

OTDK-Dolgozat

2022

BUDAPESTI CORVINUS EGYETEM
KÖZGAZDASÁGTUDOMÁNYI KAR



**A SZÍVBETEGSÉGEK DETEKTÁLÁSA MACHINE LEARNING
MÓDSZEREKKEL**
**DETECTION OF HEART DISEASES USING MACHINE LEARNING
METHODS**

Braun Gergely
Nagybali Tamás
Vörös András

Konzulens: Madari Zoltán

Kézirat lezárásának dátuma: 2023. 01. 03.

Absztrakt

A dolgozatban egy amerikai orvosi felmérés alapján modellezzük a szívbetegségek kialakulását, annak detektálását különböző gépi tanulási módszerekkel. A modellek eredményeinek felhasználása alkalmas arra, hogy előzetesen szűrhetők legyenek a szívbetegségekre hajlamos páciensek. Elemzésünkben a téma egészségügyi és közgazdasági megalapozását követően bemutatjuk az elemzésünkhöz szükséges módszertani háttérismereteket, korábbi empirikus tanulmányokat. Ismertetjük az adatfelvételt, illetve bemutatjuk a változókat, jellemezzük azokat leíró jelleggel. Az első modell, amelyet használunk a logisztikus regresszió. Második lépésben a véletlen erdő (random forest) módszer segítségével modellezzük a szívbetegségeket. Végül a neurális hálók segítségével adunk becslést a szívbetegségek tényére. A következő fejezetben összehasonlítjuk a három modell eredményeit és megvizsgáljuk, hogy melyik a legalkalmasabb modell a kívánt eredmény elérésére. Ezek után kitekintést teszünk az ilyen modellek egészségügyi és közgazdasági alkalmazására, melynek során a magyarországi helyzetet is vizsgáljuk. Tanulmányunk végén összegezzük eredményeinket, és levonjuk a következtetéseket az alkalmazásukkal kapcsolatban.

Kulcsszavak: machine learning, egészségügyi közgazdaságtan, szívbetegségek, logisztikus regresszió, véletlen erdő, neurális hálózat

JEL klasszifikáció: C01, C15, C31, C45, C55, I15

Tartalomjegyzék

1. Bevezetés (Szerző: Vörös András)	1
2. Szakirodalmi áttekintés (Szerzők: Vörös András, Braun Gergely)	2
3. A logisztikus modell módszertana (Szerzők: Braun Gergely)	5
3.1. ROC görbe (<i>receiver operating characteristic curve</i>)	6
3.2. Vágási érték.....	7
3.3. Modellszelekció	7
4. A döntési fa módszertana (Szerző: Braun Gergely).....	8
4.1. Klasszifikációs fa.....	8
4.2. A fa elemeinek bemutatása.....	9
4.3. A modell építése	9
4.4. Végső felépítés.....	11
5. A neurális háló módszertana (Szerző: Nagyball Tamás)	11
5.1. A neurális hálózatok felépítése	11
5.2. A neurális hálózatok tanulása.....	12
6. Az adatbázis (Szerzők: Braun Gergely, Vörös András).....	15
6.1. A változók bemutatása	15
6.2. A változók értelmezése	16
7. Leíró statisztikai elemzés (Szerzők: Braun Gergely, Vörös andrás).....	18
8. A logisztikus modell (Szerzők: Braun Gergely, Vörös András).....	21
8.1. A végső logisztikus modell	21
8.2. A logisztikus modell klasszifikációja.....	23
9. A véletlen erdő modell (Szerző: Nagyball Tamás)	25
9.1. Arányok megváltoztatása	26
9.2. A végső véletlen erdő modell	27
10. A neurális háló modell (Szerző: Nagyball Tamás)	29
10.1. A végső neurális háló modell.....	32
11. A végső modellek összehasonlítása (Szerző: Nagyball Tamás)	33
12. Kitekintés (Szerző: Vörös András)	35
12.1. Magyarország helyzete	35
12.2. Orvosi szakvélemények	37
12.3. Következtetések	39
13. Összefoglalás (Szerző: Nagyball Tamás).....	40
14. Hivatkozások (Szerző: Braun Gergely).....	42

Ábrajegyzék

1. ábra; ROC görbe (saját szerk.: RStudio).....	6
2. ábra; A fa elemei (saját szerk.: Word).....	9
3. ábra; Segédábra a sorrend kiválasztásához a fa modellben (saját szerk.: Word).....	10
4. ábra; Kis neurális hálózat szerkezete (saját szerk.: Word).....	11
5. ábra; A testtömegindex hisztogramja (saját szerk.: Excel)	19
6. ábra; Az életkor megoszlása kategóriánként (saját szerk.: Excel)	19
7. ábra; Jövedelmi kategóriák megoszlása (saját szerk.: Excel)	20
8. ábra; Oktatási helyzet kategóriánként (saját szerk.: Excel).....	20
9. ábra; A végleges logisztikus modell ROC-görbéje (saját szerk.: RStudio)	24
10. ábra; Az FNR értékének változása az egészségesek arányának függvényében (saját szerk.: Excel)	27
11. ábra; A klasszifikáció pontossága az egészségesek arányának függvényében (saját szerk.: Excel)	27
12. ábra; A végleges véletlen erdő modell ROC-görbéje (saját szerk.: RStudio).....	28
13. ábra; A neurális hálózat tanulása dropout regularizáció mellett (saját szerk.: RStudio)....	30
14. ábra; A becslések pontossága tanuló és megerősítő adatsoron neurális háló esetén (saját szerk.: RStudio).....	30
15. ábra; A neurális háló pontossága különböző struktúrák mellett (saját szerk.: RStudio)....	31
16. ábra; A neurális hálózat tanulási folyamata (saját szerk.: RStudio).....	31
17. ábra; Férfiak várható élettartama születéskor (2019) (saját szerk.: Excel)	35
18. ábra; Nők várható élettartama születéskor (2019) (saját szerk.: Excel).....	36
19. ábra; Felnőttek szívbetegségeinek aránya Magyarországon (saját szerk.: Excel).....	37

Táblázatjegyzék

1. táblázat; Klasszifikációs tábla	6
2. táblázat; Az adatbázis változói	15
3. táblázat; A jövedelmi változó fokozatai.....	18
4. táblázat; A végső logisztikus modell.....	22
5. táblázat; Logisztikus modell klasszifikációs tábla (C = klasszikus)	23
6. táblázat; Logisztikus modell klasszifikációs tábla (C = Youden-szabály).....	24
7. táblázat; Első véletlen erdő modell klasszifikációs táblája	26
8. táblázat; A végső véletlen erdő modell klasszifikációs táblája	28
9. táblázat; A végső neurális hálózat.....	32
10. táblázat; A végső neurális háló modell klasszifikációs táblája	32

11. táblázat; A modellek fontosabb mutatói.....	33
---	----

Egyenletjegyzék

1. egyenlet; Odds képlete	5
2. egyenlet; A feltételes valószínűség képlete.....	5
3. egyenlet; Logisztikus regresszió	5
4. egyenlet; Valódi pozitív ráta	6
5. egyenlet; Ál negatív ráta	6
6. egyenlet; Klasszifikáció pontossága.....	6
7. egyenlet; Pszeudo- R^2	7
8. egyenlet; AIC képlete.....	8
9. egyenlet; BIC képlete	8
10. egyenlet; Wald-F próbafüggvény	8
11. egyenlet; Gini Impurity Index	10
12. egyenlet; Total Gini Impurity Index.....	10
13. egyenlet; i-edik réteg neuronjainak aktivitása.....	12
14. egyenlet; ReLU aktivációs függvény	12
15. egyenlet; Softmax aktivációs függvény	12
16. egyenlet; Bináris kereszt-entrópia	12
17. egyenlet; Regularizált költség függvény	13
18. egyenlet; Testtömegindex	17

1. Bevezetés

A huszonegyedik században a globalizált társadalom egyre egészségtelenebb életmódja miatt az orvoslásban egyre inkább a krónikus betegségek megelőzése és ezek mielőbbi szűrése kerül előtérbe. Sok betegség esetén a detektálás ideje a legfontosabb tényező a páciens gyógyulási esélyeiben. Az emberek nagy része azonban nem szívesen szakít időt, pénzt vagy energiát ilyen vizsgálatok elvégzésére, sokszor már csak az egyértelmű tünetek jelentkezésekor veszik a fáradságot a cselekvésre. Legtöbbször azonban ilyenkor már a betegség jóval előrehaladottabb stádiumában van, így nehezebb a gyógyítása is.

Magyarországon egy magánklinika által elvégzett, mindenre kiterjedő kardiológiai vizsgálat ára átlagosan 60.000 forint is lehet, ami természetesen az emberek nagy hányadának már túl van fizetési hajlandóságának határán. Az állami egészségügyben azonban inkább az időfaktor a befolyásoló tényező. A hosszú sorok és lassú adminisztráció miatt kevesen mennek el a közegészségügyi létesítmények vizsgálataira. Véleményünk szerint ezen egy ingyenes, akár online is kitölthető kérdőívvel könnyen javíthatunk, melyben általános egészségügyi kérdések szerepelnek. A digitalizáció terjedése a magyar állami egészségügyi szektorba is elért az elmúlt időszakban, amely hatást a koronavírus logisztikai megoldási kényszere is segítette. A házi orvosi rendszer is rendelkezik már olyan platformmal (Erodium), amellyel online intézkedéseket tehetnek a páciensek, például időpontfoglalást, ezzel is hatásosabbá téve a folyamatot. Ezt a tényezőt felhasználva érdemes lenne olyan mechanizmusokat kitalálni, amelyek hatékony és gyors választ adnak arra, hogy egyáltalán érdemes-e a szűrési vizsgálatokat elkezdni, egy gyors modell segítségével. Néhány adat megadásával a modell számításokat végez, majd különösebb beavatkozás nélkül diagnosztizál megbetegedési kockázatokat. Az általános adatokon kívül szükséges lehet még vér és vizeletvizsgálat hiszen ezek alapján lehet megkapni például a pontos koleszterinszintet és vércukorszintet, melyen értékek a dohányzás, magas vérnyomás, fizikai aktivitás és a túlsúlyon kívül a szívbetegségek kialakulásának fő rizikófaktorai (National Health Service, 2018). Véleményünk szerint az itthon használt Erodium vagy EESZT rendszerbe könnyen beültethető lenne egy erre a célra kifejlesztett algoritmus, mely az egyének adatai alapján számol kockázatot. Ráadásul, ha egy páciens tudja, hogy veszélyeztetett kategóriába tartozik, akkor valószínűleg hajlandó elvégeztetni az időigényesebb szűrővizsgálatokat is. Szerintünk ez a magánklinikáknak is előnyére is válna, mivel ezzel a módszerrel potenciális vásárlókat nyerhetnek, ám elsősorban terhet venne le a túlterhelt államilag végzett szűrési vizsgálati folyamatokról. Csökkenne az aszimmetrikus információs probléma orvos és beteg között így az orvosok munkája is

egyszerűbbé válna. Ezzel egyfajta Pareto-javítás is elérhető, hiszen az állami egészségügy tehertől mentesül, a magánklinikák számára pénzügyileg előnyös, a betegek pedig hamarabban kaphatnak pontos diagnózist.

Dolgozatunk célja, hogy egy a fentiekben leírtakhoz hasonló, amerikai páciensek adatait tartalmazó pilot adatbázissal megvizsgáljuk mennyire lehet jól modellezni a szívbetegségek kialakulását. Ehhez a feladathoz három gépi tanulási módszert vizsgálunk. Első körben a logisztikus regresszióval, majd a véletlen erdő és neurális hálózat algoritmusokkal igyekszünk modellezni a kardiológiai probléma meglétének tényét. Véleményünk szerint a big data korában az ilyen modellek alkalmazásának nincs akadálya, egy jól implementált modell használata költséghatékony, segít a megelőzésben és viszonylag egyszerű is. Tárgyalásunk végén adunk egy rövid kitekintést is a magyar és külföldi kardiológiai viszonyokról, összekapcsoljuk a mi kutatási modellünk alkalmazási lehetőségeivel, majd összegezzük elemzésünket.

2. Szakirodalmi áttekintés

Ebben a fejezetben a kutatásunk szempontjából releváns szakirodalmak feldolgozását végeztük el.

Egy 2008-ban az Acta Otorrinolaringológica Española-ban íródott publikációt tartunk fontosnak kiemelni, amely az egyoldalú garatmandula megnagyobbodásban szenvedő betegek között vizsgálta a rosszindulatú daganat kockázati tényezőit, illetve a mandula eltávolítás relevanciáját.

Ebben a tanulmányban egy saját klinika betegeinek az adatait vették fel, akikben egy közös volt, hogy mindenkinek eltávolították a garatmanduláját. Ezen a csoporton belül pedig külön azokra a páciensekre fókuszáltak, akiknek egyoldalú garatmandula megnagyobbodásuk volt a műtét előtt, és azért műtötték meg őket, mert azt gondolták, hogy ez a megnagyobbodás egy rossz indulatú rákos daganatot is hordozhat titokban.

Logisztikus regresszióval állapították meg a kockázati tényezőket, ahogyan mi is tettük az első modellünk során.

Végül konklúzióként elmondható, hogy voltak olyan esetek, amelyeknél kiderült, hogy nem volt indokolt a műtét abban az esetben hogyha a kockázati tényezők alapján mérték volna fel a beavatkozás szükségességét, elég lett volna szoros megfigyelés. A mi munkánk is erre próbál rávilágítani, hogy egy orvosok és statisztikusok által összeállított modell sok betegség esetén megelőzhet elkerülhető beavatkozásokat, vagy éppen figyelmeztet a veszélyre.

Egy másik szakirodalom, amit érdemesnek tartunk kiemelni, az egy pszichológiai tanulmány a betegségek szűrésének és narrálásának mentális módszertanáról, hiszen ez releváns lehet a tematikánkhoz. A cikk a Journal of Health Psychology-ban jelent meg, és az alaptémája a narratíva, illetve a statisztika meggyőző erejének az összehasonlítása az egészségügyi szűrések népszerűsítésében.

A cikk öt különböző kutatás eredményeit összegzi, és azok alapján von le következtetéseket. A tanulmányok alapján az elmondható, hogy az emberek gondolkodását, viszonyulási módját jobban befolyásolja egy hatékonyan felépített verbális kommunikáció, mint a statisztikai tények ismertetése a betegségekről. Az is megállapításra kerül, hogy különböző személyiség típusokra másfajta narratíva hathat dominánsabban, e tény kifejtése azonban további vizsgálatokat igényel.

Azt mondhatjuk tehát, hogy nem elég, hogy egy matematikailag megbízható egészségügyi modellt készítünk, megfelelő kommunikáció is kell a társadalom felé, ha azt akarjuk, hogy minél hatékonyabbak legyenek a szűrések, és minél több ember komolyabban vegye a témát.

Releváns lehet még számunkra Howard G. Stoa 1989-es tanulmánya, amely a Journal of the Royal College of General Practitioners című folyóiratban jelent meg. Ebben egy kisebb összefoglalót ad azzal kapcsolatban, hogy milyen mentális következményei lehetnek az egészséges emberek egészségügyi szűréseinek.

A kutatás egy olyan mintavételezésen alapult, ahol meghívtak 215 egészséges embert szívbetegségek szűrésére, és egy szubjektív mentális állapotra vonatkozó kérdőívet kellett kitölteniük. Szintén kitöltött egy ilyen tesztet 225 másik ember, akik azonban nem mentek el szűrésre. Későbbi vizsgálatok során az volt látható, hogy a szűrésen részt vett emberek stressz szintje szignifikánsan növekedett, a másik csoporttal ellentétben, ahol ez nem volt megfigyelhető. Összegezve tehát a tanulmány azt vélte felfedezni, hogy a szűrési vizsgálatok negatív hatással lehetnek az egészséges emberek mentális állapotára.

