Klasifikacija Heart Disease UCI dataseta

1st Selma Borovac
Elektrotehnički fakultet Sarajevo
Univerzitet u Sarajevu
Sarajevo, BiH
sborovac1@etf.unsa.ba

2nd Selma Hasanbegović
Elektrotehnički fakultet Sarajevo
Univerzitet u Sarajevu
Sarajevo, BiH
shasanbego1@etf.unsa.ba

***Sažetak*—Bolesti srca predstavljaju jednu od glavnih zdravstvenih problema sa kojima se danas susrećemo, a pokazalo se da je veliki broj bolesti srca moguće spriječiti ranim otkrivanjem sklonosti ka istim. Upotreba tehnika mašinskog učenja je od izuzetne koristi pri otkrivanju sklonosti ka srčanim oboljenjima na vrijeme, a na osnovu odgovarajućih rizičnih faktora. Stoga će u ovom seminarskom radu biti izvršena analiza pojedinih rizičnih faktora na postojanje sklonosti ka srčanim oboljenjima kod pojedinaca, kao i predviđanje šanse za dobijanje istih na osnovu dostupnog seta podataka (Heart Disease UCI dataset) i to kreiranjem modela mašinskog učenja uz pomoć klasifikacijskih algoritama: kNN (engl. *k-Nearest Neighbors*) i SVM (engl. *Support Vector Machine*). Rješenje u obliku modela mašinskog učenja će biti realiziran unutar Python programskog okruženja.**

***Ključne riječi* - bolesti srca, Heart Disease UCI, mašinsko učenje, klasifikacija, kNN, SVM**

***Abstract* - Heart disease is one of the main health problems we face today, and it has been shown that a large number of heart diseases can be prevented with an early detection of having a predisposition to them. The use of machine learning techniques can be extremely useful in detecting predisposition to heart disease problems in a timely manner, based on appropriate risk factors. Therefore, in this paper, an analysis of individual risk factors for the existance of predisposition to heart disease in individuals will be performed, as well as the predicting the chance of obtaining them based on the available dataset (Heart Disease UCI dataset) by creating a machine learning model using classification algorithms: kNN (*k-Nearest Neighbors*) and SVM (*Support Vector Machine*).** Machine learning models will be implemented within Python programming environment.

***Index Terms*—heart disease, Heart Disease UCI, machine learning, classification, kNN, SVM**

I. Uvod

Srčane bolesti su jedan od glavnih briga današnjice, a predstavlja jedan od vodećih uzročnika smrti u svijetu u proteklih 20 godina (World Health Organization, WHO 2019). Bolesti srca predstavljaju skup bolesti i stanja koja uzrokuju kardiovaskularne probleme. Neki od tipova srčanih bolesti su:

- Aritmija
- Ateroskleroza
- Koronarna bolest srca (CHD)
- Kardiomiopatije
- Srčane infekcije
- Otkazivanje srca
- Urođena srčana mana

Postoje mnogi faktori rizika za bolesti srca pri čemu se neki mogu kontrolisati, a neki ne. Najčešći faktori rizika oboljenja od srčanih bolesti uključuju [1]:

- Visok krvni pritisak
- Visok holesterol
- Konzumiranje duhana
- Pretilost
- Fizička neaktivnost
- Godine
- Spol
- Porodična historija

Kako se srčane bolesti ne mogu skroz izliječiti ili poništiti, potrebno je što ranije utvrditi postojanje sklonosti ka istima te na vrijeme reagovati, odnosno spriječiti ih. Pokazalo se da oko 80% kardiovaskularnih bolesti, uključujući bolesti srca, je moguće spriječiti ukoliko se otkrije postojanje sklonosti ka istim na vrijeme.

Kako je ključno na vrijeme otkriti postojanje sklonosi ka srčanim oboljenjima, potrebno je okrenuti se tehnologiji i metodama koje to i omogućiti. Predikcija kardiovaskularnih bolesti je jedan od najvažnijih predmeta u sklopu analize kliničnih podataka [2]. Kako je količina podataka koji se generiše u medicini odgromna, poprilično je teško ručno odrediti vjerovatnoću da će pojedinac oboliti od srčanih bolesti na osnovu odgovarajućih rizičnih faktora. Također, kako se više različitih simptoma povezuje sa bolestima srca, dijagnoza istih je još više otežana. Međutim, tehnike mašinskog učenja mogu biti izrazito korisne u tu svrhu, što predstavlja osnovnu motivaciju za izradu ovog rada.

