KÖZGAZDASÁGI SZEMLE, LXV. ÉVF., 2018. JÚNIUS (592–618. o.)

TAKÁCS OLGA–VINCZE JÁNOS

# Bérelőrejelzések – prediktorok és tanulságok

**Ebben az írásban a magyarországi „bérfüggvényt" elemezzük bizonyos adatbányász – induktív statisztikai – technikák segítségével. A klasszifikációs és regressziós fák *(Classification and Regression Trees, CART)* módszerének eredeti célja elsősorban a predikció. Nagyon jól értelmezhető eredményt ad, ami az előzetes adatelemzési funkcióban lényeges. Ezért a CART-elemzés alapján megfogalmaztunk bizonyos sejtéseket az alapvető bérezést érintő problémákkal kapcsolatban is. „Véletlenerdő-algoritmus" felhasználásával ellenőriztük a változók magyarázó ereje szerinti fontosságának robusztusságát. Mindkét módszer alapján a béreket „meghatározó" két legfontosabb tényező a képzettség és a vállalatméret.**
**Journal of Economic Literature (JEL) kód: C14, J31.**

## Bevezetés

A bérek alakulása számos munkagazdaságtani tanulmány problémája. Vannak-e ágazati vagy területi különbségek a bérezésben? Létezik-e nemek közötti diszkrimináció? Mekkora a képzettségi prémium? Van-e hatása a vállalat nagyságának, a munkavállalók szervezettségének vagy a velük kötött szerződések típusának a bérekre? Igaz-e, hogy a külföldi tulajdonú cégek nagyobb bért fizetnek, vagy ez inkább az állami vállalatokra igaz? Az utóbbi évtizedekben több, Magyarországra vonatkozó empirikus munkagazdaságtani elemzés foglalkozott a bérekkel kapcsolatos fenti problémákkal. Mi itt egy látszólag egyszerű, predikciós problémát vizsgálunk: ha adottak valakinek bizonyos egyéni (életkor, képzettség, szolgálati idő stb.) és munkapiaci jellemzői (a vállalat mérete, ágazat, földrajzi elhelyezkedés), akkor hogyan tippelnénk meg a bérét a lehető legpontosabban? Azt gondoljuk, hogy ha jó predikciós módszert találunk, az informatív lehet e kérdések megválaszolásához is.

A következőkben először a bérezéssel kapcsolatos néhány magyar empirikus tanulmány számunkra érdekesnek tűnő eredményeit foglaljuk össze. Majd leírjuk azt az

*Takács Olga,* Budapesti Corvinus Egyetem.
*Vincze János,* Budapesti Corvinus Egyetem és MTA KRTK KTI.
A kézirat első változata 2017. december 6-án érkezett szerkesztőségünkbe.
DOI: http://dx.doi.org/10.18414/KSZ.2018.6.592

adatbányász-technikát, amelyet prediktorunk előállításához használtunk. Az algoritmus általános bemutatását követően a módszer működését tárgyaljuk, valamint azt, hogy milyen módon nyerhetünk belőle ki információkat. Az adatok ismertetése után az általunk felvetett problémákkal összefüggésben bemutatjuk és értelmezzük az eredményeket. Ezt követően a kapott eredmények robusztusságát ellenőrizzük, végül pedig összefoglaljuk megállapításainkat és kitekintést nyújtunk.

## Empirikus béregyenletek Magyarországon

A bevezetésben bemutatott kérdések és problémák megválaszolásához kapcsolódó magyar eredményeket tekintjük át az alábbiakban, érintjük a nemek közötti különbségeket, az iskolai végzettség és a gyakorlati idő hatását, a regionális és ágazati bérkülönbségeket, valamint a tulajdonjog kérdését.

### *Nemek közti különbségek*

A nemi diszkriminációt illetően *Lovász* [2008] arra a következtetésre jutott, hogy 1986–2003 között Magyarországon a rendszerváltást követő piaci liberalizáció következtében a férfiak és a nők közötti bérkülönbségek csökkentek. Érvelése szerint a versenypiaci helyzet kialakulása a hatékonyság növelésére ösztönözte a vállalatokat, így egyre kevésbé lehetett fenntartani a nemi megkülönböztetést. Lovász Anna egy későbbi tanulmányában (*Lovász* [2013]) vizsgálta a bérek eloszlásánál jelentkező különbségeket a férfiak és a nők esetében. Itt azt találta, hogy a bértáblán felfelé haladva, egyre inkább növekedett a nemi diszkrimináció, ami az úgynevezett „üvegplafon" jelenség meglétére utal, vagyis arra, hogy a nők számára a szervezeti hierarchiában való előrejutásnak van egy bizonyos korlátja.

### *Végzettségi prémium, gyakorlati idő*

Az életpálya-kereseti profilok vizsgálatánál *Gábor* [2008] arra az eredményre jutott, hogy a magasabb iskolai végzettséggel rendelkezők kezdő jövedelmei nagyobbak. A férfiak kezdő bére pedig minden iskolázottsági szinten magasabb a nőkénél. Emellett megállapította, hogy a kereset és a gyakorlati idő között konkáv kapcsolat áll fenn. Ez a konkáv kapcsolat a magasabb és az alacsonyabb iskolai végzettség, valamint a férfiak és nők vonatkozásában erősebb. *Galasi* [2008] eredményei szerint a felsőfokú végzettséghez kapcsolódó bérprémium mértéke folyamatosan emelkedett 1994 és 2004 között. Az emelkedést a szerző elsősorban a munkakeresleti görbe eltolódására vezeti vissza, vagyis arra, hogy a munkáltatók bizonyos munkakörökben egyre nagyobb bért hajlandók fizetni. Az ezredfordulóig ez az átsorolási hatás jelen volt a verseny- és az állami szférában is, azonban az előbbiben erőteljesebben jelentkezett. Az ezredforduló után az átsorolási hatás veszített jelentőségéből.

*Regionális és települések közötti különbségek*

A településtípusok és a regionális tagozódás bérekre gyakorolt hatását vizsgálta *Szabó* [2006] a versenyszférában 1998 és 2004 között. A szerző a településtípusokkal kapcsolatban arra a megállapításra jutott, hogy a bérezésben található eltérések jelentéktelenek a megyeközpontok, a kisebb a városok és a falvak esetében, azonban Budapest – a többi településtípussal összevetve – öt százalék alatti nettó bérelőnnyel rendelkezik. A régiók vizsgálatakor a szerző a leggazdagabb régiókat – Közép-Magyarországot, Közép-Dunántúlt és Nyugat-Dunántúlt – hasonlította össze a legszegényebb Észak-Alföld régióval. A szerző megállapította, hogy a gazdagabb régiók 3–5 százalékos bérelőnnyel rendelkeznek Észak-Alföldhöz képest.

*Köllő* [2003] vizsgálta a költségvetési szférában található regionális és településtípusokhoz kötődő bérkülönbségeket. Arra a következtetésre jutott, hogy nincsenek jelentős különbségek a költségvetési szférában. Itt a régiók közötti eltérések nem szignifikánsak, ami a közalkalmazotti és a köztisztviselői bértábla meglétével magyarázható. A megyeközpontok, a kisebb városok és a falvak közötti bérkülönbségek elhanyagolhatók. Ezekkel a településtípusokkal összehasonlítva Budapest jelentős bérelőnnyel rendelkezik. Ez az eltérés arra vezethető vissza, hogy a kisebb településeken főként olyan intézmények működnek, amelyeknek a bérkiadásai kisebbek. Ilyen intézmények az alap- vagy középfokú iskolák, a kevésbé komplex egészségügyi létesítmények, valamint az alapfokú közigazgatási intézmények.

*Ágazati bérkülönbségek, kollektív szerződések*

*Kertesi–Köllő* [2003] az ágazati bérkülönbségeket tanulmányozta. A szerzőpáros célja elsősorban az ágazatok, azon belül is az ágazati koncentráció, valamint a vállalati szintű szervezettség bérekre gyakorolt hatásának mérése volt. Így a becslés során az egyéni és a vállalati jellemzők mellett kontrolláltak az előbb felsorolt változókra is. A változók közül az ágazati koncentrációt szerepeltették a monopoljáradék nagyságának, a szervezettséget pedig a munkavállalók alkuerejének közelítő változójaként. Mivel vélhetően a monopoljáradék nagysága és a munkavállalók alkuereje kölcsönösen hat egymásra, így szimultaneitási probléma lép fel. Ennek feloldására a szerzők kétlépcsős eljárást alkalmaztak. Arra a megállapításra jutottak, hogy a szervezettség növekedése szignifikánsan hat a bérekre, ha a piac közepesen vagy erősen koncentrált. Továbbá erős koncentráltság mellett nagyobb a szervezettség hatása, mint közepes koncentráció mellett. A kapcsolat fordítva is igaz: magasabb szervezettség mellett nagyobb a koncentráció bérre gyakorolt hatása. Emellett azt találták, hogy ha magas a szervezettség egy ágazatban, akkor azoknál a vállalatoknál is szignifikánsan magasabbak a bérek, amelyek nem kötnek béralkukat.

*Rigó* [2012] megvizsgálta a kollektív és az egyéni bérmegállapodások bérekre gyakorolt hatását 1992 és 2008 között. A szerző az egyéni és a vállalati jellemzőkre, valamint a vállalati fix hatásokra is kontrollált. A vállalati fix hatások szerepeltetésével figyelembe vette azokat az időben állandó, de az adatbázisban nem szereplő

tényezőket, amelyek befolyásolhatják a bérek alakulását. Ilyen tényező lehet a vállalat tőkefelszereltsége, a vezetés hatékonysága, a vállalati struktúra vagy a jövőbeli profitvárakozások. Amennyiben csak az egyéni és a vállalati jellemzőkre kontrollált – és a szerződések típusait külön-külön szerepeltette a béregyenletben –, akkor a kollektív szerződések esetében 1 százalékponttal nagyobb a bérelőny, mint a bérmegállapodások esetében. A vállalati fix hatásokat is figyelembe véve, a bérelőnyök csökkennek, azonban továbbra is 1 százalékos szinten is szignifikánsak maradnak. Amennyiben a szerző a két szerződéstípus (kollektív szerződések, bérmegállapodások) kétértékű változóit együtt szerepeltette, csak az egyéni és a vállalati jellemzőkre kontrollálva, a bérmegállapodások hatása meghaladta a kollektív szerződésekét. A hatás 5 százalékos szinten továbbra is szignifikáns volt. A vállalati fix hatások figyelembevételével a szerződések bérekre gyakorolt hatása mindkét szerződéstípus esetében csökkent, sőt a kollektív szerződések bérekre gyakorolt hatása meghaladta a bérmegállapodásokét. Továbbá csak a kollektív szerződés hatása maradt – 10 százalékos szinten – szignifikáns. Vagyis azok a vállalatok, ahol olyan erős a szakszervezet, hogy képes kollektív szerződések kikényszerítésére, ott az egyéni bérmegállapodások hatása elhanyagolható. Ez pedig megerősíti *Kertesi–Köllő* [2003] eredményeit.

