

UNIVERZITET U BEOGRADU  
MATEMATIČKI FAKULTET



Katarina Savičić

PREPOZNAVANJE VISOKORIZIČNIH PACIJENATA OBOLELIH  
OD HIPERTROFIČNE KARDIOMIOPATIJE PRIMENOM  
METODA KLASIFIKACIJE

master rad

Beograd, 2023.

**Mentor:**

prof. dr Nenad Mitić  
Univerzitet u Beogradu, Matematički fakultet

**Članovi komisije:**

prof. dr Saša Malkov  
Univerzitet u Beogradu, Matematički fakultet

dr Mirjana Maljković  
Univerzitet u Beogradu, Matematički fakultet

## Sadržaj

1. Uvod.....	5
1.1. Kardiomiopatija.....	5
1.1.1. Hipertrofična kardiomiopatija.....	5
2. Cilj rada .....	7
2.1. Podaci.....	7
2.2. Priprema podataka.....	12
3. Algoritmi klasifikacije .....	15
3.1. Drveta odlučivanja .....	16
3.1.1. Random Trees.....	18
3.1.2. CART .....	19
3.1.3. C5.0 .....	20
3.1.4. CHAID .....	21
3.1.5. QUEST.....	22
3.2. Metod potpornih vektora .....	23
3.2.1. Linearni metod potpornih vektora.....	24
3.3. Veštačka neuronska mreža.....	25
4. Dobijeni rezultati.....	27
4.1. Mere kvaliteta .....	27
4.1.1. Matrica konfuzije.....	28
4.1.2. AUC.....	29
4.2. Globalni kvalitet algoritama .....	31
4.3. Procena kvaliteta po algoritmima .....	34
4.3.1. C5.0 .....	35
4.3.2. CART.....	36
4.3.3. CHAID .....	38
4.3.4. Linearni metod potpornih vektora.....	39
4.3.5. Neuronska mreža .....	40
4.3.6. QUEST.....	41
4.3.7. Random Trees.....	42
4.4. Najznačajniji atributi u klasifikaciji .....	44

5. Zaključak.....	46
Dodatak A.....	47
Dodatak B.....	66
Literatura.....	68

## 1. Uvod

Kardiologija je grana medicine koja se bavi prevencijom, dijagnostikom i lečenjem bolesti srca i krvnih sudova. Broj bolesti kojima se kardiologija bavi je veliki, a jednu od grupa bolesti čine različite vrste kardiomiopatija.

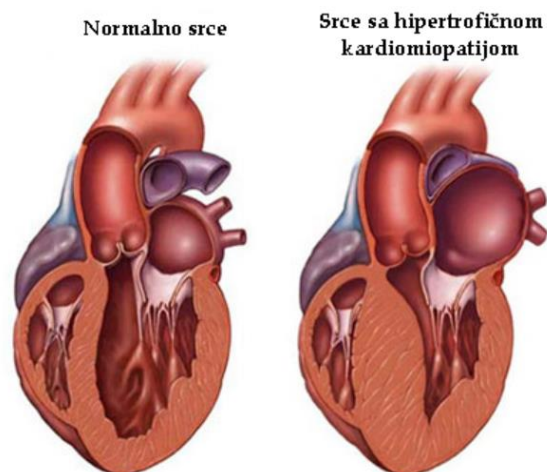
### 1.1. Kardiomiopatija

Pojam kardiomiopatija se odnosi na različite bolesti srčanog mišića. Ove bolesti imaju mnogo uzroka, znakova i simptoma. Isto tako, mogućnosti lečenja se razlikuju. U većini slučajeva kardiomiopatija uzrokuje da srčani mišić postane zadebljan i odlikuje se, najčešće, većom krutošću. U ređim slučajevima obolelo tkivo srčanog mišića zamenjuje se ožiljnim tkivom. Kako se kardiomiopatija pogoršava, srce postaje slabije u funkcionalnom smislu, postaje manje sposobno da pumpa krv po telu. Istovremeno se i normalna električna aktivnost srca menja tako da se odlikuje nepravilnim srčanim radom. Rezultat može biti nastanak i razvoj srčane slabosti ili nepravilni otkucaji srca koji se nazivaju aritmijama. Oslabljeno srce, takođe, može izazvati i druge komplikacije, poput problema sa srčanim zaliscima [1].

Osnovni tipovi kardiomiopatija uključuju: dilatativnu, hipertrofičnu i restriktivnu kardiomiopatiju. Neke druge vrste kardiomiopatije nazivaju se neklasifikovanom kardiomiopatijom. Još jedan oblik je kardiomiopatija izazvana stresom, poznata i kao sindrom slomljenog srca. Kardiomiopatija se može steći, što znači da se razvija zbog druge bolesti, stanja ili uzroka, ili se može naslediti, što znači da je gen za bolest nasleđen od roditelja. U mnogim slučajevima uzrok kardiomiopatije nije poznat. Kardiomiopatija može zahvatiti sve starosne grupe, mada je verovatnije da određene starosne grupe imaju učestaliju pojavu pojedinih vrsta ove bolesti [1].

#### 1.1.1. Hipertrofična kardiomiopatija

Hipertrofična kardiomiopatija je bolest kod koje srčani mišić postaje abnormalno debeo (Slika 1), tj. hipertrofičan. Srčani mišić postaje krut i manje elastičan, tako da se ne može proširiti i napuniti krvlju između otkucaja srca. Ova vrsta kardiomiopatije zahvata i provodni sistem srca.



Slika 1 - Srce osobe obolele od hipertrofične kardiomiopatije

*Slika je preuzeta sa [2]*

Hipertrofična kardiomiopatija može se pojaviti u bilo kom uzrastu, ali najčešće pogađa starije ljude. Ukoliko je nasledna, smatra se da je uzrokovana mutacijama na genima, i da te mutacije izazivaju abnormalno uvećanje srčanog mišića. Kod pojave bolesti usled defekta na genima, tip hipertrofične kardiomiopatije koji će se razviti može varirati unutar porodice. Iako je defekt na genima uslov za razvoj nasledne hipertrofične kardiomiopatije, ne mora da znači da će je imati svaka osoba koja nasledi neki od gena za koje je ustanovljeno da mogu biti uzroci oboljevanja od hipertrofične kardiomiopatije, jer je moguće da se bolest nikada ne ispolji. Ovaj oblik kardiomiopatije je obično nasledna bolest. Postoji 50% šanse da deca naslede ovu bolest od svojih roditelja. Oni kojima članovi porodice boluju od hipertrofične kardiomiopatije, treba da urade kardiološki pregled [1].

Kako bi se postavila dijagnoza potrebno je uraditi ekokardiografiju (na osnovu koje se može videti zadebljanje srčanog mišića), EKG, 24h Holter EKG monitoring (koji detektuje promene srčanog ritma u toku 24 časa), kateterizaciju srca (merenje pritiska krvi u srcu), MRI srca, skrining testove (preventivni pregledi kod pacijenata koji imaju dijagnostikovanu hipertrofičnu kardiomiopatiju u porodici) [1].

Hipertrofična kardiomiopatija se može lečiti lekovima, promenom stila života, ugradnjom pejsmejкера, septalnom miektomijom (operacija na otvorenom srcu kojom se odstranjuje zadebljali deo srčanog mišića), i drugim metodama.

## 2. Cilj rada

Pri dijagnostikovanju bilo koje bolesti, lekari na raspolaganju imaju određene podatke o pacijentu. Ti podaci mogu biti opšte informacije o pacijentu, o njegovom načinu života, o postojećim zdravstvenim problemima i o trenutnom zdravstvenom stanju. Takođe, na osnovu ovakvih podataka, može se odrediti i stepen bolesti, način lečenja, nivo rizičnosti pacijenta i druge. Jedan od vidova primene informatike u medicini jeste korišćenje različitih tehnika za obradu ovakvog tipa podataka i dobijanje informacija iz njih. Cilj rada je primena različitih metoda klasifikacije, nad podacima koji opisuju zdravstveno stanje pacijenata obolelih od hipertrofične kardiomiopatije, kako bi se prepoznali visokorizični pacijenti.

### 2.1. Podaci

Podaci koji se koriste u radu dobijeni su od univerzitetske bolnice u Firenci za potrebe već sprovedenog istraživanja i predstavljaju zdravstveno stanje pacijenata obolelih od hipertrofične kardiomiopatije. Podaci sadrže ukupno 13386 instanci i mogu se podeliti u tri osnovne grupe.

Prvu grupu čine neke od osnovnih informacija o pacijentu, kao što su pol, visina, težina, da li je pacijent pušač, da li se drogira, da li je u drugom stanju i slično, zatim informacije o postojećim mutacijama na određenim genima, primarna dijagnoza koja je postavljena pacijentu i, na kraju, vrednosti različitih kardioloških testova i slične informacije. Jedna instanca ove grupe podataka predstavlja jedan dolazak pacijenta na pregled. Prilikom svakog dolaska pacijenta na pregled beleže se sve navedene informacije, jer su opšte informacije o pacijentu podložne promenama u periodu između dva pregleda, a zdravstvene analize i testove je svakako neophodno raditi pri svakom dolasku pacijenta na pregled. Na slikama (Slika 2, Slika 3, Slika 4) se mogu videti statistike nekih od atributa vezanih za opšte informacije o pacijentu, postojanje mutacija na genima i vrednosti kardioloških testova. Na osnovu ovih slika, takođe, možemo primetiti da su atributi većinom prepoznati kao kontinualni, iako neki od njih imaju vrednosti nula ili jedan. Pored toga, možemo primetiti da je jedna primarna dijagnoza zastupljenija od drugih i to je dijagnoza da pacijent ima hipertrofičnu kardiomiopatiju, kao i da ima veliki broj instanci koje nemaju vrednosti u atributima koji predstavljaju kardiološke testove.

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Age		Continuous	-0.082	94.354	50.619	18.624	-0.288	—	13376
Gender		Categorical	—	—	—	—	—	2	13386
Primary_Diagnosis		Categorical	—	—	—	—	—	9	13386
FHx_DCM		Continuous	0.000	1.000	0.018	0.135	7.157	—	13386
FHx_HCM		Continuous	0.000	1.000	0.334	0.472	0.705	—	13386
FHx_SCD		Continuous	0.000	1.000	0.186	0.389	1.615	—	13386
FHx_CAD		Continuous	0.000	1.000	0.040	0.196	4.708	—	13386
Alcohol		Continuous	0.000	1.000	0.011	0.103	9.486	—	13386
Drug		Continuous	0.000	1.000	0.002	0.040	25.191	—	13386

Slika 2 - Deo skupa atributa prve grupe koji predstavlja opšte informacije o pacijentu

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Gene_Test_Negative		Continuous	0.000	1.000	0.664	0.472	-0.696	—	13386
Gene_Name__ACTC1		Continuous	0.000	1.000	0.001	0.037	27.218	—	13386
Gene_Name__CSRP3		Continuous	0.000	1.000	0.002	0.041	24.065	—	13386
Gene_Name__MYBPC3		Continuous	0.000	1.000	0.233	0.423	1.266	—	13386
Gene_Name__MYH7		Continuous	0.000	1.000	0.124	0.330	2.276	—	13386
Gene_Name__MYL2		Continuous	0.000	1.000	0.008	0.090	10.896	—	13386
Gene_Name__MYL3		Continuous	0.000	1.000	0.002	0.046	21.417	—	13386
Gene_Name__OTHER		Continuous	0.000	1.000	0.075	0.264	3.225	—	13386
Gene_Name__PRKAG2		Continuous	0.000	1.000	0.003	0.050	19.769	—	13386

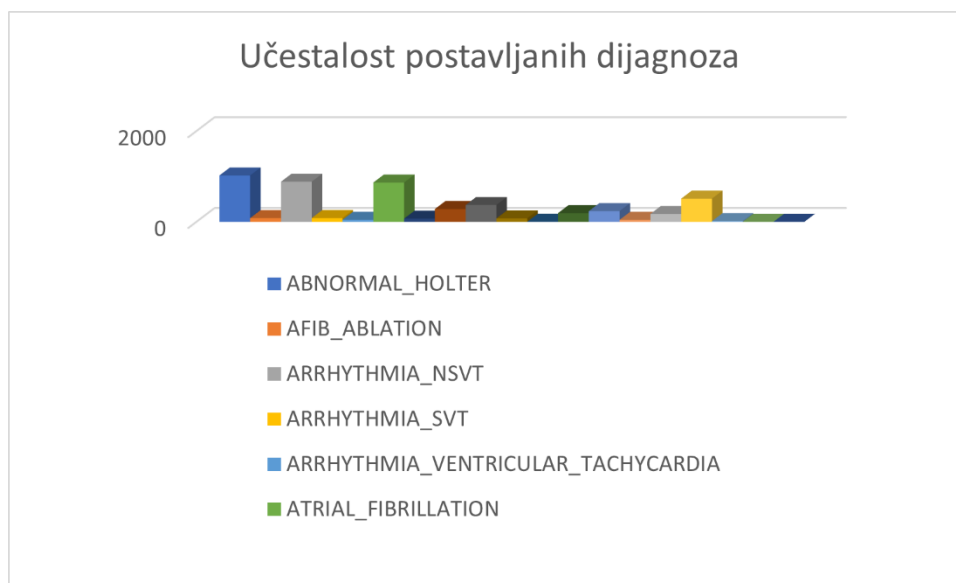
Slika 3 - Deo skupa atributa prve grupe koji predstavlja indikatore na postojeće mutacije na određenim genima

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
ECG_Pathological_Q_Waves		Continuous	0.000	1.000	0.009	0.096	10.246	—	13386
ECG_PR		Continuous	1.000	2968.000	176.266	82.514	18.387	—	3902
ECG_QRS		Continuous	12.000	886.000	113.833	42.440	5.074	—	2878
ECG_QTc		Continuous	28.000	9367.000	449.694	311.576	26.981	—	3733
ECG_Rate		Continuous	8.000	810.000	64.658	19.909	18.480	—	10430
ECG_Rhythm		Categorical	—	—	—	—	—	5	10979
ECG_P		Continuous	10.000	1222.000	117.895	42.874	18.986	—	1822
ECG_QT		Continuous	47.000	666.000	442.042	56.406	-0.474	—	336
Ech_Echo_IVS		Continuous	4.000	115.000	15.905	5.983	1.072	—	11609

Slika 4 - Deo skupa atributa prve grupe koji predstavlja vrednosti kardioloških testova i analiza



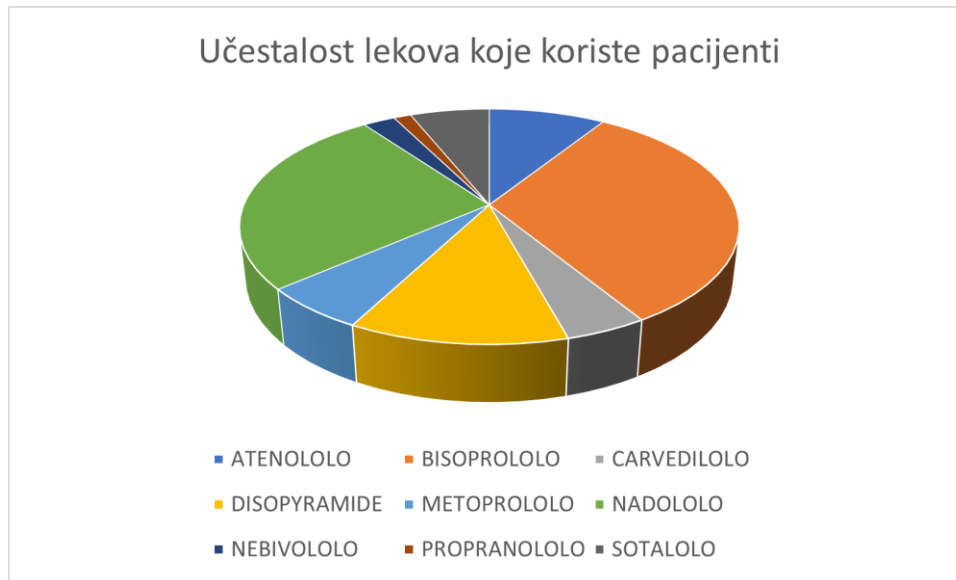
Drugu grupu čine informacije o dijagnozama koje su postavljane pacijentima i datum postavljanja dijagnoze. Prilikom dolaska pacijenta na pregled, postavlja mu se neka od dijagnoza i beleži u odgovarajuću datoteku. Ukoliko je neka dijagnoza već bila postavljena u prošlosti, neće se ponovo beležiti informacije o njoj. Jednom pacijentu može biti postavljeno više dijagnoza tokom lečenja. Kao što možemo videti na slici (Slika 5), među najzastupljenijim dijagnozama su: abnormalni holter, aritmija nsvt i atrijalna fibrilacija.



Slika 5 - Grafički prikaz učestalosti postavljenih dijagnoza

Treću grupu čine informacije o lekovima koje pacijent uzima. Za svakog pacijenta su, uz ime leka, navedeni datumi kada je pacijent počeo i prestao da uzima lek. Prilikom svakog dolaska pacijenta na pregled beleže se informacije da li je pacijent prestao sa korišćenjem nekog leka, ili je počeo da koristi novi lek. U jednom trenutku pacijent može uzimati više lekova. Na osnovu slike (Slika 6) možemo videti da su najviše korišćeni lekovi bisoprolol, koji usporava rad srca i povećava efikasnost srčane pumpe, i nadolol koji se koristi za kontrolu visokog krvnog pritiska.

Podaci ne mogu da se distribuiraju javno zbog zaštite podataka o pacijentima.



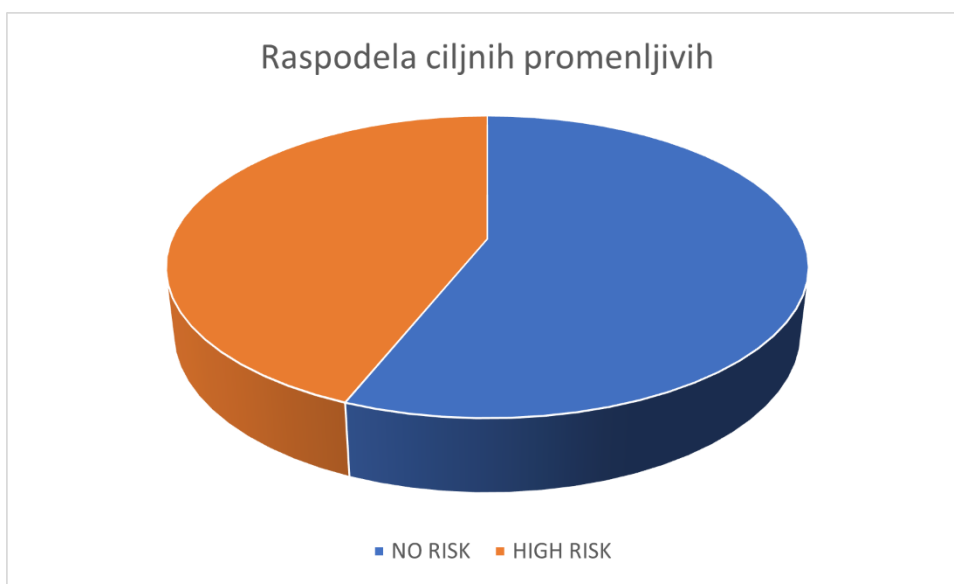
*Slika 6 - Grafički prikaz učestalosti lekova koje koriste pacijenti*

Svakom pacijentu dodeljen je njegov ID, koji se ne menja pri dolasku na ponovni pregled. Na osnovu njega moguće je objediniti ove tri grupe podataka. Objedinjene podatke čini 5840 instanci koje opisuje 130 atributa. Jedna instanca predstavlja jedan dolazak pacijenta na pregled - njegove opšte i zdravstvene informacije, kao u prvoj grupi, kojima su dodate informacije o lekovima koje pacijent koristi, i koliko ih dugo koristi. Na osnovu druge grupe podataka, koju čine informacije o dijagnozama koje su tokom vremena uspostavljane pacijentima, formiraju se indikatori da li je pacijentu bila uspostavljena neka dijagnoza u prošlosti ili da će biti u budućnosti. Objedinjavanjem podataka dobija se kompletna slika stanja pacijenta, gde jedna instanca predstavlja jedan dolazak jednog pacijenta na pregled. Nakon obavljenog pregleda, za svakog pacijenta će postojati informacije o analizama i testovima koji su mu rađeni, da li mu je ranije postavljena neka dijagnoza, da li postoji mogućnost da mu neka dijagnoza bude postavljena u budućnosti, koje lekove pije i koliko dana, da li ima mutacije na nekim od gena koji se smatraju uzrocima oboljenja od hipertrofične kardiomiopatije, kao i opšte informacije o pacijentu. Sve ove informacije o jednom dolasku na pregled su predstavljene jednom instancom u objedinjenom skupu podataka. Ovako objedinjeni podaci predstavljaju informacije o lečenju pacijenata u periodu od 5 godina i oni će kasnije biti korišćeni za odabir modela.

