Izveštaj
Analiza podataka - Žitarice

Ljiljana Popović, EE72/2014, biskvitcokolada@gmail.com

I. Baza podataka

Baza podataka pod nazivom „Žitarice“ predstavlja uzan skup podataka o osnovnim karakteristikama koje se vezuju za neke vrste žitarica. U bazi se nalazi 210 različitih uzoraka sa po 7 atributa i po jednom oznakom klase. Atributi su sledeći: Površina_A, Okolina_P, Kompaktnost_C, Dužina_jezgra, Širina_jezgra, Koeficijent_asimetrije, Dužina_useka_u_jezgru i svi su numerički. Oznaka klase je jedina kategorička varijabla. Ovakva baza podataka može poslužiti, na primer, za svrhe razvrstavanja zrna po kvalitativnim grupama.

II. Analiza podataka

Baza je podeljena u 3 klase prema oznaci labele. Nakon provere, utvrđeno je da nema nedostajućih podataka, što govori da ne treba brisati neke podatke i da je baza potpuna.

Zatim je utvrđeno da je najveći interkvartilni opseg kod šestog obeležja, što ukazuje na to da se kod atributa Koeficijent_asimetrije najveći broj podataka nalazi u opsegu od 25% do 75%. Najmanji dinamički i interkvartilni opseg ima obeležje Kompaktnost_C, za sve tri klase, što znači da su svi podaci tog atributa smešteni u uskom opsegu vrednosti i ne rasipaju se.

Pošto su sve date vrednosti podataka necelobrojne, potrebno je raditi metodu histograma kako bismo utvrdili da li za neke od atributa postoje vrednosti koje se pojavljuju u više od 10% uzoraka. Dobijeno je da postoje, i to u prvoj klasi kod tri, u drugoj klasi kod dva atributa, a u trećoj kod četiri.

Potom je pomoću boxplot dijagrama primećeno da gotovo svako obeležje sadrži outlier vrednosti. To su vrednosti koje odskaču od drugih vrednosti neke karakteristike. Primećeno je da je najveći broj kod šestog obeležja u trećoj klasi i iznosi 9, što znači da će ove vrednosti kod obeležja Koeficijent_asimetrije uticati da median i srednja vrednost ne budu isti.

III. Normalizovanje podataka

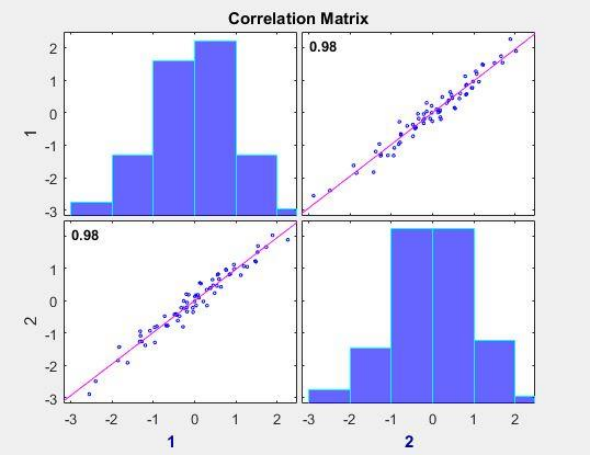
Normalizacija podataka je postupak transformacije koji se primenjuje nad podacima ukoliko njihova raspodela odstupa statistički značajno od normalne raspodele.

Nakon primene postupka *Z-normalizacije* nad podacima, svako obeležje ima nultu srednju vrednost i jediničnu standardnu devijaciju. Ovakav pristup nam može, na primer, olakšati rad sa podacima ili poslužiti za neku dublju analizu.

IV. Vizuelizacija podataka

Radi lakšeg razumevanja problema, podaci se prikazuju na različitim dijagramima, odakle se sve mnogo lakše zaključuje.

Potrebno je nad normalizovanim podacima utvrditi koja dva atributa imaju najveću korelaciju. Ovo je utvrđeno na osnovu korelacione matrice koja iscrtava međusobne zavisnosti za sva obeležja i daje vrednost njihove korelacije. Zaključeno je da su to prvi (Površina_A) i drugi (Okolina_P) atribut, sa korelacijom od čak **0.98** za prve dve klase i **0.91** za treću, što znači da oni najviše zavise jedan od drugog i da jedan od njih ne bi bitno uticao na analizu ukoliko bi se uklonio iz seta podataka. Na slici 1 je ilustrovana njihova zavisnost za prvu klasu pomoću scatterplot dijagrama.

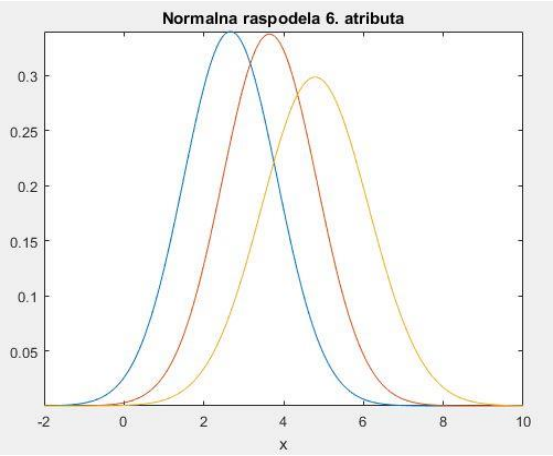


Sl. 1. Korelaciona matrica i zavisnost Površine_A od Okoline_P.

Ostvarenom korelacijom podaci skoro u potpunosti leže na pravoj $y=x$, čija bi korelacija iznosila 1.

