Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Jaka Koren

Napovedovanje ishemije iz simuliranih podatkov

diplomsko delo

univerzitetni študijski program
prve stopnje
računalništvo in informatika

Mentor: prof. dr. Igor Kononenko
Somentor: prof. dr. Marko Robnik Šikonja

Ljubljana, 2018

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Ishemična bolezen srca (tudi koronarna srčna bolezen) se pogosto konča s srčnomišičnim infarktom in je po vsem svetu eden najpogostejših vzrokov smrti. Ishemijo, ki pomeni nezadostno preskrbo tkiv s krvjo, večinoma povzročajo bolezni ožilja, kar povzroča poškodbe in odmiranje prizadetega tkiva. Sodobni, izpopolnjeni pristopi k modeliranju ishemij z metodo končnih elementov (FEM) zaenkrat še niso primerni za urgentne situacije, kot je npr. takojšnje ukrepanje v primeru ishemije, saj zahtevajo poznavanje specifičnosti anatomije posameznika, ki zahteva natančne in časovno dolgotrajne meritve. Naloga je napovedovanje ishemij s pomočjo strojnega učenja. Končna rešitev bo izbrala takšne klasifikatorje, ki dobro napovedujejo ishemične utripe. Prvostopenjski klasifikator napove, če bo utrip ishemičen, drugostopenjski klasifikator pa bo napovedoval lokacijo ishemičnega področja.

Zahvaljujem se mentorjema Igorju Kononenku in Marku Robniku Šikonji za potrpljenje, podporo in vodstvo, ter profesorici Damjani Hrovat za lektoriranje.

Družini.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
EKG	electrocardiograph	elektrokardiograf, naprava za branje električnih potencialov v srcu in obsrčnem tkivu
CA	classification accuracy	klasifikacijska točnost
RF	random forests	metoda naključnih gozdov
SVM	support vector machine	metoda podpornih vektorjev
GB	gradient boosting	metoda gradientnega pospeševanja
IG	information gain	informacijski pridobitek
MDL	minimum description length	minimalna opisna dolžina

Povzetek

Naslov: Napovedovanje ishemije iz simuliranih podatkov

Avtor: Jaka Koren

V diplomskem delu smo raziskovali možnosti uporabe metod strojnega učenja v napovedovanju ishemičnih dogodkov na podlagi meritev elektrod iz simulacije EKG. Zanimala nas je klasifikacijska točnost pri napovedovanju samih dogodkov, napovedovanju območja v srcu, kjer se je dogodek zgodil, ter možnost zmanjšanja števila elektrod, potrebnih za uspešno zaznavanje ishemij. Na simuliranih podatkih meritev elektrod smo preizkusili naključne gozdove, metodo podpornih vektorjev ter metodo gradientnega pospeševanja in merili njihovo klasifikacijsko točnost. Nato smo z njimi iskali optimalne podmnožice atributov po metodi ovojnice. Poleg tega smo primerjali uspešnost tega pristopa z metodo filtriranja na podlagi ocen pomembnosti atributov. Odkrili smo, da izbrane metode dosegajo dobre rezultate na obeh stopnjah in da klasifikacijska točnost ostaja visoka tudi pri mnogo manjših podmnožicah atributov. Metoda podpornih vektorjev s pomočjo ovojnice dosega najvišjo točnost, glede izbora atributov pa ovojnica pri zelo majhnih izborih atributov dosega boljše rezultate kot filtriranje.

Ključne besede: strojno učenje, napovedovanje, ishemija, klasifikacija, naključni gozdovi, SVM, gradientno pospeševanje.

Abstract

Title: Prediction of Ischemia on simulated data

Author: Jaka Koren

The thesis explores the machine learning approaches for ischemia prediction based on ECG electrode data. We were interested in classification accuracy at prediction of ischemia, prediction of pathological zones in heart, and possibility of reducing the number of atributes neccesary for successful detection. We used simulated data to train and test random forests, support vector machines and gradient boosting. We used these approaches to determine optimal attribute subsets using a wrapper approach, and compared how well methods perform on subsets of various sizes. We also compared the performance of our wrapper approach with a filter-based feature selection approach. Results show high classification accuracy of all methods, even on small attribute subsets. Wrapper assisted support vector machines outperform other methods, and wrapper achieves better results than filtering on small-sized subsets.

Keywords: machine learning, prediction, ischemia, classification, random forests, SVM, gradient boosting.

Poglavje 1

Uvod

Z napredkom računalništva se na vedno več področjih, kjer je potrebna analiza velikih količin podatkov, uporabljajo pristopi strojnega učenja. Velika podjetja že izrabljajo računalniške algoritme za napovedovanje cen, za samodejno vodenje strojev, samovozeče avtomobile itd. Nasprotno pa se v medicini strojno učenje uveljavlja razmeroma počasi. Podatkov za raziskovanje ne primanjkuje, obstaja veliko raziskav, ki aplicirajo metode strojnega učenja. Čeprav si lahko obetamo uspeh, se le malo izsledkov teh raziskav dejansko uporablja v praksi [5].

Eden izmed problemov, kjer bi lahko strojno učenje bistveno pripomoglo k reševanju, je napovedovanje tveganja za ishemijo pri pacientih. Ishemija ali obolenje koronarnih arterij je trenutno eden izmed vodilnih vzrokov smrti v zahodnem svetu. Do obolenja pride zaradi nezadostne oskrbe srca s krvjo, npr. ob zamašitvi srčne arterije, kar povzroči odmiranje mišičnega tkiva. Zgodnji simptom je bolečina v prsih, ki se lahko razširi v druge dele telesa, hujša posledica pa so motnje srčnega utripa. Dolgoročno lahko vodi do zastoja srca [16]. Za uspešno zdravljenje ishemije je nujno, da se obolenje prepozna in locira v zgodnjih fazah. Danes se za diagnozo obolenja najpogosteje uporablja EKG – orodje z elektrodami, ki merijo drobne spremembe v električni napetosti na koži, ki so posledica delovanja srčne mišice. Kardiologi na podlagi meritev tega orodja napovejo tveganje za obolenje; sam

postopek delovanja EKG in diagnoze smo podrobneje opisali v razdelku 2.1.

Takšen način dela je dokazano zanesljiv [17] in zadostuje v zgodnjih fazah obolenja, ko še imamo čas za dolgotrajno merjenje in interpretacijo meritev. Za točno lociranje ishemij si danes pomagamo tudi z metodami končnih elementov [18]. A te metode niso primerne v urgentnih primerih, npr. ob srčnem zastoju zaradi prej nezaznane ishemije, saj zanje potrebujemo podrobno poznavanje anatomije posameznika, kar lahko pridobimo le z dolgotrajnimi in zamudnimi meritvami. Želimo si torej hitro, samodejno orodje, ki bi na podlagi meritev EKG zanesljivo napovedalo prisotnost in območje ishemije. Strojno učenje bi lahko tukaj predstavljalo učinkovito in zadovoljivo točno rešitev.

Namen naloge je bil preizkusiti nekaj metod strojnega učenja na simuliranih podatkih meritev EKG in zgraditi dvostopenjski klasifikator. Na prvi stopnji je ta napovedoval prisotnost ishemije, na drugi stopnji pa lokacijo obolelega območja. Poleg tega smo poskušali zmanjšati število atributov, ki jih dajemo modelom za učenje, in poskušali ugotoviti, kateri so pomembni za naš problem. V 2. poglavju smo opisali delovanje EKG in naših izbranih metod strojnega učenja. 3. poglavje podrobneje predstavlja našo podatkovno množico in ostale metode, s katerimi smo si pomagali pri raziskavi. V 4. poglavju smo predstavili rezultate ocenjevanja atributov in preverjanja metod, v 5. poglavju pa naše sklepne ugotovitve. Na koncu so priložene tabele z dejanskimi meritvami ocen atributov, točnosti metod in izločenih atributov.

Poglavje 2

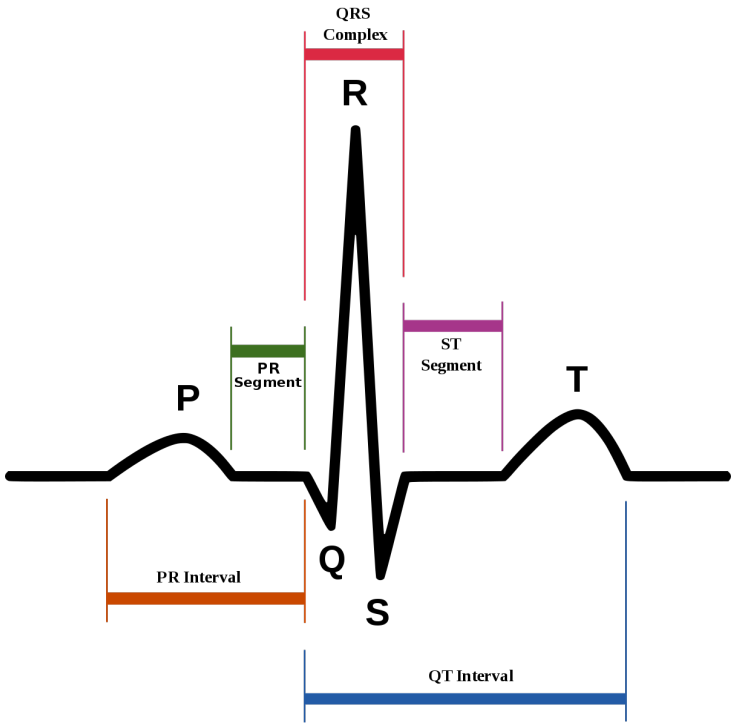
Metode

2.1 Opis EKG

EKG ali elektrokardiograf je naprava za izvajanje elektrokardiografije – procesa snemanja električne dejavnosti v srcu. Osrednji del naprave so elektrode z več odvodi, ki jih prilepimo na kožo pacientovega prsnega koša in udov. Najpogosteje se uporabljajo 10-elektrodni EKG z 12 odvodi. Naprava preko teh meri mikrospremembe električne napetosti na koži, na podlagi katerih zgradi elektrokardiogram – zapis električnega potenciala srca skozi čas. Zdrav srčni utrip se prične z električnim impulzom v preddvorno-prekatnem vozlu, ki se razširi v oba prekata. Hitra polarizacija – dvig električne napetosti v celicah povzroči krčenje prekatov, s čimer srce črpa kri po telesu. Celoten cikel se beleži kot elektrokardiogram – graf električne napetosti v srcu skozi čas [7].

Utrip zdravega srca se na elektrokardiogramu izriše kot značilna krivulja (slika 2.1). Na podlagi magnitude, sprememb ali motenj v krivulji lahko kardiologi dobijo veliko informacij o strukturi srca, pravilnosti srčnega delovanja, napovejo tveganje za različna srčna obolenja ter spremljajo učinke zdravil ali srčnih spodbujevalnikov. Postopku rekonstruiranja dogajanja znotraj telesa na podlagi elektrogardiograma pravimo inverzni elektrokardiografski problem [9].

EKG je relativno nizkocenovno in neinvazivno orodje, zato je bil v prete-



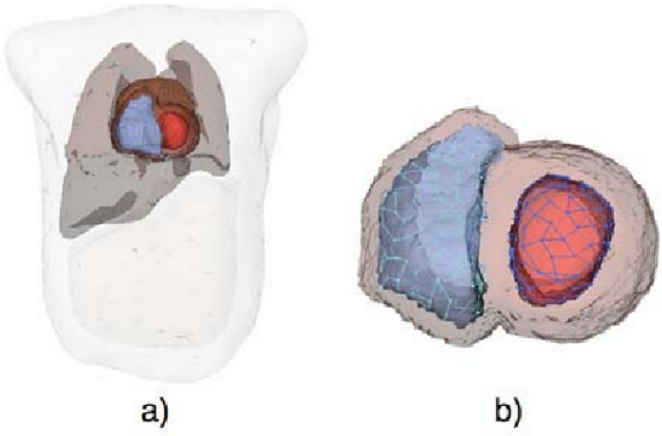
Slika 2.1: Shema krivulje EKG pri zdravem srcu. Interval PR predstavlja depolarizacijo preddvorov, kompleks QRS hitro depolarizacijo prekatov, celoten interval QT pa ponikanje električne napetosti v srcu [19].

klosti temelj več raziskav na tem področju. S. Stern, D. Tzivoni in Z. Stern so na 50 živih primerih že dokazali zanesljivost metode pri diagnosticiranju ishemije v zgodnjih stopnjah [17]. Od 32 pacientov, ki so jih na podlagi EKG prepoznali kot obolele, je bilo 28 dejansko obolelih za ishemijo. Med 18 pacienti, ki glede na EKG niso bili prizadeti, so bili le trije s hujšim koronarnim obolenjem. EKG torej ima potencial kot diagnostično orodje, a prostora za izboljšave je še veliko.

Nekatere raziskave se ukvarjajo z uporabo metod strojnega učenja za bolj točno interpretacijo elektrokardiogramov. C. Papaloukas, D. I. Fotadis, A. Likas in L. K. Michalis so v 2002 preizkusili različne tehnike za samodejno zaznavanje ishemij na podlagi meritev dolgotrajnih EKG [15]. Končni cilj raziskave je bil sistem, ki naj bi prejel signal EKG, odstranil šum, prepoznal

pomembne značilnosti in na podlagi teh klasificiral signal kot ishemičen ali zdrav. Na koncu so pokazali, da lahko pravilno naučene umetne nevronske mreže bolj točno napovejo ishemijo kot drugi sistemi. Prostora za izboljšave je še veliko in sisteme bi bilo treba preizkusiti v praksi.

Drugi raziskovalci poskušajo mehansko izboljšati EKG z iskanjem boljših načinov merjenja. Nekateri se obračajo k umetnim simulacijam. Y. Jiang je z ekipo leta 2009 raziskoval možnosti izboljšanja položajev elektrod EKG za zgodnje zaznavanje ishemij [10]. V ta namen je bil s pomočjo slikanja MRI ustvarjen računalniški anatomski model človeškega trupa (slika 2.2).



Slika 2.2: Grafični prikaz računalniškega modela trupa (a) in srca (b) pacienta [10].

Model vključuje pomembne organe in tkiva z ločljivostjo vokslov velikosti 2 mm. Vsak od teh vokslov predstavlja skupek celic, ki ga sosednji voksli električno vzburijo s svojo dejavnostjo. Model upošteva različno upornost tkiv in tako lahko precej točno simulira prenos napetosti do zunanje plasti trupa. Ishemije se določajo kot območja v trupu, ki se jih ne da vzburiti in ne prenašajo električne napetosti naprej, računalniški model pa omogoča poljubno postavitev in velikost takih območij v srčnem tkivu. S pomočjo te simulacije so reševali problem optimizacije položajev elektrod EKG po telesu. Najbolj optimalni položaj elektrod naj bi pokril območja, kjer je magnituda

površinskih vektorjev električnih potencialov največja. V študiji so razdelili levi prekat na 17 segmentov in v vsakem simulirali tri tipe ishemije v različnih velikostih. Tako so dobili slike površinske električne napetosti na koži, iz katerih so nato izločili območja, kjer so bile razlike v napetosti največje. V njihovem izboljšanem modelu EKG je več elektrod postavljenih na zgornji del prsnega koša, levo ramo in hrbet, rekonstrukcija ishemičnih položajev pa je bolj točna kot pri originalnih položajih elektrod.

