Adatbányászat beadandó projekt

Készítette: Molnár Gábor (K56G28)

Adathalmaz: Gombák (Mushrooms)

Adathalmaz karakterisztikája:	Multivariáns
Attribútumok karakterisztikája:	Kategorikus
Hozzárendelt feladat:	Osztályozás
Adatok száma:	8124
Attribútumok száma:	22
Vannak-e hiányzó értékek?	Igen
Terület:	Élet
Létrehozás dátuma:	1987.04.27.
Webes találatok száma:	153745

Forrás:

http://archive.ics.uci.edu/ml/datasets/Mushroom

The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf

Információ az adathalmazról:

Az adathalmaz az Agaricus és a Lepiota családba tartozó 23 lemezes gombáról tartalmaz leírást. Minden fajt külön-külön azonosítottak úgy, hogy határozottan ehető, határozottan mérgező, ismeretlen a hatása és nem ajánlott. Az utóbbi tartalmazhat mérgező gombákat is. Az útmutató világosan elmondja, hogy nincs egy egyszerű szabály, ami alapján meg tudjuk határozni, hogy a gomba ehető-e.

Információ az attribútumokról:

#	Elnevezés	Érték (angol)	Érték (rövid)	Érték (magyar)
1	sapka alakja	bell	b	harang

	(cap-shape)	conical	c	kúpos
		convex	x	konvex
		flat	f	lapos
		knobbed	k	göcsörtös
		sunken	s	beesett
2	sapka felszíne *(cap-surface)*	fibrous	f	szálas
		grooves	g	barázdás
		scaly	y	pikkelyes
		smooth	s	sima
3	sapka színe *(cap-color)*	brown	n	barna
		buff	b	sárgásbarna
		cinnamon	c	fahéj
		gray	g	szürke
		green	r	zöld
		pink	p	rózsaszín
		purple	u	lila
		red	e	piros
		white	w	fehér
		yellow	y	sárga
4	zúzódások *(bruises?)*	bruises	t	zúzódások
		no	f	nincs

5	szag *(odor)*	almond	a	mandula
		anise	l	ánizsos
		creosote	c	karbonsav
		fishy	y	hal
		foul	f	büdös
		musty	m	dohos
		none	n	nincs
		pungent	p	szúrós
		spicy	s	fűszeres
6	lemez illeszkedés *(gill-attachment)*	attached	a	csatolt
		descending	d	csökkenő
		free	f	szabad
		notched	n	bemetszett
7	lemez távolság *(gill-spacing)*	close	c	zárt
		crowded	w	zsúfolt
		distant	d	távoli
8	lemez méret *(gill-size)*	broad	b	széles
		narrow	n	keskeny
9	lemez szín *(gill-color)*	black	k	fekete
		brown	n	barna
		buff	b	barnássárga

		chocolate	h	csokoládé
		gray	g	szürke
		green	r	zöld
		orange	o	narancs
		pink	p	rózsaszín
		purple	u	lila
		red	e	piros
		white	w	fehér
		yellow	y	sárga
10	szár alakja *(stalk-shape)*	enlarging	e	bővülő
		tapering	t	szükülő
11	szár alja *(stalk-root)*	bulbous	b	hagymaszerű
		club	c	golfütő
		cup	u	csésze
		equal	e	egyenletes
		rhizomorphs	z	gyökérszerű
		rooted	r	gyökeres
		missing	?	hiányzó
12	szár felszín felső gyűrű *(stalk-surface-above-ring)*	fibrous	f	szálas
		scaly	y	pikkelyes
		silky	k	selyemszerű

		smooth	s	sima
13	szár felszín alsó gyűrű (stalk-surface-below-ring)	fibrous	f	szálas
		scaly	y	pikkelyes
		silky	k	selyemszerű
		smooth	s	sima
14	szár szín felső gyűrű *(stalk-color-above-ring)*	brown	n	barna
		buff	b	barnássárga
		cinnamon	c	fahéj
		gray	g	szürke
		orange	o	narancs
		pink	p	rózsaszín
		red	r	piros
		white	w	fehér
		yellow	w	sárga
15	szár szín alsó gyűrű *(stalk-color-below-ring)*	brown	n	barna
		buff	b	barnássárga
		cinnamon	c	fahéj
		gray	g	szürke
		orange	o	narancs
		pink	p	rózsaszín
		red	e	piros