Rad je podijeljen u nekoliko dijelova. Prvi dio je posvećen razumijevanju mašinskog učenja i modela koji se baziraju na mašinskom učenju. U drugom dijelu rada je pažnja posvećena algoritmima mašinskog učenja, kNN i SVM, koji će biti korišteni pri daljoj izradi. Treći dio rada obuhvata opis Heart Disease UCI dataseta koji će biti korišten za izradu modela za klasifikaciju na osnovu prethodno opisanih algoritama. Zatim su je dat opis eksperimenta i prikazani su rezultati klasifikacije. Na kraju je dat generalni zaključak kao i međusobna usporedba korištenih klasifikacijskih algoritama.

II. Mašinsko učenje

Mašinsko učenje predstavlja granu vještačke inteligencije koja se bavi tehnikama i metodama koje omogućavaju da računati i druge mašine uče na osnovu iskustva, bez eksplicitnog programiranja. Korištenjem mašinskog učenja je moguće razviti sistem koji će na osnovu odgovarajućeg algoritma imati sposobnost adaptacije dinamičnoj okolini. Samo učenje se vrši na osnovu iskustva i poznatih podataka te što je više podataka poznato mašini veća je mogućnost da se sa boljom tačnošću riješi problem.

Klasifikacija velikih količina podataka koje se svakodnevno generišu je jedan od osnovnih primjera mašinskog učenja, a bazira se na učenju modela. Učenje modela može biti nadgledano (supervizorsko, engl. *supervised learning*), nenadgledano (nesupervizorsko, engl. *unsupervised learning*), polunadgledano (engl. *semisupervised learning*) te podržano učenje (engl. *reinforcement learning*). Nadgledano učenje podrazumijeva da se uz ulazne podatke poznaju i njima odgovarajuća izlazna vrijednost. Kod nenadgledanog učenja se ne poznaju izlazne vrijednosti, samo ulazne, te je osnovni problem nenadgledanog učenja pronaći pravilnosti u ulaznim podacima. Polunadgledano učenje predstavlja kombinaciju nadgledanog i nenadgledanog učenja. Podržano

učenje se koristi za učenje agenta koje vrši interakciju sa svojim okruženjem na osnovu odabranih akcija, a pomoću signala podrške agent uči koje akcije su željene a koje ne.

Kako će u ovom radu biti korišteno isključivo nadgledano učenje modela, ono će biti detaljnije objašnjeno i nastavku.

A. Nadgledano (supervizorsko) učenje

Cilj nadgledanog učenja jeste da se na osnovu poznatih parova (ulazni uzorak, njemu pridružena izlazna vrijednost) odredi funkcija koja preslikava ulazne podatke u željene izlazne vrijednosti za te podatke. Na ovaj način se dobiva takav model koji će dati dobre rezultate ne samo nad podacima koji su korišteni pri treniranju modela (trening skup) već i na kasnijim, do tada neviđenim podacima, koji nisu korišteni prilikom treniranja. Ovakva sposobnost modela se naziva sposobnost generalizacije.

Primjeri tehnika koje spadaju u grupu nadgledanog učenja jeste klasifikacija i regresija. U ovom radu će fokus biti na klasifikaciji kako je cilj rada klasifikacija 'heart disease uci' dataseta. Klasifikacija podrazumijeva raspoređivanje podataka u neku od prethodno definisanih klasa ili grupa s obzirom na njihove osobine. Klasifikacija se odnosi na diskretnu predikciju, odnosno klasifikovanje podataka u konačan broj klasa. Općenito, riječ je o višeklasnoj klasifikaciji, dok ukoliko je broj klasa jednak 2 onda je to binarna klasifikacija.

Najpoznatiji i najčešće korišteni algoritmi mašinskog učenja koji spadaju u tehniku nadgledanog učenja jesu kNN algoritam i SVM algoritam. Oba algoritma se uspješno primjenjuju na rješavanje izazovnih problema u medicini [3]. Kako je fokus ovog rada na klasifikaciji 'heart disease uci' dataseta, oba algoritma će biti objašnjena i korištena pri izradi rada.