### *Tulajdon*

A külföldi és a belföldi tulajdon hazai bérekre gyakorolt hatását vizsgálta *Earle–Telegdy* [2012*a*]. A szerzők arra a kérdésre keresték a választ, hogy a külföldiek által megvásárolt belföldi vállalatoknál mekkora bérelőnyre tesznek szert a munkavállalók, összehasonlítva a belföldi kézben maradt cégekkel. Megállapították, hogy a külföldi befektetők leginkább olyan vállalatokat vásároltak fel, amelyek már eleve magasabb béreket fizettek. Ez feltehetően arra vezethető vissza, hogy a felvásárolt vállalatok – amelyek addig belföldi tulajdonban voltak – jobban képzett munkaerőt alkalmaztak. A jobban képzett munkavállalóknak pedig a belföldi tulajdonosok is magasabb béreket fizettek.

A szerzők vizsgálták annak a hatását, hogy mi történik, ha a külföldi tulajdonos később egy hazai befektetőnek adta el a vállalatot. Ebben az esetben a külföldi felvásárlás egyértelműen növelte a béreket, és ez a bérelőny bár csökken az eladással, mégsem tűnt el teljesen. A külföldi felvásárlás béremelő hatása megjelent mindkét nemnél, az összes végzettségi szintnél és az összes munkatapasztalat esetében. Vagyis összefoglalva: bár különböző mértékben, de a felvásárlással minden munkavállaló jobban járt. A szerzők a külföldi felvásárlás hatására történt béremelkedést elsősorban az országok közötti termelékenységkülönbséggel magyarázták. Ezenfelül a béremelés mértéke nagyobb volt, ha a felvásárlás a rendszerváltást követő időszakban – megközelítően 1998-ig – ment végbe, és ha az adott vállalat előtte állami kézben volt.

*Earle–Telegdy* [2012*b*] a rendszerváltást követő privatizáció bérekre gyakorolt hatását számszerűsítette annak függvényében, hogy belföldi vagy külföldi tulajdonos vásárolta-e fel a vállalatot. A szerzők megállapították, hogy amennyiben belföldi vállalat vásárolta meg a rendszerváltásig állami kézben lévő vállalatot, akkor ott a bérek csökkentek az állami tulajdonban maradt vállalatokhoz képest. Ezzel szemben,

ha külföldi cég vásárolta fel az adott vállalatot, akkor ott a bérek emelkedtek. Amennyiben a szerzők figyelembe vették a vállalati fix hatásokat is, akkor a bérek közötti eltérések mérséklődtek, de fennmaradtak.

## Regressziós fák

A hagyományos empirikus módszer a fenti kérdések vizsgálatánál – mint általában az ökonometriában – valamilyen paraméteres regressziós egyenlet becslése. A paramétereknek igyekszünk „kauzális" vagy „strukturális" értelmezést adni, ami nem nélkülözhet bizonyos feltevéseket. A legrugalmasabb modellben is meglehetősen kötött a függvényforma, ami ellenőrizhetetlen, és gyakran gyanús hipotéziseket involvál az „adatgeneráló" folyamatról. A regressziós elemzés sikerét általában a szignifikáns paraméterek és a mintán belüli jó illeszkedés méri, a hangsúly főleg a hipotézisek tesztelésén van.

A statisztikai tanulási – adatbányász-, gépi tanulási – irodalom és annak alkalmazói, ha nem is ortogonálisan, de több tekintetben eltérő alapelvekből és kérdésekből indulnak ki. Egyfelől sokkal nagyobb fontosságot tulajdonítanak a becslések mintán kívüli teljesítményének – úgynevezett prediktív képességének – az eredmények elbírálásánál. Másrészt kevésbé érdeklődnek az egyedi paraméterek „szignifikanciája" – és általában elméleti hipotézisek tesztelése – iránt. Azt, amit a közgazdászok gyakran előzetes adatelemzésnek neveznek, nem tekintik feltétlenül „bevezetésnek", hanem célnak is, amelynek önmagában is hasznos mondanivalója van. A függvényformával kapcsolatban nagy rugalmasságra törekednek, nem szégyellik azt, hogy indukcióra használják a módszereket, azaz az adatokból nem törekednek kizárólag deduktív következtetéseket levonni. Központi fogalmuk a torzításvariancia-átváltás *(bias-variance tradeoff)* és az a felismerés, hogy bár a komplexebb modellek könnyebben érhetnek el torzítatlan becslést, ezek előrejelzési és általánosítási képessége gyakran rosszabb, mint a kevésbé komplex modelleké.

Az utóbbi években egyre több közgazdász törekszik ezek a módszerek intenzívebb használatára. Ez részben összefüggésben van a *big data*-jelenséggel (lásd például *Varian* [2014]), amely egyfelől a korábbiaknál nagyobb tömegű adat jelenlétét, másrészt pedig a kezelésükhöz szükséges számítástechnikai kapacitás megnövekedését jelenti. Fontos látni, hogy az induktív statisztikai filozófia alkalmazásának nem szükséges feltétele a *big data* jelenléte. A cél lehet rejtett tények felfedezése, amelyek nem feltétlenül csak több petabájtos adathalmazban maradnak rejtve. Nem szükséges a hagyományos ökonometriai módszerek helyettesítőjének tekinteni ezeket a módszereket, hiszen ezek szolgáltathatnak inputot a hagyományosabb modellek specifikációjához is. Az induktív statisztikai megközelítés a hagyományos ökonometriai megközelítéshez képest alternatív nézőpontból tekint az adatokra, ami hasznos lehet, mivel kevés a remény arra, hogy bármely konkrét nézőpont tökéletesen kielégítő lehetne.

Ebben a cikkben a magyarországi „bérfüggvényt" elemezzük egy induktív statisztikai adatbányász-technikával. Ez a klasszifikációs és regressziós fák *(Classification and Regression Trees, CART)* módszere, amely mintegy 40 éves algoritmus, tehát a *big*

*data* előtti korban is jelen volt. Tudjuk, hogy a CART előrejelző funkcióban javítható, és vannak olyan algoritmusok, amelyek megvalósítják ezt a javítást. Viszont nagyon jól értelmezhető eredményt ad, ami az előzetes adatelemzési funkcióban lényeges. Érdemes megjegyezni azt is, hogy a CART-algoritmus ugyanúgy egy általános függvényközelítő eljárás, mint például a neurális hálók, tehát potenciálisan bármilyen függvény hozzátartozik az értelmezési tartományához.

Az eredmények értelmezésénél a bevezetés elején említett problémaköröket, kérdéseket tartottuk szem előtt. Az ott említett empirikus tanulmányok oksági, strukturális kérdéseket tettek fel ezekkel kapcsolatban. Egy regressziós fa önmagában nem alkalmas ilyen kérdések megválaszolására, ehhez az egyes konkrét jelenségek alapos, nem csak statisztikai jellemzőinek ismeretére van szükség. Egyetlen ilyen elemzés is külön tanulmányt igényelne, ezért igyekszünk is ellenállni annak a csábításnak, hogy merész, de felületes következtetéseket vonjunk le a CART-outputból. Azt szeretnénk megmutatni, hogy ez az alternatív szemléletű vizsgálat milyen érdekes megfigyelésekkel gazdagíthatja az egyes jelenségekről való tudásunkat.

## *A CART-ról általában*

A CART-algoritmus az adatbányász-irodalom egyik legnépszerűbb eljárása (lásd *Wu és szerzőtársai* [2008]). Az induktív statisztika „alapkövetelményének" megfelelően a CART predikciós, generalizációs teljesítményét pozitívan szokás értékelni. Például *Razi–Athappilly* [2005] általánosságban jobbnak találta a nemlineáris regresszióhoz képest, különösképpen, amikor nagyszámú redundáns magyarázó változó van.[1] A predikción kívüli célokra is számos komplex rendszert vizsgáló tudomány alkalmazta a CART-algoritmust. Például *Choi és szerzőtársai* [2013] CART-modellekkel becsülte bizonyos szennyező tevékenység légszennyezettségre való hatását olyan körülmények között, amikor a hatásokat számos szezonális és regionális „zavaró" változó is befolyásolta. (Egyéb ökológiai alkalmazásokra lásd *De'ath–Fabricius* [2000].) Kicsit közelebb a közgazdaságtanhoz *King–Resick* [2014] pszichológiai kutatásokra alkalmazott CART-elemzést, prediktorok identifikálására, illetve a köztük levő interakciók felfedezésére. Orvosi-egészség-gazdaságtani alkalmazások sem ritkák. *Lemon és szerzőtársai* [2003] a páciensek szegmentálási problémájának megoldására javasolta a CART-ot, vagyis arra, hogy azonosítsák azokat a csoportokat, amelyek számára bizonyos kezelések megfelelők.

A közgazdasági irodalomban *Durlauf–Johnson* [1995] már több mint 20 éve alkalmazták ezt a módszert keresztmetszeti növekedési regressziók nemlinearitásának a vizsgálatára. Következtetésük szerint a CART segített felfedni több adatrezsim létét, amelyek összhangban voltak egy olyan növekedési elmélettel, amelyben többszörös állandósult állapot van. *Minier* [2003] egy hasonló kérdésfeltevésre keresve a választ, a részvénypiacok gazdasági fejlődésben játszott szerepének szegmentáltságát látta igazolva.

[1] Természetesen olyankor, amikor tudjuk, hogy vannak redundáns változóink, de nem tudjuk, melyek azok.

*Galletta* [2016] ugyanebben az irányban használta a CART-ot „boldogságbecslések" nemlinearitásának és heterogenitásának vizsgálatára. Magyar vonatkozású alkalmazás is létezik már (lásd *Schiltz és szerzőtársai* [2017]), ahol a magyar középfokú oktatási termelési függvény becslésére történt kísérlet – többek között – CART segítségével.

### *A CART-algoritmus alapjai*

A CART valójában egy algoritmuscsalád, amelyet három lépésben lehet összegezni. Az alábbi leírás a regressziós változatra koncentrál, mivel ebben a tanulmányban ezt alkalmazzuk. Csupán egy intuitív összefoglalót adunk, részletesebb és matematikailag pontos leírások számos hivatkozásban megtalálhatók, kezdve a klasszikus referenciától (*Breiman és szerzőtársai* [1984]), de létezik friss áttekintő tanulmány is (*Loh* [2014]). Az általunk használt algoritmus az 'rpart' R-ben írt program. Az egyes lépésekről és fogalmakról részletes és pontos leírás található *Therneau–Atkinson* [2018]-ban.

### *Építsünk fát!*

Induljunk ki a teljes adathalmazból, ahol összesen $n$ megfigyelés van. A (magyarázandó) célváltozó kvantitatív változó, és létezik $K$ számú (magyarázó) inputváltozó. A fa kialakítása lényegében azt jelenti, hogy a teljes $n$ elemű halmazt minden lépésben 1-gyel növelt számú diszjunkt részhalmazra bontjuk, mégpedig az inputtér partícionálásával. Maga a fa egy olyan gráf, ahol az egyes csúcspontok (amelyeknek megfelel a megfigyelések egy részhalmaza) utód–szülő kapcsolatban vannak egymással. Minden felmenőnek pontosan két utóda van, egészen a végpontokig, amelyeknek nincs utódjuk, és a végső partíciót reprezentálják.

