Economica New Vol 9, No 2 (2018) 49-54.
ISSN 2560-2322

SZEGMENTÁLÁS DÖNTÉSI FA ALKALMAZÁSÁVAL

SEGMENTATION USING A DECISION TREE

Dudás Péter [1*]

[1] Közgazdasági, Pénzügyi és Menedzsment Tanszék - Gazdaságelemzés, módszertan
Gazdálkodási Kar, Neumann János Egyetem, Magyarország

Kulcsszavak:
döntési fa, szegmentálás

Keywords:
segmentation, decision tree

Összefoglalás
A gazdaságtudományok területén az igényes elemzéshez magas szintű matematikai, statisztikai módszerek ismerete szükséges. A különböző tételek, mutatók, eloszlások napjainkban már számítógépes könnyen kezelhető programokba beépített elemzési eszközök. Bizonytalan döntések esetén egyre több területen alkalmazzák a döntési fákat. A módszerrel döntési szabályok alkothatók, így szegmentálásra is használható. A tanulmány egy ilyen csoportosításra alkalmas klasszifikációt mutat be.

Abstract
In the field of economics, the knowledge of high-level mathematical and statistical methods is necessary for a demanding analysis. Today the various items, indexes, and distributions are user-friendly analytical tools that have been built into computer programmes. In the case of uncertain decisions, decision trees are being applied in more and more areas. Using this method, decision rules can be created, so they can be used for segmentation as well. The study presents a classification suitable for such a grouping.

1. Bevezetés

A matematika sajátos helyet foglal el a tudományok rendszerében. Önálló kategória, ugyanakkor sok tudomány segédtudománya. A gazdaságtudományok területén igényesebb elemzésekhez szükséges sok tétel, mutató, eloszlás, napjainkban már számítógépes könnyen kezelhető programokba beépített elemzési eszköz. Megfelelő használatukhoz kellő mélységű ismeret szükséges.

A társadalomtudományi kutatások egyik legnehezebb kérdése a megfigyelés és az elmélet közötti kapcsolat megteremtése. Megfigyelni és mérni csak ritkán lehet közvetlenül. A folyamatok állandóan változó körülmények között mennek végbe, ezért a jelenségek egyszeri megfigyelése lehetséges. A matematikai, statisztikai módszerekkel alkotott modellek egy része bizonytalan körülmények között készül, melyek a lényegesebb tényezőket veszik figyelembe, ugyanakkor a valószínűsíthető változókat is valamilyen formában kifejezésre kell juttatni. Egy ilyen modellt alkotó technika a döntési fák módszere. A döntési fa egy olyan, a döntéshozatalban használt grafikus modell, amikor több választási lehetőség is rendelkezésre áll, és a kimeneteik bizonytalanok.

* Kapcsolattartó szerző. E-mail cím: dudas.peter@gk.uni-neumann.hu

2. Alkalmazott módszer

A döntési fák alapvető funkciója, hogy egy komplex döntési problémát néhány kisebb problémára bontson. A módszer olyan döntéselőkészítő eszköz, amely segít vázolni a döntési helyzet teljes képét. A fa megmutatja, hogy az egyes utak milyen következményekhez vezetnek (WINSTON 2003).

KÁSA (2013) úgy fogalmaz, hogy a döntési fák olyan problémák megoldásában nyújtanak segítséget, amikor a döntéshozónak egymással összefüggő, láncolt döntéseket kell meghoznia kockázatos körülmények között. Tehát egyrészt a döntések véges számú sorozatát kell végrehajtani ahhoz, hogy a teljes döntési folyamat lezáruljon, másrészt ismerni kell a döntés során felmerülő összes lehetséges opciót és azt, hogy ezek mekkora eséllyel következnek be.

A döntési fák előállításának több módszere ismert, a legtöbb algoritmus bináris fát tud előállítani. Alkalmazásának előnye, hogy az algoritmus felismeri a lényegtelen változókat. Ez a probléma megértését is segíti, mert kiderül, hogy mely változók fontosak, és melyek nem. Így a döntések biztosabban meghozhatók.

