UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO
FAKULTETA ZA MATEMATIKO IN FIZIKO

Aljaž Košmerlj

KONSTRUKCIJA KRIVULJ PREŽIVETJA IZ CENZURIRANIH PODATKOV Z METODAMI STROJNEGA UČENJA

Diplomska naloga
na univerzitetnem študiju

Mentor: akad. prof. dr. Ivan Bratko

Ljubljana, 2008

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Zahvala

Zahvalil bi se rad svojemu mentorju profesorju Bratku za vse nasvete ob izdelavi tega dela. Velika zahvala gre tudi Aleksandru Sadikovu in Juretu Žabkarju iz laboratorija za umetno inteligenco, ki sta mi oba veliko pomagala in mi prva predstavila področje analize preživetja. Onkološkemu inštitutu v Ljubljani se zahvaljujem za podatke, katerih uporabo so mi omogočili. Nenazadnje hvala družini in prijateljem za oporo in vse opomine, ki sem jih potreboval.

Kazalo

Seznam uporabljenih kratic in simbolov

ROC - sprejemnikova operativna značilnost (Reciever Operating Characteristic), mera za ocenjevanje uspešnosti klasifikatorja v strojnem učenju

AUC - ploščina pod ROC krivuljo (Area Under Curve)

SVM - metoda podpornih vektorjev (Support Vector Machine)

RBF - funkcija z radialno bazo (Radial Basis Function)

DSS - za bolezen značilno preživetje (Disease Specific Survival)

DFS - preživetje brez bolezni (Disease Free Survival)

Povzetek

V diplomi predstavim in obravnavam novo metodo strojnega učenja za ocenjevanje funkcije preživetja iz podatkov analize preživetja.

Na začetku predstavim samo področje analize preživetja in problematiko s katero se ukvarja. Uvedem osnovne pojme analize preživetja, kot sta funkcija preživetja in krivulja preživetja, ter jih definiram. Predstavim pojem cenzuriranih podatkov, posebnost naborov podatkov analize preživetja, in razložim njihovo pomembnost ter težave, ki jih predstavljajo za učenje. Opišem tudi Kaplan-Meierjev model, razširjeno statistično metodo za ocenjevanje krivulj preživetja, ki je sposobna delati na cenzuriranih podatkih in na kateri idejno sloni predlagana nova metoda. Uvod zaključim s krajšim pregledom dosedanjega napredka strojnega učenja v analizi preživetja. Ugotovim, da zaenkrat ni pomembnejših metod strojnega učenja na tem področju.

V nadaljevanju podrobno opišem predlagano metodo strojnega učenja ter njene potencialne prednosti in jo temeljito preizkusim. Prvo skupino testov opravim na umetno generiranih podatkih iz fizikalne domene. Ugotovim, da je predlagana metoda uporabna in se po natančnosti lahko kosa s Kaplan-Meierjevim modelom. Obravnavam še problem nemonotonih približkov krivulj, ki jih lahko dobimo s predlagano metodo. Vse teste ponovim na realnih medicinskih podatkih o napovedni vrednosti proteinskih markerjev za preživetje bolnikov z metastatskim rakom dojke. Ti testi še utemeljijo uporabnost nove metode. V zaključku predstavim možnosti izboljšav metode in naštejem nekaj novih možnosti uporabe tehnik strojnega učenja v analizi preživetja.

Ključne besede:

analiza preživetja, krivulja preživetja, cenzurirani podatki, podatki o času do dogodka, Kaplan-Meierjev model, strojno učenje

Abstract

In the present thesis I introduce and evaluate a new machine learning method for estimating survival functions from survival analysis data.

Firstly, I describe the field of survival analysis and the problems it deals with. I introduce and define the basic terms of survival analysis, like survival function and survival curve. I also define censored data, a speciallity of survival analysis data, and explain their importance and the learning problems they cause. As a reference method I describe the Kaplan-Meier estimator, a well-known statistical method for estimating survival curves, that serves as a conceptual basis for the new proposed method. I close the introduction with a short overview of the advances of machine learning in the field of survival analysis, concluding that so far there are no well established meachine learning methods in this field.

I continue with an in depth description of the proposed method and it's potential advantages. To test the new method thoroughly I start with a series of tests on artificially generated data from a physics domain. The new method proves itself useful and can match the accuracy of the Kaplan-Meier estimator. I discuss the problem of nonmonotonic survival curve estimations, that can be obtained using the proposed method. All the tests are repeated on a set of real medical data describing the prognostic value of protein markers for survival of metastatic breast cancer patients. The results further confirm the proposed method as useful. In conclusion I present the possibilities of improving the proposed method and suggest other prospects of using machine learning techniques in survival analysis.

Keywords:

survival analysis, survival curve, censored data, time to event data, Kaplan-Meier estimator, machine learning

Poglavje 1

Uvod

Na mnogih področjih, tako znanosti, kot industrije, pogosto opazujemo stanje neke skupine osebkov tekom časa. Zanimajo nas, na primer, napredek zdravljenja pri bolniku, delovna sposobnost neke naprave ali zadovoljstvo naših strank. Za taka opazovanja so pomembni dogodki, ki spremenijo stanje posameznega osebka. Ti so za prejšnje primere lahko recimo ponovitev bolezni pri bolniku, okvara naprave ali izguba stranke. Za podatke, ki jih zberemo s tovrstnimi opazovanji, se je uveljavil izraz podatki o času do dogodka (time to event data). Modeliranje tovrstnih podatkov je zelo koristno in se uporablja v mnogih panogah. Potrebno je v medicinskih raziskavah, kjer se preučuje vpliv zdravil na populacijo bolnikov ali razvija nove postopke za zdravljenje, katerih uspešnost je potrebno nadzorovati. Podobno je za učinkovito vzdrževanje strojev ali obratov potrebno poznati življenjsko dobo posameznih strojev in njihovih delov. Uporablja se tudi v sociioloških raziskavah, recimo za modeliranje časa, po katerem nekdanji kaznjenci spet zagrešijo zločin. Zaradi tega širokega spektra uporabe, so metode za delo s podatki o času do dogodka znane pod različnimi imeni. V inženirstvu je to področje znano kot teorija zanesljivosti (reliability theory), v družbenih vedah pa kot modeliranje trajanja (duration modelling), najbolj splošen izraz pa je gotovo analiza preživetja, kot se imenuje veja statistike, ki se ukvarja s takimi podatki. Pomembno je, da se zavedamo, da je izraz preživetje tu uporabljen splošno, in ne pomeni nujno biološkega preživetja. Isto velja za izrazoslovje uporabljeno v nadaljevanju te diplomske naloge.

Z analizo preživetja modeliramo preživetje v odvisnosti od neke zvezne monotone spremenljivke, ponavadi časa. Ob tem moramo upoštevati cenzurirane podatke, ki so posebnost takih naborov. Cenzuriran podatek je podatek o osebku, katerega opazovanje se je zaključilo, preden bi se mu zgodil dogodek.

Znano je torej le, do katere točke je osebek gotovo preživel, ne pa tudi, kdaj se mu je zgodil dogodek. Cenzurirani podatki se v takih naborih pojavijo naravno. Bolnik se lahko recimo odseli in ni več del naše študije, ali pa se študija zaključi, preden bi se pacientu ponovila bolezen. Podobno je, če kot podjetje zbiramo podatke o tem, kdaj nas zapustijo stranke. Točne podatke imamo le za pretekle stranke, ki so že odšle. Podatki o strankah, ki so nam trenutno še zveste, so cenzurirani, ker ne vemo kdaj v prihodnosti nas bodo te stranke zapustile. Delež cenzuriranih podatkov v naboru je lahko velik in dobra metoda ocenjevanja preživetja mora znati izkoristiti informacije, ki jih nudijo.

Kljub pogostosti in uporabnosti naborov podatkov o času do dogodka na področju strojnega učenja še ni bilo razvitih pomembnejših metod za delo z njimi. Namen tega diplomskega dela je razviti tako metodo in jo preizkusiti. V nadaljevanju najprej predstavim teoretične osnove analize preživetja in statistične metode za obravnavo naborov podatkov o času do dogodka. Sledi opis predlagane metode strojnega učenja in test njenega delovanja na umetnih podatkih, z obravnavo prednosti in težav metode. Nazadnje metodo preizkusim še na realnih medicinskih podatkih. Rezultati testov na obeh domenah pokažejo, da je predlagana metoda tako po točnosti napovedi, kot po napaki oblike približka krivulje, sposobna doseči in celo preseči Kaplan-Meierjev model, a je njena uspešnost zelo odvisna od klasifikatorja, ki ga uporabimo v njej. Testi internega vrednotenja metode pa so pokazali, da lahko z vrednostmi AUC klasifikatorjev uporabljenih v metodi ocenimo zanesljivost približka krivulje preživetja, ki ga vrne model.

Poglavje 2

Analiza preživetja

Za razumevanje tega diplomskega dela je potrebno poznavanje osnovnih pojmov analize preživetja. V tem poglavju predstavim te osnove ter opišem Kaplan-Meierjev model, uveljavljeno statistično metodo za obravnavo podatkov o času do dogodka. V zaključku poglavja opravim še krajši pregled dosedanjega napredka strojnega učenja v analizi preživetja.

2.1 Posebnosti naborov podatkov o času do dogodka

Nabori podatkov o času do dogodka poleg ostalih potrebujejo še dva atributa, časovnega in cenzorskega. Časovni atribut izraža dolžino preživetja. Čeprav je za to lahko uporabljena poljubna zvezna monotona spremenljivka oz. količina, bom v splošni obravnavi zanjo vedno uporabil čas. Cenzorski atribut, pa je binarna spremenljivka, ki pove, ali se je na koncu časa iz časovnega atributa zgodil dogodek ali pa je bil primer takrat cenzuriran.

Nabor praviloma vsebuje še druge atribute, katerih vrednosti so ponavadi določene ob začetku opazovanja. Tako navadno nimamo podatkov o stanju osebkov med začetkom opazovanja in dogodkom. Dogodek, ki ga opazujemo, je praviloma odvisen od teh atributov. To je lahko biološka smrt organizma, ponovitev bolezni, odpoved mehanskega dela in podobno.

Omeniti velja še, da so v analizi preživetja možni tudi problemi, ko nas ne zanima čas dogodka, temveč če se je ta sploh zgodil. Bolnike lahko recimo po določenem času po posegu smatramo kot ozdravele in nas zanima katerim se bo ponovila bolezen. Ta problem sicer zveni kot normalen klasifikacijski problem, vendar ga moramo zaradi cenzuriranih podatkov, če so seveda ti prisotni,

obravnavati kot problem analize preživetja. Spet drugačen tip problema je, če želimo modelirati, da se dogodek enemu osebku lahko zgodi več kot enkrat. Na primer, bolnik lahko večkrat zboli, oseba lahko večkrat preneha kaditi. V nadaljevanju se osredotočam na probleme kjer nas zanima čas preživetja do dogodka, ki se zgodi natanko enkrat.

2.2 Cenzura

V statistiki za neko spremenljivko velja, da je cenzurirana, če je njena vrednost le delno znana. Cenzura se lahko pojavi iz več razlogov. Podatek o preživetju pacienta je recimo cenzuriran, če se študija konča preden ta umre. Cenzurirana je tudi vrednost, ki je izven merilnega območja instrumenta, s katerim jo določamo. Poznamo tri tipe cenzure:

leva cenzura - cenzurirana vrednost je pod neko znano mejo, a ni znano koliko.

cenzura intervala - cenzurirana vrednost je nekje med dvema znanima mejnima vrednostma.

desna cenzura - cenzurirana vrednost je nad neko znano mejo, a ni znano koliko.

V analizi preživetja je najbolj tipična desna cenzura časovnega atributa. Tako pri cenzuriranem primeru ne moremo izvedeti točnega časa dogodka, vemo pa, da je zagotovo preživel do nekega časa.

Cenzure ne smemo mešati z rezanjem (truncation), ki pomeni, da nobena vrednost ne mora biti nad oz. pod nekim fiksnim pragom. Primer tega je recimo, če bi zbirali podatke o premijah, ki jih izplača zavarovalnica in ima ta omejeno največjo višino premije, ki jo še izplača. Podatki o dejanski škodi so tako rezani, ker je vsa škoda, večja od maksimalne premije, zmanjšana na vrednost maksimalne premije.

2.3 Funkcija in krivulja preživetja

Naj bo T naključna spremenljivka, ki pomeni čas dogodka. Funkcija preživetja spremenljivke T izraža verjetnost, da je vrednost T večja od neke določene vrednosti t. Povedano drugače, izraža verjetnost, da je preživetje dolgo vsaj t časa. Če ima T na intervalu $[0, \infty)$ porazdelitev $f(t)$ in kumulativno porazdelitveno funkcijo $F(t)$, se funkcija preživetja izraža po enačbi 2.1.

$$S(t) = P(T > t) = \int_t^\infty f(u)du = 1 - F(t) \tag{2.1}$$

Graf funkcije preživetja po času imenujemo krivulja preživetja. Ker je $F(t)$ monotono naraščajoča in velja $0 \leq F(t) \leq 1$, je $S(t)$ očitno monotono padajoča in nenegativna. $S(0)$ je navadno 1, vendar to ni nujno, če obstaja možnost, da se je dogodek ob začetku opazovanja že zgodil in ni celotna populacija živa. Tako je recimo v primeru, ko ocenjujemo življenjsko dobo nekih strojev, za katere je možno, da so že takoj ob izdelavi pokvarjeni.

