UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Dajra Šabić

Podatkovno rudarjenje in odkrivanje zakonitosti v zdravstvu: predvidevanje preživetja raka dojke s tehnikami podatkovnega rudarjenja

Magistrsko delo

Ljubljana, 2017

UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Dajra Šabić

Mentor: izr. prof. dr. Damjan Škulj

Podatkovno rudarjenje in odkrivanje zakonitosti v zdravstvu: predvidevanje preživetja raka dojke s tehnikami podatkovnega rudarjenja

Magistrsko delo

Ljubljana, 2017

Zahvaljujem se izr. prof. dr. Damjanu Škulju za mentorstvo in strokovno pomoč pri nastajanju diplomskega dela.

Tini Žagar se zahvaljujem za pripravo podatkov za analizo ter strokovno razlago medicinske terminologije.

Podatkovno rudarjenje in odkrivanje zakonitosti v zdravstvu: predvidevanje preživetja raka dojke s tehnikami podatkovnega rudarjenja

Povzetek

Odkrivanje zakonitosti in podatkovno rudarjenje veljata za hitro razvijajoči se področji računalniške znanosti. Njun pomen se veča zaradi povečane potrebe po metodologiji in orodjih za analiziranje in razumevanje velike količine podatkov. Danes vedno več organizacij, vključno s sodobnimi zdravstvenimi ustanovami, ustvarja in zbira velike količine podatkov, kar pa zahteva avtomatiziran način analiziranja in pridobivanja znanja. Odkrivanje zakonitosti v podatkovnih bazah postaja vedno bolj priljubljeno raziskovalno orodje za zdravstvene raziskovalce pri prepoznavanju vzorcev in povezav med velikim številom spremenljivk, ki pa ravno omogočajo napovedovanje izidov bolezni s pomočjo preteklih primerov. V magistrski nalogi smo opravili predstavili proces odkrivanja zakonitosti in podatkovno rudarjenje, opravili pregled literature na temo diagnosticiranja in napovedovanja preživetja raka dojke, ter opravili praktičen primer na podali podatkov iz Registra raka Republike Slovenije. Pri kreiranju napovedovalnih modelov se je v primerjavi z RBF nevronskim omrežjem in REP odločitvenim drevesom metoda logistične regresije izkazala kot najuspešnejša.

Ključne besede: proces odkrivanja zakonitosti, podatkovno rudarjenje, rak dojke.

Data mining and knowledge discovery in health care: predicting breast cancer survivability with data mining techniques

Abstract

Knowledge discovery process and data mining are two fast growing computer science domains. Their importance is growing due to increasing need for methodology and tools for analyzing and understanding big data. Nowadays, all organizations collect huge amounts of data in datasets, which also requires automatic ways of analysis and knowledge discovery. Knowledge discovery in datasets is becoming increasingly popular research tool for medical researchers in order to identify patterns and relationships between large numbers of variables, which enables prediction of outcomes of diseases with the help of past cases. In this thesis, we present the knowledge discovery process and data mining, and review literature on diagnosing and predicting survivability of breast cancer. We also carried out an analysis on the dataset provided by the Cancer Registry of the Republic of Slovenia. The results indicate that logistic regression is the best predictor in comparison with other two performed methods - RBF neural network and REP decision tree.

Key words: Knowledge discovery process, data mining, breast cancer.

KAZALO

KAZALO SLIK

KAZALO TABEL

KAZALO GRAFOV

1 UVOD

Odkrivanje zakonitosti in podatkovno rudarjenje veljata za hitro razvijajoči se področji računalniške znanosti. Njun pomen se veča zaradi povečane potrebe po metodologiji in orodjih za analiziranje in razumevanje velike količine podatkov, ki jih dnevno pridobivajo razne institucije kot so bolnišnice, raziskovalni laboratoriji in banke. Vse to je rezultat povečane uporabe elektronskih medijev (Cios in Kurgan 2005).

Splošno znano je, da je znanje najpomembnejše sredstvo vseh organizacij v času današnje družbe, ki jo vodi informacijska tehnologija (Kaur in Wasan 2006). Danes vedno več organizacij, vključno s sodobnimi zdravstvenimi ustanovami, ustvarja in zbira velike količine podatkov, kar pa zahteva avtomatiziran način analiziranja in pridobivanja znanja (Ngan in drugi 1999). Zdravstveno področje velja kot "bogato z informacijami", a z dokaj "slabega znanja''. Bogastvo se odraža v veliki količini zdravstvenih podatkov (ang. »medical data«), ki so edinstveni (imajo posebne karakteristike), in moramo z njimi ravnati zelo pozorno (Cios in Moore 2006). Primanjkujejo učinkovita orodja za analizo teh podatkov, s katerimi se odkrivajo skrite povezave in trendi (na primer napovedovanje razvoja bolezni ipd.) (Kaur in Wasan 2006).

Opazovanje značilnosti populacije pripomore k odkrivanju faktorjev povezanih z določenimi izidi. Študije opazovanja, kot sta statistično učenje in podatkovno rudarjenje, odkrivajo povezave spremenljivk z izidi, a ne morejo vedno odkriti tudi vzročno-posledičnih povezav. Statistična raziskovanja podatkov postajajo v vedno večji meri del ostalih znanosti, kot tudi medicine in biotehnologije (Bellaachia in Guven 2006).

Medicina ima posebno vlogo v znanosti, filozofiji in vsakdanjiku. Pri izidih zdravstvene oskrbe ge za življenje ali smrt in se aplicira na vsakogar. Medicina je nujnost in ne zgolj neobvezni luksuz, užitek ali udobje. Med vsemi poklici je najdaljše izobraževanje ravno v medicini. Od zdravnikov se pričakuje, da so etični in skrbni. Medicina je priljubljena tema v popularnih medijih. Zdravstvena oskrba je tvegana, v primeru spodrsljajev pa je želja po pravnem maščevanju (tožbah) intenzivna in kaznovalna. Zdravstvene informacije o bolniku veljajo kot izredno zaupne, v javnosti pa je prisoten veliki strah pred razkritjem. Vsi z veseljem sprejemamo koristi zdravstvenih raziskav opravljenih na drugih pacientih, zelo redko pa smo

sami pripravljeni sodelovati in prispevati svoje osebne podatke v te namene (Cios in Moore 2002).

V zadnjih letih se podatkovno rudarjenje vedno bolj uporablja v zdravstveni literaturi. Na splošno ni podane neke točne definicije, ampak je razumljeno kot uporaba relativno novih metod in orodij za analiziranje velike količine podatkov (Bellazzi in Zupan 2006). Metode podatkovnega rudarjenja so še posebej priljubljene med raziskovalci, ki se ukvarjajo z opredeljevanjem in odkrivanjem vzorcev in modelov predvidevanja (Thongkam in drugi 2009). Te metode so se izkazale za veliko učinkovitejše kot tradicionalne statistične metode (Ohno-Machado 2001; Xiong in drugi 2005; Han in Kamber 2006).

Rak dojke je najpogostejši rak pri ženskah in prizadene 10 odstotkov vseh žensk v različnih fazah njihovega življenja. Stopnja preživetja v zadnjih letih opazno narašča. Po podatkih naj bi bila stopnja preživetja raka dojke pri ženskah okoli 88 odstotna v petih letih po postavljeni diagnozi in 80 odstotna po desetih letih od postavljene diagnoze. Pri spremljanju raka dojke je najpomembnejša zgodnja napoved (Ahmad in drugi 2013). Rak dojke je tudi primarni razlog umrljivosti žensk v primerjavi z ostalimi vrstami raka in je najbolj nevarna vrsta raka pri ženskah po vsem svetu. Zgodnje odkrivanje te bolezni je bistvenega pomena za zmanjšanje umrljivosti (Kumar in drugi 2013). Odkrivanje zakonitosti v podatkovnih bazah, kar vključuje podatkovno rudarjenje, postaja vedno bolj priljubljeno raziskovalno orodje za zdravstvene raziskovalce pri prepoznavanju vzorcev in povezav med velikim številom spremenljivk, ki pa ravno omogočajo napovedovanje izidov bolezni s pomočjo preteklih primerov (Gupta in drugi 2011).

V magistrski nalogi bomo raziskali, katere tehnike podatkovnega rudarjenja se najpogosteje uporabljajo v zdravstvu, in sicer za predvidevanje preživetja raka dojke v prvi vrsti. Pred tem se bomo osredotočili tudi na same zdravstvene podatke, ter raziskali njihove posebne karakteristike, zaradi katerih jih raziskovalci pojmujejo kot unikatne. V zadnjem delu magistrske naloge bomo v empiričnemu delu s pomočjo podatkov, ki smo jih pridobili iz Registra raka Republike Slovenije, prikazali konkreten primer aplikacije določenih tehnik podatkovnega rudarjenja za predvidevanje preživetja raka dojke. Za izvedbo raziskave in uporabo podatkov iz Registra raka Republike Slovenije smo pridobili pisno soglasje Komisije

za medicinsko etiko (KME). V magistrski nalogi bomo odgovore na raziskovalna vprašanja pridobili s pomočjo analize sekundarnih virov. V empiričnemu delu bomo prikazali konkreten primer podatkovnega rudarjenja s pomočjo prosto dostopnega programskega orodja za podatkovno rudarjenje WEKA.

2 PODATKOVNO RUDARJENJE IN ODKRIVANJE ZAKONITOSTI

V procesu *odkrivanja zakonitosti* (ang. »Knowledge discovery«) se uporabljajo različne metode *podatkovnega rudarjenja* (ang. »Data mining«) z namenom odkrivanja novih znanj iz ogromnih podatkovnih baz. Pojma odkrivanje zakonitosti in podatkovno rudarjenje se prvič pojavita v poznih osemdesetih letih in se od takrat redno uporabljata (Cios in drugi 2012).

2.1 Odkrivanje zakonitosti (ang. »Knowledge discovery«)

2.1.1 Opredelitev odkrivanja zakonitosti

Področje odkrivanja zakonitosti in podatkovnega rudarjenja sta inherentno povezana s podatkovnimi bazami. Proces odkrivanja zakonitosti poskrbi, da so podatkovne baze uporabniku prijazne in tako pripomore k ugodnejšemu barantanju z veliko količino podatkov in njihovo uporabo. Podatkovne baze so izjemno uporabne, saj omogočajo uporabnikom shranjevanje velikih količin podatkov. Težava pri tem pa je, da nobeno človeško bitje ni zmožno učinkovito uporabiti tolikšnih količin podatkov in razumeti osnovnih trendov ter na osnovi tega opraviti racionalne odločitve. Vse shranjene informacije postanejo vedno manj uporabne zaradi težav pri pridobivanju in dostopu do le-teh v nekem razumljivem formatu na višji ravni povzemanja (Cios in drugi 2012). S poskusom reševanja teh težav se razvije področje odkrivanja zakonitosti. Pojem odkrivanja zakonitosti se pojavi okoli leta 1989, avtorji Fraweley, Piatetsky-Shapiro in Matheus ga definirajo kot netrivialen proces identifikacije veljavnih, novih, potencialno uporabnih in končno razumljivih vzorcev (povezav, korelacij, trendov in tako naprej) v podatkih (Cios in drugi 2012).

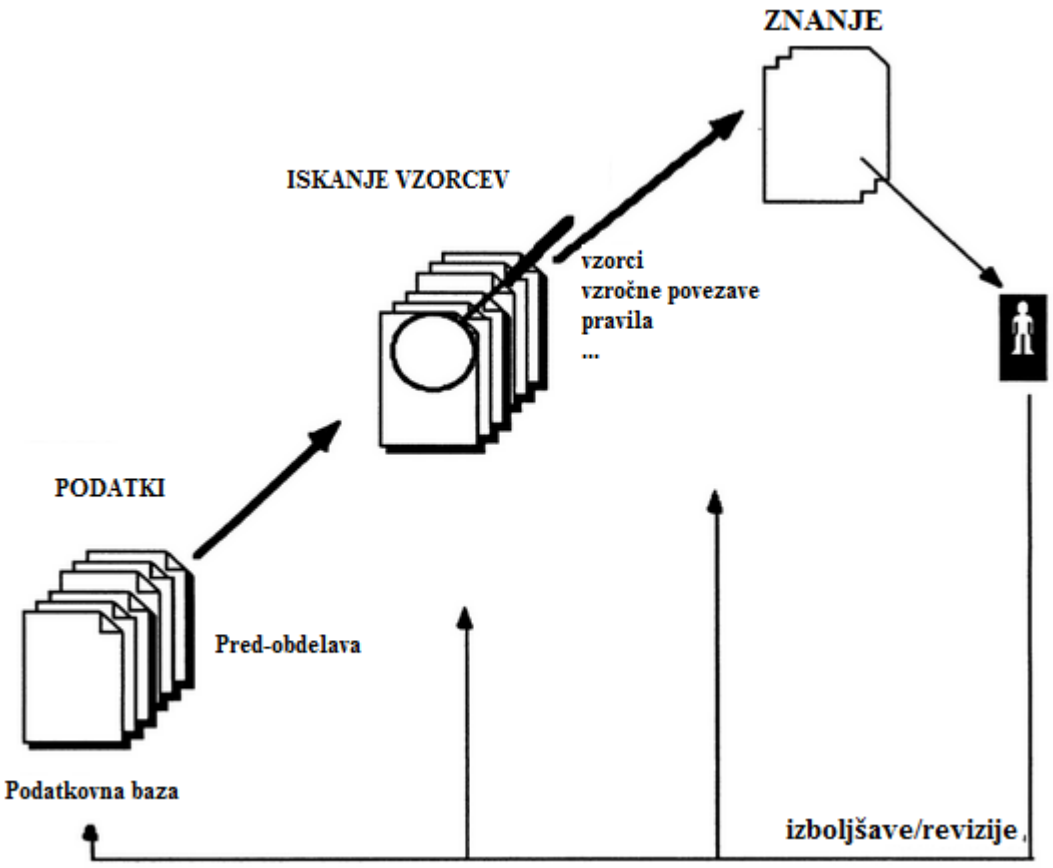
Fayyad in drugi (1996) definirajo odkrivanje zakonitosti kot proces uporabe podatkovnih baz skupaj z želenimi izbirami, pred-obdelavo, pod-vzorčenjem in transformacijo baze, pa tudi kot uporabo metod podatkovnega rudarjenja za pridobivanje vzorcev iz podatkovne baze, ter kot vrednotenje rezultatov podatkovnega rudarjenja za identifikacijo podskupin naštetih vzorcev pridobljenega znanja.

Pojma podatkovno rudarjenje in odkrivanje zakonitosti se včasih zamenjujeta. Vendar ne gre za en sam proces. Proces odkrivanja zakonitosti zajema metode podatkovnega rudarjenja kot podpodročja, ki prispevajo k končnemu uspehu (Cios in drugi 2012).

2.1.2 Potek odkrivanja zakonitosti

Odkrivanje zakonitosti je inherentno povezano s podatkovnimi bazami. Ko se srečamo s podatkovnimi bazami, izvajamo iskanje vzorcev oziroma povezav in tako ustvarjamo pomenljive dele zakonitosti (Slika 2.1).

