Klasifikacija ispitanika u odnosu na postojanje karidiovaskularnih bolesti

Iva Marković, BI-20/2020, iva.kg.01@gmail.com
Milica Tomić, BI-1/2020, milica.tomic.kg@gmail.com

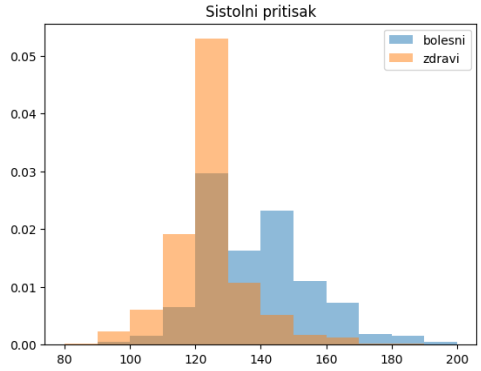
I. Uvod

Kardiovaskularne bolesti su grupa oboljenja koja karakterišu oštećenja srca i krvnih sudova. U ovom radu rešava se problem klasifikacije kardiovaskularnih bolesti. Klasifikovaće se ispitanici iz baze koji imaju bilo koji tip kardiovaskularnih bolesti i oni koji nemaju nijedan tip, odnosno postoji problem binarne klasifikacije.

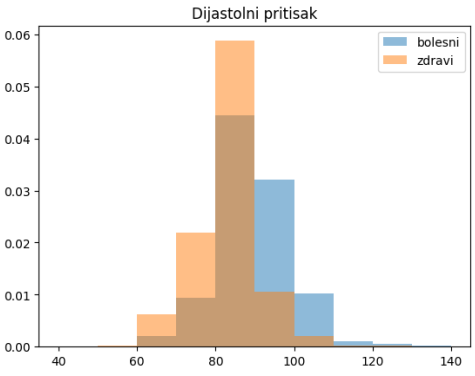
II. Baza

Baza ima 70000 uzoraka i 13 obeležja. Jedan uzorak predstavlja jednog ispitanika sa podacima o njegovim godinama, visini, polu, masi, sistolnom i dijastolnom pritisku, nivou holesterola i glukoze uz podatke o tome da li ispitanik konzumira alkohol ili cigare i da li je fizički aktivan. Svi navedeni podaci su obeležja u bazi, od kojih je 6 numeričkih obeležja.

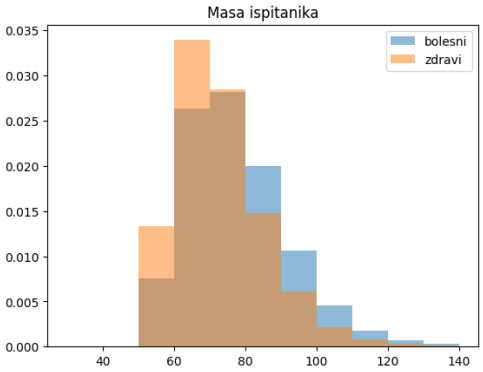
Na histogramima ispod mogu se videti obeležja koja najbolje razdvajaju klase.



Sl. 1. Raspodela bolesnih i zdravih ispitanika na osnovu sistolnog pritiska



Sl. 2. Raspodela bolesnih i zdravih ispitanika na osnovu dijastolnog pritiska



Sl. 3. Raspodela bolesnih i zdravih ispitanika na osnovu mase

III. Obrada baze

Obeležje *ID* je izbačeno jer ne sadrži korisne informacije za obučavanje modela, a obeležje *prisustvo kardiovaskularnih bolesti* se izbacuje jer se koristi za proveru uspešnosti modela.

Baza sadrži nevalidne vrednosti kao što su suviše male vrednosti za visinu i masu ispitanika u bazi, negativne i ekstremne vrednosti sistolnog i dijastolnog pritiska. U daljem nastavku opisano je kako su ovi problemi rešeni.

Visina ispitanika - izbacuju se sve vrednosti ispod 120 cm, a iznad te visine se vrednosti zadržavaju (uzeta je prosečna visina osobe patuljastog rasta koja iznosi 120 cm).

Masa ispitanika - masa ispitanika se ograničava uzimajući u obzir indeks telesne mase osobe prosečne visine u bazi (165 cm) izračunate BMI kalkulatorom. Sve vrednosti mase koje se nalaze ispod izračunate vrednosti (50 kg) se uklanjaju.

Sistolni pritisak - negativne vrednosti sistolnog pritiska se pretvaraju u apsolutne vrednosti zato što njihova apsolutna vrednost predstavlja normalan sistolni pritisak (- se smatra kao greška u upisu vrednosti u bazi). Vrednosti koje su iznad 370 mmHg i ispod 50 mmHg se izbacuju.

