Domača naloga 4

Neža Kržan, Tom Rupnik

1 Opis podatkov

Za analizo sva si izbrala podatkovni okvir z naslovom *Wine*, ki vsebuje 178 vrstic. Enoto oz. vrstico podatkov predstavlja posamezno vino. Podatki vsebujejo vrednosti kemijske analize treh različnih sort vina iz enake regije v Italiji. Analiza vsebuje vrednosti 13 različnih komponent (lastnosti) najdenih v vsaki izmed treh sort vina. Iz teh komponent želimo določiti kateri sorti pripada vino (stolpec *Class*). Komponente oz. stolpci v podatkovne okvirju so:

Tabela 1: Spremenljivke v podatkovnem okvirju.

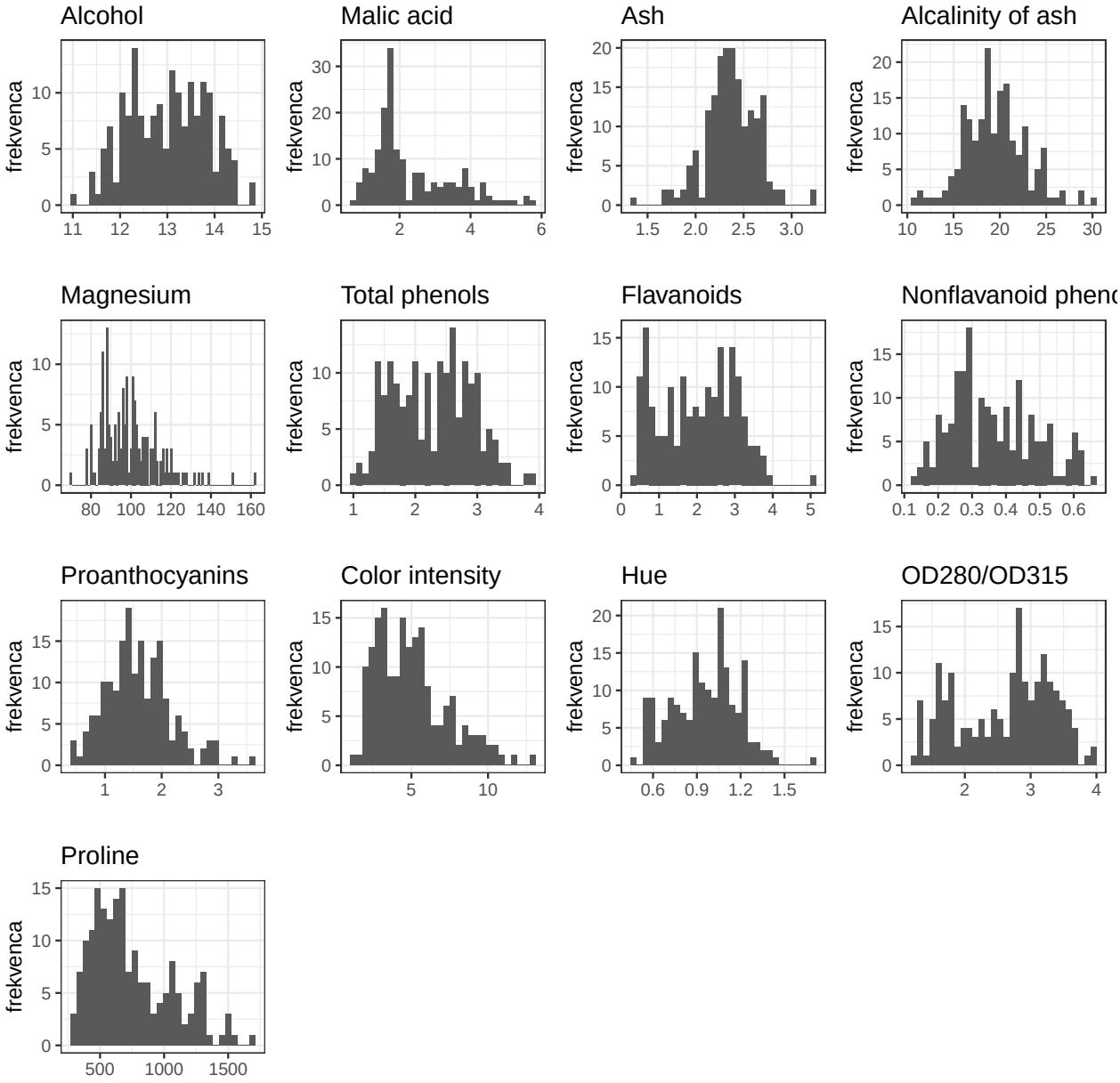
Ime	Pomen
Class	razred (uporabimo za analizo)
Alcohol	% alkohola
Malic_acid	jabolčna kislina
Ash	vsebnost pepela
Alcalinity_of_ash	alkalnost pepela
Magnesium	Magnezij
Total_phenols	skupaj fenolov
Flavanoids	rastlinske kemikalije(fenoli)
Nonflavanoid_phenols	fenolne spojine(za okus, aromo, barvo)
Proanthocyanins	kondenzirani tanini
Color_intensity	intenziteta barve
Hue	barvni odtenek vina
OD280_OD315	razmerje ocene vsebnosti fenolnih spojin
Proline	aminokislina