Pričakovana življenjska doba je integral funkcije preživetja (enačba 2.2).

$$E[T] = \int_0^\infty S(u)du \tag{2.2}$$

2.4 Kaplan-Meierjev model

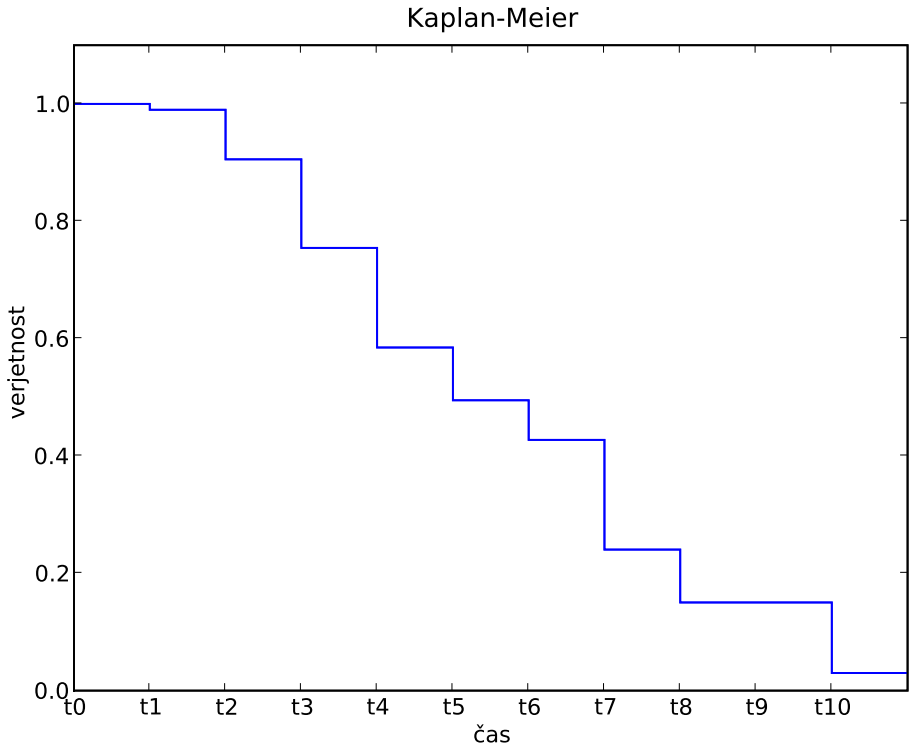
Kaplan-Meierjev model je neparametrična statistična metoda za ocenjevanje funkcije preživetja po času, ki sta jo Kaplan in Meier predstavila v svojem članku [6] leta 1958. Predlagana metoda idejno sloni na tem modelu, ki mi bo kasneje pri analizi rezultatov služil tudi kot referenčna metoda.

Imamo vzorec populacije velikosti N in z $S(t)$ označimo verjetnost, da osebek iz populacije živi vsaj do časa t. Čas diskretiziramo na časovne intervale z mejnimi časi $t_1 \leq t_2 \leq t_3 \leq \cdots \leq t_k$. Diskretizacija je poljubna, pogosta način je, da se za mejne vzame čase, ko se osebkom zgodijo dogodki. Za vsak t_i definiramo:
n_i - Število še živih osebkov tik pred časom t_i. V n_i niso šteti primeri, cenzurirani pred t_i.
d_i - Število dogodkov v času $[t_{i-1}, t_i]$ za $i > 1$ oz. $[0, t_i]$ za $i = 1$. Cenzure primera ne štejemo kot dogodek.
Oceno verjetnosti preživetja do časa t, $\hat{S}(t)$, izračunamo kot produkt:

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i} \tag{2.3}$$

Z dobljenimi ocenami lahko izrišemo približek krivulje preživetja, ki se navadno riše stopničaste oblike, ker privzamemo, da je verjetnost preživetja v diskretiziranem časovnem intervalu konstantna (Slika 2.1).

Kaplan-Meierjev model očitno ne upošteva nobenih atributov osebkov pri oceni verjetnosti preživetja. Če želimo z njim opazovati vpliv atributov na



Slika 2.1: Primer približka krivulje preživetja, dobljenega s Kaplan-Meierjevim modelom.

preživetje, moramo vzorec populacije razdeliti na podmnožice glede na vrednosti opazovanih atributov, ter med seboj primerjati krivulje preživetja, izrisane za vsako podmnožico posebej. Tako moramo, če želimo krivuljo preživetja za posamezen primer, upoštevati le primere, ki imajo popolnoma enake vrednosti atributov, kot primer, katerega krivuljo preživetja želimo. Zanesljivost dobljene krivulje pada s časom, ker pada tudi število še živih primerov, s katerimi ocenjujemo verjetnosti. Zato se krivuljo na neki točki navadno odreže, ker je preveč nezanesljiva. Ker Kaplan-Meierjev model nima nobene mere za zanesljivost, mora točko rezanja določiti domenski strokovnjak. To je pomankljivost Kaplan-Meierjevega modela. Bolje bi bilo, da bi metoda sama avtomatsko določila točko rezanja.

2.5 Strojno učenje in analiza preživetja

Metode strojnega učenja so bile v raziskavah analize preživetja zaenkrat malo uporabljene, med drugim zaradi pomanjkanja sposobnosti učenja iz cenzuriranih podatkov. Ponavadi so uporabljene posredno ali pa kot oblika predprocesiranja, recimo za razvrščanje v skupine, katerih preživetje se potem opazuje (na primer v [8]). Možen pristop je tudi, da se cenzurirane podatke priredi za uporabo v klasičnih metodah strojnega učenja. Najekstremnejši primer tega je, da iz nabora odstranimo vse cenzurirane primere, kar večinoma ni sprejemljivo. Bolje je, če skušamo preživetje ali verjetnost dogodka (če nas zanima ta, in ne čas preživetja) nekako določiti (imputirati). Tak pristop je opisan v [12]. Edini primer metode strojnega učenja, ki je razvita posebej za delo na cenzuriranih podatkih, ki sem ga uspel najti, je opisan v [7]. Gre za algoritem učenja regresijskih pravil, ki za hevristično oceno preživetja uporablja statistične metode. Tudi ta metoda tako ni popolnoma neodvisna od statističnih metod analize preživetja.

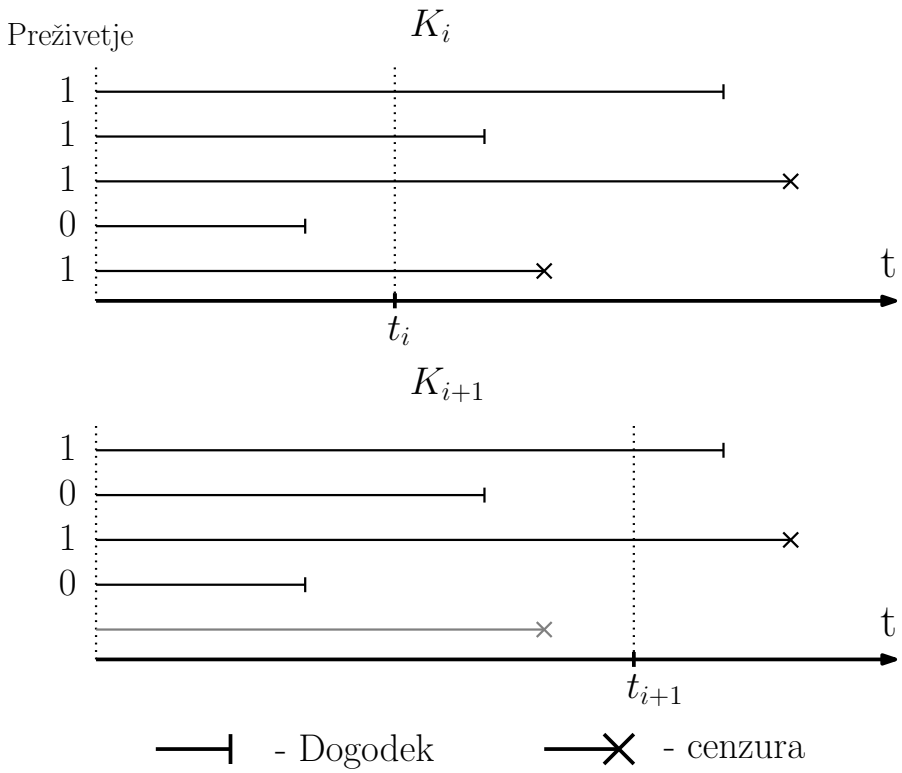
Poglavje 3

Opis predlagane metode

Osnovna ideja predlagane metode je zelo podobna Kaplan-Meierjevemu modelu. Čas diskretiziramo na intervale z mejnimi vrednostmi $t_1 \leq t_2 \leq t_3 \leq \cdots \leq t_k$ do katerih bomo ocenili verjetnost preživetja. Kot pri Kaplan-Meierjevem modelu, je tudi tu diskretizacija lahko poljubna. Za vsak čas t_i z izbrano metodo strojnega učenja zgradimo klasifikator K_i, s katerim ocenimo verjetnost preživetja do časa t_i. Klasifikator je lahko poljuben, edini pogoj za delovanje metode je, da mora biti sposoben napovedati verjetnost pripadnosti razredu in ne le vrednosti razreda. Kot bo kasneje razvidno iz testnih rezultatov, pa lahko izbira klasifikatorja zelo vpliva na točnost metode. Kateri klasifikator je najustreznejši, je odvisno od podatkov, ki jih imamo, tako da sta poznavanje domene in razumevanje metod strojnega učenja bistveni za dobro izbiro.

Za vsak klasifikator K_i posebej pripravimo učno množico D_i. Iz D_i odstranimo vse primere, cenzurirane pred časom t_i, ostalim pa določimo binaren razred "preživetje". Primerom, ki so preživeli čas t_i, določimo pozitivno vrednost razreda, primerom, ki se jim je zgodil dogodek pred časom t_i, pa negativno (glej sliko 3.1). Na tako tvorjeni učni množici D_i zgrajeni klasifikator K_i bo napovedoval verjetnost preživetja do časa t_i. Celotno zaporedje klasifikatorjev (slika 3.2) nam da zaporedne verjetnosti preživetja, ki jih lahko izrišemo v približek krivulje preživetja. Če uporabljeni klasifikator deluje na diskretnih vrednostih in ob nekem času ni več referenčnih primerov, to je, primerov, ki bi imeli vrednost kateregakoli atributa enako kot primer, ki ga klasificiramo, na tej točki nehamo ocenjevati nadaljne verjetnosti preživetja in krivuljo odrežemo. Edina izjema je, če ima primer, ki ga klasificiramo, neko vrednost nedefinirano. Takrat, glede na ta atribut, upoštevamo vse primere.

Čeprav je okvirno metoda podobna Kaplan-Meierjevemu modelu, se od

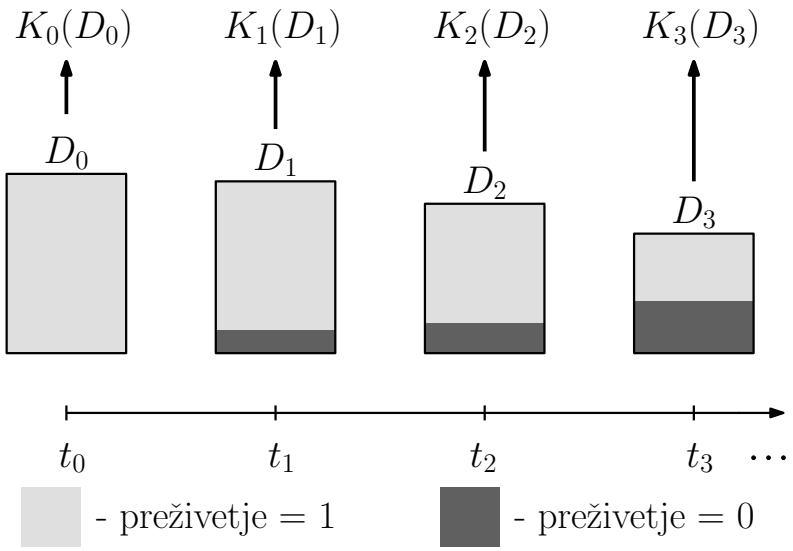


Slika 3.1: Shema izbora primerov in določitve razredov glede ne preživetje ob časih t_i in t_{i+1}.

njega razlikuje v nekaj pomembnih lastnostih. Najpomembnejše je to, da ta metoda upošteva atribute osebkov, če jih seveda upoštevajo uporabljeni klasifikatorji (kar je v strojnem učenju praviloma res). To pomeni, da za analizo vpliva atributov na preživetje ni potrebno deliti učne množice. Model lahko le enkrat zgradimo in ga potem uporabimo za klasifikacijo poljubnih primerov.

Prav tako je predlagani model, za razliko od Kaplan-Meierjevega, sposoben oceniti krivuljo preživetja za posamezen primer, ne le za skupino. Če pa želimo oceniti verjetnost preživetja neke skupine primerov, lahko pustimo nekatere vrednosti atributov nedefinirane in model bo, pod pogojem, da je uporabljeni klasifikator tega sposoben, vrnil povprečno verjetnost po vseh možnih vrednostih nedefiniranih atributov.

Prednost pa je tudi, da je v predlaganem modelu možno interno vrednotenje natančnosti krivulje. Za vsak čas t_i lahko izbrani klasifikator ovrednotimo tako, da na naboru podatkov D_i poženemo križno preverjanje ali kako drugo obliko



Slika 3.2: Shematičen prikaz konstrukcije klasifikatorjev po času.

internega vrednotenja. Izbira mere za kvaliteto klasifikatorja je zelo pomembna. Klasifikacijska natančnost in njej sorodne mere, ki ocenjujejo kalibracijsko točnost klasifikatorja so neustrezne. Njihova težava je v tem, da je s časom v učnih množicah vedno manj pozitivnih primerov in te mere pretirano nagrajujejo preproste modele, ki večino ali celo vse primere označijo za negativne. Veliko ustreznejše so mere, ki ocenjujejo diskriminacijsko točnost klasifikatorja, recimo ploščina pod ROC krivuljo (area under curve; AUC) [11]. Očitno je namreč, da klasifikator, ki slabo loči med pozitivnimi in negativnimi primeri, ne mora dobro oceniti verjetnosti za preživetje. Na podlagi AUC, lahko realiziramo avtomatično rezanje krivulj, ko postanejo preveč nezanesljive.

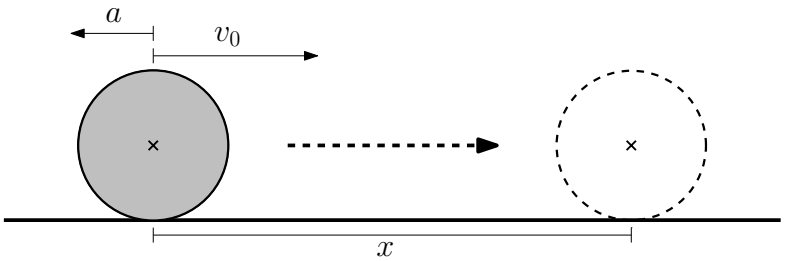
Poglavje 4

Preizkus na umetnih podatkih

Za analizo delovanja predlagane metode in kontrolirano primerjavo različnih metod med seboj sem najprej opravil vrsto testiranj na preprostih umetnih podatkih. Predlagano metodo in Kaplan-Meierjev model sem implementiral v programskem jeziku Python z uporabo knjižnic paketa Orange [3]. V tem poglavju sprva opišem testno domeno in metodo tvorjenja umetnih podatkov. Sledi še seznam in opis primerjanih metod ter razlaga metod testiranja in rezultati.

4.1 Opis testne domene

Za kontrolirano testiranje metode sem uporabil nabor umetnih fizikalnih podatkov. Šlo je za domeno podatkov o pojemajočem gibanju žogice (slika 4.1), ki ima stalen pojemek in jo zakotalimo z neko začetno hitrostjo. Namesto časa je bila opazovana količina razdalja, dogodek je bil ustavitev žogice.



Slika 4.1: Pojemajoče gibanje žogice.

