoroktól függő várható értéket transzformáljuk egy kapocsfüggvénnyel, ami így már felírható lineáris függvény segítségével. A legegyszerűbb példa GLM-re a lineáris regresszió, ahol kimenet eloszlása normális, a kapocsfüggvény pedig az identitás függvény. Esetünkben azonban a Bernoulli eloszlású kimeneteket kell vizsgálni, illetve olyan kapocsfüggvényt kell alkalmazni, amely a Bernoulli eloszlás várható értékét a $[0,1]$-ről transzformálja a $(-\infty,\infty)$ intervallumra. Erre a legelterjedtebb függvények a logit, probit, cloglog, de más függvények is használhatók, amik a valós számok halmazára képeznek a $(0,1)$ intervallumról. Az első függvény használata esetén kapjuk meg a logisztikus regressziót, melynek elméletét Kovács [2014] alapján ismertetem.

A logit transzformáció szerint az esélyhányados logaritmusára illesztünk egy lineáris modellt. Ha Y eredményváltozó várható értéke p, és n darab magyarázó változó van a modellben, akkor az alábbi módon írható fel az egyenlet:

$$logit(\mathbf{E}(Y)) = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Az egyenlet paramétereit Maximum Likelihood becslés alapján határozzuk meg. A becsült β együtthatók alapján pedig meg lehet mondani, hogy a j-edik magyarázó változó mennyivel növeli logit értékét, vagy hányszorosára növeli az esélyhányadost, valamint azt is ki lehet számolni Wald-teszt segítségével, hogy mennyire szignifikáns az adott együttható a becslés szempontjából. A kategóriába esés becsült valószínűsége az alábbi képlet szerint számítható:

$$\hat{p} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

A dolgozatban a kategóriát a halál bekövetkezte jelenti, vagyis minél nagyobb az esélyhányados értéke, annál valószínűbb, hogy halált fog becsülni a modell. A becsült valószínűségekkel és egy vágási érték meghatározásával pedig elkészíthetünk egy klasszifikációs táblát, amelynek segítségével meg lehet állapítani a modell jóságát, illetve pontosságát. A modellek teljesítményére és összehasonlítására azonban más mérőszámokat is szokás használni. A modell egészét Pearson-féle χ^2 teszt segítségével minősíthetjük, míg összehasonlítás esetén a különböző pszeudó R^2-ek alapján dönthetünk. McFadden, Cox és Snell, valamint Nagelkerke mutatója is hasonlóan a regressziós R^2-hez, egy 0 és 1 közötti számmal jellemzik a modellt. Számítási módjuk azonban eltér, hiszen az utóbbi esetben az átlaggal való becslés szórásához viszonyítunk, az előbbi esetekben pedig a nullmodell likelihood függvényéhez viszonyítunk, éppen ezért önmagukban nem szokás értelmezni őket. Egy másik szokásos mérőszám az AUC görbe alatti terület nagysága, amelyre részletesebben kitérek a 3.4. alfejezetben.

3.2. Döntési fa / véletlen erdő

A fa alapú modellekről Breiman et al. [1984] értekeztek először. Ezek szintén nagyon népszerűek a tudományos életben, hiszen hasonlóan a logisztikus regresszióhoz könnyen értelmezhető az algoritmus és intuitív módon történik a csoportokba való besorolás. Mind regresszióra, mind pedig klasszifikációra alkalmasak, és matematikai előfeltevései sincsenek. Hátránya viszont, hogy hajlamos a túlillesztésre, aminek köszönhetően előrejelző képessége elmaradhat a többi modelltől.

A döntési fa elnevezés abból adódik, hogy a megfigyeléseket egy fa alapján bontjuk meg, ahol az egyes ágakon és leveleken a célváltozó szerinti minél homogénebb csoportokat szeretnénk létrehozni, amihez a magyarázó változókat használjuk fel. Az algoritmus a törzsből indul ki, ahol minden megfigyelés szerepel, majd úgy próbál ágaztatni, hogy a heterogenitás minél jobban csökkenjen. Klasszifikáció esetén ez a Gini-mutató alapján történik, ami a következő módon számolódik egy ágra, ahol f jelzi a kategória részarányát:

$$G = 1 - f_{élt}^2 - f_{meghalt}^2$$

Ez minden ágra kiszámolódik, majd ezeknek veszi az algoritmus a ágakon vett megfigyelésekkel való súlyozott átlagát. Kategorikus magyarázó változók esetén minden lehetséges kettébontás alapján, míg numerikus változók esetén minden decilis mentén nézi meg az algoritmus, hogy mikor érhető el a legnagyobb csökkenés az előbb ismertetett mutatóban.

A fa építése során számos paramétert meghatározhatunk, amivel javíthatjuk a modell teljesítményét, és védekezhetünk a túlillesztés ellen. Az egyik legfontosabb ilyen paraméter a *cp*, vagyis a *complexity parameter*. Ennek meghatározásával beállíthatjuk, hogy mi az a minimális javulás, amit elvárunk a modelltől az újabb elágazások esetén. Így jelentősen csökkenthetjük a fa méretét, és a futási idő nagymértékben javulhat. A túlillesztést más paraméterek segítségével is el lehet kerülni. Ilyen például az ágakon lévő minimális egyedszám, amivel el lehet kerülni, hogy túl kicsi csoportokat hozzon létre az egyes ágakon az algoritmus, illetve a fa maximális mélysége is jelentősen mérsékelheti ezt a problémát, azáltal hogy nem engedi nagyon szerteágazni a fát. Az 5.3.2. alfejezetben én is ezen paraméterek kalibrálásával határozom meg a legjobban teljesítő modelleket.

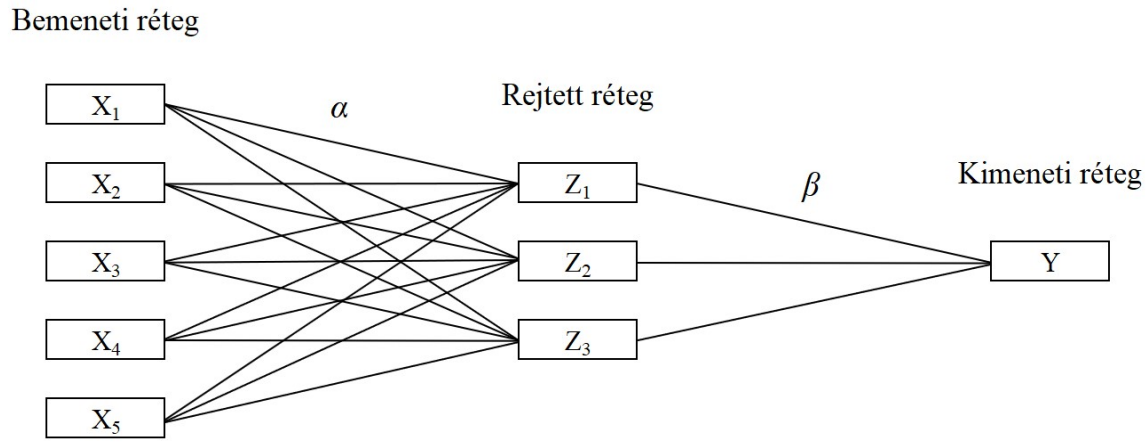
Véletlen erdő módszer esetén is van lehetőség klasszifikációra és regresszióra is. Ennél az algoritmusnál sok véletlen fát hozunk létre, és ezen modellek becslése alapján vonjuk le a következtetést a végső becslést illetően. Regresszió esetén a szavazatok átlagolásával, klasszifikáció esetén pedig a több szavazatot kapó kategória kiválasztásával történik az előrejelzés. Fontos megjegyezni, hogy az ilyen típusú modellezés akkor hasznos, ha az egyes modellek nem azonosak. Ezt el lehet érni a megfigyelésekből vett véletlen visszatevéssel vagy a változók számának véletlen meghatározásával. Elemzésem során az utóbbit választottam, mellyel tehát azt szabtam meg, hogy az egyes fáknál mennyi véletlenül kiválasztott változó alapján történjen a bekategorizálás. Egy másik fontos paraméter ebben az esetben a fák száma, vagyis hogy mennyi szavazat alapján szülessen meg a döntés. Minél több fát használunk, annál pontosabb az előrejelzés, bár a futási időt is szem előtt kell tartani ebben az esetben, és ez alapján meghatározni a fák számát. Ezenfelül a döntési fához hasonlóan, itt is meg lehet határozni a fák maximális mélységét, és az ágakon lévő minimális esetszámot, amivel itt is védekezhetünk a túlillesztés ellen. Ahogy azonban láthatjuk majd az 5.3.2. alfejezetben, nem törvényszerű, hogy kevésbé jelentkezik a túlillesztés problémája véletlen erdők esetén.

A módszerek jellegéből adódóan tehát a véletlen erdő általában kevésbé hajlamos a túlillesztésre és pontosabb becslésre is képes, mint a döntési fa, azonban ez azzal jár, hogy nem lehet könnyen értelmezni az algoritmust. Attól függően, hogy milyen problémát vizsgálunk ez akár döntő tényező is lehet a modellválasztást illetően, hiszen például ebben a dolgozatban is a változók hatásának vizsgálata az elsődleges szempont, ami véletlen erdő esetén nem visszakövethető.

3.3. Neurális hálók

Harmadik algoritmusnak az egyre népszerűbb neurális hálókat választottam, melyet Beale et al. [1996] alapján ismertetek. Ezzel a módszerrel is lehetséges a bináris klasszifikáció, és komplex struktúrájának köszönhetően olyan kapcsolatokat is fel lehet vele fedezni a változók között, melyeket a többi algoritmus nem tud megragadni. Elnevezését az agyban lévő neurális hálózatokhoz való hasonlóságáról kapta. Komplexitása miatt ez egy fekete doboz algoritmus, vagyis nem lehet visszafejteni, hogy mi alapján történik meg a besorolás, mert sokszor olyan jellemvonásra tanul rá, ami az ember számára értelmezhetetlen. Az algoritmus szemléltetéséhez az 1. ábrát használom.

1. ábra. Neurális hálózat



Forrás: Saját szerkesztés (Excelben)

A neurális hálózatokat három részre lehet osztani: bemeneti réteg, rejtett réteg és kimeneti réteg. Az 1. ábrán az első oszlop a bemeneti réteg, vagyis a magyarázó változók, és egy darab kimeneti réteg van, amin belül lehet akár több neuron is az ábrával ellentétben. Legegyszerűbb alakjában, vagyis mikor nincsen rejtett réteg, akkor a probléma jellegétől függően vagy a lineáris vagy a logisztikus regressziót kaphatjuk vissza megfelelő transzformáló függvénnyel. Rejtett rétegek szerepeltetésével azonban bonyolultabb kapcsolatok is megragadhatóak. Az ábrán három darab neuron látható a rejtett rétegben, az ezekbe vezető nyilak mindegyikén egy α súly szerepel, amikkel figyelembe vesszük a bemeneti paramétereket a következő képlet alapján:

$$Z_i = \sigma(\alpha_{0i} + \alpha_i^T X), i = 1, 2, 3$$

Mint látható, a bemeneti paraméterek súlyozottan aggregált értékét vesszük figyelembe, amit egy aktivációs függvény (σ) segítségével transzformálunk. Az α_0 értékek tulajdonképpen a

torzítást jelentő paraméterek, amikkel az aktivációs függvényt lehet eltolni. Az így kapott értékeket szintén aggregáljuk β súlyok figyelembe vételével és egy függvénnyel olyan alakra transzformáljuk, amit a kimeneti változó megkövetel. Ezt a következő formula mutatja:

$$Y = h(\beta_0 + \beta^T Z)$$

Fontos megjegyezni, hogy mindegy milyen értéket vesznek fel az egyes változók a rendszeren belül, csupán az számít, hogy az utolsó transzformáló függvénnyel (jelen esetben a h függvény) a megfelelő alakra hozzuk. Éppen ezért nem is szükséges, hogy az aktivációs függvények azonosak legyenek. Számos formája létezik, melyeket lehet alkalmazni neurális hálózatok építése során: Ilyen például a lépcsős függvény, ami egy bizonyos érték alatt nullát, felette egyet ad eredményül; a lineáris függvény, ami az input értéket adja vissza; és a szigmoid függvény is, amit már a logisztikus regresszió során is alkalmaztunk a becsült valószínűségek kiszámítására. Így ezeket az értékeket úgy lehet interpretálni, hogy minél nagyobb, annál aktívabban jelez, illetve továbbítja a jelet az adott neuron. Az algoritmus futtatása előtt érdemes a bemeneti változókat azonos mérési skálára transzformálni, mivel az aktivációs függvények sok esetben nem tudnak jelentős különbséget tenni az értékek között, ha azok túl nagyok vagy túl kicsik. Több rejtett réteg alkalmazásával tovább lehet bonyolítani a rendszert. Hasonlóan az előbbi képletkehez, itt is összegeznénk az előző réteg kimeneti értékeit megfelelő súlyokkal, majd transzformálnánk őket. Az alkalmazott rejtett rétegek és az azokban szereplő neuronok számára nincsen előzetes szabály, ennek meghatározása a probléma jellegétől függ, és próbálgatás útján lehet megtalálni az optimális kombinációt. Ebben a példában csak egy darab rejtett réteg volt, ahol a Z_i neuronok a bemeneti paraméterek egy tulajdonságára tanultak rá. Szemléletesen ez például azt jelenthetné a dolgozatban tárgyalt problémára levetítve, hogy a Z_1 neuron akkor aktivizálódik, ha egy személy aktív, egészséges életet él, míg a Z_2 neuron akkor, ha az illetőnek sok betegsége, korlátozottsága van.

Az algoritmus során tehát az elsődleges cél a megfelelő súlyok megtalálása, amit nem hagyományos paraméterbecslési eljárásokkal határozunk meg, hanem jellemzően hibavisszaterjesztéses algoritmussal. A módszer lényege, hogy egy véletlenül kiválasztott súlyrendszerből kiindulva kiszámolja, hogy mekkora hibát követ el a becslés során, majd ezt visszatáplálva a modellbe úgy próbálja megváltoztatni a súlyokat az algoritmus, hogy a legnagyobb mértékben csökkenjen a hiba.

A következő alfejezetben arra térek ki, hogy hogyan lehet, illetve hogyan érdemes összehasonlítani az ismertetett modellek eredményeit.

3.4. Modellek értékelése

A modellek teljesítményét elsősorban az 1. táblázatban látható igazságmátrixszal (*confusion matrix*), és a ROC görbéhez (*receiver operating characteristic curve*) hasonló PR görbe (*precision recall curve*) segítségével határozom meg. Az 1. táblázatban a szokásos elnevezéseket használom, illetve jelen esetben az esemény bekövetkeztét jelző „*Positive*" szó a halál bekövetkeztét jelenti.