Pored svih navedenih informacija u podacima koji će biti korišćeni, nalaze se i indikatori da li je pacijent niskorizičan ili visokorizičan, odnosno ciljne promenljive. Ciljne promenljive su predstavljene pomoću dva atributa: No\_Risk (niskorizični pacijenti) i High\_Risk (visokorizični pacijenti). Svaki atribut predstavlja indikator na stepen rizika pacijenta. Dakle, ciljne promenljive ne ukazuju na to da li je pacijent bolestan ili ne, jer u skupu podataka nema zdravih pacijenata. Ciljna promenljiva određuje koliko je stanje obolelog pacijenta rizično.

Prilikom rada sa podacima, kao ciljnu promenljivu moguće je uzeti bilo koji od ova dva atributa, obzirom da su komplementni, i unutar njih posmatrati dve klase 0 i 1. Ukoliko bismo kao ciljnu promenljivu uzeli atribut High\_Risk, instance koje pripadaju klasi 1 biće visokorizični pacijenti, a instance koje pripadaju klasi 0 niskorizični pacijenti. Ako bismo kao ciljnu promenljivu uzeli atribut No\_Risk, situacija bi bila obrnuta.

Razlika u raspodeli klasa nije velika, kao što možemo primetiti na osnovu grafika na slici (Slika 7). Treba uzeti u obzir da pri određivanju raspodele klasa nisu uzimane u obzir instance kojima su ID pacijenta i klase iste, odnosno nisu uzimani u obzir pacijenti koji dolaze na kontrolu ako im status nije promenjen. Takođe, treba imati u vidu i da je moguće da je pacijent promenio status tokom lečenja, odnosno da je niskorizičan pacijent u toku lečenja postao visokorizičan, i obrnuto, pa je moguće da se jedan pacijent broji u obe klase.



Slika 7- Grafički prikaz raspodele visokorizičnih i niskorizičnih pacijenata

## 2.2. Priprema podataka

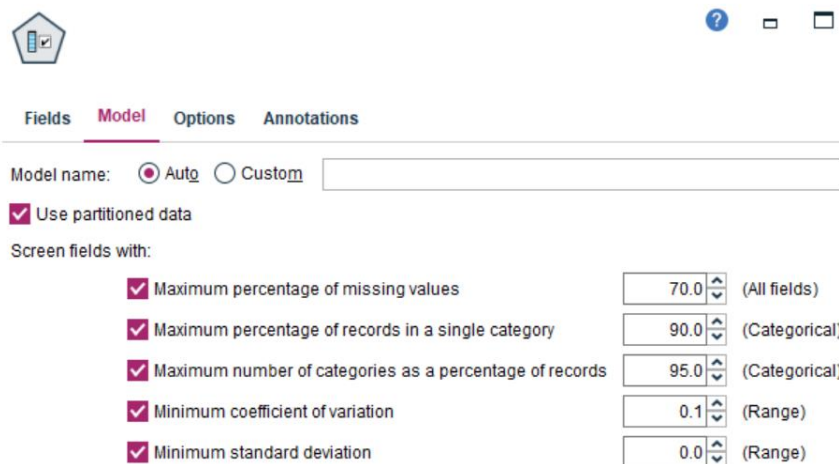
U podacima se nalazi veliki broj instanci sa nedostajućim vrednostima za neki atribut, odnosno nalazi se veliki broj atributa sa visokim procentom nedostajućih vrednosti. Atributi sa velikim brojem nedostajućih vrednosti uglavnom pripadaju grupi atributa koji opisuju rezultate kardioloških testova ili analiza. Iz tog razloga nije zahvalno popunjavati vrednosti koje nedostaju, jer su rezultati testova individualni i nepredvidivi.

Ako bi se nedostajuće vrednosti popunjavale bilo kojom metodom, mogla bi se dobiti pogrešna slika o pacijentu i moglo bi se doći do pogrešnog zaključka. Pored atributa koji imaju veliki procenat nedostajućih vrednosti, postoje atributi koji imaju veliki procenat vrednosti samo jedne kategorije. Takvi atributi takođe nisu od velikog značaja.

Kako bi se otklonili ovakvi, i slični atributi koji ne doprinose mnogo u odlučivanju, pre korišćenja modela podaci su filtrirani. Za pripremu podataka i primenu algoritama biće korišćen alat IBM SPSS Modeler.

U okviru IBM SPSS Modelera korišćen je čvor Feature Selection za filtriranje atributa (Slika 8). Filter napravljen na osnovu specifikacija čvora Feature Selection filtrira attribute koji ne zadovoljavaju ni jedan od sledećih uslova:

- Procenat nedostajućih vrednosti je iznad 70%.
- Procenat zastupljenosti jedne kategorije je iznad 90%.
- Koeficijent varijacije je iznad 0.1.
- Standardna devijacija je ispod 0.
- Procenat vrednosti koje pripadaju različitim kategorijama je iznad 95%.

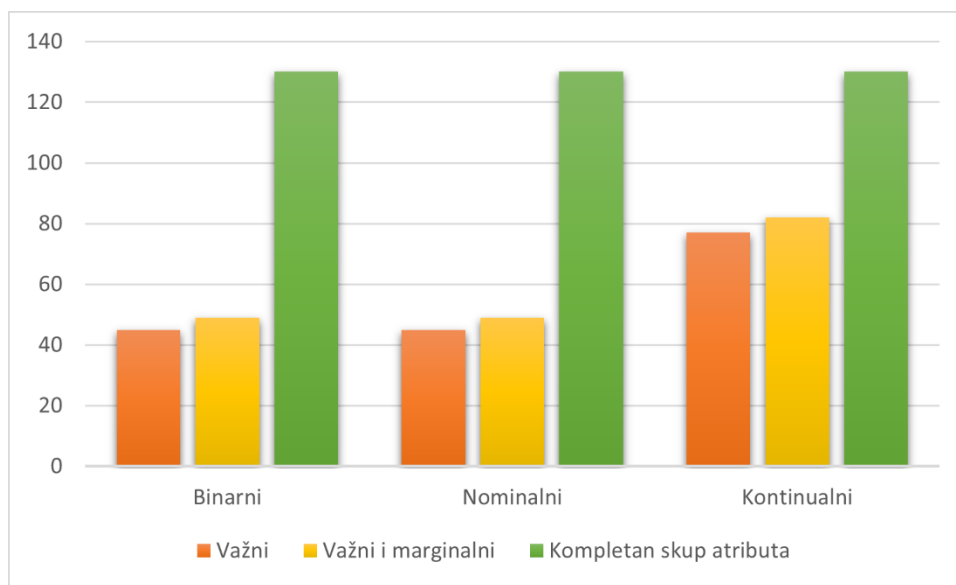


Criteria	Value	Scope
<input checked="" type="checkbox"/> Maximum percentage of missing values	70.0	(All fields)
<input checked="" type="checkbox"/> Maximum percentage of records in a single category	90.0	(Categorical)
<input checked="" type="checkbox"/> Maximum number of categories as a percentage of records	95.0	(Categorical)
<input checked="" type="checkbox"/> Minimum coefficient of variation	0.1	(Range)
<input checked="" type="checkbox"/> Minimum standard deviation	0.0	(Range)

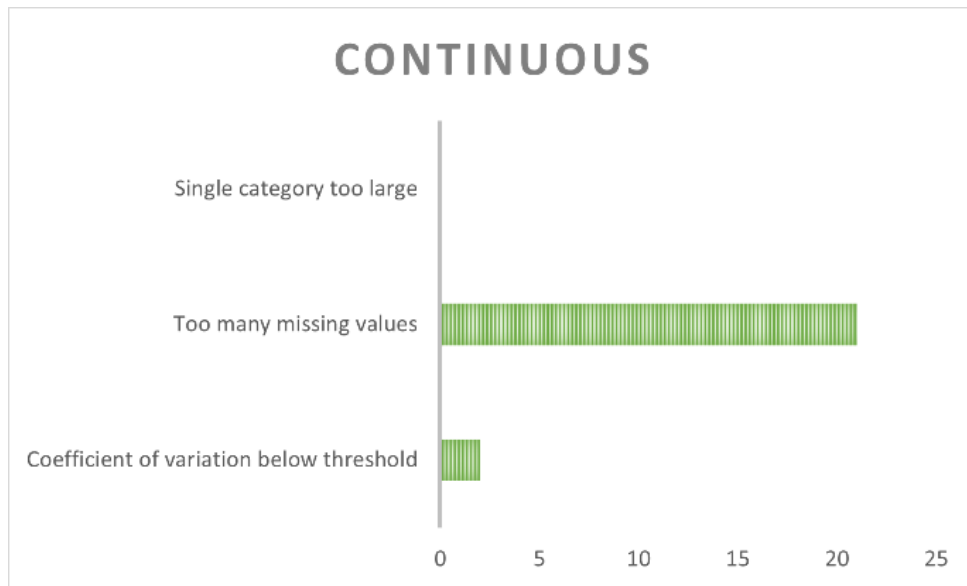
Slika 8 - Primer Feature Selection čvora

Filtrirani skup atributa je rangiran po važnosti korišćenjem Pirsonove mere, koja predstavlja veličinu linearnog odnosa promenljivih. Atributi su rangirani kao važni ( $> 0.95$ ), marginalni ( $\leq 0.95$  i  $> 0.9$ ) i nevažni ( $< 0.9$ ). Za treniranje modela biće korišćena tri skupa atributa. Prvi skup čine samo važni atributi, drugi skup čine važni i marginalni atributi, a treći skup je originalan skup atributa.

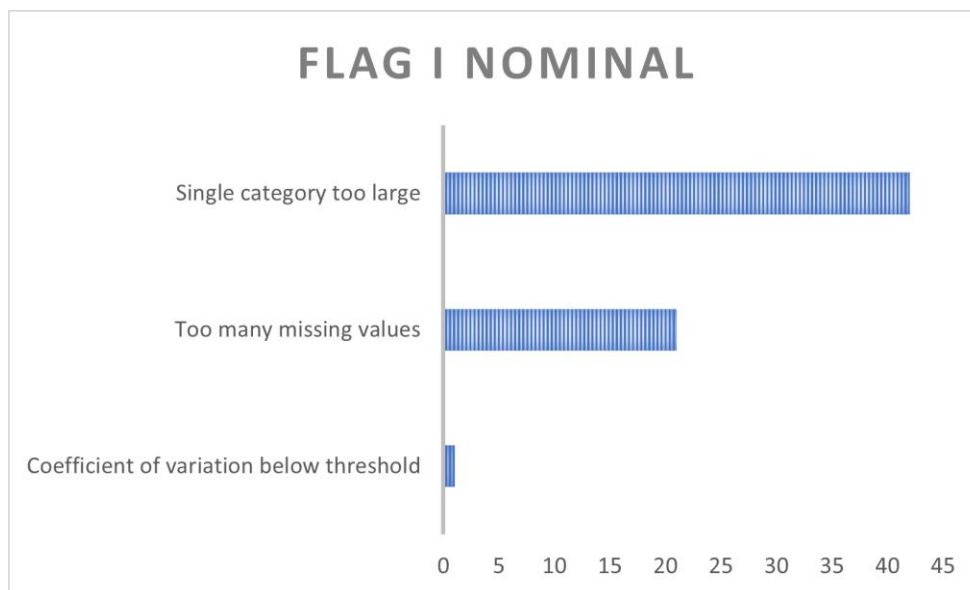
Veliki broj atributa ima vrednosti 0 ili 1, odnosno binarni su. Modeler daje mogućnost da se takvi atributi posmatraju kao različiti tipovi. Odnosno, moguće je posmatrati ih kao binarne (eng. *Flag*), kontinualne (eng. *Continuous*) i nominalne (eng. *Nominal*). U zavisnosti od tipa takvih atributa, menja se skup atributa koji se uključuju u treniranje modela, odnosno, menja se skup atributa koji se dobija primenom filtera (Slika 9). Kao što možemo videti na slikama (Slika 10, Slika 11), ukoliko se ovakvi atributi posmatraju kao binarni ili nominalni, skup atributa koji se isključuje filtriranjem je skoro tri puta veći nego skup atributa koji se isključuje kada bi se ovakvi atributi posmatrali kao kontinualni. Ukoliko bismo posmatrali past i label attribute kao binarne ili nominalne, veliki broj ovih atributa je isključen zato što više od 90% vrednosti pripada jednoj kategoriji. Ovakvi atributi su ili previše zastupljeni, ili nezastupljeni, pa ne utiču dovoljno na to da li je pacijent nisko ili visoko rizičan. Različiti algoritmi rade drugačije u zavisnosti od tipa atributa.



Slika 9 - Grafički prikaz promene broja atributa prilikom filtera u zavisnosti od tipa atributa



*Slika 10 - Atributi koji ne zadovoljavaju kriterijume filtera kada su indikatori kontinualni*

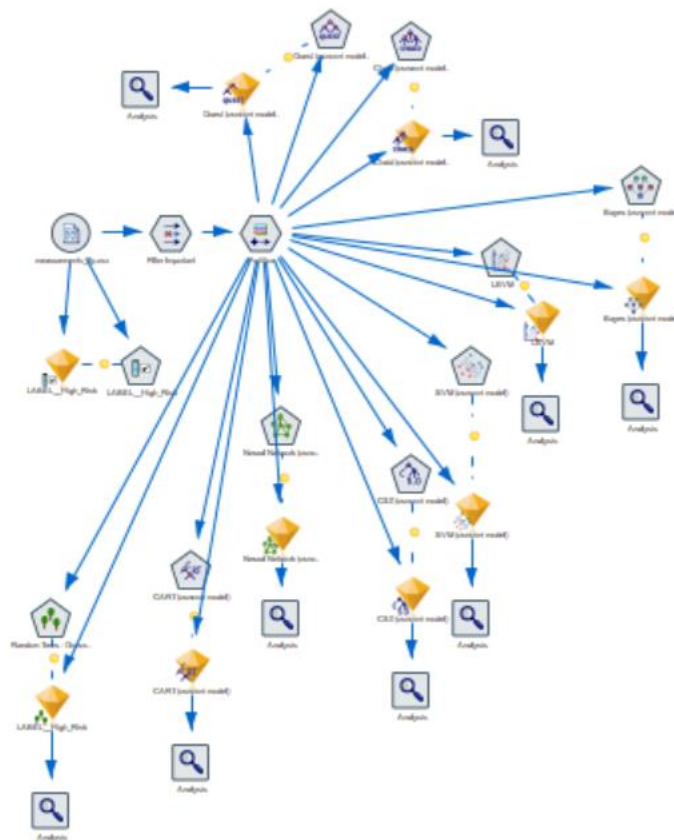


*Slika 11 - Atributi koji ne zadovoljavaju kriterijume filtera kada su indikatori binarni ili nominalni*

### 3. Algoritmi klasifikacije

Nakon što su podaci pripremljeni, i redukovan im je skup atributa, dele se na dva skupa – trening i test skup. Za trening modela koristi se 70% skupa, dok se za test koristi 30% skupa podataka.

Nakon podele podataka na trening i test skup, primenjuju se različiti modeli klasifikacije: Random Trees (slučajna drvetva), CART, C5.0, SVM (metod potpornih vektora), LSVM (linearni metod potpornih vektora), CHAID, QUEST i neuronska mreža.



Slika 12 - Primena osnovnih modela svih algoritama

### 3.1. Drveta odlučivanja

Drveta odlučivanja predstavljaju tip algoritama klasifikacije koji određuju klasu instance postavljanjem pitanja pri prolasku instance kroz strukturu koja ima oblik drveta. Svaki čvor drveta može biti unutrašnji čvor ili list. Ukoliko je čvor list, on određuje klasu instance, u suprotnom, čvor je unutrašnji, i u njemu se postavlja određeno pitanje. Pitanje u unutrašnjem čvoru se odnosi na vrednost instance u jednom atributu, odnosno svaki čvor je vezan za jedan atribut skupa podataka i vrednost instance u tom atributu određuje dalje kretanje instance kroz drvo. Koji atribut će pripadati kom čvoru, i koliko će biti unutrašnjih čvorova, određuje mera čistoće čvora. Cilj svakog čvora je da što bolje podeli skup podataka na klase, odnosno da bude što je moguće čistiji.

Na početku se svakom atributu dodeli čvor i među njima se traži najčistiji čvor kao početni. Iz početnog čvora se drvo grana na dva (ili više) nova čvora. Za svaki novi čvor se bira atribut koji će ga činiti najčistijim u odnosu na podatke koji se u njemu nalaze. Ako je u čvoru zastupljena samo jedna klasa, onda taj čvor postaje list. Drvo prestaje sa razvijanjem kada više nema novih unutrašnjih čvorova.

Drveta mogu imati više podčvorova, ali uglavnom su binarna. Ukoliko je drvo binarno, a atribut ima više od dve vrednosti, ili je kontinualan, kriterijum podele u čvoru se bira tako što se svaka vrednost uzima u obzir i bira se ona koja daje najčistiji čvor. Neke od mera čistoće čvora su Entropija i Ginijev indeks.

Jedan od načina da kontrolišemo klasifikaciju određene klase ili da joj posvetimo više pažnje, je korišćenje matrice cena. Polja matrice predstavljaju cene klasifikacije. Što su cene veće, to je veća kazna loše klasifikacije. Cene mogu biti i negativne i u tom slučaju predstavljaju nagradu za dobru klasifikaciju [3].

		Klasa predviđena modelom	
		Klasa = +	Klasa = -
Stvarna klasa	Klasa = +	-1	100
	Klasa = -	1	0

Slika 13 - Primer matrice cena za klasifikaciju instanci koje mogu pripadati klasama + ili -



Ako je pozitivna instanca dobro klasifikovana, model će biti nagrađen cenom -1. Ako je negativna instanca klasifikovana kao pozitivna, cena loše klasifikacije će biti 1, a u obrnutom slučaju 100. Ovo nam govori da je stavljen veći akcenat na klasifikaciju instanci pozitivne klase.

Matrice cena se uzimaju u obzir pri građenju modela i dobijamo model koji ima najmanju cenu. Najčešće se koriste kod drvetu odlučivanja. U ovakvim algoritmima mogu imati uticaj na izbor atributa za podelu, odlučivanje da li poddrvo treba da se iseče, modifikaciju pravila odlučivanja u svakom čvoru, može da manipuliše trening instancama tako da algoritam konvergira ka drvetu koje ima najmanju cenu [3].

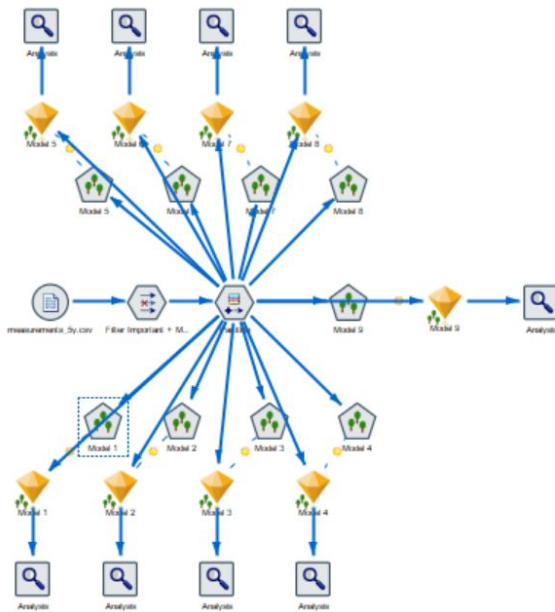
Postoje različiti algoritmi za klasifikaciju koji se zasnivaju na drvetima odlučivanja i razlikuju se po izgledu drveta, načinu rada sa nedostajućim vrednostima, izboru skupa podataka za razvoj drveta i kriterijumima zaustavljanja. Algoritmi koji se zasnivaju na drvetima odlučivanja, koji će biti korišćeni za obradu podataka su: Random Trees, C5.0, CART, CHAID i QUEST.