Slika 2.1: Osnovna shema odkrivanja zakonitosti

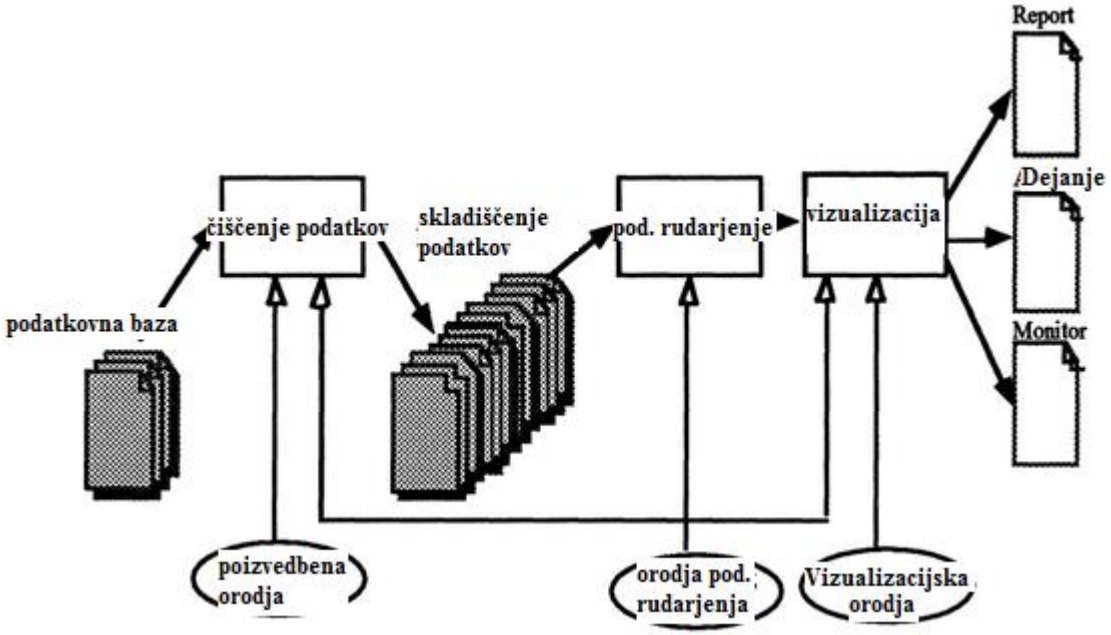


Vir: Cios in drugi (2005).

Cios in drugi (2012) nadalje izpostavijo osnovne funkcionalne faze procesa odkrivanja zakonitosti (Slika 2.2). Faza procesiranja iz osnovne sheme se pogosto nanaša na *čiščenje*

podatkov (ang. »data cleaning«). Očiščeni podatki so shranjeni v podatkovna skladišča, sledi podatkovno rudarjenje, s pomočjo katerega izhodni generator ustvarja poročila, sezname dejanj in poročila zaslona. Vsak korak je podprt z drugačno metodologijo. Kot bomo videli v nadaljevanju, gre pri podatkovnem rudarjenju za množico algoritmičnih orodij, kot so na primer statistika, regresijski modeli, razvrščanje v skupine in drugi. Mehanizmi, ki se uporabijo na ravni izhodnega generatorja, so odvisni od številnih vizualizacijskih orodij. V osnovni shemi vidimo, da je ključna interakcija med uporabnikom in sistemom odkrivanja zakonitosti. S tem je poudarjena tudi *dinamična povratna zanka* (ang. »feedback loop«), znotraj katere generiramo številne zaporedne izboljšave, ki so ključne za delovanje celotnega procesa odkrivanja zakonitosti.

Slika 2.2: Temeljne funkcionalne faze procesa odkrivanja zakonitosti



Vir: Cios in drugi (2012).

Proces odkrivanja zakonitosti je sestavljen iz naslednjih korakov:

- razumevanje domene, v kateri bo izvedeno odkritje,
- formiranje nabora podatkov, čiščenje in skladiščenje podatkov,
- pridobivanje vzorcev, kar je bistvo podatkovnega rudarjenja,
- post-obdelava odkritih zakonitosti,

- uporaba rezultatov procesa odkrivanja zakonitosti (Cios in drugi 2012).

Pri odkrivanju zakonitosti prihaja do temeljnih težav zaradi same narave podatkovnih baz in podatkov. Do težav prihaja zaradi:

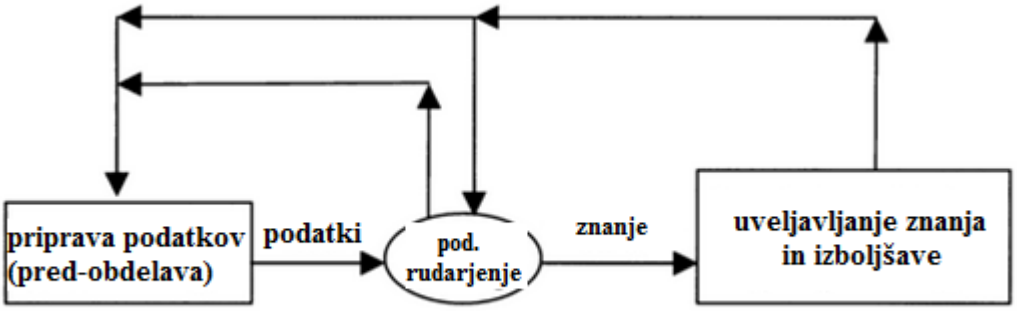
- ogromne količine podatkov – nekatere tehnike podatkovnega rudarjenja so zelo občutljive na velikost podatkov v smislu časovne kompleksnosti;
- dinamične narave podatkov – baze podatkov se konstantno posodabljajo z dodajanjem novih in z zamenjavo obstoječih podatkov, zato je potrebno postopoma posodabljati tudi pridobljeno znanje;
- nepopolnosti ali nenatančnosti podatkov – v mnogih praktičnih primerih so informacije v podatkovnih bazah nepopolne ali nenatančne, za takšne primere pa so strokovnjaki razvili različne metode znotraj metod strojnega učenja;
- šumnih podatkov – pri zbiranju podatkov je zelo težko eliminirati ali zreducirati število nesistematičnih napak tj. šumnih podatkov, na katere so mnoge metode podatkovnega rudarjenja dokaj občutljive;
- manjkajočih vrednosti – ničelne vrednosti, ki so neznane, neuporabne vrednosti, prinašajo težave, saj večina tehnik zahteva določenost dimenzij podatkovnega objekta;
- odvečnih ali nepomembnih podatkov – nabor podatkov lahko vsebuje odvečne ali nepomembne podatke, katere je potrebno eliminirati z optimalnimi in kvazi-optimalnimi algoritmi (Cios in drugi 2012).

Odkrivanje zakonitosti je dinamičen, visoko interaktiven, iterativen in polno vizualiziran proces. Glavni cilji procesa so pridobitev uporabnih poročil, prepoznavanje zanimivih dogodkov in trendov, podpora procesu odločanja ter izkoriščanje podatkov za doseganje znanstvenih, poslovnih in operativnih ciljev (Cios in drugi 2012).

Proces odkrivanja zakonitosti je tako iterativen kot tudi interaktiven (Slika 2.3). Iterativen je, saj je izhod (output) vsakega koraka povratna informacija (feedback) prejšnjega koraka in so mnoge iteracije procesa dejansko potrebne za ekstrakcijo visoko kvalitetnega znanja iz samih podatkov. Interaktiven je, saj mora v tej zanki (loop) biti vključen uporabnik, ki pomaga pri

pripravi podatkov, validaciji odkritega znanja ter pri izboljšavah in ostalih opravilih (Freitas 2013).

Slika 2.3: Iterativna narava procesa odkrivanja zakonitosti



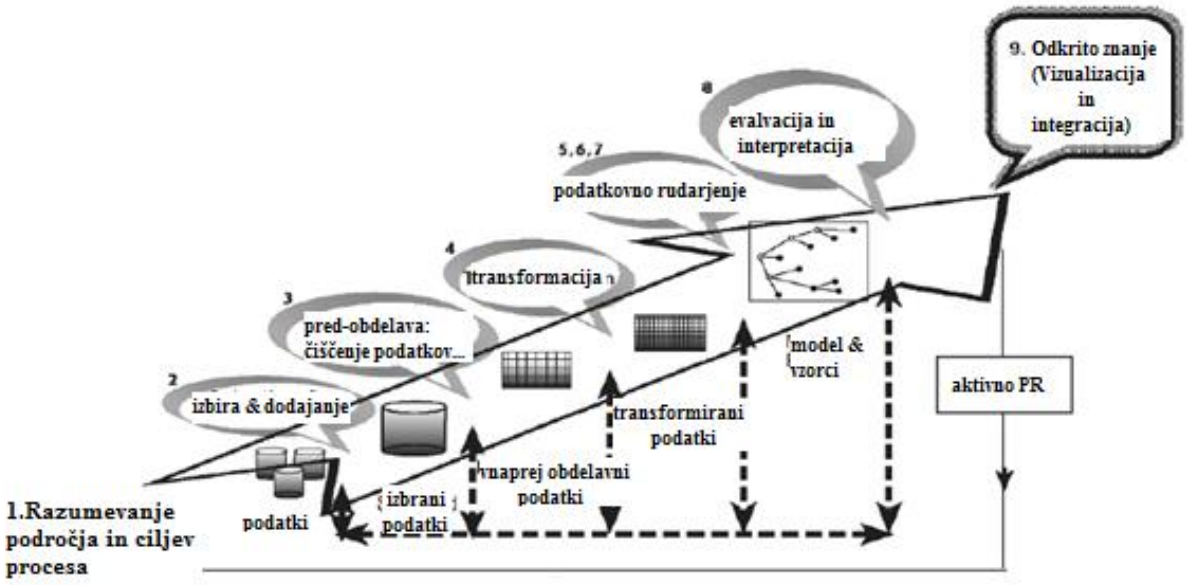
Vir: Freitas (2013).

Maimon in Rokach (2010) sta prav tako izpostavila, da je odkrivanje zakonitosti iterativen in interaktiven proces. Proces se prične z določanjem ciljev in se zaključi z implementacijo odkritega znanja. Kot rezultat sledijo spremembe na področju uporabe, kar zapira zanko in se proces odkrivanja zakonitosti lahko znova zažene na novih podatkih. Fayyadov model procesa zakonitosti je sestavljen iz devetih korakov (Slika 2.4):

1. Razvoj razumevanja področja uporabe

V tem začetno pripravljalnem koraku gre za pripravo prizorišča za razumevanje tega, kaj naj bi se storilo z mnogimi odločitvami. Osebe odgovorne za proces morajo razumeti in definirati cilje končnega uporabnika in okolja, v katerem se proces izvaja. Potrebno je ustrezno predznanje in po izvajanju procesa morda tudi revizija in prilagajanje tega koraka. Po razumevanju ciljev procesa sledi v naslednjih treh korakih pred-obdelava podatkov.

Slika 2.4: Proces odkrivanja zakonitosti



Vir: Maimon in Rokach (2010).

2. Izbira in kreiranje nabora podatkov, na katerih se bo izvajal proces

Po določanju ciljev procesa moramo določiti katere podatke bomo uporabili. Pri tem gre za ugotavljanje dostopnosti in pridobivanje potrebnih podatkov ter nato povezovanje vseh teh v skupni nabor podatkov vključno z atributi, ki jih bomo pri tem upoštevali. Ta korak je izredno pomemben, saj podatkovno rudarjenje deluje na podlagi dostopnih podatkov in se lahko v primeru pomanjkljivosti zgodi, da celotna raziskava ne uspe. Potrebno je upoštevati kar se da več atributov (čim več spremenljivk), samo zbiranje, organiziranje in delovanje podatkovnih zbirk pa je drago. V tej točki pripomore interaktivno-iterativni vidik procesa, saj nam je tako omogočeno, da začnemo z najboljšo možno razpoložljivostjo nabora podatkov, kasneje pa se ta razširi in upošteva učinke v smislu odkrivanja znanja in modeliranja.

3. Pred-obdelava in čiščenje podatkov

V tem koraku gre za okrepitev zanesljivosti podatkov. Vključuje čiščenje podatkov in s tem spopadanje z manjkajočimi vrednostmi ter odstranjevanje šumnih podatkov in odstopanj.

4. Transformacija podatkov

Po čiščenju podatkov sledi priprava podatkov za podatkovno rudarjenje. Metode pri tem vključujejo redukcijo dimenzij in transformacijo atributov. Ta korak je ključen za uspeh celotnega procesa, a je navadno specifičen glede na posamezen projekt.

5. Izbira ustreznih tehnik podatkovnega rudarjenja

V petem koraku smo pripravljeni za izbiro ustrezne tehnike podatkovnega rudarjenja, kot so na primer klasifikacija, regresija, razvrščanje v skupine. Izbira je odvisna predvsem od samih ciljev procesa odkrivanja zakonitosti, ter od prejšnjih korakov. Podatkovno rudarjenje ima dva cilja: predvidevanje in opisovanje. Predvidevanje se pogosto nanaša na nadzorovano podatkovno rudarjenje, medtem ko opisno podatkovno rudarjenje vključuje nenadzorovan in vizualizacijski vidik. Večina tehnik podatkovnega rudarjenja temelji na induktivnem učenju, kjer je model eksplicitno ali implicitno konstruirana posplošitev iz zadostnega števila učnih primerov.

6. Izbira algoritma podatkovnega rudarjenja

Ta faza vključuje izbiro specifične metode za iskanje vzorcev. Na primer – če bolj kot razumevanje želimo upoštevamo preciznost, je ustreznejša tehnika nevronsko omrežje kot pa odločitvena drevesa. Vsaka strategija meta-učenja se lahko izvede na več načinov. Meta-učenje se osredotoča na razlago dejavnikov (ne)uspeha algoritmov podatkovnega rudarjenja pri določenem problemu. Gre torej za poskus razumevanja pogojev, pod katerimi je nek algoritem podatkovnega rudarjenja najbolj primeren. Vsak algoritem pa ima parametre in taktike učenja.

7. Uporaba algoritma podatkovnega rudarjenja

Tukaj dosežemo implementacijo algoritma, a je včasih v tem koraku potrebno večkratno izvajanje algoritma, da bi dosegli zadovoljive rezultate (na primer nastavljanje kontrolnih parametrov algoritmov).

8. Evalvacija

V predzadnji fazi vrednotimo in interpretiramo pridobljene vzorce (pravila, zanesljivost itd.) glede na opredeljene cilje procesa. Tukaj premislimo o korakih pred-obdelave glede na njihov vpliv na rezultate podatkovnega rudarjenja (lahko se vrnemo na četrti korak in kaj dodamo ter ponovimo postopek). Ta korak je osredotočen na razumljivost in uporabnost inducirane ga

modela. Tukaj je odkrito znanje tudi dokumentirano za nadaljnjo uporabo. Sledi še zadnji korak, ki je uporaba in splošna povratna informacija o vzorcih in rezultatih pridobljenih s podatkovnim rudarjenjem.

9. Uporaba odkritega znanja

Na samem koncu procesa smo pripravljeni vključiti pridobljeno znanje v nek drugi sistem za nadaljnje ukrepe. Znanje je aktivno v smislu, da lahko uvedemo spremembe v procesu in izmerimo učinek tega. Uspešnost tega koraka določa učinkovitost celotnega procesa odkrivanja zakonitosti.

Poznamo 5 različnih modelov procesa odkrivanja zakonitosti. Zgoraj opisani Fayyadov model je najbolj priljubljen in tudi največkrat uporabljan. Ostali modeli avtorjev Cios, Annand in Buchner, Cabena, ter CRISP-DM model se razlikujejo po številu korakov. Vsi imajo dobre in slabe lastnosti, ki se nanašajo na področje uporabe in posebne poslovne cilje (Cios in drugi 2007).

