**Prvi deo projektnog zadatka – analiza baze podataka**

**Autori: Aleksa Simeunović IN 43/2020,**

**Nemanja Ranitović IN 8/2020**

**Baza podataka: Smoking and drinking dataset with body signal**

1. **Definisati u 2-3 rečenice problem koji će se u projektu rešavati.**

Rešavamo klasifikacioni problem svrstavanja ispitanika u pušače ili alkoholičare, zavisno od vrednosti parametara koje njihova krvna slika i dodatna lična obeležja sadrže. Bazu je popunila Nacionalna zdravstveno osiguravajuća institucija u Koreji. Sastoji se od 24 obeležja (pola, godina, visine, težine, holesterola i hemoglobina u krvi, triglicerida, proteina u urinu, parametara krvnog pritiska. . . ), koja predstavljaju razne telesne markere po kojima može da se utvrdi da li su ispitanici bili pušači, da li i dalje puše kao i da li konzumiraju alkohol.

1. **Koliko ima uzoraka u bazi?**

U bazi ima 991.345 uzoraka.

1. **Jednom rečenicom objasniti šta predstavlja jedan uzorak u konkretnoj bazi.**

Jedan uzorak predstavlja jednog ispitanika Nacionalne zdravstveno-osiguravajuće institucije u Koreji. Obeležja opisuju krvnu sliku, kao i dodatne lične podatke ispitanika.

1. **Koliko ima obeležja u bazi?**

U bazi podatka ima 24 obeležja.

1. **Navesti sva obeležja (jasnim imenom na srpskom ili opisno, nebitan je naziv u samoj bazi).**

|  |  |
| --- | --- |
| Sex | Pol |
| Age | Godine |
| Height | Visina |
| Weight | Težina |
| Wasitline | Širina struka |
| Sight Left | Vid na levo oko (Dioptrija) |
| Sight Right | Vid na desno oko (Dioptrija) |
| Hear Left | Sluh na levo uho |
| Hear Right | Sluh na desno uho |
| SBP | Krvni pritisak u arterijama tokom otkucaja |
| DBP | Krvni pritisak u arterijama između otkucaja |
| BLDS | Nivo glukoze u krvi (mg/dL) |
| tot\_chole | Količina holesterola (mg/dL) |
| HDL\_chole | Količina HDL holesterola (mg/dL) |
| LDL\_chole | Količina LDL holesterola (mg/dL) |
| Triglyceride | Trigliceridi u krvi (mg/dL) |
| Hemoglobin | Hemoglobin u krvi (g/dL) |
| Urine\_protein | Nivo proteina u urinu 1(-), 2(+/-), 3(+1), 4(+2), 5(+3), 6(+4) |
| serum\_creatinine | Kreatinin u krvi |
| SGOT\_AST | Glutamate-oxaloacetate transaminase |
| SGOT\_ALT | Alanine transaminase |
| gamma\_GTP | y-glutamyl transpeptidase |
| SMK\_stat\_type\_cd | Pušački staž(1-nikad, 2-pušio ali prestao, 3 – puši) |
| DRK\_YN | Alkoholičar(Y-Pije, N-Ne pije) |

1. **Koliko ima numeričkih obeležja?**

Age, Height, Weight, Waistline, Sight Left, Sight Right, SBP, DBP, BLDS, tot\_chole, HDL\_chole, LDL\_chole, triglyceride, hemoglobin, serum\_creatinine, SGOT\_AST, SGOT\_ALT, gamma\_GPT

1. **Ako ima kategoričkih obeležja, navesti ono koje ima najmanji broj kategorija I koje su, I navesti ono koje ima najveći broj kategorija i koliko ih je.**

Najmanji broj kategorija: hear\_left i hear\_right sa po dve kategorije. (0-normal, 1-abnormal)

Najveći broj kategorija: urine\_protein, 6 kategorija

1. **Ako se rešava klasifikacioni problem: navesti procentualno koliko ima uzoraka u svakoj od klasa.**

**Pije** – 495488 (49.98%)

**Ne pije** – 495859 (50.02%)

**Pusi** – 213954 (21.58%)

**Prestao** – 174951 (17.64%)

**Ne** **pusi** – 602441 (60.77%)

1. **Da li postoje obeležja u bazi koja smatraš da treba izbaciti iz baze? Koja su to i zašto smatraš da ih treba izbaciti?**

Smatramo da ne treba izbacivati obeležja, zato što su osnovni lični podaci bitni, a krvna slika sadrži parametre o kojima nismo dovoljno stručni da sudimo, a pretpostavka je da se nalaze u bazi podataka zato što su relevantni za istraživanje.

1. **Da li u bazi ima nedostajućih vrednosti? Ako ima, navesti za svako od obeležja koliko vrednosti mu procentualno nedostaje?**

Nema. Svako obeležje sadrži 0 nedostajućih vrednosti, čak i pre čišćenja podataka.