Ez szintén olyan tényező, amely figyelembe vehető egy általunk is felvetett rizikó modellezés kialakításánál, hiszen a mentális állapot nagyban befolyásolhatja a további betegségekre való hajlamot.

Szintén témába illő szakirodalom egy 2017-es dán kutatás (Hansen és szerzőtársai), amely 1231 emberen tesztelte azt, hogy különböző rizikójú kardiovaszkuláris szűrési módok között hogyan

válogatnak, mit tartanak maguknak előnyösnek és értékesnek, és mi az, amit már nem kockáztatnának.

Az eredmény az lett, hogy az emberek általában tisztában voltak a preferenciájukkal, és hogy mi az, amit megengednének maguknak szűrési kísérletek alatt. Számunkra ez releváns lehet, hiszen ez azt jelenti, hogy a betegek preferenciáinak megismerésével is lehet népszerűsíteni, hatékonyabbá tenni a prevenció szűréseket.

3. A logisztikus modell módszertana

A következő fejezetekben az elemzés során felhasznált modellek és tesztek módszertanát foglaljuk össze. Minden próbát 5 százalékos szignifikanciaszinten végzünk.

Jelen esetben az eredményváltozónk bináris, ezért nem becsülhetjük OLS-sel (legkisebb négyzetek módszere) (Wooldridge, 2012). Az első modellünk építése során tehát logisztikus regressziót alkalmazunk az ML (Maximum likelihood) becslés segítségével (Ferenci, 2018). Az 1-es kategória a szívbetegségekben szenvedő, a 0-s az egészségeseket jelöli.

Ez a módszer nem az eredményváltozó adott csoportjába való tartozásnak tényét, hanem a feltételes valószínűségét modellezi. Annak érdekében, hogy ezt becsülni tudjuk, néhány transzformációt kell alkalmaznunk az eredményváltozón. Az odds, más szóval az esély fogalma az egyik csoport valószínűsége a másik csoport valószínűségéhez viszonyítva (Ferenci, 2018), és ebből fejezzük ki az adott csoportba tartozás feltételes valószínűségét (esetünkben az egyes, vagyis a veszélyeztetett csoportba tartozását).

$$odds_{\underline{x}} = \frac{\mathbb{P}_{\underline{x}}}{1 - \mathbb{P}_{\underline{x}}}$$

1. egyenlet; Odds képlete

$$\mathbb{P}_{\underline{x}} = \frac{odds_{\underline{x}}}{1 + odds_{\underline{x}}}$$

2. egyenlet; A feltételes valószínűség képlete

Végül vesszük az odds-nak a természetes alapú logaritmusát, és ezt már lehetséges lineárisan modellezni, ezt nevezzük logisztikus regresszióknak (Ferenci, 2018).

$$logit_{\underline{x}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k = \underline{x}^T \beta$$

3. egyenlet; Logisztikus regresszió

Ez alapján már könnyen kiszámítható a $\mathbb{P}_{\underline{x}}$, vagyis a magyarázó változók ismeretében annak valószínűsége, hogy az adott megfigyelés az 1-es csoportba tartozik. A β paraméterekkel kiszámítható, hogy az adott magyarázóváltozó változása milyen irányban befolyásolja a megfigyelés egyes kategóriákba esésének feltételes valószínűségét (Hanck, et al., 2019).

3.1. ROC görbe (receiver operating characteristic curve)

Ahogy a fentebb leírtakban olvasható, nem azt becsüljük meg hogy egy adott esemény bekövetkezik vagy sem, hanem a feltételes valószínűséget. Kell egy határ, ami alapján dönteni tudunk (Ferenci, 2018), ezt a határt nevezzük vágási értéknek (Cut-off value, innentől C-vel jelöljük). Az alábbi táblázatban látható egy klasszifikációs mátrix.

		Becsült értékek	
		0	1
Valós értékek	0	Valódi negatív (TN)	Ál pozitív (FP)
	1	Ál negatív (FN)	Valódi pozitív (TP)

1. táblázat; Klasszifikációs tábla

$$TPR = \frac{TP}{TP + FN}$$

4. egyenlet; Valódi pozitív ráta

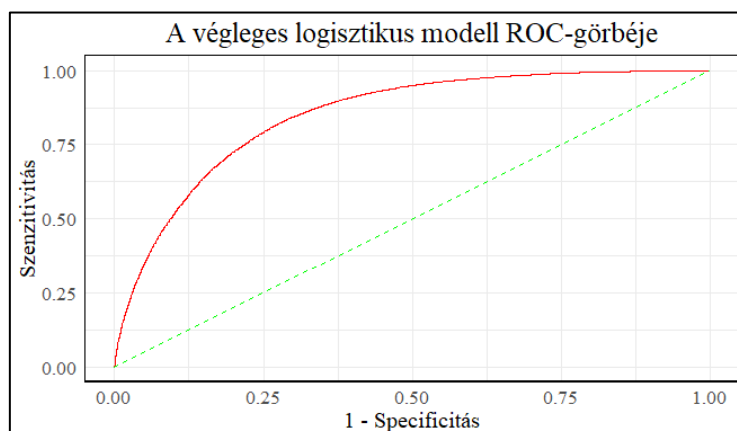
$$FNR = 1 - TPR$$

5. egyenlet; Ál negatív ráta

$$Pontosság = \frac{TN + TP}{TN + FP + FN + TP}$$

6. egyenlet; Klasszifikáció pontossága

A használt modell becsléseit 4 kategóriába sorolhatjuk, ezek a valódi negatív és pozitív, valamint az ál negatív és pozitív. Ezekből számolhatóak a valós pozitív és ál negatív arányok a táblázat alatt látható módon (Fawcett, 2005). A ROC görbén a 0-tól 1-ig végigfutó C értékek mellett ábrázoljuk a két mutató értékét.



1. ábra; ROC görbe (saját szerk.: RStudio)

Ezzel a módszerrel könnyen összehasonlítható két logisztikus modell klasszifikációs képessége, egyszerűen kiszámoljuk a görbe alatti területet (AUC – Area Under Curve), nagyobb érték esetén az adott modell várhatóan minden C érték mellett pontosabban csoportosítja a megfigyeléseket.

3.2. Vágási érték

A döntéshez legmegfelelőbb konkrét C érték azonban nem határozható meg egyértelműen a ROC-görbe alapján. Ehhez ismernünk kellene a hibás és jó döntések pontos költségfüggvényét (Ferenci, 2018). Ennek hiányában két heurisztikát említünk a vágási érték meghatározására:

Az első egy általános 50 százalékos szabály, ami annyit jelent, hogyha a feltételes valószínűség nagyobb mint 50 százalék akkor a modellünk az 1-es kategóriába sorolja a megfigyelést.

A másik módszer pedig a Youden-szabály alapján határozza meg a vágási értéket, vagyis a specificitás és a szenzitivitás összegét maximalizáljuk (Ferenci, 2018), és a kapott C érték alapján döntünk.

3.3. Modellszelekció

Annak érdekében, hogy több logisztikus regressziós modellt össze tudjunk hasonlítani többféle módszer áll a rendelkezésünkre.

3.3.1. Pszeudo- R^2 (McFadden-féle)

Ezzel a mutatóval összehasonlítjuk a modellünket egy teljesen üres modellel, mely nem tartalmaz semmilyen magyarázó változót, a megfigyeléseket a csoportok aránya alapján kategorizálja (Ferenci, 2018).

A pszeudo- R^2 gyakorlatilag azt mutatja meg, hogy a jelenlegi modellünk mennyire áll távol az üres modelltől. Ha a mutató értéke egyhez közeli akkor a modell jó kategorizáló erővel bír.

$$R^2 = \frac{(-2\ln L_{null}) - (-2\ln L_{aktuális})}{-2\ln L_{null}}$$

7. egyenlet; Pszeudo- R^2

3.3.2. Akaike információs kritérium

Az Akaike kritérium, röviden „AIC”, becslést ad arról az információ mennyiségről, amely elveszne, ha az adott modellt használjuk becslésre (Ferenci, 2018). Vagyis azt méri, hogy a modell mennyire illeszkedik az adatokhoz. Minél kisebb az AIC érték annál jobb a modell.

$$\frac{ESS}{n} \cdot e^{\frac{2k}{n}}$$

8. egyenlet; AIC képlete

Az egyenletben az n a megfigyelések száma, k a magyarázóváltozók száma, ESS pedig a hibanégyzetösszeg.

3.3.3. Bayesi információs kritérium

A Bayesi információs kritérium, röviden „BIC”, az Akaike kritériumhoz hasonló elven működik, csak szigorúbban bünteti a szabadságfokvesztést (Ferenci, 2018). Szintén, mint az AIC-nél, itt is az érték minimalizálására törekszünk. Egyenlete a következő:

$$\frac{ESS}{n} \cdot n^{\frac{k}{n}}$$

9. egyenlet; BIC képlete

Ahol n , k és az ESS szintén a megfigyelések száma, magyarázóváltozók száma és a hibanégyzetösszeg

3.3.4. Wald-teszt

A Wald-F teszt egy magyarázóváltozókra vonatkozó próba (Ferenci, 2018). Képlete az alábbi:

$$\frac{(ESS(R) - ESS(U))/m}{ESS(U)/(n - q - m)} \sim F(m, n - q - m)$$

10. egyenlet; Wald-F próbafüggvény

Ahol az $ESS(R)$ a szűkített modell, az $ESS(U)$ a bő modell hibanégyzetösszege, n a megfigyelések száma, q az elhagyott változók száma, m pedig a mindkét modellben szereplő magyarázóváltozók száma.

A Wald-teszt nullhipotézisét elfogadva arra a következtetésre jutunk, hogy az eredeti modellből elhagyott változók nem bírnak szignifikáns magyarázóerővel.

4. A döntési fa módszertana

4.1. Klasszifikációs fa

A döntési fának két fajtája van, a regressziós fa, amely numerikus, és a klasszifikációs fa, mely kategoriális eredményváltozók becslésére képes (James és szerzőtársai, 2013). Az elemzésünkben klasszifikációs fát használunk.

Természetesen a klasszifikációs fa modell is képes kezelni mind a numerikus mind pedig a kategoriális magyarázó változókat.

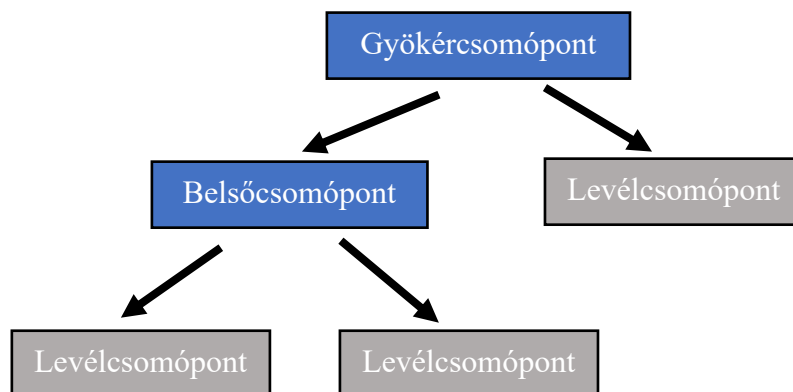
4.2. A fa elemeinek bemutatása

A klasszifikációs fa képe nem egy klasszikus fára, hanem inkább egy piramisra vagy fordított fára hasonlít. A döntési fák matematikailag gráfok (James és szerzőtársai, 2013). Három elemből állnak melyek a gyökércsomópont (root node), a belső csomópontok vagy ágak (internal nodes vagy branches) és végül a levél csomópont (leaf node vagy leafs).

A döntési fa rekurzív algoritmusok segítségével modellezi a kívánt változót. A gyökércsomópontból mindig egy elem létezik, ez a „fa” kiindulópontja, az első változó, ami alapján a modell ketté bontja a tanuló adathalmazt. Az adott csomópontoknál a felbontáshoz kapcsolódó kérdéseket úgy „teszik fel”, hogy a kisebb részek homogénebben legyenek a magyarázandó változó szempontjából, mint az egész.

A belsőcsomópontok szintén egy magyarázó változó alapján bontják részekre az adathalmazt, hasonlítanak a gyökércsomópontra, annyi különbséggel, hogy a fa bármely szintjén lehet kivéve a legutolsó vagy a legelső.

Végül a levélcsomópont, ebben jelöljük a döntést, hogy minek klasszifikáljuk az adott megfigyelést, innen már nem nyílik több elágazás, mindig a gyökér, vagy a belsőcsomópont alatt helyezkedik el.



2. ábra; A fa elemei (saját szerk.: Word)

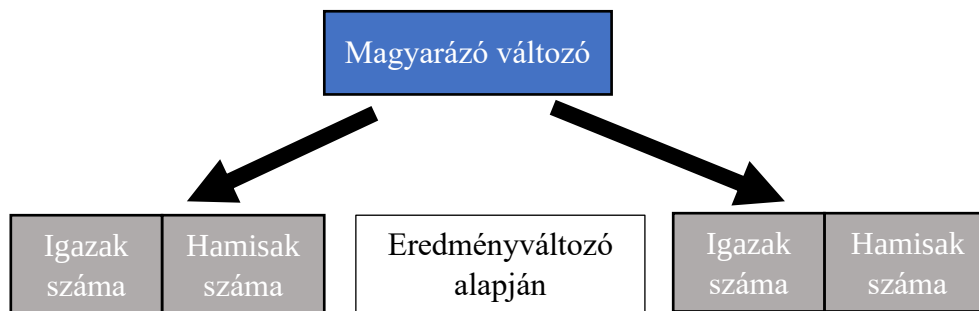
4.3. A modell építése

Egy klasszifikációs fa modell építése nem bonyolult dolog, csak nagyon hosszadalmas. Felvetődik a kérdés, hogy vajon melyik változó legyen legfelül, illetve mely sorrendben legyenek a változók, vagy érdemes-e felhasználni az összes változót egyáltalán. Ezek mind állítható paraméterek az R szoftvercsomagban (James és szerzőtársai, 2013).

Annak érdekében, hogy képesek legyünk eldönteni a sorrendet, minden változóra külön-külön meg kell vizsgálni a következő folyamatot:

Kategória változó esete

Minden megfigyelésre és változóra ki kell számolni az ábra alapján látható értékeket.



3. ábra; Segédábra a sorrend kiválasztásához a fa modellben (saját szerk.: Word)

Az ábrán látható számolás után meg kell vizsgálni a tisztaságát (impurity) a két levélnek és utána magának a változónak.

A tisztaság mértékének megállapítására többféle módszer is létezik, például a Gini Impurity Index, az Entropy és az Information Gain. A mi modellünkben a Gini Impurity Indexet használjuk, mely az egyes levelekre a következőképpen áll össze (James és szerzőtársai, 2013).

$$Gini\ Impurity\ Index_{levél} = 1 - (P(Igazak\ száma)^2) - (P(Hamisak\ száma)^2)$$

11. egyenlet; Gini Impurity Index

Ezután a változóra pedig az úgynevezett Total Gini Impurity Indexet számoljuk ki, amely a levelek Gini indexének súlyozott átlaga.

Total Gini Impurity Index

$$= P(baloldali\ ág) * Gini\ Impurity\ Index_{bal\ levél} + P(jobboldali\ ág) * Gini\ Impurity\ Index_{jobb\ levél}$$

12. egyenlet; Total Gini Impurity Index

Numerikus Változó esetében

Numerikus vagyis folytonos változó révén nincsen eredetileg két részre bontva az adatsor ezért ezt nekünk kell megtennünk.

Kiszámoljuk minden egyes megfigyelés közötti átlagot és az adott átlagok lesznek a választópontok, minden választó ponttal elvégezzük a faktor esetében is említett Gini-eljárást és amelyik választóponttal tudjuk a legkisebb Gini-t produkálni az alapján fogjuk ketté osztani az adatsort, így úgymond faktor változóra alakítjuk a folytonos adatsort (James és szerzőtársai, 2013).

4.4. Végző felépítés

A gyökércsomópontba az a változó kerül, amelynek a legkisebb a Total Gini Impurity indexe. Ezután az a változó kiesik és így folytatjuk végig ismételve a folyamatot és építjük fel a fát.