Fayyadov model procesa odkrivanja zakonitosti zagotavlja podroben tehničen opis glede na analizo podatkov, primanjkuje pa mu poslovni vidik. Model se največkrat uporablja na področju medicine, inženirstva, programske opreme in e-poslovanja. Model Annand in Bucher zagotavlja podrobno razčlenitev začetnih fazah postopka, najbolj pa se uporablja v marketingu in prodaji. Ciosov model črpa iz akademskih in industrijskih modelov in poudarja iteracijske vidike, prav tako pa identificira in opisuje več eksplicitno povratnih zank. Tako kot Fayyadov model se najbolj uporablja v medicini in na področju programske opreme. Model Cabena je poslovno usmerjen in enostaven za razumevanje tudi za ostale strokovnjake in ne samo podatkovne analitike. Uporablja se na področju marketinga in prodaje. CRISP-DM model uporablja besedišče, ki je enostavno za razumevanje in ima dobro dokumentacijo - deli vse korake v pod-korake, ki zagotavljajo vse potrebne podatke. Ta model se uporablja tako na področju medicine in inženirstva kot tudi v marketingu in prodaji (Cios in drugi 2007).

Pal in Jain (2005) izpostavljata pomanjkljivost Fayyadovega modela v primerjavi s Ciosovim, v tem, da se peti in šesti korak procesa (izbira ustreznih tehnik in algoritma podatkovnega rudarjenja) izvajata prepozno v procesu. Ta dva koraka bi se morala izvesti že v začetnem delu

procesa, saj je cilj priprave podatkov ravno priprava podatkov, da bi bili uporabni za izbrane tehnike podatkovnega rudarjenja. Nasprotno pa se v Fayyadovem modelu orodja podatkovnega rudarjenja izbirajo šele v šestem koraku glede na izide priprave podatkov. To lahko povzroči težave pri izbiri orodij podatkovnega rudarjenja, saj so lahko pripravljeni podatki neprimerni za določeno orodje.

Model, sestavljen iz pet korakov (Cabena), ki se uporablja v poslovni domeni, je podoben Ciosovemu modelu. Razlikuje se po tem, da ne vsebuje koraka razumevanja podatkov, kar pa je s strani drugih avtorjev izpostavljeno kot velika pomanjkljivost. Ciosov model ima to prednost, da je podoben modelu CRISP-DM, ki je bil potrjen na velikih poslovnih aplikacijah (Cios in drugi 2007).

Tabela 2.1: Primerjava modelov procesa odkrivanja zakonitosti

Model	**Fayyad in drugi**	**Anand in Buchner**	**Cios in drugi**	**Cabena in drugi**	**CRISP-DM**
Domena ozadja	Akademsko	Akademsko	Hibridno akadamsko in industrija	Industrija	Industrija
Koraki	1. Razvoj in razumevanje področja uporabe	1. Identifikacija človeških virov	1. Razumevanje domene problema	1. Določanje poslovnih ciljev	1. Poslovno razumevanje
	2. Izbira in kreiranje nabora podatkov	2. Specifikacija problema	2. Razumevanje podatkov	2. Priprava podatkov	2. Razumevanje podatkov
	3. Pred-obdelava in čiščenje podatkov	3. Raziskovanje podatkov	3. Priprava podatkov	3. Podatkovno rudarjenje	3. Priprava podatkov
	4. Transformacija podatkov	4. Pridobivanje znanja domene	4. Podatkovno rudarjenje	4. Analiza rezultatov	4. Modeliranje
	5. Izbira ustreznih tehnik podatkovnega rudarjenja	5. Pred-obdelava podatkov	5. Evalvacija pridobljenega znanja	5. Asimilacija znanja	5. Evalvacija
	6. Izbira algoritma podatkovnega rudarjenja	6. Identifikacija metodologije	6. Uporaba odkritega znanja		6. Uvajanje
	7. Uporaba algoritma podatkovnega rudarjenja	7. Odkrivanje vzorcev			
	8. Evalvacija	8. Obdelava pridobljenega znanja			
	9. Uporaba odkritega znanja				

Vir: Cios in drugi (2007).

2.2 Podatkovno rudarjenje (ang. »Data mining«)

2.2.1 Opredelitev podatkovnega rudarjenja

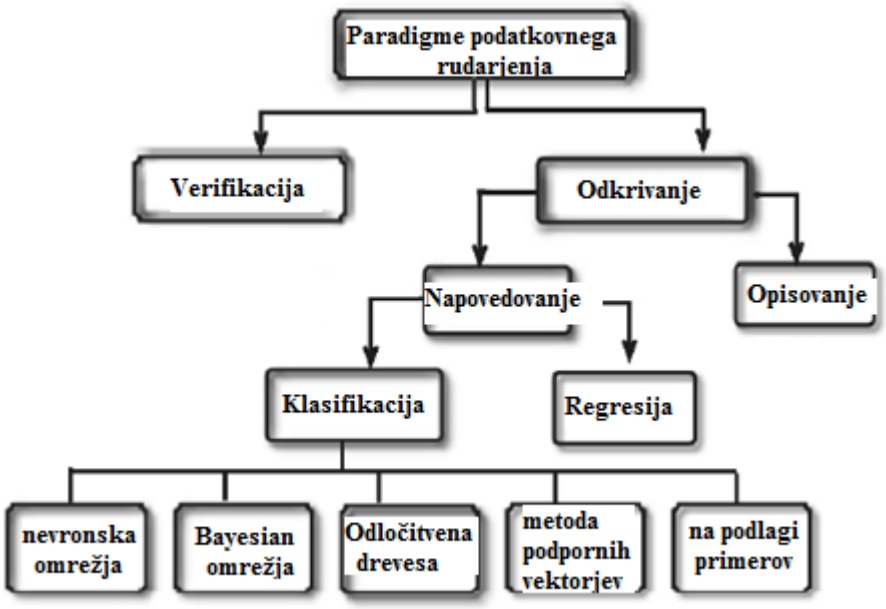
Bistvo podatkovnega rudarjenja je v osmišljanju velike količine največkrat nenadzorovanih podatkov v neki domeni oziroma področju. Lastniki podatkov navadno razumejo podatke, ki jih posedujejo in način, na kateri zbirajo podatke. Poslovneži so največja skupina, ki uporablja podatkovno rudarjenje in rutinsko zbirajo ogromne količine podatkov ter imajo interes za osmišljanje le-teh. Njihov cilj je, da njihova podjetja postanejo čim bolj konkurenčna in dobičkonosna. Tako lastniki podatkov želijo poleg splošnega razumevanja podatkov tudi pridobiti novo znanje s pomočjo le-teh v neki domeni, in sicer z namenom reševanja problemov ali odkrivanja boljših načinov poslovanja (Cios in drugi 2007).

Ciosova (2007) definicija izpostavi najprej pojem osmišljanja, kar ima različen pomen glede na izkušenost uporabnika. Vsekakor mora biti pridobljeno znanje razumljivo, veljavno, novo in uporabno. Najpomembnejše je, da je novo znanje razumljivo lastnikom podatkov, ki ga želijo uporabiti za neko korist. Drugi pojem je pojem velike količine podatkov. Pri podatkovnem rudarjenju gre vedno za analiziranje velike in ne majhne količine podatkov, kar lahko opravimo z nekimi standardnimi metodami ali manualno. Gre torej za analizo tako velike količine podatkov, ki je ne more ne človek ne najboljši algoritem analizirati brez pomoči posebnih tehnik podatkovnega rudarjenja. Prav tako gre večinoma za nenadzorovane podatke. Zbiranje nenadzorovanih podatkov je lažje in cenejše, a je bolj problematično. To rešujemo z uporabo algoritmov za iskanje naravnega razvrščanja v skupine in povezav v podatkih. Zadnji pojem je domena. Uspeh podatkovnega rudarjenja je odvisen od dostopnosti znanja domene oziroma znanja na nekem področju. Ključno je sodelovanje s strokovnjaki področja raziskave. Odkrivanje novega znanja je visoko iterativen in interaktiven proces, in ne moremo prevzeti določenega sistema podatkovnega rudarjenja za neko drugo domeno in pričakovati dobrih rezultatov (Cios 2007).

2.2.2 Tehnike podatkovnega rudarjenja

Obstaja veliko različnih metod podatkovnega rudarjenja, ki se uporabljajo za različne namene in cilje. To raznolikost metod, njihovo medsebojno povezanost in grupiranje imenujemo tudi taksonomija podatkovnega rudarjenja. Razlikovati moramo med dvema vrstama podatkovnega rudarjenja: podatkovno rudarjenje usmerjeno v preverjanje uporabnikovih hipotez ter podatkovno rudarjenje usmerjeno v odkrivanje novih pravil in vzorcev (glej Slika 2.5) (Maimon in Rokach 2005).

Slika 2.5: Taksonomija podatkovnega rudarjenja



Vir: Maimon in Rokach (2005).

Raziskovalne metode avtomatično identificirajo vzorce v podatkih, ter se nadalje delijo na metode predpostavljanja in metode opisovanja. Metode opisovanja so usmerjene na interpretacijo podatkov, ki se fokusira na razumevanje povezav med podatki. Metode predvidevanja težijo k avtomatični izgradnji vedenjskega modela, ki pridobiva nove, neznane primere, ter so sposobne predvideti vrednosti spremenljivk povezanih s primerom. Prav tako razvijajo vzorce, kateri tvorijo odkrito znanje tako, da je razumljivo in enostavno za uporabo. Nekatere metode predvidevanja oziroma napovedovanja prav tako pomagajo pri zagotavljanju

razumevanja podatkov. Večina raziskovalnih metod podatkovnega rudarjenja temelji na induktivnem učenju (model je eksplicitno ali implicitno zgrajen s posploševanjem iz zadostnega števila učnih primerov). Temeljna predpostavka induktivnega pristopa je, da se vzpostavljen model lahko aplicira za prihodnje nove primere.

Metode preverjanja ali verifikacije se ukvarjajo z evalvacijo predlagane hipoteze. Sem spadajo najpogostejše metode tradicionalne statistike kot na primer test skladnosti, test hipotez (t-test) in analiza variance (ANOVA). Le-te so manj povezane s podatkovnim rudarjenjem kot raziskovalne metode, saj je večina problemov podatkovnega rudarjenja osredotočena na odkrivanje hipotez (med velikim naborom hipotez) in ne na testiranje ene same. Ocenjevalni modeli tradicionalnih statističnih metod so v nasprotju s ključnim ciljem podatkovnega rudarjenja, i.e. identifikacijo in konstrukcijo modela temelječega na dokazih (Maimon in Rokach 2005).

V terminologiji strojnega učenja se metode predpostavljanja nanašajo na *nadzorovano učenje* (ang. »supervised learning«), ki je nasprotno nenadzorovanemu učenju. Nenadzorovano učenje se najbolj nanaša na tehnike, ki grupirajo primere brez vnaprej določenih odvisnih atributov in tako ta izraz zajema le majhen del opisnih metod (metode razvrščanja zajema, metod vizualizacije pa ne). Nadzorovane metode so tiste, ki skušajo odkriti povezavo med vhodnimi atributi (včasih imenovani kot neodvisne spremenljivke) in ciljnimi atributi (včasih definirani kot odvisne spremenljivke). Povezava med atributi je predstavljena kot struktura, ki postane model. Modeli opisujejo in razlagajo pojave, ki se skrivajo v podatkovnih nizih in se uporabljajo za predvidevanje vrednosti odvisnih spremenljivk ob znanih vhodnih spremenljivkah. Nadzorovane metode se lahko implementirajo na različnih področjih, kot so na primer marketing, finance in proizvodnja. Razlikujemo dva nadzorovana modela: klasifikacijski in regresijski model (Maimon in Rokach 2005).

V nadaljevanju bomo na kratko predstavili kriterije, ki se najpogosteje uporabljajo za evalvacijo metod oziroma modelov podatkovnega rudarjenja.

Kappa statistika (ang. »Kappa Statistics«) se pogosto uporablja za preverjanje zanesljivosti testov. Računanje temelji na razlikovanju med tem, koliko je dejanskega ujemanja (opazovano) v primerjavi s tem, kolikšno je slučajno ujemanje (pričakovano). Opazovano ujemanje (ang.

»observed agreement«) je odstotek evalvacijskega ujemanja. Kappa koeficient meri standardizirano razliko v razponu od -1 do 1, kjer vrednost 1 pomeni odlično ujemanje, 0 ravno pričakovano vrednost, negativne vrednosti pa kažejo, da je ujemanje manj možno (Viera in Garrett 2005).

Izračun opravimo s pomočjo naslednjih formul (Viera in Garrett 2005):

		Opazovanje 1 rezultati		
		Da	Ne	skupaj
Opazovanje 2	Da	*a*	*b*	m_1
Rezultati	Ne	*c*	*d*	m_0
	skupaj	n_1	n_0	*n*

$$P_e = \left[\left(\frac{n_1}{n}\right) * \left(\frac{m_1}{n}\right)\right] + \left[\left(\frac{n_0}{n}\right) * \left(\frac{m_0}{n}\right)\right]$$

Kappa statistika: $\kappa = \frac{(p_0 - p_e)}{1 - p_e}$

Kappa statistiko interpretiramo s pomočjo naslednje skale (Viera in Garrett 2005):

	Slabo	*Rahlo*	*Pošteno*	*Zmerno*	*Znatno*	*Skoraj popolno*
Kappa	0.0	0.2	0.4	0.6	0.8	1.0

Povprečna absolutna napaka (ang. Mean absolute error – MAE«) je metrika za ocenjevanje modela. Povprečna absolutna napaka modela glede na testni niz je povprečje absolutnih vrednosti posameznih napovednih napak za vse primere v testnem nizu. Vsaka napovedna napaka je razlika med dejansko vrednostjo in predvideno vrednostjo za primer.

Formula za izračun je

$$MAE = \sum_{i=1}^{n} \frac{abs\,|y_i - \lambda(x_i)|}{n}$$

kjer je y_i dejanska ciljna vrednost za testni primer x_i , $\lambda(x_i)$ je predvidena ciljna vrednost za testni primer x_i, in n je število testnih primerov (Sammut in Webb 2011).

Relativna absolutna napaka (ang. »Relative absolute error – RAE«) je relativna meritev za enostavno napovedno vrednost y_i. RAE skupno absolutno napako normalizira tako, da jo razdeli na povprečni absolutno napako (Kulkarni 2012).

$$RAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{\sum_{i=1}^{n}|y_i - \overline{y_i}|}$$

V enem izmed naslednjih poglavij bomo opravili pregled aplikacij tehnik podatkovnega rudarjenja za diagnosticiranje in predvidevanje preživetja raka dojke. V nadaljevanju se bomo tako podrobneje dotaknili metod, ki se bodo izkazale za najbolj primerne in najpogosteje uporabljane za tovrstno analizo ter bomo te metode tudi uporabili v praktičnem delu magistrske naloge.

3 Zdravstveni podatki (ang. »Medical data«)

Mnogi raziskovalci, ki izvajajo podatkovno rudarjenje na različnih področjih, se včasih ne zavedajo omejitev in težav pri obdelavi posebnih podatkov. Zdravstveni podatki so eni izmed teh in je pri njihovi analizi treba upoštevati da so heterogeni in zasebno-občutljivi. Potrebno je upoštevati tudi etični, varnostni in legalni vidik rudarjenja zdravstvenih podatkov. Med vsemi biološkimi podatki je ravno zdravstvene podatke ljudi najtežje analizirati in na njih opraviti rudarjenje, ampak po drugi strani je le-to najbolj obrestujoče (Cios in Moore 2002).

3.1 Heterogenost zdravstvenih podatkov

Zdravstveni podatki se vsakodnevno zbirajo na različne načine: iz različnih slik, pogovorov s pacienti, laboratorijskih izvidov ter s samim opazovanjem in interpretacijami. Teh podatkov je ogromno in so izredno heterogeni, vse pa je potrebno upoštevati za pravilno in uspešno diagnozo, prognozo in zdravljenje pacientov (Cios in Moore 2002).