1. táblázat. Klasszifikációs tábla

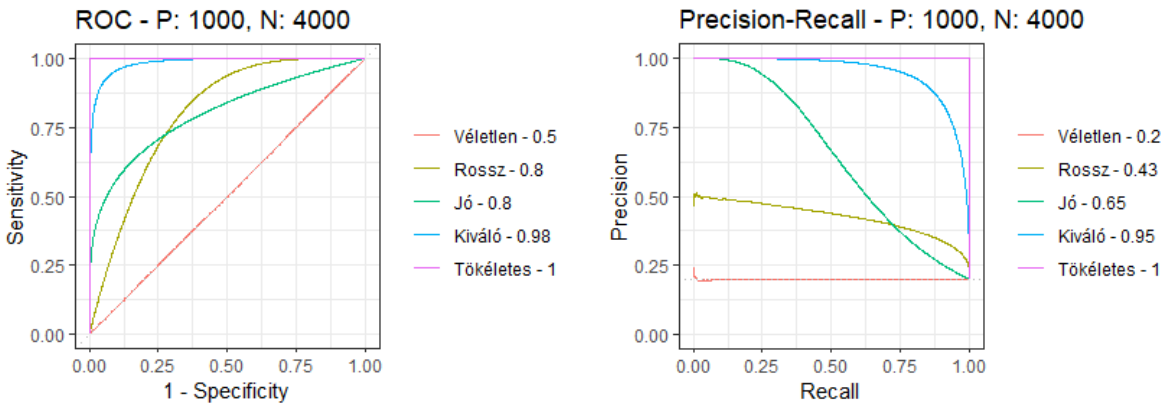
Megfigyelt \ Becsült	Élt	Meghalt
Élt	*True Negative* (TN)	*False Positive* (FP)
Meghalt	*False Negative* (FN)	*True Positive* (TP)

Forrás: Saját készítés

A gyakorlati használhatóságot szem előtt tartva két szemszögből is érdemes megvizsgálni a kapott eredményeket. A modellek jóságának mindenképpen meghatározó mutatója, hogy mennyi tényleges halálesetet tud előrejelezni. Ha konkrét előrejezésre szeretnénk használni a modelleket, akkor ezenkívül még azt érdemes figyelembe venni, hogy minél kevesebb meghalt emberre jelezze azt a modell, hogy életben maradt (FN), mivel ennek a hibának nagyobb a költsége. Azonban ha a modellt csak egy előzetes szűrőként akarjuk használni a 2.1.2. alfejezetben ismertettek szerint, akkor a meghaltnak becsült, de életben levő emberek számát (FP) érdemes minimalizálni, hiszen ezáltal úgy tudunk rossz ügyfeleket elutasítani, hogy kevés jó ügyfelet küldünk el. A feldolgozott irodalom alapján az utóbbi módszer alkalmazása tűnik életszerűbbnek, ezért az 5.3. alfejezetben ezt vizsgálom. A csoportokba történő besorolás valószínűségek alapján történik, ami egy mögöttes paramétertől függ. A küszöbérték alapértelmezetten 0.5, ami azt jelenti, hogy csak akkor jósol halált a modellt, ha több mint 50% ennek a valószínűsége. Ennek az értéknek a változtatásával változnak az igazságmátrixban szereplő értékek, így a tényleges összehasonlításhoz ROC görbét szokás használni, illetve az abból származtatható AUC értéket. Ez a görbe a helyesen becsült pozitív esetek arányát (*Sensitivity*= $\frac{TP}{TP+FN}$) és a hamis pozitív arány ($1 -$ *Specificity*= $\frac{FP}{FP+TN}$) kapcsolatát ábrázolja. Azonban ha egy csoport felismerése fontosabb a másiknál vagy nagyon aránytalan a csoportok eloszlása, akkor érdemes lehet a „*Precision-Recall*" görbét használni. Ez a görbe nem veszi figyelembe a TN csoportot, ami jellemzően kiugróan nagy tud lenni azoknál a problémáknál, amelyeknél kevés a pozitív kimenetel. Ezért a „*Specificity*" mutató helyett a „*Precision*" (= $\frac{TP}{TP+FP}$) mutató függvényében ábrázolja az előbb ismertetett „*Sensitivity*" értéket, melynek neve ebben az esetben

„*Recall*”. A görbék közötti kapcsolatot 2. ábrán szemléltetem, ahol 1000 pozitív, és 4000 negatív megfigyelésünk van.[3] A különböző modellek teljesítményét jelmagyarázatban feltüntetett görbe alatti területekkel lehet mérni.[4]

2. ábra. ROC és PR görbék közötti összefüggés



(a) ROC görbék

(b) PR görbék

Forrás: Saját szerkesztés (R-ben), package: „*precrec*” 2017

A vágási érték csökkentésével egyre több megfigyelést becsülünk pozitívnak, míg végül mindenkit annak becsülünk. Ekkor a „*Recall*” 1-es értéket vesz fel, miközben a „*Precision*” a mintában lévő pozitív megfigyelések arányához tart. Jelen esetben 0,2-höz. A vágási érték növelésével a „*Recall*” 0-hoz tart, viszont a „*Precision*” értéke [0,1] intervallumon bárhova tarhat, aminek köszönhetően nagy szórás jellemzi a görbe ezen részét. A PR görbe jellegzetességeivel és az AUCPR mutató becslésével Boyd et al. [2013] foglalkoznak, akik különböző módszereket tekintenek át a mutató meghatározására és becslésére generált adatok alapján.

Jól látható a görbék közötti összefüggés. A diagonális ROC görbe felel meg a jobb oldali ábrán levő vízszintes egyenesnek, amelyek a teljesen hasztalan modelleket jelzik. A vízszintes vonal éppen mintában szereplő pozitív megfigyelések arányát adja vissza. A lehetséges legjobb PR görbe 1-es értéket vesz fel és kicsúcsosodik, hasonlóan a legjobb ROC görbéhez. A hasonlóság elmondható a kiválóan teljesítő modellek esetében is. Azonban az is látható, hogy a PR görbék ábráján szereplő „Rossz” és „Jó” esetek jól megkülönböztethetők a görbe alatti terület alapján is, viszont a ROC görbe alatti terület szerinti teljesítményük alapján ugyanolyan jól szerepelnek. Ez teszi szükségessé PR görbe alatti terület használatát, ugyanis a „*Precision*” mutató maximalizálása a fontos a biztosítók számára előzetes szűrés esetén.

[3]Ez közel azonos a SHARE adatok között szereplő élők és meghaltak arányával a teszt mintámban.

[4]ROC görbe esetén AUC, PR görbe esetén AUCPR az elnevezése

3.4.1. Gyakorlati szempont

A gyakorlati alkalmazhatóság azonban konkrét vágási értéket követel meg, illetve azt, hogy a biztosítónak nyereséges legyen használata. Vagyis optimális érték mellett a biztosító a lehető legjobban csökkenti a költségeit, amellett hogy a lehető legkisebb veszteség éri az eredeti állapothoz képest. A vizsgálatot tovább bonyolítja, ha a kizárás lehetőségével is számolunk. Ugyanakkor a biztosító minden halálesetnél megnézi, hogy olyan betegséggel összefüggésben halt-e meg a biztosított, ami már a szerződéskötéskor ismert volt. Így ha előzetesen nem is kérdeznek rá az ügyfél minden betegségére, utólag kiderülhetnek a kizáró okok. Ezért a modellek során nem veszem figyelembe, viszont utólag vizsgálom a hatását.

A költségek csökkenése elsősorban az automatizált előszűrésből adódik, ahol nem kell a kezdeti kockázatelbírálás költségeit állni a rossznak jósolt ügyfeleknél, illetve a hagyományos kockázatelbírálás során tévesen elfogadott vagy elutasított ügyfelek megtalálása is kedvezően hat. A veszteség pedig abból fakad, hogy tévesen elutasításra kerülhetnek, akik a hagyományos kockázatelbírálás során nem lettek volna elutasítva, valamint tévesen élőnek becsülhetjük, azokat akiket hagyományos módon elutasítottunk volna. Az áttekinthetőség érdekében a 2. táblázatban szemléltetem ez előbb leírtakat, ahol az „Él” és „Meghal"" az automatizált-, a „Jó” és „Rossz” pedig a hagyományos kockázatelbírálásra vonatkozik.

2. táblázat. Automatizált rendszer okozta többletköltség és többletnyereség esetei

Megfigyelt \ Becsült / Elbírált	Él / Jó	Meghal / Jó	Él / Rossz	Meghal / Rossz
Élt	∘	+−	+	+
Meghalt	∘	++	−	+

Forrás: Saját készítés

Ahhoz hogy ilyen szempontból tudjam vizsgálni a modellek teljesítőképességét, egy olyan biztosítói adatállományra lenne szükség, ahol azok is szerepelnek, akik elutasításra kerültek a kockázatelbírálás során, valamint az is, hogy mennyibe került a kockázatoknak a felmérése. Ha pedig nem a hagyományos elbírálást tekintjük alapértelmezettnek, hanem a valóban bekövetkezett halálozásokat, akkor SHARE adatokhoz hasonlóan arra is szükség lenne, hogy mindenkiről tudjuk utólag, hogy meghaltak-e vagy sem. Ezeknek az adatoknak a hiányában azonban nem teszek kísérletet konkrét számításokkal meghatározni az optimális vágási értékeket és alátámasztani a modellek létjogosultságát, helyette inkább a modellek pontosságát szemléltetem a különböző esetekben.

4. SHARE adatok elemzése

A Survey of Health, Ageing and Retirement in Europe (SHARE) adatbázisa kutatási és oktatási célokat szolgál, melyhez bárki hozzáférhet a megfelelő nyilatkozat kitöltése után. A webhelyen[5] számos dokumentáció érhető el az adatfelvétellel és a feltett kérdésekkel kapcsolatban, melyek folyamatosan változnak, bővülnek. 2004 óta hét különböző adatfelvétel történt, melyeket hullámoknak neveztek el. Ezeknek a hullámoknak az elsődleges célcsoportja az Európa-szerte lakó 50 évnél idősebb emberek, akiket az egészségükről, társadalmi kapcsolataikról, gazdasági helyzetükről és mindennapjaikról kérdeznek meg egy személyes interjú formájában. Mivel az interjúalanyokhoz egyedi azonosítót rendelnek, így lehetővé válik az egyének életútjának követése, ami számos kutatás alapjául szolgál. Az egyes hullámokban való részvételről és feltett kérdések köréről az A függelékben adok részletes tájékoztatást.

4.1. Adatállomány bemutatása

4.1.1. Adat-előkészítés

A SHARE (2019) adatbázis 4. hullámban gyűjtött adatait használtam, mivel ez volt az első hullám, amiben Magyarország is szerepelt. A következő hullám a 7. volt, amiben Magyarország részt vett, ezért az elemzéshez azokat az országokat választottam ki, amik szintén szerepeltek a 4. és 7. hullámokban. Ezek megtekinthetők az A függelék 9. táblázatában.[6] Mivel azonban a többi ország jellemzően az 5. és 6. hullámban is részt vett, ezért az ezekben a hullámokban bekövetkezett halálesetekről sem szabad megfeledkezni, ugyanis ezeket csak a halál utáni első hullámban szerepeltetik a SHARE készítői. Vagyis ha ezektől eltekintenék, akkor csak a 6. hullám után elhunytakról lenne információm, amit nem tudnék összehasonlítani a magyar adatokkal. Az adattáblák kezeléséhez és az elemzéshez az R 2019 adatelemző szoftvert használtam.

4.1.2. Hiányzó adatok kezelése

Az adatok között sok hiányzó érték volt, azonban a SHARE készítői sok változó esetében elvégezték a hiányzó adatok pótlását. Öt külön szcenáriót futtattak, melyekben "Hot-deck" és "Fully conditional specification" módszereket alkalmaztak az imputálásra. Az előbbi az

[5] http://www.share-project.org

[6] Hollandia kivételével, mivel a 7. hullámban gyűjtött adataik még nem kerültek feldolgozásra.

adott megfigyeléshez "leghasonlóbb" megfigyelést keresi meg, és annak az értékével pótol, míg az utóbbi feltételes modelleket illeszt változónként, ahol az összes többi elérhető magyarázó változóval becsüli meg a hiányzó értékeket. Az esetek többségében öt azonos éréket becsültek a modellek (főleg a kategorikus változók esetében), így a modellekben ezen szcenáriók közül az elsőt választottam ki. Azonban nem minden változó szerint határoztak meg lehetséges értékeket a hiányzó megfigyelésekre, amiknek kezelését több lépésben oldottam meg. Először is megnéztem változónként a hiányzó megfigyelések számát, és amelyeknél több, mint 30% hiányzott, azokat töröltem. Sok ilyen változó volt az adattáblában, amiket így nem tudtam felhasználni az elemzésemhez. Többek között ezért sem tudtam közvetlenül figyelembe venni a szülők által megélt maximális életkorokat, pedig ahogy azt a 2.3.1. alfejezetben is láthattuk, a gyakorlatban már ez is döntő tényezőnek minősül. Különben a sok hiányzó adatnak lehet oka az is, hogy néhány országban nem tettek fel bizonyos kérdéseket az interjú során, illetve az is előfordulhatott, hogy egy korábbi kérdés miatt értelmetlenné vált egy bizonyos kérdés feltétele. Például ha valaki nem volt kórházban, akkor nem kérdezik meg, hogy mennyi estét töltött ott az elmúlt egy évben, valamint nem dohányosoktól sem kérdezik meg, hogy mennyi cigarettát szívnak el naponta. Az ilyen jellegű változókat nem töröltem, hanem egyesítettem, ami jelentősen lecsökkentette a hiányzó adatok arányát. Ezután megfigyelésenként is megnéztem, hogy kinek mennyi adata hiányzik. Általában az volt jellemző, hogy ha nem volt teljes az adatfelvétel valakinél, akkor egy egész kérdéskör hiányzott nála. Ha több kérdéskörből is hiányoztak adatok, akkor nehezen lehet becsülni őket, ezért ha 20-nál több változó esetében nem álltak redelkezésre értékek, akkor töröltem az adott megfigyelést. Végül pedig a még megmaradt hiányzó értékeket az R-ben található „*mice*” 2011 package segítségével pótoltam, amellyel kategorikus és numerikus változók becslése is lehetséges.

4.1.3. Magyarázó változók

A 20 változócsoportban szereplő közel 2500 változóból, először kiválasztottam azt a körülbelül 50 darabot, amelyekkel a 2.2. alfejezetben lévő tanulmányokban találkozhattunk, illetve amiket kockázatértékelő lapokon is láthatunk. Ezáltal egyszerre vizsgálhatjuk, hogy milyen új változók kerülhetnének be a kockázatelbírálás folyamatába és milyen változókat válthatnának le. A változócsoportok ismertetése megtalálható az A függelékben. A kiválasztás során az is szempont volt, hogy a gyakorlatban mennyire tudna hasznos információval szolgálni, ha megkérdeznénk az ügyféltől. Többek között ezért is került be az elemzésbe a kognitív képessé-

geket felmérő kérdések közül például a memória-, számolás,- beszédfolytonosság-teszt, valamint olyan közvetlenül mérhető képességek mint a tüdőkapacitás és szorítóerő. A fő kockázati tényezőt jelentő krónikus betegségeket, panaszokat, korlátozottságokat egy külön táblázatban kezelem részletesen, ahol megtudható, hogy egy adott embernek milyen konkrét problémái voltak. Ezeket arra használom később, hogy megnézzem milyen mértékben javítják a használatuk a modellek előrejelző képességét. Fontos megjegyezni, hogy a modellek során nem veszem figyelembe a kizárást, vagyis hogy a biztosító nem feltétlen vállal át minden kockázatot, viszont utólag megvizsgálom, hogy a halálnak volt-e köze valamilyen korábbi betegséghez, amiről tudhatott volna a biztosító. Ehhez az elhalálozás körülményeiről szóló kérdéseket is kigyűjtöttem, amiből megtudhatjuk, hogy miben halt meg egy adott illető. Előzetesen azonban csak összesített adatokat használtam fel az elemzéshez, amiből az derül ki, hogy mennyi problémája volt az adott illetőnek az adott változótípuson belül. A kiválasztott változókat és azok leírását a B függelékben ismertetem.