### 3.1.1. Random Trees

Random Trees je algoritam koji umesto jednog drveta pravi više njih. Koristi tehniku pakovanja (eng. *bagging*) za pravljenje proširenog skupa podataka za trening. Na slučajan način se iz skupa za trening bira određeni broj instanci, što znači da se neka instanca može ponoviti. Ovo se radi za svako drvo, tako da drveća nemaju disjunktne skupove na kojima se treniraju, jer bi u tom slučaju skupovi bili previše mali. Na ovaj način su skupovi na kojima se drveća treniraju veći i dobija se kvalitetniji model [4].

Drveća su binarna, odnosno mogu imati dva podčvora. Kao i u standardnoj proceduri, za svaki čvor se bira atribut koji daje najčistiju poddelu, ali u ovom slučaju ne uzima se u obzir ceo skup atributa. Random Trees uzima nasumičan (eng. *random*) skup atributa i među njima bira najboljeg kandidata za poddelu. Kako drvo raste do maksimalne dubine, i kako se kandidati za unutrašnji čvor biraju na slučajan način, drveća imaju tendenciju da budu veoma duboka. Razlog tome je što se nasumičnim izborom kandidata može preskočiti atribut koji bi dao najčistiji čvor, i na taj način produžiti dubina drveta [4].

Pri radu sa nedostajućim vrednostima, drveća ih sama popunjavaju srednjom vrednošću ili najzastupljenijom kategorijom, ako su u pitanju kategorički atributi. Klasa instance određuje se metodom glasanja. Instanca prolazi kroz sva drveća i dodeljuje joj se klasa koja je najzastupljenija među svim drvetima. Zbog načina uzorkovanja podataka za treniranje drveta, mnogo je manja verovatnoća da će doći do prilagođavanja modela, i da rezultati testiranja neće biti slični rezultatima treniranja [4].



Slika 14 - Modeli algoritma Random Trees

Prilikom korišćenja ovog algoritma napravljeno je devet različitih modela. Modeli se razlikuju po broju drveta, dubini drveta, maksimalnom broju čvorova u drvetu, kriterijumu zaustavljanja, matrici cena i maksimalnoj veličini podčvora. Detaljan opis mogućnosti implementirane verzije algoritma se može naći u [4].

### 3.1.2. CART

Classification And Regression Tree (CART) je još jedan algoritam zasnovan na drvetima odlučivanja.

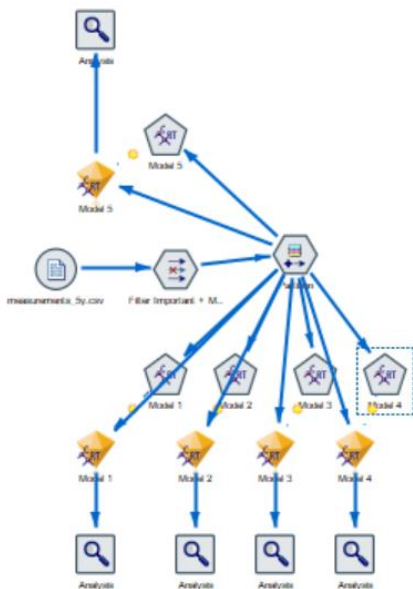
Drveta su isključivo binarna, i podržava opcije pakovanja i pojačavanja (eng. *boosting*) kako bi se povećale stabilnost i preciznost modela. Atributi se u unutrašnjim čvorovima određuju po istom principu, odnosno bira se atribut koji daje najčistiji čvor, a jedna od mere nečistoće čvora koja se koristi je Gini. U obzir za odabir atributa u unutrašnjem čvoru uvek dolaze svi atributi, a instance se kreću po drvetu uvek po istom principu – ako je uslov ispunjen prati se leva grana, ako nije, prati se desna grana. Smatra se da je ovaj princip dobar, jer binarnim drvetom neće doći do prilagođavanja modela test podacima, a opet će biti ispitano više različitih uslova za jedan atribut, jer je moguće više puta birati jedan atribut za unutrašnji čvor [5] [6].

Na ovaj način može doći do razvoja dubljeg drveta nego što je potrebno. Kako bi se ovo sprečilo, nakon što se drvo formira primenjuje se potkresivanje drveta. U odnosu na test podatke potkresuju se grane koje najmanje doprinose odlukama u drvetu. Drveta se mogu skraćivati do određene dubine, ili dok se ne dobije najbolje drvo [5] [6].

Kada su u pitanju nedostajuće vrednosti, koristi metod surogata. Ukoliko se u toku odabira najboljeg atributa za unutrašnji čvor nađe instanca sa nedostajućim vrednostima, ona se zanemari. Ako se prilikom treninga ili testiranja modela pojavi instanca koja ima nedostajuću vrednost u nekom atributu, bira se drugi atribut koji će ga zameniti i odrediti u kom poddrvetu treba da se nađe instanca. Zamenski atribut predstavlja surogat atribut. Za svaki atribut određena je lista surogata. Prvo se bira najbolji, a ako i za njega instanca nema vrednost, bira se prvi sledeći.

U slučaju da instanca nema vrednost ni za jednog surogata, dodeljuje se detetu koje ima više instanci [4] [6] [5].

Prilikom primene CART algoritma napravljeno je 5 različitih modela. Modeli se razlikuju u dubini drveta, matrici cene, broju surogata, kriterijumima zaustavljanja i korišćenju metoda pakovanja i pojačavanja, kao i metoda unakrsne provere. Detaljan opis mogućnosti implementirane verzije algoritma se može naći u [4].



Slika 15 - Modeli algoritma CART

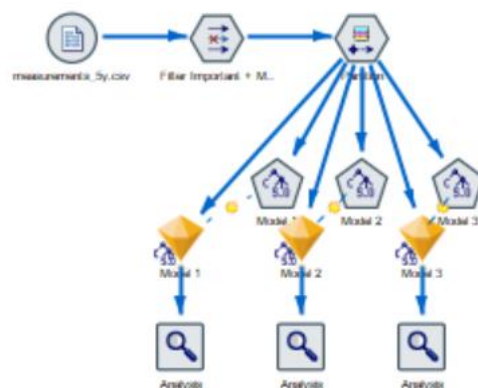
### 3.1.3. C5.0

Algoritam C5.0 predstavlja moderniju verziju algoritma C4.5. C4.5, za razliku od CART algoritma, dopušta nebinarna drveća. Jedan unutrašnji čvor može imati više izlaznih grana, u zavisnosti od broja različitih vrednosti kategoričkih atributa. Može imati jednu granu za svaku vrednost, a može i grupisati više vrednosti na jednoj grani. Kada su u pitanju numerički atributi, bira se određeni interval u odnosu na koji se deli čvor. Primenjuje se isti princip za izgradnju drveća, odnosno bira se atribut koji daje najbolju podelu, ali kod ovog algoritma kriterijum podele nije čistoća čvora, već Entropija i odnos informacione dobiti [7] [5].

Obzirom na to da se na pohlepan način bira najbolji atribut za unutrašnji čvor, i da se atributi mogu ponavljati, moguće da se ponove delovi drveća, da drvo bude previše duboko i preprilagođeno test podacima. Kako bi se to izbeglo, primenjuje se naknadno potkresivanje drveća. Potkresivanje može biti zasnovano na smanjenju nivoa greške, pesimističkoj proceni greške i intervalima poverenja [7] [5].

Kada su u pitanju nedostajuće vrednosti, C4.5 nudi nekoliko načina za prevazilaženje problema. Ukoliko je u pitanju izbor kandidata za čvor, instance koje nemaju vrednost se zanemaruju, a pri proceni kvaliteta atributa može se uzeti u obzir procenat instanci sa nedostajućim vrednostima. Ako je u pitanju klasifikacija instance koja nema vrednost, ta vrednost se može dopuniti (prosečna ili najčešća vrednost), može se odrediti posebna grana za nedostajuće vrednosti, mogu se istražiti sve grane, kombinovati ishodi kako bi se odredila klasa, može se dodeliti svim granama ili se može prekinuti proces i dodeliti najčešća klasa [7] [5].

Pored drveća odlučivanja, moguće je dobiti i skup pravila koja instanca mora da zadovolji kako bi pripadala nekoj klasi. U zavisnosti od korišćenja metoda pojačavanja ili unakrsne provere, ili korišćenja naprednih opcija algoritma, biće napravljena tri različita modela. Detaljan opis mogućnosti implementirane verzije algoritma se može naći u [4].



Slika 16 - Modeli algoritma C5.0

### 3.1.4. CHAID

Chi-squared Automatic Interaction Detection (CHAID) je još jedan od algoritama koji se zasnivaju na drvetu odlučivanja. Glavna karakteristika ovog algoritma je način na koji bira najboljeg kandidata za unutrašnji čvor. CHAID koristi meru drugačiju od ostalih algoritama,  $\chi^2$  meru. Njegov odabir najboljeg kandidata se ne zasniva na određivanju nečistoće čvora, niti dobiti podele, već na statističkoj meri koja traži najznačajniji atribut za podelu. Na ovaj način se vrši potkresivanje drveta pre, odnosno tokom izgradnje drveta [4].

Kao u standardnoj proceduri, prvo se bira koreni čvor, a zatim se procedura ponavlja za svako dete. Drvo koje se dobija ne mora biti binarno. Ovo je jedan od algoritama koji garantuje drvo koje nije binarno, ukoliko je to korisniku potrebno. Kada je u pitanju rad sa nedostajućim vrednostima, ne nudi mnogo opcija kao ostali algoritmi. Moguće je napraviti posebnu granu za nedostajuće vrednosti, što uzrokuje čestim promenama tokom pravljenja modela. Kako bi algoritam što bolje radio, poželjno je da se nedostajuće vrednosti javljaju što dalje od korena, da ne budu u velikom procentu i da ne budu slučajne [4].

I ovaj algoritam daje mogućnost pakovanja i pojačavanja. Pored toga, moguće je podešavati pravila zaustavljanja, određivati minimalan broj instanci u čvoru i menjati druge parametre. Detaljan opis mogućnosti implementirane verzije algoritma se može naći u [4]. Na osnovu vrednosti pomenutih parametara biće formirano pet modela za rad sa podacima.



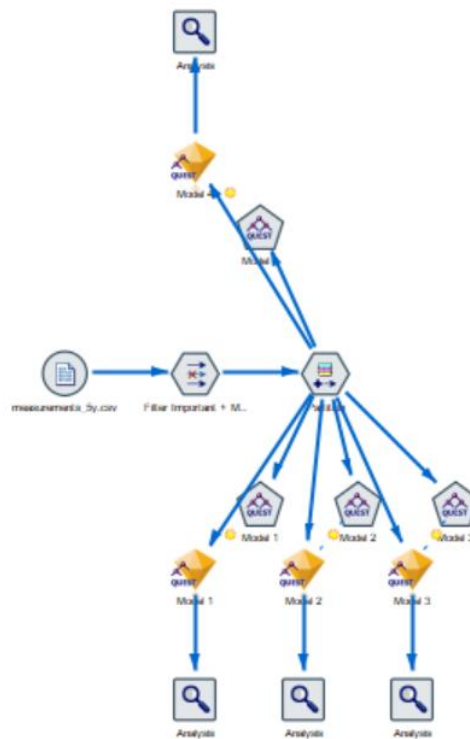
Slika 17 - Modeli algoritma CHAID

### 3.1.5. QUEST

Quick, Unbiased, Efficient Statistical Tree (QUEST) je metod klasifikacije koji gradi drveta odlučivanja koristeći statističke metode. Kao i CHAID koristi statističku metodu da odabere najboljeg kandidata za podelu i za odabir vrednosti u podeli. Za razliku od CHAID -a i CART-a ne uzima u obzir podelu kategorija atributa pri izboru najboljeg kandidata, već to bira naknadno. Za izbor najbolje podele koristi se diskriminantna analiza koja značajno ubrzava proces jer ne proverava sve kombinacije kategorija pri traženju najbolje podele [4].

Gradi isključivo binarno drvo, što se smatra prednošću, jer broj kategorija atributa ne dovodi do velikog drveta. Radi samo sa kontinualnim atributima, tako da pre izbora kandidata, kategoričke attribute prevodi u kontinualne. Kada je u pitanju potkresivanje drveta, pruža iste mogućnosti kao CART algoritam [4].

Pri pravljenju modela moguće je koristiti metode pakovanja i pojačavanja, određivati broj surugata, birati kriterijume zaustavljanja, dubinu drveta i menjati matricu cena. Na osnovu vrednosti pomenutih parametara biće napravljena 4 modela za rad sa podacima, a detaljan opis mogućnosti implementirane verzije algoritma se može naći u [4].

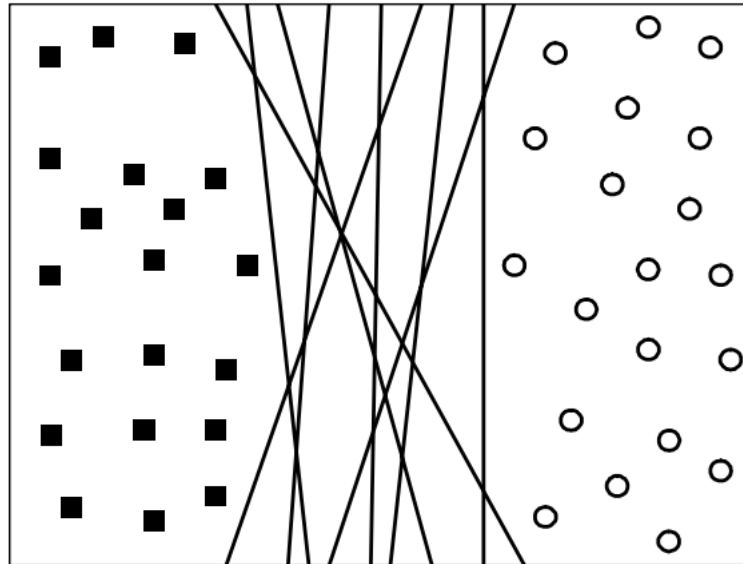


Slika 18 - Modeli algoritma QUEST

### 3.2. Metod potpornih vektora

Ideja metoda je naći hiper ravan koja će najbolje podeliti instance u zavisnosti od klase kojoj pripadaju, a za određivanje same ravni koriste se potporni vektori, odnosno trening instance.

Kao što se može videti na slici (Slika 19), podaci imaju dve klase, i moguće je linearno ih razdvojiti u odnosu na klasu kojoj pripadaju, odnosno moguće je naći hiper ravan koja će to da uradi [3].



Slika 19 - Primer linerano razdvojivih klasa

*Slika je preuzeta iz [3]*

Svaki klasifikator, odnosno hiper ravan, ima margine. Margine određuju koliko bi klasifikator mogao da odstupa od svoje pozicije u nekom od pravaca dok ne naiđe na prvu instancu klase, odnosno margine predstavljaju udaljenost klasifikatora od prve instance klase. One daju fleksibilnost pri malim promenama na klasifikatoru, inače bi promene imale značajan uticaj na klasifikaciju. Tačke, odnosno instance, koje određuju kolika će biti margina nazivamo potpornim vektorima. Ako tako posmatramo, možemo reći da su klasifikatori koji proizvode granice odlučivanja sa malim marginama osetljiviji na preprilagođene modele. Cilj je odabrati hiper ravan sa maksimalnim marginama na osnovu potpornih vektora. Na slici ispod (Slika 20) možemo videti dve hiper ravni B1 i B2 i njihove margine koje se znatno razlikuju u veličini. Veće su šanse da dobro klasifikujemo podatke ako uzmemo hiper ravan B1, nego B2 [3].

Metod potpornih vektora sa različitim jezgrima je zbog velikog broja slogova sa nedostajućim vrednostima u nekom atributu dao jako loše rezultate. Iz tog razloga će biti korišćena verzija algoritma koja koristi linearnu funkciju, odnosno linearni metod potpornih vektora.





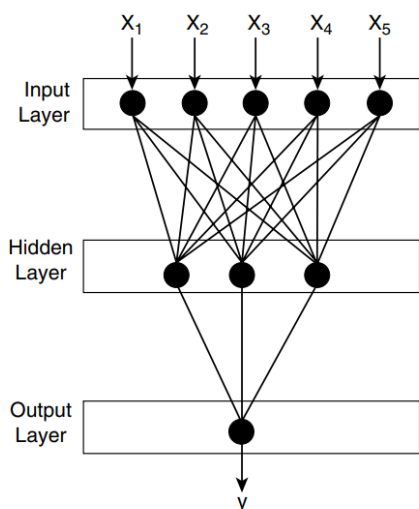
Prilikom korišćenja algoritma, biće napravljena četiri modela koji se razlikuju po funkciji kazne i parametru kazne za greške. Detaljan opis mogućnosti implementirane verzije algoritma se može naći u [4].

### 3.3. Veštačka neuronska mreža

Veštačka neuronska mreža je nastala po ugledu na biološku neuronsku mrežu. Biološka neuronska mreža ima neurone, koji su povezani vlaknima. Ta vlakna prenose impulse od jednog neurona do drugog. Isti je princip i u veštačkoj neuronskoj mreži – imamo čvorove koji su međusobno povezani.

Najjednostavniji oblik veštačke neuronske mreže je perceptron. Perceptron sadrži samo ulazne i izlazne čvorove, odnosno sadrži ulaznih čvorova onoliko koliko ima atributa i jedan izlazni čvor. Uloga ulaznih čvorova je samo da proslede podatke, a uloga izlaznog da odredi klasu podataka. Svaki ulazni čvor je povezan direktno sa izlaznim i svaka ta veza ima određenu težinu. Težine na vezama utiču na određivanje klase podataka [3].

U izlaznom čvoru se na osnovu određene funkcije računa klasa. Ta funkcija se naziva funkcija aktivacije i njen argument je skalarni proizvod podataka i težina u koji je uključen i otklon. Cilj treniranja modela je pronaći najbolje vrednosti za težine veza kako bi klasifikacija bila što je preciznija moguća. Proces treniranja modela funkcioniše tako što se za svaku instancu trening skupa, u više krugova, menjaju vrednosti težina dok se ne nađu najbolje. U svakom krugu se prethodna težina umanjuje ili uvećava u zavisnosti od rezultata klasifikacije i stepena učenja koji određuje u kojoj meri prethodna težina utiče na novu [3].



Slika 22 - Primer višeslojne neuronske mreže

Slika je preuzeta iz [2]

Perceptron omogućava klasifikaciju linearno razdvojivih podataka, a ukoliko to nije moguće, koristi se višeslojna neuronska mreža. Glavne dve razlike višeslojne neuronske mreže u odnosu na perceptron su:

- 1) Višeslojna neuronska mreža ima dodatne unutrašnje slojeve, sa unutrašnjim čvorovima. Unutrašnji čvorovi su povezani tako da samo čvorovi dva susedna sloja mogu biti direktno povezani svaki sa svakim.
- 2) Nema jednu funkciju aktivacije, već više njih. Na korisniku je da izabere neke od funkcija koje mogu biti linearne ili ne [3].

Princip treniranja je isti – traže se najbolje vrednosti za težine veza. Razlika je u funkciji koja računa težine, jer je u slučaju višeslojne mreže kompleksnija. U višeslojnoj mreži se svaki sloj može posmatrati kao poseban perceptron, a izlazni čvor kao čvor koji objedinjuje sve perceptrone [3].

Pri korišćenju veštačke neuronske mreže važno je dobro odabrati strukturu mreže kako ne bi došlo do preprilagođavanja. Mreža dobro rukuje redundantnim podacima, jer su težine automatski naučene tokom treninga i teže da budu male. Osetljive su na prisustvo šuma u trening podacima, traže mnogo vremena za treniranje, ali izgrađen model brzo klasifikuje podatke [3].

Algoritam koji će biti korišćen za rad sa podacima ne daje mogućnost rada sa nedostajućim vrednostima, već ih sam popunjava. Prilikom primene veštačke neuronske mreže na podacima, biće napravljeno pet modela. Modeli se razlikuju po funkciji aktivacije, upotrebi metoda pakovanja i pojačavanja. Detaljan opis mogućnosti implementirane verzije algoritma se može naći u [4].