3.1.1 Kompleksnost in obseg zdravstvenih podatkov

Pri podatkovnem rudarjenju in postavljanju diagnoze, ter pri napovedovanju in zdravljenju pacientov moramo upoštevati, da so surovi zdravstveni podatki heterogeni in jih je ogromno. Eno izmed najpriljubljenejših diagnostičnih orodij je zagotovo slikanje. Zaradi tega je pomembno, da obstaja učinkovita metoda za zbiranje podatkov na podlagi slik, kar pa je veliko bolj zahtevno kot analiziranje čistih numeričnih podatkovnih baz. Slikovne tehnike (na primer SPECT, MRI, PET, zbiranje EKG in EEG signalov) zbirajo ogromno gigabajtov podatkov na dan. Vsem tem slikam človeških organov pa so vedno priloženi tudi drugi klinični podatki, kot na primer zdravnikova interpretacija (diagnoza, klinični vtis). Takšna raznolikost podatkov zahteva visoko-kapacitetne naprave za zbiranje podatkov in nova orodja za analiziranje vseh teh podatkov. Čeprav je ljudem lažje interpretirati slike in smo zmožni prepoznavati neke vzorce in trende v teh podatkih in tako postaviti neko diagnozo, je še vedno težavna ogromna količina podatkov. Vsi ti shranjeni podatki postanejo neuporabni, če niso dostopni v enostavno

razumljivi obliki. Tukaj imajo veliko vlogo vizualizacijske tehnike, saj so slike ljudem najlažje razumljive in zagotavljajo ogromno informacij na enem samem posnetku (Cios in Moore 2002).

3.1.2 Zdravniška interpretacija zdravstvenih podatkov

Zdravniška interpretacija slik, signalov in ostalih kliničnih podatkov je v obliki zapisa, kar pa je zelo težko standardizirati in tudi uporabiti pri podatkovnem rudarjenju. Do težav prihaja celo na istem področju, kjer strokovnjaki znotraj iste discipline uporabljajo različne sinonime za opis ene same bolezni in različne slovnične konstrukcije za opis povezav med medicinskimi subjekti. Kot eden izmed načinov reševanja teh težav je predlagano računalniško prevajanje (ang. »computer translation«). Strojno prevajanje je navadno sestavljeno iz treh korakov: analiziranje izvirnega jezika stavka, prevajanje iz enega v drug jezik in generiranje stavka v ciljnem jeziku. V vseh jezikih imamo velik nabor različnih izrazov in je potrebno zbrati vse te izraze v slovar, ki ga pri prevajanju uporabljamo. Dela pri tem pa je seveda neskončno, kar predstavlja veliko težavo. Druga težava je ta, da vhodni stavek nikoli ni edinstven. Poleg tega predstavlja težavo tudi prevajanje dolgih stavkov (Cios in Moore 2002).

3.1.3 Analiza občutljivosti in specifičnosti

Mnoge diagnoze in zdravljenja v medicini so neprecizni in so predmet stopnje napake. Stopnja napake se meri z analizo občutljivosti in specifičnosti. V medicini je potrebno razlikovati med testom in diagnozo. Test je ena izmed mnogih vrednosti za določanje zdravstvenega stanja pacienta, medtem ko je diagnoza sinteza več testov in opazovanj, ki opisujejo pato-pshihološke procese v pacientu. Tako testi kot tudi diagnoze so del analize občutljivosti in specifičnosti (Cios in Moore 2002).

Pri analizi občutljivosti in specifičnosti dobimo testne rezultate in neodvisne meritve resnice oziroma hipoteze. *Natančnost testa* primerja, kako blizu je nova testna vrednost vrednosti, ki jo predvidevamo, če velja neki določeni pogoj in po tem vsem velja neko pravilo (Cios in Moore 2002).

Natančnost testa je definirana kot: $natančnost = \frac{TP}{total}\ 100\%$

kjer TP pomeni *true positive*, ki določa število točno prepoznanih testnih primerov, total pa je skupno število testnih primerov. Takšno merjenje je zelo uporabljano pri strojnem učenju in pri prepoznavanju vzorcev, ni pa sprejemljivo v medicini, saj prikriva bistvene detajle doseženih rezultatov. V primeru, da imamo le dva možna izida testa (pozitivni ali negativni) in da postavimo pozitivno ali negativno hipotezo, potem učinkovitost generiranih pravil merimo s tremi evalvacijskimi kriteriji. True positive (TP) se nanaša na število pravilnih pozitivnih predvidevanj; true negative (TN) je število pravilnih negativnih predvidevanj; false positive (FP) je število napačnih pozitivnih predvidevanj; in false negative (FN) je število napačnih negativnih predvidevanj (Cios in Moore 2002).

Tri meritve učinkovitosti so:

$$občutljivost = \frac{TP}{pozitivna\ hipoteza}\ 100\% = \frac{TP}{TP + FN}\ 100\%$$

$$specifičnost = \frac{TN}{negativna\ hipoteza}\ 100\% = \frac{TN}{FP + TN}\ 100\%$$

$$predvidena\ natančnost = \frac{TP + TN}{total}\ 100\% = \frac{TP + TN}{TP + TN\ + FP + FN}\ 100\%$$

Občutljivost meri zmožnost testa, da je pozitiven, kadar je izpolnjen nek pogoj, ali koliko pozitivnih testnih primerov je prepoznanih. Lahko rečemo, da z občutljivostjo merimo, kako pogosto najdemo tisto, kar iščemo in je podobno kot lažno-negativna stopnja, napaka II. vrste, β napaka, alternativna hipoteza, ali napaka opustitve (Cios in Moore 2002).

Specifičnost meri zmožnost testa, da je negativen, kadar pogoj ni izpolnjen, ali koliko negativnih testnih primerov je vključenih. Specifičnost meri, kako pogosto je tisto, kar najdemo, dejansko to, kar smo iskali. Približni sinonimi za to so napačno-pozitivna stopnja, preciznost oziroma natančnost, napaka I. stopnje, α napaka, ničelna hipoteza, ali napaka izvedbe dejanja. Predvidena natančnost poda skupno oceno. Visoko stopnja zaupanja je lahko določena samo za rezultate, ki podajo visoke vrednosti za vsa tri merjenja (Cios in Moore 2002).

Pri analizah občutljivosti in specifičnosti je potrebno oblikovati vprašanja, na katera se odgovarja z da ali ne. To niti ni tako enostavno, saj se srečamo z mnogo težavami – pride lahko na primer do podajanja pričakovanih odgovorov, namesto da dobimo pravilne odgovore. Prisotna je velika nepripravljenost za uporabo teh meritev za določanje napak, saj rezultati niso prikazani zelo prepričljivo in primerni za objavo, prav tako pa je sama analiza včasih zelo obremenjujoča, cenovno draga in včasih v raziskavi nenatančno ocenjuje vsak primer posebej (Cios in Moore 2002).

3.1.4 Slaba matematična karakterizacija zdravstvenih podatkov

Poleg že naštetih je ena izmed unikatnih značilnosti zdravstvenih podatkov ta, da so le-ti v primerjavi z ostalimi področji, najmanj matematično značilni. Tako na primer fiziki zbirajo podatke in le-te vstavijo v formule, enačbe in modele, ki potem odražajo povezave med temi podatki. Po drugi strani je konceptualna struktura medicine sestavljena iz besednih opisov in slik. Fundamentalni subjekti v medicini kot so na primer vnetja, ishemije ali neoplazije, so enako realni kot dolžina, teža in sila v fiziki, a v medicini ni primerjalne formalne strukture, v kateri bi lahko podatkovni raziskovalec organiziral informacije, da bi jih lahko modelirali z razvrščanjem v skupine, regresijskim modelom in sekvenčno analizo. Vse to je razumljivo, saj se medicina spopada z stotinami različnih anatomskih področij in tisočimi bolezni. Prav tako velja, da se logika medicine bistveno razlikuje od logike fizikalnih ved. S pomočjo vedno hitrejših računalnikov in novejših, naprednejših orodij za podatkovno rudarjenje in odkrivanje zakonitosti pa se ti problemi rešujejo (Cios in Moore 2002).

3.1.5 Kanonična oblika zdravstvenih podatkov

V matematiki je kanonična oblika zaželena in povzema vse ekvivalentne oblike istega koncepta. Kanonična oblika je na primer za eno polovico ½, algoritem za zmanjšanje neskončnosti ekvivalentnih zapisov pa je 2/4, 3/6, 4/8, 5/10, vse do ½. Mnoge intelektualne discipline so sprejele sporazum po kanonični obliki. Že v 18. stoletju se je kanonična oblika izpostavila kot nujna za doslednost pravopisa pri pripravi slovarja (Cios in Moore 2002).

Na področju medicine pa na žalost ne moremo doseči kanonične oblike zapisa niti za osnovne pojme. Kot primer lahko vzamemo "adenokarcinom kolona, metastatski na jetrih". Tukaj nimamo konsistentne oblike za izražanje in vsaka zdravstvena beseda ima edinstven zapis in pomen, a obstaja veliko različnih izrazov, ki so popolnoma zdravstveno ekvivalentni, na primer:

- adenokarcinom debelega črevesa, metastatski na jetrih,
- adenokarcinom debelega črevesa, z metastazo na jetrih,
- adenokarcinom kolona, metastatski na jetrih,
- adenokarcinom debelega črevesa, z okvaro na jetrih.

V medicini obstajajo še veliko bolj kompleksi izrazi. To, da ni kanonične oblike na medicinskem področju, prinaša največ težav ravno pri indeksiranju in statistikah (Cios in Moore 2002).

3.2 Etični, pravni in družbeni vidik zdravstvenih podatkov

Pri zdravstvenih podatkih gre za zbiranje podatkov na ljudeh, zato obstaja ogromno etičnih in legalnih tradicij za preprečevanje zlorabe bolnikov in njihovih podatkov (Cios in Moore 2002).

Pri podatkovnem rudarjenju na zdravstvenem področju se srečujemo z vprašanjem o lastništvu podatkov. V pravu je lastništvo definirano s tem, da je lastnik pooblaščen za prodajo določene lastnine. V medicini je vprašanje lastništva zdravstvenih podatkov nejasno, saj je tudi samo prodajanje podatkov o ljudeh nekako neprimerno. Količina zdravstvenih podatkov primernih za podatkovno rudarjenje je ogromna – samo v Severni Ameriki in v Evropi je ustvarjeno na tisoče terabajtov podatkov letno. Ti podatki se shranjujejo v heterogenih podatkovnih bazah in so razpršeni glede na vzpostavljeno zdravstveno nego brez neke skupne oblike in principov organizacije shranjevanja. Vprašanje lastništva teh podatkov je nejasno – ali posamezni bolniki posedujejo podatke, ki so zbrani o njih, ali pa so lastniki teh podatkov njihovi zdravniki oziroma njihove zdravstvene zavarovalnice? Nekatere zdravstvene zavarovalnice ne želijo prispevati za sodelovanje bolnikov v protokolih eksperimentalnega kliničnega zdravljenja. In če zdravstvene zavarovalnice ne posedujejo zdravstvenih podatkov svojih zavarovancev, potem lahko zavrnejo

plačevanje zbiranja in shranjevanja teh podatkov. Če je zmožnost za obdelavo in prodajo zdravstvenih podatkov neprimerno, kako lahko potem nadomestiti podatkovne menedžerje, ki urejajo in obdelujejo te podatke? Bi potemtakem morali izredno bogat vir za potencialno izboljšanje človeštva pustiti neizkoriščen (Cios in Moore 2002).

Poleg vprašanja lastništva zdravstvenih podatkov je tukaj pomembno omeniti strah pred tožbami proti zdravnikom in ostalim zdravstvenim oskrbovalcem. Cios in Moore izpostavljata kot najboljši primer zdravstveno zavarovanje v Združenih državah Amerike, kjer je zdravstveno zavarovanje za okoli 30 odstotkov dražje kot v Evropi in v Kanadi. Teh 30 odstotkov naj bi v prvi meri bilo namenjenih ravno za zdravstveno legalne potrebe, torej za pravni stoške ali ta tako imenovano obrambno medicino (na primer za izvajanje testov, s pomočjo katerih se lahko obranijo pri morebitnih prihodnjih tožbah). Potemtakem je razumljivo, da zdravniki in drugi zdravstveni izvajalci neradi predajajo zdravstvene podatke svojih pacientov nekomu drugemu. Podatkovni rudar lahko pri svojem delu izbrska podatke o neželenih dogodkih, očitne nepravilnosti v zdravstveni zgodovini bolnikov pa lahko sprožijo preiskave. Pogosto se lahko zgodi, da je neka navidezna nepravilnost v resnici posledica opustitve ali napake pri vnosu podatkov in ne nujno vedno rezultat malomarnosti zdravnika ali drugega zdravstvenega izvajalca. Takšne preiskave zahtevajo ogromno časa in čustvene energije, pri tem pa so zdravstveni izvajalci izpostavljeni velikemu tveganju (Cios in Moore 2002).

Naslednja pomembna edinstvena značilnost je skrb za zasebnost in varnost zdravstvenih podatkov. Obstajajo pravila in smernice, ki določajo prekrivanje posameznih identifikatorjev bolnikov. Potrebno je preprečiti kršitev zaupnosti podatkov o bolniku, ki lahko vodi v sodne postopke, a tudi samo kršitev odnosa med zdravnikom in bolnikom, saj je slednji popolnoma iskren ravno zaradi dejstva o strogi zaupnosti podatkov. Nekatera določila za prekrivanje identifikatorjev bolnikov so nepreklicna in stroga do te mere, da včasih tudi v primeru, ko bi lahko bolnike informirali o možnih novih načinih zdravljenja, tega ne storimo, saj ne smemo pregledovati preteklih zaupnih podatkov in moramo spoštovati njihovo zasebnost. Težave se pojavljajo tudi pri zagotavljanju varnosti pri ravnanju s podatki ter s prenosom podatkov. Dostop do zdravstvenih podatkov, kjer identifikatorji bolnikov niso prekriti, smejo imeti samo pooblaščene osebe. Elektronski prenos podatkov preko internetne povezave vsekakor ni varen. Tudi za sam prenos podatkov znotraj posamezne zdravstvene ustanove iz enega oddelka v

drugega je potrebno poskrbeti za prikrivanje identifikatorjev bolnikov. Po drugi strani pa je pri zdravstvenih podatkih pomembno zagotoviti, da nimamo podvojenih podatkov o istem bolniku, prav tako pa se je včasih potrebno nanašati na originalne zdravstvene zapise, da bi preverili pravilnost ali pa zagotovili potrebne dodatne informacije. Takšne zahteve morajo seveda izvajati pooblaščene osebe, to pa ne bi bilo možno v primeru popolne anonimnosti zdravstvenih podatkov. Obstajajo štiri oblike identifikacije podatkov o bolnikih:

1. Anonimni podatki – tukaj gre za podatke, ki so bili zbrani na način, da so pri samem zbiranju informacij bili identifikatorji pacienta takoj odstranjeni (kot primer lahko podamo odvzem tkiva v času avtopsije pri bolniku z določeno boleznijo, ki služi za kontrolo v histološkem laboratoriju: bolnikovi identifikatorji niso zabeleženi in jih ne moremo nikoli pridobiti);

2. Anonimizirani podatki – to so podatki, ki so v začetku zbrani z identifikatorji bolnikov, kateri so kasneje nepreklicno odstranjeni iz kartoteke. To pomeni da ne obstaja možnost vrnitve in kasnejšega pridobivanja dodatnih informacij. Ta način se je pogosto uporabljal v preteklosti, pri tem pa obstaja možnost neželene podvojitve podatkov in tako ne moremo popraviti ali pa dodati podatkov.

