

DOKTORI ÉRTEKEZÉS

Ágoston Norbert

Pécs, 2024

PÉCSI TUDOMÁNYEGYETEM
KÖZGAZDASÁGTUDOMÁNYI KAR

REGIONÁLIS POLITIKA ÉS GAZDASÁGTAN
DOKTORI ISKOLA

Ágoston Norbert

A gépi tanuláson alapuló fizetéseképtelenséget
előrejelző modellek háttere és gyakorlati
alkalmazása magyar kis- és középvállalatok körében

DOKTORI ÉRTEKEZÉS

Témavezetők: Dr. Bedőházi Zita-Rozália

Dr. habil. Kruzslicz Ferenc

Pécs, 2024

Tartalomjegyzék

1. Bevezetés	1
2. A vállalati fizetéseképtelenség előrejelzésének kerete.....	10
2.1. A fizetéseképtelenség fogalma	10
2.2. A vállalati fizetéseképtelenség és csőd okai	14
2.3. A fizetéseképtelenség és csődbe jutás folyamata	15
2.4. A fizetéseképtelenség előrejelzésének szerepe.....	17
2.5. A fizetéseképtelenségi eljárások jogi háttere	19
3. A fizetéseképtelenség- és csődelőrejelzés korábbi empirikus kutatásai	21
3.1. Nemzetközi időrendi áttekintés	21
3.2 A magyar csődelőrejelzés mérföldkövei	28
3.3. A fizetéseképtelenség előrejelzésének alternatív változókra épülő modelljei.....	31
4. A fizetéseképtelenség előrejelzésének módszerei.....	35
4.1. Klasszikus módszerek	35
4.1.1. Diszkriminanciaanalízis	35
4.1.2. Logisztikus regresszió	36
4.2. Gépi tanuláson alapuló eljárások.....	39
4.2.1. Döntési fa módszercsalád.....	40
4.2.2. Support Vector Machine (SVM)	42
4.2.3. Neurális hálózatok.....	48
4.3.3. Bagging.....	51
4.3.4. Boosting	53
4.3.5. Random forest	56
5. Az előrejelző modellekhez kapcsolódó adatműveletek	59
5.1. A felhasznált adatok.....	59
5.2. Az adatgyűjtés szempontjai	60

5.3. A minta összeállítása	61
5.4. Magyarázó változók.....	66
5.5. Az adatok elemzésre való előkészítése	69
5.5.1. A kiugró értékek kezelése	71
5.5.2. Változószelekció.....	73
5.6. A módszerek paramétereit	77
5.7. Validáció.....	79
5.8. A modellek teljesítményének értékelése.....	81
6. A vállalati fizetéseképtelenség előrejelzésének empirikus vizsgálata	84
6.1. A fizetéseképtelenség előrejelzése kiegyensúlyozatlan adatmintán	84
6.2. Több év adatain épült modellek.....	92
6.3. A nem pénzügyi változók beépítése a modellekbe	99
6.4. A gépi tanuláson alapuló módszerek összemérése.....	103
7. Összegzés.....	111
7.1 A tanulmány korlátai és új kutatási irányok.....	114

Táblázatok jegyzéke

1. táblázat: A reprezentatív mintában szereplő megfigyelések száma.....	62
2. táblázat: Az azonos osztálylétszámú mintában szereplő megfigyelések száma	66
3. táblázat: A különböző mintákban szereplő vállalatok működési évének (X50) átlaga, középértéke és szórása	85
4. táblázat: A különböző mintákban szereplő X39* likviditási mutató átlaga, középértéke és szórása	86
5. táblázat: A különböző mintákban szereplő pénzeszközök aránya (X37) mutató átlaga, középértéke és szórása	87
6. táblázat: A különböző mintákban szereplő tartozások lefedettsége mutató (X28) átlaga, középértéke és szórása	88
7. táblázat: A különböző mintákban szereplő bérarányos eredmény mutató (X9) átlaga, középértéke és szórása	89
8. táblázat: A reprezentatív és azonos osztálylétszámú mintán készült modellek eredményei.....	90
9. táblázat: A reprezentatív- és az azonos mintavétellel készült modellek eredményeinek összehasonlítása a Wilcoxon próbával	91
10. táblázat: A több év adatain épül modellek és az egyéves adatokon épült modellek tanuló és tesztelő mintáinak felépítése.....	93
11. táblázat: A többéves és egyéves modellek osztályozási eredményei	94
12. táblázat: A többéves és egyéves modellek AUC értékei	95
13. táblázat: A többéves- és egyéves tanulómintával készült modellek eredményeinek összehasonlítása a Wilcoxon próbával	96
14. táblázat: A mezoszintű változókkal és kizárólag pénzügyi adatokkal épülő modellek tanuló- és tesztelő mintáinak felépítése	100
15. táblázat: A különböző években az egy vállalatra jutó árbevétel (Y1) és munkanélküliségi ráta (Y2) átlaga, középértéke és szórása.....	101
16. táblázat: A mezoszintű változókkal és kizárólag pénzügyi változókkal készült modellek AUC értékei	102
17. táblázat: A mezoszintű változókkal és pénzügyi mutatókkal képzett modellek eredményeinek összehasonlítása a Wilcoxon próbával	103
18. táblázat: A többéves előrejelzési horizontokon tesztelt modellek AUC értékei.....	104

19. táblázat: Az együttes és egyéb modellek teljesítményének összehasonlítása egyéves előrejelzési horizonton a Mann-Whitney teszt segítségével	105
20. táblázat: Az együttes és egyéb modellek teljesítményének összehasonlítása többéves előrejelzési horizonton a Mann-Whitney teszt segítségével	106
21. táblázat: Az egyes módszerek teljesítményének átlaga és középértéke	107
22. táblázat: A standardizálás teljesítményre gyakorolt hatásának összehasonlítása a Mann-Whitney teszt segítségével.....	108
23. táblázat: A standardizálás tanítási időre gyakorolt hatásának összehasonlítása a Mann-Whitney teszt segítségével.....	109
24. táblázat: A vizsgált hipotézisek.....	112

Ábrák jegyzéke

1. ábra A pénzügyi egészség változásának folyamata	15
2. ábra: A döntési fa általános alakja (a) és a rekurzív particionálás (b) két független változó két osztályú elválasztására	41
3. ábra: Az SVM kétdimenziós szeparációja	43
4. ábra: Egy egyszerű előrecsatolt neurális háló felépítése	48
5. ábra: A bagging eljárás algoritmus	52
6. ábra: A boosting eljárás algoritmus	54
7. ábra: A random forest eljárás algoritmus	57
8. ábra: Az eloszlási arányok alapján reprezentatív minta megoszlása osztály és gazdasági szektor szerint	63
9. ábra: A kiugró értékek meghatározása dobozdiagramm segítségével	72
10. ábra: A változószelekció lépései	77
11. ábra: A validációs mintafelosztás és a tesztminta.....	80
12. ábra: Klasszifikációs tábla és a hozzá kapcsolódó teljesítménymutatók	82
13. ábra: A többéves és egy év adatain képzett modellek AUC értékeinek alakulása ...	98

Függelékek jegyzéke

F1. függelék: Az azonos osztálylétszámú minta megoszlása osztály és gazdasági szektor szerint	144
F2. függelék: Az empirikus kutatáshoz kijelölt magyarázó változók	146
F3. függelék: A reprezentatív mintán épített modellek paraméterei	149
F4. függelék: Az azonos osztálylétszámú mintán épített modellek paraméterei	151
F5. függelék: A reprezentatív és azonos osztálylétszámú mintával készült modellek bemeneti változói	153
F6. függelék: A több év adatain épített modellek paraméterei	155
F7. függelék: A többéves és egy év adatain épített modellek osztályozási pontosságainak trendje a különböző tesztmintákon	157
F8. függelék: A mezoszintű adatokon épített modellek paraméterei	159
F9. függelék: A mezoszintű változókkal és kizárólag pénzügyi változókkal készült modellek osztályozási eredményei	161
F10. függelék: A többéves előrejelzési horizonton mért modellek osztályozási eredménye	162

Absztrakt

A doktori értekezés szerzője: Ágoston Norbert

Címe: A gépi tanuláson alapuló fizetésképtelenséget előrejelző modellek háttere és gyakorlati alkalmazása magyar kis- és középvállalatok körében

Témavezetők: Dr. Bedőházi Zita-Rozália

Dr. habil. Kruzslicz Ferenc, egyetemi docens

A vállalkozások kényszerű megszűnése, illetve csődje visszafordíthatatlan költségeket generál az érintettek körében. Ezért a gazdasági társaságok vezetőit, a pénzügyi intézményeket és biztosítókat folyamatosan foglalkoztatja a vállalati ügyfelek fizetőképessége. Emellett a csődkutatók is folyamatosan keresik a legjobban interpretálható módszereket és a legkiemelkedőbb teljesítményű eljárásokat és ezzel összhangban egy szüntelen vizsgálódás és tesztelés jellemzi a szakterületet. Mindez pedig módot kínál az újabb és újabb módszertani kísérletre és fejlesztésre, amely a jelen tanulmány egyik fő motivációjaként is szolgált.

A disszertáció egyik célja olyan fizetésképtelenséget előrejelző modellek kifejlesztése, amelyek a magyar korlátolt felelősségű kis- és középvállalatokon, univerzálisan és hatékonyan alkalmazhatóak. Ennek kivitelezése pedig olyan módszertani finomításokat tartalmaz, amelyeknek egy része a kutatások főáramában található, és a disszertáció során hipotézisekkel teszteltem. Az első ilyen a mintavételi kérdések közé sorolható azonos vagy kiegyensúlyozatlan mintavétel problémája. A második kérdésköre a disszertációnak a tanuló adathalmaz időbeli terjedelmére reflektál. Egy következő tétele a disszertációnak, a pénzügyi változóktól eltérő, alternatív magyarázó változók tárgyköre. Ez utóbbi egy kevésbé kutatott terület, a magyar szakirodalomban nem is igazán találunk említést róla. Meg kell említeni azt a tényt is, hogy a fokozott elemzési igény és a megnövekedett számítástechnikai kapacitások lendületet adtak a hagyományostól eltérő módszerek elterjedésének.

A doktori értekezés másik fő célja a vállalati fizetésképtelenség előrejelzésére alkalmas, a legkorszerűbb gépi tanuláson alapuló módszerek ismertetése és a magyar adatokra történő kalibrálása. A széles adatominta és a több párhuzamos modell illesztése robusztus eredményeket biztosított. Az első megállapításom szerint a magyar vállalati adatomintán az azonos osztálylétszámú modellfejlesztés javasolt. Továbbá a fizetésképtelenséget előre jelezni kívánt időszakhoz legközelebb eső egy év adatai elegendőnek bizonyultak a

hatékony teljesítmény eléréséhez. Figyelembe vettem a vállalati működés külső, mezoszintű változóit, amelyek közül lehetett találni olyant, amelyik magyarázta a fizetésképtelenséget, azonban az előrejelzési teljesítményt nem javította. Végezetül a legmagasabb teljesítményű gépi tanuláson alapuló modellt keresve az együttes módszercsalád eljárásai bizonyultak a leghatékonyabbnak.

Journal of Economic Literature (JEL) kódok: C45, C53, C58, F65, G33, L25

Tárgyszavak: *felszámolás, fizetésképtelenség előrejelzés, pénzügyi teljesítmény, gépi tanulás, metamódszerek, boosting, bagging, random forest*

1. Bevezetés

A vállalati nyereségeket hajtó pezsgő lét mellett létezik a vállalatok sötét oldala, a kudarcokkal és sikertelenségekkel jellemezhető felszámolások területe. A felszámolások nem csak a vállalatok megszűnését, hanem sok esetben személyes célok és küzdelmek felbomlását, bizalmi viszonyok megszűnését, társadalmi konfliktusok kialakulását eredményezik. Minden sikertelen vállalatnak sajátos története van. Azonban ezeknek vannak olyan közös jellemzőik, amelyek által azonosítani tudunk bizonyos mintázatokat.

A felszámolás a vállalati életciklus utolsó szakasza, amelyet számos tényező előidézhet. Az egyik legfőbb kiindulópontja a fizetéseképtelenség, amelyben a gazdasági tevékenység pénzügyi problémák miatt állt le, és gyakran a felszámolási eljárással végződik. Magát a fizetéseképtelenséget is számos mögöttes erő befolyásolhatja: környezet, iparág, méret, makrogazdasági körülmények, vállalatvezetés, verseny, életciklus, kultúra stb. A vállalatok fizetéseképtelensége jelentős költségeket generál az érintettek számára. Ezért a jelenség kialakulásának mielőbbi detektálása kiemelten fontos. Ennek feloldására tesznek kísérletet a csőd- és fizetéseképtelenségi előrejelzések és modellek.

A fizetéseképtelenségi kutatások hosszú múltra tekintenek vissza. Az alkalmazott módszerek kezdetben egyváltozós elemzések voltak, amit idővel a többváltozós módszerek helyettesítettek. Manapság a számítástechnikai kapacitások lehetővé teszik, hogy a fejlett mesterséges intelligencia és gépi tanuláson alapuló módszerek elterjedjenek, és innovatív módszertani fejlesztéseket eredményezzenek. Mivel az üzleti kultúra rendkívül sokszínű, a működési környezet állandóan változó, a probléma pedig nagymértékben összetett, ezért vég nélkül vannak feltáratlan területek és kihívások a csődelőjelzés kutatásában (Fejér-Király, 2015).

A tanulmányom tárgya a vállalati fizetéseképtelenség, és annak előrejelzése. A dolgozat során ismertetem a fizetéseképtelenséghez kapcsolódó fogalmakat, irányelveket és a keretrendszert, amelyben értelmezhetjük. A vállalati fizetéseképtelenség előrejelzése a pénzügyi kockázatkezelés egyik alapvető szempontja. Magába foglalja a vállalat pénzügyi helyzetének módszeres értékelését és a fizetéseképtelenség vagy csőd valószínűségének előrejelzését. A vállalatok rendszeres időközönkénti felülvizsgálata pedig alapvetően a pénzügyi kimutatások elemzéséből indul ki.

A fizetéseképtelenség- és csődelőjelzés szakirodalma egy nagyon gazdag kutatási terület, számtalan empirikus kutatással és megközelítéssel. A terület

tanulmányozása lehetővé tette az interdiszciplinaritást, amelyben a közgazdaságtan, a pénzügy és számvitel, valamint informatika együttműködésében számos különböző perspektívából közelíthetjük meg a problémát. A kutatások szüntelen lendületét magyarázza az is, hogy a csőd és fizetéseképtelenség sokrétű problémája miatt, az elméleti keretét adó empirikus kutatások nem álltak össze egy egységes egészzé, hanem egy folyamatos keresés, kísérletezés jellemzi a területet. A kutatások nemzetközi szinten megfigyelhetőek, azonban a gyakorlatban széles körben alkalmazott modellek a híres amerikai kutatások eredményeinek köszönhetőek. Azonban a neves külföldi modellek változatlan formában történő átvételével körültekintően kell eljárni. Az amerikai vagy más külföldi vállalatokra készült csödelőrejelző modellek sajátos vállalati, működési és környezeti tulajdonságai okán nem feltétlen működnek jól hazai vállalatok fizetőképességének azonosítására. Ezzel összhangban Kristóf és Virág (2019) kihangsúlyozta, hogy magyar vállalatokra nem jók a külföldi csödmmodellek, sokkal előnyösebb eredményeket érünk el, ha saját modellek építünk. Mindez pedig megalapozta a kutatásom egyik alapvető célját, mely szerint a magyar vállalatok széles körére érvényes, fizetéseképtelenséget előrejelző modelleket készítsék, amelyek a módszerek részletezése mellett az érdekelt feleket is hasznos információval látja el, a pénzügyi kockázatkezeléssel kapcsolatos döntéshozatali folyamatokhoz. Ez pedig lehetőséget teremt módszertani kérdések tesztelésére illetve vizsgálatára, amelyekből értékes következtetések vonhatók le, segítve a modellépítés döntéseinek meghozatalát.

Az egyik módszertani vita, amely figyelmet érdemel, a mintavételhez kapcsolható. A csödelőrejelzésben hagyományosan kialakult az azonos osztálylétszámú mintavétel, ahol a fizetéseképtelenek és egészségesek egyforma arányban szerepelnek a modellezési folyamatokban. Ezt egyrészt a kezdeti és széles körben elismert csödtanulmányok alapozták meg (Altman, 1968; Edmister, 1972; Gombola et al., 1987), másrészt az azonos mintavételnek technikai és kényelmi okai is vannak. A kiegyensúlyozatlan mintavétellel csupán néhány módszertani tanulmány foglalkozott, és azok sem mindig az valós eloszlások alapján határozták meg a mintát, hanem sok esetben egy köztes, önkényesen kiválasztott arány alapján (Berg, 2007; Oommen – Baise – Vogel, 2011; Alaka et al., 2018). Ezekhez kapcsolódik a tanulmány egyik kérdése, amelyben az azonos és valós eloszlásnak megfelelő mintavétel megbízhatóságát elemzem, amellyel a mintavételi döntés megalapozását kívánom megerősíteni.

Egy másik kérdésköre a disszertációnak a tanuló adathalmaz időbeli terjedelmére vonatkozik. Ha a modellek több évet átfogó adatok felhasználásával készülnek, akkor az

előrejelzési teljesítmény javulását várjuk el, mivel az információk szélesebb spektrumát ölelik fel. Arra találunk empirikus kutatásokat, hogy a megfigyelések többéves múltbeli adatainak a felhasználása javítja a modellek teljesítményét, szemben azzal, hogy csak egyetlen év adataira támaszkodunk, mivel figyelembe veszi a pénzügyi mutatók időbeli alakulását (Berg, 2007; Nyitrai, 2015; Fejér-Király – Ágoston – Varga, 2019). Azonban arra a kérdésre, hogy ez a teljesítmény javulás számottevő-e vagy csak egy-egy sajátos helyzetben érvényesül, nem kapunk rá egyértelmű utalást, ezért a probléma kibontása kitüntetett helyet kapott a disszertációban.

A csőd- és fizetéseképtelenséget előrejelző modellek egyik folyamatosan megújuló keresése a bemeneti magyarázó változók körül forog. A fizetéseképtelenségi tanulmányok szinte kizárólag a vállalati pénzügyi adatokra támaszkodnak. Az éves beszámolókból kalkulált pénzügyi arányszámok adják a gerincét az előrejelző modelleknek. Az alkalmazandó pénzügyi mutatókra azonban nincs minden adatállományra egységesen érvényes szabály, továbbá az is zavart okoz, hogy a kutatók tengernyi mutató kalkulálását végezhetik el. A tanulmányban ötven előre definiált pénzügyi változó közül kiemelem azt a néhányat, amely a magyar kkv-k körében megbízhatóan és több perióduson keresztül, magyarázza a fizetéseképtelenséget. Emellett felmerül a pénzügyi változóktól eltérő, alternatív magyarázó változók kérdésköre. A tágabb működési környezet változóinak integrálása a fizetéseképtelenséget előrejelző modellekbe nem túl gyakori, a magyar csőd kutatásban kifejezetten hiányzik. Így kísérletet teszek külső, mezoszintű magyarázó változók előrejelző modellekbe való integrálására.

A csődelőrejelző modellek építésére számos módszer és eljárás létezik a szakirodalomban. A gépi tanulás eszköztárából egy változatos palettát mutatok be és alkalmazok a modellezés során. Az egyik a tartóvektor gépek (support vector machines, SVM), amelyek az osztályok elválasztására az optimális hipersíkot keresik, és összetett problémák nagy pontosságú elválasztására képesek. A másik, nagy népszerűségnek örvendő módszer, a neurális hálózatok, amelyek összekapcsolt csomópontok komplex rétegeinek segítségével olyan kapcsolatokat és mintákat tudnak feltárni, amelyeket a hagyományos vagy lineáris módszerek kevésbé tudnak. Emellett az együttes módszereket (ensemble) vagy metamódszerként ismert eljárásokat alkalmaztam, nevezetesen a bagging, a boosting és a random forest algoritmusokat, amelyek az alaposztályozók kombinálásával kiemelkedő eredményeket érnek el. A felvonultatott módszerek teljesítményének egy megbízható összemérése a disszertáció végén kap helyet. A módszerek összehasonlító erősségeinek megvilágításával arra törekszem, hogy hasznos

információval lássam el az érdekelt feleket, amely végső soron hozzájárulhat, hogy megerősítsék stratégiai döntéshozatali folyamataikat egy állandóan változó pénzügyi környezetben.

A következőkben ismertetem a hipotéziseimet a magyar vállalatok fizetéseképtelenségének előrejelzésére, modellezésére és módszertanára vonatkozóan. A négy hipotézisem közül először a minta reprezentativitását veszem a fókuszba, és azt feltételezem, hogy a fizetéseképtelenség előrejelzésére kialakult azonos mintavétel nem önkényes és kényelmi szempontok miatt alakult ki, hanem a modellek végső teljesítményére van döntő hatással. Ez pedig nem mindig van explicit kimondva, a kutatók sokszor a szokásra hivatkozva végzik ezt a gyakorlatot.

H1: A reprezentatív mintával készült fizetéseképtelenséget előrejelző modell az azonos osztálylétszámú mintavétellel szemben kisebb hatékonyságú előrejelzésre képes a fizetéseképtelenek helyes besorolási arányát tekintve.

Másodsorban a többéves adatok előrejelzési erejét vizsgálom. A hipotézis megválaszolásával arra a kérdésre keresem a választ, hogy vajon a többéves pénzügyi adatok felhasználása javítja-e a modellek előrejelző teljesítményét az egyetlen pénzügyi év adataiból épült modellekhez képest? A többéves megközelítés a pénzügyi változékonyság és tendenciák szélesebb spektrumát ragadja meg, ezáltal árnyaltabb és pontosabb képet adhat a vállalatok pénzügyi állapotáról. Ezzel szemben az egyéves megközelítés egyszerű megoldást kínál és a legfrissebb, legrelevánsabb információkat foglalja magába.

H2: A több év adataiból épített, fizetéseképtelenséget előrejelző modellek átlagosan magasabb előrejelző teljesítményűek egyéves előrejelzési horizonton, mint az egyetlen üzleti év adatain épült modellek.

A harmadik hipotézissel a vállalat belső pénzügyi adataitól eltérő, külső, mezoszintű inputok hatását vizsgálom a fizetéseképtelenséget előrejelző modellek teljesítményére. A hagyományos megközelítés szerint az előrejelző modellek a vállalati pénzügyi adatokra támaszkodnak és csak néhány esetben találunk makrogazdasági, iparági és vállalatspecifikus változóknak a felszámolásra gyakorolt hatáselemzését a szakirodalomban. Ezekből az egyik felmerülő kérdés: vajon a külső inputok hatást gyakorolnak-e a fizetéseképtelenségre, és ha igen, akkor az előrejelzési előny

statisztikailag kimutatható? Azt várnák, hogy a makrogazdasági, piaci és egyéb változóban olyan információk rejtőznek, amelyek nem tükröződnek a vállalati pénzügyi adatokban. A szakirodalmi megfigyelésekben azonban az alternatív típusú változók, a pénzügyi mutatókkal együtt gyakorolt hatása mérsékelt, ezért az előrejelző modellekben szinte észrevehetetlenek.

H3: A mezoszintű bemeneti adatok közül lesznek olyanok, amelyek szignifikáns hatással vannak a fizetésképtelenségre, azonban a hatásuk a pénzügyi mutatókkal szemben olyan kismértékű, hogy a mezoszintű változókat tartalmazó előrejelző modellek hasonló arányban képesek besorolni az egészséges és fizetésképtelen vállalatokat, mint a kizárólag vállalati pénzügyi adatokon alapuló fizetésképtelenséget előrejelző modellek.

Végül a fizetésképtelenség előrejelzés örökös törekvésével összhangban - a leghatékonyabb módszer után kutatva - fogalmazom meg az utolsó hipotézisem. Ebben a különböző módszerek teljesítményét mérem össze, és keresem a szakirodalomban gyakran felmerülő módszerek közül azt, amely a magyar vállalati adatokon a legjobban működik. A gépi tanulás tárházából az együttes módszerek ígéretesnek tűnnek, mivel egyesítik több tanuló eljárás képességét és ezáltal az összetett és nemlineáris problémákat könnyebben teljesítik. Feltételezem, hogy ez az előny állandó marad a logisztikus regresszió, az SVM és a neurális hálóval szemben, több adatkészleten és különböző előrejelzési horizonton.

H4: A gépi tanuláson alapuló együttes (ensemble) módszerek előrejelzési pontossága szignifikánsan magasabb, mint az egyéb (logisztikus regresszió, SVM, neurális háló) előrejelzési módszereké egy- és többéves előrejelzési horizonton.

A hipotézisek tesztelésén keresztül a doktori értekezés több ponton hozzájárul a szakirodalmi vitához és újszerű megállapításokat tesz, amelyek hozzájárulnak a tudományterület előremozdulásához.

1. A magyar vállalati adatmintán, a fizetésképtelen és egészséges eloszlás alapján a reprezentatív mintavétel nem célravezető egy átlagosan jól teljesítő fizetésképtelenséget előrejelző modell kifejlesztéséhez, sokkal inkább ajánlott az azonos osztálylétszám alkalmazása. A reprezentatív minta azt eredményezi, hogy a nagyobb számban lévő egészséges vállalatokra idomulnak a tanuló algoritmusok, a statisztikai alapokon nyugvó módszerek és a mesterséges intelligencia alapú eljárások egyaránt. Ebből adódóan a

modellek átlagos osztályozása kiváló lesz, azonban a fizetésképtelen megfigyeléseket képtelen helyesen osztályozni. Ezzel pedig az egyik leglényegesebb cél vesz el, a problémás vállalatok azonosítása. A szűkös adatállományoknál, vagy a kis mintából dolgozó elemzéseknél az azonos osztálylétszámon kívül nincs is egyéb opció. Bár a reprezentativitásra törekvő mintavétel megbízható képed ad a populációról, a fizetésképtelenség pontos előrejelzésében nem kívánatos. Az eredmény és a hozzá kapcsolódó módszertani ajánlás hozzájárul a szakirodalomban tapasztalható mintavételi vita előmozdításához.

2. A több pénzügyi év adatain készült modellek előrejelző teljesítménye átlagosan nem bizonyult magasabb teljesítményűnek a mindössze egy év adataival készült modellhez mérten. Ezek a megállapítások az előzetes feltevással ellentétesek, mely szerint a több év adatainak beépítése szükségszerűen jobb előrejelzési pontosságot eredményez. Ehelyett a modellek átlagos előrejelző ereje nem növekszik további múltbeli adatok bevonásával, hanem mintegy megegyezik az előrejelzést közvetlen megelőző egy év adatain mért eredményekkel. Ez pedig döntő fontosságú a csődelőrejelző modellfejlesztési stratégiák kialakításának szempontjából és arra ösztönözi a téma művelőit, hogy a mennyiség helyett a minőségre fektessék a hangsúlyt. Érdeemes tehát a legfrissebb és legrelevánsabb adatokra támaszkodni és azokra építeni a modelleket, ami hatékonyabb modellépítést és időszerű eredményeket biztosít. Az is körvonalazódik, hogy a halmozódó, egyre bővülő adatállomány nem rendelkezik hozzáadott értékkel az előrejelzés tekintetében. Ennek a kérdésnek a bizonyítása további elemzést igényel és egy későbbi tanulmányban kerülhet kifejtésre.

3. A mezoszintű külső tényezők segítségével egy tágabb gazdasági összefüggésben értelmezhetjük a fizetésképtelenség kialakulását. A településszintű munkanélküliségi ráta szignifikáns pozitív hatással van a fizetésképtelenségre. Azonban átlagosan a pénzügyi mutatók és más mezoszintű változók mellett az előrejelzési teljesítményt nem fokozza. A munkanélküliségi ráta, mint a helyi üzleti környezet fokmérő mutatója minden adatállományon szignifikáns magyarázó változója a fizetésképtelenségnek. Emellett a többi beválogatott mezo-változó, a helyi árbevétel, a fizikai bankfiókok száma és a település jogállása nem bizonyult szignifikánsnak, és együttesen az előrejelző teljesítményt sem növelték. Sőt a modellek átlagos teljes besorolási aránya és az AUC-érték alapján a mezo-változókkal készült modellek alacsonyabb teljesítményűnek bizonyultak, mintha a pénzügyi mutatókat csak önmagukban használtam volna. Az eredmények tükrében a vállalatok

fizetéseképtelenségének előrejelzésére inkább a vállalat belső pénzügyi változói ajánlhatóak. Természetesen fontolóra vehetők az olyan nem pénzügyi mutatók is, mint az iparági tényezők, a menedzsment minősége, a piaci hangulat és trendek mutatói, amelyek hatással lehetnek a vállalatok fizetéseképtelenségére, de ugyanakkor a pénzügyi mutatókban nem tükröződnek, így általuk akár növelhető lehet az előrejelző pontosság.

4. Az alkalmazott módszerek közül átlagosan a legkiválóbb teljesítményt az együttes (ensemble) vagy metamódszereként ismert eljárások érték el, egy- és többéves előrejelzési horizonton egyaránt. Ezek közül nem lehet egyértelmű győztes módszert megállapítani, azonban ki lehet emelni a random forest-et, amely több szempontot figyelembe véve átlagosan egy árnyalattal hatékonyabb teljesítményt ért el, mint a módszercsaládba tartozó Adaboost és bagging. A logisztikus regresszió, mint a fizetéseképtelenség klasszikus módszere meglehetősen jól teljesített. A gyors tanulási képesség és könnyű interpretálhatósági előny mellett viszonyítási alapként kifejezetten ajánlott. A módszer eredmények mellé kiegészítendő, hogy a bemeneti változók skálázására érzékeny módszereknél, mint amilyen a lineáris SVM, javasolt a változók standardizálása, amellyel nem csak a tanulási ciklus gyorsítható fel, hanem a módszer teljesítménye is szignifikánsan javul.

5. A fizetéseképtelenség azonosítására voltak olyan mutatók, amelyek valamennyi vagy csaknem mindegyik modellben konzisztensen szerepeltek, mint magyarázó változók. Az adott minta jellemzői mellett a következőket tekinthetjük a legjelentősebb változóknak a fizetéseképtelenség magyarázatára, mivel a sokrétű változószelekciós eljárások gyakran fontos változónak jelölték meg. Ezek közül az első számú a vállalatok működési éveinek a száma (X50), amely az összes modellben szerepelt magyarázó változóként. A mutatónál a kisebb érték jelent fizetéseképtelenségi kockázatot. A megfigyelések között a fizetéseképtelen osztály középértékei 5-6 év volt, míg a fizetőképeseké 8-10 év. A számos adatbázison futtatott logisztikus regresszió esélyhányadosai alapján a működési évek egységnyi növekedése 4%-7%-al csökkenti a fizetéseképtelenség esélyét ceteris paribus. A pénzügyi mutatók közül a leggyakrabban kiválasztott volt a készletekkel csökkentett forgóeszközök és összes eszköz aránya (X39), a pénzeszközök aránya a forgóeszközökön belül (X37) és a saját tőke összes kötelezettség aránya (X28). A készletekkel csökkentett forgóeszközök és összes eszköz aránya (X39) gyakorlatilag a pénzeszközök, követelések és értékpapírok együttes arányát mutatja az eszközök összértékén belül. A mutató egészséges középértéke 45-50% között helyezkedik el, míg a fizetéseképteleneké 56-72% között, tehát ezek fokozott értéke növeli a

fizetéképtelenség esélyét. Arra a kérdésre, hogy a likvid eszközök/összes eszköz mutatónak a számlalójában milyen arányok érvényesülnek az egészséges és fizetéképtelen osztályban, két megállapítást tehetünk. A rövid lejáratú likvid értékpapírok birtoklása a mintában szereplő vállalatok nagy kis hányadára jellemző (4%-5%), szinte valamennyi vállalatnál ez nulla értékű. A pénzeszközök súlyára pedig az X37-el jelölt, szintén fontosnak értékelt magyarázó változó ad iránymutatást. Ez pedig rávilágít, hogy a pénzeszközök aránya a forgóeszközökön belül a fizetéképteleneknél számottevően alacsonyabb, mint az egészségeseknél. Ebből következően pedig visszautalhatunk az X39 változóra is, hogy a magasabb arányú likvid eszközök mögött a követelések nagy aránya szerepel, így követelésmenedzsment problémákra utal a fenti két változó, amely miatt a fizetéképtelenség kialakul. Arra a kérdésre viszont, hogy ez a hitelpolitikai stratégiák elégtelenségéből, a követelés kezelés problémáiból, ügyfélkommunikációs hiányosságokból, körbetartozásokból vagy netán kevésbé diverzifikált ügyfélkörből és gyenge vagy teljesen hiányzó partnerkockázat értékelésből adódik, egy következő, sokkal fókuszáltabb kutatás adhat választ. Végül a saját tőke/idegen tőke arány (X28) bizonyos modellekben fellelhető volt, amelynek a 100% feletti értéke jelölte az egészséges osztályhoz való tartozást. A fizetéképtelenekre jellemző volt a mutató negatív értéke, amely egyben a sorozatos üzleti veszteségek jelzőszáma is.

Köszönetnyilvánítás

A doktori disszertáció elkészítése egy hosszú út volt és sokak fele nagyon hálás vagyok, akik támogattak és irányítottak a folyamat során. Mindenekelőtt szeretném kifejezni őszinte hálámat témavezetőimnek *Dr. Bedőházi Zitának* és *Dr. Kruzslíc Ferencnek* a folyamatos támogatásért, motivációért és mentorálásért. Az ők magas szintű szakmai felkészültsége és útmutása segített a kutatás megírásában, kifejezetten a kritikus pontokon való átlendülések idején, amely mellett személyes téren is folyamatos támogatást kaptam.

Kiemelt köszönettel tartozom a Pécsi Tudományegyetem Közgazdaságtudományi Karán működő Regionális Politika és Gazdaságtan Doktori Iskola valamennyi oktatójának. Köszönetemet fejezem ki a doktori iskola vezetőjének *Dr. Szerb László professzor úrnak*, aki a felívelést biztosító, következetes vitafórumok során mindig szakmailag támogató lelkesítést nyújtott. A rendelkezésemre bocsátott vállalati adatbázisa és kutatásainak eredményei nagymértékben előmozdították a disszertációm

elkészülését. Sokat köszönhetek a 2023-ban elhunyt *Dr. Varga Attilának*, aki a PhD képzésem első éveiben mentorként kapukat nyitott meg előttem a kutatás és szakmai fejlődés irányába. Köszönettel tartozom *Dr. Rideg Andrásnak*, aki a kutatáshoz szükséges adatigényemet mindig aktívan elősegítette.

Köszönöm a Sapientia Erdélyi Magyar Tudományegyetem Csíkszeredai Karának, hogy tanársegédként alkalmaztak és ösztönöztek a doktori fokozat megszerzésére. Köszönöm a Kar dékánjának, *Dr. Lázár Edének*, az Üzleti Tudományok Tanszék vezetőjének, *Dr. Nagy Benedeknek* és az oktató kollégáknak a folyamatos támogató kollegiális kapcsolatot, amellyel biztosították a szakmai fejlődésem és előmozdították a doktori fokozatszerzésem. Kiemelt köszönettel tartozom *Fejér-Király Gergely* kollégámnak, aki a legkorábban bátorított a tudományos pálya felé, ösztönzött a továbbhaladásra és végigkísért az utamon. Köszönöm a *LAM Alapítványnak*, hogy támogatóan és odafigyeléssel értékelték a kutatói előmenetelem, és anyagi támaszt nyújtottak, hogy teljes elmélyüléssel a munkámra koncentrálhassak.

Hálával tartozom a *doktorandusz kollégáknak, kutatótársaknak*, hogy szakmailag inspiráló és ugyanakkor baráti közösségben tölthettem a doktori képzés időszakát. Köszönetemet fejezem ki *hallgatóim* felé, akiket doktoranduszként és tanársegédként tanítottam, hogy megtapasztaltam a tanítás örömeit és kihívásait, amely által szakmailag jobbá váltam. Köszönöm a *barátaimnak*, hogy számíhattam rájuk és állandó támogatást nyújtottak a számukra sokszor végeleáthatatlan utamon. Mély köszönettel tartozom a családomnak. Őszintén hálás vagyok a *szüleimnek* és a *testvéremnek* minazért az áldozatért, amit értem hoztak. Végül, de nem utolsó sorban, szeretnék köszönetet mondani *házastársamnak* a végtelen támogatásáért és megértéséért. A szeretete, türelme és belém vetett rendületlen hite lehetővé tette, hogy megírjam a doktori disszertációm.

2. A vállalati fizetéseképtelenség előrejelzésének kerete

A bevezető után az első fejezet a vállalati fizetéseképtelenség előrejelzéséhez szorosan kapcsolódó összefüggéseket ismerteti, annak érdekében, hogy megteremtse egy keretet, amelyben a csődelőrejelzés értelmezhető. Ez ezért is fontos, mert a csődelőrejelzés tekintetében szűkös az elméleti megalapozás és az is szerteágazó forrásból származik. A tudományterület inkább folyamatos empirikus kutatásokon és kísérletezésen alapszik (Lim et al., 2012). Ezért a keretrendszer magába foglalja az értékelési szempontokat, korábbi kvantitatív modellek tanulságait és modellfejlesztéshez kapcsolódó stratégiai meglátásokat. A fejezetben a csőd és fizetéseképtelenség fogalmainak ismertetése után a kialakulás okait és a csődbejutás folyamatát ismerhetjük meg, amelyet a fizetéseképtelenség előrejelzés céljainak szempontjaival folytatunk. A fejezet végén a fizetéseképtelenséghez kapcsolható jogi eljárások részleteit tekintjük át.

2.1. A fizetéseképtelenség fogalma

A fizetéseképtelenség, a csőd, a gazdasági kudarc, a pénzügyi nehézség és a fizetés elmulasztása gyakran használt kifejezések a sikertelen vállalatok leírására. Ezek jelentése gyakran összemosódik, amelyet a szakirodalom sem mindig tisztáz megnyugtatóan. Constand és Yazdipour (2011) állítása szerint a csőd fogalmát illetően nincs egy egyetemesen elfogadott meghatározás. Bruno és Leidecker (1988) kissé túlzó véleménye alapján nem létezik olyan két szakértő, aki az üzleti csőd fogalmában egyetértene. Sharma és Mahajan (1980) úgy véli a vállalati kudarc fogalmának megállapítása az egyik legnehezebb feladat a csődkutatók számára.

Az üzleti kudarc fogalma Greenwald (1973) gazdasági szótára szerint, az üzleti tevékenység felhagyása önkéntesen vagy bírósági eljárások következtében, amely veszteséget okoz a hitelezőknek. Az üzleti tevékenység leállításának számos különböző oka lehet, például tőkevesztés, elégtelen nyereség vagy visszavonulás. Azonban ha a hitelezői követelések maradéktalanul teljesítve lettek, akkor a megszűnő vállalat nem tekinthető kudarcosnak (The Business Failure Record, 1978 idézi Sharma – Mahajan, 1980). Más megfogalmazás szerint az üzleti kudarc a jogi eljárások előtt már bekövetkezik, amikor a vállalati tartalékok kimerülnek, és az adott cég képtelen eleget tenni a soron következő fizetési kötelezettségeinek (Van Horne, 2002). A pénzügyi nehézség megfogalmazására a szakirodalom a csőd és fizetéseképtelenséget szinonimaként használja (Wruck, 1990), annak ellenére, hogy a jelentésük különböző.

Beaver (1966) sikertelennek jellemezte azt a vállalatot, amely ellen csődeljárás indult, negatív bankszámlaegyenlege volt, késedelmek jelentkeztek a kötvények törlesztésében vagy az osztalékfizetés elmaradt. Pénzügyi nehézségről beszélhetünk, amikor a vállalat nem képes teljesíteni kötelezettségeit. A jelenség első jelei akkor mutatkoznak, amikor a fizetési kötelezettségeket megsértik, ami az osztalékfizetés csökkenésével vagy elmaradásával párosul (Baldwin – Mason, 1983). Whitaker (1999) szerint pedig a pénzügyi nehézség akkor jelentkezik, amikor a nettó pénzáramlás kisebb, mint az esedékessé vált hosszú lejáratú adósság.

A szakirodalom a bajba jutott vállalatok állapotának leírására négy kifejezést használ, amelyek a következők (Altman – Hotchkiss – Wang, 2019):

- **Kudarc (failure):** Gazdasági értelemben a kudarc azt jelenti, hogy a vállalat megvalósult megtérülési rátája jelentősen alacsonyabb, mint a hasonló kockázatú befektetésektől elvárt hozam. Más esetben kudarcról beszélhetünk, ha a vállalat a költségeket tartósan nem tudja a bevételeivel fedezni vagy a tőkeköltségtől elmaradó alacsony átlagos tőkehozamot realizál. Hozzá kell tenni, hogy a kudarcos vállalat jellemzően nem áll le, hanem folytatni tudja a működést.
- **Fizetéseképtelenség (insolvency):** A vállalat alacsony teljesítményét mutatja és jellemzően technikai jellegű, olyan értelemben, hogy a vállalat nem tudja teljesíteni esedékessé vált kötelezettségeit. Jelenthet csupán átmeneti cash flow vagy likviditáshiányt. Ezzel szemben a számviteli értelemben vett fizetéseképtelenség (balance sheet insolvency) súlyosabb probléma, amelyben az összes kötelezettség meghaladja az összes eszköz valós értékét, azaz a vállalat nettó vagyona negatívvá válik. Wruck (1990) megkülönbözteti az állományi (stock) fizetéseképtelenséget, amely azt takarja, hogy a vállalati pénzáramlások jelenértéke alacsonyabb, mint az összes kötelezettség. Az áramlás (flow) típusú fizetéseképtelenség esetén a vállalat pénzbeáramlásaiból nem tud eleget tenni a folyó fizetési kötelezettségeinek. Ha egy vállalat flow szempontból fizetőképességű, de fizetéseképtelen állományi tekintetben, akkor a hitelezők nem tudnak eljárást kezdeményezni, mivel a követeléseik teljesülnek (Wruck, 1990).
- **Nemteljesítés (default):** Azt jelenti, hogy a hitelfelvevő megszegi a hitelezővel kötött szerződéses megállapodását. Ennek a legjellemzőbb esete a tőke és a kamatok megfizetésének az elmulasztása, amely a tartozás átszervezéséhez vagy formális csődeljáráshoz vezethet. Technikai nemteljesítésről beszélhetünk, ha a fizetéstől eltérő egyezményt sért meg az adós fél, például egy előírt minimális

likviditási szint be nem tartása, vagy egy maximálisan meghatározott adósságráta túllépése. Ennek gyakori előfordulása rendszerint az adós vállalat csökkenő teljesítményének mutatója, amely a szerződési feltételek újratárgyalását vonja maga után, ritkább esetben a tartozás visszafizetésének azonnali követelésével zárul. A Bázeli Tőkeegyezmény a default fogalmát a 90 napon túli nemteljesítő hitelekkel ragadja meg (Basel Committee on Banking Supervision, 2017).

- Csőd (bankruptcy): A formális jogi eljárásra utal, amelyben a vállalat fizetéseképtelensége miatt felszámolási eljárás, csődeljárás indul. Ugyanakkor a csőd fogalma nem azonos a felszámolással, mivel a csőd keretében egy vállalati reorganizáció után a vállalat túlélővé válhat. Azt is hozzá kell tenni, hogy a csőd jogi jelentése országonként eltérő lehet. Bizonyos országokban a csődeljárás magában hordozza a reorganizációt, máshol pedig a megszűnés előtti végső fázist jelenti a vállalat számára, ahonnan nincs visszatérés.

Brealey és Myers (2005) alapján a pénzügyi nehézség (financial distress) mindig költségekkel jár együtt, és akkor lép fel, ha a hitelfelvevő vállalat a hitelezőknek tett ígéretét nem tartja be, vagy nehezen tudja teljesíteni. Ezért az ilyen vállalat veszélyes helyzetbe kerül, ahonnan nem mindig van visszaút, számos esetben a végkifejlet a csőd és a megszűnés. A minősített könyvvizsgálók az egyes vállalatokról szóló beszámolóikban a vállalat tevékenységének folytatásával (going concern) kapcsolatos véleményt is adhatnak, amelyben esetenként részletezik a pénzügyi nehézségek okait. Azonban nem szabad figyelmen kívül hagynunk, hogy e vélemények megalapozásához csődelőrejelző modelleket használhatnak (Wahlen et al., 2018). Barboza és Altman (2024) a pénzügyi nehézséget új, sajátos módon határozta meg. Az egyik módszerük szerint csődös az a vállalat, amelynek két egymást követő üzleti évben a kamatok, adózás és értékcsökkenési leírás előtti eredménye (EBITDA) kisebb, mint a pénzügyi ráfordítások összege. A másik megközelítésük szerint pénzügyileg problémás a vállalat, amelynek a saját tőke piaci értéke két éves időtávon folyamatos csökkenést mutat.

Egy saját empirikus vizsgálat keretében, a szakirodalomban alkalmazott meghatározások alapján, a csődöt négy fogalom közé soroltam (Ágoston, 2022a):

- Jogi megközelítés szerinti csőd: hivatalos jogi eljárások útján lejátszódó csődeljárás, felszámolási eljárás, fizetéseképtelenségi eljárás, csődeljárás során történő felvásárlás vagy összeolvadás.

- Technikai csőd: ebbe a kategóriába tartozik a hitel nemteljesítés, a hitel újrastrukturálás, a kötelezettségek elmulasztása, a tőkeemelés az elegendő működő tőke biztosításához, valamint az osztalék elmaradása.
- Gazdasági megközelítés szerinti csőd: ebben a kategóriában a tanulmányok bizonyos mutatók alapján határozták meg a csődöt, mint negatív eredmény, negatív működési pénzáramlás, elégtelen eszközarányos megtérülés (ROA) és elégtelen sajáttőke arányos megtérülés (ROE). Például Asquith et al. (1994) a kamatfedezettségi mutatót alkalmazta a pénzügyi nehézség azonosítására. Az a vállalat amelynek kamatok, adózás és értékcsökkenési leírás előtti eredménye (EBITDA) kevesebb, mint 80%-a a kamatráfordításoknak, problémásnak minősül.
- Jelentés alapú megközelítés: bizonyos hivatalos közlemények alapján azonosítható a csőd úgy, mint tőzsdei beszámolók, könyvvizsgálói jelentések, pénzügyi jelentések, vagy éppen az éves pénzügyi jelentés elmaradása esetén. Ebbe a kategóriába sorolható a kötvény- és részvényértékelő (rating) szolgáltatók alapján meghatározott üzleti kudarc és nehézség is.

A tanulmányban a csődeljárást a magyar jogi terminológia szerint értelmezem, amelyben a vállalat fizetési haladék mellett lehetőséget kap tartozásainak rendezésére. Figyelembe véve, hogy a csőd, a fizetésképtelenség és a pénzügyi nehézség vizsgálata rokon kutatási területek, azért a következőkben lényegesen nem különítem el a megfogalmazásokban, hanem rokon jelentésűnek tekintem. Így a csőd, fizetésképtelenség és üzleti kudarc kifejezéseket a sikertelen vállalatok jellemzésére használom, amelyek fizetésképtelenné váltak. Ezek kimenetelére nem teszek utalást, így lehetnek véglegesen megszűnők vagy akár tovább működők is. Az empirikus részben a fizetésképtelenséget kifejező függő változót a felszámolási eljárás esetéből származtatom. Ezzel ellentétben az egészséges vagy fizetőképes kifejezéssel fogok élni azon vállalatokra vonatkozóan, amelyek nem tartoznak a csődös csoporthoz. A fizetési hajlandóságot maximálisnak feltételezem, így aki fizetőképes az hajlandó is fizetni és a fizetésképtelenség nem a fizetési hajlandóság hiányából ered.

2.2. A vállalati fizetéseképtelenség és csőd okai

A vállalati fizetéseképtelenség és csőd kialakulásának okai sokrétűek lehetnek. A szervezetelméleti megközelítés szerint többek közt közrejátszanak a kedvezőtlen vezetői tulajdonságok, a megalapozatlan döntések, a válságokra adott esetlen válaszok, az információs aszimmetria, a képesség- és erőforráshiány, a törvényi és szabályozási előírásoknak nem megfelelő működés, a környezeti evolúciós változások okozta hátrányos szelekció, a szervezeti rutinok merevsége, a szervezeti tőke helytelen kezelése, a marketing és pénzügyi funkciók rossz működése és az elégtelen szervezeti beágyazottság (Kristóf, 2005).

A fizetéseképtelenség szempontjából célszerű a pénzügyi elégtelenségek és gazdasági nehézségek kombinációját keresni. Altman et al. (2019) szerint az általános gazdasági okok a következők:

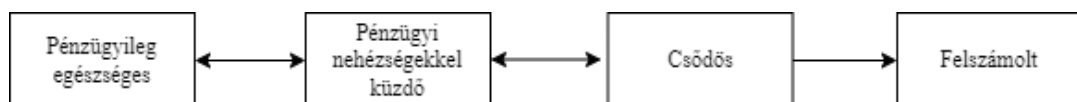
- Gyenge működési teljesítmény és magas tőkeáttétel: A gyenge teljesítmény fakadhat a nemzetközi versenyből, rosszul végrehajtott beszerzésekből, indokolatlanul magas kapacitásokból, a verseny kiéleződéséből, ársokkokból vagy ciklikus iparágakból, amelyeket a magas tőkeáttétel csak súlyosbít.
- Új vállalkozások alapításának magas aránya: Az új vállalkozások sokkal gyakrabban buknak meg, mint a tapasztalattal rendelkező vállalatok. Ha egy bizonyos időszakban magas arányban jelentek meg új vállalkozások, akkor a következő években várható, hogy a csőd aránya is megugrik.
- Technológiai innováció hiánya: Az új technológiai megoldások megjelenése veszélyt jelent azon vállalatok túlélésére, amelyek egy korábbi, kevésbé versenyképes technológiát birtokoltak.
- Finanszírozási sokk: A hitelezési csatornákhöz való hozzáférés csökkenése likviditás problémákat vet fel, amely gátja az esedékes tartozások teljesítésének. Ilyent láthattunk a 2008-2009-es válság idején, amikor a hitelpiacok kínálata meredeken csökkent és a refinanszírozási kockázat megemelkedett.
- Iparágak deregulációja: A dereguláció az iparágak versenyét fokozza, és elősegíti, hogy nagy számban legyenek belépő és kilépő vállalatok.
- Váratlan kötelezettségek: A csőd egy másik okozója lehet, ha a mérlegen kívüli tételek hirtelen mérleg szerinti kötelezettségé válnak, például peres ügyek miatt. Egy másik példa a felvásárlás, amely által megörökölhető egy nem várt kötelezettség.

Lizal (2002) a pénzügyi nehézségek három okát jelölte meg. Az első a neoklasszikus megközelítés, amely szerint a csődöt követő megszűnés az erőforrások újraelosztásának elkerülhetetlen és elemi eszköze. Ebben a tekintetben a csődesetek száma és mérete a gazdasági szerkezetátalakítás sebességét mutatja. A második, tisztán pénzügyi megközelítés szerint, a vállalat eszközszerkezete megfelelő, ugyanakkor tőkeszerkezete és likviditása elégtelen, amely oda vezet, hogy a hosszú távon fenntartható vállalat rövid távon csődbe juthat. Az elv fő kiindulópontja a tőkepiaci tökéletlenségeken és az örökölt tőkestruktúrán alapszik, ugyanakkor e nézőpont szerint nem egyértelmű, hogy a csőd előnyös vagy hátrányos a szerkezetátalakítás szempontjából. Az utolsó megközelítés alapján a vállalatnak megfelelő eszköz és tőkeszerkezete van azonban elégtelen a vezetése, így az alacsony hatékonyságú működés csődhez vezet. Ilyenkor a tulajdonosnak el kellene bocsátania a vezetőt mielőtt csődbe vinné a vállalatot. Ennek korlátja lehet a megbízó-ügynök probléma és az információs aszimmetria. Ebben a szemléletben a vállalat működőképes lenne, ezért a csőd egyértelműen kedvezőtlen a szerkezetátalakítás és a társadalmi jólét szempontjából.

2.3. A fizetésektelenség és csődbe jutás folyamata

Stickney (1990) a pénzügyi egészség romlását az 1. ábrának megfelelően egy folyamatként jellemezte, amely balról jobbra, az egészségestől a problémás és felszámolt irányba tart. A kockázatelemzésnek és csődelőrejelzésnek pedig pont az a célja, hogy időben jelzést adjon arról, ha egy vállalat pénzügyi nehézségek vagy a csődös állapot felé halad. Hozzá kell tenni, hogy az utolsó felszámolási fázis kivételével, minden szakaszban van még átjárás bármelyik irányba, tulajdonképpen megfelelő intézkedésekkel a csődös állapotból is vissza lehet térni a pénzügyileg egészséges állapotba.

1. ábra A pénzügyi egészség változásának folyamata



Forrás: (Stickney, 1990, 245. p.) alapján saját szerkesztés

A fenti megközelítést részletesebben kibontva, tekintsünk vissza Fitzpatrick (1934) munkájához, aki a csődbe jutás folyamatát öt szakaszra bontotta.

1. Lappangás időszaka: Ekkor az első kellemetlen pénzügyi nehézségek alattomosan kialakulnak, de a tulajdonos ekkor még nem szerez tudomást róla. Amikor idővel a vezetés megtudja a valós helyzetet, törekedhet a probléma megoldására vagy akár figyelmen kívül is hagyhatja. Ez attól is függ, hogy mennyire tapasztalt és kompetens a vezetés. Az üzleti életben való jártasság és hozzáértés hiánya a vállalati nehézségek széles skáláját generálhatja, amely sokszor láthatatlan módon jelentkezik. Mindamellet a tapasztalt vezetés is belecsúszhat hibákba, például egy túlzottan optimista hozzáállással vagy a figyelem más területekre történő túlzott összpontosításával a háttérben rejtőzködő rendellenességek kibontakozhatnak.
2. Pénzügyi zavar időszaka: Még a normálisan működő vállalatok életében is felmerülhet egy-egy váratlan likviditáshiány. A vállalat fundamentálisan rendben lehet, ha az eszközei értéke nagyobb, mint az összes kötelezettsége és elégséges jövedelemmel is rendelkezik. Azonban előfordulhat, hogy az eszközei nem elég likvidek ahhoz, hogy teljesíteni tudja az azonnali pénzszükségleteket. A kialakulásának számos elkerülhető vagy elkerülhetetlen tényezője lehet, amely legtöbbször visszavezethető az értékesítési volumen átmeneti csökkenéséhez. A kritikus időszak lehet mindössze néhány nap vagy akár egy-két hónap. Ebben a szakaszban orvosolható a rendellenesség külső forrás bevonásával. A kellemetlen működési helyzet és a hozzá kapcsolódó kiigazítások még nem kerülnek nyilvánosságra, így csak a vezetés és néhány belső érintett tudomásával történik.
3. Pénzügyi fizetéképtelenség: Ebben a fázisban a vállalat nem képes teljesíteni a folyó kötelezettségeit. Viszont ekkor még van lehetőség talpra állni hosszú távon, a menedzsment újjászervezésével, a pénzügyi stratégia megváltoztatásával vagy hosszú távú külső forrás megszerzésével. Ebből a fázisból sok vállalat vissza tud térni az egészséges vállalatok sorába.
4. Teljes fizetéképtelenség: Ekkor már a kötelezettségek meghaladják a vállalat összes eszközeinek az értékét. Ebben a szakaszban a vállalat legtöbb érintettje és a hitelezők értesülnek a valós állapotokról. Ettől a ponttól kezdve tekintjük a vállalatot csődösnek, habár még nem biztos, hogy elindult a csődeljárás. Még van lehetőség lépéseket tenni a megmentésre, egyrészt a hitelezőkkel való egyezség keretében a tartozások újraütemezésére. A vállalat még tehet egy utolsó próbálkozást, hogy a kockázatnak megfelelő magas árszintű külső finanszírozást és forrásokat szerezzen. Ha ezek nem bizonyulnak elég kielégítőnek, akkor a csődeljárás jelenthet egy végső talpraállási módot.