2.2 Opis metod strojnega učenja

Strojno učenje je definirano kot samodejen proces modeliranja povezav med podatki. Rezultat postopka je model pravil ali funkcij, s katerim lahko poskušamo razložiti podatke ali napovedati nek izid [12]. Pogost problem, ki ga rešujemo s strojnim učenjem, je uvrščanje ali klasifikacija. Objekt ali primer, predstavljen z naborom spremenljivk ali atributov, želimo uvrstiti v enega od možnih razredov. Naš problem napovedovanja ishemije je klasičen klasifikacijski problem, kjer postavljanje diagnoze pomeni določanje razreda: na prvi stopnji imamo dva razreda za prisotnost/odsotnost obolenja, na drugi stopnji pa več razredov za možno obolelo območje.

Model predstavlja funkcijo, ki preslika vrednosti atributov v končni razred. Za gradnjo modela rabimo algoritem, ki na podlagi prej rešenih problemov istega tipa prepozna povezave med atributi in razredom. Tem problemom pravimo učni primeri. Točnost modela preverjamo s prej neznanmi rešenimi problemi in spremljamo razliko med napovedjo modela in dejanskim razredom. Tem problemom pravimo testni primeri.

Na voljo je mnogo algoritmov strojnega učenja – metod, ki te probleme rešujejo z različnimi pristopi. V tem poglavju bomo predstavili metode strojnega učenja, ki smo jih izbrali za preizkušanje na naši podatkovni množici.

2.2.1 Naključni gozdovi

Ansambelske metode (angl. *Ensemble methods*) so način kombiniranja preprostih klasifikatorjev, kot so npr. odločitvena drevesa, v bolj kompleksne modele z boljšimi rezultati. Predlaganih načinov povezovanja je veliko – kombiniranje po Bayesovi metodi, uteženo oziroma neuteženo glasovanje, dinamično izbiranje ... Kot zelo uspešna metoda se je izkazal *bagging* (okrajšava za *Bootstrap aggregating*) – iz učne množice naključno z vračanjem izberemo N primerov, na katerih poženemo učni algoritem, npr. odločitveno drevo. Z več takimi množicami generiramo več dreves, ki pri klasifikaciji novih primerov glasujejo, v kateri razred bo primer uvrščen [1].

Naključni gozdovi (angl. *Random forests* – RF) so nadgradnja *bagginga*, ki doda več naključnosti v sam proces učenja. Poleg učnih primerov zdaj še za vsako drevo posebej izberemo naključno podmnožico atributov. Vozlišča v drevesu določamo samo na podlagi atributov iz te podmnožice. Velikost naključnih podmnožic atributov za učenje naključnih gozdov je lahko poljubna, ravno tako število dreves v gozdu. Breiman v svojem članku [2] za velikost podmnožic predlaga število, enako logaritmu števila atributov plus ena. Za velikost gozda se ponavadi uporablja 100 dreves, lahko pa tudi več.

Zaradi same količine dreves so naključni gozdovi nepregledni, a hkrati robustni na šum v učni množici. Več kot je dreves v gozdu, manjša je klasifikacijska napaka modela. V praksi modeli RF dejansko dosegajo visoko točnost, primerljivo z najboljšimi metodami. V naši nalogi smo naključne gozdove implementirali s pomočjo paketa *randomForest* za okolje R.

2.2.2 SVM

Support Vector Machines oz. metoda podpornih vektorjev je model strojnega učenja, ki se loti problema dvorazrednega klasificiranja z uporabo linearne diskriminantne funkcije [4]. Druge metode strojnega učenja pogosto iščejo ustrezno podmnožico atributov s čim višjo pomembnostjo glede na končni razred, SVM pa daje prednost kombinacijam atributov. V praksi metoda

dosega dobre rezultate, tudi na področju medicinske diagnostike.

Osnovna ideja algoritma je iskanje hiperravnine, ki najbolj jasno ločuje dva razreda v prostoru atributov. Naših n učnih primerov z t atributi predstavimo kot vektorje (x_i, y_i), $i = 1 \ldots n$, kjer x predstavlja vektor dolžine t z vrednostmi zveznih atributov primera, y pa razred, ki mu primer pripada. Če pripada prvemu, je $y_i = 1$, če drugemu, pa y_i = -1. Nas zanima optimalna hiperravnina, kjer je razdalja (margin) med vsemi vektorji primerov enega in drugega razreda ter samo ravnino maksimalna. Imamo torej problem maksimizacije funkcije skalarnih produktov, ki ga rešujemo z vstavljanjem primerov učne množice v funkcijo.

Problem takega pristopa je računska zahtevnost, še posebej v množicah z velikim številom atributov. Vsak primer posebej je treba eksplicitno preslikati v atributni prostor, kjer se število atributov še dodatno poveča, in nato računati njihove skalarne produkte z novimi primeri. SVM ta problem reši tako, da poišče učne primere, ki so potencialni hiperravnini najbljižje – tem primerom pravimo podporni vektorji. V praksi ti predstavljajo le 3 – 5 odstotkov celotne učne množice, a lahko iz njih izpeljemo optimalno hiperravnino. Za izpeljave lahko uporabljamo različne jedrne funkcije.

Ker računamo le z implicitnimi transformacijami vektorjev, ima metoda SVM nizko računsko zahtevnost in je zato primerna za množice podatkov z velikim številom manj pomembnih atributov. V osnovi je metoda narejena samo za dvorazredno klasifikacijo, a obstajajo metode za nadgradnjo v to smer [8], kot npr. "one-aganist-one"pristop. Metoda ustvari več binarnih klasifikatorjev za vsako možno kombinacijo razredov ($r(r-1)/2$ kombinacij), nove primere pa klasificira z glasovanjem. Ta pristop je implementiran v knjižnici *e1071*, ki smo jo uporabljali tudi mi.

2.2.3 Gradient Boosting

Pospeševanje (angl. *boosting*) je, podobno kot *bagging*, metoda združevanja preprostih klasifikatorjev v bolj kompleksen model, le da jih združuje zaporedno na podlagi prej uporabljenih učnih primerov. V vsakem naslednjem

koraku določamo učno podmnožico na podlagi "pomembnosti"primerov, ki je določena z utežmi.

Na začetku imajo vsi primeri v celotni učni množici enako utež, vsak ima enako verjetnost, da se bo pojavil v učni podmnožici. Na prvem koraku naučen model nato napove razrede za primere v tej podmnožici – primerom, ki so pravilno uvrščeni, zmanjšamo utež, ostalim pa jo povečamo. V sledečih korakih pri izbiranju nove učne podmnožice damo prednost primerom z večjo težo – primerom, ki so jih prejšnji modeli v zaporedju uvrstili narobe. Vsak naslednji člen v zaporedju modelov tako naučimo, da čim bolje dopolnjuje prejšnje modele. Postopek lahko ponavljamo, dokler napaka ne ostane dovolj majhna ali pa postane prevelika, ker se preostalih problemov ne da rešiti. Pri napovedovanju novih primerov uporabimo vse modele, a jih še dodatno utežimo glede na položaj v zaporedju – bolj točni imajo večjo težo pri glasovanju. Končni razred se ne napove eksplicitno, model poda verjetnost, da se primer nahaja v nekem razredu.

Breiman opaža, da gre pri tem načinu povezovanja za problem iskanja minimuma v funkcionalu klasifikacijske napake modela [6]. Takšen minimum se da iskati z algoritmom gradientnega spusta – v prostoru atributov požrešno premikamo trenutni učni izbor glede na napako, ocenjeno v prejšnjih korakih.

Drugi raziskovalci so sčasoma razvili algoritme, ki to idejo izrabljajo v praksi [13]. Ti omogočajo poljubno nastavitev različnih parametrov, med drugim uporabo drugačnih funkcij napake – tipično povprečno kvadratno napako (angl. *Mean Squared Error*). Prilagodljivost in robustnost omogočata modelu zelo visoko natančnost tudi v praksi.

Poglavje 3

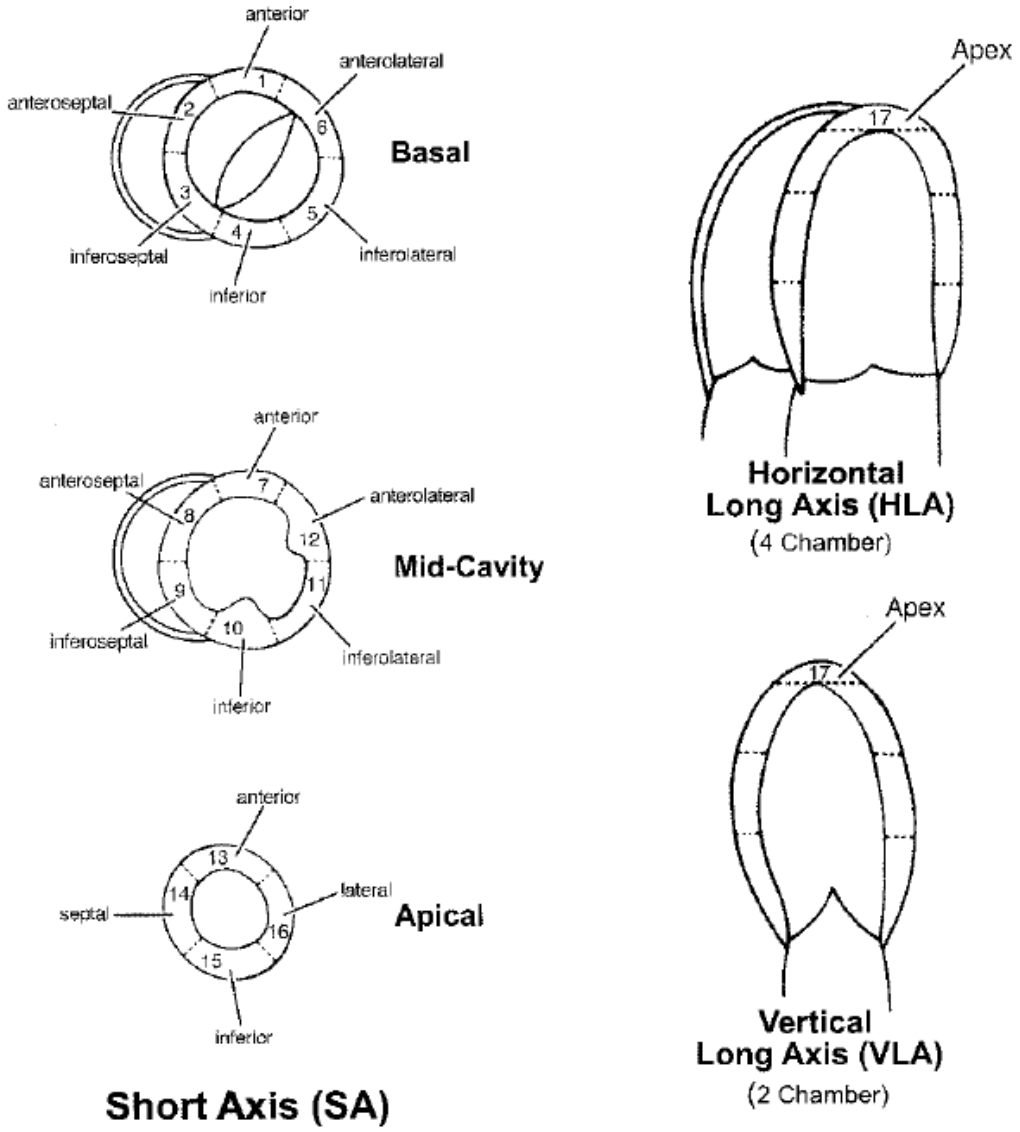
Metodologija

V tem poglavju smo podrobneje predstavili našo podatkovno množico ter metode, s katerimi smo si pomagali pri učenju metod in iskanju podmnožic atributov.

3.1 Podatki

Podatki, ki smo jih uporabljali v raziskavi, so pridobljeni z računalniškim modelom človeškega trupa. S pomočjo metode končnih elementov (angl. Finite elements method) ustvarjen model simulira učinek srčnega utripa na električni potencial na površini kože [14]. Vsak primer v množici predstavlja eno simulirano meritev z elektrodami na 100 različnih položajih. Te elektrode so naši atributi – vsak predstavlja električni potencial, izmerjen na položaju elektrode. Za vsako vrstico sta podana dva izhodna razreda: prvi razred določa samo prisotnost ishemije na srcu, drugi razred pa, če je ishemija prisotna, položaj obolelega območja. Ta je predstavljen kot eno izmed 17 območij (slika 3.1) v levem prekatu srca v skladu s priporočili Ameriške zveze za srčna obolenja [3].

Vseh primerov skupaj je v tabeli 1700, od tega polovica predstavlja primere brez ishemije, druga polovica pa primere z obolenjem in območjem. Med temi 850 primeri so razredi s 17 možnimi območji obolenja razporejeni



Slika 3.1: Shema delitve levega prekata srca na segmente [3].

enako, kar pomeni 50 primerov potrjene ishemije v nekem območju.

Ker so metode strojnega učenja, še posebej SVM, zelo občutljive na razlike v vrednostih atributov med posameznimi primeri, smo najprej podatkovno množico normalizirali. Meritve v vsakem atributnem stolpcu posebej smo matematično skalirali na interval [0,1]. Množico smo razdelili v dve tabeli

glede na klasifikacijo, ki smo jo izvajali. Prva množica vsebuje vseh 1700 primerov s prisotnostjo ishemije – stolpcem ISCHEMIA kot končnim razredom. Druga množica vsebuje 850 primerov s prisotno ishemijo, kot končni razred pa je nastavljeno območje – stolpec ZONE.

3.2 Preizkušanje

Za ocenjevanje modelov in testiranje smo uporabili klasifikacijsko točnost (angl. *Classification Accuracy* – CA). Če je N število vseh primerov v testni množici, naš model pa pravilno klasificira Np primerov, je CA enaka razmerju:

$$CA = Np/N \tag{3.1}$$

Metrika torej izraža delež pravilnosti pri napovedovanju razreda. Porazdelitev razredov lahko sicer vpliva na uporabnost metrike, ampak v našem primeru so razredi enakomerno porazdeljeni.

Na začetku smo modele ocenili s prečnim preverjanjem (angl. *cross-validation*) [12]. Ta metoda učinkovito izrabi celotno množico podatkov za učenje in testiranje modela:

1. Primere naključno razdelimo v k enako velikih rezin podatkovne množice.

2. Za vsako rezino zgradimo hipotezo na podlagi preostalih k– 1 rezin in jo preverimo s trenutno izbrano rezino.

3. Rezultate testiranj za vsako rezino povprečimo, da dobimo končno oceno modela.

V praksi se ta pristop pogosto uporablja za ocenjevanje modelov, še posebej v primerih, ko je množica podatkov majhna. Mi smo prečno preverjanje izvajali posebej na množici za klasifikacijo ishemije in posebej na množici za klasifikacijo ishemičnih območij.

