Adatbányászati technikák órai feladatsor
2018. február 26

1. Az alábbi táblázatban arról vannak adatok, hogy adott nemű (Nő/Férfi), adott kocsitípussal rendelkező (Családi/Sport/Luxus), adott ruhaméretű (S/M/L/XL) emberek két osztály közül ($C0$ és $C1$) melyikbe tartoznak.
   (a) Mennyi az egész tanítóhalmaz inhomogenitása, ha Gini-indexet használunk?
   (b) Mennyi a nyereség, ha ID alapján vágunk?
   (c) Mennyi a nyereség, ha a Nem alapján vágunk?
   (d) Mennyi a nyereség, ha a Kocsi alapján vágunk multiway splittel?
   (e) Mennyi a nyereség, ha a Ruha alapján vágunk multiway splittel?
   (f) Melyik vágást választja a döntési fát építő algoritmus?
   (g) Melyik vágást választja a döntési fát építő algoritmus, ha classification errort használunk?

ID	**Nem**	**Autó**	**Ruha**	**Osztály**
1	F	Cs	S	C0
2	F	S	M	C0
3	F	S	M	C0
4	F	S	L	C0
5	F	S	XL	C0
6	F	S	XL	C0
7	N	S	S	C0
8	N	S	S	C0
9	N	S	M	C0
10	N	L	L	C0
11	F	Cs	L	C1
12	F	Cs	XL	C1
13	F	Cs	M	C1
14	F	L	XL	C1
15	N	L	S	C1
16	N	L	S	C1
17	N	L	M	C1
18	N	L	M	C1
19	N	L	M	C1
20	N	L	L	C1

2. Adjon példát olyan tanítóhalmazra, melyben két attribútum, A és B, alapján próbáljuk a bináris célváltozót előre jelezni és a tanult döntési fa építő algoritmusnál sem A, sem B esetén vágva először nem kapunk pozitív nyereséget, de ha előbb A, aztán B alapján (vagy fordítva) két vágást hajtunk végre, akkor a levelek teljesen homogének lesznek.