4.1.4. Megfigyelések kiválasztása

Első lépésként le kellett szűrni azokat a megfigyeléseket, amelyek mind a 4. és 7. hullámban szerepeltek, mivel az is előfordult, hogy vagy csak a 4. vagy csak a 7. hullámban kérdeztek meg egy adott illetőt. Ezenfelül még azokat a megfigyeléseket is ki kellett választani, akik a 4. hullámban szerepeltek, de nem élték meg az 5. illetve 6. hullámokat. Erre azért volt szükség, mert azt szeretnénk modellezni, hogy a két hullám között eltelt időszakban kik haltak meg. Erre a minden adattáblában megtalálható „*mergeid*” változót használtam, amely személyhez kötött, vagyis a hullámok között nem változik. Az adatfelmérést a 4. hullámban az 1960-ig született embereken végezték el, azonban ahogy említettem a 2.2. alfejezetben is, a 60 évnél idősebb emberek kockázatainak a felmérése a célom, ezért az 1951 után született embereket töröltem a megfigyelések közül, mivel a 4. hullám adatfelvételére 2011-ben került sor. Ez azonban azt a fontos korlátozást jelenti a modelleket tekintve, hogy csak feltételesen fogalmazhatjuk meg az állításainkat. Pontosabban a mellett a feltétel mellett vizsgálom a populációt, hogy megélték-e a 60 éves kort. A 4.1.2. alfejezetben elvégzett módosítások és a 60 évnél fiatalabbak szűrése után 26333 megfigyelés maradt, amiből 5073 haláleset történt. Ezt az adattáblát elemzem a következő alfejezetben.

4.2. Feltáró elemzés

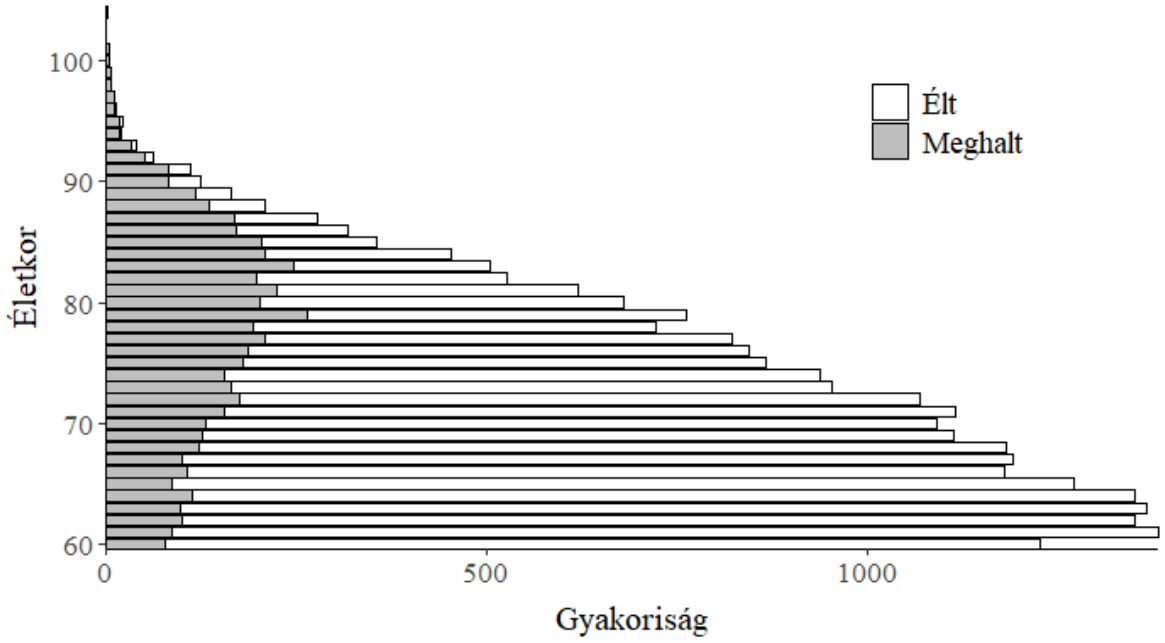
A modellek futtatása előtt a változók közötti kapcsolatokat vizsgálom adatvizualizációs eszközökkel és statisztikai tesztekkel, valamint új változókat vezetek be.

4.2.1. A változók és a halál közötti kapcsolatok

A dichotom célváltozó miatt csak asszociációs és vegyes kapcsolat merülhet fel, melyek statisztikai teszteléséhez a χ^2-tesztet, illetve a Wilcoxon-féle rangösszeg tesztet használhatjuk. Az előbbi nullhipotézise, hogy a változók függetlenek egymástól, az utóbbié pedig, hogy a megfigyelések azonos eloszlásból származnak.

Természetesnek tűnik a gondolat, hogy elsőként az életkor elhalálozásra gyakorolt hatását vizsgáljuk. Ezt szemléltetem egy halmozott sávdiagramon a 3. ábrán, ahol látható, hogy 97 éves kortól nincs élő, ami nem meglepő, hiszen tulajdonképpen azt vizsgálom, hogy a két hullám között eltelt 6 év alatt bekövetkezik-e a halál. Ebből is látszik, hogy az életkor mindenképpen szignifikáns változó lesz a modellekben.

3. ábra. Élők és meghaltak gyakorisága életkoronként

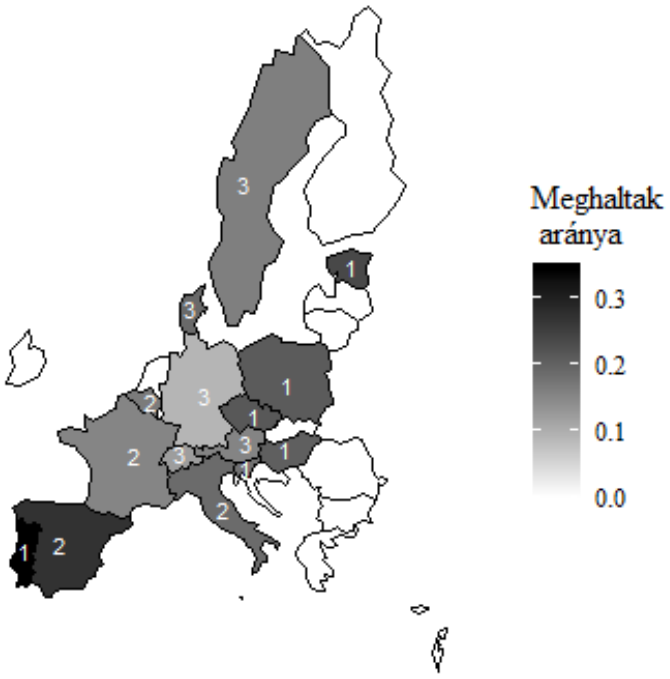


Forrás: Saját szerkesztés (R-ben)

A koronként bekövetkező halálesetek száma mellett más változóval való kapcsolatot is érdemes előzetesen vizsgálni A 4. ábrán az országonkénti halálozási arányok láthatók. A szürkített országok szerepeltek mind a 4. és 7. hullámokban, fehér háttérrel pedig azokat az

országokat ábrázoltam, ahol már végeztek SHARE felmérést, de nem vettek részt a 4. hullámban. Az eltérő halálozási arányokból látszik, hogy érdemes országspecifikusan is elvégezni az elemzéseket, vagy legalábbis csoportokba sorolni az országokat. Ezt a 2011-es egy főre jutó GDP alapján tettem meg, ami alapján az országokat három csoportra osztottam[7], amit az ábrán is feltüntettem. A csoportosításból látszik, hogy minél magasabb az egy főre jutó GDP, jellemzően annál alacsonyabb a halálozási ráta, így ez a változó is fontos lehet a nemzetközi piacon a kockázatelbírálás szempontjából. Valamint az elemzés szempontjából is fontos, mert a változó szerepeltetése nélkül más tulajdonságokkal próbálnánk magyarázni az országok közötti különbségeket. Továbbá az is látható, hogy Portugália halálozási rátája nagyon kiugró, 30% feletti, azonban ez részben annak is köszönhető, hogy a 7. hullámban rögzített adatokat még csak részben dolgozták fel Portugália esetében, ami miatt sok életben lévő ember nem szerepel még az adatbázisban.

4. ábra. Halálozási arányok a 4-es és 7-es hullám felmérései között



Forrás: Saját szerkesztés (R-ben), package: „*rworldmap*” 2011

Az elemzés szempontjából az is fontos választóvonal, hogy van-e az illetőnek életbiztosítása. A gyakorlatban megfigyelhető, hogy a biztosító állományának sokkal jobbak a várható életkilátásai, mint azt a statisztikai hivatalok által készített halandósági táblákból tudni lehetne, és ennek megfelelően sokszor kérdésként is szerepel a kockázatértékelő lapokon, hogy rendelkezik-e

[7] 1-es kategória: <30 000$; 2-es kategória: 30 000$ - 40 000$; 3-as kategória: 40 000$<

egyéb biztosítással az ügyfél.

Azonban közvetlenül nem tudtam szerepeltetni az elemzésben ezt a változót, mivel a megfigyelések között 8071 esetben hiányzott / nem tudták / nem akarták megmondani, hogy rendelkeznek-e életbiztosítással. Ezeket nem imputáltam a túl sok hiányzó adat miatt, hiszen abból lehet, hogy csak téves következtetéseket tudtam volna levonni. Viszont csoportokként kezelve őket[8] az látható a 3. táblázat alapján, hogy az életbiztosítással rendelkezők kivételével közel azonosak a halálozási ráták. Ez alapján akár egy csoportként is lehetne őket kezelni, ennek ellenére úgy láttam helyesnek, hogy megkülönböztetem azokat, akikről nem tudunk biztosat.

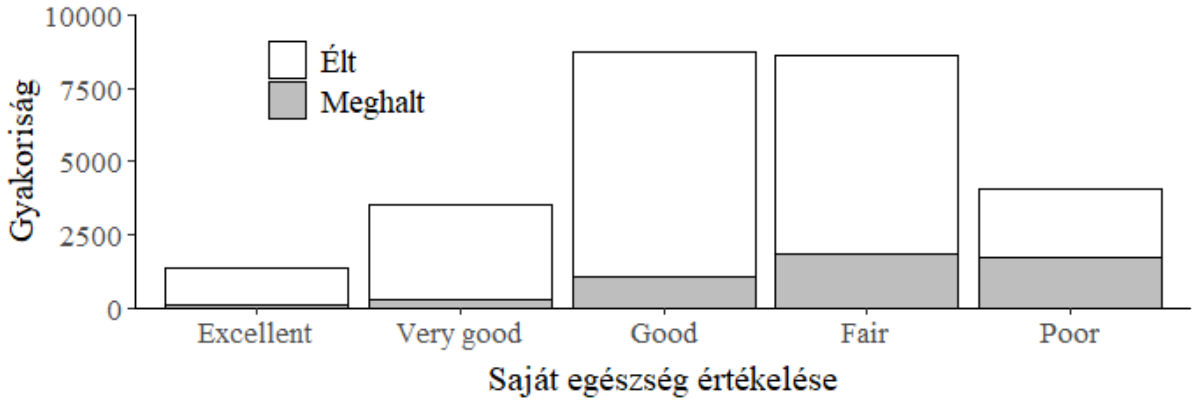
3. táblázat. Törölt megfigyelések halálozási aránya

Van-e életbiztosítása?	Élt	Meghalt	Halálozási arány
Igen	2496	282	10,2%
Nem	12359	3125	20,2%
Nem tudni	6405	1666	20,6%

Forrás: Saját készítés

Az új változók bevezetése után további kapcsolatokat mutatok be, amiknek szignifikáns lehet a magyarázó ereje. Elsőként a saját egészségi állapotról alkotott véleménnyel való kapcsolatot szemléltetem az 5. ábrán.

5. ábra. Saját egészségről alkotott vélemény és a halál kapcsolata



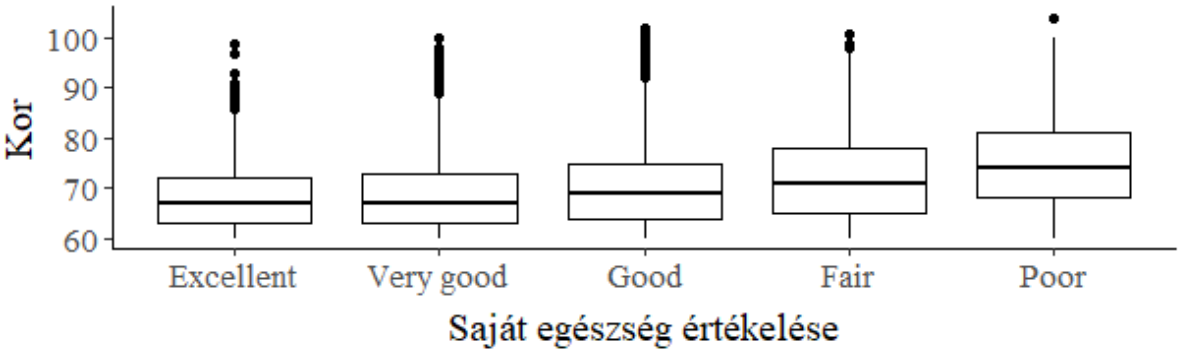
Forrás: Saját szerkesztés (R-ben)

Jól látható egyrészt, hogy az emberek jelentős része közepesre értékelte az egészségi állapotát (*Fair* vagy *Good*), valamint az is, hogy akik ettől egészségesebbnek / egészségtelenebbnek gondolták magukat, azok valóban sokkal kisebb / nagyobb arányban haláloztak el. Az oszlopdiagram alapján is látszik a változók közötti összefüggés, amit χ^2-teszttel is vizsgáltam. A

[8] van életbiztosítása, nincs életbiztosítása, nem lehet tudni

tesztstatisztika értéke 2164, ami alapján minden szignifikanciaszint mellett elutasítjuk a nullhipotézist, vagyis hogy a változók függetlenek. Bár gyakorlati használhatóság szempontjából kérdéses ennek a változónak a használata, hiszen valószínűleg nem sokan mondanák azt magukról kockázatelbírálás során, hogy nagyon rossz az egészségi állapotuk. Ezért feltehetően eltolódna az egészséges állapot felé a válaszok eloszlása, viszont indirekt kérdésként feltéve elképzelhető a használata, tehát nem hagyom ki az elemzésből. Továbbá a 6. ábráról látható, hogy az idősebbek gondolják rosszabbnak az egészségi állapotukat, vagyis a rosszabb állapotokban egyre nő a medián életkor, amit a Kruskal-Wallis teszt is alátámasztott. A tesztstatisztika értéke 1336, ami alapján a nullhipotézis (mediánok egyezősége) elutasításra kerül.