Zatim je neophodno svaki od atributa modelovati normalnom raspodelom kako bismo uočili njihove karakteristike u svakoj od tri klase. S obzirom da ovde nismo uzeli u obzir normalizovane podatke, srednja vrednost ovih raspodela neće iznostiti 0, a standardna devijacija 1.

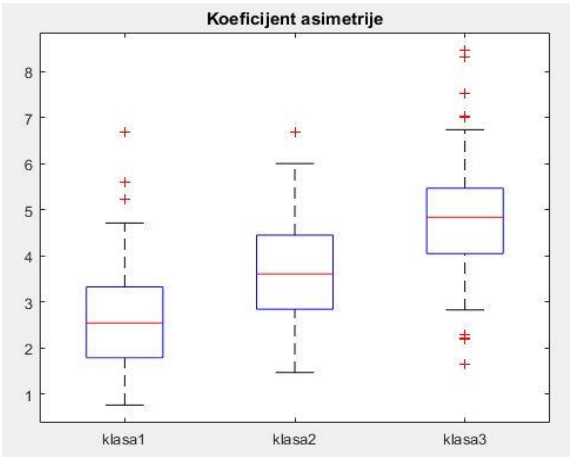
Za vizuelizaciju je izabran šesti atribut. Na slici 2 je prikazana njegova normalna raspodela za sve tri klase na istom grafiku.



Sl. 2. Normalna raspodela 6. obeležja za sve tri klase.

Primećuje se da obeležje Koeficijent_asimetrije ima približnu srednju vrednost kod sve tri klase i samim tim se ne bi moglo lako klasifikovati na osnovu ovog obeležja.

Potom je napravljeno poređenje za jedan, na slučajan način odabran atribut i vizuelizovani su rezultati. Izabrano je takođe šesto obeležje da bi se prikazali rezultati boxplot dijagrama i povezali sa odgovarajućim normalnim raspodelama prikazanim na slici 2. Slika 3 ilustruje pomenute dijagrame za šesto obeležje.



Sl. 3. Boxplot vizuelizacija 6. obeležja

Ovde se primećuje da su opsezi u kojima su smešteni podaci takođe približni, kao i na slici 2. Isto tako, primećuje se i da obeležje Koeficijent_asimetrije ima nekoliko outlier vrednosti koje neće uticati na median već samo na srednju vrednost ovog obeležja.

V. LINEARNA REGRESIJA

Linearna regresija je najjednostavniji pristup nadgledanom učenju i tehnika za predviđanje kontinualne izlazne promenljive.

U zavisnosti od odabranog skupa podataka pravi se model linearne regresije koji predviđa jedan od numeričkih atributa. Za ovo je izabran prvi atribut-Površina_A jer je za njega primećeno da ima najbolju korelaciju sa ostalim atributima, te je kao takav uzet kao najbolji primer za predikciju.

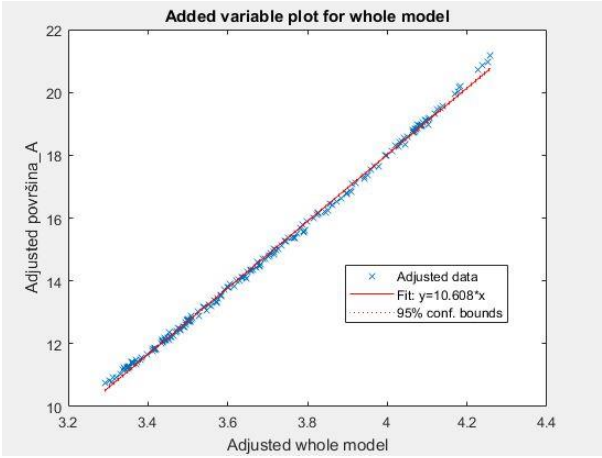
Od ukupnih 210 uzoraka, 10% je nasumično odabrano za test skup, a preostalih 90% je iskorišćeno za pravljenje modela.

Za odabir obeležja koja su važna korišćen je *Metod selekcije unapred* i RSS (Residual Sum of Squares) mera kao kriterijum. Proverom koji od atributa najviše smanjuju RSS dobijen je model liearne regresije sa 5 obeležja koji ima sledeću formulaciju:

površina_A~1+okolina_P+kompaktnost+dužina_useka_u_jezgru+širina_jezgra+dužina_jezgra

Drugi način pravljenja modela bio bi *Metod selekcije unazad* i p vrednost kao mera.

Na slici 4 je dat grafički prikaz dobijenog modela linearne regresije.



Sl. 4. Grafički prikaz modela linearne regresije

VI. PREDIKCIJA OBELEŽJA

Dobijeni model nadalje koristimo za predviđanje izabranog obeležja. Predikcija se vrši pomoću test skupa podataka i kao konačna mera uspešnosti data je MSE (Mean Squared Error)- suma kvadrata razlike originalnih test podataka i podataka koje je predvideo dati model linearne regresije i ona iznosi **0.0124**. Smatraćeno da je ovaj model „dobar" jer se greška nalazi u prihvatljivo malom opsegu vrednosti i odstupa za 0.0001 od ukupne greške modela. U tabeli 1 su date vrednosti estimiranih koeficijenata dobijenog modela.

TABELA 1: VREDNOSTI ESTIMIRANIH KOEFICIJENATA MODELA.

	Estimate	SE	tStat	pValue
(Intercept)	-24.414	1.0571	-23.095	4.0084e-56
okolina_P	1.83	0.072424	25.268	1.4249e-61
kompaktnost	10.406	1.246	8.3516	1.6292e-14
dužina_useka_u_jezgru	0.29275	0.049072	5.9658	1.2327e-08
širina_jezgra	0.88267	0.20295	4.3492	2.2653e-05
dužina_jezgra	-0.16123	0.13006	-1.2397	0.21666