Domeno so sestavljali naslednji atributi:

- v_0 - začetna hitrost, izbrana naključno, enakomerno z intervala $[10, 20]$

- a - pojemek, izbran naključno, enakomerno z intervala $[0.05, 0.2]$
- *survival* - preživetje, razredni atribut, katerega vrednost je bila določena odvisno od časa, kot opisano v opisu metode

Iz v_0 in a je končna razdalja do katere se prikotali žogica izračunljiva po formuli

$$x = \frac{v_0^2}{2a} \tag{4.1}$$

Vsak generiran primer sem z verjetnostjo 0.5 cenzuriral. To pomeni, da sem izračunano razdaljo x na naključno izbranem mestu odrezal in to mesto uporabil kot zadnjo znano razdaljo. V domeni so bili tako še naslednji meta atributi:

- x - po formuli izračunana končna razdalja, zaloga vrednosti je interval $[250, 4000]$
- x_k - modelu znana razdalja, enaka x in nato z verjetnostjo 0.5 skrajšana do naključnega mesta
- *cens* - binaren atribut, ki pove ali je primer cenzuriran (*cens*=1) ali ne (*cens*=0)

Vse teste sem opravil še na zašumljeni verziji podatkov, kjer sem po generiranju atributu v_0 dodal beli Gaussov šum s standardno deviacijo 1 in atributu a beli Gaussov šum s standardno deviacijo 0.015. Na obeh naborih podatkov, brezšumnem in zašumljenem, sem teste ponovil, tako da sem pred konstrukcijo modela iz učne množice odstranil cenzurirane podatke. Na ta način sem preveril, koliko prispevajo k natančnosti rezultata.

4.2 Primerjane metode

V vseh testih sem primerjal enak nabor metod za oceno krivulje preživetja, Kaplan-Meierjev model in predlagano metodo, realizirano s tremi različnimi klasifikatorji. Uporabil sem svojo implementacijo Kaplan-Meierjevega modela, katere edina posebnost je, da deluje na poljubno gosti ekvidistančni delitvi časa. Klasifikatorji, uporabljeni v predlagani metodi, so v paketu Orange implementirane verzije treh znanih klasifikatorjev:

Naivni Bayesov klasifikator - Klasifikator, ki oceni verjetnost pripadnosti razredu na osnovi naivne predpostavke o neodvisnosti atributov in Bayesove formule [4]. Uporabljena implementacija za primer $(a_1, a_2, \ldots, a_n)$ iz

domene z binarnim razredom C in atributi $A_1, A_2, \ldots, A_n$ izračuna verjetnost pripadnosti razredu $(C = 1)$ po enačbi 4.2.

$$P'(C = 1|A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n) = P(C = 1) \cdot \prod_{i=1}^{n} \frac{P(C = 1|A_i = a_i)}{P(C = 1)} \tag{4.2}$$

Ta oblika formule za Naivni Bayesov klasifikator je ekvivalentna bolj ustaljeni formuli s členi oblike $\frac{P(A_i=a_i|C=1)}{P(A_i=a_i)}$, če uporabimo relativne frekvence za ocene verjetnosti, ki nastopajo v njej.

Ker izpeljava zgornje ocene ni pravilna, je možno, da se verjetnosti pripadnosti posamezni vrednosti razreda ne seštejejo v 1. Zato je približek potrebno še normirati (enačba 4.3).

$$P(C = 1|A_1 = a_1, \ldots, A_n = a_n) =$$

$$= \frac{P'(C = 1|A_1 = a_1, \ldots, A_n = a_n)}{P'(C = 1|A_1 = a_1, \ldots, A_n = a_n) + P'(C = 0|A_1 = a_1, \ldots, A_n = a_n)} \tag{4.3}$$

Klasifikacijsko Drevo - Uporabil sem klasično, neobrezano klasifikacijsko drevo, zgrajeno tako, da minimizira entropijo. Postopek gradnje je, da se za podmnožico v posamezni veji drevesa izbere glede na informacijski prispevek najbolj informativen atribut in se podmnožico deli glede na njegove vrednosti. Ta postopek se ponavlja, dokler niso v vseh listih ali določene vrednosti vseh atributov, ali pa so vsi primeri iz lista iz istega razreda. Verjetnost pripadnosti razredu se določi glede na relativno frekvenco primerov v listu kamor se klasificira primer. Poglobljen opis klasifikacijskih dreves skupaj z eno najbolj znanih implementacij, ki pa ni tista, ki sem jo uporabil, lahko bralec najde v [9].

Metoda Podpornih Vektorjev (Support Vector Machine, SVM) - Metoda podpornih vektorjev primere z n atributi obravnava kot vektorje v n-dimenzionalnem prostoru in skuša z optimizacijsko metodo čim bolje ločiti vektorje, ki pripadajo različnim vrednostim razreda. Uporabil sem nelinearno verzijo SVM algoritma [1], ki originalni vektorski prostor pretvori z nelinearno jedrno funkcijo in v pretvorjenem prostoru loči primere s hiperravnino, ki jo potem preslika v nelinearno mejo v originalnem prostoru. Za jedrno funkcijo sem vzel funkcijo z radialno bazo (radial basis function, RBF), z enačbo 4.4, kjer je γ parameter jedra.

$$d(x, y) = \exp\left(-\gamma||x - y||^2\right) \tag{4.4}$$

Za Kaplan-Meierjev model in naivni Bayesov klasifikator sem atributa v_0 in a predhodno ekvidistančno diskretiziral na 10 intervalov.

Našteti klasifikatorji so bili izbrani, ker so med najbolj razširjenimi in delujejo na zelo različnih principih. Glavni cilj pri njihovem izboru niso bili nujno najboljši rezultati, temveč zastopanost različnih metod klasifikacije. Tako lahko opazujemo razlike v obnašanju predlagane metode, glede na uporabljen klasifikator. Od uporabljenega klasifikatorja je seveda odvisna časovna zahtevnost celotne metode. Ker so umetni podatki precej preprosti in ker sem hotel opraviti veliko testov, sem se odločil, da približek krivulje preživetja ocenim v dvajsetih ekvidistantnih intervalih časovnega atributa. Natančnost se je izkazala za dovolj dobro za primerjavo, hkrati pa časovna zahtevnost, predvsem na velikih naborih, ni bila pretirana.

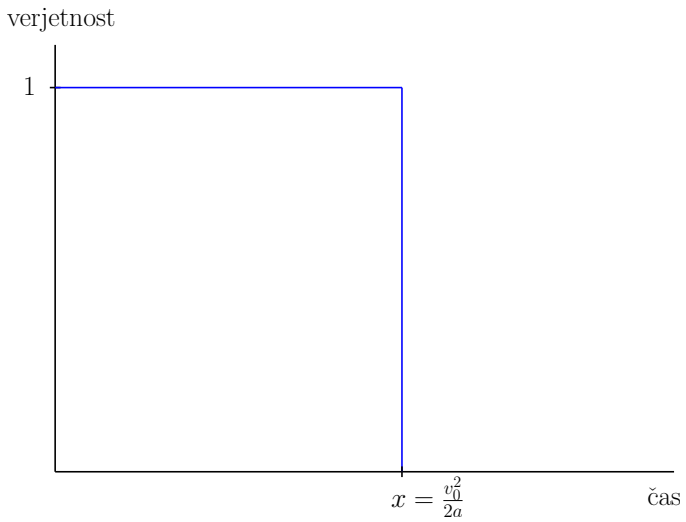
4.3 Oblika krivulje

Prva mera uspešnosti metode je, koliko se dobljena oblika krivulje prilega pravi krivulji preživetja. Ker so podatki umetni, poznamo obliko prave krivulje. Tak test na realnih problemih ni možen, ker oblike krivulje preživetja praktično nikoli ne poznamo.

Če bi izrisali pravo krivuljo preživetja za nek primer kotaljenja žoge, bi imela obliko stopnice. Verjetnost preživetja bi bila 1 do, po enačbi izračunane, razdalje x, potem pa bi padla na 0 (slika 4.2). Pomembno je, da ločimo med pravo in idealno krivuljo. Idealna krivulja se od prave razlikuje v tem, da upoštevamo, da modeli ocenjujejo verjetnosti v diskretiziranem času in se pravi krivulji lahko približajo le do idealne krivulje, glede na diskretizacijo, natančno.

Za vtis, kako se obnašajo posamezne metode sledijo krivulje preživetja za izbran primer z učnimi množicami različnih velikosti. Vse krivulje so bile izrisane za nešumne podatke, tako da sem najdaljšo razdaljo v učni množici razdelil na dvajset ekvidistantnih intervalov in v njihovih mejnih točkah z modelom ocenil verjetnost preživetja. Poleg napovedane krivulje je za primerjavo na vseh grafih s prekinjeno črto vrisana še prava krivulja preživetja.

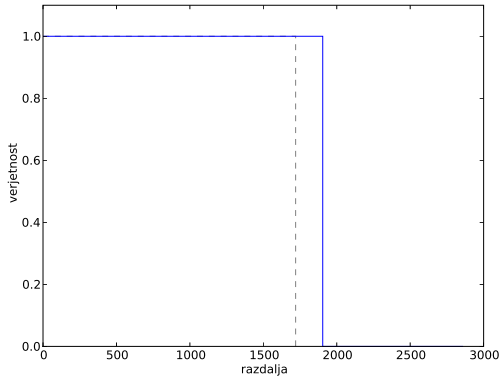
Krivulje, dobljene s Kaplan-Meierjevim modelom (Slika 4.3), po obliki lepo ustrezajo stopničasti obliki prave krivulje preživetja. Napaka, ki je najbolj razvidna na najnatančnejši krivulji (Slika 4.3(c)), je posledica predhodne diskretizacije atributov v_0 in a, ker ima izbrani primer enako diskretno vrednost posameznega atributa, kot primeri, ki živijo dlje ali manj od njega. To je sicer edini model, ki ne more vrniti nemonotonega približka krivulje preživetja. Že iz izpeljave modela je namreč očitno, da bo približek vedno monotono padajoč.



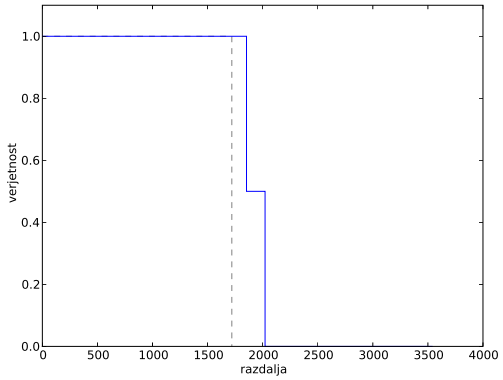
Slika 4.2: Prava krivulja preživetja.

Krivulje, ki jih dobimo s predlagano metodo realizirano z naivnim Bayesovim klasifikatorjem (Slika 4.4), po obliki precej bolj odstopajo od prave. Razlog za to je močna interakcija med atributoma primerov, kar pomeni, da naivna predpostavka, ki jo naredi naivni Bayesov klasifikator, ne drži. Model tako upošteva primere, ki imajo enako začetno hitrost, a veliko večji ali manjši pojemek, kar pomeni večje oz. manjše preživetje. Podobno velja tudi za primere z enakim pojemkom in drugačno začetno hitrostjo. Posledica te napake je, da modelov približek počasi pada, namesto da bi imel obliko stopnice. Iz slik 4.4(a) in 4.4(b) je razvidno, da lahko ta model vrne nemonotone približke krivulj preživetja. Nemonotomne krivulje bom posebej obravnaval v razdelku 4.6.

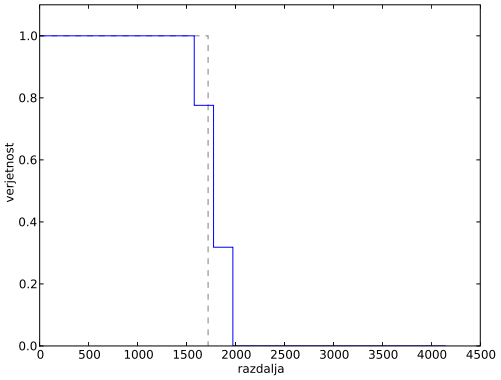
Predlagana metoda, realizirana s klasifikacijskimi drevesi, deluje na nediskretiziranih podatkih in tako krivulje (Slika 4.5) nimajo napak, ki bi bile posledica diskretizacije, kot je posebej razvidno iz krivulje, dobljene z največjo učno množico (Slika 4.5(c)). Če pa je učna množica majhna (Slika 4.5(a)), pride do izraza relativno slaba sposobnost klasifikacijskih dreves, da ocenijo verjetnost pripadnosti razredu. Tudi ta model lahko vrne nemonoton približek krivulje preživetja.



(a) Velikost učne množice 100 primerov

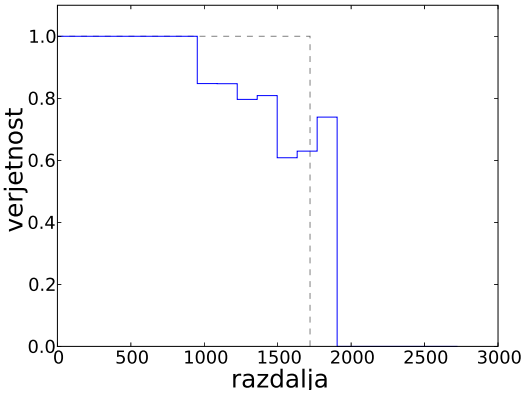


(b) Velikost učne množice 1000 primerov

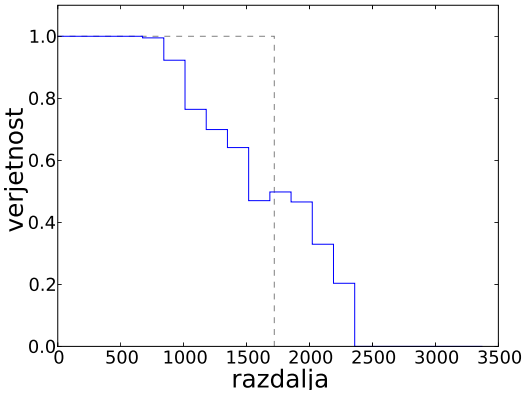


(c) Velikost učne množice 10000 primerov

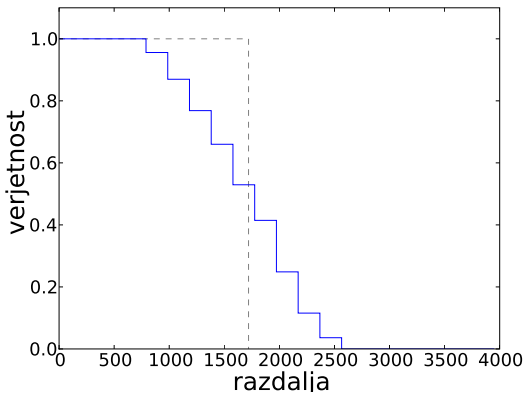
Slika 4.3: Krivulje preživetja primera z začetno hitrostjo $v_0 = 15.359$ in pojemkom $a = 0.069$ ter razdaljo oz. preživetjem $x = 1720.084$, zgrajene s Kaplan-Meierjevim modelom in učnimi nabori različnih velikosti. S prekinjeno črto je izrisana prava krivulja preživetja za ta primer.