B. KNN algoritam

Algoritam k-najbližih susjeda ili kNN (engl. *k-nearest neighbors*)je algoritam koji pretpostavlja da su slični primjeri međusobno blizu. Uzorcima se dodjeljuje klasa koja preovladava u skupu k najbližih primjera iz trening skupa podataka, pri čemu se skup od k najbližih primjera određuje na osnovu udaljenosti primjera iz trening skupa podataka od testnog primjera. Za k=1 testnom primjeru se dodjeljuje klasa koja odgovara njegovom najbližem susjedu (NN algoritam) [4].

KNN algoritam se može jednostavno opisati kroz sljedeći niz koraka:

1. Izračunati udaljenosti između testnog primjera i svih primjera iz trening skupa podataka
2. Ponaći k primjera sa najmanjom proračunatom udaljenosti (k najbližih primjera)
3. Testnom primjeru dodijeliti klasu koju se najviše ponavlja unutar k skupa primjera

Na slici 1 je prikazan princip rada kNN algoritma za slučaj kada je k=5. Uočava se da od 5 najbližih susjeda četiri pripadaju klasi ω_1, a jedan pripada klasi ω_3. Shodno prethodno navedenom, testni primjer X_e će, zbog većinskog 'glasanja', biti klasificiran unutar klase ω_1.

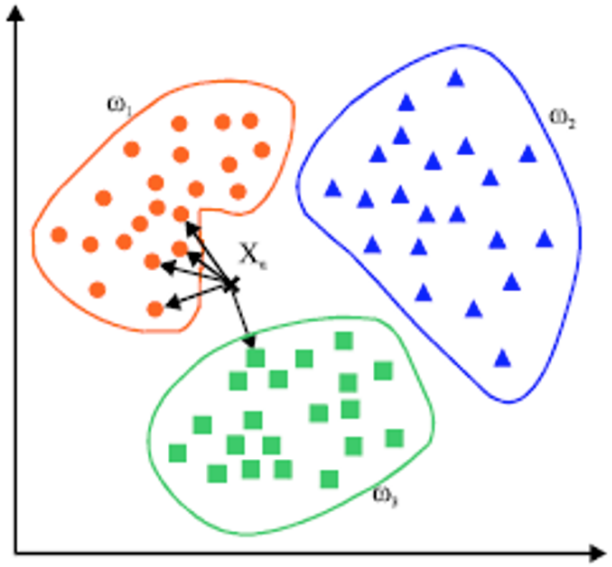
Mjere udaljenosti su ključna komponenta kNN algoritma, kako one određuju koji primjeri su slični a koji nisu, na osnovu čega se vrši sama klasifikacija. Neke od mjera udaljenosti koje se najčešće koriste su:

- Euklidska udaljenost

$$D(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

- Manhattan udaljenost

$$D(x,y) = \sqrt{\sum_{i=1}^{N}|x_i - y_i|}$$



Slika 1. Princip rada kNN algoritma

- Minkowski udaljenost

$$D(x,y) = \sqrt[p]{\sum_{i=1}^{N}(x_i - y_i)^p}$$

KNN je jednostavan i intuitivan algoritam koji je lako implementirati. Prirodno se podvrgava višeklasnim problemima, međutim testna faza zahtjeva da imamo pristup trening podacima, što memorijske zahtjeve čini visokim. Također, odabir vrijednosti k može značajno uticati na performanse kNN algoritma. Ukoliko je odabrana vrijednost k previše visoka, doći će do nedovoljnog treniranja modela (*underfitting*), a ukoliko je vrijednost parametra k nedovoljno visoka dolazi do pretreniranja modela (*overfitting*), gdje je model naučio ne samo pravilnosti koje postoje među ulaznim podacima, nego i šum i slučajne pravilnosti. Da bi se navedeno izbjeglo, potrebno je kreirati skup za treniranje i skup za validaciju kako bi odabrali k koje dalje najbolje rezultate i najveću sposobnost generalizacije.

C. SVM

SVM (engl. *support vector machine*) predstavlja skup linearnih vektora koji se mogu koristiti za klasifikaciju i regresiju, a prvobitno je osmišljen za rješavanje probleme klasifikacije.

SVM modeli se definišu kao n-dimenzionalni vektori pri čeku svaka dimenzija predstavlja značajku određenog objekta. Pokazalo se da je SVM efikasan pristup pri rješavanju višedimenzionalnih problema, a zbog svoje komputacijske efikasnosti se koristi za klasifikaciju velikih datasetova [5].