Induljunk ki a gyökérből, ami a teljes megfigyelés halmaz! Természetes, hogy a célváltozó teljes átlagát tekintjük a gyökérhez tartozó legjobb „predikciónak". Tegyük fel, hogy a teljes inputhalmazt valahogyan két részhalmazra bontjuk, és mindkét részhalmaz átlagát tekintjük a megfelelő részhalmazhoz tartozó „becslésnek". Ez a „becslés" finomabb, mint a legdurvább első becslés volt, vagyis gyakorlatilag biztos, hogy a négyzetes becslési hiba *(residual sum of squares, RSS)* csökken. A két részhalmazra bontás nyilván nagyon sokféleképpen megtörténhet, a CART (egyik) lényeges tulajdonsága az az elv, amelynek alapján elvégezzük ezt a felbontást. A cél az, hogy minél inkább csökkentsük a négyzetes becslési hibát. Egy teljes leszámlálása a lehetőségeknek és az optimális felbontás *(split)* választása csak elvben valósítható meg, és nem is célszerű. A CART a következőképpen jár el: veszi az első magyarázó változót és annak összes lehetséges bináris megbontását (ha a változó rendezett, akkor csak a rendezett felbontásokat). Minden egyes lehetséges felbontásra kiszámolja a négyzetes becslési hiba csökkenését, és kiválasztja azt, amelyik a legnagyobb javulást éri el. Ugyanezt megteszi a második, harmadik, $k$-adik változóval is. Majd azt a változót és azt a felbontást választja, amelyre a legnagyobb a négyzetes becslési hiba csökkenése. Ez a felbontás (vágás) egy három csúcspontú fát eredményez, ahol most két végpont van.

A következő vágásnál már mind a két végpontra el kell végezni az egyes változók összes lehetséges felosztását, de az újabb vágás most is csak egy végpontot érint, vagyis a fa mérete újra kettővel, a végpontok száma pedig 1-gyel nő. A következő lépésben most már a három végpont valamelyikén haladunk tovább a négyzetes becslési hiba legnagyobb csökkenése elvének figyelembevételével. Mivel véges számú adatunk van, a faépítés egyszer meg fog állni, de a gyakorlatban az algoritmusok már akkor is leállnak, amikor a végpontokhoz tartozó megfigyelések száma az összes megfigyelés számához képest nagyon kicsivé válik.[2] Minden végpontnak megfelel az inputtér egy diszjunkt részhalmaza, és ezek uniója kiadja az inputteret. Ez a módszer az inputtér homogenizálásának fogható fel, mivel ugyanahhoz a végponthoz hasonló elemek tartoznak abban az értelemben, hogy a hozzájuk tartozó célváltozóértékek közel lesznek egymáshoz. A lényeg persze az, hogy azonosítjuk az inputtér azon részhalmazait, ahol ez a homogenitás érvényesül.

*A fa metszése*

Az így épített fa egy nem paraméteres becslésnek tekinthető, ahol az inputtér diszjunkt részhalmazaihoz a benne lévő megfigyelések célváltozóinak átlagát rendeljük. Ez a becslés „túlilleszt", vagyis túlságosan pontosan adja vissza az empirikus adatokat, aminek következtében rossz az általánosítóképessége, azaz a mintán kívül pontatlan lesz az előrejelzés. A metszési művelet a nagy és nagyon komplex fát egyre kevésbé komplex alfákra „metszi" vissza, amelyek adott komplexitási feltétel (végpontok száma) mellett optimálisak, vagyis a legkisebb négyzetes becslési hibával rendelkeznek. Belátható, hogy ez a metszési műveletsor ekvivalens azzal, hogy definiálunk egy új célfüggvényt, amely tartalmazza nemcsak a legkisebb négyzetes becslési hibát, hanem egy komplexitási büntetőfüggvényt is, és egy adott komplexitási büntetőparaméter mellett ezt a módosított célfüggvényt minimalizáló részfát választjuk (lásd *Therneau–Atkinson* [2018] 12–13. o.). Így egy részfasorozatot kapunk, amelynek egyik végén áll a maximális fa (ahol a komplexitás büntetőparamétere 0), a másik végén pedig az osztatlan fa (ahol a büntetőparaméter végtelen nagy). Belátható az is, hogy az így alkotott részfasorozat egymásba ágyazott.

*Validáció – a legjobb részfa kiválasztása*

A CART-algoritmus a legjobb részfát (ami ekvivalens az optimális komplexitással) keresztvalidációval határozza meg (*Hastie és szerzőtársai* [2009]). Az általunk használt algoritmusban „10-szeres keresztvalidációt" használtunk. A validáció célja, hogy az egyes modellek (azaz fák) mintán kívüli előrejelzési képességeit hasonlítsuk össze, és azt a részfát válasszuk, amely ezen ismérv alapján a legjobban teljesít. A „10-szeres

[2] Az általunk használt algoritmusban 20 volt az megfigyelésszám, amely alatt további vágásra nem került sor.

kereszt" kifejezés arra utal, hogy a minta 10 darab 1/10-ed részét szisztematikusan félretesszük a tesztadathalmaz számára, és a becslést a maradék 9/10-ekből végezzük el. A gyakorlati alkalmazásoknál gyakran megelégednek „közel" optimális választással, amennyiben nem a predikciós funkció az elsődleges.

*Hogyan használhatjuk a CART-algoritmust?*

Milyen módon használhatunk egy CART-algoritmussal létrehozott, „közel" optimális fát, ha nem kimondottan a predikció a célunk? A számos szóba jöhető lehetőség közül mi a CART-„output" alábbi jellemzőit fogjuk használni.

1. A fastruktúra, azaz az inputtér diszjunkt részhalmazokra bontása. Ez jelezheti számunkra, hogy melyek a fontos sajátos szegmensei a populációnak, és milyen értelemben sajátosak.

2. Az egyes vágási lépéseknél a „szurrogátum" (helyettesítő) változók listája. Minden egyes vágásnál csak egy változó alapján történik felbontás, de az nyilván részben véletlenszerű – mintafüggő –, hogy legjobban pontosan melyik változó teljesít. Fontos információtartalma van annak, hogy ha valamely változó esetleg nem állna rendelkezésre, akkor melyik másik változó tudná majdnem ugyanazt a felbontást produkálni, vagyis lenne jó szurrogátuma (helyettesítője) a kérdéses változónak. (A szurrogátum pontos meghatározását lásd *Therneau–Atkinson* [2018] 18–19. o.) Ez a fajta, változók közötti helyettesíthetőség is hozzájárul a változók fontosságának mértékéhez. A szurrogátum változó pontosíthatja/módosíthatja elképzelésünket a szegmentációt illetően.[3]

3. A változófontossági mérték. Ez egy százalékosan megadott összefoglaló adat arról, hogy melyik magyarázó változó milyen mértékben járul hozzá az adott fa kialakításához. Ha egy változóval valamely csomópontnál vágunk, akkor a négyzetes becslési hiba így elért csökkenését a változó érdemének tulajdonítjuk. Ezekhez az érdemekhez súlyozva hozzáadjuk az adott változó szurrogátumként elért potenciális érdemeit is (a pontos definíciót lásd *Therneau–Atkinson* [2018] 12. o.).

Összefoglalva: a CART lényege számunkra az inputtér felbontása részhalmazokra, ami elkülöníti hasonló egyedek szegmenseit a populációban, de egy regressziós elemzés számára sugallhat változótranszformációkat (beleértve interakciókat is). A változók fontosságáról adott információja összevethető a hagyományos szignifikanciaanalízissel, és kérdéseket vethet fel olyankor, amikor ezek eltérnek egymástól.

## Adatok

Az elemzéshez a Nemzeti Munkaügyi Hivatal által gyűjtött 2015-ös bértarifaadatokat használjuk. Célváltozónk a havi keresetek logaritmusa. A havi kereset az alapilletmény mellett tartalmazza az egyéb kiegészítő juttatásokat is, mint például a műszak- vagy az

[3] Minden csomópontnál maximum öt szurrogátummal számol az algoritmus.

éjszakai pótlékot. A bérekkel kapcsolatban szerettük volna elkerülni a rövidebb munkaidőben foglalkoztatottak kisebb havi bérének torzító hatását. Emiatt a vizsgálataink során csak a teljes munkaidőben foglalkoztatottak bérét vizsgáljuk. Emellett az állami szektorban számos esetben nem piaci alapon, hanem bértábla szerint alakulnak a bérek. Itt most csak a versenyszférában foglalkoztatottak bérét elemezzük.

A munkavállalók egyéni jellemzőinél a következő változókat vesszük figyelembe: NEM, ÉLETKOR, BECSÜLT GYAKORLATI IDŐ, SZOLGÁLATI IDŐ, ISKOLAI VÉGZETTSÉG, ÚJ BELÉPŐ-E.

A vállalati jellemzők: TULAJDONOS „NEMZETI" HOVATARTOZÁSA, TULAJDON MAGÁN- VAGY KÖZTULAJDONJELLEGE, MÉRET, A TELEPHELY KÖZIGAZGATÁSI EGYSÉG SZERINTI ELHELYEZKEDÉSE, TERÜLETI ELHELYEZKEDÉSE, ÁGAZAT, KOLLEKTÍV SZERZŐDÉS MEGLÉTE.

Az egyes változók pontos leírását a *Függelék F1. táblázata* tartalmazza. Az indikátorváltozókat nem tekintettük ordinálisan rendezettnek, holott több közülük kézenfekvően ilyen. Feltételezésünk szerint a rendezettségnek „ki kell jönnie" a becslésből.

## Eredmények

A validációs hibák minimuma 0,39 körül mozog, de ez olyan kis komplexitási paraméternél található, ahol a regressziós fa több ezer végpontból áll. Ha megelégszünk közel optimális megoldással, ahol a validációs hiba mutatójának értéke 0,4, a komplexitási paraméter értéke akkor is 0,0001. Az ennek megfelelő regressziós fa még mindig óriási, részleteiben elemezhetetlen. Ezért elemzési célra két kisebb fát választunk 0,01-es és 0,001-es komplexitásiparaméter-értékekkel. A legkisebb fa vizuálisan is ábrázolható, és végig fogjuk nézni valamennyi végpontját, illetve azt, hogy miként jutunk el a végpontokhoz. A közepes fa egészében már vizuálisan élvezhetetlen, de bizonyos érdekes részeit kiemeljük. Ennek a két fának és a nagy, közel optimális fának a változófontossági adatait közöljük az *1. táblázatban,* ami azt mutatja, hogy az egyre bővülő fák bizonyos aggregált tulajdonságai kvalitatíve stabilak.

### *A legegyszerűbb fa*

A legegyszerűbb, legkisebb fa leírását a *2. táblázat* tartalmazza. Minden egyes szülő két utóddal rendelkezik. Ha a szülő számjele *n,* akkor a két utód számjele $2n$ és $2n + 1$. Például a gyökérnek (1-es csomópont) két utódja a 2-es és a 3-as csomópont. Az első információ, amely a számjelet követi, megmutatja, hogy mely változó mely értékei alapján vágva jutunk ebbe a csomópontba. (Ehhez lásd a *Függelék F2. táblázatát.*) A második információ egy egész szám, amely azt mutatja meg, hogy hány megfigyelés tartozik az adott csomóponthoz. Például a 14-es csomóponthoz 6180. Az utolsó két szám a csomópont négyzetes becslési hibája (RSS) és a keresetek logaritmusának átlaga. A csillaggal jelölt csomópontok végpontok.