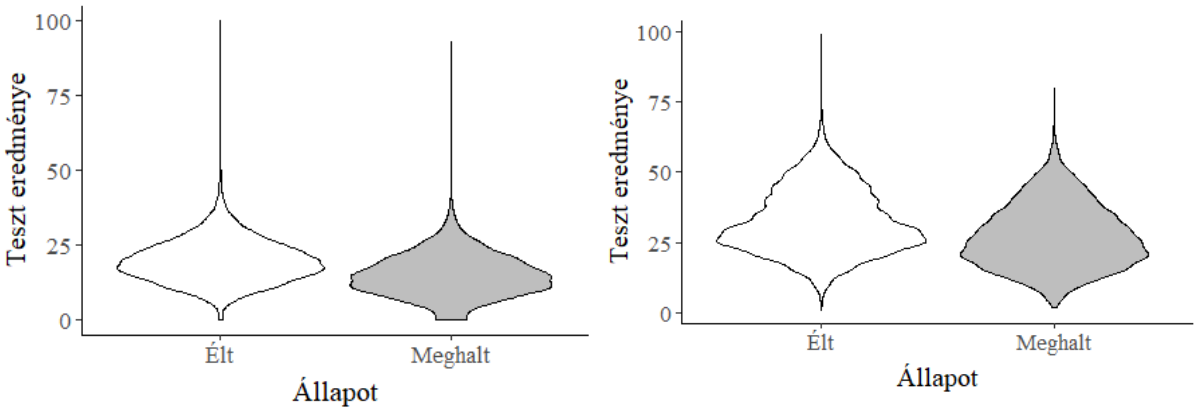
6. ábra. Saját egészségről alkotott vélemény és a kor kapcsolata



Forrás: Saját szerkesztés (R-ben)

A képességeket felmérő tesztekkel is megfigyelhető a kapcsolat, amit az eloszlások kirajzolásával szemléltetek a 7.a és a 7.b ábrákon.

7. ábra. Képességtesztekkel való kapcsolat



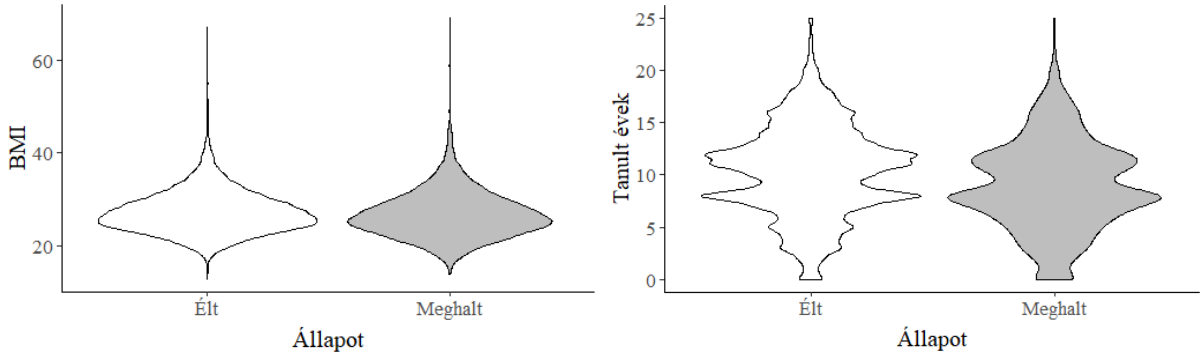
(a) Beszéd gördülékenység

(b) Szorítóerő

Forrás: Saját szerkesztés (R-ben)

Emellett a BMI-vel (8.a) és a tanulással töltött évek (8.b) számával is vizsgáltam a kapcsolatot. A Wilcoxon-féle rangösszegteszt alapján mind a négy esetben azt mondhatjuk, hogy nem azonos eloszlásból származnak a megfigyelések, mivel minden szokásos szignifikanciaszint mellett elutasítjuk a nullhipotézist. Ebből következik, hogy mindegyik változónak lehet magyarázó ereje a modellekben. Mint az látható az ábrákról, mindegyik esetben a várt irányban torzul az eloszlás az elhunytak esetében: a beszéd gördülékenysége, és a szorítóerő esetében kisebb a ferdeségi mutató értéke; BMI-vel való kapcsolatnál azt láthatjuk, hogy kicsit vastagabb mind a két széle az eloszlásnak, vagyis a súlyos sovány és súlyos elhízás többször fordult elő ebben a kategóriában; oktatásban töltött évek számánál pedig azt láthatjuk, hogy több ember nem fejezte be az általános iskolát, illetve kevesebb ember jutott el a magasabb tanulmányi szintekre.

8. ábra. Testalkattal és iskolázottsággal való kapcsolat



(a) Testtömegindex

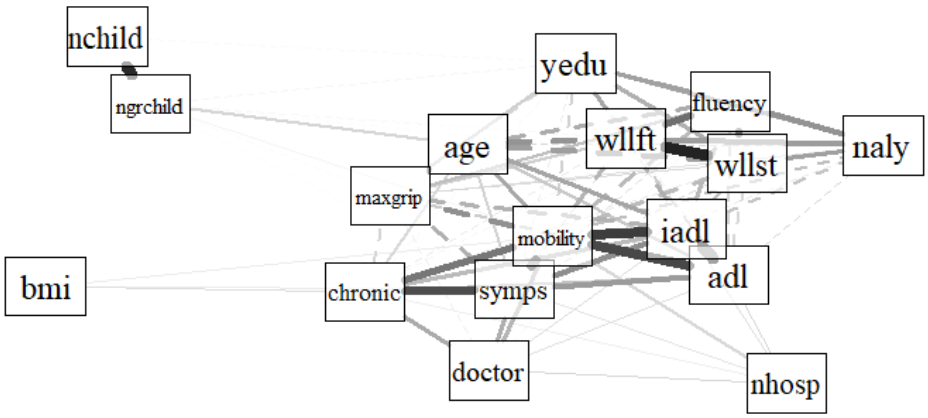
(b) Tanulmányok hossza

Forrás: Saját szerkesztés (R-ben)

4.2.2. Kapcsolat a magyarázó változók között

A magyarázó változók és célváltozó kapcsolatainak feltérképezése mellett, a magyarázó változók egymással vett kapcsolatából is fontos következtetéseket lehet leszűrni. A numerikus változók közötti szignifikáns korrelációt a 9. ábra mutatja, amin a maximális korreláció (legvastagabb, legsötétebb vonal) 0,7 körüli, míg a leghalványabb vonal 0,1-es korrelációt mutat. A folytonos vonal jelzi a kapcsolat pozitivitását, míg a szaggatott vonalak negatív kapcsolatot jeleznek. A változók elhelyezkedése és távolsága nem tartalmaz többletinformációt.

9. ábra. Numerikus változók közötti korreláció



Forrás: Saját szerkesztés (R-ben)

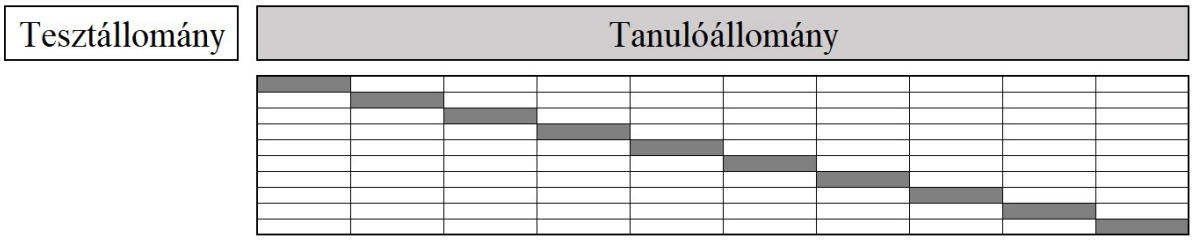
Rögtön látható, hogy a legerősebb korreláció a betegségekkel összefüggő változók (*chronic*), panaszok (*symps*), korlátozottságok (*mobility, adl, iadl*), orvosi látogatások (*doctor*) között van, amikkel nem meglepő módon az életkor (*age*) is közepes pozitív korrelációt mutat. Ez azt vetíti előre, hogy valószínűleg a modellekben nem lesz szükség minden ilyen változóra. Emellett egy másik pozitívan korreláló változócsoport is látható, amik a képességeket veszik figyelembe, mint például a memóriát (*wllft,wllst*) és a beszéd gördülékenységét (*fluency*). Ezek továbbá pozitívan korrelálnak az oktatásban eltöltött évek számával (*yedu*) és a szabadidős tevékenységek számával (*naly*) is. A két változócsoport között jellemzően közepesen erős negatív korreláció tapasztalható, vagyis minél több betegsége, panasza van valakinek, annál rosszabbul fog teljesíteni a képességeket felmérő teszteken. A változók kapcsolatát Reshef et al. [2011] által javasolt *Maximal Information Coefficient (MIC)* alapján is teszteltem, ami azon az ötleten alapszik, hogy ha van kapcsolat két változó között, akkor találhatunk olyan rácsfelosztást, ami körülhatárolja a ponthalmazt. A tetszőleges rácsfelosztásnak köszönhetően pedig nemlineáris kapcsolatok felfedezését is lehetővé teszi ez a mutató. Mindazonáltal az általam vizsgált változók között nem volt olyan jelentős kapcsolat észlelhető, amelyet a 9. ábra ne mutatott volna.

5. Modellezés

5.1. Modellszelekció folyamata

A modellek összehasonlíthatóságának érdekében olyan adatokon kell mérni a modellek teljesítményét, amelyek nem szerepeltek a betanítás folyamatában. Erre a legmegfelelőbb eszköz a keresztvalidáció, amely során k egyenlő részre osztjuk a megfigyeléseket, és mindig más részhalmazt választunk tesztállománynak. A megmaradt adatokon tanul a modell, melynek teljesítményét a kiválasztott részállományon mérjük. A következő lépésben egy másik részhalmazt választunk tesztállománynak, és szintén a megmaradt adatokon építjük fel a modellt. Ezt ismételjük meg k alkalommal, vagyis az összes részhalmazt egyszer választjuk ki tesztállománynak. Az eljárás előnye, hogy minden adat pontosan egyszer fog szerepelni tesztelő adatok között, amivel elkerülhető az egyetlen véletlenül kiválasztott tesztállományból bekövetkezhető torzítás. Ugyanis ha csak egy részhalmazon mérjük le a modell teljesítményét, akkor előfordulhat, hogy az átlagosnál jobb / rosszabb eredményeket kapunk attól függően, hogy könnyen / nehezen becsülhető egyedek kerültek be a tesztállományba. Azonban a kellően sok megfigyelésnek köszönhetően és az áttekinthetőséget szem előtt tartva csak egy tesztállományon vetettem össze a modellek teljesítményét, ahol az adatok 20%-át választottam ki véletlenszerűen tesztelésre. Keresztvalidációt csak a tanulóállományon belül alkalmaztam, a megfelelő paraméterek és változók kiválasztására a 10. ábrán szemléltetett módon 10 részre osztva. Így a paraméterek optimalizálása és a végső modellek tesztelése más részhalmazokon történik, ami elengedhetetlen a torzítatlan becslések szempontjából.

10. ábra. Adatállomány felosztása



Forrás: Saját szerkesztés (Excelben)

Vagyis az egyes paramétereket 10 diszjunkt részhalmazon validálom, majd ezekkel a paraméterekkel újrabecsülöm a modellt a teljes tanulóállományon, és végezetül a tesztállományon mérem le az eredményességét.

5.2. Változószelekció

A modellek futtatása előtt különböző változószelekciós eljárásokkal szűröm ki a legfontosabb változókat, amihez a teljes tanulóállományt használom. A szelekcióra azért van szükség, mert a célom a lehető legkevesebb adatból minél pontosabb előrejelző képességgel rendelkező modellek meghatározása, ugyanis ezáltal egyszerre csökkenhetnek a biztosítók és az ügyfelek terhei is.

5.2.1. Legjobb részhalmaz

Elsőként a legjobb részhalmaz (*best subset*) módszerrel vizsgáltam a legfontosabb változókat. Az eljárás során a bayesi információs kritérium minimalizálásával határoztam meg a változók fontosságát a $BIC = ln(n) \cdot k - 2 \cdot ln(\hat{L})$ képlet alapján, ahol n a megfigyelések száma, k a szabadságfok, $\hat{L}$ pedig a likelihood függvény értéke. Annak érdekében, hogy belátható időn belül lefusson az algoritmus, meghatároztam a maximum felhasználható változók számát, ugyanis 44 változó esetén 2^{44} féle modellt kellene lefuttatni. Az eljárás a következő változókat választotta, ahol a számok jelzik a kiválasztás sorrendjét.

1. age	4. sphus	7. nhosp	10. esmoked
2. iadl	5. phinact	8. wllft	11. naly
3. gender	6. maxgrip	9. BMI	12. ngrchild

5.2.2. Lépésenkénti szelekció

Mivel ezzel a módszerrel csak korlátozott számú változót tudtam leszűrni, ezért *backward-forward* változószelekciós eljárással is meghatároztam a szignifikáns változókat. Ehhez az Akaike információs kritériumot használtam, ami csak annyiban tér el a BIC-től, hogy a büntetőparaméter nem veszi figyelembe a megfigyelések számát ($AIC = 2 \cdot k - 2 \cdot ln(\hat{L})$), ezért várhatóan itt is szerepleni fognak az előbbi változók. Az így becsült eredményt a 11. táblázatból lehet látni, negatív együtthatókkal azokat a változókat, amelyek csökkentik a halál bekövetkeztének valószínűségét, pozitívval pedig azokat, amelyek növelik azt. A szelekció során az üres modellből indultam ki és lépésenként megvizsgáltam, hogy valamely változó használata vagy mellőzése milyen mértékben javítja a modellt. Ebből adódóan a változók sorrendje azt mutatja, hogy milyen sorrendben kerültek be a modellbe. A táblázat mellett jelölöm, hogy melyik változók együtthatóját tekinthetjük nullától különbözőnek a megszokott szignifikanciaszintek mellett (·: 10%; *: 5%; **: 1%; ***: 0,1%). Ezenfelül az együtthatók nagyságából azt is láthatjuk,

hogy melyik kategorikus változók játszanak döntő szerepet az osztályozás során, amiket félkövér betűtípussal jelzek.

