Statistica mintapélda dokumentumok osztályozására

Mintapéldaként egy 10 rövid dokumentumból álló korpuszt definiálunk. A dokumentumokat aszerint címkézzük fel, hogy a nyaralásról szólnak-e vagy sem. A forrásadatokat tartalmazó táblázatot az 1. ábra mutatja.

	1 Var1	2 Var2
1	autó jegy labda sapka uszoda	Igen
2	könyv dolgozni ebédidő csengő	Nem
3	nap sétalni inni enni	Igen
4	uszoda pihenni olvasni feküdni	Igen
5	számolni munkaruha dolgozni műszak	Nem
6	nyaralni labda pihenni sétálni	Igen
7	munkarend főnök kolléga dolgozni rohanni	Nem
8	hétköznap iskola komoly	Nem
9	nyár meleg aludni utazni	Igen
10	család utazni pihenni meleg	Igen

1. ábra. Minta induló adatai

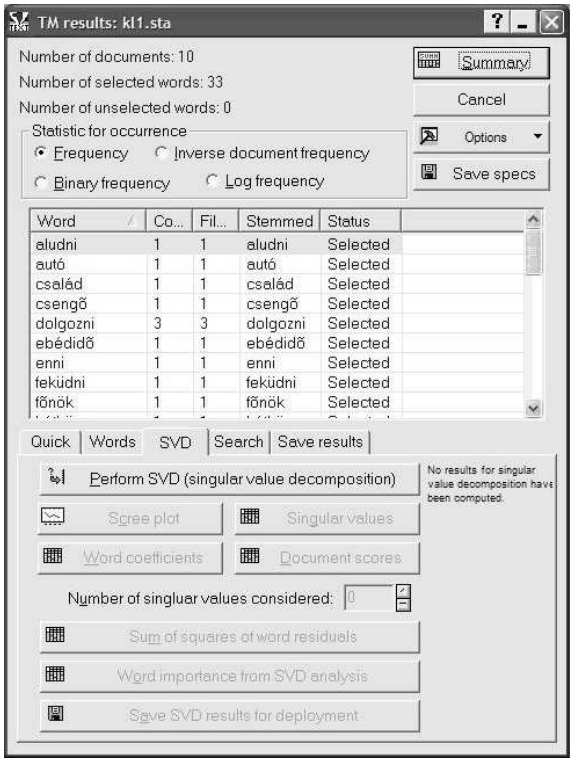
A minta feldolgozásához megnyitjuk a Text mining ablakot. A panelban a szöveg forrásaként változót adunk meg (Retrieve text contents from variable). Nem árt indulás előtt ellenőrizni, hogy a szövegben szereplő betűk mindegyike benne van-e az értelmezett karakterek halmazában. A panel Text variable nyomógombja mögött lehet a szöveget tartalmazó változókat kijelölni. Esetünkben a Var1 nevű változót adjuk meg. Az eredményül kapott szó–dokumentum gyakorisági mátrixot a 2. ábra mutatja be.

A kapott táblázatot külön lapon mentjük le. Ehhez a Save Results menüpontból a Save Statistic Values to Stand-Alone Spreadsheet funkciót aktivizáljuk. A kimentés során megadhatjuk, hogy a szógyakorisági adatok mellett mely forrásmezőt kívánjuk az új táblába átvinni. Példánkban a Var2 változót hozzuk át az eredménybe. A kapott táblázat első néhány oszlopát mutatja be a 3. ábra.

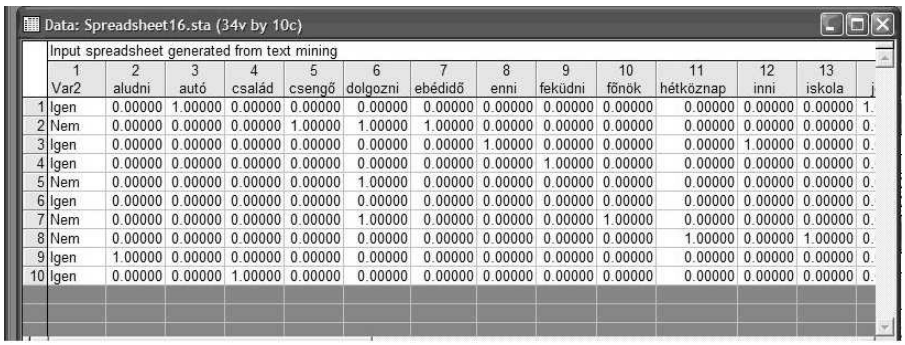
A következő lépésben meghatározzuk, hogy mely szavak játszanak fontosabb szerepet a kategória meghatározásában. Ehhez már a Statistica rendszer általános *Data Miner* adatbányászati modulját használjuk fel.