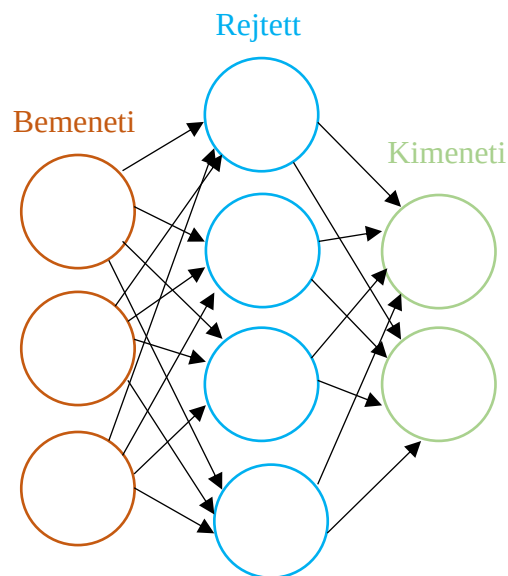
5. A neurális háló módszertana

5.1. A neurális hálózatok felépítése

Az emberi agyban az információ áramlása az agy neuronjainak végtelenül komplex kapcsolatán keresztül történik. Ez a folyamat ihlette a neurális hálózatokat (Fazekas, 2013).

A neurális hálózatok három fő részre bonthatóak, ezek a bemeneti réteg, rejtett réteg vagy rétegek és a kimeneti réteg (Nielsen, 2019). Minden réteget neuronok képeznek, melyek összeköttetésben állnak a szomszédos rétegek összes neuronjával egy nem lineáris aktivációs függvényen keresztül.

Egy kis hálózat szerkezete az alábbi:



4. ábra; Kis neurális hálózat szerkezete (Saját szerk.: Word)

A neurális hálózatok csak numerikus adatokat tudnak feldolgozni, tehát a kategoriális változók átalakításra szorulnak a tanulási folyamat előtt. Az ilyenkor használt kódolás angol neve one-hot encoding, mely majdnem megegyezik a regresszióból ismert dummy-kódolással azt leszámítva, hogy nincs referenciakategória. A megfelelően kódolt változók a bemeneti rétegbe érkeznek, ezen rétegben elhelyezkedő neuronok száma megegyezik a magyarázóváltozók dimenziójával (Bengio és szerzőtársai, 2016). Innen a rejtett rétegekbe megy az információ, majd a kimeneti rétegben jelenik meg a modell becslése, mivel mi bináris klasszifikációs problémával szembesülünk ezért az utolsó rétegben két neuron található.

Minden rejtett és kimeneti rétegben elhelyezkedő neuronnak van egy aktiválási szintje, ez 0-tól 1-ig terjed. Minden ilyen neuron kapcsolatban áll az előtte lévő rétegben elhelyezkedő összes neuronnal, ezt a kapcsolatot számszerűsítik a bemeneti kapcsolatokhoz rendelt súlyok (weights). Ezen kívül az egyes neuronoknak van egyedi úgynevezett eltolássúlyuk (bias). Egy adott neuron aktiválási szintjét tehát három tényező befolyásolja; az előző réteg neuronjainak aktiválási szintje, az összekötő súlyok és az eltolássúlyuk (Altrichter és társai, 2006). Az i -edik réteg neuronjainak aktivitásának matematikája így az alábbi képlettel írható le (Nielsen, 2019):

$$\mathbf{a}^{(i)} = g(\mathbf{W}\mathbf{a}^{(i-1)} + \mathbf{b})$$

13. egyenlet; i -edik réteg neuronjainak aktivitása

Ahol $\mathbf{W}$ a súlymátrix, $\mathbf{a}^{(i-1)}$ az előző réteg aktivitási vektora $\mathbf{b}$ az eltolássúly-vektor, g pedig az aktivációs függvény. A rejtett rétegekben az úgynevezett ReLU „Rectified Linear Unit” a kimeneti rétegben pedig a Softmax aktivációs függvényeket használjuk, melyek képlete a következő (Nielsen, 2019):

$$\max(0, x)$$

14. egyenlet; ReLU aktivációs függvény

$$\frac{e^{x_i}}{\sum_{j=1}^k e^{x_j}}$$

15. egyenlet; Softmax aktivációs függvény

ahol x_i a neuronba érkező megfelelő inputok, k a kimeneti rétegben lévő neuronok száma. A neurális hálók a komplex struktúrájuknak és a nem lineáris aktivációs függvényeknek köszönhetően bármilyen függvényt tetszőleges pontossággal meg tudnak becsülni így kiválóan alkalmazhatóak különféle problémákra.

5.2. A neurális hálózatok tanulása

A regresszióhoz hasonlóan a neurális hálózatok optimalizálásához is szükség van egy költségfüggvényre, amelyet minimalizálunk. Ezekből szintén nagyon sok létezik, bináris klasszifikációhoz azonban a bináris kereszt-entrópia a legalkalmasabb (Nielsen, 2019). Ennek képlete:

$$-\frac{1}{n} \sum_x y * \ln(a) + (1 - y) * \ln(1 - a)]$$

16. egyenlet; Bináris kereszt-entrópia

Ahol y a valós érték, x az input érték, a pedig a kimeneti rétegben lévő neuronok aktiválási szintjei közül a legnagyobb. A költségfüggvény nagy előnye, hogy a tanulás mértéke függ a hiba nagyságától így gyorsítva gyenge kiindulópontok esetén a folyamatot.

Természetesen az egyik legfontosabb változó az optimalizálás folyamatában a modell struktúrájának megválasztása. Komplex adatbázisok nagy dimenziókkal több réteget és neuront igényelnek (James és szerzőtársai, 2013). Az adatbázisunkhoz optimális struktúra megtalálásához sok kísérletezés kell. Mivel a neuronok aktiválási szintje határozza meg a végső outputot a neurális hálók a neuronokat összekötő súlyok és az eltolássúly állítgatásával tanulnak (Nielsen, 2019). Az optimalizáló algoritmusok a gradiens vektorok ($\partial C/\partial w$ és $\partial C/\partial b$) segítségével minimalizálják a költségfüggvényt. A legtöbb optimalizáló eljárás, mint például az általunk használt Adam is az úgynevezett sztochasztikus gradiens süllyedés egy továbbfejlesztése (Reddi, 2018). Ennek lényege, hogy a gradiens vektorokat egy tanuló adathalmazból vett véletlen mintából számolja minden iterációban, így felgyorsítva a folyamatot. Fontos optimalizációs paraméter ezen kívül a tanulási ráta, ez egy konstans melyet a süllyedés mértékének és a számítási idő egyensúlyozására használunk (Nielsen, 2019). Ha a modell megakad valahol érdemes a tanulási rátát kisebbre venni, hiszen előfordulhat, hogy átugrunk egy optimális minimum helyet. A regresszióval ellentétben azonban ebben az esetben az optimalizálási problémának nincs zárt megoldása, így különböző algoritmusokat több iteráción át futtatva fokozatosan kell megtalálni a költségfüggvény minimumát. A nehézséget az okozza, hogy a folyamat könnyen megakad a lokális minimumokon (Gonzalez, 2000).

Gyakran előfordul azonban főleg komplexebb, sok paraméteres hálózatok esetében, hogy a modell túlilleszkedik a tanulóhoz használt adatsorra. Ennek hátránya, hogy a tanuló adatsoron elért nagy pontosság mellett korábban nem látott adatokon a modell nagyon gyengén teljesít. Ennek elkerülésére alkalmaznak úgynevezett regularizációs technikákat. A három legismertebb a L1, L2 és a dropout regularizáció (Nielsen, 2019).

Az L1 és L2 esetében a minimalizálandó költségfüggvényhez adódik egy függvény mely a súlyok méretét hivatott büntetni, így nehezítve a modellnek a tanuló adatsor „memorizálását”. Ennek köszönhetően gyakran javul a modell általános teljesítménye korábban nem látott adatokon. A L2 módszerrel regularizált költségfüggvény az alábbi alakot ölti (Nielsen, 2019):

$$C = C_0 + \frac{\lambda}{2n} \sum_w w^2$$

17. egyenlet; Regularizált költség függvény

Ahol a C regularizált költségfüggvény az alap C_0 függvényből és a regularizáló tagból áll, w pedig az egyes súlyok. $L1$ regularizáció esetében a súlyok abszolútértékeinek összegét adjuk hozzá az alap költségfüggvényhez (Nielsen, 2019).

Dropout regularizáció esetében viszont magán a hálózat felépítésén változtatunk. Minden iteráció előtt a rejtett rétegek neuronjainak egy adott hányadát véletlenszerűen ideiglenesen eltávolítjuk, majd így futtatjuk az algoritmust. Egy heurisztikus magyarázat ezen módszer működésére az, hogy a bennmaradó neuronok nem támaszkodhatnak bizonyos hiányzó neuronokra így az adatsornak robusztusabb „vonásait” kell megtanulniuk (Nielsen, 2019).

A neurális hálók alkalmazásánál tehát jóval több paraméter helyes kiválasztására van szükség, mint a regresszió vagy a döntési fák esetében. Optimalizálásuk így időigényesebb, azonban a legtöbb esetben jobb eredményeket érnek el, mint a többi gépi tanulási algoritmus.

6. Az adatbázis

Adatbázisunkat a nyílt információkkal rendelkező Kaggle nevű honlapról töltöttük le. Az adatok egy 2015-ös felmérés során lettek rögzítve az Amerikai Egyesült Államok összes államában, valamint a különleges státuszú területein (Guam, Puerto Rico, Columbia kerület). Ezek az információk „The Behavioral Risk Factor Surveillance System (BRFSS)” című projektet hívatottak segíteni, a Centers for Disease Control and Prevention (CDC) állami intézmény közreműködésével. A projekt lényege, hogy az ország lakóit viselkedési rizikó faktorok alapján elemzik, és következtetéseket próbálnak levonni a különböző tényezők egészségre gyakorolt hatásáról. A mintavétel aránytalanul rétegzett, tehát az adatbázisban szereplő csoportok valószínűleg nem reprezentatívak.

6.1. A változók bemutatása

<i>Típus</i>	<i>Név</i>	<i>Szintek</i>
Faktor	Szívbetegség vagy szívroham	1-szívbetegség, 0 - nincs
Faktor	Magas vérnyomás	1-magas vérnyomás, 0 - nem magas
Faktor	Magas Koleszterinszint	1-magas koleszterinszint, 0 - nem magas
Faktor	Koleszterin vizsgálat az elmúlt 5 évben	1-volt vizsgálat, 0 - nem volt
Numerikus	Testtömegindex	Folytonos változó
Faktor	Dohányzó	1-dohányzik, 0 - nem dohányzik
Faktor	Korábbi agyvérzés	1-volt korábbi agyvérzés, 0 - nem volt
Faktor	Diabétesz	0 - nem cukorbeteg, 1 - cukorbeteg, 2 - terhesség alatti cukorbeteg
Faktor	Fizikai aktivitás	1- mozgott az elmúlt 30 napban, 0 - nem mozgott
Faktor	Rendszeres zöldség vagy gyümölcs fogyasztás	1 - fogyaszt naponta, 0 - nem fogyaszt naponta
Faktor	Magas alkoholfogyasztás	1-magas alkoholfogyasztás, 0 - nem magas
Faktor	Egészségügyi biztosítással rendelkezik	1 - rendelkezik, 0 - nem rendelkezik
Faktor	Orvosi vizsgálat elmaradása magas költség miatt	1 - igen, 0 - nem
Faktor	Általános egészségi állapot	5 kategória, 1 nagyon rossz, 5 nagyon jó
Numerikus	Mentális állapot	Elmúlt 30 napból hány nap érezte magát rosszul mentálisan
Numerikus	Fizikai állapot	Elmúlt 30 napból hány nap érezte magát rosszul fizikailag
Faktor	Nehézségek sétá közben	1-nehézségek, 0 - nincs
Faktor	Nem	1-férfi, 0 - nő
Faktor	Kor	13 kategória, 18 éves kortól 4 évenként
Faktor	Iskolázottság	6 fokozat, középiskolát nem végzettől egyetemi diplomáig
Faktor	Jövedelem	8 kategória, lásd: 3. táblázat

2. táblázat; Az adatbázis változói

Az adatbázis eredetileg 400 000 megfigyelést tartalmazott és 330 változót. Az adatsor, amivel mi dolgozunk tisztított, nincsenek benne a hibásan kitöltött vagy üresen hagyott válaszokat tartalmazó sorok, valamint a változók száma is szűkített, a felhasznált változók tartalmazzák valamilyen formában a CDC által megállapított fő rizikófaktorokat (Centers for Disease Control and Prevention), melyek a vérnyomás, koleszterinszint, vércukorszint, túlsúly, dohányzás, túlzott alkoholfogyasztás és fizikai aktivitás hiánya.

A fentebb látható táblázatban foglaltuk össze az összes felhasznált változót a modelljeinkben, kezdve a sort a magyarázott változóval, vagyis a szívbetegség vagy szívroham múltbéli tényével. Az első oszlopban az adott változók típusát jelöljük meg, a harmadikban pedig a faktor típusú változók szintjeit ismertetjük.

A folytonos változók a testtömegindex, a mentális, illetve fizikai állapot. A maradék magyarázók kivétel nélkül faktor változók, illetve a magyarázott változó is faktor jellegű.

Úgy gondoljuk, hogy némelyik változó értelmezése nem derül ki egyértelműen a táblázatból, illetve bizonyos esetekben változtatásokat kell végrehajtanunk az egyszerűbb interpretáció érdekében, ezeket magyarázzuk a következőkben.

6.2. A változók értelmezése

A magyarázott változó a „szívbetegség vagy szívroham”, két gyakori szívbetegségre kérdez rá a páciens múltjából. Az egyik a szívkoszorúér-betegség (coronary heart disease), míg a másik miokardialis infarktus (myocardial infarction). Az előbbi abban az esetben fejlődik ki, amennyiben szívtartériák nem képesek elég oxigénben dús vért szállítani a szívbe, ez a betegség az egyik vezető halálozási ok a tengerentúlon. Utóbbit a köznyelvben szívrohamnak is nevezik, akkor fordul elő, amikor a szívizom vérellátásában hiány lép fel, általában az egyik szívtartéria elzáródása okozhatja, sajnos sok esetben ez is halállal végződik. A megkérdezettek 9,42 százaléka szenvedte már el legalább az egyik betegséget.

Fontos megemlíteni, hogy a magas vérnyomás, magas koleszterinszint, korábbi agyvérzés, cukorbetegség esetében akkor került rögzítésre valamely betegség múltbéli ténye az adatbázisba amennyiben egy szakember állapította meg a diagnózist. A cukorbetegség esetében a terhesség alatti cukorbetegség is jelölve van. A magas vérnyomás, illetve a magas koleszterinszinttel rendelkező betegek aránya egyaránt 42 százalék, az agyvérzésen átesetteknek aránya 4 százalék, illetve az egyének 84 százaléka nem szenved cukorbetegségben.

A testtömegindex, ismertebb nevén BMI index, az illető testfelépítéséről ad egy értéket. Kiszámítása az alábbi egyszerű képlet alapján történik. A megkérdezettek átlagos testtömegindexe 28, a medián 27, viszonylag alacsony 23 százalékos relatív szórással.

$$BMI = \frac{Testtömeg[kg]}{testmagasság^2[m]}$$

18. egyenlet; Testtömegindex

A rendszeres zöldség és gyümölcs fogyasztás eredetileg két különálló változó volt, mi viszont összevonva vizsgáljuk őket. Abban az esetben tekintünk valakit rendszeres fogyasztónak, ha naponta legalább egy gyümölcsöt vagy zöldséget fogyaszt. A válaszadók 88 százaléka fogyaszt naponta zöldséget vagy gyümölcsöt.