3. De-identificirani podatki – podatki, ki so začetno pridobljeni z bolniškimi identifikatorji, kateri so pozneje ustrezno kodirani ali šifrirani. Bolniki so tako lahko ponovno identificirani pod pogoji, ki jih določa ustrezna ustanova (v ZDA Institutional review board, pri nas Komisija Republike Slovenije za medicinsko etiko).

4. Identificirani podatki – takšni podatki so lahko zbrani le pod nadzorom ustreznih ustanov, z upoštevanjem predpisanih smernic ter s pisnim privoljenjem bolnikov (Cios in Moore 2002).

Šifrirani podatki, ki se prenašajo preko javnega internetnega kanala, in so preneseni enkrat v neko bazo podatkov, so dokaj varni pred napadalci. Podatki iz ene ustanove, v kateri prihaja do več posodobitev podatkovne baze v nekem časovnem obdobju pa so manj varni pred določenimi napadalci (Berman 1996). V namen zaščite podatkov pred napadalci obstaja več protokolov šifriranja:

- dvojno šifriranje (ang. »double-brokered encryption«)
- OTP šifriranje z Lookup tabelami (ang. »One-time-pad encryption«)

- Javno-zasebno šifriranje (ang. »Public-private encryption«) (Schneier 1996).

Za namen zdravstvenih raziskav so najbolj ustrezni in najmanj tvegani de-identificirani podatki bolnikov. V primeru, da raziskovalec upošteva le podatke, ki so bili zbrani v času diagnoze in zdravljenja določene bolezni, in ni sprememb in zahtevanja dodatnih podatkov o bolnikih (s poskusom komuniciranja z bolniki z namenom spraševanja), potem je tukaj edino možno tveganje izguba zaupanja do bolnika. Pri takšnem načinu raziskovanja (de-identificiranih) zdravstvenih podatkov pravimo, da gre za podatke minimalnega tveganja, ki jih lahko uporabimo v raziskavah le ob odobritvi ustrezne institucije (v Sloveniji Komisija Republike Slovenije za medicinsko etiko) (Cios in Moore 2002).

Vsaka uporaba de-identificiranih zdravstvenih podatkov mora biti upravičena in odobrena s strani Komisije za medicinsko etiko in mora imeti neke pričakovane koristi. Pravno in etično ne moremo izvajati nobenih podatkovnih analiz zlonamerno ali iz neresnih namenov. Internet je še vedno najbolj dostopna, najcenejša in najbolj konvencionalna oblika za distribucijo podatkov. Na ta način najbolje dostopajo do podatkov vsi tisti ki imajo legitimne razloge za njihovo uporabo (razne interesne skupine, zdravstvene skupine in raziskovalne skupine z nekonvencionalnimi znanstvenimi perspektivami), a je na žalost na ta način dostop omogočen tudi za zlonamerno uporabo. Na vprašanje, kako rešiti to težavo, še vedno iščemo odgovor (Cios in Moore 2002).

Za zagotavljanje varnosti in zasebnosti bolnikov obstajajo različne smernice, administrativna pravila in postopki, ki običajno niso zahtevani za ne-zdravstvene podatke. Morajo obstajati pravila za ocenjevanje in potrjevanje, da so v raziskovalni instituciji upoštevani ustrezni varnostni ukrepi. Med ustrezno organizacijo in zunanjimi uporabniki podatkov mora obstajati pravno veljavna pogodba, ki omogoča dostop do (individualno prepoznavnih) zdravstvenih informacij in zahteva zaščito podatkov s strani zunanje stranke. Za podatke mora obstajati načrt v primeru nesreč, kar vključuje varnostno kopiranje podatkov in načrt za odpravo posledic nesreč. Mora obstajati sistem kontroliranja dostopa do podatkov, ki vključuje avtorizacijo, vzpostavitev in modifikacijo pravic za dostop do podatkov. Potreben je stalni notranji pregled pravic za dostop do podatkov, da bi preprečili morebitne varnostne kršitve. Organizacija mora zagotoviti nadzor osebja, ki opravlja tehnično-vzdrževalne aktivnosti z namenom urejanja

zapisov dostopnosti, da zagotovimo, da imajo le-ti primeren dostop do podatkov. Prav tako moramo zagotoviti, da so uporabniki podatkovnega sistema ustrezno usposobljeni in seznanjeni o sistemski varnosti. Natančno morajo biti določeni postopki, ki se izvajajo v primeru, ko nekdo od zaposlenih izgubi dostop do podatkov. Zagotovljena morajo biti usposabljanja za varnost podatkov za vse zaposlene, vključno z usposabljanjem za ozaveščanje, periodičnimi varnostni opomniki, izobraževanjem uporabnikov v zvezi z zaščito pred virusi in s pomembnostjo prijavljanja neskladij in napak pri prijavi v sistem ter o upravljanju gesel. Vsa ta in mnoga druga pravila ovirajo raziskovalce pri podatkovnem rudarjenju (Cios in Moore 2002). Raziskovalci morajo previdno razmisliti o zaznani potrebi o določenih podatkih (na primer poštna številka bivališča bolnika), saj lahko nekateri podatki v kombinaciji z drugimi podatki razkrijejo oziroma ponovno identificirajo podatke (Sweeney 2001).

3.3 Statistika zdravstvenih podatkov

Splošno znano je, da se osnovne predpostavke metod podatkovnega rudarjenja in posebej statistike bistveno razlikujejo v primeru zdravstvenih podatkov. Človeška medicina je primarno aktivnost za skrb bolnikov, in samo sekundarno velja kot raziskovalni vir. V osnovi se podatki v medicini zbirajo z namenom, da nam le-ti prinesejo koristi na individualni ravni bolnikov. Nekateri bolniki se strinjajo, da se jih vključi v raziskovalne projekte, od katerih nimajo neposredne koristi, a takšnih podatkovnih zbirk je zelo malo, ter so ozko usmerjena in strogo urejena s pravnimi in etičnimi vidiki.

Klasična statistika je zasnovana na ideji ponovljivosti eksperimenta z vnaprej določenimi pravili. Sredi eksperimenta ni smiselno spreminjati pravil, saj bi na ta način postale formule in porazdelitve nesmiselne. Glede na to dejstvo vidimo, da so klasični statistični testi v medicini lahko predmet prekinitve. Intelektualna paradigma klasične statistike je odvisna ne samo od dejanskih zbranih številk, ampak tudi od predpostavk na samem začetku statistične raziskave. Če nekdo spremeni mnenje in s tem predpostavko sredi raziskave, potem s tem škoduje interpretaciji podatkov, čeprav se nobena izmed opazovanih vrednosti ne spremeni. V teoriji se

klinične raziskave načrtujejo z vnaprej določeno ničelno hipotezo in vnaprej določeno velikostjo vzorca, testiranje pa se izvaja, dokler dogovorjena velikost vzorca ni dosežena. Raziskava se ne sme prekiniti, ko nekdo doseže številčne vrednosti za statistično značilnost, in sicer dokler se ne doseže tudi vnaprej določeno velikost vzorca. To velja zato, ker matematično sklepanje interpretira eksperimentalne rezultate, ki temeljijo na prvotnem eksperimentalnem načrtu in na to nanašajočih se pričakovanjih (te imenujemo priorne obrazložitve). Ni možno preoblikovati osnovnih predpostavk statističnega testa med tem, ko je eksperiment že v teku. Ta paradigma ustvarja dilemo, saj se lahko zgodi, da imamo prepričljive dokaze o zmoti pri priornih predpostavkah, in to veliko prej, kot je dosežena vnaprej določena velikost vzorca, te napačne predpostavke pa škodujejo bolnikom (Cios in Moore 2002).

Pod temi zgoraj opisanimi pogoji so bile prekinjene mnoge večje in pomembnejše medicinske raziskave vključno z na primer kemoterapijo raka prostate (US Veterans Administration Co-operative Urological Research Group 1967), kemoterapijo raka dojke (Mansour in drugi 1989), oralnim hipoglikemičnim zdravljenjem odraslih sladkornih bolnikov (Goldner in drugi 1971) in steroidnim zdravljenjem cistične friboze (Ali in drugi 2000).

Podobne težave s prekinitvijo se pojavljajo tudi pri orodjih za podatkovno rudarjenje kot so na primer nevronske mreže s paradigmo testnih množic. Velja pravilo, da ne smemo uporabiti enakih opazovanj, ampak moramo izvajati teste na način, da uporabljamo elemente iz različnih testnih množic. V nasprotnem primeru gre za goljufanje pri izvajanju testa. Pri zdravstvenih podatkih se pojavi problem, saj ne moremo ustvarjati novih testnih množic, vendar moramo etično zagotoviti uporabo istih opažanj in stališč vedno znova. Podobna težava je, da različni rezultati pri statistični porazdelitvi niso popolnoma naključni, saj jih omejuje dejstvo, da so kombinacije zdravstvenih dogodkov običajne in pogoste ali pa redke. Ti dogodki so zdravnikom znani kot pogosti ali redki, vendar natančne verjetnosti za njih niso znane. Pojavi se vprašanje, ali lahko določimo statistično značilnost rezultata, če naravna značilnost bolezni in verjetni dogodki v tej hipotezi niso bili uporabljeni v ničelni hipotezi. Poleg tega je pri statističnih evalvacijah med konkurenčnimi medicinskimi hipotezami zahtevana poštenost, kar je naslednji problem. Nekateri statistični testi se ne načrtujejo za odkrivanje neke resnice, ampak z namenom iskanja zmagovalca. Tukaj gre za to, da mnogi zdravstveni raziskovalci, ki želijo dobiti nepovratna sredstva za raziskovanje, tipično formulirajo svoje ideje v skladu s tem,

kar je pravično in dobro za njih, in ne v skladu s tem, kar je dejansko pošteno in potrebno za samo raziskavo. Za reševanje te težave bi bilo primerno v raziskave vključiti in zaposliti filozofe, teoretične statistike in podatkovne analitike (Cios in Moore 2002).

Podatkovno rudarjenje ima veliko skupnega s statistiko, saj gre v obeh primerih za odkrivanje neke strukture v podatkih, a v veliki meri črpa iz številnih drugih disciplin, predvsem iz strojnega učenja in tehnologije podatkovnih baz. Od statistike se razlikuje v tem, da se pri podatkovnem rudarjenju ukvarjamo s heterogenimi podatkovnimi polji, in ne samo s heterogenimi števili kot pri statistiki. Heterogenost podatkov najbolje ponazorimo z zdravstvenimi podatki, ki vsebujejo slike npr. CT, signale kot EKG, klinične informacije kot so temperatura, nivo holesterola, analize urina, itd., kot tudi zdravniške interpretacije zapisane v nestrukturiranem jeziku. Uspeh podatkovnega rudarjenja leži v napredku tehnologije podatkovnih baz (Cios in Moore 2002).

V nadaljevanju bomo predstavili še nekaj unikatnih značilnosti zdravstvenih podatkov.

Zaradi obsežnosti in heterogenosti zdravstvenih podatkov je malo verjetno zagotoviti uspešnost orodij podatkovnega rudarjenja pri obdelavi teh surovih podatkov. Orodja zahtevajo ekstrahiranje vzorca iz podatkovne baze v upanju, da so tako pridobljeni rezultati reprezentativni za celotno bazo podatkov. Redukcijo dimenzionalnosti lahko dosežemo z vzorčenjem pri zdravstvenih zapisih enega pacienta (zapisi so izbrani naključno in uporabljeni pri podatkovnem rudarjenju) ali z vzorčenjem značilnosti v prostoru, kjer so izbrane samo določene značilnosti pri vseh podatkovnih zapisih.

Baze zdravstvenih podatkov se konstantno posodabljajo, na primer z dodajanjem novih slik CT slik (za obstoječega ali novega pacienta) ali z zamenjavo obstoječih slik (na primer zamenjava zaradi tehničnih težav ipd.). To zahteva primerne metode, ki so zmožne postopoma posodobiti tudi že pridobljeno znanje na podlagi teh podatkov.

Zdravstveni podatki zbrani v podatkovnih bazah so pogosto nepopolni ali nenatančni, saj na primer določeni testi pri enem obisku niso izvedeni ali pa niso izvedeni zaradi drugih težav pacienta.

Pri zbiranju zdravstvenih podatkov se je zelo težko v celoti izogniti podatkovnim šumom. Tako je potrebno izbrati metode podatkovnega rudarjenja, ki so manj občutljive na šume, ali pa moramo paziti, da je količina šumnih podatkov v prihodnjih podatkih vsaj približno enaka količini v trenutni podatkovni bazi.

Pri velikih podatkovnih bazah zdravstvenih podatkov se srečujemo s težavo manjkajočih podatkov. Manjkajoča vrednost je lahko posledica tega, da le-teh nismo vnesli po nesreči, ali pa tudi namerno zaradi tehničnih, ekonomskih ali etičnih razlogov. Eden izmed načinov reševanja problema je nadomestitev manjkajočih podatkov z najbolj verjetnimi vrednostmi. Drugi način je nadomestitev manjkajočih vrednosti z vsemi možnimi vrednostmi za določen atribut. Tretji način je vmesna rešitev z določitvijo verjetnega razpona vrednosti in ne ene same najbolj verjetne vrednosti. Pri tem pa je problematično vprašanje, kako nepristransko določiti ta razpon. Problem manjkajočih podatkov je zelo pogost v zdravstvenih podatkovnih bazah, saj so zdravstveni podatki zbrani stranski produkt pri aktivnostih za skrb bolnikov, in ne v sklopu nekega organiziranega raziskovalnega protokola, kjer je možno izvršiti izčrpno zbiranje podatkov.

Nabor zdravstvenih podatkov lahko vsebuje odvečne, nepomembne ali nekonsistentne podatkovne objekte in/ali atribute. O neskladnosti podatkov govorimo, kadar isti podatek spada v več kot eno medsebojno izključujočo kategorijo. Kot primer lahko podamo povišane določene vrednosti pri nekem zdravem pacientu, kar ni verjetno, a lahko do tega pride zaradi nepozornosti pri transportu vzorcev v laboratorij. Tega pa seveda ne moremo predvidevati brez dodatnih raziskav in podatkov, kar je nepraktično za podatkovno rudarjenje.

Zelo pogosto želimo najti 'naravne' skupine v ogromnih zdravstvenih podatkovnih bazah. Objekti so podobno razvrščeni v skupine, če so si podobni (glede na neko meritev), istočasno pa se razlikujejo od objektov, razvrščenih v druge skupine. Glavna skrb pri tem je, kako vključiti zdravstveno znanje v mehanizme razvrščanja v skupine. Brez tega in vsaj delnega človeškega nadzora lahko kmalu pride do težave pri razvrščanju v skupine, ki so računalniško neizvedljive ali pa nesmiselne.

V medicini je glavni interes kreirati človeku razumljive opise zdravstvenih konceptov ali modelov. To najbolje dosežemo s pomočjo strojnega učenja, konceptualnega razvrščanja v

skupine, genetskih algoritmov in mehkih množic (ang. »fuzzy sets«), saj te metode lahko ustvarijo modele na osnovi vzročno-posledičnih pravil. Na tem področju so druge metode, kot so umetne nevronske mreže, manj zanimive (Cios in Moore 2002).