*1. táblázat*
Változófontossági mértékek a különböző komplexitási paraméterrel rendelkező fák esetében (százalék)

| Változó | Kis fa | Közepes fa | Nagy fa |
|---|---|---|---|
| Képzettség | 64 | 53 | 46 |
| Vállalatméret | 13 | 12 | 13 |
| Külföldi tulajdon | 4 | 5 | 5 |
| Kor | 2 | 3 | 4 |
| Ágazat | 7 | 8 | 10 |
| Tapasztalat | 3 | 4 | 4 |
| Kollektív szerződés | 3 | 2 | 3 |
| Állami tulajdon | 1 | 1 | 1 |
| Szolgálati idő | 1 | 3 | 4 |
| Nem | 1 | 3 | 3 |
| Régió | 0 | 2 | 3 |
| Új belépő | 0 | 1 | 1 |
| Településforma | 0 | 2 | 2 |

*Forrás:* Bértarifa-felvétel, saját számítás.

A gyökér tartalmazza az osztatlan mintát, ahol 158 461 megfigyelés van. A fa először a legfeljebb közép- (122 185 megfigyelés) és a felsőfokú végzettséggel rendelkezőket (36 276 megfigyelés) különbözteti meg. Érdekes, hogy a második legjobb metszésnél, amit nem hajt végre az algoritmus, a végzettség helyett az ágazat lenne a metszési elv (a legjobb szurrogátum), ekkor 78 százalékban egyezne meg a két fa, és csak az információ, kommunikáció ágazatát különböztetnénk meg a többitől. A legfeljebb érettségivel rendelkezők átlagosan 206 583 forintot keresnek, míg az egyetemi vagy főiskolai diplomával rendelkezők átlagosan 524 901 forintos jövedelemmel rendelkeznek.[4] Így az átlagok közötti különbség több mint 300 000 forint.

A fa a továbbiakban mindkét iskolázottsági csoportnál a létszám alapján osztódik tovább. Az algoritmus a maximum középfokú végzettséggel rendelkezőknél 126,5 fős, a diplomásoknál 64,5 fős vállalati méretnél bontotta tovább a mintát. Mindkét esetben a nagyobb vállalatoknál nagyobb jövedelemre tesznek szert a munkavállalók. Az alacsonyabb végzettség esetén a „kollektív szerződés" kétértékű változója jó szurrogátumnak minősül: 68,3 százalékban ugyanazt a modellt kapnánk, ha ezen változó alapján vágna az algoritmus.

A legfeljebb érettségivel rendelkezőknél a létszám szerinti bontás alapján a kisebb vállalatoknál átlagosan 175 741, míg a nagyobb vállalatoknál átlagosan 238 889 forintot

[4] Magyarázott változónak a keresetek logaritmusát vettük. Ezért a táblázatban a keresetek logaritmusának átlagai szerepelnek. A szövegben viszont a forintátlagértékeket közöljük, amit úgy számoltunk ki, hogy a logaritmusok átlagértékéhez hozzáadtuk a variancia felét, majd ennek vettük az exponenciális értékét, azaz az *e* számra emeljük.

*2. táblázat*
A legegyszerűbb fa

| Csomópont | Vágási pont | Megfigyelések száma | RSS | A keresetek logaritmusainak átlaga |
|---|---|---|---|---|
| 1. | Gyökér | 158 461 | 57883,11 | 12,33 |
| 2. | iskveg9f = 07,8, szak, szakm, szakkoz, gim, tech | 122 185 | 23501,00 | 12,14 |
| 4. | letszam_bv1 < 126,5 | 61 986 | 9538,10 | 12,00 |
| 8. | iskveg9f = 07,8, szak, szakm* | 34 105 | 3298,85 | 11,91 |
| 9. | iskveg9f = szakkoz, gim, tech* | 27 881 | 5593,71 | 12,11 |
| 5. | letszam_bv1 ≥ 126,5 | 60 199 | 11409,42 | 12,29 |
| 10. | iskveg9f = 07,8, szak, szakm* | 32 394 | 4136,57 | 12,15 |
| 11. | iskveg9f = szakkoz, gim, tech* | 27 805 | 5920,19 | 12,45 |
| 3. | iskveg9f = fois, egyet | 36 276 | 16301,40 | 12,95 |
| 6. | letszam_bv1 < 64,5 | 13 439 | 6844,40 | 12,67 |
| 12. | kraf = 0%* | 10 642 | 4631,55 | 12,55 |
| 13. | kraf = 100%, tobbs, kisebbs* | 2 797 | 1439,70 | 13,14 |
| 7. | letszam_bv1 ≥ 64,5 | 22 837 | 7808,18 | 13,11 |
| 14. | kor < 31.5* | 6 180 | 1130,92 | 12,84 |
| 15. | kor ≥ 31,5* | 16 657 | 6072,72 | 13,21 |

*Megjegyzés:* a csillaggal jelölt csomópontok végpontok.
*Forrás:* Bértarifa-felvétel, saját számítás.

keresnek. Ezt követően mind a kis-, mind a nagyvállalatoknál a végzettség szerint vágott az algoritmus. Itt mindkét esetben megkülönböztette az érettségit adó és nem adó oktatási formákat. Az érettségivel nem rendelkezők minden esetben kevesebbet keresnek, mint azok, akik rendelkeznek. A legfeljebb érettségivel rendelkezők esetében a bérkülönbség a legtávolabbi levelek között megközelítően 130 000 forint.

A felsőfokú végzettséggel rendelkezőknél a létszám alapján megkülönböztetett kisvállalatok átlagosan 409 633, a nagyvállalatok átlagosan 585 830 forintot fizetnek. A felsőfokú végzettségnél az átlagos bérkülönbség sokkal nagyobb a kis- és nagyvállalatok között, mint alacsonyabb végzettség esetén. A *2. táblázatban* közölt logaritmusok összehasonlítása azt bizonyítja, hogy a különbség százalékosan is lényegesen nagyobb.[5]

A felsőfokú végzettséggel rendelkező, kisvállalatoknál dolgozókat az algoritmus a továbbiakban a külföldi részesedés alapján különbözteti meg. Az egyik csoport a teljesen magyar tulajdonú vállalatok, a másik a részben vagy egészében külföldi

[5] A logaritmusok különbsége közelítően a százalékos eltérésnek felel meg. A közelítés nagy logaritmikus különbségekre már pontatlan, de a százalékok összehasonlítására alkalmas.

tulajdonban lévő cégek. Az átlagos bérek közötti különbség a két csoport esetében több mint 300 000 forint.

A nagyvállalatoknál azonban nem a tulajdon, hanem az életkor alapján végzi a megkülönböztetést a CART. Itt a vágási pont a 31,5. életév volt. Közel azonos fa adódna, ha az életkor helyett a 10,5 éves munkatapasztalatnál vágna az algoritmus, ami érthető, mivel a munkatapasztalat az életkor és a végzettség alapján számolódik (lásd *Függelék F1. táblázat*). A magasabb életkor majdnem 240 000 forintos bérelőnyt jelent.

A felsőfokú végzettség esetén a fa legtávolabbi végpontjai közötti bérkülönbség több mint 300 000 forint. Vagyis alacsonyabb végzettség esetén kisebb volt a bérkülönbség a két legtávolabbi levél között, mint a felsőfokú végzettség esetén.

Azt gondolhatjuk, hogy a felbontások alapján négy kifejezetten fontos változónk van: képzettség, vállalatméret, külföldi tulajdon és életkor. Tekintsük most újra az *1.* (változófontossági) *táblázatot!*

Az *1. táblázatból* kitűnik, hogy az ágazati hovatartozás fontosabb, mint akár a külföldi tulajdon, akár az életkor, és megjelennek további változók is, amelyekkel eddig nem találkoztunk. A magyarázat abban rejlik, hogy ezek számos helyen jobb szurrogátumok, mint a kor vagy a külföldi tulajdon. Az output részletes elemzése valóban azt mutatja, hogy az ágazati hovatartozás az 1., 2., 3., 4. csomópont felosztásánál is hatékony szurrogátum lehetne. Ugyanis 65–78 százalékban ugyanazt kapnánk, ha az ágazatok szerint bontanánk. A tapasztalat szintén relatíve jó szurrogátum: jó szurrogátuma az életkornak (7. csomópont), de jobb szurrogátum, mint az életkor egyéb csomópontokban. Ez indokolja, hogy összességében a becsült gyakorlati idő fontosabb változó, mint az életkor. A kollektív szerződés az alacsonyabb végzettség esetén a létszámnak jó szurrogátuma, azonban e változó szerint nem történik vágás. Feltehetően alacsonyabb végzettség esetén a kollektív szerződések hatása nagyobb a bérekre.

Látható, hogy még ennek az egyszerű fának az elemzése is érdekes információkat nyújt. Tekintsük most a közepes fát, azaz hogyan bontjuk tovább a kis fát, ha a komplexitási paraméter értékét egy tizedére csökkentjük, vagyis kevésbé törekszünk egyszerű modellt előállítani!

### *Közepes fa*

A közepes fa kis fához hasonló részletes leírását nem közöljük. Itt csak a végpontokhoz tartozó csoportok jellemzőit ismertetjük, majd az egyes változók fontosságát elemezzük. A közepes fa leírását a *Függelék F2. táblázata* tartalmazza.

Az egész fa legkisebb átlagjövedelmű csoportja ahhoz a végponthoz tartozik, amely úgy jellemezhető, hogy az alacsony képzettségű, 127 fősnél kisebb vállalatnál dolgozó nők. A felső ág „királyai" azok a középfokú képzettségű és 127 fős vagy annál nagyobb vállalatnál dolgozó férfiak, akik legalább hatéves szolgálati idővel rendelkeznek, és az B, C, D, F, H, J, K, M, P, Q, S ágazatokban dolgoznak. Az ágazatkódokat a *3. táblázat* tartalmazza.

*3. táblázat*
Ágazatok összehasonlítása*

| Ágazat | Alsóág | Felsőág |
|---|---|---|
| A Mezőgazdaság, erdőgazdálkodás, halászat | L | L |
| B Bányászat, kőfejtés | R | L |
| C Feldolgozóipar | R | L |
| D Villamosenergia-, gáz-, gőzellátás, légkondicionálás | R | R |
| E Vízellátás; szennyvíz gyűjtése, kezelése, hulladékgazdálkodás, szennyeződésmentesítés | L | L |
| F Építőipar | R | L |
| G Kereskedelem, gépjárműjavítás | L | L |
| H Szállítás, raktározás | R | L |
| I Szálláshely-szolgáltatás, vendéglátás | L | L |
| J Információ, kommunikáció | R | L |
| K Pénzügyi, biztosítási tevékenység | R | R |
| L Ingatlanügyletek | L | L |
| M Szakmai, tudományos, műszaki tevékenység | R | L |
| N Adminisztratív és szolgáltatást támogató tevékenység | L | L |
| P Oktatás | – | L |
| Q Humán-egészségügyi, szociális ellátás | R | L |
| R Művészet, szórakoztatás, szabad idő | R | L |
| S Egyéb szolgáltatás | R | L |

* Az R a nagyobb átlagos bér felé irányuló, az L pedig a kisebb átlagos bér felé tartást mutatja.
*Forrás:* Bértarifa-felvétel, saját számítás.