5. Megerősített fizetéseképtelenség: Ez a szakasz általában gyorsan követi az előzőt. Hivatalosan bejelentik a fizetéseképtelenséget és jogi lépéseket tesznek a vállalat hitelezőinek a védelmében. A reorganizáció útján lehetőség van még a vállalat helyreállítására, azonban a legtöbb vállalat, amelyik eléri ezt a szakaszt, a felszámolása majdnem biztosra vehető.

2.4. A fizetéseképtelenség előrejelzésének szerepe

A globális gazdasági tér magával hozta, hogy a világ minden piacgazdaságában szükség van vállalati fizetéseképtelenség előrejelzésére és a csődmodellekre. A régi aforizma, mely szerint a nagyvállalatok igen biztonságosak és a csődbe jutás esélye csekély, mára megdőlni látszik, így valamennyi méretű vállalat célját képezi a fizetéseképtelenség felmérésének és a csődelőrejelzésnek. A gazdasági rendszer egyik-másik szektorának problémái könnyen tovaterjednek a többi szektorra, kifejezetten ha a pénzügyi szektor a van válságban (Virág, 2007). A rendszerszintű kockázatok csökkenthetők a pénzügyi nehézségek előrejelzésével, stabilabb gazdasági teret hozva létre. A recessziós környezetben pedig, amikor a finanszírozók kockázatkörülése dominál, a csődelőrejelzés fontossága hatványozottan felértékelődik (Nyitrai, 2014a).

A vállalati csődelőrejelzésre és a pénzügyi nehézségek felderítésére a technikák és módszerek széles skálája áll a pénzügyi szakemberek rendelkezésére. A megalapozott csődmodellek mellett a javasolható módszerek közt szerepel az éves beszámoló értékelése, a vállalati stratégia- és iparág-elemzés, a részvények alakulásának vizsgálata, a kötvényhozamok értékelése, a várható piacra lépők hatásának elemzése, a folyó- és jövőbeli időszakok pénzáramlásának értékelése (Samuels – Brayshaw – Craner, 1995).

A vállalatok bukása, a kapcsolódó magas költségek következtében, majdnem mindenik érintettre hatással van. A vállalati szinten túl társadalmi költségek is keletkeznek a csőd fertőző hatásának következtében, amely a foglalkoztatottságot és a társadalmi jólétet veszélyezteti (Virág et al., 2013a). A vállalat szintjén a menedzsment, a részvényesek és hitelezők érdeke, hogy időben észleljék egy esetleges csőd kialakulását és lépéseket tegyenek a helyzet súlyosságának csökkentésére vagy elkerülésére (Keasey – Watson, 1991). A társadalmi szintű problémák megelőzéséhez a kormányzatok szabályozási eszközökkel mérsékelhetik a csőd és felszámolások súlyos hatásait.

Az elmúlt időszakban a vállalatok működési környezete átalakuláson ment keresztül, a globális gazdaság kihívásainak valamint a széleskörű negatív történések következtében, különösen kiemelhetjük a COVID-19 és az orosz-ukrán háború

következményeit. A verseny kiéleződött, a vállalati sérülékenység növekedett, az állami beavatkozások fokozott jelentőséget kaptak. Ezekkel a radikális változásokkal a csődértékelési rendszereknek is tartaniuk kell a lépést (Virág et al., 2013a).

A független könnyvizsgálók és hitelminősítők egyedi vállalatok pénzügyi állapotának helytálló értékelését tudják biztosítani. Azonban ha a vállalatok tömegét kell osztályozzuk, a csődmodellek hatékonyabb előrejelzésre képesek, és általánosító képességei is megbízhatóbbak kifejezetten, ha kizárólag nyilvánosan elérhető adatok állnak a rendelkezésre (Keasey – Watson, 1991; Virág et al., 2013a).

A piaci tökéletlenségek elve mentén a rendelkezésre álló szabad tőke nem elégséges az összes pozitív nettó jelenértékű projekt finanszírozására, ezért számos potenciálisan nyereséges beruházás megvalósítása elmarad. A befektetők az információs aszimmetria okán nem juttatnak tőkét a jövedelmező projektekbe, csak a magas várható megtérülésű és relatív alacsony kockázatú beruházások fognak megvalósulni. Ebben a tekintetben a csődelőjelző modellek kettős előnnyel rendelkeznek. Egyfelől alkalmasak a kockázatok elemzésére, másfelől csökkentik a tulajdonosok és menedzserek közti információs aszimmetriát (Virág et al., 2013a).

Az alapvető felhasználói a csődmodelleknek a bankok, hitelintézetek, befektetők, hitelminősítő ügynökségek, biztosítók, könyvvizsgálók és szabályozási szervek. Számukra a vállalati csődelőjelzésből származó információ kulcsfontosságú az eredményes működéshez. A bankok kockázatkezelésének fontos eszköze, amely mentén a veszteségek minimalizálásához vezető döntések hozhatók. Egy befektető pedig jó ha ismeri a csőd kockázatokat egy kötvény vásárlása előtt (Lifschutz – Jacobi, 2010). A szabályozói szervek a pénzügyi stabilitás megerősítésére és korai figyelmeztető rendszerként vehetik igénybe a csődmodelleket. A csődesetek és okainak megértése segíti a szabályozókat jobb szakpolitikák kidolgozásában, hatékonyabb szabályozási keretet lérehozva a jövőbeli válságok megelőzése érdekében. Az sem utolsó szempont, hogy a vállalatok stratégiai tervezésében felhasználható és a partnerkockázat értékelésére ad egy eszközt, hozzájárulva egy megbízható ellátási lánc és fizetőképes ügyfélkör kialakításához. Végeredményben a csődelőjelzés lehetővé teszi, hogy a vállalatvezetés, a befektető, a hitelező és a részvényes megfelelő intézkedéseket hajtson végre, minimalizálva a várható veszteségeket (Brabazon – O’Neill, 2004; Lim et al., 2012).

2.5. A fizetéseképtelenségi eljárások jogi háttere

A gazdálkodó egységek folytonos ki- és belépése a gazdasági rendszerben egy természetes, ciklikus jelenségként értelmezhető. Azonban tekintettel arra, hogy a kilépés számos költséggel párosul, ezért van jelentősége annak, hogy bizonyos jogi eljárások és törvények védik az érintetteket. Majdnem minden országban alkalmazzák a fizetéseképtelenségi eljárások valamely formáját.¹ Magyarországon a csődeljárás és a felszámolási eljárás célja a fizetéseképtelen helyzetek méltányos kezelése. Az eljárások jogi előzménye volt az 1986. évi 11. számú törvényerejű rendelet, bevezette és szabályozta a felszámolási eljárás intézményét. Ez jelentős mérföldkőnek számított, azonban nem tudta betölteni szerepét, elsősorban a rendszerváltás okozta piacgazdasági körülmények közt. Így aztán az 1991. évi XLIX törvény hatályon kívül helyezte és bevezette a csődeljárás intézményét (Kerek, 1992). E törvény többször módosított változata² ma is szabályozza a csődeljárás, felszámolási eljárást és a végelszámolást.

A csődeljárás az adós fél kezdeményezésére indul, amelyben fizetési moratóriumot kérelmez a hitelezőkkel szemben. Az adós egyezségi javaslatot (teljesítési határidők módosítása, követelések átvállalás vagy elengedése) és a fizetéseképtesség helyreállítására alkalmas reorganizációs tervet készít, amelyet a hitelezőkkel megvitat. Ha a megállapodás nem jön létre a bíróság megindítja a felszámolási eljárást (Kerek, 1992).

Felszámolási eljárást a bíróság hivatalból, illetve az adós vagy a hitelezők kérelmére indíthatnak. Célja a hitelezők törvényben előírt sorrendben való kielégítése a fizetéseképtelen adós megszűnése során, vagyonának értékesítésével. A kielégítési sorrend elején a felszámoláshoz kapcsolódó költségek vannak, amelyek többek közt a munkabérek, bérjellegű juttatásokat, a felszámoló díját és a vagyon értékesítésének díját tartalmazzák. Ezt követi a zálogjoggal biztosított követelések, majd a gazdálkodó szervezetet terhelő tartásdíjak és különböző járadékok. Negyedik helyen a kötvény alapú követelésektől eltérő, magánszemélyek nem gazdasági tevékenységből eredő követelése, így például kártérítések és séremdíjak, ezen kívül a kis- és mikrovállalkozások és mezőgazdasági őstermelők követelése szereplenek. Az ötödik a sorrendben a társadalombiztosítási alapokhoz kapcsolódó tartozások, adók és köztartozások, majd a

¹ Az Európai Unió egyes nemzeteinek fizetéseképtelenségi eljárásait az alábbiakban találjuk: <https://e-justice.europa.eu/447/EN/insolvencybankruptcy>

² 2017. évi XLIX. törvény.

hatodik az egyéb követelések, amely kategóriába a hitelezők nagyobb része tartozik. Hetedik helyen a késedelmi kamatok és bírság jellegű tartozások, valamint utolsó helyen az adós vállalat tulajdonosainak és vezetőinek a követelései. A felszámolási eljárás sem visszafordíthatatlan, lehetőség van a felszámolás megszüntetésére egyezség megkötésével, a követelések maradéktalan teljesítésével, a felszámolási díjak megfizetésének biztosítékával.

A megszűnés egy további esete a végelszámolás, amelyben a vállalat saját elhatározása alapján nem folytatja a tevékenységét, tartozásait ki tudja egyenlíteni és jogutód nélkül megszűnik³. Ebben az esetben alapvetően nem áll fenn a fizetéseképtelenség problémája. A kényszertörlési eljárást a cégbíróság indítja jogsértés miatti szankcióként a gazdasági társaságok jogutód nélküli megszűnésének céljával. A kényszertörlés megindításának feltétele az, hogy a törvényes működés nem biztosítható, vagy a végelszámolást három éven belül nem fejezték be és a törlés iránti kérelem nem valósult meg. Az eljárás szerepe abban áll, hogy a vagyon nélküli vállalatokat gyorsabban és egyszerűbb lefolyással, jogutód nélkül szűnhessenek meg (Dzsula, 2018).

³ A végelszámolás és kényszertörlési eljárás törvényi keretének részleteit, a cégnyilvánosságról, a bírósági cégeljárásról és a végelszámolásról szóló 2006. évi V. törvényben találja az Olvasó.

3. A fizetésképtelenség- és csődelőrejelzés korábbi

empirikus kutatásai

A következő fejezet a csődelőrejelzés korábbi szakirodalmi kutatásait mutatja be. Tekintettel arra, hogy a csődelőrejelzés témájában megjelent publikációk száma nagyon magas, ezért a következő áttekintés nem tér ki az összes kutatásra, inkább csak azokra, amelyek mérföldkőnek tekinthetők, vagy módszertani megfontolásból kiemelkedőek a többi közül. Az átfogó jellegű irodalomelemző tanulmányokból bizonyos szempontok mentén rendszerezett képet kapunk a meglévő publikációk főbb megállapításairól. Ilyen tekintetben ajánlom az Olvasó figyelmébe a csődelőrejelzés területén végzett szisztematikus irodalomkutatásom, amely a nemzetközi szakirodalom legtöbbet hivatkozott csőd kutatásainak tapasztalatait foglalja össze (Ágoston, 2022a). A következő bemutatás azonban időrendi sorrendben igyekszik ismertetni a főbb eredményeket, megállapításokat, a fontosabb módszertanok alkalmazási példáját, illetve tanulságait.

3.1. Nemzetközi időrendi áttekintés

A csődelőrejelzésben az első feljegyzett próbálkozások az 1930-as évekig nyúlnak vissza, amikor még a többváltozós statisztikai módszereket mellőzve, az egészséges és fizetésképtelen vállalatok pénzügyi mutatói közt eltéréseket kerestek (Fitzpatrick, 1932; Winakor – Smith, 1935). A csődelőrejelzés alapjait Beaver (1966) fektette le, aki kutatásában az egyváltozós diszkriminanciaanalízis módszerével 90%-os pontossággal jelezte előre a fizetésképtelenséget egy évvel a csőd bekövetkezése előtt. Az első többváltozós csődelőrejelző modell Altman (1968) nevéhez köthető. Úttörő diszkriminanciaanalízis alapú modellje mindössze 33 fizetőképessé és ugyanennyi fizetésképtelen vállalat adataiból épült fel és öt pénzügyi mutatón alapult. A kifejlesztett Z-érték 95%-os pontossággal különítette el a mintában szereplő vállalatokat egy évvel a csőd bekövetkezése előtt. Azonban a két éves előrejelzési horizonton már csak 78%, a három éves előrejelzésnél a véletlen találgatásnál rosszabb, 48% átlagos pontossággal rendelkezett (Altman, 1968). Mindemellett Edward Altman az egyik legtöbbet hivatkozott szerző a csődelőrejelzés szakirodalmában (Bellovary – Giacomino – Akers, 2007) és a Z-modellt hosszú időn keresztül etalonként kezelte a szakirodalom, valamint az intézményi hitelminősítők, és máig számos modellfejlesztés összehasonlítási alapjául szolgál (Agarwal – Taffler, 2008; Bauer – Agarwal, 2014; Giannopoulos et al., 2022;

Manh – Nguyen, 2024). Egy másik híres modell is Altman nevéhez köthető, az úgynevezett ZETA-modell, a Z-érték második generációs továbbfejlesztett változata (Altman – Haldeman – Narayanan, 1977), amely hét pénzügyi mutatón alapult, 53 fizetésképtelen és 58 egészséges vállalati mintából épült fel. A csődöt megelőző évben a modell 96% pontossággal sorolta be a fizetésképtelen vállalatokat a tanuló mintán, ugyanakkor egy független tesztelő mintán átlagosan 89% pontosságú volt.

Az 1970-es években egy új modellcsalád jelent meg a szakirodalomban, a logisztikus regresszió elemzés. Az első logit módszeren alapuló csődelőrejelző modell Martin (1977) nevéhez köthető, aki amerikai bankok csődjét vizsgálta. A vállalati fizetésképtelenség előrejelzését logisztikus regresszió módszerrel Ohlson (1980) vezette be. Adatmintájában 105 csődbe ment és 2058 egészséges vállalatot vizsgált az 1970-1976 közötti időszakban. A modell előrejelző teljesítménye és megbízhatósága Altman (1968) modelljénél kedvezőbb volt (Ohlson, 1980). Ezt követően a logit módszer a csődkutatók körében rendkívüli módon elterjedt és nagy népszerűségnek örvendett. Még a 2000-es évek végéig is a leggyakrabban használt technika volt a hitelminősítések területén (Thomas, 2009), így a 2008-as globális gazdasági válság utáni csődelőrejelzés nagyfokú megugrása mellett is a logisztikus regresszió maradt az egyik leggyakoribb módszer (Shi – Li, 2019). A logisztikus regresszió analízis módszerének népszerűségét Bellovary – Giacomino – Akers (2007), Alaka et al. (2018) és Ágoston (2022a) is megerősíti, akik a szakirodalmi tanulmányok egy jelentős részét átvizsgálva arra jutottak, hogy a logisztikus regresszió az egyik leggyakrabban használt módszer a kutatók körében.

Az 1980-as években a csődelőrejelzés történetében megjelent a logit alternatív módszere, a probit-analízis, ami elsősorban Zmijewski (1984) nevéhez köthető. Zmijewski három pénzügyi mutatóra épülő modelljét 840 elemszámú vállalaton tanította be és ugyanennyi független vállalaton tesztelte. Kimutatta, hogy a valós csődaránytól jelentősen eltérő mintavétel nem torzítja jelentősen az osztályozási pontosságokat, sem a statisztikai következtetéseket. A szakirodalomban találunk néhány probit alkalmazási példát csődelőrejelzésre (Theodossiou, 1991; Boritz – Kennedy – Albuquerque, 1995; Ryu – Yue, 2005; Lin, 2009), azonban a módszer a szigorú feltételrendszere és nehézkes alkalmazása okán nem igazán terjedt el a gyakorlatban (Bellovary – Giacomino – Akers, 2007). 1985-ben megjelent a nemparaméteres módszerek közé sorolható rekurzív particionáló algoritmus, amely döntési fa segítségével egyszerű szabályok alapján ábrázolja a változókat és a hozzájuk tartozó küszöbértékek kombinációit (Frydman – Altman – Kao, 1985). A módszert alkalmazó - a szakirodalom által jelentősen hivatkozott

- csődkutatások McKee és Greenstein (2000), Chen (2011b), Delen et al. (2013), Korol (2013) munkájához köthetőek.

Az 1990-es évek elejétől a számítógépes technika gyors fejlődésének köszönhetően az adatbányászati és mesterséges intelligencia módszerek a csődelőrejelzésben is egyre inkább elterjedtek. Megjelentek a kiemelkedő módszernek számító neurális hálók, amelyek a biológiai idegrendszer elvére épülnek, és tanulás útján szerzik problémamegoldó képességüket. Odom és Sharda (1990) voltak az elsők, akik a csődelőrejelzésre alkalmaztak neurális hálózatot. Altman klasszikus modelljével vetették össze a neurális háló alapú modellt összesen 129 vállalat adatain. A tesztelő mintán elért legjobb eredményük az 50-50%-os tanuló-tesztelő mintafelosztással jelentkezett, ahol a csődösök helyes besorolási pontossága 82% volt a neurális hálónál és 59% a diszkriminanciaanalízis esetén. Feltcher és Goss (1993) Egyesült Államokbeli vállalatok kis mintás elemzésével tesztelték a háromtól hétig terjedő rejtett csomópontú előrecsatolt (feed forward) neurális hálók és a logit modell pontosságát. Arra jutottak, hogy a négy rejtett rétegű neuronháló a legkiválóbb, és több mint 10%-al pontosabb előrejelzésre képes a logisztikus regresszióhoz viszonyítva. A neuronháló előnyét mutatta ki a diszkriminanciaanalízissel szemben Coats és Fant (1993), Wilson és Sharda (1994), Tsukuda és Baba (1994) valamint Leshno és Spector (1996). Ebben az időszakban Boritz et al. (1995) tanulmányában kifejtette, hogy a neuronháló nem csodaszer, abban az esetben, ha kevés csődös példából tanulhat az algoritmus, akkor az új adatokon gyakorlatilag képtelen lesz megkülönböztetni a csődös eseteket az egészségesektől.

Szintén a 90-es években, teret nyertek a mesterséges intelligencia-alapú klaszterelemzési eljárásnak számító önszerveződő térképek (self organizing maps, SOM) vagy más néven Kohonen hálók. A neurális hálók családjába tartozó módszer dimenziócsökkentő eljárással vizuálisan különíti el az osztályokat, azért különösen látványos eredményt ad. A felülvizsgálat nélküli tanulási tulajdonságainak köszönhetően nincs szükség előre megadott outputokra, és nem kell eloszlási feltételeknek sem megfelelni. E módszert csődesetek vizsgálatára először Martin-del-Brio és Serrano-Cinca (1993) használta spanyol bankok és vállalatok mintáján és azt találták, hogy a számítási algoritmus egyszerűsége okán gyorsabb, mint a hibavisszaterjesztő neuronháló, és a fentebb említett néhány speciális tulajdonsága bizonyos esetben versenyképessé teszi más módszerekkel szemben. Ezt bizonyítja, hogy Chen (2011a) hibrid SOM modelljei, valamint du Jardin és Severin (2011) modelljei felülmúlták a benchmark módszereket. Ugyanakkor egyes vélemények szerint, további módszerekkel, statisztikákkal, szakértői

rendszerrel kombinálva fejtheti ki az igazi előnyét a vállalati fizetőképesség megítélésében (Serrano-Cinca, 1993). Lee et al. (2005) összehasonlító tanulmányában azt mutatta ki, hogy a módszer nem biztosít magas előrejelzési pontosságot a neuronhálózhoz és a logit-hoz viszonyítva, azonban praktikusabb eszköz a fizetéseképtelenség időbeni alakulásának vizsgálatára főként a 2-3 évet meghaladó elemzéseknél.

Mar Molinero és Ezzamel (1991) a többdimenziós skálázást (multidimensional scaling, MDS) vezette be a pénzügyi nehézséggel küzdő vállalatok vizsgálatára. A szerzők 1976-1982 időszakban 47 csődös és 235 egészséges vállalat pénzügyi mutatóinak eltéréseit elemezték az MDS módszerrel. A szerzők értékes következtetéseket vontak le. Az egészséges vállalatok esetében a jövedelmezőség szempontja az elsődleges, míg más dimenziók mint a likviditás és a működő tőke, háttérbe szorultak. A csődös vállalatok esetén a csőd előtti 2-5 években a működő tőke javítása kapott nagy hangsúlyt, annak érdekében, hogy a romló helyzetben időt nyerjen a vállalat, lemondva a potenciális rövid távú nyereségről, és helyrehozzák az értékesítési hanyatlást. A csőd előtti első évben azonban a likviditás dimenziója volt a leginkább meghatározó, amely a csőd elkerülése érdekében tett kétségbeesett kísérletként értelmezhető. Neophytou és Mar Molinero (2004) 50 csődös és 50 egészséges részvénytársaságot vizsgált, amelynek 90%-át az MDS modell tanulási folyamataira használt, a maradék 10%-ot tesztelésre. Ezen a téren az MDS legnagyobb hátránya, hogy önmagában előrejelzésre nem alkalmas. A szerzők az MDS szignifikáns dimenzióit a logit módszer segítségével azonosították, ezt követően pedig az új adatokat egyenként bevitték a teljes tanuló adatállományba és újratestelték a modellezést. Azután azt vizsgálták, hogy az új vállalati megfigyelés a pénzügyi változók struktúrája alapján melyik osztályhoz hasonlít, azaz a szignifikáns dimenziókban melyik régiókban helyezkedik el. Ezzel a megközelítéssel a tesztmegfigyeléseket 80% pontossággal sorolták a helyes osztályba. A szerzők véleménye szerint a módszer paradigmaváltást idéz majd elő a csödelőrejelzésben, az eredmények vizualizációs szemléltetésének, a könnyű értelmezhetőségnek, a kvalitatív információk bevonásának és az adatokkal kapcsolatos csekély elvárásoknak köszönhetően. Azonban a megállapítás húsz év távlatában még nem igazolódott be. További alkalmazási példája a módszernek Mar Molinero és Serrano-Cinca (2001) munkája, amelyben spanyol bankcsődöket elemeztek. Javaslatuk alapján a módszer előrejelzés helyett inkább a pénzügyi nehézségben lévő bankok pénzügyi pályájának nyomon követésére alkalmazandó. Khoja et al. (2016) arab tőzsdén jegyzett részvénytársaságok elemzésére a logit mellett az MDS módszert alkalmazták. Négy dimenziót azonosítottak, és megfigyelésük alapján, a

fizetéképtelen vállalatok csak egyik dimenzióknak tulajdonítottak jelentőséget, míg a fizetőképesek a további háromnak.

Slowinski és Zopounidis (1995) durva halmazok elméletén (rough set theory) nyugvó módszert dolgozott ki a csőd előrejelzésére Pawlak (1982) elvei mentén, amely szorosan kapcsolódik a fuzzy halmazok elméletéhez. A kutatók egy görög befektetési bank ügyfeleinek 39 elemű mintáját felhasználva, a pénzügyi mutatók mellett kvalitatív mutatókat is alkalmaztak úgy, mint a menedzser tapasztalata, a vállalat piaci pozíciója és a piaci rugalmasság. A szabályrendszerük tökéletes pontossággal becsülte a vállalatok fizetőképességét (Slowinski – Zopounidis, 1995). Az ellenük felmerülő kritika lehet, hogy a végső modelljüket nem értékelték a tanuló mintától eltérő, külső adatmintán. Ezt követően azonban egyes kutatók validálták a módszert, mivel más adatokon sikerrel alkalmazták a vállalati fizetőképesség döntési szabályaként (Dimitras et al., 1999; McKee, 2000; Beynon – Peel, 2001; Xiao et al., 2012).

Az esetlapú következtetés (case based reasoning, CBR) a csődkutatásokban, mint adatbányász technika, vegyes eredményeket hozott. A korai modellek teljesítménye jellemzően alulmaradt a korábbi statisztikai módszerekhez viszonyítva (Bryant, 1997; Jo – Han – Lee, 1997). Annak ellenére, hogy kevésbé alkalmasnak bizonyult a csődelőrejelzésre, mégis hasznos eszköznek mutatkozott abban az esetben, ha kis elemszámú tanuló minta áll a rendelkezésre (Jo – Han – Lee, 1997). A későbbi empirikus kutatásokban, főként egy hibrid előrejelző rendszer részeként, a módszer teljesítménye sokkal inkább kiegyensúlyozottá vált (Li – Sun, 2008; Ahn – Kim (2009); Li et al., 2014).

A neurális háló versengő módszerének tekinthető a gépi tanulási elveken alapuló support vector machine (SVM) eljárás, amely sokoldalú, összetett problémák megoldására alkalmas, így a lineáris és nemlineáris osztályozásra is. Fan és Palaniswami (2000) alkalmazták elsőként az SVM módszert csődelőrejelzésre. A szerzők 174 Ausztrál kv-k mintáján azt találták, hogy az SVM modell jobb eredményre képes, mint a diszkriminanciaanalízis és a neurális hálók családjába tartozó többrétegű preceptron valamint az LVQ (Learning Vector Quantization). Az ezt követő tanulmányok is az SVM színvonalas teljesítményét bizonyították (Shin – Lee – Kim, 2005; Chaudhuri – De, 2011), és a módszer számos esetben 90% feletti pontossággal határozta meg a fizetéképtelenség tényét (Hua et al., 2007; Xu – Wang, 2009; Kim, 2011; Li – Sun, 2012). Kirkos (2015) irodalomkutatása alapján a csődelőrejelzés területén a modern, intelligens módszerek közül messze a leggyakrabban alkalmazott algoritmus az SVM.

Shumway (2001) túlélés-analízis módszerrel fejlesztett hazard modelljét (hazard rate model), összevetette Altman (1968) és Zmijewski (1984) klasszikus modelljeivel. Shumway az elsők közt volt, aki a pénzügyi mutatók mellett piacorientált magyarázó változókat is alkalmazott és kifejezte a hazard modell egyszerű becslését és könnyű magyarázhatóságát. A modellje Altman (1968) és Zmijewski (1984) megoldásától eltérő statisztikai következtetésekre jutott, a hagyományos változók közel fele nem volt hatással a fizetéképtelenségre. Rámutatott, hogy a csőd előrejelzése statikus modellel torzításokhoz vezethet, mivel nem veszi figyelembe, hogy a vállalatok a csődöt több évvel megelőzően már kedvezőtlen állapotban voltak, vagy rövid idő alatt mentek csődbe. Chava és Jarrow (2004) havi és éves megfigyelések alkalmazásával validálta Shumway (2001) modelljének kiváló teljesítményét Altman (1968) és Zmijewski (1984) modelljével szemben. Kiemelték az iparági hatások fontosságát az előrejelző pontosságra, valamint a piaci változók jelentős információtartalmát a csőd vonatkozásában. Hillegeist et al. (2004) hazard modell segítségével összehasonlította Altman (1968), Ohlson (1980) és Black – Scholes – Merton piaci alapokon nyugvó opció-árazási modelljét. Az eredmények következtében az utóbbi modellt javasolta csődelőrejelzésre. Beaver, McNichols és Rhie (2005) hazard modelljének robusztusságát bizonyította negyven éves időszak adatain, valamint megerősítette az előző tanulmányok eredményeit abban, hogy a pénzügyi mutatók előrejelző képességének romlását hosszabb időhorizontokon a piaci mutatók párhuzamos alkalmazása kompenzálja. A fentiekkel némileg ellentmondó következtetésre jutott Tian, Yu és Guo (2015) hazard csődelőrejelző modellje, amely LASSO változószelekció után a számviteli mutatókból nyert információ jelentőségét magasabbra értékelte, mint a piaci alapú változókét. Mayew et al. (2015) hazard modellek segítségével az éves beszámolókból található vezetői jelentések és elemzések nyelvi tónusát, a könyvvizsgálói jelentések szöveges tartamának szerepét vették figyelembe a vállalatok fennmaradásának előrejelzésére, pénzügyi és piaci változók kíséretében. A szerzők azt tapasztalták, hogy önmagában a vezetői beszámolók hasonló mértékben képesek a fizetéképtelenség előrejelzésére, mint a modellezés során alkalmazott pénzügyi mutatók.

Jones és Hensher (2004) szerzőpáros random paraméter modellt (mixed logit model) fejlesztett és összehasonlította a multinominális logit modellel. Háromkimenetelű osztályozási modellt dolgoztak ki: egészséges vállalatok, fizetéképtelen vállalatok és csődeljárás alatt lévő vállalatok. Az eredményeik azt mutatták, hogy a random paraméter modell minden vizsgálati horizonton fölülmúlta a multinomiális logit modellt. Trabelsi et

al. (2015) Bayes-i, hazard és mixed logit modelleket állított párhuzamba és arra következtetésre jutott, hogy a Bayes-i és a mixed logit a legjobb előrejelző teljesítményre képes módszerek.

A csőd kutatás szakirodalmában három gyakori metamódszert vagy más néven együttes (ensemble), gépi tanuláson alapuló módszert fedezhetünk fel. Ezek a gyorsítás vagy boosting, a zsákolás vagy bagging és a véletlen erdő vagy random forest néven ismert eljárások (Barboza et al., 2017). Mivel az eljárások egyesítik több tanuló módszer előnyét, ezért ezt követően az egyetlen legjobb módszer keresése némileg teret vesztett a kutatók körében. West et al. (2005) neurális háló alaptanuló osztályozóra három együttes eljárást alkalmazott: bagging, boosting és keresztvalidálást. A csődös és egészséges vállalatok osztályozására átlagosan a bagging eljárás eredményében volt a legkisebb a hiba mértéke.

Alfaro Cortés et al. (2007) az Adaboost.M1 algoritmust alkalmazta annak érdekében, hogy a döntési fa csődelőrejelző képességét javítsa. Spanyol vállalati adatok felhasználásával a módszer 94% pontossággal osztályozta a tesztminta vállalatait. A legfontosabb magyarázó változónak a működő tőke és árbevétel aránya, az adó és kamatfizetés előtti eredmény és összes eszköz aránya, valamint a cash flow és összes kötelezettség aránya bizonyult. Ugyanezen szerzők egy következő munkájukban a boosting és neurális háló összehasonlítását végezték el (Alfaro Cortés et al., 2008). Következtetéseik szerint az Adaboost eljárás minden tesztmintán előnyösebb eredményt ért el, mint a neurális háló, mivel átlagosan 30%-al kevesebb hibát ejtett. Hasonló eredményre jutott Sun et al. (2011) kínai vállalatok csődelőrejelzésénél. A boosting előrejelző teljesítménye fölülmulta a neurális háló mellett az SVM módszert is. Koreai vállalatok mintáján Kim – Kang (2010) szerzőpáros neurális háló alapú bagging és boosting módszert fejlesztett és kimutatták az együttes módszerek magasabb teljesítményét a hagyományos neurális hálóval szemben. A tesztmintán az átlagos előrejelző pontosság 76% volt a bagging módszernél, 75% a boostingnál és 71% volt a hagyományos neurális háló esetében. Chandra et al. (2009) random forest alapú hibrid módszereket vonultatott fel dotcom vállalatok csődjének előrejelzésére. Szintén random forest technikával készült döntési fákat tesztelt Figini et al. (2016) német kkv-k csődelőrejelzésére. Barboza et al. (2017) Észak-Amerikai vállalatok adatain bizonyította a random forest magasabb előrejelző pontosságát és kisebb hibaarányát hét alternatív csődelőrejelző módszerrel szemben. Lombardo et al. (2022) következtetése alapján a random forest kiváló előrejelző teljesítményt nyújt és aránylag kevés a bemeneti

paraméter igénye, így gyorsabb a tervezése és a tanítása is. A neurális hálónak általában nehezebb a tervezése és hosszabb a számítási ideje, azonban a szerzők a csődelőrejelzésre hatékonyabbnak találták, mivel új adatok esetén a legjobb általánosító képességeikkel rendelkeznek.

A neuronháló módszereknél a mélytanulás (deep learning) alkalmazása az utóbbi években kezdett kibontakozni. A csődelőrejelzésben két fő mélytanulási módszer fedezhető fel, a rekurrens- és a konvolúciós neuronháló (Qu et al., 2019). A rekurrens neuronháló (Recurrent Neural Network, RNN) főként szekvenciális és idősoros adatok elemzésére alkalmas, amelynek speciális típusa a Long Short-Term Memory (LSTM). Az RNN és LSTM módszertanra épülő csődelőrejelzést találunk Kim – Cho – Ryu (2021) munkájában. Mai et al. (2019) mélytanuló csődelőrejelző modellje numerikus valamint szöveges adatok felhasználásával valósult meg, több mint tízezer amerikai vállalat mintáján. A mélytanulás hatékonynak bizonyult a szöveges adatok osztályozásánál, azonban a numerikus adatokkal a pontossága nem volt jobb, mint az SVM-nek vagy a random forest-nek. A legjobb modellek a szöveges és numerikus adatok együttes alkalmazásával jelentkeztek. A szerzők bebizonyították a szöveges beszámoló információ értékét, és kiemelték, hogy a hagyományos kvantitatív modellekbe beépítve számottevő teljesítménynövekedés érhető el. A konvolúciós neuronháló mélytanuló algoritmussal du Jardin (2022) vizsgálódott. A bemeneti pénzügyi adatok képként történő reprezentációval létrejött modell magasabb előrejelző képességekkel rendelkezett, mint az SVM és az ensemble modellek. A mélytanulással szemben felhozható ellenérv a nehéz magyarázhatóság, ami végső soron nem elhanyagolható a csődelőrejelzés területén (Mai et al., 2019; Soui et al., 2020).

3.2 A magyar csődelőrejelzés mérföldkövei

A magyar csőd kutatás kezdete az 1990-es évekre vezethető vissza, és az első magyar csődmodell Virág Miklós nevéhez fűződik, aki a rendszerváltás utáni első két év adatain, többváltozós diszkriminanciaanalízis és logisztikus regresszió alapú modelleket készített (Virág, 2004). Az úttörő tanulmányok közé tartozik Virág Miklós és Hajdu Ottó 1996-os nagy terjedelmű kutatása, amelyben a gazdaság egészének, a nemzetgazdasági ágaknak és az ágazatoknak a csődmodelljét készítették el úgy, hogy valamennyi modelljük 90% feletti besorolási pontosságokat ért el (Virág – Hajdu, 1996; Hajdu – Virág, 2001). Hámori (2001b) logisztikus regresszió-alapú modelljében 685 egészséges és 72 fizetéképtelen vállalat adatait használta fel. A szerző módszertani előrelépéseket

tett, többek között a változókat négy faktorba tömörítette, amelyek segítségével a modell osztályozási pontossága 95% volt. Virág és Kristóf (2005) neurálisháló-alapú modellt hoztak létre az első hazai csődmodell adatbázisán, majd eredményeiket a diszkriminanciaanalízissel, illetve a logisztikus regresszió módszerével kapott eredményekkel vetették össze. A szerzők különböző hálóstruktúrákat teszteltek, melyek közül a legjobbnak a négyrétegű neuronháló bizonyult backpropagation⁴ algoritmus használata mellett. A neurális háló osztályozási pontossága átlagosan 86,5%-os volt, így a diszkriminanciaanalízis teljesítményét 8,6 százalékponttal, a logisztikus regresszió eredményét pedig 4,7 százalékponttal haladta meg. Virág és Kristóf (2006) az iparági ráták alkalmazásával négy különböző csődelőrejelző módszer, a diszkriminanciaanalízis, a logisztikus regresszió, a döntési fa és a neurális háló, összehasonlítását végezte el. Szintén Virág és Kristóf (2009) a megfigyelések többdimenziós skálázása után kapott koordináták segítségével logisztikus regressziót futtatott, és a kismintás csődelőrejelzésük 94%-os besorolási pontosságot ért el.

A csődbejutás kialakulásának okait és folyamatait magyar nyelven Kristóf (2005) tárgyalja, aki szervezetelméleti megközelítések részletezésével támogatja a csődelőrejelzés empirikus kutatásait. A csődelőrejelzés modellépítés nélküli módszertani összefoglalóját Szabadosné Németh és Dávid (2005), Oravetz (2007) valamint Rattig (2015) tanulmányában olvashatjuk.

Virág és Nyitrai (2013b) alkalmazta először az SVM-módszert magyar vállalatok adatain. Azt találták, hogy ugyanazon az adatbázison az SVM közel 5%-al magasabb átlagos osztályozási pontosságot ért el, mint a neurális háló. Különböző kernel függvényeket teszteltek, és a kevésbé használatos varianciaanalízis függvényt (analysis of variance, ANOVA) találták a leghatékonyabbnak. Ugyanezen szerzők (Virág – Nyitrai, 2014a) két együttes metamódszer, az Adaboost és a bagging teljesítményét hasonlították össze C4.5 döntési fák alkalmazásával. A modellépítéshez 976 magyar vállalat 2001 és 2012 közötti adatait használták fel. A legjobb eredményt a bagging biztosította, átlagosan 81%-os osztályozási pontossággal, az Adaboost 79%, a C4.5 73% pontosságot ért el. A szerzőpáros a durva halmazelméleten (rough set theory) alapuló módszert is tesztelte (Virág – Nyitrai, 2014b). A „ha-akkor” típusú szabályokat használó eljárást összehasonlították az SVM, illetve a neurális hálóval, és arra az eredményre jutottak, hogy az előbbi két módszer azonos, 89%-os osztályozási pontosságot ért el a

⁴ backwards propagation of error = hiba-visszaterjesztés

tesztmintán, míg a neurális háló kevéssel elmaradva, 88%-os pontosságot mutatott. Magyarországi adatokon a k-legközelebbi szomszéd módszert Nyitrai (2015) alkalmazta először csődelőrejelzésre. Az ezer elemű kiegyensúlyozott vállalati mintán a több időszakot, a csőd előtti 1, 2 és 3 évet, egyidejűleg tartalmazó modellek teljesítettek a legjobban. Nyitrai (2017) a csődelőrejelzésben stock és flow típusú bemeneti adatok hatásait vizsgálta az előrejelző teljesítményre.

Kristóf (2018) hazai mikrovállalkozások csődelőrejelzésén keresztül az esetalapú következtetés (case-based reasoning, CBR) módszertanát és alkalmazási feltételeit ismertette. Emellett egy összehasonlító elemzés keretében a CBR módszer előrejelző erejét vetette össze a döntési fa, a logisztikus regresszió és a neurális háló módszerekkel. A tesztminta megfigyelései alapján a CBR előrejelző teljesítménye elmaradt a neurális háló és logisztikus regresszió eredményétől. Nyitrai (2018) a vállalati méret és a tevékenységi kör hatását vizsgálta a csődelőrejelző modellek szerkezetére és teljesítményére. Az eredményei arra utaltak, hogy a vállalat mérete tekintetében jobb eredmény érhető el, ha a főbb méretkategóriákra külön modelleket építünk, vagy legalább bináris változó formájában szerepeltetjük az átfogó modellekben. Az egyes nemzetgazdasági ágakban működő vállalatokra pedig célszerű önálló modelleket felállítani a magasabb teljesítmény érdekében. Nyitrai (2019) a dinamizált csődmodellekkel végzett empirikus kutatásában a hagyományos statikus mutatószámok mellé a pénzügyi mutatók időbeli trendindikátorát is bevonta, amely az osztályozási teljesítmény javulását eredményezte a hagyományos, statikus modellekhez viszonyítva.

A magyarországi csődelőrejelzés harminc éves fejlődéstörténetéről széles körű elemzést készített Kristóf és Virág (2019; 2020). A szerzők megállapításai alapján a magyar szakirodalom leggyakoribb eljárásai a logisztikus regresszió, a neurális háló és a döntési fák, ugyanakkor a módszertani előrelépéseket a mesterséges intelligencia, gépi tanulás, hibrid modellezés és adatbányászat jellemzi. A klasszikus diszkriminanciaanalízis módszer viszonylag egyszerű interpretálhatóságát beárnyékolja a szigorú előfeltétel rendszer és a mérsékelt teljesítmény. A szerzők ajánlják a jövőbeli kutatások során az újabb módszerfejlesztések mellett viszonyítási alapként szerepeltetni legalább az egyiket a logisztikus regresszió, a neurális háló és a döntési fa módszerek közül (Kristóf – Virág, 2019).

Molnár (2019) az önkormányzati tulajdonban álló vállalatok működőképességét, fizetőképességét vizsgálta külföldi csődmodellek segítségével. A vállalatok 55%-a elfogadható fizetőképességgel rendelkezett, 45%-a stabil tőkeállapotban volt. Ehhez

viszonyítva az amerikai vállalatokra fejlesztett Altman-modell (1968) pesszimista becslést adott, a Zmijewski-modell (1984) előrejelzése pedig túl optimista volt. Patyi – Szóka – Pataki (2023) három nyugat-magyarországi megyére öt klasszikus csődelőrejelző modellt tesztelt az élelmiszerkereskedelmi ágazatban. Az öt modellből kettőt alkalmatlannak találtak az élelmiszerkereskedelmi vállalatok csődjének előrejelzésére. Szántó (2023) a logit alapú csődelőrejelzés modellek kiugró értékeinek kezelésére tett kísérletet és azt találta, hogy a legelőnyösebb módszer a helyettesítés, mégpedig a legközelebbi, nem kiugró, értékkel.

Kristóf – Virág (2023) a visegrádi országok nagyvállalatainak csődelőrejelzését végezték el egy széles adatmintán és hosszú időtávra, 2013-tól 2021-ig, belefoglalva a COVID-19 alatti pénzügyi eseményeket. A CHAID és logisztikus regresszió alapú modelljeik kedvezőek voltak egészen a 2020-as előrejelzésig, azonban a COVID-19 alatti 2021-es előrejelzés teljesítménye jelentősen alacsonyabb volt. Ennek feloldására egy új, csak a legfrissebb adatokat figyelembe vevő modell specifikálását javasolták. A szerzők ajánlása alapján a korábban kidolgozott, kevésbé hatékony modellek helyett újakat kell fejleszteni és a jelenlegi dinamikus gazdasági környezetben a csődmodelleket legalább évente felül kell vizsgálni.

3.3. A fizetéseképtelenség előrejelzésének alternatív változókra épülő modelljei

Megfigyelhető, hogy a vállalati csőd és fizetéseképtelenség nem mindig belső okokra, működési hibákra vezethető vissza, hanem külső gazdasági és környezeti válságok is előidézhetnek csődhelyzeteket. Azonban elmondható, hogy ez az irány nem tekinthető a csődelőrejelzés főágának. A vállalat belső pénzügyi adatai domináns szerepet játszanak a csődmodellezésben (Nyitrai, 2017). Ha elegendően széles körben merítünk a szakirodalmi kutatásokból, akkor is csak néhány tanulmány tesz említést az alternatív változók alkalmazásáról (du Jardin, 2009; Ágoston, 2022a). Ezen alternatív változók leggyakrabban piaci és makrogazdasági változók, amelyek nem egyedileg, hanem a bevált pénzügyi magyarázó változók mellett kapnak helyet (Ágoston, 2022a).

Néhány fizetéseképtelenséget kutató tanulmány említést tesz iparági, vállalatirányítási (Linag et al., 2016), menedzsmentattitűd- (Ciampi, 2015) és egyéb nem pénzügyi mutatók alkalmazásáról (Lussier, 1995; Teng – Bhatia – Anwar, 2011), azonban ezek inkább ritka kivételnek számítanak. Lussier (1995) kérdőíves felmérésében 1-től 7-

ig terjedő Likert-skálán mérte a vállalatok belső jellemzőit és arra jutott, hogy a jövő tervezése, a szakmai tanácsadás minősége, az alapító vállalkozó végzettségi szintje valamint az alkalmazottak megszerzésének nehézségi szintje szignifikánsnak bizonyult a sikeres és a csődös vállalatok elkülönítésére. Teng – Bhatia – Anwar (2011) a túlélés legfontosabb tényezői közé sorolta többek közt az alkalmazottak megtartásának képességét, a kiváló ügyfélkapcsolatot és a kiemelkedő vezetői képességekkel rendelkező felső vezetők meglétét.

Az alternatív változókat alkalmazó tanulmányokban az egyik leggyakoribb bemeneti mutató a munkanélküliségi ráta, amely rendszerint pozitív kapcsolatban áll a fizetéseképtelenséggel és a megszűnéssel (Everett – Wattson, 1998; Buehler – Kaiser – Jaeger, 2012; Figlewski – Frydman – Liang, 2012; Mare, 2015). A másik gyakori magyarázó változó a bruttó hazai termék és annak változatai, ami a fizetéseképtelenséggel negatív összefüggésben áll (Ilmakunnas – Topi, 1999; Box, 2008; Hol, 2007; Aly – Daly, 2010; Rezende et al., 2017). Továbbá a csődöt a kamatlábak alakulásával (Hilgeist et al., 2004; Hernandez Tinoco – Wilson, 2013), valamint a valutaárfolyam-változással is magyarázza a szakirodalom (Holmes – Hunt – Stone, 2010; Nam et al., 2008; Karas 2022). Nem utolsósorban a részvénypiaci hozamok segítségével is igyekeztek a csődvalószínűséget interpretálni (Shumway, 2001; Hernandez Tinoco – Wilson, 2013; Agrawal – Maheshwari, 2014).

Kis számban fordulnak elő azok a tanulmányok, amelyek a csőd és a földrajzi elhelyezkedés közötti kapcsolatot próbálják feltárni. Keeble – Walker (1994) kifejezi ennek egyik fő okát, mely szerint súlyos hiányosságok vannak a vállalati csőd térbeliségének elméleti megalapozásában: „... *the almost complete lack of theory relating to the location of business failures*”. (Keeble – Walker, 1994 p. 423). Buehler – Kaiser – Jaeger (2012) kifejezetten hangsúlyozza a földrajzi elhelyezkedés fontosságát a vállalatok csődjével kapcsolatban. A szerzők széles svájci adatmintán kimutatták, hogy a nagyobb városokban és a turisztikai településeken alacsonyabb a csődarány. A kedvező üzleti feltételekkel rendelkező régiókban, ahol alacsony szintű a munkanélküliségi ráta és a vállalati adók, valamint magas az állami beruházás, ott mérsékelt a csődarány. Fotopoulos – Louri (2000) az agglomerációs előnyöket hangsúlyozza. Szerintük az újonnan megalakuló vállalatok nagyobb eséllyel válnak túlélővé nagyvárosi környezetben. A szerzőpáros agglomerációs pontnak kizárólag a fővárost, Athént, tekintette. Empirikus eredmények bizonyítják, hogy a térbeli elhelyezkedés számít a csőd szempontjából, és kimutatható a szomszédoktól tovaterjedő fertőző hatás (Andreano et al., 2018; Mate-

Sánchez-Val, 2021; Rico – Cantarero – Puig, 2021; Serra et al., 2022). A vállalatok megszűnésére a külső finanszírozási adottságokat nem hagyhatjuk figyelmen kívül. Térbeli elemzések kimutatták, hogy a kisvállalatok csődjére kifejezetten negatív hatással van a helyi pénzügyi szolgáltatások fejlettsége és koncentrációja (Arcuri – Brunetto – Levratto, 2019; Arcuri – Levratto, 2020). Néhány módszertani kérdés mellett, a térbeliség és a felszámolások kapcsolatát saját empirikus kutatásban is kimutatom, nyomatékosítva a fenti kutatások eredményét főként a munkanélküliség és a fizikai bankfiókok hatásainak tekintetében (Ágoston, 2024).

Az alternatív változók ritka alkalmazásának lehetséges magyarázata, hogy nem mindig mutatkozik kifejezett hatás a fizetéseképtelenségre. Kim – Sohn (2010) modelljében tíz makrogazdasági változó közül csupán kettő bizonyult szignifikánsnak. Hasonló tendencia figyelhető meg Sousa et al. (2022) tanulmányában, ahol a hat kiválasztott makrogazdasági mutatóból csupán kettő volt hatással a csőd kialakulására: a bruttó hazai termék növekedési üteme és az új vállalkozások aránya mutatók. Hernandez Tinoco és Wilson (2013) empirikus tanulmányukban kifejtik, hogy a piaci alapú mutatók jelentős hozzájárulással rendelkeznek a csődelőrejelzés teljesítményéhez, mivel a pénzügyi mutatókban nem tükröződő többletinformációt hordoznak. Továbbá, a szerzőpáros megállapította, hogy a makrogazdasági mutatók bevonása a modellekbe csak kis mértékben járult hozzá a modellek fejlesztéséhez.

A makrogazdasági változók alulreprezentáltságát mutatja Nam et al. (2008) tanulmánya, amelyben 33 pénzügyi változó mellé csupán két makrogazdasági változót vettek fontolóra, amelyből végül egyet alkalmaztak. Chen (2011a) számításba vett 33 pénzügyi változót, 8 nem pénzügyi változót és egy makrogazdasági indexet. A végkövetkeztetése szerint a pénzügyi előrejelzésre a vállalat belső pénzügyi mutatói nagyobb hatással bírnak, mint a nem pénzügyi és makrogazdasági változók.

Altman et al. (2016) megerősíti a nem pénzügyi változók alkalmazásának fontosságát azzal, hogy a pénzügyi és nem pénzügyi változók csak gyenge korrelációt mutatnak egymással. Ez azt jelenti, hogy a nem pénzügyi változók a pénzügyi változókkal kombinálva fontos többletinformációt tudnak nyújtani a csőd előrejelzésében. A bizonyítás azonban elmaradt.

A pénzügyi arányszámok mellett narratív jellegű beszámolók is alkalmasak lehetnek a fizetőképesség becslésére (Cecchini et al. 2010). Hopwood – McKeown – Mutchler (1989) kimutatta, hogy a szöveges könyvvizsgálói jelentés, egyéves előrejelzési horizonton jelentős magyarázóerővel bír, illetve olyan többletinformációt hordoz, amely

nem található meg a pénzügyi mutatókban. Mayew et al. (2015) a szöveges közlemények szerepét vizsgálták a vállalatok fennmaradásának előrejelzésére, és azt tapasztalták, hogy a menedzseri beszámolók 15%-ban magyarázták a jövőbeli fizetőképességet. Shirata et al. (2011) szövegbányászat segítségével térképezte fel, hogy a fizetőképes vállalatok szöveges jelentésében a „kutatás és fejlesztés” illetve „beruházás” egy mondatban szerepel az „osztalék” és „eredménytartalék” kifejezésekkel. Ezzel szemben, amennyiben a „sajnálattunkra” és a „nem teszi lehetővé” kifejezések egy mondatban szerepeltek az „osztalék” kifejezéssel, akkor a vállalat nagy eséllyel ment csődbe, még akkor is, ha a pénzügyi adatok stabilitást tükröztek. Korábbi saját empirikus kutatásban a LDA (látens Dirichlet-allokáció) szövegbányász klaszterezéssel vizsgáltam a vállalatok éves beszámolókhöz tartozó kiegészítő mellékletek szöveges tartalmát. A mellékletekben található pozitív és negatív szavak gyakorisága nem adott magyarázatot a fizetőképesség tekintetében. Mindössze a szöveges beszámolók terjedelme bizonyult magyarázó erejűnek. Az eredményeim alapján a fizetéseképtelen vállalatok beszámolóí rövidebbek és szűkszavúbbak voltak, az egészséges vállalatok beszámolójához képest (Ágoston, 2022c).

A piaci és makrogazdasági változók alternatív tényezőként való bevonása a csődmodellekbe továbbra is nyitott kérdés marad. Bár egyesek mellett érvelnek, hogy segítségükkel sokkal szélesebb gazdasági tendencia és információ ragadható meg, indokoltságukra nem találunk egyenes választ. A fentiekből az körvonalazódik, hogy a pénzügyi mutatókkal ellentétben a piaci és makrogazdasági változók csekély mértékben jeleznek közvetlen kapcsolatot a fizetőképességgel, így aztán kevésbé hatékonyak az előrejelzési feladatokra. Jóllehet az alternatív változók kiegészítő betekintést biztosítanak a vállalati fizetéseképtelenségre vonatkozóan, az elsődleges hangsúly továbbra is a pénzügyi mutatókon marad, mivel azok magasabb teljesítménnyel képesek előre jelezni a pénzügyi nehézséget és fizetéseképtelenséget.

4. A fizetésképtelenség előrejelzésének módszerei

A következő fejezetben azokat a - szakirodalomban általánosan alkalmazott - módszereket és eljárásokat ismertetem, amelyeket relevánsnak ítélt, és így az empirikus vizsgálatokban is helyet kaptak. Ezek a módszerek a kutatás gerincét képezik, és egy keretet adnak a soron következő elemzésekhez.

4.1. Klasszikus módszerek

A csőd- és fizetésképtelenség statisztikai alapokon nyugvó klasszikus módszerei régóta fontos szerepet játszanak a vállalatok pénzügyi egészségének értékelésében. E módszerek közül most megemlítem az úttörő módszernek számító diszkriminanciaanalízist, amely Altman híres Z-score modelljével vált híressé. A másik klasszikus, de rendkívül széles körben alkalmazott eljárás a logisztikus regresszió, amely máig használatos eszköz a csödelőrejelzésre a rugalmasságának és magas hatékonyságának köszönhetően.

4.1.1. Diszkriminanciaanalízis

A többváltozós diszkriminanciaanalízis módszere a megfigyeléseket előre meghatározott osztályokba sorolja be (Altman, 1968). A módszer megkeresi a jellemvonások olyan lineáris kombinációját, amely a legjobban elkülöníti a meghatározott csoportokat (Virág, 2001). Alkalmazása során azonosíthatjuk azokat a változókat, amelyek megkülönböztetik a megfigyelt osztályokat, azaz megvizsgálhatjuk, hogy a független változókkal becsülhető-e az osztályhoz való tartozás (Sajtos – Mitev, 2007). Az eljárás során új változók, úgynevezett diszkriminancia függvények jönnek létre úgy, hogy a lehető legnagyobb különbség legyen a csoportok között (Székelyi – Barna, 2008). A diszkriminancia függvény általános alakja a következő módon írható fel:

$$Z = c + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (1)$$

ahol Z a diszkriminancia-érték, amely által a besorolás megvalósítható, c a konstans, a β_i -k a diszkriminancia súlyok vagy együtthatók, az x_i -k pedig a független változók. A diszkriminancia függvény (Z) egy kritikus érték, amely elválasztja a csődös és nem csődös vállalatokat egymástól. A modell akkor válik előrejelzővé, ha az eredeti adatoktól eltérő, új mintát viszünk be (Virág et al., 2013a).

A diszkriminanciaanalízis számos feltételt támaszt az adatokkal kapcsolatban: a magyarázó változók normális eloszlása és függetlensége, valamint a független változók varianciájának a függő változók csoportjaiban megegyezőnek kell lennie (Ooghe et al., 1999; Virág – Kristóf, 2005). Továbbá a csoportok nagysága megközelítőleg azonos kell legyen. Ugyanakkor lehetőleg a teljes minta nagysága legalább tízszer nagyobb legyen, mint a független változók száma, valamint ajánlott, hogy a független változók száma kevesebb legyen, mint a legkisebb osztályban levő megfigyelések száma. Az osztályok teljesen ki kell zárják egymást, a függő változó egyes megfigyelései csak egyetlen osztályba lehetnek besorolva. A független változók között lineáris kapcsolatnak kell lennie, ennek vizsgálatára segítség lehet egy szóródási kép az egyes osztályokról. Multikollinearitás sem lehet, az eredményváltozók egymás közt nem függhetnek össze, csak a magyarázó változóval lehet összefüggés (Sajtos – Mitev, 2007).

A módszer egyszerű kezelhetősége mellett, a gyakorlati alkalmazása során a fizetésképtelenséget kutatók azzal szembesülnek, hogy a magyarázó változók, azaz pénzügyi mutatók között jellemzően nincs függetlenség, ezért multikollinearitás lép fel, ami nem teljesíti a módszer egyik feltételét. Ennek megoldása lehet a bemenő változók racionális csökkentése, összevonás és a szakértői felülvizsgálat (Altman, 1993).