4 Rak dojke

4.1 Rak dojke

Do raka dojke pride, ko celice v dojki začnejo rasti izven nadzora. Te celice običajno tvorijo tumor, ki je pogosto viden s pomočjo rentgenskega slikanja ali pa ga zatipamo v dojki v obliki zatrdline. Tumor je maligen (rakast), če se celice lahko razrastejo v obdajajoča se tkiva ali pa se širijo (metastaze) na ostale dele telesa. Rak dojke se pojavlja najpogosteje pri ženskah, a se lahko pojavi tudi pri moških (American Cancer Society 2017).

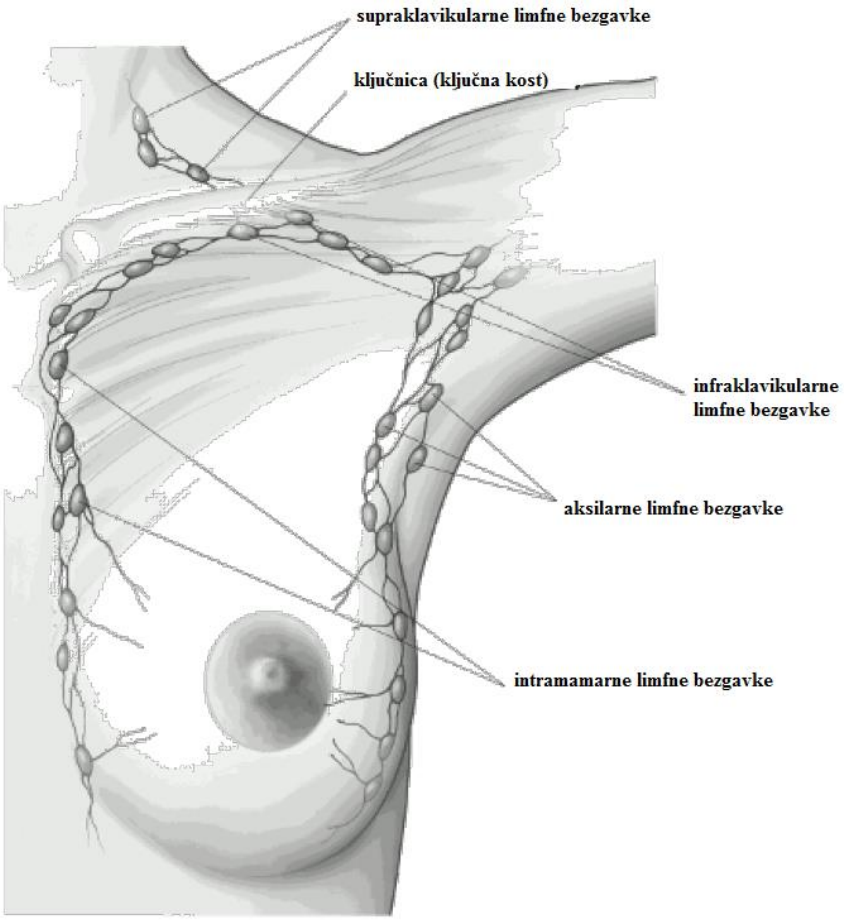
Rak dojke se lahko začne širiti iz različnih delov dojke. Večinoma se rak razvije v kanalih, kjer se pretaka mleko do bradavice (*duktalni karcinom*). Nekateri se začnejo v žlezah, ki proizvajajo materino mleko (*lobularni karcinom*). Poleg teh dveh tipov obstajajo še drugi, ki pa so manj pogosti (American Cancer Society 2017). Obstajajo še druge redkejše vrste karcinoma dojke: *sarkomi*, ki se razvijejo iz celic strome in *maligni limfomi* iz limfatičnega tkiva (Onkološki inštitut 2006).

Večina vrst raka dojke povzroči pojav zatrdlin v dojkah, a ne vse. Tako je potrebno pozornost usmeriti tudi na ostale simptome. Pomembno je tudi zavedanje, da niso vse zatrdline v dojkah kancerogene, ampak so tudi benigne. Benigni tumorji dojke so neobičajne tvorbe, ampak se ne širijo izven dojke in niso življenjsko nevarne. Nekatere benigne zatrdline pa lahko povečajo tveganje za nastanek raka dojke, zato mora vsaka sprememba na dojki biti ustrezno pregledana (American Cancer Society 2017).

Rak dojke se širi, ko rakaste celice pridejo v kri ali limfni sistem in se prenesejo na druge dele telesa. Limfni sistem je omrežje limfnih žil, ki so razvejane po celotnem telesu. Limfna odtočila ali mezgovnice prenašajo limfno tekočino in povezujejo limfne bezgavke. Limfne bezgavke ali vozli so majhne zbirke celic imunskega sistema. Limfne mezgovnice so kot majhne žile, po katerih se iz dojke prenaša čista tekočina imenovana limfa (in ne kri). Limfa vsebuje tekočino s tkivi, odpadne snovi in tudi celice imunskega sistema. Celice raka dojke lahko vstopijo v limfne mezgovnice in začnejo rasti v limfnih bezgavkah. Večina limfnih žil oziroma mezgovnic se iz dojke razveja v (Slika 4.1): limfne bezgavke pod roko (aksilarno vozlišča), limfne bezgavke okoli ključne kosti (supraklavikularne in infraklavikularne bezgavke) in limfne

bezgavke znotraj prsnega koša okoli prsne kosti (intramamarne limfne bezgavke) (American Cancer Society 2017).

Slika 4.1: Limfne bezgavke oziroma vozlišča dojke



Vir: American Cancer Society (2017).

Če se kancerogene celice razširijo na limfne bezgavke oziroma vozlišča, obstaja večje tveganje, da te celice potujejo po limfnem sistemu in se razširijo (metastazirajo) na druge dele telesa. Več kot je limfnih bezgavk z rakastimi celicami, večje so možnosti, da bomo karcinome našli tudi v drugih organih. Zaradi tega ugotovitev karcinoma v eni ali več bezgavkah vpliva na načrt zdravljenja bolnika. Razširitev le-teh po navadi ugotavljamo s kirurškimi posegi. Ni nujno, da pri vseh bolnikih rakaste celice metastazirajo, prav tako pa pri nekaterih v limfnih bezgavkah niso prisotne, a se kasneje razvijejo in širijo (American Cancer Society 2017).

Spremembe ali mutacije DNK lahko povzročijo, da normalne prsne celice postanejo kancerogene. Nekatere mutacije DNK so podedovane od staršev in lahko močno povečajo tveganje za rak na dojki. Tudi drugi dejavniki, povezani z življenjskim stilom (hrana, ukvarjanje s športom, alkohol, tobak ipd.), lahko povečajo možnosti za razvoj raka dojke, a še vedno ni ugotovljeno, kako natančno nekateri izmed dejavnikov povzročajo to, da normalne zdrave celice postanejo rakaste. Tudi hormoni igrajo pomembno vlogo pri razlogih za nastanek, a tudi ta povzročitelj ni popolnoma jasen (American Cancer Society 2017).

Najpomembnejši dejavniki tveganja za raka dojke so:

- Spol: bolezen je pogostejša pri ženskah;
- Starost: ženske, starejše od 50 let pogosteje zbolijo za rakom dojke;
- Prejšnji rak dojke: bolnice, ki so se že zdravile zaradi raka dojke, so dva- do tri-krat bolj ogrožene, da bodo ponovno zbolele (na isti dojki, če ni bila operativno odstranjena, ali pa na drugi);
- Benigne spremembe v dojki: ogroženost je največja pri tistih z atipično hiperplazijo;
- Rak dojke v družini: v primeru, da je sorodnik oziroma sorodnica prvega kolena zbolela za rakom dojke, to predstavlja dva- do tri-krat večjo ogroženost, da bo oseba tudi sama zbolela;
- Starost ob prvi menstruaciji: bolj ogrožene so ženske, ki so dobile prvo menstruacijo pred 11. letom starosti, izgubile pa so jo po 50. letu starosti;
- Rodnost, starost ob prvem porodu in število porodov: največrat zbolijo ženske, ki niso nikoli rodile in pa tiste, ki so rodile po 30. letu starosti;
- Kontracepcijske tablete in hormonski nadomestki za lajšanje menopavznih težav: oboji večajo nevarnost raka dojke;
- Debelost: predvsem po menopavzi, ker v maščevju nastajajo spolni hormoni;
- Alkohol: pri tistih, ki konzumirajo 30 do 60g alkohola, je 1.4-krat večja ogroženost kot pri abstinentih (Borštnar in drugi 2006).

Dejavniki, ki zmanjšujejo nevarnost raka dojke:

- Dojenje: če ženska doji dlje kot eno leto;

- Število porodov: če ženska rodi najmanj petkrat;
- Telesna dejavnost: pozitivno vpliva tako v mladostniškem kot tudi odraslem obdobju (Borštnar in drugi 2006).

4.2 Prognostični dejavniki pri raku dojke

Prognostične dejavnike za predvidevanje preživetja raka dojke delimo na dve skupini:

- kronološki (glede na časovno prisotnost);
- biološki (glede na potencialno obnašanje tumorja) (Bundred 2001).

Stanje limfnih bezgavk, velikost tumorja in histološka ocena so nekateri izmed napovednih dejavnikov, ki se v današnjem času uporabljajo. Stanje limfnih vozlišč je časovno odvisen dejavnik, ki je neposredno povezan s prognozo. Večje kot je število vključenih limfnih bezgavk, slabša je prognoza raka dojke. Tudi velikost tumorja je časovno odvisen dejavnik in je neposredno povezan s preživetjem raka. Preživetje je obratno sorazmerno povezano z velikostjo tumorja. Verjetnost dolgoročnega preživetja je večja v primeru manjših tumorjev kakor v primerjavi z večjimi tumorji. Patološki TMN klasifikacijski model (ang. »tumor size, number of positive regional lymph nodes and distant metastatis«) je model za oceno raka dojke na podlagi velikosti tumorja, števila pozitivnih limfnih bezgavk in oddaljene metastaze (Burke in drugi 1995). Histološka ocena je biološki dejavnik, ki temelji na treh različnih dejavnikih: mitotična stopnja, nuklearna stopnja oziroma gradus in arhitekturno morfološki izgled (Bundred 2001; Rampaul 2001). Histološka ocena (1, 2 ali 3) je izredno povezana z dolgoročnim preživetjem. Bolniki z oceno 1 imajo veliko večje možnosti za preživetje v primerjavi s tistimi, ki imajo oceno 3. Poleg teh dejavnikov so razviti tudi različni indeksi za napoved preživetja raka dojke. Najpogosteje uporabljan je NPI indeks (Nottigham prognostic index), ki je sestavljen iz kombinacije zgoraj naštetih dejavnikov.

$$NPI = TS + LS + HS$$

TS je velikost tumorja v centimetrih (ang. »tumor size«) x 0.2; LP je limfno stanje (ang. »lymph stage«), ki zavzema vrednosti 1, 2 ali 3; in HS je histološka ocena (ang. »histological stage«) z vrednostjo 1, 2 ali 3. S pomočjo NPI indeksa razvrščamo bolnike v tri napovedne skupine:

dobra, vmesna in slaba. Dokazano je, da NPI indeks zagotavlja relativno dobre napovedne vrednosti z rezultati, ki so podobni tistim, ki jih dobimo z multivariatno analizo (Bundred 2001; Rampaul 2001).

Poleg tega se uporablja tudi BCSS indeks (ang. »breast cancer severity score«). Ta indeks temelji na premeru tumorja, številu pozitivnih limfnih bezgavk, esterogenskih receptorjih in progesteronskih receptorjih. S pomočjo BCSS indeksa je možno doseči višjo stopnjo napovedne natančnosti kakor s konvencionalnim ocenjevalnim sistemom (Jimenez-Lee 2003).

4.3 Rak dojke v Sloveniji

V Sloveniji poteka redno zbiranje podatkov o raku dojke in vseh ostalih vrstah raka. Zbiranje podatkov vodi Register raka Republike Slovenije (RRRS), ki je eden izmed najstarejših populacijskih registrov raka v Evropi in je bil ustanovljen leta 1950 na Onkološkem inštitutu Ljubljana. Podatki se redno obdelujejo in so predstavljeni v obliki letnih poročil. Zadnje končano poročilo je Letno poročilo za leto 2013, saj je čas objave običajno dve do tri leta. Podrobne analize izhajajo v obliki monografij in člankov. V Sloveniji je prijavljanje raka obvezno in zakonsko predpisano od ustanovitve RRRS. Podatke o bolnikih z rakom sporočajo RRRS vse fizične osebe, ki opravljajo zdravstveno dejavnost v Sloveniji, in sicer na predvidenem posebnem obrazcu, imenovanem Prijavnica rakave bolezni (dostopna na njihovi spletni strani). Prispeli podatki na prijavnicah so ustrezno kodirani v skladu z mednarodnimi in v RRRS dogovorjenimi pravili. Podatki so dostopni slovenskim zdravnikom, raziskovalcem in širši javnosti. Za pridobitev in uporabo podatkov v raziskavi je potrebno pridobiti odobritev s strani Komisije Republike Slovenije za medicinsko etiko. Podatki RRRS so vključeni v mnoge mednarodne podatkovne zbirke in projekte. Objavljeni so v knjigi Cancer Incidence in Five Continents, zbirkah ECO, GLOBOCAN in ACCISS. Podatki o preživetju slovenskih bolnikov so obdelani tudi v mednarodnih raziskavah EUROCARE, RARECARE in EUNICE (Onkološki inštitut Ljubljana, Epidemiologija in register raka, Register raka 2016).

V Sloveniji je rak dojke najpogostejši rak pri ženskah, kar velja tudi za ostale razvite države. Obolevnost za rakom dojke se povečuje od leta 1950, Slovenija pa se z 1228 novimi bolnicami

v enem letu (povprečje za obdobje od 2009 do 2013) uvršča v sredino svetovne lestvice (Slika 4.2). Po podatkih Registra raka za Slovenijo se rak dojke redko pojavi pred 30. letom starosti, večina žensk zboli po 50. letu starosti (Borštnar in drugi 2006). V letu 2013 je za rakom dojke zbolelo 1268 bolnikov, od tega približno 1% moških. Pojavnost v zadnjih desetletjih narašča (Onkološki inštitut 2016).

Slika 4.2: Statistike raka dojke za obdobje 2009 do 2013

INCIDENCA (povprečje v obdobju 2009–2013)	**Ženske**
Število novih primerov v enem letu	1.228
Odstotek med vsemi raki (%)	20,4
Mesto po pogostnosti med vsemi raki	1.
Odstotek med vsemi raki razen kožnega (%)	25,2
Tveganje raka do 75. leta starosti (KT) (%)	7,2
Groba incidenčna stopnja na 100.000	118,5
Starostno standardizirana incidenčna stopnja (SSS) na 100.000 (W)	65,9
Ocenjeni delež letne spremembe grobe inc. stopnje zadnjih 10 let (%)	1,3
Ocenjeni delež letne spremembe SSS zadnjih 10 let (%)	0,7
UMRLJIVOST (povprečje v obdobju 2009–2013)	
Število smrti v enem letu	413
Odstotek med vsemi smrtmi zaradi raka (%)	15,9
Tveganje smrti za rakom do 75. leta starosti (KT) (%)	1,6
Groba umrljivostna stopnja na 100.000	39,9
Starostno standardizirana umrljivostna stopnja (SSS) na 100.000 (W)	15,7
Ocenjeni delež letne spremembe grobe umr. stopnje zadnjih 10 let (%)	0,4
Ocenjeni delež letne spremembe SSS zadnjih 10 let (%)	- 2,3
PREVALENCA (na dan 31. 12. 2013)	
Število živih oseb z diagnozo raka ob koncu leta 2013 (prevalenca)	15.294
Število živih oseb z diagnozo raka na 100.000 (prevalenčna stopnja)	1.471,3
1-letna prevalenca	1.220
5-letna prevalenca	4.140

Vir: Slora (2016).