A CHAID alapú döntési fák

A CHAID (Chi-squared Automatic Interaction Detector), egy többváltozós rekurzív osztályozó eljárás, amely eredetileg kategorikus változókról szólt (KASS 1980). Az algoritmus fejlesztésével később alkalmassá vált a függő változó, és a független változók esetében is folytonos ismérvek kezelésére (HÁMORI 2001).

A CHAID azért is kedvelt szegmentációs technika, mert alternatívája a hagyományos klaszteranalízisnek, amely alapvetően mennyiségi változókkal leírható megfigyelések csoportosítására alkalmas.

A döntési fák azért is népszerűek, mert a kijelölt függő változó és a magyarázó változók közötti kapcsolatrendszert vizuális formában, könnyen értelmezhető fastruktúrában lehet látni, és könnyen interpretálható. A modellkészítés szempontjából az eljárás előnye, hogy a változók mérési skálájára és azok eloszlására vonatkozóan nincs megkötés, az algoritmus kibővítésével a folytonos és kategóriás függő és független változókat egyaránt képes kezelni.

Megjegyzendő, hogy a skálák gyakorlati alkalmazásával kapcsolatban módszertani szempontból gyakran felmerülő probléma, hogy milyen szintű mérésűnek tekintsünk egy változót. A statisztikai programok elterjedése is felveti a kérdést, hogy a mérési skálák befolyásolják-e, és ha igen, mennyiben az alkalmazható módszereket. A programok használata a számításokat gyorsítja, de a kutatónak kell észben tartania, hogy milyen adatokkal dolgozik.

A metrikus változókkal kapcsolatban több kutató is igazolta a valószínűségszámításból ismert központi határeloszlás tétele alapján, hogy mindegy ordinális vagy intervallum típusúnak tekintjük-e az adatokat (pl. BORGATTA–BOHRNSTEDT 1980).

A CHAID-modellt alkotó rekurzív algoritmus fő lépései:

- a függő változó kijelölése,
- minden magyarázó változó esetén, azon kategóriák *egyesítése*, amelyek legkevésbé különböznek egymástól a függő változóra vonatkozóan,
- a megfigyelések magyarázó változó kategóriái szerinti *felosztása* /a felosztás utáni részadatbázisok jelentik a fastruktúra következő szintjét/,
- az algoritmus addig folytatja a kategóriák egyesítését és az esetek felosztását, míg el nem ér valamely előre definiált *megállítási* kritériumot.

A leírtak szemléltetésére a turizmus területéről vettem egy példát. A turizmus a gazdaság egyik legjelentősebb szektora. Igény van a turizmus tudományos kutatására, és az is tény, hogy a turizmus olyan komplex rendszer, amely szinte mindenkit érint. A KSH szerint 2016-ban a lakosság közel fele vett részt szabadidős utazáson, ami kedvezőbb a néhány évvel ezelőtt tapasztalt ⅓-nál, ugyanakkor az átlagos tartózkodási idő gyakorlatilag nem változott.

Az adatok feldolgozásához, az elemzésekhez az SPSS programot használtam.

Economica New Vol 9, No 2 (2018) 49-54.
ISSN 2560-2322

3. Eredmények

Egy előző kutatásban feltételeztem, hogy hagyományos szociodemográfiai ismérvekkel elkülöníthető a turisztikai motiváció, így az életkor vagy az ehhez köthető életciklus lehet a szegmentáció alapja. A számítások alapján a csoportok jól elkülönültek, ugyanakkor a hipotézist csak részben találtam igazoltnak.