4.1.2. Logisztikus regresszió

A logisztikus regresszió (vagy röviden logit) alkalmazása a csődelőrejelzés szakirodalmában nagyon elterjedt, már a csőd kutatás kezdeti szakaszában is egy igen gyakran használt módszer volt (Bellovary – Giacomino – Akers, 2007). Nem mellesleg a 2010-es évek után is nagy népszerűségnek örvendett a hitelkockázati modellezésnél (du Jardin, 2016). A kétértékű (dichotóm) logisztikus regresszió eredménye egy $[0, 1]$ intervallumban lévő érték, amely megmutatja, hogy egy adott megfigyelés milyen valószínűséggel tagja egy bizonyos osztálynak (de Menezes et al., 2017). A multinomiális logit esetén a függő változó kettőnél több értékkel rendelkezik. Ez utóbbi azonban kevésbé elterjedt a fizetésképtelenség becslésére, kivételes példákat felhozva Lau (1987) öt állapotú pénzügyi helyzetet modellezett, Jones és Hensher (2004) három állapotú osztályozási modellt fejlesztett.

A kétértékű logit modellnél jelöljük a fizetésképtelenséget $y=1$ -el az egészséges állapotot $y=0$ -val. A függő változó (y) valamely bekövetkezése nem becsülhető a független változókból (x) a legkisebb négyzetek módszerével, mivel a függő változó becslése nem mindig a $[0,1]$ intervallumban maradna (Kovács, 2014). Mivel a két

esemény együttes valószínűsége 100%, az egyik esemény valószínűségéből visszavezethető a másik esemény valószínűsége. A $Pr(y = 1|x)$ fizetési képesség valószínűségét jelöljük P_x -el, amelyet a x magyarázó változók valamilyen kombinációja okoz, a fizetőképességét pedig jelöljük $(1 - P_x)$ -el (Hajdu, 2004). Az esélyhányados (odds arány) alapján:

$$P_x = \frac{P_x/(1 - P_x)}{1 + P_x/(1 - P_x)} = \frac{odds_x}{1 + odds_x} \quad (2)$$

Az esélyhányados logaritmus:

$$\ln(odds_x) = \text{logit}(P_x) = \beta_0 + \sum_{i=1}^n (\beta_i * x_i) \quad (3)$$

ahol β_i regressziós együtthatók, x_i a bementi független változók. Másként felírva az esélyhányados:

$$odds_x = \frac{P_x}{(1 - P_x)} = e^{\beta_0 + \sum(\beta_i * x_i)} \quad (4)$$

A fentiek alapján a következő egyenlettel felírható a csőd bekövetkezésének becsült valószínűsége:

$$P_x = \frac{e^{\beta_0 + \sum(\beta_i * x_i)}}{1 + e^{\beta_0 + \sum(\beta_i * x_i)}} = \frac{1}{1 + e^{-(\beta_0 + \sum(\beta_i * x_i))}} \quad (5)$$

A paraméterek becsüléséhez a logisztikus regresszió a maximum likelihood becslést használja, ami olyan együtthatókat keres, melyek a modell legjobb illeszkedését biztosítják. A módszer több lépéses eljárás során határozza meg a legjobb paramétereket (Székelyi – Barna, 2008). A (4) egyenletben a független változók paramétereinek becsüléséhez a következő likelihood függvény írható fel:

$$L(\beta_0, \beta_1, \dots, \beta_n) = \prod_{i=1}^n (P_i)^{y_i} \cdot (1 - P_i)^{(1-y_i)} \quad (6)$$

A fenti egyenletbe a valószínűségeket behelyettesítve és egyszerűsítve, az egyenlet a következő formában írható fel:

$$L(\beta_0, \beta_1, \dots, \beta_n) = \prod_{i=1}^n \left(\frac{e^{\beta_0 + \sum \beta_i \cdot x_i}}{1 + e^{\beta_0 + \sum \beta_i \cdot x_i}} \right)^{y_i} \cdot \left(\frac{1}{1 + e^{\beta_0 + \sum \beta_i \cdot x_i}} \right)^{(1-y_i)} \quad (7)$$

Az úgynevezett Newton – Raphson iteratív folyamat során, az egyenletbe a paraméterek különböző értékkel helyettesítődnek be, aminek a végére az eljárás megtalálja a

likelihood maximális értéket (Kovács, 2014). A regressziós paraméterek értelmezéséhez tekintsük a következő egyenletet:

$$odds_{x_j+1} = odds_{x_j} \cdot e^{\beta_j} \quad (8)$$

ahol $e^{\beta_j} = \exp(\beta_j)$ tényező adja meg az x_j magyarázó változó egységnyi változásának az esélyhányadosra gyakorolt hatását, a többi független változó kontroll alatt tartásával (Hajdu, 2004).

A magyarázó változók paramétereinek ismeretében, a regressziós függvénybe behelyettesítve mindenik megfigyelésnek lesz egy valószínűségi értéke (0 és 1 között) a fizetőképességre vonatkozóan. A megfigyelések osztályozása úgy történik, hogy meghatározunk egy úgynevezett vágóértéket (cut of value) vagy kritikus értéket, esetünkben a fizetőképesség valószínűségére vonatkozóan, és ha a számított feltételes valószínűség meghaladja ezt, akkor az egyed fizetőképes, ellenkező esetben fizetéképtelen osztályba sorolódik. Alapesetben a vágópont értéke 0,5 szokott lenni, amitől el lehet térni lehetőleg úgy, hogy a téves besorolás minél kisebb legyen, valamint a rossz osztályozással járó veszteség minimális legyen (Hajdu, 2004).

A diszkriminanciaanalízissel szemben számos előnyös tulajdonsággal rendelkezik. A logit modell magyarázó változója lehet nominális, ordinális, intervallum- és arányskála (Kovács, 2014). Nem követeli meg a bemeneti változók többdimenziós normális eloszlását, és azonos variancia-kovariancia mátrixot (Virág – Kristóf, 2005). Az előfeltevései közt az sem szerepel, hogy a lineáris kapcsolat legyen a kimeneti érték és a független változók között. Azonban a kimeneti eredmény kétértékű (bináris) és egymást kizáró változó kell legyen, továbbá nem lehet korreláció a független változók között. Tehát a logit érzékeny a multikollinearitásra, ami akkor jelentkezik, ha a magyarázó változók együtt mozgása statisztikailag jelentős. A jelenséget nem csak egy változó, hanem a változók egy csoportja is okozhatja. A multikollinearitás számos negatív következményekkel jár úgy, mint a megbízhatatlan paraméter becslés, a magas standard hiba és a helytelen eredmény a magyarázó változók egyéni hatásának vizsgálatokor (Kovács, 2008). A multikollinearitás detektálására vannak bizonyos tesztek és mutatók, például a varianciainflációs tényező (VIF), ugyanakkor figyelemmel kell lennünk arra, hogy a tesztek elbírálása ne legyen önkényes (Burns – Burns, 2008). Továbbá a logisztikus regresszió jó teljesítéséhez nagy elemszámú minta szükséges. A kis mintaméret a statisztikai tesztek erejét rontja. Egyes szakértők szerint a logit módszer esetén az adatbázis minimális nagysága 500 kell legyen (Bujang et al., 2018). Más

vélemény alapján a minimális megfigyelés legalább 60 esetet kell tartalmazzon (Sajtos – Mitev, 2007). Peduzzi et al. (1996) különböző megfigyelés és független változó arányt vizsgáltak és azt találták, hogy minimum tízszer annyi megfigyelés kell legyen, mint a függő változók száma. A logisztikus regresszió nagyon érzékeny a kiugró (outlier) adatokra, ezért még a modellezés előtt kezelni kell a kiugróan magas vagy alacsony értékeket.

Az utóbbi évek empirikus kutatásai alátámasztják, hogy a logit módszer már nem a leghatékonyabb a hitelkockázat osztályozására (Kruppa et al., 2013; Barboza – Kimura – Altman, 2017; Nguyen – Viviani – Ben Jabeur, 2023), azonban továbbra is kiváló benchmark eszközként szolgálhat az újabb módszerek mellett.

4.2. Gépi tanuláson alapuló eljárások

A fent említett diszkriminanciaanalízist és logisztikus regressziót a gépi tanuláson alapuló algoritmusok között említhetjük, azonban nem soroljuk az intelligens mesterséges intelligencia módszerek közé. Gépi tanulási eljárások között inkább a nemparametrikus módszereket értem, és ilyenformán ismertetem a döntési fán alapuló algoritmusokat, az SVM módszert és a neurális hálózatokat. Általános meghatározást tekintve a gépi tanulás a mesterséges intelligencia azon részterülete, amely számítástechnikai alapokon nyugvó eljárások segítségével lehetővé teszi olyan elemző adatmodellek létrejöttét, amelyek automatikusan, rejtett információk feltárására alkalmasak anélkül, hogy egyértelműen kódolnánk őket (Wang, 2017).

A gépi tanulás története visszanyúlik 1943-ig amikor McCulloch és Pitts matematikai modelljével megalkotta a neuron alapjait. Később Alan Turing 1950-ben kiadott nagy hatású munkájában a mesterséges intelligencia koncepcióját alapozta meg. Nem sokra rá, 1952-ben Samuel sakkjátéka volt a gépi tanulás egyik első gyakorlati példája. 1958-ban Frank Rosenblatt megalkotta a perceptront, amely megalapozta a neuronháló alapú gépi tanulást és a következő évtizedben jelentős kutatási érdeklődést váltott ki (Russel – Norvig, 2005; Gupta et al., 2021). Az 1970-es években az egyrétegű perceptron sikertelenségeinek következtében csökkent az érdeklődés a mesterséges intelligencia és neurális hálózatok irányába, amelyet a mesterséges intelligencia első visszaeső ciklusának (AI Winter) emlegetnek (Wooldridge, 2021). Az 1980-as évek közepén ismét újjáéledt az érdeklődés, egyrészt a megnövekedett számítási teljesítménynek valamint a többrétegű neurális hálóknak köszönhetően, amelynek alapjait Rumelhart – Hinton – Williams (1986) fektette le (Russel – Norvig, 2005). A 80-

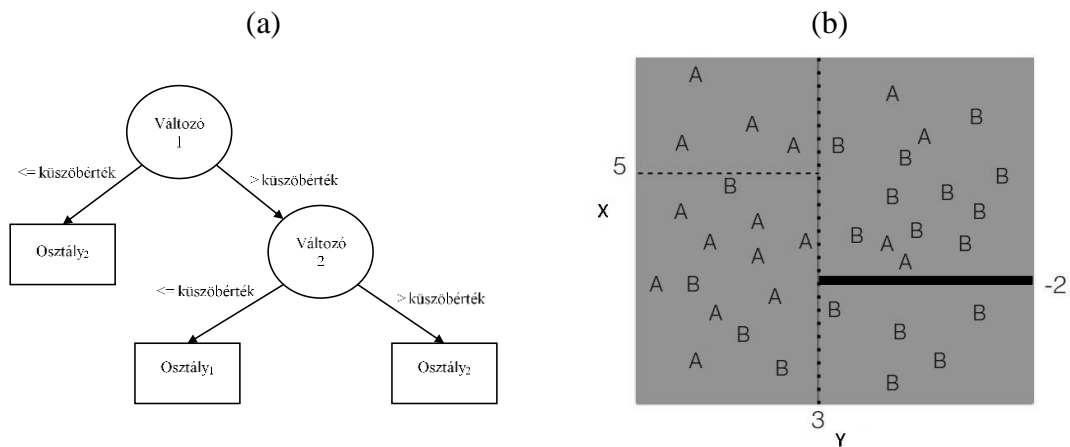
as évek végétől a 90-es évek közepéig egy második visszaeső ciklusát élte meg a mesterséges intelligencia kutatása (Wooldridge, 2021). Azt követően a több feldolgozó rétegű, mélytanuló neuronháló lett a kutatási érdeklődés egyik fő meghatározója, amely kitolta a mesterséges intelligencia módszerek teljesítményét számos tudományterületen (Schmidhuber, 2015; Wooldridge, 2021).

A gépi tanulás rengeteg probléma megoldására használható, amelyek közül csak az egyik az osztályozás. A gépi tanulást három fő típusba sorolhatjuk: felügyelt-, felügyelet nélküli- és megerősítéses tanulás. Felügyelt tanulásnál az algoritmus a bemeneti érték mellett kimeneti értékeket is kap. Ezek segítségével az osztályozó módszer tanul és felismeri az azonos jellemzőkkel bíró megfigyelések mintázatait valamint azonosítja azokat a jellemzőket, amelyek megkülönböztetik az eltérő csoportok megfigyeléseit (Barboza – Kimura – Altman, 2017). A felügyelet nélküli tanulásnál az algoritmus világosan kifejezett útmutatás nélkül fedezi fel a mintákat, struktúrákat és kapcsolatokat az adatokon belül (Gupta et al., 2021). A megerősítéses tanulásnál az algoritmus a környezettel való kölcsönhatáson keresztül, próbálgatások és hibákból tanulva az optimális eredményre összpontosít, a maximális jutalom érdekében (Gupta et al., 2023).

4.2.1. Döntési fa módszercsalád

A döntési fa módszer a rekurzív particionáló algoritmust alkalmazza, amely által klasszifikációs fákat állít elő egyszerű szabályok alapján. A fastruktúra létrehozása több lépéses folyamat, amelynek során a teljes mintát kétfelé osztja egyszerű szabályok alapján, hogy a lehető leghomogénebb osztályok jöjjenek létre. Az elválasztást úgy valósítja meg, hogy a függő változó szempontjából az elválasztott csoporton belüli variancia a lehető legkisebb legyen, míg a csoportok közti variancia a legnagyobb (Hámmori, 2001a). Addig állítja elő az elágazásokat, amíg van elválasztásra alkalmas változó. A fa tetején lévő első változó kiválasztása kulcsfontosságú. A módszer a magyarázó változók rangsorát elkészíti, külön-külön kipróbálja a változókat particionálásra, és miután mindenik elválasztás megtörtént, az lesz a fa csúcsán, amelyik a legkisebb hibát követi el. Ezt követően a többi változó ugyanezzel a módszerrel kerül kiválasztásra, amíg fel nem épül a teljes fa (Virág et al., 2013a).

2. ábra: A döntési fa általános alakja (a) és a rekurzív particionálás (b) két független változó két osztályú elválasztására



Forrás: (Virág et al., 2013a; Berk, 2020) alapján saját szerkesztés

A döntési fa grafikus ábrázolásával egy könnyen értelmezhető döntéshozatali rendszert kapunk (2. ábra, a-panel). Általánosan körökkel illusztrálják a változókat (csomópontok) és négyszögekkel a címkéket (levelek), amely a fizetéseképtelenség vizsgálata esetén két osztály lehet: fizetéseképtelen vagy egészséges. A körökből kiinduló nyilakon (a fa ágain) a kalkulált küszöbértékek vannak. Ha a változó értéke nagyobb, mint a küszöbérték akkor az egyik oldali ágra kerül, ha pedig kisebb vagy egyenlő, akkor a másik oldali ágra (Kristóf, 2008).

A 2. ábra (b) paneljén egy szórásdiagram segítségével a rekurzív particionálásra láthatunk egy példát, ahol A és B kimeneti osztályok vannak, amelyeket csupán x és y független változókkal magyarázunk. A függőleges $y = 3$ pontozott vonal az első particionálás, mivel a lehetséges elválasztások közül a legjobb homogenitást biztosítja, amely után az A-osztály a bal oldalon, míg a B-osztály a jobb oldalon koncentrálódik. A következő vágás a szaggatott vonallal jelölt $x = 5$, amely a baloldalon a legjobb elválasztás mivel a bal felső sarokban csak A osztály szerepel. A jobb oldalon a legjobb elválasztás $x = -2$ folytonos vastagított vonal a jobb alsó sarokban homogén B osztályt hoz létre. Így az $y \leq 3$ és $x > 5$ mindig A-osztályt eredményez. Hasonlóan $y > 3$ és $x \leq -2$ mindig B-t eredményez (Berk, 2020).

A feltételrendszere a statisztikai módszerekhez viszonyítva megengedőbb, nem igényli a független változók normál eloszlását és a csoportok azonos kovarianciáit sem

(Frydman – Altman – Kao, 1985). A döntési fa eljárása során felépített modellen statisztikai próbát, szignifikancia tesztet nem lehet végrehajtani. Egy másik korlátja, miszerint a felépült modellből nem lehet egyértelműen megállapítani, a független változók relatív hozzájárulását a fizetéseképtelenség kialakulásához (Virág et al., 2013a).

A módszer fejlődéstörténetében a fák létrehozásához számos algoritmust kifejlesztettek a kutatók. Az egyik legkorábbi a CHAID (Chi-squared Automatic Interaction Detector) döntési fa, amely chi-négyzet statisztikával végzi az optimális elosztást a csoportok között (Kass, 1980; Hámori, 2001a). A CART (Classification And Regression Trees) algoritmus folytonos célértékek becslésre, valamint osztályozásra alkalmas bináris fákat állít elő úgy, hogy a Gini-index segítségével határozza meg a legjobb vágást majd második lépésként metszi a gyengén teljesítő ágakat (Breiman – Friedman – Olshen, 1984). Az úgynevezett ID3 (Iterative Dichotomiser3) az információs előnyt (information gain) használja az elválasztás hasznosságának mérésére (Quinlan, 1986). Ennek újabb változatai jelentek meg, előbb Classifier 4.5 néven, amely a nyereségarányt (gain ratio) használja az optimális elválasztáshoz (Quinlan, 1993). Nem sokra rá, megjelent a továbbfejlesztett kereskedelmi változat C5.0 néven (Quinlan, 1997). Az OC1 (Oblique Classifier 1) algoritmus a CART-ra épülő döntési fákat Murty et al. (1993) fejlesztette ki, amely nem csak a tengelyekkel párhuzamos elválasztásra képes, hanem bármilyen irányú hipersík szóba jöhet.

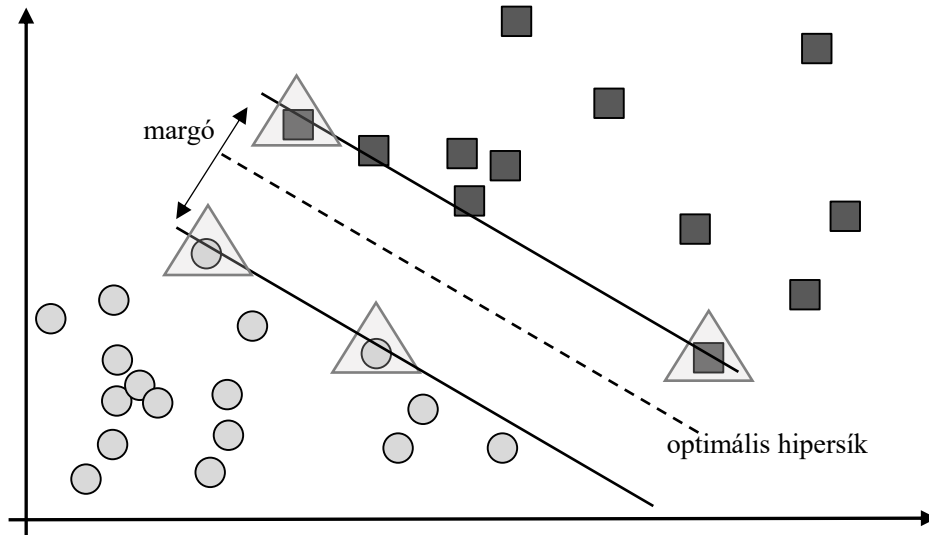
A döntési fa módszer családba sorolhatók azok az együttes (ensemble) módszerek, amelyek döntési fákon alapulnak: a bagging (zsákolás), a boosting (gyorsítás) és a random forest (véletlen erdő) metamódszerek. Ezeket külön módszereknek tekintem és a fejezet későbbi részében részletesen ismertetem.

4.2.2. Support Vector Machine (SVM)

A Support Vector Machine általánosan kernelgépeknek nevezett tanulómódszert takar, amely a magyar terminológiában is főként SVM-ként használatos, azonban a „tartó vektor gépek” vagy „szupport vektorgépek” néven is ismert (Russell – Norvig, 2005). A módszer kifejlesztése Cortes – Vapnik (1995) szerzőpárhoz köthető. Az eljárás alkalmas lineáris és nem-lineáris elválasztásra. A 3. ábra szemlélteti a kétdimenziós lineáris szeparációt, amelyen látható, hogy a margót a hipersík alatt és felett párhuzamosan, a hipersíktól egyenlő távolságra elhelyezkedő egyenesek határozzák meg. Ezek a lehető legnagyobb távolságra vannak egymástól úgy, hogy a legkisebb hibával

határozzák meg az osztályok határát. A margóhoz legközelebb eső pontok a tartó vektorok, a 3. ábrán háromszöggel jelölt pontok.

3. ábra: Az SVM kétdimenziós szeparációja



Forrás: (Cortes – Vapnic, 1995, 275. p.) alapján saját szerkesztés

A módszer algoritmus, egy tanuló minta felhasználásával kalibrálja és létrehozza az osztályozási mechanizmust. Két jól elkülöníthető osztály lineáris szétválasztásához vegyük az (\mathbf{x}_i, y_i) ($i = 1, 2, \dots, N$) tanítópontokat, egységnyi skálázással az optimális hipersík $\mathbf{w}^T \mathbf{x} + b = 0$, amely megfelel a következő feltételeknek:

$$\mathbf{w}^T \mathbf{x}_i + b \geq +1, \quad \text{ha } y_i = +1 \quad (9)$$

$$\mathbf{w}^T \mathbf{x}_i + b \leq -1, \quad \text{ha } y_i = -1 \quad (10)$$

amely felírható

$$y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad (i = 1, 2, \dots, N) \quad (11)$$

ahol \mathbf{w} a hipersík súlyvektora, b az eltolásérték (bias), $\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \dots, \mathbf{x}_i^{(n)})^T$ a bemeneti vektor és $y_i \in \{1, -1\}$ a tanuló halmaz címkéi, azaz az ismert osztályok, valamint N a tanuló minta elemszáma. Ha \mathbf{x}_i megfigyelés a fizetőképes osztályból származik, akkor $y_i = 1$, amennyiben a fizetéképtelen osztályból akkor $y_i = -1$. Legyen d a margó, amely az optimális hipersík alatt és fölött párhuzamosan elhelyezkedő hipersíkok közti távolságot jelöli. Tehát az osztályozási feladat során a \mathbf{w} és b által meghatározott hipersíkot

keressük, amelynek távolsága a két osztálytól a legnagyobb, valamint a margója a lehető legszélesebb legyen.

A két hipersík, $\mathbf{w}^T \mathbf{x} + b = 1$ és $\mathbf{w}^T \mathbf{x} + b = -1$ közti maximális távolság $\frac{1}{\|\mathbf{w}\|} - \frac{-1}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$, így az optimális hipersík margója $\frac{\frac{2}{\|\mathbf{w}\|}}{2} = \frac{1}{\|\mathbf{w}\|}$, amely akkor lesz maximális ha $\|\mathbf{w}\|$ minimális értékű, formalizálva az optimalizációs probléma:

$$\text{Min}_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad (12)$$

amely megfelel a (II) feltételnek (Altrichter et al., 2006). Ennek a szélsőérték keresési problémának a feltételei egyenlőtlenségek formájában vannak felírva. Az optimalizálási problémánál egyensúlyt kell találni a margó maximalizálása és a tanulási hiba minimalizálása között. A feladat megoldását duális feladatra történő átírással kereshetjük. Ehhez Lagrange kritériumot írunk fel (primal problem) és bevezetjük az $\alpha_i \geq 0$ Lagrange multiplikátorokat:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^N \alpha_i [y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1] \quad (13)$$

A fenti Lagrange kritériumfelület nyeregpontját (saddle point) keressük, úgy hogy minimalizáljuk \mathbf{w} és b szerint, valamint maximalizáljuk α_i szerint. Az optimumra \mathbf{w} és b szerinti parciális deriválásából a következő egyenleteket kapjuk:

$$\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = 0 \quad \frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = 0 \quad (14)$$

amelyek arra utalnak, hogy

$$\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \quad \sum_{i=1}^N \alpha_i y_i = 0 \quad (15)$$

$$\alpha_i \geq 0$$

Ha a Lagrange multiplikátorokat ismerjük a feladat megoldható, hiszen \mathbf{w} ezektől, valamint a tanítópontok értékétől függ. Behelyettesítve L -be a (15)-öt kapjuk a másodlagos kvadratikus programozással megoldható feladatot:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^N \alpha_i y_i \mathbf{w}^T \mathbf{x}_i - b \sum_{i=1}^N \alpha_i y_i + \sum_{i=1}^N \alpha_i \quad (16)$$

amelyben a harmadik tag (14 és 15) alapján egyenlő nullával és így

$$\mathbf{w}^T \mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{w}^T \mathbf{x}_i = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \quad (17)$$

Az $L(\mathbf{w}, b, \boldsymbol{\alpha}) = Q(\boldsymbol{\alpha})$ duális optimalizálási feladat:

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \quad (18)$$

azzal a feltétellel, hogy

$$\sum_{i=1}^N \alpha_i y_i = 0 \quad (19)$$

$$\alpha_i \geq 0$$

Ezen alapelveken túl a valós problémáknál ritkán fordul elő az, hogy két osztály lineárisan szeparálható és a margón belül nincsenek megfigyelések. Ebből kifolyólag a (11) egyenlőtlenség nem áll fenn azoknál a pontoknál, amelyek a margón belül, de a „jó” oldalon helyezkednek el. Ezért egy úgynevezett gyengítő (slack, ξ) változó alkalmazásával a (11) összefüggés módosul és a biztonsági sávban lévő tanító megfigyelések gyengítve érvényesülnek (Altrichter et al., 2006). A minimalizálandó kifejezés pedig a (12)-höz viszonyítva kibővül egy második taggal (Min – Lee, 2005):

$$\text{Min } \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi_i \quad (20)$$

azzal a feltétellel, hogy

$$\begin{aligned} y_i(\mathbf{w}^T \mathbf{x}_i + b) &\geq 1 - \xi_i, \\ \xi_i &\geq 0, \quad (i = 1, 2, \dots, N) \end{aligned} \quad (21)$$

Ha $\xi_i = 0$ akkor a kezdeti feladathoz jutunk. Ha $1 \geq \xi_i > 0$ akkor az adott pont a margón belül helyezkedik el de még a helyes oldalon. Ha $\xi_i > 1$ akkor a megfigyelés az optimális hipersík ellenkező oldalán található. C a szabadon választható hiperparaméter, amely a büntetés költsége, vagy másként az adott vektor választási költsége. A magas dimenziószámú térben a tökéletes lineáris szeparáció nem minden esetben érhető el, ezért a C konstans érték megadásával bizonyos hibát megengedünk. Ha C nulla, a gyengítő változó nem érvényesül. Amennyiben C kis értékű, akkor a (20) egyenletben a második tag súlya kicsi lesz, ami azt eredményezi, hogy több megfigyelés kerülhet a margókon

belülre vagy az ellenkező oldalra, azaz a margó szélesebb sávú lehet és kevésbé büntetjük a hibát (Altrichter et al., 2006). Minél nagyobb C érték növeli a büntetést, azaz kevesebb megfigyelés helyezkedhet a margók közt vagy a hipersík ellenkező oldalán, amely által csökken a modell általánosíthatósága. Ugyanakkor megjelenik a túlilleszkedés problémája, amelyhez az alacsony osztályozási pontosság társul a tanuló mintától különböző, kívülálló mintán (Virág – Nyitrai, 2013b).

Mivel a valós osztályozási problémák legtöbbször lineárisan nem szeparálható, ezért egy $\varphi(\mathbf{x})$ nemlineáris függvény alkalmazása szükséges a bemeneti tér egy magasabb dimenziószámú leképezéséhez, ahol aztán elvégezhető a szeparáció (Li – Wang – He, 2013). Ez egyszerűen a lineáris megoldásból $\mathbf{x} \rightarrow \varphi(\mathbf{x})$ helyettesítéssel megoldható, amelyben az optimális hipersík $\mathbf{w}^T \varphi(\mathbf{x}) + b = 0$ (Altrichter et al., 2006).

A (20)-as minimalizálási feladat megoldásához szintén Lagrange multiplikátorok kellenek, mivel feltételes szélsőérték keresési problémát oldunk meg:

$$L(\mathbf{w}, \xi, \gamma, \alpha, b) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi_i - b \sum_{i=1}^N \alpha_i [y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^N \gamma_i \xi_i \quad (22)$$

A Lagrange kritériumot minimalizáljuk \mathbf{w} , b és ξ_i szerint, valamint maximalizáljuk α_i és γ_i szerint. Az utolsó tag, egy új tag, és azt fejezi ki, hogy az ξ_i értékek nagyon kis értékek legyenek, lehetőleg nullák. γ_i szintén Lagrange multiplikátorok.

A kernel trükk alapján a kernel függvény $K(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\varphi}(\mathbf{x}_i)^T \boldsymbol{\varphi}(\mathbf{x}_j)$ azzal a feltétellel, hogy $K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$. A kernel függvény a magasabb dimenziószámú térbe történő vetítést helyettesítő távolságfüggvény, előre definiált az algoritmusban, és megoldást ad az optimalizálási problémára. A szakirodalomban kernel trükk néven ismert technika lényege, hogy az egyes pontok magasabb dimenziószámú térbeli távolságát a magasabb dimenzióba történő vetítés nélkül kapjuk meg (Bodon – Buza, 2014). Különböző kernel függvények jöhetnek szóba, a gyakorlatban jellemző a lineáris-, polinomiális-, Gauss- és szigmoid típusú (tangens hiperbolikus) függvény (Altrichter et al., 2006; Sun – Li, 2012).

- Lineáris kernel függvény: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
- Polinomiális függvény: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + 1)^d$ ahol d a függvény foka.

- Gauss-féle radiális alapfüggvény (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right)$ ahol σ pozitív konstans.
- Tangens hiperbolikus (szigmoid) függvény: $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \mathbf{x}_i^T \mathbf{x}_j + \theta)$, ahol γ és θ olyan konstansok, amelyek magfüggvényt eredményeznek.

A kvadratikus programozással megoldható duál feladat a $\sum_{i=1}^N \alpha_i y_i = 0$ és $C \geq \alpha_i \geq 0$ feltételekkel a következő lesz:

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (23)$$

A megoldása után azok a tanuló pontok, amelyekhez a Lagrange multiplikátor (α_i) értéke nem egyenlő nullával tartó vektoroknak nevezzük. Mivel a (23) megoldása után jellemzően számos α_i egyenlő lesz nullával, ezért N-nél kevesebb tanuló pont fog részt venni a végső eredmény kialakításában. A végső SVM osztályozó nemlineáris szeparálásra a következő szignumfüggvénnyel fejezhető ki:

$$\text{sign} \left[\sum_{i=1}^N \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \right] \quad (24)$$

A SVM módszernél a felhasználó első lényeges döntése a helyes kernel függvény kiválasztása. A lineáris kernel nem biztosít jó előrejelzést, a lineárisan nem szétválasztható adathalmazon, azonban az előnye, hogy az eredmények könnyen értelmezhetőek. Az RBF függvény értelmezése és magyarázata nehéz, viszont empirikus eredmények alapján jobban teljesít a többi függvénytípusnál, és kiváló előrejelzést biztosít a nehezen osztályozható adatmintán (Min – Lee, 2005; Barboza – Kimura – Altman, 2017). További paraméteroptimalizálási lehetőség a C „büntető” konstans megválasztása és a kernel γ paraméter megadása, azaz a sugárfüggvény sávszélessége.

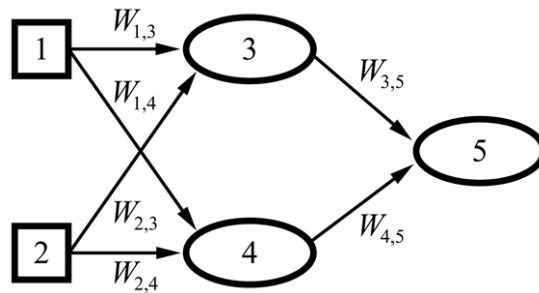
Az SVM módszer a mesterséges intelligenciára jellemző előnyökkel és hátrányokkal rendelkezik. A módszer egyik hátránya, hogy nagy adathalmaz esetén jelentősen megnövekszik a megoldási idő. Másik gyengesége a fekete doboz jelleg, például a csődmodelleknél nem kapunk választ arra, hogy mely pénzügyi mutatók milyen módon befolyásolják a modellt (Marques – Garcia – Sanchez, 2012). Előnyeként hozható fel, hogy a változók száma nem befolyásolja az optimalizálási eljárást (Trustoff – Konrad – Leker, 2011).

4.2.3. Neurális hálózatok

A neurális hálózatok, a mesterséges intelligencia fő képviselői, amelyek alkalmazása a számítástechnika gyors fejlődésének köszönhetően jelentősen elterjedt. Felépítésük az emberi idegrendszer működésén alapszik, éppen ezért számos területen alkalmazható összetett problémák megoldására, például adatbányászat, jelfeldolgozás vagy karakter felismerés. A neurális háló matematikai modelljét McCulloch – Pitts (1943) dolgozta ki, alapul véve a biológiai neuron felépítését. Rosenblatt (1958) kifejlesztette az egyrétegű neuronhálót a perceptron algoritmussal, amely gyakorlatilag meghatározta a mesterséges neuronháló alkalmazásának kiindulópontját.

A neurális háló alapegysége a neuron, amely a műveleti vagy feldolgozó komponensnek tekinthető, és amely bemeneti illetve kimeneti oldalon kapcsolatban áll más neuronokkal, így létrehozva a hálózatot. A 4. ábrán egy egyszerű előrecsatolt (feed forward) rétegzett neuronháló látható. A bemeneti rétegben két neuron található (1 és 2) amelyek egyenrangúak és memória nélküliek, nem végeznek műveletet csak a magyarázó változókat testesítik meg. A neuronháló egyetlen rejtett réteggel rendelkezik, ami két neuronból áll (3 és 4), és egy kimeneti neuron (5) látható.

4. ábra: Egy egyszerű előrecsatolt neurális háló felépítése



Forrás: (Russel – Norvig, 2005, 652. p)

Mindenik szomszédos neuron összeköttetésben áll egymással, és mindenik összeköttetéshez egy súly tartozik (W), amely meghatározza a kapcsolat előjelét és erősségét. A súlyok a tanulási folyamat során kezdetben véletlenszerűek majd egy leállási feltétel teljesülésig változnak. Az előrecsatolt neuronhálónál az információ terjedése csak egy irányú, a bemeneti neuronoktól a kimeneti neuronok felé. A visszacsatolt hálózatoknál (recurrent networks) a hálózat topológiájában hurkot találhatunk. A neurális

hálózatok működése két mozzanatra bontható. Először van a tanulási fázis melyben kialakulnak a hálózat súlyai, és ez jellemzően az időigényesebb mozzanat. A második fázisban az előhívás történik, a korábban kialakított háló használatával (Altrichter et al., 2006). A tanulási fázisban a hibavisszaterjesztés (backpropagation) által az észlelt hibákat visszafele haladva a belső kapcsolatok erősségét módosítja, a meglévő súlyokat figyelembe véve úgy, hogy a tanítási hiba csökkenjen (Russel – Norvig, 2005).

A szakirodalmi összefoglalóban láthattuk, hogy a neuronhálók alkalmazása, a fizetésképtelenség előrejelzésére nagyon népszerű lett már az 1990-es években, és egészen napjainkig kedvelt módszer a kutatók körében (Balcaen – Ooghe, 2004). A módszer számos előnyös tulajdonsággal rendelkezik. Az egyik figyelemreméltó jellemzője, hogy kis elemszámú minta esetén is hatékonyan alkalmazható (Cybinski, 2000). A közgazdasági alkalmazásban a leggyakoribb a háromrétegű előrecsatolt perceptron (Uliha, 2015), amely az egyik legkézenfekvőbb a fizetésképtelenség vizsgálatára is. A bemeneti rétegben az input adatok, azaz a pénzügyi mutatók és más fizetésképtelenséggel összefüggésbe hozható mutatószámok kerülnek. A kimeneti rétegben csak egy egység neuron szükséges, ez a függő változó előrejelzése, azaz a vállalatok fizetőképes vagy fizetésképtelen osztálya. Az eljárás során a bemeneti rétegekből az információ a rejtett rétegekbe kerül, ahol szabadon megválasztott számú neuron található (Uliha, 2015). A rejtett köztes rétegek neuronjai egy-egy függvényt jelölnek, ami a bemeneti rétegből kapott, N dimenziós input értékeket átalakítja:

$$S_i = f \left(\sum_{j=0}^N w_{i,j} x_j \right) = \mathbf{w}^T \mathbf{x} \quad (25)$$

ahol S_i a rejtett rétegben található i -edik neuron által adott függvényérték, $w_{i,j}$ az input változók i -edik neuronhoz tartozó súlyai vagy koefficiensei, x_j a megfigyelhető értéket hordozó input, c_j konstans tag a neuronhoz tartozó torzítás. A $w_{0,j}$ az eltolássúly, az x_0 a rögzített bemenet. $f(\cdot)$ a köztes rétegben alkalmazott aktivizációs függvényforma, amely a leggyakrabban:

- rectified linear (ReLU):

$$f(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases} \quad (26)$$

- tangens hiperbolikus függvény (Tanh):

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (27)$$

- szigmoid vagy logisztikus függvény:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (28)$$

A ReLu a szakaszonként lineáris függvények közé tartozik, kifejezetten a mély hálózatokhoz alkalmas függvény, annak is a Leaky ReLu vagy PReLu változata kiküszöböli az inaktívra váló neuronok okozta hibákat. A Tanh nemlineáris függvény nulla közepű [-1,1] tartományban képi le a valós értékű bemeneteket. A szigmoid függvény 0 és 1 közötti értékeket ad outputként ezért bináris klasszifikációs probléma megoldására kifejezetten hasznos. Az egyszerű lépcsősfüggvény, amelynek kimenetele -1 és 1 értékek egyikét veheti fel, a korai neurális hálózatokban volt elterjedt, azonban a mélytanuló neurális hálók építésére már nem alkalmas.

A rejtett réteg J darab neuronja J darab output értéket ad eredményül. A rejtett réteg neuronjainak értéke inputként szolgál a kimeneti réteg neuronjának, ahol egy g függvény átranszformálja egyetlen számmá, a függő változó előrejelzésévé (Uliha, 2015).

$$y_i = g \left(\sum_{k=1}^J \mathbf{w}_k S_{k,i} \right) \quad (29)$$

ahol \mathbf{W}_k a rejtett réteg k -adik neuronjából érkező input súlyvektorát jelöli, g a kimeneti függvény. A csödelőrejelzésre általánosan alkalmazott a kimeneti rétegben a normalizált exponenciális aktivizációs függvény (softmax), amely 0 és 1 közötti értéket ad vissza, azért a kimeneti érték valószínűségeloszlásként értelmezhető és a következő módon írható fel:

$$Softmax(\mathbf{z}_i) = \frac{e^{z_i}}{\sum_{j=1}^L e^{z_j}} \quad (30)$$

ahol \mathbf{z} a bemeneti vektor, L a függő változó osztályainak a száma, ami a bináris csödelőrejelzés esetében $L = 2$, egészséges vagy fizetésképtelen.

A köztes, feldolgozó rétegek számát, illetve a rétegekben található csomópontok számát tekintve, nincs egységes szabály és a szakirodalomban sincs egyetértés (Lam, 2004). Azonban vannak olyan gyakorlati lépések, amelyek segítségével a hálózat méretét optimalizálhatjuk (Altrichter et al. 2006). Ha egyre több rejtett réteget használunk, akkor egyre nő annak az esélye, hogy túltanulás lépjen fel, azaz a modell a megfigyelt mintára specializálódik és külső megfigyelések tesztelésére vagy előrejelzésre nem lesz alkalmas

(Balcaen – Ooghe, 2004). A tútanulás ellen az egyik megoldása lehet, hogy a modellt addig kell hagyni tanulni, amíg a tesztelő és tanulási minta hibája közel van egymáshoz. Amikor a tesztelő minta hibája romlani kezd, abba kell hagyni a tanulást (Virág et al., 2013a). Továbbá a rejtett neuronok számának meghatározására sincs szabály, többnyire esetleges a megállapításuk. Az egyre több rejtett neuron számításigénye magas, azonban a kevés rejtett neuron nem képes megfelelő pontosságot elérni (Zhao et al., 2015).

A módszer erősségeként említhető, hogy az univerzális approximátor tulajdonságának köszönhetően, bármilyen típusú függvény utánzására képes, magukból az adatokból ismeri fel a kapcsolatok természetét (Virág – Kristóf, 2005). A neurális hálók fő gyengesége, hogy „fekete dobozként” működnek. A gépi tanulási módszerek között a neuronháló az egyik, amelynél a legnehezebb megállapítani, hogy mely bemeneti tényezők mekkora súllyal befolyásolják a végső outputot, a fizetésképtelenségi modelleknél mely pénzügyi mutatók mekkora súllyal befolyásolják az osztályokba való besorolást (Balcean – Ooghe, 2004).

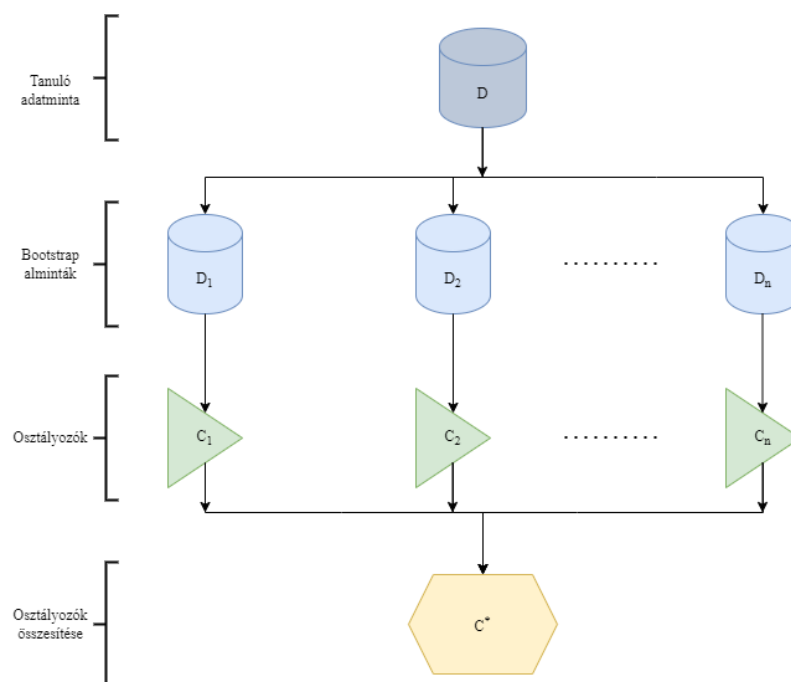
A neurális háló „fekete doboz” jellegének feloldására az egyik lehetőség a változók relatív fontosságának kiszámítása Garson (1991) ajánlása alapján. Az eljárás, kizárólag az egy rejtett réteget tartalmazó neuronhálónál alkalmazható, ezért a korlátai miatt érdemes megfontolni más fontosságmutatókat is (Olden et al., 2004). Szóba jöhetnek a lokális magyarázhatóságra alkalmas LIME (Local Interpretable Model-Agnostic Explanations) és a SHAP (SHapley Additive exPlanations) eljárások. A LIME, helyettesítő, könnyen értelmezhető modelleket hoz létre, amely az eredeti modell viselkedését leírja, ezáltal képes megmagyarázni az input változó előrejelzését (Riberio – Singh – Guestrin, 2016). A SHAP a Shapley-értéken alapul és a bemeneti változók marginális hozzájárulását méri (Lundberg – Lee, 2017; Bhattacharya, 2022; Cho – Shin, 2023).

4.3.3. Bagging

A bagging vagy „zsákolás” együttes eljárás, a „dinamikus” döntési fák módszercsaládba sorolható algoritmus, amennyiben az alaptanuló algoritmus homogén döntési fa. A döntési fán kívül egyéb alaptanuló algoritmus is szóba jöhet például logisztikus regresszió, neurális háló vagy SVM (Wang et al., 2011). Az együttes osztályozókban rejlő előny pedig az, hogy a teljesítményük jellemzően magasabb, mint azok az egyéni osztályozók teljesítménye, amelyek alkotják (Dietterich, 1997).

A bagging név a „bootstrap aggregálás” technika rövidítése, amely egy bizonyos független osztályozó eljárást foglal magába. A teljes adatmintából több különböző, véletlenszerűen kiválasztott almintát használ fel, visszatevéses eljárással, osztályozó eljárást alkalmaz rajta, majd végül a több különböző eredményt átlagolja. Ez pedig általában hatékonyabb osztályozáshoz vezet, mint egy egyedülálló modell (Bodon – Buza, 2014). A létrehozott alminták bootstrap másolatok, amelyekre jellemző lehet, hogy az eredeti mintából egyes elemek többször előfordulhatnak, mások meg egyáltalán nem jelennek meg. Így átlagosan az eredeti tanulóadat 63,2%-a használódik a bootstrap alminták létrehozásánál. Jellemzően elegendő 25, 50, legfeljebb 100 bootstrap másolat (Breiman, 1996).

5. ábra: A bagging eljárás algoritmusá



Forrás: (Yang et al., 2019) alapján saját szerkesztés

A bagging célja az, hogy csökkentse a magyarázó változók varianciáját és egyetlen osztály túlzott illeszkedését úgy, hogy a tanuló minta újrakombinálásával állítja elő a jobb osztályozókat (Barboza – Kimura – Altman, 2017). A technika előnye, hogy jóval alacsonyabb előrejelzési kockázatot hordoz, mint egyetlen alaptanuló eljárás

becslése (Bauer – Kohavi, 1999; Quinlan, 1996). A Breiman (1996) által kifejlesztett bagging algoritmust az 5. ábrán láthatjuk és a következő lépésekkel foglalhatjuk össze.

1. A tanuló mintából n darab véletlenszerű bootstrap alminta létrehozása (D_1, D_2, \dots, D_n).
2. Az 1. lépésnél létrehozott D_i almintákon egy-egy osztályozó eljárás (C_1, C_2, \dots, C_n) lefuttatása.
3. Az 1. és 2. lépések ismétlődnek $t = 1$ -től n -ig.
4. Végül az osztályozók eredményeit átlagoljuk a következő módon:

$$C^*(x) = \frac{1}{n} \sum_{i=1}^n C_i(\mathbf{x}) \quad (31)$$

ahol \mathbf{x} az egyes elemek adatai a tanuló mintából. Az utolsó lépésben az az osztályozó kerül kiválasztásra, amely a legtöbb szavazatot kapja, és az lesz a teljes adatminta osztályozója.

A módszer hátránya az eddig említett eljárásokhoz hasonlóan a „fekete doboz” jelleg, tehát nem kapunk világos választ arra, hogy mely mutatók milyen módon befolyásolják a modellt. Ennek feloldására regressziós fáknál a hiba eltérésnégyzet összeg (Residual Sum of Square, RSS), osztályozási fáknál a Gini-index segítségével végezhetjük a magyarázó változók fontosságmérését (James et al., 2013).

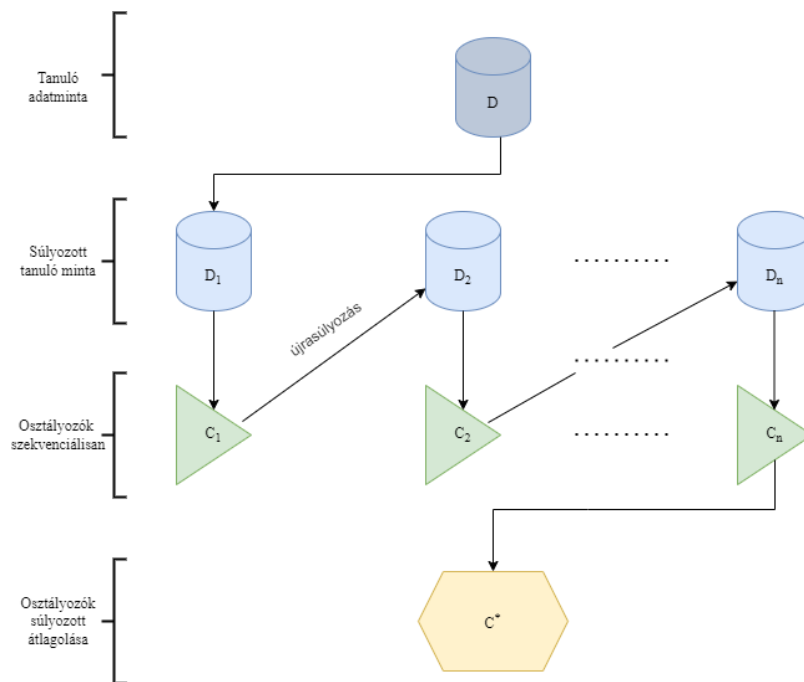
4.3.4. Boosting

A boosting alapjait Schapire (1990) fektette le, amely mentén Freund és Shapire (1996) dolgozta ki az Adaboost (Adaptive Boost) algoritmust, amelyet a következőkben ismertetek. Az alapgondolata hasonló a bagging eljáráshoz, hogy gyenge (weak) tanító eljárásokat használ erős tanító létrehozásához. Az adaptív elnevezés onnan ered, hogy a súlyozási paramétereket iteratív módon, az osztályozási teljesítmény alapján adaptív módon változtatja. Az osztályozóknak azonban bizonyos paramétereit rendszerint nem változtatjuk, például döntési fáknál azonos metszési módszert és vágási függvényt használunk (Bodon – Buza, 2014).

A boosting-gal a magyar terminológiában turbózás, fokozás vagy gyorsítás néven találkozhatunk (Russel – Norvig, 2005). A boosting a tanítóhalmazhoz hozzárendel egy $\omega_i \geq 0$ súlyt, amely az adott minta fontosságát reprezentálja. Az egyes osztályozások után a tanuló minta súlya változik az osztályozási teljesítménynek megfelelően. A következő osztályozó eljárás ezután az újrásúlyozott tanuló minta felhasználásával épül fel (Kim – Upneja, 2014). Hasonlóan, mint a bagging, a boosting is számos statisztikai

módszert és gépi tanuláson alapuló eljárást alkalmazhat osztályozásra, azonban gyakorta egy vágású döntési fa (döntési tönk, decision stump) alaptanulóval épül fel (James et al., 2013). Az eljárás strukturális felépítése a 6. ábrán látható.

6. ábra: A boosting eljárás algoritmusja



Forrás: (Yang et al., 2019) alapján saját szerkesztés

Az algoritmus kezdetén minden megfigyelésnek azonos súlya van $\omega_1 = 1/m$, ahol m a tanító halmaz elemeinek a száma. A D adathalmazt a kezdeti súlyokat felhasználva D_1 súlyozott mintán egy C_1 osztályozót képzünk. Ezt követően kiszámoljuk a hibaarányt (ε_1), majd az egyes megfigyelések újrásúlyozása történik a C_1 osztályozó hatékonysága alapján. Amennyiben az osztályozó teljesítménye kielégítő, a megfigyelések súlyai az osztályozó fontosságának (α) megfelelően változnak, a helyesen osztályozottak súlya csökken, a helytelenül osztályozottaké növekszik, azzal a reménnyel, hogy az új osztályozó helyesen fogja osztályozni a korábban hibásan szeparált megfigyelést (Bodon – Buza, 2014; Barboza – Kimura – Altman, 2017). A következőkben röviden áttekintjük az Adaboost algoritmusát Altrichter et al. (2006) és Heo – Yang (2014) alapján. A C_i osztályozó hibaaránya a következő módon formalizálható:

$$\varepsilon_i = \sum_{j=1}^m \omega_j I(C_i(\mathbf{x}_j) \neq y_j) \quad (32)$$

ahol m a tanítópontok száma, I egyenlő 1 ha a $C_i(\mathbf{x}_j) \neq y_j$ kifejezés igaz, vagy 0 ha hamis. \mathbf{x}_j a bemeneti magyarázó változó, y_j az osztálycímke amely a csődös vállalatok esetén -1, az egészségeseknél +1 értéket kap. Legyen a C_i osztályozó fontossági mutatója α_i , amely a súlyozásért felel és tulajdonképpen a hibaaránytól függ:

$$\alpha_i = \frac{1}{2} \ln \left(\frac{1 - \varepsilon_i}{\varepsilon_i} \right) \quad (33)$$

Ha az osztályozás a véletlen találgatásnak felel meg, azaz $\varepsilon_i = 0,5$, akkor az osztályozó fontossága $\alpha_i = 0$. Egyre kisebb hiba esetén a fontosságmutató növekszik, 0,5-nél nagyobb hibát nem fogadunk el. A $\omega_{i,t}$ jelölje az (x_i, y_i) tanítópontnak a t -edik gyorsítási menet során hozzárendelt súlyát, ahol $t=1, \dots, T$, amelyben T szabadon megválasztott érték, például 1000. A súlyok módosítása az alábbiak szerint történik:

$$\omega_{i,t+1} = \frac{\omega_{i,t}}{\sum_i \omega_{i,t+1}} \begin{cases} e^{-\alpha_t}, & \text{ha } C_t(\mathbf{x}_i) = y_i \\ e^{\alpha_t}, & \text{ha } C_t(\mathbf{x}_i) \neq y_i \end{cases} \quad (34)$$

ahol a nevezőben található $\sum_i \omega_{i,t+1}$ normalizáló tényező arra szolgál, hogy a súlyok összege az újrásúlyozás után is egységnyi legyen.

A végső osztályozó a következőképpen adható meg:

$$C^*(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t C_t(\mathbf{x}) \right) \quad (35)$$

ahol a függvény értéke +1 ha egészséges a vállalat és -1 ha fizetésektelen. A bemutatott módszer újrásúlyozás alapú (boosting by reweighting), amely a teljes tanulóadatot használja. Mindez olyan esetben alkalmazható, amikor az alaptanuló eljárás képes a minta súlyait beépíteni a tanulási eljárásba. A másik típus az újramintavételezéses gyorsítás (boosting by resampling) amely az eloszlási súlyok alapján véletlen visszatevéses eljárással az osztályozási ciklus után újabb tanulóadatot készít. Az utóbbi előnye, hogy bármely alaptanuló eljárásba illeszthető és a gyakorlatban általában jobban teljesít, mint a súlyozáson alapuló eljárás (Botta, 2001; Seiffert et al., 2008).

Mint ahogy a gépi tanuláson alapuló algoritmusok többsége, az Adaboost is fekete doboz jellegű előrejelző rendszer. Az eljárásnál az egyik fő hiperparaméter a tanulási

ciklusok száma (T), valamint a döntési fa jellegének köszönhetően a felhasználó feladata definiálni az döntési fák és az elágazási csomópontok számát. A tanulási ráta (learning rate) hiperparaméter egy pozitív, nulla és egy közötti érték, jellemzően 0,01 vagy 0,001 amely végső osztályozási szabály meghatározásához szükséges. Ez a konstans nagyobb jelentőséget ad a kisebb hibát generáló osztályozóknak (James et al., 2013).

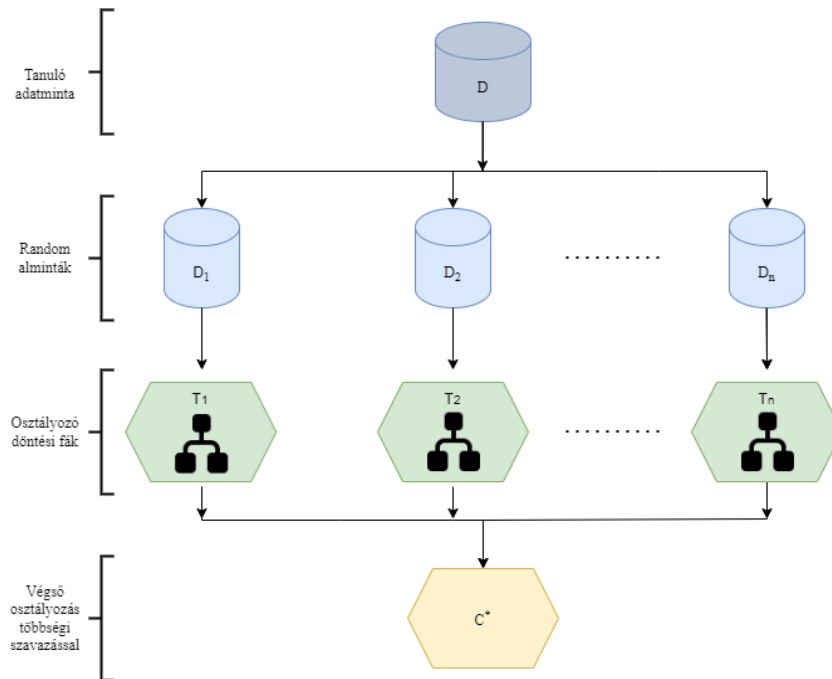
Az Adaboost módosított változata a gradiens boosting (Gradient Boosting Machine, GBM), amely alkalmas klasszifikációs és regressziós feladatok elvégzésére. Ebben a kontextusban a gradiens az elvárt hibától való eltérést jelenti. A GBM koncepcióját először Friedman (2001) ismertette. Az iteratív eljárás során a veszteségfüggvény minimalizálására törekszik, amely előzőleg tetszőlegesen megválasztható. Az osztályozók nem a hibásan osztályozott megfigyelések újrasúlyozásával, hanem az előző osztályozásban magas hibával számolt megfigyeléseket azonosítja és a szükséges elmozdulási iránnyal korrigálja az alaptanuló függvényt (Binder et al., 2014).

