Budapesti Corvinus Egyetem
Közgazdaságtudományi Kar

Eötvös Loránd Tudományegyetem
Természettudományi Kar

# Big Data elemzés az egészségbiztosításban

Készítette: Tóth Péter

Biztosítási és pénzügyi matematika mesterszak

Aktuárius szakirány

2020

Szakszemináriumvezető: Vékás Péter

Operációkutatás és Aktuáriustudományok Tanszék

**Tartalom**

## Ábrák jegyzéke



## Táblázatok jegyzéke

## Bevezetés

Az adatelemzés és a gépi tanulási módszerek integrált részévé váltak a modern tudományos módszertannak, automatizált eljárásokat kínálva egy jelenség előrejelzésére múltbeli megfigyelések alapján. A nagy mennyiségű jelentős részben nem struktúrált adatok, vagy ismertebb nevén Big Data felhasználása különböző elemzésekre egyre elterjedtebb napjainkban. A vállalatok számára elérhető adatok mennyisége rendkívüli gyorsasággal növekszik. E növekedéssel a közönséges elemzési technikák nem tudják tartani a lépést. A modern digitális korban a Big Data technológiák elősegítik a hatalmas mennyiségű információ feldolgozását, növelik a munkafolyamat hatékonyságát és csökkentik az üzemeltetési költségeket.

A biztosítási ágazat mindig is meglehetősen konzervatív volt; az új technológiák bevezetése azonban nem csupán modern tendencia, hanem a verseny fenntartásának szükségességes lépése. Az Európai Biztosítás- és Foglalkoztatói nyugdíj-hatóság (EIOPA) közleménye szerint a biztosítótársaságok 31%-a használ Big Data feldolgozást elősegítő olyan eszközöket, mint például a mesterséges intelligencia vagy különböző gépi tanulási módszerek, valamint további 24%-uk bizonyítottan a tervezési szakaszban tart (EIOPA, 2019)

Diplomamunkám témája Big Data elemzés az egészségbiztosításban. Kutatásom célja, különböző változók alapján minél pontosabban előrejelezni egy egyén várható egészségügyi kiadásait a következő évre nézve. Ehhez három gépi tanulási módszert használok fel és ezek közül a legjobban teljesítőt választom a végső modellnek. Továbbá kivizsgálni, hogy milyen szorosan mozognak együtt a várható kifizetések és az olyan tényezők, mint a kor, a testtömegindex, vagy a dohányzás.

A téma aktualitása abban rejlik, hogy a biztosítótársaságok más vállalatokhoz hasonlóan a gyorsuló technológiai fejlődésnek köszönhetően egyre több adattal rendelkeznek, így ezek feldolgozásához előbb utóbb elkerülhetetlenné válnak a Big Data elemző technológiák.

A témára azért esett a választásom, mert a biztosítási és pénzügyi matematika mesterszakos tanulmányaim alatt az általam legkedveltebb tárgyak az egészségbiztosítás, a többváltozós statisztikai modellezés és a többváltozós statisztikai modellezés II. Emiatt úgy gondoltam, hogy kombinálhatnám azt a tudást, amit ezek alatt a kurzusok alatt szereztem egy kutatásban. Miután rátaláltam az EIOPA által kiadott rendeletre, amely beszámol a Big Data

elemzés előnyeiről az egészségbiztosításban (Eiopa, 2019), még jobban bebizonyosodott, hogy érdemes ezzel a témával foglalkoznom.

Úgy gondolom e fajta kutatásokat azért fontos végeznünk, mivel a Big Data elemzésnek és a Big Data elemzési technikáknak, mint a gépi tanulási módszerek, egyre nagyobb a térnyerésük a világban, ami annak köszönhető, hogy általuk pontosabban és gyorsabban elvégezhetőek a kívánt számítások, melyek így nem csak költség-, de időhatékonyabbá is válnak.

A szakirodalomban többféle felhasználását találhatjuk a Big Data elemzésnek és a gépi tanulási módszereknek az egészségbiztosításban. Egyesek biztosítási csalások kiszűrésére alkalmazzák(Kirlidog & Asuk, 2012), míg mások egy felhő alapú ún. „marketplacet" hoznának létre, amelyről mindenki eléri a számára megfelelő biztosítási formát(Abbas et al., 2015). Akadnak olyan munkák is, amelyek rettegett betegségek kiszűrésére használják a véletlen erdő gépi tanulási módszert (Vekeman et al., 2019), vagy éppen egy ország egészségügyi kiadásait próbálják előrejelezni mesterséges neurális hálók segítségével historikus adatok alapján (Yeh et al., 2014).

A munkámban három gépi tanulási módszert használok a modellezéshez. Ezek a mesterséges neurális hálók, a véletlen erdő és a Support Vector Machine (SVM). A neurális hálók a gépi tanulási módszerek egyik fő eszköze. Amint azt a neve („neurális") is sugallja, ezt a módszert az agyi neuronok és az agy működése inspirálta. A neurális hálók bemeneti és kimeneti rétegekből, valamint rejtett rétegekből állnak. A rejtett rétegekben mennek végbe a számítások, amelyek olyan bonyolultak, hogy sokszor ún. „black-box" módszernek, azaz fekete doboznak is nevezik. Mivel nem tudni pontosan mi történik benne, de eredményül kapunk egy, klasszifikáció esetén több, értelmezhető kimeneti értéket. A véletlen erdő, ami egyszerűsítve több döntési fa átlagolva, a fa alapú algoritmusok egyik legjobb és leginkább alkalmazott gépi tanulási módszerek egyike. Véletlen erdő használata esetén egyszerű lépések segítségével hajthatunk végre bonyolult döntéseket. A véletlen erdő egyik nagy hátránya, hogy hajlamos a túlillesztésre. Az támaszvektor-gépek (SVM) egy viszonylag új gépi tanulási módszer. Használható klasszifikációra, regresszióra és klaszterezésre is. Ez egy nem valószínűség alapú módszer, mely csak numerikus prediktorokat kezel. Alapjai ahogyan a nevében is szerepel az ún. támaszvektorok, amelyek segítségével létrehozható egy hipersík, amely szeparálja a változókat két külön kategóriába, így az egyik kategóriához tartozó megfigyelések a hipersík egyik, míg a másik kategóriához tartozók a hipersík másik oldalán helyezkednek el, a támaszvektorok pedig a hipersíkhoz legközelebb álló pontok halmaza. Optimális esetben a

hipersík és a támaszvektorok közötti távolság maximális. Mivel általában két csoport nem szeparálható lineárisan, ezért az SVM egy magfüggvény segítségével magasabb dimenzióba transzformálja az adatokat, ahol ezt már megtehetjük.

Az elemzéshez egy egészségbiztosítási adatállományt használtam, amelyet a Kaggle adatelemzési platformról töltöttem le. 1338 egyén adatait tartalmazza, köztük az egy év alatt kifizetett egészségügyi kiadásokat is. E mellett olyan további változókat, mint a kor, a testtömegindex, a lakhely az USA négy régióján belül, a dohányzás és a gyerekszám. Az adathalmaz egy valós demográfiai statisztikai adatokon alapuló szimulált adatbázis, melyet az Amerikai Egyesült Államok Népszámlálási Irodája hozott létre.

A dolgozatom három fejezetből áll. Az elsőben röviden bemutatom a szakirodalomban fellelhető kutatások eredményeit. A másodikban ismertetem a már említett három gépi tanulási módszer elméleti hátterét. A harmadikban pedig részletesen bemutatom a felhasznált adatbázist, az elemzés menetét és az eredményeket.

# 1. Big Data elemzés és az egészségbiztosítás

Ebben a fejezetben szeretném bemutatni, hogy a Big Data elemzést és a gépi tanulási módszereket milyen feladatokra használták eddig az egészségbiztosításban, vagy épp milyen elméletek születtek felhasználási lehetőségeikről. Az említett gépi tanulási technikákról a következő fejezetben írok részletesen.

2019 tavaszán az EIOPA (Európai Biztosítás- és Foglalkoztatói nyugdíj-hatóság) közzétette a széles körű, Big Data elemzés gépjármű- és egészségbiztosítás területén történő felhasználását vizsgáló jelentését (EIOPA, 2019). A jelentésben kiemelik, hogy a Big Data elemzési technikák lehetővé teszik a vállalatok számára, hogy jobban megértsék a fogyasztók igényeit, tulajdonságait és életmódját, lehetővé téve számukra a pontosabb kockázatértékelések kidolgozását. Ez lehetővé teszi a cégeknek, hogy személyre szabottabb és kényelmesebb termékeket és szolgáltatásokat fejlesszenek ki a fogyasztók számára, figyelembe véve azt a tényt, hogy ezek az intézkedések javítják az ügyfelek elkötelezettségét és felhasználói élményét.

Az EIOPA úgy véli, hogy a Big Data elemzési technikák területén az egyik kulcsfontosságú fejlemény a mobiltelefon-technológia fokozott használata új adatforrások gyűjtésére és a fogyasztókkal való kapcsolattartásra, példákat gyűjtöttek arra, hogy a biztosítótársaságok milyen típusú szolgáltatásokat nyújtanak ügyfeleiknek mobiltelefon-alkalmazásokon keresztül. Ezek leginkább a gépjármű biztosításokkal kapcsolatosak, de szó esik arról is, hogy például az ügyfelek orvosi és fogorvosi rendeléseket kezdeményezhetnek mobiltelefon-alkalmazásukon keresztül.

Az innováció fejlődésének gyorsasága, a verseny dinamikája a piacokon és a cégek stratégiai üzleti tervei azt mutatják, hogy bár az új adatforrások használata és a Big Data elemzési technikák elfogadása a biztosítási ágazatban még mindig nem elterjedt, ennek ellenére várhatóan ezek jelentősen növekedni fognak az elkövetkező években. Ugyanakkor kiemelik, hogy figyelembe kell venni azt a lehetőséget is, hogy egyes fogyasztói csoportok rosszabb helyzetbe is kerülhetnek, ha nem működnek megfelelő kormányzati szabályozások.

A Big Data elemzési technikák lehetővé teszik nagyon pontos számítások kidolgozását, korlátozott emberi beavatkozással vagy anélkül, növelve a döntéshozatal hatékonyságát és sebességét, és ezáltal csökkentve a működési költségeket. A historikus adatokkal kapcsolatos esetleges torzulásokat azonban a gépi tanulási algoritmusok fogják

megerősíteni, ha a cégek nem rendelkeznek megfelelő irányítási szabályokkal. Ez a kérdés akkor válik jelentősebbé, ha a „black box” algoritmusok bizonyos ítéleteit nem lehet pontosan értelmezni (EIOPA, 2019).

Az adatbányászat módszerei széleskörűen alkalmazhatóak különböző problémák megoldására, egyike a legnagyobb problémáknak, amelyekkel az egészségbiztosítók és úgy általában a biztosító intézmények szembesülnek az a csalás. A következőben ismertetem Kirlidog és Asuk (2012) munkáját, amely Support Vector Machine (SVM) gépi tanulási módszert alkalmazva próbálja detektálni a nagy méretű egészségbiztosítási csalásokat. Kanadában például az éves biztosítási kifizetések 10-15%-a csalás miatt lesz kifizetve (Gill K. M., 2009), ezek az esetek megelőzése költséges, időigényes és nem hatékony, így sok biztosító inkább fizet.

A csalások felismerésének az egyik leghatékonyabb módszere az „anomaly detection” – anomália észlelés, ami kiugró értékek észlelésére alapszik és a csalás valószínűségét számolja ki múltbéli biztosítási esetek adataira alapozva. Az adatbányászat e fajtája statisztikai, matematikai, Mesterséges intelligencia (MI) és Machine learning – gépi tanulás (ML) technikákat alkalmaz, hogy lényeges információkat nyerjen ki. A potenciálisan csalárd állítások megjelölésével a nyomozók összpontosíthatnak olyan eseményekre, amelyek valószínűleg csalók, és csökkentik a hamis pozitív elbírálások számát.

Minden biztosítási vállalatnak megvannak a technikái a csalás szűrésére bizonyos minták alapján, ami rendszerint a vállalat tapasztalatából adódik. Ezek kiszűrése egyszerűen végezhető SQL segítségével is, de a találatok nagy része nem igazi csalás és közelről kell őket megvizsgálni. Ehhez sok segítséget ad az anomália észlelés, klaszterezés és klasszifikáció, hiszen elsődleges funkciója ezen találatok szűkítése, ami kevesebb manuális vizsgálattal jár.