Abban az esetben kategorizálunk valakit magas alkohol fogyasztóként, ha férfiként hetente több mint tizennégy pohár, illetve nőként több mint hét pohár italt fogyaszt (National Institute on Alcohol Abuse and Alcoholism (NIAAA)). A magas alkohol fogyasztók aránya rendkívül alacsony, 5,6 százalék.

Egy válaszadót dohányzónak tekintünk, ha élete során már több mint száz szál cigarettát elszívott, ami öt doboznak felel meg. Az emberek 44 százaléka tekinthető dohányzónak az általunk vizsgált adatbázisban.

A mentális és fizikai állapot változók azt mutatják, hogy a válaszadó a kérdőív kitöltését megelőző 30 napban hány napon érezte magát rosszul mentálisan, illetve fizikailag.

Az általános egészségi állapot 5 kategóriát tartalmaz. A válaszadók önbevallásos alapon választottak szintet, sorrendben a kiváló, nagyon jó, jó, elfogadható, rossz lehetőségek közül választhattak.

Az iskolai végzettség változója 6 kategóriába van rendezve. Az első szint az iskolát nem végeztek, a második az általános iskolát végeztek csoportja. Ezután a valamennyi időt középiskolában töltők (9.-11. tanévük között) következnek, majd a középiskolát teljesen elvégeztek kategóriája. Az ötödik szint az egyetemi, főiskolai, technikai alapképzést jelöli, az utolsó csoport pedig a 4, vagy annál több éves egyetemi képzés vagy képzésekkel (összeadva) rendelkező személyeket tartalmazza.

A jövedelem faktor változóinak dollárban mért szintjeit az alábbi táblázatban foglaljuk össze, éves adatokkal.

1. szint	kevesebb, mint 10 000 \$
2. szint	10 000 \$-tól 15 000 \$-ig
3. szint	15 000 \$-tól 20 000 \$-ig
4. szint	20 000 \$-tól 25 000 \$-ig
5. szint	25 000 \$-tól 35 000 \$-ig
6. szint	35 000 \$-tól 50 000 \$-ig
7. szint	50 000 \$-tól 75 000 \$-ig
8. szint	több, mint 75 000 \$

3. táblázat; A jövedelmi változó fokozatai

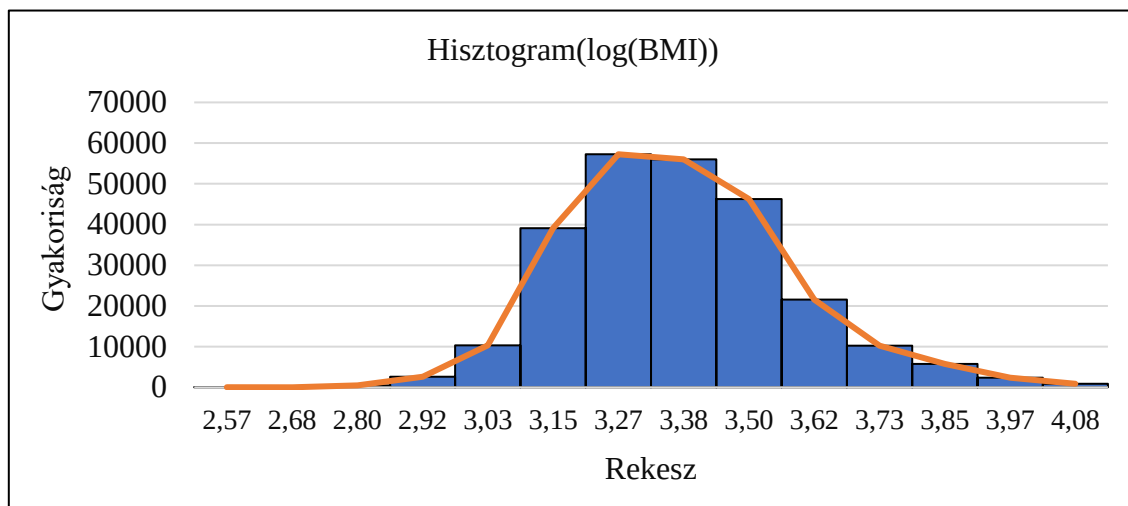
7. Leíró statisztikai elemzés

Az adatok bővebb ismertetése érdekében ezen fejezetben egy rövid áttekintést nyújtunk a megfigyelések leíró statisztikai elemzéséről. Mivel három változónk van csak, amelyek numerikusak, ezért legtöbbször csak arányokat mutatunk be.

A kutatáson megkérdezettek 9,42 százaléka szenvedett már el legalább az egyik típusú szívbetegséget. Ez a CDC becslése alapján körülbelül reális arány, hiszen országos szinten 8,3 százalékot mértek.

A magas vérnyomás, illetve a magas koleszterinszinttel rendelkező betegek aránya egyaránt 42 százalék, az agyvérzésen átesetteknek aránya 4 százalék, illetve az egyének 84 százaléka nem szenved cukorbetegségben. Ezek szintén nagyjából megközelítik a valósi arányokat.

A megkérdezettek átlagos testtömegindexe 28, a medián 27, viszonylag alacsony 23 százalékos relatív szórással. Mivel az átlag nagyobb, mint a medián, ezért jobbra elnyúló a BMI eloszlása, ami logikus tekintve, hogy nagyon magas index csak ritkán fordul elő. Mivel a testtömegindex logaritmikus alakjának hisztogramját is megvizsgáltuk, mely normális eloszlást közelített meg, így a továbbiakban a sikeres modellezés érdekében ebben a formájában használjuk a változót.

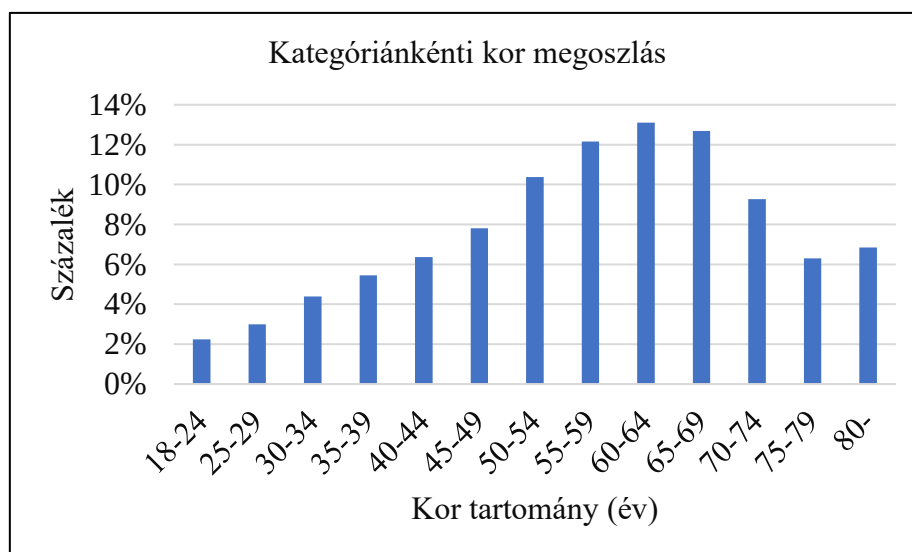


5. ábra; A testtömegindex hisztogramja (saját szerk.: Excel)

A válaszadók 88 százaléka fogyaszt naponta zöldséget vagy gyümölcsöt. Ez véleményünk szerint nem reális arány (tekintve a fejlett országok modern étkezési szokásait), így ez torzító hatással lehet az eredményekre.

A magas alkohol fogyasztók aránya rendkívül alacsony, 5,6 százalék, ez szintén alacsonynak tűnik a valósághoz képest. Az emberek 44 százaléka tekinthető dohányzónak, ez az alkohollal szemben viszont igencsak felülbecsülheti a realitást.

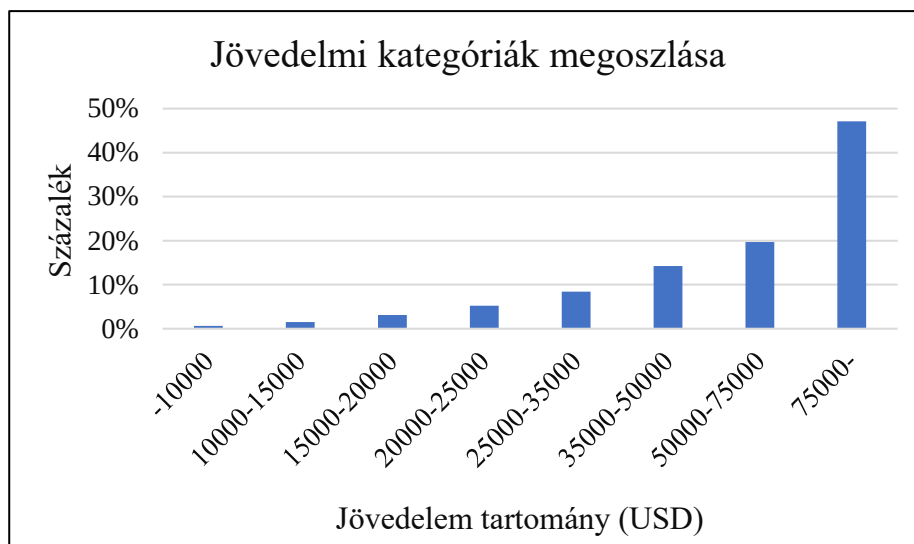
Az emberek 69 százaléka egyetlen napon sem érezte magát negatívan mentálisan, illetve a válaszadók 63 százaléka sem érezte magát rosszul fizikailag. Általánosan elmondható, hogy mindkét kategóriában kevesebb azon páciensek száma, akik az elmúlt 30 napban számottevő ideig voltak rossz állapotban fizikai vagy mentális szempontokból.



6. ábra; Az életkor megoszlása kategóriánként (saját szerk.: Excel)

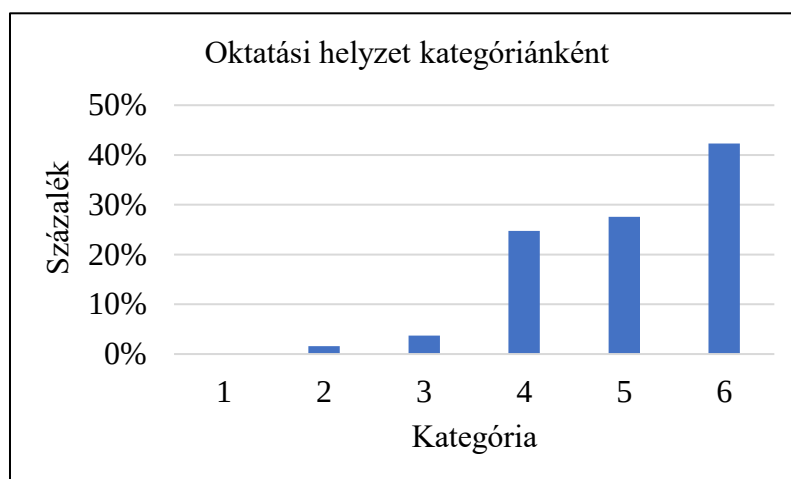
A hatodik ábra bemutatja a kor kategóriánkénti megoszlását egy hisztogramon. Az eloszlás balra elnyúlik, és 55 és 69 év között vesz fel magasabb értékeket, így arra következtethetünk, hogy a középkorúak, és az idősebbek vannak felülreprezentálva az adatbázisunkban.

A jövedelmi kategóriákat szintén hisztogramon ábrázoljuk, ahogyan az fentebb látható is. A



7. ábra; Jövedelmi kategóriák megoszlása (saját szerk.: Excel)

meglepő eredmény az, hogy ahogyan nő a jövedelem szintje, annál nagyobb a hozzátartozó emberek aránya, ez véleményünk szerint nem reprezentatív, azonban az intervallumok eltérő szélessége is közrejátszik a kimenetben.



8. ábra; Oktatási helyzet kategóriánként (saját szerk.: Excel)

Végül az oktatási helyzet kimeneteit is ábrázoljuk hisztogramon. Az ábrára tekintve egyből szembe tűnhet, hogy a 3 magasabb szinthez sokkal több ember tartozik. Ez nem meglepő hiszen középiskolai végzettsége a fejlett országok középkorú lakosainak túlnyomó részben van, ám érdekes, hogy a hosszabb egyetemi képzést kapott személyek aránya a mintánkban meghaladja a 40 százalékot.

A változók rövidebb leíró típusú áttekintése átfogó képet nyújt a fontosabb statisztikai tulajdonságaikról. Ezek vizsgálata után elkezdhető a modellek kialakítása, elemzése.

8. A logisztikus modell

Ebben a fejezetben a logisztikus regressziós modellünket mutatjuk be. Mivel az eredményváltozónk, a szívbetegség kialakulása, binomiális kimenetelű, ezért választottuk ezt az elemzési formát. A következőkben a modell előkészítését, kialakítását, végeredményeit analizáljuk az R szoftver segítségével, a módszertani részben leírt ökonometriai formulákkal.

Az adatbázisunkat egy teszt és tanuló halmazra bontjuk, redundáns változók szűrése után logisztikus regressziókat futtatunk, majd szignifikancia, McFadden R^2 , AUC és közgazdasági logika alapján alakítjuk a magyarázó változókat a modellben.

Mielőtt elkészítenénk a regressziós modellt az adatbázist két részre bontjuk, egy úgynevezett tanuló és egy tesztelő részre. A gyakorló rész a megfigyelések 90 százaléka melyet véletlen mintavétel segítségével kapunk meg az eredeti mintánkból, vagyis ez a rész 228 312 megfigyelést tartalmaz és értelemszerűen a maradék 25 368 adat pedig a teszt részbe kerül. Erre a módszerre azért van szükség, hogy elkerüljük a túlillesztést, mert a modellt a gyakorló adatokon építjük meg, viszont a klasszifikációt már a tesztelő adatokon végezzük.

8.1. A végső logisztikus modell

Az előzetes számítások alapján az alábbi ábrán látható a végső, felhasználásra alkalmas modellünk, változók szerint lebontva. A kihagyott változókat Wald-F próbával teszteltük, a számítások alapján nem veszített magyarázóerőt a modellünk.

<i>Végső logisztikus modell</i>	
	Eredményváltozó: Szívbetegség
Konstans	-7,195*** (0,247)
Magas vérnyomás	0,526*** (0,019)
Magas koleszterin	0,609*** (0,017)
Kol. vizsgálat	0,527*** (0,069)
Dohányzó	0,363*** (0,017)
Stroke	1,000*** (0,026)
Cukorbeteg	0,033 (0,049)
Terhesség alatti cukorbeteg	0,278*** (0,019)
Fizikai aktivitás	0,044** (0,018)
Gyümölcs/zöldség fogy.	0,021 (0,023)

Magas alkohol fogyasztás	-0,312*** (0,042)
Vizsgálat mulasztása költs. miatt	0,253*** (0,028)
Ált. egészség 2	0,429*** (0,039)
Ált. egészség 3	0,978*** (0,038)
Ált. egészség 4	1,479*** (0,041)
Ált. egészség 5	1,880*** (0,048)
Mentális egészség	0,002** (0,001)
Fizikai egészség	0,003*** (0,001)
Nehézségek séta közben	0,296*** (0,020)
Nem	0,757*** (0,017)
Kor 2	0,154 (0,240)
Kor 3	0,491** (0,213)
Kor 4	0,593*** (0,205)
Kor 5	0,896*** (0,199)
Kor 6	1,238*** (0,195)
Kor 7	1,494*** (0,193)
Kor 8	1,703*** (0,192)
Kor 9	1,979*** (0,192)
Kor 10	2,254*** (0,192)
Kor 11	2,537*** (0,192)
Kor 12	2,687*** (0,192)
Kor 13	2,981*** (0,192)
Jövedelem 2	-0,004 (0,043)
Jövedelem 3	-0,046 (0,042)
Jövedelem 4	-0,072* (0,041)
Jövedelem 5	-0,143*** (0,041)
Jövedelem 6	-0,144*** (0,040)
Jövedelem 7	-0,189*** (0,040)
Jövedelem 8	-0,267*** (0,039)
Log (BMI)	0,035 (0,041)
Megfigyelések	228,312
Log Likelihood	-54,334,050
Akaike Inf. Kritérium	108,748,100
Megjegyzés:	*p<0,1; **p<0,05; ***p<0,01

4. táblázat; A végső logisztikus modell

Az ábrára tekintve láthatjuk az egyes változókat, valamint a hozzájuk tartozó becsült paramétereket és standard hibákat. Ahogyan az a megjegyzésben is látható (a módszertanban leírtak alapján) a kettő és három csillaggal jelölt bétákat tekintjük szignifikánsnak.