Dijastolni pritisak - negativne vrednosti dijastolnog pritiska se pretvaraju u apsolutne vrednosti zato što njihova apsolutna vrednost predstavlja normalan dijastolni pritisak (- se smatra kao greška u upisu vrednosti u bazi). Vrednosti koje su iznad 360 mmHg i ispod 20 mmHg se izbacuju.

Nakon obrade baze podataka, baza sadrži 67776 uzoraka i 11 obeležja.

IV. OBUKA MODELA

Baza podataka je podeljena na skup za trening i skup za test. Trening skup sadrži 90% nasumičnih uzoraka, dok 10% nasumičnih uzoraka se nalazi u test skupu.

GridSearchCV funkcija u *Pythonu* se koristi za automatsko pretraživanje kroz sve moguće kombinacije hiperparametara koje su definisane u okviru funkcije i moddl evaluira koristeći unakrsnu validaciju. Unakrsnom validacijom se dostupni skup uzoraka za testiranje deli na 10 jednakih podskupova i u svakoj rundi u datom podskupu se izdvajaju uzorci za testiranje i validaciju. Parametri ove funkcije su:

- klasifikator koji se koristi,
- parametri i njihove vrednosti zadatog klasifikatora,
- metrika za procenu performanse modela,
- broj podskupova za unakrsnu validaciju.

Kao metrika za procenu performanse modela koristi se osetljivost koja predstavlja udeo ispravno klasifikovanih uzoraka iz klase bolesnih. Što je veća osetljivost, to znači da je veći broj bolesnih otkriveno, a mali broj nije otkriven. Takođe, veća je greška bolesnog ispitanika svrstati u klasu zdravih, nego zdravog ispitanika u klasu bolesnih, što predstavlja još jedan razlog za korišćenje osetljivosti.

U ovom projektu korišćena su 4 algoritma za klasifikacione probleme:

- *Logistička regresija,*
- *Klasifikator metodom k najbližih suseda (kNN),*
- *Metoda slučajne šume (Random Forest Tree),*
- *Mašina na bazi vektora nosača (SVM).*

A. *Logistička regresija*

Logistička regresija je statistički model koji se koristi za predviđanje aposteriorne verovatnoće da će neki događaj pripadati određenoj klasi. Koristi se za binarnu klasifikaciju u datom problemu. Prednost logističke regresije u medicinskim istraživanjima su otpornost na šum i autlajere i to što je izlaz u obliku verovatnoće da ispitanik ima određenu bolest.

Parametri i njihove vrednosti korišćene pri rešavanju klasifikacionog problema su:

- *solver* sa vrednostima *lbfgs*, *sag*, *saga*,
- *fit_intercept* postavljena na vrednost *True.*

Parametar *solver* se odnosi na algoritam koji se koristi za rešavanje optimizacionog problema minimizacije funkcije cene. Podrazumevana vrednost je *lbfgs* i preporučuje se korišćenje kod baze podataka koja ima veliki skup podataka, kao i vrednosti *sag*, *saga*. Parametar *fit_intercept* odredjuje da li model ima slobodan član ili ne u zavisnosti od toga da li je vrednost postavljena na *True* ili *False*, respektivno.

Najbolji dobijen rezultat je 0.669, sa parametrom solver čija je vrednost *lbfgs.*

B. *Klasifikator metodom k najbližih suseda (kNN)*

Metoda k najbližih suseda koristi se kao neparametarska metoda za procenu gustine raspodele koja jednostavno aproksimira Bajesov klasifikator. Algoritam klasifikuje nepoznati uzorak na osnovu klasne pripadnosti k susednih uzoraka iz skupa za obuku. Prednost ove metode je ta što ona koristi lokalne informacije uzoraka, odnosno ne zavisi od raspodele uzoraka u klasama, već posmatra samo rastojanja k najbližih suseda od nepoznatog uzorka.

Parametri i njihove vrednosti korišćene pri rešavanju klasifikacionog problema su:

- n_*neighbors* sa vrednostima *9,11,13,15,17,*
- *metric* sa vrednostima *chebyshev, euclidean, manhattan.*

Parametar n_*neighbors* određuje koliko suseda će biti uzeto u obzir prilikom klasifikacije (parametar *k*), dok parametar *metric* u zavisnosti od njegove vrednosti meri udaljenost između dva uzorka (vektora obeležja). Pošto je reč o binarnoj klasifikaciji, preporučuje se da vrednosti za n_*neighbors* budu neparni brojevi kako bi se sprečio nerešen rezultat pri glasanju za dodelu klase. Izabrane vrednosti za parametar *metric* su dobre za viskodimenzionalni prostor, *manhattan* ima dodatnu otpronost na autlajere, odnosno manju osetljivost na ekstremne vrednosti u pojedinim dimenzijama, *euclidean* se najčešće koristi kod obeležja sa realnim vrednostima. *Chebyshev* metrika ne zahteva normalizaciju ili standardizaciju podataka, i dozvoljava atribute merene u različitim jedinicama ili opsezima vrednosti, što je čest slučaj kod medicinskih podataka.