SVM omogućava rješavanje problema binarne klasifikacije gdje se instance dvije klase ne mogu linerano razdvojiti. Ovakva funkcionalnost je moguća jer treniranje i klasifikacija nepoznatih instanci zavisi isključivo od:

1. Skalarnog proizvoda instanci koje se nalaze na (unutar) margini
2. Skalarnog proizvoda nepoznate i instanci na (unutar) margini

SVM proširuje Support Vektor klasifikatora pomoću tzv. kernel funkcija. Uobičajne kernel funkcije su:

- Bez promjene (linearni kernel)

$$\phi(x_1) \cdot \phi(x_2) = K(x_1, x_2) = x_1 \cdot x_2$$

- Polinomialni kernel

$$K(x_1, x_2) = (x_1 \cdot x_2 + 1)^n$$

- Radial-basis kernel

$$K(x_1, x_2) = e^{\frac{-|x_1 - x_2|^2}{2\delta^2}}$$

Parametri n i δ su hiper-parametri, tj. učimo ih npr. pomoću metode cross-validacije.

III. Heart Disease UCI dataset

Heart Disease UCI dataset je jedan od mnogih dostupnih datasetova na kaggle.com, a dostupan je na linku [6].

Set podataka koji će se korisiti čini 76 kolona koje predstavljaju razne značajke, rizične faktore, dok je njih 14 izvojeno za korištenje. Za ovaj set podataka ispitano je 303 pacijenta sa 14 osobina, a sami set podataka je pretpocesiran, pri čemu je izdvojeno da postoje nedostajuće vrijednosti. Izdvojene osobine su navedene u nastavku:

- age
- sex
- chest pain type (1=typical angina; 2=atypical angina; 3=non-anginal pain; 4=asymptopmatic)
- resting blood pressure (testbps)
- serum cholestoral
- fasting blood sugar
- resting electrocardiographic results (0=normal; 1=ST-T wave abnormality, 2=probable of definite left ventricular hypertrophy)
- maximum heart rate achieved
- exercise induced angina
- oldpeak = ST depression induced by exercise relative to rest
- number od major vessels (0-3) colored by flourosopy
- the slope of the peak exercise ST segment
- thal: 3=normal; 6=fixed defect; 7=reversable defect
- target

Potrebno je ukratko objasniti uticaj pojedinih atributa na krajnji rezultat klasifikacije (*target*), odnosno na koji način navedeni atributi utiču na pojavu srčanih bolesti.

A. Uticaj atributa na rezultat klasifikacije, target

Godine su značajan rizični faktor kod srčanih oboljenja kako starenjem se povećava rizik od oštećenih i suženih arterija, ali i slabljenja srčanog mišića.

Spol, zajedno sa godinama utiče na vjerovatnoću obolijevanja od srčanih bolesti. Muškarci su generalno izloženi većem riziku od ovih obolijenja, dok kod žena se on povećava nakon menopauze.

Visok krvni pritisak ili hipertenzija je jedan od glavnih rizičnih faktora za pojavu bolesti srca. Prema studijima iz 2010. pokazalo se da više od 76% muškaraca i 80% žena iznad 75 godina imaju visok krvni pritisak koji, ukoliko se ne kontroliše, može dovesti do zadebljavanja arterija, sužavanja krvih sudova, kao i do srčanog udara [13].

Holesterol pomaže tijelu pri izgradnji novih ćelija, izoliranju nerava i proizvodnji hormona. Ukoliko se u krvi nazali previsok nivo holesterola, dolazi do njegovog nakupljanja u zidovima arterija, čime se one sužavaju i dolazi do usporavanja ili čak blokiranja protoka krvi prema srcu, uzrokujući bol u prsima zbog nedostatka krvi i kisika ili srčani udar, respektivno.

Nivo šećera u krvi u stanju gladi pokazuje kako ljudsko tijelo upravlja šećerom u krvi. Nerelugaran, visok nivo šećera u krvi može upućivati na dijabetes. Normalni nivo šećera u krvi jeste ispod 100 mg/dl, dok osobe kod kojih je ovaj nivo viši imaju povećan rizik od srčanih oboljenja za 300% [12].

Maksimalno ostvaren broj otkucaja srca kod zdravih osoba se dobije ukoliko se oduzme broj godina osobe od 220. Ukoliko postoje problemi sa srcem u smislu njegovog otkazivanja, srce ne pupma adekvatno, te tijelo ne dobiva dovoljno krvi i kisika. Nervni i hormonalni sistem pokušavaju navedeno kompenzirati povećavanjem krvnog pritiska i broja otkucaja srca. Međutim, na ovaj način se srčani mišić dodatno umara, uzrokujući dugoročno dodatne probleme (gubitak daha, umor, otok abdomena i članaka, oštećenje drugih organa, često bubrega) [14].