3.3 Izbira atributov

Pogosto je bolj kot izbira primernega modela za reševanje klasifikacijskih problemov pomembna izbira atributov, ki jih dajemo modelu za učenje. Prednost izbiranja pomembnih spremenljivk je trikratna: model se na manjši množici podatkov hitreje nauči zakonitosti razredov, izboljša se natančnost pri napovedovanju in s tem tudi naše razumevanje delovanja modela. V našem primeru bi v bodoče potrebovali manj senzorjev, s katerimi bi napovedovali ishemijo in njeno lokacijo. Vprašanje, kako iskati primerne podmnožice atributov, je bilo deležno mnogih raziskav, v praksi pa sta se uveljavila dva pristopa.

Prvi pristop vsaki spremenljivki dodeli oceno pomembnosti glede na razred. Na podlagi ocene določimo n najpomembnejših, ki jih uporabimo za treniranje modela. Temu pristopu pravimo filtriranje. Za ocenjevanje atributov se uporablja mnogo različnih mer, večina jih je osnovana na količini informacije atributa. Da izvemo, kateri izid izmed n možnih nezdružljivih izidov X se je zgodil, potrebujemo

$$H(X) = -\sum_{i=1}^{n} P(X_i) \log_2 P(X_i) \tag{3.2}$$

bitov informacije. Tej meri pravimo entropija razreda. Mere, kot so informacijski prispevek, razmerje prispevka in MDL [12], poskušajo predvsem izračunati, kolikšen delež te informacije prispeva ta atribut glede na vrednosti v svoji domeni. Njihov problem je v sklepanju, da so atributi medsebojno neodvisni. Rešitev ponuja metrika Relief, ki za nek učni primer oceni pomembnost atributov na podlagi podobnih primerov iz istega razreda (najbližji zadetek) in podobnih primerov iz nasprotnega razreda (najbližji pogrešek). S takšno lokalnostjo lahko implicitno ocenjuje kvaliteto atributov v odvisnosti od drugih. Izboljšane inačice algoritma lahko ocenjujejo tudi atribute v večrazrednih množicah in so bolj odporne na šum ali manjkajoče podatke [12].

Filtriranje ima to prednost, da ga lahko uporabimo v kombinaciji s katerimkoli modelom. Ocenjevanje je v primerjavi s strojnim učenjem časovno

nezahtevno. Kljub temu ni zagotovljeno, da bo izbor atributov primeren za naš model. Drugi pristop naslavlja ta problem z metodo *Wrapper* (ovojnica). Kohavi in John sta leta 1997 predstavila algoritem, ki v podatkovni množici išče optimalno podmnožico atributov s pomočjo modela strojnega učenja [11]. Znotraj izbrane množice podatkov za učenje algoritem na poljuben način izbere več podmnožic atributov in z njimi trenira naš model. Tako postavi več hipotez, na podlagi katerih oceni, katere podmnožice dajejo najboljše rezultate za naš model. Najboljša podmnožica atributov se uporabi kot končna množica, na kateri bomo zgradili naš model. Končni klasifikator se preizkusi na ločeni podatkovni množici, ki ni bila uporabljena v procesu iskanja atributov.

Ker je implementacija preprosta in se atributi dobro prilegajo izbranim modelom, so metode ovojnice pogoste in v praksi dosegajo zelo dobre rezultate. Njihova slabost pa je zamudnost algoritma. Za vsako podmnožico atributov moramo naučiti in preizkusiti model, kar pa se ne obnese pri podatkih z veliko atributi. Za učinkovitost je treba najti primeren iskalni algoritem, kar pozitivno vpliva na točnost končnega klasifikatorja.

V naši raziskavi smo za izbiranje atributov implementirali naslednji algoritem ovojnice. Celotno podatkovno množico najprej razdelimo na učno in testno množico tako, da se ohrani enakomerna porazdelitev razredov. Velikost testne množice je enaka eni petini celotne. Pri učni množici nato poženemo naslednji algoritem:

Algorithm 1 Ovojnica za iskanje podmnožic atributov

Attr_subsets ← prazen seznam najboljših podmnožic atributov
Tab_CA ← prazna tabela CA najboljših podmnožic med iskanjem
Tab_FCA ← prazna tabela CA najboljših podmnožic pri končnem testiranju
train_set ← podatkovna množica za treniranje
test_set ← podatkovna množica za testiranje
Attr ← množica vseh atributov
$N \leftarrow$ velikost(*Attr*)
while $N >= 1$ **do**
    $m \leftarrow N$
    subtrain_set ← podmnožica *train_set* za treniranje modelov med iskanjem
    subtest_set ← *train_set* − *subtrain_set*
    CA_Best ← 0
    Attr_Best ← prazen seznam atributov
    while $m >= 1$ **do**
        *Attr*2 ← Attr - Attr[m]
        *subtrain_set*2 ← podmnožica *subtrain_set*, določena z atributi *Attr*2
        *subtest_set*2 ← podmnožica *subtest_set*, določena z atributi *Attr*2
        CA ← *treniraj_model* z *subtrain_set*2 in *subtest_set*2
        if *CA* > *CA_Best* **then**
            CA_Best ← *CA*
            Attr_Best ← *Attr*2
        $m \leftarrow m - 1$
    Tab_CA[*N*] ← *CA_Best*
    Attr_Subsets[*N*] ← *Attr_Best*
    $N \leftarrow N - 1$
$N \leftarrow$ velikost(*Attr*)
while $N >= 1$ **do**
    *train_set*2 ← podmnožica *train_set*, določena z atributi *Attr_Subsets*[*N*]
    *test_set*2 ← podmnožica *test_set*, določena z atributi *Attr_Subsets*[*N*]
    Tab_FCA[*N*] ← *treniraj_model* z *train_set*2 in *test_set*2
    $N \leftarrow N - 1$

Poglavje 4

Rezultati

4.1 Ocene atributov

Najprej smo ocenili pomembnost atributov glede na končna razreda. Uporabili smo informacijski prispevek (angl. *Information Gain* – IG), minimalno dolžino opisa (angl. *Minimum description length* – MDL), Relief in Gini-indeks. Ocene atributov so podrobneje predstavljene v dodatku A.

Informacijski prispevek in MDL podobno ocenjujeta atribute, veliko se jih pri obeh podobno uvrsti. Pri ocenjevanju glede na razred za prisotnost ishemije (ISCHEMIA) podajo IG, MDL in Gini-indeks podobne ocene. Pri prvih dveh mnogi atributi dosežejo enako razvrstitev glede na pomembnost. Atributi "Node 471", "Node 16389", "Node 494", "Node 4511", "Node 17236", "Node 17268" ter "Node 330" dosegajo pri vseh najvišja mesta. Relief tem po drugi strani daje manjšo prednost, "Node 330" se na njegovi lestvici znajde na 14. mestu. Na prvih pet mest postavlja atribute "Node 3921", "Node 3818", "Node 3588", "Node 3658" in "Node 343", ki pri drugih ocenah dosegajo precej slabše rezultate. Prvouvrščeni "Node 3921" pri MDL in IG dosega 17., pri Gini-indeksu pa 16. mesto, "Node 3588" pa je pri teh uvrščen še nižje. Zanimivo pa se pri vseh ocenah atribut "Node 949" uvršča na zadnje mesto. Ocene, ki jih prejme, so opazno nižje od ocen ostalih atributov.

Podobno je pri ocenjevanju glede na razred za območje ishemije (ZONE).

IG in MDL razvrščata atribute zelo podobno, na prvih 7 mest postavljata enake atribute celo v enakem vrstnem redu. Ocene Gini-indeksa se tokrat bolj razlikujejo, tako Gini kot Relief postavljata prej prvouvrščene atribute na precej nižja mesta. Relief daje atributu "Node 792", ki je glede na IG in MDL najbolj pomemben, drugo najnižje mesto. Najslabšo oceno vse mere ponovno dajo atributu "Node 949".

Glede na razpon ocen lahko sklepamo, da je veliko atributov nepomembnih in bi jih lahko odstranili iz množice. Razpored pomembnosti pa se med različnimi metodami ocenjevanja precej razlikuje. Pri iskanju idealne podmnožice atributov smo zato raje izbrali metodo ovojnice, rezultate pa smo opisali v razdelku 4.4.

4.2 Rezultati prečnega preverjanja

Naše metode smo ločeno prečno preverjali na obeh problemih. Tabela 4.1 prikazuje izmerjeno CA za vsako metodo na množici za klasifikacijo ishemije in na množici za klasifikacijo ishemičnih območij. Vrednosti v tabeli predstavljajo povprečene meritve pri prečnem preverjanju – naši množici podatkov smo razdelili na 10 enakih rezin.

Vidimo, da vsi modeli na obeh problemih dosegajo visoko točnost. Pri klasificiranju prisotnosti ishemije dosegata SVM in gradientno pospeševanje najvišjo točnost, naključni gozdovi pa so jima zelo blizu. Večja razlika pa se opazi pri klasificiranju območja ishemije, kjer SVM dosega opazno višjo klasifikacijsko točnost. Drugi metodi dosegata podoben rezultat, okoli 86

Metoda	CA - ISCHEMIA	CA - ZONE
Naključni gozdovi	0.9476	0.8611
SVM	0.9511	0.9552
Gradientno pospeševanje	0.9511	0.8682

Tabela 4.1: Izmerjena klasifikacijska točnost metod pri prečnem preverjanju

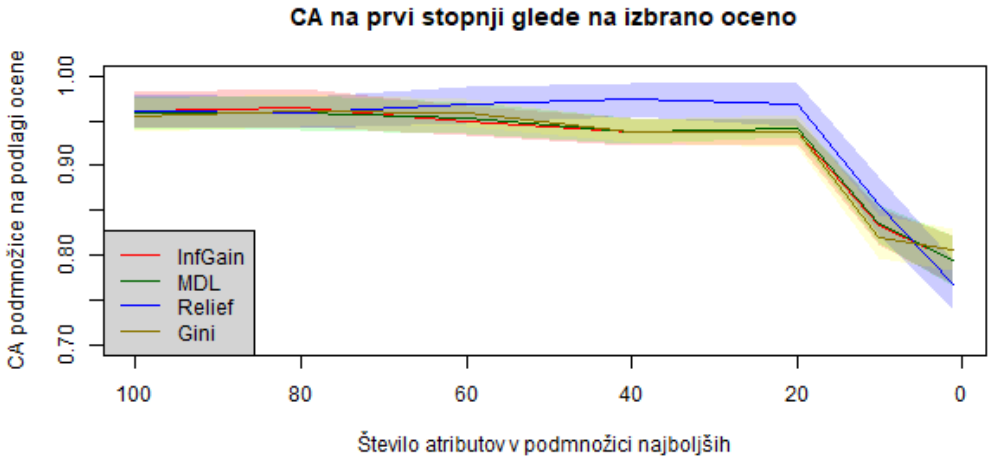
odstotkov, kar pa je še vedno precej dobro.

4.3 Rezultati filtriranja atributov

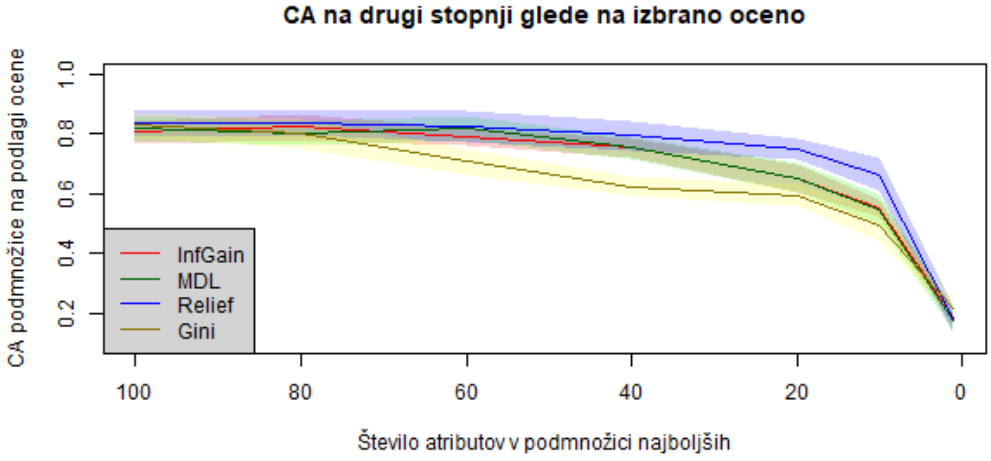
Zanimalo nas je, kako se naše metode obnesejo pri manjših podmnožicah atributov oz. koliko atributov lahko odstranimo in še vedno ohranimo primerljivo klasifikacijsko točnost. Na podlagi naših ocen atributov smo sestavili podmnožice velikosti 100, 80, 60, 40, 20, 10 in 1 najboljši atribut. Naši množici podatkov smo razdelili v učno in testno podmnožico v razmerju 4:1. Na njiju smo nato zaporedoma učili in preizkušali metode strojnega učenja s podmnožicami atributov, kot so jih določile mere za ocenjevanje. Grafi 4.1– 4.6 prikazujejo CA na podmnožicah različnih velikosti pri različnih merah, osenčena območja pa standardni odklon CA pri prečnem preverjanju. Upoštevati je treba, da vrednost *Število atributov v podmnožici* predstavlja samo velikost podmnožice. Sam izbor atributov za podmnožico velikosti N se med posameznimi merami razlikuje. Tabele z izmerjenimi CA za vsako metodo pri različnih podmnožicah so v dodatku B.

Na vseh grafih pri obeh stopnjah opazimo, da podmnožice, izbrane z mero Relief, dosegajo primerljivo točnost tudi pri manjših velikostih. Pri naključnih gozdovih in *boostingu* so CA pri večjih podmnožicah bolj primerljive, šele pri naborih 40 in manj se pokažejo večje razlike. Pri SVM ima Relief pri obeh problemih že na začetku opazno prednost. Zanimivo je, da dajejo na prvi stopnji vse ostale ocene zelo podobne rezultate, pri drugem problemu pa mera Gini vidno zaostaja tudi za MDL in informacijskim prispevkom.