4. táblázat. *Backward/Forward* regresszió az összes kiválasztott adatot felhasználva

Változó	**Együttható**	**Szignifikancia**
Konstans	-6,633	***
age	0,089	***
sphus_Very Good	0,214	
sphus_Good	**0,307**	*
sphus_Fair	**0,562**	***
sphus_Poor	**0,898**	***
iadl	0,140	***
gender_Female	**-1,016**	***
maxgrip	-0,023	***
phinact_Yes	**0,400**	***
nhosp	0,014	***
wllft	-0,075	***
GDPcat_2	**-0,438**	***
GDPcat_3	-0,188	**
naly	-0,094	***
esmoked_Yes	0,220	***
areabldgi_The suburb or outskirts of a big city	-0,218	*
areabldgi_A large town	-0,049	
areabldgi_A small town	0,003	
areabldgi_A rural area or village	-0,207	**
eat_No	0,222	***
doctor	0,011	***
symps	-0,061	***
nchild	-0,046	**
fluency	-0,014	***
mobility	0,038	**
drinking_Yes	0,187	**
bmi	-0,013	**
mstat_Registered partnership	0,410	·
mstat_Married, not living with spouse	0,192	
mstat_Never married	0,258	**
mstat_Divorced	0,276	**
mstat_Widowed	0,238	***
lifeins_Yes	**-0,292**	***
lifeins_No	-0,184	***
lifehap_Sometimes	-0,115	*
lifehap_Rarely	0,060	
lifehap_Never	-0,042	
gali_Limited	0,131	*
yedu	0,011	·
adl	0,046	·

Forrás: Saját készítés

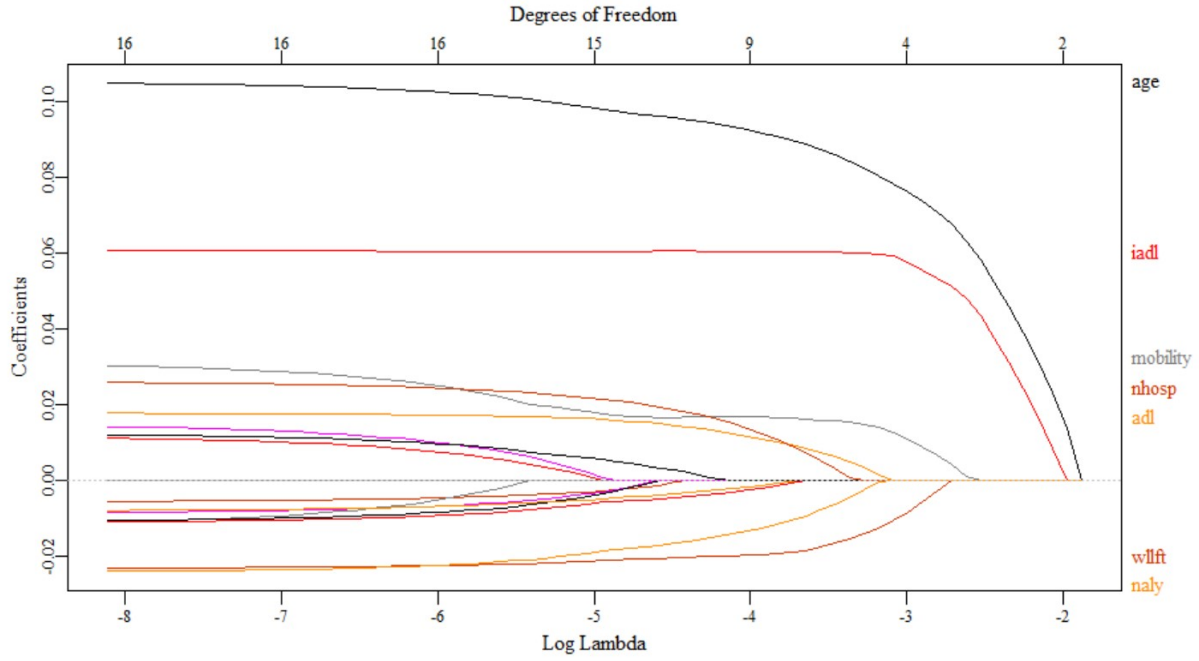
A szelekció során más változók is beválasztásra kerültek, azonban nem tüntettem fel őket, mert nem voltak szignifikánsak. Ilyenek voltak a jelenlegi foglalkozás (*cjs*), hallás (*hearing*), tüdőkapacitás (*lung*) és hogy milyen kezes az illető (*hand*). A táblázatból látható, hogy a

legjelentősebb kategorikus változók a nem (*gender*), a saját egészség rossznak történő besorolása (*sphus*), a fizikai inaktivitás (*phinact*), a gazdasági környezet (*GDPcat*) valamint az életbiztosítás birtoklása (*lifeins*). Ilyen kérdésekkel találkozhatunk előzetes online kockázatfelmérések során is, ahogy azt a 2.3. alfejezetben is említettem. Vagyis ezek a kérdések valóban jó szűrőként szolgálnak. A szignifikáns változók között szerepelnek továbbá a 2.2. alfejezetben említett tényezők is, melyek az egyén élethelyzetére, tevékenységeire és képességeire kérdeznek rá.

5.2.3. Zsugorító módszerek

Végül a numerikus változókat egy másfajta szelektálási módszernek is alávetettem, amihez a változók sztenderdizálására volt szükség. A zsugorító módszerek jellemzője, hogy nem kizárják a változókat, hanem az együtthatóikat csökkentik le egy λ büntetőparaméter függvényében, aminek köszönhetően robusztusabb modelleket kapunk. *Ridge* regresszió esetében $\lambda \cdot \sum_{j=1}^{p} \beta_j^2$, *Lasso* regresszió esetében pedig $\lambda \cdot \sum_{j=1}^{p} |\beta_j|$ a büntetőtag, ahol β jelöli az együtthatókat, p pedig a változók számát. Az optimális λ-t próbálgatással és keresztvalidációval keresi meg a program, ami egyben azt is megmondja, hogy melyik változókat érdemes bent hagyni a modellben. A 11. ábrán látható a λ paraméter hatása az együtthatók nagyságára *Lasso* regresszió esetén.

11. ábra. *Lasso* regresszió együtthatóinak változása λ függvényében



Forrás: Saját szerkesztés (R-ben)

Ez alapján a kiválasztott változók fontosságára is lehet következtetni. Egyrészt a nagy együttható, másrészt pedig a nagy büntetőparaméter melletti bekerülés is a változók fontosságát jelzik. Az ábráról látható, hogy melyik változók a fontosak: *age, iadl, mobility, nhosp, adl, wllft, naly*. *Ridge* regresszió esetén valóban csak zsugorítás történik, nem nullázódnak ki az együtthatók, azonban *Lasso* esetében el is tűnnek a változók λ növelésével. Így a *Lasso* regresszió esetében az optimális λ meghatározásával is tudunk változókat szűrni. Az optimális λ mellett még az előbb említett 7 változó mellett további 3 változó (*wllst, fluency, doctor*) szerepeltetését találta szükségesnek az eljárás, azonban ezek együtthatója csupán tized akkora. További érdekesség, hogy a *chronic* változó egyedüliként, az ábrán feltüntetett legkisebb λ mellett sem kerül be a modellbe, ami valószínűleg annak köszönhető, hogy ma már a krónikus betegségekben szenvedők is sok esetben teljes életet élhetnek.

Összességében elmondható, hogy a különböző módszerek hasonló változókat találtak fontosnak, amik elsősorban alapvető adatokra, egészségre és aktivitásra kérdeznek rá: *age*, *sphus*, *gender*, *iadl*, *phinact*, *nhosp*, *wllft*, *esmoked*, *naly*. Azok a változók, amelyek semmilyen eljárás mellett sem kerültek kiválasztásra, valószínűleg felesleges szerepeltetni a modellekben, mert nem tud annyi többletinformációval szolgálni, mint amennyit azok megkérdezése jelent. Ide tartozik az előbb említett *chronic* és különböző képességeket mérő tesztek: *writing*, *reading*, *numeracy*, *numeracy2*, *orienti*, *memory*, *eyesightr*. Érdemes megjegyezni, hogy ezek a változók nem függetlenek egymástól (χ^2-tesztet minden lehetséges párosításra elutasíthatjuk), valamint az is fontos közös vonásuk, hogy mindegyik ordinális skálán mérhető. Ebből arra lehet következtetni, hogy az adatok ilyesfajta rögzítése nem bizonyul jó módszernek előrejelzés szempontjából. Mindamellett volt néhány olyan változó is, amelyek fontosságáról az egyes eljárások mást mondtak[9], így a modellezés során érdemes kísérletezni ezek szerepeltetésével, hogy mennyi plusz információt tudnak nyújtani. Éppen ezért a következő alfejezetben nem csak a különböző algoritmusok összehasonlítása a célom, hanem az is, hogy az egyes algoritmusok milyen adatokat felhasználva tudnak a legpontosabb előrejelzést adni. Ennek teszteléséhez három különböző adathalmazon is mérem a teljesítményüket:

1. csak azokkal a változókkal, amik kiemelkedően fontosak a változószelekció szerint;
2. ezeket kiegészítve és keresztvalidálva azokkal, amik csak részben tűntek fontosnak;
3. a 4.1.3. alfejezetben említett részletes változókkal.

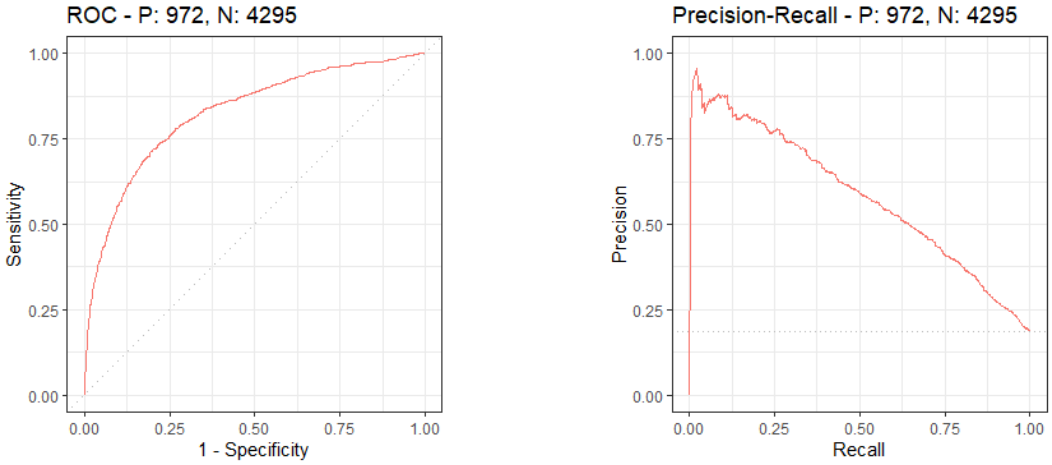
[9]A változók módszerek szerinti fontossága megtekinthető a B függelék 11. táblázatában.

5.3. Modellek

5.3.1. Logisztikus regresszió

Elsőként a logisztikus regressziónak vetettem alá az adatokat, kezdve a legfontosabb változók csoportjával. A tesztadatokon kiértékelt modell eredménye a 12. ábrán látható.

12. ábra. ROC és PR görbék ábrája logisztikus regresszió esetén



Forrás: Saját szerkesztés (R-ben)

A modell AUC értéke 0,827, ami a becslési képességet tekintve jónak mondható, azonban a 3.4. alfejezetben is említett AUCPR érték jobban leírja a probléma sajátosságait. Ennek értéke 0.582. Az ábra konkrét értelmezését és használhatóságát az 5. táblázat alapján mutatom be, ahol különböző vágási értékek mellett ismertetem a PR görbe pontjait.

5. táblázat. Logit modell klasszifikációs táblája különböző vágási értékek mellett

Megfigyelt \ Becsült	Élt	Meghalt	Élt	Meghalt	Élt	Meghalt
Élt	4205	90	4294	1	2162	2133
Meghalt	703	269	949	23	111	861
Vágási érték	58,5%		93,11%		9%	

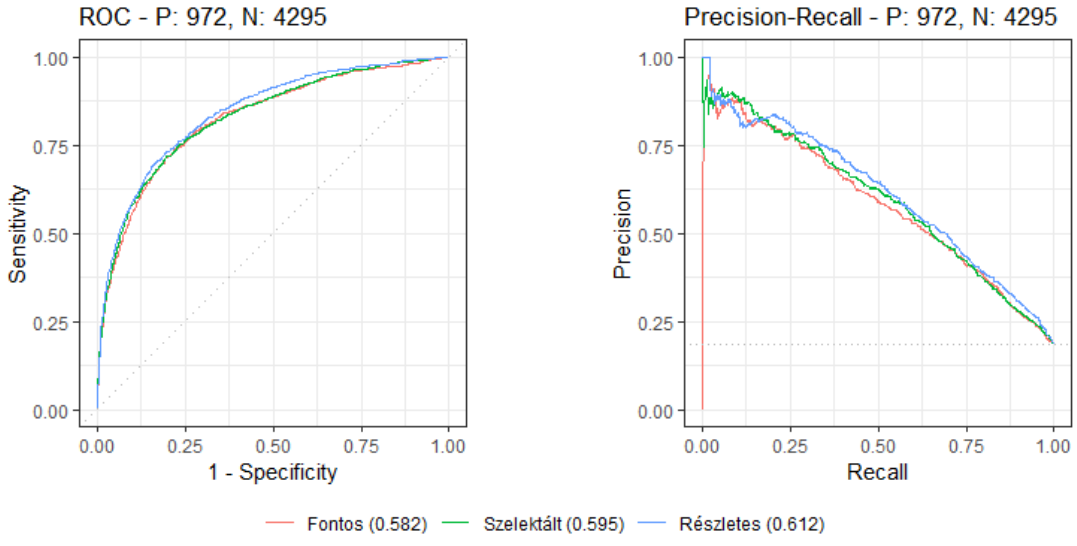
Forrás: Saját készítés

A példákon keresztül be lehet mutatni az előzetes szűrőként való felhasználási lehetőségeket, melynek vizsgálatát tűztem ki célul a dolgozatomban. 58,5%-os vágási érték mellett a biztosító úgy tudott volna elutasítani hagyományos kockázatelbírálás nélkül 359 ügyfelet, hogy közülük 75%, vagyis 269 ember valóban meg is halt. Továbbá az is látható, hogy a rossz ügyfelek 28%-át tudja beazonosítani a modell, ami összességében meglehetősen jónak tűnik, tekintve hogy csak a néhány alapvető változót használtam fel a becsléshez. Ezeknek a százalékoknak az aránya rajzolódik ki PR görbén, melyen a vágási érték megválasztásával lehet „mozogni”, attól függően,

hogy mennyi rossz ügyféltől és milyen hibázási arány mellett szeretnénk megszabadulni. Ha például szinte biztosra akarunk menni, hogy csak rossz ügyfelet utasítunk el, akkor érdemes magas vágási értéket meghatározni. Jelen esetben nem fordult elő, hogy csak helyesen utasítanánk el ügyfeleket, de például 93,11%-os vágási érték mellett, ahol felveszi a görbe a maximumát, mindösszesen 1 embert jósol tévesen halottnak a modell, míg 23 másikat helyesen. Alacsony vágási értéket választva pedig el lehet érni, hogy szinte minden kockázatos ügyfelet megtaláljon a modell, azonban csak azon az áron, hogy sok jó ügyfelet is elutasítunk. 9%-os vágási érték mellett azt láthatjuk, hogy a jó ügyfelek majdnem felét elutasítjuk, de ugyanakkor a megmaradt állományban jelentősen lecsökkent a kockázatos ügyfelek aránya, ami a kezdeti 18,5%-os halálozási rátát 4,9%-ra módosította. Az optimális vágási érték meghatározása fontos lehet a biztosítóknak profitabilitás szempontjából, azonban ahogy a 3.4.1. alfejezetben is írtam, ennek vizsgálata a biztosítók éles adatain múlik.