Poglejmo si osnovne opisne statistike spremenljivk.

Tabela 2: Opisna statistika spremenljivk.

Ime	min	max	povprecje
Alcohol	11.03	14.83	13.00
Malic_acid	0.74	5.80	2.34
Ash	1.36	3.23	2.37
Alcalinity_of_ash	10.60	30.00	19.49
Magnesium	70.00	162.00	99.74
Total_phenols	0.98	3.88	2.30
Flavanoids	0.34	5.08	2.03
Nonflavanoid_phenols	0.13	0.66	0.36
Proanthocyanins	0.41	3.58	1.59
Color_intensity	1.28	13.00	5.06
Hue	0.48	1.71	0.96
OD280_OD315	1.27	4.00	2.61
Proline	278.00	1680.00	746.89

Za boljšo predstavo o spremenljivkah, si poglejmo še njihove histograme.



Slika 1: Porazdelitev spremenljivk v podatkovnem okviru vin.

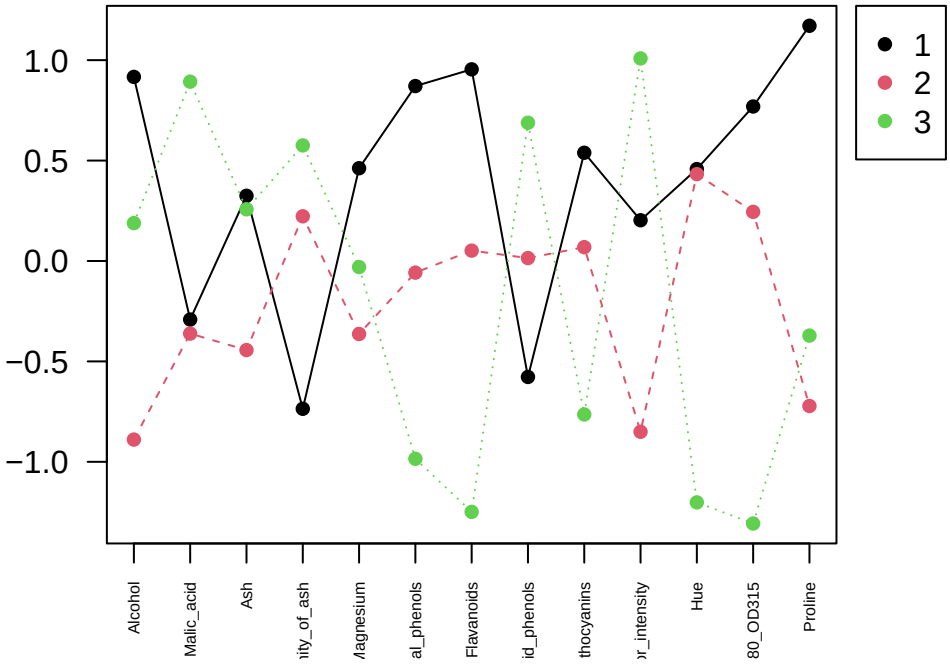
2 Skupine in povezanost

Za skupine torej imamo sorte vina, označene so z številkami 1, 2 in 3.

Iz zgornje tabele vidimo, da je največ enot prisotnih v skupini **2** (71) in najmanj v skupini **3** (48). Preostanek enot je v skupini **1** in sicer 59.