Nézzük meg, kik a felsőfokú végzettségűek „páriái"! Ők a főiskolai végzettségű, 16 főnél kisebb, belföldi tulajdonú, az A, B, C, E, F, G, H, I, J, L, M, N, P, Q, R, S ágazatba tartozó vállalatoknál dolgozók. Összesen 2113 ilyen személy van a mintában. A *3. táblázat* alapján az látszik, hogy az ágazati vágásnál nagy az átfedés. Tehát számos olyan ágazat van, amely relatíve nagyobb béreket kínál a kevésbé képzetteknek, mint a főiskolát végzett és kisvállatoknál dolgozóknak.

Végül pedig tekintsük a magyar munkapiac királyait (és királynőit)! A legmagasabb átlagjövedelmű csoportba tartoznak az egyetemi végzettségű, 38 évnél idősebb, 65 főnél nagyobb vállalatnál a közép-magyarországi régióban dolgozók. Ezek száma 2816 a mintában.

Ha a legnagyobb fát is figyelembe vesszük, akkor a változófontossági táblázat azt mutatja, hogy a további finomítás a sorrenden alig módosít, de újabb változók is képbe kerülnek. Eközben a képzettség megőrzi vitathatatlan vezető helyét, habár a relatív súlya csökken. A legnagyobb fának összesen 355 végpontja van, ami nyilván áttekinthetetlen, de amennyiben az aggregált változófontosságot tekintjük, akkor kvalitatívan nincs nagy változás az előzőkhöz képest.

*Mit mondanak a fák?*

A kis, közepes és nagy fák áttekintése után ebben az alfejezetben sorra vesszük, hogy a bevezetés első bekezdésében említett problémakörökről milyen következtetéseket vonhatunk le a fák alapján.

KÉPZETTSÉG • A képzettség minden kétséget kizáróan a legfontosabb „magyarázó" változó. Ez nagy vonalakban alátámasztja azt, ami az irodalom áttekintéséből is kiderült. Az első vágás a képzettség szerint történik, és ez önmagában 32 százalékkal növeli az $R^2$-et. A következő vágás már csak mintegy 4,4 százalékos, a többi pedig természetesen egyre csökkenő mértékű növekedést okoz az $R^2$-ben. A kis fában még két képzettség szerinti vágás van, ahol elkülönülnek az alsó- és középfokú végzettségűek.

A közepes fában két új vágást találunk. A fentiekben már volt arról szó, hogy a főiskolai és egyetemi végzettségűek elkülönülnek a legmagasabb munkajövedelműek között, azonban a felső ágon is viszonylag gyorsan elválik a legfeljebb általános iskolai végzettség a szakmunkás végzettségtől (32. csomópont). Ez a megkülönböztetés még többször előfordul a fa mélyebb részein is. A középfokú végzettség további bontása is néha megjelenik a fa mélyebb részein, de kevésbé gyakran, mint az alacsony képzettség bontása. Ugyanez igaz az egyetemi és főiskolai diploma szerinti bontásra.

Összefoglalva a CART-elemzés azt sugallja, hogy az iskolai végzettség kilencfokozatú felosztását durvíthatjuk: képzetlen, szakmunkás, középfokú, főiskolai és egyetemi végzettségi szintekre. Ez némiképpen eltér a szokásos ötszintű (0–7 osztály–általános iskolai 8 osztály–szakiskola, szakmunkás–érettségi–diploma) csoportosítástól. Ebben az új felosztásban nem különböztetjük meg a nyolc osztályt végzetteket az általános iskolát nem befejezőktől, viszont a diplomásokat szétbontjuk főiskolát és egyetemet végzettekre.

VÁLLALATMÉRET • Láttuk, hogy létszám (vállalatméret) tekintetében a felső ágon a 126., az alsó ágon a 62. szintnél van a nagy- és a kisvállalatok legdurvább elkülönítése. A közepes fában két újabb vágás jelenik meg, az egyik a 16., a másik a 17. szinten, mindkét esetben a felső ágon. Ez arra utal, hogy a képzettség és a méret közötti „interakció" is releváns jellemzője a bérezésnek.

A nagy fában van egy furcsa jelenség: egy bizonyos csoportban a 4150 főt meghaladó vállalatoknál dolgozók bére kisebb, mint a több száz főt foglalkoztató vállalatok átlagbére, és a csoporthoz 1642 megfigyelés tartozik. Ez utalhat arra, hogy a vállalatnagyság és a bérek közti összefüggés nem monoton. Egy ehhez hasonló furcsa „nagyon nagy" vállalati hatást találunk a felső ágon is, de ott viszonylag kevés elemű a részmintánk. Ezek a megfigyelések arra utalhatnak, hogy itt bizonyos „kilógó" *(outlier)* vállatokról van szó, amelyek speciális kezelést igényelnének egy regresszióban.

A vizsgálataink alapján azt mondhatjuk, hogy vélhetően a vállalatméret hatása nem független a képzettségtől. Az azonban nem egyértelmű, hogy mennyi nemlinearitás és/vagy nemfolytonosság van a vállalatméret és a bérek közötti összefüggésben.

Vagyis csak egy alapos elemzés mondhatja meg, hogy mennyire indokolt bizonyos vállalatokat kivenni a mintából mint kilógókat.[6]

Ágazat • Az első ágazatok szerinti vágás a 18. csomópontnál történik a közepes fában, a D (Villamosenergia-, gáz-, gőzellátás, légkondicionálás), a J (Információ, kommunikáció) és a K (Pénzügyi, biztosítási tevékenység) ágazatok kerülnek egy csoportba, ahol jobbra, azaz a magasabb bér irányba történik a megbontás. A 12. csomópontnál is van egy olyan felbontás, ahol a D és a K ágazatok jobbra tartanak a többiekhez képest. A közepes fában található többi felbontás eléggé egyenlően oszlik el az ágazatok között. Az ágazatokat osztályozhatjuk a *3. táblázat* alapján, ha az azonos mintájú ágazatokat – (L, L), (R, L), (R, R) – azonos csoportba tartozónak tekintjük. Érdekes, hogy a negyedik lehetséges minta (L, R) egyik ágazatra sem jellemző. Vagyis létezik két ágazat, a D és a K, ahol az alacsonyabb és a magasabb képzettségűeknek is nagyobb bér „jut". Van hat ágazat, ahol mindkét képzettségi szinten inkább alacsonyabbak a bérek (L, L), és a többi ágazatban a magasabb képzettségen magasabbak, alacsony képzettség mellett pedig alacsonyabbak (R, L). Vagyis „egalitariánus" (L, R) ágazat nem létezik.

Földrajz, településjelleg • A középső fában öt, régiók szerinti felbontás van, mindegyik esetben a Közép-Magyarország régió jobbra, azaz a magasabb bér felé tart. A nagy fában már komplikáltabb a helyzet. Itt előfordul, hogy ez a régió valamely csúcspontnál balra halad, de azért a leggyakrabban itt is igaz marad az, hogy jobbra tart. A Nyugat- és a Közép-Dunántúl régiók gyakran egy csoportba kerülnek Közép-Magyarországgal, de nem minden esetben, és van úgy, hogy ezek jobbra, miközben Közép-Magyarország balra tart.

A modellben van egy másik településföldrajzi változó is, amely „Budapest", „város" és „falu" értékekkel rendelkezik. Ennek a változónak általános fontossága kicsi, és a vágásokat figyelve nem igazán látható benne konzisztencia. Van, amikor „Budapest" és „város" kerülnek egy csoportba, de olyan is akad, amikor „Budapest" és „falu", és olyan is, amikor „város" és „falu". Ha esetleg azt várnánk, hogy Budapest mindig jobbra tart, akkor is csalódnunk kellene.

Tehát levonhatjuk azt a következtetést, hogy inkább érdemes a KSH-régiókat használni, mint a hármas településföldrajzi felosztást. Budapest, esetleg Nyugat- és Közép-Dunántúl esetében érdemes interakciókat is keresni. Nem kizárt az sem, hogy a „Budapest és közép-magyarországi falu" elkülönítés releváns lehet az egyszerű Közép-Magyarország kategóriával szemben, amelyikbe Budapest is beletartozik.

Kor, tapasztalat, új belépés • Demográfiai változó több is van. Tekintsük először az életkort! A közepes fában csak három darab „kor" szerinti vágás van, mindhárom az alsó ágon. Mindhárom esetben 30 és 40 közötti a vágás, és a várakozásnak

[6] Az *outlier* kifejezést úgy értelmezzük, mint olyan megfigyeléseket, amelyek nem ugyanazt az összefüggést elégítik ki, mint a minta többi része. Ez a megkülönböztetés természetesen nagymértékben „feltevés". A kérdés az, hogy a feltevéseink csak a konkrét adatoktól független – *a priori* – feltevések lehetnek, vagy megengedjük azt is, hogy az adatok által sugallt feltevések legyenek.

megfelelően nagyobb életkornál jobbra tartunk. A nagy fában több „kor" szerinti vágás is van. A felső ágon ezek 25 és 30 éves kor közötti vágások, míg az alsó ágon 25 és 51 éves közöttiek. Ez arra utalhat, hogy 30 év fölött a nem felsőfokú képzettségnél nem nagyon számít az életkor. Felsőfokú végzettségnél körülbelül 31 év a nagy választóvonal, de utána is vannak finom különbségek a bérezésnél az életkor függvényében. Tehát egy hagyományos regresszióban megfontolandó lenne a „végzettség" és a „kor" interakciójának a vizsgálata.

A „tapasztalat" szerinti vágás fontosabbnak látszik összességében, mint az „életkor" szerinti, de ez elsősorban a szurrogátumszerepnek tulajdonítható. A „tapasztalat" definíciója alapján az „életkor" és a „tapasztalat" majdnem lineáris kapcsolatban vannak egymással.[7] Az eredmények azonban azt mutatják, hogy ha csekély is, de van különbség a két változó között. A nagy fában előfordul olyan vágás, amikor a növekvő tapasztalat egy kis szakaszon csökkenő bérrel jár, ami érdekes, és megint utalhat *outlier* jellegre is vagy a bérek és a tapasztalat között fennálló konkáv kapcsolatra, ahogyan *Gábor* [2008] kimutatta. A „tapasztalat" szerinti vágások főként az alsó ágon vannak, és különösképpen 8 és 16 év között, illetve 35 év után.

A „szolgálati idő" a vágási pontokban nagyságrendileg 3–5 évnek felel meg. A szolgálati időt csak a maximum középfokú végzettséggel rendelkezőknél használja az algoritmus. A szolgálati idő együtt mozog a korral, a tapasztalattal, de az ezekkel való helyettesítés információveszteséget okoz. Amennyiben a szolgálati idő szerint megkülönböztetjük az egy évnél rövidebb tapasztalattal rendelkezőket, vagyis az új belépőket, akkor e szerint nincsen egyetlen vágás sem. Mindez arra utalhat, hogy a szolgálati időnél is lényeges lehet a képzettséggel való interakció, illetve az, hogy az „új belépés" elhanyagolható változó, ha rendelkezünk „szolgálati idő" információval.