4.3.5. Random forest

A random forest a magyar terminológiában véletlen erdő eljárásként is ismert, kifejezetten a döntési fa osztályozóhoz tervezett együttes módszer. A módszer alapja a bináris fákat előállító klasszifikációs és regressziós fa (Classification and Regression Tree, CART). Több döntési fa előrejelzéseit átlagolja és az osztályba kerülést többségi szavazással dönti el (Breiman, 2001). A módszer a lehető legnagyobb metszés nélküli fákat hozza létre úgy, hogy mindenik elágazásnál a változók egy bizonyos halmazát használja fel, majd az elkészült nem korreláló fák eredményeit aggregálja (Figini – Savona – Vezzoli, 2016; Takács, 2021). A véletlen erdő a Bagging eljárashoz hasonló, azonban annak továbbfejlesztett változatának tekinthető. Az alaptanuló algoritmus kizárólag döntési fa, amely több különböző bootstrap almintán fut le, ezért általában a teljes minta egyharmada nem kerül felhasználásra a fák kiépítésekor (Yeh – Chi – Lin, 2014). A döntési fával létrehozott bagging-nél a fák erősen korreláltak lehetnek, tehát ezek átlagolásával sem fog lényegesen csökkenni a variancia. A random forest megoldja ezt a problémát azzal, hogy minden vágásnál a tanulóadatok egy részét használja, így korrelálatlan fákat hoz létre minden tanulási ciklusban (James et al., 2013). A kétszeres véletlenszerűsítés egyértelműen előnyösebb módszerré teszi a bagging módszernél. Először a megfigyelések véletlenszerű kiválasztása történik a döntési fa létrehozásakor, amely a bagging-nél is megjelenik. Majd a független változók köréből a döntési fa

mindenik elágazásánál véletlenszerű mintavételre kerül sor (Nyitrai, 2021), és a legjobb elválasztást biztosítót meghagyjuk. Az eljárás szerkezetét a 7. ábra szemlélteti.

7. ábra: A random forest eljárás algoritmusja



Forrás: (Yang et al., 2019) alapján saját szerkesztés

A véletlen erdő algoritmusát Tattar (2018) és Berk (2020) alapján a következő:

1. Vegyünk a tanuló adatokból N méretű véletlen részmintát visszatevéses módszerrel.
2. Vegyünk a magyarázó változókból m méretű véletlen részmintát visszatevés nélkül, úgy hogy $m < p$, ahol p az összes magyarázó változó száma. Osztályozási feladatokra a javasolt méret $m = \sqrt{p}$.
3. Hozzunk létre egy bináris particionálást a 2. lépésben kiválasztott független változókkal.
4. Ismételjük meg a 2. lépést mindenik elágazásánál, amíg a döntési fa a kívánt méretű nem lesz. Osztályozásra javasolt minimum csomópontok száma 1. Fontos, hogy metszés nélküli fát hozzunk létre.

5. A kihagyott adatokat (out of the bag, OOB) vigyük végig a fán, és tároljuk el a valós és az előrejelzett osztályt mindenik megfigyeléshez.
6. Ismételjük meg a 1-5 pontokat T -szer, ahol T elegendően nagy érték például 500 vagy 1000.
7. Csak a modellépítésre nem használt megfigyelésekre (OOB), számoljuk meg, hogy a fák kimenetele hány alkalommal jelezte előre az egyik és a másik osztályt.
8. Többségi szavazással döntsük el a végső osztályozást.
9. Az előrejelzéshez az új megfigyeléseket végigvisszük a fákon, amely után többségi szavazással eldöntjük a győztes módszert.

A módszer számos paraméter előzetes meghatározását igényli, amelyre vannak javaslatok, azonban a végső optimalizáció a kutató saját megítélésére van bízva. A hiperparaméterek optimalizációja magába foglalja a fák számát, az egyes fáknál a levelek méretét és a vágóváltozók számát (Loh, 2009; Nyitrai, 2021; Takács, 2021).

A módszernek számos kedvező jellemzője van, ilyen a kimagasló precizitás, de legalább olyan pontosságra képes, mint az Adaboost eljárás (Figini – Savona – Vezzoli, 2016). Továbbá azonosítja a kiugró adatokat, de nem érzékeny rájuk, és gyorsabb, mint a bagging illetve a boosting eljárás. A módszer előnyére szolgál, hogy azonosítja a változók fontosságát az előrejelzésben, és a döntési fa együttes módszerekhez mérten, mérsékli az adatok túlzott illeszkedését (Breiman, 2001). A „black-box” jelleg feloldása a bootstrap mintavételnek köszönhető, mivel az adatok csak egy részét használja fel, a megfigyelések körülbelül egyharmada véletlenszerűen kimarad a tanulófázisból. Ezeket a fák építésére fel nem használt „out-of-bag” (OOB) megfigyelések tesztelő készletként szolgálnak, amelyek segítségével kiszámolható a változó fontossága (Archer – Kimes, 2008). Első lépésként valamennyi fára vonatkozóan azonosítjuk az OOB megfigyeléseket, és mindenik fára az OOB megfigyelések becslését elvégezzük, valamint a helyes osztályozást feljegyezzük. A bemeneti változót az OOB megfigyelésekre vonatkozóan permutáljuk és így is készítünk becslést az eredeti fákkal. Ekkor megnézzük, hogy a helyes osztályozás aránya mennyivel változott. Az OOB megfigyelések helyes szavazatának, és a permutált OOB megfigyelések helyes szavazatának különbségeit átlagoljuk, így jutunk hozzá az egyes bemeneti változók fontosságértékeihez. Az a változó lesz fontosabb az osztályozás során, amelynek fontosságértéke magasabb. Azon változók, amelyek fontosságértéke nulla, nem járulnak hozzá a fák elágazásaihoz (Hastie et al., 2008).

5. Az előrejelző modellekhez kapcsolódó adatműveletek

A bevezetés, a szakirodalmi áttekintés és a módszerek ismertetése után rátérek az adatállományra és az azzal kapcsolatos kérdések tisztázására. A fizetéseképtelenség- illetve a csődelőrejelzéshez szükséges adatbázis összeállítása jellemzően munkaigényes, és kihívás a kutatók számára. Ez kifejezetten érvényesül a pénzügyi nehézségben lévő fizetéseképtelen (csődös) vállalatok esetén, mivel számos esetben már a csőd vagy más fizetéseképtelenségi eljárást megelőzően nem hoznak nyilvánosságra hiteles pénzügyi beszámolókat.

5.1. A felhasznált adatok

A Magyarországon nyilvánosan elérhető térítésmentes cégszűrés adatbázis az Igazságügyi Minisztérium Céginformációs és az Elektronikus Cégeljárásban Közreműködő Szolgálat⁵ adatbázisa, amely - a cégszűrés azonosító értékeinek megadása után - hozzáférést biztosít a keresett vállalat összes közzétett beszámolójához. A részletes adatok elérése mellett szembesülnünk kell azzal, hogy nincs mód az adatok tömeges letöltésére. Így aztán egy kellően nagy adatbázis egyesével történő adatrögzítése jelentős időráfordítást igényel. Az üzleti adatszolgáltatók⁶ adatbázisában történő díjmentes keresés a vállalatok alapadatait eredményezi, ha ennél többet szeretnénk megtudni, például a pénzügyi adatokat, akkor azt térítés ellenében tehetjük meg.

A tanulmány során felhasznált adatok több különböző forrásból származnak. Az empirikus kutatáshoz összeállított adatbázis kiinduló pontját a Pécsi Tudományegyetem Közgazdaságtudományi Kar tulajdonában álló vállalati adatbázis adta, amely többek közt a kutatáshoz lényeges pénzügyi adatokat is tartalmazta. Emellett az OPTEN, valamint az EMIS⁷ által szolgáltatott magyarországi vállalatok éves közzétett beszámolóit és nyilvános cégkivonat adatait bővítették az adatbázist.

Az adatbázis összeállításánál a fizetéseképtelen vállalatok kiválasztása a legösszetettebb feladat, amit a pénzügyi adatok hiányos rendelkezésre állása csak tovább fokoz. A fizetéseképtelenség meghatározása jogi megközelítés alapján történt, azaz a hivatalos felszámolási eljárások alatt lévő vállalatokat tekintetem fizetéseképtelennek. A csődeljárási eseteket kis számukra való tekintettel nem vettem figyelembe, valamint

⁵ Igazságügyi Minisztérium Céginformációs és az Elektronikus Cégeljárásban Közreműködő Szolgálat: e-beszamolo.im.gov.hu

⁶ Például az OPTEN Informatikai Kft.: opten.hu vagy a Céginformáció.hu Kft.: ceginformacio.hu

⁷ EMIS, ISI EMERGING MARKETS GROUP: emis.com

más fizetéseképtelenségre és lezárásra vonatkozó jogi eljárásokat is figyelmen kívül hagytam. Amennyiben a vizsgálati időszakban nem indult fizetéseképtelenségi eljárás egy bizonyos vállalat ellen, akkor „egészségesnek” azaz fizetőképesnek címkéztem. A fizetéseképtelen vállalatok alapadatainak a forrása a Cégek Közlöny⁸ volt, amely a cégnyilvánosságról és a bírósági cégeljárásokról nyújt közvetlen hozzáférést. A gazdasági ügyszak hirdetményei között időrendi sorrendben megtalálható a különböző eljárások alá került cégek megnevezése, cégjegyzékszám és az ügyel kapcsolatos hivatalos határozat. Ezen túlmenően az egészséges vállalatokat és a fizetéseképtelenségi eljárások tényét a Fizetéseképtelenségi Nyilvántartás⁹ elektronikus felületén ellenőriztem.

A lokális külső hatások mérésére a településeken érvényesülő gazdasági, önkormányzati, közlekedési adottságokat jellemző adatok szolgáltak. Ezek az Országos Területfejlesztési és Területrendezési Információs Rendszer (TEIR) adatbázisából származtak, amely mellett párhuzamosan a Központi Statisztikai Hivatal (KSH) adatait is felhasználtam. A tanulmányban a vizsgált változók és a felszámolás közötti pozitív, illetve negatív kapcsolatot/hatást/összefüggést matematikai értelemben használom az azonos, illetve az ellentétes irányú kapcsolat/hatás kifejezésére.

5.2. Az adatgyűjtés szempontjai

A rendelkezésre álló adatokból egy olyan minta összeállítása volt a cél, amely általánosan alkalmazható a magyar vállalatok fizetéseképtelenségének előrejelzésére, valamint felhasználható bizonyos módszertani kérdések megválaszolására. Ugyanakkor élve néhány korlátozással a következő keretek közt értelmezhetőek az eredmények.

Az ilyen kutatásoknál szokás szerint a fizetéseképtelen vállalatokból indulunk ki mivel azok jelentik a szűk keresztmetszetet. Számukat tekintve kevesebb van belőlük, és sok esetben hiányosak a hozzájuk köthető információk. Ennek mentén az egyik megkötésem az volt, hogy a fizetéseképtelen vállalatok esetén a fizetéseképtelenségi eljárást megelőző közvetlen első évből álljanak a rendelkezésre pénzügyi adatok. Ez azért lényeges, mert a fizetéseképtelenségre jellemző tulajdonságok rendszerint egyre inkább érvényüket veszítik, ahogy távolodunk a fizetéseképtelenség időpontjától.

A vállalati méret alapján a mikro, kis és közepes méretű vállalatokra korlátozódtam, gazdasági forma szerint csak a korlátolt felelősségű társaságokat

⁸ Igazságügyi Minisztérium Cégek Közlöny portál: cegkozlonny.hu

⁹ Igazságügyi Minisztérium Fizetéseképtelenségi Nyilvántartása: fizeteskeptelenseg.im.gov.hu

vizsgáltam, ezzel megtartva egy elég nagy merítési méretet. A foglalkoztatotti létszám tekintetében a mintába a minimum 2 és maximum 249 foglalkoztatott személlyel rendelkező vállalat kerülhetett be. További bekerülési kritérium volt, hogy a vállalat mérlegfőösszege és árbevétele legkevesebb 1-1 millió forint értékű legyen és legalább két lezárt üzleti évvel rendelkezzen. Csak az év végi, teljes üzleti évre vonatkozó adatokat vettem figyelembe, a részleges félévközi adatokat elvettem.

Az elemzett iparágak tekintetében nem volt kizárási feltétel, mindössze annyi, hogy kimaradtak a pénzügyi intézményekre vonatkozó, sajátos számviteli szabályokkal működő bankok és biztosítási társaságok, mivel a nem pénzügyi vállalatoktól eltérő mérleg- és eredménykimutatás-szerkezettel rendelkeznek. Az iparági elhatárolás érdekében a nemzetgazdasági ágakat ismertetem, amelyeket az elsődleges tevékenység TEÁOR¹⁰ száma alapján határoztam meg. Hozzá kell tenni, azt hogy ilyen tekintetben a vállalatokat nagyon erőltetetten lehet besorolni egyetlen iparágba, mivel jellemzően egyszerre több tevékenységet folytathatnak. Gyakran előfordul, hogy bizonyos másodlagos tevékenységek nagyobb árbevételű biztosítanak, mint a statisztikai főtevékenység, ezért az iparági besorolás félrevezető lehet (Deák, 2000). Mivel nem álltak rendelkezésre részletesebb adatok, ezért a statisztikai célra kijelölt főtevékenységet vettem alapul a vállalatok iparági besorolásánál.

5.3. A minta összeállítása

A mintavételezés a modellezési folyamatok egyik alapvető feladata. A jelentősége abban nyilvánul meg, hogy - figyelembe véve a potenciális módszerek korlátait - a populációból egy olyan megfelelő részhalmazt állítsunk elő, amely később következtetések levonására és előrejelzésre alkalmas a teljes sokaságra tekintve. A mintavételezés általános kérdésköréről és technikáiról részletes útmutatást ad Lázár (2022) munkája.

A minta összeállításánál a kutatási kérdéseknek megfelelően két - a csődelőrejelzésben általános - mintafelosztásra törekedtem a fizetésektelen és egészséges vállalatok csoportja között. Egyrészt a valós sokasági arányok alapján a reprezentativitást biztosító mintavételre törekedtem, amely statisztikai szempontból helyénvaló megközelítés. Az osztályok évek szerinti eloszlása az *1. táblázatban* látható. Mint ahogy már említettem a fizetésektelen adatbázis összeállítása az igazi kihívás az

¹⁰ Gazdasági tevékenységek egységes ágazati osztályozási rendszere, 2008

ilyen jellegű kutatásoknál. A fizetésektelen adatállomány összeállítása után az megoszlási reprezentatívitást (később az azonos létszámot) szem előtt tartva határoztam meg az egészségesek számát. Az egészségesek kiválogatása egy nagyobb adatállományból történt véletlenszerű eljárással. Ennek megfelelően az egyes évek összes megfigyelése 15-27 ezer közötti, a teljes reprezentatív adatállomány pedig több mint 100 ezer megfigyelésből áll. Az *1. táblázat* utolsó oszlopában a kft.-k felszámolási aránya egy kalkulált mutató, amely a KSH adatai alapján a regisztrált korlátolt felelősségű társaságok és a felszámolás alá került kft.-k számának a hányadosa (Központi Statisztikai Hivatal, 2014; 2015; 2016; 2017; 2018; 2019).

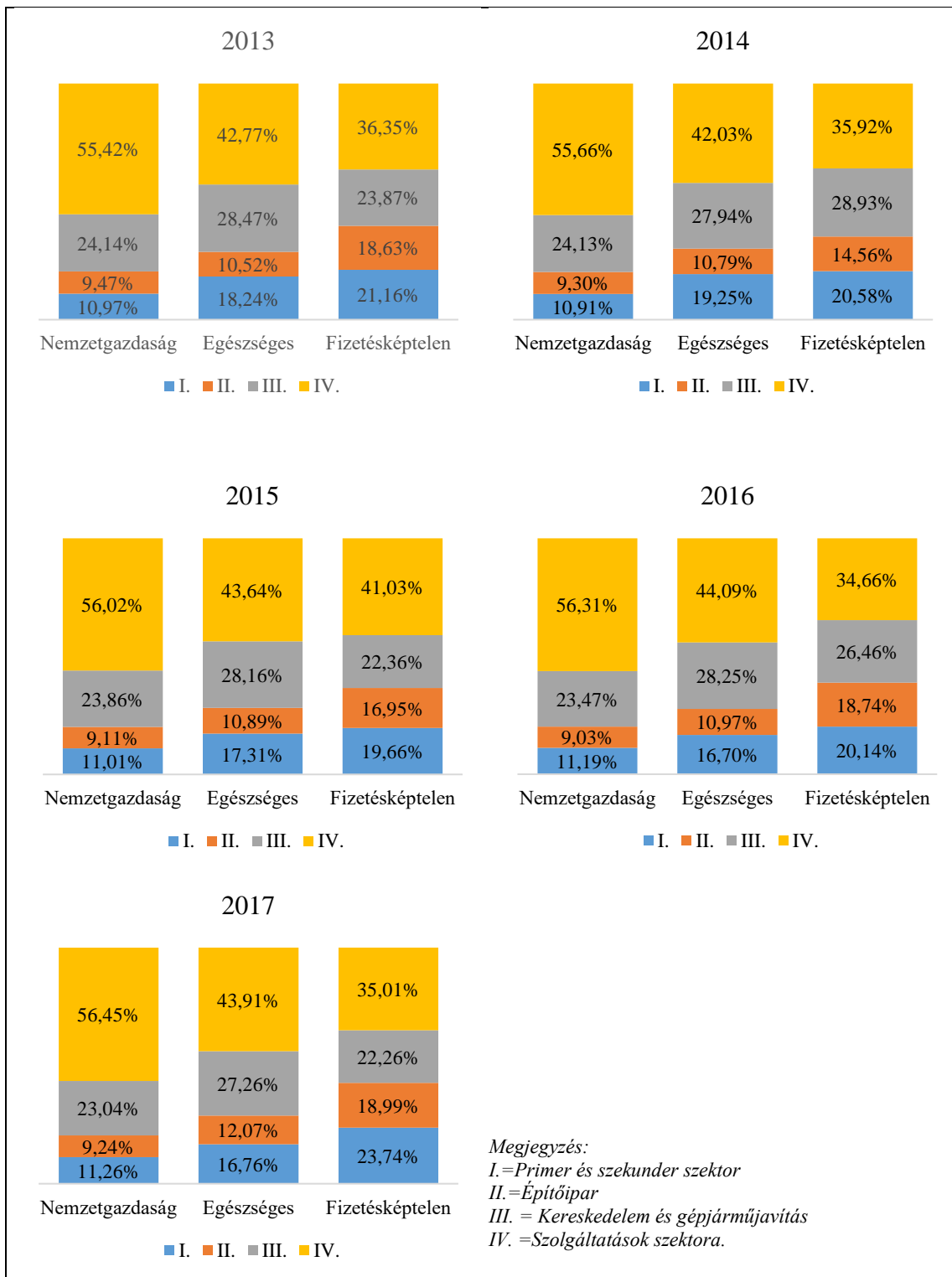
1. táblázat: A reprezentatív mintában szereplő megfigyelések száma

Évek	Fizetésektelen	Egészséges	Összesen	Kft.-k felszámolási aránya
2013	553	26 558	27 111	2,086%
2014	515	15 098	15 613	3,411%
2015	407	19 290	19 697	2,110%
2016	427	23 777	24 204	1,776%
2017	337	21 062	21 399	1,566%
Összesen	2 239	105 785	108 024	

Forrás: saját szerkesztés

A mintában négy gazdasági szektort különítettem el: I. a primer és szekunder szektort, II. az építőipar szektort, III. a kereskedelem és gépjárműjavítás szektorát, valamint IV. a szolgáltatások és egyéb tevékenységek szektorát. A *8. ábra* a szektorokban lévő vállalatok eloszlását tartalmazza a vizsgált években a teljes nemzetgazdaságban, valamint a mintában szereplő két osztályban. Ha tökéletesen nem is egyeznek az arányok, mindkét osztály eloszlása nagymértékben közelíti a teljes populáció eloszlását, azzal az észrevétellel, hogy az egészségesek eloszlása jobban igazodik a nemzetgazdasági megoszláshoz.

8. ábra: Az eloszlási arányok alapján reprezentatív minta megoszlása osztály és gazdasági szektor szerint



Forrás: saját szerkesztés

Alapvetően a csődelőrejelzést megvalósító tanulmányok nagyobb része az egyenlő osztálylétszámú mintavételt részesíti előnyben. Alaka et al. (2018) irodalomösszefoglaló tanulmányában rámutat, hogy csak néhány módszertani munka foglalkozik a kiegyensúlyozatlan osztálylétszám kérdéseivel. Zmijewski (1984) kifejti, hogy ha a csődös és egészséges vállalatok aránya nem reprezentálja a valós populációt, akkor torzítás léphet fel, azonban ez nem befolyásolja a végső statisztikai következtetéseket. Berg (2007) kiemeli a reprezentatív minta fontosságát és alkalmazza, azonban az eredmények értékelésénél csak a modellek átlagos teljesítményét közli, a csődös megfigyelések előrejelzési pontosságát viszont nem. McKee és Greenstein (2000) mintavétele a valós csődaránynak megfelelt, azonban az eredményei a fizetéképtelenek gyenge osztályozását mutatták. Oommen, Baise és Vogel (2011) szerint, ha a minta osztály szerinti eloszlása minél jobban közelít a valós eloszláshoz, annál pontosabb lesz a logisztikus regresszió becslése. Verganzoles és Severin (2018) az azonos osztálylétszámú mintavétel mellett kipróbált több arányú mintát különböző módszereken. Következtetéseik alapján, ha csődös vállalatok aránya 20% alatti, jelentősen csökken mindegyik modell előrejelző képessége a csődös vállalatok tekintetében. Ezek és a szakirodalmi ellentmondások adtak indítékot a reprezentatív minta megfontolására a saját csődmodell fejlesztésénél.

A reprezentatív adatminta alkalmazása kiegyensúlyozatlan osztályokat jelent, amely a modellezés során problémát okozhat. Ilyen helyzetben a klasszifikációs eljárások a teljes hiba minimalizálása érdekében a többségi osztályon kimagaslóan, a kisebbségi osztályon pedig rosszul teljesítenek. Ennek kezelésére a modellezés során gyakoriság alapú súlyokat állítottam be, amely által a tanulási fázisban nagyobb figyelmet kaptak a kisebbségi osztály megfigyelései úgy, hogy a bemeneti adatok eloszlásában nem történt változás. Az osztályszúlyok beállításához az algoritmus szintű kiigazítás a következő módon történt:

$$\omega_i = \frac{N}{2C_i} \quad (36)$$

ahol ω_i az i osztályhoz tartozó súlyérték, N az összes megfigyelés, C_i pedig az i osztályban lévő összes megfigyelés száma. Szokásos módon az osztályozási algoritmusok hibás osztályozási költsége azonos mértékű a függő változó mindenik kategóriájához (Thabtah et al., 2020). A (36) segítségével azonban az algoritmusok költségfüggvényébe az egyes osztályokhoz különböző súlyok rendelődnek hozzá, úgy hogy a kisebbségi

osztály hibája nagyobb hatással lesz a teljes osztályozási költségre, ezzel fokozva a modellek helyes besorolását a kisebbségi osztályban.

A reprezentatív eloszlású mintavétel után az azonos osztálylétszámú mintavételre összpontosítottam. Ha visszatekintünk a szakirodalmi feldolgozásra, akkor azt látjuk, hogy a csődelőrejelzésre hagyományosan az azonos osztálylétszám alkalmazása alakult ki (Beaver, 1966; Altman, 1968; Edmister, 1972; Gombola et al., 1987), ugyanakkor ez a koncepció a későbbi kutatásokban is fennmaradt (Lin, 2009; Barboza – Kimura – Altman, 2017; Brygała, 2022). A fizetésképtelen megfigyelések felülreprezentálásának praktikus okai vannak. Egyfelől a csődös vállalatok adatai nem mindig elérhetőek, így a szűkösen hozzáférhető adatokon a véletlen mintavételen alapuló technikák akadályba ütköznek, vagy torzításokat tartalmazhatnak. A rendelkezésre álló teljes adatbázis nagysága is meghatározza, mivel a kis mintás elemzéseknél a fizetésképtelen megfigyelések felülreprezentálása elkerülhetetlen (Altman – Haldeman – Narayanan, 1977; Platt – Platt – Pedersen, 1994). Másfelől a reprezentativitásra törekvő mintavételnél a felszámolásra került vállalatok aránya nagyon alacsony. Ez pedig a modellezésnél azt eredményezi, hogy a nagy számban rendelkezésre álló egészséges vállalatokra specializálódhat az előrejelzési algoritmus és ebből adódóan a fizetésképtelen megfigyeléseket gyakorlatilag nem képes helyesen megbecsülni (du Jardin, 2015). A neurális háló például nem tudja megkülönböztetni a fizetésképtelen eseteket, ha azok aránya a mintában 20% alatti (Boritz et al., 1995). Ilyen szempontokat figyelembe véve érdekesebb a mintát egyenlő arányban felosztani (Jo – Han – Lee, 1997).

Az azonos mintavétel teljesítményre gyakorolt hatásainak feltárása érdekében a mintát egyenlő arányban is felosztottam, amelyet a 2. táblázat mutat. A fizetésképtelen megfigyelések adott száma alapján az egészséges megfigyeléseket véletlenszerűen választottam ki. A fizetésképtelen megfigyelések megegyeznek a reprezentatív mintában szereplőkkel. A kezdeti években összesen ezret is meghaladó megfigyelés volt az elemzési adatbázisban, a 2018-as évtől kezdődően azonban kevesebb adat állt a rendelkezésre. Összesen több mint ötezer megfigyelést tartalmazott az osztály szerint kiegyensúlyozott adatbázis. A táblázatban látható adatállomány több célt szolgált, ezért egyes esetben ennek csak egy részhalmazát használtam fel. Ilyen módon a reprezentatív mintával való összeméréshez csak a 2013-2017 közötti éveket vettem figyelembe, azonos időszakokat, mint ami a reprezentatív adatállományban a rendelkezésre állt.

2. táblázat: Az azonos osztálylétszámú mintában szereplő megfigyelések száma

Évek	Fizetésképtelen	Fizetőképes	Összesen
2013	553	553	1106
2014	515	515	1030
2015	407	407	814
2016	427	427	854
2017	337	337	674
2018	75	75	150
2019	91	91	182
2020	52	52	104
2021	206	206	412
Összesen	2663	2663	5326

Forrás: saját szerkesztés

Az *F1. függelékben* található az azonos osztálylétszámú minta megoszlása osztály és gazdasági szektorok szerint az egyes vizsgálati években. A minta megoszlása relatíve jól illeszkedik a teljes populáció megoszlásához. Kevésbé egyező megoszlásokat a 2018-as évtől kezdődően tapasztalunk, ahol a minta mérete számottevően leszűkült. Így egyes éveknél megfigyelhető, hogy a mintában nem a szolgáltatások szektorában van a legtöbb vállalati megfigyelés, ahogy a kft.-k populációja mutatja, hanem a III.-al jelölt kereskedelem és gépjárműjavítás szektorban.

5.4. Magyarázó változók

A fizetésképtelenséget előrejelző modellek építőkövei a magyarázó változók, amelyek összetett kölcsönhatásainak feltárásával fontos betekintés nyerhetünk a fizetésképtelen és az egészséges vállalatok rejtett pénzügyi mintázataiba. Mindez pedig lehetőséget teremt az általánosításra és az egyedi vállalatok problémás területein történő beavatkozásra.

A modellezés során a függő változó a fizetésképtelenség ténye, amely a felszámolás jogi eljárás alapján került megállapításra. Ez a kétértékű változó 1-es értéket kapott, amennyiben a vállalat fizetésképtelen és 0-át, ha fizetőképes, azaz egészséges.

A csödelőrejelzés szakirodalmában a magyarázó változók köre hangsúlyosan a pénzügyi adatok felhasználásából adódik. A csödkutatások több mint 90%-ban a vállalatokhoz kapcsolódó pénzügyi bemeneti változók figyelhetők meg du Jardin (2009)

szerint. Ezzel egybevágó következtetést tudtam levonni a saját irodalomelemző kutatásban is (Ágoston, 2022a). A makrogazdasági változók alkalmazásának az aránya 3-5% közötti, amelynél némileg magasabb arányú a piaci alapú bementi változók megjelenése a kutatásokban¹¹ (du Jardin, 2009; Ágoston, 2022a).

A felhasznált független magyarázó változók körét a kutatási kérdéseknek megfelelően válogattam ki, így két fő csoportba sorolhatók. Az egyik csoportban vannak a vállalatok belső attribútumait leképező pénzügyi és méretre vonatkozó jellemzők. A másik csoportban vannak a külső környezeti tényezőket megvilágító mutatók, amelyek a vállalatok székhelyén érvényesülő mezoszintű magyarázó változók. A pénzügyi magyarázó változók kiválasztásában segítségemre voltak a magyar csődelőrejelzés szakirodalmában megalapozott mutatók (Virág, 2001; Virág – Kristóf, 2005; Nyitrai, 2014a; Virág – Nyitrai, 2014a). Ezen kívül Bellovary et al. (2007) irodalomösszefoglaló munkája, Altman (1968), Carton és Hofer (2006), du Jardin (2015), Linag et al. (2016), Zięba et al. (2016) és a saját empirikus tapasztalatok alapján jelöltem ki a magyarázó változókat (Ágoston, 2022b; Ágoston, 2023; Ágoston, 2024). A mezoszintű változók kiválasztását Buehler, Kaiser és Jaeger (2012) valamint Arcuri, Brunetto és Levratto (2019) munkája inspirálta.

A kiválasztott mutatókat, a jövedelmezőség, likviditás, eladósodottság, hatékonyság, eszköz- és tőkeszerkezeti, cash flow, méret és egyéb típusokba sorolva az *F2. függelékben* ismertetem, ahol a megnevezés mellett a számítási módjuk is megtalálható. Összesen ötven pénzügyi és egyéb, a vállalat belső információjához köthető változót vettem figyelembe (jelölésük X1-X50), emellett hét mezoszintű magyarázó változót választottam az elemzésekhez (jelölésük Y1-Y7). A pénzügyi magyarázó változók mindegyike az év végi záró mérleg- és eredménykimutatásból kalkulált hányados típusú arányszám. A vállalat belső jellemzőit leíró pénzügyi és méretre vonatkozó mutatókat nyolc típusba sorolhatjuk. A jövedelmezőséget tükröző mutatókból kilencet számítottam ki, köztük a ROA, a ROE, a ROS több változatát, a forgótőke arányos árbevételt és a bérarányos eredményt. A likviditást jellemző mutatók száma nyolc, ezek közül kalkuláltam a likviditása rátát, a gyorsrátát, a pénzhányad mutatót és a működő tőkéhez kapcsolódó mutatókat. Az eladósodottság tekintetében tizenkét mutatót

¹¹ A gondolatmentből kiolvasható, hogy a három változó típus (pénzügyi, piaci, makrogazdasági) arányainak összege meghaladhatja a 100%-ot, ráadásul nem is vettük figyelembe az egyéb, szóba jöhető bemeneti változókat. Ez abból adódik, hogy bizonyos tanulmányok nem csak egy típusú változót használnak, hanem jellemzően a pénzügyi változók mellett piaci és/vagy makrogazdasági változókat is alkalmaznak.

vettem figyelembe, többek közt a bonitás, a saját vagyon aránya, a kötelezettségek aránya mutatókat. A hatékonyság mutatói közül hatot számítottam ki, köztük az eszközök-, készletek és vevők forgási sebesség mutatóit. Az eszköz és tőkeszerkezeti mutatók száma szintén hat, a cash flow alapú mutatók száma öt. A pénzügyi mutatókon túlmenően a vállalat mérlegfőösszegének és árbevételének a természetes alapú logaritmusát alkalmaztam, mint a vállalatok méretére jellemző mutatókat. Az egyéb típusú változók között emlitem az alkalmazottak átlagos létszámát a vizsgált üzleti évben és a vállalat működésének korát években kifejezve.

A mezoszintű mutatók a vállalatok székhelye alapján kerültek kiválasztásra. Ezek a vállalat külső körülményeit leíró, adott településre jellemző változók. Az első közülük a település vállalatainak jövedelmezőségét jelző, egy vállalatra jutó árbevétel (Y1), amelynek feltételezett hatása a vállalati fizetéképtelenségre tekintve negatív. A helyi munkanélküliségi ráta (Y2) az álláskeresők és a gazdaságilag aktív korú népesség hányadosa, amely egyben a helyi gazdaság egészségi állapotának a mutatója is. A munkanélküliségi ráta alacsony szintje kedvező gazdasági környezetet, valamint fokozott gazdasági aktivitást jelez, amelyben a lakossági fogyasztás által megnövekedett vállalati bevétel várható. Ezzel szemben az emelkedő munkanélküliségi ráta a fogyasztáson alapuló ágazatokban a kereslet csökkenéséhez és ezzel együtt a bevételek zsugorodásához vezet.

A településen az ezer főre jutó bankfiókok és takarékszövetkezeti fiókok száma (Y3) a pénzügyi infrastruktúra szintjét méri. A banki források elérhetőségei és költségei fontos szerepet játszanak abban, hogy a vállalatok finanszírozni tudják a működésüket. Azon kívül, hogy alapvető pénzügyi támogatást nyújtanak, lehetővé teszik a növekedést a befektetési lehetőségek finanszírozásával, valamint erősíthetik az ellenálló képességet gazdasági visszaesések idején. Távoli hozzáféréssel (mobilbank, internetbank) rendkívül sok banki tevékenység elvégezhető, azonban a bankfiókok jelentősége hosszú ideig még megmarad (Helmeczi, 2010).

A település önkormányzatához kapcsolódó mutatót is bevontam a modellezésbe, mivel a vállalatok telephelyválasztásának lényeges eleme. Az egy főre jutó önkormányzati adóbevételek (Y4), az önkormányzat pénzügyi erejét és a település fejlettségét tükrözi. A beszedett adók mértéke az ingatlanpiactól, a gépkocsiparktól illetve az ott élők gazdasági erejétől függ. Minél magasabb az érték, annál függetlenebb a helyi önkormányzat a központi költségvetéstől.

A település megközelíthetőségét, a közúti közlekedés és szállítási lehetőségek helyzetét képviselő változók a főváros távolsága (Y5), és a legközelebbi gyorsforgalmi csomópont távolsága (Y6). A főváros mutató percben fejezi ki egy adott település távolságát a leggyorsabb közúton, a legközelebbi gyorsforgalmi csomópont mutató a legrövidebb úton fejezik ki a kilométerben mért távolságot. A város, kétértékű változó (Y7), a település jogállását tükrözi, amely a város és község elválasztásának mérésére szolgál azzal a céllal, hogy a városi/községi elhelyezkedés jelentőségét mérje a fizetésképtelenség szempontjából.

5.5. Az adatok elemzésre való előkészítése

Az adatok elemzésre való előkészítése számos lépést foglal magába annak érdekében, hogy kellően tiszták és a kutatás céljának megfelelően strukturáltak legyenek. Elsőként az adatgyűjtésnél említett szűrési kritériumokat érvényesítettem az árbevételre, eszközállományra, alkalmazotti létszámra és a vállalat életkorára vonatkozóan. Az ezt követő adattisztítás magába foglalta a hiányzó értékekkel rendelkező adatrekordok törlését. Amennyiben a vállalatnak a mutatószámok kiszámításához szükséges pénzügyi adatainak bármelyike hiányzott, akkor a teljes megfigyelést töröltem az adatbázisból. Ha strukturális hibákat, adatfelvételi tévedéseket észleltem, akkor szintén töröltem a teljes vállalati megfigyelést. Például abban az esetben, amikor az eszköz és forrásállomány összege nem egyezett meg, vagy a forgóeszközök állománya negatív értéket vett fel, esetleg a kötelezettségek összértéke negatív volt, eltávolítottam az adott vállalatot az adatbázisból. Azoknál a megfigyeléseknél, ahol a személyi jellegű ráfordítások egy üzleti év alatt nulla értéket mutattak, a vállalati adatsort maradéktalanul töröltem az adattáblából, amelyet adatfelvételi hibának vagy ellentmondásosnak értékeltem az olyan vállalatoknál, ahol a korábbi szűrési feltételem szerint legalább két foglalkoztatott munkavállaló kell legyen.

A tanulmány során felhasznált bemeneti adatok jelentős része hányados típusú pénzügyi mutató volt. Ennek következtében különös figyelmet kell fordítanunk, az olyan számítási problémák orvoslására, mint a kettős negatív osztás vagy a nulla értékű nevező, különben téves következtetésre jutunk. Vegyük elsőnek a kettős negatív osztás esetét. Az általam összeállított adatbázisban három mutatónál merült fel a probléma. Az egyik a sajáttőke arányos megtérülés (ROE, jelölése X2), ahol az adózás előtti eredmény és a saját tőke is negatív értékeket vehet fel. Ilyen formán a kettős negatív osztás hányadosa pozitív, amely látszólagosan egy kedvező összképet mutat a vállalatról, viszont ez teljes

mértékben félrevezető eredmény. A másik figyelmet érdemlő mutató az adózott eredmény és nettó vagyon hányadosa (jelölése X6), ahol a nettó vagyon a teljes eszközállomány és az összes kötelezettség különbsége, amely szintén negatív értéket vehet fel, amennyiben a kötelezettségek meghaladják az eszközök összértékét. A harmadik mutató a működő tőke és nettó vagyon hányadosa (jelölése X16), ahol a forgóeszközökből kivonva a rövid lejáratú kötelezettségeket kapjuk a működő tőkét. Ez szintén lehet negatív, ha a rövid kötelezettségek meghaladják a forgóeszközök állományát. Megfigyelhető volt, hogy a kettős negatív osztás jelenség többnyire a problémákkal küzdő, fizetésképtelen vállalatoknál jelentkezett. Minden hasonló esetben a beavatkozás létfontosságú az eredményeink hitelessége szempontjából. Az egyik lehetőség a mutatószámot a többi megfigyelés minimum értékével helyettesíteni, ezzel egyidőben azonban számolnunk kell egy torzító hatás jelentkezésével. A másik korrekciós lehetőség a változót teljes egészében kihagyni a vizsgálatokból. A torzító hatás korlátozásának érdekében az utóbbi megoldást választottam az adatelőkészítés során, tekintettel arra, hogy a kezdeti rendelkezésre álló adatállomány relatív nagy volt. Ebből kifolyólag az X2, X6 és X16 jelölésű mutató végül nem vett részt a modellezésben.

A hányados típusú mutatók másik korrigálható jelensége a nulla értékű nevező, amely értelmetlen eredményt ad. Ez a jelenség több mutatót érintett az adatállományban, így például a X27-el jelölt befektetett eszközök saját finanszírozása (saját tőke/befektetett eszközök) mutatónál bizonyos szolgáltató vállalatok csak forgóeszközökkel gazdálkodtak azaz nem rendelkeztek befektetett eszközökkel. Vagy egy másik az X47-el jelölt mutató, amely a működési cash flow és a hosszú lejáratú tartozások aránya, amelyben számos vállalatnak nem volt megfigyelhető hosszú forrása. Néhány kivételes esetben a saját tőke éppen nulla értékű lett így a ROE (X2) és a bonitás mutatót (X18, összes kötelezettség/saját tőke) korrigálni kellett. Más vállalatoknál az adózás előtti eredmény volt nulla, ekkor a pénzügyi ráfordítások/adózás előtti eredmény (X29) mutató korrekcióra szorult. Egy következő esetben, a nevező egy kivonás eredményéből adódott és a nulla értékű különbség javításra szorult. Ez kifejezetten az X25 jelölésű összes kötelezettségek/(összes eszköz-összes kötelezettség) mutatónál jelentkezett. A nulla nevező orvoslására az egyik lehetőség, hogy az értelmetlen eredményt adó mutatókat a többi megfigyelés releváns mutatójából számított mediánnal vagy valamelyik szélső percentilisével helyettesítjük (Han – Kamber, 2006). A másik megoldás a nevezőt egy nagyon kis számmal helyettesíteni például 0,0001-el. Ennél egy fokkal jobb praktikanak tekintetem a Virág és Nyitrai (2014a) által javasolt megoldást, amelyben a mutató

nevezőjének 0 értékét 1-el helyettesítették. Ez utóbbi ugyan növeli a kiugró megfigyelések számát, azonban nem olyan nagymértékben mint a nevezőben 0,0001-et használnánk. A saját adatbázisnál, ahol egy arányszám nevezője 0 értéket vett fel, 1-re korrigáltam, amely gyakorlatilag 1000 Forintnak felelt meg mivel az adataim ezres nagyságrendben voltak.

5.5.1. A kiugró értékek kezelése

A kiugró értékek (outliers) olyan megfigyelések, amelyek jelentősen eltérnek az adathalmaz többi elemétől. Ezek a szélsőséges megfigyelések esetenként elferdíthetik a parametrikus statisztikai elemzéseket és torzíthatják az értelmezéseket, ha nem megfelelően azonosítjuk és kezeljük (Zimmerman, 1994). A csődelőrejelző modellek építésénél a kiugró értékek jelenléte pedig általános jelenség (McLeay – Omar, 2000).

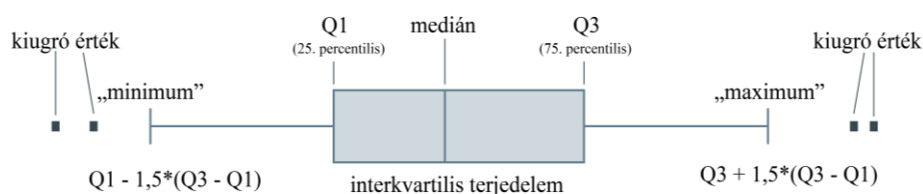
A szakirodalomban nincs konszenzus a kiugró értékek azonosítása és kezelése tekintetében. A statisztikai elemzésekben hüvelykujjszabályként fellelhetjük az átlagtól kettő- (Min – Lee, 2008) és három szórás terjedelemnél nagyobb értékek outliernek való azonosítását (Li – Sun 2011; Sun – Li, 2012). A kiugró értékek kezelésének a legegyszerűbb módja a megfigyelés kihagyása a modelltől (Sánchez-Lasheras et al., 2012; Szántó, 2023). Ehhez a megoldáshoz lehetőleg nagy adatállomány szükséges, hogy a modellépítésre megmaradt adathalmaz elegendő legyen konkluzív következtetések levonására.

A csődkutatásban a winsorizálás módszer gyakran fellelhető, hatékony eszköz a kiugró értékek kezelésére (Wu – Gaunt – Gray, 2010; Mansi – Maxwell – Zhang, 2010; Fejér-Király – Ágoston – Varga, 2019; Nyitrai – Virág, 2019). Ez a helyettesítési módszer több lépéses folyamat szokott lenni, mivel a helyettesítés után az új szórásérték alapján is jelentkezhetnek outlierok, ezért egészen addig kell végezni amíg már nem tapasztalunk kiugró értékeket (Nyitrai – Virág, 2017a). A szélsőséges értékek kezelésének további módja a folytonos változók kategoriális változókká való transzformálása (Sun – Shenoy, 2007). Fejér-Király – Ágoston – Varga (2019) a kvartilis helyzetmutatókon alapuló kategorizálást javasol, Nyitrai – Virág (2019) a CHAID alapú kategorizálását ajánlja a kiugró értékek kezelésére.

A kiugró adatok bármilyen kezelése információvesztéssel jár és lehetőség szerint érdemes azokat az eredeti értékükön a modellben tartani. Ezáltal a modell a későbbi futtatás során is alkalmas lesz a szélsőséges értékeket tartalmazó megfigyelések helyes értékelésére (Nyitrai, 2014b). A gépi tanulási módszerek az adathalmazzal

szemben nem támasztanak előfeltevéseket, így a kiugró értékeket tartalmazó adatok használatának nincs akadálya. Azonban a parametrikus módszereknél, mint a logisztikus regresszió, szükségeszerű a kiugró értékek kezelése, ezért a klasszikus módszerek futtatása előtt kezeltem a szélsőséges értékeket. A 9. ábra a dobozdiagram segítségével mutatja be az kiugró értékek meghatározásának felépítését.

9. ábra: A kiugró értékek meghatározása dobozdiagram segítségével



Forrás: (Központi Statisztikai Hivatal, 2022) alapján saját szerkesztés

Kiugró értékeknek azonosítottam azokat a megfigyeléseket, amelyek az alsó kvartilis (Q1) alatt másfélszeres interkvartilis terjedelmen kívül voltak, valamint amelyek a felső kvartilis (Q3) felett másfélszeres interkvartilis terjedelmen kívül álltak. Az interkvartilis terjedelem a felső és az alsó kvartilis különbsége vagy másképp a 75. és 25. percentilis különbsége. A kiugró értékeket iteratív módon helyettesítettem a másfélszeres interkvartilis terjedelem értékével.

Az kiugró értékek kezelésén túl, az adatok transzformációja vagy normalizálása a ferdeség korrigálásán és a variancia stabilizálásán keresztül számos előnyt hordoz a modellezésre nézve. Ebből kifolyólag a csődtanulmányokban is előfordul a bemeneti értékek normalizálása a [0, 1] intervallumban (Kennedy – Namee – Delany, 2013) vagy a [-1,1] terjedelemben (Zhong et al., 2014). A kutatás során alapvetően arra törekedtem, hogy az összes bemeneti változót az eredeti értékén használjam, nem történt transzformáció, normalizálás, abból a megfontolásból, hogy ez akár növelheti is az osztályozási pontosságot (lásd Jo – Han – Lee, 1997). Azonban bizonyos módszertani kérdések megválaszolásához egyes módszerekhez standardizált mutatóértékeket használtam, amely a (37) képlettel fejezhető ki.

$$z = \frac{x - \mu}{\sigma} \quad (37)$$

ahol x egy tetszőleges érték az \mathbf{X} változóban, μ a változó átlaga, σ pedig a változó szórása. Ennek segítségével a változók átlaga megközelítőleg nulla és szórása egységnyi lesz,

ugyanakkor megőrzi az eredeti adatok eloszlási tulajdonságait. Az így kapott értékeket z-értéknek is nevezzük, amely a standardizálás egyik alapvető módszere. A standardizálás által az egyes attribútumok értéktartományai közti különbségek mérséklődnek és bizonyos osztályozó módszerek hatékonyabb működésre lehetnek képesek. Luor (2015) kifejezi, hogy a standardizálás hatása az SVM teljesítményére nem mindig nyilvánvaló, azonban empirikus elemzésében több standardizálási eljárás tesztelése során a z-érték transzformációval készült modellek osztályozási teljesítménye javult. Szintén a standardizálás klasszifikációra ható előnyét emeli ki az SVM módszernél Ali, Smith és Miles (2006). Shanker, Hu és Hung (1996) a standardizálás hasznosságát emeli ki a neurális háló teljesítményére vonatkozóan, ugyanakkor hozzáteszi, hogy ez az előny csökken, amint egyre nagyobb adathalmazt alkalmazunk vagy a neurális háló rejtett rétegeinek a számát növeljük. A csődelőrejelzés kutatásában nem terjed el normaként, azonban találunk kiváló gyakorlati példákat a standardizálás alkalmazására (Jo – Han – Lee, 1997; Min – Jeong, 2009; Nyitrai – Virág, 2019). Mindezek alapján a modellezés során megfontoltam a standardizált változók alkalmazását az SVM módszereknél.

5.5.2. Változószelekció

A csőd és fizetéképtelenség előrejelzésénél a megfelelő változók kiválasztása kulcsfontosságú a jól teljesítő modellek kifejlesztéséhez. A változószelekciós módszerek segítenek kiszűrni a megfelelő paramétereket, amelyek a leginkább hozzájárulnak az előrejelző teljesítményhez ezzel együtt pedig rámutatnak, hogy mely változók értelmetlenek, redundánsak, korreláltak, irrelevánsak és fölöslegesek tehát kihagyhatóak a modellezésből.

A változó (variable) és jellemző (feature) kiválasztás az angol terminológiában használatos fogalmak, amelyek jelentése kissé eltér, de sok esetben szinonimaként használatos. A szelekciós eljárásoknál azonban teljesen egyenértékűek. Elkülönítésre a neurális hálónál és a gépi tanulás kernel módszereknél van szükség. A kernel gépek magas dimenziószámú térben működnek és a bemeneti nyers változók (variable) alapján magasabb szintű jellemvonásokat, jellemzőket (feature) számolnak ki implicit módon (Guyon – Elisseeff, 2003), például a fizetéképtelen és egészséges vállalatok közti komplex mintázatot vagy határvonalat.

Mivel a csődelőrejelzés ezen területén sincs egy általánosan elfogadott elmélet vagy megközelítés ezért egy folyamatos kísérletezés és próbálgatás jellemzi a változószelekciós eljárásokat (Kirkos, 2015). Így aztán vannak külön tanulmányok,

amelyek a változószelekció kérdéskörével foglalkoznak (du Jardin, 2009; Tsai, 2009; du Jardin, 2012; Papíková – Papík, 2022). A következőkben említést teszünk az általánosan alkalmazott változószelekciós módszerekről a fizetéképtelenség előrejelzésének területén. A statisztikai teszteken alapuló változószelekciónál, a t-teszt, az F-próba vagy a korreláció analízis p-értéke alapján döntenek el a kutatók, hogy mennyire fontos egy változó a csőd szempontjából (du Jardin, 2009; Tsai, 2009; Kirkos, 2015). A lépésenkénti (stepwise) regresszió egy iteratív módszer, amely statisztikai szignifikancia alapján hozzáad vagy eltávolít változókat a modellből. Ez lehet előrefelé (forward) lépcsőzetes, változókat egyenként hozzáadva vagy hátrafelé lépcsőzetes (backward), amely változókat egyenként eltávolítva határozza meg a lényeges inputok körét (Shin – Lee – Kim, 2005; Tsai, 2009).

A dimenziócsökkentő eljárások, mint például a főkomponens elemzés és faktor analízis, a kezdeti bementi változókat korrelációmentes új változókká (faktorokká, főkomponensekké) alakítja, csökkentve a változók számát úgy, hogy az információ jelentős része megmarad (Tsai, 2009). A szűkítő (shrinkage) regressziós eljárások többek közt változószelekciós feladatokra is alkalmasak. A LASSO (Least Absolute Shrinkage and Selection Operator, legkisebb abszolút értékű szűkítés és szelekciós operátor) regressziós eljárás optimalizációs folyamatában csökkenti a változók együtthatóinak értékét. Így az együtthatók akár 0 értéket is felvehetnek, ez pedig azt eredményezi, hogy csak a nem nulla együtthatóval rendelkező változókat érdemes a modellben hagyni. A Ridge regresszió hasonló működésű, mint a LASSO, az egyik különbség az, hogy nem az együtthatók abszolút összegét, hanem az együtthatók négyzetösszegét bünteti, így a koefficiensek egyike sem csökkenhet nullára (Tibshirani, 1996; Hunyadi, 2012; Horváth, 2021). Emiatt a Ridge regresszió számos kritikát kapott, hogy nem képes a változószelekcióra, azonban a regularizációs paraméterek finomhangolásával végrehajtható a változószelekció (Wu, 2021). Ugyancsak ehhez a módszercsaládhoz tartozik a rugalmas hálók (elastic net) módszere, amely az előző kettő kombinációjaként értelmezhető (Zou – Hastie, 2005).

Egy másik csoportosítása a változószelekciós algoritmusoknak a szűrő (filter) típusú, a wrapper (csomagoló) típusú és a beépített (embedded) típusú változószelekció. A szűrő típusú változószelekciós módszerek jellemzője, hogy az osztályozó algoritmustól függetlenül, szelektálja ki vagy rangsorolja a változókat egy bizonyos statisztika vagy pontszám alapján. Előnyük, hogy mérsékelt a számításigényük és kevésbé hajlamosak a túlilleszkedésre, azonban hátrányként hozható fel, hogy nem veszik figyelembe a

változók közötti kölcsönhatásokat. A wrapper típusú algoritmusok a változók egy részhalmazával kezdik, majd egy modell teljesítményét mérő kritérium alapján hozzáad vagy eltávolít változókat. Hasznos eljárás abban az esetben, ha az előrejelző teljesítményt maximalizálni szeretnénk. Ugyanakkor hiányosságként hozható fel a túlilleszkedési hajlam, a megnövekedett számításigény, mivel figyelembe veszi a változók kölcsönhatásait és a modell teljesítményét (Wah et al., 2018). A harmadik csoport a beágyazott (embedded) típusú algoritmusok, amelyek a klasszifikációs eljárás tanulási fázisával egyidőben határozzák meg a változók fontosságát. Ez lehetővé teszi, hogy a változószelekció a modelloptimalizáció részévé váljon és megtanulja a modell, hogy mely jellemzők a leghatékonyabbak. Ennek előnye, hogy a modellépítésbe integrált változószelekció nagyobb pontosságú modelleket eredményez, ezzel szemben a változószelekció lépései kevésbé átláthatóak a szűrő vagy wrapper típusú módszerekhez viszonyítva (Chen et al., 2020; Papíková – Papík, 2022). A csődelőjelzés kutatásai közt legfőképp a szűrő és wrapper típusú eljárásokat fedezhetjük fel (Sun et al., 2014; Lin – Lu – Tsai, 2019).

Van azonban egy másik kutatási kategória, amely nem kifinomult változószelekciós módszereket próbál kidolgozni. A csőd kutatások egy jelentős része, megközelítőleg a fele, a modellezéshez felhasznált változókat egy előző tanulmány eredményesnek bizonyult pénzügyi változóiból származtatja, ezen belül is túlnyomórészt Altman (1968) elismert mutatóit adaptálják. Irodalomkutató cikkében du Jardin (2009) azt találta, hogy a megvizsgált tanulmányok 40%-a használt olyan mutatót, amely a szakirodalomban népszerű vagy korábbi tanulmány hatékonynak értékelte. A saját szisztematikus irodalomelemző kutatásban is hasonló eredményt tapasztaltam. A több mint száz, legtöbbször hivatkozott tanulmány 54%-a egy korábbi tanulmányra utalt a változók kiválasztásánál. A tanulmányok 26%-a diszkriminanciaanalízist és logisztikus regressziót alkalmazott egy előzetes változószelekcióra, valamint 15% ad hoc változószelekciót alkalmazott, 5% pedig korrelációs statisztika alapján döntötte el, mely változókat tartsa meg a modellezésre (Ágoston, 2022a).

A kiválasztás módozatai mellett az inputok számossága is szóba jöhet. Du Jardin (2009) irodalomelemzéséből az derül ki, hogy a csődelőjelzésre az elmúlt négy évtizedben több mint 500 különböző pénzügyi mutatót használtak a modellekben. A saját irodalomelemző eredmények alapján a szakirodalomban figyelembe vett változók száma 2 és 190 közötti, a változószelekció után átlagosan 16 magyarázó változót használtak a modellépítésre (Ágoston, 2022a). Bellovary et al. (2007) a csődelőjelzés kezdetétől

egészen a 2000-es évek elejéig készített széleskörű irodalomelemzést és megállapításuk szerint a modellenként alkalmazott átlagos változós szám a tíz. A fentiek alapján úgy vélem, a bemeneti változók száma a 10 ± 5 intervallumban optimális. Ez a tartomány elegendő rugalmasságot biztosít az adatállomány figyelembevételéhez és a kulcsfontosságú változók megragadásához anélkül, hogy a modell túlságosan bonyolulttá és értelmezhetetlenné válna.