Indulásként aktivizáljuk a Feature Selection and Variable Screening opciót. A megjelenő ablakban (ld. 4. ábra) kell kijelölni a függő és független mennyiségeket. Jelen esetben a függő mennyiség egy kategóriaváltozó (Var2), a többi mező folytonos értékű független mennyiség.

Az elemzés lefutása után megjelenik az eredménypanel, amelyen most a 10 legjobb szó kapott helyet (ld. 5. ábra).

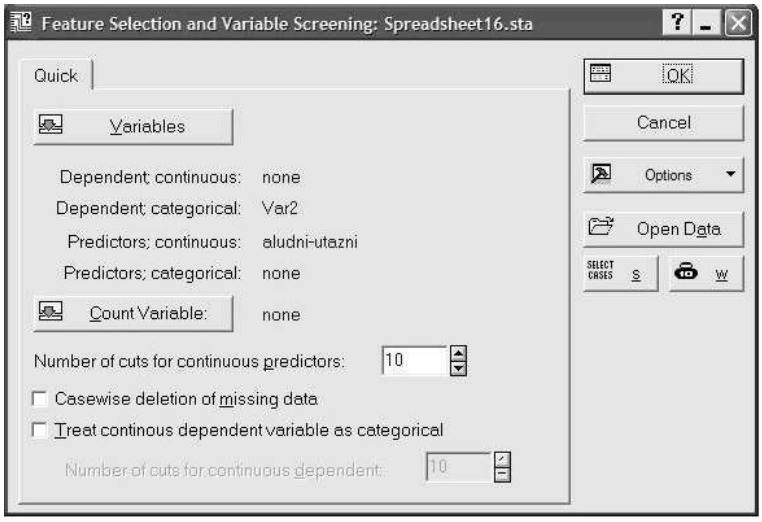


2. ábra. Gyakorisági mátrix

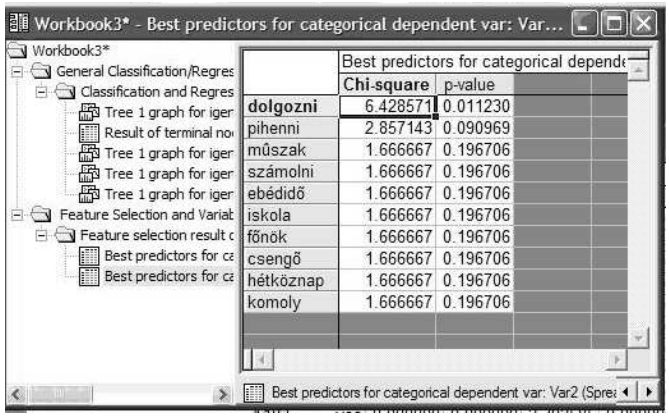


	1 Var2	2 aludni	3 autó	4 család	5 csengő	6 dolgozni	7 ebédidő	8 enni	9 feküdni	10 főnök	11 hétköznap	12 inni	13 iskola	j
1	Igen	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.
2	Nem	0.00000	0.00000	0.00000	1.00000	1.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.
3	Igen	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.
4	Igen	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.
5	Nem	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.
6	Igen	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.
7	Nem	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.
8	Nem	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	1.00000	0.
9	Igen	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.
10	Igen	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.