(a) Velikost učne množice 100 primerov

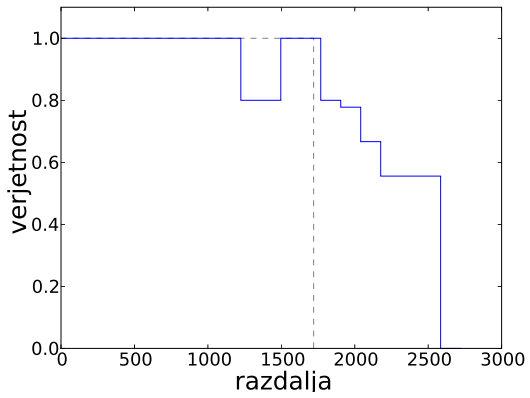


(b) Velikost učne množice 1000 primerov

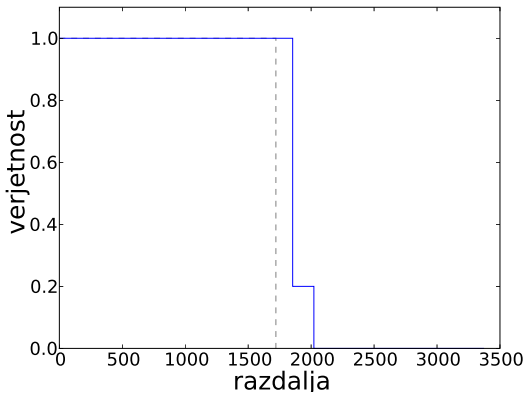


(c) Velikost učne množice 10000 primerov

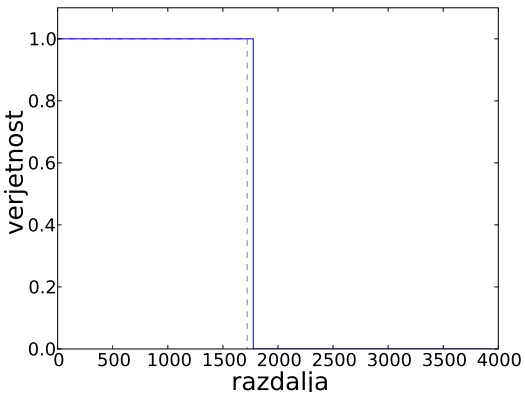
Slika 4.4: Krivulje preživetja primera z začetno hitrostjo $v_0 = 15.359$ in pojemkom $a = 0.069$ ter razdaljo oz. preživetjem $x = 1720.084$, zgrajene s predlagano metodo z naivnim Bayesovim klasifikatorjem in učnimi nabori različnih velikosti. S prekinjeno črto je izrisana prava krivulja preživetja za ta primer.



(a) Velikost učne množice 100 primerov

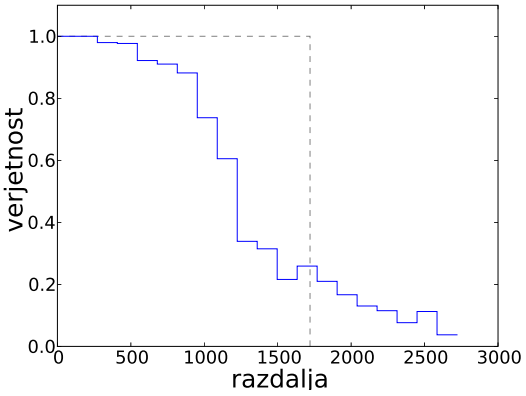


(b) Velikost učne množice 1000 primerov

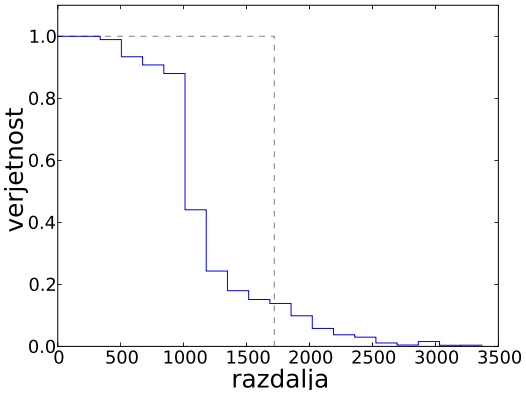


(c) Velikost učne množice 10000 primerov

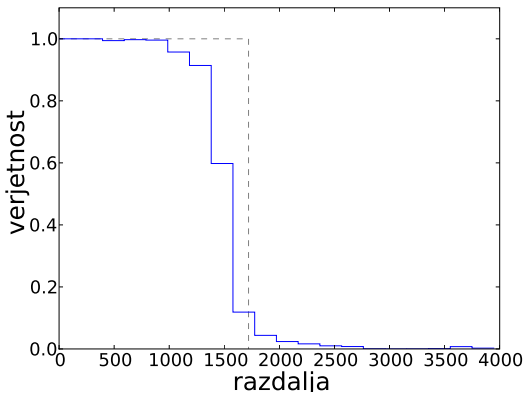
Slika 4.5: Krivulje preživetja primera z začetno hitrostjo $v_0 = 15.359$ in pojemkom $a = 0.069$ ter razdaljo oz. preživetjem $x = 1720.084$, zgrajene s predlagano metodo s klasifikacijskimi drevesi in učnimi nabori različnih velikosti. S prekinjeno črto je izrisana prava krivulja preživetja za ta primer.



(a) Velikost učne množice 100 primerov



(b) Velikost učne množice 1000 primerov



(c) Velikost učne množice 10000 primerov

Slika 4.6: Krivulje preživetja primera z začetno hitrostjo $v_0 = 15.359$ in pojemkom $a = 0.069$ ter razdaljo oz. preživetjem $x = 1720.084$, zgrajene s predlagano metodo s SVM in učnimi nabori različnih velikosti. S prekinjeno črto je izrisana prava krivulja preživetja za ta primer.

št. učnih primerov	KM		NB		KD		SVM	
	avg Δs	max Δs	avg Δs	max Δs	avg Δs	max Δs	avg Δs	max Δs
100	15.90	407.25	65.18	428.79	156.89	1196.63	209.31	460.55
200	10.91	329.51	65.17	355.57	58.64	867.24	230.39	518.66
500	7.49	175.42	59.49	331.89	39.61	1150.21	236.83	513.89
1000	16.98	376.79	61.42	549.49	17.81	293.48	180.03	431.64
2000	10.43	251.50	55.77	242.50	6.34	152.66	129.12	314.78
5000	10.41	168.76	49.58	141.18	3.54	254.80	81.84	304.81
10000	11.88	275.23	48.68	177.41	1.39	38.73	50.80	172.98

Tabela 4.1: Absolutne in maksimalne razlike površin za Kaplan-Meierjev model (KM) in predlagano metodo, realizirano z naivnim Bayesovim klasifikatorjem (NB), klasifikacijskim drevesom (KD) in SVM (SVM) na brezšumnih podatkih.

št. učnih primerov	KM		NB		KD		SVM	
	avg Δs	max Δs	avg Δs	max Δs	avg Δs	max Δs	avg Δs	max Δs
100	11.81	407.25	53.96	440.48	115.97	985.11	175.15	400.54
200	11.30	329.51	63.66	355.62	60.79	880.00	227.45	495.19
500	7.50	175.42	54.26	294.85	42.34	1121.15	209.61	415.10
1000	13.16	297.64	56.35	549.49	11.33	237.50	160.62	432.53
2000	8.25	231.79	54.07	239.90	7.97	152.66	111.85	299.94
5000	6.33	96.02	47.30	128.92	3.28	230.85	80.02	297.77
10000	9.26	231.58	46.62	168.35	1.04	2.40	49.91	164.27

Tabela 4.2: Absolutne in maksimalne razlike površin za Kaplan-Meierjev model (KM) in predlagano metodo, realizirano z naivnim Bayesovim klasifikatorjem (NB), klasifikacijskim drevesom (KD) in SVM (SVM) na brezšumnih podatkih brez cenzuriranih primerov

Tudi predlagana metoda, realizirana s SVM, deluje na nediskretiziranih podatkih. Ker SVM verjetnost pripadnosti razredu oceni glede na razdaljo primera do meje med razredoma, ima krivulja, ki jo vrne ta model, obliko dokaj zmehčane stopnice (4.6(c)). Ta model je zaradi optimizacijske narave SVM klasifikatorja bolj nagnjen k majhnim spremembam med verjetnostmi ob zaporednih časih. Iz istega razloga so tudi pri tem modelu možne nemonotone krivulje.

Za kvantitativno primerjavo oblik krivulj sem naštete metode naučil na učnih množicah velikosti 100, 200, 500, 1000, 2000, 5000 in 10000 ter vsakič pridobil krivulje preživetja 100 testnih primerov. Za vsak primer sem izračunal absolutno razliko ploščin med napovedano in idealno krivuljo, Δs. V tabelah 4.1, 4.2, 4.3 in 4.4 so navedena povprečna in maksimalna odstopanja površin krivulj, avg Δs in max Δs, za vse modele za vsak tip podatkov posebej.

št. učnih primerov	KM		NB		KD		SVM	
	avg Δs	max Δs	avg Δs	max Δs	avg Δs	max Δs	avg Δs	max Δs
100	57.48	1086.00	88.06	610.99	141.35	943.04	210.50	463.61
200	49.56	576.64	93.55	538.64	72.75	547.27	246.44	580.98
500	84.75	1225.63	99.77	443.71	68.33	654.63	242.71	612.79
1000	65.89	732.65	75.10	457.73	42.04	538.95	210.30	706.46
2000	87.43	557.22	103.14	508.84	97.19	1312.10	187.31	609.82
5000	71.66	621.45	86.88	488.53	40.08	547.27	99.91	429.42
10000	70.18	577.78	86.48	369.31	37.75	396.81	90.55	456.38

Tabela 4.3: Absolutne in maksimalne razlike površin za Kaplan-Meierjev model (KM) in predlagano metodo, realizirano z naivnim Bayesovim klasifikatorjem (NB), klasifikacijskim drevesom (KD) in SVM (SVM) na zašumljenih podatkih

št. učnih primerov	KM		NB		KD		SVM	
	avg Δs	max Δs	avg Δs	max Δs	avg Δs	max Δs	avg Δs	max Δs
100	56.22	1086.00	76.93	657.22	143.70	1005.57	167.87	380.46
200	46.97	494.26	85.87	459.33	74.06	704.36	238.55	536.51
500	77.98	1225.63	92.86	464.13	63.80	670.21	216.94	589.39
1000	55.82	732.65	71.64	452.41	40.55	532.16	186.08	430.13
2000	65.21	531.80	94.86	444.89	85.19	1273.00	166.63	624.15
5000	53.44	558.33	80.01	453.79	39.44	436.60	92.27	394.52
10000	49.70	530.08	80.03	353.12	38.05	299.26	87.34	488.63

Tabela 4.4: Absolutne in maksimalne razlike površin za Kaplan-Meierjev model (KM) in predlagano metodo, realizirano z naivnim Bayesovim klasifikatorjem (NB), klasifikacijskim drevesom (KD) in SVM (SVM) na zašumljenih podatkih brez cenzuriranih primerov

št. učnih primerov	KM		NB		KD		SVM	
	avg Δx	max Δx	avg Δx	max Δx	avg Δx	max Δx	avg Δx	max Δx
100	507.55	2191.98	275.90	1445.36	266.54	1174.70	433.33	1746.47
200	202.95	1505.19	198.62	750.14	178.19	902.90	445.53	1886.09
500	123.27	435.78	193.70	622.07	157.03	1112.54	464.49	1618.20
1000	146.54	612.58	193.82	749.20	116.34	345.43	338.45	1254.63
2000	127.28	473.77	181.30	722.09	100.80	344.44	330.42	1090.46
5000	138.74	405.80	164.87	404.01	100.99	375.00	234.95	767.09
10000	134.29	570.48	171.43	477.35	104.69	285.12	220.80	775.97

Tabela 4.5: Absolutne in maksimalne napake napovedi za Kaplan-Meierjev model (KM) in predlagano metodo, realizirano z naivnim Bayesovim klasifikatorjem (NB), klasifikacijskim drevesom (KD) in SVM (SVM) na brezšumnih podatkih.

4.4 Napovedna točnost

Vse modele sem testiral še kot napovedne modele. Za iste nabore učnih in testnih podatkov kot v razdelku 4.3 sem izračunal povprečno in maksimalno absolutno odstopanje napovedanega preživetja (razdalje kotaljenja) od pravega, avg Δx in max Δx. Napovedana razdalja je pričakovana vrednost razdalje glede na krivuljo, ki jo vrne model, se pravi ploščina pod njo. Podatki o napakah so zbrani v tabelah 4.5, 4.6, 4.7 in 4.8. Pri teh napakah je potrebno upoštevati, da metode nekaj napake naredijo že zaradi diskretizacije časovnega atributa (razdalje). Tako se lahko celo popoln model v povprečju naredi za pol intervala diskretizacije napake. Če upoštevamo, da je razpon možnih razdalj kotaljenja v podatkih brez šuma 3750, je torej zgornja meja za širino intervala $\frac{3750}{20} = 187.5$, kar pomeni v povprečju napako 93.75. Ker se za razpon pri diskretizaciji vzame interval med minimalno in maksimalno vrednostjo razdalje med primeri v učnem naboru, je pri manjših učnih naborih sicer ta napaka lahko nekaj manjša od zgornje meje, vendar tam ponavadi ne pride do izraza zaradi siceršnje napake modela.

št. učnih primerov	KM		NB		KD		SVM	
	avg Δx	max Δx	avg Δx	max Δx	avg Δx	max Δx	avg Δx	max Δx
100	658.49	2734.98	268.82	1445.36	206.94	1122.33	407.64	1781.88
200	233.15	1505.19	194.47	808.28	186.59	939.51	447.06	2026.38
500	143.37	2294.50	186.98	772.44	149.11	1316.50	439.44	1706.84
1000	128.63	612.58	194.00	749.20	114.26	325.58	314.50	1257.45
2000	114.36	684.45	179.15	767.07	103.96	420.15	314.76	1114.75
5000	117.46	323.43	163.24	451.34	103.50	351.05	231.69	753.10
10000	113.56	523.75	168.89	546.62	99.69	285.12	220.96	802.20

Tabela 4.6: Absolutne in maksimalne napake napovedi za Kaplan-Meierjev model (KM) in predlagano metodo, realizirano z naivnim Bayesovim klasifikatorjem (NB), klasifikacijskim drevesom (KD) in SVM (SVM) na brezšumnih podatkih brez cenzuriranih primerov.