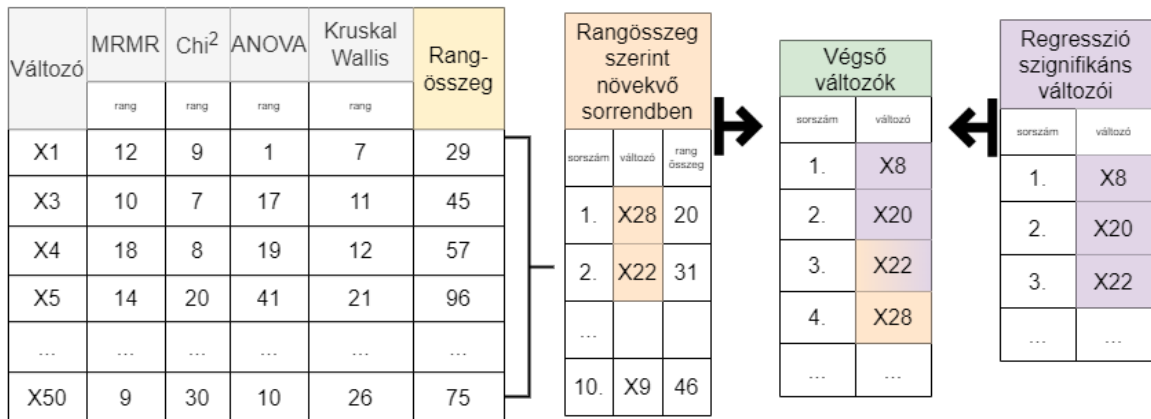
A tanulmány során alkalmazott változószelektációs eljárás alapját a lépésenkénti regresszió adta. Számos különböző próbafuttatást végezve forward és backward eljárásban kiszűrtem a regresszió szerinti szignifikáns változókat. Ezt követően a szignifikáns változókat még Pearson-féle korrelációs analízisnek vettem alá. Berry, Feldman és Stanley Feldman (1985) számítása szerint, a magas korrelációs együttható akkor okozhat multikollinearitási problémát, amikor a korrelációs együttható abszolút értéke túllépi a 0,7 értéket kis minta esetén, vagy 0,8-0,85 értéket nagy minta esetén. Ezért ha a korrelációs együttható abszolút értéke meghaladta a 0,7-es értéket, akkor az összefüggő változók közül csak az egyiket hagytam meg a végső modellben. Ez további próbafuttatásokat igényelt, amelyek szempontja az volt, hogy a tanuló mintán a fizetésképtelen vállalatok besorolási aránya minél magasabb legyen.

Ezt követően kiegészítő változószelektációs eljárásokat futtattam az összes változóra vonatkozóan. Statisztikai teszteken alapuló változószelektációt végeztem az MRMR algoritmus, a Chi-négyzet teszt, az ANOVA és a Kruskal Wallis teszt segítségével. Az MRMR (minimum redundancy maximum relevance) változószelektációs algoritmus arra törekszik, hogy azokat a változókat jelölje ki, amelyek a leginkább relevánsak, azaz fokozottan összefüggnek a célváltozóval, miközben a magyarázó változók egymás között nem korrelálnak (Peng – Long – Ding, 2005). A Chi-négyzet és ANOVA tesztek a függő és független változók közti kapcsolatot vizsgálják és a p-érték alapján adnak rangszámot a változóknak. A Kruskal Wallis módszer az ANOVA nem parametrikus megfelelője, a Wilcoxon teszt kiterjesztése, amely a különböző osztályokban a medián értékeket hasonlítja össze. A módszer elsősorban az egyes csoportok közti különbségek feltárására használatos, azonban változószelektációs célokra is alkalmazható (Wang – Xie, 2014).

A statisztikai teszteken alapuló változószelektáció rangjait összegeztem és rangösszeg szerint növekvő sorrendbe állítottam. Így az első helyen a legfontosabb változó szerepelt, az utolsó helyen pedig a legkevésbé lényeges változó. Az összesített rangok alapján a tíz legkisebb értékűt, azaz legfontosabb változót felhasználtam a

modellépítéshez, mégpedig úgy, hogy az előzőleg kiválasztott szignifikáns változók mellé illeszttem. Másképpen megfogalmazva egyesítettem a logit lépésenkénti szignifikáns változóit és a statisztikai tesztek változószelekciós eredményeit. A változószelekció folyamata a 10. ábrán látható.

10. ábra: A változószelekció lépései



Forrás: saját szerkesztés

Az egyesítés után, a végző változók kiválasztása előtt még egy utolsó Pearson-korrelációs elemzést végeztem és a 0,7-es korrelációs együtthatót meghaladó változópárok közül az egyiket elvettem. Megtartottam azt a változó, amely szignifikáns volt a regresszió szerint, szemben a nem szignifikánssal, valamint azt változót, amely használatával a tanuló mintán a fizetésektelen vállalatok helyes osztályozása magasabb arányú volt. A változószelekció eredményeként kiválasztott mutatókat mindegyik módszernél egyformán alkalmaztam, tehát azonos változókészlettel dolgoztak az egyes módszerek. Azonban a változószelekciós lépéseket mindegyik újabb vizsgálati évben előlről kezdtem és új változókészletet hoztam létre.

5.6. A módszerek paramétereit

A disszertáció során a gépi tanulás módszereit arra használtam fel, hogy különbséget tegyek a fizetésektelen és egészséges vállalatok között, olyan pénzügyi dimenziók mentén, mint a likviditás, a jövedelmezőség, a pénzügyi szerkezet és a tőkeáttétel. A modellezési eredmények megismételhetősége érdekében az alkalmazott módszerbeállításokat az F3., F4., F6. és F8. függelékben ismertetem. Az alkalmazott gépi

tanulási modellek felügyelt tanulási eljárással készültek, mivel a helyes kimenetek mindvégig ismertek voltak.

A gépi tanulás területén a paraméterek finomhangolása vagy másként a hiperparaméter optimalizáció segítségével a különböző módszerek részleteibe elmélyülve tökéletesíthetjük a modellek teljesítményét egy-egy konkrét feladat megoldására vonatkozóan. Ez ugyanakkor eszköz is, hogy megtaláljuk az egyensúlyt a túltanulás és az általánosíthatóság között. A paraméterek optimalizációjára három alapvető eljárás vettem figyelembe: a rácskeresést (grid search), a véletlen keresést (random search) és a Bayes optimalizációt. A rácskeresés a rácson belül minden kombinációt tesztl, ezért számításigényes eljárás (Alibrahim – Ludwig, 2021). A véletlen keresés a rácson belül véletlenszerűen pontok halmazát tesztl, ezért hatékonyabb és előnyösebb a rácskeresésnél főleg olyan helyzetekben, amikor a hiperparaméter tér sokdimenziós és csak néhány hiperparaméter fontos. Hatékonysága ellenére nem garatált, hogy a globális optimumot, azaz a legjobb paramétereket választja ki (Bergstra – Bengio, 2012). A Bayes optimalizálás egy kifinomultabb, valószínűség alapú technika, amely iteratív módon tanul a korábbi tesztlésekből és az alapján választja ki a következő mintát a hiperparaméter térben (Xia et al., 2017). Empirikus kutatások bizonyították, hogy a Bayes optimalizáció felülmúlja a rácskeresés és véletlen keresés teljesítményét (Li – Kanoulas, 2018; Wu et al., 2019). Ezért az SVM, a neuronháló, a bagging, a random forest és az Adaboost módszerek hiperparamétereinek finomhangolására a Bayes optimalizálást használtam. Az együttes módszereknél klasszifikációs fákat használtam alaptanulóként. A neuronháló tanítási folyamatában az optimalizálást, a kvázi-Newton módszercsaládba tartozó, LBFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) algoritmus végezte, ahol a leállási kritérium 1000 tanulási ciklus volt. A lambda regularizációs hiperparaméter a túlilleszkedést mérsékli úgy, hogy a veszteségfüggvényhez egy büntetést rendel hozzá. A lambda növelése kisebb mértékű túlilleszkedést, de ugyanakkor nagyobb torzítást okoz a becslésben, ezért egy egyensúlyt kell megtalálni a lambda és a túlilleszkedés között. A neurális háló köztes rétegeiben a (26), (27), (28) alapján fontoltam meg a Relu, Tanh és szigmoid függvényeket különböző modellek tanításánál. A kimeneti rétegben a (30) alapján a softmax függvényt alkalmaztam, mivel az outputokat valószínűségi értéként adja meg, illetve a szigmoid függvényt, amely 0 és 1 közötti értéket ad meg eredményül, így mindkettő könnyen értelmezhető a kétkimenetelű osztályozási problémánál.

A módszerek paramétereinek kiválasztásához bizonyos fokig a szakirodalmi empiriára is támaszkodhatunk. Az egyes módszereknél nincsenek mindent átfogó egyértelmű iránymutatások csupán javaslatok. Például a neuronháló esetén komoly erőfeszítések történnek az általánosítható megoldásokra, azonban a kutatónak saját magának kell optimalizálnia a háló méretét. Erre vonatkozó részletes útmutatás Altrichter et al. (2006) munkájában olvasható. A csődkutatásokban gyakori az egy köztes rétegű, tíz neuront tartalmazó háló (Wilson – Sharda, 1994; Kim – Kang, 2010; Zhao et al., 2015; Barboza – Kimura – Altman, 2017), ezért külön megfontoltam ennek az alkalmazását is. Más módszerek hiperparamétereinek finomhangolására utalásokat a *4. fejezetben* a módszerismertetések közt talál az Olvasó. A szakirodalmi tapasztalatok alapján felépült modellek eredményét összevettem az optimalizációs eljárások átesett modellek eredményével, és végül csak azokat a modelleket, illetve a hozzájuk kapcsolódó paramétereket közöltem, amelyek a legmagasabb teljesítményt nyújtották.

5.7. Validáció

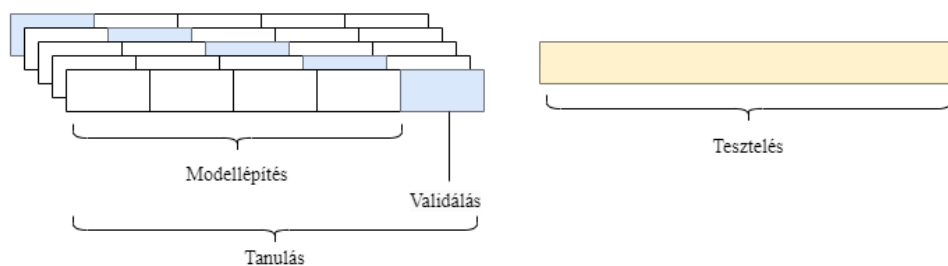
A gépi tanuláson alapuló modellek fejlesztésénél kihagyhatatlan lépés a validálás, annak érdekében, hogy a létrejövő modellek kellően hatékonyak legyenek és ugyanakkor ne csak a tanuló mintára simulva érjenek el eredményeket, hanem általánosítási képességgel is rendelkezzenek, új, nem ismert adatokon. A gyakorlatban a tanuló mintán tanított modellt a véglegesítés előtt egy másik adathalmazon, a validációs mintán futtatjuk, annak érdekében, hogy optimalizáljuk a hiperparamétereket, és értékeljük a modellt. Ezt követően a kellően finomhangolt modellt egy teljesen különálló adathalmazon, a tesztmintán értékeljük, annak érdekében, hogy felmérjük az általánosíthatóságot és az előrejelző képességet.

Számos validációs stratégia létezik, azonban nincs egy abszolút iránymutatás, amely mentén meg kellene tenni a validálást. A stratégiák közt szerepel a kívülálló (holdout) mintafelosztás, keresztvalidáció, az egyet kihagyó módszer és a rétegzett keresztvalidálás. A kívülálló (holdout) validáció az egyik legalapvetőbb megoldás főként akkor, ha elegendően nagy adatbázis áll a rendelkezésre. Ekkor az adathalmazt két részre osztjuk, tanulómintát a modellfejlesztésre, és a tanuló mintától eltérő validáló/tesztelő mintát, amely nem vesz részt a modellépítésben, csak a modell teljesítményének értékelését szolgálja. A véletlenszerű felosztási arány bármilyen lehet, gyakori az 50:50, 60:40, 70:30 és a 80:20-as tanuló-tesztelő arány. A három részre való felosztás is elképzelhető például a 60:20:20 tanuló-validáló-tesztelő mintára (Yadav – Shukla, 2016)

A keresztvalidáció (k-fold cross-validation) a teljes adatállományt k darab egyenlő részre osztja fel, amelyből $k-1$ darabot használ a modellépítésre, és egyet a validálásra. A keresztvalidáló eljárás felosztási k -értékére nincs egyértelmű ajánlás, azonban gyakori a tízszeres és ötszörös felosztás (Mahmood – Kahn, 2009). A modellek tanítása k -szor történik iteratív módon úgy, hogy mindegyik almintá szerepeljen egyszer validáló mintaként. Végül a modell teljesítményét úgy kapjuk meg, hogy az egyes alminták eredményeit átlagoljuk. Ezt a stratégiát főként akkor szokták javasolni, amikor kisméretű az adathalmaz. Az egyet kihagyó stratégia a keresztvalidációnak egy sajátos esete, amikor az adathalmazból egy elemet validációként használunk fel, az összes többi elemet a tanulásra. A tanuló iterációk száma az elemek összes számával lesz egyenlő, annak érdekében, hogy mindegyik elem egyszer validáló szerepben legyen. Szintén relatív kisméretű adathalmazokra ajánlott, mivel a számos iteráció miatt magas a számítási igénye (Marcot – Hanea, 2021).

A rétegzett keresztvalidálás (stratified cross-validation) stratégiája olyan osztályozási feladatoknál javasolt, ahol az egyes osztályok eloszlása kiegyensúlyozatlan. A technika alapjaiban ugyanolyan, mint a keresztvalidáció, azzal a különbséggel, hogy a teljes adatállomány felosztásakor figyelembe veszi az osztályok eloszlási arányát, így mindenik almintában azonos lesz az osztályok eloszlási aránya (Mathai – Chen – Kirchmair, 2020; Tougui – Jilbab – El Mhamdi, 2021).

11. ábra: A validációs mintafelosztás és a tesztminta



Forrás: (Mathai – Chen – Kirchmair, 2020) alapján saját szerkesztés

A modellek tanulási eljárásának validálására és a legjobb osztályozási teljesítményt nyújtó paraméterek finomhangolására az ötszörös keresztvalidációt használtam, amelyet a 11. ábrán szemléltetek. Az adathalmaz öt azonos méretű részhalmazából a tanulás négy részhalmazon történt, a fennmaradó egy pedig validálta a

modelleket. Ezt a folyamatot ötször kellett megismételni és minden alkalommal más-más részhalmazt használni validáló halmazként. Az egyes ismétlésekből kapott teljesítményt átlaga adta az átfogó értékelését a modellnek. Ezek után egy teljesen új adatmintán, a tesztmintán ellenőriztem a modell előrejelző képességét. Választásom azért esett a keresztvalidációra, mert az nagyobb mennyiségű adatot használ fel tanulásra, mintha a tanuló halmazból elkülönítenénk egy részt validálásra. Emellett egyes empirikus elemzések alapján hatékonyabbnak is bizonyult, mint a „holdout” validáció (Blum – Kalai – Langford, 1999; Yadav – Shukla, 2016).

5.8. A modellek teljesítményének értékelése

A modellek teljesítményének objektív értékelésére és a különböző modellek összehasonlításához a következő teljesítménymutatókat használtam: egészségesek helyes besorolási aránya, a fizetéképtelen megfigyelések helyes besorolási aránya, átlagos besorolási pontosság mutató, elsőfajú hibaarány, másodfajú hibaarány, F₁-érték, ROC-görbe alatti terület.

A besorolási pontosságok tekintetében a klasszifikációs tábla nyújt segítséget (*12. ábra*), amely alapján számos teljesítményt reprezentáló arányszámot lehet kalkulálni. Pozitív ténynek tekintetem az egészséges vállalatokat és negatívnak a fizetéképteleneket. Így az igaz pozitív érték (TP) az egészséges megfigyelések helyesen besorolt egyedeinek a számát mutatja. Az igaz negatív (TN) érték mutatja a fizetéképtelenek helyes besorolásainak a számát. A hamis negatív (FN) vagy álnegatív az egészséges megfigyelések fizetéképtelennek való azonosítása, amit az összes egészséges megfigyeléssel osztva másodfajú hibának nevezünk (Type II error). A hamis pozitív (FP) vagy álpozitív a fizetéképtelen megfigyelések téves besorolását mutatja az egészséges csoportban, amely az összes valós fizetéképtelen megfigyelés hányadosaként az elsőfajú hibát jelenti (Type I error). Hitelezői szempontból az elsőfajú hiba fontosabb és magasabb költséggel jár, mivel a fizetéképtelen adós is hitelt kap. A másodfajú hiba csupán egy jó lehetőség elszalasztását jelenti, azaz egy működőképes és egészséges ügyfél elutasítását (Virág et al., 2013a).

12. ábra: Klasszifikációs tábla és a hozzá kapcsolódó teljesítménymutatók

		előrejelzett osztály			
		egészséges	fizetéképtelen		
valós osztály	egészséges	TP	FN	szenzitivitás: $\frac{TP}{(TP+FN)}$	másodfajú hiba: $\frac{FN}{(TP+FN)}$
	fizetéképtelen	FP	TN	specifitás: $\frac{TN}{(TN+FP)}$	elsőfajú hiba: $\frac{FP}{(FP+TN)}$
				átlagos pontosság: $\frac{(TP+TN)}{(TP+TN+FP+FN)}$	

Megjegyzés: TP (true positive) = igaz pozitív, FN (false negative) = hamis negatív,
 FP (false positive) = hamis pozitív, TN (true negative) = igaz negatív.

Forrás: (Virág, 2001, 76 p.) alapján saját szerkesztés

A szenzitivitás ebben a megközelítésben az egészséges megfigyelések egészségesnek osztályozott száma és az összes egészséges megfigyelés hányadosa. Ez a helyesen besorolt egészségesek arányát mutatja. A specifitás a ténylegesen fizetéképtelen helyes becslése és az összes fizetéképtelen megfigyelés aránya. Az átlagos besorolási pontosságmutató a helyesen osztályozott egészségesek és fizetéképtelenek, valamint az összes megfigyelés hányadosa. Ez a mutató az előrejelző modellek átlagos teljesítményét mutatja. Az elsőfajú hibaarány a fizetéképtelenek hibásan osztályozott aránya, amely kifejezhető az 1-specifitás összefüggésből. A másodfajú hibaarány kifejezi, hogy az egészséges vállalatok hány százalékát osztályoztuk fizetéképtelennek (1-szenzitivitás).

Az általánosított F_{β} -érték a precizitás (precision) és a visszahívás (recall) közötti egyensúly beállítása mellett egyetlen számba tömörítve ad visszajelzést a klasszifikáció hatékonyságáról. A gyakran alkalmazott F_1 -érték (ahol a $\beta=1$) harmonikus átlagot számol, és azonos fontosságot tulajdonít a precizitásnak és a visszahívásnak a következő képlet alapján (Kubat, 2017):

$$F_1 = \frac{2pr}{p+r} \quad (38)$$

ahol az egészségeseknél a $p = \frac{TP}{TP+FP}$ a precizitást jelöli, az $r = \frac{TP}{TP+FN}$ a szenzitivitás vagy másképpen visszahívás. Ugyanezt számolva a fizetésképteleneknél a $p = \frac{TN}{TN+FN}$ a negatív prediktív érték, amely a valóban fizetésképtelenek arányát mutatja meg az összes fizetésképtelennek előrejelzett vállalat közül, valamint az $r = \frac{TN}{TN+FP}$ a specifitást jelöli. Az F_1 -érték nulla és egy közötti számot ad megoldásul, ahol a magas érték azt jelzi, hogy a modell a legtöbb esetet helyesen azonosítja és kevés hamis hibát vét.

A modellek teljesítményét, továbbá a kumulált osztályozási pontosságot (receiver operating characteristic, ROC) a görbe alatti terület nagyságával fejeztem ki. A ROC-görbe grafikusán ábrázolva megmutatja a vágóértékek figyelembevételével, hogy a modell során kapott osztályvalószínűségek mennyire pontosak a valós osztályozáshoz mérten. Segítségével a vágóértékek kalibrációját is el lehet végezni (Yin – Tian, 2014). Tehát az ROC-görbe az elsőfajú hiba arányát, valamint a fizetőképes vállalatok helyesen osztályozott arányát ábrázolja. A koordináta-rendszerben a 45 fokos referenciaegyenesen a helyesen osztályozott fizetőképesek aránya és a fizetőképesnek becsült fizetésképtelen vállalatok aránya egyenlő, ami a véletlen találgatást szemlélteti, és ezen a szinten a modell nem fogadható el. A ROC-görbe alapján akkor értékelhetünk hatékonyabbnak egy modellt, ha egyre inkább eltávolodik a referenciaegyenesestől. Az ROC-görbe alatti terület nagyságával (area under curve, AUC) számszerűen is mérhető a modell teljesítménye. A mutató értéke [0, 1] közötti, ahol a 0,5 a véletlen találgatásnak, az 1 a tökéletes modellnek felel meg. Az AUC előnyös mutató több szempontból is, annak köszönhetően, hogy az egy osztályra való túlzott illeszkedést kiszűri, illetve az osztály kiegyensúlyozatlanságának problémáját feloldja (Chava – Jarrow, 2004; Agarwal – Taffler, 2008; Kristóf, 2008; Nyitrai – Virág, 2017b). A következő ökölszabály alkalmazható a modellek értékelésére (Kovács, 2014):

- AUC = 0,50-0,60: nem alkalmas a modell a megkülönböztetésre
- AUC = 0,60-0,70: gyenge modell
- AUC = 0,70-0,80: közepes modell
- AUC = 0,80-0,90: jó modell
- AUC = 0,90-1,00: kiváló modell

6. A vállalati fizetéseképtelenség előrejelzésének

empirikus vizsgálata

A modellezésre, próbafuttatásokra valamint tesztelésre számos különböző programcsomagot használtam. Alapvetően a MATLAB R2023b verzióját alkalmaztam a Neural Network Toolbox, a Statistical Toolbox és a Classification Learner kiegészítő programcsomagokkal. Ezen kívül az R 4.2.2, az SPSS v26, Python 3.12 és a Weka 3.8.6 szoftvereket használtam különböző tesztelésekhez. A modellek kalibrációja és optimalizációja számos próbafuttatást igényelt, azonban csak a kérdések megválaszolása szempontjából releváns és legjobban teljesítő végső modellek eredményeit ismertettem.

6.1. A fizetéseképtelenség előrejelzése kiegyensúlyozatlan

adatmintán

Ahogy a csődelőrejelzés szakirodalmában tapasztaljuk, az 50%-50%-os fizetéseképtelen és egészséges megfigyeléseken alapuló kiegyensúlyozott adatminta alkalmazása modellépítésre általános. Eközben a valóságban a csődös és egészséges vállalatok sokasági aránya jelentősen eltérő ettől a hipotetikus eloszlástól. A fizetéseképtelen vállalatok felülreprezentálása praktikus okokból történik, azonban az előnye nem mindig egyértelmű. A következőkben megvizsgálom a valós sokasági arányok alapján reprezentatív adatminta és az azonos mintafelosztás előrejelzési teljesítményre gyakorolt hatását.

H1: A reprezentatív mintával készült fizetéseképtelenséget előrejelző modell az azonos osztálylétszámú mintavétellel szemben kisebb hatékonyságú előrejelzésre képes a fizetéseképtelenek helyes besorolási arányát tekintve.

A vizsgálathoz egy teljes üzleti év adataiból illesztettem a modelleket, a tesztelő minta pedig a rá következő egy év volt. Ilyenformán a 2013-as modellen a 2014-es adatokat teszteltem, a 2014-es modellen a 2015-öst, és így tovább 2017-es tesztmintáig befejezőleg. Összesen 4-4 modell épült a reprezentatív és kiegyensúlyozott mintára. A futtatott módszerek a logisztikus regresszió, az SVM (lineáris és Gaussi), a neuronháló, a bagging, a random forest és a boosting volt. Végül a klasszikus diszkriminanciaanalízis módszerét elvettem tekintettel a számos korlátozó előfeltételre, amelyeknek a pénzügyi mutatók nem feleltek meg. A reprezentatív mintával készült modellek paramétereit az $F3$.

függelékben, az azonos osztálylétszámú mintán épített modelleket az *F4. függelék A-paneljében* láthatjuk. A modellekhez felhasznált bemeneti mutatószámok az *F5. függelékben* láthatóak, amelyek kiválasztásának menete az *5.5.2 fejezetben* található. Az azonos évben, de különböző mintákhoz saját változókészletet használtam. Abból a célból jelöltem ki sajátos változókat az egyes vizsgálati mintafelosztások teszteléséhez, az azonos és reprezentatív modellekhez, hogy a lehető legmagasabb teljesítményt ériék el, eltekintve az esetleges torzulástól, amit a különböző darabszámú változó okozhat.

Tekintsük át a leggyakrabban előforduló változókat. Az X50-el jelölt, a vállalatok működési évét jelölő mutató mindenik modellben szerepelt, a logisztikus regresszió stepwise algoritmus rendszeresen szignifikáns negatív hatású változónak azonosította a fizetéseképtelenségre nézve. A *3. táblázat* a működési év (X50) alapstatisztikáit mutatja be a különböző mintafelosztásokban és osztályokban. Ez alapján a fizetéseképtelenek középvértékei alacsonyabbak (5-6 év), mint az egészséges megfigyelések ugyanezen mutatói (8-10 év). A reprezentatív és azonos minták dimenzióját tekintve nincsenek túl nagy eltérések, ami a két mintafelosztás hasonlóságát jelzi a mutató tekintetében. A fizetéseképtelen megfigyelések pedig megegyeznek a reprezentatív és az azonos osztálylétszámú mintában.

3. táblázat: A különböző mintákban szereplő vállalatok működési évének (X50) átlaga, középvértéke és szórása

Év	Minta	Egészséges			Fizetéseképtelen		
		átlag	medián	szórás	átlag	medián	szórás
2013	Reprezentatív	10,36	9	6,42	7,86	5	6,24
	Azonos	9,76	8	6,41	7,86	5	6,24
2014	Reprezentatív	11,53	10	6,42	7,88	5	5,97
	Azonos	10,94	10	6,34	7,88	5	5,97
2015	Reprezentatív	10,82	9	6,77	8,24	6	6,26
	Azonos	10,72	10	6,6	8,24	6	6,26
2016	Reprezentatív	11,14	9	6,91	8,04	6	6,04
	Azonos	11,61	9	7,32	8,04	6	6,04
2017	Reprezentatív	10,98	9	7,38	7,16	6	6,00
	Azonos	10,87	8	7,41	7,16	6	6,00

Forrás: saját szerkesztés

A másik gyakran előforduló változó az X39-el jelölt mutató, amely a nyolc modellből hatban előfordult és a készletekkel csökkentett forgóeszközök arányát mutatja

meg az összes eszközhöz mérten. Tulajdonképpen az összes eszközökön belül a leggyorsabban pénzzé tehető eszközök arányával fejezi ki a vállalat likviditási képességét, mivel a számlálóban a pénzeszközök, követelések és értékpapírok szerepelnek. Az egészséges és fizetéseképtelen megfigyelések átlagértékei a 4. táblázatban láthatóak.

4. táblázat: A különböző mintákban szereplő X39* likviditási mutató átlaga, középértéke és szórása

Év	Minta	Egészséges			Fizetéseképtelen		
		átlag	medián	szórás	átlag	medián	szórás
2013	Reprezentatív	0,47	0,45	0,67	0,58	0,66	0,37
	Azonos	0,50	0,48	0,31	0,58	0,66	0,37
2014	Reprezentatív	0,47	0,45	0,30	0,55	0,56	0,37
	Azonos	0,47	0,44	0,31	0,55	0,56	0,37
2015	Reprezentatív	0,49	0,47	0,31	0,59	0,67	0,37
	Azonos	0,48	0,46	0,30	0,59	0,67	0,37
2016	Reprezentatív	0,49	0,47	0,30	0,62	0,67	0,35
	Azonos	0,49	0,46	0,31	0,62	0,67	0,35
2017	Reprezentatív	0,50	0,49	0,31	0,61	0,72	0,35
	Azonos	0,49	0,49	0,31	0,61	0,72	0,35

*X39 = (forgóeszközök-készletek)/összes eszközök

Forrás: saját szerkesztés

Alapvetően a magasabb arányhoz kedvezőbb likviditást párosíthatunk, olyan megközelítésből, hogy a vállalat eszközeinek nagyobb részét likvid eszközökbe kötötte le. Ezzel szemben az alacsonyabb arány mérsékelt likviditást és relatív magasabb kockázatokat jelezhet hitelezői szempontból. Alaposabban megfontolva, egy egészséges egyensúly megtalálása volna a cél, mivel egy bizonyos szint alatt likviditási gondok jelentkezhetnek, de hatékonysági problémákra utal, ha túl magas az érték.

Az eredmények alapján a fizetéseképtelen vállalatok középértékei 56-72% közöttiek és legalább tíz százalékponttal magasabbak, mint az egészségeseké (4. táblázat). Tehát ebben az esetben a fizetéseképtelenek magas aránya azt jelzi, hogy a vállalat feleslegesen sok likvid eszközt tart, nem hatékonyan gazdálkodik az eszközeivel ezáltal az esetleges nyereséget termelő beruházásokat elszalasztja. Hozzá kell tenni, hogy a mintában szereplő vállalatok között nem jellemző a likvid értékpapírok tartása mindössze 5%-6%-nak van nem tartós befektetésként értékpapírja a mérlegében. Ebből kifolyólag a mutató számlálójában túlnyomórészt pénzeszközök és követelések szerepelnek. A változó hatásait tekintve az X39-et a logisztikus regresszió szignifikáns

pozitív hatásúként értelmezte, ami azt jelenti, hogy a változó növekedése - a többi változó változatlansága mellett - növeli a fizetésképtelenség esélyét.

Ha tovább megyünk, és a következő rendszeresen beválogatott magyarázó változót tekintjük, akkor részletesebb rálátást kapunk a korábbi változóra. Az X37-el jelölt mutató a pénzeszközök arányát fejezi ki a forgóeszközökön belül. A különböző minták és osztályok alapstatisztikáit az 5. táblázatban láthatjuk.

5. táblázat: A különböző mintákban szereplő pénzeszközök aránya (X37) mutató átlaga, középértéke és szórása

Év	Minta	Egészséges			Fizetésképtelen		
		átlag	medián	szórás	átlag	medián	szórás
2013	Reprezentatív	0,29	0,20	0,28	0,19	0,04	0,30
	Azonos	0,31	0,21	0,28	0,19	0,04	0,30
2014	Reprezentatív	0,30	0,21	0,28	0,19	0,04	0,28
	Azonos	0,30	0,21	0,29	0,19	0,04	0,28
2015	Reprezentatív	0,34	0,25	0,30	0,24	0,08	0,32
	Azonos	0,35	0,28	0,29	0,24	0,08	0,32
2016	Reprezentatív	0,34	0,25	0,30	0,24	0,06	0,32
	Azonos	0,35	0,28	0,31	0,24	0,06	0,32
2017	Reprezentatív	0,36	0,28	0,31	0,24	0,08	0,32
	Azonos	0,36	0,25	0,32	0,24	0,08	0,32

Forrás: saját szerkesztés

A mutató értékei alapján a fizetésképtelen vállalatok átlagosan nem tartanak arányaiban olyan nagy pénzeszköz-állományt, mint az egészséges osztályban lévő vállalatok. Az egészségesek középértéke 20-28%, míg a fizetésképteleneké 4-8% közötti. Így a fenti X39 mutatóra is jobb magyarázatot kapunk, mivel körvonalazódott, hogy annak a nevezőjében nem kifejezetten pénzeszközök, hanem inkább követelések szerepelnek. Következésképpen a korábbi X37 változó tulajdonképpen követeléskezelési, hiányosságokra utal. Ha még tovább mennénk, és megnéznénk az X32-el jelölt vevők forgási sebessége mutatót, akkor megerősítést nyerne a feltételezésünk, mivel azt látnánk, hogy a fizetésképteleneknél átlagosan alacsonyabb értékek vannak, ami lassabb követelés behajtást eredményez.

A pénzeszközök aránya mutató (X37) a logisztikus regresszió alapján szignifikáns negatív kapcsolatban áll a fizetésképtelenséggel, ceteris paribus. Az ebből levonható következtetés az, hogy a pénzeszközök arányának növekedése csökkenti a

fizetéseketelenség esélyét. Emellett úgy gondolom, hogy a pénzeszközök állományának meghatározásánál fokozottan az optimális egyensúlyra kellene törekednie a vállalatvezetésnek, az elegendő mértékű likviditás és a pénztöbblet hozamának maximalizálása között. A kedvező vállalati profil fenntartása érdekében azt is lényeges szem előtt tartani, hogy az alacsony arányú pénzállomány kedvezőtlen képed ad a hitelezőknek, beszállítóknak mivel fokozza a rövid lejáratú kötelezettségek nem teljesítésének kockázatát.

Több modellben felbukkant a saját tőke és összes kötelezettség aránya (X28), amely megmutatja, hogy a saját tőke értéke mennyire fedi le a rövid és hosszú lejáratú kötelezettségek értékét. A 6. táblázatban a medián értékeket figyelve a fizetésképtelen vállalatoknál nagyobb arányban jelentkezett negatív előjelű változó, amely a saját tőke negatív értékéből adódik. A negatív saját tőke pedig a többéves, tartós veszteségekből származik, és súlyos pénzügyi kockázatokat jelent a vállalat életképességére. Ez egyben azt is jelenti, hogy a vállalat összes tartozása meghaladja az eszközeinek az összértékét, ebből kifolyólag a vállalat vezetésének azonnali és hatékony lépéseket szükségszerű tennie a problémák kezelésére.

6. táblázat: A különböző mintákban szereplő tartozások lefedettsége mutató (X28) átlaga, középértéke és szórása

Év	Minta	Egészséges			Fizetésképtelen		
		átlag	medián	szórás	átlag	medián	szórás
2013	Reprezentatív	35,53	0,9	1018,67	25,95	-0,20	979,32
	Azonos	4,25	1,01	13,82	25,95	-0,20	979,32
2014	Reprezentatív	24,20	1,00	692,99	0,04	-0,16	1,68
	Azonos	3,65	1,05	10,24	0,04	-0,16	1,68
2015	Reprezentatív	42,56	0,98	1289,38	51,25	-0,19	816,36
	Azonos	5,02	0,99	34,80	51,25	-0,19	816,36
2016	Reprezentatív	47,22	1,28	1299,98	0,26	-0,28	3,10
	Azonos	57,17	1,15	757,52	0,26	-0,28	3,10
2017	Reprezentatív	60,30	1,34	1574,23	22,34	-0,09	2862,25
	Azonos	28,53	1,54	277,81	22,34	-0,09	2862,25

Forrás: saját szerkesztés

Említést érdemel még a bérarányos eredmény (X9), amely a működési eredmény és a személyi jellegű ráfordítások hányadosa, két-két modellben lényeges változóként kiválasztásra került. A 7. táblázat adatai alapján a fizetésképtelenek fokozottan a negatív tartományban helyezkednek el, így átlagosan a működési veszteségük nagyobb, mint a

személyi jellegű ráfordításaik. Az egészségesek esetén a mutató 20%-29% közötti középvértéke azt mutatja, hogy a működési eredményük átlagosan pozitív, azonban ennél nagyobb értékű a személyi ráfordítás. A logisztikus regresszió a bérarányos eredményt negatív hatású változóként azonosította a fizetésképtelenségre nézve, azaz az egyre nagyobb mutató csökkenti a fizetésképtelenség esélyét, a többi változó változatlansága mellett. Ily módon a bérek sokszerű növekedése többszörösen csökkenti a mutató értékét a számláló és a nevező kedvezőtlen alakulásával, ezáltal növelve a fizetésképtelenség esélyét.

7. táblázat: A különböző mintákban szereplő bérarányos eredmény mutató (X9) átlaga, középvértéke és szórása

Év	Minta	Egészséges			Fizetésképtelen		
		átlag	medián	szórás	átlag	medián	szórás
2013	Reprezentatív	0,51	0,20	8,73	-1,18	-0,62	5,41
	Azonos	0,54	0,26	1,92	-1,18	-0,62	5,41
2014	Reprezentatív	0,58	0,24	22,08	-1,22	-0,67	5,44
	Azonos	0,52	0,21	1,65	-1,22	-0,67	5,44
2015	Reprezentatív	0,79	0,25	11,64	-2,34	-0,62	14,01
	Azonos	0,47	0,23	2,55	-2,34	-0,62	14,01
2016	Reprezentatív	0,69	0,23	6,56	-1,46	-0,72	4,69
	Azonos	0,51	0,21	1,43	-1,46	-0,72	4,69
2017	Reprezentatív	0,78	0,24	9,54	-3,87	-0,53	27,13
	Azonos	0,71	0,29	2,54	-3,87	-0,53	27,13

Forrás: saját szerkesztés

A magyarázó változók tárgyalása után rátérek az egyes modellek eredményére a különböző arányú mintakészleteknél (8. táblázat). Csak a tesztmintán elért eredményeket közöltem, nem találtam magyarázó értékűnek a tanuló mintán elért teljesítményeket ismertetni. Az Adaboost módszer belső súlyozási eljárása nem teszi lehetővé az osztályszűrés, ezért ennél a módszernél nem volt mód a reprezentatív minta újrasúlyozására. A reprezentatív kiegyensúlyozatlan adatmintára való tekintettel nem a besorolási pontosság szimmetrikus mértékeket közöltem, hanem az F_1 -értéket az egészséges és fizetésképtelen megfigyelések tekintetében. Az egészséges megfigyelések F_1 -értéke a reprezentatív adatmintán majdnem minden modellnél magasabb, mint az azonos osztálylétszámmal készült modellek hasonló mutatójánál. Ez alól kivételt képzett a Gaussi SVM és a neuronháló, amely bizonyos adatokon teljesen a fizetésképtelenekre

idomult. A reprezentatív mintánál az egészségesek átlagos F_1 -értéke 88%, az azonos osztálylétszámánál 78%.

8. táblázat: A reprezentatív és azonos osztálylétszámú mintán készült modellek eredményei

Módszer	Tanuló minta 2013, tesztelő minta 2014					
	Minta: reprezentatív			Minta: azonos osztálylétszám		
	Egészségesek F_1 -érték	Fizetéseképtelenek F_1 -érték	AUC	Egészségesek F_1 -érték	Fizetéseképtelenek F_1 -érték	AUC
Logit	95,22%	32,13%	0,8653	79,32%	75,51%	0,8423
SVM lineár	94,10%	28,74%	0,8657	79,91%	76,07%	0,8531
SVM RBF	98,32%	0,00%	0,4836	83,00%	80,29%	0,8819
Neuronháló	92,81%	24,17%	0,8448	83,30%	81,67%	0,8919
Bagging	98,37%	5,66%	0,5146	84,58%	83,87%	0,9110
Random Forest	98,40%	11,85%	0,5317	84,29%	83,25%	0,9170
Adaboost	98,41%	30,57%	0,6013	84,79%	83,30%	0,9122
Módszer	Tanuló minta 2014, tesztelő minta 2015					
	Minta: reprezentatív			Minta: azonos osztálylétszám		
	Egészségesek F_1 -érték	Fizetéseképtelenek F_1 -érték	AUC	Egészségesek F_1 -érték	Fizetéseképtelenek F_1 -érték	AUC
Logit	89,45%	13,46%	0,8311	80,88%	80,54%	0,8805
SVM lineár	88,67%	13,11%	0,8430	80,77%	80,15%	0,8423
SVM RBF	0,00%	4,05%	0,5003	80,99%	81,41%	0,8346
Neuronháló	80,60%	9,12%	0,8173	80,44%	80,00%	0,8518
Bagging	93,94%	10,64%	0,6126	76,79%	77,04%	0,9090
Random Forest	94,80%	11,93%	0,6180	76,37%	76,94%	0,9143
Adaboost	92,88%	12,31%	0,6622	77,88%	79,09%	0,9129
Módszer	Tanuló minta 2015, tesztelő minta 2016					
	Minta: reprezentatív			Minta: azonos osztálylétszám		
	Egészségesek F_1 -érték	Fizetéseképtelenek F_1 -érték	AUC	Egészségesek F_1 -érték	Fizetéseképtelenek F_1 -érték	AUC
Logit	90,92%	13,67%	0,8592	81,41%	78,85%	0,8748
SVM lineár	93,29%	16,68%	0,8566	16,31%	67,48%	0,8723
SVM RBF	81,13%	7,49%	0,7761	75,79%	74,03%	0,8797
Neuronháló	0,00%	3,47%	0,5000	76,68%	74,48%	0,8680
Bagging	99,14%	11,99%	0,5325	83,93%	83,51%	0,8074
Random Forest	99,16%	11,79%	0,5315	83,49%	83,25%	0,8571
Adaboost	98,52%	20,59%	0,5989	83,80%	82,67%	0,8564
Módszer	Tanuló minta 2016, tesztelő minta 2017					
	Minta: reprezentatív			Minta: azonos osztálylétszám		
	Egészségesek F_1 -érték	Fizetéseképtelenek F_1 -érték	AUC	Egészségesek F_1 -érték	Fizetéseképtelenek F_1 -érték	AUC
Logit	89,86%	10,10%	0,8068	81,39%	78,66%	0,8727
SVM lineár	94,73%	14,84%	0,8074	47,25%	80,60%	0,7211
SVM RBF	99,20%	0,00%	0,6542	80,51%	78,44%	0,8718
Neuronháló	96,44%	17,90%	0,8108	81,34%	80,67%	0,8764
Bagging	99,21%	40,11%	0,5103	83,33%	81,56%	0,9083
Random Forest	99,18%	13,90%	0,5406	83,45%	81,74%	0,9088
Adaboost	99,07%	20,52%	0,5731	84,12%	82,90%	0,9049

Forrás: saját számítás

A 8. táblázatban kiemelve láthatjuk a fizetéseképtelenek eredményeit mivel azok szerepelnek a kérdéses fókuszpontjában. A színskála vörös oldala a gyenge teljesítményt, az egyre sötétebb zöld szín pedig a kitűnő eredményeket jelzi. Az azonos osztálylétszámánál átlagosan 80%-os értékeket látunk. Ehhez mérten a reprezentatív mintában a kisebb számban megfigyelt fizetéseképteleneket a modellek korlátozott mértékben tudták azonosítani, átlagosan 15% F_1 -értéket kapunk, legjobb esetben is mindössze 40%-os értéket. A fizetéseképtelenek gyenge osztályozási teljesítményével pedig pont a lényeg vesz el, mivel ezen a területen kifejezetten a fizetéseképtelenek

érdekelnek igazán. Ha az összes vállalatot egészségesnek becsüljük az egy nagyon egyszerű - és ráadásul téves - osztályozási mechanizmust takar, amely azon alapszik, hogy a legnépesebb osztályba sorolunk mindenkit, így végül az átlagos osztályozásunk kedvező képet mutat.

Az azonos mintavételen alapuló modellek sokkal kiegyensúlyozottabb képet adtak. Itt egyetlen illeszkedési probléma ismerhető fel, a lineáris függvényformával készült SVM 2015-ös és 2016-os tanuló mintán készült modelleknél, ahol a tanuló algoritmus túlságosan rásimult a fizetésképtelen megfigyelésekre, ezzel egyidőben az egészségeseket is fizetésképtelennek azonosította.

A görbe alatti terület (AUC) segítségével egy következő számszerű mértékkel összemérhetjük a teljesítményeket (8. táblázat). A kezdeti várakozásokkal összhangban, az azonos osztálylétszámú minta esetén az AUC majdnem mindenütt magasabb értéket képviselt, következésképpen átlagosan jobb előrejelző képességű modellek készültek. Ez alól kivétel néhány esetben a logisztikus regresszió és a lineáris SVM.

Annak bizonyítására, hogy az azonos osztálylétszámú mintavételnél a fizetésképtelenek pontosság mértéke, valamint a modellek AUC értéke jelentősen jobb-e a reprezentatív minta eredményeihez mérten, a következő statisztikai tesztekkel végeztem el. Először a páros T-próba alkalmazását fontoltam meg, azonban az F_1 -értékek és az AUC értékek egyike sem normális eloszlású ezért a T-próba nemparametrikus megfelelőjével a páros Wilcoxon próbával hasonlítottam össze az eredményeket, amely rangtranszformációt végez a megfigyelések különbségein. A különbségek előjeleit nem veszi figyelembe a rangok kiosztásánál, csak a rangok összegzésénél. A Wilcoxon próba eredményei a 9. táblázatban láthatóak.

9. táblázat: A reprezentatív- és az azonos mintavétellel készült modellek eredményeinek összehasonlítása a Wilcoxon próbával

Teljesítménymutató	Negatív rangok gyakorisága	Pozitív rangok gyakorisága	Azonos rangok gyakorisága	Z-érték	p-érték
Reprezentatív, egészséges F_1 -érték – Azonos, egészséges F_1 -érték	2	26	0	-3,393	<0,001
Reprezentatív, fizetésképtelen F_1 -érték – Azonos, fizetésképtelen F_1 -érték	28	0	0	-4,623	<0,001
Reprezentatív, AUC – Azonos, AUC	24	4	0	-4,190	<0,001

Forrás: saját szerkesztés

Az alacsony, 0,05-nél kisebb p-értékek azt jelzik, hogy a rangsorokba rendezett medián értékek statisztikailag szignifikánsan különböznek egymástól. A negatív rangok gyakorisága azt jelzi, hogy hány megfigyelésnél volt a kivonandó - a második tag, amely az azonos osztálylétszámú mintát jelenti - magasabb értékű mind a kisebbítendő, így a köztük lévő különbség negatív. Ennek a fordítottja a pozitív rangok előfordulása. Az azonos rangok gyakorisága pedig akkor jelentkezik, amikor a két megfigyelés megegyezik. A 9. táblázat alapján tehát, a Wilcoxon rang teszt azt mutatja, hogy az egészségesek F_1 -értékénél a pozitív rangok érvényesülnek, ami azt jelenti, hogy a reprezentatív mintavételnél magasabb volt és ez statisztikailag szignifikáns. Továbbá a fizetéképtelenek F_1 -értéke az azonos mintavételnél szignifikánsan magasabb volt, mint a reprezentatív mintavételnél, a negatív rangok alapján. Hasonlóan a negatív rangok érvényesülnek az AUC értéknél is, tehát az azonos osztálylétszámú mintavételnél szignifikánsan magasabb az AUC érték, mint a reprezentatív mintavételnél.

Az eredmények alapján a fizetéképtelenek relatív magas arányú helyes osztályozásához, valamint egy átlagosan jó előrejelző modellhez a fizetéképtelenek felülreprezentálása szükséges, ez pedig az azonos osztálylétszám létrehozásával megvalósítható. Különböző modellek nem tudják jól megbecsülni a fizetéképtelenséget, ami a hitelező vállalatoknak jelentős költséget okozhat. Az ismertetett eredmények mentén az első hipotézisem elfogadom, amely szerint a reprezentatív mintával készült fizetéképtelenséget előrejelző modell az azonos osztálylétszámú mintavétellel szemben átlagosan kisebb hatékonyságú előrejelzésre képes a fizetéképtelenek helyes besorolási arányát tekintve.

6.2. Több év adatain épült modellek

A tanulmány másik célja megvizsgálni a gépi tanuláson alapuló fizetéképtelenségi modellek megbízhatóságát és előrejelző képességét, amelyek több megfigyelési év pénzügyi mutatóit egyidejűleg tartalmazzák. Ezzel a kérdéssel a modellépítési elgondolások alapjait szeretném megerősíteni. A több év adatain épült modellek vélhetően több információt szolgálnak és ennek érdekében az előrejelző teljesítmény várhatóan javulni fog. Az ehhez kapcsolódó hipotézisem a következő:

H2: A több év adataiból épített, fizetéképtelenséget előrejelző modellek átlagosan magasabb előrejelző teljesítményűek egyéves előrejelzési horizonton, mint az egyetlen üzleti év adatain épült modellek.

A modellek építéséhez a mintavétel az azonos osztálylétszámú eljárás szerint történt, mivel bebizonyosodott, hogy az hatékonyabb megoldásra vezet. A vizsgálati időszak az előbbi modellekhez mérten kitolódik, és a 2013-2021 közötti évekre korlátozódik. Tisztázandó, hogy a több év nem az egyes vállalatok több éves múltbeli adatait jelenti, hanem több üzleti évből származó különböző vállalatok adatait. Ezzel biztosítva van, hogy a fizetésképtelenek adatai mindig a legfrissebbek legyenek, mivel mindvégig a fizetésképtelenség bekövetkezése előtti első évből származnak. A tanuló-tesztelő minták évek szerinti felépítése a 10. táblázatban látható, az egyes évek minta elemszáma a korábban bemutatott 2. táblázatban található.

10. táblázat: A több év adatainak épülő modellek és az egyéves adatokon épült modellek tanuló és tesztelő mintáinak felépítése

Nr.	Többéves modell		Egyéves modell	
	Tanuló	Tesztelő	Tanuló	Tesztelő
1.	2013-2014	2015	2014	2015
2.	2013-2015	2016	2015	2016
3.	2013-2016	2017	2016	2017
4.	2013-2017	2018-2019	2017	2018-2019
5.	2013-2019	2020-2021	2018-2019	2020-2021

Forrás: saját szerkesztés

Összesen 5-5 modellt illesztettem a többéves és a viszonyítási alapként szolgáló egyéves tanuló mintára. A többéves modellek legalább kettő- és legtöbb hét üzleti évig tartalmaznak adatokat a tanuló mintában. A többéves modellek tehát legkevesebb egy, legtöbb hat üzleti év adataival többet használhatnak tanulásra, mint az egyéves modellek. Tekintettel arra, hogy a 2018-2021 közötti időszakban relatív kevés adat állt a rendelkezésre, kétévenként összevontam az adatokat. Így alakult ki a 2018-2019 és a 2020-2021-es tesztelő minta készlet az utolsó két-két modellhez (a 2018-2019-es tanulásra is felhasznált állomány volt). Az utolsó modell egy további érdekes kérdésre is rávilágít. A COVID-19 válsággal sújtott 2020-2021-es időszak jelentős külső sokkot tartalmazott, amely a tanuló modellekbe nem épül be, de ugyanakkor a modellek előrejelző képességét próbára teszi. Ezért felmerül a kérdés, hogy ilyen feltételek mellett mennyire jól teljesítenek a hagyományos előrejelzési megközelítések.

A futtatott módszerek hasonlóan a korábbiakhoz a logisztikus regresszió, az SVM (lineáris és Gaussi), a neuronháló, a bagging, a random forest és a boosting volt. Az egyéves adatállományokon készült modellek paramétereit az F4. függelék A és B-

paneljében, valamint a többéves adatokon épített modelleket az *F6. függelékben* találjuk. A 11. táblázatban a tesztmintán elért eredmények láthatóak, az eredmények mérésére az osztályozási pontosságot használtam. Zöld színnel vannak kiemelve azok az értékek, amelyek egy adott módszer esetén magasabb besorolási arányt értek el az ellentétes mintavétellel szemben, valamint vörössel, ami alacsonyabb teljesítményű, mint a viszonyítási modell. A kiemelés ellenére a táblázat áttekinthetősége összetett és nehéz belőle egyértelmű mintázatokat kiolvasni.

11. táblázat: A többéves és egyéves modellek osztályozási eredményei

	Többéves modellek			Egyéves modellek		
	Tanuló minta 2013-2014, tesztelő minta 2015			Tanuló minta 2014, tesztelő minta 2015		
	Egészségesek helyes besorolási aránya	Fizetéképtelenek helyes besorolási aránya	Teljes helyes besorolási arány	Egészségesek helyes besorolási aránya	Fizetéképtelenek helyes besorolási aránya	Teljes helyes besorolási arány
Logit	83,78%	68,30%	76,04%	81,57%	79,85%	80,71%
SVM lineár	86,00%	66,09%	76,04%	82,06%	78,87%	80,47%
SVM RBF	85,50%	72,48%	78,99%	80,10%	82,31%	81,20%
Neuronháló	80,10%	79,12%	79,61%	81,33%	79,12%	80,22%
Bagging	82,56%	76,90%	79,73%	76,41%	77,45%	76,90%
Random forest	81,33%	78,13%	79,73%	75,43%	77,89%	76,66%
Adaboost	83,05%	75,92%	79,48%	75,68%	81,33%	78,50%
	Tanuló minta 2013-2015, tesztelő minta 2016			Tanuló minta 2015, tesztelő minta 2016		
Logit	90,16%	73,07%	81,62%	86,65%	73,77%	80,21%
SVM lineár	92,04%	71,19%	81,62%	9,13%	97,19%	53,16%
SVM RBF	74,00%	73,54%	73,77%	78,45%	71,43%	74,94%
Neuronháló	83,14%	81,73%	82,44%	80,09%	71,14%	75,64%
Bagging	83,84%	83,84%	83,84%	85,01%	82,44%	83,72%
Random forest	84,78%	83,37%	84,07%	84,07%	82,67%	83,37%
Adaboost	83,61%	81,50%	82,55%	86,65%	79,86%	83,26%
	Tanuló minta 2013-2016, tesztelő minta 2017			Tanuló minta 2016, tesztelő minta 2017		
Logit	85,46%	69,14%	77,30%	86,94%	73,29%	80,12%
SVM lineár	89,91%	69,44%	79,67%	35,61%	84,87%	60,24%
SVM RBF	84,87%	78,93%	81,90%	84,57%	74,48%	79,53%
Neuronháló	85,76%	81,01%	83,38%	82,79%	79,23%	81,01%
Bagging	83,68%	81,01%	82,34%	87,54%	77,45%	82,49%
Random forest	83,98%	81,31%	82,64%	87,54%	77,74%	82,64%
Adaboost	83,38%	80,71%	82,05%	87,24%	79,82%	83,53%
	Tanuló minta 2013-2017, tesztelő minta 2018/19			Tanuló minta 2017, tesztelő minta 2018/19		
Logit	95,78%	59,04%	77,41%	89,82%	67,66%	78,74%
SVM lineár	97,59%	54,22%	75,90%	4,79%	98,80%	51,80%
SVM RBF	90,96%	68,67%	79,82%	89,22%	71,86%	80,54%
Neuronháló	89,76%	72,89%	81,33%	89,82%	79,64%	84,73%
Bagging	93,37%	73,49%	83,43%	93,41%	77,84%	85,63%
Random forest	92,77%	73,49%	83,13%	92,22%	79,04%	85,63%
Adaboost	92,77%	75,30%	84,04%	93,41%	76,65%	85,03%
	Tanuló minta 2013-2019, tesztelő minta 2020/21			Tanuló minta 2018/19, tesztelő minta 2020/21		
Logit	96,51%	41,09%	68,80%	86,54%	64,62%	75,58%
SVM lineár	96,51%	42,25%	69,38%	87,69%	63,46%	75,58%
SVM RBF	95,74%	50,78%	73,26%	88,08%	67,69%	77,88%
Neuronháló	93,02%	51,94%	72,48%	77,69%	67,69%	72,69%
Bagging	93,02%	53,10%	73,06%	88,08%	62,31%	75,19%
Random forest	93,41%	53,88%	73,64%	86,92%	61,92%	74,42%
Adaboost	90,70%	52,71%	71,71%	85,38%	62,31%	73,85%

Forrás: saját számítás

Mindemellett, hogy a többéves modelleket győztes megoldásnak vártam, az adatok alapján nem tűnik ez magától értetődőnek. Sőt, az egyik talán legérdekesebb eredmény, ami elsőre szembetűnik, hogy az egyéves modellek pontosabban osztályozták a fizetésekteleneket az utolsó két modellnél (tanuló minta 2017 és 2018/19), ezáltal pedig az átlagos besorolási pontosságok is előnyösebbek lettek. Ezek a modellek mondhatni két éves előrejelzési horizonton teljesítették túl a több év adatain képzett modelleket mivel a tesztminta rendre 2018/19 és 2020/21 volt. Nem csupán a gazdaságilag normálisnak tekinthető 2018-2019 időszakra volt jó az egy év adataiból épült modellek előrejelzése, hanem a gazdaságilag kihívásokkal teli 2020-2021 évekre is előnyösebb eredmények mutatkoztak. Ugyanakkor a 2020-2021-re előrejelző összes modellnél jelentős teljesítménycsökkenés volt tapasztalható. A fizetésektelen megfigyelések 62%-67% körüli besorolási pontossága nem tekinthető színvonalasnak, viszont a többéves modell, a véletlen találgatásnak megfelelő, 50% közelében lévő eredményeket hozott, ami elfogadhatatlan.