Angina predstavlja bol u predijelu prsa uzrokovana smanjenjem protoka krvi u srce, a obično je opisana kao stezanje, pritisak, težina ili bol u prsima. Drugi simptomi koji se vežu za anginu su umor, mučnina, vrtoglavica, znojenje i gubitak daha. Potrebno je uzeti u obzir sve simptome da bi se utvrdilo o koem tipu boli u prsima je riječ, te da li je riječ o stabilnoj ili nestabilnoj angini [15]. Nestabilna angina može dovesti do srčanog udara, a mogu je uzrokovati krvni ugrušci koji blokiraju ili djelimično blokiraju krvne sudove. S druge strane, stabilna angina je česta pojava i ne nužno vezana za ozbiljne bolesti srca. Često je uzrokovana fizičkom aktivnošću, ali i stresom, hladnom temperaturom i teškim obrocima.

IV. Implementacija

Za implementaciju klasifikacije Heart Disease dataseta korišteni su klasifikacijski algoritmi kNN i SVM napisanih u programskom jeziku Python. Prvi korak u implementaciji jeste priprema samih podataka za klasifikaciju.

A. Priprema podataka

Za potrebe implementacije su korišteni sljedeći python paketi: *pandas*, *matplotlib* i *sklearn* iz kojih su preuzete sve potrebne funkcije. Nakon preuzimanja i učitavanja odgovarajućeg dataseta, koji je u CSV formatu, vrši se njegova podjela na set podataka za treniranje i na set podataka za testiranje, pri čemu je omjer podjele 0.85:0.15, u svrhu što boljeg treniranja modela. Nakon normalizacije i trening i test skupa podataka, potrebno testirati model te odrediti matricu konfuzije i ROC krivu u svrhu vizuelizacije rezultata klasifikacije.

1) Matrica konfuzije: Matrica konfuzije (engl. *confusion matrix*) predstavlja tabelarni prikaz informacija o stvarnim klasama i predviđenim klasifikacijama koje je izvršio klasifikacijski model, a koristi se za vizuelizaciju rezultata klasifikacije. Na slici 2 je prikazana struktura 2x2 matrice konfuzije, pri čemu je:

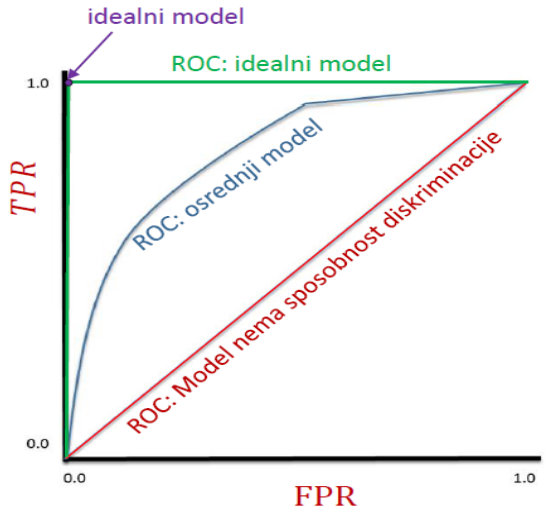
- TP (engl. *True Positive*) - broj tačno klasificiranih pozitivnih uzoraka
- FP (engl. *False Positive*) - broj netačno klasificiranih pozitivnih uzoraka
- FN (engl. *False Negative*) - broj netačno klasificiranih negativnih uzoraka
- TN (engl. *True Negative*) - broj tačno klasificiranih negativnih uzoraka

Matrica konfuzije		Klasa predviđena modelom		
		P' Pozitivna	N' Negativna	Ukupno
Stvarna klasa	P Pozitivna	TP	FN	P
	N Negativna	FP	TN	N
	Ukupno	P'	N'	P+N

Slika 2. Matrica konfuzije

Na osnovu matrice konfuzije se mogu odrediti različite mjere za ocjenu performansi klasifikatora: tačnost, preciznost, osjetljivost, specifičnost, F1 mjera i slično.