A munkában (Kirlidog & Asuk, 2012) egy török biztosító állományát használták fel, amelyben 2001-től 2009-ig szerepeltek rekordok. Összesen 808 ezer megfigyelést tartalmazott az adatbázis. Az elemzéshez az SVM-et egy Oracle rendszeren végezték el. A szoftver kiszámolta minden egyes egyénre annak a valószínűségét, hogy a szerződésben észlelhető e anomália. Ezekből 6595 rekord esetén volt az anomália valószínűsége 50% fölött, ekkor tekinthetők csalásnak. Ez az eljárás csökkenti a biztosítási díjakba beárazott csalási költségeket, amiket az ügyfelek kell kifizessenek.

Hasonló kutatást végeztek egy Ghánai állományon is (Sowah et al., 2019). Ők is felvetik a biztosítási csalások jelentőségét a biztosítási piacon, melynek gyorsulása mélyen

befolyásolta az egészségügyi ellátást. A csalások miatti pénzügyi veszteség következtében azon ügyfelek, akik valóban ellátásra szorulnak nem kapják meg a megfelelő kezelést, mivel a szolgáltatókat nem fizetik ki időben a biztosítók, a biztosítási igények mélyebb felülvizsgálata miatt, így a késedelemre hivatkozva a szolgáltatók nem hajlandók elvégezni a szükséges vizsgálatokat.

Ezen folyamat gyorsításaként a szerzők felvetik egy szoftver szükségességét a biztosítási csalások detektálására. Így született meg a döntéstámogató rendszer (DSS – decision support system). A szoftverbe be kell táplálni az adatállományt, amely SVM segítségével felismeri a csalásokat, így gyorsítva a folyamatot. A teszt folyamán több magfüggvénnyel is futtatták a modellt, melyek közül a Gauss (radiális) adta a legjobb találati arányt (87,91%).

Abbas et al., 2015-ös munkájukban felhívják a figyelmet arra, hogy a technológia és az internet adta lehetőségek fejlődésével és a Big Data térnyerésével az egészségügy is elkezdett az elektronikus rendszerek felé fordulni. Így mára az e-egészségügy magába integrálja egyik oldalon az egészségügyi szolgáltatók, korházak, klinikák és laboratóriumok, másik oldalon a biztosítók információit egy nagy felhőalapú egészségügyi adatbázisba, különféle mutatókkal. Ennek egyik mellékága a biztosítók által nyújtott szolgáltatások és annak költségeinek felhőalapú összehasonlítása. Erre léteznek különböző weboldalak, de, mivel rengeteg különféle információt kell összehasonlítsanak és sok szempont alapján kell személyre szabhatónak lenniük, minőségük fejlesztést kíván. A probléma megoldására a szerzők az e-egészségügyi felhőt bővítenék ki úgy, hogy az személyre szabottan biztosítási programokat ajánljon az embereknek a létező biztosítási „marketplace”-ről a felhasználóról már meglévő adatokra alapozva.

Ez a megoldás költségcsökkentő hatással is járhat, ha áttérnek a felhő alapú számítástechnika alkalmazására annak költségeit a felhasználókra lehet hárítani és az IT kapacitásokat se kell növelni, hiszen ez a technológia nagy adathalmazok kezelésére is alkalmas.

Vekeman et al 2019-es cikkükben klasszifikációra használt véletlen erdő segítségével modellezni próbálták a Lennox-Gastaut szindróma (LGS[1]) előfordulását. Igaz, ez nem

[1] A Lennox – Gastaut szindróma (LGS) komplex, ritka és súlyos gyermekkori epilepsziás megbetegedés. Jellemzője több és egyidejű rohamok, kognitív diszfunkció. 3–5 éves gyermekeknél jelentkezik, és felnőttkorban is fennmaradhat. Számos génmutációval, perinatális sérülésekkel, veleszületett fertőzésekkel, agydaganatokkal és genetikai rendellenességekkel, például West-szindrómával társulhat.(Markand, 2003)

közvetlenül egészségbiztosítási alkalmazás, azonban az egészségbiztosítók is felhasználhatják a hasonló modelleket, hogy a rettegett betegségeket kiszűrjék.

A véletlen erdő módszert használták, hogy egy olyan rendszert hozzanak létre, ami felismeri a lehetséges LGS betegeket. A véletlen erdő gépi tanulási módszer segítségével működő prediktív módszer, ami bináris eredményű eseményeket modellez. Minden egyes levél egy „ha – akkor" feltételt tartalmaz az inputból kiindulva és végig menve a megfelelő fák levelein (úton) egy végső kimenetelhez vezet. Minden egyes fa a megfigyelések adataira alapoz, így a véletlen erdő az összes fa bináris szavazatát aggregálva saját valószínűségi előrejelzést tesz.

A véletlen erdő módszert egy logisztikus regresszióhoz és egy naiv Bayes módszerhez hasonlították, amiből kiderült, hogy a véletlen erdő módszer felülmúlta az utóbbi kettő teljesítményét. Magas érzékenységet és részletességet mutatott a módszer, minden fontosnak határozott változó szignifikáns hatású volt a logisztikus regressziós modellben is.

2014-ben egy Tawiani kutatócsoport backpropagationt használó neurális háló segítségével próbálták előrejelezni a nemzeti egészségbiztosítási kiadásokat (Yeh et al., 2014). A modellükben 1996 és 2007 közötti havi egészségügyi kiadásokat vettek figyelembe. Az adathalmazuk 144 megfigyelésből állt. Arra a következtetésre jutottak, hogy a három fő magyarázó változó, ami meghatározza az egészségügyi kiadások mértékét az infláció, az öregedési index, valamint a biztosított lakosság száma. A neurális háló hatékonyságát összevetették ugyanezen adatbázison végzett Monte Carlo szimuláció és többváltozós regresszió teljesítményével. E három modell közül a neurális háló adta a legpontosabb becsléseket, ezzel is bizonyítva, hogy a gépi tanulási módszereknek helyük van az egészségbiztosítási számításokban.

Egy 2018-as kutatás a biztosítási alkuszok perspektívájából közelíti meg a problémát. Felhívják a figyelmet rá, hogy nem csak a biztosító, de a brókerek szempontjából is fontos az, hogy a szerződéskötés után a biztosított minél tovább az egészségbiztosító ügyfele maradjon. Ennélfogva a szerződések várható megszűnésének időpontját előrejelezve optimalizálható a biztosító kereskedelmi folyamata. A munkában ezt a feladatot egy súlyozott véletlen erdő módszerrel oldják meg. Megvizsgálnak különböző cenzúrázási esetetek, majd összehasonlítják az elért eredményeiket a hasonló sztenderd módszerekével, majd következtetésként kijelentik, hogy az általuk felállított modell helyénvaló ezen problémafelvetésben (Gerber et al., 2018).

A Mumbai-i egyetemen végzett tanulmánynak az a célja, hogy feltárja a hordozható technológiák bevezetésének hatásait az Indiai egészségbiztosító társaságokra. Megjegyzik, hogy a versenyképesség szempontjából fontos ezen technológiák alkalmazása, általuk előállított Big Data elemzése. A tanulmányt 53 indiai egészségbiztosítási szakértő bevonásával végezték el, akik félig strukturált kérdőívet töltöttek ki. Majd az adatokat elemezték s így kapták eredményeiket. Az eredményekből elmondható, hogy sikerült egy olyan empirikus modellt létrehozniuk, amellyel az összes feltételezés erősen megalapozottá vált, kivéve a mérsékelt kapcsolatot a hordozható technológiák adaptációja és a termékinnováció között. Megállapították ezen technológiák interakciójának természetét a technológiai politika, a vállalati kultúra, a stratégiai filozófia, a termékinnováció, a tudás menedzsment és az ügyfélszolgálat terén, valamint hatását a cég teljesítményére és versenyképességére. (Nayak et al., 2019)

Összességében elmondható, hogy a Big Data elemzés és a Big Data elemzési technikák, mint a gép tanulási módszerek széleskörűen alkalmazhatóak az egészségbiztosítási intézményekben. Számos tanulmány született ezen technikák lehetséges alkalmazhatóságáról, amelyek mindegyike egy kicsit más perspektívából közelíti meg a kérdést. Azonban minden esetben elmondható, hogy a módszerek használata a jövőben csakis a megfelelő szabályozással működhet megfelelően.

# 2. Gépi tanulási módszerek

Ebben a fejezetben bemutatom az általam választott három gépi tanulási módszer elméletét, melyeket az elemzéshez használok az egészségügyi kiadások előrejelzésére.

## 2.1. Mesterséges hálózatok

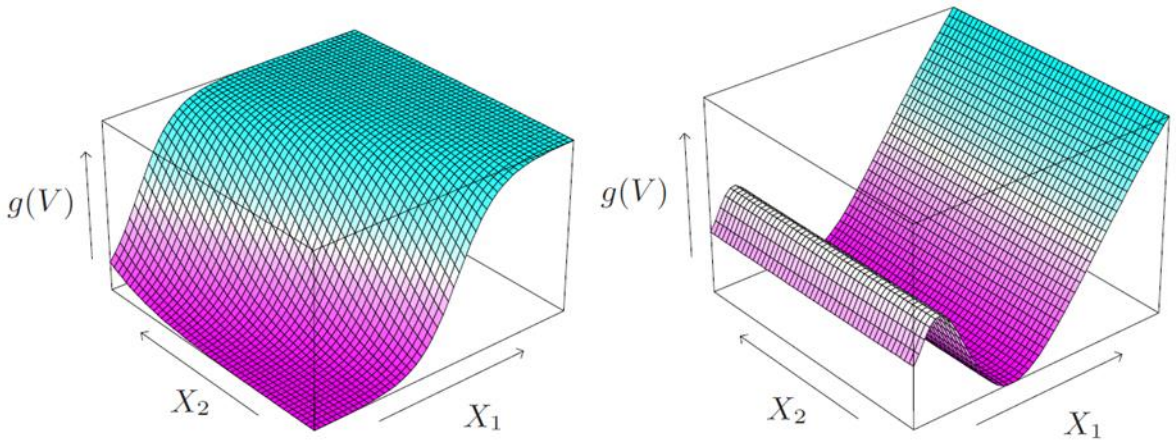
### 2.1.1. Projekció kereső regresszió (Projection pursuit regression – PRR)

A PRR a statisztikában a korai neurális hálózatok egy párhuzamosan fejlődött félparametrikus, simító eljárások közé tartozó modellje. A hálók elterjedésével részesedése a statisztikai elemzésekben jelentősen csökkent. A következőben bemutatom a PRR elméleti hátterét.

Legyen $X$ egy p elemű input vektor, és $Y$ egy célváltozó. Legyen $\omega_m$ egy $M$ elemű ismeretlen vektor. Akkor a PRR modell a következő lesz:

$$f(x) = \sum_{m=1}^{M} g_m(\omega_m^T X).$$

Ez egy additív modell, melyben az inputokat $\omega_m$-el súlyozzuk. A $g_m(\omega_m^T X)$ egy ismeretlen zsugorítófüggvény $\mathbb{R}^p$-ben, amely csak az $\omega_m$ vektortól függ.




**1. ábra. Két zsugorító függvény példája (Hastie et al., 2001)**

(**Bal oldali**): $\boldsymbol{g(V) = 1/\left[1 + \exp\left(-5(V - 0.5)\right)\right]}$, $\boldsymbol{ahol\ V = (X_1 + X_2)/\sqrt{2}}$.

(**Jobb oldali**): $\boldsymbol{g(V) = (V + 0.1)\sin(1/(V/3 + 0.1))}$, $\boldsymbol{ahol\ V = X_1}$.

Bevezetjük a következő skalárt $V_M = \omega_m^T X$, amely projektálja az X input vektort $\omega_m$-be. A feladatunk az, hogy megtaláljuk azt az $\omega_m$-et, amely mellett a modell jól illeszkedik *Y*-ra. Innen ered a név „projekció kereső regresszió”(Hastie et al., 2001).

Az 1. ábra bemutat két példát a zsugorító függvényekre. A bal oldali példa esetén az $\omega = \left(\frac{1}{\sqrt{2}}\right)(1,1)^T$, tehát a függvény csak $X_1 + X_2$ irányba változhat. A jobb oldali példán $\omega = (1,0)$.

A modell illesztése szokásos módon az RSS (residual sum of squares – négyzetes hibaösszeg) minimalizálásán keresztül történik a $g_m$ függvény és az $\omega_m$ irányvektorokon keresztül. A $g_m$-re korlátozásokat kell bevezetni azért, hogy elkerüljük a túlillesztést.

A projekció kereső modell nagyon általános, mivel a lineáris kombinációk nem lineáris függvényei meglepően nagyszámú modellcsaládot eredményeznek.

