Analiza podataka

Odabrati jedan od ponuđenih skupova podataka i izvršiti analizu:

1. Ukoliko postoje nedostajući podaci, izbaciti ih iz razmatranja (ukoliko nije naglašeno drugačije kod detaljnijih objašnjenja baze).
2. Utvrditi dinamički i interkvartilini opseg za svaki od atributa i prokomentarisati zapažanja.
3. Utvrditi da li za neke od atributa postoji vrednost koja se javlja u više od 10% uzoraka. UKOLIKO JE ATRIBUT SA DISKRETNIM VREDNOSTIMA, OVO SE MOŽE URADITI PREKO MODUSA, A UKOLIKO JE SA KONTINUALNIM (NECELOBROJNIM) VREDNOSTIMA, POTREBNO JE NAPRAVITI HISTOGRAM PA TAKO DOĆI DO REZULTATA (ODNOSNO NPR. UMESTO DA SE PROVERI KOLIKO SE PUTA JAVLJA VREDNOST 3.54, PROVERAVA SE KOLIKO SE PUTA JAVLJAJU VREDNOSTI IZ INTERVALA 3.4 DO 3.6). REZOLUCIJA HISTOGRAMA JE PROIZVOLJNA.
4. Nacrtati boxplot za svaki od atributa i utvrditi da li sadrži outliere ukoliko su outlieri definisani kao vrednosti veće od $q_3+w*(q_3-q_1)$ i manje od $q_1+w*(q_3-q_1)$, gde je q_3 gornji kvartil, q_1 je donji kvartil, a w treba da bude jednako 1.
5. Potom nad podacima izvršiti z-normalizaciju, a onda uraditi ~~sledeće analize~~ ZADATKE 6 I 7.
6. Utvrditi koja dva atributa imaju najveću korelaciju.
7. Nacrtati *scatterplot* za ta dva atributa kao i pravu koja najbolje opisuje njihovu zavisnost.
8. Svaki od atributa modelovati normalnom raspodelom (NAD ORIGINALNIM PODACIMA).
   Pomenute analize uraditi za sve klase posebno. ODNOSI SE NA ZADATKE 1 DO 8.
   Potom napraviti poređenje za jedan od atributa PO ŽELJI na sledeći način:
9. Odabrati jedan od atributa i na istom grafiku za sve klase nacrtati boxplot koji ga opisuje i prokomentarisati zapažanja (nad originalnim podacima). OVO JE VEĆ URADJENO U ZADATKU 4, ALI SA CILJEM PROVERE POSTOJANJA OUTLIERA, A OVDE RADIMO DETALJNIJE UPOREDNO POREĐENJE ISTOG ATRIBUTA ZA RAZLIČITE KLASE. SLIKA MOŽE DA SADRŽI SUBPLOTOVE ZA SVAKI OD BOXPLOTOVA.
10. Na drugom grafiku nacrtati normalne raspodele kojima su modelovani u svakoj od klasa i prokomentarisati zapažanja (nad ~~normalizovanim~~ORIGINALNIM podacima). OVO JE URADJENO U ZADATKU 8, SAMO TREBA SAD NACRTATI I IZVRŠITI UPOREDNU ANALIZU ZA SVE KLASE.

U zavisnosti od odabranog skupa podataka, napraviti model linearne regresije koji predviđa jedan od numeričkih atributa.

1. Potrebno je 10% nasumično izabranih uzoraka ostaviti kao test skup, a preostalih 90% koristiti za pravljenje modela.
2. Koristeći RSS (Residual Sum of Squares) meru kao kriterijum i metod selekcije obeležja unapred ili p-vrednost i metod selekcije obeležja unazad, odrediti najbolji model linearne regresije sa 5 obeležja (uz 5 različitih obeležja mogu postojati njihovi kvadrati ili kombinacije).
3. Izvršiti predikciju nad test skupom i kao konačnu meru uspešnosti modela dati MSE (sumu kvadrata razlike originalnih test podataka i podataka koje je predvideo tvoj konačni model linearne regresije.

Bol u kičmi

Atributi: pelvic_incidence (karlična incidenca), pelvic_tilt (karlična nakrivljenost), lumbar_lordosis_angle (ugao lumbalne lordoze), sacral_slope (nagib sakruma), pelvic_radius (prečnik karlice), degree_spondylolisthesis (stepen spondilolisteze), pelvic_slope (nagib karlice), direct_tilt (direktna nakrivljenost), thoracic_slope (nagib toraksa), cervical_tilt (nakrivljenost vratne kičme), sacrum_angle (ugao sakruma), scoliosis_slope (nagib skolioze), Attribute class {Abnormal, Normal}.

Za analizu koristiti sve atribute. Podrazumeva se postojanje 2 klase – postojanje bola (*Abnormal*) i odsustvo bola (*Normal*).

Dijabetes

Atributi: Broj trudnoća, Glukoza, Krvni pritisak, Debljina kože, Insulin, BMI, *DiabetesPedigreeFunction*, Godine, Klasa. *Missing values* su one gde je 0 a nije logično da je 0 (npr. pritisak, dok je kod broja trudnoća vrednost 0 opravdana): ako fali više od 20% podataka, tu varijablu izbaciti iz razmatranja; ako fali do 20% podataka, tu varijablu popuniti srednjom vrednošću izračunatom bez nedostajućih podataka.

Analizirati sve atribute. Podrazumeva se postojanje 2 klase – ima dijabetes (*Klasa=1)* i nema dijabetes (*Klasa=0)*. Npr. predvideti broj trudnoća.

Napuštanje posla

Atributi: nivo zadovoljstva, poslednja evaluacija, broj projekata, prosečan broj sati mesečno, vreme provedeno u kompaniji, problem na poslu, napustio/la ili ne, unapređenje u poslednjih 5 godina, departman, plata.

Analizirati sve numeričke atribute (prvih 5). Klasama smatrati departmane, a u okviru njih posebno izanalizirati one koji su napustili posao i one koji nisu. Potom smatrati visinu plate klasom, a u okviru te tri klase izanalizirati one koji su napustili posao i one koji nisu. Zanemariti podatak o unapređenju u proteklih 5 godina. Npr. predvideti prosečan broj sati mesečno.

Žitarice

Atributi: površina A, okolina P, kompaktnost C = 4*pi*A*P^2, dužina jezgra, širina jezgra, koeficijent asimetrije, dužina useka u jezgru, klasa.

Analizirati sve varijable za svaku od klasa. Npr. predvideti dužinu useka u jezgru.

Vina

Atributi: Vinarija , Alkohol, *Malic* kiselina, Suvi ostatak, Alkalnost suvog ostatka, Magnezijum, Ukupno fenola, Flavanoidi, Neflavoidni fenoli, *Proanthocyanins*, Intenzitet boje, Nijansa, OD280/OD315 razređenog vina, Proline.

Analizirati 10 atributa za vina svake od 3 vinarije. Npr. predvideti intenzitet boje.