TULAJDONFORMA • Először tekintsük az állami tulajdon hatását! Az állami tulajdoni arány lényegtelen változónak tűnik. Csak a nagy fában van néhány, ennek megfelelő vágás, de a vágás iránya nem szisztematikus.

Ezzel szemben a külföldi tulajdon aránya viszonylag fontos változó, és a vágásoknál általában (de nem mindig) a nagyobb külföldi tulajdoni arány azt jelenti, hogy jobbra, vagyis magasabb bérek felé tartunk. Külföldi tulajdon szerinti vágások vannak mind a felső, mind pedig az alsó ágon.

NEM • A kis fában nincs nemek szerinti vágás, viszont a közepes fában a felső ágon négy és az alsó ágon egy elágazás van. Úgy tűnik, mintha a nemek közti bérkülönbség az alacsonyabb végzettség esetén lenne fontosabb. Ezt igazolja, hogy az alsó ágon csak egy olyan alágon van vágás a nemek szerint, ahol már csak főiskolai végzettségűek maradtak. Ez alapján alacsonyabb végzettség esetén karakterisztikusabbnak látszik a nemek közti jövedelemkülönbség. A nagy fában természetesen már számos vágás van nemek szerint, de ezek mindig a főiskolai végzettségűeknél jelentkeznek. A szakirodalommal összhangban olyan esetet nem találunk, ahol a vágás után a nők

[7] A „tapasztalat" változó meghatározása: az életkorból levonják az iskolaérettséget jelző hat évet és az iskolában töltött évek számát.

átlagos keresete magasabb, mint a férfiaké. A bérek nemek szerinti eltérése tehát továbbra is létező, de mintha azt sugallná a CART-elemzés, hogy alacsony képzettségnél nagyobb az eltérés, mint magasnál, ami első látásra ellentmondani látszik az üvegplafon-jelenség meglétének 2015-ben (*Lovász* [2013]). Amennyiben regressziós modellben gondolkodunk, akkor érdemes ezek alapján kereszthatásokkal dolgozni és a nemeket legalábbis a képzettséggel interakcióba hozni.

Szerződések • A „kollektív szerződés léte" változója nem okoz vágást a kis és közepes fában. A nagy fában először a 81. csomópontban – és összesen négyszer – fordul elő vágás a kollektív szerződés megléte alapján, mindig a felső ágon. Mindezen esetekben a kollektív szerződés léte határozottan pozitív hatású a bérekre. Azt gondolhatjuk ennek alapján, hogy itt is lehet interakcióval kísérletezni, mert úgy tűnik, hogy a kollektív szerződés elsősorban a közepes vagy alacsony végzettségűek bérét befolyásolja.

Összefoglalva a *képzettség* az az egyéni jellemző, amely a leginkább befolyásolja a kereseteket. A magasabb iskolai végzettséggel rendelkezők magasabb bérre tesznek szert, ami egybevág *Gábor* [2008] eredményeivel. A második legfontosabb változó a *vállalatméret,* amely leginkább a felsőfokú végzettséggel nem rendelkezők bérére hat. Az *ágazat* is jelentős mértékben befolyásolja a béreket. Itt négy csoportot lehetett meghatározni aszerint, hogy az alacsonyabb iskolai végzettségű legjobban keresők és a magasabb végzettségű legrosszabbul keresők átlagos bérei adott iparágban jobbra vagy balra tartanak. A *földrajzi elhelyezkedést* figyelembe véve a Közép-Magyarország régió mindig jobbra tart. Ehhez a régióhoz néha csatlakozik Nyugat- és Közép-Dunántúl. Ez megfelel *Szabó* [2006] eredményének, amely szerint ebben a három régióban jelentős a bérelőny. A *településtípusoknál* pedig nem volt egyértelmű irány sem Budapest, sem a városok, sem a falvak tekintetében. Ez ellentmond annak, amit *Köllő* [2003] talált. Itt azonban fontos megjegyezni, hogy *Köllő* [2003] a költségvetési szférát vizsgálta, míg mi csak a versenyszférával foglalkoztunk. Így jelenleg csak annyit tudunk mondani, hogy a versenyszférában látottak nem vágnak egybe az állami szektorban látottakkal. Az *életkor* és a *becsült gyakorlati idő,* vagyis a *tapasztalat* nagymértékben együtt mozog, ami a tapasztalat definíciójára vezethető vissza. A két változó közül a tapasztalat a fontosabb, amit a szurrogátumjelleg magyaráz. Általában a *külföldi tulajdonú* vállalatok nagyobb bért fizetnek a munkavállalóknak. Ez teljes mértékben megegyezik azzal, amit *Earle–Telegdy* [2012*a*] és [2012*b*] talált. Alacsonyabb képzettség esetén a *nem* fontosabb változó, mint egyetemi vagy főiskolai végzettségnél. A nők minden pontban kevesebbet keresnek, mint a férfiak. Vagyis lehet, hogy csökkentek a nemi különbségek – mint ahogyan *Lovász* [2008] kimutatta –, azonban továbbra is fennállnak.

Eddig bemutattuk, hogy az egyéni és a vállalati jellemzők hogyan befolyásolják a béreket, valamint rámutattunk arra, hogy milyen esetleges interakciók létezhetnek. Most a CART-elemzés alapján kapott változófontossági mértékek robusztusságát teszteljük.

## Robusztusságvizsgálat

A robusztusságvizsgálatához véletlen erdőt használunk. Először áttekintést adunk a véletlen erdőkről, majd pedig összevetjük a CART- és a véletlenerdő-algoritmusok átlagolásából származó változófontossági mértékeket.

A véletlen erdő előre meghatározott számú fa átlagaként áll elő (a fák száma az algoritmus egy paramétere). A fák készítésénél az algoritmus először *bootstrap* eljárással készít egy mintát. Erre a mintára aztán egy regressziós fa épül. A fa építése a fent bemutatottól annyiban tér el, hogy minden csomópontnál az algoritmus csak a változók egy véletlen részhalmazából választ változót a vágáshoz. (Az algoritmus egyik paramétere az, hogy hány elemű legyen ez a véletlen részhalmaz.) Az erdő által kínált predikció a sok fa átlaga.

A véletlen erdő célja a minimális variancia elérése. Ezt úgy éri el, hogy egyenként zajos, de közel torzítatlan és viszonylag független fákat átlagol. A torzítatlanságot az biztosítja, hogy az átlag várható értéke feltehetően megegyezik az egyes fákra számolt várható értékkel, mivel a fák ugyanabból az eloszlásból származnak. A fák közötti alacsony korrelációt a változók véletlen kiválasztása okozza. Minél kevesebb a kiválasztott változók száma, annál inkább biztosított az előállított fák közötti korrelálatlanság (*Hastie és szerzőtársai* [2009]).

Saját számításainkhoz az R „randomForest" programcsomagját használjuk. Készítettünk egy 40 000 elemből álló tanuló mintát, és erre a mintára illesztettünk 50 darab véletlen erdőt. Mivel a célunk nem a modell mintán kívüli előrejelző képességének javítása volt, így csak a tanuló mintán számolt véletlen erdők eredményeit hasonlítjuk össze és átlagoljuk. Amiatt tartjuk szükségesnek több véletlen erdő eredményének összevetését, mert az erdő kialakításánál a véletlen nagy szerepet játszik mind a minta, mind a változók kiválasztásában. Itt azonban – nem úgy, mint a CART esetében – többszöri futtatás során nem mindig kapjuk ugyanazt az erdőt vissza. Ha több erdőt vizsgálunk, akkor képet kapunk arról, hogy a változófontossági mértékek mennyire stabilak.

A számítási kapacitást és az algoritmus által elkövetett hiba nagyságát figyelembe véve, a kiválasztott változók számát ötben határoztuk meg, és egy erdő 200 fát tartalmaz. Feltételezésünk szerint az öt darab változó nagy függetlenséget biztosít a fák között. A fák számának növelésével a hiba egyre csökken, és konvergál a minimális hibanagysághoz. Egy 200 elemű véletlen erdőnél a hiba már elég alacsony, és még gyorsan le is fut az algoritmus.

A változófontossági mértékeket a két módszer esetében a *4. táblázat* mutatja. A véletlenerdő-algoritmusnál a változófontossági mérték analóg a CART-algoritmusban használt mértékkel, amennyiben az adott változó négyzetes becslési hibát (RSS) csökkentő érdemeinek relatív (100-ra normalizált) értékével számolunk. A különbség az, hogy itt az érdemeket átlagoljuk az összes fára, és az érdemek meghatározásánál nem vesszük figyelembe a szurrogátumokat (lásd *Ishwaran* [2007]).[8] A CART az RSS-csökkenés mellett az adott változó szurrogátumjellegét is figyelembe

[8] Mivel a véletlenerdő-algoritmus nem számol szurrogátumokat.

veszi. Vagyis ha egy változó sok esetben jó szurrogátum, akkor a CART esetében nagyobb a változó fontossága, mint a véletlen erdő esetében.

*4. táblázat*
A véletlen erdőkből és a CART-ból számolt változófontossági mértékek, illetve különbségük

| Változó | CART legnagyobb fa | Véletlen erdő | Különbség |
|---|---|---|---|
| Képzettség | 46,12 | 37,41 | 8,71 |
| Vállalatméret | 13,09 | 16,76 | –3,67 |
| Szolgálati idő | 3,59 | 9,82 | –6,23 |
| Életkor | 3,55 | 6,11 | –2,56 |
| Külföldi tulajdon | 5,27 | 6,07 | –0,80 |
| Ágazat | 10,49 | 5,81 | 4,68 |
| Tapasztalat | 4,40 | 5,41 | –1,01 |
| Régió | 3,09 | 4,64 | –1,56 |
| Nem | 3,42 | 2,95 | 0,47 |
| Településforma | 2,22 | 2,12 | 0,11 |
| Kollektív szerződés | 2,53 | 1,60 | 0,94 |
| Állami tulajdon | 1,46 | 0,88 | 0,58 |
| Új belépő | 0,64 | 0,42 | 0,22 |

*Forrás:* Bértarifa-felvétel, saját számítás.

Az 50 darab véletlenerdő-szimulációból számolt átlagos változófontossági mértékeket és a hozzájuk kapcsolódó leíró statisztikákat a *Függelék F3. táblázata* tartalmazza. A szimulációs eredmények alig szóródnak, ami arra utal, hogy a változófontossági mértékek viszonylag stabilak. Így a változófontossági mérték tekintetében hiába különböznek az egyes erdők, az átlaguk mégis jól jellemzi a változók fontosságát. A CART-algoritmusnak a legnagyobb fához tartozó változófontossági mértékeit közöljük. Az utolsó oszlop a véletlen erdők eredményének CART-tól való eltérését mutatja az egyes változók esetében.

Összességében megállapítható, hogy a két módszer esetében, bár a változófontossági mértékek eltérnek, a változók egymáshoz viszonyított sorrendje nagyjából azonos. Az eredmények összevetésénél azt találjuk, hogy a legfontosabb magyarázó változók továbbra is a képzettség, a vállalatméret, a szolgálati idő, a kor, a külföldi tulajdon, az ágazat és a tapasztalat. Ezeknek a változóknak az összsúlya több mint 85 százalék mindkét esetben.