Ha *M* kellően nagy, akkor $g_m$ megfelelő megválasztása szinte bármilyen folytonos függvény approximációjára képes $\mathbb{R}^p$-ben, ezért az ilyen modelleket univerzális approximátornak nevezik. Az illesztett modellek interpretációja azonban általában nagyon nehéz, mivel minden egyes input elég komplex átalakuláson mennek keresztül, míg a modellbe kerülnek. Ennek eredményeképp a PRR-nek a legnagyobb hasznosíthatósága az előrejelzésnél van, ezzel szemben nem túl hatékony érthető modellek létrehozására.

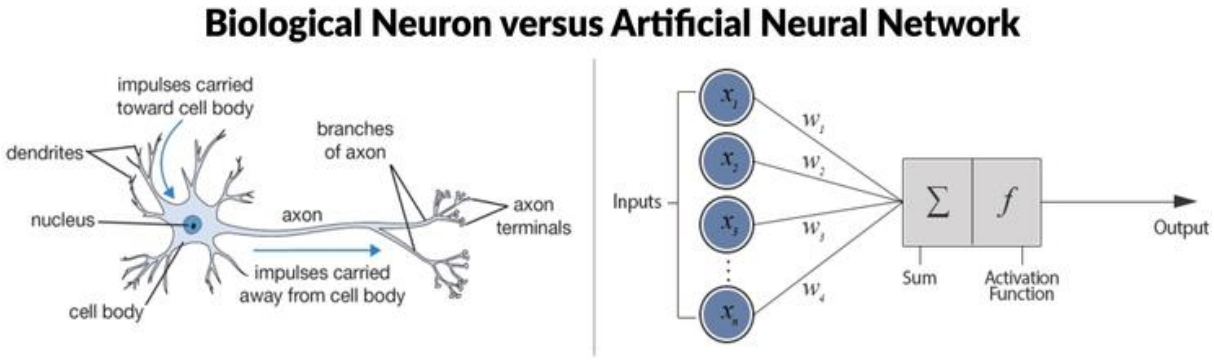
Az *M=1* esetet az ökonometriában „single index” modellnek nevezik. Ebben az esetben a súlyvektor egy egyelemű skalárrá egyszerűsödik. Ekkor egydimenziós simítási feladatunk van, amit bármely simító eljárással meg tudunk oldani, hogy megkapjuk a $g$-t. Kissé általánosabb, mint a lineáris regresszió, de hasonló értelmezést ad. (Hastie et al., 2001)

### 2.1.2 Mesterséges neurális hálózat

A neurális hálózat számos modellt és tanulási módszert magába foglaló kifejezéssé nőtte ki magát. A mesterséges neurális hálózat egy speciális információfeldolgozó rendszer. Sok egyszerű feldolgozóelemből áll, amelyek egymással irányított, súlyozott kapcsolatokkal vannak összekötve. Működésekor az egyes elemek a hozzájuk érkező információt feldolgozzák és a kapcsolatokat megfelelően súlyozva más feldolgozóelemekhez továbbítják (Borgulya, 1998). A neurális hálózat működését és felépítését az emberi agy működéséhez hasonlíthatjuk, mindkettőre jellemző, hogy párhuzamosan működő, központi irányítás nélküli egységekből (neuronokból) áll, melyek összeköttetésben állnak egymással (a neuronok a bemeneti

információt továbbítják más neuronok számára), az elemek közti összeköttetéseket (kapcsolatokat) pedig a kapcsolat erősségét kifejező súlyok (weights) jellemzik.

Az ember cselekvéseinek csak kisebb része ösztönös, döntő többségük tudatos, tanult mozzanat. Az agy, mint irányító központ, egy életen keresztül döntéseket hoz, amelyek külső vagy belső ingerekre érkező válaszreakciók. Hogy egy ingerre ki, hogyan reagál szinte teljesen egyedi – minél komplexebb az impulzus, annál bonyolultabb a válasz és annál nehezebb előrejelezni az egyén reakcióját. A 2. ábra bal oldala egy természetes neuron felépítését szemlélteti a központi sejttesttel és az abból kiinduló nyúlványokkal. A nyúlványok közül egy, a tengelynyúlvány (axon) vezeti el az ingerterületet a neuronból. A neuron többi nyúlványa, a dendritek a sejttesttel együtt az ingerület átvételére szolgálnak. Az ingerület átadása a szinapszisban történik sajátos vegyületek közvetítésével. A szinapszisban egy axon továbbítja az ingerületet egy másik neuronnak, melyben a dendritek, vagy maga a sejttest veszi fel az ingerületeket (Borgulya, 1998). Az ábra bal oldala pedig a mesterséges neuront ábrázolja, a $w$-k a szinapszist, az aktivációs függvény az axont, az öszegző (mátrix) pedig a sejttestet helyettesíti. A mesterséges neuron outputja lehet egy darab regresszió esetén, vagy több darab klasszifikáció esetén.




**2. ábra. Az agyi és a mesterséges neuron felépítése (Devopedia, 2019)**

Az emberi agy hozzávetőleg $10^{11}$ idegsejtet, és $10^{14}$ - 5 ·$10^{14}$ szinapszist tartalmaz; az idegsejt bemenő kapcsolatainak száma dendritjei kiterjedésétől függ, a szinaptikus bemenetek száma széles határok között, 1 és nagyjából 100000 között változhat. A neuronok közötti kapcsolatok nem állandóak, új kapcsolatok jöhetnek létre, meglévők szűnhetnek meg, a szinaptikus kapcsolatok erőssége módosulhat, idegsejtek egyes csoportjai pedig akár helyet is változtathat. Az idegrendszer korlátozott keretek között új idegsejtek előállítására, valamint sérült részek helyreállítására is képes. Ezek a jelenségek, valamint az idegrendszer felépítése

feltételezések szerint kapcsolatban állnak a memóriával, az agy tanulási, érzékelési, valamint az alkalmazkodás és az intelligencia képességével (Nebehaj, 2010).

A rövid felvezető után, amivel megpróbáltam szemléltetni az agyi neuronok és a mesterséges neuronok közötti hasonlóságot, áttérek a mesterséges neurális hálók matematikai modelljére.

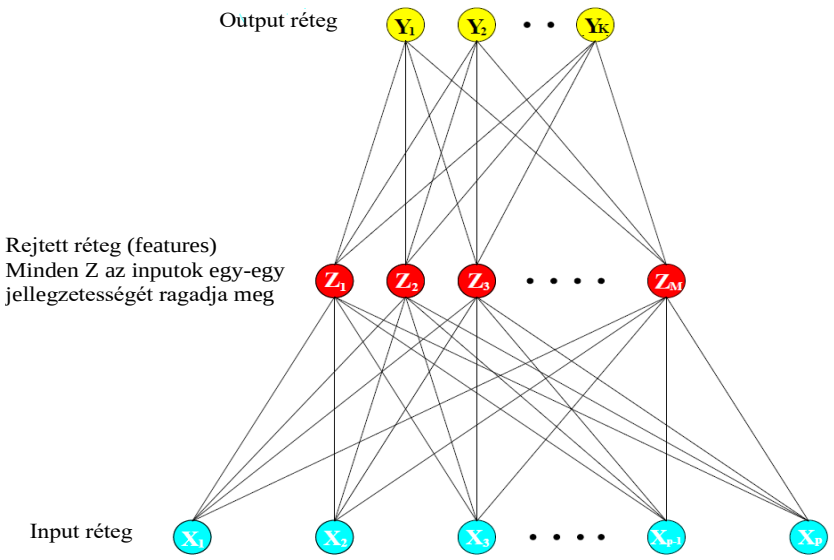
Definiáljuk az ún. $Z_m$ feature -öket, amelyek az $X_p$ lineáris kombinációjaként állnak elő, majd pedig az $Y_k$ outputok a $Z_m$ lineáris kombinációjaként állíthatók elő.

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), \qquad m = 1, \dots, M,$$

$$T_k = \beta_{0k} + \beta_k^T Z, \qquad k = 1, \dots, K,$$

$$f_k(X) = g_k(T), \qquad k = 1, \dots, K,$$

ahol $Z = (Z_1, Z_2, \dots, Z_M)$, $T = (T_1, T_2, \dots, T_K)$ és $\sigma$ az aktivációs függvény, amely az információ továbbítását szolgálja és elvégzi a transzformációt, ami végül az outputot adja. Az aktivációs függvény n-nek egy lineáris vagy nem lineáris függvénye, amelyet úgy választunk meg, hogy az alkalmas legyen a probléma megoldására. A $\sigma(v)$ aktivációs függvénynek leggyakrabban a szigmoid $\sigma(v) = 1/(1 + e^{-v})$ függvényt választják, viszont e mellett gyakori még: az úgynevezett hard limit (szigorú korlátos aktivációs függvény), melynek értéke 0, ha a függvény argumentuma kisebb, mint 0 és 1 ha az argumentum 0 vagy annál nagyobb, valamint a lineáris, amelynek értéke megegyezik az inputtal. (Vakhal, 2019).




**3. ábra. Egy rejtett rétegű mesterséges neurális háló (Hastie et al., 2001)**

A $g_k(T)$ output függvény a $T$ vektor utolsó transzformációja. Lineáris regresszió esetén általában az identitás függvényt válasszuk, ami $g_k(T) = T$. Korábban a klasszifikációs ($K > 1$) problémáknál is identitás függvényt használtak, de ezt később felváltotta a softmax függvény:

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}}.$$

A háló középső részén lévő $Z_m$ réteget rejtett rétegnek nevezzük (melyekből általában több is lehet), mivel a $Z_m$ értékeit közvetlenül nem figyeljük meg, rájuk úgy is gondolhatunk, mint az eredeti inputok bázis kiterjesztésére.

A neurális hálózat modellnek ismeretlen számú paramétere van, melyeket súlyoknak neveznek. Feladatunk megtalálni azokat az értékeket, amelyekkel a modell a legjobban illeszkedik az adatokra. Legyen $\theta$ a súlyok és a konstans tagok halmaza. Feladatunk az $R(\theta)$ minimalizálása gradiens eljárás segítségével. Ezt backpropagationnek (visszacsatolás) nevezzük. (Hastie et al., 2001). A gradiens eljárást alkalmazó illesztési mechanizmus miatt szükség van a differenciálhatóságra.

Folytonos $R(\theta)$ esetben:

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} \left(y_{ik} - f_k(x_i)\right)^2$$

Diszkrét $R(\theta)$ esetben (kereszt-entrópia eltérés):

$$R(\theta) = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} log f_k(x_i)$$

A kereszt-entrópia egy hibaértékelési, modellilleszkedési mutató, amely a következőképp adható meg: $R(\Theta) = -\frac{1}{N}\sum y^i lna\left(x^i\right) + \left(1 - y^i\right)\ln\left(1 - a\left(x^i\right)\right)$, ahol $x$ az input adat, $y$ a megfelelő output, $a(x)$ a modell által adott output. Ez csak diszkrét esetben használható fel (Vakhal,2019).

Felhasználva a korábban definiáltakat, legyen $z_m = \sigma(\alpha_{0m} + \alpha_m^T x_i)$ és $z = (z_{1i}, z_{2i}, \dots, z_{Mi})$ akkor

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} \left(y_{ik} - f_k(x_i)\right)^2 = \sum_{i=1}^{N} R_i$$

deriválás után

$$\frac{\partial R_i}{\partial \beta_{km}} = -2\big(y_{ik} - f_k(x_i)\big)g_k'(\beta_k^T z_i)z_{mi}$$

$$\frac{\partial R_i}{\partial \alpha_{ml}} = -\sum_{k=1}^{K} 2\big(y_{ik} - f_k(x_i)\big)g_k'(\beta_k^T z_i)\beta_{km}\sigma'(\alpha_m^T x_i)x_{il}$$

A deriváltak $r+1$-ik iterációja után megkapjuk

$$\beta_{km}^{r+1} = \beta_{km}^{r} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_{km}^{(r)}}$$

$$\alpha_{ml}^{r+1} = \alpha_{ml}^{r} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \alpha_{ml}^{r}}$$

A $\gamma_r$ együtthatókat tanulási rátáknak nevezzük, amely a tolerálható hibát szabályozza. Értéke általában 0,0001 és 0,4 között van.