		white	w	fehér
		yellow	y	sárga
16	fátyol típus *(veil-type)*	partial	p	részleges
		universal	u	teljes
17	fátyol szín *(veil-color)*	brown	n	barna
		orange	o	narancs
		white	w	fehér
		yellow	y	sárga
18	gyűrű szám *(ring-number)*	none	n	nincs
		one	o	egy
		two	t	kettő
19	gyűrű típus *(ring-type)*	cobwebby	c	pókhálós
		evanescent	e	tünékeny
		flaring	f	lobogó
		large	l	nagy
		none	n	nincs
		pendant	p	függő
		sheathing	s	köpeny
		zone	z	zóna
20	spóra szín *(spore-print-color)*	black	k	fekete
		brown	n	barna

		buff	b	barnássárga
		chocolate	h	csokoládé
		green	r	zöld
		orange	o	narancs
		purple	u	lila
		white	w	fehér
		yellow	y	sárga
21	populáció *(population)*	abundant	a	bőséges
		clustered	c	csoportosított
		numerous	n	számos
		scattered	s	elszórt
		several	v	néhány
		solitary	y	magányos
22	előhely *(habitat)*	grasses	g	füvek
		leaves	l	levelek
		meadows	m	rétek
		paths	p	ösvények
		urban	u	városok
		waste	w	hulladék
		woods	d	fák

Feladat:

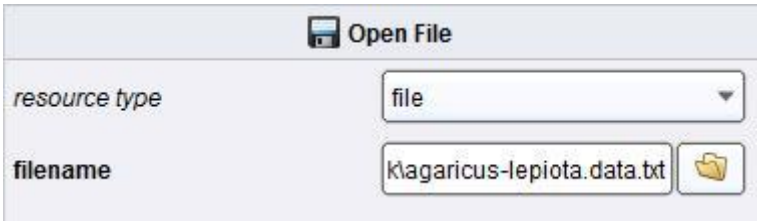
Egy osztályozási modell felállítása, mely az attribútumok alapján az egyes gombákról szóló rekordokat az ehető (**edible = e**) és a mérgező (**poisoness = p**) osztályokba sorolja.

Feladat megoldásának lépései:

Az adatokat tartalmazó fájl **agaricus-lepiota.data.txt**, amely egy egyszerű csv állományként tárolja az egyes gombákat reprezentáló rekordokhoz tartozó 22 attribútum értéket vesszővel elválasztva. Az adatokhoz tartozó metadatokat a **agaricus-lepiota.names.txt** fájl tartalmazza, melyből többek között az attribútumok neveit és az egyes értékek valódi jelentését is megtudhatjuk.

A fájl megnyitása:

1. ábra: Open File operátor

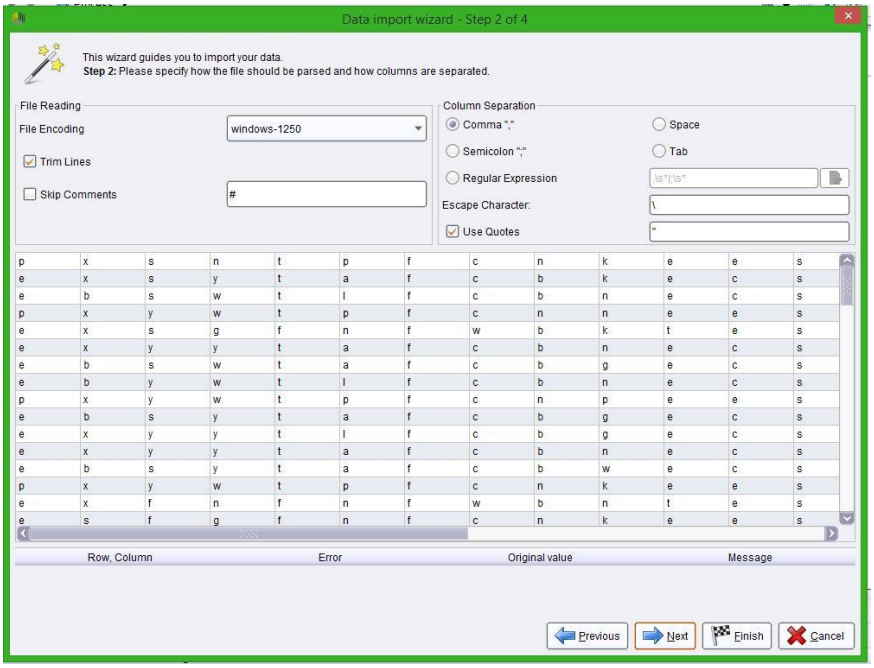


2. ábra: Open File operátor beállításai

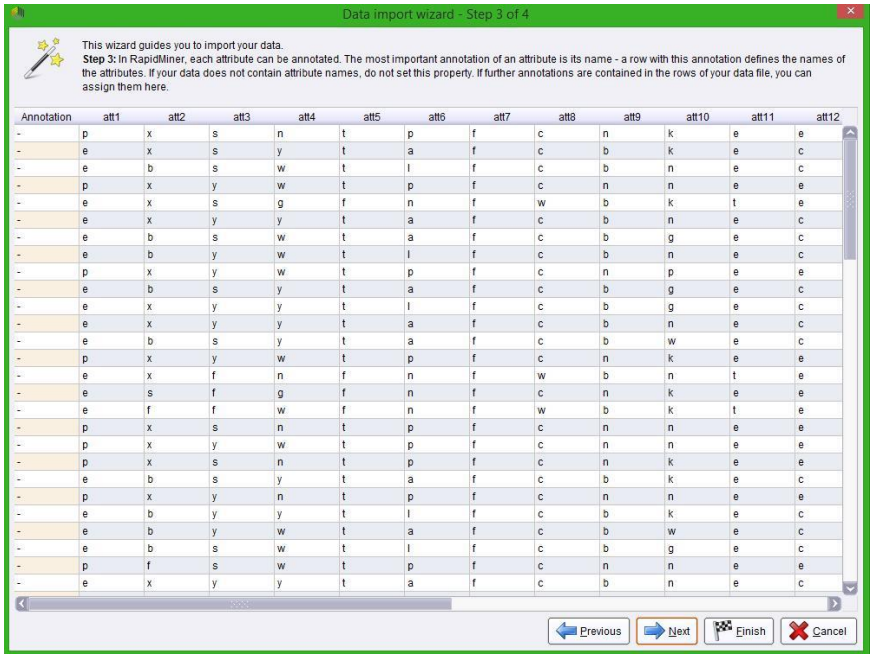
Az **Open File** operátor segítségével tudjuk az adatokat tartalmazó fájlt megnyitni. A beállításokban a **resource type** tulajdonságot **file** típusúra kell állítani, illetve a **filename** attribútum esetén meg kell adnunk a fájl elérési útvonalát. Ezen operátor segítségével, akár közvetlenül az UCI szerveréről URL cím alapján is megnyithatjuk a fájlt.

Az adatok beolvasása:

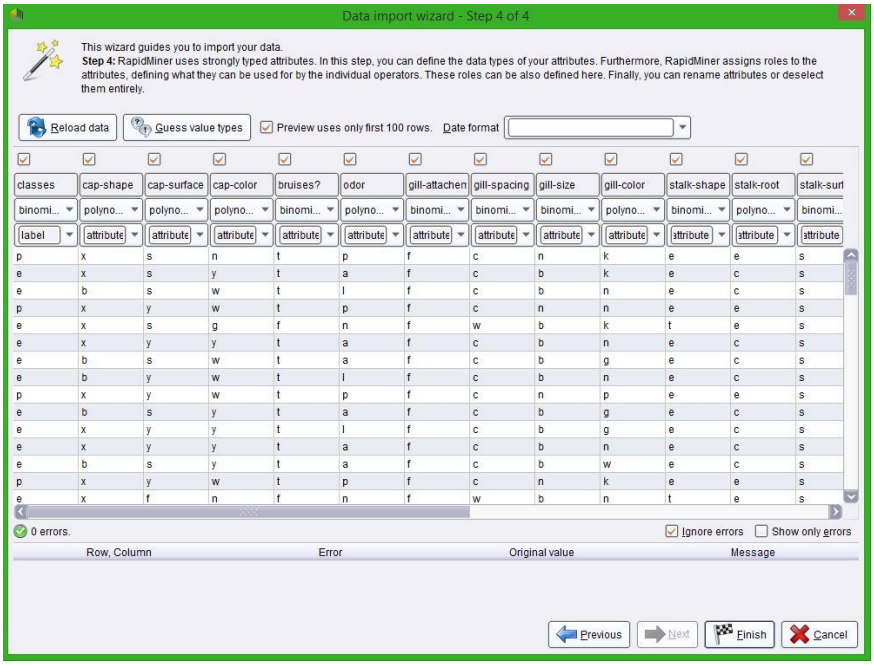
3. ábra: Read CSV operátor



4. ábra: Read CSV operátor wizard 1. lépés



5. ábra: Read CSV operátor wizard 2 lépés



6. ábra: Read CSV operátor wizard 3. lépés

Az adatokat valamilyen módon a Rapidminer-nek értelmezhetővé kell tennünk. Ezt teszi lehetővé a **Read CSV** operátor, melynek segítségével könnyedén értelmezhetünk csv típusú fájlokat. A fájl beolvasása után meg kell adnunk, hogy az egyes értékeket milyen elválasztó karakter választja el (jelen esetben vessző), aztán meg kell adnunk, hogy a fájl tartalmazza-e az oszlopneveket. Esetünkben nem tartalmazta, ezért a beolvasó varázsló (wizard) 3. lépésében manuálisan írtam be az egyes oszlopok nevét a meta adatokat tartalmazó fájlból. Továbbá ezen a ponton állíthatjuk be, hogy az egyes oszlopokban milyen típusú és szerepű adatok találhatóak. Általában a Rapidminer helyesen ismeri fel az adatok típusát, mint ennél a fájlnál is, egyedül a **classes** oszlop szerepét kellett **attribute** értékről **label** értékre állítani.