A nem szignifikáns változók a zöldség és gyümölcs fogyasztás, illetve a logaritmizált testtömegindex ($\text{Log}(\text{BMI})$), viszont mi mégsem távolítjuk el őket a végső modellből, mivel úgy gondoljuk, hogy a valóságban befolyásolják a szívbetegségek kialakulásának valószínűségét, illetve kivételük esetén a klasszifikációs pontosság csökkent, de annak a valószínűsége is fennáll, hogy másodfajú hibát követünk el a szignifikanciaszintnél.

A cukorbetegség, a kor, illetve a jövedelmi magyarázók bizonyos szintjei nem térnek el szignifikánsan a referencia kategóriától, azonban a többi szignifikáns 5 százalékos szignifikancia szinten, ezért nem távolítjuk el őket a modellből. Az is fontos tényező ebben az esetben, hogy egyes kategóriákban nem a valóságnak megfelelő arányban szerepelnek adatok.

A paramétereket tekintve vannak olyanok, amelyek szembemennek az előzetes elvárásainkkal. Ilyen például a fizikai aktivitás pozitív, az alkohol fogyasztás negatív, valamint az általános egészségi állapot pozitív hatása a szívbetegség kialakulásának feltételes valószínűségeire. Ezeken kívül minden változó általunk logikusnak gondolt irányba befolyásolja az eredményváltozót.

A modell eredményességét globális szinten is vizsgáljuk, erre McFadden-típusú R^2 -t használunk. Ennek a kiszámított értéke 23,68 százalék. Ez azt jelenti, hogy a modellünk ennyiben ad jobb becslést, mint az üres modell. Ez nem mondható jó értéknek, azonban annyira rossznak sem, mint ha csak az üres modell (a betegek és egészségesek aránya) alapján számítanánk.

8.2. A logisztikus modell klasszifikációja

Végül annak érdekében, hogy a teszt adatokat meg tudjuk becsülni, vagyis, hogy a változók alapján be tudjuk sorolni őket szívbeteget vagy egészséges kategóriába, szükség van egy vágási értékre, ami alapján a feltételes valószínűségeket konkrét becsléssé változtatjuk.

8.2.1. Klasszikus vágási érték

Klasszikusan az 50 százalékos vágási értéket szokták alkalmazni, így mi is ezt használjuk először.

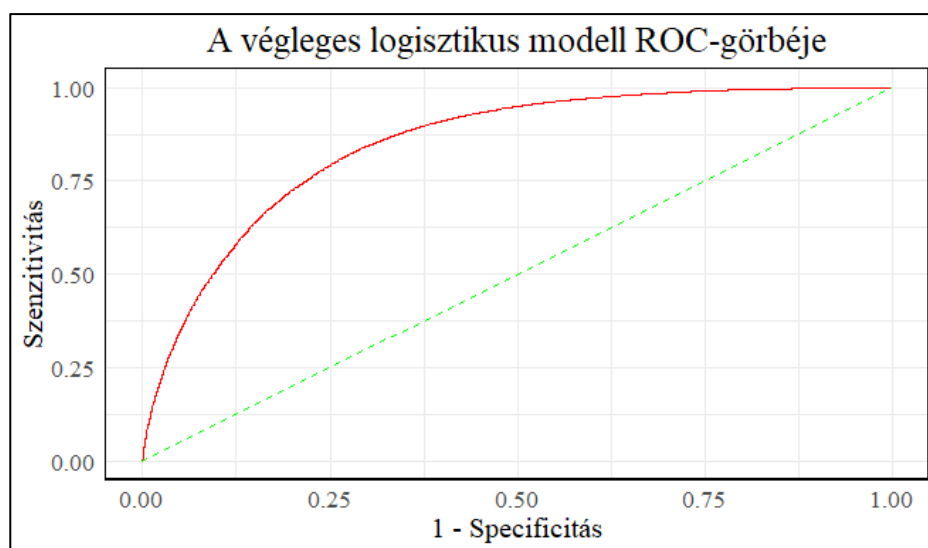
		Becsült értékek			
		0	1	Pontosság	90,95%
Valós értékek	0	22768	250	1-TPR = FNR	87,11%
	1	2047	303	Szenzitivitás (TPR)	12,89%

5. táblázat; Logisztikus modell klasszifikációs tábla (C = klasszikus)

Látható a táblázat alapján, hogy habár a klasszifikáció pontossága viszonylag magas, a szenzitivitás, vagyis valódi pozitív ráta viszont nagyon alacsonyon, ezáltal a hamis negatív pedig magas. Optimális esetben fordított értéket kellene ezen két mutatónak felvennie.

8.2.2. Youden-szabály

Az alábbi jobbnak tekinthető módszer a Youden-szabály alapján választja meg a vágási értéket, vagyis úgy, hogy a specificitást és a szenzitivitást maximalizálja. Az alábbi ábrán látható a logisztikus regresszió ROC-görbéje:



9. ábra; A végleges logisztikus modell ROC-görbéje (sajátszerk.: RStudio)

A görbe alatti terület 0,8474, tehát viszonylag jó klasszifikációs erővel bír a modell. A Youden-szabály alapján az optimális vágási érték 8,22 százalék, vagyis azon pácienseket, akiknek a változók alapján a feltételes valószínűségük a szívbetegségre magasabb, mint ez az érték, szívbetegként klasszifikáljuk.

		Becsült értékek			
		0	1	Pontosság	72,91%
Valós értékek	0	16536	6482	1-TPR = FNR	16,60%
	1	390	1960	Szenzitivitás (TPR)	83,40%

6. táblázat; Logisztikus modell klasszifikációs tábla (C = Youden-szabály)

Az 50 százalékos klasszifikációhoz képest láthatóan csökkent a pontosság, viszont a valódi pozitív ráta óriási mértékben megnőtt, amely jóval lényegesebb, illetve ezáltal a hamis negatív ráta lecsökkent.

Elmondható tehát, hogy a Youden-szabály alapján megválasztott vágási érték az FNR szempontjából jobb klasszifikációt eredményez, azonban a pontosság lecsökkenése és a hamisan beteg kategóriába sorolt egészségesek száma miatt valószínűleg nem a logisztikus regresszió a legmegfelelőbb a problémához.

Véleményünk szerint a tévesen betegnek minősített (FP) alanyok magas száma azt eredményezheti, hogy az emberek nem veszik kellően komolyan a teszt kitöltését és valószínűleg kevésbé hajlandóak egy ilyen teszt alapján változtatni az életmódjukon vagy további vizsgálatokon átesni.

9. A véletlen erdő modell

A logisztikus regressziós modell után döntési fákkal közelítjük meg a klasszifikációs problémát. Az ilyen algoritmusok előnye, hogy nagyon könnyen lehet intuíciót szerezni működésükről, jól ábrázolhatók így laikusok által is értelmezhetők. Jól kezelik a kategoriális változókat dummy változók létrehozása nélkül. Potenciális hátrányuk, hogy gyakran nem robusztusak, vagyis egy apró eltérés az inputban nagy különbségekhez vezet az outputban.

Ebben a fejezetben a véletlen erdő (random forest) nevű algoritmus használjuk. Ez az általános döntési fák egy továbbfejlesztett változata. Ahogy a neve is sugallja itt több fával dolgozunk, kategorizációs probléma esetén a „növesztett” fák szavaznak arról, hogy az adott megfigyelés melyik csoportba tartozik. Véletlennek nevezzük őket mert minden egyes fának minden ágánál a magyarázó változók egy adott elemszámú találomra választott halmazát vizsgálja, a felosztásnál csak ezen halmaz elemeit veheti figyelembe. Így a fák valamilyen mértékben egymástól függetlenné tehetők. Az ilyen erdők pontosabb becslést kínálnak, és a szavazásnak köszönhetően a legtöbb esetben robusztusabbak, mint egy fa (James és szerzőtársai, 2013). A módszer hátránya, hogy nagy számítási teljesítményt igényel.

A logisztikus regresszióhoz hasonlóan az adatbázist itt is két részre bontottuk, az adatok 90 százalékán futtatjuk a modellt, 10 százalékát pedig becslésre használjuk. A véletlen erdő építéskor két fontos paraméterünk van, a fák száma és a változók halmazának nagysága. Empirikus eredmények alapján az ágaknál vizsgált változók legtöbbször optimális száma a magyarázóváltozók számának négyzetgyöke. 20 magyarázóváltozónk van, felfele kerekítés után 5-tel dolgozunk ebben a fejezetben. Ezt a számot keresztvalidációval is megfelelőnek találjuk. A fákat illetően pedig általánosan kijelenthető, hogy számuk növekedésével nő a becslés pontossága, azonban egy tanulmány melyben 29 adatbázist vizsgáltak (Oshiro – Perez Baranauskas, 2012) kimutatta, hogy 128 fa felett nagyjából 4000 fáig nem nő szignifikánsan a

pontosság, így mi is ennyit használunk. A modell becslését (a 0 ismét egészséges, 1 a beteg) a teszt adatsoron az alábbi klasszifikációs táblázat foglalja össze:

		<i>Becsült értékek</i>			
		0	1	Pontosság	89,99%
<i>Valós értékek</i>	0	22466	538	1-TPR = FNR	84,64%
	1	2001	363	Szenzitivitás (TPR)	15,36%

7. táblázat; Első véletlen erdő modell klasszifikációs táblája

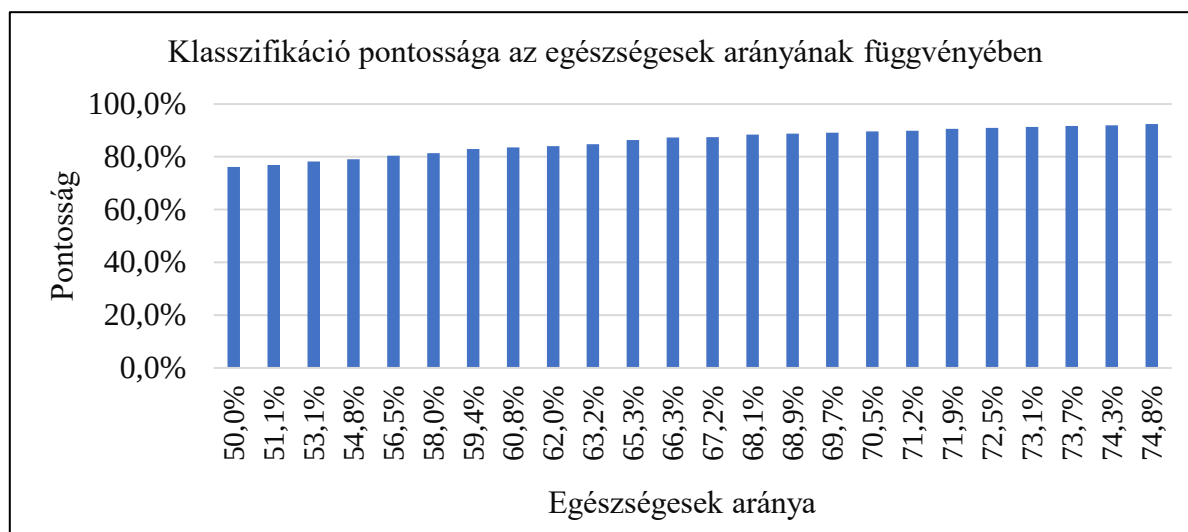
Láthatjuk, hogy a modell 90 százalékos pontossággal klasszifikálja az alanyokat, azonban ez az adat félrevezető, mivel a tesztelésre használt adatsorban 90,68 százaléka a megfigyeléseknek egészséges. Ez azt jelenti, ha minden egyes megfigyelést az egészséges kategóriába sorolunk akkor ennél nagyobb pontosságot érünk el. Ebből is sejthető, hogy a modellünk nem megfelelő. A modell nehezen birkózik meg a betegek helyes klasszifikálásával, nagyon sok másodfajú hibát ejtünk. Fontos ismét megemlíteni, hogy a klasszifikáció pontos költségfüggvényét nem ismerjük, de mivel szívbetegek kialakulásának észlelését szeretnénk megkönnyíteni pont az efféle hiba járhat a legsúlyosabb következményekkel. Az eredmények alapján elmondható, hogy sima véletlen erdővel a probléma a jelenlegi formájában nem megoldható.

A gyenge teljesítményt az adathalmazban lévő beteg és egészséges emberek erősen eltérő aránya okozza. A fák ágainak létrehozásakor használt Gini Impurity Indexet sikerül minimalizálni, azonban az egészséges alanyok nagy száma miatt ezt úgy érjük el, hogy a legtöbb megfigyelést egészségesnek minősítjük. Ezen okból a modell nem használható a probléma megoldásához, érdemes változtatásokat eszközölnünk.

9.1. Arányok megváltoztatása

Az előző bekezdésben tárgyalt problémát úgy lehet megoldani, hogy kiegyenlítjük a két csoport arányát a tanuló adatbázison belül. Ezt nyugodtan megtehetjük, mivel a véletlen erdőknek nincsenek olyan szigorú modellfeltételei mint a logisztikus regresszióknak, valamint a változók fontossága ebben az esetben másodlagos számunkra, így a jobb klasszifikáció érdekében használhatunk ilyen módszereket. A módosított tanuló adatbázisban megtartjuk az összes beteg megfigyelést, az egészségesek számát pedig 20891-ről – vagyis 50 százalékról – indulva növeljük. Minden arány mellett tízszer futtatjuk a modellt tíz különböző azonos nagyságú teszt mintán, majd ezek eredményét átlagoljuk.

Az alábbi két diagrammon a pontosság és az FNR értékeinek változását láthatjuk az egészségesek arányának függvényében:



10. ábra; A klasszifikáció pontossága az egészségesek arányának függvényében (saját szerk.: Excel)



11. ábra; Az FNR értékeinek változása az egészségesek arányának függvényében (saját szerk.: Excel)

Nem meglepő módon ahogyan közeledünk az eredeti arányok felé egyre jobban nő a pontosság. Az FNR értéke nem korrelál ilyen erősen az egészségesek arányával. A minimuma 54,8 százalékos arány, vagyis 26 ezer egészséges mellett van, azonban a modell kiértékelésénél a 58 ezer egészséges megfigyelést tartalmazó adathalmaz mellett döntöttünk, mivel az FNR értéke itt nem tér el jelentősen a minimumtól, de a pontosság jóval nagyobb.