Következő lépésben azt vizsgáltam, hogy mekkora javulás érhető el, ha ezeket a változókat kiegészítem a B függelékben ismertetett változókkal. Ehhez keresztvalidációt alkalmaztam az 5.1. alfejezetben leírt módon. A legjobb modellt az alapján választottam ki, hogy milyen átlagos AUCPR értéket tudnak elérni a keresztvalidáció során. Tehát tulajdonképpen lépésenkénti változószelekciót hajtottam végre manuálisan, tekintve hogy a beépített R függvények AUCPR érték szerint nem tudnak szelektálni. Végül azzal az állománnyal is elvégeztem a vizsgálatot, amiben részletesen szerepel, hogy kinek milyen betegségei, panaszai voltak korábban. A modellek által generált görbéket a 13. ábrán hasonlítom össze, zárójelben az AUCPR értékekkel.

13. ábra. Különböző változók által generált ROC és PR görbék



Forrás: Saját szerkesztés (R-ben)

Az AUCPR értékeket nézve, csak minimális különbséget láthatunk, azonban az ábrán látszik, hogy szinte minden vágási érték mellett a részletes változókkal futtatott modell eredményezi a legjobb arányokat. Ennek az eltérésnek a hatását mutatom be a 6. táblázatban, ahol 58,5%-os vágási érték mellett nézem meg az egyes modellek klasszifikációs tábláját.

6. táblázat. Változók használatának hatása a klasszifikációs táblára

Megfigyelt \ Becsült	Élt	Meghalt	Élt	Meghalt	Élt	Meghalt
Élt	4205	90	4191	104	4187	108
Meghalt	703	269	674	298	645	327
Változók	Fontos		Szelektált		Részletes	

Forrás: Saját készítés

Ebben a speciális esetben a Szelektált változók modellje 29 új ügyfelet tudott helyesen beazonosítani az eredeti modellhez („Fontos") képest, viszont 14 ügyfél esetében helytelenül tette ezt. Így összességében csökkent is a „*Precision*" mutató értéke. Részletes változók modellje esetén ez az érték nőtt, csakúgy mint a „*Recall*". Fontos azonban látni, hogy egy bizonyos vágási érték mellett nem azonos x- vagy y tengely menti érték szerint hasonlítjuk össze a görbéket, mert egyszerre változik mind a kettő arány. Az is látható továbbá, hogy magas vágási értékek mellett nagyon szóródnak a görbék értékei, ezért ezen a szakaszon nem érdemes őket vizsgálni. A vágási érték csökkentésével azonban egyre jobban kirajzolódik, ha valamelyik görbe dominálja a másikat. Például 25,99%-os vágási érték mellett a Szelektált és a Részletes változók modellje esetében is csak helyesen történt átsorolás a Fontoshoz képest. A kérdés azonban az, hogy ez az extra pontosság megéri-e, hogy jelentősen több a kérdések mennyisége. Fontos esetén 9, Szelektált esetén 27, míg Részletes esetén 130 változó alapján készült az előrejelzés, ami alapján jogosan merülhet fel ez a kérdés. Ezért a részletes változók esetén is végeztem változószelekciót, hogy lássam milyen változók bizonyulnak a legfontosabbnak. Az első ilyen változó, ami bekerült a modellbe a meleg étel elkészítésével kapcsolatos nehézségek voltak, vagyis jelentősen növeli a kockázatot, ha valaki ilyen szintű napi teendőket nem tud ellátni. Ezenkívül még bekerült a rákos megbetegedés, a cukorbetegség, ízületi panaszok jelenléte és különböző gyógyszerek szedése. Közvetett módon ezek a változók szerepeltek a szűkebb modellekben is összesített formában, viszont ezek közül csak az „*iadl*" került is be a fontos változók közé. Tehát mérlegelni kell, hogy a részletesebb információ az ügyfél állapotáról mennyivel javítja a modellt, aminek meghatározásához biztosítói adatállomány lenne szükséges. A következő alfejezetben fa alapú modellek segítségével nézem meg a változók előrejelző képességét.

5.3.2. Döntési fa / véletlen erdő

Ezeknek a modelleknek a keretein belül is hasonlóan jártam el, mint logisztikus regresszió esetében. Vagyis megvizsgáltam a 3 változócsoport által generált fáknak a becslési pontosságát az AUCPR értékek alapján, amihez minden esetben keresztvalidáció által meghatározott paramétereket használtam. A keresztvalidáció során mindig a tanulóállományon belül létrehozott 10 tesztalmazon elért átlagos AUCPR érték alapján döntöttem a paraméterek megválasztásáról. Döntési fánál az összes változó használata mellett figyelembe vettem a „*complexity parameter*" (cp) értékét, a leveleken és ágakon minimálisan előforduló megfigyelések számát, és a fa mélységét. Elsőként alapbeállítás mellett az optimális cp értéket kerestem meg, majd ennek használatával kísérleteztem a többi paraméterrel. A módszer nem alkalmas minden paraméter-kombináció leellenőrzésére, de egy átfogó képet kaphatunk arról, hogy milyen irányba érdemes azokat változtatni, amit újrahangolással tovább lehet javítani. Véletlen erdő esetében pedig a fák- és a felhasznált változók számával, valamint szintén az ágokan szereplő minimális egyedszámmal és a fa mélységével optimalizáltam a modelleket. Annak ellenére, hogy a döntési fa sok elágazás esetén erősen hajlamos a túltanulásra, a paraméterek finomhangolása során jellemzően nagyon komplex fákat kaptam vissza. Az egyes esetekben megválasztott paramétereket foglalom össze a 7. táblázatban.

7. táblázat. Döntési fák és véletlen erdők paraméterei és teljesítményük

Döntési fa	változók száma	cp	max mélység	ágak min egyedszáma	átlagos AUCPR
Fontos	9	0,002	9	30	0,546
Szelektált	43	0,0001	9	70	0,538
Részletes	130	0,0004	7	40	0,529
Véletlen erdő	**véletlen változók**	**fák száma**	**max mélység**	**ágak min egyedszáma**	**átlagos AUCPR**
Fontos	3	199	4	40	0,576
Szelektált	6	199	4	40	0,582
Részletes	30	199	3	80	0,555

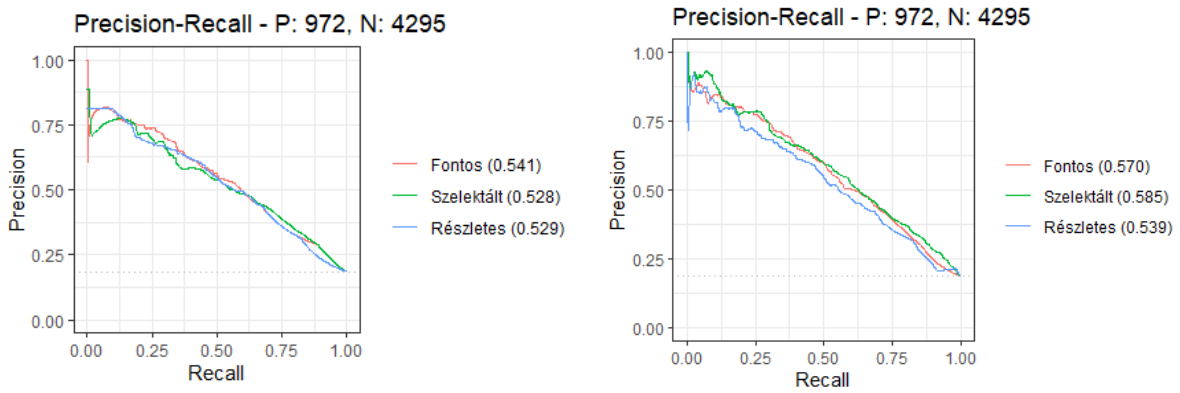
Forrás: Saját készítés

A táblázat alapján látható, hogy minél több változó áll rendelkezésre döntési fa esetén, annál inkább csökken a modell teljesítménye új adatokon, vagyis hajlamos a túltanulásra. A teljesítménybeli különbségek minimálisak, de észlelhetőek. Mindegyik esetben nagyon kicsi cp érték eredményezte a legjobb becslést, ami rengeteg elágazáshoz vezetett. A többi paraméter ezt a hatást tudta korrigálni azáltal, hogy több egyedet kívánt meg az egyes ágakon vagy kisebb faméretet állított be az alapbeállításhoz képest. Ezeknek a paramétereknek a finomhangolásával azonban csak minimális javulást lehetett elérni, viszont jelentősen lehetett csökkenteni a futási időt. Véletlen erdő esetén még jelentősebb volt a túltanulás mértéke, azonban ennek ellenére is

jobb eredményeket ért el a keresztvalidáció során. Például a legjobban teljesítő („Szelektált”) modell esetében a teszt- és tanulóállomány közötti eltérés az AUCPR értékben, több mint 0,2 volt. Itt is elmondható, hogy leginkább a fák- és a használt változók számával lehetett lényeges javulást elérni. A fák számát tekintve azt láttam, hogy 200-on felül már minimális volt ez a változás, ezért mindenhol 199-et használtam. További hasonlóság a döntési fa modellel, hogy a legtöbb változó mellett („Részletes”) született a leggyengébb modell, vagyis nem feltétlen előnyös, ha sok információ áll rendelkezésre ezeknél a modelleknél.

Végül a 14. ábrán összevetettem a döntési fa és a véletlen erdő modellek teljesítményét a tesztállományon, ahol zárójelben jelzem a megfelelő AUCPR értékeket. A bal oldali ábráról látható, hogy a szűkebb modell tudott elérni nagyobb pontosságot, azonban teljesítményben még ez is elmarad a logisztikus regresszió során tapasztalt pontosságtól. A jobb oldali ábrán szereplő véletlen erdő modellek jobban meg tudták közelíteni ezt a szintet, de ezek sem érték el.

14. ábra. Döntési fa és véletlen erdő által generált PR görbék



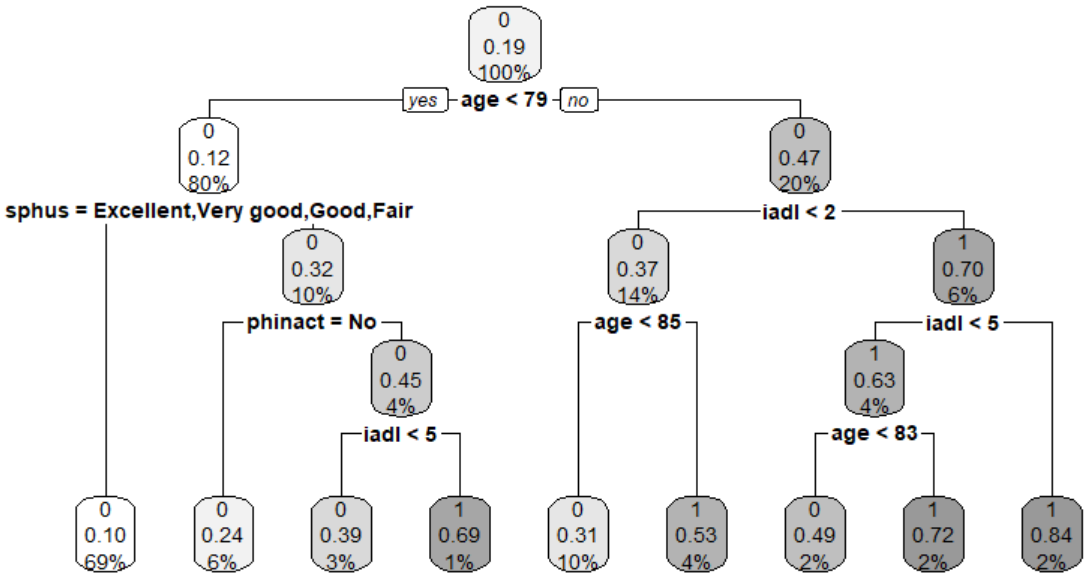
(a) Döntési fa

(b) Véletlen erdő

Forrás: Saját szerkesztés (R-ben)

A túltanulás ellenére azonban a döntési fa használata mellett szóló érv, hogy könnyű interpretálni az eredményt. Ezt a 15. ábrán mutatom be a legszűkebb modellt használva, a fa mélységét 4-re állítva. Ezen a leegyszerűsített példán jól látható, hogy azokat tekinti a modell legnagyobb arányban élőnek, akik 79 évesnél fiatalabbak, és nem a legrosszabb véleménnyel voltak saját egészségi állapotukról. Hasonlóan elmondható, hogy a legrizikósabb csoportba azok tartoznak bele, akik 79 évesnél idősebbek, és több mint 5 mindennapi teendő során tapasztalnak nehézségeket.

15. ábra. Döntési fa elégazásai



Forrás: Saját szerkesztés (R-ben)

5.3.3. Neurális hálók

Harmadik algoritmusként a neurális hálókat választottam. Ennél a módszernél szintén szükséges a paraméterek finomhangolása, illetve a változók normalizálása is segíti az algoritmust, ahogy azt a 3.3. alfejezetben említettem, ezért ebben az esetben a $[0,1]$ intervallumra transzformáltam minden magyarázó változót. Emellett még arra is szükség volt, hogy a kategorikus változók esetén *dummy* változókat hozzak létre a változón belüli szintek szerint. Továbbá a hosszú futási idő miatt kénytelen voltam a tanuló adatok véletlenül kiválasztott 20%-án tanítani a modelljeimet. Próbálgatás útján azt láttam, hogy a rétegek számának növelésével nem lehet elérni jelentős javulást, ezért csak 1 réteget alkalmaztam modelljeimben. Illetve az is megfigyelhető volt, hogy jelentősen túltanult a modell több neuron alkalmazásával, ezért a paraméterek keresztvalidálása során elsősorban a neuronok optimális számát kerestem egyetlen rejtett rétegen belül. Az eredményeket a 8. táblázatban foglalom össze.