Hiányzó értékek:

Az adatok beolvasása után láthatjuk, hogy az egyes rekordokhoz tartozó attribútumokban találhatunk hiányzó értékeket. Ezt az információt beolvasás után a **Meta Data View** fülön láthatjuk.

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

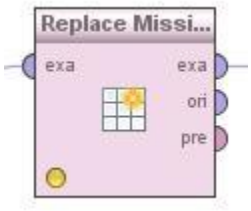
ExampleSet (8124 examples, 1 special attribute, 22 regular attributes)

Role	Name	Type	Statistics	Range	Missings
label	classes	binominal	mode = e (4208), least = p (3916)	p (3916), e (4208)	0
regular	cap-shape	polynominal	mode = x (3656), least = c (4)	x (3656), b (452), s (32), f (3152), k (828), c (4)	0
regular	cap-surface	polynominal	mode = y (3244), least = g (4)	s (2556), y (3244), f (2320), g (4)	0
regular	cap-color	polynominal	mode = n (2284), least = u (16)	n (2284), y (1072), w (1040), g (1840), e (150(	0
regular	bruises?	binominal	mode = f (4748), least = t (3376)	t (3376), f (4748)	0
regular	odor	polynominal	mode = n (3528), least = m (36)	p (256), a (400), l (400), n (3528), f (2160), c ('	0
regular	gill-attachement	binominal	mode = f (7914), least = a (210)	f (7914), a (210)	0
regular	gill-spacing	binominal	mode = c (6812), least = w (1312)	c (6812), w (1312)	0
regular	gill-size	binominal	mode = b (5612), least = n (2512)	n (2512), b (5612)	0
regular	gill-color	polynominal	mode = b (1728), least = r (24)	k (408), n (1048), g (752), p (1492), w (1202),	0
regular	stalk-shape	binominal	mode = t (4608), least = e (3516)	e (3516), t (4608)	0
regular	stalk-root	polynominal	mode = b (3776), least = r (192)	e (1120), c (556), b (3776), r (192), ? (2480)	0
regular	stalk-surface-above-ring	binominal	mode = s (5176), least = f (552)	s (5176), f (552)	2396
regular	stalk-surface-below-ring	polynominal	mode = s (4936), least = y (284)	s (4936), f (600), y (284), k (2304)	0
regular	stalk-color-above-ring	binominal	mode = w (4464), least = g (576)	w (4464), g (576)	3084
regular	stalk-color-below-ring	binominal	mode = w (4384), least = p (1872)	w (4384), p (1872)	1868
regular	veil-type	binominal	mode = p (8124), least = p (8124)	p (8124)	0
regular	veil-color	binominal	mode = w (7924), least = n (96)	w (7924), n (96)	104
regular	ring-number	binominal	mode = o (7488), least = t (600)	o (7488), t (600)	36
regular	ring-type	binominal	mode = p (3968), least = e (2776)	p (3968), e (2776)	1380
regular	spore-print-color	polynominal	mode = w (2388), least = u (48)	k (1872), n (1968), u (48), h (1632), w (2388),	0
regular	population	polynominal	mode = v (4040), least = c (340)	s (1248), n (400), a (384), v (4040), y (1712), c	0
regular	habitat	polynominal	mode = d (3148), least = w (192)	u (368), g (2148), m (292), d (3148), p (1144),	0

7. ábra: A Meta Data View táblázat

Az táblázat utolsó oszlopa alapján a 22 attribútumból csak 6 attribútum esetén van hiányzó érték.