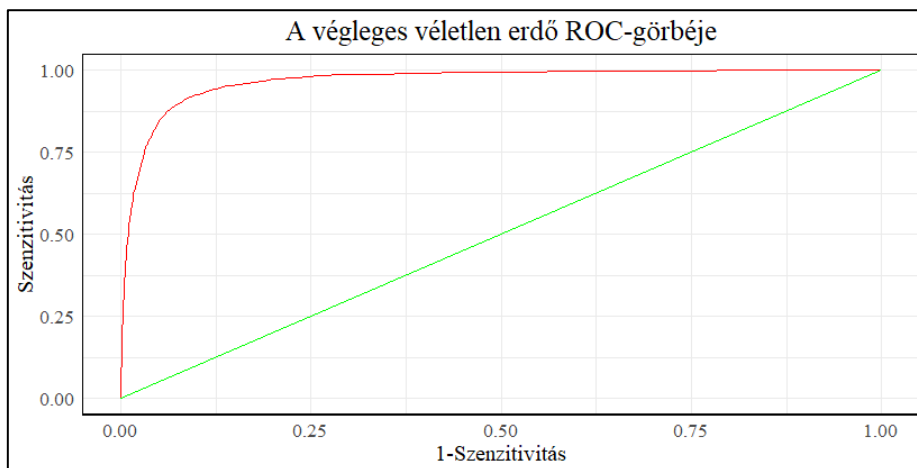
9.2. A végső véletlen erdő modell

A modellt ezen adathalmazon futtatva majd az eddig használt teszt adatsoron ellenőrizve az alábbi klasszifikációs táblát kapjuk:

		Becsült értékek		Pontosság	91,08%
		0	1		
Valós értékek	0	20870	2047	1-TPR = FNR	8,85%
	1	217	2234	Szenzitivitás (TPR)	91,15%

8. táblázat; A végső véletlen erdő modell klasszifikációs táblája

A végső modell ROC-görbéje pedig a következő:



12. ábra; A végleges véletlen erdő modell ROC-görbéje (saját szerk.: RStudio)

A görbe alatti terület 0,9462 ez az érték jóval nagyobb, mint a logisztikus regresszió esetében számolt 0,8474 tehát ez alapján a modell jó klasszifikáció erővel bír.

Összevetve az eredményeket az eredeti adatbázison futtatott véletlen erdővel láthatjuk, hogy mindhárom vizsgált mutató javult. A becslés pontossága növekedett, jobban klasszifikálja a modell a beteg megfigyeléseket. Az FNR értéke a korábbi eredményekhez képest elenyésző tehát kevés beteg embert sorolunk rossz csoportba. A helytelenül betegek csoportjába sorolt alanyok száma növekedett, így egy trade-offal szembesülünk. Az efféle hiba azonban várhatóan sokkal kisebb költségekkel járhat, mint ha valakiről azt mondjuk, hogy egészséges miközben nem, tehát ezt a modellt tekintjük a legjobbnak.

10. A neurális háló modell

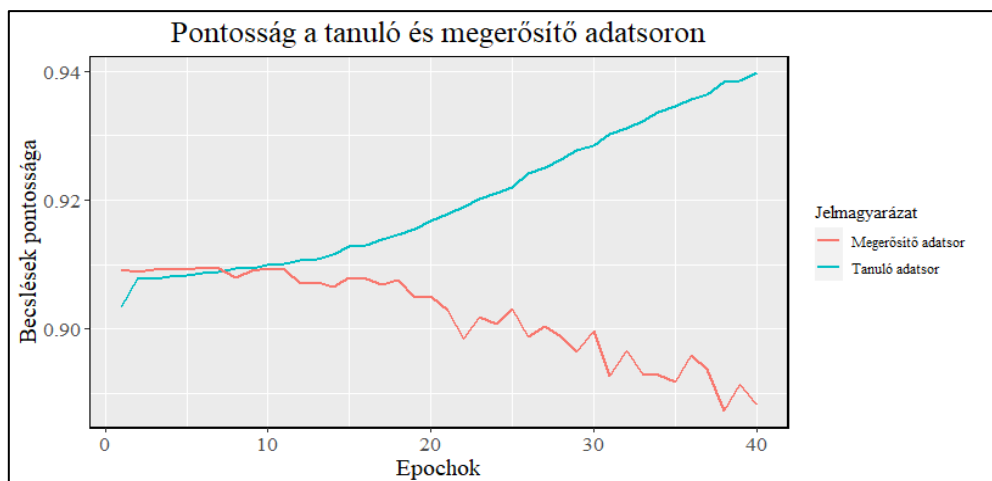
A harmadik és egyben utolsó módszer, amit alkalmazunk a klasszifikációs problémákra a neurális hálózat, melyet napjainkban a legkülönbözőbb problémák megoldására használnak, mint például képfelismerés, önvezető autók, nyelvfeldolgozás stb. A neurális hálózatok komplex, rengeteg paramétert tartalmazó gráf alapú rendszerek, melyekben az információ rétegekbe rendezett mesterséges neuronokon keresztül halad át (Nielsen, 2019). Az előző két fejezetben tárgyalt modellekkel ellentétben a paraméterek megadására és a tanulás optimalizálására nincs egzakt módszer, matematikáját a mai napig kutatják ezért is hívják gyakran „fekete dobozoknak”. Tökéletesítésükhöz sok kísérletezés és heurisztikus megközelítés kell.

Az ebben a fejezetben alkalmazott neurális hálózatban használt struktúra és paraméterek kiválasztása rengeteg empirikus tanulmány és kísérletezés eredménye. A rejtett rétegek és ezzel együtt a tanítható paraméterek számának növekedésével a modell egyre komplexebb adatsoron képes tanulni, azonban a túl nagy szerkezet használata a számítási idő megnövekedésén túl az általánosítási képesség romlásával járhat. Kis hálózatok esetén pedig fennáll a lehetőség, hogy megragadnak kedvezőtlen lokális minimumoknál így nem képesek tanulni az adathalmazból (Gonzalez, 2000).

A modellhez szükséges átalakításokat elvégezve a magyarázó változók mátrixa 20×253680 -ról 62×253680 dimenziósra módosult, tehát a bemeneti rétegben 62 neuron található. A dimenzió nagyságából azt feltételezhetjük, hogy egy kis modell nem alkalmas a probléma megoldásához.

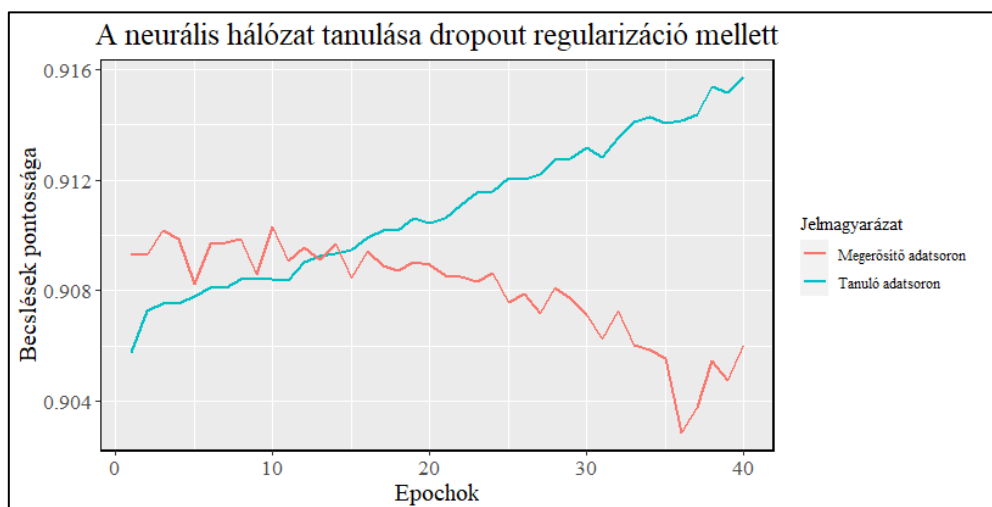
Minden futtatott modell esetében a rejtett rétegekben ReLU, a kimeneti rétegben pedig Softmax aktivációs függvényt alkalmazunk az Adam optimalizálóval. Használni fogjuk ezen kívül AMSgrad optimalizációs módszert (Reddi-Kale-Kumar, 2018), mely az Adam és ehhez hasonló optimalizálók konvergenciáját hivatott javítani.

Ahogy a döntési fáknál itt is beleütközünk abba a problémába, hogy az egészséges és beteg páciensek aránya nagyon különböző. Ennek köszönhetően egy egyszerűbb hálózat a teszt adatsorban lévő alanyokat nagyrészt az egészséges kategóriába sorolja. Egy komplex, sok paraméterű hálózatnál azonban az alábbi ábrán látható problémát tapasztaljuk:



13. ábra; A becslések pontossága tanuló és megerősítő adatsoron neurális háló esetén (saját szerk.: RStudio)

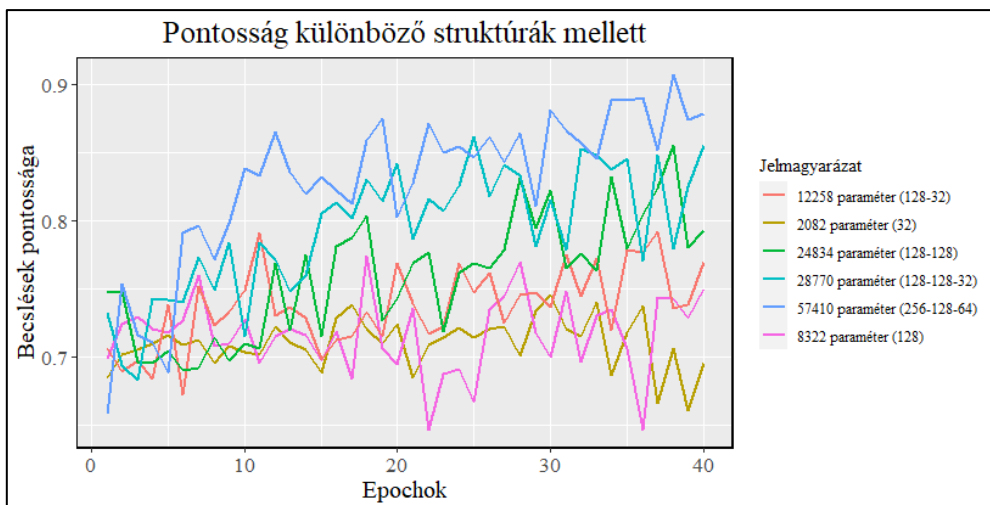
Látható, hogy a tanuló és megerősítő adatsorokon számolt pontosság hamar elválk, a hálózat elkezd „megtanulni” a tanuló adatsort így az általánosító képessége jelentősen lecsökken, ez látható a piros vonalon. Regularizációs módszerek (L1, L2, dropout) alkalmazása mellett pedig nem figyelhetünk meg valódi tanulási folyamatot, ráadásul kisebb mértékben ugyan, de egy pont után a megerősítő adatsoron ebben az esetben is egyre gyengébb eredményeket érünk el iterációként. Az alábbi ábra ezt szemlélteti dropout regularizáció mellett:



14. ábra; A neurális hálózat tanulása dropout regularizáció mellett (saját szerk.: RStudio)

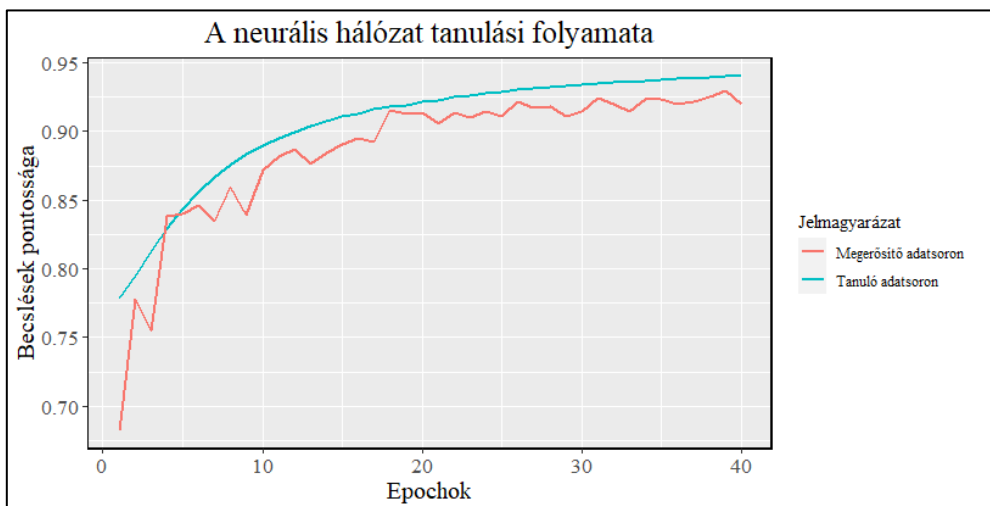
A probléma orvoslására ismételten módosítjuk az adatsort, ebben az esetben a betegekből veszünk véletlen mintákat amíg ki nem egyenlítődnek az arányok. Ehhez az R ROSE (Random Over-Sampling Examples) könyvtárát alkalmazzuk. Az új tanuló adathalmazunk így 413616 elemű lesz, benne fele-fele arányban vannak egészségesek és betegek. A neurális háló futtatásakor a tanuló adathalmaz 20 százalékát megerősítő adatsorként használjuk, hogy a modell futtatása közben lássuk a hálózat által nem látott adatokon a pontosság alakulását. Az

itt alkalmazandó struktúra szintén nem egyértelmű, néhány opciót a megerősítő adatsoron mért pontossággal az alábbi ábra tartalmaz:



15. ábra; A neurális háló pontossága különböző struktúrák mellett (saját szerk.: RStudio)

A jelmagyarázatban fel van tüntetve az egyes struktúrákban a paraméterek száma, valamint a zárójelekben a neuronok elrendezése a rejtett rétegekben. Látható, hogy a legjobb növekvő trendet a komplexebb struktúrájú modellek mutatják, míg az egyszerűbbek egy alacsonyabb érték körül ingadoznak. Ez alapján érdemes a bonyolultabb hálózatokkal kísérletezni. A végső neurális hálózat tanulási folyamatát ábrázolja az alábbi ábra:



16. ábra; A neurális hálózat tanulási folyamata (saját szerk.: RStudio)

Látható, hogy a tanuló és megerősítő adatsoron mért pontosság iterációként növekvő trendet mutat és szépen együtt mozognak. Ezek alapján a modellnek jó az általánosító képessége, nem beszélhetünk túlillesztésről így nem kell regularizációt alkalmazni.

10.1. A végső neurális háló modell

A végső modellt először 60 iteráción át fejlesztjük, megakadás esetén a tanulási rátát csökkentve és a véletlen minták elemszámát növelve. Így végül 160 iteráció alapján elkészült modell összefoglaló táblázata és a klasszifikációs táblázat az alábbiak:

Végső neurális hálózat		
Struktúra:	3 rejtett réteg; 512-256-128 neuron	
Paraméterek száma:	196738	
Optimalizáló:	Adam	
Véletlen minta nagyság:	32, 60 iteráció után 200	
Tanulási ráta:	0.001, 60 iteráció után 0.00001	
Pontosság	FNR	Szenzitivitás
95,92%	4,10%	95,90%

9. táblázat; A végső neurális hálózat

		<i>Becsült értékek</i>	
		0	1
<i>Valós értékek</i>	0	21763	926
	1	98	2293

10. táblázat; A végső neurális háló modell klasszifikációs táblája

Ezek az eredmények természetesen a változatlan, eredeti arányok melletti teszt adatsoron elért klasszifikációt mutatják. Láthatjuk, hogy a pontosság elég magas, 95,92 százalék, az FNR pedig 4,1 százalék tehát kevés a szívbetegedések szempontjából feltehetőleg súlyosabb másodfajú hibát vétünk. Az elsőfajú hiba aránya is viszonylag alacsony, tehát kevés alanyak jelezné a modellt a további vizsgálatok szükségességét miközben az illető nem veszélyeztetett. Költséghatékonysági szempontból pontos költségfüggvény ismerete nélkül nem lehet több megállapítást tenni a modell eredményeiről.

11. A végső modellek összehasonlítása

Ebben a fejezetben a három különböző módszertannal rendelkező modellt, a logisztikus, a véletlen erdő és a neurális hálós algoritmust fogjuk összehasonlítani releváns mutatók alapján. Ehhez a korábban ismertetett klasszifikációs táblák pontossági mutatóit használjuk fel.

Mindhárom modellnek megvannak a maga előnyei és hátrányai. A logisztikus regresszió egyszerűen létrehozható, valamint az optimalizációs problémának zárt megoldása van, pontosan kiszámolhatóak a feltételes valószínűségek, és a magyarázó változók hatása is egyértelműen megadható az eredményváltozóra. Hátránya viszont, hogy szigorúak a modellfeltevései, amelyek nem biztos, hogy mindig teljesülnek, valamint nem igazán robosztus a viselkedése outlier adatok esetén (Wooldridge, 2012).