## Összefoglalás

Milyen eredményeket könyvelhetünk el ebből a megszokottól eltérő statisztikai elemzési megközelítésből? Kaptunk egy, az előrejelzés szempontjából releváns változófontossági sorrendet, ami várakozásainknak nem teljesen felel meg. Vannak

majdnem kollineáris változóink, ilyen az életkor és a tapasztalat. A regressziókban nem lehetne mindegyikük számára szimultán paramétereket becsülni. A CART ezek között mégis talál különbséget, és úgy tűnik, hogy összességében a tapasztalat az, ami – ha választani kell a kettő között – relevánsabb, és amelynek ráadásul érezhető nemlineáris hatása is van.

A létszámmal mért vállalati méret olyan numerikus változó, amelynél bonyolult nemlinearitásokat tapasztaltunk. Meg lehetne vizsgálni, hogy a nagyon nagy vállalatnál dolgozók bérével kapcsolatban tapasztaltak mennyiben jelentnek nemlinearitást, és mennyiben kilógó értékek. A CART-megközelítés egyik potenciális előnye például egy regressziós modellel szemben az, hogy bár a regresszióban is a kilógó értékek figyelmeztethetnek arra, hogy valamilyen nemlinearitással van dolgunk, de nem szolgáltatnak arról információt, hogy a változó tér mely szegmensében kell keresnünk a problémát.

Mint már említettük, a kereszthatások vizsgálatánál találtunk érdemleges összefüggéseket, amelyek érdekes további kutatásoknak adhatnak lökést. Például a képzettség–nem és a képzettség–kollektív szerződés kétértékű változók relevanciája olyan hipotéziseket vet fel, amelyeket más módszerekkel, más vagy éppen újfajta adatokon lehet vizsgálni.

Ez a kutatás nyilvánvalóan nem oldott meg problémákat, legfeljebb felvetett. Kézenfekvő továbbhaladási irány több év adatainak vizsgálata. Hasonló kvalitatív tényeket látunk az elmúlt évtizedek adataiban, vagy esetleg van valamilyen kivehető változási irány? A különböző évekre becsült modellek használhatók-e más évek adatainak leírására? A bevezetésben áttekintett empirikus tanulmányok egy része időbeli változásokat vizsgált (*Gábor* [2008], *Galasi* [2008], *Lovász* [2008], *Köllő* [2003]), tehát az azokban felvetett problémákhoz csak több év adatainak feldolgozásával tudnánk hozzászólni.

Egy másik fontos irány lehet a változótér felbontásának részletes tanulmányozása. Tudunk-e olyan időben stabil változókonfigurációkat találni, amelyek lehetőséget adnak arra, hogy szegmentáljunk, és például aszerint különböztessünk meg munkapiaci szegmenseket, hogy bennük a bérek alakulása mennyire jól vagy kevésbé jól előrejelezhető.

*Hivatkozások*

Breiman, L.–Friedman, J.–Stone, C. J.–Olshen, R. A. [1984]: Classification and regression trees. Chapman and Hall/CRC, London–New York.

Choi, W.–Paulson, S. E.–Casmassi, J.–Winer, A. M. [2013]: Evaluating meteorological comparability in air quality studies: Classification and regression trees for primary pollutants in California's South Coast Air Basin. Atmospheric Environment, Vol. 64. 150–159. o. https://doi.org/10.1016/j.atmosenv.2012.09.049.

De'ath, G.–Fabricius, K. E. [2000]: Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology, Vol. 81. No. 11. 3178–3192. o. https://doi.org/10.2307/177409.

Durlauf, S. N.–Johnson, P. A. [1995]: Multiple regimes and cross-country growth behaviour. Journal of Applied Econometrics, Vol. 10. No. 4. 365–384. o. https://doi.org/10.1002/jae.3950100404.

Earle, J. S.–Telegdy Álmos [2012*a*]: A külföldi beruházások hatásai a munkavállalók béreire. Megjelent: *Fazekas Károly–Benczúr Péter–Telegdy Álmos* (szerk.): Munkaerőpiaci tükör 2012. MTA Közgazdaságtudományi Intézet–Országos Foglalkoztatási Közalapítvány, Budapest, 215–230. o.

Earle, J. S.–Telegdy Álmos [2012*b*]: Privatizáció, foglalkoztatás és bérek. Megjelent: *Fazekas Károly–Benczúr Péter–Telegdy Álmos* (szerk.): Munkaerőpiaci tükör 2012. MTA Közgazdaságtudományi Intézet–Országos Foglalkoztatási Közalapítvány, Budapest, 231–274. o.

Gábor R. István [2008]: A hiányzó láncszem? Életpálya-keresetek és keresetingadozás. Közgazdasági Szemle, 55. évf. 12. sz. 1057–1074. o.

Galasi Péter [2008]: A felsőfokú végzettségű munkavállalók munkaerő-piaci helyzete és foglalkozásuk-iskolai végzettségük illeszkedése. Budapesti Munkagazdaságtani Füzetek, 3. sz. http://www.econ.core.hu/file/download/BWP/BWP0803.pdf.

Galletta, S. [2016]: On the determinants of happiness: A classification and regression tree (CART) approach. Applied Economics Letters, Vol. 23. No. 2. 121–125. o. https://doi.org/10.1080/13504851.2015.1054066.

Hastie, T.–Tibshirani, R.–Friedman, J. [2009]: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York. https://doi.org/10.1007/978-0-387-21606-5.

Ishwaran, H. [2007]: Variable importance in binary regression trees and forests. Electronic Journal of Statistics, Vol. 1. 519–537. o. https://doi.org/10.1214/07-ejs039.

Kertesi Gábor–Köllő János [2003]: Ágazati bérkülönbségek Magyarországon, II. rész. Járadékokon való osztozkodás koncentrált ágazatokban, szakszervezeti aktivitás jelenlétében. Közgazdasági Szemle, 50. évf. 12. sz. 1049–1074. o.

King, M. W.–Resick, P. A. [2014]: Data mining in psychological treatment research: A primer on classification and regression trees. Journal of Consulting and Clinical Psychology, Vol. 82. No. 5. 895–905. o. https://doi.org/10.1037/a0035886.

Köllő János [2003]: Regionális kereseti és bérköltségkülönbségek. Megjelent: *Fazekas Károly* (szerk.): Munkaerőpiaci tükör 2003. MTA Közgazdaságtudományi Intézet–Országos Foglalkoztatási Közalapítvány, Budapest, 65–78. o.

Lemon, S. C.–Roy, J.–Clark, M. A.–Friedmann, P. D.–Rakowski, W. [2003]: Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. Annals of Behavioral Medicine, Vol. 26. No. 3. 172–181. o. https://doi.org/10.1207/s15324796abm2603_02.

Loh, W. Y. [2014]: Fifty years of classification and regression trees. International Statistical Review, Vol. 82. No. 3. 329–348. o. https://doi.org/10.1111/insr.12016.

Lovász Anna [2008]: Competition and the gender wage gap: New evidence from linked employer-employee data in Hungary, 1986–2003. BWP, 4. sz. http://www.econ.core.hu/file/download/BWP/BWP0804.pdf.

Lovász Anna [2013]: Jobbak a nők esélyei a közszférában? A nők és a férfiak bérei közötti különbség és a foglalkozási szegregáció vizsgálata a köz- és magánszférában. Közgazdasági Szemle, 60. évf. 7–8. sz. 814–836. o.

Minier, J. A. [2003]: Are small stock markets different? Journal of Monetary Economics, Vol. 50. No. 7. 1593–1602. o. https://doi.org/10.2139/ssrn.250473.

Razi, M. A.–Athappilly, K. [2005]: A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. Expert Systems with Applications, Vol. 29. No. 1. 65–74. o. https://doi.org/10.1016/j.eswa.2005.01.006.

Rigó Mariann [2012]: A szakszervezeti bérrés becslése Magyarországon. Megjelent: *Fazekas Károly–Benczúr Péter–Telegdy Álmos* (szerk.): Munkaerőpiaci tükör 2012. MTA Közgazdaságtudományi Intézet–Országos Foglalkoztatási Közalapítvány, Budapest, 200–214. o.

Schiltz, F.–Masci, C.–Agasisti, T.–Horn, D. [2017]: Using Machine Learning to Model Interaction Effects in Education: A Graphical Approach. BWP, 4. sz. http://www.econ.core.hu/file/download/bwp/bwp1704.pdf.

Szabó Péter András [2006]: Regionális kereseti és bérkülönbségek. Megjelent: *Fazekas Károly–Kézdi Gábor* (szerk.): Munkaerőpiaci tükör 2006. MTA Közgazdaságtudományi Intézet–Országos Foglalkoztatási Közalapítvány, Budapest, 70–79. o.

Therneau, T. M.–Atkinson, E. J. [2018]: An Introduction to Recursive Partitioning Using the RPART Routines. Mayo Foundation, https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

Varian, H. R. [2014]: Big data: New tricks for econometrics. The Journal of Economic Perspectives, Vol. 28. No. 2. 3–27. o. https://doi.org/10.1257/jep.28.2.3.

Wu, X.–Kumar, V.–Quinlan, J. R.–Ghosh, J.–Yang, Q.–Motoda, H.–Zhou, Z. H. [2008]: Top 10 algorithms in data mining. Knowledge and Information Systems, Vol. 14. No. 1. 1–37. o. https://doi.org/10.1007/s10115-007-0114-2.

## Függelék

*F1. táblázat*
A felhasznált változók listája

| Változó | A változó tartalma |
|---|---|
| Nem (nem) | 0 = nő, 1 = férfi |
| Életkor (kor) | |
| Becsült gyakorlati idő (exp) | Az életkor mínusz az iskolaérettséget jelző hat év és az iskolában töltött évek száma |
| Szolgálati idő (szolgho) | Adott vállalatnál eltöltött idő hónapban |
| Iskolai végzettség (iskveg9) | 0–7 osztály<br>Általános iskola 8 osztály<br>Szakiskola<br>Szakmunkásképző iskola<br>Szakközépiskola<br>Gimnázium<br>Technikum<br>Főiskola<br>Egyetem |
| Új belépő-e (ujbel) | 0 = nem, 1 = igen |

*F1. táblázat folytatása*

| Változó | A változó tartalma |
|---|---|
| Tulajdonos „nemzeti" hovatartozása (KRA) | 100% külföldi<br>Többségi külföldi<br>Kisebbségi külföldi<br>0% külföldi |
| A tulajdon magán- vagy köztulajdonjellege (ARA) | 100% állami-önkormányzati<br>Többségi állami-önkormányzati<br>Kisebbségi állami-önkormányzati<br>0% állami-önkormányzati |
| Méret (LETSZAM_BV1) | Vállalati létszám |
| Telephely közigazgatási egység szerinti elhelyezkedése (TTIP) | Budapest<br>Város<br>Egyéb |
| Területi elhelyezkedés (KSHREG) | Közép-Magyarország<br>Közép-Dunántúl<br>Dél-Alföld<br>Nyugat-Dunántúl<br>Dél-Dunántúl<br>Észak-Magyarország<br>Észak-Alföld |
| Ágazat (AG1) | A Mezőgazdaság, erdőgazdálkodás, halászat<br>B Bányászat, kőfejtés<br>C Feldolgozóipar<br>D Villamosenergia-, gáz-, gőzellátás…<br>E Vízellátás, szennyvíz gyűjtése, kezelése<br>F Építőipar<br>G Kereskedelem, gépjárműjavítás…<br>H Szállítás, raktározás<br>I Szálláshely-szolgáltatás, vendéglátás<br>J Információ, kommunikáció<br>K Pénzügyi, biztosítási tevékenység<br>L Ingatlanügyletek<br>M Szakmai, tudományos és műszaki tevékenység<br>N Adminisztratív és szolgáltatást támogató tevékenység<br>P Oktatás<br>Q Humán-egészségügyi, szociális ellátás<br>R Művészet, szórakoztatás, szabad idő<br>S Egyéb szolgáltatás |
| Kollektív szerződés megléte (KOL) | Adott évre van-e érvényes kollektív szerződése a munkavállalónak? |

*Forrás:* Bértarifa-felvétel.