2) ROC kriva: ROC (engl. Receiver Operating Characteristics) kriva prikazuje osjetljivost (TPR) kao funkciju FPR-a (1-specifičnost) (slika 3). U tu svrhu je potrebno definisati navedeno. Osjetljivost (engl. *True Positive Rate*, TPR) klasifikatora je broj korektno klasificiranih pozitivnih uzoraka u odnosu na ukupan broj pozitivnih primjera. FPR (engl. *False Positive Rate*) klasifikatora predstavlja broj nekorektno klasificiranih pozitivnih uzoraka u odnosu na ukupan broj primjera koji su stvarno negativni. FPR se definiše i kao 1-specifičnost,TNR, što je broj korektno klasificiranih negativnih primjera u odnosu na ukupan broj negativnih primjera.



Slika 3. ROC kriva

ROC kriva, odnosno tačnije površina ispod ROC krive (engl. *Area Under the Curve*, AUC) predstavlja mjeru diskriminacije posmatranog modela. Ona pokazuje sposobnost modela da izvrši ispravnu klasifikaciju podataka. Za model kod kojeg je TPR=FPR se kaže da on nema sposobnost diskriminacije, a za model kod kojeg je TRP=1 i FRP=0 se kaže da je on idealan. Što AUC ima veću vrijendost, odnosno što je ROC kriva bliža idealnoj, model je bolji.

U nastavku su prikazani rezultati klasifikacije Heart Disease dataseta korištenjem kNN i SVM modela za klasifikaciju, a za prikaz istih je korištena matrica konfuzije, ROC kriva, a dat je i prikaz izvještaja klasifikacije. Pri čemu za rezultat klasifikacije, target, vrijedi:

- target = 0 - vjerovatnoća da osoba ima srčane bolesti <50%, klasifikuje se kao nepostojanje srčanih oboljenja
- target = 1 - vjerovatnoća da osoba ima srčane bolesti >50%, klasifikuje se kao postojanje srčanih oboljenja

B. KNN model klasifikacije

Na slici 4 je prikazana KNN matrica konfuzije, gdje se uočava da poprilično veliki broj ispravno klasificiranih i pozitivnih i negativnih uzoraka u odnosu na ukupan broj pozitivnih, odnosno negativnih primjera. Korištenjem ove matrice konfuzije prikazuje se izvještaj same klasifikacije (slika 5), pri čemu su pozitivni uzorci označeni sa *target=0*, odnosno oni koji su klasificirani da nemaju srčanih bolesti. Sa izvještaja se očitava preciznost, F1 mjera, tačnost, makro mjere i težinska suma mjera za svaku klasu.

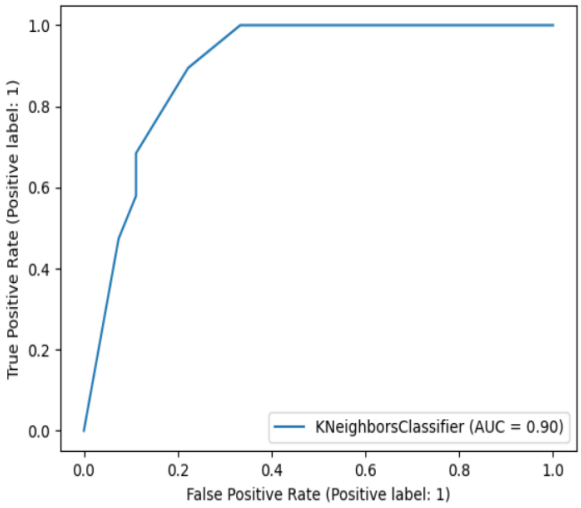
```
KNN Prediction

[[21  6]
 [ 2 17]]
```

Slika 4. KNN matrica konfuzije

	precision	recall	f1-score	support
0	0.91	0.78	0.84	27
1	0.74	0.89	0.81	19
accuracy			0.83	46
macro avg	0.83	0.84	0.82	46
weighted avg	0.84	0.83	0.83	46

Slika 5. KNN izvještaj klasifikacie

Na slici 3 je prikazana ROC kriva sa koje se očitava AUC vrijednost od 0.9. Analizom ove ROC krive se može zaključiti da ovaj model ima jako dobru sposobnost generalizacije, kako je AUC vrijendost biizu vrijednosti 1, a samim time sama krivulja je blizu idealne.