Jelen tanulmányban a rendelkezésre álló adatbázis alapján kategorizáltam a megkérdezetteket, hogy érdekli-e valamilyen turizmusforma, és ha igen milyen kategóriák szerint alakulnak ki lényegesebb csoportok. Az adatbázis 302 esetet tartalmaz, öt kategóriás magyarázó változóval. A kiemelt változók az előzetes kutatás változóiból kerültek kiválasztásra. Ezek a következők:

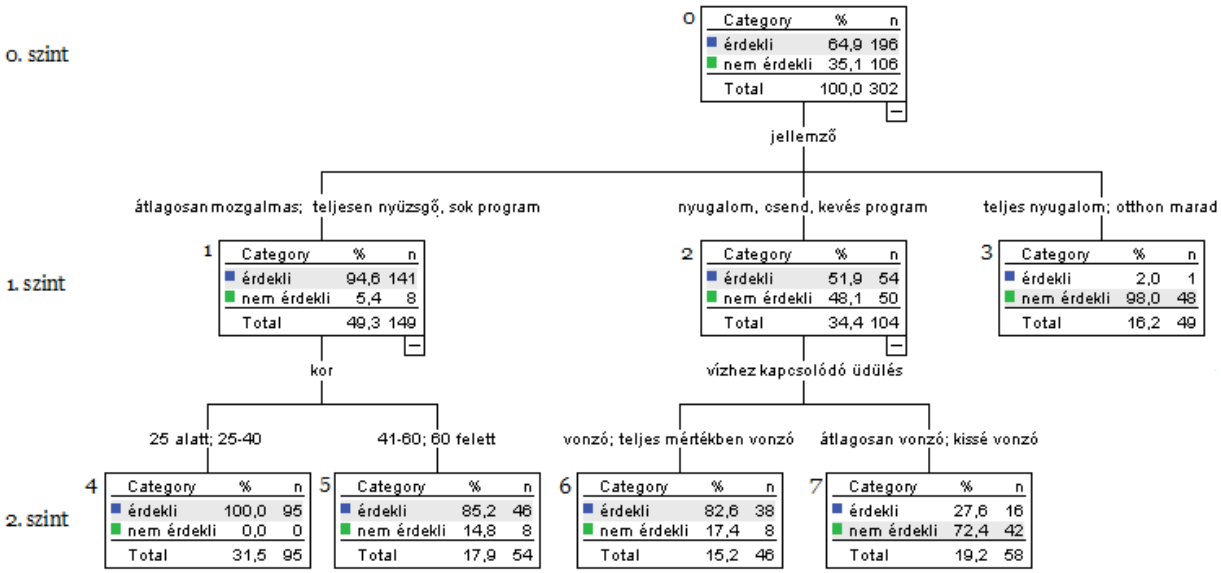
- 1: kor (25 alatt; 25-40; 41-60; 60 felett)
- 2: jellemző turizmus intenzitás (otthon marad; teljes nyugalom, pihenés; csendes hely kevés program; átlagosan nyüzsgő, mozgalmas hely; teljesen nyüzsgő hely, sok program)
- 3: vízhez kapcsolódó turizmus (nem vonzó; kismértékben vonzó; átlagosan vonzó; teljes mértékben vonzó)
- 4: egészségturizmus (nem vonzó; kismértékben vonzó; átlagosan vonzó; teljes mértékben vonzó)
- 5: rokonok, barátok meglátogatása kötelező ünnepeken kívül is (nem vonzó; kismértékben vonzó; átlagosan vonzó; teljes mértékben vonzó)

Célváltozó a turizmus valamilyen kiemelt formájáért való olyan érdeklődés, amely a részvételt indukálja. Az előzetes kutatás alapján ennek a változónak két összevonható kategóriája van: *érdekli* és *nem érdekli*.

A CHAID modell felépítéséhez elfogadtam az egyesítés és a felosztás küszöbszintjeinek a program által alapbeállításként ajánlott α= 0,05 szintet.

A megállási kritériumokat tekintve a program által kínált automatikus beállítást fogadtam el, ugyanakkor a felosztásra kerülő részadatbázisok minimális esetszámát 50-re, a felosztás során keletkező új részadatbázisok minimális esetszámát 30-ra (ez még a megkérdezettek 10%-a) állítottam be.

Az alapparaméterek beállítása után a program az 1. ábrán látható CHAID modellt alakította.



