Izveštaj
Analiza podataka – dijabetes

I. Opis baze podataka

Ovaj izveštaj se bavi analizom podataka vezanih za dijabetes. Baza sadrži podatke o 768 ispitanika, od kojih je 500 klasifikovano kao zdravo (nema dijabetes), dok je 268 klasifikovano kao bolesno (ima dijabetes). Kod ispitanika je posmatrano 8 atributa : broj trudnoća, nivo glukoze, krvni pritisak, debljina kože, nivo insulina, indeks telesne mase (BMI), funkcija verovatnoće dijabetesa (Diabetes Pedigree Function – DPF) i godine starosti. Svi atributi su numerički.

Analiziranje ovih podataka i uočavanje eventualne pravilnosti i zavisnosti između atributa bi moglo biti dobra teorijska osnova za projektovanje klasifikatora koji će osobu nepoznatog zdravstvenog stanja svrstati u grupu zdravih ili obolelih od dijabetesa. Takav klasifikator bi odlučivao na osnovu vrednosti pomenutih atributa kod određene osobe i mogao bi biti koristan lekaru u procesu postavljanja dijagnoze.

II. Analiza podataka

Prilikom analiziranja baze izostavljeni su ispitanici kod kojih je nedostajalo više od 20% podataka (više od jednog atributa), kako bi rezultati bili tačniji. Kod ispitanika sa manje od 20% nedostajućih podataka, kao vrednost atributa koji nedostaje uzeta je srednja vrednost tog atributa izračunata bez nedostajućih podataka. Zbog ove korekcije se baza svela na 534 ispitanika, 357 zdravih i 177 bolesnih.

A. Dinamički i interkvartilni opseg

Tabela 1 : Dinamički opseg atributa kod zdravih osoba

Atribut	**Dinamički opseg**
Broj trudnoća	13
Nivo glukoze	141
Krvni pritisak	86
Debljina kože	53
Nivo insulina	729
Indeks telesne mase (BMI)	39,1
Funkcija verovatnoće dijabetesa (DPF)	2,244
Godine starosti	60

Tabela 2 : Dinamički opseg atributa kod bolesnih osoba

Atribut	**Dinamički opseg**
Broj trudnoća	17
Nivo glukoze	121
Krvni pritisak	80
Debljina kože	92
Nivo insulina	832
Indeks telesne mase (BMI)	44,2
Funkcija verovatnoće dijabetesa (DPF)	2,293
Godine starosti	49

Tabela 3 : Interkvartilni opseg (IQR opseg) kod zdravih i bolesnih osoba

Atribut	**IQR opseg – zdrave osobe**	**IQR opseg – bolesne osobe**
Broj trudnoća	3	7
Nivo glukoze	30,25	53,25
Krvni pritisak	16	16
Debljina kože	14	12,25
Nivo insulina	59	64
BMI	9,6	6,95
DPF	0,3455	0,4648
Godine starosti	10,25	17,25

Iz ovih podataka se može zaključiti da dinamički opseg nije potpuno merodavan za procenu intervala koji vrednosti nekog atributa zauzimaju. Bolje informacije daje interkvartilni opseg. Na primer, dinamički opseg vrednosti nivoa insulina kod bolesnih osoba je 832, ali se 50% tih vrednosti zapravo nalazi u opsegu 64. Sa nivoom glukoze je nešto drugačiji slučaj – sve vrednosti se nalaze u opsegu 121, a 50% vrednosti se nalazi u opsegu 53,25, što znači da su vrednosti nivoa glukoze ravnomerno raspoređene po celom opsegu.

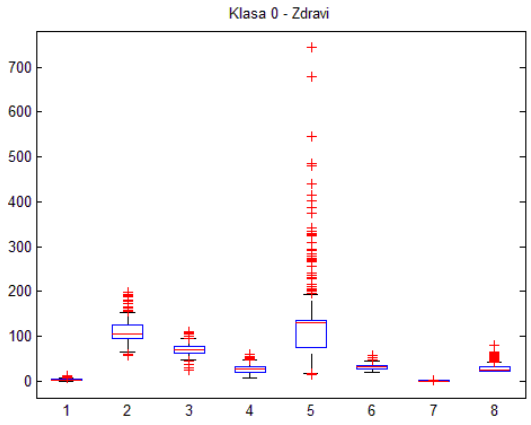
B. Vrednosti koje se često pojavljuju

Kod određenih atributa postoje vrednosti koje se javljaju u više od 10% slučajeva. Te vrednosti su određene pomoću histograma za kontinualne atribute, odnosno, pomoću funkcije *mode* za diskretne atribute.