12. táblázat: A többéves és egyéves modellek AUC értékei

Teszt-minta	Módszer	Modell		Teszt-minta	Módszer	Modell	
		Többéves	Egyéves			Többéves	Egyéves
2015	Logit	0,8330	0,8748	2016	Logit	0,8930	0,8805
	SVM lineár	0,8380	0,8723		SVM lineár	0,8929	0,8423
	SVM RBF	0,8515	0,8797		SVM RBF	0,8080	0,8346
	Neuronháló	0,8734	0,8680		Neuronháló	0,8990	0,8518
	Bagging	0,8821	0,8074		Bagging	0,9015	0,9090
	Random Forest	0,8866	0,8571		Random Forest	0,9068	0,9143
	Adaboost	0,8717	0,8564		Adaboost	0,9011	0,9129
2017	Logit	0,8533	0,8727	2018/19	Logit	0,8758	0,8842
	SVM lineár	0,8633	0,7211		SVM lineár	0,8766	0,8934
	SVM RBF	0,8834	0,8718		SVM RBF	0,8899	0,8990
	Neuronháló	0,9000	0,8764		Neuronháló	0,9031	0,9281
	Bagging	0,9013	0,9083		Bagging	0,9267	0,9186
	Random Forest	0,9044	0,9088		Random Forest	0,9302	0,9271
	Adaboost	0,9014	0,9049		Adaboost	0,9176	0,9267
2020/21	Logit	0,8211	0,8144				
	SVM lineár	0,8169	0,8216				
	SVM RBF	0,8618	0,8609				
	Neuronháló	0,8739	0,7731				
	Bagging	0,8762	0,8390				
	Random Forest	0,8721	0,8499				
	Adaboost	0,8654	0,7385				

Forrás: saját szerkesztés

Hasonlóan nehéz egyértelmű mintázatot kiolvasni az ROC-görbe alatti terület alapján mért teljesítménymutató értékeiből (12. táblázat). Kiemelten szerepeltettem az előnyösebb AUC értékeket az egyes módszerek mentén. Azt a megállapítást lehet tenni,

hogy átlagosan a közepes és jó teljesítményű modellek mellett, a neuronháló és az együttes (ensemble) módszereknél megjelenik 0,9 fellelti AUC érték is, ami kiváló teljesítményt takar.

13. táblázat: A többéves- és egyéves tanulómintával készült modellek eredményeinek összehasonlítása a Wilcoxon próbával

Teljesítménymutató		Negatív rangok gyakorisága	Pozitív rangok gyakorisága	Azonos rangok gyakorisága	Z-érték	p-érték
A-panel	Többéves modell, egészségesek helyes besorolása – Egyéves modell, egészségesek helyes besorolása	11	24	0	-3,260	0,001
	Többéves modell, fizetéseképtelenek helyes besorolása – Egyéves modell, fizetéseképtelenek helyes besorolása	23	11	1	-3,214	0,001
	Többéves modell, teljes helyes besorolási arány – Egyéves modell, teljes helyes besorolási arány	22	12	1	-0,915	0,360
	Többéves modell, AUC – Egyéves modell, AUC	17	18	0	-0,885	0,376
B-panel	Többéves modell, egészségesek helyes besorolása – Egyéves modell, egészségesek helyes besorolása	8	13	0	-1,547	0,122
	Többéves modell, fizetéseképtelenek helyes besorolása – Egyéves modell, fizetéseképtelenek helyes besorolása	9	11	1	-0,579	0,563
	Többéves modell, teljes helyes besorolási arány – Egyéves modell, teljes helyes besorolási arány	9	11	1	-0,859	0,390
	Többéves modell, AUC – Egyéves modell, AUC	11	10	0	-0,643	0,520

Megjegyzés: A-panel: tesztminta eredmények 2015-től 2020/21-ig. B-panel: tesztminta eredmények 2015-től 2017-ig.

Forrás: saját szerkesztés

Annak bizonyítására, hogy a több év adatain képzett modellek jobb teljesítményűek, ismét a statisztikai tesztek felé fordultam és összehasonlítottam a többéves modellek eredményeit az egyéves modellekével. Mivel az értékek egyik esetben sem voltak normális eloszlásúak ezért a T-próbát elvettem és a Wilcoxon páros rangpróbát futtattam. Ennek eredményei a 13. táblázat A-paneljében láthatóak, ahol az összes tesztminta eredményeit figyelembe vettem. A pozitív rangok gyakorisága mutatja, hogy hány módszernél volt a többéves modell magasabb eredményű, a negatív rangok előfordulása pedig az egyéves modell fölényét mutatja a többévessel szemben. A teszt p-értéke alapján megállapítható, hogy az egészséges megfigyelések helyes besorolási aránya a többéves modelleknél szignifikánsan magasabb volt, mint az egyéves modelleknél. Ezzel szemben a fizetéseképtelenek helyes osztályozási aránya átlagosan

szignifikánsan magasabb volt az egyéves modelleknél. Ugyanakkor a teljes képet nézve, az átlagos besorolási arány és az AUC értéke nem különbözött szignifikánsan a többéves modellek és egyéves modellek tekintetében. A várt eredmény ezzel ellentétes, a több év adatain épülő modell ereje csupán az egészségesek pontosabb felderítésében mutatkozik. A 13. táblázat B-panelje a kétéves előrejelzési horizont nélküli eredmények összemérését mutatja. A próbák p-értéke alapján azonban nincs szignifikáns különbség a teljesítménymutatókban a két mintavétel között.

Tekintsük meg közelebbről az eredményeket és módszerenként vizsgáljuk meg a két mintavétel különbségeit. Az F7. függelékben módszerenként ábrázolva láthatjuk a többéves tanuló eljárás és az egy év adataira támaszkodó tanuló eljárás előrejelzési pontosságainak a trendjét. Az egészségesek osztályozását tekintve a többéves modell előnye a logisztikus regresszió és a lineáris függvényformával készült SVM esetén érzékelhető. A többi módszernél az egészségesek magasabb arányú besorolása nem szembetűnő, a 2018/19-es tesztmintán a közel azonos pontosság után a 2020/21-es előrejelzésnél mutatkozott meg a többéves módszer hatékonyabb besorolása. Közel ennek az ellentéte látszik a fizetésképtelenek besorolását tekintve, ahol a logit és a lineáris SVM-nél egyértelmű az egyéves adatok alapján készült modellek felülteljesítése. A többi módszernél az utolsó két tesztmintán mutatkozott az egyéves tanulóminta előnye, így összeségében ezért kaptunk szignifikánsan magasabb értékeket, amely eredmény a 13. táblázat B-paneljében is igazolást nyert. A teljes osztályozási pontosság-trendvonalak pedig jelentősen egymásra simulnak, tehát vizuálisan sem különböznek egymástól az átlagos teljesítmények. A módszerek közül feltűnően szélsőséges eredményeket adott a lineáris SVM. A torzító hatását mérsékelhetjük, ha kihagyjuk az elemzésből, azonban a Wilcoxon próba újrafuttatásával, a 13. táblázatban korábban közölt konklúziókkal megegyező eredményre jutunk.

13. ábra: A többéves és egy év adatain képzett modellek AUC értékeinek alakulása



Forrás: saját szerkesztés

A modellek átlagos teljesítményét mérő AUC értékek trendjét a 13. ábrán láthatjuk, ahol a vízszintes tengelyen a tesztminta éve, a függőleges tengelyen az AUC érték egységesen skálázva szerepel. Éles különbséget így sem tudunk kimutatni a különböző tanuló mintával felépült modellek között. Mindössze olyan árnyalati elérések észlelhetők, hogy a neuronháló és az együttes módszereknél, a 2020/21-es tesztmintán jobb modelleket kaptunk a többéves adatok felhasználásával.

Az eredmények alapján az körvonalazódott, hogy az egyre bővülő adatállományból nem származik tanulási előny az adatok egyre nagyobb és egyre távolabbi halmozásával. A túl távoli adatok figyelembevétele irrelevánsnak mutatkozott és jelentős hozzáadott értéket nem képviselt az egyetlen üzleti év adatain felépült viszonyítási modellekhez mérten, azt leszámítva, hogy az egészséges vállalatok azonosításában egyes periódusokban eredményes volt. A csődelőrejelző modellek fontos célja a fizetéképtelen csoport minél hatékonyabb előrejelzése, azonban ebből a szempontból elmaradt a tesztadatokat közvetlenül megelőző egy év adatain képzett modellek teljesítményétől. Az előrejelezni kívánt időhorizonthoz legközelebb lévő egyetlen üzleti év adatainak a felhasználása bizonyos feltételek mellett hatékonyabb megoldásnak bizonyult. Azonban ha a két éves előrejelzési horizont (2018/19 és 2020/2021) eredményeit nem vesszük figyelembe akkor statisztikailag nem volt kimutatható különbség a két mintavétel között. Az eredmények tükrében a második hipotézist elutasítom. A több év adataiból épített modellek átlagos teljesítménye nem bizonyult kedvezőbbnek egyéves előrejelzési horizonton, mint az egyetlen üzleti év adatain épült előrejelző modellek.

6.3. A nem pénzügyi változók beépítése a modellekbe

Relatív kevés tanulmány foglalkozik olyan csődelőrejelző modellek építésével, amelybe a vállalaton kívüli, makrogazdasági, piaci és egyéb tényezőket vesznek számításba. Ennek legfőbb oka abban keresendő, hogy az ilyen típusú magyarázó változók egyáltalán nem, vagy legfeljebb kis mértékben járulnak hozzá a fizetéképtelenséget előrejelző modellek eredményéhez. Ehhez kapcsolódó hipotézisem a következő:

H3: A mezoszintű bemeneti adatok közül lesznek olyanok, amelyek szignifikáns hatással vannak a fizetéképtelenségre, azonban a hatásuk a pénzügyi mutatókkal szemben olyan kismértékű, hogy a mezoszintű változókat tartalmazó előrejelző modellek hasonló arányban képesek besorolni az egészséges és fizetéképtelen vállalatokat, mint a kizárólag vállalati pénzügyi adatokon alapuló fizetéképtelenséget előrejelző modellek.

Utalok a mezoszintű gazdasági skálára, amely alatt a vállalatok közvetlen működési környezetében érvényesülő gazdasági jellemzőket értem, specifikusan településszintű sajátosságokat. Ez a vállalati mikro szintű pénzügyi adatoknál egy fokkal

tágabb megközelítés, és ugyanakkor a nemzetgazdasági szintű, makrogazdasági megfigyelésektől szűkebb skálát jelent.

A mezoszintű modellek alapjául a korábban ismertetett azonos osztálylétszámú, az előrejelzési évet követlenül egy évvel megelőző adatokból tanuló modellek adták, amelyekhez hozzákapcsoltam a településszintű változókat. Az így létrejövő mezoszintű modellek az előzőleg kiválasztott pénzügyi változók mellé a vállalat működési környezetével összefüggő változókat tartalmaznak, amelyeket az *5.4 fejezetben* a magyarázó változók közt ismertettem. A viszonyítási modellek pedig a korábban bemutatott, településszintű adatok nélküli, azonos osztálylétszámú modellek voltak. A településszintű adatok nem álltak rendelkezésre minden megfigyelési évre, ezért összesen öt mezoszintű adatot tartalmazó modellt sikerült illesztetnem. Ezek tanuló- és tesztelő mintáinak évek szerinti eloszlása a *14. táblázatban* látható.

14. táblázat: A mezoszintű változókkal és kizárólag pénzügyi adatokkal épülő modellek tanuló- és tesztelő mintáinak felépítése

Nr.	Mezoszintű modell		Pénzügyi adatok modellje	
	Tanuló	Tesztelő	Tanuló	Tesztelő
1.	2013	2014	2013	2014
2.	2014	2015	2014	2015
3.	2015	2016	2015	2016
4.	2016	2017	2016	2017
5.	2017	2018	2017	2018

Forrás: saját szerkesztés

A módszerek továbbra is a logisztikus regresszió, az SVM (lineáris és Gaussi), a neuronháló, a bagging, a random forest és a boosting volt, amelyeknek a paramétereit az *F8. függelékben* láthatjuk. A mezoszintű magyarázó változók közül a változószelekciós eljárás során az egy vállalatra jutó árbevétel (Y1), a munkanélküliségi ráta (Y2), az egy főre jutó bankfiókok száma (Y3) és a település jogállása (Y7) mutató került be a modellekbe. A logisztikus regresszió szignifikánsnak azonosította a munkanélküliségi rátát az öt modelltől négyben, a hatás pedig mindig pozitív volt a fizetéseképtelenségre nézve, amely az előzetes várakozásoknak megfelelő. A munkanélküliségi ráta ceteris paribus növekedésével a felszámolás esélye is növekszik. Egyetlen modellben az egy vállalatra jutó árbevétel is szignifikáns negatív hatást jelzett a fizetéseképtelenségre. A két mutató átlag és medián értékeit tekintve (*15. táblázat*), kedvezőbb képet láthatunk az

egészséges osztályban mivel az egy vállalatra jutó árbevétel magasabb, a munkanélküliségi ráta alacsonyabb, mint a fizetésektelen osztály megfigyeléseinél.

15. táblázat: A különböző években az egy vállalatra jutó árbevétel (Y1) és munkanélküliségi ráta (Y2) átlaga, középértéke és szórása

Év	Mutató	Egészséges			Fizetésektelen		
		átlag	medián	szórás	átlag	medián	szórás
2013	Y1	55429	47329	50667	49803	32213	67009
	Y2	4,18%	2,98%	2,28%	4,99%	4,35%	2,76%
2014	Y1	60975	48701	71260	51188	29654	68800
	Y2	4,31%	3,00%	2,46%	5,21%	4,41%	3,30%
2015	Y1	66123	51711	62314	56349	36006	56636
	Y2	3,84%	2,61%	2,45%	4,36%	3,35%	3,16%
2016	Y1	61027	49242	52937	56315	36008	51924
	Y2	3,13%	2,15%	2,37%	3,50%	2,58%	2,63%
2017	Y1	65702	51441	61138	56264	38558	71528
	Y2	2,93%	2,15%	2,25%	3,32%	2,58%	2,44%

Megjegyzés: Y1 = Egy vállalatra jutó árbevétel (millió Ft), Y2 = Munkanélküliségi ráta.

Forrás: saját szerkesztés

A tesztmintán elért eredmények az *F9. függelékbe* kerültek, mivel nagy kiterjedésük mellett vizuálisan nem nyújtottak könnyen értékelhető eredményeket. Az egészségesek helyes besorolása tekintetében a mezoszintű változókkal készült modellek 73% átlagos pontosságot, a pénzügyi változók modelljei 79% pontosságot értek el. A fizetésektelenek osztályozásában szintén a pénzügyi változók modelljei teljesítettek átlagosan jobban 78% pontossággal, szemben a mezo-modellek 71%-os pontosságával. Így a teljes átlagos osztályozási eredmény a pénzügyi modelleknél 79%, a mezo-változókkal készült modelleknél 72% volt. A *16. táblázatban* az AUC értékek vannak feltüntetve, kiemelve zölddel azokat az értékeket, amelyek egy adott módszer esetén magasabbak voltak. Első ránézésre szembetűnik, hogy a kizárólag pénzügyi adatokkal készült modellek átlagosan jobbnak bizonyultak a mezoszintű mutatókat is alkalmazó modelleknél, előbbieik AUC átlagértéke 0,88 míg a mezoszintű változókkal épült modelleké 0,77. Egyedül a bagging eljárás tudta megragadni a mezoszintű változók információtartalmát, mivel több adathalmazon magasabb AUC értéket ért el, mint a csak pénzügyi adatokkal készült modell.

16. táblázat: A mezoszintű változókkal és kizárólag pénzügyi változókkal készült modellek AUC értékei

Minta	Módszer	Változók		Minta	Módszer	Változók	
		Mezo	Pénzügyi			Mezo	Pénzügyi
Tanuló 2013, tesztelő 2014	Logit	0,8483	0,8423	Tanuló 2014, tesztelő 2015	Logit	0,8382	0,8748
	SVM lineár	0,6625	0,8531		SVM lineár	0,6145	0,8723
	SVM RBF	0,5019	0,8819		SVM RBF	0,5381	0,8797
	Neuronháló	0,8563	0,8919		Neuronháló	0,8359	0,8680
	Bagging	0,9118	0,9110		Bagging	0,8125	0,8074
	Random Forest	0,9176	0,9170		Random Forest	0,8385	0,8571
	Adaboost	0,9211	0,9122		Adaboost	0,8295	0,8564
Tanuló 2015, tesztelő 2016	Logit	0,8810	0,8805	Tanuló 2016, tesztelő 2017	Logit	0,6408	0,7211
	SVM lineár	0,5854	0,8423		SVM lineár	0,5270	0,8718
	SVM RBF	0,5440	0,8346		SVM RBF	0,8435	0,8764
	Neuronháló	0,7741	0,8518		Neuronháló	0,9081	0,9083
	Bagging	0,9006	0,9090		Bagging	0,9091	0,9088
	Random Forest	0,9019	0,9143		Random Forest	0,9035	0,9049
	Adaboost	0,8876	0,9129		Adaboost	0,6408	0,7211
Tanuló 2017, tesztelő 2018/19	Logit	0,8790	0,8842				
	SVM lineár	0,4792	0,8934				
	SVM RBF	0,4886	0,8990				
	Neuronháló	0,4849	0,9281				
	Bagging	0,9205	0,9186				
	Random Forest	0,9230	0,9271				
	Adaboost	0,9165	0,9267				

Forrás: saját szerkesztés

Annak érdekében, hogy a különbségeket statisztikailag is megvizsgáljuk a páros Wilcoxon próbát alkalmaztam, a korábbi tesztek mintájára. Az eredmények a 17. táblázatban vannak összefoglalva. Az egészségesek besorolásában nincs statisztikailag igazolható különbség a csak pénzügyi és a mezoszintű változókkal készült modellek között. A fizetéképtelenek besorolásának p-értéket tekintve marginálisan szignifikáns a különbség a kizárólag pénzügyi mutatókat alkalmazó modellek javára (10%-os szignifikancia szinten). Azonban szigorúan véve 5%-os szignifikancia szinten nincs különbség a modellek között. Továbbá a teljes besorolási pontosság és az AUC szerinti modell teljesítmény tekintetében a pénzügyi adatokból épült modellek szignifikánsan jobbnak bizonyultak, mint a mezoszintű változókat is integráló modellek a Wilcoxon próba szerint.

17. táblázat: A mezoszintű változókkal és pénzügyi mutatókkal képzett modellek eredményeinek összehasonlítása a Wilcoxon próbával

Teljesítménymutató	Negatív rangok gyakorisága	Pozitív rangok gyakorisága	Azonos rangok gyakorisága	Z-érték	p-érték
Mezo, egészségesek helyes besorolása – Pénzügyi egészségesek helyes besorolása	17	18	0	-0,016	0,987
Mezo, fizetésképtelenek helyes besorolása – Pénzügyi, fizetésképtelenek helyes besorolása	22	13	0	-1,802	0,072
Mezo, teljes helyes besorolási arány – Pénzügyi, teljes helyes besorolási arány	25	10	0	-3,128	0,002
Mezo, AUC – Pénzügyi, AUC	26	9	0	-4,095	<0,001

Forrás: saját szerkesztés

A mezoszintű külső tényezők beépítésével a fizetésképtelenséget előrejelző modellek egy tágabb gazdasági kontextusban helyeződnek, azonban ez nem ad pontosabb képet a vállalatok csőd kockázatáról. Az eredmények tükrében a fizetésképtelenségnek vannak szignifikáns külső okai, mint a munkanélküliségi ráta, azonban ezek hatása az előrejelző teljesítményre elhanyagolható. A fizetésképtelenséget előrejelző modellek a felhasznált mezoszintű adatforrásokból nem tudtak olyan információt kinyerni, ami általánosan és egyértelműen hasznosabb lenne az előrejelzésre, mint ami a pénzügyi adatokból kinyerhető. A harmadik hipotézisem részben elfogadom. A mezoszintű magyarázó változók közt egyértelműen a munkanélküliségi rátát lehet kiemelni, mint a fizetésképtelenségre szignifikánsan ható tényező. A szignifikáns hatás ellenére a mezoszintű változókkal felépült modellek bizonyos teljesítménymutatók alapján átlagosan alulmaradnak a kizárólag vállalati pénzügyi adatokból felépült modellek előrejelzési teljesítményével szemben.

6.4. A gépi tanuláson alapuló módszerek összemérése

A nemzetközi csődelőjelzés szakirodalmát áttekintve nem találunk általánosan legjobb módszert, amely többek közt a gazdasági, pénzügyi, ipari és szabályozói komplexitásokból eredő különbözőségeket feloldja és a csődelőjelzés problémájára egyeduralkodó megoldást kínál. Az bizonyos, hogy a hagyományos statisztikai eljárásoktól elfordult a figyelem a gépi tanulási technikák felé, amelyek általában fokozottabb teljesítményt ajánlanak. A következőkben a fizetésképtelenség előrejelzésére

használatos gépi tanulás összehasonlító elemzését ismertetem magyar vállalati keretek közt, amely magába foglalja a neurális háló, az SVM módszerek, a bagging, a random forest és a boosting metamódszereket a viszonyítási logisztikus regresszió mellett. Ehhez kapcsolódó hipotézisem a következő:

H4: A gépi tanuláson alapuló együttes (ensemble) módszerek előrejelzési pontossága szignifikánsan magasabb, mint az egyéb (logisztikus regresszió, SVM, neurális háló) előrejelzési módszereké egy- és többéves előrejelzési horizonton.

18. táblázat: A többéves előrejelzési horizontokon tesztelt modellek AUC értékei

Minta	Módszer	AUC	Minta	Módszer	AUC	Minta	Módszer	AUC
tanuló 2013, tesztelő 2015	Logit	0,8265	tanuló 2013, tesztelő 2016	Logit	0,8662	tanuló 2014, tesztelő 2016	Logit	0,872
	SVM lineár	0,8377		SVM lineár	0,8754		SVM lineár	0,8722
	SVM RBF	0,8619		SVM RBF	0,8865		SVM RBF	0,8644
	Neuronháló	0,8658		Neuronháló	0,8879		Neuronháló	0,875
	Bagging	0,8839		Bagging	0,9013		Bagging	0,8323
	Random Forest	0,8905		Random Forest	0,9056		Random Forest	0,8873
	Adaboost	0,8829		Adaboost	0,9029		Adaboost	0,8879
tanuló 2014, tesztelő 2017	Logit	0,8608	tanuló 2015, tesztelő 2017	Logit	0,8637	tanuló 2015, tesztelő 2018/19	Logit	0,8631
	SVM lineár	0,8535		SVM lineár	0,8170		SVM lineár	0,8351
	SVM RBF	0,8555		SVM RBF	0,8305		SVM RBF	0,8631
	Neuronháló	0,8561		Neuronháló	0,8332		Neuronháló	0,8697
	Bagging	0,8338		Bagging	0,8915		Bagging	0,9238
	Random Forest	0,8733		Random Forest	0,8966		Random Forest	0,9261
	Adaboost	0,8678		Adaboost	0,9001		Adaboost	0,9385
tanuló 2016, tesztelő 2018/19	Logit	0,8598	tanuló 2016, tesztelő 2020/21	Logit	0,7764	tanuló 2017, tesztelő 2020/21	Logit	0,8551
	SVM lineár	0,5685		SVM lineár	0,6933		SVM lineár	0,8325
	SVM RBF	0,8946		SVM RBF	0,8139		SVM RBF	0,8218
	Neuronháló	0,8967		Neuronháló	0,8517		Neuronháló	0,8867
	Bagging	0,9161		Bagging	0,8869		Bagging	0,8379
	Random Forest	0,9174		Random Forest	0,8878		Random Forest	0,8741
	Adaboost	0,9085		Adaboost	0,8756		Adaboost	0,8730

Forrás: saját szerkesztés

A korábbi hipotézisek teszteléséhez több számszerűsíthető futtatás történt, így azok végső eredményét a módszerekre fókuszálva összefoglalom és kiegészítem az előrejelzési horizont kettő- illetve három éves kiterjesztésével. A hét módszer segítségével 140 modell 203 teljesítményértékét mérem össze, és bizonyos modelleknél a többéves előrejelzések eredményeit is számításba veszem. Ilyen tekintetben az azonos

osztálylétszámú mintán képzett egy- és többéves előrejelzések eredményét, a reprezentatív adalmintával készült előrejelzéseket, a többéves tanuló adatokon épült modellek előrejelzési teljesítményét és a mezoszintű változók alkalmazásával készült modellek eredményét következtetések levonása érdekében összesítem. A legtöbb modell eredményét korábban már ismertettem. Ezen felül a többéves előrejelzések eredményei az *F10. függelékben* kaptak helyet. A tesztelő minta kiválasztásának szempontja volt, hogy legalább két évvel későbbi legyen, mint a tanuló minta. A *18. táblázatban* a többéves előrejelzés AUC értékei láthatóak. Áttekintve a táblázatot az AUC értékekből nem tűnik ki, hogy a modellek osztályozása egyre kevésbé pontos lenne az előrejelzési évek eltolásával.

19. táblázat: Az együttes és egyéb modellek teljesítményének összehasonlítása egyéves előrejelzési horizonton a Mann-Whitney teszt segítségével

Teljesítménymutató	Módszer-csoport	Átlag	Szórás	Átlag rang	Mann-Whitney U	p-érték
Egészségesek helyes osztályozási aránya	Együttes	88,32%	6,41%	108,20	4506,50	0,193
	Egyéb	80,07%	24,29%	97,35		
Fizetéképtelenek helyes osztályozási aránya	Együttes	66,85%	22,43%	117,83	3668,50	0,001
	Egyéb	61,11%	26,53%	90,13		
Teljes átlagos helyes osztályozási arány	Együttes	83,06%	6,52%	127,26	2848,00	<0,001
	Egyéb	76,81%	12,68%	83,05		
AUC	Együttes	0,8841	0,0432	137,94	1919,50	<0,001
	Egyéb	0,8057	0,1161	75,05		

Megjegyzés: Együttes módszerek: bagging, random forest, adaboost. Egyéb módszerek: logisztikus regresszió, SVM, neuronháló.

Forrás: saját szerkesztés

Az összehasonlítandó módszereket két csoportba soroltam. Az első csoportba kerültek az együttes módszerek közé sorolható bagging, random forest és Adaboost. A második csoportba szerepelnek az egyéb, viszonyítási módszerek, köztük a logisztikus regresszió, a lineáris SVM, a gaussi függvényformával készült SVM és a neurális háló. Mivel az eredmények továbbra sem a normális eloszlást követik ezért a kétmintás t-próbát elvettem és a Mann-Whitney-féle nemparametrikus tesztet alkalmaztam. A nullhipotézise azt mondja ki, hogy a két sokaság, esetünkben az együttes és az egyéb módszerek eredménye, azonos eloszlásba tartozik. Az összes modell eredményeinek összehasonlító értékei a *19. táblázatban* láthatóak. Az egészségesek helyes osztályozási arányának tekintetében az együttes és alternatív módszerek nem különböznek egymástól

($p=0,193$). Azonban a fizetéképtelenek tekintetében az együttes módszerek szignifikánsan magasabb átlagértékkel rendelkeznek ezáltal pontosabbak, mint az egyéb módszerek átlagosan. Emellett a teljes osztályozási arány és az AUC mutatóval mért modell teljesítmény is szignifikánsan magasabb az együttes módszereknél ($p<0,001$).

A többéves előrejelzés eredményeinek összehasonlítását a 20. táblázatban foglaltam össze. Hasonlóan az egyéves előrejelzéshez, a többéves osztályozási képességben nincs különbség a két módszercsoport között az egészségesek helyes osztályozását tekintve. Emellett a fizetéképtelenek helyes osztályozási képességében sem mutatható ki eltérés a két módszercsoport között 5%-os szignifikancia szinten. Ugyanakkor egy átfogó képet nézve, az együttes modellek eredményessége megmutatkozik, mivel a teljes átlagos előrejelző pontosság valamint az AUC teljesítménymutató szignifikánsan magasabb, mint a viszonyítási modelleknél.

20. táblázat: Az együttes és egyéb modellek teljesítményének összehasonlítása többéves előrejelzési horizonton a Mann-Whitney teszt segítségével

Teljesítménymutató	Módszercsoport	Átlag	Szórás	Átlag rang	Mann-Whitney U	p-érték
Egészségesek helyes osztályozási aránya	Együttes	87,29%	5,65%	36,26	371,00	0,110
	Egyéb	78,56%	22,44%	28,81		
Fizetéképtelenek helyes osztályozási aránya	Együttes	72,86%	9,65%	36,50	364,50	0,091
	Egyéb	70,89%	10,80%	28,63		
Teljes átlagos helyes osztályozási arány	Együttes	80,07%	3,47%	41,83	220,50	<0,001
	Egyéb	74,73%	8,16%	24,63		
AUC	Együttes	0,8890	0,0264	44,80	140,50	<0,001
	Egyéb	0,8429	0,0598	22,40		

Megjegyzés: Együttes módszerek: bagging, random forest, adaboost. Egyéb módszerek: logisztikus regresszió, SVM, neuronháló.

Forrás: saját szerkesztés

A módszercsoportokat kibontva kaphatunk választ arra, hogy az átlagok mögött mely módszerek egyénileg kimagaslóak. Az egyes módszerek szintjére fókuszálva az átlagos teljesítmények a 21. táblázatban vannak összefoglalva. A zöld kiemelés az adott teljesítménymutatón mért legmagasabb érték, a piros pedig a legalacsonyabb. A logisztikus regresszió átlagosan az egészségesek felderítésében volt kiemelkedő a többi módszer közül. A lineáris kernellel készült SVM minden teljesítménymutatóban alulmaradt, ez azzal magyarázható, hogy a felszámolás osztályozási problémája lineárisan nehezen szeparálható. Ennél némileg jobb eredményt láthatunk a Gauss-féle

radiális bázis függvényekkel működő SVM esetén, azonban ez a várakozásaimtól még így is jelentősen elmarad. Meglepő az is, hogy mérsékelt eredmény figyelhető meg a neuronhálónál is. Azt vártam volna, hogy SVM RBF és a neuronháló minden tekintetben megelőzi a logisztikus regressziót. Az együttes módszerek az azonos jellegüknek köszönhetően nagyon hasonló átlagos teljesítményt értek el. Ha egyetlen győztes módszert kellene hirdetni, akkor az eredmények alapján a random forest-et emelném ki a többi közül.

21. táblázat: Az egyes módszerek teljesítményének átlaga és középértéke

Teljesítménymutató		Logisztikus regresszió	SVM lineáris	SVM RBF	Neuronháló	Bagging	Random forest	Adaboost
Egészségesek helyes osztályozási aránya	átlag	89,21%	68,32%	74,90%	87,83%	88,47%	88,08%	88,45%
	medián	86,94%	86,00%	85,50%	85,83%	87,57%	86,99%	88,43
Fizetéképtelenek helyes osztályozási aránya	átlag	60,36%	59,71%	63,70%	60,64%	66,91%	67,29%	66,34%
	medián	68,06%	68,54%	72,48%	71,14%	77,15%	77,74%	75,92%
Teljes helyes osztályozási arány	átlag	80,56%	70,54%	75,90%	80,25%	83,12%	83,12%	82,94%
	medián	78,49	75,58%	78,99%	79,94%	82,93%	82,94%	82,55%
AUC	átlag	0,8529	0,7647	0,7783	0,8268	0,8839	0,8931	0,8753
	medián	0,8631	0,8216	0,8515	0,8680	0,9006	0,9056	0,8989
Tanulási idő (másodperc)	átlag	7,93	5836,25	1097,82	3167,83	3151,93	3026,41	3165,32
	medián	7,77	1576,55	191,37	1065,15	987,36	1134,45	1135,85

Forrás: saját szerkesztés

A tanítási időt mint teljesítménymutatót figyelembe vettem az összehasonlítás során. A modellek tanítása egyetlen számítógépen azonos hardver és szoftver konfiguráció mellett történt, a párhuzamos futtatást mellőzve egymást követő sorozatokban. A 21. táblázatban az átlagos tanítási időket tekintve a statisztikai alapokon nyugvó logisztikus regresszió előnye rendkívüli, mindössze néhány másodperc alatt a legnagyobb adathalmazzal is megtanulja a modelleket. Ezzel szemben a lineáris kernellel működő SVM a legidőigényesebb, továbbá a neuronháló és az együttes módszerek is átlagosan jelentős időráfordítást igényeltek.

Arra a kérdésre, hogy a bemutatott két SVM erősen mérsékelt képességét jobban ki tudjuk-e aknázni a változók skálázásával, standardizálásával, a következők alapján igyekszek választ adni. A két módszert, a lineáris SVM-et és a Gaussi SVM-et az eddig ismertett modellspecifikációk mindegyikével megmértem standardizált és nem standardizált változók segítségével. Így tehát a reprezentatív adatmintán, az azonos adatmintán, a több évet magába foglaló tanulómintán, és a mezoszintű változók integrálásával készült modellek segítségével, egy- és többéves előrejelző teljesítményt

mértem le, összesen 58-at módszerenként, amelynek fele standardizált változókkal készült. Összehasonlítottam az egészségesek- valamint a fizetésektelenek helyes osztályozási arányát, az átlagos osztályozási pontosságot és az AUC értéket a Mann-Whitney-féle nemparametrikus teszt segítségével, amelynek eredményei a 22. táblázatban láthatóak.

22. táblázat: A standardizálás teljesítményre gyakorolt hatásának összehasonlítása a Mann-Whitney teszt segítségével

Módszer	Bemeneti változók	átlagérték			
		A	B	C	AUC
SVM lineáris	Standardizált	90,74%	58,55%	81,19%	0,83
	Nem standardizált	68,32%	59,71%	70,54%	0,76
	Mann-Whitney U	285,50	394,50	234,50	243,50
	p-érték	0,036	0,686	0,004	0,006
SVM RBF	Standardizált	87,20%	62,92%	81,45%	0,83
	Nem standardizált	74,90%	63,70%	75,90%	0,78
	Mann-Whitney U	374,00	418,00	323,00	319,00
	p-érték	0,469	0,969	0,129	0,114

Megjegyzés: A = Egészségesek helyes osztályozási aránya, B = Fizetésektelenek helyes osztályozási aránya, C = Teljes helyes osztályozási arány

Forrás: saját szerkesztés

A lineáris SVM módszer standardizált változókkal készült modelljeinél az egészségesek helyes osztályozási arányának átlaga közel 91%, ami a nyers – nem standardizált - változóknál mindössze 68%. A Mann-Whitney próba szerint szignifikáns különbség van a két csoport között ($p=0,036$). Ugyanez nem mondható el a fizetésektelenek besorolási pontosságánál, mivel nincs statisztikailag igazolható különbség a standardizált és nem standardizált változókkal készült modellek teljesítménye között. Sőt, a nyers adatokkal dolgozó modell átlagértéke több mint egy százalékponttal magasabb, mint a standardizált modell teljesítménye. A modellek átlagos teljesítményeit tükröző teljes helyes besorolási pontosság és AUC érték szignifikáns különbségét látjuk a p-érték alapján, úgy hogy a standardizált modelleknél magasabb átlagértékeket tapasztalunk.

A SVM RBF módszernél az egészségesek helyes osztályozása, a teljes besorolási pontosság valamint az AUC érték a standardizált modelleknél átlagosan magasabb. Azonban a fizetésektelenek helyes osztályozását tekintve a nem standardizált adatokkal

elért teljesítmény kedvezőbb. Ugyanakkor ezek a különbségek a megválasztott próba szerint statisztikailag nem kimutathatóak, tehát az SVM RBF módszernél nincs lényeges teljesítménybeli különbség abban, hogy a bemeneti változókat standardizáljuk vagy sem.

23. táblázat: A standardizálás tanítási időre gyakorolt hatásának összehasonlítása a Mann-Whitney teszt segítségével

Módszer	Bemeneti változók	Tanítási idő (másodperc)
SVM lineáris	Standardizált	2373,75
	Nem standardizált	5836,25
	Mann-Whitney U	138,00
	p-érték	0,096
SVM RBF	Standardizált	1069,42
	Nem standardizált	1097,82
	Mann-Whitney U	151,00
	p-érték	0,185

Forrás: saját szerkesztés

A továbbiakban a tanulási idő tekintetében mértem össze a standardizált és nyers adatokkal készült modelleket. A 23. táblázatban látható a két módszer és a különböző adatállományok átlagos tanítási ideje. A nem standardizált változókkal készült modelleknek átlagosan hosszabb tanulási időt mértem. Ezek közül a lineáris SVM betanítása az időigényesebb, azonban a Mann-Whitney próba alapján nincs szignifikáns különbség a standardizált vagy a nem standardizált változókkal készült modellek tanulási ideje között. A standardizálás előnyeinek kiaknázásával a lineáris SVM a módszer összehasonlítás során egyes teljesítménymutatóknál javít és előrébb mozdul a rangsor utolsó helyről, azonban összességében továbbra is korlátozott képességekkel rendelkezik a fizetésképtelenség előrejelzésére.

A módszercsoportok eredménye alapján az együttes módszercsaládba tartozó eljárások átlagosan jobban teljesítettek az adatokon, mint más alkalmazott módszerek. Ez annak köszönhető, hogy az együttes módszereknél több tanuló eljárás követi egymást, továbbá a döntési fa alapú módszerek hatékonyan megragadják a nem lineáris összefüggéseket azáltal, hogy a bemeneti jellemzőket több részre felosztják, és mindeniken előrejelzést végeznek. Az eredmények azt mutatják, hogy az előny némileg

mérséklődik a hosszabb távú előrejelzés tekintetében, de az átlagos modellteljesítmény így is magasabb az egyéb módszercsoporthoz viszonyítva.

Kiegészítendő, hogy a felsorakoztatott módszerek nem fedik le az összes potenciálisan szóba jöhető módszert a fizetéseképtelenség előrejelzésére, mindössze a szakirodalmi alkalmazásban a legelterjedtebbeket. Ezek mellett számos más, kiváló teljesítményt ajánló alternatív módszer szóba jöhet, azonban ezek érdemi alkalmazására a dolgozat keretei határt szabtak, ezért legfeljebb említés szinten jelentek meg más fejezetekben.

Az együttes módszerek mindegyike javasolt a fizetéseképtelenség kockázatának értékeléséhez, amelyek közül több teljesítménymutatót alapul véve kiemelhető a random forest eljárás. Hozzá kell tenni, hogy az alternatív módszerek közé sorolt logisztikus regresszió teljesítménye létjogosultságot biztosít a módszernek. Legkevesebb benchmark módszerként okvetlenül ajánlott, kifejezetten ha figyelembe vesszük, hogy a felépítése relatív egyszerű, a tanítási ideje nagyon gyors és az eredményei könnyen interpretálhatóak. Mindezek alapján az együttes módszerek magasabb előrejelző pontosságára vonatkozó hipotézist elfogadom.

7. Összegzés

Tanulmányom céljaként tűztem ki egy kellően széles, magyar vállalati adatmintára épülő, általános érvényű fizetésképtelenséget előrejelző modellek felállítását és értékelését. A felépült modellek nem csak az előrejelző módszerek megértését gazdagítják, hanem akár a pénzügyi- akár a kereskedelmi hitelminősítések gyakorlati alkalmazása során segítséget nyújthatnak. A vállalatok tulajdonosai, vezetői és érdekeltjei számára szintén eszköztár lehet a saját- vagy partner vállalat fizetőképességének felmérésére. A további célok közt szerepelt a fizetésképtelenség előrejelzés módszertani kérdéseinek tisztázása valamint újító jellegű előrelépések megtétele, főként olyan területen, ahol a szakirodalomban nincsenek egyértelmű utalások vagy ellentmondások mutatkoznak. Nem utolsó sorban pedig a gépi tanulás eszköztárának részletes bemutatása, gyakorlatba való ültetése és hasznosságának értelmezése volt a kutatás központi fókuszában.

A csődelőrejelzés szakirodalmában nincs egy egységesen elfogadott elmélet, sem egyetlen univerzális fogalom a csőd meghatározására. A tudományterület képviselői azonosan értelmezik a csőd fogalmát a vállalati fizetésképtelenség, a pénzügyi nehézség, a fizetés elmulasztása és a kudarc fogalmával. Ez alapvetően a kutatások célját, rendeltetését és eredményét érdemben nem befolyásolja, azonban a bizonyos fogalmak közötti finom eltéréseket érdemes szem előtt tartani. Ezért a tanulmány első részében a fizetésképtelenség és csőd keretrendszerét ismertettem, amelyben értelmezni tudjuk az előrejelzés szempontjait, szerepét és fontosságát.

Ezt követően a tanulmány terjedelmesebb része, a szakirodalmi áttekintés következett. Ez időrendi sorrendben a csődelőrejelzés és fizetésképtelenséghez kapcsolódó tanulmányok mérföldköveit ismerteti a kezdetektől a napjainkig. Mivel a csőd- és fizetésképtelenség előrejelzés egy nagyon gazdag és széles kutatási terület, számos mellékággal, ezért a témában írt összes tudományos írás ismertetése nem volt célja a dolgozatnak, hatékonysági és áttekinthetőségi szempontból sem. Mindössze az angol nyelven írt, magasan hivatkozott és jelentős hatású művek kerültek bemutatásra, különösen kiemelve az újszerű módszertani megoldásokat és azok lényeges elemeit. A témában folytatott, magyar nyelven írt kutatások jelentős részét ismertettem, amelyek tapasztalataihoz kapcsolódva és eredményeiből merítve épült a jelen dolgozat is.

A disszertáció harmadik fejezete a fizetésképtelenség módszereinek áttekintését adta, részletesen kifejtve a mesterséges intelligencia és azon belül a gépi tanulás fejlett

eszköztárát. A klasszikus módszerek mellett, az SVM, a neurális háló és a döntési fa módszercsaládba tartozó bagging, boosting és random forest módszerek működését és specifikációit ismerhettük meg, rávilágítva az előnyökre és az esetleges hátrányokra.

A következőkben összefoglalom a disszertáció során elért eredményeket. A 24. táblázat tartalmazza a megfogalmazott hipotéziseimet és az eredmények alapján hozott döntéseket.

24. táblázat: A vizsgált hipotézisek

Jelölés	Hipotézis	Eredmény
H1	A reprezentatív mintával készült fizetésképtelenséget előrejelző modell az azonos osztálylétszámú mintavétellel szemben kisebb hatékonyságú előrejelzésre képes a fizetésképtelenek helyes besorolási arányát tekintve.	Elfogadva
H2	A több év adataiból épített, fizetésképtelenséget előrejelző modellek átlagosan magasabb előrejelző teljesítményűek egyéves előrejelzési horizonton, mint az egyetlen üzleti év adatain épült modellek.	Elutasítva
H3	A mezoszintű bemeneti adatok közül lesznek olyanok, amelyek szignifikáns hatással vannak a fizetésképtelenségre, azonban a hatásuk a pénzügyi mutatókkal szemben olyan kismértékű, hogy a mezoszintű változókat tartalmazó előrejelző modellek hasonló arányban képesek besorolni az egészséges és fizetésképtelen vállalatokat, mint a kizárólag vállalati pénzügyi adatokon alapuló fizetésképtelenséget előrejelző modellek.	Részben elfogadva
H4	A gépi tanuláson alapuló együttes (ensemble) módszerek előrejelzési pontossága szignifikánsan magasabb, mint az egyéb (logisztikus regresszió, SVM, neurális háló) előrejelzési módszereké egy- és többéves előrejelzési horizonton.	Elfogadva

Forrás: saját szerkesztés

A szakirodalomban a mintavétel vitájához járul hozzá a disszertáció első hipotézisének az elfogadása. Gyakorlati okokból és történelmi hagyományokra támaszkodva sokan az egyenlő számosságú osztályok kialakítását szorgalmazzák. A másik tábor a fizetésképtelen és egészséges eloszlás alapján reprezentatív mintát javasol, de olyan is előfordul, hogy egy tetszőleges arányt határoz meg a kutató. Elemzésemben rámutattam, hogy a valós fizetésképtelen és egészséges vállalati eloszláson alapuló mintavétel nem hatékony a fizetésképtelenség előrejelzésére, ehelyett az egyenlő osztálylétszámú mintavétel javasolható, amely módszertani szempontból is előnyös. Ez a technika statisztikai szempontból éppenséggel felülreprezentálja a fizetésképteleneket, de egyúttal elegendő fizetésképtelen megfigyelést is biztosít a módszereknek ahhoz, hogy az egészségesek és a fizetésképtelenek jellemzőit meg tudja tanulni és ismeretlen adatmintán is el tudja választani.

Egy másik kihívása az előrejelzéseknek, hogy elegendő adat álljon a rendelkezésre. A második hipotézisemben azt fogalmaztam meg, hogy a több év pénzügyi adataira épülő modellek magasabb előrejelző pontosságot biztosítanak a nagyobb mennyiségű adatnak köszönhetően, mint az egy év adatain nyugvó modellek. Ezt a hipotézist azonban el kellett utasítanom, mivel azonos eredményeket tapasztaltam a fizetésképtelenséget közvetlenül megelőző egy év adatainak, valamint több összevont év adatainak előrejelző teljesítményében. Javaslatom szerint a kutatók, mellőzve a túlságosan távoli adatok integrálását, a legfrissebb, legrelevánsabb adatokra építsenek prediktív modellt, amely így hatékonyabb és relevánsabb eredményeket biztosít.

A harmadik hipotézisem bizonyítása során igyekeztem meghatározni, hogy a vállalatok belső pénzügyi magyarázó változóitól eltérő, mezoszintű input változók milyen mértékben növelik a modellek előrejelző erejét. Az eredmények pedig vegyes képet adtak. Egyrészt vannak olyan külső tényezők, amelyek magyarázzák a fizetésképtelenséget, például a munkanélküliségi ráta, ezek azonban nem járulnak hozzá az előrejelző teljesítményhez a vállalati pénzügyi mutatók mellett. Az eredmény magyarázatot ad arra, hogy a tágabb működési környezet változóinak integrálása a fizetésképtelenséget előrejelző modellekbe miért kevésbé gyakori a szakirodalomban.

A fejlett statisztikai technikák és gépi tanulási algoritmusok alkalmazásával számos modell létrejött, amelyek lehetővé tették, hogy egy objektív összehasonlító elemzést végezzek a módszerek tekintetében. A negyedik hipotézist igazoltam, amely szerint az együttes (ensemble) osztályozó módszerek számos különböző adatbázison és előrejelzési horizonton, több teljesítménymutatót figyelembe véve átlagosan jobban teljesítettek, mint a viszonyítási módszerek, köztük az SVM, a neuronháló és a logisztikus regresszió. Az együttes metamódszerek közül külön kiemelhető a random forest eljárás a fokozott teljesítményének köszönhetően.

Az ötven figyelembe vett pénzügyi magyarázó változó közül megközelítőleg 10-10 került felhasználásra modellenként. A bemeneti magyarázó változók számának meghatározására nincs szabály, azonban több szakirodalmat és elemzést alapul véve, az 5-15 közötti tartományt javasoltam. A kiválasztott magyarázó változó közül volt 3-4 olyan, amelyik stabilan majdnem mindenik előrejelző modellbe bekerült. A kkv-ra összpontosító elemzéseknél ajánlott a vállalat működésének évét tényezőként bevonni, emellett a likvid eszközök-, pénzeszközök arányának vizsgálata, valamint az eladósodottsági mutatók figyelembevétele is indokolt a fizetésképtelenség általános feltérképezésére.

További eredménye a disszertációnak, hogy a skálázásra érzékeny módszerek bemeneti értékeit célszerű standardizálásnak alávetni, ez kifejezetten a lineáris elválasztást megvalósító SVM-re érvényes. Ugyanakkor a magyarázó változók standardizálásával átlagosan mérsékeltebb tanítási idő érhető el, amelyet különösképp nem szoktak kiemelni a csődmodellek fejlesztésekor, azonban lényeges tényező. Vegyük figyelembe, hogy ha nagy mennyiségű adat mellett egy gépi tanulási algoritmus napokig vagy akár hetekig tanul, az egy hitelebírálnál kevésbé lesz hatékony az elvesztegetett idő tekintetében. A beérkezett adatok időben történő előrejelzésével optimálisabb döntések hozhatók, amelyek mérsékelhetik a költségeket. Nem utolsó sorban kutatói szempontból a rövidebb tanulási időtartam felgyorsítja a kísérletek ütemét, lehetőséget adva az újszerű és innovatív megoldásoknak.

Hangsúlyozni kell, hogy olyan általános érvényű előrejelző modelleket ismertetek, amelyek széles, többszörös adatmintán készültek, ezért az eredmények robusztusnak minősíthetőek. A 75%-83%-os átlagos osztályozási arány nem mondható szokatlanul magas értéknek, inkább egy átlagos jó teljesítményként értelmezhető. A mérsékelt előrejelző pontosság abból adódik, hogy az eredmények megbízhatóságára és a magyar kkv-k szempontjából az általánosíthatóságra törekedtem, amely kompromisszumokat jelentett a teljesítmény tekintetében. Minden bizonnyal, jobban specifikált, egy-egy iparágra finomhangolt modellek ennél magasabb osztályozási eredményt érhetnek el. A következtetéseim a fizetésképtelenség előrejelzésére vonatkozóan vontam le. Azonban az eredmények tükrében számos modellépítési stratégia finomhangolható, a tapasztalati eredmények akár más témájú dichotóm típusú döntési mechanizmushoz is adaptálhatóak.

7.1 A tanulmány korlátai és új kutatási irányok

A fizetésképtelenség kutatásnál az egyik sarkalatos pont az adatok elérhetősége. A jelen tanulmány kérdéseinek megválaszolásához az elmúlt évtized adatait vettem figyelembe. Ezzel együtt a legutolsó megfigyelési periódusokat, úgy mint a 2018-2019 és a 2020-2021 időszakot két évenként összevontan alkalmaztam, mivel nem állt rendelkezésre megfelelő mennyiségű adat. Ráadásul a legfrissebb adatok hiányában, a 2021 utáni évek vizsgálatát kénytelen voltam mellőzni. A big data korában is vannak olyan területek, mint a fizetésképtelenség előrejelzése, ahol nehéz hozzáférni a nagy mennyiségű és jó minőségű adathoz.

A tanulmány több módszert ismertet és validál, amelyek jó előrejelzési teljesítményre képesek, azonban nem meríti ki az ezekben rejlő összes kapacitásokat.

Nem beszélve az alternatív módszerek és hibrid megoldások teszteléséről. Az együttes módszereknél például a döntési fa alaptanuló helyett meg lehetne fontolni a neurális háló alkalmazását. Meglepő eredménye a tanulmánynak, hogy a legnépszerűbb módszerek közé sorolható neurális háló csupán mérsékelt teljesítményt ért el. Ennek fokozására a legkorszerűbb mélytanuló neuronháló specifikálása nyújthat megoldást, amelynek végrehajtása szintén javasolható későbbi kutatások témájának.

A tanulmányok túlnyomó többségében a fizetésképtelenség vagy csőd állapotát a jogi eljárások alapján határozzák meg. A jövőbeni kutatásoknál javasolt a pénzügyi nehézség más definícióját is vizsgálni, amely a jövedelmezőség mutatóin vagy a lehetőségekhez mérten piaci mutatókon alapszik. Ez pedig hatékonyabb megoldást is biztosíthat olyan vállalatok esetén, amelyek a fizetésképtelenség szélén állnak, de ellenük még nem indult meg a jogi eljárás.

A modellek bemeneti adatát képző változók az éves jelentésből nyert pénzügyi arányszámok. Az éves jelentésekben a mérleg csak egy adott pillanatban, az üzleti év végén érvényes értékeket mutatja, így az adott évben az állományok időbeli alakulása teljes mértékben ismeretlen. Továbbá az a kérdés is felmerülhet, hogy a kkv-k beszámolóit mennyire megbízhatóak, lehetnek-e benne hibák, torzítások, amelyek a végső következtetéseinket is elferdítik. Ugyanakkor egyelőre csak ezekre támaszkodhatunk, a pénzügyi mutatók kiváltására nincs más mód és úgy látszik, hogy nem is érdemes például külső, mezoszintű változókkal helyettesíteni. A külső, mezoszintű változók alkalmazásával a pénzügyi változóktól egy fokkal távolabbi perspektívából vizsgáltam a vállalatok túlélését. Ettől még távolabbi fókuszpont is elképzelhető, amelyekre szintén nincs közvetlen hatásuk a vállalatoknak, ugyanakkor e külső jelenségek és tényezők erőteljesen befolyásolhatják a vállalatok teljesítményét és pénzügyi egészségét. Ilyenek például a makrogazdasági események és sokkok, szabályzási változások vagy geopolitikai események.

Ajánlatos figyelembe venni a korlátozott számítási kapacitásokat is. A mai modern számítástechnikai eszközök kiváló teljesítményre képesek, azonban szem előtt kell tartani, hogy a gépi tanulás algoritmusai rendkívül időigényesek lehetnek, és jelentős erőforrást emésztenek fel. A több tízezres rekordok összetett gépi tanulási folyamatokra az otthoni vagy irodai használatra kiadott számítógépek kevésbé alkalmasak.

A tanulmány térbeli szempontból statikus, figyelmen kívül hagyja a földrajzi elhelyezkedést a fizetésképtelenség tekintetében. Jövőbeni kutatásoknál megfontolandó lehet a csőd és fizetésképtelenség vizsgálatát térbeli alapokra helyezni, kísérletet tenni a

földrajzi tér hatásának beágyazására az előrejelző modellekbe. Érdekes lehet megvizsgálni annak a lehetőségét, hogy a magyarázó változók hatásainak kiszűrése után marad-e térbeli függőség a közeli és távoli elhelyezkedésű vállalatok között a gazdasági-társadalmi interakcióknak köszönhetően. Ilyen irányba történtek már saját kezdeményezésű előrelépések, tágabb perspektívájú elemzések segítségével, egyrészt módszertani kifejtéssel a középpontban, másrészt a megszűnést vizsgáló rokon kutatási területen (Ágoston, 2023; Ágoston, 2024), azonban mikroszintű vizsgálatok még nem készültek.

Felhasznált irodalom

- Agarwal, V.** – Taffler, R. (2008): Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance*, Vol. 32. No. 8. pp. 1541-1551.
- Ágoston N.** (2022a): Külföldi csődelőrejelző módszerek szisztematikus irodalomelemzése. *Vezetéstudomány*, 53. évf. 1. sz. pp. 69-89.
- Ágoston N.** (2022b): Mesterséges intelligencia és gépi tanulási módszerek a vállalati fizetéseképtelenség becslésére. *Statisztikai Szemle*, 100. évf. 6. sz. pp. 584-609.
- Ágoston N.** (2022c): A kiegészítő melléklet szerepe a hitelezési döntéstámogatásban. In: Benke, Mariann; Schmuck, Roland; Riedelmayer, Bernadett (szerk.) *3. Farkas Ferenc Nemzetközi Tudományos Konferencia: „Menedzsment forradalmak”*, Pécsi Tudományegyetem Közgazdaságtudományi Kar Vezetés- és Szervezéstudományi Intézet, pp. 551-563.
- Ágoston N.** (2023): Spatial analysis of firm exit in European regions. *Journal of East European Management Studies*, The CEE Region on a New Wave of Transition, Special Issue, pp. 165-188.
- Ágoston N.** (2024): A többskálás földrajzilag súlyozott regresszió alkalmazása a településszintű vállalati felszámolások hazai vizsgálatához, 2019. *Területi Statisztika*, 64. évf. 2. sz. pp. 125-149.
- Agrawal, K.** – Maheshwari, Y. (2014): Default risk modelling using macroeconomic variables. *Journal of Indian Business Research*, Vol. 6. No. 4. pp. 270-285.
- Ahn, H.** – Kim, K. J. (2009): Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. *Applied Soft Computing*, Vol. 9. No.2. pp. 599-607.
- Alaka, H. A.** – Oyedele, L. O. – Owolabi, H. A. – Kumar, V. – Ajayi, S. O. – Akinade, O. O. – Bilal, M. (2018): Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, Vol. 94. pp. 164-184.
- Alfaro Cortés, E.** – Gámez Martínez, M. – García Rubio, N. (2007): Multiclass corporate failure prediction by Adaboost. M1. *International Advances in Economic Research*, Vol. 13. No. 3. pp. 301-313.