Slika 23 - Modeli algoritma neuronske mreže

## 4. Dobijeni rezultati

### 4.1. Mere kvaliteta

Nakon treniranja modela, potrebno je odrediti koji model najbolje radi. Jedna od mera za ocenu modela je preciznost. Iako se često koristi, postoje slučajevi u kojima ne možemo dobiti prave informacije o kvalitetu modela samo na osnovu nje. Preciznost predstavlja procenat tačno klasifikovanih instanci. Što je više instanci tačno klasifikovano - model je bolji. Problem kod ovakvog načina ocene kvaliteta modela je što ne uzima u obzir raspodelu klasa u skupu podataka. Zašto to predstavlja problem? Pretpostavimo da 99% podataka čini klasa A, a 1% klasa B. Ako bi naš model klasifikovao dobro klasu A, a skroz promašio klasu B, on bi i dalje imao preciznost 99%. Na osnovu ovakvih rezultata mogli bismo reći da naš model radi jako dobro, ali u stvari on uopšte ne klasifikuje jednu klasu.

Naši podaci imaju dve klase, i na osnovu njih pacijenti mogu biti niskorizični i visokorizični. Ako bismo niskorizičnog pacijenta klasifikovali kao visokorizičnog, on bi verovatno uradio temeljnije analize, započeo proces lečenja i lekari bi u nekom trenutku zaključili da on ipak nije visokorizičan. Ako bismo, u obrnutom slučaju, klasifikovali visokorizičnog pacijenta kao niskorizičnog, neke od posledica bolesti bi možda uticale na njegov život pre nego što bi počeo da se leči. Ovo je razlog zašto nam je bitno da znamo ne samo da li klasifikator greši, već i kod koje klase greši.

Zbog svih ovih problema, važno je koristiti i druge mere za ocenu kvaliteta modela. Neke od njih su AUC i matrica konfuzije.

#### 4.1.1. Matrica konfuzije

Za razliku od preciznosti koja svaku klasu tretira jednako, ova metrika posvećuje pažnju svakoj klasi i pogodna je za analizu modela kada klase nisu jednako zastupljene ili kada nas zanima kakve greške klasifikator pravi. Matrica konfuzije je matrica koja nam daje informacije o tome kako je model klasifikovao instance svake klase i kakve je greške pravio.

U matrici se za svaku klasu nalazi broj dobro klasifikovanih instanci i broj loše klasifikovanih instanci, kao i informacija kojoj klasi je model dodelio loše klasifikovane instance.

		Klasa predviđena modelom	
		Klasa = +	Klasa = -
Stvarna klasa	Klasa = +	$f_{++}$ (TP)	$f_{-+}$ (FN)
	Klasa = -	$f_{+-}$ (FP)	$f_{--}$ (TN)

Slika 24 - Primer matrice konfuzije za instance koje mogu pripadati klasama + ili -

Na slici (Slika 24) vidimo opšti model matrice konfuzije za skup podataka koji ima dve klase. Na osnovu opšteg primera matrice, možemo videti najčešće korišćene oznake:

- Stvarno pozitivno (TP – true positive), što odgovara broju instanci koje pripadaju pozitivnoj klasi i tako su i klasifikovane od strane modela.
- Lažno negativno (FN – false negative), što odgovara broju instanci koje pripadaju pozitivnoj klasi ali ih je model klasifikovao kao negativne.
- Lažno pozitivno (FP – false positive), što odgovara broju instanci koje pripadaju negativnoj klasi ali ih je model klasifikovao kao pozitivne.
- Stvarno negativno (TN – true negative), što odgovara broju instanci koje pripadaju negativnoj klasi i tako su i klasifikovane od strane modela [3].

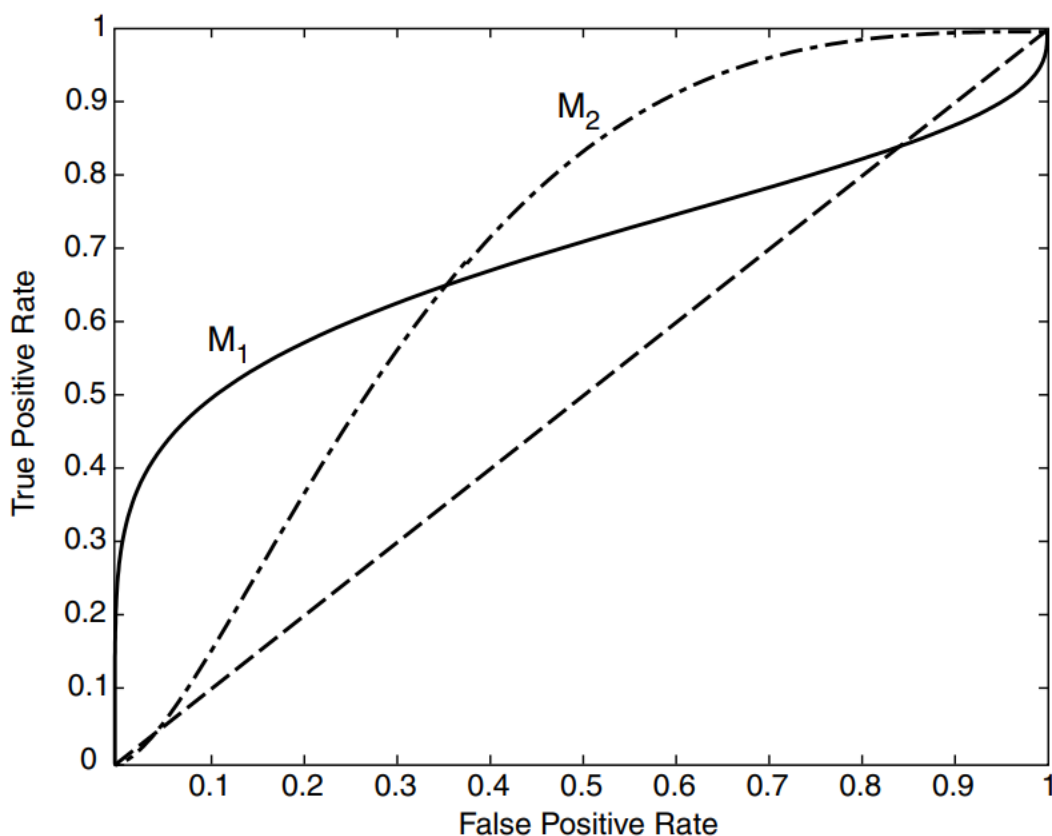
Na osnovu ovih oznaka možemo izvući dodatne, koje predstavljaju mere matrice u procentima:

- Stopa stvarno pozitivnih (TPR – true positive rate).
- Stopa stvarno negativnih (TNR – true negative rate).
- Stopa lažno pozitivnih (FPR – false positive rate).
- Stopa lažno negativnih (FNR – false negative rate).

Primena matrice konfuzije nije ograničena samo na podatke sa dve klase. Dimenzija matrice zavisi od broja klasa u skupu podataka, razlika je samo u tome što postoji više opcija za pogrešno klasifikovane podatke, pa će i oznake biti malo drugačije i biće ih više, ali suština ostaje ista. Korišćenjem ove metrike možemo da vidimo kako model klasifikuje instance važne klase i koliko greši, što nam je od velikog značaja pri ponovnom izboru parametara modela [3].

#### 4.1.2. AUC

AUC (Area Under the ROC Curve) predstavlja prostor ispod ROC krive. ROC (Receiver Operating Characteristic) kriva predstavlja grafički prikaz odnosa stope tačno pozitivnih i stope lažno pozitivnih instanci, tj. TPR i FPR. Pri iscrtavanju krive na y-osi se nalazi stopa stvarno pozitivnih, a na x-osi stopa lažno pozitivnih. Svaka tačka na krivoj odgovara jednom modelu klasifikatora, odnosno jednoj primeni modela. Na slici možemo videti ROC krive dva različita klasifikatora,  $M_1$  i  $M_2$  [3].



Slika 25 - Primer dve ROC krive,  $M_1$  i  $M_2$

Slika je preuzeta iz [3]

Postoji nekoliko kritičnih tačaka na krivoj:

(TPR=0, FPR=0): Model predviđa da svaka instanca pripada negativnoj klasi.

(TPR=1, FPR=1): Model predviđa da svaka instanca pripada pozitivnoj klasi.

(TPR=1, FPR=0): Idealan model.

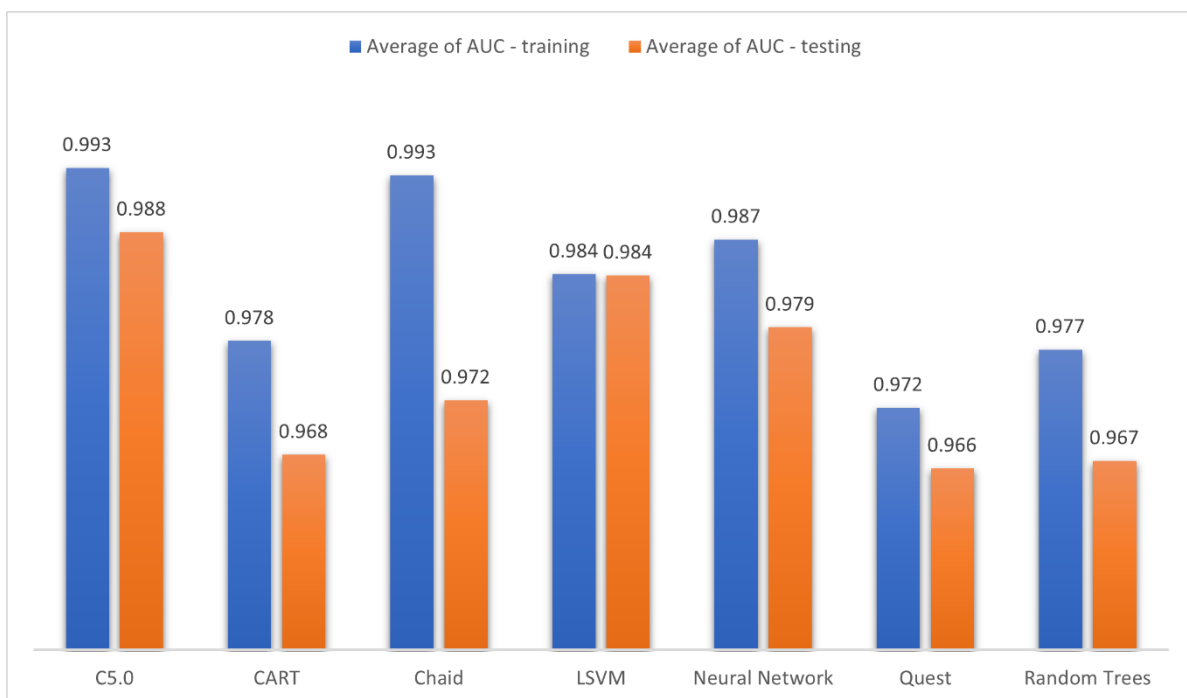
U slučaju idealnog modela sve pozitivne instance su klasifikovane kao pozitivne, i ni jedna negativna instanca nije klasifikovana kao pozitivna, što znači da su sve instance dobro klasifikovane [3].

Ako je model savršen, vrednost AUC će biti 1. Što je model bolji, veću će površinu kriva zahvatati. Ako model slučajno pogađa, vrednost AUC će biti 0.5. Ako poredimo dva modela, strogo bolji model će imati veću vrednost AUC.

## 4.2. Globalni kvalitet algoritama

Prethodno opisane mere korišćene su kako bi se procenio kvalitet korišćenih algoritama. U procenu kvaliteta algoritama je pored preciznosti uzimana u obzir matrica konfuzije kao jedna od najbitnijih mera. Razlog za to je pomenuta priroda podataka i važnost da visoko rizični pacijenti budu prepoznati kao takvi. Obzirom na to da softver daje mogućnost da atribute, indikatore postojećih dijagnoza u prošlosti ili budućnosti, posmatramo kao tri različita tipa podataka, i da se u zavisnosti od tipa koji odaberemo menja i skup bitnih atributa koji se koriste za treniranje modela, performanse modela mogu zavistiti od tipa tih atributa.

Za početak ćemo uzeti globalni slučaj za poređenje algoritama. U globalnom slučaju nije nam bitno da li su nam indikatori kontinualni, nominalni ili binarni, kao ni da li smo koristili samo atribute koji su procenjeni kao važni za klasifikaciju ili sve. Globalni slučaj je rad algoritma na svim mogućim kombinacijama uslova i procena njegovog kvaliteta na osnovu AUC mere.

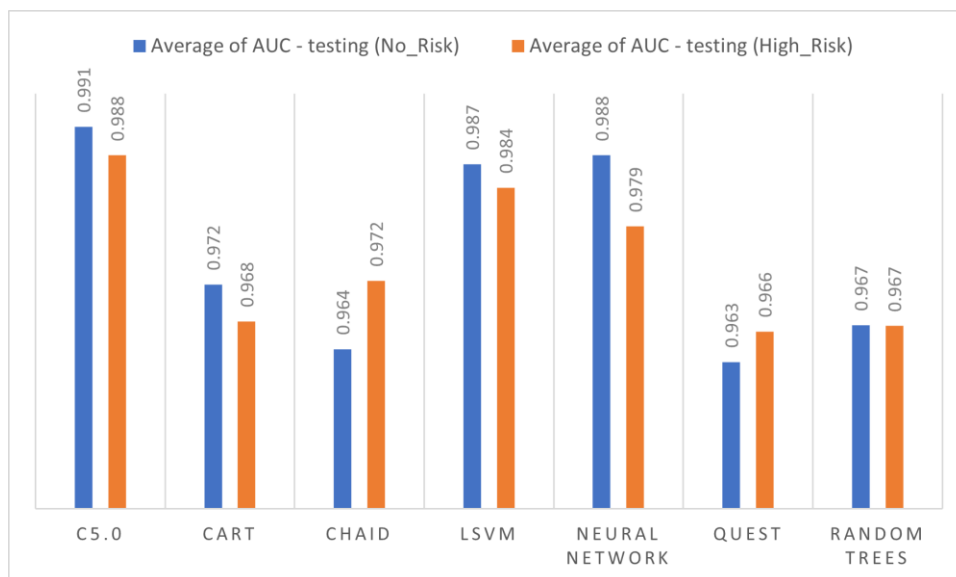


Slika 26 - Grafički prikaz odnosa performansi algoritama na trening i test podacima na osnovu AUC mere

Na slici (Slika 26) prikazan je globalni uspeh algoritama meren AUC merom. Prilikom procene uspeha algoritama uzeti su u obzir rezultati iz svih strimova, gde svaki strim predstavlja jednu kombinaciju uslova (tip atributa koji predstavljaju indikatore određene bolesti i skup atributa). U svakom strimu je za uspeh algoritma uzet model koji je dao najbolje rezultate. Detaljnije informacije o performansama algoritma mogu se naći u tabelama u dodatku A.

Pored algoritama koje vidimo na slici (Slika 26), u planu je bilo i korišćenje Bajesove mreže poverenja. Bajesova mreža poverenja je dala jako loše rezultate, jer nije mogla da klasifikuje instance koje imaju nedostajuće vrednosti, što je razlog zašto nije dalje korišćena i prikazana u radu. Pored Bajesove mreže, ni Metod potpornih vektora (SVM) nije mogao da radi sa nedostajućim vrednostima, ali zato linearna verzija ovog metoda (LSVM) jeste i dala je dobre rezultate. Algoritam koji se pokazao kao najbolja opcija u globalnom slučaju je C5.0, što je bilo i očekivano. U radu sa ovakvim podacima u najvećem broju slučaja najbolje rezultate daju algoritmi zasnovani na drvetima odlučivanja i neuronska mreža. Vidimo i da neuronska mreža nije mnogo lošija, čak je i bolja od ostalih algoritama zasnovanim na drvetima odlučivanja. Pri poređenju rezultata algoritama ova dva tipa, ipak treba uzeti u obzir da neuronska mreža sama popunjava nedostajuće vrednosti, dok većina algoritama zasnovanih na drvetima odlučivanja ima mehanizme za rad sa podacima kojima fale vrednosti bez da ih popunjavaju. Sem linearnog metoda potpornih vektora, rezultati algoritama na test podacima se ne razlikuju mnogo od rezultata na trening podacima, pa možemo zaključiti da model dobro radi, odnosno da nije prilagođen samo trening podacima.

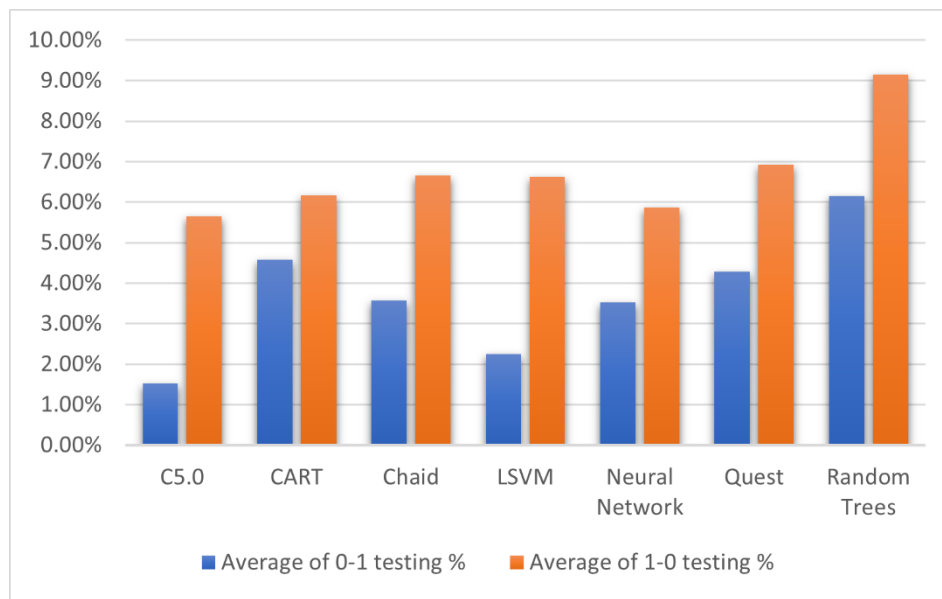
Prilikom korišćenja algoritama, nije uvek uzimana ista ciljna promenljiva. Kako je ciljna promenljiva, odnosno rizičnost, predstavljena preko dva atributa, pri korišćenju algoritama imamo dva slučaja. Prvi slučaj je kada je ciljna promenljiva No\_Risk, a drugi kada je ciljna promenljiva High\_Risk. Na slici (Slika 27) možemo videti performanse algoritama predstavljene AUC merom u odnosu na ciljnu promenljivu, i možemo zaključiti da algoritmi rade približno jednako.



Slika 27 - Grafički prikaz odnosa performansi algoritama na skupovima sa High\_Risk i No\_Risk ciljnom promenljivom, izražen AUC merom



U skupu podataka za test imamo 975 instanci klase 0 (niskorizični pacijenti) i 796 instanci klase 1 (visokorizični pacijenti). Prilikom klasifikacije neke instance koje pripadaju klasi 0 biće klasifikovane kao instance klase 1, i obrnuto. Uzimajući u obzir ceo skup uslova za korišćenje algoritama, na slici (Slika 28) možemo videti prosečan broj loše klasifikovanih instanci na globalnom nivou, izražen u procentima. Odnosno, možemo videti procenat instanci koji pripada klasi 0, a klasifikovan je kao klasa 1, i obrnuto.



Slika 28 - Grafički prikaz proseka loše klasifikovanih test instanci izražen u procentima

Možemo videti da neki algoritmi zasnovani na drvetima odlučivanja prave razliku u pogrešnom klasifikovanju instanci klasa 0 i 1, iako pružaju mogućnost korišćenja matrice cena. Matrica cena nam može pomoći pri usmeravanju klasifikatora na neku klasu, ali nam ne može omogućiti da u potpunosti dobro klasifikujemo instance te klase. Ako bismo napravili model koji uvek dobro klasifikuje instance jedne klase, velika je verovatnoća da bi bio prilagođen samo toj klasi i da bi u velikoj meri loše klasifikovao instance druge klase.