Najbolji dobijen rezultat je 0.691 za vrednosti *11* i *euclidean* parametra n_*neighbors* i *metric*, respektivno.

C. *Metoda slučajne šume (Random Forest Tree)*

Metoda slučajne šume predstavlja metodu ansambalskog učenja koja koristi stabla odluke kao jednostavne klasifikatore. Obučava se mnoštvo stabala odluke, a donošenje krajnje odluke o klasi vrši se glasanjem. Pomoću bootsrap metode ponovnog uzorkovanja iz skupa za obuku formira se više novih skupova za obuku tako što se nasumično izvlače uzorci sve dok formirani skup ne bude iste veličine kao i originalni skup za obuku. Prednost ove metode je nezavisnost od tipa podataka, kao i robusnost na uzorke koji znatno odstupaju od ostalih.

Parametri i njihove vrednosti korišćene pri rešavanju klasifikacionog problema su:

- *n_estimators* sa vrednostima *100, 325, 550, 725, 1000,*
- *max_depth* sa vrednostima *10, 50, 100.*

Parametar *n_estimators* predstavlja broj stabala koji se obučava. Podrazumevana vrednost je 100. Veći broj stabala dovodi do bolje preformanse modela, ali može i povećati vreme obučavanja. *Max_depth* predstavlja dubinu svakog stabla u šumi. Velike vrednosti ovog parametra mogu dovesti do natprilagođenja, dok važi obrnuto ako su male vrednosti.

Najbolji dobijen rezultat je 0.703 za vrednosti *775* i *100* parametra *n_estimators* i *max_depth*, respektivno.

D. *Mašina na bazi vektora nosača (SVM)*

Mašina na bazi vektora nosača je algoritam koji za cilj ima da indetifikuje hiperravan koja treba da razdvoji

uzorke tako da između oblasti popunjenih uzoraka različitih klasa postoji što širi prostor. Ovaj algoritam je dobar za konstrukciju nelinearnih granica odlučivanja zato što nelinearno preslikava uzorke u višedimenzionalni prostor nakon čega sledi konstrukcija optimalne hiperravni razdvajanja u tom višedimenzionalnom prostoru.

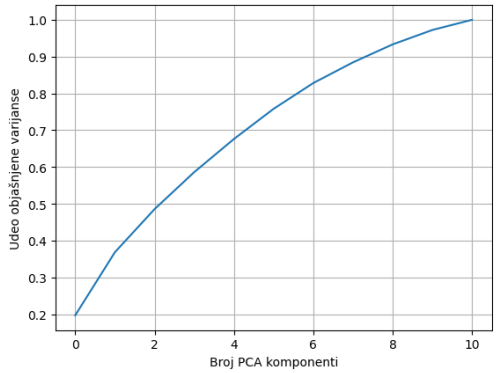
Parametri i njihove vrednosti korišćene pri rešavanju klasifikacionog problema su:

- *C* sa vrednostima *0.1, 1, 10*,
- *gamma* sa vrednostima *0.1, 1, 10*,
- *kernel* sa vrednositma *linear*, *poly*, *rbf*.

Parametar C predstavlja regularizacioni parametar, odnosno predstavlja parametar koji definiše toleranciju prelaska uzorka na pogresnu stranu margine (vektora nosača). Velike vrednosti parametra *C* mogu dovesti do natprilagođenja. *Gamma* je parametar za *rbf* kernel, slično važi kao i kod parametra *C*, velike vrednosti dovode do natprilagođenja. *Linear* kernel transformiše podatke u linearno separabilan prostor, dok *poly* kernel transformiše podatke u prostor visoke dimenzionalnosti. Parametar *degree* određuje red polinoma koji se koristi za preslikavanje. Podrazumevana vrednost za parametar *degree* je 3.

V. Smanjenje dimenzionalnosti

PCA je metoda koja se koristi za smanjenje dimenzionalnosti. Ovo je metoda nenaglednog učenja i biranjem pravca najvećeg rasipanja uzoraka formira se novi prostor sa manjim brojem dimenzija i uzorci se projektuju na njega. Na *slici 4*. je prikazan grafik udela objašnjene varijanse od broja zadržanih PCA komponenti. Cilj je zadržati preko 90% varijanse, zato se čuva 7 komponenti, a dodavanjem preostalih komponenti ostavaruje se mali doprinos u porastu objašnjene varijanse.



Sl. 4. Grafik zavisnosti udela objašnjene varijanse od broja PCA komponenti.