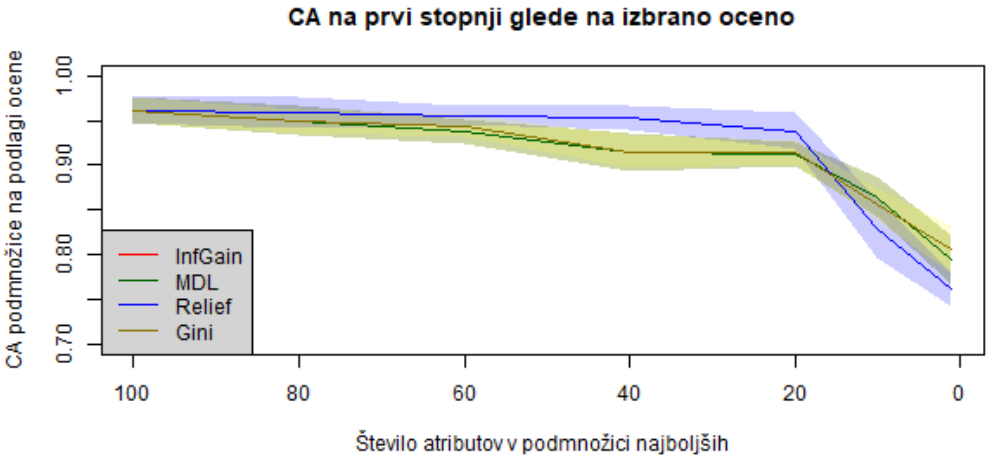
Na prvi stopnji najboljši rezultat dosežejo naključni gozdovi, s CA 0.97 pri 40 najboljših atributih na podlagi ocene Relief. GB z Reliefom doseže enak rezultat pri 80 atributih. Na drugi stopnji z Reliefom prednjači SVM z 91-odstotno točnostjo pri 40 atributih in 80-odstotno pri 10 najboljših atributih. Opažamo torej, da je mogoče ohraniti visoko CA teh metod tudi pri manjšem izboru atributov. Ostaja pa vprašanje, ali so izbori atributov res najbolj primerni za naše metode in ali lahko z drugačnim izborom še izboljšamo



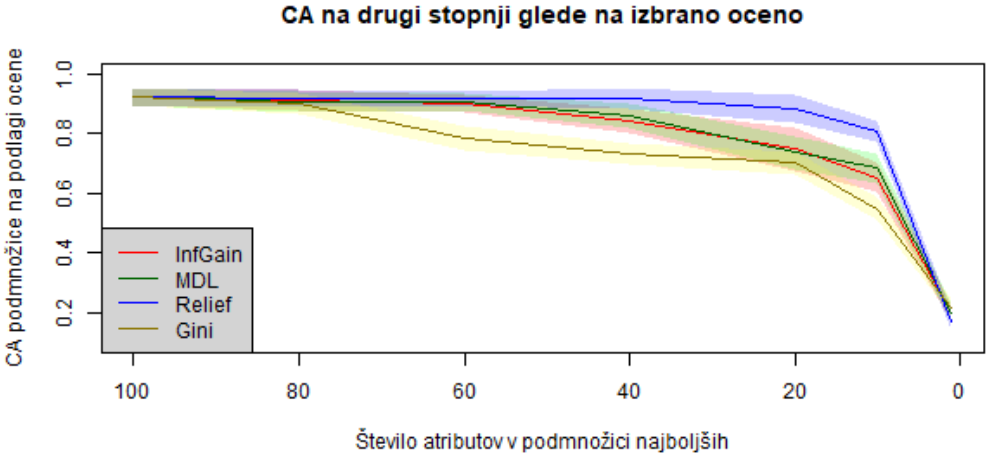
Slika 4.1: CA metode naključnih gozdov pri razredu ISCHEMIA na podmnožicah najboljših n atributov, izbranih na podlagi različnih algoritmov ocenjevanja.



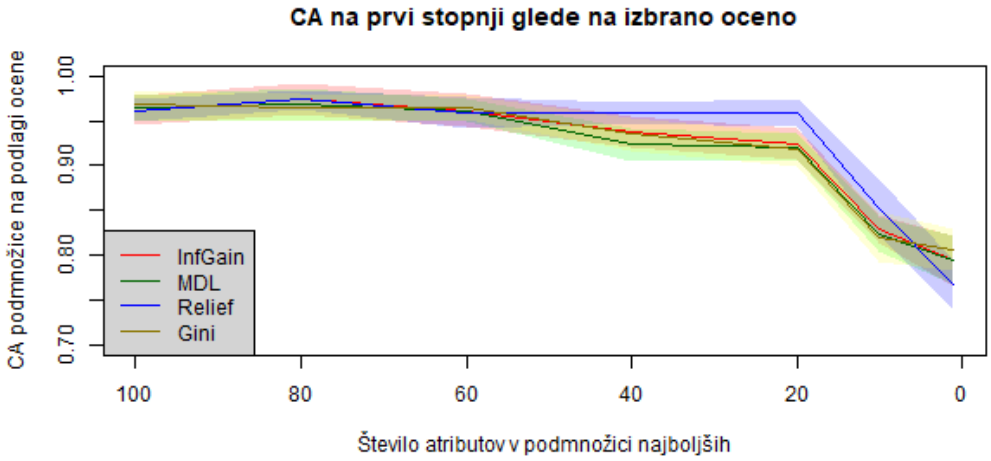
Slika 4.2: CA metode naključnih gozdov pri razredu ZONE na podmnožicah najboljših n atributov, izbranih na podlagi različnih algoritmov ocenjevanja.



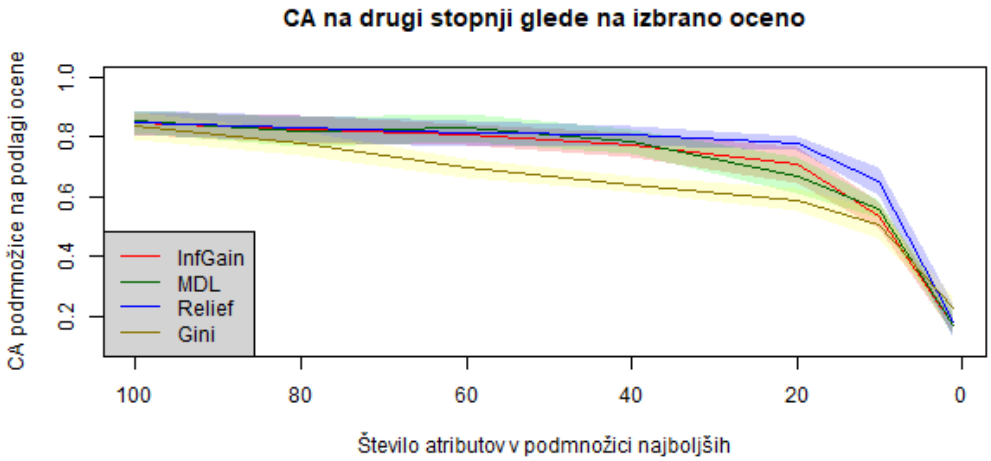
Slika 4.3: CA metode SVM pri razredu ISCHEMIA na podmnožicah najboljših n atributov, izbranih na podlagi različnih algoritmov ocenjevanja.



Slika 4.4: CA metode SVM pri razredu ZONE na podmnožicah najboljših n atributov, izbranih na podlagi različnih algoritmov ocenjevanja.



Slika 4.5: CA metode Gradient Boosting pri razredu ISCHEMIA na podmnožicah najboljših n atributov, izbranih na podlagi različnih algoritmov ocenjevanja.



Slika 4.6: CA metode Gradient Boosting pri razredu ZONE na podmnožicah najboljših n atributov, izbranih na podlagi različnih algoritmov ocenjevanja.

rezultate.

4.4 Rezultati ovojnice za izbiro atributov

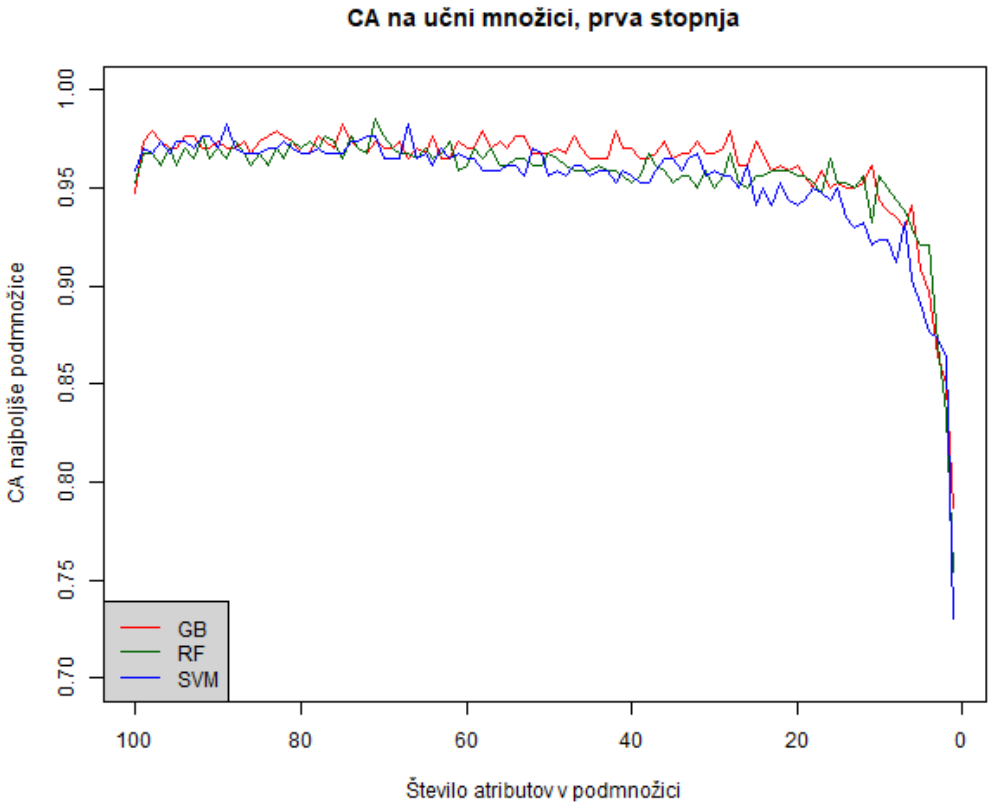
V ta namen smo implementirali metodo ovojnice, ki izbira podmnožice atributov na podlagi klasifikacijske točnosti metod. Naša ovojnica sprejme obe podatkovni množici in ju loči na učno in testno podmnožico, nato pa nad izbrano metodo izvaja ločene teste.

Grafi na slikah 4.7–4.10 prikazujejo spreminjanje CA metod pri različnih velikostih podmnožic atributov. Upoštevati je treba, da vrednost Število atributov v podmnožici predstavlja samo velikost podmnožice. Sam izbor atributov za podmnožico velikosti N se med posameznimi metodami razlikuje. Dodatek D vsebuje točne meritve CA za vse metode, dodatek C pa zaporedja izločenih atributov.

4.4.1 Prva stopnja – klasificiranje ishemij

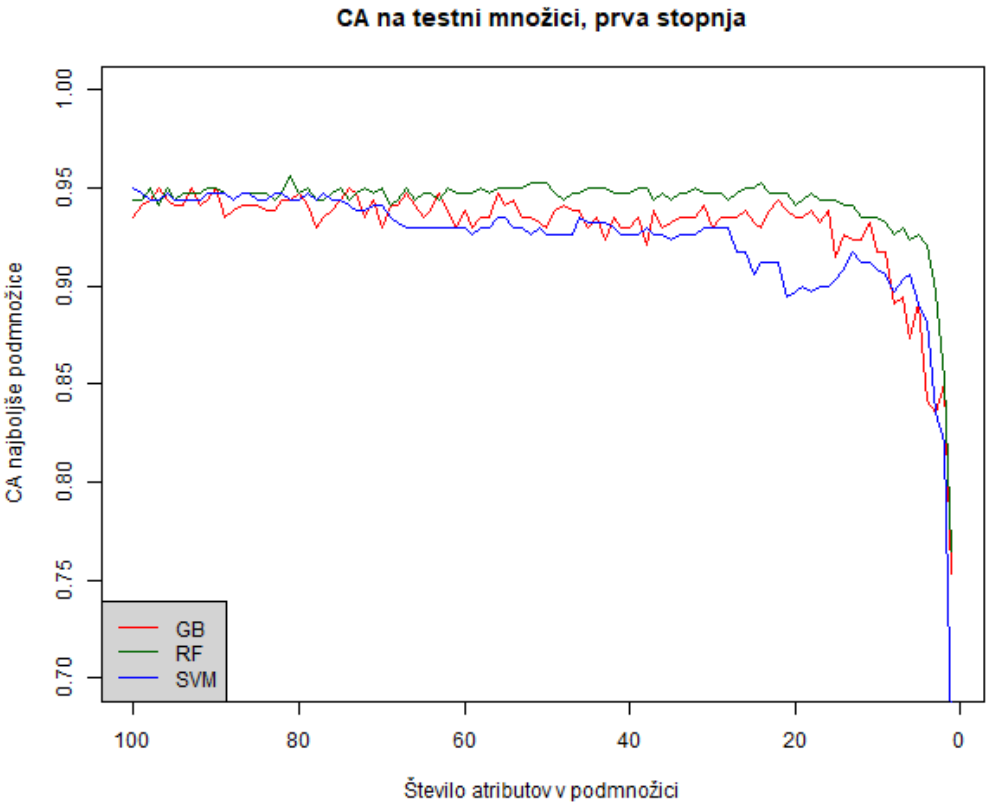
Pri prvi stopnji (slika 4.7) vidimo, da metode ohranjajo visoko točnost tudi pri veliko manjšem izboru atributov. Na učni množici se pri izborih, večjih od 40 atributov, CA giblje okoli 0.97. Opazimo celo, da je ta metrika boljša kot pri učenju na vseh 100 atributih. Razlike med posameznimi metodami so pri velikih izborih atributov zelo majhne, a v povprečju lahko vidimo, da največjo točnost dosega GB, ki mu sledita SVM in RF. Ko se velikost izbora manjša, se CA SVM in RF postopoma spušča za nekaj odstotkov, GB pa ohranja približno enake povprečne vrednosti. Pri izborih atributov, manjših od 30, začne pri vseh metodah CA pričakovano močno padati. A tudi na podlagi enega samega izbranega atributa dosega GB točnost 0.77, SVM in RF pa 0.75. Na učni množici torej GB opazno vodi med izbranimi metodami.

Na testni množici (slika 4.8) na prvi stopnji je pri velikih izborih vzorec podoben, klasifikacijska točnost začne bistveno padati šele pri izborih, manjših od 40. Kot pričakovano je za vse metode klasifikacijska točnost tukaj malo nižja kot pri učni množici, a sprva še vedno višja od 0.9. Je pa tokrat



Slika 4.7: CA v odvisnosti od velikosti izbrane množice atributov na učni množici za prvo stopnjo.