### 4.3. Procena kvaliteta po algoritmima

Atributi koji predstavljaju indikatore na neku od dijagnoza, ili atributi koji imaju identičan skup vrednosti kao indikatori, mogu biti različitih tipova. U zavisnosti od toga da li ih posmatramo kao kontinualne, nominalne ili binarne, mogu biti isključeni zato što imaju preveliki procenat instanci jedne kategorije, ili ne ispunjavaju neki od ostalih kriterijuma pri analizi značajnih atributa. Skup atributa je veći za 50% kada su atributi kontinualnog tipa, nego kada su binarni ili nominalni.

Pored razlike u tipu ovih atributa, postoji razlika i u skupu atributa koje biramo za trening modela. Možemo odabrati samo attribute koji su rangirani kao važni za predviđanje klase, možemo i važnim dodati marginalne. Možemo i uzeti sve attribute u obzir, bez obzira na njihovu korelaciju sa ciljnom promenljivom, ili na kriterijume koje zadovoljavaju.

Uzimajući u obzir sve kombinacije, skup uslova nad kojim se primenjuje svaki od algoritama čine sledeći uslovi:

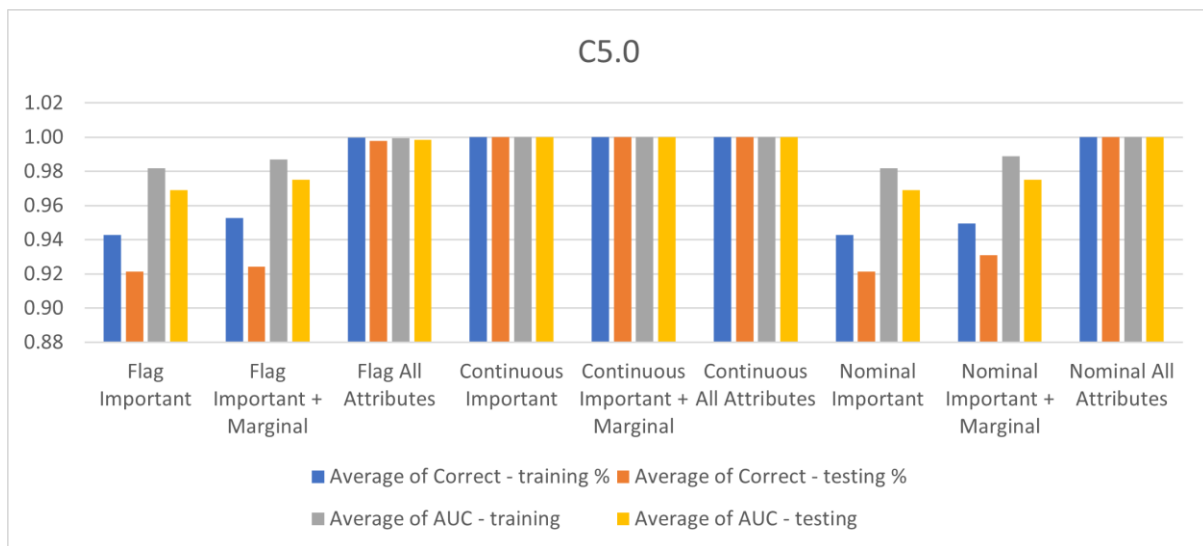
- Tip indikatora je binaran, skup atributa čine samo važni atributi – Flag Important.
- Tip indikatora je binaran, skup atributa čine važni i marginalni atributi – Flag Important + Marginal.
- Tip indikatora je binaran, skup atributa čine svi atributi – Flag All Attributes.
- Tip indikatora je kontinualan, skup atributa čine samo važni atributi – Continuous Important.
- Tip indikatora je kontinualan, skup atributa čine važni i marginalni atributi – Continuous Important + Marginal.
- Tip indikatora je kontinualan, skup atributa čine svi atributi – Continuous All Attributes.
- Tip indikatora je nominalan, skup atributa čine samo važni atributi – Nominal Important.
- Tip indikatora je nominalan, skup atributa čine važni i marginalni atributi – Nominal Important + Marginal.
- Tip indikatora je nominalan, skup atributa čine svi atributi – Nominal All Attributes.

U nastavku će za svaki algoritam biti predstavljene performanse u odnosu na skup uslova. Prilikom predstavljanja rezultata, za svaki skup uslova će biti predstavljene performanse najboljeg modela algoritma čije se performanse predstavljaju. Kako se za svaki skup uslova, odnosno skup podataka, bira najbolji model nekog algoritma, moguće je da će se među predstavljenim performansama nekog algoritma naći rezultati više različitih modela. Obzirom da smo videli da su performanse algoritama približno jednake bez obzira koju ciljnu promenljivu koristimo (Slika 27), za predstavljanje performansi algoritama biće korišćeni strimovi u kojima je ciljna promenljiva u skupu podataka atribut High\_Risk. Detaljnije informacije o performansama svih algoritama se mogu naći u tabelama u dodatku A.

#### 4.3.1. C5.0

Prilikom primene algoritma C5.0 napravljena su tri modela. Na osnovu tabela u dodatku A možemo videti da je od ova tri modela najbolji bio model 1. Model 1 koristi metod pojačavanja praveći 10 modela, a kao meru kvaliteta modela koristi preciznost. Ovaj model je jedini od tri modela koji koristi metod pojačavanja, što je razlog njegovih dobrih performansi, iako su i ostala dva modela imala približno dobre rezultate. U nastavku će biti prikazani rezultati ovog modela na različitim skupovima podataka.

Ako pogledamo sliku (Slika 29) i prvo obratimo pažnju na performanse algoritma kada je u pitanju samo skup atributa, možemo videti da algoritam radi odlično kada se koristi kompletan skup atributa. Kada se model primenjuje na filtriranom skupu atributa, performanse modela zavise od tipa atributa koji predstavljaju indikatore i nekoliko kardioloških testova. Ukoliko su atributi kontinualnog tipa, algoritam takođe radi odlično, ali ukoliko su atributi nominalnog ili binarnog tipa, performanse algoritma su lošije. Dakle, možemo zaključiti da model odlično radi kada su atributi kontinualnog tipa ili kada se koristi ceo skup atributa, dok u ostalim uslovima to nije slučaj.



Slika 29 - Grafički prikaz performansi algoritma C5.0

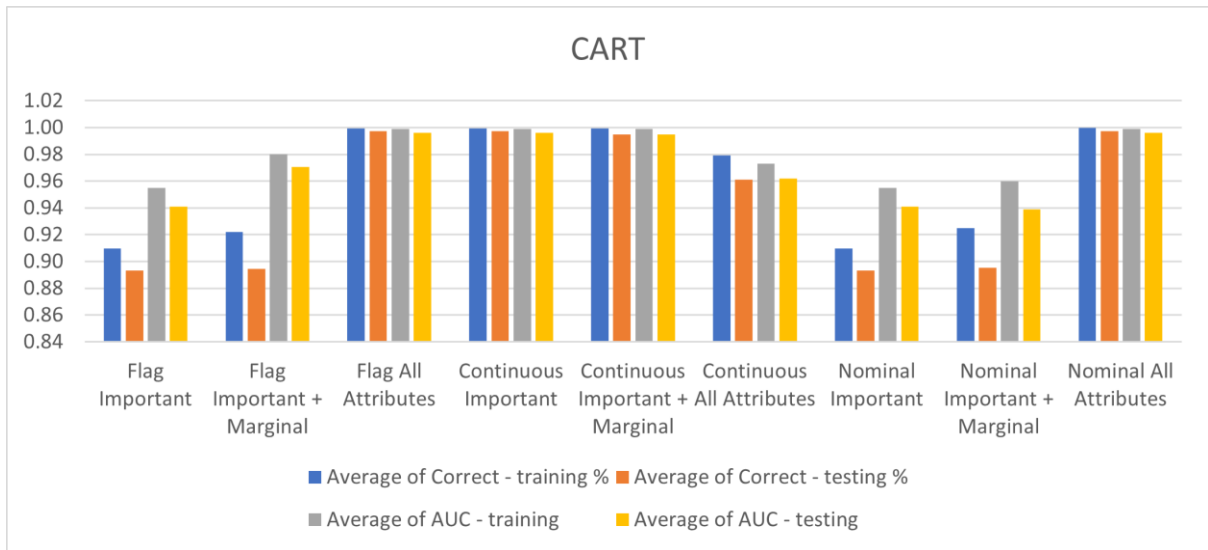
#### 4.3.2. CART

Od pet modela, koliko je napravljeno prilikom primene algoritma CART, najbolje rezultate su dali model 4 i model 5. Model 4 je češće izdvajan kao najbolji model u strimovima gde je ciljna promenljiva High\_Risk, dok je model 5 češće izdvajan kao najbolji model u strimovima gde je ciljna promenljiva No\_Risk.

Kod modela 4 maksimalna dubina drveta je 8, primenjuje se potkresivanje, a broj surogata je 10. Za zaustavljanje su uzeti kriterijumi da se u roditeljskoj grani nalazi najviše 2%, a grani deteta 1% instanci. Koristi se metod pojačavanja, a za odabir klase instance koristi se metod glasanja. Razlika modela 5 u odnosu na model 4 je u broju surogata, koji u modelu 5 iznosi 5, i u matrici cena, gde cena klasifikacije instanci klase 1 kao instance klase 0 iznosi 1,5, dok je vrednost cene obrnute greške u klasifikaciji 1. Razlog boljih performansi modela 5 u odnosu na model 4 na podacima gde je ciljna promenljiva No\_Risk može biti broj instanci koje pripadaju klasama 1 i 0. Kod podataka gde je ciljna promenljiva High\_Risk, klasi 0 pripada veći broj instanci nego klasi 1. Kod podataka kojima je ciljna promenljiva No\_Risk veći broj instanci će pripadati klasi 1, pa će samim tim i forsiranje klasifikacije instanci te klase matricom cena dati bolje rezultate, nego klasifikacija algoritmom koji ne koristi matricu cena.

Na slici (Slika 30) su u okviru predstavljanja performansi algoritma predstavljene performanse i modela 4 i modela 5. Model 5 je izdvojen kao najbolji na skupovima podataka u kojima su indikatori binarnog ili kontinualnog tipa, a skup atributa čine samo važni i marginalni atributi. Model 4 je izdvojen kao najbolji u ostalih 7 slučajeva. Takođe, treba uzeti u obzir da su u ovom delu rada prikazani rezultati samo za skupove podataka u kojima je ciljna promenljiva atribut High\_Risk, jer, kako smo već videli, algoritmi približno isto rade nezavisno od odabrane promenljive. Detaljnije informacije o tome na kom skupu podataka se koji model bolje pokazao se mogu naći u tabelama u dodatku A.

Kao što možemo videti (Slika 30), CART algoritam ima performanse slične algoritmu C5.0, odnosno skoro odlično radi kada su indikatori kontinualni ili skup atributa potpun, a lošije radi kada su indikatori binarni ili nominalni, a skup atributa je filtriran. Razlika u odnosu na performanse algoritma C5.0 je u slučaju kada su indikatori kontinualni, a skup atributa kompletni. U tom slučaju, performanse algoritma CART su za par procenata slabije.

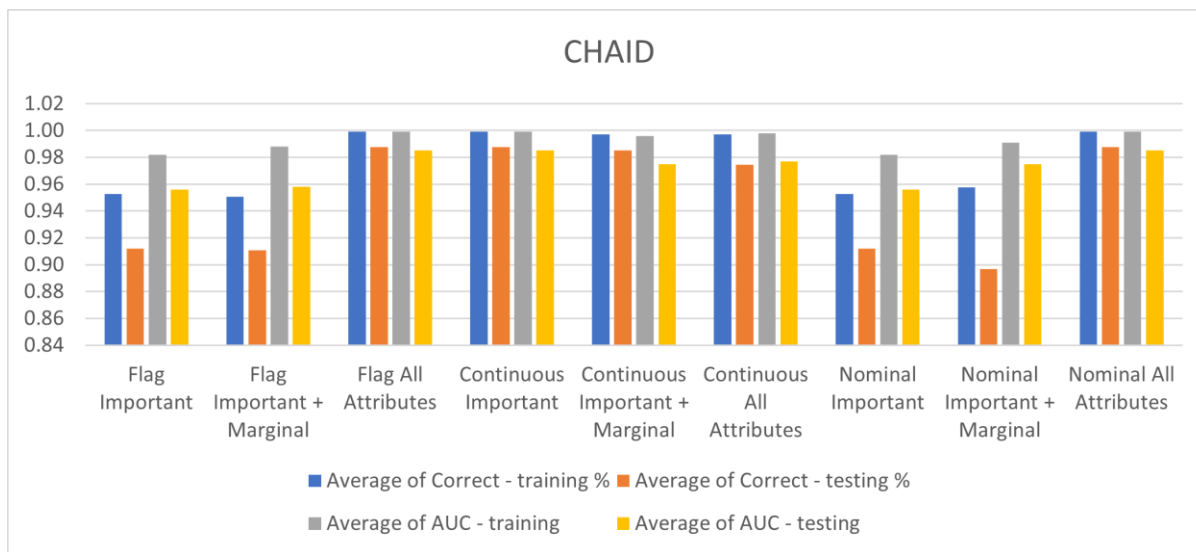


*Slika 30 - Grafički prikaz performansi algoritma CART*

### 4.3.3. CHAID

Kod CHAID algoritma, od pet modela najbolje performanse su imali model 1 i model 5. Model 1, koji se više puta pojavljivao kao najbolji model, za algoritam formiranja drveća koristi CHAID algoritam, a ne Exhaustive CHAID, maksimalna dubina drveta je 5, ne koristi matricu cena, koristi metod pojačavanja gde je broj drveta 10 i klasa instance se određuje glasanjem. Pored ovih opcija, postoje i napredne opcije o kojima više možete pogledati u nekom od strimova. Jedina razlika modela 5 u odnosu na model 1 je u matrici cena. Model 5 za klasifikaciju instanci koje pripadaju klasi 1 kao instance koje pripadaju klasi 0 koristi kaznu u vrednosti od 1,5, dok je vrednost kazne za obrnutu grešku 1. Moglo bi se pretpostaviti da je razlog boljih performansi modela 5 u odnosu na model 1 isti kao u slučaju algoritma CART, ali to bi bila loša pretpostavka jer se u ovom slučaju model 5 javlja kao bolji model kod obe varijante ciljne promenljive.

Kada su u pitanju performanse algoritma u odnosu na prethodno pomenut skup uslova, na osnovu slike (Slika 31) možemo zaključiti da efikasnost algoritma zavisi od istih uslova kao i efikasnost algoritma C5.0. Jedina razlika je u samom kvalitetu. CHAID ne daje odlične rezultate kao C5.0, ali ne daje ni mnogo lošije, jer razlika nije veća od 3%. Kada su u pitanju uslovi pri kojima algoritam daje lošije rezultate, odnosno binarni ili nominalni indikatori i filtriran skup atributa, vidimo da je takođe razlika svega nekoliko procenata u odnosu na algoritam C5.0.

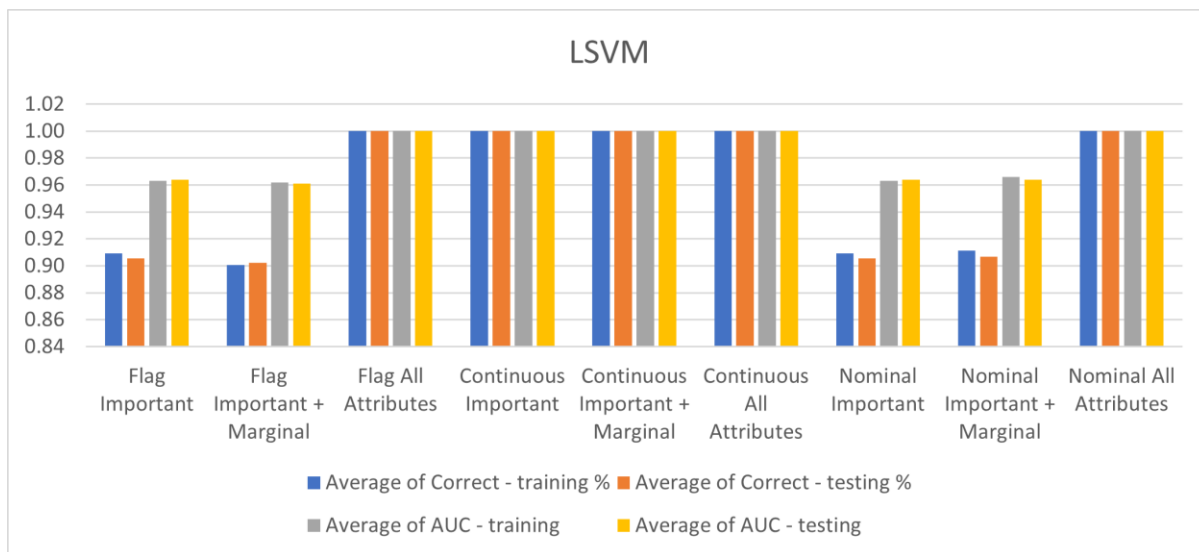


Slika 31 - Grafički prikaz performansi algoritma CHAID

#### 4.3.4. Linearni metod potpornih vektora

Od četiri modela, koliko je napravljeno korišćenjem linearnog metoda potpornih vektora (LSVM), najviše su se istakli modeli 2 i 3. U najvećem broju strimova se ova dva modela izdvajaju zajedno, ali se u nekim strimovima izdvaja kao najbolji samo jedan od ova dva modela. Oba modela kao vrednost parametra greške (lambda parametar) imaju vrednost 0,1, razlika između ova dva modela je u funkciji greške koja se koristi. Model 2 koristi funkciju L2, dok model 3 za funkciju greške koristi funkciju L1.

Performanse linearnog metoda potpornih vektora su identične performansama algoritma C5.0. Performanse na isti način zavise od uslova kao i kod algoritma C5.0. Oba algoritma daju i odlične i lošije rezultate pod istim uslovima. Jedna od razlika je u lošijim rezultatima, gde se lošiji rezultati linearnog metoda potpornih vektora razlikuju za manje od 2% od lošijih rezultata algoritma C5.0, uzimajući, naravno, iste uslove u obzir. Druga razlika je takođe u lošijim rezultatima, ali ne konkretno u razlici vrednosti između ova dva algoritma, već u razlici rezultata na trening i test podacima. Kao što možemo videti na slici (Slika 32), kod linearnog metoda potpornih vektora vrednosti AUC mere i preciznost su isti na trening i test skupovima podataka, odnosno algoritam radi jednako dobro na oba skupa podataka, dok kod C5.0 to nije slučaj. Za sada je linearni metod potpornih vektora najpribližniji algoritmu C5.0.

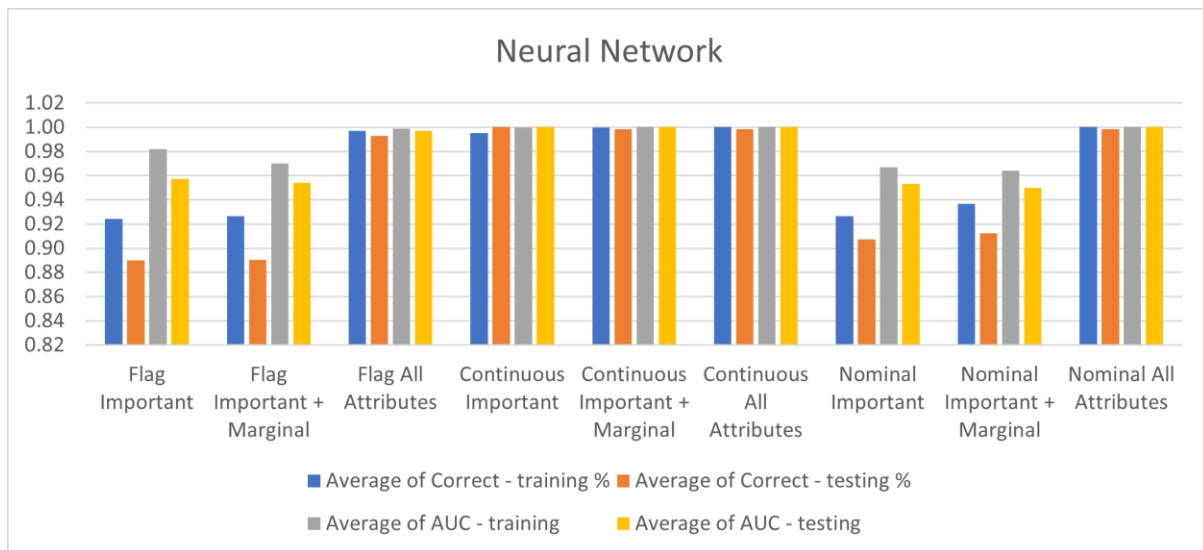


Slika 32 - Grafički prikaz performansi linearnog metoda potpornih vektora

#### 4.3.5. Neuronska mreža

Kod neuronske mreže je slična situacija kao kod linearnog metoda potpornih vektora kada je odabir najboljih modela u pitanju. Od pet modela koliko je pravljeno pri radu sa podacima, najviše se kao najbolji modeli u strimovima izdvajaju zajedno model 4 i model 5, dok se u manjem broju strimova izdvajaju zasebno. Oba modela su višeslojna neuronska mreža, koriste metod glasanja pri određivanju klase instance i prestaju sa radom ukoliko je vreme rada algoritma duže od 15 minuta. Razlika ova dva modela je u metodama pakovanja i pojačavanja. Model 4 koristi metod pojačavanja sa 20 modela, a model 5 metod pakovanja sa 10 modela. Na osnovu rezultata u tabelama u dodatku A, može se videti da oba modela rade identično.