št. učnih primerov	KM		NB		KD		SVM	
	avg Δx	max Δx	avg Δx	max Δx	avg Δx	max Δx	avg Δx	max Δx
100	644.74	2918.10	253.32	1105.99	252.93	1145.20	417.68	1751.41
200	359.04	2631.26	237.11	1424.38	228.89	1187.95	470.76	1793.28
500	309.03	1782.03	242.29	738.99	209.78	865.11	464.06	1805.58
1000	234.33	947.64	225.70	850.05	195.41	741.47	394.83	1392.14
2000	237.98	853.38	234.23	950.85	259.52	1351.42	383.76	1210.65
5000	221.30	975.07	228.90	1067.19	176.93	903.54	253.83	950.56
10000	241.40	1014.62	264.76	1275.10	221.55	817.16	286.79	956.73

Tabela 4.7: Absolutne in maksimalne napake napovedi za Kaplan-Meierjev model (KM) in predlagano metodo, realizirano z naivnim Bayesovim klasifikatorjem (NB), klasifikacijskim drevesom (KD) in SVM (SVM) na zašumljenih podatkih.

št. učnih primerov	KM		NB		KD		SVM	
	avg Δx	max Δx	avg Δx	max Δx	avg Δx	max Δx	avg Δx	max Δx
100	739.40	2918.10	304.89	2101.48	240.60	1078.84	400.16	1849.43
200	407.31	2631.26	227.31	1428.99	224.32	1195.80	461.64	1894.67
500	312.77	1782.03	237.86	851.39	199.66	877.55	436.84	1836.71
1000	226.57	947.64	216.90	842.23	189.49	673.29	374.22	1413.14
2000	219.37	877.47	225.99	1055.75	264.56	1491.96	370.25	1273.34
5000	189.55	996.25	218.17	1110.32	175.08	946.03	243.28	950.17
10000	220.40	1119.25	259.75	1350.88	226.83	841.16	283.15	1006.86

Tabela 4.8: Absolutne in maksimalne napake napovedi za Kaplan-Meierjev model (KM) in predlagano metodo, realizirano z naivnim Bayesovim klasifikatorjem (NB), klasifikacijskim drevesom (KD) in SVM (SVM) na zašumljenih podatkih brez cenzuriranih primerov.

4.5 Interno ovrednotenje

Modele, zgrajene s predlagano metodo, sem interno ovrednotil kot opisano v poglavju 3. Vsak klasifikator sem ovrednotil z petkratnim križnim preverjanjem in kot mero za uspešnost uporabil AUC. Tega sem izračunal kot povprečno vrednost AUC vseh delitev križnega preverjanja. Če so bili v nekem naboru le primeri iz enega razreda, sem vrednost AUC nastavil na 1. To sicer ni popolnoma v skladu z definicijo mere, a ustreza uspešnosti modela, saj vsak od uporabljenih klasifikatorjev v tem primeru vse primere klasificira pravilno.

Vrednost mere AUC je enaka verjetnosti, da klasifikator naključnemu primeru iz razreda napove večjo verjetnost, da je v razredu, kot naključnemu primeru izven razreda, zato je očitno, da je visoka vrednost mere AUC pogoj za zanesljivo delovanje modela na vseh primerih. Rezultati internega vrednotenja za posamezen tip podatkov in klasifikator so v dodatku A.

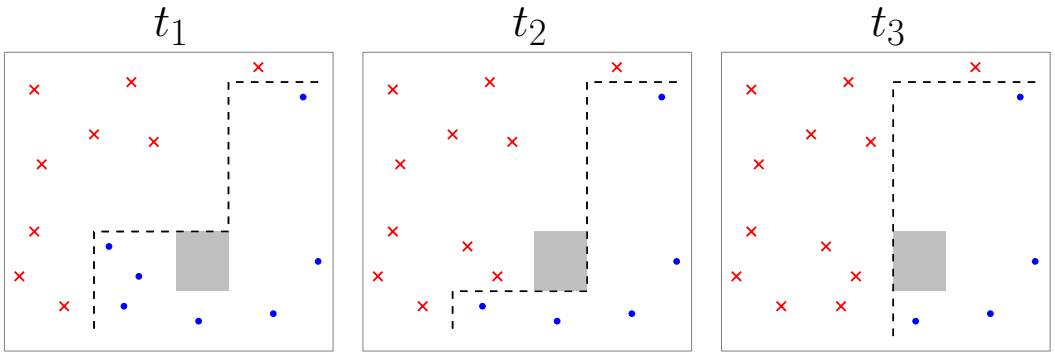
Iz vrednosti AUC se jasno vidi, da zanesljivosti klasifikatorjev naraščajo s številom primerov v učni množici, kar se ujema z napovedno točnostjo modelov in manjšo razliko v obliki krivulj. Šum po pričakovanjih zmanjša zanesljivost modelov, a ne spremeni trenda AUC, tako da uporabnost mere ostaja. Zelo zanimiva je tudi opazna razlika v zanesljivosti modelov zgrajenih s SVM, ki se po napakah oblike in napovedi najslabše obnese. To nakazuje možnost, da lahko ta način vrednotenja uporabimo kot pomoč pri izbiri najustreznejšega klasifikatorja.

4.6 Nemonotonost krivulj

Kot so pokazali testi, je mogoče s predlagano metodo, ne glede na uporabljen klasifikator, dobiti približek krivulje preživetja, ki ni monotono padajoč, kljub temu, da se s časom število pozitivnih primerov zmanjšuje ali v najboljšem primeru ostane enako, število negativnih primerov pa narašča ali ostane enako. Med v času zaporednima naboroma podatkov sta možni dve spremembi posameznega primera. Prva je, da je prej pozitiven primer sedaj postal negativen, druga pa, da je prej pozitiven primer bil vmes cenzuriran in zato odstranjen. Meja med pozitivno in negativno vrednostjo razreda se tako s časom premika glede na preživetje primerov. Vzroki za nemonotonost so pri vsakem klasifikatorju drugačni, vendar so pri vseh posledica tega, kako klasifikator modelira to mejo. Izkaže se celo, da pri nobenem klasifikatorju to obnašanje ni odvisno od cezuriranih primerov in se lahko pojavi tudi v naborih brez njih.

Nemonotonosti pri klasifikacijskih drevesih je morda še najlažje razumeti. Drevesa skonstruirajo približek meje med pozitivnimi in negativnimi primeri

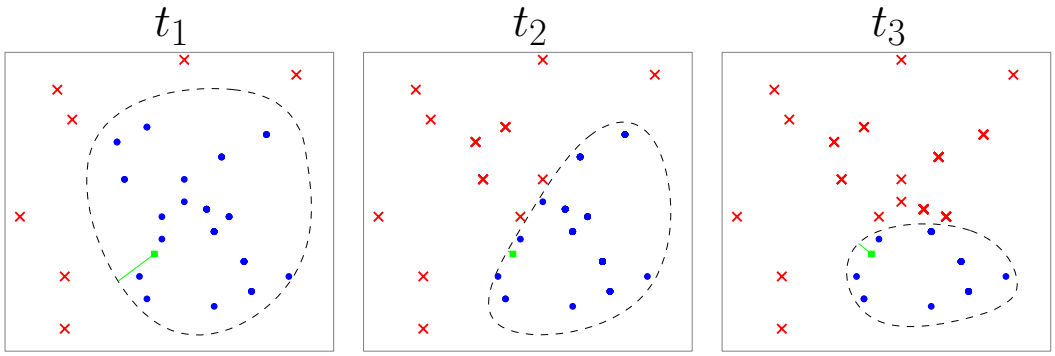
tako, da prostor primerov razdelijo z ortagonalnimi hiperravninami. Ker hkrati stremijo k najpreprostejšemu modelu, je možno, da so zaporedni približki meje taki, da se neko področje prestavlja z ene strani meje na drugo. Tako obnašanje ni odvisno od cenzuriranih podatkov, možno je tudi samo s prehajanjem primerov iz pozitivne v negativno vrednost razreda. Skica primera takega obnašanja je na sliki 4.7.



Slika 4.7: Shema približka meje med pozitivnimi in negativnimi primeri, ki ga naredi klasifikacijsko drevo, ob treh zaporednih časih t_1, t_2, t_3. Če s temi tremi klasifikatorji določamo verjetnost pripadnosti razredu primeru, ki leži kjerkoli v osenčenem področju, verjetnosti ne bodo padajoče po času.

SVM z RBF kot bazno funkcijo primere loči tako, da najprej z RBF preslika prostor primerov in jih potem v preslikanem prostoru z optimizacijsko metodo skuša čim bolje ločiti ter potem mejo preslika nazaj v originalni prostor. Rezultat je nelinearna sklenjena meja v originalnem prostoru. Ker se verjetnost določi glede na razdaljo od te meje, ki je določena z optimizacijo in je precej občutljiva na spremembe v podatkih, obstaja možnost nemonotonosti v zaporednih klasifikatorjih. Na sliki 4.8 je primer zaporednih naborov necenzuriranih podatkov z vrisano mejo, ki jo skonstruira SVM, pri katerih pride do nemonotonosti.

Pri naivnem Bayesovem klasifikatorju pa možnost nemonotonosti sledi že iz formule modela. Recimo, da računamo verjetnosti pripadnosti razredu za primer X z vrednostmi atributov $(A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n)$ ob dveh zaporednih časih t in t'. V učnem naboru označimo z N število vseh primerov in z N_{a_i} število primerov, ki imajo vrednost atributa A_i enako a_i, s C število primerov, ki imajo pozitivno vrednost razreda, s $\overline{C}$ število primerov, ki imajo negativno vrednost razreda, C_{a_i} in $\overline{C_{a_i}}$ naj označujeta primere, ki imajo vrednost atributa A_i enako a_i in imajo vrednost razreda pozitivno oz. negativno. Oznake $N', N'_{a_i}, C', \overline{C'}, C'_{a_i}, \overline{C'_{a_i}}$ naj označujejo iste vrednosti za čas t'. Naivni



Slika 4.8: Shema približka meje med pozitivnimi in negativnimi primeri, ki ga naredi SVM, ob treh zaporednih časih t_1, t_2, t_3. Če s temi tremi klasifikatorji določamo verjetnost pripadnosti razredu primeru, ki je označen z zelenim kvadratom.

Bayesov klasifikator z relativnimi frekvencami oceni verjetnost, da ima X pozitivno vrednost razreda po formuli 4.5 in po podobni izpeljavi verjetnost, da ima primer negativno vrednost razreda, po formuli 4.6.

$$P = P(C = 1|A_1 = a_1, \ldots, A_n = a_n) = P(C = 1) \cdot \prod_{i=1}^{n} \frac{P(C = 1|A_i = a_i)}{P(C = 1)} =$$

$$= \frac{C}{N} \frac{\frac{C_{a_1}}{N_{a_1}}}{\frac{C}{N}} \cdots \frac{\frac{C_{a_n}}{N_{a_n}}}{\frac{C}{N}} = \frac{\prod_{i=1}^{n} C_{a_i}}{\prod_{i=1}^{n} N_{a_i}} \cdot \frac{N^{n-1}}{C^{n-1}} \tag{4.5}$$

$$\overline{P} = P(C = 0|A_1 = a_1, \ldots, A_n = a_n) = \frac{\prod_{i=1}^{n} \overline{C_{a_i}}}{\prod_{i=1}^{n} N_{a_i}} \cdot \frac{N^{n-1}}{\overline{C^{n-1}}} \tag{4.6}$$

Sedaj lahko izpeljemo pogoj za to, da je verjetnost pripadnosti razredu ob času t' večja kot ob času t.

$$\frac{P}{P + \overline{P}} < \frac{P'}{P' + \overline{P'}}$$

$$\cancel{P P'} + P\overline{P'} < \cancel{P P'} + \overline{P} P'$$

$$\frac{\prod_{i=1}^{n} C_{a_i}}{\cancel{\prod_{i=1}^{n} N_{a_i}}} \cdot \frac{\cancel{N^{n-1}}}{C^{n-1}} \cdot \frac{\prod_{i=1}^{n} \overline{C'_{a_i}}}{\cancel{\prod_{i=1}^{n} N'_{a_i}}} \cdot \frac{\cancel{N'^{n-1}}}{\overline{C'^{n-1}}} < \frac{\prod_{i=1}^{n} \overline{C_{a_i}}}{\cancel{\prod_{i=1}^{n} N_{a_i}}} \cdot \frac{\cancel{N^{n-1}}}{\overline{C^{n-1}}} \cdot \frac{\prod_{i=1}^{n} C'_{a_i}}{\cancel{\prod_{i=1}^{n} N'_{a_i}}} \cdot \frac{\cancel{N'^{n-1}}}{C'^{n-1}}$$

$$\frac{\prod_{i=1}^{n} \frac{C_{a_i}}{\overline{C_{a_i}}}}{\left(\frac{C}{\overline{C}}\right)^{n-1}} < \frac{\prod_{i=1}^{n} \frac{C'_{a_i}}{\overline{C'_{a_i}}}}{\left(\frac{C'}{\overline{C'}}\right)^{n-1}} \tag{4.7}$$

V neenačbi 4.7 upoštevamo še neenakosti, ki sledijo iz tega, da se število pozitivnih primerov manjša ali ostane enako zaradi cenzure ali dogodkov, število negativnih primerov pa narašča ali ostane enako. Velja torej še:

$$C \leq C', \overline{C} \geq \overline{C'} \text{ in } C_{a_i} \leq C'_{a_i}, \overline{C_{a_i}} \geq \overline{C'_{a_i}};\ \forall i \in \{1, 2, \ldots, n\}$$

Lahko je videti, da situacija, ko med časoma t in t' noben od primerov, ki se z X ujema v vrednosti kateregakoli atributa, ne doživi dogodka in ni cenzuriran, sicer pa nekaj primerov doživi dogodek, ustreza enačbi 4.7 in vsem zgornjim pogojem. Ta primer tudi pokaže, da je nemonotonost predlagane metode, realizirane z negativnim Bayesovim klasifikatorjem, neodvisna od cenzuriranih primerov.

Poglavje 5

Preizkus na realnih podatkih

Delovanje predlagane metode sem testiral na naboru realnih podatkov o času do dogodka, ki so že bili predmet študije analize preživetja. Tako je bila mogoča primerjava mojih rezultatov z izsledki prejšnje raziskave na teh podatkih. V tem poglavju opišem domeno realnih podatkov in navedem rezultate testov metod, skupaj z njihovo primerjavo s predhodnimi rezultati.