Tabela 3: Velikost skupin.

skupina	število.vin	delež...
1	59	33.15
2	71	39.89
3	48	26.97



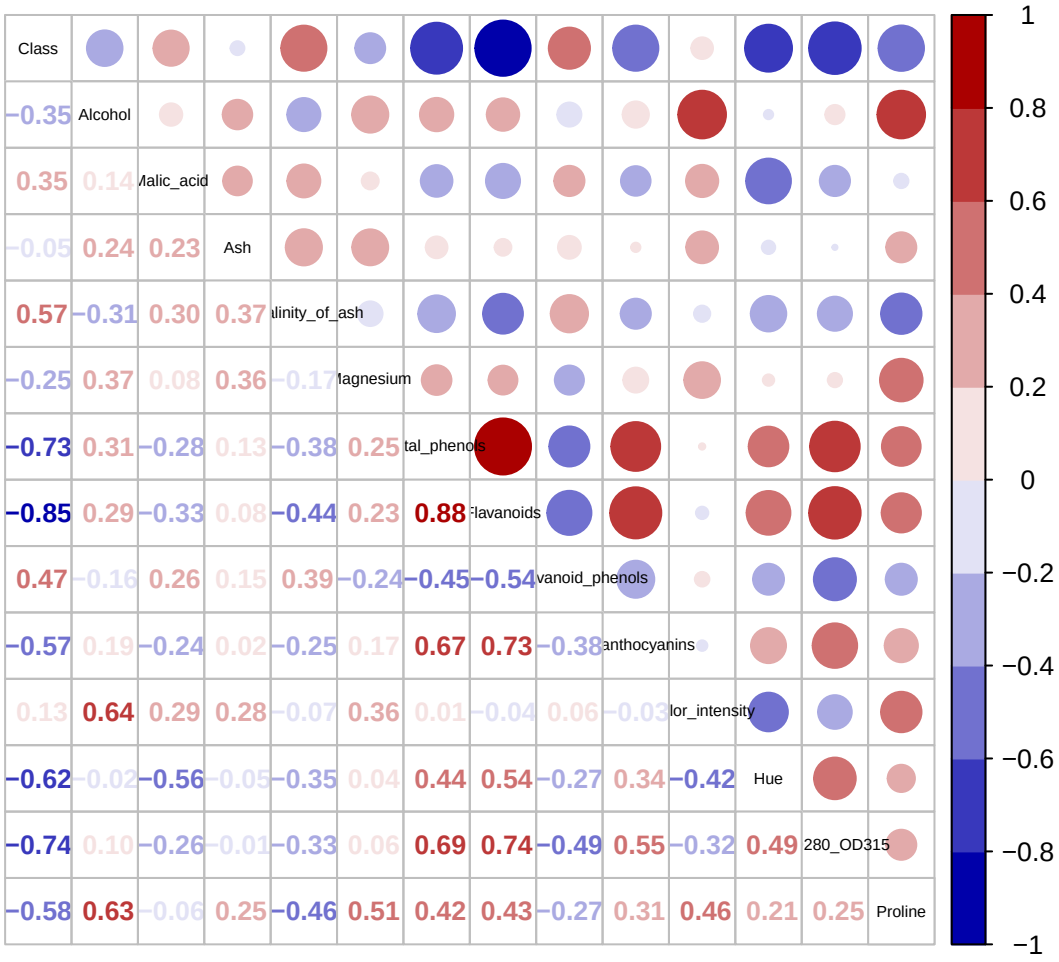
Slika 2: Povprečja neodvisnih spremenljivk po skupinah (standardizirane vrednosti).

Iz grafičnega prikaza lahko vidimo, da večina spremenljivk dobro loči med skupinami, je pa nekaj takih kjer to ne velja. Primer take spremenljivke je na primer *Hue* ali *Ash*. z grafa tudi vidimo, da posamezna spremenljivka ima različna povprečja v posamezni skupini(sorti vina). Če so razlike v povprečjih majhne, lahko pričakujemo, da bomo z diskriminatnimi funkcijami slabo ločevali med skupinami.

3 Predpostavke

Pri diskriminantni analizi imamo nekaj predpostavk, katerim morajo podatki zadoščati. V najinem podatkovnem okviru imava vsaj dve skupini - imava 3 skupine(3 sorte vina) **1**, **2**, **3**, in zgoraj smo videli v tabeli 3, da ima vsaka skupina več kot dve enoti. Imava tudi manjše število spremenljivk kot enot(zapisano je v začetku naloge, koliko enot in koliko spremenljivk imamo v podatkovnem okviru). Ker pa skupine niso enako velike, ne bova predpostavila enako velikih skupin.