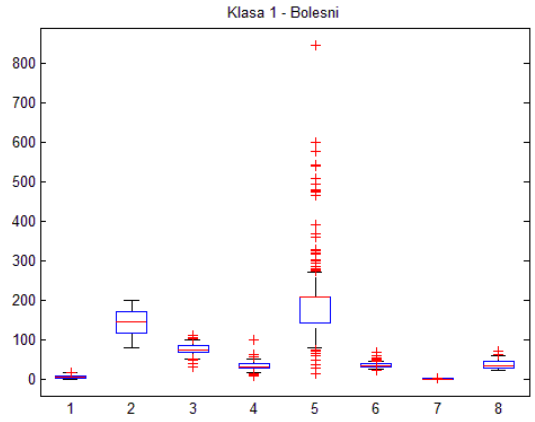
TABELA 4 : VREDNOSTI KOJE SE POJAVLJUJU U VIŠE OD 10% SLUČAJEVA

Atribut	Česta vrednost
Broj trudnoća – zdrave osobe	0
	1
	2
Broj trudnoća – bolesne osobe	0
	1
	3
Nivo insulina – zdrave osobe	130,2879
Nivo insulina – bolesne osobe	206,8462
BMI – bolesne osobe	Opseg 33,5 – 34,5
DPF – zdrave osobe	Opseg 0,15 – 0,25
	Opseg 0,25 – 0,35
	Opseg 0,35 – 0,45
	Opseg 0,45 – 0,55
DPF – bolesne osobe	Opseg 0,15 – 0,25
	Opseg 0,25 – 0,35
	Opseg 0,35 – 0,45
	Opseg 0,65 – 0,75
Godine starosti – zdrave osobe	21
	22

C. *Boxplot podataka*



Slika 1. Uporedni prikaz boxplot-ova za broj trudnoća, nivo glukoze, krvni pritisak, debljinu kože, nivo insulina, BMI, DPF i godine starosti kod zdravih osoba

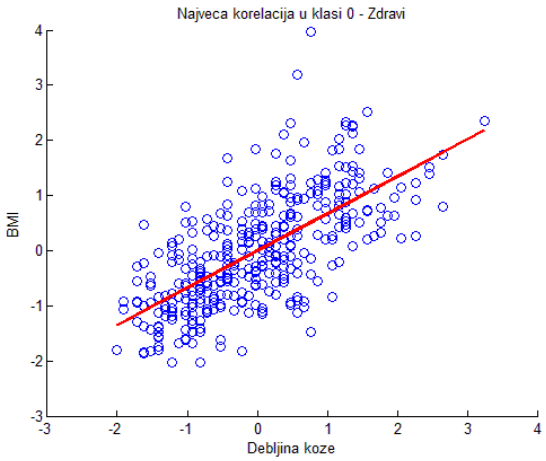


Slika 2. Uporedni prikaz boxplot-ova za broj trudnoća, nivo glukoze, krvni pritisak, debljinu kože, nivo insulina, BMI, DPF i godine starosti kod bolesnih osoba

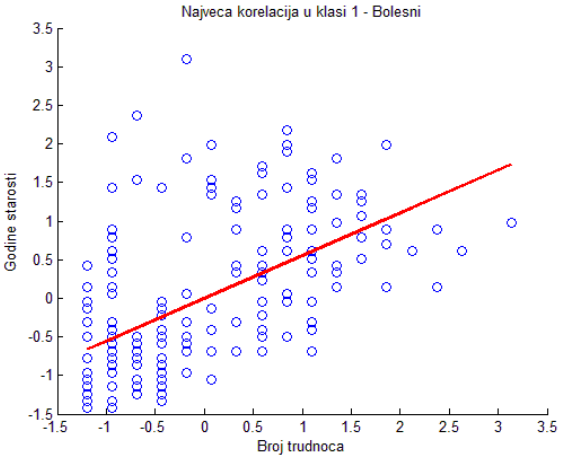
Boxplot-ovi daju pregled medijana, interkvartilnih opsega i prisustva outlier-a za svaki atribut. Može se primetiti da svi atributi, izuzev nivoa glukoze kod bolesnih osoba, sadrže outlier-e. Naročito je zanimljiva njihova raspodela kod nivoa insulina, gde outlier-i zauzimaju višestruko veći opseg od interkvartilnog opsega.

D. *Atributi sa najvećom korelacijom*

Nakon urađene Z-normalizacije podataka, primenom funkcije *corrplot* utvrđeni su parovi atributa sa najvećom korelacijom, tj., međusobnom zavisnošću. U klasi zdravih osoba, to su debljina kože i indeks telesne mase, dok u klasi bolesnih najveću međuzavisnost imaju broj trudnoća i godine starosti.