Ha átírjuk az első deriváltak egyenleteit

$$\frac{\partial R_i}{\partial \beta_{km}} = \delta_{ki} z_{mi}$$

$$\frac{\partial R_i}{\partial \alpha_{ml}} = s_{mi} x_{il}$$

ahol a $\delta_{ki}$ és $s_{mi}$ értékek a hiba tagok a modell kimeneti értékénél és a rejtett rétegeinél. Ezek a hibák kielégítik a következő egyenletet:

$$s_{mi} = \sigma'(\alpha_m^T x_i) \sum_{k=1}^{K} \beta_{km}\delta_{ki}$$

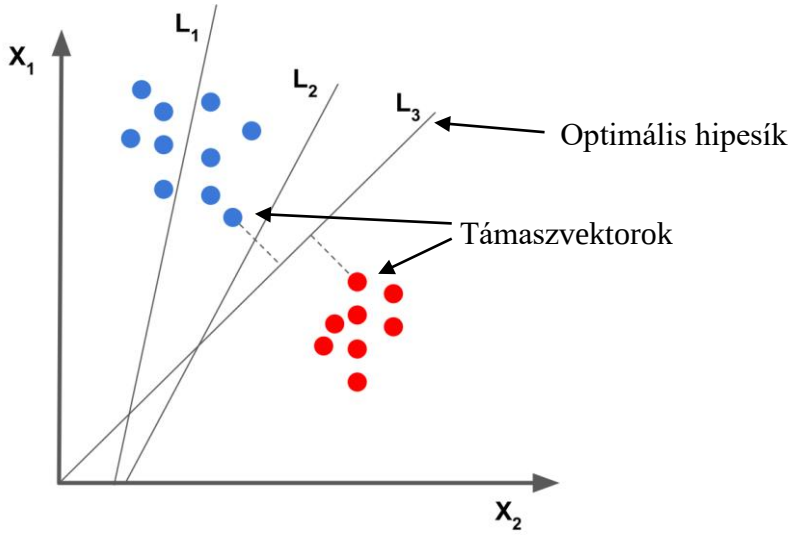
melyet backpropagation egyenletnek nevezünk. Ezt alkalmazva az első lépésben megtörténik a becslés az aktuális súlyokkal, majd miután a modell kiszámolja a hibát az visszacsatolódik és új hibát ad, azután a folyamat kezdődik elölről. Ezt alkalmazva az első lépésben az aktuális súlyok rögzítettek, az $\hat{f}_k(x_i)$ prediktorok pedig az egyenletből becsülhetők vissza. A következő lépésben a $\delta_{ki}$ hibák számolódnak az egyenletből, majd visszacsatolódnak az egyenletbe, megadva az új $s_{mi}$ hibákat, majd a folyamat kezdődik elölről. Ezt a két lépéses eljárást nevezzük backpropagationnek, vagy delta szabálynak (Widrow, Hoff, 1960).

## 2.2. Támaszvektor-gépek (Support Vector Machine – SVM)

Ebben a részben bemutatom a Support Vector Machine (SVM) elméletét. Az SVM egy viszonylag új gépi tanulási (machine learning) módszer, de a többihez hasonlítva ez az egyik legpontosabb. Felhasználható regresszióra, klasszifikációra, klaszterezésre. Csak numerikus prediktorok kezelésére alkalmazható. Az SVM kialakulása Vapnik és Chervonenskis nevéhez fűződik, akik 1963-ban bemutatták a lineáris SVM-et. Később Boser-Guyon-Vapnik 1992-es munkájukban bevezették a nemlineáris SVM-et, a modern SVM, amelyet napjainkban is használunk Cortes−Vapnik 1995-ös munkájukban fejlesztették ki (Vékás, 2019).

### 2.2.1. Lineáris SVM

A lineáris SVM kiinduló feladatban a prediktorok $\mathbb{R}^p$ terében egy optimális $p-1$ dimenziós hipersíkot keres, amellyel tökéletesen elválasztható a kimenet két kategóriája egymástól. A következő ábra a kétdimenziós esetet illusztrálja:




**4. ábra. Lineáris szeparálás kétdimenzióban (Mallick, 2018)**

Jól látható, hogy az $L_1$ egyenes nem jól választja el a változókat, az $L_2$ és $L_3$ már elfogadható, viszont a megfelelő egyenes az $L_3$ lesz, mivel az „tisztábban” szeparálja a csoportokat. Optimális hipersík esetén maximális a hipersík és a hozzá legközelebb eső pont távolsága. Ezeket a pontokat támaszvektoroknak – ahonnan az eljárás neve is ered – nevezzük (Awad et al., 2015).

A hipersík egyenlete felírható a következőképp:

$$w \cdot x - b = 0,$$

ahol egy $x_0$-ról behelyettesítéssel eldönthető, hogy a hipersík melyik oldalán helyezkedik, így a megkapott érték eldönti, hogy az input melyik csoportba sorolható. Felmerül a probléma, hogy ha létezik a térnek egy ilyen felbontása, akkor létezik végtelen sok.

A megoldáshoz normalizálni kell a $w$ súlyvektort $\left(\frac{w}{\|w\|}\right)$, ekkor a támaszvektorokra igaznak kell lennie a következő egyenletek egyikének:

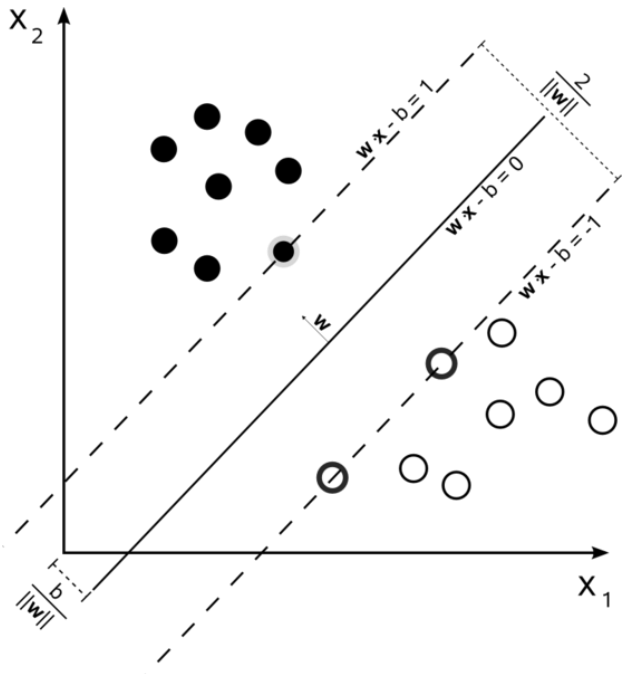
$$w \cdot x - b = +1$$

$$w \cdot x - b = -1$$

A feladat a támaszvektorok és a hipersík közötti távolság maximalizálása, tehát hogy a hipersík minél jobban szétválassza a változókat. Erről a távolságról belátható, hogy az értéke $\frac{1}{\|w\|}$, ebből következik, hogy a maximalizálási feladat helyett, a $\|w\|$ minimalizálása a feladat.

$$\min_{w,b} \|w\|:$$

$$y_i(w \cdot x_i - b) \geq 1 \qquad (i = 1,2, \dots, n)$$




**5. ábra. A határmezsgye (margin) ábrázolása (Ismeretlen szerző, 2015)**

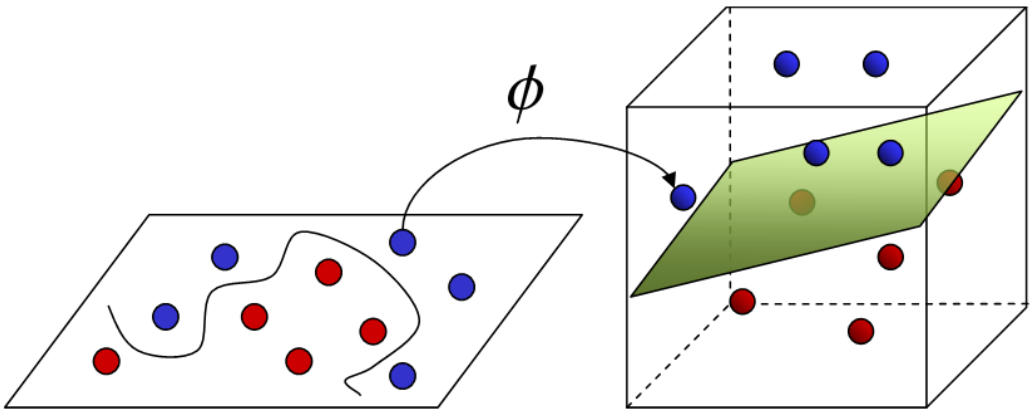
A hipersík két oldalán lévő támaszvektorok és a hipersík közötti sávot határmezsgyének (margin) nevezzük. A határmezsgye szélessége $\frac{2}{\|w\|}$.

Abban az esetben, ha a két csoport lineárisan nem szeparálható megoldást jelenthet a nemlineáris SVM (Vékás, 2019).

### 2.2.2. Nemlineáris SVM

A való életben a legtöbb adathalmaz lineárisan nem szeparálható. Ekkor megfogalmazható a Cover-tétele.

*Cover-tétel.* Két csoport nagyobb valószínűséggel lineárisan szeparálható, ha a prediktorokat egy $\phi$ nemlineáris transzformációval magasabb dimenziójú térbe transzformáljuk (Cover, 1965).




**6. ábra: Nemlineáris transzformáció $\phi$: $\mathbb{R}^2 \rightarrow \mathbb{R}^3$ (ismeretlen szerző)**

Ekkor optimalizációs feladat a következőre módosul, ahol a $\phi$ az említett transzformációs függvény:

$$\min_{w,b} \|w\|:$$

$$y_i(w \cdot \phi(x_i) - b) \geq 1 \qquad (i = 1,2,\dots,n)$$

A $\phi$ transzformációs függvényt bázisfüggvénynek nevezzük.

Felmerül a kérdés, hogy milyen transzformációt használjunk. Erre bevezetjük az úgynevezett magfüggvényt (kernel) (Vékás, 2019).

*Definíció.* Legyen $\phi: \mathbb{R}^p \to \mathbb{R}^q$ a magasabb dimenziójú térbe transzformáló nemlineáris függvény. Akkor a $\kappa: \mathbb{R}^{2p} \to \mathbb{R}$ magfüggvény az eredeti tér két vektorára megadja az új, transzformált térben vett skaláris szorzatukat $\kappa(u, v) = \phi(u) \cdot \phi(v)$.

A következő magfüggvények a leggyakrabbak a gyakorlatban:

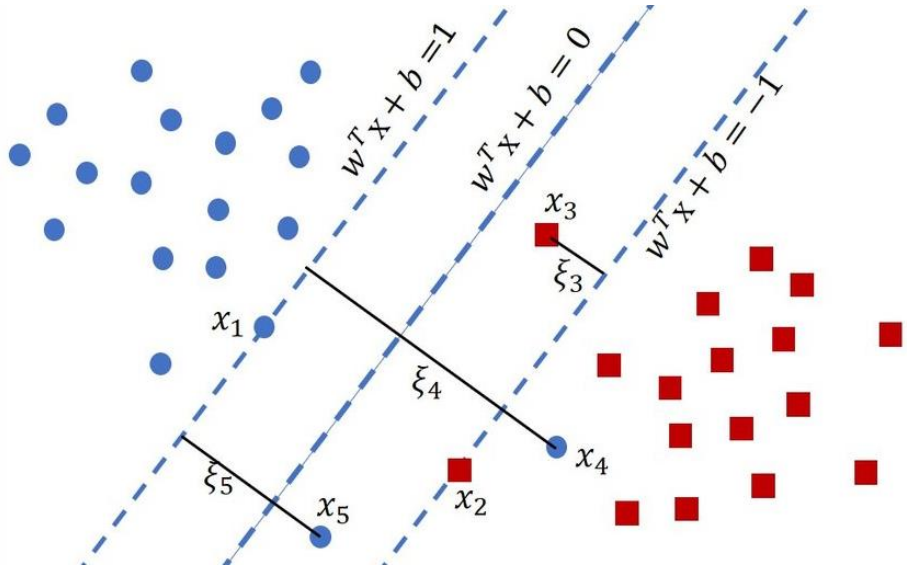
| **Név** | $\boldsymbol{\kappa(u, v) =}$ |
|---|---|
| **Lineáris** | $u \cdot v$ |
| **Polinomiális** | $\gamma(u \cdot v + c_0)^d$ |
| **Gauss (Radiális)** | $\exp(-\gamma\|u - v\|^2)$ |
| **Szigmoid** | $\tanh(\gamma u \cdot v + c_0)$ |

**1. táblázat. Gyakori magfüggvények (saját táblázat)**

A magfüggvények használatára azért van lehetőség, mert a hipersík egyenletében csak azokat a pontokat kell figyelembe venni, amelyek a támaszvektorokat határozzák meg (Virtás, 2019).

### 2.2.3. Modern SVM

A nemlineáris SVM csak abban az esetben működik, ha a magasabb dimenziójú, transzformált térben már lineárisan szeparálható a két csoport. Azonban ez a gyakorlatban nem mindig van így, gyakran nem lehetséges, vagy ha lehetséges is, nem biztos, hogy a kívánt eredményhez vezet. E probléma leküzdésére fejlesztették ki a modern SVM-et, amely nem törekszik tökéletes szeparálásra és megengedi, hogy a megfigyelések közül, néhány a határmezsgyébe vagy a hipersík rossz oldalára kerüljön és ezt majd a célfüggvényben bünteti.