Sedaj moramo preveriti predpostavko o odsotnosti multikolinearnosti, torej, da nobena od spremenljivk ni popolna linearna kombinacija ostalih. Težavo imamo, ko je katera od korelacij med spremenljivkama zelo blizu 1(npr. nad 0.95). V ta namen si najprej poglejmo korelacijsko matriko.



Slika 3: Korelacijska matrika spremenljivk v podatkovnem okviru.

Iz matrike vidimo, da nimamo večjih korelacij, kar bi lahko pomenilo, da multikolinearnost ni prisotna(sicer se le ta zelo pozno vidi, npr. šele, ko imamo koeficiente nad 0.95). Če sestavimo matriko XX', kjer imamo v matriki X podatke naših spremenljivk, je njen rang enak 13 in njena občutljivost oz. pogojenostno število je enako 31.202, kar je dokaj majhno, torej res lahko rečemo, da je predpostavka o odsotnosti multikolinearnosti izpolnjena.

Naslednja predpostavka pa je homogoenost, ki zahteva, da so vse kovariančne matrike skupin na populaciji enake. Pri prevrjanju te predpostavke si pomagamo z Box M-testom, kjer preverjamo H_0 : *kovariančne matrike po skupinah(po sortah vin) so enake.*

```
## ------------------------------------------------
##       MBox     Chi-sqr          df             P
## ------------------------------------------------
##   764.8065   684.2031         182        0.0000
## ------------------------------------------------
## Covariance matrices are significantly different.
```

Iz zgornjega zapisa vidimo, da je $p < 0.05$, torej zavračamo ničelno hipotezo, kar pomeni, da pogoj ni izpolnjen, ampak nadaljujemo z diskriminantno analizo.

4 Analiza

Izvedimo sedaj linearno diskriminantno analizo na standardiziranih podatkih.

S prvo diskriminantno funkcijo pojasnimo 68% razlik med povprečji spremenljivk, z drugo diskriminantno funkcijo pa 31% razlik med povprečji spremenljivk.

Poglejmo si še ali sta diskriminantni funkciji statistično značilni, kjer preverjamo domnevo H_0 : *povprečja diskriminantnih funkcij po skupinah(po sortah vin) enaka.*

Tabela 4: Statistična značilnost diskriminantnih funckij.

	WilksL	F	df1	df2	p
1 to 2	0.01934	77.61987	26	326	0
2 to 2	0.19499	56.42241	12	164	0

Pri stopnji značilnosti 0.05 lahko na podlagi $p - vrednost$ za obe diskriminantni funkciji zavrnemo domnevo H_0 in sprejememo, da so povprečja po sortah vin za obe diskriminantni funkciji različna.

Sedaj nas seveda zanima med katerimi skupinami bolje ločuje posamezna diskriminantna funkcija.

```
##                              LD1          LD2
## Alcohol              -0.3274906  0.707744750
## Malic_acid            0.1846135  0.341153776
## Ash                  -0.1012536  0.643569825
## Alcalinity_of_ash     0.5169574 -0.488847900
## Magnesium            -0.0309001 -0.006609312
## Total_phenols         0.3868085 -0.020160425
## Flavanoids           -1.6592953 -0.491436530
## Nonflavanoid_phenols -0.1861596 -0.202977648
## Proanthocyanins       0.0767491 -0.175764297
## Color_intensity       0.8231206  0.587061123
## Hue                  -0.1869798 -0.346430951
## OD280_OD315          -0.8218561  0.036340126
## Proline              -0.8474810  0.898426186
```