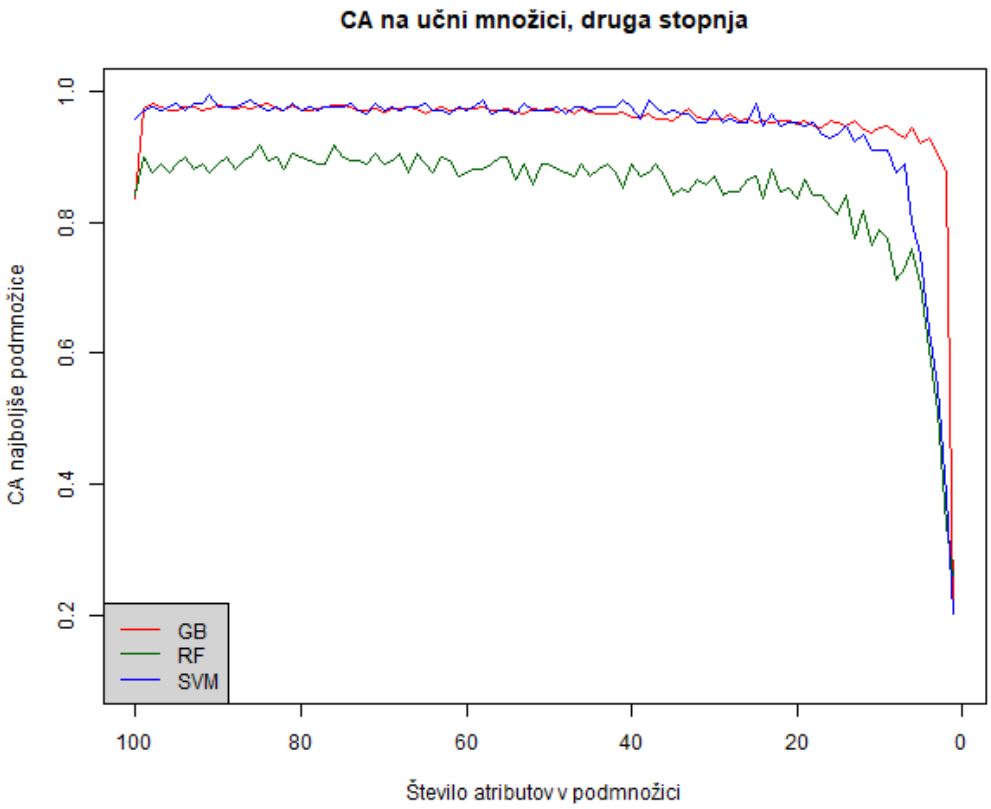
vidna prednost RF, ki v povprečju dosega višje in bolj stabilne točnosti kot GB in SVM, ki je tokrat na zadnjem mestu. Ko velikost množice pade na 70, se klasifikacijska točnost SVM spusti opazno nižje kot RF, pri 30 atributih pa vidno pade in doseže lokalno dno pod 0.9 pri 21 atributih. RF in GB začneta bistveno izgubljati točnost šele pri izborih, manjših od 15 atributov, SVM pa se tukaj še zadnjič povzpne do 0.91 CA. Pri enem samem atributu dosega RF točnost 0.76, GB 0.75, SVM pa 0.67.



Slika 4.8: CA v odvisnosti od velikosti izbrane množice atributov na testni množici za prvo stopnjo.

4.4.2 Druga stopnja – klasificiranje območij ishemij

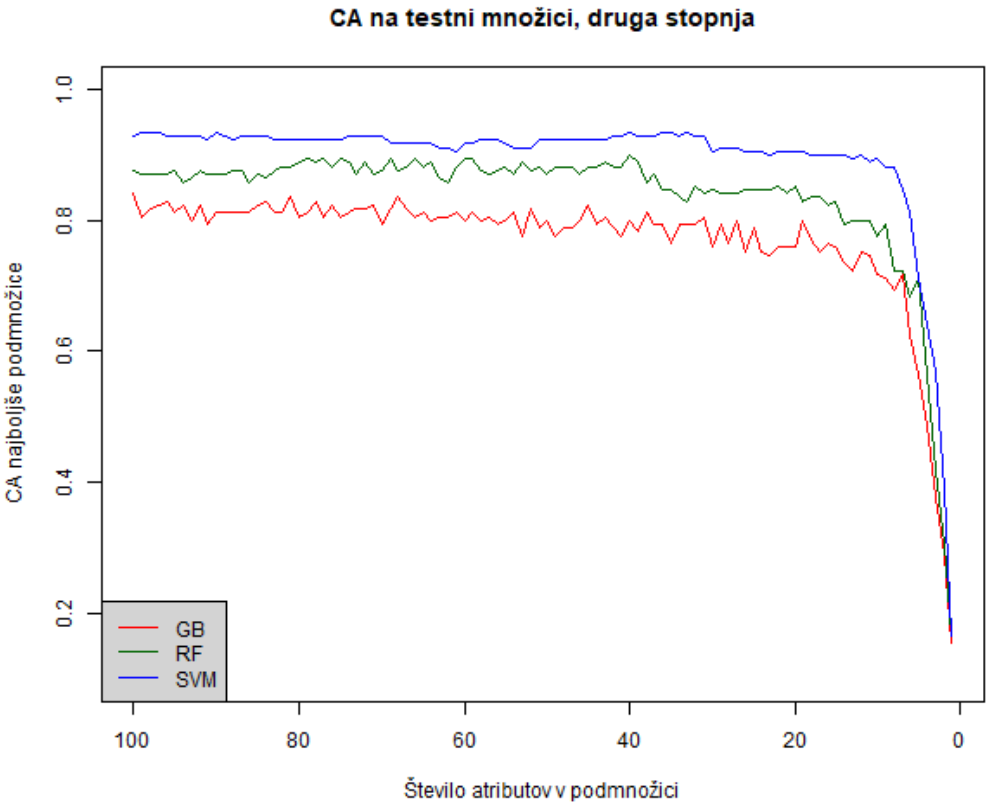
Pri klasifikaciji območja ishemije se pojavi podoben vzorec, a razlike med posameznimi metodami so tokrat bolj očitne. Na učni množici ponovno vidimo strm dvig klasifikacijske točnosti takoj po odstranitvi nekaj motečih atributov. SVM in GB tukaj dosežeta točnost 0.98, ki ne začne bistveno padati, dokler ne zmanjšamo izbora pod 20 atributov. Tam SVM prvi popusti in se postopoma spusti do točnosti 0.2 pri enem samem atributu. GB pri dveh atributih še vedno pravilno klasificira 88 odstotkov primerov, pri enem pa 20, enako kot SVM. RF se na učni množici ves čas drži pod prejšnjima



Slika 4.9: CA v odvisnosti od velikosti izbrane množice atributov na učni množici za drugo stopnjo.

metodama, pri naborih, večjih od 20, se CA giblje malo pod 0.90, pod 20 atributi pa pade. Pri enem samem atributu se obnese malo bolje kot SVM in GB s točnostjo 0.25 .

Na naši testni množici se ponovno SVM izkaže kot boljši pri izborih atributov, večjih od 30, kjer drži povprečno točnost 0.92, pri velikostih med 30 in 10 se spusti na 0.90. Ko se velikost zmanjša pod 10, začne CA strmo padati in se ustavi pri točnosti 0.16 pri enem atributu. Sledi mu RF, ki ima pri izborih, večjih od 40 atributov, povprečno točnost 0.88, zatem pade na 0.82 in pri izborih pod 10 atributi začne padati enako kot SVM na 0.16 pri enem atributu. GB ponovno pade najnižje, CA postopoma pade na 0.8 pri



Slika 4.10: CA v odvisnosti od velikosti izbrane množice atributov na testni množici za drugo stopnjo.

40 atributih, nadalje pa se giblje med 0.8 in 0.75, dokler pod 20 atributi ne začne padati in se ustavi pri 0.15 na enem samem atributu.

4.4.3 Opažanja

Prečno preverjanje nam je nakazalo vzorec, ki se pri izločanju atributov ohranja. Vse naše metode dosegajo nad 94-odstotno CA pri klasificiranju ishemije ter nad 86-odstotno CA pri klasificiranju območja. Na prvi stopnji prednjačita RF in GB z majhno prednostjo, na drugi stopnji pa SVM dosega precej višjo točnost od drugih dveh.

Naša domneva, da veliko atributov ni bistvenih za naš problem, je potrjena. Pri vseh metodah vidimo, da se, dokler se velikost nabora ne zmanjša pod 20, ohranja blizu začetne vrednosti. Opažamo tudi vrhove, kjer je CA pri manjših izborih atributov celo višja kot pri polnem izboru.

Na prvi stopnji na koncu po točnosti prevlada RF. GB mu sledi z malo nižjo, a še vedno sorazmerno konsistentno krivuljo. SVM sicer dosega primerljive, mestoma boljše rezultate pri večjih podmnožicah na učni množici podatkov, na testni pa natančnost pade nižje kot pri RF in GB, kar kaže na preveliko prilagajanje učnim podatkom (angl. *overfitting*).

Na drugi stopnji se SVM strmo povzpne in ohrani nad 90-odstotno CA tudi pri manjših izborih. Na učni množici se GB tudi ohrani visoko malo dlje pri zelo majhnih izborih, medtem ko na testni množici začne nižje in tam ostane, dokler se izbor atributov ne zmanjša na manj kot 10.

V primerjavi z metodo filtriranja na podmnožicah enakih velikosti vidimo, da daje metoda ovojnice boljše rezultate pri majhnih podmnožicah atributov. Na razredu ISCHEMIA vse metode s filtriranjem dosegajo malo višjo točnost, npr. pri 20 atributih RF doseže 96 odstotkov s filtriranjem, z ovojnico pa 94 odstotkov. Pri manjših izborih se situacija obrne, RF z ovojnico tukaj doseže 8 odstotkov več, SVM tudi, GB pa 6 odstotkov več. Na razredu ZONE RF in SVM pri vseh velikostih izborov z ovojnico dosežeta višjo točnost kot s filtriranjem, RF pri 40, 20 in 10 atributih celo za 11 odstotkov več. SVM dosega pri večjih izborih primerljivo točnost, pri 40 in 20 atributih 2 odstotka več, pri 10 atributih pa celo 9 odstotkov več. GB pa tudi z ovojnico na testni množici povečini ne presega rezultatov filtriranja, le pri 10 atributih doseže 6 odstotkov višjo točnost. Naša metoda ovojnice ima torej na drugi stopnji in pri iskanju majhnih podmnožic atributov prednost pred filtriranjem.

Zaporedja izločanja atributov po tem postopku se precej razlikujejo med metodami. Ravno tako ni videti povezave med vrstnim redom in ocenami atributov. "Node 949", ki je ocenjen najslabše, se izloča pozneje, kot bi pričakovali. Na drugi stopnji ga najhitreje izloči SVM, in sicer pri iskanju množice velikosti 69, sledita mu RF in GB. Na prvi stopnji, pri klasifikaciji

razreda ISCHEMIA, ga GB izloči pri iskanju množice velikosti 89, RF in SVM pa ga izločita veliko kasneje.

Atribut "Node 471", ki ga IG in MDL na prvi stopnji ocenjujeta najbolje, RF izloči hitro, GB pri iskanju podmnožice velikosti 54, SVM pa pri iskanju predzadnje. "Node 3921", po Reliefu najbolje ocenjen atribut, je odstranjen hitro pri GB, pri RF se ohrani do podmnožice velikosti 40, pri SVM pa do velikosti 28. "Node 16389" se ohrani dlje, pri RF do velikosti 10, pri GB pa do velikosti 16. SVM ga izloči pri iskanju podmnožice velikosti 77.

Na drugi stopnji se med najboljše atribute uvrščajo "Node 792" po IG in MDL, "Node 3548" po Reliefu in "Node 3133" po Giniju. Prvi se pri RF ohrani do podmnožice velikosti 12, pri GB do velikosti 39, pri SVM pa komaj do velikosti 73. "Node 3548" se pri vseh izloči kmalu, pri SVM že pri velikosti 84, pri GB na velikosti 79, le RF ga ohranja do velikosti 36. "Node 3133" po GB izpade iz množice velikosti 77, po RF iz množice velikosti 63, pri SVM pa iz velikosti 40.

Bistvenih razlik med podmnožicami izborov med metodami ni, ravno tako ni vidne povezave med našimi ocenami atributov in dejanskim izborom atributov po metodi ovojnice.

Na podlagi teh meritev smo za vsako metodo na obeh stopnjah izbrali "idealno"podmnožico atributov, ki daje na učni množici najboljši rezultat. Z njimi smo izmerili točnost metod, ki bi jo dobili pri avtomatskem izboru najboljše podmnožice. Izmerjen rezultat samodejnega izbora najboljše podmnožice atributov na učni podatkovni množici prikazuje tabela 4.2.

Po CA pri razredu ISCHEMIA vodi SVM s 94-odstotno točnostjo na testni množici pri 67 atributih, RF in GB pa zaostajata za 1 odstotek z 71 in 75 atributi. Pri razredu ZONE doseže na testni množici najvišjo točnost SVM s 95 atributi, najmanj atributov pa izbere RF, ki na testni množici z njimi doseže 84-odstotno CA. Treba je upoštevati, da so te množice izbrane na podlagi maksimalnih izmerjenih vrednosti znotraj naše ovojnice. Grafi na slikah 4.7, 4.8, 4.9 in 4.10 pokažejo, da obstajajo manjše podmnožice atributov, kjer se CA izbranim vrednostim v drugem stolpcu močno približa.

Metoda in razred	Velikost	CA na learn	CA na test
RF na ISCHEMIA	71	0.9853	0.9382
SVM na ISCHEMIA	67	0.9824	0.9441
GB na ISCHEMIA	75	0.9824	0.9353
RF na ZONE	76	0.9176	0.8412
SVM na ZONE	91	0.9941	0.9588
GB na ZONE	84	0.9824	0.9353

Tabela 4.2: Tabela CA, pridobljenih z "avtomatskim" izborom najboljše podmnožice atributov. Velikost pove, koliko atributov je v izboru, learn poda izmerjeno CA na učni množici, test pa izmerjeno CA na testni množici.

Poglavje 5

Sklepne ugotovitve

V diplomski nalogi smo preizkusili tri metode strojnega učenja za namen napovedovanja ishemije na podlagi simuliranih meritev EKG. Hoteli smo tudi zmanjšati izbor atributov v podatkovni množici. Za implementacijo in testiranje smo uporabili programsko okolje R, podatkovno množico pa smo pridobili s pomočjo računalniškega modela človeškega trupa [14].

Naša podatkovna množica je vsebovala dva problema – samo prisotnost ishemije in, v primeru da je ta pozitivna, območje v srčni mišici, kjer je prišlo do obolenja. Vsak primer vsebuje 100 atributov, ki smo jih ocenili po pomembnosti glede na končni razred. Različne metode ocenjevanja so atribute razvrstile različno, a filtriranje je pokazalo očitno prednost metode Relief. Podmnožice, izbrane na podlagi te ocene, pri naših metodah dosegajo podobno visoko klasifikacijsko točnost kot izbor vseh atributov. Kljub temu nas je zanimalo, ali obstajajo podmnožice, ki bi izboljšale ta rezultat, zato smo se odločili za uporabo metode ovojnice. Ustvarili smo ovojnico, ki išče dobre izbore atributov s pomočjo metod strojnega učenja. Z njo smo poganjali in preizkušali naključne gozdove, metodo podpornih vektorjev in metodo gradientnega pospeševanja.