Kada su u pitanju performanse neuronske mreže (Slika 33), one su odlične kada su indikatori kontinualni, ili je skup atributa kompletan. Kada su indikatori nominalni ili binarni, a skup atributa filtriran, performanse su lošije, kao i kod svih algoritama do sada. Gledajući dobre rezultate, neuronska mreža se nalazi odmah iza linearnog metoda potpornih vektora, ali ako bismo gledali lošije rezultate, neuronska mreža se nalazi između algoritma C5.0 i linearnog metoda potpornih vektora.



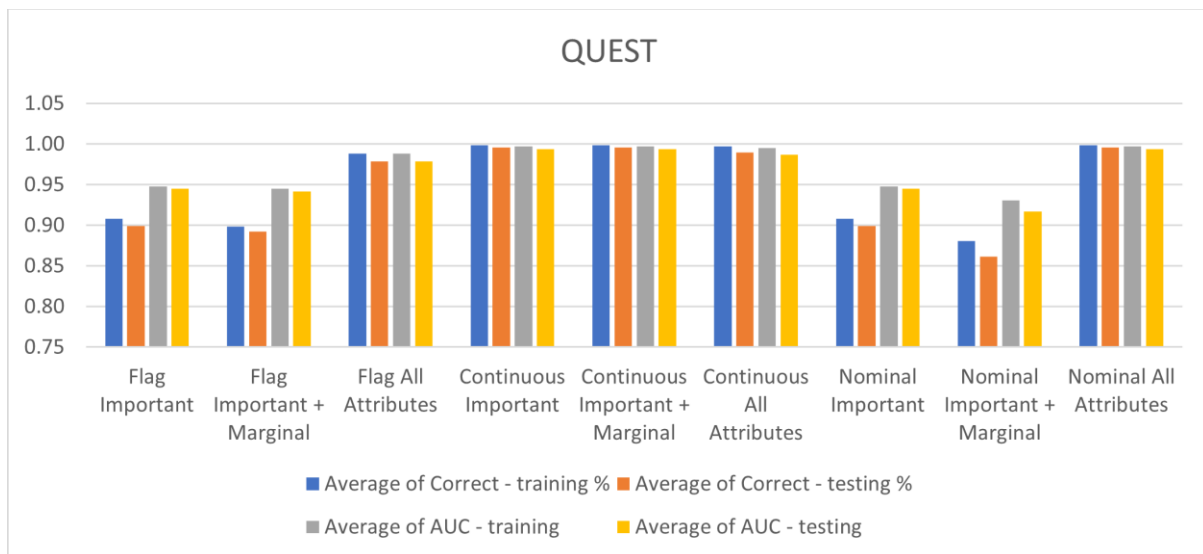
Slika 33 - Grafički prikaz performansi neuronske mreže



#### 4.3.6. QUEST

Od četiri modela, koliko je napravljeno korišćenjem algoritma QUEST, u najvećem broju strimova kao najbolji model izdvojio se model 1. Pored modela 1, u nekoliko strimova se kao najbolji model izdvojio model 4. Model 1 koristi metod pojačavanja praveći deset modela, maksimalna dubina drveta je pet, kao i maksimalan broj surogata, klasa instance se određuje glasanjem, a kriterijumi zaustavljanja su kao kod algoritama C5.0 i CART, ako se u roditeljskoj grani nalazi manje od 2% instanci skupa, ili ako se u grani deteta nalazi manje od 1% instanci. Model 4 je vrlo sličan modelu 1. Jedina razlika između ova dva modela je u tome što model 4 koristi matricu cena. Cena klasifikacije instance klase 1 u klasu 0 iznosi 3, dok u obrnutom slučaju cena klasifikacije iznosi 1.

Kao što možemo videti na slici (Slika 34), algoritam pokazuje dobre i loše rezultate pod istim uslovima kao i svi ostali rezultati. U slučaju binarnih indikatora i kompletnog skupa atributa, rezultati odstupaju za do 2% od odličnih, dok su u ostalim slučajevima performanse algoritma QUEST odlične. Kada su u pitanju lošiji rezultati, ovaj algoritam ima najlošije rezultate do sada, a najmanja vrednost zabeležena je na skupu podataka u kom su indikatori nominalnog tipa, a skup atributa filtriran na važne i marginalne attribute.

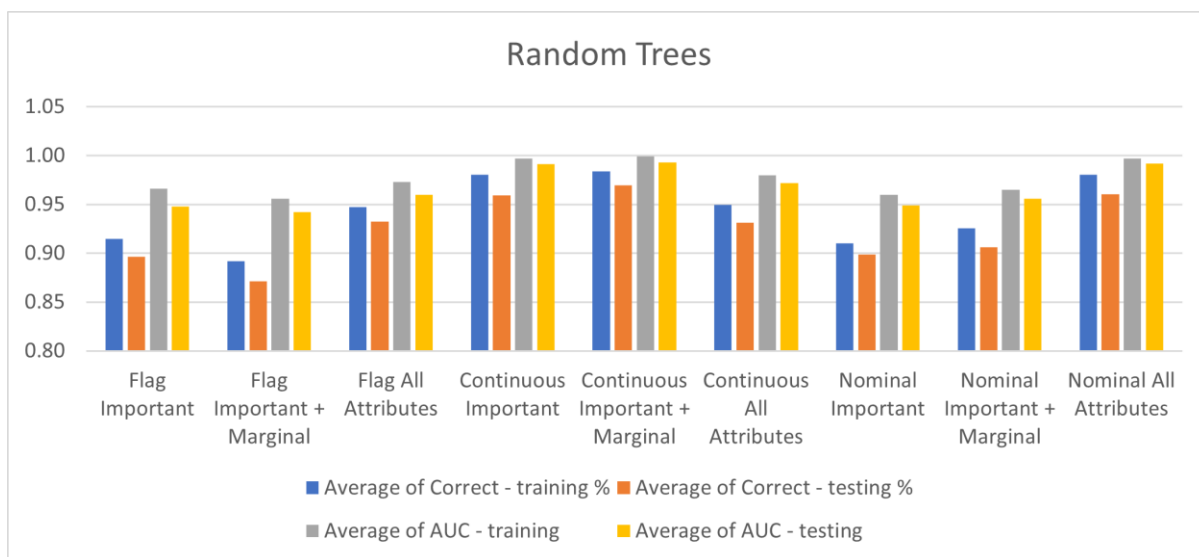


Slika 34 - Grafički prikaz performansi algoritma QUEST

#### 4.3.7. Random Trees

Prilikom korišćenja algoritma Random Trees, napravljeno je devet modela, od kojih se skoro svaki izdvajao kao najbolji model u nekom strimu. Model koji se najčešće izdvajao kao najbolji je model broj 8. Model 8 pravi sto modela, odnosno drveta. Maksimalna dubina drveta je 15, maksimalan broj čvorova 5000, a maksimalna veličina podčvora 4. Model koristi matricu cena koje kažnjava klasifikaciju instance klase 1 u klasu 0 vrednošću 1.5, a obrnutu klasifikaciju vrednošću 1. Algoritam se zaustavlja kada preciznost modela više ne može da se poboljša. Napredna podešavanja su većinom slična filteru podataka koji je primenjen pre upotrebe algoritma, a ostala napredna podešavanja se mogu videti u strimovima. Model 7 koji se prvi sledeći najčešće izdvaja kao najbolji model u strimovima se razlikuje od modela 8 samo u matrici cena. Kod ovog modela duplo jače se kažnjava klasifikacija instance koja pripada klasi 0 kao instance klase 1, nego obrnuto. Ostali modeli koji se u nekim strimovima izdvajaju kao najbolji, imaju veće razlike u postavci modela. O tome koji modeli su u kojim strimovima izdvojeni kao najbolji mogu se naći u tabelama u dodatku A, a detalji o samim modelima i njihovoj građi se mogu naći u strimovima.

Na osnovu slike (Slika 35), može se zaključiti da algoritam Random Trees ima najslabije performanse od svih algoritama. Kod ovog algoritma ne važe uslovi za dobre i loše rezultate kao kod ostalih algoritama. Rezultati su najlošiji kada su indikatori binarni ili nominalni, a skup atributa filtriran, kao i kod ostalih algoritama. Iako su malo bolji od rezultata pod prethodnim uslovima, rezultati su lošiji i kada su indikatori binarni ili kontinualni, a skup atributa kompletan. Najbolje rezultate algoritam daje kada su indikatori kontinualni, a skup atributa filtriran, ili kada su indikatori nominalni, a skup atributa kompletan.



Slika 35 - Grafički prikaz performansi algoritma Random Trees

Na osnovu grafika performansi svih modela, može se zaključiti da modeli najbolje rade kada su atributi koji predstavljaju indikatore na neku dijagnozu, ili još par atributa koji predstavljaju rezultate nekog kardiološkog testa, kontinualnog tipa. U tom slučaju nije bitno kakav je skup atributa, da li je kompletan ili je filtriran. U tim uslovima svi algoritmi sem algoritma CHAID daju odlične rezultate, ali i CHAID daje bolje rezultate nego u drugim uslovima. Takođe najbolje rezultate daje i kada su atributi binarni ili nominalni, a skup atributa nije filtriran na važne i marginalne. Lošije rezultate daje kada su atributi binarni ili nominalni, a skup atributa filtriran na važne i marginalne. Svi modeli pokazuju identično ponašanje u istim uslovima, uz manje varijacije u kvalitetu. Na osnovu ovoga možemo zaključiti da je za rad sa ovim skupom podataka najbolje posmatrati attribute kao kontinualne, dok skup atributa nije bitan. Uzevši ove uslove u obzir, najbolji algoritmi za rad sa podacima su C5.0, linearni metod potpornih vektora, neuronska mreža i odmah za njima QUEST. Najgore performanse imao je algoritam CHAID, dok je CART negde između ove dve grupe.

Kao izuzetak među algoritmima trebalo bi uzeti algoritam Random Trees. Ovaj algoritam je jedini čiji kvalitet rada ne prati u potpunosti iste uslove kao i ostali algoritmi. Pored toga, ovaj algoritam je jedini algoritam kod koga se kao najbolji model u strimovima izdvaja više od dva modela. Izdvaja se čak sedam modela od korišćenih devet. Na osnovu performansi svakako se može izdvojiti kao najlošiji algoritam. Uzevši u obzir nekonzistentnost tipa modela pri radu sa podacima i performanse algoritma, ovaj algoritam bi možda trebalo razmatrati kao poslednju opciju pri radu sa ovim podacima.

#### 4.4. Najznačajniji atributi u klasifikaciji

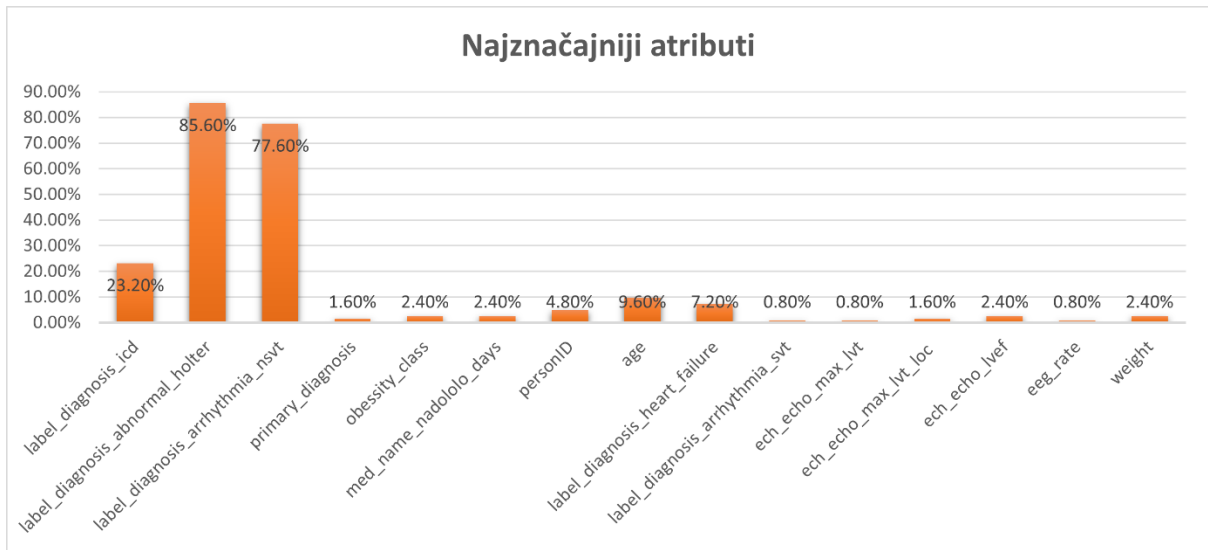
Najznačajniji atributi za predviđanje da li je pacijent nisko ili visoko rizičan su (Slika 36):

- label\_diagnosis\_abnormal\_holter
- label\_diagnosis\_arrhythmia\_nsvt
- label\_diagnosis\_icd .

Tri najznačajnija atributa su iz label grupe atributa. Ovi atributi predstavljaju indikatore da je pacijentu bila uspostavljena neka dijagnoza u periodu od 5 godina od trenutka lečenja. Sve tri dijagnoze su vezane za aritmije, odnosno za nepravilan rad srca. Na osnovu ovoga možemo zaključiti da su ovi atributi stvarno značajni za klasifikaciju podataka.

Atributi PersonID, Age, Ech\_Echo\_LVEF, i drugi, koji su se retko izdvajali kao značajni, ali ipak jesu, su se najčešće izdvajali kao značajni kod algoritma Random Trees. I drugi algoritmi su povremeno prepoznavali neki od retko značajnih atributa kao značajni, ali sa velikom razlikom u odnosu na algoritam Random Trees. Razlika je tome što u skupu značajnih atributa kod drugih algoritama, pored atributa koji se retko izdvajao kao značajan nalazio i bar jedan atribut koji se često izdvajao kao značajan. Kod algoritma Random Trees se često dobijao skup značajnih atributa u kome su svi atributi oni atributi koji se retko izdvajaju kao značajni.

Izuzetak među atributima koji su imali značaja u prepoznavanju visokorizičnih pacijenata bi mogao da bude atribut PersonID koji predstavlja ID pacijenta. ID je jedinstven za svakog pacijenta i, tako gledano, ne bi trebalo da ima uticaja na rezultate klasifikacije. Ali pacijenti dolaze na kontrole i podaci se ponavljaju, pa bismo tako mogli prepoznati visokorizičnog pacijenta na osnovu njegovog ID-ja. Ako bismo ipak želeli da dobijemo model koji je nepristrasan i koji ne uzima u obzir nikakve podatke o pacijentu sem njegovog trenutnog zdravstvenog stanja, onda bi atribut PearsonID trebalo koristiti prilikom klasifikacije.



Slika 36 - Grafički prikaz najznačajnijih atributa

## 5. Zaključak

U cilju prepoznavanja visokorizičnih pacijenata obolelih od hipertrofične kardiomiopatije primenom metoda klasifikacije korišćeni su algoritmi C5.0, CART, CHAID, QUEST, LSVM, Random Trees i neuronska mreža. Pri radu sa podacima kao ciljnu promenljivu bilo je moguće odabrati jedan od dva atributa: High Risk (indikator visokog rizika) ili No Risk (indikator niskog rizika). Algoritmi su dali dobre rezultate u oba slučaja, pri čemu su rezultati bili približno jednaki bez obzira na izbor ciljne promenljive. Na osnovu ovoga možemo zaključiti da se ovi algoritmi mogu koristiti u dijagnostici i za predviđanje visokorizičnih i za predviđanje niskorizičnih pacijenata. Za razliku od ciljne promenljive koja nije imala značajan uticaj na rezultate, izbor tipa atributa jeste. Atributi koji predstavljaju indikatore na postojeću dijagnozu bolesti u prošlosti ili mogućnost postavljanja iste u budućnosti, kao i nekoliko atributa koji predstavljaju rezultate kardioloških testova, mogli su imati jedan od tri tipa: binarni, nominalni ili kontinualni. Izbor tipa ovih atributa značajno utiče na performanse algoritama. Svi algoritmi su dali najbolje rezultate kada su ovi atributi bili kontinualnog tipa. Pored tipa ovih atributa, značajan ulogu u kvalitetu rezultata imao je i odabir skupa atributa. Algoritmi su takođe dali najbolje rezultate kada je korišćen kompletan skup atributa. Može se zaključiti da pri radu sa ovakvim podacima treba koristiti kompletan skup atributa ili posmatrati attribute koji predstavljaju indikatore kao kontinualne. Od svih algoritama, najbolje rezultate imali su algoritmi C5.0, linearni metod potpornih vektora, neuronska mreža i odmah za njima QUEST. Najgore performanse imao je algoritam CHAID, dok je CART negde između ove dve grupe. Algoritam Random Trees je pokazao odstupanja u rezultatima pod istim uslovima kao i ostali algoritmi, ali je svakako imao najlošije rezultate, pa se pored algoritma CHAID svakako može smatrati najlošijim algoritmom za rad sa ovakvim podacima.

Kao najznačajniji atributi za predviđanje rizičnosti pacijenta izdvojeni su atributi: label\_diagnosis\_abnormal\_holter, label\_diagnosis\_arrhythmia\_nsvt i label\_diagnosis\_icd. Sva tri atributa ukazuju na postojeći problem sa radom srca, pa se može zaključiti da su atributi zaista značajni za prepoznavanje visokorizičnih pacijenata. Uključivanje atributa PearsonID, koji je takođe izdvojen kao jedan od značajnih atributa, treba razmotriti prilikom klasifikacije.

Kao što smo videli na primeru podataka o pacijentima koji su oboleli od hipertrofične kardiomiopatije, mogu se definisati uslovi pod kojima algoritmi najbolje rade, ali skup podataka neće uvek ispunjavati te uslove. Od podataka koji su nam na raspolaganju i uslova koje ti podaci zadovoljavaju, zavisice i koji algoritam koristimo. Bez obzira na moguća mnogobrojna odstupanja u skupovima podataka i potrebnom prilagođavanju svakom, metod klasifikacije bi trebalo jednog dana da bude uključen u dijagnostiku i druge grane medicine.

## Dodatak A

U ovom dodatku biće predstavljene tabele sa podacima vezanim za performanse algoritama. Svaka tabela odgovara jednom strimu (spisak strimova se nalazi u dodatku B), a za svaki algoritam izdvojen je model koji je imao najbolje rezultate. Zaglavlje tabele se nalazi u prvoj koloni, a polja tabele će biti opisana u nastavku. Izvorna verzija tabele se nalazi u xls formatu u dodatku u elektronskoj verziji rada.

*Ciljna promenljiva* – definiše koji atribut predstavlja ciljnu promenljivu u skupu podataka. Može imati vrednosti High Risk (indikator visokorizičnih pacijenata) i No Risk (indikator niskorizičnih pacijenata).

*Tip podataka* – definiše kog tipa su atributi koji predstavljaju indikatore na postojeću dijagnozu u prošlosti ili budućnosti, i još nekih atributa koji predstavljaju rezultate nekog kardiološkog testa. Može imati vrednosti Continuous (neprekidan tip), Flag (binaran tip) i Nominal (nominalan tip).