5 Podatkovno rudarjenje in odkrivanje zakonitosti v zdravstvu

Tehnologija podatkovnega rudarjenja zagotavlja pristop do skritih vzorcev v podatkih. Odkrito znanje se lahko uporabi v zdravstveni administraciji za izboljšanje kvalitete storitev.

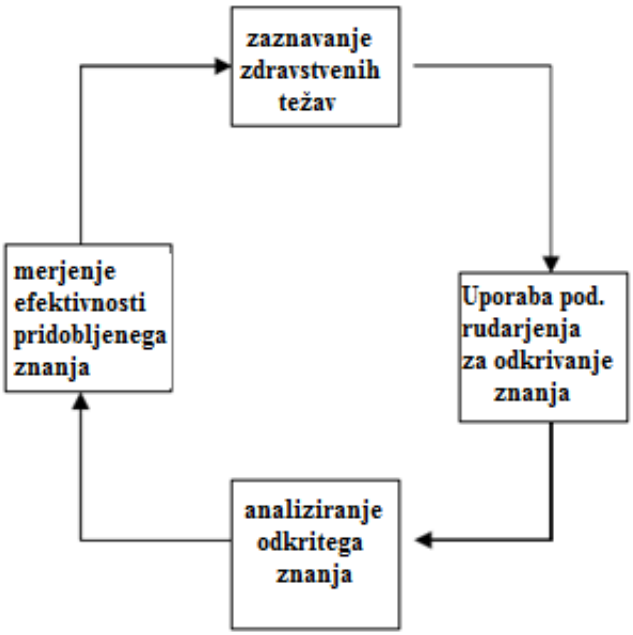
Pridobljeno znanje lahko uporabijo tudi zdravniki – na primer za zmanjšanje škodljivih učinkov zdravil, tako da predlagajo druge cenejše terapevtsko enakovredne alternative. Podatkovno rudarjenje se lahko na področju medicine uporabi na področju:

- modeliranje podatkov za zdravstvene aplikacije,
- izvršilni informacijski sistem za zdravstveno varstvo,
- napovedovanje stroškov zdravljenja,
- predvidevanje obnašanja bolnika glede na njegovo zdravstveno zgodovino,
- javna zdravstvena informatika,
- elektronsko upravljanje v zdravstvu in
- zdravstveno zavarovanje (Kaur in Wasan 2006).

V zdravstvu se tradicionalno odloča na podlagi osnovnih informacij, pridobljenih izkušnjah v preteklih primerih in glede na finančna sredstva. Vsekakor se tehnike podatkovnega rudarjenja in proces odkrivanja zakonitosti lahko uporabijo in ustvarijo podatkovno bogato zdravstveno okolje. Zdravstvene ustanove lahko implementirajo proces odkrivanja zakonitosti s pomočjo usposobljenih zaposlenih, ki dobro razumejo in poznajo zdravstveno industrijo. S pomočjo tega procesa lahko iz velike količine podatkov določimo smiselne vzorce in razvijemo strateške rešitve. Analitiki zdravstvenih podatkov lahko uporabijo primere iz drugih področij in na podlagi teh aplicirajo uporabo procesa odkrivanja zakonitosti na zdravstvenem področju. Zdravstvenih podatkov je ogromno, zdravstvene organizacije pa morajo imeti možnost analiziranja le-teh. Podatki o zdravljenju bolnikov so računalniško shranjeni, podatkovno rudarjenje pa na ta način lahko pripomore k reševanju mnogih pomembnih in kritičnih zdravstvenih vprašanj (Kaur in Wasan 2006).

Brez podatkovnega rudarjenja si je težko predstavljati popolni potencial zbranih podatkov v zdravstvu, saj je teh podatkov ogromno, so večdimenzionalni, porazdeljeni in negotovi. Ti podatki morajo biti pretvorjeni v informacije in znanje, ki bo pripomoglo h kontroli, stroškom in vzdrževanju visoke kontrole zdravstvene oskrbe pacientov (Kaur in Wasan 2006).

Slika 5.1 Proces odkrivanja zakonitosti in podatkovnega rudarjenja v zdravstvu



Vir: Kaur in Wasan (2006).

Za uspeh zdravstvenih organizacij je ključna možnost zajema, shranjevanja in analize podatkov (Slika 5.1). Sprotna analitična obdelava, OLAP (ang. »Online Analytical processing«), predstavlja enega izmed načinov za analiziranje večdimenzionalnih podatkov. Skladiščenje podatkov in analiza le-teh z OLAP orodji pripomore k boljšemu odločanju. Razširjena uporaba zdravstvenih informacijskih sistemov, vključno s podatkovnimi bazami, katere so vsak dan večje, prinaša zdravnikom in drugemu osebju težave pri (ne)uporabi teh vseh zbranih podatkov. Tradicionalne podatkovne analize niso več dovolj, s tem pa so postale metode za učinkovito računalniško podprte analize nujno potrebne (Kaur in Wasan 2006).

Tehnike podatkovnega rudarjenja so zaradi svoje napovedovalne sposobnosti pogosto uporabljane v diagnostiki in zdravstvenih aplikacijah. Algoritmi podatkovnega rudarjenja povzemajo znanje iz preteklih primerov v kliničnih podatkih in oblikujejo modele pogostih ne-linearnih povezav med odvisnimi in neodvisnimi spremenljivkami. Dobljeni modeli predstavljajo formalizirano znanje, katero pogosto zagotavlja dobro diagnostično mnenje.

Metode *klasifikacije* so najpogosteje uporabljane v zdravstvenem podatkovnem rudarjenju (Chen in drugi 2005). Dreiseitl in drugi (2001) so primerjali pet klasifikacijskih algoritmov za diagnosticiranje pigmentnih kožnih sprememb. Logistična regresija, nevronske mreže in metoda podpornih vektorjev so se izkazali kot primerljivo dobri, medtem ko sta se metodi najbližjih sosedov in odločitveno drevo izkazali kot nekoliko slabši. S pomočjo klasifikacijskih tehnik so se analizirali tudi različni signali in njihova povezava z določenimi boleznimi ali simptomi (Chen in drugi 2005). Acir in Guzelis (2004) sta uporabila metodo podpornih vektorjev pri avtomatični detekciji signala v elektroencefalografiji (EEG), ki se uporablja pri diagnosticiranju nevroloških motenj povezanih z epilepsijo. Kandaswamy in drugi (2004) so uporabili umetna nevronska omrežja za klasifikacijo pljučnih zvočnih signalov v šest kategorij za pomoč pri postavljanju diagnoze.

Podatkovno rudarjenje se uporablja tudi za odkrivanje znanja iz zdravstvenih podatkov. Kot primer lahko podamo uporabo tehnik podatkovnega rudarjenja za pridobivanje diagnostičnih pravil iz podatkov o raku dojke (Kovalerchuk in drugi 2001). Podatkovno rudarjenje je bilo aplicirano tudi na področje kliničnih podatkovnih baz z namenom prepoznavanja novih zdravstvenih znanj (Prather in drugi 1997; Hripcsak in drugi 2002).

Eno izmed zdravstvenih področij, kjer se veliko uporablja podatkovno rudarjenje, je predvidevanje raka. Mangasarian (1995) obravnava uporabo linearnega programiranja za klinično klasifikacijo pacientov z rakom dojke. Bellazzi in drugi (2001) so predstavili uporabo orodij podatkovnega rudarjenja za izpeljavo napovedovalnega modela za napovedovanje izidov raka jetrnih celic. Land (2001) opisuje novo tehnologijo nevtronskega omrežja za izboljšanje diagnoze raka dojke s pomočjo mamografije. Walter in Mohan (2000) sta predstavila algoritem, s pomočjo katerega pridemo do klasifikacijskih pravil iz usposobljenih nevronskih omrežij, ter opisala apliciranje tega pri diagnozi raka dojke. Poleg tega sta navedla, kako točnost teh omrežij in pravil, ki jih izpeljemo iz le-teh omrežij, lahko izboljšamo z enostavno pred-obdelavo podatkov. Zupan in drugi (2000) sta oblikovala shemo, ki omogoča uporabo klasifikacijskih metod za analizo preživetja zbolelih za rakom prostate. Zhang in Zhang (1999) sta razvila in uveljavila ProstA-sure, ki je algoritem nevronskega omrežja, ki analizira profile multiplih označevalcev tumorja in vrača vrednost diagnostičnega indeksa, kateri se lahko uporabi za zgodnje odkrivanje raka prostate.

5.1 Pregled aplikacij tehnik podatkovnega rudarjenja za diagnosticiranje in predvidevanje preživetja raka dojke

Klinična diagnoza pomembno pripomore k predvidevanju malignih primerov. Različne pogoste metode za diagnosticiranje raka dojke so mamografija, biopsija, tomografija in magnetna resonanca. Rezultati, ki jih pridobimo s pomočjo teh metod, se uporabljajo za prepoznavanje vzorcev, kateri pomagajo zdravnikom pri razlikovanju malignih in benignih primerov. Raziskovalci so uporabili veliko različnih tehnik podatkovnega rudarjenja za te namene (Gupta in drugi 2011).

Jerez in drugi (2005) so analizirali podatke o bolnikih z velikim tveganjem za nastanek raka dojke s pomočjo različnih metod podatkovnega rudarjenja (oziroma procesa odkrivanja zakonitosti) in s tradicionalnimi statističnimi metodami. Metode podatkovnega rudarjenja so se izkazale kot bolj uspešne pri napovedni analizi raka dojke. Razavi in drugi (2007) so primerjali učinkovitost procesa odkrivanja zakonitosti in področnih strokovnjakov za predvidevanje raka dojke. Rezultati so pokazali, da so je proces odkrivanja zakonitosti pri tem uspešnejši kot področni strokovnjaki. Thongkam in drugi (2009) so v svoji raziskavi dokazali, da je za doseganje najvišje stopnje uspešnosti procesa odkrivanja zakonitosti pri napovedovanju raka dojke nujno imeti veliko količino podatkov, kateri morajo biti visoko-kvalitetno pripravljeni za analizo.

Sarvestani in drugi (2010) so naredili primerjavo zmogljivostjo različnih nevronskih omrežij za raziskovanje diagnosticiranja in predvidevanja preživetja raka dojke. Ta raziskava je pokazala, da so statistična nevronska omrežja učinkovita pri diagnosticiranju raka dojke, saj so se mnoga aplicirana nevronska omrežja, ki so sestavljala diagnostični sistem, delovala dobro.

Anunciacao in drugi (2010) so raziskali uporabnost metode odločitvenega drevesa za detekcijo visoko-rizičnih skupin z rakom dojke. Na podlagi 164 kontrolnih podatkov in 94 primerov so s pomočjo orodja WEKA odkrili, da 13 primerov spada v visoko-rizično skupino. Rezultati so pokazali, da je možno najti statistično značilno povezanost z rakom dojke s pomočjo odločitvenega drevesa in izbiro najboljšega atributa.

Abdelaal in Frouq (2010) sta raziskovala sposobnost klasifikacijske metode podpornih vektorjev pri analiziranju podatkovnih baz za ekstrakcijo mamografskih značilnosti in starosti, ki diskriminira pravilne in napačne rezultate. Metoda podpornih vektorjev je pokazala obetajoče rezultate za povišanje točnosti diagnostične klasifikacije.

Chang Pin in Ming Der (2008) sta ugotovila, da genetski algoritemski model prinaša boljše rezultate v primerjavi z ostalimi modeli podatkovnega rudarjenja za analiziranje podatkov pacientov z rakom dojke (v smislu splošne natančnosti klasifikacije bolnikov ter izražanja in kompleksnosti klasifikacijskih pravil). V komparativni raziskavi so bile vključene metode umetnih nevronskih omrežij, odločitveno drevo, logistična regresija in genetski algoritem. Rezultati so pokazali, da je genetski algoritem zmožen proizvesti točne rezultate v klasifikaciji podatkov o raku dojke, ugotovljeno klasifikacijsko pravilo pa je sprejemljivejše in razumljivejše od ostalih.

Gandhi in drugi (2010) so v svoji raziskavi ustvarili klasifikacijsko pravilo za podatke o raku dojke in pri tem uporabili PSO algoritem oziroma optimizacijo s kolonijo mravelj (ang. »Particle Swarm Optimization«).

Padmavati (2011) je izvedel komparativno raziskavo na temo predvidevanja raka dojke, kjer je uporabil metodo umetnih nevronskih omrežij (omrežja RBF in MLP) ter logistično regresijo. Logistična regresija je bila izvedena s pomočjo statističnega programa SPSS, metode nevronskih omrežij pa z uporabo MATLAB-a. Ugotovljeno je bilo, da nevronska omrežja zahtevajo več časa kot logistična regresija, a je občutljivost in specifičnost pri obeh tipih modelov nevronskih omrežij pokazala boljšo napovedno moč. RBF (ang. »Radial basis function (RBF) neural network) ima v primerjavi z MLP (ang. »Multilayer perceptron (MLP) network models«) boljšo napovedno sposobnost in zahteva manj časa.

Sawarkar in drugi (2006) so v svoji raziskavi uporabili metodo podpornih vektorjev in metodo umetnih nevronskih omrežij za analizo podatkov raka dojke. Oba napovedna modela sta se izkazala za bolj natančna od človeških napovedi. 97-odstotna natančnost teh modelov je lahko uporabljena za odločitev o izognitvi biopsiji.

Jamarani in drugi (2005) so predstavili pristop za zgodnje diagnosticiranje raka dojke z uporabo kombinacije umetnih nevronskih omrežij in zbirke slik (mamografija in ostale slike). Pristop je testiran s pomočjo mamografskih podatkovnih zbirk in ostalih slik iz lokalnih bolnišnic, ugotovljeno pa je bilo, da ta pristop lahko pomaga radiologom pri mamografskih analizah in diagnostičnem odločanju.

Po postavljeni diagnozi raka dojke je potrebno maligne zatrdline odstraniti. Med samim postopkom morajo zdravniki določiti napoved oziroma prognozo bolezni. Gre za napoved pričakovanega poteka bolezni, kar je izredno pomembno, saj sta vrsta in intenzivnost zdravil odvisna od tega. Problem napovedi imenujemo tudi analiza preživetja ali življenjski podatki. Okrnjenost podatkov predstavlja v tem primeru večjo težavo kot pri diagnosticiranju. V podatkih imamo le nekaj primerov, kjer je opažena ponovitev bolezni. V tem primeru lahko razvrstimo bolnika kot ponovno zbolelega in imam informacijo o času ponovitve. Po drugi strani pa ponovitve bolezni ne opazimo pri večini bolnikov. Pri teh primerih nimamo neke določene točke, na kateri lahko določimo pacienta kot ne-ponovno zbolelega. Nimamo časa ponovitve bolezni in tako so podatki okrnjeni. Vse, kar v teh primerih vemo, je le čas njihovega zadnjega zdravniškega pregleda. To se imenuje čas preživetja po bolezni. Napovedovanje z predvidevanjem izida bolezni pomaga pri vzpostavitvi načrta zdravljenja. Pri napovedovanju raka dojke gre za:

- napoved občutljivosti raka dojke (ocena tveganja),
- napoved ponovitve raka dojke in
- napoved preživetja raka dojke (Gupta in drugi 2011).