Kot najboljša modela se izkažeta RF na prvi stopnji in SVM na drugi stopnji. GB se na drugi stopnji pri učenju kosa s SVM, a izgubi nekaj odstotkov CA na testni množici, kar kaže na preveliko prilagajanje učnim

podatkom (angl. *overfitting*). Na prvi stopnji se podobno zgodi s SVM pri majhnih izborih atributov. Sicer smo pri vseh izbranih metodah dosegli visoko klasifikacijsko točnost na obeh stopnjah, ki se bolj ali manj ohranja tudi pri manjših izborih atributov. Pri nekaterih se CA po umiku nekaterih atributov celo poviša. To potrjuje začetno domnevo, da je mnogo atributov nepomembnih ali odvečnih (redundantnih) za klasifikacijo ishemije in njenih območij in lahko metode strojnega učenja v ta namen izvajamo na veliko manjših izborih.

Zaporedje izločanja atributov se ne ujema z začetnimi ocenami atributov. Filtriranje pokaže, da lahko visok rezultat dosežemo z izločanjem na podlagi ocen, a ga metoda ovojnice preseže pri majhnih podmnožicah na prvi stopnji in skoraj povsod na drugi stopnji. Res pa je, da smo filtriranje za primerjavo izvajali z majhnim naborom velikosti podmnožic. Za bolj temeljito primerjavo metod bi morali ustvariti podmnožice enakih velikosti, kot jih ustvari ovojnica.

Metoda ovojnice daje dobre rezultate, je pa naša implementacija te metode zelo osnovna. Nove podmnožice požrešno išče z najboljšo podmnožico prejšnje velikosti kot izhodiščem, s čimer se sicer izogne veliki časovni porabi, a hkrati spregleda velik iskalni prostor atributskih podmnožic. Z implementacijo drugačnih iskalnih algoritmov bi mogoče lahko izboljšali CA modelov in našli primernejše podmnožice atributov. Pri prihodnjih izboljšavah in interpretiranju rezultatov te naloge je treba upoštevati tudi določeno mero naključnosti v procesu učenja metod in iskanju podmnožic.

Naša podatkovna množica je bila pridobljena znotraj računalniške simulacije. V praksi lahko sicer pričakujemo dobre rezultate, a da bi vedeli zagotovo, bi bilo potrebno testiranje metod na realnih primerih.

Dodatek A

Ocene atributov

A.1 Glede na razred ISCHEMIA

Tabela A.1: Tabela ocen pomembnosti atributov glede na razred ISCHEMIA. Atributi so razvrščeni po pomembnosti, kot jih je ocenil algoritem Relief.

Atribut	IG	MDL	Relief	Gini
Node 3921	0,234	0,231	0,254	0,148
Node 3818	0,236	0,233	0,247	0,148
Node 3588	0,229	0,226	0,244	0,146
Node 3658	0,230	0,228	0,237	0,144
Node 343	0,232	0,230	0,232	0,145
Node 379	0,204	0,201	0,228	0,134
Node 3078	0,230	0,227	0,221	0,146
Node 445	0,183	0,180	0,203	0,120
Node 457	0,180	0,177	0,197	0,118
Node 465	0,186	0,183	0,197	0,123
Node 3548	0,209	0,206	0,192	0,134
Node 4232	0,188	0,185	0,191	0,120
Node 463	0,244	0,241	0,182	0,150

Atribut	IG	MDL	Relief	Gini
Node 330	0,246	0,243	0,177	0,155
Node 488	0,246	0,243	0,172	0,150
Node 208	0,216	0,213	0,170	0,133
Node 207	0,216	0,213	0,170	0,133
Node 245	0,239	0,236	0,169	0,150
Node 219	0,235	0,232	0,168	0,148
Node 32	0,232	0,230	0,167	0,146
Node 501	0,177	0,174	0,166	0,112
Node 471	0,273	0,270	0,165	0,164
Node 651	0,201	0,198	0,162	0,125
Node 4844	0,199	0,196	0,158	0,122
Node 5707	0,191	0,188	0,156	0,122
Node 588	0,201	0,198	0,155	0,125
Node 792	0,209	0,206	0,155	0,134
Node 382	0,123	0,121	0,154	0,082
Node 268	0,224	0,221	0,153	0,145
Node 539	0,195	0,192	0,152	0,117
Node 572	0,208	0,205	0,151	0,124
Node 531	0,178	0,175	0,150	0,111
Node 17657	0,205	0,202	0,148	0,122
Node 541	0,206	0,203	0,148	0,123
Node 5004	0,212	0,209	0,146	0,127
Node 494	0,264	0,261	0,145	0,159
Node 5003	0,206	0,203	0,141	0,124
Node 3853	0,141	0,138	0,139	0,092
Node 595	0,187	0,184	0,138	0,118
Node 538	0,168	0,165	0,137	0,104
Node 769	0,185	0,182	0,136	0,117
Node 17653	0,188	0,185	0,135	0,115
Node 692	0,189	0,186	0,135	0,116

Atribut	IG	MDL	Relief	Gini
Node 3133	0,211	0,208	0,133	0,125
Node 16389	0,270	0,268	0,128	0,170
Node 585	0,177	0,174	0,128	0,110
Node 4552	0,118	0,115	0,127	0,075
Node 611	0,243	0,240	0,125	0,153
Node 577	0,172	0,169	0,124	0,111
Node 4239	0,141	0,138	0,123	0,089
Node 19281	0,227	0,224	0,123	0,143
Node 5652	0,176	0,173	0,123	0,109
Node 552	0,185	0,182	0,121	0,116
Node 780	0,223	0,221	0,121	0,142
Node 789	0,221	0,218	0,121	0,140
Node 759	0,209	0,206	0,120	0,133
Node 752	0,224	0,221	0,118	0,142
Node 727	0,231	0,228	0,114	0,145
Node 17236	0,252	0,249	0,114	0,158
Node 3252	0,185	0,182	0,113	0,111
Node 383	0,240	0,237	0,109	0,151
Node 496	0,138	0,136	0,101	0,090
Node 17268	0,248	0,246	0,100	0,156
Node 534	0,143	0,141	0,095	0,093
Node 4511	0,255	0,252	0,094	0,160
Node 459	0,132	0,129	0,092	0,082
Node 624	0,225	0,222	0,092	0,141
Node 5180	0,164	0,162	0,092	0,105
Node 590	0,229	0,226	0,090	0,144
Node 337	0,207	0,204	0,090	0,131
Node 336	0,205	0,202	0,086	0,129
Node 544	0,154	0,151	0,084	0,099
Node 381	0,217	0,214	0,083	0,141

Atribut	IG	MDL	Relief	Gini
Node 675	0,186	0,183	0,080	0,118
Node 690	0,211	0,208	0,080	0,133
Node 542	0,238	0,236	0,079	0,150
Node 3825	0,138	0,135	0,078	0,089
Node 441	0,237	0,234	0,077	0,149
Node 5538	0,175	0,172	0,075	0,111
Node 377	0,208	0,205	0,073	0,136
Node 4886	0,194	0,191	0,068	0,126
Node 390	0,206	0,203	0,068	0,135
Node 5280	0,167	0,164	0,067	0,106
Node 739	0,191	0,188	0,067	0,125
Node 656	0,213	0,210	0,067	0,138
Node 537	0,154	0,151	0,067	0,098
Node 5116	0,164	0,161	0,065	0,104
Node 5169	0,219	0,216	0,065	0,142
Node 401	0,212	0,209	0,064	0,135
Node 4264	0,207	0,204	0,063	0,135
Node 529	0,161	0,158	0,062	0,102
Node 17293	0,226	0,223	0,062	0,146
Node 436	0,216	0,213	0,060	0,140
Node 17215	0,219	0,216	0,060	0,142
Node 16798	0,220	0,217	0,059	0,143
Node 439	0,215	0,213	0,058	0,137
Node 499	0,221	0,218	0,057	0,144
Node 16985	0,223	0,221	0,056	0,145
Node 506	0,214	0,212	0,055	0,139
Node 949	0,002	0,001	0,031	0,001

A.2 Glede na razred ZONE

Tabela A.2: Tabela ocen pomembnosti atributov glede na razred ZONE. Atributi so razvrščeni po pomembnosti, kot jih je ocenil algoritem Relief.

Atribut	IG	MDL	Relief	Gini
Node 3548	0,577	0,515	0,188	0,044
Node 379	0,539	0,476	0,187	0,041
Node 538	0,523	0,463	0,182	0,042
Node 207	0,570	0,507	0,182	0,049
Node 488	0,485	0,425	0,180	0,041
Node 3133	0,550	0,486	0,172	0,059
Node 3078	0,609	0,541	0,167	0,045
Node 463	0,470	0,413	0,167	0,038
Node 552	0,567	0,494	0,159	0,046
Node 4232	0,519	0,455	0,157	0,040
Node 3853	0,513	0,451	0,155	0,058
Node 4552	0,511	0,444	0,148	0,058
Node 3588	0,512	0,449	0,148	0,038
Node 3252	0,499	0,431	0,147	0,059
Node 501	0,661	0,591	0,139	0,045
Node 457	0,490	0,429	0,137	0,035
Node 3658	0,492	0,437	0,135	0,035
Node 3818	0,446	0,393	0,135	0,031
Node 496	0,577	0,505	0,133	0,059
Node 531	0,653	0,579	0,132	0,049
Node 343	0,476	0,423	0,127	0,034
Node 32	0,557	0,484	0,124	0,055
Node 577	0,609	0,540	0,123	0,052
Node 4239	0,523	0,445	0,121	0,059
Node 459	0,505	0,442	0,120	0,059
Node 465	0,477	0,422	0,118	0,036

Atribut	IG	MDL	Relief	Gini
Node 534	0,543	0,475	0,118	0,059
Node 5172	0,601	0,530	0,116	0,051
Node 5180	0,532	0,464	0,114	0,059
Node 445	0,462	0,404	0,113	0,033
Node 3921	0,440	0,388	0,111	0,030
Node 219	0,517	0,454	0,110	0,058
Node 675	0,492	0,433	0,107	0,059
Node 245	0,512	0,449	0,101	0,058
Node 3825	0,457	0,401	0,101	0,059
Node 5003	0,515	0,455	0,101	0,038
Node 330	0,453	0,395	0,100	0,059
Node 5538	0,476	0,413	0,100	0,059
Node 471	0,440	0,387	0,099	0,056
Node 544	0,509	0,443	0,098	0,059
Node 268	0,516	0,453	0,097	0,059
Node 539	0,588	0,523	0,088	0,040
Node 494	0,453	0,398	0,087	0,057
Node 541	0,533	0,470	0,087	0,036
Node 611	0,512	0,451	0,087	0,055
Node 585	0,606	0,533	0,085	0,048
Node 19281	0,542	0,471	0,083	0,050
Node 5280	0,462	0,399	0,082	0,059
Node 5652	0,587	0,519	0,080	0,049
Node 336	0,461	0,403	0,080	0,059
Node 588	0,604	0,537	0,079	0,040
Node 789	0,585	0,514	0,079	0,051
Node 752	0,541	0,471	0,078	0,053
Node 377	0,455	0,401	0,078	0,059
Node 739	0,459	0,389	0,078	0,059
Node 537	0,482	0,416	0,077	0,059

Atribut	IG	MDL	Relief	Gini
Node 4844	0,596	0,533	0,077	0,040
Node 337	0,464	0,406	0,077	0,059
Node 759	0,585	0,513	0,076	0,054
Node 651	0,656	0,586	0,074	0,044
Node 390	0,444	0,390	0,074	0,059
Node 381	0,451	0,397	0,074	0,059
Node 769	0,629	0,554	0,074	0,047
Node 5116	0,459	0,394	0,074	0,059
Node 17657	0,573	0,511	0,073	0,038
Node 595	0,577	0,508	0,073	0,045
Node 529	0,455	0,393	0,073	0,059
Node 727	0,548	0,481	0,071	0,056
Node 4264	0,438	0,386	0,071	0,059
Node 17653	0,580	0,510	0,071	0,045
Node 5004	0,576	0,513	0,071	0,038
Node 572	0,571	0,508	0,069	0,040
Node 780	0,579	0,506	0,068	0,047
Node 690	0,468	0,408	0,067	0,059
Node 401	0,437	0,384	0,066	0,059
Node 382	0,335	0,287	0,066	0,027
Node 692	0,616	0,545	0,065	0,048
Node 436	0,434	0,382	0,065	0,059
Node 4886	0,441	0,377	0,065	0,059
Node 4511	0,405	0,354	0,064	0,059
Node 499	0,437	0,379	0,064	0,059
Node 439	0,439	0,387	0,064	0,059
Node 17293	0,432	0,374	0,063	0,059
Node 383	0,433	0,379	0,063	0,059
Node 624	0,505	0,442	0,063	0,059
Node 16798	0,437	0,384	0,062	0,059

Atribut	IG	MDL	Relief	Gini
Node 656	0,442	0,386	0,062	0,059
Node 17236	0,466	0,408	0,062	0,059
Node 590	0,483	0,421	0,061	0,059
Node 542	0,459	0,400	0,061	0,059
Node 5169	0,435	0,377	0,061	0,059
Node 16985	0,432	0,378	0,061	0,059
Node 17215	0,426	0,370	0,060	0,059
Node 441	0,429	0,377	0,060	0,059
Node 506	0,425	0,371	0,059	0,059
Node 16389	0,427	0,376	0,057	0,059
Node 5707	0,647	0,576	0,057	0,044
Node 17268	0,454	0,400	0,056	0,059
Node 792	0,696	0,616	0,042	0,048
Node 949	0,055	0,020	0,019	0,005

Dodatek B

CA, izmerjena s filtriranjem

B.1 Naključni gozdovi

Tabela B.1: Tabela CA, izmerjenih z metodo RF na podmnožicah atributov, izbranih s filtriranjem, na razredu ISCHEMIA. Stolpci predstavljajo različne mere ocen, vrstice pa velikosti izbranih podmnožic (100, 80 ... najbolje ocenjenih atributov).

Velikost	IG	MDL	Relief	Gini
100	0,962	0,959	0,962	0,956
80	0,965	0,959	0,959	0,962
60	0,950	0,953	0,968	0,959
40	0,938	0,938	0,974	0,938
20	0,938	0,941	0,968	0,938
10	0,832	0,835	0,856	0,821
1	0,794	0,794	0,768	0,806

Tabela B.2: Tabela CA, izmerjenih z metodo RF na podmnožicah atributov, izbranih s filtriranjem, na razredu ZONE. Stolpci predstavljajo različne mere ocen, vrstice pa velikosti izbranih podmnožic (100, 80 ... najbolje ocenjenih atributov).