Tabela 5: Povprečja po skupinah

skupina	LD1	LD2
1	-3.422	1.692
2	-0.080	-2.473
3	4.325	1.578

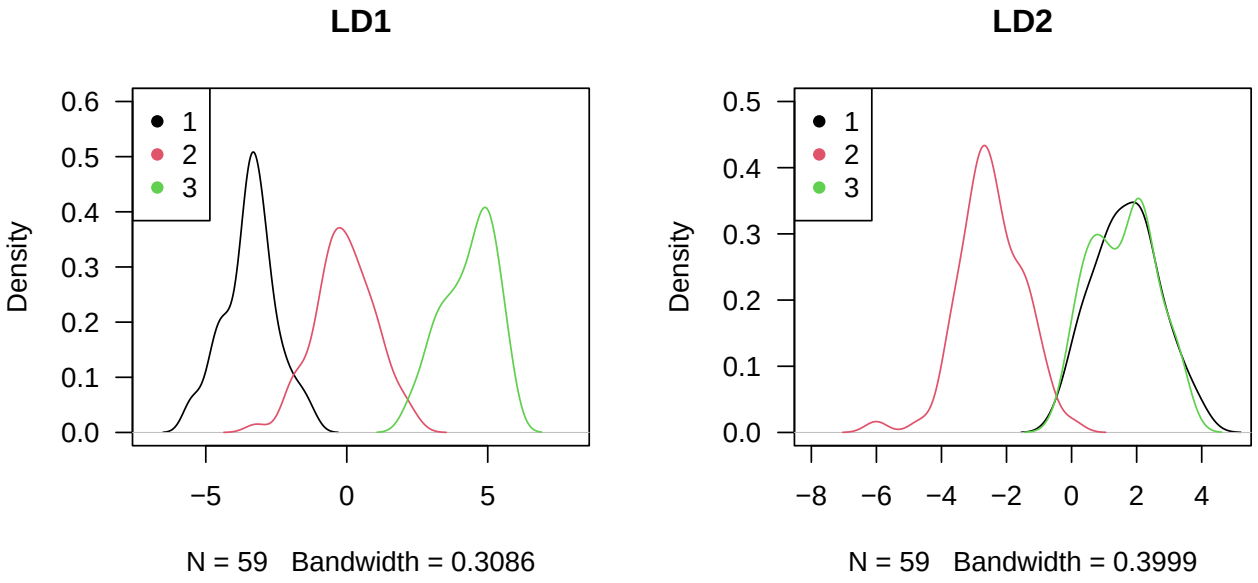
Prva linearna diskriminantna funkcija najbolj ločuje med sorto vina **1** in sorto vina **3**. Druga linearna diskriminantna funkcija pa najbolje ločuje med sorto vina **1** in **2**.

Poglejmo si spremenljivke, ki imajo največji vpliv na vrednosti diskriminantne funkcije, ker so z posamezno funkcijo močno korelirane. Torej spremenjivke, ki najbolj ločujejo med skupinami.

V največji meri se nam splača govoriti o rastlinskih kemikalijah(fenoli) *Flavanoids*, prav tako pa tudi o intenziteti barve(*Color_intensity*), razmerju ocene vsebnosti fenolnih spojin(*OD280_OD315*), aminokislinah(*Proline*) pri prvi diskriminantni funkciji.

Prav tako se splača govoriti o aminokislinah(*Proline*), % alkohola(*Alcohol*), vsebovanost pepela(*Ash*) pri drugi diskriminantni funkciji. Šibko koreliranost z drugo diskriminantno funkcijo pa imajo tudi intenziteta barve(*Color_intensity*) in spremenljivka *Flavanoids*.

Poglejmo si še grafično kako je z ločevanjem po skupinah. Na spodnjem grafu vidimo, da se pri LD1 porazdelitve skupin ne prekrivajo tako zelo, s tem ko pri LD2, ne ločujemo med skupinama **1** in **3**, sorta vina **1** pa ima težišče precej v levo. Vidimo, da ne glede na to kaj smo opisali, da dejanska sposobnost ločevanja po sortah pri drugi funkciji zelo mahna, pri prvi pa bi to bilo možno, ampak druga funkcija vseeno dobro ločuje med sorto 2 od sort 1 in 3, kar je tudi dobro.



Slika 4: Grafični prikaz porazdelitev vrednosti po skupinah.

5 Kvaliteta ocenjenega modela

Če si ponovno ogledamo tabelo iz sklopa 2.

Tabela 6: Velikost skupin.

skupina	število.vin	delež...
1	59	33.15
2	71	39.89
3	48	26.97

V našem primeru imamo največ enot v skupini **2** (71). V primeru, da bi vse enote razvrstili v največjo skupino, bi bil delež pravilno razvrščenih enot enak 39.89%.

Tabela 7: Klasifikacijska tabela.