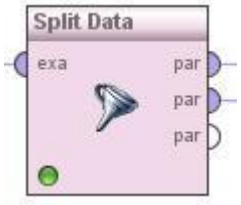
Attribútum neve	**Hiányzó érték száma**
stalk-surface-above-ring	2396
stalk-color-above-ring	3048
stalk-color-below-ring	1868
veil-color	104
ring-number	36
ring-type	1380



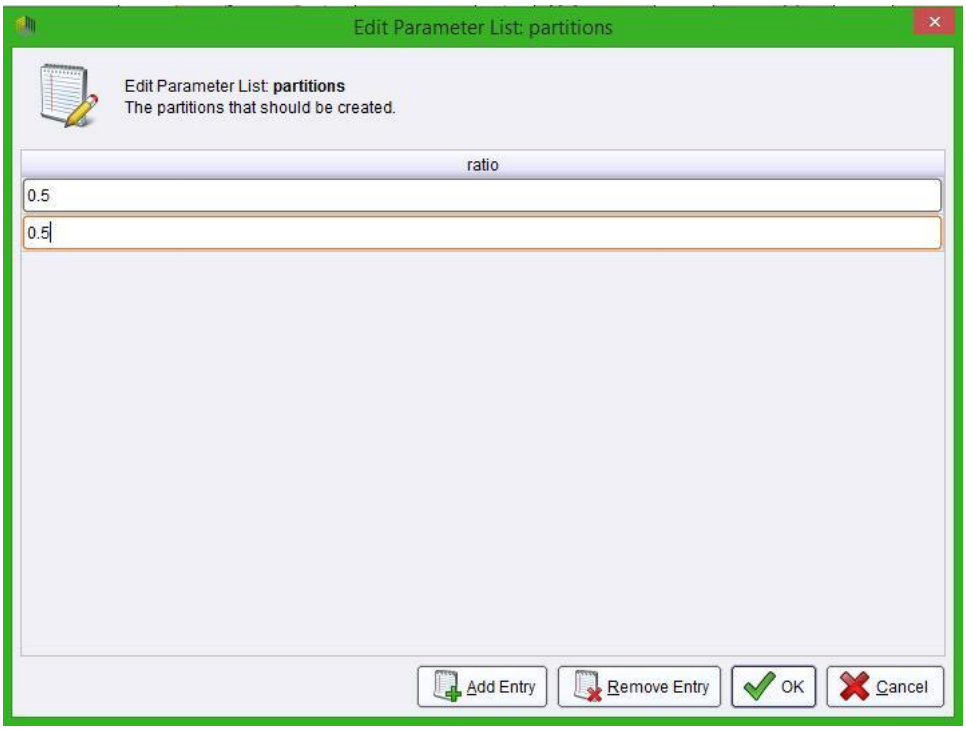
8. ábra: Replace Missing Values operátor

A **Replace Missing Values** operátort használjuk a hiányzó adatok helyettesítésére. Ezt az operátort úgy állítottam be, hogy a hiányzó értékeket az adott tulajdonságbeli értékek átlagával helyettesítse.

Az adatok felosztása:



9. ábra: Split Data operátor



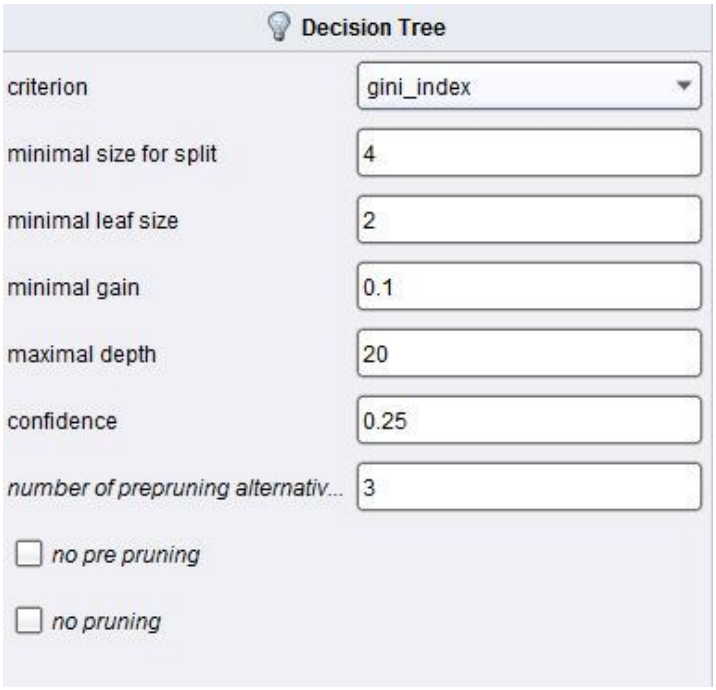
10. ábra: Split Data operátor arányok beállítása

A **Split Data** operátort alkalmazva létrehozhatunk az input adatokból több adatállományt. A gombákat leíró adatokat két részre osztottam, 50%-50% arányban. Az első fél tartalmazza a tanító adatokat, míg a második fele a teszt adatállomány, melyen validáljuk a tanító adatokon illesztett modellt.

Döntési fa:

11. ábra: Decision Tree operátor

A **Decision Tree** operátor az inputként kapott tanító adatállomány és a megfelelő paraméterek segítségével egy döntési fát generál. Ez a döntési fa szolgál a modellünk alapjául, amelyre következő operátor fogja (próbálja) ráilleszteni a teszt adatokat.