A véletlen erdő modell nagyon könnyen interpretálható, akár laikusok számára is. Szintén előnye, hogy nem kell dummy változókat létrehozni, azonban ez a módszer sem igazán robosztus, valamint sok számítási kapacitást vesz igénybe (James és szerzőtársai, 2013).

Végül pedig a neurális hálózat hátránya, hogy a kategoriális változók átalakításra szorulnak, így nagyon megnő a magyarázó változók dimenziója. Szintén negatívuma, hogy a paraméterek végtelen kombinációja miatt rendkívül időigényes optimalizálni, a jó klasszifikációhoz pedig több megfigyelésre van szükség, mint a másik két modell esetén, valamint hajlamos a túlillesztésre. Előnye azonban, hogy komplexitása miatt a legbonyolultabb kapcsolatokat, összefüggéseket is képes modellezni, ezzel általában sokkal pontosabb becslések érhetők el, mint a másik módszereink esetén, és kevésbé érzékeny a kiugró adatokra (Nielsen, 2019).

Az elkészített modelljeink fontosabb klasszifikációs mutatóit az alábbi táblázatban foglaljuk össze.

	Logisztikus regresszió	Véletlen erdő	Neurális hálózat
Pontosság	90,95%	91,08%	95,92%
1-TPR = FNR	87,11%	8,85%	4,10%
Szenzitivitás (TPR)	12,89%	91,15%	95,90%
AUC	0,8474	0,9462	-

11. táblázat; A modellek fontosabb mutatói

A számokra tekintve az mondható el, hogy a neurális hálózat készítette el számunkra a leginkább pontos modellt, a legmagasabb pontosság, a legalacsonyabb ál negatív ráta és a legmagasabb valós pozitív ráta is.

A véletlen erdő értékei viszonylag sokkal maradtak el a neurális háló modellétől. A logisztikus regresszió klasszifikációs értékei jelentősen rosszabbak a másik két modell eredményeinél, ez azért lehetséges véleményünk szerint, mert a logisztikus módszertan alapvető modellfeltevéseinek sérülése nélkül nem változtathatunk a betegek és egészségesek mintabeli arányán.

Előzetesen a neurális hálót gondoltuk a legeredményesebbnek a komplexitása miatt, ezt sikerült is eltalálnunk, azonban a logisztikus és a véletlen erdő modellek értékeitől nagyjából ugyanazt vártuk, ez az előbb említett alapfeltevés miatt azonban nem bizonyult helyes predikciónak.

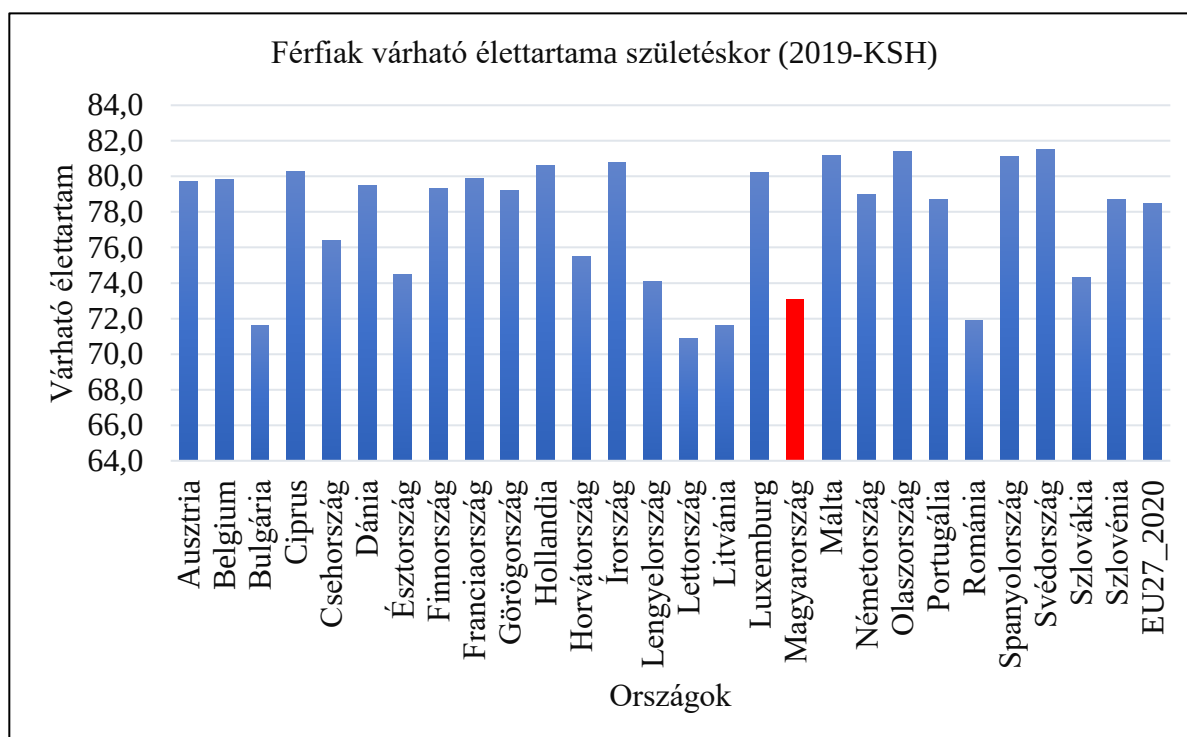
12. Kitekintés

Ebben a fejezetben kitekintést teszünk kutatásunk, modelljeink magyarországi egészségügyi alkalmazásának lehetőségeiről. Ismertetjük a magyarországi helyzetet, orvosi szakvéleményeket, majd következtetéseket vonunk le a témánkkal kapcsolatban.

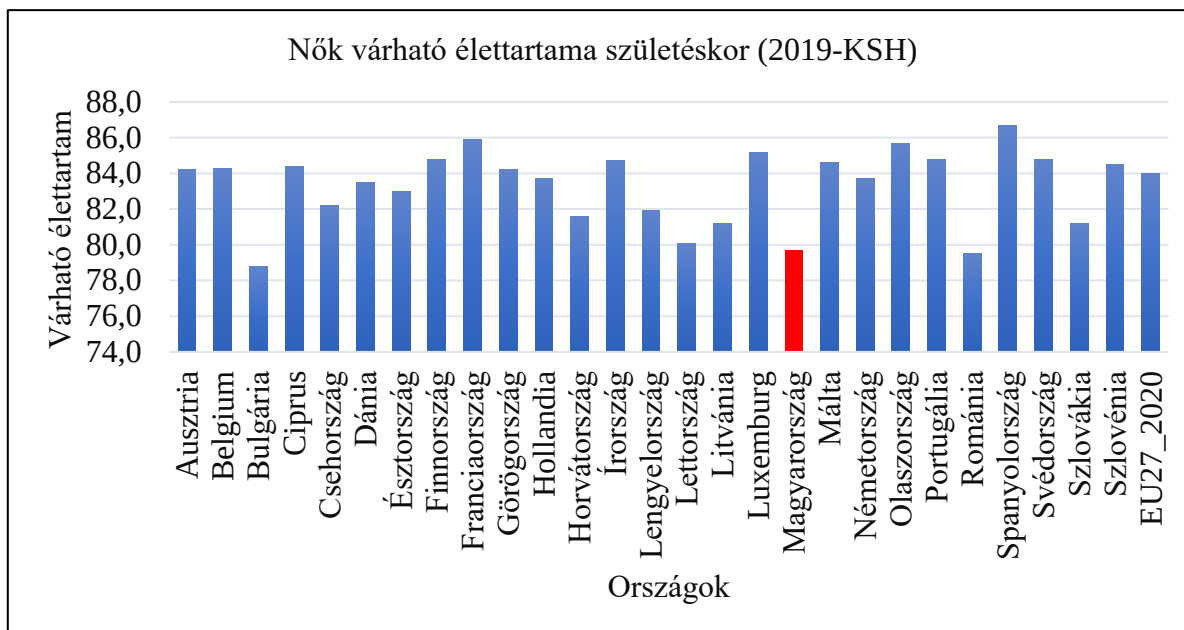
12.1. Magyarország helyzete

Mindenekelőtt érdemesnek tarjuk a szívmegetegedések kérdését kontextusba helyezni. Mint az teljesen logikusnak is hangzik, a szívmegetegedések a várható élettartam csökkenésével járnak, hiszen az alapvető vérkeringés legfontosabb, központi eleméről van szó. A hatékony szűrés célja tehát alapvetően ennek a kockázatnak a lehető legnagyobb mértékű csökkentése. Ahhoz tehát, hogy a magyar helyzetet elemezni tudjuk, fontos, hogy áttekintést nyújtsunk néhány elengedhetetlen várható élettartamra vonatkozó adatról regionális szinten.

Az összehasonlíthatóság kedvéért az Európai Unió átlagát szemléltetjük.



17. ábra; Férfiak várható élettartama születéskor (2019-KSH) (saját szerk.: Excel)

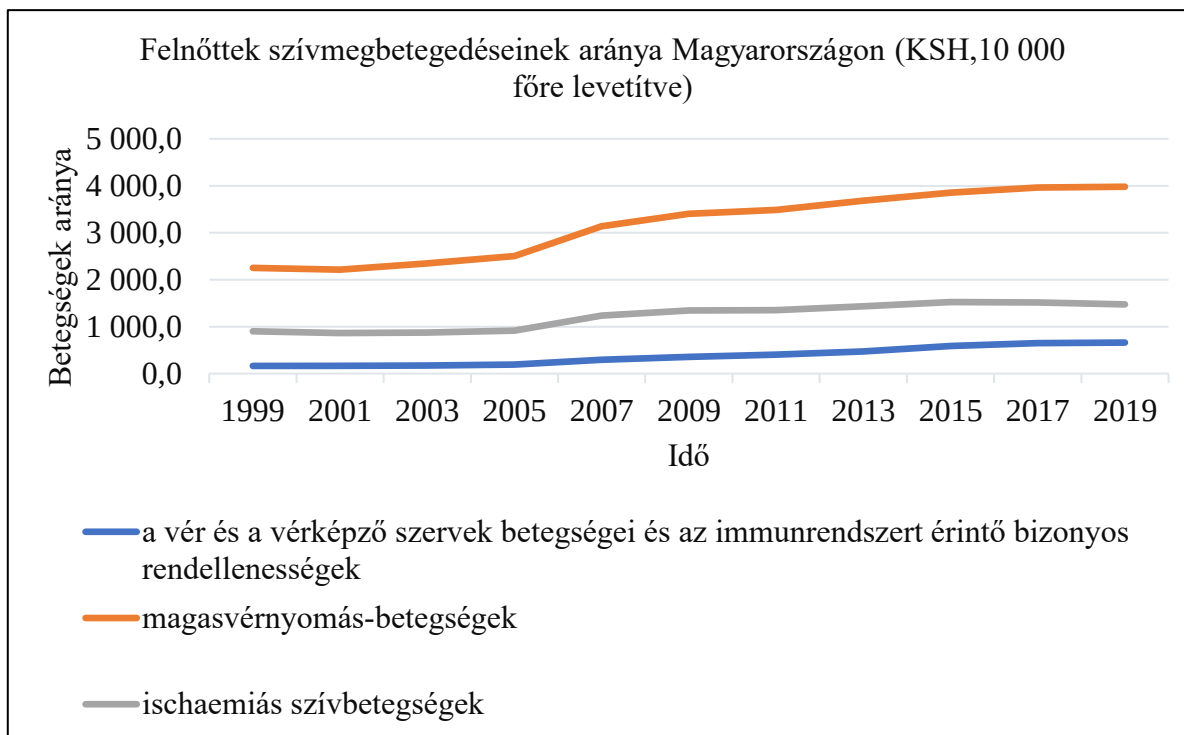


18. ábra; Nők várható élettartama születéskor (2019-KSH) (saját szerk.: Excel)

Az ábrán az Európai Unió régióba tartozó országok születéskor várható élettartamát ábrázoljuk, lebontva nemek szerint. Elsőre szembeütő lehet, hogy a férfiak várható élettartama országtól függetlenül szignifikánsabban alacsony, mint a nőké. Mélyebb elemzések nélkül feltételezhető, hogy ez a férfiak általános életmódjával magyarázható, hiszen több káros anyagot fogyasztanak általában, és akár a mindennapi életvitelük, munkájuk is életveszélyes faktorokat is tartalmazhat, hiszen a fizikai munkában felülreprezentáltak. Ezek alapján az első következtetés, amit levonhatunk, hogy egy ilyen előre kialakított szűrési modell, amelyet létrehozunk, érdemleges szerepet játszhat abban, hogy a férfiak károsabb életvitelét detektálja, szűrje a betegségek hajlamait, a kockázatosabbak esetében akár többször, gyorsan, hatékonyan, akár orvos jelenléte nélkül.

Ha külön a magyarországi helyzetre tekintünk, azt látjuk, hogy a férfiak esetén 73,1, nők esetén pedig 79,7 a születéskor várható élettartam a 2019-es adat alapján. Látszik, hogy Nyugat-Európa országai jóval magasabb adatszámokkal rendelkeznek, és az Európai Unió keletibb országai közül is többnek kedvezőbbek az élettartam eredményei. Ez újabb következtetést von maga után: regionális szinten le van maradva az ország egy olyan mutató esetén, amely közvetetten, proxy változó módjára indikálja, hogy milyen hatékony az adott állam egészségügyi rendszerének betegségeket, valamint azokra való hajlamokat felismerő mechanizmusa. Ez újabb fontos indok egy ilyen hatékony egészségügyi modell, és rendszer kialakítására.

A továbbiakban a magyar kardiológiai megbetegedések helyzetét is bemutatjuk idősoros adatok segítségével.



19. ábra; Felnőttek szívbetegedéseinek aránya Magyarországon (KSH) (saját szerk.: Excel)

Az ábra az elmúlt két évtized leginkább előforduló szívbetegedéseinek arányát mutatja be Magyarországon. Az adatok kirajzolódásából jól látszódik, hogy sajnos egy alapvetően enyhén növekvő tendencia mutatkozik meg az elmúlt időszakban, amely több tényező együttes eredménye, azonban jelen esetben számunkra nem is ez a legfontosabb kérdés, hanem az, hogy mire lehet ebből következtetni. Az első dolog, amely nyilvánvaló, hogy egy egészségesebb életmód ösztönzése fontos intézkedés volna az állami szervektől. A megelőzéseket figyelmen kívül hagyva azonban a másik fontos kérdés, ami felmerül, az a megbetegedések szűrési terheinek csökkentése a növekedő nyomás alatt. Egyértelmű, hogy egy ilyen trend esetén egy kisebb ország, amely kevesebb humán és technikai erőforrással is rendelkezik, nem képes ezeket a szűrési orvosi munkákat kellő gyorsasággal elintézni önmagában. Tehát továbbra is az látszódik az empirikus számításokból, hogy egy ilyen szűrési modellnek hatalmas szerepe lenne az egészségügy hatékonysági fejlődésében.

12.2. Orvosi szakvélemények

A továbbiakban az orvosi szakvéleményeket is áttekintjük. Ehhez egy budapesti körzeti orvost, valamint egy kardiológust is megszólaltatunk, anonim formában. A témához kapcsolódóan a magyar helyzetről és a szűrési mechanizmusokról, és azok hatékonyságairól kérdezzük őket. Gondolataikat egymás után, összekapcsolva osztjuk meg.