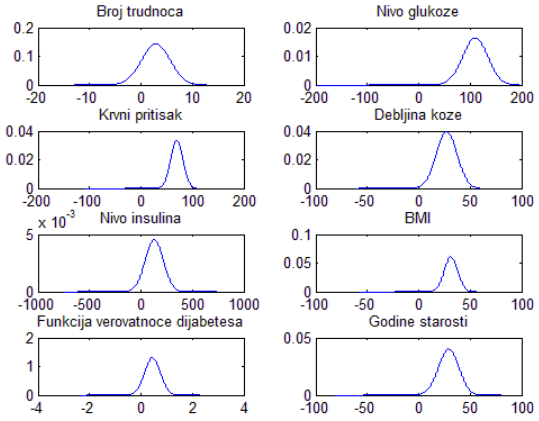


Slika 3. Scatterplot debljine kože i indeksa telesne mase kod zdravih osoba

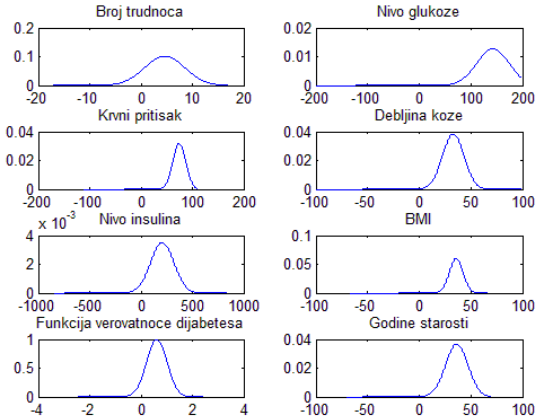


Slika 4. Scatterplot broja trudnoća i godina starosti kod bolesnih osoba

E. Atributi modelovani normalnom raspodelom

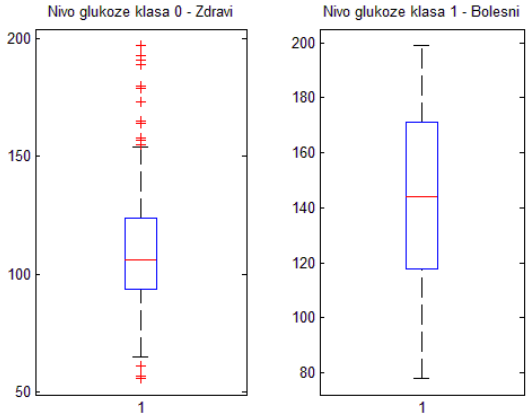


Slika 5. Prikaz atributa modelovanih normalnom raspodelom u klasi zdravih osoba



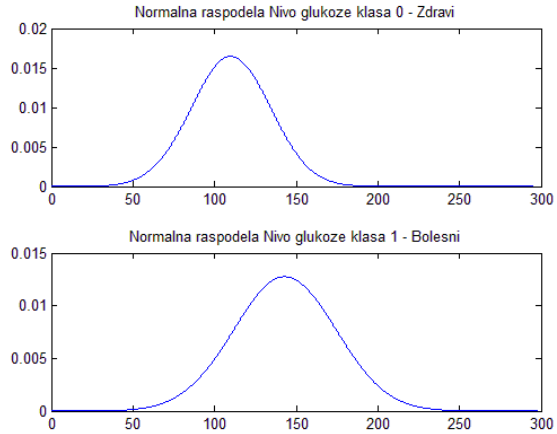
Slika 6. Prikaz atributa modelovanih normalnom raspodelom u klasi bolesnih osoba

F. Poređenje nivoa glukoze kod zdravih i bolesnih osoba



Slika 7. Uporedni prikaz boxplot-ova za nivo glukoze kod zdravih i bolesnih osoba

Sa boxplot-ova se jasno vidi da nivo glukoze kod zdravih osoba, pored toga što ima značajno manju centralnu vrednost, takođe i dosta manje varira nego kod bolesnih osoba. Visoke vrednosti (preko 150) su kod zdravih osoba svrstane u outlier-e, dok se kod bolesnih smatraju za „normalne“, ili čak pripadaju interkvartilnom opsegu.



Slika 8. Uporedni prikaz normalne raspodele nivoa glukoze kod zdravih i bolesnih osoba

Posmatranjem grafika normalne raspodele nivoa glukoze dolazi se do sličnih zaključaka kao i analizom boxplot-ova. Raspodela nivoa glukoze kod bolesnih osoba ima veću standardnu devijaciju i veću srednju vrednost nego kod zdravih osoba.

III. ZAKLJUČAK

Dijabetes se definiše kao stanje hronično povišenog nivoa glukoze u krvi, tako da bi svako klasifikovanje osoba na zdrave i obolele od dijabetesa trebalo da se zasniva na praćenju i analizi tog parametra. Jedno merenje glikemije (nivoa glukoze u krvi) nije dovoljno da se dobije potpuna slika o zdravstvenom stanju osobe, već se takva

merenja ponavljaju više puta, i to u različitim situacijama (pre obroka, posle obroka...).

Zbog toga bi klasifikator koji bi delio osobe na zdrave i obolele od dijabetesa trebalo projektovati tako da pri odlučivanju najviše uzima u obzir upravo glikemiju. Takođe bi mu trebalo obezbediti veću bazu podataka o ovom parametru (ponoviti merenja više puta). Naravno, uključivanje ostalih parametara (nivoa insulina, indeksa telesne mase, debljine kože...) u proces klasifikacije je poželjno, s obzirom na to da i ovi parametri imaju određeni uticaj na patogenezu dijabetesa.