12. ábra: A Decision Tree operátor alapbeállítási a Gini indexszel

Megnevezés	**Leírás**	**Típus**
criterion	Kritérium kiválasztása, hogy mely attribútumok esetén történjen a vágás. Választható értékek: **information_gain**, **gain_ratio**, **gini_index**, **accuracy**.	lista
minimal size for split	A minimális csomópont szám, amellyel ha megegyezik az adott csomópontbeli érték, vagy nagyobb tőle, akkor vágásra kerül sor.	egész
minimal leaf size	A minimális levélelemek száma, amely a fa generálásánál szükséges.	egész
minimal gain	Minimális gain, amely a csúcs vágása előtt számolunk. Ha ettől nagyobb a szám, akkor megtörténik a vágás, ellenben nem. Nagy minimális gain szám esetén kisebb fa jöhet létre.	valós
maximal depth	A fa maximális mélységét állíthatjuk be.	egész
confidence	Megad egy olyan értéket, amely egy pesszimista hiba számolást nyújt a metszéshez.	valós
number of prepruning...	Párhuzamosan fut a fa generálásakor.	egész
no pre pruning	Nincs előmetszés.	logikai
no pruning	Nincs metszés.	logikai

Gini index

A vágások esetén szennyezettségi mérőszámra van szükségünk. A szennyezettség mérésére és ez által mérőszám előállítására több módszer is rendelkezésünkre áll, például: **Gini index, Entrópia, Téves osztályozási hiba.** A feladat elvégzése során a Gini index módszert használtam fel, és ennek alapján készült el a döntési fa. Ez a módszer lehetőséget ad a vágások jóságának mérésére. A CART, SLIQ, SPRINT algoritmusok is a Gini indexet használják.

Modell alkalmazása:

13. ábra: Apply Model operátor

A tanító adatokra alkalmazva a **Decision Tree** operátor létrehozott egy modellt. A modellt és a teszt adatokat felhasználva az **Apply Model** operátor futtatása után olyan címkézett (**labeled**) adatokat kapunk, amelyekkel már végezhetünk méréseket.

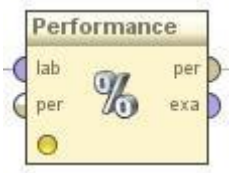
ExampleSet (4062 examples, 4 special attributes, 22 regular attributes)

Row No.	classes	confidence(p)	confidence(e)	prediction(classes)	cap-shape	cap-surface	cap-color	bruises?	odor	gill-attache
1	e	0	1	e	x	s	g	f	n	f
2	e	0	1	e	x	y	y	t	a	f
3	p	1	0	p	x	y	w	t	p	f
4	e	0	1	e	b	s	y	t	a	f
5	e	0	1	e	x	y	y	t	l	f
6	p	1	0	p	x	y	w	t	p	f
7	e	0	1	e	s	f	g	f	n	f
8	e	0	1	e	f	f	w	f	n	f
9	p	1	0	p	x	s	n	t	p	f
10	e	0	1	e	b	s	y	t	a	f
11	e	0	1	e	b	y	y	t	l	f
12	e	0	1	e	x	y	y	t	a	f
13	e	0	1	e	f	f	n	f	n	f
14	e	0	1	e	b	s	y	t	l	f
15	p	1	0	p	x	y	w	t	p	f
16	e	0	1	e	x	y	y	t	l	f
17	e	0	1	e	b	y	y	t	l	f
18	e	0	1	e	s	f	g	f	n	f

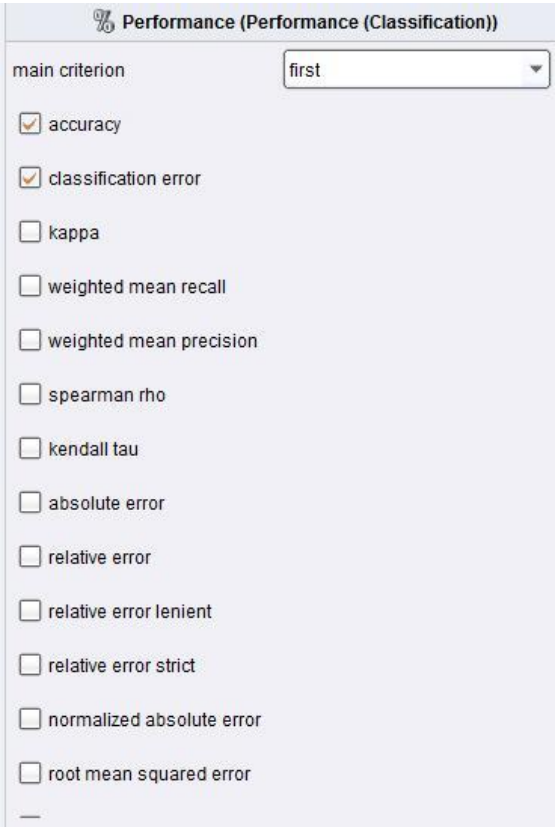
14. ábra: Adatok az Apply Model operátor alkalmazása után