1. ábra. A CHAID modell

Az ábra a függő változó kategóriáinak megfelelően mutatja az adott részhalmazba tartozó esetek számát és azon belüli százalékos megoszlását. A kis táblákban a becsült érték az a kategória

lesz, amelyet az adott halmazba kerülő esetek több mint 50%-a képvisel. Ezt a program a háttér színezésével jelöli.

A fa *0*. szintje mutatja a teljes adatbázis megoszlását a célváltozó kategóriái szerint. A jobb oldali oszlop az egyes kategóriák elemszámait, a középső oszlop a százalékos megoszlást mutatja. Amennyiben csak ez az eredmény ismert, az állapítható meg, hogy a megkérdezettek 65%-a vesz részt a turizmus valamilyen formájában, és 35%-a nem. Ez is kedvezőbb, mint a KSH jelentés szerinti arány, ugyanakkor adódik a kérdés, hogy valóban állítható-e az, hogy a megkérdezettek 35%-a kizárható, vagy van olyan rejtett információ, ami alapján ez az arány csökkenthető.

Ehhez az algoritmus először minden magyarázó változóra elvégzi a lehetséges összevonásokat, és a program kiválasztja a magyarázó változók közül a legerősebb hatásút.

Esetünkben ez a turizmus intenzitását leíró változó. Ez azt is jelenti, hogy ennek a változónak a kategóriái gyakorolják a legnagyobb hatást a csoportok alakulására. A kiinduló adatbázis felosztása ennek a változónak a kategóriái mentén történt meg. Ez a fastruktúra *1*. szintje. A felosztás eredményeképpen előálló három részadatbázisban (1-3) megfigyelhetők a célváltozó kategóriáinak eloszlásai az egyes részadatbázisokat reprezentáló kis táblákban.

Látható, hogy az ismérvkategóriához való tartozás ismerete a bizonytalanságot lényegesen csökkenti. Ha mindenkit, aki az előzetes kutatásban a turizmus iránt nem érdeklődők (106 fő) csoportjába került, és ezért kizárnánk, akkor 58 olyan főt zárnánk ki, akiket tulajdonképpen érdekel a turizmus és részt is vesz valamilyen formájában. (A nyugodt csendes nyaralást választókból 50 fő, az ennél is mozgalmasabb szabadidő eltöltést kedvezőnek megítélőkből 8 fő.) Meg kell említeni, hogy az előzetes csoportosításban az érdeklődők közül 1 fő a turizmus egyetlen formáját sem választók közé, a nem érdeklődők értékelési halmazába került.

Ez együtt azt jelenti, hogy a megkérdezettek ([50+8+1]/302) 19,4 százalékában helytelenül döntenénk csak az előkutatás alapján. Így a kezdeti 35,1 százalékos passzivitást sikerült csökkenteni azáltal, hogy ismert a turizmus intenzitásához való hozzáállás.

Az összevonó–felosztó algoritmust tovább folytatva a *2*. szinten látható, hogy két azonos hatású magyarázó változó van, így a két részadatbázison a felosztás különböző módon történt. Akik az átlagosnál jobban kedvelik a nyüzsgő, mozgalmas nyaralást sok programmal, azokat az életkor osztja szét kb. ⅔-⅓ arányban (4,5 részhalmazok). A vízhez kapcsolódó üdülés, a nyugalmas, csendes kevés programot tartalmazó szabadidő eltöltést választókat 56%-44% arányban bontotta aszerint, hogy kevésbé, vagy jelentősen vízhez kapcsolják-e ezt az üdülési formát (6,7 részhalmazok).

A két részadatbázis szétválasztása után a döntési bizonytalanság ([8+8+16]/302) 7,9 százalékra csökkent. Ezen a szinten négy diszjunkt részadatbázisra lett felosztva az eredeti adatmátrix.

Az algoritmus a teljes nyugalommal történő csendes időtöltés, illetve az otthonmaradás ágat nem bontotta tovább, így ez egy levél.