3. ábra. Eredménygyakorisági táblázat

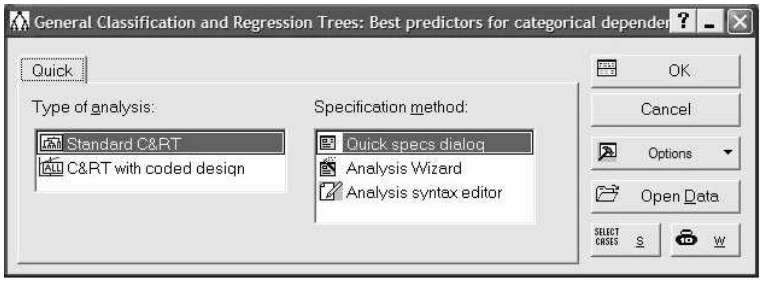


4. ábra. Dimenzió szelekció paraméterablaka



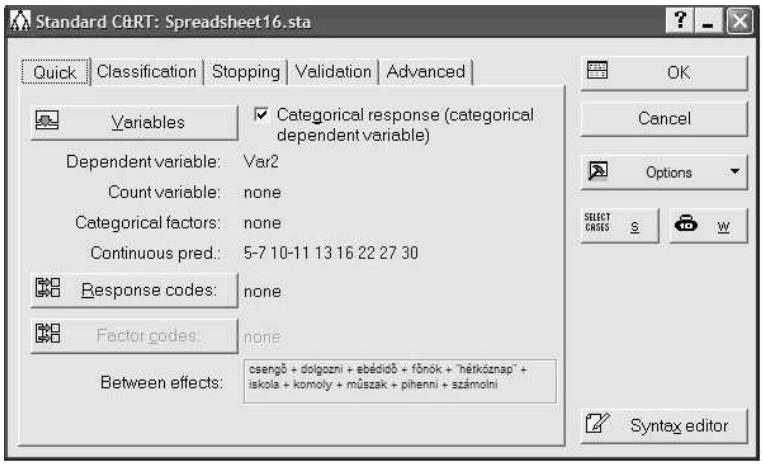
5. ábra. Kiválasztott dimenziók

Ezen szavakat fogjuk felhasználni a döntési fával végrehajtott osztályozásnál. A döntési fa felépítése a következő lépésekben megy végbe. A kapott lapot kijelöljük, mint alapértelmezési bemeneti forrást és elindítjuk a Data Mining menüpontból a General Classification/Regression Tree Models modult. A kapott ablakot a 6. ábra mutatja.



6. ábra. Az osztályozási módszer kijelölése

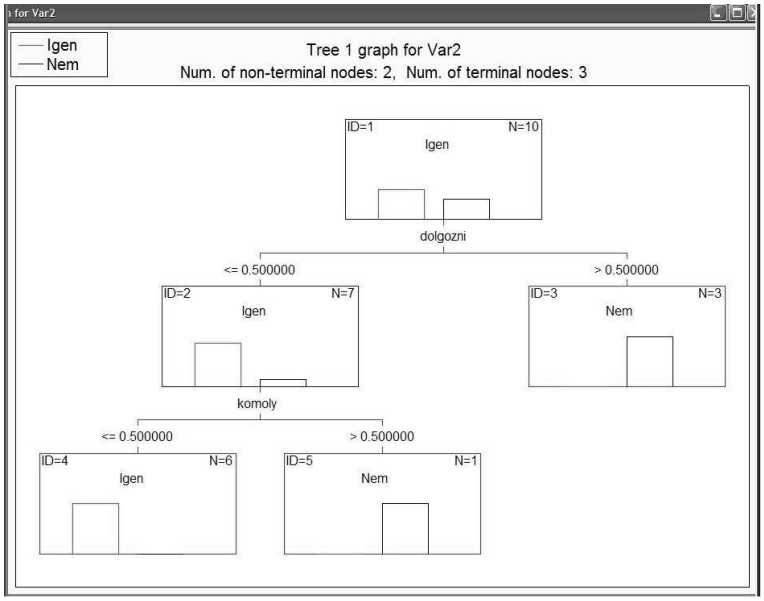
A lehetőségek közül a Standard CART[1] módszert választva a CART paramétereit beállító ablak jelenik meg. Itt is ki kell jelölni a figyelembe veendő változókat. Most csak azon változókat jelöljük ki, amelyek szerepelnek a legjobb szavak között. A Var2 mező itt is kategória típusú függő mennyiség lesz.



7. ábra. Osztályozási paraméterek kijelölése

A fa felépítés megállási feltételeként a FACT-style módot válasszuk ki. A döntési fa előállítása után megkapjuk a megjelenítő ablakot. Ebben a Summary fülön lévő Tree Graph opciót aktivizáljuk. Ekkor grafikus formában megkapjuk a fa szerkezetét, mint azt a 8. ábra is mutatja.

[1] A program a módszerre a CRT rövidítést használja.



8. ábra. Az eredményként kapott döntési fa

Az eredményből látható, hogy előbb a *dolgozni* szót kell vizsgálni. Ha szerepel, akkor *Nem* értékű a kategória. Ellenkező esetben a *komoly* szót kell ellenőrizni. Ha szerepel a dokumentumban, akkor a dokumentum *Nem* kategóriát fog kapni, ha pedig nem, akkor *Igen* kategóriaérték lesz az eredmény.