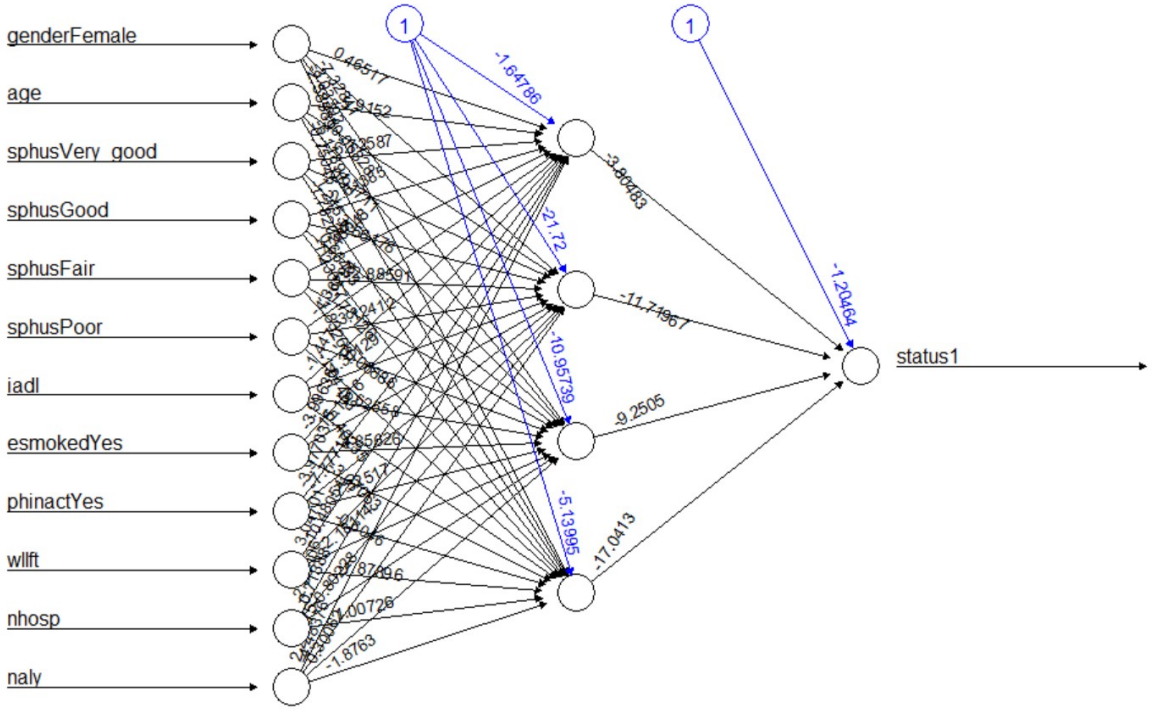
8. táblázat. Neurális hálók paraméterei és teljesítményük

Változócsoport	változók száma	rejtett rétegek száma	neuronok száma	AUCPR tesztállományon
Fontos	9	1	4	0,589
Szelektált	43	1	1	0,597
Részletes	130	1	2	0,563

Forrás: Saját készítés

Az eredményekből látható, hogy ezzel a módszerrel sem sikerült elérni jelentős javulást a becslési pontosságban. A legfontosabb változókra lefuttatva 4 neuron alkalmazása adta vissza a legpontosabb becslést, de az sem jelentett jelentős javulást a sima logisztikus regresszióhoz képest. A hálózatot a 16. ábrán szemléltetem, ahol a nyilakon szereplő számok a súlyokat, a csúcspontokban lévő számok pedig a torzító paramétereket jelentik.

16. ábra. Fontos változókra kapott neurális háló



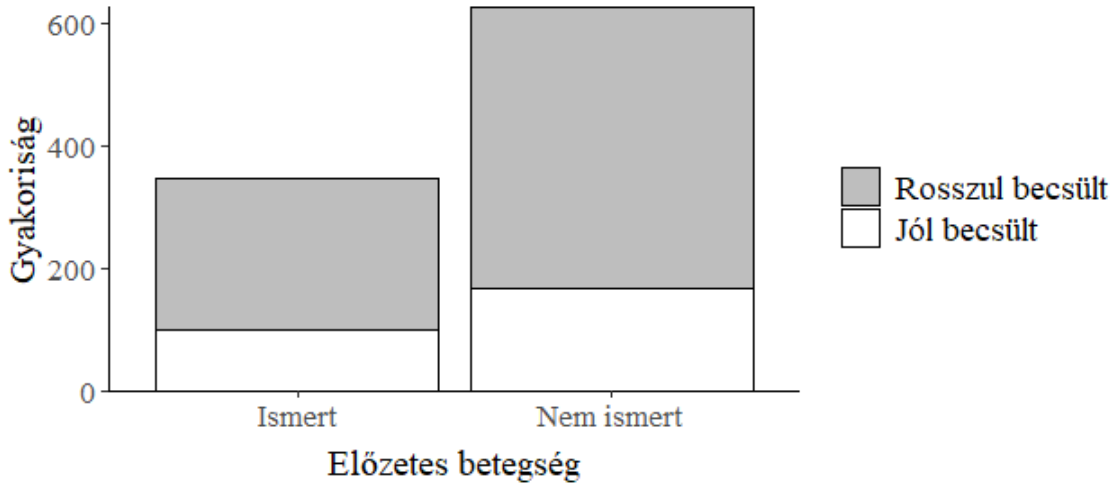
Forrás: Saját szerkesztés (R-ben)

A szelektált változók esetén azzal szembesültem, hogy csupán 1 neuron alkalmazása vezetett a legjobb becsléshez, amivel tulajdonképpen a logisztikus regresszióhoz nagyon hasonló algoritmust kaptam vissza, hiszen csak annyiban tér el, hogy a kimeneti változóhoz hozzáadunk egy torzító paramétert és azt újra transzformáljuk egy megfelelő függvénnyel. Látható is, hogy mindössze 0,002-es eltérés van a két módszer AUCPR értéke között. A részletes változók használata során pedig azt tapasztaltam, hogy elmaradt a teljesítménye a logisztikus regresszióétól, ami szerintem annak köszönhető, hogy ilyen sok változó esetén sokkal nehezebben találja meg az algoritmus a megfelelő súlyokat a véletlen próbálkozásoknak köszönhetően. Ezért ebben az esetben nem volt jól alkalmazható az algoritmus. Azonban nagyobb számítási kapacitás használatával, és több idő igénybevételével tovább lehetett volna javítani a becslés pontosságát, és valószínűleg bonyolultabb összefüggéseket is fel lehetett volna fedezni további rétegek és neuronok alkalmazásával mindhárom változócsoport esetében.

5.4. Kizárások vizsgálata

Ahogy azt említettem az 3.4.1. alfejezetben a kockázatelbírálás szerepe fontos a kizáró tényezők meghatározása szempontjából, ezért megvizsgálom, hogy milyen arányban haltak meg olyan okokból az emberek, amelyekről már korábban is tudni lehetett. Ehhez a fontos változók használatával számított logisztikus regresszió becslését használom. Az 5. táblázatból már láthattuk, hogy milyen besorolási pontossággal rendelkezik a modell 58,5%-os vágási érték mellett. A 17. ábrán ezt az esetet vizsgálom meg, vagyis megnézem mennyi ember halt meg olyan betegségben, ami szerepelt a kórtörténetében.

17. ábra. Előzetes betegségek vizsgálata



Forrás: Saját szerkesztés (R-ben)

Látható, hogy összességében a tesztállományon belül a halálesetek körülbelül harmadában volt összefüggésben a halál egy korábbi betegséggel. 348 esetben volt ismert és 624 esetben nem. Továbbá azt is meg lehet állapítani, hogy a modell nem tudja szignifikánsan jobban előrejelezni egyik csoportban sem, hogy kik fognak meghalni, csupán 2,5%-kal nagyobb az arány az ismert kategóriában. Ez kicsit ellentmond a várakozásaimnak, hiszen akiknek haláluk nem volt összeköthető korábbi betegséggel, ott nagyobb szerepe van a véletlen hatásnak, mint például baleset vagy fertőzés okozta halál esetén. Az arány minimálisan javul a várt irányba a vágási érték csökkentésével, mivel már csak azokat becsli élőnek a modell, akiknek minden értékük jó, így nem is volt korábbi betegségük. A legtöbben és a legnagyobb arányban szív- és érrendszeri betegségekben haltak meg úgy, hogy a betegségről már korábban tudni lehetett, ezért kiemelten fontosnak látszik ennek a betegségnek a felderítése a kockázatelbírálás során.

6. Összefoglalás

Dolgozatomban az életbiztosítások során történő adatalapú kockázatelbírálással foglalkoztam. Először áttekintettem, hogy milyen pozitív hatásai lehetnek, illetve hogy milyen problémák merülhetnek fel a folyamat automatizálásából adódóan. Ehhez áttekintettem a témában íródott cikkeket és ismertettem azok eredményeit, amikből arra lehetett következtetni, hogy nem feltétlenül a teljesen automatizált folyamatoké a jövő.

Különösképp fontosnak tartottam azon korosztály kockázatainak a vizsgálatát, akik már hagyományos módon nem juthatnak biztosításhoz. Ezért a 60 évesnél idősebb emberek kockázataival foglalkoztam, illetve ezen kockázatok felmérésének lehetőségeit tekintettem át. Néhány országtól eltekintve egyáltalán nincs kifejezetten idősekre szabott kockázatfelmérés, így a terület vizsgálatával fontos piaci rést lehet megszüntetni, hiszen a demográfiai tendenciák miatt egyre többen szeretnének idősebb korukban biztosítást kötni.

A klasszifikációs algoritmusok elméletének áttekintése után a SHARE adatbázis adatait felhasználva készítettem el a modelljeimet. Az adatbázis kiválóan szolgálta a dolgozat kérdéseinek megválaszolását, hiszen néhány évente ugyanazokon a személyeken végzik el a felmérést Európa-szerte, ami lehetővé teszi az egyén életútjának követését. Az interjún feltett kérdések köre kellően széleskörű ahhoz, hogy lefedje a jelenlegi kockázatelbírálási kérdéseket és lehetőséget adjon új fontos változók felfedezésére.

Először a rendelkezésre álló adatok alapján változószelekciós eljárásokkal vizsgáltam, hogy mely változóknak van a legnagyobb hatásuk a halál bekövetkeztének szempontjából. A legfontosabb változók között szerepeltek a kockázatelbíráló lapokról ismerős változók, mint például a kor, nem, kórházi látogatások száma, dohányzási szokások; de ezeken kívül új változók is fontosnak bizonyultak az idősebb korosztálynál: szabadidős tevékenységek, mindennapi tevékenységekben való korlátozottságok és képességfelmérő tesztek is. Ezek az eredmények összhangban vannak, illetve alátámasztják a már alkalmazott időskori kockázatelbírálás során fontosnak tartott kérdéseket. Ezenkívül még két bővebb változócsoportot hoztam létre, majd összehasonlítottam teljesítményüket a különböző algoritmusok keretein belül. Ehhez a biztosító számára is releváns mutatószámot, az AUCPR értéket használtam. Ez a mutató kiküszöböli a ROC görbe alatti terület mutatójának azt a hibáját, hogy nem fektet kellő hangsúlyt a pozitív esetek megtalálására, azáltal hogy figyelembe veszi a helyesen és tévesen pozitívnak becsültek arányát.

A modellezés során nem volt megfigyelhető jelentős különbség a változócsoportok teljesítményei között, vagyis nem feltétlen az az optimális, hogy minden lehetséges adatot bekér a biztosító az ügyféltől, de ez algoritmusonként változott. Fa alapú modellek és neurális hálók esetében például túl sok adatnak köszönhetően túltanult a modell, és így kevésbé adott pontosabb előrejelzést. Logisztikus regresszió esetén pedig csak minimális javulás volt megfigyelhető több változó használata esetén. Ezeknél a modelleknél az is fontos szempont lehet, hogy vissza tudjuk követni, hogy mi alapján történik a besorolás, hiszen nem feltétlenül tudnának az emberek megbízni egy olyan rendszerben, aminek folyamatába nem látnak bele. Ezeket összevetve a logisztikus regressziót látom a legesélyesebb algoritmusnak, amelyet a biztosító használhatna, mivel kellően pontos és áttekinthető a működése.

Gyakorlati szempontból fontos tényező még a kizárások detektálása, ezért utólag megvizsgáltam, hogy a tesztállományon belül meghalt emberek milyen arányban haltak meg olyan betegség következtében, amiknek voltak már korábban jelei. A várakozásaimmal ellentétben nem volt szignifikáns különbség a becslés pontosságában attól függően, hogy ismert vagy nem ismert betegségben hunyt el az adott illető. Azonban az adatokból lehetett látni, hogy a legtöbb ismert betegségből fakadó halál szív- és érrendszeri megbetegedés következményeként történt, ezért szerepeltetése indokolt lehet a kockázatelbíráló programokban.

Összességében izgalmas és releváns témának gondolom az automatizált kockázatelbírálás vizsgálatát, amit alátámasztanak a biztosítási piacon megfigyelhető tendenciák is, illetve az ügyfelek elégedetlensége is a régi rendszerrel szemben. A dolgozat keretein belül konkrét számpéldával nem volt lehetőségem alátámasztani a modellek létjogosultságát, de fontosnak tartom ennek a kérdésnek a további vizsgálatát, vagyis hogy az egyszerűbb, automatizált kockázatfelmérésnek köszönhető előnyök meghaladják-e a pontatlanabb becslések okozta veszteségeket.

Hivatkozások

Abrokwah, S. (2015). Predictive analytics in the life insurance process. *The Actuary*. Elérhető: (2020-02-23).

Arora, N. és Vij, S. K. (2012). A hybrid neuro-fuzzy network for underwriting of life insurance. *International Journal of Advanced Research in Computer Science*, 3(2).

Batty, M., Tripathi, A., Kroll, A., Wu, C.-s. P., Moore, D., Stehno, C., Lau, L., Guszcza, J., és Katcher, M. (2010). Predictive modeling for life insurance, ways life insurers can participate in the business analytics revolution. *Deloitte Consulting LLP*.

Beale, H. D., Demuth, H. B., és Hagan, M. (1996). *Neural network design*. Pws, Boston.

Bennett, A. K. (2004). Older age underwriting: frisky vs frail. *Journal of Insurance Medicine-New York then Denver*, 36(1):74–83.

Bernico, M. L. és Myers, J. (2019). Using images and voice recordings to facilitate underwriting life insurance. US Patent App. 10/296,982.

Berthelé, E. (2018). Using big data in insurance. In Corlosquet-Habart, M. és Janssen, J., editors, *Big data for insurance companies*, chapter 5, 130–161. Wiley Online Library.

Biddle, R., Liu, S., Tilocca, P., és Xu, G. (2018). *Automated Underwriting in Life Insurance: Predictions and Optimisation*, 135–146.

Boobier, T. (2016). Underwriting. In *Analytics for insurance: The real business of Big Data*, chapter 4, 51–59. John Wiley & Sons.

Boodhun, N. és Jayabalan, M. (2018). Risk prediction in life insurance industry using supervised learning algorithms. *Complex & Intelligent Systems*, 4(2):145–154.

Boyd, K., Eng, K. H., és Page, C. D. (2013). Area under the precision-recall curve: point estimates and confidence intervals. In *Joint European conference on machine learning and knowledge discovery in databases*, 451–466. Springer.

Breiman, L., Friedman, J., Olshen, R., és Stone, C. (1984). Classification and regression trees. wadsworth & brooks. *Cole Statistics/Probability Series*.

Börsch-Supan, A. (2019). Survey of health, ageing and retirement in europe (share) wave 4. Release version: 7.0.0. SHARE-ERIC. Data set. DOI: 10.6103/SHARE.w4.700.

Dr. Halász Katalin (2019). Személyes interjú (CIG Pannónia Életbiztosító Zrt.). Kockázatelbíráló orvos szerepe egy biztosítónál (2019.12.11.).

Eling, M. és Kraft, M. (2017). The impact of telematics on the insurability of risks.

Finkelstein, E. A., Haaland, B. A., Bilger, M., Sahasranaman, A., Sloan, R. A., Nang, E. E. K., és Evenson, K. R. (2016). Effectiveness of activity trackers with and without incentives to increase physical activity (trippa): a randomised controlled trial. *The lancet Diabetes & endocrinology*, 4(12):983–995.