Velikost	IG	MDL	Relief	Gini
100	0,806	0,818	0,835	0,829
80	0,824	0,800	0,835	0,800
60	0,788	0,818	0,824	0,706
40	0,753	0,753	0,794	0,624
20	0,653	0,653	0,747	0,594
10	0,553	0,547	0,665	0,494
1	0,188	0,176	0,182	0,218

B.2 SVM

Tabela B.3: Tabela Ca, izmerjenih z metodo SVM na podmnožicah atributov, izbranih s filtriranjem, na razredu ISCHEMIA. Stolpci predstavljajo različne mere ocen, vrstice pa velikosti izbranih podmnožic (100, 80 ... najbolje ocenjenih atributov).

Velikost	IG	MDL	Relief	Gini
100	0,962	0,962	0,962	0,962
80	0,950	0,950	0,959	0,950
60	0,938	0,938	0,956	0,944
40	0,915	0,915	0,953	0,915
20	0,912	0,912	0,938	0,915
10	0,865	0,865	0,829	0,856
1	0,794	0,794	0,762	0,806

Tabela B.4: Tabela Ca, izmerjenih z metodo SVM na podmnožicah atributov, izbranih s filtriranjem, na razredu ZONE. Stolpci predstavljajo različne mere ocen, vrstice pa velikosti izbranih podmnožic (100, 80 ... najbolje ocenjenih atributov).

Velikost	IG	MDL	Relief	Gini
100	0,924	0,924	0,924	0,924
80	0,912	0,906	0,918	0,900
60	0,900	0,906	0,918	0,782
40	0,841	0,859	0,918	0,729
20	0,747	0,735	0,882	0,700
10	0,653	0,682	0,806	0,547
1	0,200	0,200	0,171	0,218

B.3 Gradient Boosting

Tabela B.5: Tabela CA, izmerjenih z metodo GB na podmnožicah atributov, izbranih s filtriranjem, na razredu ISCHEMIA. Stolpci predstavljajo različne mere ocen, vrstice pa velikosti izbranih podmnožic (100, 80 ... najbolje ocenjenih atributov).

Velikost	IG	MDL	Relief	Gini
100	0,962	0,965	0,962	0,968
80	0,974	0,968	0,974	0,965
60	0,962	0,962	0,959	0,965
40	0,938	0,924	0,959	0,935
20	0,924	0,921	0,959	0,918
10	0,829	0,824	0,853	0,821
1	0,794	0,794	0,768	0,806

Tabela B.6: Tabela CA, izmerjenih z metodo GB na podmnožicah atributov, izbranih s filtriranjem, na razredu ZONE. Stolpci predstavljajo različne mere ocen, vrstice pa velikosti izbranih podmnožic (100, 80 ... najbolje ocenjenih atributov).

Velikost	IG	MDL	Relief	Gini
100	0,847	0,853	0,847	0,835
80	0,824	0,818	0,829	0,776
60	0,806	0,829	0,812	0,694
40	0,771	0,782	0,806	0,641
20	0,706	0,671	0,776	0,588
10	0,535	0,559	0,653	0,506
1	0,171	0,171	0,182	0,229

Dodatek C

Vrstni red izločanja atributov

Tabela C.1: Tabela vrstnih redov izločanja atributov znotraj metode ovojnice. Številka v prvem stolpcu pove, v katerem krogu metode je bil atribut izločen.

	RF		SVM		GB	
	ISCHEMIA	ZONE	ISCHEMIA	ZONE	ISCHEMIA	ZONE
1	Node 377	Node 541	Node 3252	Node 377	Node 752	Node 539
2	Node 330	Node 3818	Node 459	Node 336	Node 5169	Node 32
3	Node 541	Node 5707	Node 207	Node 537	Node 17653	Node 459
4	Node 4552	Node 651	Node 534	Node 692	Node 3548	Node 496
5	Node 337	Node 4552	Node 3658	Node 16985	Node 656	Node 17268
6	Node 769	Node 471	Node 4552	Node 219	Node 459	Node 3658
7	Node 5172	Node 379	Node 727	Node 436	Node 727	Node 739
8	Node 436	Node 377	Node 401	Node 752	Node 3818	Node 572
9	Node 5004	Node 506	Node 441	Node 17268	Node 789	Node 207
10	Node 692	Node 5280	Node 4511	Node 611	Node 439	Node 759
11	Node 5652	Node 17268	Node 537	Node 690	Node 949	Node 471
12	Node 572	Node 5538	Node 692	Node 3588	Node 268	Node 4844
13	Node 534	Node 4886	Node 17657	Node 488	Node 441	Node 651
14	Node 3548	Node 465	Node 3853	Node 383	Node 3921	Node 16985

	RF		SVM		GB	
	ISCHEMIA	ZONE	ISCHEMIA	ZONE	ISCHEMIA	ZONE
15	Node 4264	Node 577	Node 3548	Node 4232	Node 499	Node 585
16	Node 439	Node 539	Node 17236	Node 577	Node 245	Node 465
17	Node 471	Node 595	Node 445	Node 3548	Node 457	Node 494
18	Node 739	Node 383	Node 383	Node 769	Node 17236	Node 501
19	Node 16798	Node 436	Node 377	Node 471	Node 445	Node 541
20	Node 381	Node 330	Node 19281	Node 4886	Node 5003	Node 4264
21	Node 445	Node 16389	Node 595	Node 330	Node 219	Node 544
22	Node 3252	Node 381	Node 494	Node 245	Node 759	Node 3548
23	Node 611	Node 401	Node 16389	Node 531	Node 17215	Node 4232
24	Node 539	Node 499	Node 3133	Node 3921	Node 5538	Node 3133
25	Node 5003	Node 5003	Node 588	Node 268	Node 675	Node 439
26	Node 3818	Node 3078	Node 245	Node 445	Node 383	Node 330
27	Node 441	Node 780	Node 381	Node 759	Node 336	Node 537
28	Node 585	Node 382	Node 624	Node 792	Node 488	Node 656
29	Node 656	Node 542	Node 769	Node 19281	Node 552	Node 5180
30	Node 3133	Node 17653	Node 463	Node 441	Node 5280	Node 690
31	Node 496	Node 5004	Node 501	Node 949	Node 769	Node 552
32	Node 792	Node 590	Node 390	Node 5004	Node 739	Node 531
33	Node 465	Node 439	Node 379	Node 3078	Node 3078	Node 499
34	Node 463	Node 17236	Node 5004	Node 624	Node 3658	Node 17657
35	Node 17293	Node 5169	Node 541	Node 390	Node 585	Node 780
36	Node 4844	Node 16798	Node 3818	Node 4552	Node 494	Node 4552
37	Node 401	Node 588	Node 3588	Node 207	Node 501	Node 379
38	Node 494	Node 3133	Node 690	Node 499	Node 3853	Node 245
39	Node 506	Node 3252	Node 739	Node 739	Node 3133	Node 377
40	Node 3078	Node 752	Node 3078	Node 501	Node 534	Node 5538
41	Node 577	Node 17657	Node 611	Node 381	Node 207	Node 769
42	Node 5707	Node 789	Node 651	Node 4511	Node 5652	Node 19281
43	Node 5280	Node 544	Node 5169	Node 780	Node 542	Node 3252

	RF		SVM		GB	
	ISCHEMIA	ZONE	ISCHEMIA	ZONE	ISCHEMIA	ZONE
44	Node 245	Node 4844	Node 656	Node 459	Node 5180	Node 445
45	Node 457	Node 5180	Node 577	Node 534	Node 690	Node 611
46	Node 759	Node 501	Node 552	Node 457	Node 471	Node 5116
47	Node 5169	Node 390	Node 337	Node 5280	Node 17293	Node 390
48	Node 17657	Node 5652	Node 17293	Node 17653	Node 544	Node 506
49	Node 219	Node 494	Node 675	Node 5172	Node 4232	Node 336
50	Node 752	Node 537	Node 4232	Node 3818	Node 792	Node 457
51	Node 542	Node 4239	Node 5116	Node 542	Node 3252	Node 3588
52	Node 32	Node 585	Node 4886	Node 5652	Node 572	Node 4511
53	Node 5538	Node 245	Node 268	Node 496	Node 588	Node 3818
54	Node 3825	Node 949	Node 5707	Node 595	Node 32	Node 752
55	Node 379	Node 3921	Node 5172	Node 4844	Node 577	Node 789
56	Node 531	Node 3825	Node 5652	Node 16798	Node 337	Node 5172
57	Node 5180	Node 5116	Node 457	Node 3825	Node 17268	Node 949
58	Node 343	Node 656	Node 4844	Node 337	Node 590	Node 692
59	Node 4232	Node 445	Node 32	Node 585	Node 381	Node 595
60	Node 3658	Node 337	Node 5180	Node 401	Node 390	Node 529
61	Node 3921	Node 4511	Node 542	Node 3133	Node 401	Node 16389
62	Node 544	Node 463	Node 531	Node 789	Node 5116	Node 792
63	Node 789	Node 17215	Node 330	Node 544	Node 780	Node 5707
64	Node 3853	Node 488	Node 5280	Node 3853	Node 692	Node 219
65	Node 383	Node 3548	Node 336	Node 5003	Node 651	Node 675
66	Node 17215	Node 336	Node 17653	Node 5538	Node 16985	Node 534
67	Node 268	Node 5172	Node 436	Node 382	Node 537	Node 17236
68	Node 499	Node 675	Node 4264	Node 539	Node 343	Node 3853
69	Node 3588	Node 4264	Node 343	Node 5180	Node 529	Node 436
70	Node 17268	Node 19281	Node 496	Node 552	Node 506	Node 3078
71	Node 390	Node 219	Node 3825	Node 529	Node 4239	Node 624
72	Node 690	Node 739	Node 439	Node 17293	Node 5004	Node 17653

	RF		SVM		GB	
	ISCHEMIA	ZONE	ISCHEMIA	ZONE	ISCHEMIA	ZONE
73	Node 336	Node 727	Node 3921	Node 494	Node 539	Node 590
74	Node 459	Node 769	Node 792	Node 5116	Node 4886	Node 4239
75	Node 590	Node 3588	Node 465	Node 16389	Node 611	Node 759
76	Node 780	Node 534	Node 539	Node 541	Node 463	Node 16798
77	Node 4886	Node 690	Node 529	Node 465	Node 465	Node 588
78	Node 552	Node 3658	Node 17268	Node 675	Node 496	Node 337
79	Node 529	Node 17293	Node 219	Node 590	Node 17657	Node 441
80	Node 949	Node 268	Node 789	Node 727	Node 377	Node 577
81	Node 538	Node 457	Node 16798	Node 5169	Node 531	Node 381
82	Node 675	Node 624	Node 16985	Node 5707	Node 436	Node 3921
83	Node 727	Node 611	Node 5003	Node 572	Node 4552	Node 383
84	Node 4239	Node 531	Node 752	Node 588	Node 5172	Node 5169
85	Node 19281	Node 529	Node 538	Node 656	Node 16389	Node 488
86	Node 5116	Node 552	Node 949	Node 343	Node 5707	Node 268
87	Node 16985	Node 343	Node 544	Node 439	Node 595	Node 4886
88	Node 488	Node 496	Node 590	Node 4264	Node 538	Node 534
89	Node 4511	Node 792	Node 585	Node 17236	Node 19281	Node 5280
90	Node 595	Node 692	Node 759	Node 17657	Node 541	Node 17293
91	Node 16389	Node 441	Node 499	Node 4239	Node 3825	Node 441
92	Node 588	Node 4232	Node 4239	Node 506	Node 3588	Node 727
93	Node 17653	Node 207	Node 488	Node 651	Node 624	Node 5004
94	Node 651	Node 16985	Node 506	Node 3658	Node 4264	Node 538
95	Node 537	Node 459	Node 5538	Node 463	Node 379	Node 3825
96	Node 624	Node 3853	Node 17215	Node 379	Node 4511	Node 17215
97	Node 382	Node 538	Node 572	Node 538	Node 16798	Node 382
98	Node 501	Node 572	Node 471	Node 32	Node 382	Node 5003
99	Node 207	Node 32	Node 780	Node 3252	Node 4844	Node 343
100	Node 17236	Node 759	Node 382	Node 17215	Node 330	Node 463

Dodatek D

CA, izmerjena z metodo ovojnice

D.1 Naključni gozdovi

Tabela D.1: CA, izmerjene z metodo RF. Prvi stolpec določa velikost izbora atributov, ostali pa izmerjeno CA s tem izborom pri obeh razredih na učni in testni množici.

	ISCHEMIA, Train	ZONE, Train	ISCHEMIA, Test	ZONE, Test
1	0,755	0,260	0,765	0,165
2	0,835	0,335	0,853	0,306
3	0,871	0,512	0,897	0,418
4	0,921	0,606	0,921	0,565
5	0,921	0,706	0,926	0,712
6	0,929	0,759	0,924	0,682
7	0,938	0,729	0,929	0,724
8	0,944	0,712	0,926	0,724
9	0,950	0,776	0,932	0,794
10	0,956	0,788	0,935	0,776
11	0,932	0,765	0,935	0,800

	ISCHEMIA, Train	ZONE, Train	ISCHEMIA, Test	ZONE, Test
12	0,956	0,818	0,935	0,800
13	0,950	0,776	0,941	0,800
14	0,953	0,841	0,941	0,794
15	0,953	0,812	0,944	0,829
16	0,965	0,824	0,944	0,824
17	0,947	0,841	0,944	0,835
18	0,953	0,841	0,947	0,835
19	0,956	0,865	0,944	0,829
20	0,956	0,835	0,941	0,853
21	0,959	0,853	0,947	0,841
22	0,959	0,847	0,947	0,853
23	0,959	0,882	0,947	0,847
24	0,956	0,835	0,953	0,847
25	0,956	0,871	0,950	0,847
26	0,950	0,865	0,950	0,847
27	0,953	0,847	0,947	0,841
28	0,968	0,847	0,944	0,841
29	0,956	0,841	0,947	0,841
30	0,950	0,871	0,947	0,847
31	0,959	0,859	0,947	0,841
32	0,950	0,865	0,950	0,853
33	0,956	0,847	0,947	0,829
34	0,956	0,853	0,947	0,835
35	0,953	0,841	0,944	0,847
36	0,959	0,871	0,947	0,847
37	0,962	0,888	0,944	0,871
38	0,968	0,876	0,950	0,859
39	0,956	0,871	0,950	0,888
40	0,953	0,888	0,947	0,900
41	0,956	0,853	0,947	0,882