**7. ábra. Soft-margin (puha határ) SVM (ismeretlen szerző)**

Ezt nevezzük soft-margin (puha határ) SVM-nek. Tehát a modern SVM az egy puha határral kiegészített nemlineáris SVM.

Az ábrán látható $\xi_i$-k a határsértés mértékei. A határsértés mértéke attól függ, hogy a változó hol helyezkedik el határhoz képest. $\xi = 0$, ha nem történt határsértés, $0 < \xi < 1$, ha a határmezsgyén belül található, $\xi > 1$, ha a határ rossz felén található a változó (Shalev-Shwartz, Ben-David, 2014).

Ebben az esetben az optimalizációs feladat a következő lesz:

$$\min_{w,b} \left( \|w\| + c \sum_{i=1}^{n} \xi_i \right):$$

$$y_i(w \cdot \phi(x_i) - b) \geq 1 - \xi_i \qquad (i = 1,2, \dots, n)$$

$$\xi_i \geq 0 \qquad (i = 1,2, \dots, n)$$

A c paramétert cost (költség) paraméternek nevezzük, amivel büntetjük a modellt a határsértések miatt. A c paraméter értékeit általában 2 hatványának, a $\gamma$-t pedig $\frac{1}{10}$ hatványának nagyságrendjében keressük. (Vékás, 2019)

Az SVM regressziós módszerként is használható, az algoritmus fő jellemzői megegyeznek a klasszifikációs esettel. Az Support Vector Regression (SVR) ugyanazokat az elveket használja a besoroláshoz néhány kisebb eltéréssel, mint az SVM. Mivel a kimenet

(output) valós szám, így nehezebb előrejelezni az információt, amelynek végtelen lehetősége lehet.

A következőkben bemutatom, hogyan használhatjuk fel az SVM-et regresszióra.

A lineáris regressziós modell

$$f(x) = wx^T + b.$$

A SVR esetében meghatározunk egy maximális $\varepsilon$ szórást, illetve bevezetjük a nem-negatív $\xi_i$ és $\xi_i^*$ segédváltozókat, melyek ahogyan az SVM esetében, úgy most is a határsértés mértékei, azonban ebben az esetben a hipersík $\pm\varepsilon$ sávján kívüli eltéréseket mérjük pozitív, illetve negatív irányban. Az $\varepsilon$-t veszteségfüggvénynek nevezzük.

Ekkor az optimalizációs feladat a következő lesz:

$$\min_{w,b}\left(\|w\| + c\sum_{i=1}^{n}\xi_i + \xi_i^*\right):$$

$$y_i(w \cdot x_i - b) \leq \varepsilon_i + \xi_i \qquad (i = 1,2,\dots,n)$$

$$y_i(w \cdot x_i - b) \geq -\varepsilon_i - \xi_i^* \qquad (i = 1,2,\dots,n)$$

$$\xi_i, \xi_i^* \geq 0 \qquad (i = 1,2,\dots,n)$$

Nem-lineáris esetben ugyanazt az eljárást kell követnünk, mint klasszifikációs esetben. Egy nem-lineáris transzformációval egy magasabb dimenziójú térbe transzformáljuk. Felhasználhatjuk ugyanazokat a magfüggvényeket, amelyeket az SVM-nél már megtekintettünk(Awad et al., 2015).

## 2.3. Döntési fa, Véletlen erdő

### 2.3.1. Döntési fa

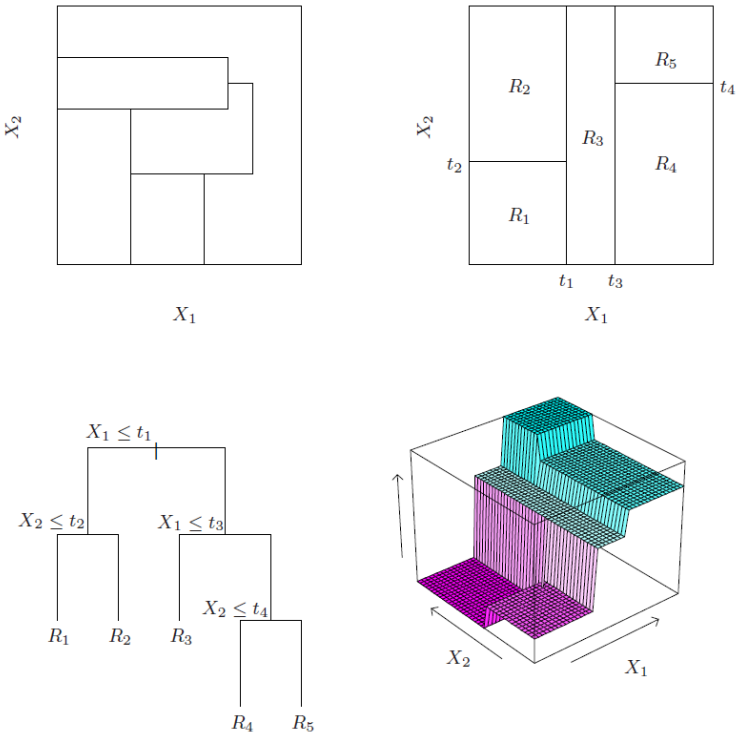
A döntési fa alapú algoritmusok az egyik legjobb és leginkább alkalmazott gép tanulási módszerek egyike, ahol egyszerű lépések segítségével hozhatunk meg bonyolult döntéseket. A fa alapú eljárás előnye, hogy lehetővé teszi a prediktív modellek nagy arányú pontosságát, stabilitását és könnyű értelmezését, valamint jól ábrázolható. Hátránya, hogy könnyen hajlamos a túlillesztésre. A lineáris modellektől eltérően, a nemlineáris kapcsolatokat is meglehetősen jól ábrázolják. Alkalmas mind regresszióra, mind pedig osztályozásra (CART - Classification And Regression Trees). Ebben a fejezetben csak a regressziós módszerről lesz szó.

Figyeljünk meg egy regressziós problémát, ahol $Y$ egy folytonos magyarázott változó, valamint $X_1, X_2 \in (0; 1)$ inputok. A 9. ábra bal felső részén a teret felosztjuk a koordináta tengelyekkel párhuzamos vonalakkal. Minden mezőben modellezhetjük az $Y$-t különböző konstansokkal. Azonban akad egy probléma, annak ellenére, hogy minden felosztó vonal könnyen leírható, mint például $X_1 = c$, némelyik mezőt bonyolult lehet leírni.

A dolgok leegyszerűsítése érdekében rekurzív bináris partíciókra korlátozzuk a figyelmünket. Ahogyan a 9. ábra jobb felső részét tekintjük, ahol először felosztjuk a teret két mezőre és a reakciót az $Y$ átlagával modellezzük. A változókat és a felosztási pontokat (split-point) úgy választjuk meg, hogy a legjobb illeszkedést érjük el. Ezután azokat további két mezőre osztjuk és ez a folyamat addig folytatódik, amíg valamilyen megállási szabályt nem alkalmazunk. Például a 9. ábra jobb felső részén először $X_1 = t_1$ esetben felosztjuk a teret két részre, majd az $X_1 \leq t_1$ tartományt $X_2 = t_2$-nél. Az $X_1 > t_1$ tartományt $X_1 = t_3$-nál osztjuk fel, majd végül az $X_1 > t_3$-at $X_2 = t_4$-nél. A folyamat eredménye egy partíció, amit felosztottunk öt mezőre, $R_1, R_2, \dots, R_5$, melyek láthatók a 9. ábrán. A megfelelő regressziós modell, amely megadja az $Y$-t egy $c_m$ konstanssal az $R_m$ mezőben:

$$\hat{f}(X) = \sum_{m=1}^{5} c_m \, I\{(X_1, X_2) \in R_m\}.$$

Ez a modell ábrázolható egy bináris fával (9. ábra bal alsó része). A teljes adathalmaz a fa tetején van. Az elágazásoknál a megfigyeléseink a feltétel teljesítése esetén balra, ellenkező esetben jobbra kerülnek. A végső meghatározások – melyeket leveleknek nevezünk – felelnek azért, hogy az $X_1$ melyik mezőbe tartozik. A jobb alsó ábra ennek a modellnek a „perspektív ábrája”, az illusztráció kedvéért a $c_1 = -5, c_2 = -7, c_3 = 0, c_4 = 2, c_5 = 4$ konstansokat vettük (Hastie et al., 2001).




**8. ábra. Döntési fa (Hastie et al., 2001)**

Regressziós fa növesztéséhez tegyük fel, hogy az adathalmazunk tartalmaz $p$ inputot és reakciót, minden megfigyelésre az $N$-ből, ezek $(x_i, y_i)$ párok lesznek, ahol $i = (1,2, \dots, N)$ és $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$. Az algoritmusnak automatikusan el kell döntenie a felosztó változókat és a felosztási pontokat, valamint azt, hogy milyen alakja legyen a fának. Tegyük fel, hogy van egy partíciónk $M$ mezőre osztva $R_1, R_2, \dots, R_M$, és a reakciót mint $c_m$ konstans modellezzük minden mezőben:

$$f(x) = \sum_{m=1}^{M} c_m \, I(x \in R_m)$$

Ha kritériumként a legkisebb négyzetek összegét használjuk $\sum(y_i - f(x_i))^2$, akkor könnyen látható, hogy a legjobb $\hat{c}_m$ az $y_i$ átlaga lesz, feltéve, hogy az $x_i \in R_m$. A legjobb bináris partíció megtalálásakor a legkisebb négyzetek összegének módszerével számítási nehézségekbe ütközhetünk, emiatt a következő algoritmust használjuk: az összes adattal kezdünk, felteszünk egy j felosztó változót és egy s felosztási pontot és definiáljuk a két félsíkot

$$R_1(j,s) = \{X | X_j \leq s\} \text{ és } R_2(j,s) = \{X | X_j > s\}$$

Ezután keressük azt a $j$ felosztó változót és azt az $s$ felosztási pontot, amely megoldása a következő egyenletnek:

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right]$$

Miután megtaláltuk a legjobb felosztást, az adatokat felosztjuk a két rész alapján, majd mindkét részben megismételjük a folyamatot és így tovább, így megkapjuk a döntési fánkat.

Felmerül a kérdés, hogy milyen nagyra növesszük a fát? Egy nagyon nagy fa könnyen túlillesztődik, ami azt jeleni, hogy a fának sok ága képződik és a levelek szinte homogének, így akár az is előfordulhat, hogy a modell tökéletesen illeszkedik az adatokra, ami más adatokon sokkal rosszabb eredményt adna. Míg egy túl kicsiből kimaradhatnak fontos elemek(Dávid, 2019).

A fa mérete meghatározza a modell komplexitását, az optimális méretét adaptívan kell kiválasztanunk az adatokból. Az ajánlott eljárás erre az, hogy növesszünk egy nagy $T_0$ fát, megállítva a felosztási folyamatot, amikor a levelek száma elér egy előre meghatározott mennyiséget (például 5-öt), ezután ezt a nagy fát megmetszük (pruning).

Definiálunk egy alfát $T \subset T_0$, egy fa, amelyet a $T_0$ metszésével kaphatunk meg. A végső leveleket $m$-el indexáljuk, az $m$-ik levél az $R_m$ mezőbe tartozást reprezentálja. Legyen $|T|$a belső levelek száma a $T$ fán. Továbbá legyen

$$N_m = \#\{x_i \in R_m\},$$

$$\hat{c}_m = \frac{1}{N_m}\sum_{x_i\in R_m} y_i,$$

$$Q_m(T) = \frac{1}{N_m}\sum_{x_i\in R_m}(y_i-\hat{c}_m)^2,$$

ekkor definiáljuk a költség komplexitási kritériumot

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m\, Q_m(T) + \alpha|T|.$$

Az ötlet az, hogy minden $\alpha$ számára megtalálni azt $T_\alpha$-t ($T_\alpha \subset T_0$), amely minimalizálja a $C_\alpha(T)$-t. Az $\alpha \geq 0$ finomhangoló paraméter optimalizálja a fa méretét,

figyelembe véve a legjobb illeszkedést az adatokra. A nagy $\alpha$ érték kis fát, míg a kis $\alpha$ nagy fát eredményez (James et al., 2013).