Govindarajula, S. G. K. (2019). Classifying risk in life insurance using predictive analytics.

Gschlössl, S., Schoenmaekers, P., és Denuit, M. (2011). Risk classification in life insurance: methodology and case study. *European Actuarial Journal*, 1(1):23–41.

Hauer, J., Góg, E., Horváth, A., Hrabár, Á., és Pálinkás, K. (2017). A használatalapú biztosítás múltja, jelene és jövője. *Biztosítás és Kockázat*, 4(2):22–39.

Jakicic, J. M., Davis, K. K., Rogers, R. J., King, W. C., Marcus, M. D., Helsel, D., Rickman, A. D., Wahed, A. S., és Belle, S. H. (2016). Effect of wearable technology combined with a lifestyle intervention on long-term weight loss: the idea randomized clinical trial. *Jama*, 316(11):1161–1171.

Joram, M., Harisson, B., és Joseph, K. (2017). A knowledge-based system for life insurance underwriting. *International Journal of Information Technology and Computer Science*, 9:40–49.

Keating, R. (2017). New tech & big data: Are they good for insurance? *The Actuary*. Elérhető: `https://www.theactuary.com/features/2017/10/new-tech-big-data-are-they-good-for-insurance/`.

Kivisaari, E. (2018). Big data is coming-are your ready?: Insurance principle and actuaries in the age of fintech. *Biztosítás és Kockázat*, 5(2):60–63.

Klein, A. M. (2013). Life insurance underwriting in the united states–yesterday, today and tomorrow. *British Actuarial Journal*, 18(2):486–502.

Kovács, E. (2014). *Többváltozós adatelemzés*. Typotex, Budapest.

Kovács, E. és Varga, V. (2019). Adathullámok egészségről, idősödésről, nyugdíjba vonulásról. *Biztosítás és Kockázat*, 6(4):42–55.

Liferay & IDC (2020). Accelerating customer experience transformation in insurance through digital experience platforms. Elérhető: `https://www.liferay.com/documents/10182/13811/AcceleratingCustomerExperienceTransformationinInsuranceThroughDigitalExperiencePlatforms` (2020-02-20).

Matt, B. (2016). Is analytics changing the underwriting we know? *Insurance Portal*. Elérhető: `https://insurance-portal.ca/article/is-analytics-changing-the-underwriting-we-know/`.

Neaţu, A. M. (2015). Behavioral economics:"nudging" people into more active health conscious behaviors through wearable technology.

Nelder, J. A. és Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Nikolopoulos, C. és Duvendack, S. (1994). A hybrid machine learning system and its application to insurance underwriting. In *Proceedings of the First IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence*, 692–695. IEEE.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., és Sabeti, P. C. (2011). Detecting novel associations in large data sets. *science*, 334(6062):1518–1524.

Saito, T. és Rehmsmeier, M. (2017). Precrec: fast and accurate precision-recall and roc curve calculations in r. *Bioinformatics*, 33 (1):145–147.

Sallai, L. (2019). Insurtech: körkép és trendek. *Biztosítás és Kockázat*, 6(4):98–109.

Schmeiser, H., Störmer, T., és Wagner, J. (2016). Unisex insurance pricing: consumers' perception and market implications. In *The Geneva Papers*, 102–138. Springer.

Seekings, C. (2018). Insurance customers happy to share data for cheap premiums. *The Actuary*. Elérhető: https://www.theactuary.com/news/2018/06/nearly-half-of-insurance-customers-happy-to-share-data-for-cheap-premiums/(2019-10-20).

South, A. (2011). rworldmap: A new r package for mapping global data. *The R Journal*, 3(1):35–43.

Spender, A., Bullen, C., Altmann-Richer, L., Cripps, J., Duffy, R., Falkous, C., Farrell, M., Horn, T., Wigzell, J., és Yeap, W. (2019). Wearables and the internet of things: Considerations for the life and health insurance industry. *British Actuarial Journal*, 24.

Van Buuren, S. és Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

Wells, V. (2017). Underwriting the elderly. Elérhető: https://www.rgare.com/docs/default-source/underwriting-university/ui---underwriting-the-elderly-2017.pdf (2020-02-24).

Yi Tan és Guo-Ji Zhang (2005). The application of machine learning algorithm in underwriting process. In *2005 International Conference on Machine Learning and Cybernetics*, 3523–3527.

Függelék

A. SHARE hullámok és kérdések

9. táblázat. Országok az egyes hullámokban

Ország	1. hullám	2. hullám	3. hullám	4.hullám	5. hullám	6. hullám	7. hullám
Ausztria	2004	2006/07	2008/09	2011	2013	2015	2017
Németország	2004	2006/07	2008/09	2011/12	2013	2015	2017
Svédország	2004	2006/07	2008/09	2011	2013	2015	2017
Hollandia	2004	2007	2008/09	2011	2013	2015	2017
Spanyolország	2004	2006/07	2008/09	2011	2013	2015	2017
Olaszország	2004	2006/07	2008/09	2011	2013	2015	2017
Franciaország	2004/05	2006/07	2009	2011	2013	2015	2017
Dánia	2004	2007	2008/09	2011	2013	2015	2017
Görögország	2004/05	2007	2008/09	-	-	2015	2017
Svájc	2004	2006/07	2008/09	2011	2013	2015	2017
Belgium	2004/05	2006/07	2008/09	2011	2013	2015	2017
Izrael	2005/06	2009/10	-	-	2013	2015	2017
Csehország	-	2006/07	2008/09	2011	2013	2015	2017
Lengyelország	-	2006/07	2008/09	2011/12	-	2015	2017
Írország	-	2007	2009/10/11	-	-	-	-
Luxemburg	-	-	-	-	2013	2015	2017
Magyarország	-	-	-	2011	-	-	2017
Portugália	-	-	-	2011	-	2015	2017/18
Szlovénia	-	-	-	2011	2013	2015	2017
Észtország	-	-	-	2010/11	2013	2015	2017
Horvátország	-	-	-	-	-	2015	2017
Litvánia	-	-	-	-	-	-	2017
Bulgária	-	-	-	-	-	-	2017
Ciprus	-	-	-	-	-	-	2017
Finnország	-	-	-	-	-	-	2017
Lettország	-	-	-	-	-	-	2017
Málta	-	-	-	-	-	-	2017
Románia	-	-	-	-	-	-	2017
Szlovákia	-	-	-	-	-	-	2017

Forrás: `http://www.share-project.org`

A modulok ismertetéséhez Kovács és Varga [2019] fordítását használom.

- **Demográfia (DN)**: Alapvető demográfiai adatok (család, iskolázottság)
- **Szociális kapcsolatok (SN)**: Személyes kapcsolatok adatai
- **Gyermekek (CH)**: A válaszadók gyermekeivel kapcsolatos információk
- **Fizikai egészség (PH)**: Krónikus betegségek, mindennapi nehézségek
- **Viselkedési kockázat (BR)**: Egészségre káros szokások
- **Kognitív képességek (CF)**: Tesztek eredményei: memória, koncentráció, beszéd
- **Mentális egészség (MH)**: Érzelmi és mentális egészségi állapot
- **Egészségügyi ellátás (HC)**: Orvos- és kórházlátogatások
- **Foglalkoztatottság és nyugdíj (EP)**: Munka- és nyugdíjjövedelmek
- **Számítógép-használat (IT)**: Számítógép-használat munkahelyen és otthon
- **Mini gyermekkor (MC)**: Gyermekkori élethelyzet és betegségek
- **Szorítóerő-mérés (GS)**: Szorítóerő mindkét kézen, mérés alapján
- **Sétasebesség (WS)**: Séta sebessége és a teszt körülményei
- **Felállás székről (CS)**: Székről való felállás sebessége és a teszt körülményei
- **Vérminta (BS)**: Vérvétel körülményei, eredmények nélkül
- **Kilégzési erő mérése (PF)**: Tüdőkapacitás mérése eszközzel
- **Szociális támogatás (SP)**: Kapott és adott segítségek
- **Pénzügyi transzfer (FT)**: Kapott és adott pénzügyi transzferek
- **Lakhatás (HO)**: Lakhellyel kapcsolatos információk
- **Háztartás jövedelme (HH)**: Előző évben szerzett összes jövedelem személyenként
- **Fogyasztás (CO)**: A háztartás ételekre fordított kiadásai
- **Eszközök (AS)**: A háztartás pénzügyi és nem pénzügyi eszközei, bevételei
- **Aktivitás (AC)**: Szabadidős tevékenységek és azokkal való elégedettség
- **Elvárások (EX)**: A jövővel kapcsolatos várakozások
- **Interjú körülményei (IV)**: Az interjúztató észrevételei
- **End of Life (XT)**: A halál körülményei, okai

10. táblázat. Kérdések köre az egyes hullámokban

Modul	Modul megnevezése	1. hullám	2. hullám	4.hullám	5. hullám	6. hullám	7. hullám
DN	Demographics and Networks	X	X	X	X	X	X
SN	Social Networks			X		X	
CH	Children	X	X	X	X	X	X
PH	Physical Health	X	X	X	X	X	X
BR	Behavioral Risks	X	X	X	X	X	X
CF	Cognitive Function	X	X	X	X	X	X
MH	Mental Health	X	X	X	X	X	X
HC	Health Care	X	X	X	X	X	X
EP	Employment and Pensions	X	X	X	X	X	X
IT	Computer Use				X	X	X
MC	Mini Childhood				X		X
GS	Grip Strength	X	X	X	X	X	X
WS	Walking Speed	X	X				
CS	Chair Stand		X		X		
BS	Blood Sample					X	
PF	Peak Flow		X	X		X	
SP	Social Support	X	X	X	X	X	X
FT	Financial Transfers	X	X	X	X	X	X
HO	Housing	X	X	X	X	X	X
HH	Household Income	X	X	X	X	X	X
CO	Consumption	X	X	X	X	X	X
AS	Assets	X	X	X	X	X	X
AC	Activities	X	X	X	X	X	X
EX	Expectations	X	X	X	X	X	X
IV	Interviewer Observations	X	X	X	X	X	X
XT	End-of-Life Interview		X	X	X	X	X

Forrás: `http://www.share-project.org`

B. Felhasznált SHARE változók

11. táblázat. Kiválasztott változók és azok fontossága az egyes módszerek szerint

Kód	Mérési skála	Fontosság			Leírás
		1	2	3*	
gender	Nominális	•	•		Nő vagy Férfi
hand	Nominális	○	○		Jobb- vagy balkezes
age	Arány	•	•	•	A megkérdezettek életkora
yedu	Arány	○	•	○	Tanulással töltött évek száma
mstat	Nominális	○	•		Családi állapot; 6 szint: *Married (living with spouse), Registered partnership, Married (not living with spouse), Never married, Divorced, Widowed*
nchild	Arány	○	•	○	Gyerekek száma
ngrchild	Arány	•	○	○	Unokák száma
esmoked	Ordinális	•	•		Dohányzott-e valaha napi szinten
drinking	Ordinális	○	•		Fogyaszt-e kettőnél több pohár alkoholt majdnem minden nap
eat	Ordinális	○	•		Eszik-e minden nap legalább három alkalommal
bmi	Arány	•	•	○	BMI - *Body Mass Index*
sphus	Ordinális	•	•		Saját egészség értékelése; 5 szint: *Exellent, Very good, Good, Fair, Poor*
chronic	Arány	○	○	○	Krónikus betegségek száma
symps	Arány	○	•	○	Tünetek száma
doctor	Arány	○	•	•	Elmúlt egy évben történt orvosi találkozások száma
nhosp	Arány	•	•	•	Elmúlt egy évben kórházban töltött esték száma
gali	Ordinális	○	•		Korlátozott-e a mindennapi tevékenységekben
mobility	Arány	○	•	•	Mozgásra vonatkozó korlátozások száma
adl	Arány	○	•	•	Korlátozottságok száma az alapvető mindennapi tevékenységekben
iadl	Arány	•	•	•	Korlátozottságok száma komplex mindennapi tevékenységekben
eyesightr	Ordinális	○	○		Olvasási képesség; 5 szint: *Exellent, Very good, Good, Fair, Poor*
reading	Ordinális	○	○		Saját olvasási képesség értékelése; 5 szint: *Exellent, Very good, Good, Fair, Poor*
writing	Ordinális	○	○		Saját írási képesség értékelése; 5 szint: *Exellent, Very good, Good, Fair, Poor*
hearing	Ordinális	○	○		Hallás; 5 szint: *Exellent, Very good, Good, Fair, Poor*
orienti	Ordinális	○	○		Időbeli orientációs teszt (pontos dátum ismerete); 5 szint: *Bad, 1, 2, 3, Good*
wllft	Arány	•	•	•	Tíz szó memorizálása elsőre
wllst	Arány	○	○	•	Tíz szó memorizálása másodszorra
fluency	Arány	○	•	•	Beszéd gördülékenység, 1 perc alatt minél több állat megnevezése
numeracy	Ordinális	○	○		Első számolási teszt (arányosítás); 5 szint: *Bad, 2, 3, 4, Good*
numeracy2	Ordinális	○	○		Első számolási teszt (kivonás); 6 szint: *Bad, 1, 2, 3, 4, Good*
memory	Ordinális	○	○		Memóriateszt; 5 szint: *Exellent, Very good, Good, Fair, Poor*
maxgrip	Arány	•	•	○	Maximum szorítóerő
lung	Ordinális	○	○		Kilégzési erő mérése; 10 szint: *No result, 100-200, . . . , 800-900*
cjs	Nominális	○	○		Jelenlegi foglalkozás; 6 szint: *Retired, Employed or self-employed, Unemployed, Permanently sick, Homemaker, Other*
phinact	Ordinális	•	•		Aktív életet él-e
naly	Arány	•	•	•	Szabadidős tevékenységek száma
saly	Ordinális	○	○		Tevékenységek hiányával való elégedettség; 11 szint: *Completely dissatisfied, 1, . . . , 9, Completely satisfied*
eurod	Ordinális	○	○		Európai depressziós skála: 13 szint: *Not depressed, 1, . . . , 11, Very depressed*
lifesat	Ordinális	○	○		Életelégedettség; 11 szint: *Completely dissatisfied, 1, . . . , 9, Completely satisfied*
lifehap	Ordinális	○	○		Boldogságérzet gyakorisága; 4 szint: *Never, Rarely, Sometimes, Often*
areabldgi	Nominális	○	•		Lakóhely elhelyezkedése; 5 szint: *A big city, The suburbs or outskirts of a big city, A large town, A small town, A rural area or village*
GDPcat	Ordinális	○	•		Országok beosztása egy főre jutó GDP alapján; 3 szint: *1, 2, 3*
lifeins	Nominális	○	•		Rendelkezik-e életbiztosítással; 3 szint *Yes, No, Don't know*
status	Nominális				Célváltozó: meghalt vagy sem az adott illető

Forrás: Saját készítés

*: 1: Legjobb részhalmaz; 2: Lépésenkénti szelekció; 3: Zsugorító módszerek

Adott módszer szerint fontos (•) vagy nem fontos (○) változó