- Alfaro Cortés E.** – García Rubio, N. – Gámez Martínez M. – Elizondo, D. (2008): Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. *Decision Support Systems*, Vol. 45. No. 1. pp. 110-122.
- Ali, A.** – Daly, K. (2010): Macroeconomic determinants of credit risk: Recent evidence from a cross country study. *International Review of Financial Analysis*, Vol. 19. No. 3. pp. 165-171.
- Ali, S.** – Smith-Miles, K. A. (2006): Improved Support Vector Machine Generalization Using Normalized Input Space. In *Proceedings of the 19th Australian Joint Conference on Artificial Intelligence, AI 2006: Advances in Artificial Intelligence*, Hobart, Australia, pp. 362-371.
- Alibrahim, H.** – Ludwig, S. A. (2021): Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization. In *2021 IEEE Congress on Evolutionary Computation*, pp. 1551-1559.
- Altman, E. I.** (1968): Financial ratios, discriminant analysis and prediction of corporate bankruptcy. *The Journal of Finance*, Vol. 23. No. 4. pp. 589-609.
- Altman, E. I.** – Haldeman, R. G. – Narayanan, P. (1977): ZETA analysis. A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, No. 1. pp. 29-54.
- Altman, E. I.** (1993): *Corporate Financial Distress and Bankruptcy. A Complete Guide to Predicting and Avoiding Distress and Profiting from Bankruptcy*. Second Edition, John Wiley & Sons, New York, 384. p.
- Altman, E. I.** – Hotchkiss, E. – Wang, W. (2019): *Corporate Financial Distress, Restructuring, and Bankruptcy*. Fourth Edition, John Wiley & Sons, Hoboken, New Jersey, 368 p.
- Altman, E. I.** – Iwanicz-Drozdowska, M. – Laitinen, E. K. – Suvas, A. (2016): Financial and non-financial variables as long-horizon predictors of bankruptcy. *Journal of Credit Risk*, Vol. 12. No. 4.
- Altrichter M.** – Horváth G. – Pataki B. – Strausz G. – Takács G. – Valyon J. (2006): *Neurális hálózatok*. Panem Könyvkiadó, Budapest, 448 p.
- Andreano, M. S.** – Benedetti, R. – Mazzitelli, A. – Piersimoni, F. (2018): Spatial autocorrelation and clusters in modelling corporate bankruptcy of manufacturing firms *Economia e Politica Industriale*, Vol. 45. No. 4. pp. 475-491.

- Archer, K. J.** – Kimes, R. V. (2008): Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, Vol. 52. No. 4. pp. 2249-2260.
- Arcuri, G.** – Brunetto, M. – Levratto, N. (2019): Spatial patterns and determinants of firm exit: an empirical analysis on France. *The Annals of Regional Science*, Vol. 62. No.1. pp. 99-118.
- Arcuri, G.** – Levratto, N. (2020): Early stage SME bankruptcy: does the local banking market matter? *Small Business Economics*, Vol. 54. pp. 421-436.
- Asquith, P.** – Gertner, R. – Scharfstein, D. (1994): Anatomy of financial distress: An examination of junk-bond issuers. *The Quarterly Journal of Economics*, Vol. 109. No.3. pp. 625-658.
- Balcaen, S.** – Ooghe, H. (2004): *Alternative methodologies in studies on business failure: do they produce better results than the classic statistical methods?* Vlerick Leuven Ghent Working Paper Series 2004/16. Vlerick Leuven Ghent Management School, University of Ghent, Ghent.
- Baldwin, C. Y.** – Mason, S. P. (1983): The resolution of claims in financial distress the case of Massey Ferguson. *The Journal of Finance*, Vol. 38. No. 2. 505-516.
- Barboza, F.** – Kimura, H. – Altman, E. (2017): Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, Vol. 83. pp. 405-417.
- Barboza, F.** – Altman, E. (2024): Predicting financial distress in Latin American companies: A comparative analysis of logistic regression and random forest models. *The North American Journal of Economics and Finance*, Vol. 72, No. 102158. pp. 1-14.
- Basel Committee on Banking Supervision.** (2017): *Basel III: Finalising post-crisis reforms* [on-line]. Bank for International Settlements. www.bis.org (2024.01.18.)
- Bauer, E.** – Kohavi, R. (1999): An empirical comparison of voting classification algorithms: bagging, boosting and variants. *Machine Learning*, Vol. 36. No. 1. pp. 105-139.
- Bauer, J.,** – Agarwal, V. (2014): Are hazard models superior to traditional bankruptcy prediction approaches? A comprehensive test. *Journal of Banking & Finance*, Vol. 40. pp. 432-442.
- Beaver, W.** (1966): Financial ratios as predictors of failure. *Journal of Accounting Research*, Vol. 4. pp. 71-111.

- Beaver, W. H.** – McNichols, M. F. – Rhie, J. W. (2005): Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies*, Vol. 10. pp. 93-122.
- Bellovary, J.** – Giacomino, D. – Akers, M. (2007): A review of bankruptcy prediction studies: 1930-present. *Journal of Finance Education*, Vol. 33. pp. 1-42.
- Berg, D.** (2007): Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry*, Vol. 23. No. 2. pp. 129-143.
- Bergstra, J.** – Bengio, Y. (2012): Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, Vol. 13. No. 2. pp. 281-305.
- Berk, R. A.** (2020): *Statistical Learning from a Regression Perspective*. Third Edition, Cham, Switzerland, Springer Nature Switzerland.
- Berry, W. D.** – Feldman, S. – Stanley Feldman, D. (1985): *Multiple regression in practice* (No. 50). Sage Publications, Thousand Oaks, California.
- Beynon, M. J.** – Peel, M. J. (2001): Variable precision rough set theory and data discretisation: an application to corporate failure prediction. *Omega*, Vol. 29. No. 6. pp. 561-576.
- Bhattacharya, A.** (2022): *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more*. Packt Publishing Ltd., Birmingham, UK.
- Binder, H.** – Gefeller, O. – Schmid, M. – Mayr, A. (2014): The Evolution of Boosting Algorithms. *Methods of Information in Medicine*, Vol. 5. No. 06. pp. 419-427.
- Blum, A.** – Kalai, A. – Langford, J. (1999): Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 203-208.
- Bodon F.** – Buza K. (2014): *Adatbányászat* [on-line]. <https://dtk.tankonyvtar.hu/handle/123456789/12379> (2023.12.18.)
- Botta, M.** (2001): Resampling vs reweighting in boosting a relational weak learner. *Proceedings of the Advances in Artificial Intelligence: 7th Congress of the Italian Association for Artificial Intelligence*, Bari, Italy, September 25-28, 2001, pp. 70-80.
- Boritz, J. E.** – Kennedy, D. B. – Albuquerque, A. D. M. E. (1995): Predicting corporate failure using a neural network approach. *Intelligent Systems in Accounting, Finance and Management*, Vol. 4. No. 2. pp. 95-111.
- Box, M.** (2008): The death of firms: exploring the effects of environment and birth cohort on firm survival in Sweden, *Small Business Economics*, Vol. 31. No. 4. pp. 379-393.

- Brabazon, A.** – O'Neill, M. (2004): Diagnosing corporate stability using grammatical evolution. *International Journal of Applied Mathematics and Computer Science*, Vol. 14. No. 3. pp. 363-374.
- Brealey, R. A.** – Myers, S. C. (2005): *Modern vállalati pénzügyek*. Panem Könyvkiadó, Budapest, Hetedik kiadás, 1175 p.
- Breiman, L.** – Friedman, J. H. – Olshen, R. A. – Stone, C. J. (1984): *Classification and regression trees*. Florida, Chapman & Hall/CRC, Taylor & Francis Group.
- Breiman, L.** (1996): Bagging predictors. *Machine Learning*, Vol. 24. No. 2. pp. 123-140.
- Breiman, L.** (2001): Random forests. *Machine Learning*, Vol. 45. No. 1. pp. 5-32.
- Bruno, A. V.** – Leidecker, J. K. (1988): Causes of new venture failure: 1960s vs. 1980s. *Business horizons*, Vol. 31. No. 6. pp. 51-56.
- Bryant, S. M.** (1997): A case-based reasoning approach to bankruptcy prediction modeling. *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol. 6. No. 3. pp. 195-214.
- Brygala, M.** (2022): Consumer bankruptcy prediction using balanced and imbalanced data. *Risks*, Vol. 10. No. 2. pp. 1-13.
- Buehler, S.** – Kaiser, C. – Jaeger, F. (2012): The geographic determinants of bankruptcy: evidence from Switzerland, *Small Business Economics*, Vol. 39. No. 1. pp. 231-251.
- Burns, R.** – Burns, R. (2008): *Business Research Methods and Statistics using SPSS*, SAGE Publications, 560. p.
- Bujang, M. A.** – Sa'at, N. – Bakar, T. M. I. T. A. – Joo, L. C. (2018): Sample size guidelines for logistic regression from observational studies with large population: emphasis on the accuracy between statistics and parameters based on real life clinical data. *The Malaysian Journal of Medical Sciences*, Vol. 25. No. 4. pp. 122-130.
- Carton, R. B.** – Hofer, C. W. (2006): *Measuring organizational performance: Metrics for entrepreneurship and strategic management research*, Edward Elgar Publishing, Cheltenham, UK.
- Cecchini, M.** – Aytug, H. – Koehler, G. J. – Pathak, P. (2010): Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, Vol. 50. No. 1. pp. 164-175.
- Chandra, D. K.** – Ravi, V. – Bose, I. (2009): Failure prediction of dotcom companies using hybrid intelligent techniques. *Expert Systems with Applications*, Vol. 36. No. 3. pp. 4830-4837.

- Chaudhuri, A.** – De, K. (2011): Fuzzy support vector machine for bankruptcy prediction. *Applied Soft Computing*, Vol. 11. No. 2. pp. 2472-2486.
- Chava, S.** – Jarrow, R. A. (2004): Bankruptcy prediction with industry effects. *Review of Finance*, Vol. 8. No. 4. pp. 537-569.
- Chen, M. Y.** (2011a): Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. *Computers & Mathematics with Applications*, Vol. 62. No. 12. pp. 4514-4524.
- Chen, M. Y.** (2011b): Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems with Applications*, Vol. 38. No. 9. pp. 11261-11272.
- Chen, C. W.** – Tsai, Y. H. – Chang, F. R. – Lin, W. C. (2020): Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems*, Vol. 37. No. 5.
- Cho, S. H.** – Shin, K. S. (2023): Feature-weighted counterfactual-based explanation for bankruptcy prediction. *Expert Systems with Applications*, Vol. 216. pp. 1-14.
- Ciampi, F.** (2015): Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms. *Journal of Business Research*, Vol. 68. No. 5. pp. 1012-1025.
- Coats, P. K.** – Fant, L. F. (1993): Recognizing financial distress patterns using a neural network tool. *Financial Management*, Vol. 23, No. 4. pp. 142-155.
- Constand, L. R.** – Yazdipour, R. (2011): Firm failure prediction models: a critique and a review of recent developments. *Advances in Entrepreneurial Finance: With Applications from behavioral Finance and Economics*, Springer Science and Business Media, pp. 185-204.
- Cortes, C.** – Vapnik, V. N. (1995): Support-Vector Networks. *Machine Learning*, Vol. 20. No. 3. pp. 273-297.
- Cybinski, P.** (2000): The path to failure: where are bankruptcy studies at now. *Journal of Business and Management*, Vol. 7. No. 1. pp. 11-39. In Balcean – Ooghe (2014): *Alternative methodologies in studies of business failure: do they produce better results than the classic statistic methods?* Vlerick Leuven Ghent Working Paper Series 2004/16, University of Ghent, Ghent.
- de Menezes, F. S.** – Liska, G. R. – Cirillo, M. A. – Vivanco, M. J. (2017): Data classification with binary response through the Boosting algorithm and logistic regression. *Expert Systems with Applications*, Vol. 69. pp. 62-73

- Deák Sz.** (2000): A Porter-féle rombusz-modell főbb közgazdasági összefüggései. In: Lengyel I. (szerk): *Versenyképesség-regionális versenyképesség*. Szegedi Tudományegyetem Gazdaságtudományi Kar Közleményei, JATEPress, Szeged, pp. 67-87.
- Delen, D.** – Kuzey, C. – Uyar, A. (2013): Measuring firm performance using financial ratios: A decision tree approach. *Expert Systems with Applications*, Vol. 40. No. 10. pp. 3970-3983.
- Dietterich, T. G.** (1997): Machine-learning research. *AI magazine*, Vol. 18. No. 4. pp. 97-136.
- Dimitras, A. I.** – Slowinski, R. – Susmaga, R. – Zopounidis, C. (1999): Business failure prediction using rough sets. *European Journal of Operational Research*, Vol. 114. No. 2. pp. 263-280.
- du Jardin, P.** (2009): Bankruptcy prediction models: How to choose the most relevant variables? *Bankers, Markets & Investors*, No. 98. pp. 39-46.
- du Jardin, P.** (2012): The influence of variable selection methods on the accuracy of bankruptcy prediction models. *Bankers, Markets & Investors*, No. 116, pp. 20-39.
- du Jardin, P.** (2015): Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research*, Vol. 242. No. 1. pp. 286-303.
- du Jardin, P.** (2016): A two-stage classification technique for bankruptcy prediction. *European Journal of Operational Research*, Vol. 254. No. 1. pp. 236-252.
- du Jardin, P.** (2023): Designing topological data to forecast bankruptcy using convolutional neural networks. *Annals of Operations Research*, Vol. 325. No. 2. pp. 1291-1332.
- Dzsula M.** (2018): *A kényszertörlési eljárások szabályozásának jogértelmezési, jogalkalmazási problémái*. In: Szikora V. – Fézer T. (szerk.): *Fizetésképtelenség és jogutód nélküli megszűnés. Tapasztalatok és reformjavaslatok*. Debreceni Egyetem, Állam- és Jogtudományi Kar, Debrecen, pp. 31-56.
- Edmister, R. O.** (1972): An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitative Analysis*, Vol. 7. No. 2. pp. 1477-1493.
- Everett, J.** – Watson, J. (1998): Small business failure and external risk factors. *Small Business Economics*, Vol. 11. No. 1. pp. 371-390.
- Fan, A.** – Palaniswami, M. (2000): Selecting bankruptcy predictors using a support vector machine approach. In *Proceedings of the IEEE-INNS-ENNS International Joint*

- Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, Vol. 6. pp. 354-359.
- Fejér-Király G.** (2015): Bankruptcy prediction: A survey on evolution, critiques, and solutions. *Acta Universitatis Sapientiae, Economics and Business*, Vol. 3. No. 1. pp. 93-108.
- Fejér-Király G.** – Ágoston N. – Varga J. (2019): Modelling the financial failure of Romanian Stock companies. *Acta Universitatis Sapientiae Economics and Business*, Vol. 7. No. 1. pp. 65-82.
- Figini, S.** – Savona, R. – Vezzoli, M. (2016): Corporate default prediction model averaging: A normative linear pooling approach. *Intelligent Systems in Accounting, Finance and Management*, Vol. 23. pp. 6-20.
- Figlewski, S.** – Frydman, H. – Liang, W. (2012): Modeling the effect of macroeconomic factors on corporate default and credit rating transitions. *International Review of Economics & Finance*, Vol. 21. No. 1. pp. 87-105.
- Fitzpatrick, P.** (1932): *A comparison of the ratios of successful industrial enterprises with those of failed companies*. The Accountants' Publishing Company, Washington, USA.
- Fitzpatrick, P. J.** (1934): Transitional stages of a business failure. *The Accounting Review*, Vol. 9. No. 4. pp. 337-340.
- Fletcher, D.** – Goss, E. (1993): Forecasting with neural networks: an application using bankruptcy data. *Information & Management*, Vol. 24. No. 3. pp. 159-167.
- Fotopoluos, G.** – Louri, H. (2000): Location and survival of new entry. *Small Business Economics*, Vol. 14. pp. 311-321
- Freund, Y.** – Schapire, R. E. (1996): Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning*, Bari, Italy.
- Friedman, J. H.** (2001): Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, Vol. 29. No. 5. pp. 1189-1232.
- Frydman, H.** – Altman, E. I. – Kao, D. L. (1985): Introducing recursive partitioning for financial classification: the case of financial distress. *The Journal of Finance*, Vol. 40. No. 1. pp. 303-320
- Garson, D. G.** (1991): Interpreting neural network connection weights. *Artificial Intelligence Expert*, Vol. 6. No. 7. pp. 47-51.

- Giannopoulos, G.** – Sardar, S. A. – Salti, R. – Sykianakis, N. (2022): Analyzing Insolvency Prediction Models in the Period Before and After the Financial Crisis: A Case Study on the Example of US Firms. *International Journal of Finance, Insurance and Risk Management*, Vol. 12. No. 3. pp. 23-45.
- Gombola, M. J.** – Haskins, M. E. – Ketz, J. E. – Williams, D. D. (1987): Cash flow in bankruptcy prediction. *Financial Management*, Vol. 16. No. 4. pp. 55-65.
- Greenwald, D.** (1973): *The McGraw-Hill Dictionary of Modern Economics: A Handbook of Terms and Organizations*. McGraw-Hill Book Company, New York, USA, 792 p.
- Gupta, R.** – Srivastava, D. – Sahu, M. – Tiwari, S. – Ambasta, R. K. – Kumar, P. (2021): Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity*, Vol. 25. pp. 1315-1360.
- Gupta, V.** – Liao, W. K. – Choudhary, A. – Agrawal, A. (2023): Evolution of artificial intelligence for application in contemporary materials science. *MRS Communications*, Vol. 13. No. 5. pp. 754-763.
- Guyon, I.** – Elisseeff, A. (2003): An introduction to variable and feature selection. *Journal of Machine Learning Research*, Vol. 3. pp. 1157-1182.
- Hajdu O.** – Virág M. (2001): A Hungarian Model for Predicting Financial Bankruptcy. *Society and Economy in Central and Eastern Europe*, Vol. 23. No. 1-2. pp. 28-46.
- Hajdu O.** (2004): A csődesemény logit-regressziójának kismintás problémái. *Statisztikai Szemle*, 82. évf. 4. sz. pp. 392-422.
- Hámori G.** (2001a): A CHAID alapú döntési fák jellemzői. *Statisztikai Szemle*, 79. évf. 8. sz. pp. 703-710.
- Hámori G.** (2001b): A fizetésektelenség előrejelzése logit-moddal. *Bankszemle*, 45. évf. 1-2. sz. pp. 65-87.
- Han, J.** – Kamber, M. (2006): *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, New York, USA.
- Helmecki I.** (2010): *A magyarországi pénzforgalom térképe*, MNB-tanulmányok 84. Magyar Nemzeti Bank, Budapest.
- Heo, J.** – Yang, J. Y. (2014): AdaBoost based bankruptcy forecasting of Korean construction companies. *Applied Soft Computing*, Vol. 24. pp. 494-499.
- Hernandez Tinoco, M.** – Wilson, N. (2013): Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis*, Vol. 30. pp. 394-419.

- Hillegeist, S. A.** – Keating, E. K. – Cram, D. P. – Lundstedt, K. G. (2004): Assessing the probability of bankruptcy. *Review of Accounting Studies*, Vol. 9. No. 1. pp. 5-34.
- Hol, S.** (2007): The influence of the business cycle on bankruptcy probability. *International Transactions in Operational Research*, Vol. 14. No. 1. pp. 75-90.
- Holmes, P.** – Hunt, A. – Stone, I. (2010): An analysis of new firm survival using a hazard function. *Applied Economics*, Vol. 42, No. 2. pp. 185-195.
- Hopwood, W.** – McKeown, J. – Mutchler, J. (1989): A test of the incremental explanatory power of opinions qualified for consistency and uncertainty. *Accounting Review*, Vol. 64. No. 1. pp. 28-48.
- Horváth G.** (2021): Vállalatok hitelkockázati modellezése a Magyar Nemzeti Bank felügyeleti stressztesztjében. *Hitelintézeti Szemle*, 20. évf. 1. sz. pp. 43-73.
- Hua, Z.** – Wang, Y. – Xu, X. – Zhang, B. – Liang, L. (2007): Predicting corporate financial distress based on integration of support vector machine and logistic regression. *Expert Systems with Applications*, Vol. 33. No. 2. pp. 434-440.
- Hunyadi L.** (2012): Tibshirani, R.: Regressziós szűkítés és változószelekció a LASSO segítségével: visszatekintés. *Statisztikai Szemle*, 90. évf. 1. sz. pp. 107-109.
- Ilmakunnas, P.** – Topi, J. (1999): Microeconomic and macroeconomic influences on entry and exit of firms. *Review of Industrial Organization*, Vol. 15. No. 3. pp. 283-301.
- James, G.** – Witten, D. – Hastie, T. – Tibshirani, R. (2013): *An introduction to Statistical Learning with Applications in R*. Springer Science+Business Media, New York, USA, 426 p.
- Jo, H.** – Han, I. – Lee, H. (1997): Bankruptcy prediction using case-based reasoning, neural network and discriminant analysis for bankruptcy prediction. *Expert Systems with Applications*, Vol. 13. No. 2. pp. 97-108.
- Jones, S.** – Hensher, D. A. (2004): Predicting firm financial distress: A mixed logit model. *The Accounting Review*, Vol. 79. No. 4. pp. 1011-1038.
- Karas, M.** (2022): The Hazard Model for European SMEs: Combining Accounting and Macroeconomic Variables. *Journal of Competitiveness*, Vol. 14. No. 3. pp. 76-92.
- Kass, G. V.** (1980): An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol. 29. No. 2. pp. 119-127.
- Keasey, K.** – Watson, R. (1991): Financial distress prediction models: a review of their usefulness. *British Journal of Management*, Vol. 2. No. 2. pp. 89-102.

- Keeble, D.** – Walker, S. (1994): New firms, small firms and dead firms: spatial patterns and determinants in the United Kingdom. *Regional Studies*, Vol. 28. No. 4. pp. 411-427.
- Kennedy, K.** – Namee, B. M. – Delany, S. J. (2013): Using semi-supervised classifiers for credit scoring. *Journal of the Operational Research Society*, Vol. 64. No. 4. pp. 513-529.
- Kerek Z.** (1992): *Csődeljárás, felszámolás, reorganizáció*. Közgazdasági és Jogi Könyvkiadó, Budapest, 175 p.
- Khoja, L.** – Chipulu, M. – Jayasekera, R. (2016): Analysing corporate insolvency in the Gulf Cooperation Council using logistic regression and multidimensional scaling. *Review of Quantitative Finance and Accounting*, Vol. 46. pp. 483-518.
- Kim, M. J.** – Kang, D. K. (2010): Ensemble with neural networks for bankruptcy prediction. *Expert Systems With Applications*, Vol. 37, No. 4. pp. 3373-3379.
- Kim, H. S.** – Sohn, S. Y. (2010): Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research*, Vol. 201. No. 3. pp. 838-846.
- Kim, S. Y.** (2011): Prediction of hotel bankruptcy using support vector machine, artificial neural network, logistic regression, and multivariate discriminant analysis. *The Service Industries Journal*, Vol. 31. No. 3. pp. 441-468.
- Kim, S. Y.** – Upneja, A. (2014): Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. *Economic Modelling*, Vol. 36. pp. 354-362.
- Kim, H.** – Cho, H. – Ryu, D. (2022): Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data. *Computational Economics*, Vol. 59. No. 3. pp. 1231-1249.
- Kirkos, E.** (2015): Assessing methodologies for intelligent bankruptcy prediction. *Artificial Intelligence Review*, Vol. 43. pp. 83-123.
- Korol, T.** (2013): Early warning models against bankruptcy risk for Central European and Latin American enterprises. *Economic Modelling*, Vol. 31. pp. 22-30.
- Kovács E.** (2014): *Többváltozós adatelemzés*. Typotex Kiadó, Budapest, 243. p.
- Központi Statisztikai Hivatal** (2014): Statisztikiai Tükör. *A regisztrált gazdasági szervezetek száma, 2013* [on-line]
<https://www.ksh.hu/docs/hun/xftp/gyor/gaz/gaz21312.pdf> (2021.06.01.)

- Központi Statisztikai Hivatal** (2015): Statisztikai Tükör. *A regisztrált gazdasági szervezetek száma, 2014* [on-line]
<https://www.ksh.hu/docs/hun/xftp/gyor/gaz/gaz21412.pdf> (2021.06.01.)
- Központi Statisztikai Hivatal** (2016): Statisztikai Tükör. *A regisztrált gazdasági szervezetek száma, 2015* [on-line]
<https://www.ksh.hu/docs/hun/xftp/gyor/gaz/gaz1512.pdf> (2021.06.01.)
- Központi Statisztikai Hivatal** (2017): Statisztikai Tükör. *A regisztrált gazdasági szervezetek száma, 2016* [on-line]
<https://www.ksh.hu/docs/hun/xftp/gyor/gaz/gaz1612.pdf> (2021.06.01.)
- Központi Statisztikai Hivatal** (2018): Statisztikai Tükör. *A regisztrált gazdasági szervezetek száma, 2017* [on-line]
<https://www.ksh.hu/docs/hun/xftp/gyor/gaz/gaz1712.pdf> (2021.06.01.)
- Központi Statisztikai Hivatal** (2019): Statisztikai Tükör. *A regisztrált gazdasági szervezetek száma, 2018* [on-line]
<https://www.ksh.hu/docs/hun/xftp/gyor/gaz/gaz1812.pdf> (2021.06.01.)
- Központi Statisztikai Hivatal** (2022): A dobozdiagram magyarázata [on-line]
<https://www.ksh.hu/s/kiadvanyok/termofoldarak-es-berleti-dijak-2022/boxplot.html>
 (2024.03.22)
- Kristóf T.** (2005): Szervezetek jövőbeni fennmaradása különböző megközelítésekben. *Vezetéstudomány*, 36. évf. 9. sz. pp. 15-23.
- Kristóf T.** (2008): A csődelőrejelzés és a nem fizetési valószínűség számításának módszertani kérdéseiről. *Közgazdasági Szemle*, 55. évf. 5. sz. pp. 441-461.
- Kristóf T.** (2018): A case-based reasoning alkalmazása a hazai mikrovállalkozások csődelőrejelzésére. *Statisztikai Szemle*, 96. évf. 11-12 sz. pp. 1109-1128.
- Kristóf T. – Virág M.** (2019): A csődelőrejelzés fejlődéstörténete Magyarországon. *Vezetéstudomány*, 50. évf. 12. sz. pp. 62-73.
- Kristóf T. – Virág M.** (2020): A comprehensive review of corporate bankruptcy prediction in Hungary. *Journal of Risk and Financial Management*. Vol. 13. No. 2. pp. 1-20.
- Kristóf T. – Virág M.** (2023): Failure Prediction of Visegrad Four Large Corporates–Is History Repeating Itself in the COVID-19 Era? *SSRN Electronic Journal*, pp. 1-16.
- Kruppa, J. – Schwarz, A. – Armingier, G. – Ziegler, A.** (2013): Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, Vol. 40. No. 13. pp. 5125-5131.

- Kubat, M.** (2017): *An introduction to machine learning*. Springer International Publishing, Cham, Switzerland, Second Edition, 348 p.
- Lam, M** (2004): Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. *Decision Support Systems*, Vol. 37. No. 4. pp. 567-581.
- Lau, A. H. L.** (1987): A five-state financial distress prediction model. *Journal of Accounting Research*, Vol. 25. No. 1. pp. 127-138.
- Lázár E.** (2022): *Közgazdasági kutatómódszertan*. Risoprint Kiadó, Kolozsvár, 202 p.
- Lee, K.** – Booth, D. – Alam, P. (2005): A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. *Expert Systems with Applications*, Vol. 29. No. 1. pp. 1-16.
- Leshno, M.** – Spector, Y. (1996): Neural network prediction analysis: The bankruptcy case. *Neurocomputing*, Vol. 10. No. 2. pp. 125-147.
- Li, H.** – Sun, J. (2008): Ranking-order case-based reasoning for financial distress prediction. *Knowledge-Based Systems*, Vol. 21. No. 8. pp. 868-878.
- Li, H.** – Sun, J. (2011): On performance of case-based reasoning in Chinese business failure prediction from sensitivity, specificity, positive and negative values. *Applied Soft Computing*, Vol. 11. No. 1. pp. 460-467.
- Li, H.** – Sun, J. (2012): Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples—Evidence from the Chinese hotel industry. *Tourism Management*, Vol. 33. No. 3. pp. 622-634.
- Li, S.** – Wang, M. – He, J. (2013): Prediction of banking systemic risk based on support vector machine. *Mathematical Problems in Engineering*, Article ID 136030.
- Li, H.** – Yu, J. L. – Yu, L. A. – Sun, J. (2014): The clustering-based case-based reasoning for imbalanced business failure prediction: a hybrid approach through integrating unsupervised process with supervised process. *International Journal of Systems Science*, Vol. 45. No. 5. pp. 1225-1241.
- Li, D.** – Kanoulas, E. (2018): Bayesian optimization for optimizing retrieval systems. In: Yi, C. – Chengxiang Z. (eds.): *WSDM'18. Association for Computing Machinery*. New York, USA, pp. 360-368.
- Liang, D.** – Lu, C. C. – Tsai, C. F. – Shih, G. A. (2016): Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, Vol. 252. No. 2. pp. 561-572.

- Lifschutz, S.** – Jacobi, A. (2010): Predicting bankruptcy: evidence from Israel. *International Journal of Business and Management*, Vol. 5. No. 4. pp. 133-141.
- Lim, T.** – Lim Xiu Yun, J. – Siwei, G. – Jiang, H. (2012): Bankruptcy prediction: theoretical framework proposal. *International Journal of Management Sciences and Business Research*, Vol. 1. No. 9. pp. 69-74.
- Lin, T. H.** (2009): A cross model study of corporate financial distress prediction in Taiwan: Multiple discriminant analysis, logit, probit and neural networks models. *Neurocomputing*, Vol. 72. No. 16-18. pp. 3507-3516.
- Lin, W. C.** – Lu, Y. H. – Tsai, C. F. (2019): Feature selection in single and ensemble learning-based bankruptcy prediction models. *Expert Systems*, Vol. 36. No.1.
- Lizal, L.** (2002): Determinants of financial distress: What drives bankruptcy in a transition economy? The Czech Republic case. *William Davidson Working Paper*, Number 451.
- Loh, B. W.-Y.** (2009): Improving the precision of classification trees. *The Annals of Applied Statistics*, Vol. 3. No. 4. pp. 1710-1737.
- Lombardo, G.** – Pellegrino, M. – Adosoglou, G. – Cagnoni, S. – Pardalos, P. M. – Poggi, A. (2022): Machine Learning for Bankruptcy Prediction in the American Stock Market: Dataset and Benchmarks. *Future Internet*, Vol. 14. No. 8. pp. 1-23.
- Lundberg, S. M.** – Lee, S. I. (2017): A unified approach to interpreting model predictions. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA.
- Luor, D.-C.** (2015): A comparative assessment of data standardization on support vector machine for classification problems. *Intelligent Data Analysis*, Vol. 19. No. 3. pp. 529-546.
- Lussier, R. N.** (1995): A nonfinancial business success versus failure prediction model for young firms. *Journal of Small Business Management*, Vol. 33. No. 1. pp. 8-20.
- Mahmood, Z.** – Khan, S. (2009): On the use of k-fold cross-validation to choose cutoff values and assess the performance of predictive models in stepwise regression. *The International Journal of Biostatistics*. Vol. 5. No. 1. pp. 1-21.
- Mai, F.** – Tian, S. – Lee, C. – Ma, L. (2019): Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, Vol. 274. No. 2. pp. 743-758.

- Manh, P. T.** – Nguyen, H. V. (2024): Applying the Z-Score Model to Predict Corporate Financial Distress: An Empirical Research on the Listed Firms in Vietnam Stock Market. *Oblik i finansi*, Vol. 1. pp. 38-48.
- Mansi, S.** – Maxwell, W. – Zhang, A. (2010): Bankruptcy prediction models and the cost of debt. *The Journal of Fixed Income*, Vol. 21. No. 4. pp. 25-42.
- Mar Molinero, C.** – Ezzamel, M. (1991): Multidimensional scaling applied to corporate failure. *Omega*, Vol. 19. No. 4. pp. 259-274.
- Mar Molinero, C.** – Serrano-Cinca, C. (2001): Bank failure: a multidimensional scaling approach. *The European Journal of Finance*, Vol. 7. No. 2. pp. 165-183.
- Marcot, B. G.** – Hanea, A. M. (2021): What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? *Computational Statistics*, Vol. 36. No. 3. pp. 2009-2031.
- Mare, D. S.** (2015): Contribution of macroeconomic factors to the prediction of small bank failures. *Journal of International Financial Markets, Institutions and Money*, Vol. 39. pp. 25-39.
- Marques, A. I.** – Garcia, V. – Sanchez, J. S. (2012): Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, Vol. 39. No. 11. pp. 10244-10250.
- Martin, D.** (1977): Early warning of bank failure: A logit regression approach. *Journal of banking & finance*, Vol. 1. No. 3. pp. 249-276.
- Martin-Del-Brio, B.** – Serrano-Cinca, C. (1993): Self-organizing neural networks for the analysis and representation of data: Some financial cases. *Neural Computing & Applications*, Vol. 1. No. 3. pp. 193-206.
- Mate-Sánchez-Val, M.** (2021): The impact of geographical positioning on agri-food businesses' failure considering nonlinearities. *Agribusiness*, Vol. 37. No. 3. pp. 612-628.
- Mathai, N.** – Chen, Y. – Kirchmair, J. (2020): Validation strategies for target prediction methods. *Briefings in Bioinformatics*, Vol. 21. No. 3. pp. 791-802.
- Mayew, W. J.** – Sethuraman, M. – Venkatachalam, M. (2015): MD&A disclosure and the firm's ability to continue as a going concern. *The Accounting Review*, Vol. 9. No. 4. pp. 1621-1651.
- McCulloch, W. S.** – Pitts, W. (1943): A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, Vol. 5. pp. 115-133.

- McKee, T. E.** (2000): Developing a bankruptcy prediction model via rough sets theory. *Intelligent Systems in Accounting, Finance & Management*, Vol. 9. No. 3. pp. 159-173.
- McKee, T. E.** – Greenstein, M. (2000): Predicting bankruptcy using recursive partitioning and a realistically proportioned data set. *Journal of Forecasting*, Vol. 19. No. 3. pp. 219-230.
- McLeay, S.** – Omar, A. (2000): The sensitivity of prediction models to the non-normality of bounded and unbounded financial ratios. *British Accounting Review*, Vol. 32. No. 2. pp. 213-230.
- Min, J. H.** – Jeong, C. (2009): A binary classification method for bankruptcy prediction. *Expert Systems with Applications*, Vol. 36. No. 3. pp. 5256-5263.
- Min, J. H.** – Lee, Y. C. (2005): Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems With Applications*, Vol. 28. No. 4. pp. 603-614.
- Min, J. H.** – Lee, Y. C. (2008): A practical approach to credit scoring. *Expert Systems with Applications*, Vol. 35. No. 4. pp. 1762-1770.
- Molnár P.** (2019): Az önkormányzati vállalatok működőképessége megyei összehasonlításban. *Területi Statisztika*, 59. évf. 3. sz. pp. 273-299.
- Murthy, S. K.** – Kasif, S. – Salzberg, S. – Beigel, R. (1993): OC1: A randomized algorithm for building oblique decision trees. In: *Proceedings of AAAI Conference on Artificial Intelligence*, Vol. 11, Book One, Machine Learning, pp. 322-327.
- Nam, C. W.** – Kim, T. S. – Park, N. J. – Lee, H. K. (2008): Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies. *Journal of Forecasting*, Vol. 27. No. 6. pp. 493-506.
- Neophytou, E.** – Mar Molinero, C. (2004): Predicting Corporate Failure in the UK: A Multidimensional Scaling Approach. *Journal of Business Finance and Accounting*, Vol. 31. No. 5-6. pp. 677-710.
- Nguyen, H. H.** – Viviani, J. L. – Ben Jabeur, S. (2023): Bankruptcy prediction using machine learning and Shapley additive explanations. *Review of Quantitative Finance and Accounting*, pp. 1-42.
- Nyitrai T.** (2014a): Validációs eljárások a csődelőrejelző modellek teljesítményének megítélésében. *Statisztikai Szemle*, 92. évf. 4. sz. pp. 357-377.
- Nyitrai T.** (2014b): Növelhető-e a csőd-előrejelző modellek előre jelző képessége az új klasszifikációs módszerek nélkül? *Közgazdasági Szemle*, 61. évf. 5. sz. pp. 566-585.

- Nyitrai T.** (2015): Hazai vállalkozások csődjének előrejelzése a csődeseményt megelőző egy, két, illetve három évvel korábbi pénzügyi beszámolók adatai alapján. *Vezetéstudomány*, 46. évf. 5. sz. pp. 55-65.
- Nyitrai T.** (2017): Stock és flow típusú számviteli adatok alkalmazása a csődelőrejelző modellekben. *Vezetéstudomány*, 48. évf. 8-9. sz. pp. 68-77.
- Nyitrai T.** – Virág M. (2017a): A pénzügyi mutatók időbeli tendenciájának figyelembevétele logisztikus regresszióra épülő csődelőrejelző modellekben. *Statisztikai Szemle*, 95. évf. 1. sz. pp. 5-28.
- Nyitrai T.** – Virág M. (2017b): Magyar vállalkozások felszámolásának előrejelzése pénzügyi mutatóik idősorai alapján. *Közgazdasági Szemle*, 64. évf. 3. sz. pp. 305-324.
- Nyitrai T.** (2018): A vállalatok tevékenységi körének és méretének hatása a csődelőrejelző modellekre. *Statisztikai Szemle*, 96. évf. 10. sz. pp. 973-1000.
- Nyitrai T.** (2019): Dynamization of bankruptcy models via indicator variables. *Benchmarking: An International Journal*, Vol. 26. No. 1. pp. 317-332.
- Nyitrai T.** – Virág M. (2019): The effects of handling outliers on the performance of bankruptcy prediction models. *Socio-Economic Planning Sciences*, Vol. 67. pp. 34-42.
- Nyitrai T.** (2021): A gépi tanulás módszereinek alkalmazása R-ben. *Statisztikai Szemle*, 99. évf. 2. sz. pp. 173-198.
- Odom, M. D.** – Sharda, R. (1990): A neural network model for bankruptcy prediction. *Proceeding of the International Joint Conference on Neural Networks*, 17-21 June 1990, San Diego, CA, USA, pp. 163-168
- Ohlson, J. A.** (1980): Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, Vol. 18. No. 1. pp. 109-131.
- Olden, J. D.** – Joy, M. K. – Death, R. G. (2004): An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, Vol. 178. No. 3-4. pp. 389-397.
- Ooghe, H.** – Claus, H. – Sierens, N. – Camerlynck, J. (1999): *International comparison of failure prediction models from different countries: an empirical analysis*. Working Paper, Department of Corporate Finance, nr. 99/79, University of Ghent, Belgium.
- Oommen, T.** – Baise, L. G. – Vogel, R. M. (2011): Sampling bias and class imbalance in maximum-likelihood logistic regression. *Mathematical Geosciences*, Vol. 43. pp. 99-120.

- Oravecz B.** (2007): Credit scoring modellek és teljesítményük értékelése. *Hitelintézeti Szemle*, 6. évf. 6. sz. pp. 607-627.
- Papíková, L.** – Papík, M. (2022): Effects of classification, feature selection, and resampling methods on bankruptcy prediction of small and medium-sized enterprises. *Intelligent Systems in Accounting, Finance and Management*, Vol. 29. No. 4. pp. 254-281.
- Patyi B.** – Szóka K. – Pataki L. (2023): Applicability of bankruptcy prediction models in the western Hungarian food retail sector. *Acta Academiae Beregsasiensis. Economics*, Vol. 4. pp. 227-239.
- Pawlak, Z.** (1982): Rough sets. *International Journal of Computer & Information Sciences*, Vol. 11. pp. 341-356.
- Peduzzi, P.** – Concato, J. – Kemper, E. – Holford, T. R. – Feistein, A. R. (1996): A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *Journal of Clinical Epidemiology*, Vol. 49. No. 12. pp. 1373-1379.
- Peng, H.** – Long, F. – Ding, C. (2005): Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 27. No. 8. pp. 1226-1238.
- Platt, H. D.** – Platt, M. B. – Pedersen, J. G. (1994): Bankruptcy discrimination with real variables. *Journal of Business Finance & Accounting*, Vol. 21. No. 4. pp. 491-510.
- Poston, K. M.** – Harmon, K. – Gramlich, J. D. (1994): A test of financial ratios as predictors of turnaround versus failure among financially distressed firms. *Journal of Applied Business Research*, Vol. 10. No. 1. pp. 41-56.
- Qu, Y.** – Quan, P. – Lei, M. – Shi, Y. (2019): Review of bankruptcy prediction using machine learning and deep learning techniques. *Procedia Computer Science*, Vol. 162. pp. 895-899.
- Quinlan, J. R.** (1986): Induction of decision trees. *Machine Learning*, Vol. 1. No. 1. pp. 81-106.
- Quinlan, J. R.** (1993): *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, California, USA, 302 p.
- Quinlan, J. R.** (1996): Bagging, boosting and C4.5. *Proceedings of the Thirteenth National Conference on Artificial Intelligence*. Menlo Park, CA: AAAI Press/MIT Press: Cambridge, MA, pp. 725-730.

- Quinlan, J. R.** (1997): Is See5/C5.0 Better Than C4.5? [on-line]. RuleQuest Research. <https://rulequest.com/see5-comparison.html> (2023.10.18).
- Ratting A.** (2015): Fizetésképtelenség-előrejelzési megközelítések. *Társadalom és Gazdaság*, 7. évf. 3. sz. pp. 53-73.
- Rezende, F. F.** – Montezano, R. M. D. S. – Oliveira, F. N. D. – Lameira, V. D. J. (2017): Predicting financial distress in publicly-traded companies. *Revista Contabilidade & Finanças*, Vol. 28. pp. 390-406.
- Ribeiro, M. T.** – Singh, S. – Guestrin, C. (2016): " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135-1144.
- Rico, M.** – Cantarero, S. – Puig, F. (2021): Regional disparities and spatial dependence of bankruptcy in Spain. *Mathematics*, Vol 9. No. 9. pp. 1-20.
- Rosenblatt, F.** (1958): The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, Vol. 65. No. 6. pp. 386-408.
- Rumelhart, D. E.** – Hinton, G. E. – Williams, R. J. (1986): Learning representations by back-propagating errors. *Nature*, Vol. 323. pp. 533-536.
- Russell, S.** – Norvig, P. (2005): *Mesterséges intelligencia modern megközelítésben*. Panem Könyvkiadó, Budapest.
- Ryu, Y. U.** – Yue, W. T. (2005): Firm bankruptcy prediction: experimental comparison of isotonic separation and other classification approaches. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 35. No. 5. pp. 727-737.
- Sajtos L.** – Mitev A. (2007): *SPSS kutatási és adatelemzési kézikönyv*. Alinea Kiadó, Budapest, 404. p.
- Samuels, J. M.** – Brayshaw, R. E. – Craner, J. M. (1995): *Financial statement analysis in Europe*. Chapman and Hall, London.
- Sánchez-Lasheras, F.** – De Andrés, J. – Lorca, P. – De Cos Juez, F. J. (2012): A hybrid device for the solution of sampling bias problems in the forecasting of firms' bankruptcy. *Expert Systems with Applications*, Vol. 39. No. 8. p. 7512-7523
- Seiffert, C.** – Khoshgoftaar, T. M. – Van Hulse, J. – Napolitano, A. (2008): Resampling or reweighting: A comparison of boosting implementations. *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence*. Dayton, OH, USA, pp. 445-451.

- Serra, L.** – Detotto, C. – Verdoy, P. – Vannini, M. (2022): Intersectoral and spatial spillovers of firms' bankruptcy in Spain. *Letters in Spatial and Resource Sciences*, Nol. 15. No. 2. pp. 197-211.
- Serrano-Cinca, C.** (1996): Self organizing neural networks for financial diagnosis. *Decision Support Systems*, Vol. 17. No. 3. pp. 227-238.
- Schapire, R. E.** (1990): The strength of weak learnability. *Machine Learning*, Vol. 5. No. 2. pp. 197-227.
- Schmidhuber, J.** (2015): Deep learning in neural networks: An overview. *Neural Networks*, Vol. 61. pp. 85-117.
- Shanker, M.** – Hu, M. Y. – Hung, M. S. (1996): Effect of data standardization on neural network training. *Omega*, Vol. 24. No. 4. pp. 385-397.
- Sharma S.** – Mahajan, V. (1980): Early Warning Indicators of Business Failure. *Journal of Marketing*, Vol. 44. No. 4. pp. 80-89.
- Shi, Y.** – Li, X. (2019): An overview of bankruptcy prediction models for corporate firms: A Systematic literature review. *Intangible Capital*, Vol. 15. No. 2. pp. 114-127.
- Shin, K.** – Lee, T. S. – Kim, H. (2005): An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, Vol. 28, pp. 127-135.
- Shirata, C. Y.** – Takeuchi, H. – Ogino, S. – Watanabe, H. (2011): Extracting key phrases as predictors of corporate bankruptcy: Empirical analysis of annual reports by text mining. *Journal of Emerging Technologies in Accounting*, Vol. 8. No. 1. pp. 31-44.
- Shumway, T.** (2001): Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, Vol. 74. No. 1. pp. 101-124.
- Slowinski, R.** – Zopounidis, C. (1995): Application of the rough set approach to evaluation of bankruptcy risk. *Intelligent Systems in Accounting, Finance and Management*, Vol. 4. No. 1. pp. 27-41.
- Soui, M.** – Smiti, S. – Mkaouer, M. W. – Ejbali, R. (2020): Bankruptcy prediction using stacked auto-encoders. *Applied Artificial Intelligence*, Vol. 34. No. 1. pp. 80-100.
- Sousa, A.** – Braga, A. – Cunha, J. (2022): Impact of macroeconomic indicators on bankruptcy prediction models: Case of the Portuguese construction sector. *Quantitative Finance and Economics*, Vol. 6. No. 3. pp. 405-432.
- Stickney, C. P.** (1990): *Financial Statement Analysis: A strategic perspective*. Harcourt Brace Jovanovich Publishers, Orlando, Florida, USA, 615 p.

- Sun, L.** – Shenoy, P. P. (2007): Using Bayesian networks for bankruptcy prediction: Some methodological issues. *European Journal of Operational Research*, Vol. 180. No. 2. pp. 738-753.
- Sun, J.** – Jia, M. Y. – Li, H. (2011): AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies. *Expert Systems With Applications*, Vol. 38. No. 8. pp. 9305-9312.
- Sun, J.** – Li, H. (2012): Financial distress prediction using support vector machines: Ensemble vs. individual. *Applied Soft Computing*, Vol. 12. No. 8. pp. 2254-2265.
- Sun, J.** – Li, H. – Huang, Q. H. – He, K. Y. (2014): Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, Vol. 57. pp. 41-56.
- Szabadosné Németh Zs.** – Dávid L. (2005): A kis- és középvállalati szegmens mulasztási valószínűségének előrejelzése magyarországi környezetben. *Hitelintézeti Szemle*, 4. évf. 3. sz. pp. 39-58.
- Szántó T. K.** (2023): Kiugró értékek kezelése logisztikus regresszió alapuló csődelőrejelzési modellek esetén. *Pénzügyi Szemle*, 69. évf. 3. sz. pp. 93-108.
- Székelyi M.** – Barna I. (2008): *Túlélő készlet az SPSS-hez*. Typotex Kiadó, Budapest, Negyedik kiadás, 453 p.
- Thabtah, F.** – Hammoud, S. – Kamalov, F. – Gonsalves, A. (2020): Data imbalance in classification: Experimental evaluation. *Information Sciences*, Vol. 513. pp. 429-441.
- Takács O.** (2021): Nemek közötti bérkülönbségek Magyarországon: a véletlenerdő-és az OLS-becslésen alapuló Blinder–Oaxaca-dekompozíció eredményeinek összehasonlítása. *Statisztikai Szemle*, 99. évf. 1. sz. pp. 5-45.
- Tattar, P. N.** (2018): *Hands-On Ensemble Learning with R*. Packt Publishing Ltd., Birmingham, UK.
- Teng, S. S. H.** – Bhatia, S. G. – Anwar, S. (2011): A success versus failure prediction model for small businesses in Singapore. *American Journal of Business*, Vol. 26. No.1. pp. 50-64.
- Theodossiou, P.** (1991): Alternative models for assessing the financial condition of business in Greece. *Journal of Business Finance and Accounting*, Vol. 18. No. 5. pp. 697-720.
- The Business Failure Record** (1978): Dun and Bradstreet Inc., New York, USA.
- Thomas, L. C.** (2009): *Consumer Credit Models: Pricing, Profit and Portfolios*. Oxford, Oxford University Press, 400 p.

- Tian, S.** – Yu, Y. – Guo, H. (2015): Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, Vol. 52. pp. 89-100.
- Tibshirani, R.** (1996): Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Vol. 58. No. 1. pp. 267-288.
- Tougui, I.** – Jilbab, A. – El Mhamdi, J. (2021): Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthcare Informatics Research*, Vol. 27. No. 3. pp. 189-199.
- Trabelsi, S.** – He, R. – He, L. – Kusy, M. (2015): A comparison of Bayesian, Hazard, and Mixed Logit model of bankruptcy prediction. *Computational Management Science*, Vol. 12. pp. 81-97.
- Trustorff, J.-H.** – Konrad, P. M. – Leker J. (2011): Credit risk prediction using support vector machines. *Review of Quantitative Finance and Accounting*, Vol. 36. No. 4. pp. 565-581.
- Tsai, C. F.** (2009): Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, Vol. 22. No. 2. pp. 120-127.
- Tsukuda, J.** – Baba, S. I. (1994): Predicting Japanese corporate bankruptcy in terms of financial data using neural network. *Computers & Industrial Engineering*, Vol. 27. No. 1-4. pp. 445-448.
- Uliha G.** (2015): Rövid távú olajár-előrejelzések teljesítményének stabilitása. *Statisztikai Szemle*, 93. évf. 3. sz. pp. 189-224.
- Van Horne, J. C.** (2002): *Financial management and policy* - 12th ed., Prentice-Hall Inc., New Jersey, USA, 832 p.
- Veganzones, D.** – Séverin, E. (2018): An investigation of bankruptcy prediction in imbalanced datasets. *Decision Support Systems*, Vol. 112. pp. 111-124.
- Virág M.** – Hajdu O. (1996): Pénzügyi mutatószámokon alapuló csődmodellszámítások. *Bankszemle*, 15. évf. 5. sz. pp. 42-53.
- Virág M.** (2001): *Pénzügyi elemzés csődelőrejelzés*, Kossuth Kiadó, Budapest, 112 p.
- Virág M.** (2004): A csődmodellek jellegzetességei és története. *Vezetéstudomány*, 35. évf. 10. sz. pp. 24-32
- Virág M.** – Kristóf T. (2005): Az első hazai csődmodell újraszámítása neurális hálók segítségével. *Közgazdasági Szemle*. 52. évf. pp. 144-162.
- Virág M.** – Kristóf T. (2006): Iparági rátákon alapuló csődelőrejelzés sokváltozós statisztikai módszerekkel. *Vezetéstudomány*, 37. évf. 1. sz. pp. 25-35.

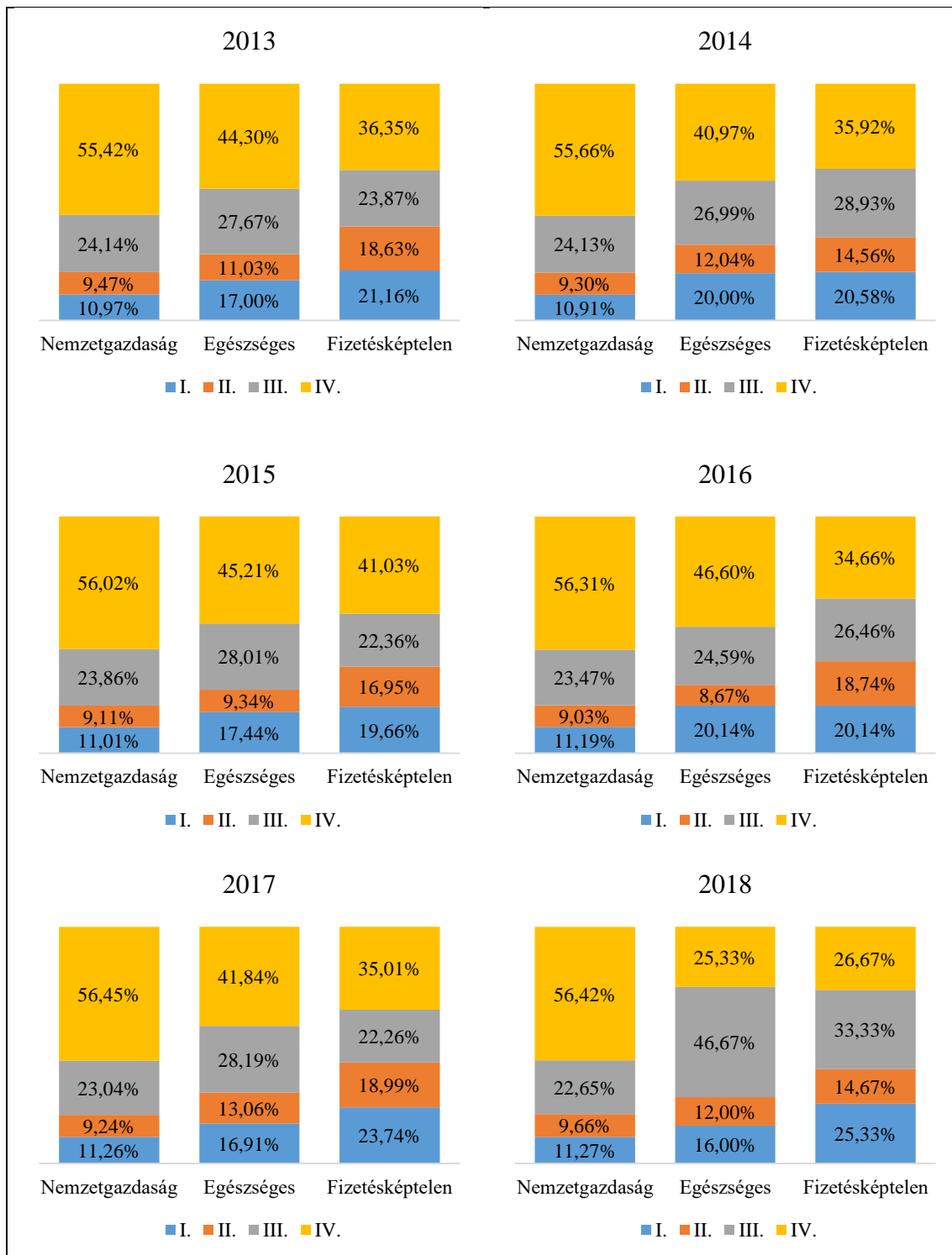
- Virág M.** (2007): A csődmodellek. In: Fazakas G. (szerk.): *Vállalati pénzügyi döntések*. Budapesti Corvinus Egyetem Pénzügyi Intézet, Tanszék Kiadó, Budapest, pp. 195-209.
- Virág M.** – Kristóf T. (2009): Többdimenziós skálázás a csődmodellezésben. *Vezetéstudomány*, 40. évf. 1. sz. 50-58.
- Virág M.** – Fiáth A. – Kristóf T. – Varsányi J. (2013a): *Pénzügyi elemzés, csődelőrejelzés, válságkezelés*. Kossuth Kiadó, Budapest, 301 p.
- Virág M.** – Nyitrai T. (2013b): Application of support vector machines on the basis of the first Hungarian bankruptcy model. *Society and Economy*. Vol. 35. No. 2. pp. 227-248.
- Virág M.** – Nyitrai T. (2014a): Metamódszerek alkalmazása a csődelőrejelzésben. *Hitelintézeti Szemle*, 13. évf. 4. sz. pp. 180-195.
- Virág M.** – Nyitrai T. (2014b): Is there a trade-off between the predictive power and the interpretability of bankruptcy models? The case of the first Hungarian bankruptcy prediction model. *Acta Oeconomica*, Vol. 64. No. 4. pp. 419-440.
- Wah, Y. B.** – Ibrahim, N. – Hamid, H. A. – Abdul-Rahman, S. – Fong, S. (2018): Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science & Technology*, Vol. 26. No. 1. pp. 329-340.
- Wahlen, J. M.** – Baginski, S. P. – Bradshaw, M. T. (2018): *Financial reporting, financial statement analysis, and valuation: A strategic perspective*. 9th Edition. Cengage Learning, Boston, USA.
- Wang, G.** – Hao, J. – Ma, J. – Jiang, H. (2011): A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, Vol. 38. No. 1. pp. 223-230.
- Wang, Z.** – Xie, Z. (2014): Infrared face recognition based on local binary patterns and Kruskal-Wallis test. In *2014 IEEE/ACIS 13th International Conference on Computer and Information Science*, pp. 185-188.
- Wang, N.** (2017): Bankruptcy prediction using machine learning. *Journal of Mathematical Finance*, Vol. 7. No. 4. pp. 908-918.
- West, D.** – Dellana, S. – Qian, J. (2005): Neural network ensemble strategies for financial decision applications. *Computers & Operations Research*, Vol. 32. No. 10. pp. 2543-2559.
- Whitaker, R. B.** (1999): The early stages of financial distress. *Journal of Economics and Finance*, Vol. 23. No. 2. pp. 123-132.