1. Táblázat. A döntési fa reprezentációja

Halmaz	*érdekli*		*nem érdekli*		*becsült érték*	*Sig[a]*
	eset	*%*	*eset*	*%*		
0	196	64,9	106	35,1	érdekli	0,000
1	141	94,6	8	5,4	érdeki	0,000
2	54	51,9	50	48,1	érdekli	0,000
3	1	2,0	48	98,0	nem érdekli	0,000
4	95	100,0	0	0,0	érdekli	0,001
5	46	85,2	8	14,8	érdekli	0,001
6	38	82,6	8	17,4	érdekli	0,000
7	16	27,6	42	72,4	nem érdekli	0,000

a. Bonferroni adjusted *Saját számítás, SPSS tábla*

Összegezve, minden magyarázó változó kapcsolata szignifikáns az eredmény változóval, tehát bekerültek a modellbe. (A szignifikancia vizsgálat a Bonferroni-kiigazítással történt. Ezt több

hipotézis egyidejű tesztelése esetén szokás alkalmazni. Amennyiben „*n*" féle különböző hipotézis van, melyeket külön-külön α szignifikanciaszinten célszerű tesztelni, együttes fennállásuk esetén a szignifikanciaszintet α/*n* szinten kell megválasztani ahhoz, hogy az első fajú hiba elkövetésének valószínűsége ne legyen nagyobb, mint α. Esetünkben a különböző és egyidejűleg fennálló hipotéziseket a fastruktúra különböző szintjein vizsgált függetlenségi hipotézisek jelentik.)

Meg kell említeni, hogy az egészségturizmus és a rokonokkal, barátokkal történő szabadidő eltöltés is szignifikáns változók, amelyek akkor láthatók, ha a részadatbázisok beállított száma 30-ra, a felosztás során keletkező új részadatbázisok minimális esetszáma 10-re csökken. Ez a részletesebb fastruktúra (3. szint) a tanulmány terjedelme miatt nem közölhető.

2. Táblázat. A bizonytalanság változása

Szint	*nem érdekli a turizmus (bizonytalanság %)*	*bizonytalanság változása (%pont)*	
		előző szinthez	*0. szinthez*
0.	35,1	–	–
1.	19,4	– 15,7	– 15,7
2.	7,9	– 11,5	– 27,2
3.	7,6	– 0,3	– 27,5

Saját számítás

(Megjegyzés a 3. szinthez: A már említett újabb két változó jelentkezik. A középkorúaknál idősebbeket az egészségturizmus kedveltsége, a vízhez kapcsolódó időtöltést kevésbé kedvelőket a rokonok, barátok „kötelező viziten" kívüli meglátogatása bontja szét. Amint említettem a változók szignifikánsak, de hatásuk már nem jelentős.)

A bizonytalanság változása azt méri, hogy egy újabb változó bevonása a magyarázó változók közé hány százalékponttal csökkenti a magyarázat bizonytalanságát. Ebben az esetben jelentése az, hogy a passzív besorolások száma az egyes szinteken hány százalékponttal csökken a korábbi szinten mérthez képest.

A táblában „nem érdekli a turizmus" oszlop jól mutatja, hogy a növekvő szintek (egyre több magyarázó változó) hogyan eredményeznek pontosabb besorolásokat. Az előző szinthez viszonyított hibacsökkenés tendenciaszerűen csökken, de egyre kisebb mértékben. Az utolsó oszlop monoton növekedéssel mutatja, hogy az induló állapothoz képest az egyes lépések után összesesen mekkora hibacsökkenés érhető el.

A 3. táblázat a végső „levelekben" lévő esetekről ad információt, ezek mutatják a végső osztályozást. A „halmaz" oszlop mutatja a levélbe került esetek számát és ezek arányát az összes megkérdezetthez (302). Az „érdekli" oszlop az egyes levelekben a cél kategóriába (turizmus iránt érdeklődő) eső esetek számát mutatja, és ezek arányát az összes turizmus iránt érdeklődőhöz képest (196).