Slika 6. KNN ROC kriva

C. SVM model klasifikacije

Na slici 7 je dat prikaz SVM matrice konfuzije, dok je iz nje prikazan izvještaj klasifikacije prikazan na slici 8. U ovom slučaju je za negativne uzorke, odnosno one koji su iznačeni sa *target=1* zbog prisustva srčanih oboljenja, izvršena još bolja klasifikacija, kako je samo jedan negativan uzorak netačno klasificiran. Kako je povećan broj korektno klasificiranih primjera u usporedbi sa istim dobivenim primjenom kNN modela, to je i preciznost SVM modela veća.

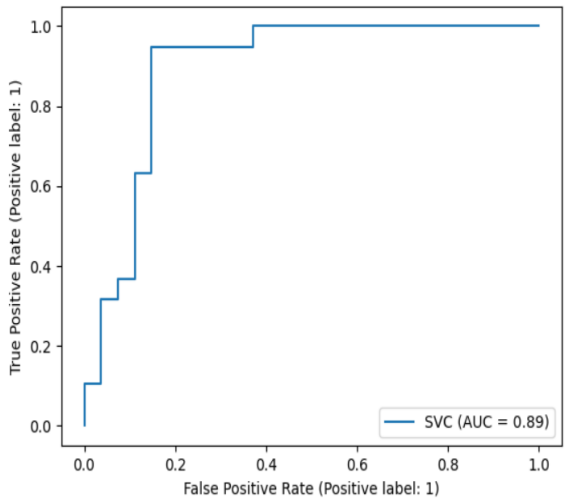
```
SVM Linear Prediction

[[21  6]
 [ 1 18]]
```

Slika 7. SVM matrica konfuzije

	precision	recall	f1-score	support
0	0.95	0.78	0.86	27
1	0.75	0.95	0.84	19
accuracy			0.85	46
macro avg	0.85	0.86	0.85	46
weighted avg	0.87	0.85	0.85	46

Slika 8. SVM izvještaj klasifikacie

ROC kriva za SVM model klasifikacije je prikazana na slici 9. Zbog prirode SVM algoritam koji se bazira na n-dimenzionalnim vektorima, ROC kriva se crta u skladu sa time, te ona ima stepenast oblik. AUC ima vrijednost 0.89, što znači da i SVM model ima dobre sposobnosni diskriminacije.



Slika 9. SVM ROC kriva

V. Diskusija

U prethodnom dijelu je izvršena implementacija kNN i SVM modela klasifikacije nad istim setom podataka. Za oba modela su korišteni isti setovi trening skupa podataka, kao i isti skup testnih podataka. Ukoliko uporedimo oba modela, možemo zaključiti da je kNN model klasifikacije nešto bliži idealnom u odnosu na SVM prilikom usporedbe ROC krivih. Međutim, kako je razlika u površini ispod krivih poprilično mala, samo 0.1, a oba modela imaju krive koje su jako blizu idealnim, za oba modela se može reći da imaju dobre sposobnosti diskriminacije.

Također, usporedbom matrica konfuzije vidi se da i kNN i SVM modeli na podjednak način klasificiraju pozitivne uzorke, pri čemu se od 27 stvarno pozitivnih uzoraka, 6 uzoraka nekorektno klasificiralo kao negativni. Navedeno daje osjetljivost, TPR, od 78% za oba modela. Međutim, SVM model daje nešto bolje rezultate prilikom klasifikacije negativnih primjera. Kod SVM modela, od 19 stvarno negativnih primjera, samo 1 primjer je pogrešno klasificiran kao pozitivan, u odnosu na 2 pogrešno klasificirana kod kNN modela. Specifičnost kNN modela jeste 89%, dok je specifičnost SVM modela nešto veća i iznosi 95%.

VI. Zaključak

Zbog činjenice da su bolesti srca jedan od vodećih uzročnika smrti u svijetu u proteklih 20 godina, javlja se sve veća potreba za što ranijim utvrđivanjem postojanja sklonosti ka srčanim bolestima, u svrhu pravovremenog reagovanja i eventualnog sprečavanja istih. Kako postoje mnogi rizični faktori oboljenja od bolesti srca, njihova pravovremena predikcija i dijagnoza je od ključne važnosti. U tu svrhu se mogu koristiti tehnike mašinskog učenja, što je iskorišteno u ovom radu.

U svrhu predikcije postojanja oboljenja srca kod pojedinaca na osnovu pojedinih faktora rizika implementirana su dva modela zasnovana na kNN i SVM algoritmima klasifikacije. Implementirani modeli su trenirani i testirani nad Heart Disease UCI datasetom, a rezultati klasifikacija koji su dobiveni korištenjem ovih modela su prikazani pomoću matrica konfuzije, izvještaja klasifikacije i ROC krivulja.