Minden $\alpha$-ra meg lehet mutatni, hogy létezik egy egyedi legkisebb $T_\alpha$, amely minimalizálja a $C_\alpha(T)$-t. Ezen $T_\alpha$ megtalálásához a leggyengébb kapcsolat metszést használjuk. Ebben az esetben tovább folytatjuk a metszést, míg nem kapunk egy egylevelű fát. Ez megad egy alfa sorozatot, amely tartalmazza a $T_\alpha$-t. Az $\alpha$ becslésére keresztvalidációval kerül sor, azt az $\hat{\alpha}$ értéket válasszuk, amelyik minimalizálja a keresztvalidált négyzetösszegeket. A végül megkapott fánk a $T_{\hat{\alpha}}$ lesz.

### 2.3.2. Véletlen erdő

Azonban egyedül a döntési fák, gyakran rosszabbul teljesítenek, mint egy egyszerű regresszió. A predikció pontosságának növelése érdekében létezik a véletlen erdők módszere. A döntési fákra jellemző a magas variancia, amiből kifolyólag, ha például az adathalmazunkat két részre osztjuk véletlenszerűen és döntési fát illesztünk mindkét felére, a kapott eredmények eltérhetnek egymástól. Erre a problémára megoldást jelent a bagging. A bagging egy általános eljárás a statisztikai tanuló eljárások varianciájának csökkentésére. A bagging (bootstrap és aggregation összeolvadása) azon az összefüggésen alapszik, hogy n darab $\sigma^2$ varianciájú független azonos eloszlású valószínűségi változó átlagolása csökkenti az átlag varianciáját. A variancia csökkentése és a predikció pontosságának növelése egy statisztikai tanuló módszer esetén arra lenne szükség, hogy a populációnak sok különböző mintájára építsünk külön modelleket és az eredményeket átlagoljuk. De mivel általában csak egy minta áll rendelkezésre ezért bootstrappelnünk kell (Virtás, 2019). A bootstrappelés különböző minták létrehozása visszatevéses mintavétellel. Ebben az esetben generálunk B darab bootstrappelt mintát. Ezeken hajtjuk végre a bagginget, a fák metszése nélkül, így végül kapunk B darab predikciót, ami átlagolva megadja a végleges predikciót

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x).$$

A véletlen erdő a fentebb írt bagging eljárás továbbfejlesztett verziója. A véletlen erdő a baggingelt fák javulását biztosítja azzal, hogy csökkenti a korrelációt a változók között. Ahogyan a bagging esetében, úgy ebben az esetben is a bootstrappelt adatokra építünk számos fát. Viszont minden esetben amikor egy fán vágásra kerül sor, a $p$ darab változó helyett, csak

$m$ darab véletlenszerűen választott változó alapján vágunk. Minden egyes vágás esetén újabb $m$ véletlenszerű változót választunk ki. Alapesetben az $m \approx \sqrt{p}$ (James et al., 2013), de keresztvalidációval megtalálható az optimális $m$ érték.

## 3. Adatgyűjtés, elemzés és eredmények

Az elméleti áttekintés után térjünk rá a gyakorlati felhasználásra. Szakdolgozatomban egy egészségbiztosítási állomány várható kárkifizetéseit jelzem előre gépi tanulási módszerek segítségével. Ezt a már fentebb leírt három módszer mindegyikével elvégzem, majd összehasonlítom a felépített modellek hatékonyságát és kiválasztom a legjobb modellt. A modellek jóságát két teljesítmény mutatóval mértem, az $R^2$-el, valamint az átlagos négyzetes hiba négyzetgyökével (RMSE – Root Mean Squared Error). Munkám első lépése az adatok tanulmányozása és leíró statisztikák készítése volt.

### 3.1. Adatok bemutatása

Az elemzéshez használt adatok egy valós demográfiai statisztikai adatokon alapuló szimulált adatbázis, melyet az Amerikai Egyesült Államok Népszámlálási Irodája hozott létre (Lantz, 2013). Magát az adatbázist a Kaggle adatelemzési platformról töltöttem le (Choi, n.d.).

Az adathalmaz 1338 egyén adatait tartalmazza. Ezek az adatok a következő ábrán láthatóak:

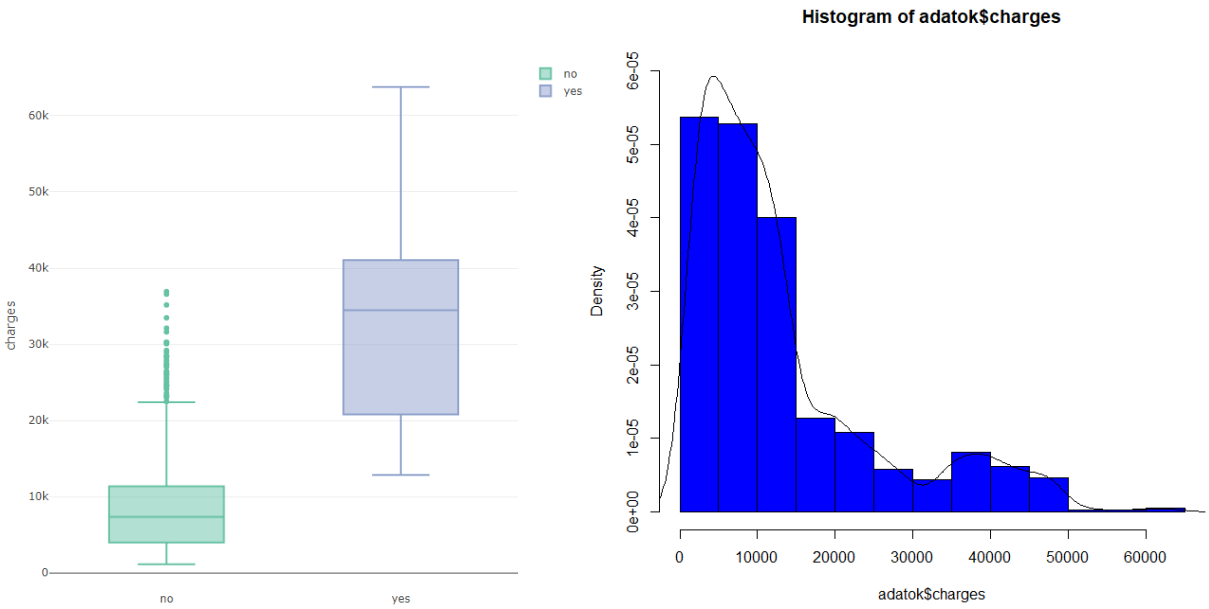
| | age | sex | bmi | children | smoker | region | charges |
|---|---|---|---|---|---|---|---|
| 1 | 19 | female | 27.900 | 0 | yes | southwest | 16884.924 |
| 2 | 18 | male | 33.770 | 1 | no | southeast | 1725.552 |
| 3 | 28 | male | 33.000 | 3 | no | southeast | 4449.462 |
| 4 | 33 | male | 22.705 | 0 | no | northwest | 21984.471 |
| 5 | 32 | male | 28.880 | 0 | no | northwest | 3866.855 |
| 6 | 31 | female | 25.740 | 0 | no | southeast | 3756.622 |

**9. ábra. Minta az adatbázisból (saját ábra)**

Ahol az age az egyén korát, a sex a nemét, a bmi a testtömegindexét[2], a children a gyerekei számát, a region az USA négy régióján belül a lakóhelyét, a charges pedig az egy év alatt bekövetkezett egészségügyi kiadások mértékét jelöli, valamint a smoker azt, hogy az ügyfél dohányzik e. Célom, hogy ezen mutatók alapján minél pontosabban meg tudjam becsülni egy-egy biztosított várható egészségügyi kiadásait egy évre, valamint kivizsgálni, hogy melyek azok, amelyek a leginkább hatnak az egészségügyi kiadások mértékére.

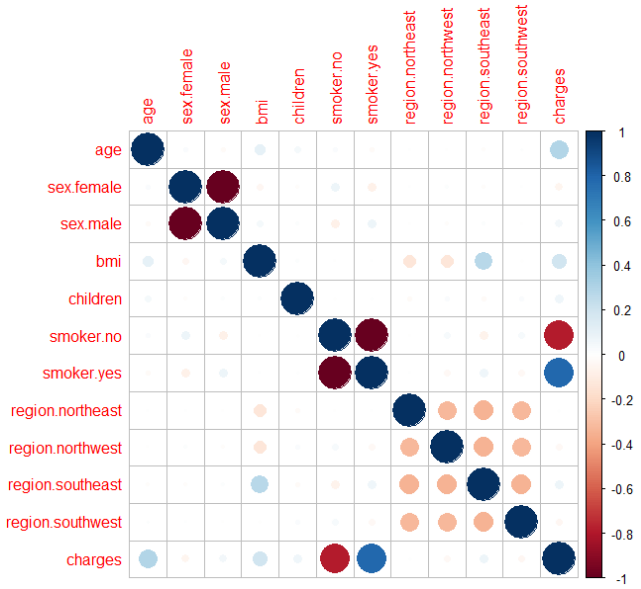
[2] A testtömegindexet ma széles körben alkalmazzák az egészséges testsúly, túlsúlyosság vagy soványság meghatározására.(*Testtömegindex* , n.d.)

Az egyének átlagéletkora 39.2 év, átlagosan 1 gyerekük van és évente átlagosan 13270 dollárt fordítanak egészségügyi kiadásokra.




**10 ábra. Az egészségügyi kiadások boxplot diagramja, feltéve, hogy az egyén dohányzik vagy sem; a kiadások hisztogramja sűrűségfüggvénnyel ábrázolva (saját ábra)**

Ahogyan a fenti boxplot diagrammon látható, a dohányzóknak átlagosan sokkal több kiadásuk van egy évben. Ez okozza azt is, ahogyan a hisztogrammon is látható, hogy a kiadások eloszlása két móduszú.



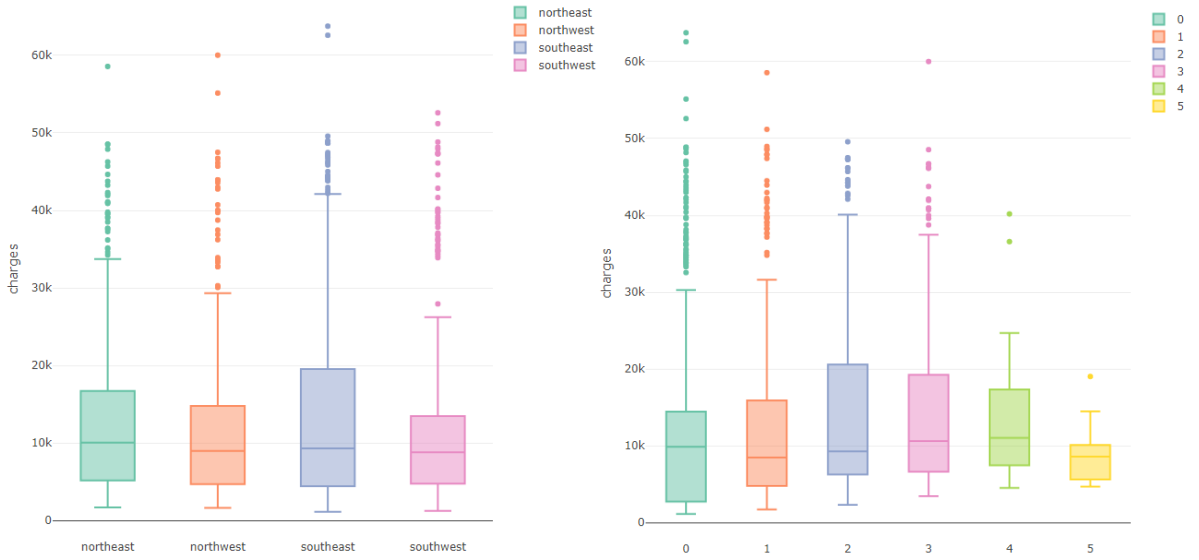

**11. ábra. A változók korrelációs diagramja (saját ábra)**

A 11. ábrán látható a változók korrelációs diagramja, ahol a piros szín a negatív, a kék pedig a pozitív korrelációt mutatja. Jól látható, hogy az egészségügyi kiadásokkal leginkább a dohányzás korrelál, utána a kor és a testtömegindex következik.

Ábrázolva a kiadásokat a testtömegindex függvényében külön választva a dohányosokat a nem dohányosoktól (12. ábra), megállapítható, hogy mind a két esetben a testtömegindex növekedésével nő a várható egészségügyi kiadások mértéke, azonban ha illesztünk rájuk egy regressziós egyenest, a dohányos állományra illesztet egyenes meredeksége sokkal nagyobb, tehát egyértelműen kijelenthető, hogy azok az egyének akik dohányoznak és emellett még túlsúlyosak is várhatóan több kiadással kell szembenézniük, mint azoknak akik csak túlsúllyal szenvednek.