*Filter* – definiše koji filter je primenjen nad skupom atributa. Može imati vrednosti Important (filtrirani su samo važni atributi), Important + Marginal (filtrirani su važni i marginalni atributi) i No Filter (nije korišćen filter).

*Algoritam* – označava koji algoritam je primenjen nad podacima.

*Model* – označava koji model je izdvojen kao najbolji model datog algoritma u strimu.

*Correct – training, Correct – testing, Wrong – training, Wrong – testing* – predstavlja procenat tačno ili pogrešno klasifikovanih instanci za trening i test skup.

*0-0 Training, 0-1 Training, 1-0 Training, 1-1 Training* – broj instanci trening skupa koji pripada klasi 0 i klasifikovan je kao instanca klase 0, pripada klasi 0 i klasifikovan je kao instanca klase 1, itd.

*0-0 Testing, 0-1 Testing, 1-0 Testing, 1-1 Testing* – broj instanci test skupa koji pripada klasi 0 i klasifikovan je kao instanca klase 0, pripada klasi 0 i klasifikovan je kao instanca klase 1, itd.

*AUC – training, AUC – testing* – AUC vrednosti algoritma primenjenog na trening i test podacima.

*Most important attribute* – najznačajniji atributi za predviđanje klase instance.

*MIA – value* – mera uticaja najznačajnijeg atributa na predviđanje klase instance.

Tabela 1 – Performanse modela koji pripadaju strimu M5y

<b>Ciljna Promenljiva</b>	High Risk	High Risk	High RRisk	High Risk	High Risk	High Risk	High Risk
<b>Tip podataka</b>	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous
<b>Filter</b>	Important	Important	Important	Important	Important	Important	Important
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 9	Model 4	Model 1	Model 2, 3	Model 1	Model 4	Model 4, 5
<b>Correct - training %</b>	0.9803	0.9995	1	1	0.9993	0.9985	1
<b>Correct - testing %</b>	0.9593	0.9972	1	1	0.9876	0.996	0.9983
<b>Wrong - training %</b>	0.0197	0.0005	0	0	0.0007	0.0015	0
<b>Wrong - testing%</b>	0.0407	0.0028	0	0	0.0124	0.004	0.0017
<b>0-0 Training</b>	2109	2150	2150	2150	2150	2150	2150
<b>0-1 Training</b>	41	0	0	0	0	0	0
<b>1-0 Training</b>	39	2	0	0	3	6	0
<b>1-1 Training</b>	1880	1917	1919	1919	1916	1913	1919
<b>0-0 Testing</b>	939	975	975	975	975	975	975
<b>0-1 Testing</b>	36	0	0	0	0	0	0
<b>1-0 Testing</b>	36	5	0	0	22	7	3
<b>1-1 Testing</b>	760	791	976	796	774	789	793
<b>AUC - training</b>	0.997	0.999	1	1	0.999	0.997	1
<b>AUC - testing</b>	0.991	0.996	1	1	0.985	0.994	1
<b>Most important attribute</b>	label_diagnosis_icd	label_diagnosis_abnormal_holter/ label_diagnosis_arrhythmia_nsvt	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt /heart_failure	label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter
<b>MIA - value</b>	0.7	0.18	0.08	0.07	0.18	0.19	0.16



Tabela 2 – Performanse modela koji pripadaju strimu M5y\_Flag

<b>Ciljna Promenljiva</b>	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk
<b>Tip podataka</b>	Flag	Flag	Flag	Flag	Flag	Flag	Flag
<b>Filter</b>	Important	Important	Important	Important	Important	Important	Important
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 6	Model 4	Model 1	Model 2	Model 1	Model 1	Model 4
<b>Correct - training %</b>	91.45%	90.96%	94.27%	90.91%	95.26%	90.81%	92.41%
<b>Correct - testing %</b>	89.61%	89.33%	92.15%	90.57%	91.19%	89.89%	88.99%
<b>Wrong - training %</b>	8.55%	9.04%	5.73%	9%	4.74%	9.19%	7.59%
<b>Wrong - testing%</b>	10.39%	10.67%	7.85%	9.43%	8.81%	10.11%	11.01%
<b>0-0 Training</b>	1993	1969	2118	2057	2091	2039	1964
<b>0-1 Training</b>	157	181	32	93	59	111	186
<b>1-0 Training</b>	191	187	201	277	134	263	123
<b>1-1 Training</b>	1728	1732	1718	1642	1785	1656	1796
<b>0-0 Testing</b>	890	872	943	933	915	902	860
<b>0-1 Testing</b>	85	103	32	42	60	73	115
<b>1-0 Testing</b>	99	86	107	125	96	106	80
<b>1-1 Testing</b>	697	710	689	671	700	690	716
<b>AUC - training</b>	0.966	0.955	0.982	0.963	0.982	0.948	0.982
<b>AUC - testing</b>	0.948	0.941	0.969	0.964	0.956	0.945	0.957
<b>Most important attribute</b>	age	label_diagnosis_abnormal_holter	age/prim_diagn/obesity_class/nadolo/icd/nsvt/holter	label_diagnosis_icd/holter   nsvt	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt
<b>MIA - value</b>	0.07	0.22	0.05/0.02(ns vt   holter)	0.17 / 0.09 / 0.07	0.21	0.22	0.22

Tabela 3 – Performanse modela koji pripadaju strimu M5y\_Flag\_MI

<b>Ciljna Promenljiva</b>	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk
<b>Tip podataka</b>	Flag	Flag	Flag	Flag	Flag	Flag	Flag
<b>Filter</b>	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 4	Model 5	Model 1	Model 3	Model 5	Model 1	Model 5
<b>Correct - training %</b>	89.19%	92.21%	95.28%	90.07%	95.06%	89.83%	92.65%
<b>Correct - testing %</b>	87.13%	89.44%	92.43%	90.23%	91.08%	89.22%	89.05%
<b>Wrong - training %</b>	10.81%	7.79%	4.72%	9.93%	4.94%	10.17%	7.35%
<b>Wrong - testing %</b>	12.87%	10.56%	7.57%	9.77%	8.92%	10.78%	10.95%
<b>0-0 Training</b>	1909	2020	2111	1988	2053	2025	2022
<b>0-1 Training</b>	241	130	39	162	97	125	128
<b>1-0 Training</b>	199	187	153	242	104	289	171
<b>1-1 Training</b>	1720	1732	1766	1677	1815	1630	1748
<b>0-0 Testing</b>	847	887	931	906	894	895	879
<b>0-1 Testing</b>	128	88	44	69	81	80	96
<b>1-0 Testing</b>	100	99	90	104	77	111	98
<b>1-1 Testing</b>	696	697	706	692	719	685	698
<b>AUC - training</b>	0.956	0.961	0.987	0.962	0.988	0.945	0.97
<b>AUC - testing</b>	0.942	0.942	0.975	0.961	0.958	0.942	0.954
<b>Most important attribute</b>	personID / age / label_diagnosis_icd	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_abnormal_label_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_icd / label_diagnosis_abnormal_holter	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / holter	label_diagnosis_abnormal_holter / nsvt
<b>MIA - value</b>	0.08 / 0.06 / 0.05	0.21	0.03	0.16 / 0.13	0.22	0.23	0.2 / 0.18

Tabela 4 – Performanse modela koji pripadaju strimu M5y\_Flag\_NoFilter

<b>Ciljna Promenljiva</b>	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk
<b>Tip podataka</b>	Flag	Flag	Flag	Flag	Flag	Flag	Flag
<b>Filter</b>	No filter	No filter	No filter	No filter	No filter	No filter	No filter
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 2	Model 4	Model 1	Model 2	Model 1	Model 1	Model 5
<b>Correct - training %</b>	94.72%	99.95%	100.00%	100.00%	99.93%	98.80%	99.68%
<b>Correct - testing %</b>	93.22%	99.72%	99.77%	100.00%	98.76%	97.85%	99.27%
<b>Wrong - training %</b>	5.28%	0.05%	0.00%	0.00%	0.07%	1.20%	0.32%
<b>Wrong - testing%</b>	6.78%	0.28%	0.23%	0.00%	1.24%	2.15%	0.73%
<b>0-0 Training</b>	2120	2150	2150	2150	2150	2150	2144
<b>0-1 Training</b>	30	0	0	0	0	0	6
<b>1-0 Training</b>	185	2	0	0	3	49	7
<b>1-1 Training</b>	1734	1917	1919	1919	1916	1870	1912
<b>0-0 Testing</b>	956	975	975	975	975	975	969
<b>0-1 Testing</b>	19	0	0	0	0	0	6
<b>1-0 Testing</b>	101	5	4	0	22	38	7
<b>1-1 Testing</b>	695	791	792	796	774	758	789
<b>AUC - training</b>	0.973	0.999	1	1	0.999	0.988	0.999
<b>AUC - testing</b>	0.96	0.996	1	1	0.985	0.979	0.997
<b>Most important attribute</b>	label_diagnosis_icd	label_diagnosis_abnormal_holter	label_diagnosis_icd / heart_failure / arrhythmia_svt	label_diagnosis_abnormal_holter	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	ech_echo_max_lvt / label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt
<b>MIA - value</b>	0.13	0.18	0.07	0.07	0.17	0.18	0.13

Tabela 5 – Performanse modela koji pripadaju strimu M5y\_MI

<b>Ciljna Promenljiva</b>	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk
<b>Tip podataka</b>	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous
<b>Filter</b>	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 5	Model 5	Model 1	Model 2, 3	Model 5	Model 4	Model 4
<b>Correct - training %</b>	98.38%	99.95%	100.00%	100.00%	99.73%	99.85%	99.98%
<b>Correct - testing %</b>	96.95%	99.49%	100.00%	100.00%	98.53%	99.60%	99.83%
<b>Wrong - training %</b>	1.62%	0.05%	0.00%	0.00%	0.27%	0.15%	0.02%
<b>Wrong - testing%</b>	3.05%	0.51%	0.00%	0.00%	1.47%	0.40%	0.17%
<b>0-0 Training</b>	2102	2150	2150	2150	2150	2150	2150
<b>0-1 Training</b>	48	0	0	0	0	0	0
<b>1-0 Training</b>	18	2	0	0	11	6	1
<b>1-1 Training</b>	1901	1917	1919	1919	1908	1913	1918
<b>0-0 Testing</b>	943	975	975	975	975	975	975
<b>0-1 Testing</b>	32	0	0	0	0	0	0
<b>1-0 Testing</b>	22	9	0	0	26	7	3
<b>1-1 Testing</b>	774	787	796	796	770	789	793
<b>AUC - training</b>	0.999	0.999	1	1	0.996	0.997	1
<b>AUC - testing</b>	0.993	0.995	1	1	0.975	0.994	1
<b>Most important attribute</b>	label_diagnosis_icd	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter
<b>MIA - value</b>	0.06	0.18	0.08	0.06	0.19	0.19	0.16

Tabela 6 – Performanse modela koji pripadaju strimu M5y\_NoFilter

<b>Ciljna Promenljiva</b>	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk
<b>Tip podataka</b>	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous
<b>Filter</b>	No filter	No filter	No filter	No filter	No filter	No filter	No filter
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 2	Model 4	Model 1	Model 2,3	Model 5	Model 1	Model 4, 5
<b>Correct - training %</b>	94.94%	97.94%	100.00%	100.00%	99.73%	99.71%	100.00%
<b>Correct - testing %</b>	93.11%	96.10%	100.00%	100.00%	97.46%	98.98%	99.83%
<b>Wrong - training %</b>	5.06%	2.06%	0.00%	0.00%	0.27%	0.29%	0.00%
<b>Wrong - testing%</b>	6.89%	3.90%	0.00%	0.00%	2.54%	1.02%	0.17%
<b>0-0 Training</b>	2035	2134	2150	2150	2149	2150	2150
<b>0-1 Training</b>	115	16	0	0	1	0	0
<b>1-0 Training</b>	91	68	0	0	10	18	0
<b>1-1 Training</b>	1828	1851	1919	1919	1909	1907	1919
<b>0-0 Testing</b>	906	954	975	975	962	975	974
<b>0-1 Testing</b>	69	21	0	0	13	0	1
<b>1-0 Testing</b>	53	48	0	0	32	18	2
<b>1-1 Testing</b>	743	748	796	796	764	778	794
<b>AUC - training</b>	0.98	0.973	1	1	0.998	0.995	1
<b>AUC - testing</b>	0.972	0.962	1	1	0.977	0.987	1
<b>Most important attribute</b>	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	ech_echo_max_lvt_loc / label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter / ech_echo_lvef
<b>MIA - value</b>	0.09	0.15	0.09	0.07	0.18	0.21	0.13

Tabela 7 – Performanse modela koji pripadaju strimu M5y\_NoFilter\_NoRisk

<b>Ciljna Promenljiva</b>	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk
<b>Tip podataka</b>	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous
<b>Filter</b>	No filter	No filter	No filter	No filter	No filter	No filter	No filter
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 8	Model 5	Model 1	Basic, 2, 3	Model 1	Model 1	Model 4, 5
<b>Correct - training %</b>	98.03%	100.00%	100.00%	100.00%	99.31%	99.46%	100.00%
<b>Correct - testing %</b>	95.93%	99.72%	100.00%	100.00%	97.40%	98.98%	99.77%
<b>Wrong - training %</b>	1.97%	0.00%	0.00%	0.00%	0.69%	0.54%	0.00%
<b>Wrong - testing%</b>	4.07%	0.28%	0.00%	0.00%	2.60%	1.02%	0.23%
<b>0-0 Training</b>	1880	1919	1919	1919	1891	1897	1919
<b>0-1 Training</b>	39	0	0	0	28	22	3
<b>1-0 Training</b>	41	0	0	0	0	0	1
<b>1-1 Training</b>	2109	2150	2150	2150	2150	2150	974
<b>0-0 Testing</b>	760	791	796	796	755	778	793
<b>0-1 Testing</b>	36	5	0	0	41	18	3
<b>1-0 Testing</b>	36	0	0	0	5	0	1
<b>1-1 Testing</b>	939	975	975	975	970	975	974
<b>AUC - training</b>	0.997	1	1	1	0.991	0.989	1
<b>AUC - testing</b>	0.991	0.998	1	1	0.968	0.977	1
<b>Most important attribute</b>	label_diagnos is_icd / label_diagnos is_abnormal_ holter / label_diagnos is_arrhythmia_ nsvt	label_diagno sis_arrhyth mia_nsvt / label_diagno sis_abnorma l_holter	label_diagno sis_arrhyth mia_nsvt / label_diagno sis_abnorma l_holter	label_diagno sis_abnormal _holter	label_diagnosi s_arrhythmia_ nsvt / label_diagnosi s_abnormal_h olter	label_diagnos is_arrhythmia_ nsvt / label_diagnos is_abnormal_ holter	label_diagno sis_arrhythmi a_nsvt / label_diagno sis_abnormal _holter
<b>MIA - value</b>	0.07	0.17	0.07	0.07	0.18	0.17	0.16

Tabela 8 – Performanse modela koji pripadaju strimu M5y\_Nominal

<b>Ciljna Promenljiva</b>	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk
<b>Tip podataka</b>	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
<b>Filter</b>	Important	Important	Important	Important	Important	Important	Important
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 7	Model 4	Model 1	Basic, 2	Model 1	Model 1	Model 5
<b>Correct - training %</b>	91.03%	90.96%	94.27%	90.91%	95.26%	90.81%	92.68%
<b>Correct - testing %</b>	89.84%	89.33%	92.15%	90.57%	91.19%	89.89%	90.74%
<b>Wrong - training %</b>	8.97%	9.04%	5.73%	9.09%	4.74%	9.19%	7.32%
<b>Wrong - testing%</b>	10.16%	10.67%	7.85%	9.43%	8.81%	10.11%	9.26%
<b>0-0 Training</b>	2017	1969	2118	2057	2091	2039	2098
<b>0-1 Training</b>	133	181	32	93	59	111	52
<b>1-0 Training</b>	232	187	201	277	134	263	246
<b>1-1 Training</b>	1687	1732	1718	1642	1785	1656	1673
<b>0-0 Testing</b>	902	872	943	933	915	902	934
<b>0-1 Testing</b>	73	103	32	42	60	73	41
<b>1-0 Testing</b>	107	86	107	125	96	106	123
<b>1-1 Testing</b>	689	710	689	671	700	690	673
<b>AUC - training</b>	0.96	0.955	0.982	0.963	0.982	0.948	0.967
<b>AUC - testing</b>	0.949	0.941	0.969	0.964	0.956	0.945	0.953
<b>Most important attribute</b>	age / ech_echo_lvef	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	nadololo / primary_diagnosis / obesity_classes / icd / age	label_diagnosis_icd	label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter
<b>MIA - value</b>	0.07	0.22	0.05	0.17	0.2	0.22	0.21

Tabela 9 – Performanse modela koji pripadaju strimu M5y\_Nominal\_NoFilter

<b>Ciljna Promenljiva</b>	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk
<b>Tip podataka</b>	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
<b>Filter</b>	No filter	No filter	No filter	No filter	No filter	No filter	No filter
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 7	Model 4	Model 1	Basic, 2, 3	Model 1	Model 4	Model 4, 5
<b>Correct - training %</b>	98.06%	99.98%	100.00%	100.00%	99.93%	99.85%	100.00%
<b>Correct - testing %</b>	96.05%	99.72%	100.00%	100.00%	98.76%	99.60%	99.83%
<b>Wrong - training %</b>	1.94%	0.05%	0.00%	0.00%	0.07%	0.15%	0.00%
<b>Wrong - testing%</b>	3.95%	0.28%	0.00%	0.00%	1.24%	0.40%	0.17%
<b>0-0 Training</b>	2110	2150	2150	2150	2150	2150	2150
<b>0-1 Training</b>	40	0	0	0	0	0	0
<b>1-0 Training</b>	39	2	0	0	3	6	0
<b>1-1 Training</b>	1880	1917	1919	1919	1916	1913	1919
<b>0-0 Testing</b>	939	975	975	975	975	975	975
<b>0-1 Testing</b>	36	0	0	0	0	0	0
<b>1-0 Testing</b>	34	5	0	0	22	7	3
<b>1-1 Testing</b>	762	791	796	796	774	789	793
<b>AUC - training</b>	0.997	0.999	1	1	0.999	0.997	1
<b>AUC - testing</b>	0.992	0.996	1	1	0.985	0.994	1
<b>Most important attribute</b>	icd / label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_abnormal_holter / icd / label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter
<b>MIA - value</b>	0.07 / 0.06 / 0.05	0.18	0.19	0.07 / 0.06 / 0.05	0.19	0.19	0.16



Tabela 10 – Performanse modela koji pripadaju strimu M5y\_Nominal\_MI

<b>Ciljna Promenljiva</b>	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk	High Risk
<b>Tip podataka</b>	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
<b>Filter</b>	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 4	Model 4	Model 1	Basic, 2	Model 5	Model 4	Model 5
<b>Correct - training %</b>	92.53%	92.50%	94.96%	91.13%	95.75%	88.06%	93.68%
<b>Correct - testing %</b>	90.63%	89.55%	93.11%	90.68%	89.67%	86.17%	91.24%
<b>Wrong - training %</b>	7.47%	7.50%	5.04%	8.87%	4.25%	11.94%	6.32%
<b>Wrong - testing%</b>	9.37%	10.46%	6.89%	9.32%	10.33%	13.83%	8.58%
<b>0-0 Training</b>	2049	2018	2126	2058	2050	1858	2095
<b>0-1 Training</b>	101	132	24	92	100	292	55
<b>1-0 Training</b>	203	173	181	269	73	194	202
<b>1-1 Training</b>	1716	1746	1738	1650	1846	1725	1717
<b>0-0 Testing</b>	913	889	950	931	876	826	925
<b>0-1 Testing</b>	62	86	25	44	99	149	50
<b>1-0 Testing</b>	104	99	97	121	84	96	101
<b>1-1 Testing</b>	692	697	699	675	712	700	694
<b>AUC - training</b>	0.965	0.96	0.989	0.966	0.991	0.931	0.964
<b>AUC - testing</b>	0.956	0.939	0.975	0.964	0.975	0.917	0.95
<b>Most important attribute</b>	age / personID / weight	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	icd / label_diagnosis_abnormal_holter	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter
<b>MIA - value</b>	0.07 / 0.06	0.2	0.04	0.16 / 0.08	0.2	0.18	0.2