Nakon primene PCA metode upotrebljeni su isti parametri u *GridSearchCV* funkciji. Najbolji rezultati i najbolje kombinacije hiperparametara za date algoritme su:

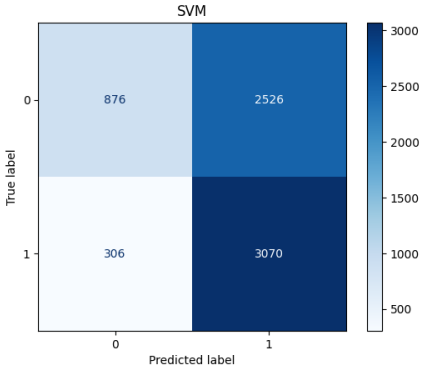
- *Logistička regresija* – rezultat *0.681* za vrednost *sag* parametra *solver*,
- *kNN* - rezultat *0.692* za vrednosti *euclidian* i *13* parametara *metric* i *n_neighbors*, respektivno,
- *Metoda slučajne šume* – rezultat *0.702* za vrednosti *100* i *775* parametara *max_depth* i *n_estimators*, respektivno,
- *SVM* – rezultat *0.882* za vrednosti *0.1* i *10* prametara *C* i *gamma*, respektivno.

VI. Testiranje

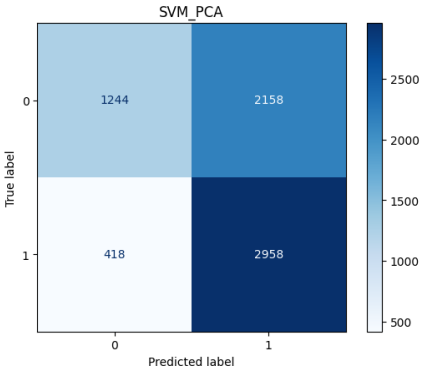
Nakon obuke modela, upoređivani su svi modeli, pre i posle smanjenja dimenzionalnosti, sa najboljom kombinacijom hiperparametara i izabarana su tri najbolja za testiranje:

- *SVM* za vrednosti *0.1* i *10* parametara *C* i *gamma*, respektivno,
- *SVM* nakon smanjenja dimenzionalnosti za vrednosti *0.1* i *10* parametara *C* i *gamma*, respektivno,
- *Random Forest Tree* za vrednosti *775* i *100* parametara *n_estimators* i *max_depth*, respektivno.

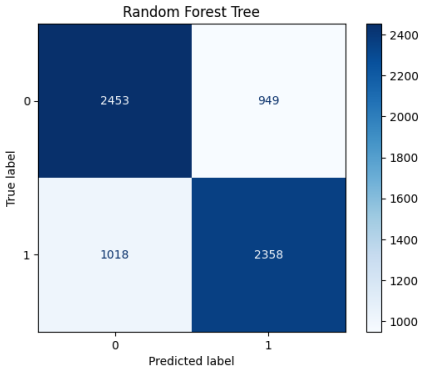
Za *mašinu na bazi vektora nosača* (*SVM*) dobijen je rezulat *0.909*, za isti model ali nakon smanjenja dimenzionalnosti rezultat je *0.876*, dok je za *Random Forest Tree* rezultat *0.704*.



Sl. 5. Matrica konfuzije za SVM



Sl. 6. Matrica konfuzije za SVM_PCA



Sl. 7. Matrica konfuzije za Random Forest Tree

Najbolji rezultat pri rešavanju ovog problema dobijen je korišćenjem *mašine na bazi vektora nosača (SVM*). Matrica konfuzije za ovu metodu pokazuje da su u velikoj meri otkriveni pacijenti koji zaista boluju od kardiovaskularnih bolesti. Takođe, veći broj zdravih pacijenata je detektovano kao bolesno, dok je mali broj pacijenata koji imaju kardiovaskularne bolesti neotkriveno, što je i najbitnije kod medicinskih problema ovog tipa. Manja je cena detekovati zdravog pacijenta kao bolesnog nego bolesnog pacijenta kao zdravog.

Tabela 1: Mere uspesnosti za 3 najbolja algoritma.

Mere uspešnosti/algoritam	SVM	SVM PCA	Random Forest Tree
tačnost	0.582	0.620	0.713
preciznost	0.549	0.578	0.716
osetljivost	0.909	0.876	0.704
F - mera	0.684	0.700	0.710

VII. LITERATURA

[1] Tijana Nosek, Branko Brkljac, Danica Despotovic, Milan Secujski, Tatjana Loncar-Turukalo*," Praktikum iz mašinskog ucenja*" Univerzitet u Novom sadu, Fakultet tehničkih nauka, 2020.
[2] https://scikit-learn.org/stable/index.html