**12. ábra. Az egészségügyi kiadások a: testtömegindex függvényében (bal), kor függvényében (jobb) (saját ábra)**

A fenti 12. ábra. jobb oldali ábráján az egészségügyi kiadások láthatók a kor függvényében. Mint az általában igaz, így ebben az esetben is megállapítható, hogy minél idősebb valaki, annál nagyobb az egészségügyi kiadásainak mértéke. Azonban az is látszik, hogy azok, akik dohányoznak átlagosan sokkal többet költenek egészségük megőrzésének érdekében. Tulajdonképpen egy ugyanolyan korú dohányos kiadása ~20 ezer dollárral több, mint nem dohányzó társáé.




**13. ábra. Boxplot diagram a (bal) lakóhely szerinti, valamint a (jobb) gyerekek száma szerinti kiadásokról (saját ábra)**

A lakóhely és a kiadások boxplot diagramján látható (14. ábra – bal), hogy a medián régiónként majdnem megegyező, az észak-keleti régióban azonban egy kicsit magasabban van ennek értéke. Minden régióban sok az outlier, ezek kiszűrésével általában javítani lehet a modell teljesítményét, azonban ebben az esetben ez csak ront rajta, így benne hagytam az adatbázisban. A délkeleti régió lakosainak felső kvartilise kiemelkedően magasabban van, mint a többi három régiónak. A gyerekek száma és a kiadások boxplot diagramján látható (14. ábra – jobb), hogy azon egyéneknek, akiknek nincs gyerekük a leginkább szóródott a kiadások mértéke, az alsó kvartilisban többen is vannak és alacsonyabban is van, mint a gyerekes személyek esetében, azonban a medián magasabban van, mint az egy vagy két gyerekesek esetében, valamint rendkívül extrém értékek is előfordulnak náluk. Az öt gyerekkel rendelkezők esetén az egész terjedelem egy szűk tartományon van, azonban ilyen egyénekből csak 18 van az adatbázisban, így nem biztos, hogy elegendő adat van annak megállapításához, hogy a sok gyerekkel rendelkezők kevesebbet költenek egészségügyre.

### 3.2. Elemzés és eredmények

Az elemzést mind a három módszer esetén az R nyílt forráskódú statisztikai szoftverrel végeztem. Minden esetben egy, az adott módszerhez fejlesztett csomagot használtam. Így a

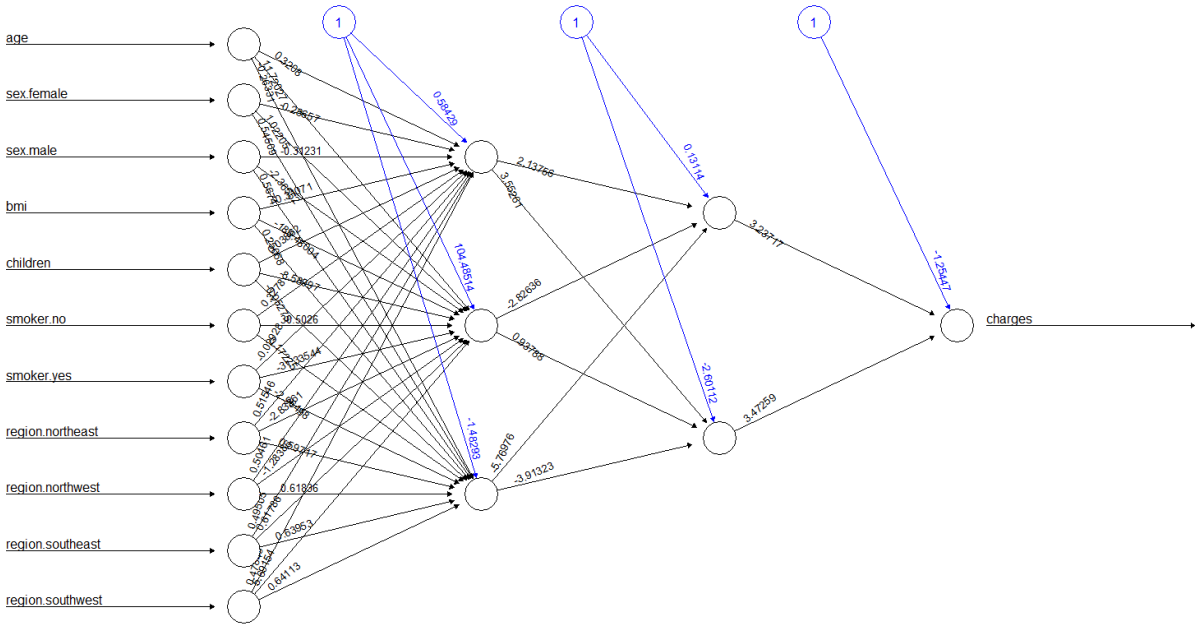
véletlen erdőt a randomForest, az SVM módszert az e1071, a mesterséges neurális hálókat a neuralnet csomag segítségével tudtam felépíteni.

Első lépésként a kategorikus változókból kellett dummy változókat képeznem. Ezt a caret csomag segítségével, az úgynevezett „one-hot” kódolással végeztem el.

Ezután két részre osztottam az adatbázisomat – 70% - 30% arányban – tanuló és tesztelő adatbázisra. A tanuló adatbázisra építettem fel a modelleket, melyek közül az első a neurális hálók volt. A tesztelő adatbázist később validációra használom a célváltozó kihagyásával.

A neurális háló gépi tanulási módszer megköveteli a sztenderdizált, vagy normalizált változókat, a gyorsabb konvergencia és a pontosabb modell érdekében. A neurális háló tanítási ideje a rejtett rétegek és a neuronok számának növelésével rendkívül megnőhet, így a futási ideje meghaladja a többi módszerét. Gyakran előfordul az a probléma is, hogy nem konvergál az optimális megoldásba, ekkor a paraméterek változtatásával újra futtatva a tanítást megkapjuk a kívánt eredményt.

Először egy egyrétegű egy neuronból álló hálót próbáltam, ami lényegében megfelel egy lineáris regressziónak, identitás aktivációs függvénnyel. Miután a tanuló adatbázison betanulta a modell, a tesztelő adatbázison lemértem a háló jóságát, úgy, hogy a tesztelő adatbázisra futtattam egy predikciót, amit összehasonlítottam a tesztelő adatbázis eredeti célváltozóinak értékével. Ez 74%-os $R^2$ – et adott, valamint 0,479 RMSE értéket.




**14. ábra. 2 rejtett rétegű 3, illetve 2 neuronnal rendelkező háló (saját ábra)**

Ezután rejtett rétegek és a neuronok számának több kombinációját is kipróbáltam és a legjobb modell, amely a legjobb eredményeket adta a két rejtett rétegű 3, illetve 2 neuronnal rendelkező úgynevezett mély tanuló háló lett (14. ábra).

Ez a mély tanuló háló már sokkal jobb eredményeket adott, így ez lett a neurális hálók közül a legjobban teljesítő modell. Az $R^2$ értéke ebben az esetben 83,61%, az átlagos négyzetes hiba négyzetgyöke pedig 0,415.

Következő módszerként az SVM-et alkalmaztam. Az SVM esetében is szükséges a sztenderdizálás, így a mesterséges neurális hálók számára már sztenderdizált tanuló és tesztelő adatbázist használtam. Alapértelmezett paraméter beállításokkal (radiális magfüggvény, $\gamma = \frac{1}{n}, \varepsilon = 0{,}1\ , c = 1$, ahol $n$ a megfigyelések száma) egy egész jó modellt kaptam, 80,95%-os $R^2$ – el, valamint 0,437 RMSE értékkel. Azonban jobb eredmények elérésének érdekében finomhangolás módszer futtatásával a következő paramétereket kaptam: radiális magfüggvény, $\gamma = 0{,}01, \varepsilon = 0{,}1,\ c = 16$ . Alkalmazva ezeket paramétereket jobb eredményeket kaptam, de még így is alulmaradt a neurális háló teljesítményétől. Szám szerint 81,91%-os $R^2$ és 0,425 átlagos négyzetes hiba négyzetgyöke értékekkel.

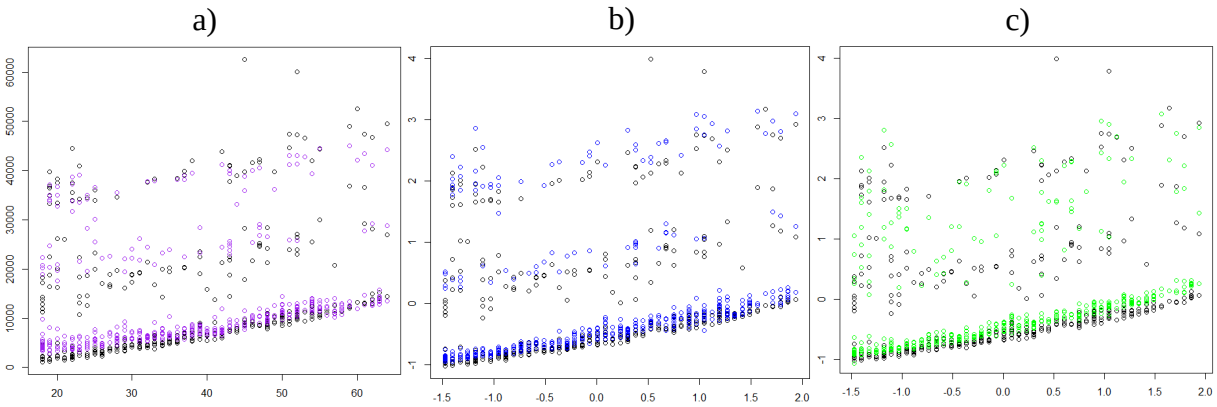
A véletlen erdő módszert a valódi adatokból készített tanuló és tesztelő adatbázison végeztem. A legjobb modell egy ezer fából álló erdő lett, amely 83,38%-os $R^2$ – et és 5031.5 átlagos négyzetes hiba négyzetgyöke értéket adott. Mivel a véletlen erdő tanításához nem szükséges sztenderdizálni az adatokat, így az átlagos négyzetes hiba négyzetgyöke teljesítmény mérő mutatót ebben az esetben nem tudjuk összehasonlítani a sztenderdizált adatokra tanított modellekkel, csak típuson belüli optimális modell kiválasztásához használtam.

Összehasonlításként a következő táblázat szolgál:

| | **Véletlen erdő** | **Neurális háló** | **SVM** |
|---|---|---|---|
| $\boldsymbol{R^2}$ | 83,38% | 83,61% | 81,91% |

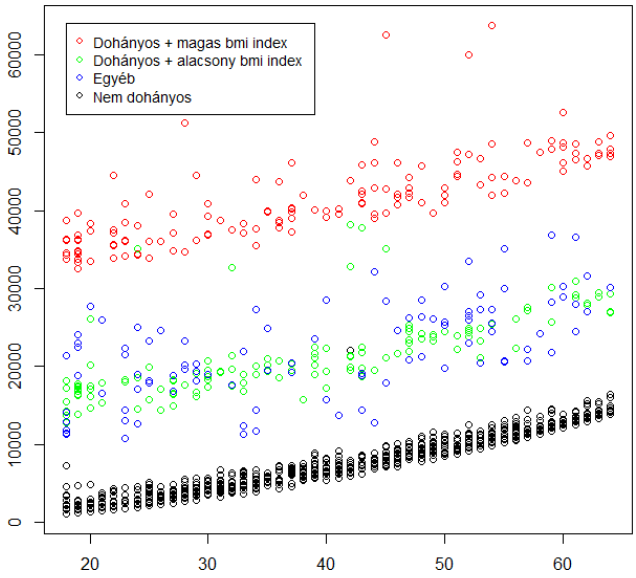
**2. táblázat. A gépi tanulási módszerek teljesítményének összehasonlítása (saját ábra)**

Jól látható, hogy mind három módszer 80% fölötti $R^2$-el rendelkezik, azonban a neurális háló valamivel pontosabb eredményeket ad, így ezt választottam a legjobb modellnek.