1. **Da li ima nevalidnih vrednosti u bazi? Ako ima, navesti za svako od obeležja koje su vrednosti?**

Vrednosti koje ćemo navesti ispod su primeri autlajera:   
1. Waistline – 999

2. Dioptrija levog oka – 9.9

3. Dioptrija desnog oka – 9.9

4. SBP – 273

5. BLDS – 852

6. tot\_chole – 2344

7. HDL\_chole – 8110

8. LDL\_chole – 5119

9. triglyceride – 9490

10. hemoglobine – 25

11. serum\_creatinine – 98

*Za svaku od vrednosti planiramo da nađemo neki validan opseg u konsultaciji sa izvorima na internetu ili nekim iz struke i da odstranimo sve nepoželjne vrednosti.*

1. **Ako ima nedostajućih i/ili nevalidnih vrednosti u bazi, za svako od obeležja navesti kako će problem biti rešen.**

Nedostajućih vrednosti nema, a nevalidne će biti zamenjene prosekom ili medijanom validnih vrednosti.

1. **Kada je završeno izbacivanje, dopuna, i drugo, navesti koliko je u sređenoj bazi ostalo uzoraka, a koliko obeležja.**

U bazi ima 991.345 uzoraka.

U bazi ima 24 obeležja.

1. **Da li neka od obeležja sadrže autlajere? Nabvesti koja obeležja ih sadrže.**

Visina, težina, dioptrija levog oka, dioptrija desnog oka, Krvni pritisak u arterijama tokom otkucaja, Krvni pritisak u arterijama između otkucaja, nivo glukoze u krvi, količina holesterola, količina HDL holesterola, količina LDL holesterola, trigliceridi u krvi, hemoglobin u krvi, kreatinin u krvi, Glutamate-oxaloacetate transaminase. Alanine transaminase, y-glutamyl transpeptidase.

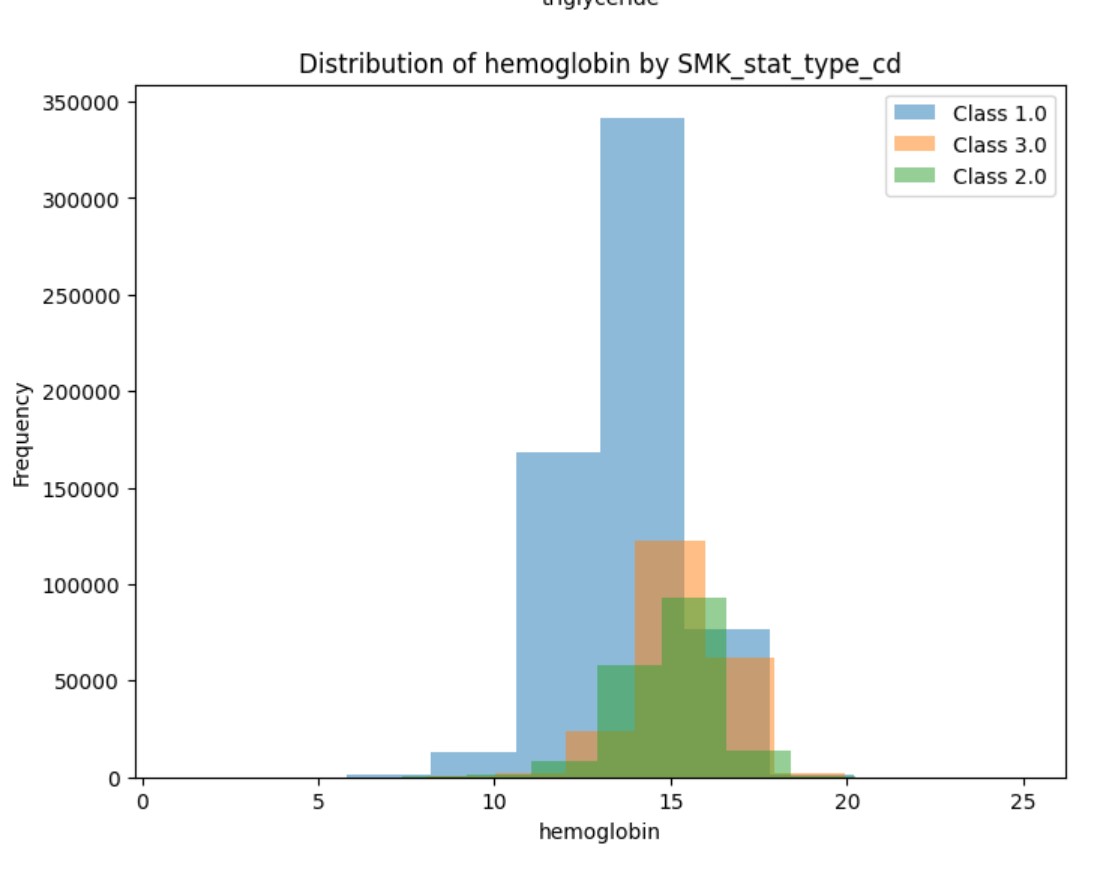
1. **Da li postoje parovi obeležja korelisani više od 0.7? Navesti takve parove obeležja.**

sex - height : Correlation = 0.7227744544957324

SBP - DBP : Correlation = 0.7411308762479987

tot\_chole - LDL\_chole : Correlation = 0.8773672716973919

1. **Ako se rešava klasifikacioni problem: iscrtavanjem histograma reći koje se obeležje izdvaja kao najdiskriminatornije (najbolje razdvaja klase)?**



Nema obeležja koje je preterano dobro razdvojilo klase za pušače, ali najbliži tome je bio **hemoglobin u krvi**, što i logički ima smisla.

Takođe, situacija za razdvajanje klasa alkoholičara nije mnogo bolja jer nemamo obeležje koje dovoljno dobro razdvaja te klase, ali **najmanje dioptrija desnog oka**.