*F2. táblázat*
A közepes fa

| Csomópont | Vágási pont | Megfigyelések száma | RSS | A bérek logaritmusának átlaga |
|---|---|---|---|---|
| 1. | gyökér | 158 461 | 57 883,11 | 12,33 |
| 2. | iskveg9f = 07,8, szak, szakm, szakkoz, gim, tech | 122 185 | 23 501,00 | 12,14 |
| 4. | letszam_bv1 < 126,5 | 61 986 | 9 538,10 | 12,00 |
| 8. | iskveg9f = 07,8, szak, szakm | 34 105 | 3 298,85 | 11,91 |
| 16. | nemf = no* | 7 841 | 385,90 | 11,79 |
| 17. | nemf = ferfi | 26 264 | 2 782,02 | 11,94 |
| 34. | kraf = 0% | 23 893 | 2 253,98 | 11,92 |
| 68. | szolgho < 51,5* | 13 712 | 1 084,01 | 11,87 |
| 69. | szolgho ≥ 51,5* | 10 181 | 1 086,26 | 11,99 |
| 35. | kraf = 100%, tobbs, kisebbs* | 2 371 | 399,47 | 12,16 |
| 9. | iskveg9f = szakkoz, gim, tech | 27 881 | 5 593,71 | 12,11 |
| 18. | kraf = 0% | 24 526 | 4 208,32 | 12,07 |
| 36. | ag1f = 1, 2, 3, 5, 6, 7, 8, 9, 12, 13, 14, 16, 17, 18, 19 | 22 205 | 3 485,45 | 12,04 |
| 72. | szolgho < 48,5* | 11 478 | 1 486,84 | 11,97 |
| 73. | szolgho ≥ 48.5* | 10727 | 1 890,28 | 12,11 |
| 37. | ag1f = 4, 10, 11* | 2 321 | 572,99 | 12,31 |
| 19. | kraf = 100%, tobbs, kisebbs | 3 355 | 934,85 | 12,46 |
| 38. | szolgho < 54,5* | 1 817 | 399,41 | 12,32 |
| 39. | szolgho ≥ 54,5* | 1 538 | 462,60 | 12,62 |
| 5. | letszam_bv1 ≥ 126,5 | 60 199 | 11 409,42 | 12,29 |
| 10. | iskveg9f = 07,8, szak, szakm | 32 394 | 4 136,57 | 12,15 |
| 20. | nemf = no* | 11 667 | 817,68 | 11,97 |
| 21. | nemf = ferfi | 20 727 | 2 741,20 | 12,25 |
| 42. | szolgho < 34,5* | 7 218 | 732,68 | 12,09 |
| 43. | szolgho ≥ 34,5 | 13 509 | 1 736,46 | 12,33 |
| 86. | ag1f = 1, 7, 9, 11, 12, 14, 17, 19* | 2 169 | 238,32 | 12,11 |
| 87. | ag1f = 2, 3, 4, 5, 6, 8, 10, 13, 18* | 11 340 | 1 369,72 | 12,38 |
| 11. | iskveg9f = szakkoz, gim, tech | 27 805 | 5 920,19 | 12,45 |
| 22. | ag1f = 1, 5, 7, 9, 12, 14 | 7 400 | 1 481,47 | 12,27 |
| 44. | kraf = tobbs, 0%* | 5 359 | 861,16 | 12,18 |
| 45. | kraf = 100%, kisebbs | 2 041 | 459,70 | 12,51 |
| 90. | ttipf = bp, varos* | 999 | 178,63 | 12,31 |

*F2. táblázat folytatása*

| Csomópont | Vágási pont | Megfigyelések száma | RSS | A bérek logaritmusának átlaga |
|---|---|---|---|---|
| 91. | ttipf = falu* | 1 042 | 203,15 | 12,70 |
| 23. | ag1f = 2, 3, 4, 6, 8, 10, 11, 13, 17, 18, 19 | 20 405 | 4 124,58 | 12,51 |
| 46. | nemf = no | 9 821 | 1 747,13 | 12,42 |
| 92. | kshregf = kd, nyd, dd, em, ea, da* | 5 012 | 737,11 | 12,30 |
| 93. | kshregf = km* | 4 809 | 863,48 | 12,54 |
| 47. | nemf = ferfi | 10 584 | 2 195,40 | 12,61 |
| 94. | szolgho<72,5 | 4 977 | 976,89 | 12,50 |
| 188. | kshregf = dd, em, ea, da* | 1 775 | 266,42 | 12,34 |
| 189. | kshregf = km, kd, nyd* | 3 202 | 643,94 | 12,58 |
| 95. | szolgho ≥ 72,5* | 5 607 | 1 103,85 | 12,70 |
| 3. | iskveg9f = fois, egyet | 36 276 | 16 301,40 | 12,95 |
| 6. | letszam_bv1 < 64,5 | 13 439 | 6 844,40 | 12,67 |
| 12. | kraf = 0% | 10 642 | 4 631,55 | 12,55 |
| 24. | ag1f = 1, 2, 3, 5, 6, 7, 8, 9, 10, 12, 13, 14, 16, 17, 18, 19 | 9 756 | 3 874,42 | 12,50 |
| 48. | iskveg9f = fois | 6 035 | 2 048,13 | 12,40 |
| 96. | letszam_bv1 < 17,5* | 2 113 | 582,81 | 12,25 |
| 97. | letszam_bv1 ≥ 17,5 | 3 922 | 1 397,87 | 12,47 |
| 194. | exp<9,5* | 835 | 149,27 | 12,23 |
| 195. | exp ≥ 9,5* | 3 087 | 1 184,98 | 12,54 |
| 49. | iskveg9f = egyet | 3 721 | 1 655,41 | 12,67 |
| 98. | letszam_bv1 < 16,5* | 1 220 | 418,42 | 12,45 |
| 99. | letszam_bv1 ≥ 16,5 | 2 501 | 1 147,96 | 12,78 |
| 198. | kor < 30,5* | 549 | 119,28 | 12,44 |
| 199. | kor ≥ 30,5* | 1 952 | 950,81 | 12,87 |
| 25. | ag1f = 4,11 | 886 | 510,26 | 13,05 |
| 50. | kshregf = kd, nyd, dd, em, ea, da* | 319 | 127,06 | 12,65 |
| 51. | kshregf = km* | 567 | 302,03 | 13,28 |
| 13. | kraf = 100%,tobbs,kisebbs | 2 797 | 1 439,70 | 13,14 |
| 26. | kor < 33,5* | 849 | 243,65 | 12,79 |
| 27. | kor ≥ 33,5 | 1 948 | 1 052,31 | 13,29 |
| 54. | ag1f = 1, 3, 5, 6, 8, 9, 10, 12, 14, 16, 17* | 864 | 399,91 | 13,06 |
| 55. | ag1f = 2, 4, 7, 11, 13, 18, 19* | 1 084 | 575,81 | 13,46 |
| 7. | letszam_bv1 ≥ 64,5 | 22 837 | 7 808,18 | 13,11 |

*F2. táblázat folytatása*

| Csomópont | Vágási pont | Megfigyelések száma | RSS | A bérek logaritmusának átlaga |
|---|---|---|---|---|
| 14. | kor < 31,5 | 6 180 | 1 130,92 | 12,84 |
| 28. | kshregf = kd, nyd, dd, ea, da* | 1 453 | 279,06 | 12,61 |
| 29. | kshregf = km, em* | 4 727 | 749,91 | 12,91 |
| 15. | kor ≥ 31,5 | 16 657 | 6 072,72 | 13,21 |
| 30. | iskveg9f = fois | 9 533 | 3 369,52 | 13,09 |
| 60. | nemf = no | 4 028 | 1 198,03 | 12,93 |
| 120. | ttipf = varos* | 1 717 | 469,31 | 12,77 |
| 121. | ttipf = bp, falu* | 2 311 | 651,32 | 13,05 |
| 61. | nemf = ferfi | 5 505 | 1 996,25 | 13,20 |
| 122. | ttipf = varos* | 2 592 | 812,85 | 13,06 |
| 123. | ttipf = bp, falu* | 2 913 | 1 078,43 | 13,34 |
| 31. | iskveg9f = egyet | 7 124 | 2 381,79 | 13,37 |
| 62. | kshregf = kd, nyd, dd, em, ea, da* | 2 202 | 749,14 | 13,20 |
| 63. | kshregf = km | 4 922 | 1 545,49 | 13,44 |
| 126. | kor < 38,5* | 2 206 | 462,00 | 13,32 |
| 127. | korr ≥ 38,5* | 2 716 | 1 019,30 | 13,55 |

*Megjegyzés:* a csillaggal jelölt csomópontok végpontok.
*Forrás:* Bértarifa-felvétel, saját számítás.

*F3. táblázat*
Véletlenerdő-szimulációk leíró statisztikái

| Változó | Átlag | Minimum | Maximum | Terjedelem | Szórás |
|---|---|---|---|---|---|
| Képzettség | 37,4127 | 36,8313 | 38,3955 | 1,5642 | 0,3198 |
| Vállalatméret | 16,7649 | 16,5221 | 17,0725 | 0,5504 | 0,1575 |
| Külföldi tulajdon | 6,0750 | 5,7566 | 6,4139 | 0,6573 | 0,1589 |
| Életkor | 6,1063 | 6,0336 | 6,2088 | 0,1752 | 0,0382 |
| Ágazat | 5,8108 | 5,6263 | 6,0435 | 0,4172 | 0,0809 |
| Tapasztalat | 5,4090 | 5,3409 | 5,4875 | 0,1466 | 0,0345 |
| Kollektív szerződés | 1,5953 | 1,4024 | 1,7486 | 0,3462 | 0,0819 |
| Állami tulajdon | 0,8793 | 0,8443 | 0,9108 | 0,0665 | 0,0157 |
| Szolgálati idő | 9,8208 | 9,6990 | 9,9852 | 0,2863 | 0,0519 |
| Régió | 4,6434 | 4,4213 | 4,8506 | 0,4292 | 0,1085 |
| Nem | 2,9491 | 2,8997 | 3,0022 | 0,1025 | 0,0221 |
| Új belépő | 0,4166 | 0,3997 | 0,4286 | 0,0290 | 0,0066 |
| Településtípus | 2,1167 | 1,9731 | 2,2701 | 0,2970 | 0,0666 |

*Forrás:* Bértarifa-felvétel, saját számítás.