Tabela 11 – Performanse modela koji pripadaju strimu M5y\_Nominal\_NoFilter\_NoRisk

<b>Ciljna Promenljiva</b>	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk
<b>Tip podataka</b>	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
<b>Filter</b>	No filter	No filter	No filter	No filter	No filter	No filter
<b>Algoritam</b>	Random Trees	CART	C5.0	Chaid	Quest	Neural Network
<b>Model</b>	Model 8	Model 5	Model 1	Model 1	Model 1	Model 4, 5
<b>Correct - training %</b>	94.37%	97.69%	100.00%	99.66%	96.73%	100.00%
<b>Correct - testing %</b>	91.30%	95.93%	99.77%	97.12%	95.77%	99.60%
<b>Wrong - training %</b>	5.63%	2.31%	0.00%	0.34%	3.27%	0.00%
<b>Wrong - testing%</b>	8.70%	4.07%	0.23%	2.88%	4.23%	0.40%
<b>0-0 Training</b>	1808	1847	1919	1905	1819	1919
<b>0-1 Training</b>	111	72	0	14	100	0
<b>1-0 Training</b>	118	22	0	0	33	0
<b>1-1 Training</b>	2032	2128	2150	2150	2117	2150
<b>0-0 Testing</b>	724	740	792	748	740	792
<b>0-1 Testing</b>	72	56	4	14	56	4
<b>1-0 Testing</b>	82	16	0	3	19	3
<b>1-1 Testing</b>	893	959	975	972	956	972
<b>AUC - training</b>	0.982	0.972	1	0.997	0.98	1
<b>AUC - testing</b>	0.961	0.95	1	0.973	0.97	1
<b>Most important attribute</b>	age / eeg_rate / personID	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter / heart_failure / icd /	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	ech_echo_max_lvt_loc / label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt
<b>MIA - value</b>	0.05	0.19	0.07	0.18	0.2	0.13 / 0.12

Tabela 12 – Performanse modela koji pripadaju strimu M5y\_Nominal\_NoRisk

<b>Ciljna Promenljiva</b>	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk
<b>Tip podataka</b>	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
<b>Filter</b>	Important	Important	Important	Important	Important	Important	Important
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 5	Model 5	Model 1	Basic	Model 5	Model 4	Model 4, 5
<b>Correct - training %</b>	89.90%	91.57%	94.27%	90.91%	94.05%	89.28%	92.82%
<b>Correct - testing %</b>	88.31%	90.23%	92.15%	90.57%	90.57%	89.50%	91.08%
<b>Wrong - training %</b>	10.10%	8.43%	5.73%	9.09%	5.95%	10.72%	7.18%
<b>Wrong - testing%</b>	11.69%	9.77%	7.85%	9.43%	9.43%	10.50%	8.92%
<b>0-0 Training</b>	1732	1660	1718	1642	1713	1513	1717
<b>0-1 Training</b>	187	259	201	277	188	406	202
<b>1-0 Training</b>	224	84	32	93	54	30	90
<b>1-1 Training</b>	1926	2066	2118	2057	2069	2120	2060
<b>0-0 Testing</b>	707	685	689	671	691	625	698
<b>0-1 Testing</b>	89	111	107	125	105	171	98
<b>1-0 Testing</b>	118	62	32	42	62	15	60
<b>1-1 Testing</b>	857	913	943	933	913	960	915
<b>AUC - training</b>	0.962	0.943	0.982	0.963	0.975	0.932	0.974
<b>AUC - testing</b>	0.949	0.929	0.969	0.964	0.948	0.925	0.964
<b>Most important attribute</b>	age / ech_echo_lvef	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	nadololo / primary_diagnosis / obesity_classes / icd / age	icd / label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt
<b>MIA - value</b>	0.07 / 0.05	0.22	0.05	0.17 / 0.9 / 0.7	0.21	0.24	0.2 / 0.18

Tabela 13 – Performanse modela koji pripadaju strimu M5y\_NoRisk

<b>Ciljna Promenljiva</b>	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk
<b>Tip podataka</b>	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous
<b>Filter</b>	Important	Important	Important	Important	Important	Important	Important
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 8	Model 5	Model 1	Basic, 2, 3	Model 1	Model 1	Model 4, 5
<b>Correct - training %</b>	98.30%	100.00%	100.00%	100.00%	99.31%	99.46%	100.00%
<b>Correct - testing %</b>	95.93%	99.72%	100.00%	100.00%	97.40%	98.98%	99.77%
<b>Wrong - training %</b>	1.97%	0.00%	0.00%	0.00%	0.69%	0.54%	0.00%
<b>Wrong - testing%</b>	4.07%	0.28%	0.00%	0.00%	2.60%	1.02%	0.23%
<b>0-0 Training</b>	1880	1919	1919	1919	1891	1897	1919
<b>0-1 Training</b>	39	0	0	0	28	22	0
<b>1-0 Training</b>	41	0	0	0	0	0	0
<b>1-1 Training</b>	2109	2150	2150	2150	2150	2150	2150
<b>0-0 Testing</b>	760	791	796	796	755	778	793
<b>0-1 Testing</b>	36	5	0	0	41	18	3
<b>1-0 Testing</b>	36	0	0	0	5	0	1
<b>1-1 Testing</b>	939	975	975	975	970	975	974
<b>AUC - training</b>	0.997	1	1	1	0.991	0.989	1
<b>AUC - testing</b>	0.991	0.998	1	1	0.968	0.977	1
<b>Most important attribute</b>	icd / label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt / heart_failure	label_diagnosis_abnormal_holter / icd / label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter
<b>MIA - value</b>	0.07 / 0.06	0.18	0.08 / 0.07	0.07 / 0.06	0.18	0.19	0.16

Tabela 14 – Performanse modela koji pripadaju strimu M5y\_Flag\_NoRisk

<b>Ciljna Promenljiva</b>	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk
<b>Tip podataka</b>	Flag	Flag	Flag	Flag	Flag	Flag	Flag
<b>Filter</b>	Important	Important	Important	Important	Important	Important	Important
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Basic	Model 5	Model 1	Basic, 2, 3	Model 1	Model 1	Model 4, 5
<b>Correct - training %</b>	94.96%	100.00%	100.00%	100.00%	99.31%	98.80%	100.00%
<b>Correct - testing %</b>	92.49%	99.72%	99.77%	100.00%	97.40%	97.85%	99.60%
<b>Wrong - training %</b>	5.04%	0.00%	0.00%	0.00%	0.69%	1.20%	0.00%
<b>Wrong - testing%</b>	7.51%	28.00%	0.23%	0.00%	2.60%	2.15%	0.40%
<b>0-0 Training</b>	1770	1919	1919	1919	1891	1870	1919
<b>0-1 Training</b>	149	0	0	0	28	49	0
<b>1-0 Training</b>	56	0	0	0	0	0	0
<b>1-1 Training</b>	2049	2150	2150	2150	2150	2150	2150
<b>0-0 Testing</b>	710	791	792	796	755	758	789
<b>0-1 Testing</b>	86	5	4	0	41	38	7
<b>1-0 Testing</b>	47	0	0	0	5	0	0
<b>1-1 Testing</b>	928	975	975	975	970	975	975
<b>AUC - training</b>	0.979	1	1	1	0.991	0.988	1
<b>AUC - testing</b>	0.962	0.998	1	1	0.968	0.979	1
<b>Most important attribute</b>	label_diagnos is_icd / age / heart_failure	label_diagno sis_abnorma l_holter / label_diagno sis_arrhyth mia_nsvt	label_diagno sis_icd / heart_failur e / label_diagno sis_arrhyth mia_svt / label_diagno sis_abnorma l_holter	label_diagno sis_abnormal _holter / label_diagno sis_icd / label_diagno sis_arrhythm ia_nsvt	label_diagnosi s_abnormal_h olter / label_diagnosi s_arrhythmia_ nsvt	label_diagnos is_abnormal_ holter / label_diagnos is_arrhythmia _nsvt	label_diagno sis_abnormal _holter / label_diagno sis_arrhythmi a_nsvt
<b>MIA - value</b>	0.09 / 0.06	0.18	0.07	0.07 / 0.06	0.18	0.18	0.16

Tabela 15 – Performanse modela koji pripadaju strimu M5y\_Flag\_MI\_NoRisk

<b>Ciljna Promenljiva</b>	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk
<b>Tip podataka</b>	Flag	Flag	Flag	Flag	Flag	Flag	Flag
<b>Filter</b>	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 8	Model 5	Model 1	Basic, 2	Model 1	Model 1	Model 5
<b>Correct - training %</b>	91.72%	91.47%	94.96%	91.13%	96.53%	90.54%	92.97%
<b>Correct - testing %</b>	89.38%	90.97%	93.11%	90.68%	89.89%	90.18%	91.30%
<b>Wrong - training %</b>	8.28%	8.53%	5.04%	8.87%	3.47%	9.46%	7.03%
<b>Wrong - testing %</b>	10.62%	9.03%	6.89%	9.32%	10.11%	9.82%	8.70%
<b>0-0 Training</b>	1709	1695	1738	1650	1837	1661	1669
<b>0-1 Training</b>	210	224	181	269	82	258	250
<b>1-0 Training</b>	127	123	24	92	59	127	36
<b>1-1 Training</b>	2023	2027	2126	2058	2091	2023	2114
<b>0-0 Testing</b>	690	704	699	675	700	695	677
<b>0-1 Testing</b>	106	92	97	121	96	101	119
<b>1-0 Testing</b>	82	68	25	44	83	73	35
<b>1-1 Testing</b>	893	907	950	931	892	902	940
<b>AUC - training</b>	0.961	0.943	0.989	0.966	0.992	0.939	0.975
<b>AUC - testing</b>	0.946	0.942	0.975	0.964	0.953	0.94	0.959
<b>Most important attribute</b>	age / personID / ech_echo_lvef	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_icd / label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter
<b>MIA - value</b>	0.07 / 0.06 / 0.05	0.21	0.04	0.18 / 0.08 / 0.06	0.2	0.2	0.2

Tabela 16 – Performanse modela koji pripadaju strimu M5y\_Flag\_NoFilter\_NoRisk

<b>Ciljna Promenljiva</b>	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk
<b>Tip podataka</b>	Flag	Flag	Flag	Flag	Flag	Flag	Flag
<b>Filter</b>	No filter	No filter	No filter	No filter	No filter	No filter	No filter
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 2	Model 5	Model 1	Basic, 2	Model 1	Model 1	Model 4, 5
<b>Correct - training %</b>	94.72%	100.00%	100.00%	100.00%	99.31%	98.80%	100.00%
<b>Correct - testing %</b>	93.22%	99.72%	99.77%	100.00%	97.40%	97.85%	99.60%
<b>Wrong - training %</b>	5.28%	0.00%	0.00%	0.00%	0.69%	1.20%	0.00%
<b>Wrong - testing%</b>	6.78%	0.28%	0.23%	0.00%	2.60%	2.15%	0.40%
<b>0-0 Training</b>	1734	1919	1919	1919	1819	1870	1919
<b>0-1 Training</b>	185	0	0	0	28	49	0
<b>1-0 Training</b>	30	0	0	0	0	0	0
<b>1-1 Training</b>	2120	2150	2150	2150	2150	2150	2150
<b>0-0 Testing</b>	695	791	792	796	755	758	789
<b>0-1 Testing</b>	101	5	4	0	41	38	7
<b>1-0 Testing</b>	19	0	0	0	5	0	0
<b>1-1 Testing</b>	956	975	975	975	970	975	975
<b>AUC - training</b>	0.973	1	1	1	0.991	0.988	1
<b>AUC - testing</b>	0.96	0.998	1	1	0.968	0.979	1
<b>Most important attribute</b>	icd / label_diagnosis_arrythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_abnormal_holter / label_diagnosis_arrythmia_nsvt	icd / heart_failure / label_diagnosis_arrythmia_svt / label_diagnosis_abnormal_holter	label_diagnosis_abnormal_holter / icd / label_diagnosis_arrythmia_nsvt	label_diagnosis_abnormal_holter / label_diagnosis_arrythmia_nsvt	label_diagnosis_abnormal_holter / label_diagnosis_arrythmia_nsvt	label_diagnosis_abnormal_holter / label_diagnosis_arrythmia_nsvt
<b>MIA - value</b>	0.13 / 0.9 / 0.7	0.18	0.07 / 0.06	0.07 / 0.06	0.18	0.18	0.16

Tabela 17 – Performanse modela koji pripadaju strimu M5y\_Nominal\_MI\_NoRisk

<b>Ciljna Promenljiva</b>	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk
<b>Tip podataka</b>	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
<b>Filter</b>	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 8	Model 4	Model 1	Basic, 2	Model 5	Model 1	Model 4
<b>Correct - training %</b>	92.01%	92.50%	94.96%	91.13%	95.31%	90.54%	92.80%
<b>Correct - testing %</b>	89.27%	89.55%	93.11%	90.68%	90.51%	90.18%	91.30%
<b>Wrong - training %</b>	7.99%	7.50%	5.04%	8.87%	4.69%	9.46%	7.20%
<b>Wrong - testing%</b>	10.73%	10.45%	6.89%	9.32%	9.49%	9.82%	8.87%
<b>0-0 Training</b>	1748	1746	1738	1650	1770	1661	1772
<b>0-1 Training</b>	171	173	181	269	149	258	197
<b>1-0 Training</b>	154	132	699	92	42	127	96
<b>1-1 Training</b>	1996	2018	2126	2058	2108	2023	2054
<b>0-0 Testing</b>	695	697	699	675	701	695	695
<b>0-1 Testing</b>	101	99	97	121	95	101	101
<b>1-0 Testing</b>	89	86	25	44	73	73	56
<b>1-1 Testing</b>	886	889	950	931	902	902	919
<b>AUC - training</b>	0.971	0.96	0.989	0.966	0.985	0.939	0.974
<b>AUC - testing</b>	0.95	0.939	0.975	0.964	0.955	0.94	0.965
<b>Most important attribute</b>	age / weight / personID	label_diagnosis_abnormal_label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter / label_diagnosis_icd	label_diagnosis_icd / label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_abnormal_holter / label_diagnosis_arrhythmia_nsvt
<b>MIA - value</b>	0.06	0.2 / 0.18	0.04 / 0.03	0.16 / 0.08 / 0.06	0.22	0.2	0.19



Tabela 18 – Performanse modela koji pripadaju strimu M5y\_MI\_NoRisk

<b>Ciljna Promenljiva</b>	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk	No Risk
<b>Tip podataka</b>	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous
<b>Filter</b>	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal	Important + Marginal
<b>Algoritam</b>	Random Trees	CART	C5.0	LSVM	Chaid	Quest	Neural Network
<b>Model</b>	Model 8	Model 4	Model 1	Basic, 2, 3	Model 1	Model 1	Model 4, 5
<b>Correct - training %</b>	98.87%	100.00%	100.00%	100.00%	99.75%	99.46%	100.00%
<b>Correct - testing %</b>	97.12%	99.38%	100.00%	100.00%	98.02%	98.98%	99.89%
<b>Wrong - training %</b>	1.13%	0.00%	0.00%	0.00%	0.25%	0.54%	0.00%
<b>Wrong - testing%</b>	2.88%	0.62%	0.00%	0.00%	1.98%	1.02%	0.11%
<b>0-0 Training</b>	1892	1919	1919	1919	1909	1897	1919
<b>0-1 Training</b>	27	0	0	0	10	22	0
<b>1-0 Training</b>	19	0	0	0	0	0	0
<b>1-1 Training</b>	2131	2150	2150	2150	2150	2150	2150
<b>0-0 Testing</b>	768	785	796	796	763	778	795
<b>0-1 Testing</b>	28	11	0	0	33	18	1
<b>1-0 Testing</b>	23	0	0	0	2	0	1
<b>1-1 Testing</b>	952	975	975	975	973	975	974
<b>AUC - training</b>	0.999	1	1	1	0.997	0.989	1
<b>AUC - testing</b>	0.994	0.996	1	1	0.977	0.977	1
<b>Most important attribute</b>	label_diagnosis_icd / personID / weight	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter / heart_failure / label_diagnosis_cardiac_arrest	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter / heart_failure / label_diagnosis_cardiac_arrest	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter	label_diagnosis_arrhythmia_nsvt / label_diagnosis_abnormal_holter
<b>MIA - value</b>	0.06 / 0.05	0.16	0.08 / 0.07	0.08 / 0.07	0.19	0.2	0.16

## Dodatak B

U ovom dodatku se nalazi spisak IBM SPSS Modeler strimova (potoka) koji su napravljeni prilikom rada sa podacima. Svi strimovi koriste iste modele, razlika je u tipu atributa koji predstavljaju indikatore, ciljnoj promenljivoj i skupu atributa. Strimovi u elektronskom obliku se nalaze u dodatku u elektronskoj verziji rada.

M5y – ciljna promenljiva je High Risk, indikatori su kontinualnog tipa, a skup atributa čine samo važni atributi.

M5y\_Flag – ciljna promenljiva je High Risk, indikatori su binarnog tipa, a skup atributa čine samo važni atributi.

M5y\_Flag\_MI – ciljna promenljiva je High Risk, indikatori su binarnog tipa, a skup atributa čine važni i marginalni atributi.

M5y\_Flag\_NoFilter – ciljna promenljiva je High Risk, indikatori su binarnog tipa, a skup atributa čine svi atributi.

M5y\_MI – ciljna promenljiva je High Risk, indikatori su kontinualnog tipa, a skup atributa čine važni i marginalni atributi.

M5y\_NoFilter – ciljna promenljiva je High Risk, indikatori su kontinualnog tipa, a skup atributa čine svi atributi.

M5y\_NoFilter\_NoRisk – ciljna promenljiva je No Risk, indikatori su kontinualnog tipa, a skup atributa čine svi atributi.

M5y\_Nominal – ciljna promenljiva je High Risk, indikatori su nominalnog tipa, a skup atributa čine samo važni atributi.

M5y\_Nominal\_NoRisk – ciljna promenljiva je No Risk, indikatori su nominalnog tipa, a skup atributa čine samo važni atributi.

M5y\_Nominal\_MI – ciljna promenljiva je High Risk, indikatori su nominalnog tipa, a skup atributa čine važni i marginalni atributi.

M5y\_Nominal\_NoFilter – ciljna promenljiva je High Risk, indikatori su nominalnog tipa, a skup atributa čine svi atributi.

M5y\_Nominal\_NoFilter\_NoRisk – ciljna promenljiva je No Risk, indikatori su nominalnog tipa, a skup atributa čine svi atributi.

M5y\_Nominal\_MI\_NoRisk – ciljna promenljiva je No Risk, indikatori su nominalnog tipa, a skup atributa čine važni i marginalni atributi.

M5y\_NoRisk – ciljna promenljiva je No Risk, indikatori su kontinualnog tipa, a skup atributa čine samo važni atributi.

M5y\_Flag\_NoRisk – ciljna promenljiva je No Risk, indikatori su binarnog tipa, a skup atributa čine samo važni atributi.

M5y\_Flag\_MI\_NoRisk – ciljna promenljiva je No Risk, indikatori su binarnog tipa, a skup atributa čine važni i marginalni atributi.

M5y\_Flag\_NoFilter\_NoRisk – ciljna promenljiva je No Risk, indikatori su binarnog tipa, a skup atributa čine svi atributi.

M5y\_MI\_NoRisk – ciljna promenljiva je No Risk, indikatori su kontinualnog tipa, a skup atributa čine važni i marginalni atributi.

## Literatura

- [1] P. Milenković, Patološka fiziologija,, Beograd: Univerzitet u Beogradu, 2003.
- [2] „[https://kardiologija.in.rs/new\\_page\\_2.html](https://kardiologija.in.rs/new_page_2.html)“.
- [3] P.-N. Tan, M. Steinbach i V. Kumar, Introduction to Data Mining, Pearson Education Limited, 2014.
- [4] SPSSModeler Algorithms Guide, IBM Corporation, 1994, 2020.
- [5] W. Xindong i K. Vipin, The Top Ten Algorithms in Data Mining, Taylor & Francis Group, 2009.
- [6] L. Breiman, J. H. Friedman, R. A. Olsen i C. J. Stone, Classification And Regression Trees, Chapman & Hall/CRC, 1998.
- [7] J. R. Quinlan, C4.5: Programs For Machine Learning, San Mateo, California: Morgan Kaufmann Publishers, 1993.