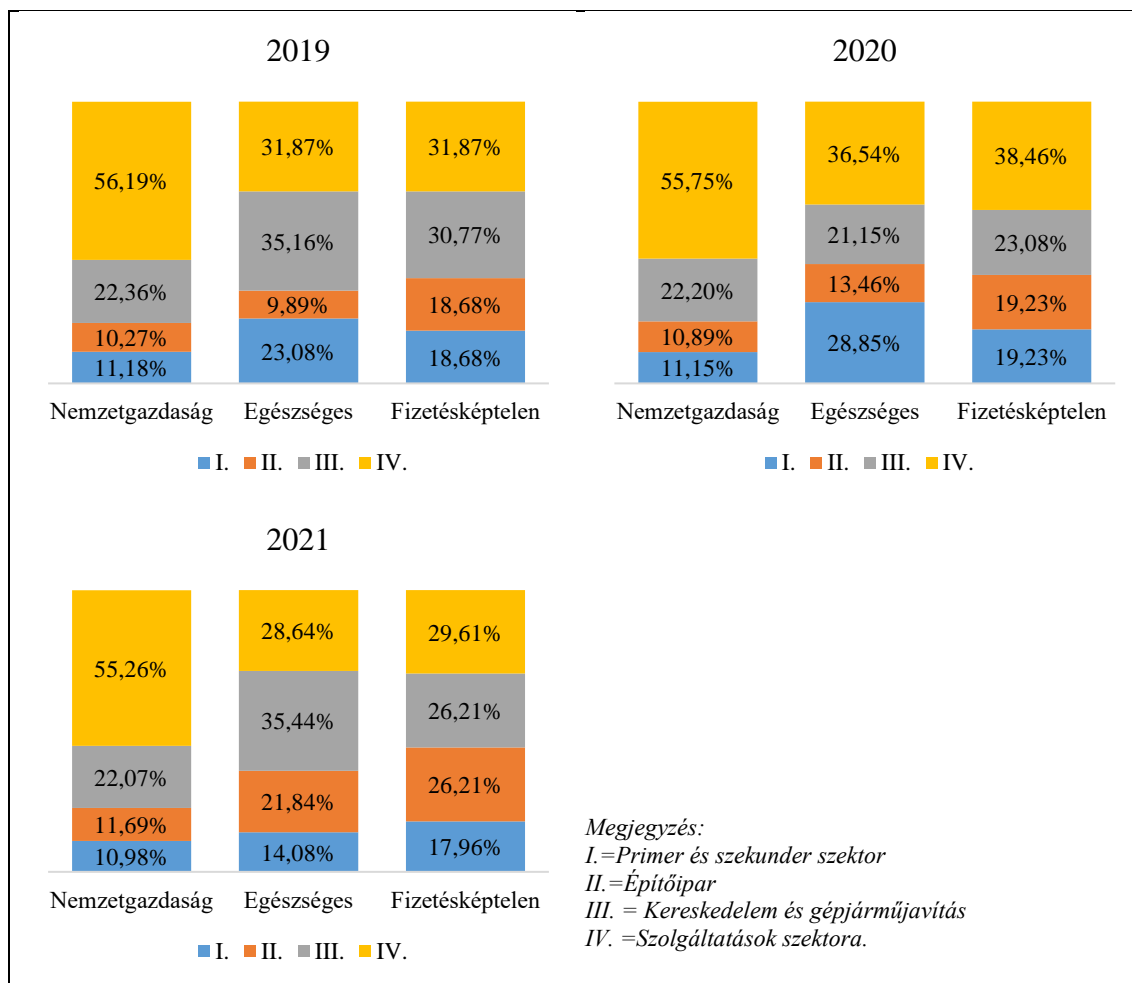
- Wilson, R. L.** – Sharda, R. (1994): Bankruptcy prediction using neural networks. *Decision Support Systems*, Vol. 11. No. 5. pp. 545-557.
- Winakor, A.** – Smith, R. (1935): Changes in the financial structure of unsuccessful industrial corporations. *Bulletin*, Vol. 51. pp. 1-41.
- Wooldridge, M.** (2021): *A brief history of artificial intelligence: what it is, where we are, and where we are going*. Flatiron Books, New York, USA.
- Wruck, K. H.** (1990): Financial distress, reorganization, and organizational efficiency. *Journal of Financial Economics*, Vol. 27. No. 2. pp. 419-444.
- Wu, X.** – Kumar, V. – Quinlan, J. R. – Ghosh, J. – Yang, Q. – Motoda, H. – McLachlan, G. J. – Ng, A. – Liu, B. – Yu, P. S. – Zhou, Z.-H. – Steinbach, M. – Hand, D. J. – Steinberg, D. (2008): Top 10 algorithms in data mining. *Knowledge and Information Systems*, Vol. 14. pp. 1-37.
- Wu, Y.** – Gaunt, C. – Gray, S. (2010): A comparison of alternative bankruptcy prediction models. *Journal of Contemporary Accounting & Economics*, Vol. 6. No. 1. pp. 34-45.
- Wu, J.** – Chen, X. Y. – Zhang, H. – Xiong, L. D. – Lei, H. – Deng, S. H. (2019): Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, Vol. 17. No. 1. pp. 26-40.
- Wu, Y.** (2021): Can't ridge regression perform variable selection? *Technometrics*, Vol. 63. No. 2. pp. 263-271.
- Xia, Y.** – Liu, C. – Li, Y. – Liu, N. (2017): A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, Vol. 78. pp. 225-241.
- Xiao, Z.** – Yang, X. – Pang, Y. – Dang, X. (2012): The prediction for listed companies' financial distress by using multiple prediction methods with rough set and Dempster-Shafer evidence theory. *Knowledge-Based Systems*, Vol. 26. pp. 196-206.
- Xu, X.** – Wang, Y. (2009): Financial failure prediction using efficiency as a predictor. *Expert Systems with Applications*, Vol. 36. No. 1. pp. 366-373.
- Yadav, S.** – Shukla, S. (2016): Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In: Raju, B. – Garg, D. – Raju, V. – Raju, R. (eds): *IEEE 6th International Advance Computing Conference*, Conference Publishing Services, pp. 78-83.

- Yang, X.** – Wang, Y. – Byrne, R. – Schneider, G. – Yang, S. (2019): Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical Reviews*, Vol. 119. No. 18. pp. 10520-10594.
- Yeh, C.-C.** – Chi D.-J. – Lin, Y.-R. (2014): Going-concern prediction using hybrid random forests and rough set approach. *Information Sciences*, Vol. 254. pp. 98-110.
- Yin, J.** – Tian, L. (2014): Joint confidence region estimation for area under ROC curve and Youden index. *Statistics in Medicine*, Vol. 33. No. 6. pp. 985-1000.
- Zhao, Z.** – Xu, S. – Kang, B. H. – Kabir, M. M. J. – Liu, Y. – Wasinger, R. (2015): Investigation and improvement of multi-layer perception neural networks for credit scoring. *Expert Systems with Applications*, Vol. 42. No. 7. pp. 3508-3516.
- Zhong, M.** – Miao, C. – Shen, Z. – Feng, Y. (2014): Comparing the learning effectiveness of BP, ELM, I-ELM and SVM for corporate credit ratings. *Neurocomputing*, Vol. 128. pp. 285-295.
- Zięba, M.** – Tomczak, S. K. – Tomczak, J. M. (2016): Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, Vol. 58. pp. 93-101.
- Zimmerman, D.** (1994): A note on the influence of outliers on parametric and nonparametric tests. *The Journal of General Psychology*, Vol. 121. No. 4. pp. 391-401.
- Zmijewski, M. E.** (1984): Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, Supplement to Vol. 22, pp. 59-82.
- Zou, H.** – Hastie, T. (2005): Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Vol. 67. No. 2. pp. 301-320.

FÜGGELÉK

F1. függelék: Az azonos osztálylétszámú minta megoszlása osztály és gazdasági szektor szerint





Forrás: saját szerkesztés

F2. függelék: Az empirikus kutatáshoz kijelölt magyarázó változók

Jelölés	Mutató számítási módja	Típusa	Mértékegység	Megnevezés
X1	Adózás előtti eredmény / Összes eszköz	jövedelmezőségi	mértékegység nélküli	Eszközarányos nyereség (ROA)
X2	Adózás előtti eredmény / Saját tőke	jövedelmezőségi	mértékegység nélküli	Sajáttőke arányos megtérülés (ROE)
X3	Adózás előtti eredmény / Értékesítés nettó árbevétele	jövedelmezőségi	mértékegység nélküli	Árbevétel arányos nyereség (ROS)
X4	Adózott eredmény / Értékesítés nettó árbevétele	jövedelmezőségi	mértékegység nélküli	Árbevétel arányos nyereség (ROS)
X5	Értékesítés nettó árbevétele / Nettó forgótőke	jövedelmezőségi	mértékegység nélküli	Forgótőke arányos árbevétel
X6	Adózott eredmény / (Összes eszköz-Összes kötelezettség)	jövedelmezőségi	mértékegység nélküli	Adózott eredmény és nettó vagyon aránya
X7	Működési eredmény / Összes eszköz	jövedelmezőségi	mértékegység nélküli	Üzleti (üzemi) eredmény és összes eszköz aránya
X8	Adózott eredmény / Összes eszköz	jövedelmezőségi	mértékegység nélküli	Adózott eredmény és összes eszköz aránya
X9	Működési eredmény / Személyi jellegű ráfordítások	jövedelmezőségi	mértékegység nélküli	Béarányos eredmény
X10	Forgóeszközök / Rövid lejáratú kötelezettségek	likviditási	mértékegység nélküli	Likviditási általános ráta
X11	(Forgóeszközök-Készletek) / Rövid lejáratú kötelezettségek	likviditási	mértékegység nélküli	Likviditási gyorsráta
X12	Pénzeszközök / Rövid lejáratú kötelezettségek	likviditási	mértékegység nélküli	Pénzhányad mutató
X13	(Pénzeszközök+piacképes értékpapírok+nettó követelések) / [(Éves működési költségek-nem készpénzjellegű ráfordítások) / 365]	likviditási	mértékegység nélküli	Defensive interval ratio (No-credit interval ratio). A napok számát adja meg ameddig a cég a forgóeszközökön kívül más tőkeforrást igénybe venne.
X14	Működő tőke / Összes eszköz	likviditási	mértékegység nélküli	Működő tőke és összes eszköz aránya
X15	Működő tőke / Értékesítés nettó árbevétele	likviditási	mértékegység nélküli	Működő tőke és árbevétel aránya
X16	Működő tőke / (Összes eszköz-Összes kötelezettség)	likviditási	mértékegység nélküli	Működő tőke és nettó vagyon aránya
X17	Készletek / Működő tőke	likviditási	mértékegység nélküli	Forgótőke lekötöttsége
X18	Összes kötelezettség / Saját tőke	eladósodottsági	mértékegység nélküli	Bonitás mutató
X19	Eredménytartalék / Összes eszköz	eladósodottsági	mértékegység nélküli	Eredménytartalék aránya

X20	Saját tőke / Mérlegfőösszeg	eladósodottsági	mértékegység nélküli	Saját vagyon aránya
X21	Összes kötelezettség / Összes eszköz	eladósodottsági	mértékegység nélküli	Összes kötelezettség és összes eszköz aránya
X22	Rövid lejáratú kötelezettségek / Összes eszköz	eladósodottsági	mértékegység nélküli	Rövid kötelezettségek és összes eszköz aránya
X23	Hosszú lejáratú kötelezettségek / Összes eszköz	eladósodottsági	mértékegység nélküli	Hosszú kötelezettségek és összes eszköz aránya
X24	(Összes eszköz-Összes kötelezettség) / Összes eszköz	eladósodottsági	mértékegység nélküli	Nettó vagyon és összes eszköz aránya
X25	Összes kötelezettségek / (Összes eszköz-Összes kötelezettség)	eladósodottsági	mértékegység nélküli	Összes kötelezettség és nettó vagyon aránya
X26	(Összes eszköz-Összes kötelezettség) / Összes kötelezettség	eladósodottsági	mértékegység nélküli	Nettó vagyon és összes kötelezettség aránya
X27	Saját tőke / Befektetett eszközök	eladósodottsági	mértékegység nélküli	Befektetett eszközök saját finanszírozása (másként fedezeti mutató)
X28	Saját tőke / Összes kötelezettség	eladósodottsági	mértékegység nélküli	Tartozások lefedettsége
X29	Pénzügyi ráfordítások / Adózás előtti eredmény	eladósodottsági	mértékegység nélküli	Pénzügyi ráfordítások és adózás előtti eredmény aránya
X30	Értékesítés nettó árbevétele/Összes eszköz	hatékonysági	mértékegység nélküli	Eszközök forgási sebessége (fordulatok száma)
X31	Értékesítés nettó árbevétele / Készletek	hatékonysági	mértékegység nélküli	Készletek forgási sebessége (fordulatok száma)
X32	Értékesítés nettó árbevétele / Követelések	hatékonysági	mértékegység nélküli	Vevők forgási sebessége (fordulatok száma)
X33	Forgóeszközök / Értékesítés nettó árbevétele	hatékonysági	mértékegység nélküli	Forgóeszközök és árbevétel aránya
X34	Készletek / Értékesítés nettó árbevétele	hatékonysági	mértékegység nélküli	Készletek és árbevétel aránya
X35	(Forgóeszközök-Készletek) / Értékesítés nettó árbevétele	hatékonysági	mértékegység nélküli	Likvid eszközök és árbevétel aránya
X36	Forgóeszközök / Összes eszköz	eszközszerkezeti/tőkeszerkezeti	mértékegység nélküli	Forgóeszközök aránya
X37	Pénzeszközök / Forgóeszközök	eszközszerkezeti/tőkeszerkezeti	mértékegység nélküli	Pénzeszközök aránya
X38	Pénzeszközök / Összes eszköz	eszközszerkezeti/tőkeszerkezeti	mértékegység nélküli	Pénzeszközök és összes eszköz aránya
X39	(Forgóeszközök-Készletek)/ Összes eszköz	eszközszerkezeti/tőkeszerkezeti	mértékegység nélküli	Likvid eszközök aránya az összes eszközökön belül

X40	Rövid lejáratú kötelezettségek / Összes kötelezettség	eszközszerkezeti/tókeszerkezeti	mértékegység nélküli	Rövid lejáratú kötelezettségek aránya
X41	Hosszú lejáratú kötelezettségek / Összes kötelezettség	eszközszerkezeti/tókeszerkezeti	mértékegység nélküli	Hosszú lejáratú kötelezettségek aránya
X42	Működési CF / Összes eszköz	cash flow	mértékegység nélküli	Működési cash flow és mérlegfőösszeg aránya
X43	Működési CF / Összes kötelezettség	cash flow	mértékegység nélküli	Működési cash flow és összes tartozások aránya
X44	Működési CF / Értékesítés nettó árbevétele	cash flow	mértékegység nélküli	Működési cash flow és nettó árbevétel aránya
X45	Működési CF / Rövid lejáratú kötelezettségek	cash flow	mértékegység nélküli	Működési cash flow és rövid távú tartozások aránya
X46	Működési CF / Hosszú lejáratú tartozások	cash flow	mértékegység nélküli	Működési cash flow és hosszú távú tartozások aránya
X47	ln(Mérlegfőösszeg)	méret	mértékegység nélküli	Eszközállomány természetes alapú logaritmus
X48	ln(Értékesítés árbevétele)	méret	mértékegység nélküli	Az árbevétel természetes alapú logaritmus
X49	Alkalmazottak átlagos létszáma	egyéb	fő	Az üzleti évben az alkalmazottak átlagos létszáma
X50	Az évek száma a beszámolóig amióta működik a vállalat	egyéb	évek	Az évek száma egészen kerekítve lefele, a beszámoló időpontjáig
Y1	Értékesítés nettó árbevétele a településen / Regisztrált vállalatok száma a településen	mezo	millió Ft/vállalat	Egy vállalatra jutó árbevétel a településen
Y2	Nyilvántartott álláskereső száma / Lakónépesség 15–64 évesek	mezo	mértékegység nélküli	Munkanélküliségi ráta
Y3	(Bankfiók és takarékszövetkezeti fiók összesen / Lakónépesség) * 1000	mezo	db/1000 fő	Ezer főre jutó bankfiók és takarékszövetkezeti fiók a településen
Y4	Helyi önkormányzatok bevételei helyi adóból / Lakónépesség	mezo	ezer Ft/fő	Egy főre jutó helyi adóbevételek
Y5	Budapest elérési ideje közúton, a leggyorsabb úton	mezo	perc	A főváros közötti távolsága percben kifejezve
Y6	Legközelebbi gyorsforgalmi csomópont távolsága a legrövidebb közúton	mezo	km	A legközelebbi gyorsforgalmi út távolsága közúton
Y7	Település jogállása dummy (1=város, 0=község)	mezo	mértékegység nélküli	Városi és községi település azonosítása

Forrás: saját szerkesztés

F3. függelék: A reprezentatív mintán épített modellek paramétere

Módszerek	Paraméterek	
	Tanuló minta: 2013	Tanuló minta: 2014
Logit	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = -1,98 - 0,06X3 - 0,003X9 - 0,03X15 + 0,28X21 + 0X28 - 2,21X37 + 1,49X39 - 0,19X48 - 0,04X50$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az aláhúzással jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 15,1\%$</p>	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = -0,37 - 1,25X8 - 0,05X20 + 0X32 - 2,27X37 + 1,36X39 - 0,25X48 - 0,07X50$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az aláhúzással jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 25,8\%$</p>
SVM Lineáris	Kernel: lineáris Büntető paraméter (C): 0,053	Kernel: lineáris Büntető paraméter (C): 0,0085
SVM RBF	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 5,54 Kernel paraméter (γ): 792,74	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 0,0061 Kernel paraméter (γ): 994,36
Neurális háló	Rejtett rétegek száma: 2 Csomópontok száma a rejtett rétegben: 4-22 Futásszám: 1000 Függvényforma: köztes rétegben ReLu, kimeneti rétegben sigmoid Lambda: 3,56	Rejtett rétegek száma: 3 Csomópontok száma a rejtett rétegben: 7-68-65 Futásszám: 1000 Függvényforma: köztes rétegben Tanh, kimeneti rétegben sigmoid Lambda: $4,14e^{-6}$
Bagging	Elágazási csomópontok: 84 Döntési fák száma: 500	Elágazási csomópontok: 92 Döntési fák száma: 119
Random Forest	Elágazási csomópontok: 234 Döntési fák száma: 236 Véletlenszerűen kiválasztott független változók száma: 3	Elágazási csomópontok: 798 Döntési fák száma: 146 Véletlenszerűen kiválasztott független változók száma: 3
Adaboost	Elágazási csomópontok: 13 Döntési fák száma: 10 Tanulási ráta: 0,07	Elágazási csomópontok: 36 Döntési fák száma: 146 Tanulási ráta: 0,02
	Tanuló minta: 2015	Tanuló minta: 2016
Logit	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = -4,01 - 0,55X7 - 0,31X14 + 0,18X23 + 0X32 - 1,77X37 + 1,59X39 - 0,04X50$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az aláhúzással jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 20,6\%$</p>	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = -1,17 - 0,72X7 - 0,02X9 - 1,35X37 - 0,2X48 - 0,05X50$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az aláhúzással jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 16,8\%$</p>
SVM Lineáris	Kernel: lineáris Büntető paraméter (C): 0,012	Kernel: lineáris Büntető paraméter (C): 0,0012
SVM RBF	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 0,0106	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 994,27

	Kernel paraméter (γ): 0,053	Kernel paraméter (γ): 640,90
Neurális háló	Rejtett rétegek száma: 3 Csomópontok száma a rejtett rétegben: 1-1-1 Futásszám: 1000 Függvényforma: köztes rétegben ReLu, kimeneti rétegben sigmoid Lambda: $2,28e^{-9}$	Rejtett rétegek száma: 2 Csomópontok száma a rejtett rétegben: 66-22 Futásszám: 1000 Függvényforma: köztes rétegben ReLu, kimeneti rétegben softmax Lambda: 0,00017
Bagging	Elágazási csomópontok: 15 Döntési fák száma: 66	Elágazási csomópontok: 8 Döntési fák száma: 123
Random Forest	Elágazási csomópontok: 132 Döntési fák száma: 107 Véletlenszerűen kiválasztott független változók száma: 1	Elágazási csomópontok: 12 Döntési fák száma: 11 Véletlenszerűen kiválasztott független változók száma: 4
Adaboost	Elágazási csomópontok: 20 Döntési fák száma: 31 Tanulási ráta: 0,05	Elágazási csomópontok: 10 Döntési fák száma: 39 Tanulási ráta: 0,02

Forrás: saját szerkesztés

F4. függelék: Az azonos osztálylétszámú mintán épített modellek paramétere

Módszerek	Paraméterek	
<i>A-panel: Modellek 2013-tól 2016-ig</i>		
	Tanuló minta: 2013	Tanuló minta: 2014
Logit	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = 1,26 - 0,41X3 - 0,20X9 - 0,081X15 + 0,43X21 + 0X28 - 1,82X37 + 1,21X39 - 0,17X48 - 0,04X50$</p> <p>Megjegyzés: A <u>félkövérrel</u> jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2=37,7\%$</p>	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = 3,05 - 1,36X8 + 0,001X10 - 0,005X15 + 0,05X19 + 0,38X23 - 0,56X26 - 0,89X38 - 1,44X41 - 0,20X48 - 0,03X50$</p> <p>Megjegyzés: A <u>félkövérrel</u> jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2=47,7\%$</p>
SVM Lineáris	Kernel: lineáris Büntető paraméter (C): 1,33	Kernel: lineáris Büntető paraméter (C): 0,41
SVM RBF	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 41,54 Kernel paraméter (γ): 15,75	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 757,42 Kernel paraméter (γ): 129,86
Neurális háló	Rejtett rétegek száma: 1 Csomópontok száma a rejtett rétegben: 230 Futásszám: 1000 Függvényforma: köztes rétegben sigmoid, kimeneti rétegben softmax Lambda: 0,032	Rejtett rétegek száma: 1 Csomópontok száma a rejtett rétegben: 1 Futásszám: 1000 Függvényforma: köztes rétegben Tanh, kimeneti rétegben softmax Lambda: 0,00012
Bagging	Elágazási csomópontok: 31 Döntési fák száma: 497	Elágazási csomópontok: 8 Döntési fák száma: 75
Random Forest	Elágazási csomópontok: 826 Döntési fák száma: 168 Véletlenszerűen kiválasztott független változók száma: 1	Elágazási csomópontok: 33 Döntési fák száma: 10 Véletlenszerűen kiválasztott független változók száma: 2
Adaboost	Elágazási csomópontok: 6 Döntési fák száma: 340 Tanulási ráta: 0,01	Elágazási csomópontok: 48 Döntési fák száma: 212 Tanulási ráta: 0,12
	Tanuló minta: 2015	Tanuló minta: 2016
Logit	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = -1,32 - 0,27X3 - 1,18X7 - 1,23X14 + 0,31X23 + 0,003X28 - 0,25X30 + 0X32 + 0,56X34 - 1,05X37 + 2,56X39 + 0,07X47 - 0,06X50$</p> <p>Megjegyzés: A <u>félkövérrel</u> jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2=51,7\%$</p>	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = 0,02 - 0,284X9 - 0,02X17 + 0,18X19 - 1,62X24 + 0X27 - 0,08X28 - 0,11X30 - 0,99X37 + 2,26X39 - 0,05X50$</p> <p>Megjegyzés: A <u>félkövérrel</u> jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2=32,4\%$</p>

SVM Lineáris	Kernel: lineáris Büntető paraméter (C): 0,055	Kernel: lineáris Büntető paraméter (C): 3,68
SVM RBF	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 264,17 Kernel paraméter (γ): 103,32	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 512,23 Kernel paraméter (γ): 82,23
Neurális háló	Rejtett rétegek száma: 2 Csomópontok száma a rejtett rétegben: 14-2 Futásszám: 1000 Függvényforma: köztes rétegben Tanh, kimeneti rétegben softmax Lambda: $5,65e^{-8}$	Rejtett rétegek száma: 1 Csomópontok száma a rejtett rétegben: 84 Futásszám: 1000 Függvényforma: köztes rétegben sigmoid, kimeneti rétegben softmax Lambda: $8,65e^{-8}$
Bagging	Elágazási csomópontok: 737 Döntési fák száma: 482	Elágazási csomópontok: 53 Döntési fák száma: 240
Random Forest	Elágazási csomópontok: 228 Döntési fák száma: 382 Véletlenszerűen kiválasztott független változók száma: 6	Elágazási csomópontok: 183 Döntési fák száma: 496 Véletlenszerűen kiválasztott független változók száma: 1
Adaboost	Elágazási csomópontok: 10 Döntési fák száma: 236 Tanulási ráta: 0,02	Elágazási csomópontok: 7 Döntési fák száma: 34 Tanulási ráta: 0,41
<i>B-panel: Modellek 2017-től 2019-ig</i>		
	Tanuló minta: 2017	Tanuló minta: 2018-2019
Logit	$P(y = 1) = \frac{1}{1 + e^{-z}}$ <p>ahol $Z = -0,58 - 1,88X_1 - 0,92X_{14} + 0X_{32} - 1,53X_{37} + 2,47X_{39} - 0,07X_{50}$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 49,3\%$</p>	$P(y = 1) = \frac{1}{1 + e^{-z}}$ <p>ahol $Z = 0,50 - 9,39X_1 + 0,008X_9 - 2,69X_{23} - 5,14X_{37} + 2,14X_{39} + 1,86X_{41} - 0,06X_{50}$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 66,2\%$</p>
SVM Lineáris	Kernel: lineáris Büntető paraméter (C): 0,519	Kernel: lineáris Büntető paraméter (C): 2,42
SVM RBF	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 925,78 Kernel paraméter (γ): 50,83	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 987,01 Kernel paraméter (γ): 22,92
Neurális háló	Rejtett rétegek száma: 3 Csomópontok száma a rejtett rétegben: 11-1-4 Futásszám: 1000 Függvényforma: köztes rétegben sigmoid, kimeneti rétegben softmax Lambda: $1,28e^{-5}$	Rejtett rétegek száma: 2 Csomópontok száma a rejtett rétegben: 127-111 Futásszám: 1000 Függvényforma: köztes rétegben Tanh, kimeneti rétegben softmax Lambda: $3,60e^{-8}$
Bagging	Elágazási csomópontok: 8 Döntési fák száma: 10	Elágazási csomópontok: 18 Döntési fák száma: 10
Random Forest	Elágazási csomópontok: 54 Döntési fák száma: 170 Véletlenszerűen kiválasztott független változók száma: 1	Elágazási csomópontok: 12 Döntési fák száma: 11 Véletlenszerűen kiválasztott független változók száma: 7
Adaboost	Elágazási csomópontok: 2 Döntési fák száma: 56 Tanulási ráta: 0,51	Elágazási csomópontok: 21 Döntési fák száma: 33 Tanulási ráta: 0,001

Forrás: saját szerkesztés

**F5. függelék: A reprezentatív és azonos osztálylétszámú mintával
készült modellek bemeneti változói**

Modell	Jelölés	Reprezentatív	Jelölés	Azonos
2013	X3	Adózás előtti eredmény/Értékesítés nettó árbevétele	X3	Adózás előtti eredmény/Értékesítés nettó árbevétele
	X9	Működési eredmény/Személyi jellegű ráfordítások	X9	Működési eredmény/Személyi jellegű ráfordítások
	X15	Működő tőke/Értékesítés nettó árbevétele	X15	Működő tőke/Értékesítés nettó árbevétele
	X21	Összes kötelezettség/Összes eszköz	X21	Összes kötelezettség/Összes eszköz
	X28	Saját tőke/Összes kötelezettség	X28	Saját tőke/Összes kötelezettség
	X37	Pénzeszközök/Forgóeszközök	X37	Pénzeszközök/Forgóeszközök
	X39	(Forgóeszközök-Készletek)/Összes eszköz	X39	(Forgóeszközök-Készletek)/Összes eszköz
	X48	ln(Értékesítés árbevétele)	X48	ln(Értékesítés árbevétele)
	X50	Az évek száma a beszámolóig amióta működik a vállalat	X50	Az évek száma a beszámolóig amióta működik a vállalat
2014	X8	Adózott eredmény/Összes eszköz	X8	Adózott eredmény/Összes eszköz
	X20	Saját tőke/Mérlegfőösszeg	X10	Forgóeszközök/Rövid lejáratú kötelezettségek
	X32	Értékesítés nettó árbevétele/Követelések	X15	Működő tőke/Értékesítés nettó árbevétele
	X37	Pénzeszközök/Forgóeszközök	X19	Eredménytartalék/Összes eszköz
	X39	(Forgóeszközök-Készletek)/Összes eszköz	X23	Hosszú lejáratú kötelezettségek/Összes eszköz
	X48	ln(Értékesítés árbevétele)	X26	(Összes eszköz-Összes kötelezettség)/Összes kötelezettség
	X50	Az évek száma a beszámolóig amióta működik a vállalat	X38	Pénzeszközök/Összes eszköz
			X41	Hosszú lejáratú kötelezettségek/Összes kötelezettség
			X48	ln(Értékesítés árbevétele)
			X50	Az évek száma a beszámolóig amióta működik a vállalat
2015	X7	Működési eredmény/Összes eszköz	X3	Adózás előtti eredmény/Értékesítés nettó árbevétele
	X14	Működő tőke/Összes eszköz	X7	Működési eredmény/Összes eszköz
	X23	Hosszú lejáratú kötelezettségek/Összes eszköz	X14	Működő tőke/Összes eszköz
	X32	Értékesítés nettó árbevétele/Követelések	X23	Hosszú lejáratú kötelezettségek/Összes eszköz
	X37	Pénzeszközök/Forgóeszközök	X28	Saját tőke/Összes kötelezettség
	X39	(Forgóeszközök-Készletek)/Összes eszköz	X30	Értékesítés nettó árbevétele/Összes eszköz
	X50	Az évek száma a beszámolóig amióta működik a vállalat	X32	Értékesítés nettó árbevétele/Követelések
			X34	Készletek/Értékesítés nettó árbevétele
			X37	Pénzeszközök/Forgóeszközök
			X39	(Forgóeszközök-Készletek)/Összes eszköz
			X47	ln(Összes eszköz)
			X50	Az évek száma a beszámolóig amióta működik a vállalat

2016	X7	Működési eredmény/Összes eszköz	X9	Működési eredmény/Személyi jellegű ráfordítások
	X9	Működési eredmény/Személyi jellegű ráfordítások	X17	Készletek/Működő tőke
	X37	Pénzeszközök/Forgóeszközök	X19	Eredménytartalék/Összes eszköz
	X48	ln(Értékesítés árbevétele)	X24	(Összes eszköz-Összes kötelezettség)/Összes eszköz
	X50	Az évek száma a beszámolóig amióta működik a vállalat	X27	Saját tőke/Befektetett eszközök
			X28	Saját tőke/Összes kötelezettség
			X30	Értékesítés nettó árbevétele/Összes eszköz
			X37	Pénzeszközök/Forgóeszközök
			X39	(Forgóeszközök-Készletek)/Összes eszköz
			X50	Az évek száma a beszámolóig amióta működik a vállalat

Forrás: saját szerkesztés

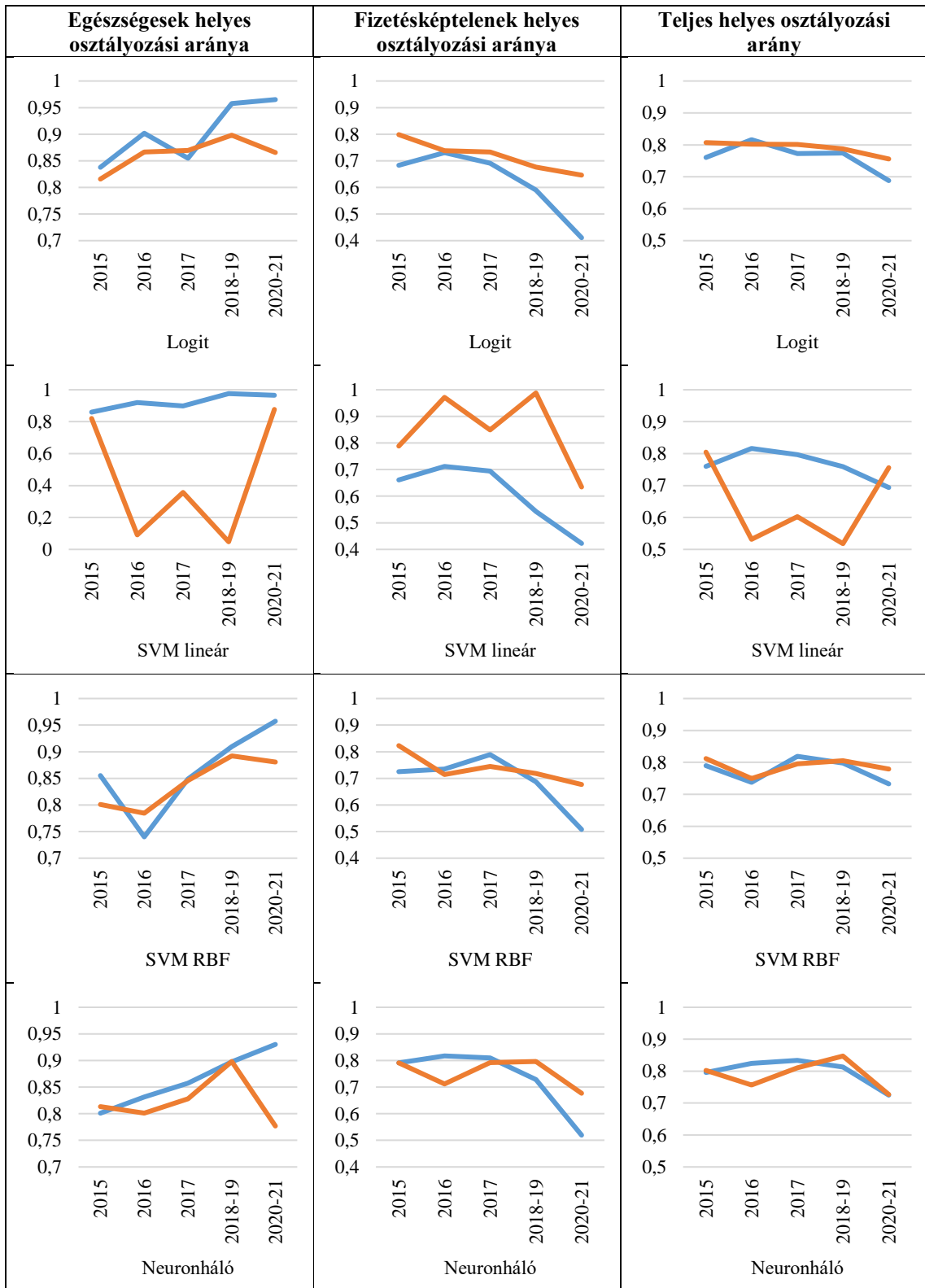
F6. függelék: A több év adatain épített modellek paramétere

Módszerek	Paraméterek	
	Tanuló minta: 2013-2014	Tanuló minta: 2013-2015
Logit	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = -0,07 - 1,08X_1 - 0,37X_3 - 0,14X_9 + 0,001X_{13} - 0,10X_{19} + 0,80X_{23} - 0,04X_{30} + 0,03X_{33} - 1,49X_{37} + 1,1X_{39} - 0,07X_{50}$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 38,9\%$</p>	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = -0,45 - 0,29X_1 - 0,2X_9 + 0,001X_{13} - 0,07X_{19} + 0,98X_{22} - 0,08X_{30} - 0,05X_{50}$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 37,5\%$</p>
SVM Lineáris	Kernel: lineáris Büntető paraméter (C): 2,99	Kernel: lineáris Büntető paraméter (C): 0,21
SVM RBF	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 990,39 Kernel paraméter (γ): 108,97	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 3,76 Kernel paraméter (γ): 5,17
Neurális háló	Rejtett rétegek száma: 1 Csomópontok száma a rejtett rétegben: 277 Futásszám: 1000 Függvényforma: köztes rétegben sigmoid, kimeneti rétegben softmax Lambda: $2,25e^{-5}$	Rejtett rétegek száma: 2 Csomópontok száma a rejtett rétegben: 284-275 Futásszám: 1000 Függvényforma: köztes rétegben sigmoid, kimeneti rétegben softmax Lambda: $2,99e^{-5}$
Bagging	Elágazási csomópontok: 42 Döntési fák száma: 86	Elágazási csomópontok: 227 Döntési fák száma: 172
Random Forest	Elágazási csomópontok: 2052 Döntési fák száma: 301 Véletlenszerűen kiválasztott független változók száma: 1	Elágazási csomópontok: 1289 Döntési fák száma: 241 Véletlenszerűen kiválasztott független változók száma: 2
Adaboost	Elágazási csomópontok: 3 Döntési fák száma: 16 Tanulási ráta: 0,54	Elágazási csomópontok: 13 Döntési fák száma: 15 Tanulási ráta: 0,30
	Tanuló minta: 2013-2016	Tanuló minta: 2013-2017
Logit	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = -1,14 - 0,53X_1 - 0,21X_9 - 0,57X_{19} + 0,03X_{33} - 1,29X_{37} + 1,29X_{39} + 0,09X_{47} - 0,06X_{50}$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 35,9\%$</p>	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = 1,83 - 0,75X_1 - 0,14X_9 + 0X_{10} - 0,63X_{24} - 1,53X_{37} + 1,42X_{39} - 0,18X_{48} - 0,04X_{50}$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 43,2\%$</p>
SVM Lineáris	Kernel: lineáris Büntető paraméter (C): 0,53	Kernel: lineáris Büntető paraméter (C): 0,80
SVM RBF	Kernel: Gaussi radiális bázisfüggvény	Kernel: Gaussi radiális bázisfüggvény

	Büntető paraméter (C): 483,13 Kernel paraméter (γ): 18,01	Büntető paraméter (C): 51,44 Kernel paraméter (γ): 8,54
Neurális háló	Rejtett rétegek száma: 1 Csomópontok száma a rejtett rétegben: 8 Futásszám: 1000 Függvényforma: köztes rétegben Tanh, kimeneti rétegben softmax Lambda: $4,82e^{-7}$	Rejtett rétegek száma: 2 Csomópontok száma a rejtett rétegben: 295-262 Futásszám: 1000 Függvényforma: köztes rétegben Tanh, kimeneti rétegben softmax Lambda: $4,32 e^{-6}$
Bagging	Elágazási csomópontok: 1147 Döntési fák száma: 499	Elágazási csomópontok: 130 Döntési fák száma: 496
Random Forest	Elágazási csomópontok: 3554 Döntési fák száma: 406 Véletlenszerűen kiválasztott független változók száma: 1	Elágazási csomópontok: 83 Döntési fák száma: 182 Véletlenszerűen kiválasztott független változók száma: 8
Adaboost	Elágazási csomópontok: 24 Döntési fák száma: 249 Tanulási ráta: 0,47	Elágazási csomópontok: 56 Döntési fák száma: 41 Tanulási ráta: 0,07
	Tanuló minta: 2013-2019	
Logit	$P(y = 1) = \frac{1}{1 + e^{-z}}$ <p>ahol $Z = 1,89 - 1,60X_1 + 0,85X_2 - 1,82X_3 + 1,46X_4 - 0,2X_5 - 0,04X_6$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 38,8\%$</p>	
SVM Lineáris	Kernel: lineáris Büntető paraméter (C): 2,44	
SVM RBF	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 962,09 Kernel paraméter (γ): 9,92	
Neurális háló	Rejtett rétegek száma: 3 Csomópontok száma a rejtett rétegben: 7-68-65 Futásszám: 1000 Függvényforma: köztes rétegben Tanh, kimeneti rétegben softmax Lambda: $1,34e^{-5}$	
Bagging	Elágazási csomópontok: 113 Döntési fák száma: 492	
Random Forest	Elágazási csomópontok: 114 Döntési fák száma: 220 Véletlenszerűen kiválasztott független változók száma: 2	
Adaboost	Elágazási csomópontok: 2 Döntési fák száma: 498 Tanulási ráta: 0,98	

Forrás: saját szerkesztés

F7. függelék: A többéves és egy év adatain épített modellek osztályozási pontosságainak trendje a különböző tesztmintákon





Forrás: saját szerkesztés

F8. függelék: A mezoszintű adatokon épített modellek paraméterei

Módszerek	Paraméterek	
	Tanuló minta: 2013	Tanuló minta: 2014
Logit	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = 0,49 - 0,39X_3 - 0,20X_9 - 0,08X_{15} + 0,42X_{21} + 0X_{28} - 1,87X_{37} + 1,26X_{39} - 0,17X_{48} - 0,04X_{50} - 0,00Y_1 + 16,75Y_2 - 0,15Y_3 + 0,19Y_7$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 40,0\%$</p>	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = 1,57 - 1,32X_8 + 0,00X_{10} - 0,03X_{15} - 0,02X_{19} + 0,57X_{23} + 0,00X_{26} - 1,45X_{38} - 0,00X_{41} - 0,17X_{48} - 0,06X_{50} - 0,00Y_1 + 11,84Y_2 + 0,012Y_3 + 0,02Y_7$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 34,0\%$</p>
SVM Lineáris	Kernel: lineáris Büntető paraméter (C): 999,99	Kernel: lineáris Büntető paraméter (C): 96,26
SVM RBF	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 981,29 Kernel paraméter (γ): 55,84	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 988,67 Kernel paraméter (γ): 49,59
Neurális háló	Rejtett rétegek száma: 1 Csomópontok száma a rejtett rétegben: 55 Futásszám: 1000 Függvényforma: köztes rétegben sigmoid, kimeneti rétegben softmax Lambda: $3,92e^{-5}$	Rejtett rétegek száma: 2 Csomópontok száma a rejtett rétegben: 5-32 Futásszám: 1000 Függvényforma: köztes rétegben sigmoid, kimeneti rétegben softmax Lambda: $5,25e^{-5}$
Bagging	Elágazási csomópontok: 46 Döntési fák száma: 148	Elágazási csomópontok: 1027 Döntési fák száma: 131
Random Forest	Elágazási csomópontok: 430 Döntési fák száma: 184 Véletlenszerűen kiválasztott független változók száma: 1	Elágazási csomópontok: 17 Döntési fák száma: 10 Véletlenszerűen kiválasztott független változók száma: 12
Adaboost	Elágazási csomópontok: 4 Döntési fák száma: 361 Tanulási ráta: 0,19	Elágazási csomópontok: 18 Döntési fák száma: 338 Tanulási ráta: 0,24
	Tanuló minta: 2015	Tanuló minta: 2016
Logit	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = -2,14 - 0,30X_3 - 1,18X_7 - 1,25X_{14} + 0,31X_{23} + 0,003X_{28} - 0,25X_{30} + 0,00X_{32} + 0,62X_{34} - 1,06X_{37} + 2,66X_{39} + 0,09X_{47} - 0,06X_{50} - 0,000004Y_1 + 6,99Y_2 + 0,62Y_3 + 0,33Y_7$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 53,0\%$</p>	$P(y = 1) = \frac{1}{1 + e^{-Z}}$ <p>ahol $Z = -0,36 - 0,28X_9 - 0,02X_{17} + 0,18X_{19} - 1,62X_{24} + 0,00X_{27} - 0,08X_{28} - 0,09X_{30} - 1,05X_{37} + 2,29X_{39} - 0,05X_{50} - 0,00Y_1 + 9,56Y_2 + 0,15Y_3 - 0,02Y_7$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 48,2\%$</p>
SVM Lineáris	Kernel: lineáris Büntető paraméter (C): 198,84	Kernel: lineáris Büntető paraméter (C): 0,23

SVM RBF	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 983,90 Kernel paraméter (γ): 61,31	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 810,16 Kernel paraméter (γ): 97,60
Neurális háló	Rejtett rétegek száma: 2 Csomópontok száma a rejtett rétegben: 71-2 Futásszám: 1000 Függvényforma: köztes rétegben sigmoid, kimeneti rétegben softmax Lambda: $8,51e^{-7}$	Rejtett rétegek száma: 2 Csomópontok száma a rejtett rétegben: 1-6 Futásszám: 1000 Függvényforma: köztes rétegben sigmoid, kimeneti rétegben softmax Lambda: 0,00026
Bagging	Elágazási csomópontok: 808 Döntési fák száma: 115	Elágazási csomópontok: 83 Döntési fák száma: 495
Random Forest	Elágazási csomópontok: 663 Döntési fák száma: 61 Véletlenszerűen kiválasztott független változók száma: 1	Elágazási csomópontok: 172 Döntési fák száma: 497 Véletlenszerűen kiválasztott független változók száma: 2
Adaboost	Elágazási csomópontok: 50 Döntési fák száma: 47 Tanulási ráta: 0,86	Elágazási csomópontok: 12 Döntési fák száma: 12 Tanulási ráta: 0,30
Tanuló minta: 2017		
Logit	$P(y = 1) = \frac{1}{1 + e^{-z}}$ <p>ahol $Z = -1,22 - 1,87X_1 - 0,98X_{14} + 0,00X_{32} - 1,57X_{37} + 2,68X_{39} - 0,06X_{50} - 0,00Y_1 + 11,76Y_2 + 0,68Y_3 + 0,17Y_7$</p> <p>Megjegyzés: A félkövérrel jelölt változó szignifikáns 0,01 szignifikancia szinten. Az <u>aláhúzással</u> jelölt változó szignifikáns 0,05 szignifikancia szinten.</p> <p>Nagelkerke $R^2 = 50,8\%$</p>	
SVM Lineáris	Kernel: lineáris Büntető paraméter (C): 176,92	
SVM RBF	Kernel: Gaussi radiális bázisfüggvény Büntető paraméter (C): 974,61 Kernel paraméter (γ): 30,88	
Neurális háló	Rejtett rétegek száma: 2 Csomópontok száma a rejtett rétegben: 1-293 Futásszám: 1000 Függvényforma: köztes rétegben Tanh, kimeneti rétegben softmax Lambda: $3,48e^{-6}$	
Bagging	Elágazási csomópontok: 12 Döntési fák száma: 10	
Random Forest	Elágazási csomópontok: 542 Döntési fák száma: 28 Véletlenszerűen kiválasztott független változók száma: 1	
Adaboost	Elágazási csomópontok: 3 Döntési fák száma: 10 Tanulási ráta: 0,52	

Forrás: saját szerkesztés

F9. függelék: A mezosztintű változókkal és kizárólag pénzügyi változókkal készült modellek osztályozási eredményei

	Mezosztintű változókkal			Pénzügyi változókkal		
	Tanuló minta 2013, tesztelő minta 2014					
	Egészségesek helyes besorolási aránya	Fizetésképtelenek helyes besorolási aránya	Teljes helyes besorolási arány	Egészségesek helyes besorolási aránya	Fizetésképtelenek helyes besorolási aránya	Teljes helyes besorolási arány
Logit	85,63%	68,74%	77,18%	86,02%	69,13%	77,57%
SVM lineár	95,15%	12,23%	53,69%	88,16%	68,54%	78,35%
SVM RBF	16,70%	79,42%	48,06%	89,13%	74,37%	81,75%
Neuronháló	85,83%	75,15%	80,49%	87,18%	77,86%	82,52%
Bagging	86,21%	81,94%	84,08%	86,80%	81,75%	84,27%
Random Forest	87,18%	82,72%	84,95%	86,99%	80,58%	83,79%
Adaboost	87,96%	80,78%	84,37%	88,74%	79,42%	84,08%
Tanuló minta 2014, tesztelő minta 2015						
Logit	77,89%	72,73%	75,31%	81,57%	79,85%	80,71%
SVM lineár	47,67%	60,69%	54,18%	82,06%	78,87%	80,47%
SVM RBF	9,58%	86,73%	48,16%	80,10%	82,31%	81,20%
Neuronháló	82,06%	73,46%	77,76%	81,33%	79,12%	80,22%
Bagging	76,66%	78,13%	77,40%	76,41%	77,45%	76,90%
Random Forest	74,20%	79,36%	76,78%	75,43%	77,89%	76,66%
Adaboost	74,45%	77,40%	75,92%	75,68%	81,33%	78,50%
Tanuló minta 2015, tesztelő minta 2016						
Logit	87,59%	76,11%	81,85%	86,65%	73,77%	80,21%
SVM lineár	91,10%	26,70%	58,90%	9,13%	97,19%	53,16%
SVM RBF	10,54%	88,99%	49,77%	78,45%	71,43%	74,94%
Neuronháló	79,63%	65,11%	72,37%	80,09%	71,14%	75,64%
Bagging	84,54%	81,50%	83,02%	85,01%	82,44%	83,72%
Random Forest	85,95%	80,33%	83,14%	84,07%	82,67%	83,37%
Adaboost	86,89%	75,64%	81,26%	86,65%	79,86%	83,26%
Tanuló minta 2016, tesztelő minta 2017						
Logit	87,54%	75,37%	81,45%	86,94%	73,29%	80,12%
SVM lineár	45,40%	72,11%	58,75%	35,61%	84,87%	60,24%
SVM RBF	13,65%	90,50%	52,08%	84,57%	74,48%	79,53%
Neuronháló	91,69%	65,58%	78,64%	82,79%	79,23%	81,01%
Bagging	89,02%	77,45%	83,23%	87,54%	77,45%	82,49%
Random Forest	90,21%	77,45%	83,83%	87,54%	77,74%	82,64%
Adaboost	90,21%	78,34%	84,27%	87,24%	79,82%	83,53%
Tanuló minta 2017, tesztelő minta 2018/19						
Logit	91,57%	65,66%	78,61%	89,82%	67,66%	78,74%
SVM lineár	47,59%	52,41%	50,00%	4,79%	98,80%	51,80%
SVM RBF	5,42%	98,19%	51,81%	89,22%	71,86%	80,54%
Neuronháló	96,99%	1,20%	49,10%	89,82%	79,64%	84,73%
Bagging	92,17%	78,31%	85,24%	93,41%	77,84%	85,63%
Random Forest	92,77%	74,10%	83,43%	92,22%	79,04%	85,63%
Adaboost	93,37%	75,90%	84,64%	93,41%	76,65%	85,03%

Forrás: saját számítás

F10. függelék: A többéves előrejelzési horizonton mért modellek osztályozási eredménye

	Egészségesek helyes besorolási aránya	Fizetéképtelenek helyes besorolási aránya	Teljes helyes besorolási arány	Egészségesek helyes besorolási aránya	Fizetéképtelenek helyes besorolási aránya	Teljes helyes besorolási arány
Tanuló minta 2013, tesztelő minta 2015			Tanuló minta 2013, tesztelő minta 2016			
Logit	85,01%	68,06%	76,54%	86,42%	70,73%	78,57%
SVM lineár	86,24%	66,34%	76,29%	92,27%	71,19%	81,73%
SVM RBF	85,50%	72,73%	79,12%	87,35%	76,58%	81,97%
Neuronháló	81,82%	74,94%	78,38%	85,48%	76,81%	81,15%
Bagging	84,77%	77,15%	80,96%	85,48%	80,80%	83,14%
Random Forest	83,78%	75,68%	79,73%	86,89%	79,39%	83,14%
Adaboost	86,49%	73,22%	79,85%	86,65%	77,75%	82,20%
Tanuló minta 2014, tesztelő minta 2016			Tanuló minta 2014, tesztelő minta 2017			
Logit	80,80%	78,45%	79,63%	83,98%	73,00%	78,49%
SVM lineár	83,14%	78,69%	80,91%	83,98%	71,81%	77,89%
SVM RBF	78,69%	79,86%	79,27%	80,12%	75,37%	77,74%
Neuronháló	81,73%	79,86%	80,80%	83,09%	74,78%	78,93%
Bagging	78,69%	81,97%	80,33%	80,42%	77,45%	78,93%
Random Forest	79,16%	84,54%	81,85%	79,23%	79,82%	79,53%
Adaboost	76,35%	85,95%	81,15%	78,04%	80,42%	79,23%
Tanuló minta 2015, tesztelő minta 2017			Tanuló minta 2015, tesztelő minta 2018/19			
Logit	86,94%	69,44%	78,19%	94,01%	60,48%	77,25%
SVM lineár	10,68%	95,25%	52,97%	18,56%	97,01%	57,78%
SVM RBF	79,53%	71,81%	75,67%	91,62%	60,48%	76,05%
Neuronháló	80,42%	68,84%	74,63%	91,02%	64,07%	77,54%
Bagging	86,05%	76,26%	81,16%	93,41%	76,05%	84,73%
Random Forest	85,46%	78,34%	81,90%	92,81%	76,65%	84,73%
Adaboost	88,43%	74,78%	81,60%	94,61%	70,66%	82,63%
Tanuló minta 2016, tesztelő minta 2018/19			Tanuló minta 2016, tesztelő minta 2020/21			
Logit	88,62%	67,66%	78,14%	89,23%	58,46%	73,85%
SVM lineár	51,50%	56,89%	54,19%	52,69%	73,46%	63,08%
SVM RBF	91,62%	70,06%	80,84%	91,54%	55,38%	73,46%
Neuronháló	92,22%	67,66%	79,94%	93,85%	58,08%	75,96%
Bagging	91,62%	74,25%	82,93%	92,31%	55,00%	73,65%
Random Forest	91,62%	74,85%	83,23%	93,08%	57,31%	75,19%
Adaboost	92,22%	73,05%	82,63%	91,92%	56,92%	74,42%
Tanuló minta 2017, tesztelő minta 2020/21						
Logit	91,15%	57,69%	74,42%			
SVM lineár	5,77%	97,69%	51,73%			
SVM RBF	89,62%	54,62%	72,12%			
Neuronháló	91,92%	58,08%	75,00%			
Bagging	92,69%	51,92%	72,31%			
Random Forest	91,92%	58,85%	75,38%			
Adaboost	92,69%	58,08%	75,38%			

Forrás: saját számítás