5.1 Opis podatkov

Realni podatki, na katerih sem testiral metode, so iz onkološke raziskave prognostične vrednosti markerjev Stathmin-1, S100A2 in SYK pri ER-pozitivnih bolnikih s primarnim rakom dojke zdravljenih z adjuvantno monoterapijo s Tamoxifenom. Raziskava je skupaj z medicinskim ozadjem podrobneje opisana v [5], od koder je tudi povzet nadaljni opis domene.

V raziskavi je bilo udeleženih 215 bolnic obolelih za rakom dojke, ki so bile med letoma 1994 in 1999 zdravljene na Onkološkem Inštitutu v Ljubljani. Vse bolnice so imele pozitivne estrogenske receptorje (bile so ER-pozitivne) in so bile zdravljene izključno z adjuvantno hormonalno terapijo s Tamoxifenom. Nobena od njih ni bila zdravljena z adjuvantno ali neadjuvantno kemoterapijo. Srednja starost bolnic ob času diagnoze je bila 67.9 let (razpon starosti je bil 36-83 let), 207 (96.3%) jih je bilo postmenopavznih, 8 (3.7%) pa predmenopavznih ob času diagnoze. Prizadetost bezgavk je bila negativna pri 109 (50.7%) bolnicah, pozitivna pri 99 (46.0%) bolnicah in neznana pri 7 3.3% bolnicah. Srednji čas opazovanja bolnic je bil 84 mesecev (razpon, 5-133 mesecev). Bolezen se je ponovila pri 54 bolnicah (25.1%) in 59 (27.4%) jih je umrlo, od tega 44 zaradi raka dojke in 15 iz drugih vzrokov. Karakteristike tumorjev: patološka velikost tumorja, histologija tumorja, gradus tumorja in status

hormonskih receptorjev so bili določeni s standardnimi metodami v uporabi na Onkološkem Inštitutu v Ljubljani v času zdravljenja. Pri primerjavi rezultatov sem se omejil le na preživetje glede na izraženost markerja stathmin. Izraženost stathmin se glede na intenziteto deli na nizko, pri 126 (58.6%) bolnicah, in visoko, pri 89 (41.4%) bolnicah. V tabeli 5.1 so navedene porazdelitve izraženosti stathmin skupaj s karakteristikami tumorjev.

Opazujemo lahko dve obliki preživetja oz. dva nabora časovnega in cenzorskega atributa. Za bolezen značilno preživetje (disease specific survival, DSS), pri katerem je dogodek smrt zaradi bolezni, v tem primeru raka dojke. Cenzorski atribut je vzrok smrti; 44 (20.5%) bolnic je umrlo zaradi raka dojke, 15 (6.9%) bolnic je umrlo iz drugih vzrokov, 156 (72.6%) bolnic je bilo ob koncu opazovanja še živih; slednji dve vrednosti sta obravnavani kot cenzura podatka. Preživetje brez bolezni (disease free survival, DFS), kjer je dogodek ponovitev bolezni. Cenzorski atribut je ponovitev bolezni; bolezen se je ponovila pri 54 bolnicah (25.1%), pri 161 (74.9%) bolnicah pa do ponovitve bolezni do konca opazovanja ni prišlo.

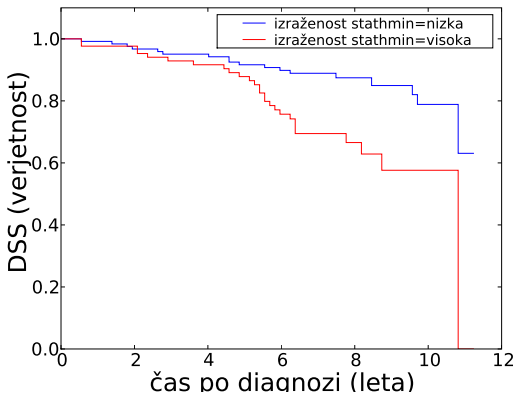
Podatki so z Onkološkega inštituta v Ljubljani. Zaradi varstva osebnih podatkov, so bili iz nabora predhodno odstranjeni vsi atributi s katerimi bi lahko neposredno identificirali posamezne bolnice. Da pa bi na Onkološkem inštitutu po potrebi vendarle lahko identificirali bolnice, je nabor vseboval MD5 izvlečke (glej [10]) identifikacijskih podatkov kot meta atribut.

Karakteristika		Izraženost stathmin	
	N(%)	Nizka N(%)	Visoka N(%)
Patološka velikost tumorja:			
pT1(0–20mm)	79(36.7%)	49(62%)	30(38%)
pT2(21–50mm)	119(55.4%)	68(57%)	51(43%)
pT3(>50mm)	17(7.9%)	9(53%)	8(47%)
Histološki tip tumorja:			
Duktalni invazivni	162(75.3%)	92(57%)	70(43%)
Drugi invazivni	53(24.7%)	34(64%)	19(36%)
Gradus malignosti:			
G1	24(11.2%)	20(83%)	4(17%)
G2	105(48.8%)	64(61%)	41(39%)
G3	85(39.5%)	42(49%)	43(51%)
Neznano	1(0.5%)		
Prizadetost bezgavk:			
Negativno	109(50.7%)	57(52%)	52(48%)
Pozitivno	99(46.0%)	65(66%)	34(34%)
Neznano	7(3.3%)		
Status hormonskih receptorjev:			
PR+	166(77.2%)	99(60%)	67(40%)
PR-	49(22.8%)	27(55%)	22(45%)

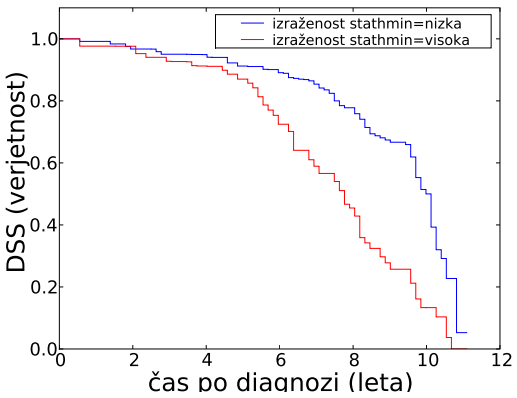
Tabela 5.1: Porazdelitev karakteristik tumorjev in izraženosti stathmin.

5.2 Oblika krivulj

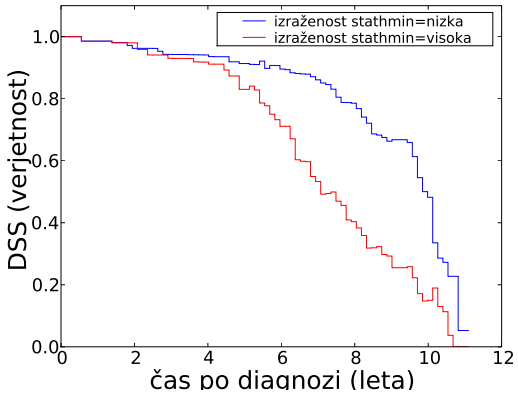
Na realni domeni ni možnosti, da bi testirali, koliko približki krivulj preživetja odstopajo od nekih pravih krivulj, ker pravih krivulj ne poznamo. Edina informacija o kvaliteti oblik je, koliko se skladajo z izsledki raziskave v [5]. Na grafih na sliki 5.1 so izrisane krivulje za bolezen značilnega preživetja (DSS) glede na različno izraženost markerja stathmin, na sliki 5.2 pa iste krivulje za preživetje brez bolezni (DFS). Vsi grafi, razen tistih dobljenih s predlagano metodo realizirano s SVM, se ujemajo z izsledki raziskave v [5]. Potrjujejo, da nizka izraženost markerja stathmin napoveduje boljše možnosti preživetja po zdravljenju. Vsi testi na realnih podatkih so bili opravljeni z delitvijo časovnega atributa na 80 intervalov.



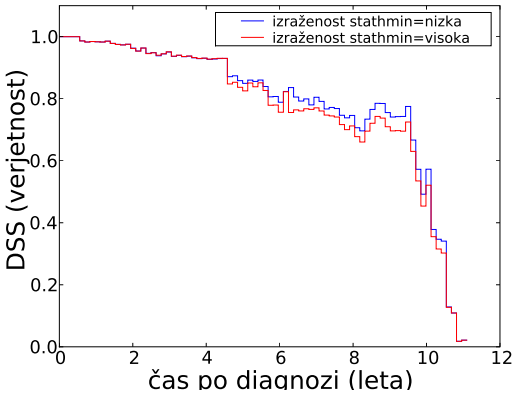
(a) Kaplan-Meier



(b) Predlagana metoda: naivni Bayesov klasifikator

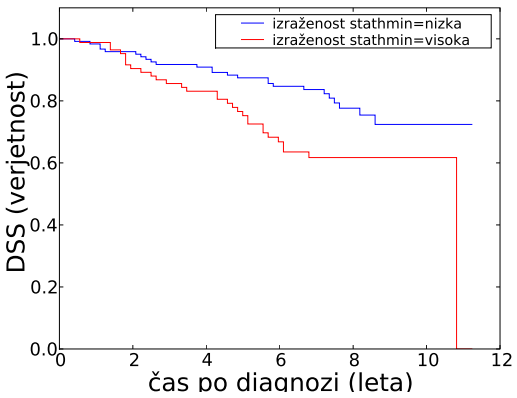


(c) Predlagana metoda: klasifikacijska drevesa

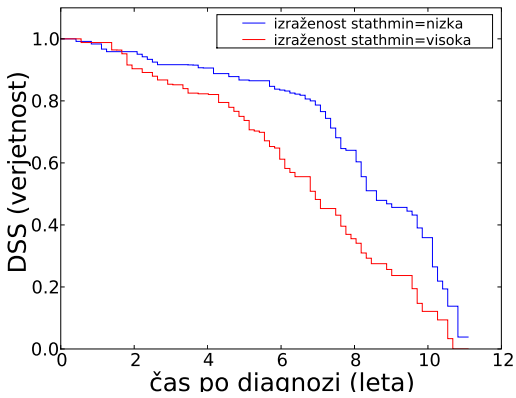


(d) Predlagana metoda: SVM

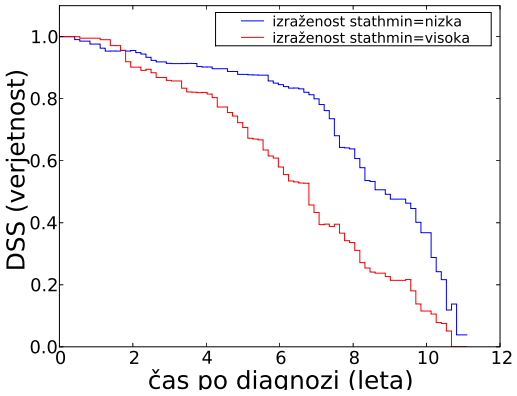
Slika 5.1: Krivulje za za bolezen značilno preživetje, dobljene z različnimi metodami.



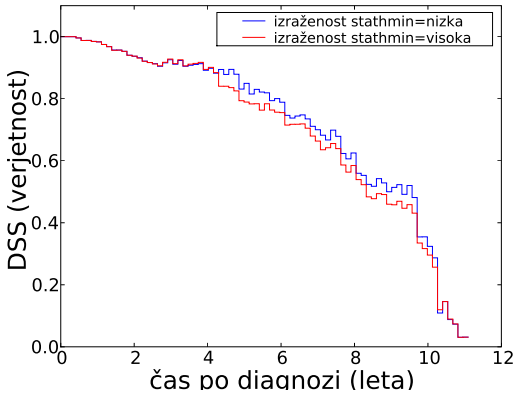
(a) Kaplan-Meier



(b) Predlagana metoda: naivni Bayesov klasifikator



(c) Predlagana metoda: klasifikacijska drevesa



(d) Predlagana metoda: SVM

Slika 5.2: Krivulje za preživetje brez bolezni, dobljene z različnimi metodami.

model	avg Δ DSS	max Δ DSS
Kaplan-Meier	4.449	10.613
predlagana metoda: naivni Bayes	3.051	7.905
predlagana metoda: klasifikacijsko drevo	2.957	6.924
predlagana metoda: SVM	3.176	7.436

Tabela 5.2: Napaka na necenzuriranih primerih pri določanju za bolezen značilnega preživetja.

5.3 Napovedna točnost

Težava pri testiranju metod kot napovednih modelov za preživetje na realnih podatkih je, da ni neodvisnih testnih podatkov. Tako sem moral uporabiti tehniko križnega preverjanja. Opravil sem petkratno križno preverjanje, pri čemer sem poskrbel, da so bile vse delitve stratificirane glede na število cenzuriranih primerov. Ker je pri obeh tipih preživetja (DSS in DFS) necenzuriranih primerov relativno malo, sem poleg povprečnega absolutnega odstopanja od prave vrednosti pri necenzuriranih primerih, meril še napako na cenzuriranih primerih. Ker odstopanja od prave vrednosti pri cenzuriranih podatkih ne moremo meriti, saj nimamo prave vrednosti, sem štel pri koliko primerih so modeli podcenili preživetje (se pravi, je model napovedal dogodek v času pred cenzuro) in za koliko je posamezen model v povprečju podcenil preživetje cenzuriranih primerov. Napake na necenzuriranih primerih so zbrane v tabelah 5.2 in 5.3, na cenzuriranih primerih pa v tabelah 5.4 in 5.5.