	1	2	3
1	59	0	0
2	0	71	0
3	0	0	48

Že iz tabele je razvidno, da vse enote razvrstimo pravilno (izven diagonale so vrednosti 0), torej je delež pravilno razvrščenih enak 1. Prav tako je popravljen Randlov indeks enak 1.

Preverimo še kakšne vrednosti bi dobili v primeru uporabe navzkrižnega preverjanja(CV), ker smo se prej učili na istih podatkih kot smo potem model tudi testirali, zato je možnost, da precenjujemo. Želimo pa si

bolj verodostojne ocene oziroma nepristranske ocene za delež pravilno razvrščenih enot.

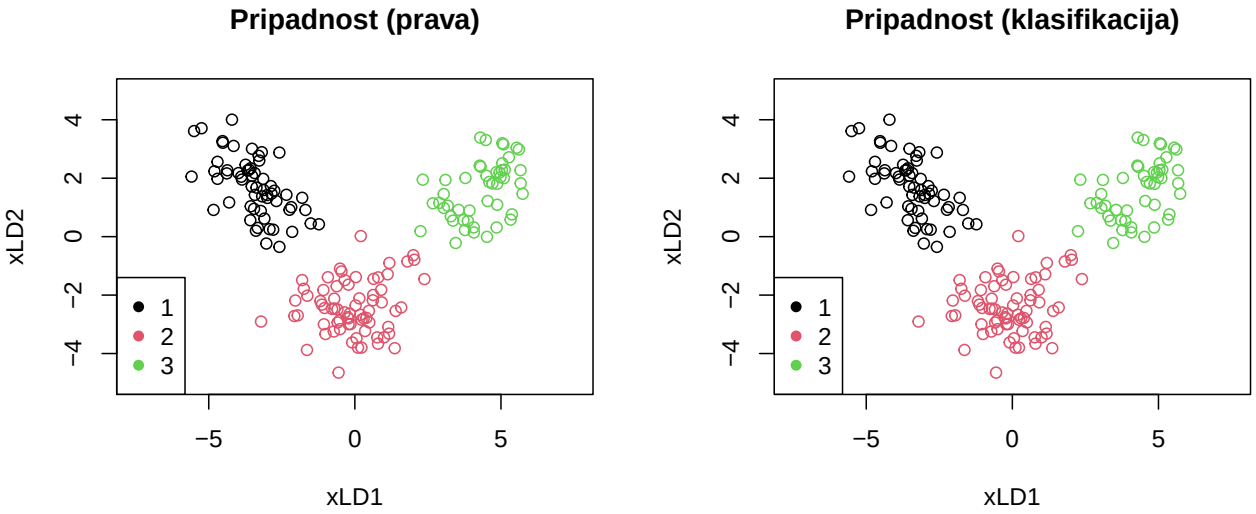
Tabela 8: Klasifikacijska tabela (CV).

	1	2	3
1	59	0	0
2	1	69	1
3	0	0	48

Iz tabele je vidno, da tokrat dobimo nekoliko slabše rezultate. V primerjavi s prejšnjo razvrtitvijo, smo tokrat dve enoti iz skupine **2** napačno razvrstili. Eno enoto smo uvrstili v skupino **1** in eno v skupino **3**. Nepristranska ocena za delež pravilno razvrščenih enot je enaka 0.989, popravljen Randov indeks pa 0.965.

6 Grafični prikaz

Naprej si oglejmo prikaz enot glede na pravo pripadnost sorti vina, nato pa še pripadnost posamezni skupini glede na klasifikacijo po LDA.



Slika 5: Prikaz enot glede na pravo pripadnost sorti vina(levo) in pripadnost posamezni skupini glede na klasifikacijo po LDA(desno).

Že iz levega grafa se vidi, da so sorte vina med seboj dobro ločene oz. se med seboj ne prekrivajo. Zato ni presenečenje, da je taka situacija tudi v primeru klasifikacije po LDA. Skupine se lepo ločujejo glede na prvo diskriminantno funkcijo(gledamo vodoravno) in sicer tudi glede na drugo(gledamo navpično), ampak slabše. Kot smo že na začetku videli s prvo diskriminantno funkcijo pojasnimo 68% razlik med povprečji spremenljivk, z drugo diskriminantno funkcijo pa 31% razlik med povprečji spremenljivk. To pa bi se delno dalo razbrati tudi iz zgornjih grafov.

Grafa sta med seboj identična zato lahko zaključimo, da diskriminantna analiza odlično ločuje med skupinami.