Iako oba modela daju slučne rezultate klasifikacije, na osnovu ovog primjera se može reći da je SVM model nešto bolji i to zbog uspješnije klasifikacije negativnih uzoraka, koji su u ovom slučaju uzorci kod kojih je vjerovatnoća prisustva srčanih bolesti iznad 50%. Uvijek je bolje napraviti takvu dijagnozu da se osobe koje zaista imaju srčane bolesti (negativni primjeri) klasificiraju kao osobe sa srčanim bolestima (ispravno klasificirani negativni primjeri). Dakle, bitno je da se dešava što je manje grešaka prilikom klasifikacije negativnih uzoraka, da je što manje lažno pozitivnih uzoraka.

Poboljšanja klasifikacije su moguća promjenom dostupnog seta podataka, njegovim nadopunjavanjem, kako postoje nedostajuće vrijednosti, ali i dodavanjem podataka novoispitanih pacijenata. Rezultat klasifikacije, ne samo da zavisi i od seta ulaznih podataka, već i od parametara korištenih algoritama klasifikacije. Kod upotrebe kNN algoritma, treba se voditi računa o vrijednosti k, odnosno o broju susjeda koji se uzimaju u obzir prilikom proračuna, da ne dođe do pretreniranja ili nedovoljnog treniranja modela. U tu svrhu se, pored promjena u trening skupu uzoraka, treba kreirati i skup za validaciju kako bi se pronašla vrijednost parametra k koja daje najbolje rezultate.

Literatura

[1] Deekshatulu, B. L., and Priti Chandra. ”Classification of heart disease using k-nearest neighbor and genetic algorithm.” Procedia technology 10 (2013): 85-94.

[2] Patel, Jaymin, Dr TejalUpadhyay, and Samir Patel. ”Heart disease prediction using machine learning and data mining technique.” Heart Disease 7.1 (2015): 129-137.

[3] Bzdok, Danilo, Martin Krzywinski, and Naomi Altman. ”Machine learning: supervised methods.” (2018): 5-6.

[4] Raikwal, J. S., and Kanak Saxena. ”Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set.” International Journal of Computer Applications 50.14 (2012).

[5] Pouriyeh, Seyedamin, et al. ”A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease.” 2017 IEEE symposium on computers and communications (ISCC). IEEE, 2017.

[6] Kaggle, Heart Disease UCI. [Online] Dostupno na: https://www.kaggle.com/ronitf/heart-disease-uci.

[7] Detrano, Robert, et al. ”International application of a new probability algorithm for the diagnosis of coronary artery disease.” The American journal of cardiology 64.5 (1989): 304-310.
[8] Buettner, Ricardo, and Marc Schunter. ”Efficient machine learning based detection of heart disease.” 2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom). IEEE, 2019.
[9] Ramalingam, V. V., Ayantan Dandapath, and M. Karthik Raja. ”Heart disease prediction using machine learning techniques: a survey.” International Journal of Engineering & Technology 7.2.8 (2018): 684-687.
[10] Đonko, Dž., Support Vector Machines (SVM), Predavanje 8, Mašinsko Učenje, Elektrotehnički Fakultet Sarajevo, Univerzitet u Sarajevu
[11] Turajlić, E. Predavanje 7, Digitalna Obrada Signala u Telekomunikacijama 2, Elektrotehnički Fakultet Sarajevo, Univerzitet u Sarajevu, 2021.
[12] Fasting Blood Sugar Above 90 Puts You At Risk of Heart Disease, 2002. [Online] Dostupno na : http://www.diabetesincontrol.com/fasting-blood-sugar-above-90-puts-you-at-risk-of-heart-disease/
[13] UTSouthwestern Medical Center. Under Pressure: How blood pressure affects heart disease risk, 2016. [Online] Dostupno na: https://utswmed.org/medblog/high-blood-pressure-heart-disease/
[14] Michigan Medicine. Heart Failure: Compensation by the Heart and Body, 2020. [Online] Dostupno na: https://www.uofmhealth.org/health-library/aa86963
[15] Mayo Clinic. Angina - Symptoms and Causes [Online] Dostupno na: https://www.mayoclinic.org/diseases-conditions/angina/symptoms-causes/syc-20369373