3. Táblázat. A modell összefoglalása

„levél"	*halmaz*		*érdekli*		*response %*	*index %*
	eset	*%*	*eset*	*%*		
4	95	31,5	95	48,5	100,0	154,1
5	54	17,9	46	23,5	85,2	131,3
6	46	15,2	38	19,4	82,6	127,3
7	58	19,2	16	8,2	27,6	42,5
8	49	16,2	1	0,5	2,0	3,1

Saját számítás, SPSS tábla

A „response” oszlop az egyes levelekben az „érdekli” arányt mutatja (%-ban) a levélbe tartozó esetek számához. Az index %, a response % és a gyökérbeli várható érdeklődés % (64,9%) aránya. Ez minden levélben 100% lenne, ha az algoritmus nem találna érdeklődést kiváltó okokat a független változókban. Minél nagyobb az index%, annál nagyobb a levélben megjelenő szabály hatása.

A modell ellenőrzését mutatja a 4. táblázat. Látható, hogy milyen a találati pontosság, tehát az egyes kategóriákhoz tartozó tényleges esetek közül hányszor volt egyezés és hányszor volt tévedés az előrejelzésben (a sorok tartalmazzák a tényadatokat, az oszlopok az előrejelzéseket). Annál jobb a modell, minél nagyobb részarányban esnek a megfigyelések a főátlóra. Ezek a helyes döntések. Ebből a teljes pontosság 89,4%, tehát a félreosztályozás kockázata 10,6%.

4. Táblázat. A klasszifikáció eredménye[a]

Tényleges besorolás	*Előrejelzett besorolás*		*Helyes döntés*
	érdekli	*nem érdekli*	
érdekli	188	8	95,9%
nem érdekli	24	82	77,4%
Pontosság	70,2%	29,8%	89,4%

az 1.ábrán nem látható 3. szinttel együtt

Saját számítás, SPSS tábla

4. Összegzés

A tanulmányban a rendelkezésre álló adatok elemzése után a következők állapíthatók meg:

- A kutatás rámutat, hogy az alkalmazott matematikai, statisztikai módszerek alkalmasak arra, hogy az adatok struktúráját felderítsük.
- Bizonytalan esetekben célszerű a döntéseket több oldalról alátámasztani.
- A döntési fán alapuló eljárás előnye, hogy alkalmazásával nemcsak részenkénti előrejelzés készíthető, hanem teljes kép alakítható az egyes szegmensek összetételről, ami a marketing kampányok tervezésénél döntő fontosságú.
- A bemutatott eljárás szabályokon alapul, amelyek szintenként növelik az adatok csoportosítását definiáló gráfot.
- A döntési fánál a szabályok egyúttal definiálják az adott feltételhez tartozó előrejelzést, ami a szabály megfogalmazása után az egyszerű többségi elven alapul.
- A döntési fák nagyméretű adathalmazokra is hatékonyan felépíthetők.

„Ne hozzunk döntést tudás nélkül.” (BERNOULLI)

Irodalomjegyzék

[1] Borgatta, E.F.–Bohrnstedt, G.W. (1980): Level of measurement – once over again. *Sociological Methods and Research*, 9 (2) pp. 147-160.
[2] Hámori G. (2001): A CHAID alapú döntési fák jellemzői. *Statisztikai Szemle*, 79. (8) pp. 703-710.
[3] Jánosa A. (2011): Adatelemzés SPSS használatával. Bp. ComputerBooks. pp. 264-285.
[4] Kása R (2013): Döntéselmélet. BGF Bp.
[5] Kass, G. (1980) An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29. (2) pp. 119–127.
[6] Rapcsák T. (2007): Többszempontú döntési problémák. BCE, pp. 4-5.
[7] Winston, W.L. (2003): Operations Research - Applications and Algorithms, 4th ed., Thomson Press, pp. 641-645.
[8] Zoltayné Paprika Z. (szerk.) (2002): Döntéselmélet, Alinea Kiadó, Budapest.