A körzeti házi orvos elsőként a téma történelmi kontextusába nyújt betekintést. Elmondása szerint nem is olyan régen még általános kötelezettség volt bizonyos szervek, funkciók szűrésére, azonban ez mára megváltozott, leginkább csak egyes állami alkalmazottaknak vannak ilyesfajta kötelezettségeik. Az orvos ezzel kapcsolatban szkepticizmusát fejezi ki, ám elmondása szerint valóban megterhelő volt ennek a lebonyolítása. A mostani eljárásról azt mondja, hogy véleménye szerint több tényező miatt bürokratikus intézménnyé vált a szűrések lebonyolítása. A várólisták hossza, az orvosok elfoglaltsága, és az egészségügy humán erőforrásának csökkenése (például a nyugati magasabb bérek miatt) egyaránt hatottak elképzelése szerint arra, hogy a mai állami rendszer eredménytelenné váljon. A bürokráciai alatt a szűrés megszervezését érti, hiszen sok telefonálás, üzenetváltás, adatfelvétel, beutalás kérés történik egy ilyen mechanizmus alatt, ami így tovább bénítja a működést. Elmondása szerint a koronavírus csak rontott a helyzeten. Az eleve nehezen megszerveződő rendszert felváltotta egyfajta „távgyógyítási” mechanizmus, amely alatt a vírus okozta intézkedéseket érti, ezek által csak rendkívül kevés személyes találkozóra nyílt lehetőség a betegek és az orvosok között. Sok kórházban, rendelőben bejárási korlátozásokat vezettek be, vagy csak mérsékelt mennyiségben engedtek be pácienseket, hangsúlyozza ki a szakember. Ez véleménye szerint csak tovább rontott az eleve nem hatékony rendszeren.

A számunkra leginkább témába illő szakember, a kardiológus orvos elsőként a magyar társadalom kardiológiai kockázatairól beszél. Elmondása szerint a felnőtt emberek igencsak nagy részét érinti szívmegebetegedési hajlam, aminek okát több tényezőben is látja. A magas zsírtartalmú ételek, dohány, alkohol fogyasztása, a kevesebb testmozgás egyaránt hathat a kockázat emelkedésére, ráadásul kihangsúlyozza, hogy egy ezek által kialakult hajlam akár örökletes is lehet genetikailag, így rendkívül fontos kezdeni valamit ezzel kapcsolatban. Kedvünkért a kérdés gazdasági vonaláról is beszél. Véleménye szerint Magyarországon valóban rendkívül nagy szerepe van a magánvállalkozások szűréseinek, hiszen az említett állami terhelő tényezők ebbe az irányba terelik a betegeket. Ezen vállalatok viszont a profit maximalizálása érdekében hajlandók magasabb árakat kérni a kétségbeesett emberektől, ezért is volna szükség egy olyan hosszútávú, stabil rendszerre, amely gyors és hatékony, akár pénzügyi értelemben is. A humán erőforrás hiányában a gépesítési potenciált véleménye szerint már felismerte az egészségügy, azonban ilyesfajta rendszerek is sok szakértői munkát igényelnek.

Az orvosok egyetértenek abban, hogy egy általunk is felvetett modell, diagnosztikai mechanizmus kialakítása rendkívül nagy jelentőséggel, hozzáadékkal bírna a szűrési problémák

megoldásában, mind hatékonysági, mind pedig pénzügyi szempontból. Elmondásaik szerint egyes helyeken már használnak ehhez hasonló gépi tanulási módszereket, azonban nem igazán elterjedtek egyelőre, ám fontos volna a helyzet javítására.

12.3. Következtetések

Összességében tehát az empirikus elemzés és az orvosi szakvélemény megerősítik azt a gondolatot, hogy az általunk felvetődő kardiológia diagnosztikai modellezés rendkívül releváns kérdés, amely sürgető választ igényel. Az állami egészségügy különböző tényezői, a magyar társadalom egészségi állapotának tendenciája, valamint a koronavírus általi bizonytalan helyzet egyre nehezebb irányba befolyásolja a szűrési folyamatok logisztikáját, azonban egy gyors és megbízható gépi rendszer sokat segíthetne a probléma enyhítésén.

13. Összefoglalás

Ebben a dolgozatban a szívbetegségek detektálásával, előrejelzésével foglalkoztunk. Ehhez három különböző machine learning módszert használtunk fel, a logisztikus regressziót, a véletlen erdőt és a neurális hálót. Előzetesen betekintést nyújtottunk a három modellezés módszertanába a könnyebb értelmezés érdekében. Az adatbázis szűrése és kiigazítása után felépítettük a modelljeinket és összehasonlítást végeztünk több szempont alapján.

A logisztikus regresszió esetében magasabbak voltak az elvárásaink, melyek végül nem igazolódtak be, hiszen klasszifikációs mutatók alapján nem sikerült megfelelően becsülni az alanyok egészségügyi állapotát.

A véletlen erdő modell esetén már pontosabb, jobb eredményeket sikerült elérnünk a modell egyszerűsége ellenére. Előzetesen nem gondoltuk, hogy eredményesebb lesz, mint a logisztikus regresszió.

Végül a neurális háló nyújtotta a legpontosabb becslést, minden vizsgált mutató alapján, amely egybevág azzal az előzetes várakozásunkkal, hogy a legkomplexebb modell bizonyul a legjobbnak, és úgy gondoljuk, hogy jobb becslést nyújthat, mint egy általános orvosi internetes oldal. Véleményünk szerint hatékonyan megkönnyítheti a egészségügyi dolgozók munkáját.

A gépi tanulási módszerek használata után egy kitekintést írtunk a szívbetegségek detektálásának hazai kontextusáról, adatairól, majd orvosok szakvéleményét is kikértük a témában.

Végül úgy gondoljuk végső konklúzióként, hogy hatékony és jó lenne hasonló, az általunk használnál részletesebb modellek használata az egészségügy több területén is, amellyel számos betegséget lehetne megelőzni, illetve tehermentesíteni vele az orvosok munkáját. A hazai egészségügyben is elterjedt a digitalizáció az Erodium és az EESZT rendszer segítségével, azonban ezen rendszerekben rejlő potenciál még nincsen teljesen kihasználva. Úgy gondoljuk, hogy egy miénkhez hasonló szűrési modell könnyen implementálható lehetne Magyarországon is, ám ehhez szükséges lenne az informatikai rendszerek folyamatos fejlesztése és az információ áramlás a különböző rendszerek között. A modellt úgy lehetne naprakészen tartani, hogy a már orvosok által pontos diagnózist megkapó betegek adatait hozzáadjuk a már meglévő adatbázishoz.

Az elemzéshez használt R kód az alábbi linken érhető el:
[https://github.com/Demok3/TDK/blob/faa51891c341f65ded9f47e935c7da9e95368037/tdk%20\(1\).r](https://github.com/Demok3/TDK/blob/faa51891c341f65ded9f47e935c7da9e95368037/tdk%20(1).r)

14. Hivatkozások

Altrichter Márta, Horváth Gábor, Pataki Béla, Strausz György, Takács Gábor, Valyon József (2006): *Neurális hálózatok*, Elérhető: [Neurális hálózatok | Mesterséges Intelligencia Almanach \(bme.hu\)](#) Letöltés dátuma: 2022. 01. 05.

Behavioral Risk Factor Surveillance System (BRFSS): 2015 BRFSS cleaned Survey Data, Elérhető: <https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset>
Eredeti teljes adatbázis helye: https://www.cdc.gov/brfss/annual_data/annual_2015.html
Letöltés dátuma: 2021. 11. 15.

Centers for Disease Control and Prevention (CDC): *Know Your Risk for Heart Disease*
Elérhető: [Know Your Risk for Heart Disease | cdc.gov](#) (online)

Fazekas István (2013): *Neurális hálózatok*, Elérhető: [Debreceni Egyetem Informatikai Kar. Fazekas István. Neurális hálózatok - PDF Ingyenes letöltés \(docplayer.hu\)](#) Letöltés dátuma: 2021. 11. 13.

Ferenci Tamás (2018): *Ökonometria I.* Elérhető: http://www.medstat.hu/oktatas/okonometria/FerenciTamas_OkonometriaI.pdf Letöltés dátuma: 2021. 09. 09.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013): *An Introduction to Statistical Learning* (second edition), Elérhető: [An Introduction to Statistical Learning \(statlearning.com\)](#) Letöltés dátuma: 2021. 09. 14.

Howard G. Stoa (1989): *Can health screening damage your health?* Elérhető: [193.full.pdf \(bjgp.org\)](#) Letöltés dátuma: 2022. 04. 19.

Ian Goodfellow, Yoshua Bengio, Aaron Courville (2016): *Deep learning*, Elérhető: [Deep Learning \(deeplearningbook.org\)](#) Letöltés dátuma: 2022. 01. 23.

Jeffrey M. Wooldridge (2012): *Introductory Econometrics: A Modern Approach*, Elérhető: https://economics.ut.ac.ir/documents/3030266/14100645/Jeffrey_M._Wooldridge_Introductory_Econometrics_A_Modern_Approach_2012.pdf Letöltés dátuma: 2021. 11. 28.

Központi Statisztikai Hivatal (KSH): A háziorvosi és a házi gyermekorvosi szolgálathoz bejelentkezettek egyes betegségei. Elérhető: [4.1.1.24. A háziorvosi és a házi gyermekorvosi szolgálathoz bejelentkezettek egyes betegségei \(ksh.hu\)](#) Letöltés dátuma: 2022. 03. 20.

Központi Statisztikai Hivatal (KSH): Születéskor várható átlagos élettartam - férfiak (2000-2019). Elérhető: [22.1.3.10. Születéskor várható átlagos élettartam – férfiak \(ksh.hu\)](https://ksh.hu/hu/index.php?lang=hu&lang2=en&lang3=en&lang4=en&lang5=en&lang6=en&lang7=en&lang8=en&lang9=en&lang10=en&lang11=en&lang12=en&lang13=en&lang14=en&lang15=en&lang16=en&lang17=en&lang18=en&lang19=en&lang20=en&lang21=en&lang22=en&lang23=en&lang24=en&lang25=en&lang26=en&lang27=en&lang28=en&lang29=en&lang30=en&lang31=en&lang32=en&lang33=en&lang34=en&lang35=en&lang36=en&lang37=en&lang38=en&lang39=en&lang40=en&lang41=en&lang42=en&lang43=en&lang44=en&lang45=en&lang46=en&lang47=en&lang48=en&lang49=en&lang50=en&lang51=en&lang52=en&lang53=en&lang54=en&lang55=en&lang56=en&lang57=en&lang58=en&lang59=en&lang60=en&lang61=en&lang62=en&lang63=en&lang64=en&lang65=en&lang66=en&lang67=en&lang68=en&lang69=en&lang70=en&lang71=en&lang72=en&lang73=en&lang74=en&lang75=en&lang76=en&lang77=en&lang78=en&lang79=en&lang80=en&lang81=en&lang82=en&lang83=en&lang84=en&lang85=en&lang86=en&lang87=en&lang88=en&lang89=en&lang90=en&lang91=en&lang92=en&lang93=en&lang94=en&lang95=en&lang96=en&lang97=en&lang98=en&lang99=en&lang100=en) Letöltés dátuma: 2022. 03. 27.

Központi Statisztikai Hivatal (KSH): Születéskor várható átlagos élettartam - nők (2000-2019). Elérhető: [22.1.3.11. Születéskor várható átlagos élettartam – nők \(ksh.hu\)](https://ksh.hu/hu/index.php?lang=hu&lang2=en&lang3=en&lang4=en&lang5=en&lang6=en&lang7=en&lang8=en&lang9=en&lang10=en&lang11=en&lang12=en&lang13=en&lang14=en&lang15=en&lang16=en&lang17=en&lang18=en&lang19=en&lang20=en&lang21=en&lang22=en&lang23=en&lang24=en&lang25=en&lang26=en&lang27=en&lang28=en&lang29=en&lang30=en&lang31=en&lang32=en&lang33=en&lang34=en&lang35=en&lang36=en&lang37=en&lang38=en&lang39=en&lang40=en&lang41=en&lang42=en&lang43=en&lang44=en&lang45=en&lang46=en&lang47=en&lang48=en&lang49=en&lang50=en&lang51=en&lang52=en&lang53=en&lang54=en&lang55=en&lang56=en&lang57=en&lang58=en&lang59=en&lang60=en&lang61=en&lang62=en&lang63=en&lang64=en&lang65=en&lang66=en&lang67=en&lang68=en&lang69=en&lang70=en&lang71=en&lang72=en&lang73=en&lang74=en&lang75=en&lang76=en&lang77=en&lang78=en&lang79=en&lang80=en&lang81=en&lang82=en&lang83=en&lang84=en&lang85=en&lang86=en&lang87=en&lang88=en&lang89=en&lang90=en&lang91=en&lang92=en&lang93=en&lang94=en&lang95=en&lang96=en&lang97=en&lang98=en&lang99=en&lang100=en) Letöltés dátuma: 2022. 03. 27.

Marie-Josée Perrier, Kathleen A Martin Ginis (2017): *Narrative interventions for health screening behaviours: A systematic review*, Journal of Health Psychology, Elérhető: [Narrative interventions for health screening behaviours: A systematic review - PubMed \(nih.gov\)](https://pubmed.ncbi.nlm.nih.gov/31811111/) Letöltés dátuma: 2021. 11. 23.

Michael Nielsen (2019): *Neural Networks and Deep Learning*, Elérhető: [Neural networks and deep learning](https://mnielsen.github.io/neural-networks-and-deep-learning/) Letöltés dátuma: 2021. 12. 20.

National Health Service (2018): *Cardiovascular disease*, Elérhető: [Cardiovascular disease - NHS \(www.nhs.uk\)](https://www.nhs.uk/) Letöltés Dátuma: 2022. 01. 31.

National Institute on Alcohol Abuse and Alcoholism (NIAAA): *Drinking Levels Defined*, Elérhető: [Drinking Levels Defined | National Institute on Alcohol Abuse and Alcoholism \(NIAAA\) \(nih.gov\)](https://www.niaaa.nih.gov/publications/brochures/drinking-levels-defined) (online)

Sashank J. Reddi, Satyen Kale, Sanjiv Kumar (2018): *On the convergence of Adam and beyond*, Elérhető: [pdf \(openreview.net\)](https://openreview.net/pdf?id=Hkz9wZk0wZ) Letöltés dátuma: 2022. 01. 31.

Steven Gonzalez (2000): *Neural Networks for Macroeconomic Forecasting: A Complementary Approach to Linear Regression Models*, Elérhető: [Neural Networks for Macroeconomic Forecasting: A Complementary Approach to Linear Regression Models | Request PDF \(researchgate.net\)](https://www.researchgate.net/publication/221111111_Neural_Networks_for_Macroeconomic_Forecasting_A_Complementary_Approach_to_Linear_Regression_Models) Letöltés dátuma: 2021. 01. 12.

Susanna Tobias Gómez, Víctor Palomar Asenjo, Montserrat Borrás Perera, Isabel Pérez Hernández, Antonio Ruiz Giner, Víctor palomar García (2008): *Clinical significance of unilateral tonsillar enlargement*, Acta Otorrinolaringológica Española, Elérhető: <https://www.elsevier.es/index.php?p=revista&pRevista=pdf-simple&pii=S2173573509701280&r=495> Letöltés dátuma: 2021. 11. 27.

Thais Mayumi Oshiro, Pedro Santoro Perez, José Augusto Baranauskas (2012) *How Many Trees in a Random Forest?* Elérhető:

https://www.researchgate.net/publication/230766603_How_Many_Trees_in_a_Random_Forest Letöltés dátuma: 2022. 02. 23.

Tina Birgitte Hansen, Jes Sanddal Lindholt, Axel Cosmus Pyndt Diederichsen, Michiel C J Bliemer, Jess Lambrechtsen, Flemming Hald Steffensen (2018). *Individual preferences on the balancing of good and harm of cardiovascular disease screening*, Elérhető: [Individual preferences on the balancing of good and harm of cardiovascular disease screening | Heart \(bmj.com\)](#) Letöltés dátuma: 2022. 02. 14.

Tom Fawcett (2005): *An introduction to ROC analysis*, Elérhető: <https://www.sciencedirect.com/science/article/abs/pii/S016786550500303X> Letöltés dátuma: 2021. 11. 19.