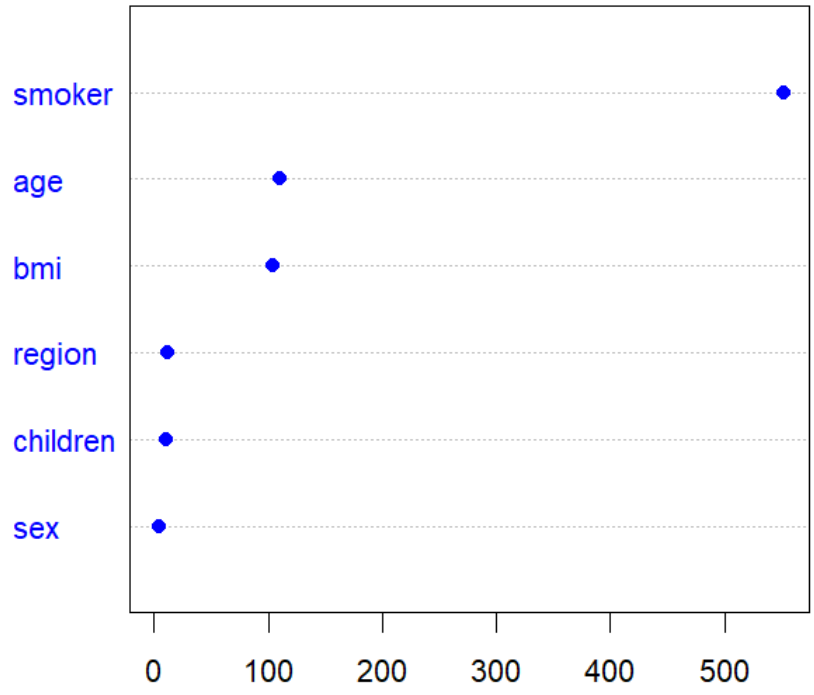

**15. ábra. Predikciók a kor és a kiadások függvényében. a) Véletlen erdő; b) Neurális háló; c) SVM esetén (saját ábra)**

Ha ábrázoljuk a predikciókat a kor függvényében, akkor láthatjuk, hogy az SVM (15. ábra. c)) értékei a leginkább szóródottak, a fekete pontok az eredeti értékek, míg a zöld színű pontok a becsült kiadások. Sok becsült érték esik a nagy és közepes kárkifizetések közé, ahol az eredeti adatbázisnak nincsenek értékei. Ezt a véletlen erdő ( a)) és a neurális háló ( b)) sokkal jobban kezelte, a véletlen erdő esetében még van néhány érték a nem kívánt tartományban, azonban a neurális háló becslései már nagyon jól illeszkednek az eredeti adatokhoz. A fenti ábrán (15. ábra) látható, hogy elkülönülő sávok szerepelnek a pontdiagramon. Ennek okát a következő ábrán (16. ábra) szemléltetem:




**16. ábra. Az egészségügyi kiadások megoszlása**

Ahogy az az ábrán is látható, a három különböző sáv jól elkülöníthető. A felső piros sáv azokból áll, akik dohányoznak és emellett I. fokú elhízással is szenvednek, ez 30-as testtömegindex értéknél magasabb esetben mondható el. A középső zöld sávban azok szerepelnek, akik dohányoznak, azonban testtömegindexük alacsonyabb 30-nál. Noha az optimális testtömegindex 25 alatt kezdődik, úgy gondoltam, hogy a 30-at választom vízválasztónak, mivel a 25-29,99 közötti indexet egy kevés túlsúllyal is el lehet érni, így az nem feltétlenül tükrözi az elhízás mértékét. A fekete pontok nagy része az alsó sávban helyezkedik el, ők azok, akik nem dohányoznak. Azonban vannak, akiket nem lehet jól beazonosítani a rendelkezésre álló információk alapján, ezek a kék pontok. Alapesetben ebbe a sávba még azok kárait várnánk, akik nem dohányoznak, de elhízottak. Ez azonban nincs, vagy csak részlegesen van így. Vannak köztük, olyanok, akik testtömegindexe 22, azonban akad olyan is, aki 40 fölötti értékkel rendelkezik, s ez így van az alsó sávban is. Ez lehet az egyik hátránya annak, hogy az adatok nem teljesen valósak. Valószínűsíthető, hogy egy valós adatbázison ezeket könnyebben el lehetne szeparálni.




**17. ábra. A változók fontossága a modellben (saját ábra)**

A 17. ábra bemutatja, hogy mely változók a legfontosabbak a modellben. Egyértelműen kijelenthető, hogy a várható éves egészségügyi kiadásokra legnagyobb hatással a dohányzás van, ez után következik a kor és a testtömegindex, majd őket követi elenyésző hatással a régió, a gyerekek száma, valamint a nem.

Összességében elmondható, hogy a gépi tanulási módszerek mind a három vizsgált típusa jól teljesít. Azonban az adatelemzők által közkedvelt mély tanuló neurális háló szignifikánsan jobb eredményt adott a két másik módszernél. Összehasonlításként futtattam az adatokra egy egyszerű lineáris regressziót is, ami 71%-os $R^2$-et adott, ami jóval elmarad a vizsgált három módszer teljesítményétől.

## Konklúzió

Munkámban ismertettem a Big Data elemzés felhasználásának módjait az egészségbiztosításban. Különböző kutatások születtek ezen módszer lehetőségeiről, amelyek bemutatják, hogy helyük van az egészségbiztosítási számításokban.

Az elemzéshez gépi tanulási módszereket használtam, mint Big Data elemző módszer. Részletesen bemutattam a három leggyakrabban használt gépi tanulási módszer elméleti hátterét regressziós feladatokra.

Célom az volt, hogy megpróbáljam minél pontosabban előrejelezni egy egészségbiztosítási ügyfél várható kárkifizetéseit egy évre, különböző paraméterek alapján. Az eredmények alapján elmondható, hogy a vizsgált prediktorok közül a dohányzás van a legrosszabb hatással az egészségügyi kiadások mértékére, ezután nem meglepően a kor, majd végső sorban a testtömegindex. Az előrejelzéshez három gépi tanulási módszert használt fel, melyek a neurális háló, a support vector machine (SVM), valamint a véletlen erdő. A modellek jóságát az $R^2$ és az RMSE (átlagos négyzetes hiba négyzetgyöke) mutatókkal mértem. Mivel a neurális háló és az SVM esetén az adatokat sztenderdizálnunk kell, így az átlagos négyzetes hiba négyzetgyöke mutató leginkább a módszerek tanításakor a típuson belüli legjobb modell kiválasztására szolgált.

A három módszer közül a legjobban egy mélytanuló két rétegű 3, illetve 2 neuronnal rendelkező neurális háló teljesített. Ez a tesztelő adatbázison 83,61%-os $R^2$ értéket adott, ami egy jó modellnek mondható.

Mivel az egészségbiztosítás árazása során is teljesülnie kell az ekvivalencia elvnek, mely szerint a bevételek várható értékének jelenértéke meg kell, hogy egyezzen a kiadások jelenértékének várható értékével. Ezért, egy olyan modellt, amely meg tudja becsülni egy egyén várható kárkifizetését egy adott évre, esetlegesen felhasználhatunk egészségbiztosítási termékek árazására is. A szakirodalomban elterjedt módszerrel ellentétben, ahol az egy éves szerződés esetén a várható követelések száma és a várható károk mértéke egy-egy random faktor, a díj pedig megegyezik ezek várhatóértékének szorzatával, egy ilyen modell az egyén paramétereitől függően becsüli meg a károk várható mértékét, ezáltal azon személyek esetében, akik egészséges életmódot folytatnak, nem dohányoznak, nem túlsúlyosak kisebb díjat lehetne szabni.

Sajnos nem sikerült olyan adatbázist találnom, amelyben több egészségügyi mutató is szerepel és valós Big Data-nak nevezhető. Egy olyan modell betanításával, amelyben több paraméter van és így hatékonyabban előrejelezhető a várható kárkifizetések mértéke esetlegesen létrehozható egy olyan szoftver, ami sokkal hatékonyabban árazza be a biztosításokat, megspórolva időt, energiát és pénzt.

## Irodalomjegyzék


Awad, M., Khanna, R., Awad, M., & Khanna, R. (2015). Support Vector Regression. In Efficient Learning Machines (pp. 67–80). Apress. Https://doi.org/10.1007/978-1-4302-5990-9_4

Az EIOPA felülvizsgálta a Big Data elemzés használatát a gépjármű- és az egészségbiztosítás területén. (n.d.). Retrieved May 14, 2020, from https://www.mnb.hu/felugyelet/felugyeleti-keretrendszer/felugyeleti-hirek/hirek-ujdonsagok/az-eiopa-felulvizsgalta-a-big-data-elemzes-hasznalatat-a-gepjarmu-es-az-egeszsegbiztositas-teruleten

Borgulya I.: Neurális hálók és fuzzy-rendszerek, Budapest-Pécs, Dialóg Campus Kiadó, 1998.

Choi, M. (n.d.). Medical Cost Personal Datasets | Kaggle. Retrieved May 6, 2020, from https://www.kaggle.com/mirichoi0218/insurance

Cover, T.M., Geometrical and Statistical properties of systems of linear inequalities with applications in pattern recognition, 1965

Dávid, V. (2019). EÖTVÖS LORÁND TUDOMÁNYEGYETEM TERMÉSZETTUDOMÁNYI KAR BUDAPESTI CORVINUS EGYETEM KÖZGAZDASÁGTUDOMÁNYI KAR.

Devopedia. (2019). Artificial Neural Network. Https://devopedia.org/artificial-neural-network

Dr. Vékás Péter. (2019). Támaszvektor-gépek (SVM).

EIOPA. (2019). Big Data Analytics in Motor and Health Insurance: A Thematic Review. In EIOPA Thematic Review. https://doi.org/10.2854/54208

EIOPA.(2019) Big data analytics in motor and health insurance Fact Sheet. https://www.eiopa.europa.eu/sites/default/files/publications/reports/eiopa_bigdataanalytics_factsheet_april2019_0.pdf

Gerber, G., Faou, Y. Le, Lopez, O., Trupin, M., Gerber, G., Faou, Y. Le, Lopez, O., & Trupin, M. (2018). The impact of churn on client value in health insurance , evaluation using a random forest under random censoring To cite this version : HAL Id : hal-01807623 The impact of churn on client value in health insurance , evaluation using a random forest under.

Gill K. M., W. K. A. & G. M. (1994). (2009). Insurance fraud: The business as a victim. Vol.

1., 73–82.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Https://doi.org/10.1007/978-1-4614-7138-7

Kirlidog, M., & Asuk, C. (2012). A Fraud Detection Approach with Data Mining in Health Insurance. In Procedia - Social and Behavioral Sciences (Vol. 62, pp. 989–994). Https://doi.org/10.1016/j.sbspro.2012.09.168

Lantz, B. (2013). Machine Learning With R. Https://doi.org/10.4018/978-1-7998-2718-4.ch015

Mallick, S. (2018). Support Vector Machines (SVM). Learn opencv. Https://www.learnopencv.com/support-vector-machines-svm/

Markand, O. N. (2003). Lennox-Gastaut Syndrome (Childhood Epileptic Encephalopathy). In Journal of Clinical Neurophysiology (Vol. 20, Issue 6, pp. 426–441). Https://doi.org/10.1097/00004691-200311000-00005

Nayak, B., Bhattacharyya, S. S., & Krishnamoorthy, B. (2019). Integrating wearable technology products and big data analytics in business strategy: A study of health insurance firms. Journal of Systems and Information Technology, 21(2), 255–275. Https://doi.org/10.1108/JSIT-08-2018-0109

Nebehaj V.: Pénzügyi és gazdasági idősorok előrejelzése neurális hálózatok segítségével, 2010

Sowah, R. A., Kuuboore, M., Ofoli, A., Kwofie, S., Asiedu, L., Koumadi, K. M., & Apeadu, K. O. (2019). Decision Support System (DSS) for Fraud Detection in Health Insurance Claims Using Genetic Support Vector Machines (gsvms). Journal of Engineering (United Kingdom), 2019. Https://doi.org/10.1155/2019/1432597

Testtömegindex . (n.d.). Retrieved May 10, 2020, from https://hu.wikipedia.org/wiki/Testtömegindex#cite_note-:1-3

Vapnik, Vladimir N. The Nature of Statistical Learning Theory, Second Edition. New York: Springer, 1999

Vekeman, F., Piña-Garza, J. E., Cheng, W. Y., Tuttle, E., Giguère-Duval, P., Oganisian, A., Damron, J., Sheng Duh, M., Shen, V., Saurer, T. B., Montouris, G. D., & Isojarvi, J. (2019).

Development of a classifier to identify patients with probable Lennox–Gastaut syndrome in health insurance claims databases via random forest methodology. Current Medical Research and Opinion, 35(8), 1415–1420. Https://doi.org/10.1080/03007995.2019.1595552

Widrow, B. And Hoff, M. (1960). Adaptive switching circuits, IRE WESCON Convention record, Vol. 4. Pp 96-104; Reprinted in An-dersen and Rosenfeld (1988).

Yeh, C., Hsu, A., & Chai, K. (2014). NEURAL NETWORK FORECASTS OF TAIWAN. 8(5), 95–114.