Predlagana metoda s poljubnim klasifikatorjem glede na oba tipa napake preseže natančnost Kaplan-Meierjevega modela. Težava Kaplan-Meierjevega modela je, da določa približek krivulje preživetja posameznega primera le na osnovi primerov, ki se s tem primerom ujemajo v vrednostih vseh atributov. V problemih z več atributi je potrebno imeti zelo veliko število primerov, da lahko zagotovimo dovolj gosto pokritost prostora, da bo takih primerov dovolj. V praktičnih aplikacijah je tolikšno število primerov navadno nemogoče ali nerealno dobiti. Zaradi tega se Kaplan-Meierjevega modela na ta način navadno ne uporablja.

model	avg Δ DFS	max Δ DFS
Kaplan-Meier	4.158	9.239
predlagana metoda: naivni Bayes	3.361	8.032
predlagana metoda: klasifikacijsko drevo	3.451	8.286
predlagana metoda: SVM	3.436	6.862

Tabela 5.3: Napaka na necenzuriranih primerih pri določanju preživetja brez bolezni.

model	št. podcenjenih primerov	povprečna podcenitev
Kaplan-Meier	137	6.395
predlagana metoda: naivni Bayes	50	1.289
predlagana metoda: klasifikacijsko drevo	48	1.297
predlagana metoda: SVM	53	1.065

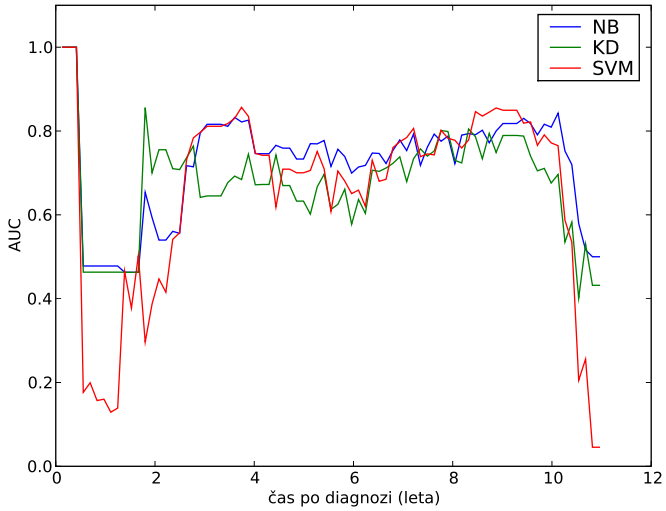
Tabela 5.4: Napaka na cenzuriranih primerih pri določanju za bolezen značilnega preživetja.

model	št. podcenjenih primerov	povprečna podcenitev
Kaplan-Meier	132	6.336
predlagana metoda: naivni Bayes	66	1.560
predlagana metoda: klasifikacijsko drevo	60	1.747
predlagana metoda: SVM	68	1.371

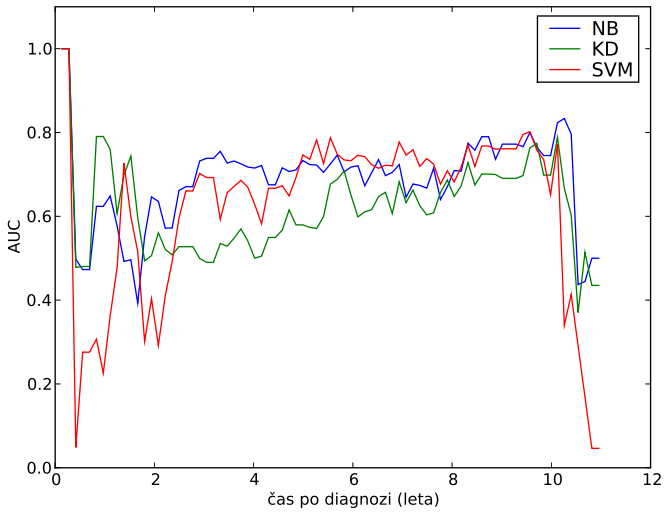
Tabela 5.5: Napaka na necenzuriranih primerih pri določanju preživetja brez bolezni.

5.4 interno ovrednotenje

Na realnih podatkih sem na enak način kot na umetnih interno ovrednotil zanesljivost posameznih klasifikatorjev, da sem dobil mero zanesljivosti krivulje. Ker je časovni atribut diskretiziran na več intervalov, kot pri umetnih podatkih, podajam rezultate v dveh grafih, 5.3 in 5.4. Krivulje jasno kažejo, da je zanesljivost krivulj preživetja najslabša na začetku in koncu časa, ko sta vrednosti razreda zelo neuravnoteženo zastopani.



Slika 5.3: Vrednosti AUC klasifikatorjev, zgrajenih za določanje za bolezen značilnega preživetja, po času. (NB - Predlagana metoda: naivniBayes, KD - Predlagana metoda: klasifikacijska drevesa, SVM - Predlagana metoda: SVM)



Slika 5.4: Vrednosti AUC klasifikatorjev, zgrajenih za določanje preživetja brez bolezni, po času. (NB - Predlagana metoda: naivniBayes, KD - Predlagana metoda: klasifikacijska drevesa, SVM - Predlagana metoda: SVM)

Poglavje 6

Zaključek

V pričujoči diplomi sem predstavil področje analize preživetja in probleme, s katerimi se to področje ukvarja. Predstavil sem razširjeno statistično metodo za obravnavo podatkov analize preživetja, Kaplan-Meierjev model, in predlagal novo metodo za ta namen, ki deluje na principih strojnega učenja. Predlagano metodo sem implementiral in jo testiral na umetno generiranih podatkih ter jo primerjal z Kaplan-Meierjevim modelom. Nazadnje sem podobne teste in primerjavo opravil še na realnih medicinskih podatkih.

Testi pokažejo, da lahko predlagana metoda po točnosti doseže in v nekaterih primerih celo preseže Kaplan-Meierjev model. Med njene prednosti je gotovo potrebno šteti tudi sposobnost upoštevanja atributov in posledično sposobnost, da izriše krivuljo preživetja za posamezen primer. Morda največja prednost pa je možnost internega ovrednotenja metode na podatkih, ki je lahko uporabna tako za določanje zanesljivosti krivulje, kot tudi kot mera, kateri klasifikator uporabiti za ocene verjetnosti preživetja.

Te prednosti imajo seveda svojo ceno. Prva je gotovo večja kompleksnost modela in večja časovna zahtevnost njegove izgradnje, vendar je predlagano metodo potrebno zgraditi le enkrat, potem pa lahko z njo hitro zgradimo posamezno krivuljo. Kaplan-Meierjev model mora za ta namen vedno znova deliti učno množico. Poleg tega je uporaba predlagane metode zahtevnejša za uporabnika, saj zahteva poznavanje področja strojnega učenja.

Daleč največja težava predlagane metode je možnost, da je dobljeni približek krivulje preživetja lahko nemonoton. To zelo zmanjšuje kredibilnost metode kot tehniko za ocenjevanje krivulje preživetja. Ni mi uspelo odkriti nikakršnega popravka metode ali pa ustreznega klasifikatorja, ki bi odpravil to težavo. Pristop, da bi dobljene krivulje naknadno naredili monotone z večanjem nekaterih verjetnosti je nevaren in zavajajoč. Popravki nekaterih

krivulj so mogoči s podrobnejšim pregledom naborov podatkov v časih, kjer se pojavi nemonotonost in klasifikatorjev zgrajenih na njih. Če na primer opazimo, da je nemonotonost pri modelu, realiziranem s klasifikacijskimi drevesi, posledica tega, da je primer, ki ga klasificiramo, ravno malo čez mejo področja, kjer bi bila dobljena verjetnost bolj smiselna, lahko upoštevamo to boljšo verjetnost. Vendar ob tem času ne znam podati niti nekih splošnih smernic za tovrstne popravke, kaj šele algoritma, ki bi to počel. Vseeno pa so krivulje dobljene s predlagano metodo lahko zelo dobre. Manjša nemonotonost niti ni tako moteča, če upoštevamo, da gre samo za približke krivulje preživetja, za katere se zavedamo, da imajo določeno mero napake. Ker lahko zanesljivost krivulje preverimo z internim vrednotenjem, je nemonotonost še lažje sprejeti, saj lahko ocenimo, kdaj je bolj nevarna. Vsekakor pa gre za težavo, katere morebitna rešitev ostaja predmet prihodnjega dela.

Način testiranja natančnosti napovedi je do Kaplan-Meierjevega modela morda malo nepošten, saj se ga za izris krivulj preživetja posameznih primerov navadno ne uporablja. Za ta namen je v statistični analizi preživetja bolj razširjen Coxov regresijski model [2]. V nadaljnem delu je gotovo potrebna eksperimentalna primerjava predlagane metode kot napovednega modela s tem modelom.

Ob opisani metodi se porajajo ideje za integracijo drugih področij strojnega učenja v analizo preživetja. Študija, kako se interakcije med atributi spreminjajo po času in kako vplivajo na verjetnost preživetja, opazovanje kako se informativnost atributov spreminja s časom, možnost podrobnejše raziskave zanimivih točk na krivulji preživetja (npr. točke kjer verjetnost preživetja nenadoma zelo pade) so le nekatere izmed njih. V celotnem področju analize preživetja s strojnim učenjem je tako še veliko možnosti za nadaljnje delo.

Dodatek A

Rezultati internega vrednotenja na umetnih podatkih

zaporedna št. klasifikatorja	100	200	500	1000	2000	5000	10000
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.823	0.97	0.999	0.974	0.993	0.996
3	0.991	0.942	0.997	0.987	0.989	0.993	0.995
4	0.85	0.972	0.991	0.985	0.99	0.991	0.993
5	0.9	0.967	0.977	0.98	0.989	0.993	0.993
6	0.842	0.949	0.966	0.984	0.99	0.994	0.993
7	0.861	0.954	0.982	0.984	0.991	0.992	0.993
8	0.913	0.953	0.971	0.987	0.99	0.994	0.993
9	0.819	0.952	0.972	0.993	0.99	0.996	0.993
10	0.875	0.907	0.996	0.975	0.993	0.996	0.996
11	0.877	0.905	0.971	0.975	0.988	0.993	0.995
12	0.925	0.875	0.98	0.997	0.991	0.996	0.996
13	0.925	0.698	0.973	0.982	0.98	0.992	0.996
14	0.925	0.698	0.97	1.0	0.997	0.998	0.997
15	1.0	0.6	0.989	0.996	0.996	0.998	0.992
16	1.0	0.5	0.918	0.994	0.946	0.996	0.997
17	1.0	0.5	0.856	0.925	0.87	0.999	0.999
18	0.638	0.5	0.657	0.996	0.998	0.996	0.999
19	0.5	0.5	0.5	0.997	0.997	0.996	0.998

Tabela A.1: Interno vrednotenje predlagane metode, realizirane z naivnim Bayesovim klasifikatorjem, na nešumnih podatkih. Tabela vsebuje vrednosti AUC za v modelu zaporedne klasifikatorje glede na velikost učnega nabora.

zaporedna št. klasifikatorja	100	200	500	1000	2000	5000	10000
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.792	0.924	0.956	0.959	0.971	0.98
3	0.929	0.845	0.938	0.964	0.98	0.986	0.989
4	0.9	0.96	0.938	0.978	0.982	0.988	0.993
5	0.901	0.911	0.935	0.957	0.981	0.99	0.992
6	0.944	0.928	0.963	0.961	0.981	0.99	0.991
7	0.926	0.934	0.95	0.964	0.982	0.989	0.994
8	0.944	0.941	0.948	0.97	0.978	0.989	0.992
9	0.898	0.891	0.931	0.968	0.979	0.988	0.989
10	0.875	0.963	0.951	0.976	0.963	0.984	0.988
11	0.842	0.951	0.91	0.948	0.984	0.979	0.987
12	0.887	0.903	0.933	0.973	0.962	0.989	0.991
13	0.887	0.869	0.839	0.964	0.959	0.989	0.981
14	0.887	0.869	0.922	0.969	0.987	0.985	0.985
15	0.955	0.763	0.794	0.949	0.944	0.964	0.978
16	0.955	0.839	0.775	0.868	0.95	0.941	0.967
17	0.955	0.794	0.862	0.923	0.935	0.949	0.949
18	0.743	0.794	0.649	0.896	0.83	0.927	0.966
19	0.458	0.473	0.488	0.49	0.994	0.999	0.95

Tabela A.2: Interno vrednotenje predlagane metode, realizirane s klasifikacijskimi drevesi, na nešumnih podatkih. Tabela vsebuje vrednosti AUC za v modelu zaporedne klasifikatorje glede na velikost učnega nabora.

zaporedna št. klasifikatorja	100	200	500	1000	2000	5000	10000
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.255	0.988	0.996	0.996	0.993	0.99
3	0.782	0.736	0.929	0.95	0.954	0.979	0.99
4	0.9	0.907	0.933	0.919	0.943	0.962	0.98
5	0.885	0.923	0.917	0.91	0.932	0.946	0.971
6	0.882	0.861	0.885	0.898	0.922	0.947	0.968
7	0.769	0.878	0.853	0.897	0.914	0.95	0.969
8	0.875	0.804	0.86	0.9	0.914	0.95	0.968
9	0.879	0.819	0.898	0.944	0.932	0.96	0.967
10	0.875	0.887	0.93	0.956	0.963	0.976	0.97
11	0.693	0.855	0.94	0.958	0.971	0.984	0.981
12	0.505	0.77	0.953	0.972	0.974	0.989	0.985
13	0.561	0.514	0.914	0.964	0.986	0.993	0.99
14	0.458	0.435	0.809	0.983	0.991	0.996	0.993
15	0.375	0.393	0.811	0.989	0.982	0.997	0.996
16	0.439	0.459	0.389	0.991	0.932	0.997	0.998
17	0.273	0.27	0.028	0.922	0.976	0.999	0.998
18	0.428	0.207	0.079	0.613	0.915	0.946	0.999
19	0.042	0.223	0.185	0.469	0.373	0.365	0.999

Tabela A.3: Interno vrednotenje predlagane metode, realizirane s SVM, na nešumnih podatkih. Tabela vsebuje vrednosti AUC za v modelu zaporedne klasifikatorje glede na velikost učnega nabora.

zaporedna št. klasifikatorja	100	200	500	1000	2000	5000	10000
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.818	0.966	0.999	0.976	0.992	0.994
3	0.991	0.939	0.995	0.981	0.987	0.99	0.994
4	0.788	0.972	0.976	0.979	0.985	0.988	0.99
5	0.883	0.947	0.968	0.976	0.984	0.99	0.991
6	0.858	0.98	0.965	0.978	0.987	0.992	0.991
7	0.717	0.934	0.967	0.983	0.987	0.991	0.991
8	0.883	0.942	0.954	0.984	0.989	0.992	0.991
9	0.925	0.927	0.956	0.992	0.989	0.995	0.991
10	0.913	0.856	0.991	0.975	0.992	0.994	0.995
11	0.826	0.863	0.948	0.972	0.984	0.991	0.995
12	0.826	0.875	0.988	0.996	0.984	0.995	0.996
13	0.826	0.698	0.96	0.981	0.981	0.99	0.996
14	0.826	0.698	0.97	0.998	0.994	0.997	0.997
15	0.895	0.6	0.936	0.991	0.996	0.997	0.992
16	0.895	0.5	0.918	0.998	0.971	0.997	0.996
17	0.895	0.5	0.856	0.999	0.916	0.998	0.989
18	0.638	0.5	0.657	0.996	0.998	0.992	0.999
19	0.5	0.5	0.5	0.997	0.997	0.996	0.998

Tabela A.4: Interno vrednotenje predlagane metode, realizirane z naivnim Bayesovim klasifikatorjem, na nešumnih podatkih brez cenzuriranih primerov. Tabela vsebuje vrednosti AUC za v modelu zaporedne klasifikatorje glede na velikost učnega nabora.