Az táblázatban 4 kiemelt oszlopot láthatunk. Az első (**classes**) oszlop az eredeti adatállományban található label szerepkörű attribútum, a másik három oszlop pedig az **Apply Model** operátor alkalmazásával létrejött konfidencia és predikciós oszlopok. A predikciós oszlopban találhatjuk annak az osztálynak a nevét, amely az elvárt, illetve a konfidencia oszlopokban láthatjuk annak az eredményét, hogy az elvárt osztályba tartozás problémája százalékosan mennyire teljesült.

Osztályozás:



15. ábra: Performance (Classification) operátor



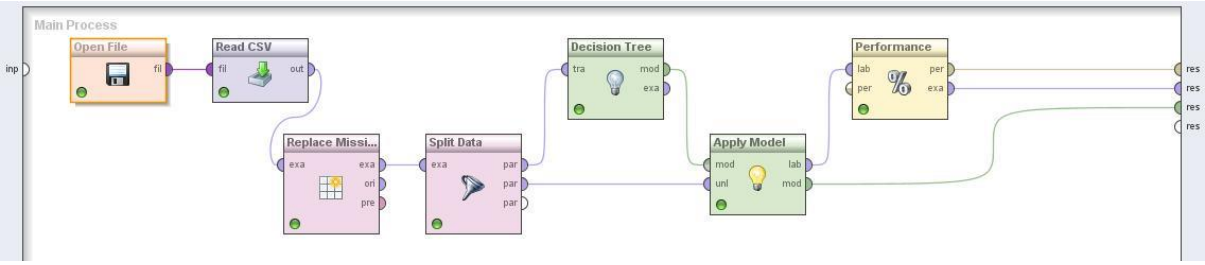
16. ábra: Performance (Classification) beállítások

A **Performance (Classification)** operátor a beállításaiban kiválasztott statisztikai adatokat szolgáltat számunkra, mellyel meg tudjuk állapítani, hogy az osztályozási feladatok és az előállított modell illesztése mennyire volt sikeres.

Kiíratni az alábbi értékeket tartottam fontosnak:

1. pontosság (**accuracy**)
2. osztályozási hiba (**classification error**)
3. korreláció (**correlation**)
4. korreláció négyzet (**squared correlation**)

A feladat megoldása:

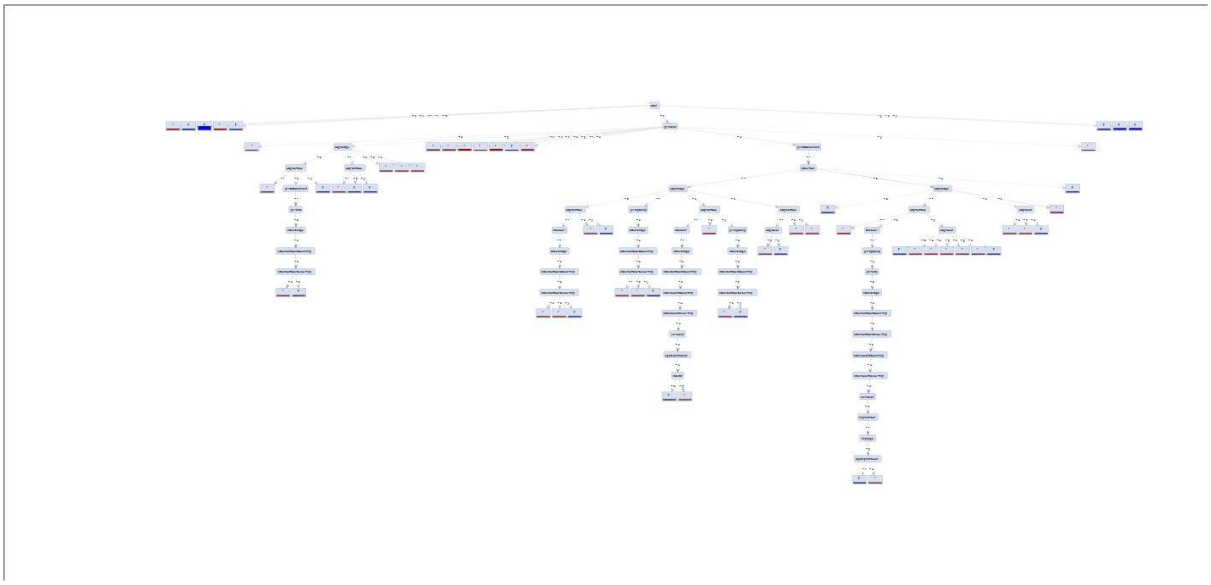


17. ábra: Az operátorok

Eredmény

Alapbeállításokkal

Döntési fa



18. ábra: A döntési fa alapbeállításokkal és Gini indexszel

Peformancia vektor

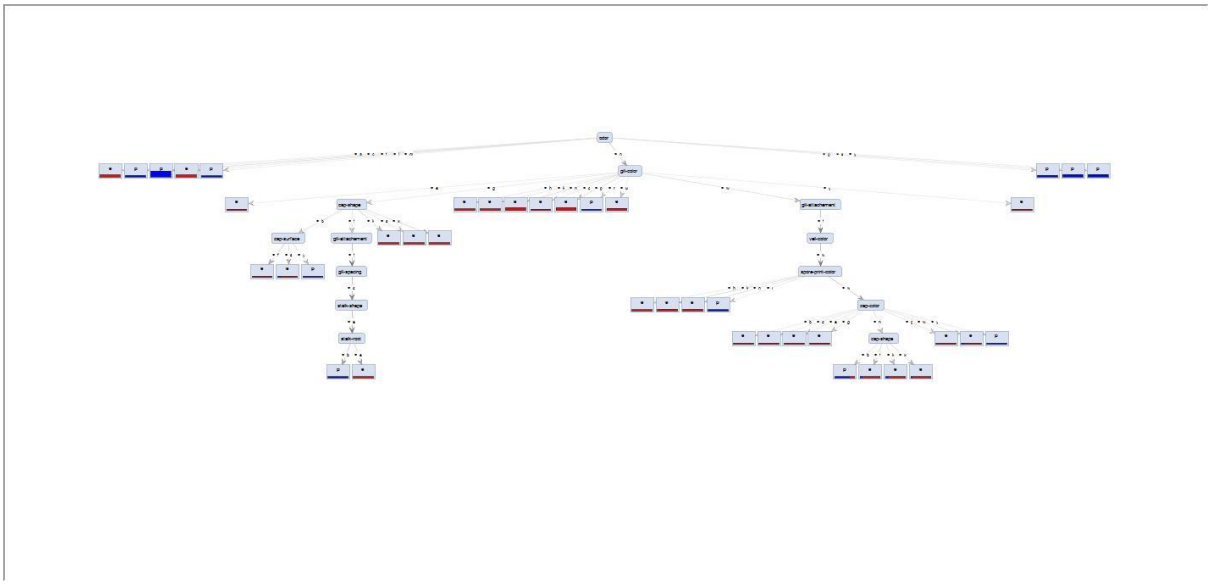
pontosság (**accuracy**)	99.34%
osztályozási hiba (**classification_error**)	0.66%
korreláció (**correlation**)	0.987
korreláció négyzet (**squared correlation**)	0.974

Az elkészült projektet az operátorok alapbeállításai mellett futtattam le először. Amely meglepően jó eredménnyel szolgált. A táblázatban láthatjuk, hogy a pontosság nagyobb, 99% és az osztályozási hiba ezzel szemben elég kicsi. Az osztály és a feltételezett osztály oszlopainak értékei nagymértékben állnak kapcsolatban egymással a korreláció értéke alapján. Egyetlen szépséghibája a komplex döntési fa, amely arra ad következtetést, hogy

talán túlillesztéssel állunk szemben, ezért megpróbáltam az alapbeállítások mellett más paraméterekkel generálni döntési fát és ezáltal új modellt.

Beállítások módosításával

Döntési fa



19. ábra: A döntési fa módosított beállításokkal és Gini indexszel

Performancia vektor

pontosság (**accuracy**)	99.46%
osztályozási hiba (**classification_error**)	0.54%
korreláció (**correlation**)	0.989
korreláció négyzet (**squared correlation**)	0.978

A **Decision Tree** operátor két paraméterét állítottam át, a **minimal leaf size**-ot 2-ről 4-re, illetve a **maximal depth**-t csökkentettem 20-ról 10-re. Ezzel a módosított beállítással sikerült elérnem, hogy a generált döntési fa nem csak, hogy kevésbé lett komplex, de ezzel szemben valamivel pontosabb eredményt is kapunk a performancia vektor táblázatában feltüntetett kapott értékek alapján.

Ábrajegyzék