	ISCHEMIA, Train	ZONE, Train	ISCHEMIA, Test	ZONE, Test
42	0,959	0,876	0,947	0,882
43	0,959	0,888	0,950	0,888
44	0,962	0,882	0,950	0,882
45	0,959	0,871	0,950	0,882
46	0,959	0,888	0,947	0,871
47	0,959	0,871	0,947	0,882
48	0,962	0,876	0,944	0,882
49	0,965	0,882	0,947	0,882
50	0,968	0,888	0,953	0,871
51	0,962	0,888	0,953	0,882
52	0,962	0,859	0,953	0,876
53	0,965	0,888	0,950	0,888
54	0,965	0,865	0,950	0,871
55	0,962	0,900	0,950	0,882
56	0,962	0,900	0,950	0,876
57	0,971	0,888	0,947	0,871
58	0,965	0,882	0,950	0,876
59	0,971	0,882	0,947	0,894
60	0,962	0,876	0,947	0,894
61	0,959	0,871	0,947	0,882
62	0,974	0,894	0,950	0,859
63	0,968	0,900	0,944	0,865
64	0,965	0,876	0,947	0,888
65	0,971	0,888	0,947	0,882
66	0,965	0,906	0,944	0,894
67	0,968	0,876	0,950	0,882
68	0,968	0,906	0,944	0,876
69	0,971	0,894	0,941	0,894
70	0,976	0,888	0,950	0,876
71	0,985	0,906	0,947	0,871

	ISCHEMIA, Train	ZONE, Train	ISCHEMIA, Test	ZONE, Test
72	0,968	0,888	0,950	0,888
73	0,971	0,894	0,947	0,871
74	0,976	0,894	0,944	0,888
75	0,965	0,900	0,950	0,894
76	0,974	0,918	0,947	0,882
77	0,976	0,888	0,944	0,894
78	0,971	0,888	0,944	0,888
79	0,974	0,894	0,950	0,894
80	0,971	0,900	0,947	0,888
81	0,974	0,906	0,956	0,882
82	0,965	0,882	0,947	0,882
83	0,971	0,900	0,944	0,876
84	0,962	0,894	0,947	0,865
85	0,968	0,918	0,947	0,871
86	0,962	0,900	0,947	0,859
87	0,968	0,894	0,947	0,876
88	0,974	0,882	0,944	0,876
89	0,965	0,900	0,947	0,871
90	0,971	0,888	0,950	0,871
91	0,965	0,876	0,950	0,871
92	0,976	0,888	0,947	0,876
93	0,965	0,882	0,947	0,865
94	0,971	0,900	0,947	0,859
95	0,962	0,888	0,944	0,876
96	0,971	0,876	0,950	0,871
97	0,962	0,888	0,941	0,871
98	0,968	0,876	0,950	0,871
99	0,968	0,900	0,944	0,871
100	0,953	0,841	0,944	0,876

D.2 SVM

Tabela D.2: CA, izmerjene z metodo SVM. Prvi stolpec določa velikost izbora atributov, ostali pa izmerjeno CA s tem izborom pri obeh razredih na učni in testni množici.

	ISCHEMIA, Train	ZONE, Train	ISCHEMIA, Test	ZONE, Test
1	0,730	0,201	0,669	0,165
2	0,865	0,382	0,821	0,388
3	0,874	0,547	0,835	0,571
4	0,876	0,635	0,882	0,641
5	0,891	0,753	0,891	0,712
6	0,903	0,800	0,906	0,812
7	0,932	0,888	0,903	0,847
8	0,912	0,876	0,897	0,882
9	0,924	0,912	0,906	0,882
10	0,924	0,912	0,909	0,894
11	0,921	0,912	0,912	0,888
12	0,932	0,935	0,912	0,900
13	0,929	0,924	0,918	0,894
14	0,935	0,947	0,909	0,900
15	0,950	0,935	0,903	0,900
16	0,944	0,929	0,900	0,900
17	0,947	0,935	0,900	0,900
18	0,950	0,953	0,897	0,900
19	0,944	0,947	0,900	0,906
20	0,941	0,953	0,897	0,906
21	0,944	0,953	0,894	0,906
22	0,953	0,947	0,912	0,906
23	0,941	0,965	0,912	0,900
24	0,950	0,947	0,912	0,906
25	0,941	0,982	0,906	0,906

	ISCHEMIA, Train	ZONE, Train	ISCHEMIA, Test	ZONE, Test
26	0,962	0,953	0,918	0,906
27	0,950	0,953	0,918	0,912
28	0,956	0,959	0,929	0,912
29	0,956	0,953	0,929	0,912
30	0,959	0,971	0,929	0,906
31	0,956	0,953	0,929	0,929
32	0,968	0,953	0,926	0,929
33	0,965	0,965	0,926	0,935
34	0,959	0,965	0,926	0,929
35	0,965	0,971	0,924	0,935
36	0,965	0,965	0,926	0,935
37	0,959	0,976	0,926	0,929
38	0,953	0,988	0,929	0,929
39	0,953	0,959	0,926	0,929
40	0,956	0,976	0,926	0,935
41	0,959	0,988	0,926	0,929
42	0,953	0,976	0,929	0,929
43	0,959	0,976	0,932	0,924
44	0,959	0,976	0,932	0,924
45	0,956	0,971	0,932	0,924
46	0,962	0,976	0,935	0,924
47	0,962	0,976	0,926	0,924
48	0,956	0,965	0,926	0,924
49	0,959	0,976	0,926	0,924
50	0,956	0,971	0,926	0,924
51	0,968	0,971	0,929	0,924
52	0,971	0,971	0,926	0,912
53	0,956	0,982	0,929	0,912
54	0,962	0,965	0,929	0,912
55	0,962	0,971	0,935	0,918

	ISCHEMIA, Train	ZONE, Train	ISCHEMIA, Test	ZONE, Test
56	0,959	0,971	0,935	0,924
57	0,959	0,965	0,929	0,924
58	0,959	0,988	0,929	0,924
59	0,965	0,976	0,926	0,918
60	0,965	0,971	0,929	0,918
61	0,968	0,976	0,929	0,906
62	0,965	0,965	0,929	0,912
63	0,971	0,971	0,929	0,912
64	0,962	0,971	0,929	0,918
65	0,968	0,982	0,929	0,918
66	0,965	0,976	0,929	0,918
67	0,982	0,976	0,929	0,918
68	0,965	0,971	0,932	0,918
69	0,965	0,976	0,935	0,918
70	0,965	0,971	0,941	0,929
71	0,976	0,982	0,941	0,929
72	0,976	0,965	0,938	0,929
73	0,974	0,971	0,938	0,929
74	0,974	0,982	0,941	0,929
75	0,968	0,976	0,944	0,924
76	0,968	0,976	0,944	0,924
77	0,968	0,976	0,947	0,924
78	0,971	0,971	0,944	0,924
79	0,968	0,976	0,947	0,924
80	0,968	0,971	0,944	0,924
81	0,971	0,982	0,944	0,924
82	0,974	0,971	0,947	0,924
83	0,971	0,976	0,947	0,924
84	0,971	0,971	0,944	0,929
85	0,968	0,976	0,944	0,929

	ISCHEMIA, Train	ZONE, Train	ISCHEMIA, Test	ZONE, Test
86	0,968	0,988	0,947	0,929
87	0,968	0,982	0,947	0,929
88	0,971	0,976	0,944	0,924
89	0,982	0,976	0,947	0,929
90	0,971	0,976	0,947	0,935
91	0,976	0,994	0,947	0,924
92	0,976	0,982	0,944	0,929
93	0,971	0,982	0,944	0,929
94	0,974	0,971	0,944	0,929
95	0,974	0,982	0,944	0,929
96	0,968	0,976	0,947	0,929
97	0,974	0,971	0,944	0,935
98	0,968	0,976	0,944	0,935
99	0,971	0,971	0,947	0,935
100	0,959	0,959	0,950	0,929

D.3 Gradient Boosting

Tabela D.3: CA, izmerjene z metodo GB. Prvi stolpec določa velikost izbora atributov, ostali pa izmerjeno CA s tem izborom pri obeh razredih na učni in testni množici.

	ISCHEMIA, Train	ZONE, Train	ISCHEMIA, Test	ZONE, Test
1	0,787	0,201	0,753	0,153
2	0,850	0,879	0,850	0,288
3	0,865	0,906	0,835	0,376
4	0,897	0,929	0,841	0,488
5	0,909	0,921	0,891	0,565
6	0,941	0,944	0,874	0,624
7	0,929	0,929	0,894	0,718

	ISCHEMIA, Train	ZONE, Train	ISCHEMIA, Test	ZONE, Test
8	0,935	0,938	0,891	0,694
9	0,938	0,947	0,918	0,712
10	0,944	0,944	0,918	0,718
11	0,962	0,938	0,932	0,747
12	0,953	0,941	0,924	0,753
13	0,950	0,956	0,924	0,724
14	0,950	0,947	0,926	0,735
15	0,953	0,953	0,915	0,759
16	0,950	0,956	0,938	0,765
17	0,959	0,944	0,932	0,753
18	0,950	0,947	0,938	0,771
19	0,956	0,956	0,935	0,800
20	0,962	0,950	0,935	0,759
21	0,959	0,956	0,938	0,759
22	0,962	0,956	0,944	0,759
23	0,959	0,953	0,938	0,747
24	0,968	0,956	0,929	0,753
25	0,974	0,953	0,932	0,788
26	0,962	0,959	0,938	0,753
27	0,962	0,956	0,935	0,800
28	0,979	0,965	0,935	0,765
29	0,971	0,959	0,935	0,794
30	0,968	0,959	0,929	0,759
31	0,968	0,959	0,941	0,806
32	0,974	0,962	0,935	0,794
33	0,968	0,974	0,935	0,794
34	0,968	0,965	0,935	0,794
35	0,965	0,956	0,932	0,765
36	0,974	0,959	0,929	0,794
37	0,968	0,959	0,938	0,794

	ISCHEMIA, Train	ZONE, Train	ISCHEMIA, Test	ZONE, Test
38	0,965	0,965	0,921	0,812
39	0,965	0,962	0,935	0,782
40	0,971	0,962	0,929	0,800
41	0,971	0,968	0,929	0,776
42	0,979	0,965	0,935	0,788
43	0,965	0,965	0,924	0,806
44	0,965	0,965	0,935	0,794
45	0,965	0,968	0,929	0,824
46	0,971	0,974	0,938	0,800
47	0,976	0,965	0,938	0,788
48	0,968	0,974	0,941	0,788
49	0,971	0,968	0,938	0,776
50	0,968	0,974	0,929	0,800
51	0,968	0,971	0,932	0,788
52	0,968	0,974	0,935	0,818
53	0,976	0,965	0,935	0,776
54	0,976	0,968	0,944	0,812
55	0,971	0,974	0,941	0,800
56	0,974	0,971	0,947	0,794
57	0,971	0,971	0,935	0,806
58	0,979	0,976	0,935	0,800
59	0,971	0,974	0,929	0,812
60	0,971	0,974	0,938	0,800
61	0,974	0,974	0,929	0,812
62	0,965	0,971	0,938	0,806
63	0,965	0,976	0,947	0,806
64	0,976	0,971	0,938	0,800
65	0,968	0,965	0,935	0,812
66	0,971	0,974	0,941	0,806
67	0,965	0,976	0,947	0,818

	ISCHEMIA, Train	ZONE, Train	ISCHEMIA, Test	ZONE, Test
68	0,974	0,974	0,941	0,835
69	0,971	0,974	0,941	0,818
70	0,971	0,968	0,929	0,794
71	0,974	0,974	0,944	0,824
72	0,968	0,971	0,935	0,818
73	0,971	0,971	0,947	0,818
74	0,974	0,976	0,950	0,812
75	0,982	0,979	0,944	0,806
76	0,971	0,979	0,938	0,824
77	0,974	0,976	0,935	0,806
78	0,976	0,974	0,929	0,829
79	0,968	0,971	0,941	0,812
80	0,968	0,971	0,947	0,806
81	0,974	0,979	0,944	0,835
82	0,976	0,971	0,944	0,812
83	0,979	0,974	0,938	0,812
84	0,976	0,982	0,938	0,829
85	0,974	0,979	0,941	0,824
86	0,968	0,974	0,941	0,812
87	0,974	0,976	0,941	0,812
88	0,971	0,974	0,938	0,812
89	0,971	0,976	0,935	0,812
90	0,974	0,979	0,950	0,812
91	0,971	0,974	0,944	0,794
92	0,971	0,971	0,941	0,824
93	0,976	0,976	0,950	0,800
94	0,976	0,976	0,941	0,824
95	0,971	0,971	0,941	0,812
96	0,971	0,971	0,944	0,829
97	0,974	0,976	0,950	0,824

	ISCHEMIA, Train	ZONE, Train	ISCHEMIA, Test	ZONE, Test
98	0,979	0,982	0,944	0,818
99	0,974	0,974	0,941	0,806
100	0,947	0,835	0,935	0,841

Literatura

[1] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Manuel D. Cerqueira, Neil J. Weissman, Vasken Dilsizian, Alice K. Jacobs, Sanjiv Kaul, Warren K. Laskey, Dudley J. Pennell, John A. Rumberger, Thomas Ryan, and Mario S. Verani. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart. *Circulation*, 105(4):539–542, 2002.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995.

[5] Rahul C. Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015.

[6] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[7] Mervin J. Goldman and Nora Goldschalager. Principles of clinical electrocardiography. *Appleton Lange*, 1989.

[8] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, Mar 2002.

[9] Mareike Hullerum. The inverse problem of electrocardiography. 2012.

[10] Yuan Jiang, Cong Qian, Raghed Hanna, Dima Farina, and Olaf Dössel. Optimization of the electrode positions of multichannel ecg for the reconstruction of ischemic areas by solving the inverse electrocardiographic problem. *International Journal of Bioelectromagnetism*, 11(1):27–37, 2009.

[11] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273 – 324, 1997.

[12] I. Kononenko and M. Robnik Šikonja. *Inteligentni sistemi.* Založba FE in FRI, 2010.

[13] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*, pages 512–518, 2000.

[14] Smiljana Đorović, Marko Robnik Šikonja, Miloš Radović, Bojana Anđelković Ćirković, and Nenad Filipović. Finite element modelling of cardiac ischemia and data mining application for ischemic detection and localization. *Proceedings*, 2(8), 2018.

[15] Costas Papaloukas, Dimitrios I Fotiadis, Aristidis Likas, and Lampros K Michalis. Automated methods for ischemia detection in long duration ecgs. *Cardiovascular Reviews and Reports*, 24(6):313–319, 2003.

[16] Joint International Society and Federation of Cardiology/World Health Organization task force on standardization of clinical nomenclature. Nomenclature and criteria for diagnosis of ischemic heart disease. report of the joint international society and federation of cardiology/world health organization task force on standardization of clinical nomenclature. *Circulation*, 59(3):607–609, 1979.

[17] S. Stern, D. Tzivoni, and Z. Stern. Diagnostic accuracy of ambulatory ECG monitoring in ischemic heart disease. *Circulation*, 52(6):1045–1049, 1975.

[18] Dafang Wang, Robert M. Kirby, Rob S. MacLeod, and Chris R. Johnson. Inverse electrocardiographic source localization of ischemia: An optimization framework and finite element solution. *Journal of Computational Physics*, 250:403 – 424, 2013.

[19] Electrocardiography. dostopno na:. `https://en.wikipedia.org/wiki/Electrocardiography`. Dostopano: julij 2018.