zaporedna št. klasifikatorja	100	200	500	1000	2000	5000	10000
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.788	0.9	0.92	0.955	0.966	0.986
3	0.988	0.855	0.955	0.964	0.971	0.986	0.989
4	0.908	0.908	0.951	0.967	0.982	0.984	0.992
5	0.908	0.906	0.948	0.96	0.971	0.983	0.991
6	0.958	0.935	0.962	0.968	0.982	0.983	0.99
7	0.867	0.916	0.933	0.966	0.974	0.983	0.99
8	0.871	0.886	0.921	0.978	0.98	0.982	0.99
9	0.842	0.915	0.908	0.949	0.969	0.986	0.987
10	0.887	0.969	0.961	0.96	0.975	0.98	0.99
11	0.831	0.964	0.935	0.943	0.966	0.977	0.986
12	0.831	0.903	0.949	0.959	0.959	0.976	0.982
13	0.831	0.869	0.918	0.94	0.988	0.988	0.985
14	0.831	0.869	0.893	0.896	0.981	0.97	0.984
15	0.955	0.763	0.812	0.947	0.947	0.931	0.957
16	0.955	0.839	0.775	0.943	0.949	0.937	0.972
17	0.955	0.794	0.862	0.827	0.829	0.949	0.977
18	0.743	0.794	0.649	0.896	0.83	0.898	1.0
19	0.458	0.473	0.488	0.49	0.994	0.999	0.95

Tabela A.5: Interno vrednotenje predlagane metode, realizirane s klasifikacijskimi drevesi, na nešumnih podatkih brez cenzuriranih primerov. Tabela vsebuje vrednosti AUC za v modelu zaporedne klasifikatorje glede na velikost učnega nabora.

zaporedna št. klasifikatorja	100	200	500	1000	2000	5000	10000
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.417	0.983	0.995	0.996	0.993	0.991
3	0.884	0.839	0.941	0.928	0.947	0.97	0.984
4	0.85	0.966	0.937	0.912	0.93	0.949	0.972
5	0.833	0.908	0.896	0.906	0.926	0.935	0.963
6	0.883	0.828	0.863	0.9	0.905	0.942	0.958
7	0.867	0.847	0.864	0.901	0.906	0.942	0.961
8	0.95	0.8	0.886	0.947	0.906	0.948	0.964
9	0.592	0.72	0.888	0.961	0.953	0.969	0.965
10	0.686	0.791	0.887	0.958	0.967	0.98	0.976
11	0.202	0.804	0.868	0.958	0.975	0.985	0.984
12	0.329	0.69	0.911	0.986	0.979	0.988	0.988
13	0.256	0.558	0.946	0.969	0.99	0.993	0.993
14	0.252	0.557	0.822	0.98	0.988	0.998	0.995
15	0.45	0.297	0.824	0.99	0.977	0.975	0.997
16	0.3	0.423	0.56	0.994	0.988	0.997	0.998
17	0.332	0.204	0.031	0.718	0.976	0.997	0.99
18	0.13	0.15	0.279	0.624	0.931	0.757	0.999
19	0.042	0.089	0.172	0.682	0.5	0.362	0.999

Tabela A.6: Interno vrednotenje predlagane metode, realizirane s SVM, na nešumnih podatkih brez cenzuriranih primerov. Tabela vsebuje vrednosti AUC za v modelu zaporedne klasifikatorje glede na velikost učnega nabora.

zaporedna št. klasifikatorja	100	200	500	1000	2000	5000	10000
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.478	0.935	0.956	0.968	0.976	0.986
3	0.818	0.872	0.965	0.962	0.969	0.974	0.977
4	0.857	0.957	0.972	0.961	0.965	0.967	0.971
5	0.781	0.934	0.945	0.943	0.955	0.961	0.967
6	0.833	0.912	0.946	0.962	0.965	0.962	0.968
7	0.826	0.905	0.938	0.973	0.969	0.969	0.969
8	0.873	0.924	0.943	0.967	0.971	0.974	0.968
9	0.888	0.88	0.907	0.963	0.97	0.974	0.968
10	0.838	0.841	0.951	0.965	0.971	0.974	0.969
11	0.814	0.872	0.93	0.973	0.971	0.976	0.972
12	0.776	0.926	0.912	0.97	0.956	0.973	0.977
13	0.776	0.759	0.899	0.965	0.95	0.978	0.976
14	0.776	0.759	0.827	0.984	0.952	0.969	0.977
15	0.695	0.669	0.883	0.964	0.899	0.974	0.976
16	0.695	0.736	0.738	0.912	0.843	0.978	0.978
17	0.695	0.486	0.603	0.912	0.785	0.899	0.982
18	0.467	0.486	0.494	0.895	0.485	0.835	0.988
19	0.5	0.496	0.496	1.0	0.492	0.476	0.892

Tabela A.7: Interno vrednotenje predlagane metode, realizirane z naivnim Bayesovim klasifikatorjem, na zašumljenih podatkih. Tabela vsebuje vrednosti AUC za v modelu zaporedne klasifikatorje glede na velikost učnega nabora.

zaporedna št. klasifikatorja	100	200	500	1000	2000	5000	10000
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.471	0.809	0.834	0.836	0.86	0.884
3	0.77	0.912	0.937	0.935	0.922	0.923	0.931
4	0.905	0.914	0.963	0.92	0.938	0.926	0.933
5	0.881	0.921	0.932	0.91	0.92	0.932	0.93
6	0.912	0.923	0.926	0.934	0.937	0.931	0.938
7	0.881	0.894	0.893	0.943	0.933	0.94	0.94
8	0.946	0.877	0.924	0.939	0.944	0.941	0.93
9	0.898	0.824	0.925	0.921	0.927	0.93	0.924
10	0.852	0.861	0.883	0.909	0.917	0.931	0.911
11	0.879	0.895	0.894	0.93	0.894	0.906	0.887
12	0.771	0.904	0.933	0.925	0.865	0.904	0.892
13	0.771	0.844	0.871	0.884	0.819	0.892	0.886
14	0.771	0.844	0.816	0.913	0.853	0.785	0.883
15	0.691	0.748	0.77	0.888	0.809	0.829	0.838
16	0.691	0.825	0.679	0.786	0.709	0.706	0.826
17	0.691	0.617	0.728	0.771	0.737	0.745	0.758
18	0.605	0.617	0.642	0.785	0.495	0.711	0.762
19	0.469	0.71	0.477	0.743	0.496	0.498	0.722

Tabela A.8: Interno vrednotenje predlagane metode realizirane klasifikacijskimi drevesi na zašumljenih podatkih. Tabela vsebuje vrednosti AUC za v modelu zaporedne klasifikatorje glede na velikost učnega nabora.

zaporedna št. klasifikatorja	100	200	500	1000	2000	5000	10000
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.183	0.965	0.924	0.84	0.91	0.91
3	0.258	0.804	0.907	0.913	0.888	0.94	0.951
4	0.78	0.85	0.921	0.907	0.919	0.935	0.952
5	0.9	0.866	0.895	0.888	0.908	0.925	0.946
6	0.753	0.844	0.853	0.889	0.893	0.922	0.946
7	0.801	0.848	0.791	0.886	0.881	0.93	0.944
8	0.9	0.699	0.777	0.88	0.888	0.941	0.939
9	0.871	0.787	0.854	0.896	0.912	0.951	0.939
10	0.663	0.822	0.933	0.926	0.92	0.947	0.946
11	0.566	0.739	0.895	0.919	0.919	0.953	0.939
12	0.512	0.803	0.891	0.959	0.909	0.919	0.934
13	0.236	0.469	0.856	0.932	0.916	0.955	0.933
14	0.349	0.295	0.504	0.98	0.882	0.918	0.917
15	0.277	0.404	0.579	0.936	0.624	0.876	0.863
16	0.193	0.561	0.226	0.938	0.44	0.908	0.887
17	0.189	0.204	0.546	0.612	0.33	0.73	0.9
18	0.167	0.156	0.536	0.666	0.135	0.832	0.948
19	0.042	0.232	0.174	0.922	0.059	0.296	0.843

Tabela A.9: Interno vrednotenje predlagane metode, realizirane s SVM, na zašumljenih podatkih. Tabela vsebuje vrednosti AUC za v modelu zaporedne klasifikatorje glede na velikost učnega nabora.

zaporedna št. klasifikatorja	100	200	500	1000	2000	5000	10000
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.45	0.92	0.952	0.959	0.969	0.981
3	0.887	0.863	0.962	0.954	0.96	0.967	0.971
4	0.875	0.942	0.945	0.952	0.958	0.96	0.965
5	0.875	0.925	0.944	0.932	0.951	0.953	0.96
6	0.892	0.907	0.933	0.95	0.956	0.957	0.963
7	0.75	0.874	0.929	0.965	0.966	0.966	0.964
8	0.808	0.915	0.934	0.961	0.972	0.972	0.964
9	0.912	0.875	0.933	0.963	0.966	0.97	0.964
10	0.793	0.817	0.948	0.958	0.97	0.971	0.967
11	0.833	0.76	0.917	0.969	0.968	0.972	0.969
12	0.833	0.926	0.837	0.972	0.949	0.967	0.974
13	0.833	0.759	0.855	0.96	0.941	0.974	0.978
14	0.833	0.759	0.81	0.987	0.939	0.967	0.976
15	0.784	0.669	0.859	0.962	0.876	0.972	0.974
16	0.784	0.736	0.738	0.942	0.853	0.973	0.978
17	0.784	0.486	0.603	0.824	0.816	0.898	0.982
18	0.467	0.486	0.494	0.895	0.485	0.667	0.988
19	0.5	0.496	0.496	1.0	0.492	0.476	0.892

Tabela A.10: Interno vrednotenje predlagane metode, realizirane z naivnim Bayesovim klasifikatorjem, na zašumljenih podatkih brez cenzuriranih primerov. Tabela vsebuje vrednosti AUC za v modelu zaporedne klasifikatorje glede na velikost učnega nabora.

zaporedna št. klasifikatorja	100	200	500	1000	2000	5000	10000
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.604	0.87	0.878	0.862	0.829	0.897
3	0.843	0.922	0.939	0.92	0.919	0.919	0.928
4	0.925	0.894	0.922	0.936	0.942	0.919	0.926
5	0.933	0.917	0.913	0.896	0.918	0.924	0.926
6	0.9	0.922	0.916	0.909	0.932	0.916	0.926
7	0.771	0.879	0.886	0.94	0.917	0.94	0.929
8	0.933	0.881	0.873	0.935	0.928	0.926	0.918
9	0.85	0.878	0.865	0.913	0.922	0.925	0.905
10	0.771	0.878	0.866	0.887	0.903	0.932	0.894
11	0.738	0.865	0.872	0.909	0.915	0.899	0.872
12	0.738	0.904	0.923	0.908	0.841	0.878	0.889
13	0.738	0.844	0.851	0.891	0.836	0.874	0.879
14	0.738	0.844	0.796	0.861	0.832	0.864	0.883
15	0.636	0.748	0.867	0.865	0.813	0.8	0.836
16	0.636	0.825	0.679	0.864	0.713	0.763	0.763
17	0.636	0.617	0.728	0.823	0.658	0.693	0.748
18	0.605	0.617	0.642	0.785	0.495	0.495	0.771
19	0.469	0.71	0.477	0.743	0.496	0.498	0.697

Tabela A.11: Interno vrednotenje predlagane metode realizirane klasifikacijskimi drevesi na zašumljenih podatkih brez cenzuriranih primerov. Tabela vsebuje vrednosti AUC za v modelu zaporedne klasifikatorje glede na velikost učnega nabora.

zaporedna št. klasifikatorja	100	200	500	1000	2000	5000	10000
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	0.429	0.935	0.92	0.869	0.894	0.918
3	0.561	0.849	0.898	0.902	0.899	0.928	0.944
4	0.817	0.866	0.902	0.892	0.902	0.925	0.944
5	0.9	0.892	0.882	0.88	0.905	0.913	0.935
6	0.9	0.841	0.818	0.878	0.881	0.912	0.937
7	0.833	0.823	0.805	0.894	0.884	0.925	0.936
8	0.9	0.749	0.832	0.908	0.903	0.942	0.933
9	0.775	0.713	0.883	0.92	0.917	0.953	0.934
10	0.552	0.797	0.893	0.94	0.929	0.945	0.932
11	0.45	0.575	0.908	0.915	0.914	0.943	0.923
12	0.403	0.863	0.769	0.912	0.895	0.901	0.922
13	0.291	0.406	0.785	0.894	0.883	0.943	0.932
14	0.434	0.276	0.452	0.979	0.866	0.949	0.907
15	0.3	0.424	0.509	0.939	0.678	0.921	0.88
16	0.391	0.525	0.395	0.927	0.811	0.902	0.912
17	0.482	0.297	0.447	0.901	0.308	0.81	0.88
18	0.145	0.123	0.45	0.805	0.133	0.334	0.886
19	0.042	0.232	0.106	0.937	0.058	0.327	0.86

Tabela A.12: Interno vrednotenje predlagane metode, realizirane s SVM, na zašumljenih podatkih brez cenzuriranih primerov. Tabela vsebuje vrednosti AUC za v modelu zaporedne klasifikatorje glede na velikost učnega nabora.

Slike

Tabele

Literatura

[1] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, New York, NY, USA, 1992. ACM Press.

[2] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, March 1972.

[3] J. Demsar, B. Zupan, and G. Leban. Orange: From experimental machine learning to interactive data mining (www.ailab.si/orange). White paper, Faculty of Computer and Information Science, University of Ljubljana, 2004.

[4] R.O. Duda, P.E. Hart, N.J. Nilsson, and SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER. *Subjective Bayesian Methods for Rule-Based Inference Systems*. Defense Technical Information Center, 1976.

[5] R. Golouh, T. Cufer, A. Sadikov, P. Nussdorfer, P.A. Usher, N. Brünner, M. Schmitt, R. Lesche, S. Maier, M. Timmermans, et al. The prognostic value of Stathmin-1, S100A2, and SYK proteins in ER-positive primary breast cancer patients treated with adjuvant tamoxifen monotherapy: an immunohistochemical study. *Breast Cancer Research and Treatment*, 110(2):317–326, 2008.

[6] E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[7] Ludvik Kos. Strojno učenje iz cenzuriranih podatkov. Magistrsko delo, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, 2006.

[8] Y.J. Lee, OL Mangasarian, and WH Wolberg. Survival-Time Classification of Breast Cancer Patients. *Computational Optimization and Applications*, 25(1):151–166, 2003.

[9] Ross J. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., 1993.

[10] R. Rivest. The MD5 Message Digest Algorithm, RFC 1321. *Internet Activities Board*, 1992.

[11] Kent A. Spackman. Signal detection theory: valuable tools for evaluating inductive learning. In *Proceedings of the sixth international workshop on Machine learning*, pages 160–163, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[12] B. Zupan, J. Demšar, M.W. Kattan, J.R. Beck, and I. Bratko. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial Intelligence In Medicine*, 20(1):59–75, 2000.

Izjava

Izjavljam, da sem diplomsko nalogo izdelal samostojno pod vodstvom mentorja akad. prof. dr. Ivana Bratka. Izkazano pomoč drugih sodelavcev sem v celoti navedel v zahvali.

Ljubljana, 17.9.2008 Aljaž Košmerlj