Najširše sprejeti napovedni dejavnik za rak dojke je American joint Commission on Cancer (AJCC), ki predstavlja sistem, ki temelji na TNS sistemu (T-tumor, N-node (slovensko bezgavke) in M-metastaza). Za preživetje se šteje katerakoli incidenca raka dojke, kjer je oseba še vedno živa od dneva diagnoze. Cilj napovedi je obravnavati primere, kjer se rak ni ponovil (cenzurirani podatki) in pa tudi primere, kjer se je rak v določenem času ponovil. Napovedni problemi pri raku dojke nastajajo v okviru ponovitvenih klasifikacijskih problemov (Gupta in drugi 2011).

Bellaachia in Gauven (2006) sta raziskala tri različne metode podatkovnega rudarjenja, in sicer C4.5 odločitveno drevo, naivno Bayesian klasifikacijo in algoritem vzvratnega razširjenja napake (umetno nevronsko omrežje) pri predvidevanju oziroma napovedovanju preživetja raka dojke. Pri raziskovanju stopnje preživetja sta uporabila podatke o raku dojke iz podatkovne zbirke Surveillance Epidemiology and End Results (SEER). V pred-klasifikacijskem procesu sta uporabila drugačen vidik, tako da sta določila tri nove kodirane spremenljivke, s pomočjo katerih se določi ali je pacient preživel rak dojke ali pa ni preživel (in je torej vzrok smrti rak dojke). Za analizo podatkov sta uporabila orodje WEKA. Avtorja sta ugotovila, da za dane podatke v primerjavi z ostalima dvema metodama najbolje deluje C4.5 algoritem, ki je algoritem za gradnjo odločitvenega drevesa.

Lundin in drugi (1999) so ocenjevali natančnost nevronskih omrežij pri predvidevanju 5, 10 in 15-letnega preživetja raka dojke. Opravili so primerjavo 82/300 napačnih napovedi z logistično regresijo in 49/300 napačnih napovednih ocen za preživetje raka dojke z nevronskimi omrežji. Ugotovili so, da je metoda nevronskih omrežij natančnejša.

Street (1998) je uporabil klasifikacijo umetnih nevronskih omrežij za Winsconsin Prognostic Breast Cancer (WPBC) in SEER podatkovne zbirke pri analizi preživetja. Razvil je novo kodiranje (dobra in slaba prognoza cenzuriranih podatkov) v strukturi umetnih nevronskih omrežij, da bi zagotovil ustrezen okvir za prognostično napoved. Chi in drugi (2007) so uporabili ta model, ki ga je razvil Street leta 1998, za napoved raka dojke na WPBC podatkih in Love podatkih. V raziskavi so uporabili ponovitev po 5-ih letih kot mejno točko za določitev stopnje tveganja. Model je uspešno predvidel ponovitveno verjetnost in razdelil bolnike na takšne z dobro (več kot 5 let) in slabo (manj kot 5 let) napovedjo.

Choi in drugi (2009) so primerjali učinkovitost umetnega nevronskega omrežja, Bayesiovega omrežja (ang. »Bayesian Network«) in hibridnega omrežja (ang. »Hibrid Network«) za napovedovanje prognoze raka dojke. Ugotovili so, da sta se metodi umetnega nevronskega omrežja in hibridnega omrežja podobno dobro odnesli, in sta imeli višjo stopnjo natančnosti v primerjavi z Bayesiovim omrežjem.

Ashar in drugi (2015) so v svoji raziskavi Predvidevanje preživetja raka dojke skozi proces odkrivanja zakonitosti v podatkovnih bazah, razvili napovedne modele in raziskovali povezavo

med določenimi napovednimi spremenljivkami in preživetjem raka dojke. Metoda podpornih vektorjev se je izkazala za najboljšo med vsemi modeli za predvidevanje preživetja raka dojke, in je zaznala deset pomembnih napovednih spremenljivk. Kot najbolj pomembna spremenljivka se je izkazala spremenljivka, ki meri obnašanje tumorja, kot najmanj pomembna pa spremenljivka, ki meri stadij raka.

Kumar in drugi (2013) so v svoji raziskavi primerjali različne klasifikacijske metode podatkovnega rudarjenja za analizo podatkov raka dojke. V programu WEKA so primerjali šest različnih klasifikacijskih tehnik, primerjava rezultatov pa je pokazala, da ima najvišjo stopnjo natančnosti metoda podpornih vektorjev.

Delen in drugi (2005) so v svoji raziskavi za predvidevanje preživetja raka dojke primerjali tri metode podatkovnega rudarjenja – umetna nevronska omrežja, odločitveno drevo in logistično regresijo. V napovednem modelu so uporabili podatke za rak dojke iz SEER podatkovne zbirke. Podatke so pred samo analizo ustrezno pripravili in pri tem poudarili, da je to eden izmed ključnih korakov pri podatkovnem rudarjenju. Potrebno je bilo preveriti porazdelitev podatkov in odstraniti manjkajoče in izstopajoče vrednosti, do katerih je prišlo tudi zaradi napak pri vnosu podatkov. Za analizo preživetja so ustvarili novo spremenljivko preživetje, ki je definirana kot incidenca raka dojke, kjer je oseba po šestih mesecih (petih letih) po datumu diagnoze še vedno živa. Končna podatkovna baza, ki so jo uporabili, je bila sestavljena iz 17 spremenljivk (16 napovednih spremenljivk in 1 odvisna spremenljivka, glej Tablo 5.1) in 202 932 podatkov.

Tabela 5.1: Napovedne spremenljivke za modeliranje preživetja raka dojke

Ime spremenljivke	**Število različnih vrednosti**		
Človeška rasa	28		
Zakonski stan	6		
Lokacija nastanka	9		
Histologija	91		
Obnašanje	2		
Histološka ocena	5		
Razširjenost bolezni	29		
Vključenost limfnih bezgavk	10		
Radiacija	10		
Stadij raka	5		
SEER koda operativnih posegov	11		
	Povprečje	**Standardni odklon**	**Rang**
Starost	60.60	13.98	10 – 106
Velikost tumorja	19.75	17.65	0 – 200
Število pozitivnih bezgavk	1.423	3.659	0 – 75
Število bezgavk	11.307	8.628	0 – 91
Število lokacij nastanka	1.23	0.491	1 – 8

Vir: Delen in drugi (2005).

Tabela 5.2: Porazdelitev odvisne spremenljivke

Odvisna spremenljivka ***Preživetje***	**Frekvenčna porazdelitev**	**Odstotki**
0 (ni preživel/-a)	109.659	54.00
1 (preživel/-a)	93.273	46.00
Skupaj	202.932	100.00

Vir: Delen in drugi (2005)

Avtorji so v raziskavi uporabili tri različne klasifikacijske metode podatkovnega rudarjenja: umetna nevronska omrežja, odločitveno drevo in logistično regresijo (Delen in drugi 2005). Za te metode so se odločili na podlagi pregleda literature, s pomočjo katere so ugotovili, da so ravno te tri metode najpogosteje uporabljane in najuspešnejše pri tovrstnih analizah. Zaradi tega bomo v zadnjem praktičnem delu magistrske naloge tudi mi uporabili ravno te tri metode, katere so na kratko opisane v nadaljevanju.

Umetna nevronska omrežja (ang. »artificial neural networks«) so biološko navdihnjena metoda, ki je sestavljena iz visoko sofisticiranih analitičnih tehnik in omogoča modeliranje ekstremno kompleksnih ne-linearnih funkcij. Nevronska omrežja so analitične tehnike ustvarjene po vzorcu procesa učenja v kognitivnem sistemu in nevroloških funkcij možganov. Metoda ima sposobnost napovedovanja novih ugotovitev na podlagi drugih ugotovitev po izvršbi tako imenovanega učenja na podlagi obstoječih podatkov. Delen in drugi (2005) so uporabili MLP arhitekturo umetnih nevronskih omrežij z algoritmom vzvratnega ponavljanja (ang. »back-propagation«). Model MLP ima pomembno funkcijo približevanja za napovedne in klasifikacijske probleme. MLP je najbolje raziskan in najpogosteje uporabljan model umetnih nevronskih omrežij. MLP je za klasifikacijski problem predvidevanja preživetja raka dojke bolj učinkovit kot model RBF in SOM (ang. »self-organizing map«) (Delen in drugi 2005).

Metoda odločitvenega drevesa je vedno bolj popularna metoda podatkovnega rudarjenja. Najpogosteje uporabljani algoritmi odločitvenega drevesa so ID3, C4.5, C5 in Breimanovo klasifikacijsko in regresijsko drevo. Ta tehnika rekurzivno loči ugotovitve v vejah z namenom izboljšanja natančnosti predvidevanja. Za identificiranje spremenljivke in ustrezne meje za spremenljivko, ki razdeli vhodne ugotovitve na več podskupin, uporablja matematične algoritme (informacijski dobiček, Ginijev indeks in Hi-kvadrat test). Ta korak se ponavlja na vsakem vozlišču vej, dokler ni konstruirano celotno drevo. Cilj razdelitvenega algoritma je najti mejni par spremenljivk, ki poveča homogenost in vrne dva ali več podskupinskih vzorcev (Delen in drugi 2005).

Tretja metoda je logistična regresija. Ta se v glavnem uporablja za napovedovanje binarnih ali večrazrednih odvisnih spremenljivk. Neodvisna spremenljivka je diskretna in tako ni možno direktno modeliranje z linearno regresijo, zato namesto da napovedujemo oceno samega

dogodka, z linearno regresijo zgradimo model za napovedovanje verjetnosti, da se bo ta dogodek zgodil. Pri dvorazrednem problemu tako 50-odstotna verjetnost pomeni, da primer razdelimo na razred '1' in '0'. Linearna regresija predpostavlja, da je povezava med odvisno spremenljivko in koeficienti napovednih spremenljivk linearna (Delen in drugi 2005).

Za najboljšo metodo za predvidevanje preživetja raka dojke se je izkazala metoda odločitvenega drevesa, in sicer z 93.6-odstotno stopnjo natančnosti. Druga po vrsti je bila metoda umetnih nevronskih omrežij z 91-odstotno natančnostjo, logistična regresija pa z najslabšo klasifikacijsko natančnostjo 89.2-odstotno. Rezultati analize so pokazali, da je najpomembnejša napovedna spremenljivka za preživetje raka dojke histološka ocena. Ta rezultat je v skladu s preteklimi raziskavami, kjer je bilo prav tako ugotovljeno, da je histološka ocena najpomembnejši napovedni dejavnik pri napovedovanju preživetja raka dojke. Kot drugi najpomembnejši napovedni dejavnik se je izpostavil Stadij raka, sledijo pa spremenljivke Radiacija, Število limfnih bezgavk in Velikost tumorja (Delen in drugi 2005).

V naslednjem poglavju bomo glede na pregled literature opravili konkreten primer predvidevanja preživetja raka dojke s pomočjo najpogosteje uporabljanih tehnik podatkovnega rudarjenja.

6 Primer predvidevanje preživetja raka dojke na podlagi podatkov v Sloveniji

V praktičnem delu magistrske naloge bomo uporabili podatke o raku dojke, ki smo jih pridobili iz Registra raka Republike Slovenije. Zaradi edinstvenosti in občutljive narave podatkov smo v prvem koraku morali najprej vložiti prošnjo za uporabo teh podatkov v raziskovalne namene. Uporabo podatkov je odobrila Komisija Republike Slovenije za medicinsko etiko.

Podatki, ki jih bomo uporabili, so podatki o prijavi raka dojke v Sloveniji za obdobje od leta 2005 do leta 2009. Začetno imamo skupaj 6168 podatkovnih zapisov oziroma primerov, za analizo pa smo med vsemi dostopnimi spremenljivkami izbrali 10 napovednih spremenljivk in eno odvisno spremenljivko, ki so predstavljene v Tabeli 6.1.

Tabela 6.1: Napovedne spremenljivke in odvisna spremenljivka za analizo

Ime spremenljivke	*Število različnih vrednosti*	*Vrednosti*
Histologija: histološka skupina karcinoma po ICDO3	48	15-Leiomyosarcoma 1-Small cell carcinoma 24-Sarcoma, unspecified 26-Malignat fibroepithelial tumors 3-Squamous cell carcinoma 45-Malignant tumor, unspecified 46-Neoplasma, no microscopic confirmation 6-Adenocarcinoma 7-Other specified carcinomas 8-Carcinoma, unspecified
Lokacija raka po MKB10: lokacija raka po Mednarodni klasifikaciji za statistične namene	12	Aksilarni predel dojke Bradavica in areola Dojka, neopredeljeno Drudi karcinomi in situ dojke Intraduktalni karcinom in situ Lobularni karcinom in situ Osrednji del dojke Praraščajoča lezija dojke Spodnji-notranji kvadrant dojke Spodnji-zunanji kvadrant dojke Zgornji-notranji kvadrant dojke Zgornji-zunanji kvadrant dojke
Gradus	4	1 - dobro diferenciran 2 - zmerno diferenciran 3 - slabo diferenciran 9 - neznano
Malignost	3	1 - benigen 2 - in situ

			3 - maligen
Lateralnost	4		1 –levo 2 – desno 3 - bilateralno
Multiplost	2		9-neznano 0-ne 1-da
UICC stadij			0 I II IIA IIB IIIA IIIB IIIC IV Nepalpabilne 9 - neznano
Radiacija (kemoterapija)	2		1 - da 0 - ne
Operacija	2		1 - da 0 - ne
	Povprečje	***Std. odklon***	
Starost: starost bolnika ob ugotovitviraka dojke	61.58	13.81	23 – 103 let
Odvisna spremenljivka: Preživetje bolnika: ali je pacient živ ali ne			
Vrednosti	Frekvenčna porazdelitev		Odstotek
1 - živ 9 - mrtev	4364 1804		70.8 29.2

Vir: Register raka Republike Slovenije (2015).

Za izvedbo praktičnega dela bomo uporabili prosto dostopno programsko orodje WEKA za podatkovno rudarjenje. Vse analize bomo opravili s pomočjo algoritmov iz knjižnične zbirke

orodja. Dostop do orodja je enostaven in hiter, orodje pa ponuja številne funkcionalnosti podatkovnega rudarjenja – na primer razvrščanje v skupine, klasifikacijo, združevanje, izbiro atributov in vizualizacijo. Za uporabo tega orodja smo se odločili na podlagi tega in pa tudi zaradi pogoste uporabe s strani mnogih raziskovalcev. WEKA se je izkazala kot uporabno in ključno orodje pri analizi realnih podatkov. Zmanjšuje stopnjo kompleksnosti pri vključevanju realnih podatkov v različne sheme strojnega učenja in ocenjevanju izidov teh shem. Prav tako zagotavlja fleksibilno pomoč za raziskave strojnega učenja in je odlično orodje za uvajanje ljudi pri uporabi strojnega učenja v izobraževalne namene. WEKA je bila razvita na Univerzi Waikato v Novi Zelandiji, njeno ime WEKA pa je kratica za Waikato Environment of Knowledge Analysis. Sistem je zapisan v Javi (objektno usmerjen programski jezik, ki je dostopen za večino računalniških platform), WEKA pa je testirana za Linux, Windows in Macintosh operacijske sisteme. Java omogoča enoten vmesnik za več različnih algoritmov učenja skupaj z metodami za pred in post-obdelavo in za ocenjevanje rezultatov shem za dane podatke (Shrivastava in drugi 2013). Pri uvozu podatkov v WEKO je predviden format ARFF. Podatki, ki smo jih dobili iz RRRS, so bili za nas pripravljeni v Excelovi datoteki, to je v xlsx formatu. Te bomo najprej shranili v CSV formatu in jih bomo potem s pomočjo orodja, ki ga nudi WEKA, pretvorili v ARFF format.

Na Sliki 6.1 vidimo uporabniški vmesnik orodja WEKA, in vidimo, da imamo štiri aplikacije (Shakil in drugi 2015):

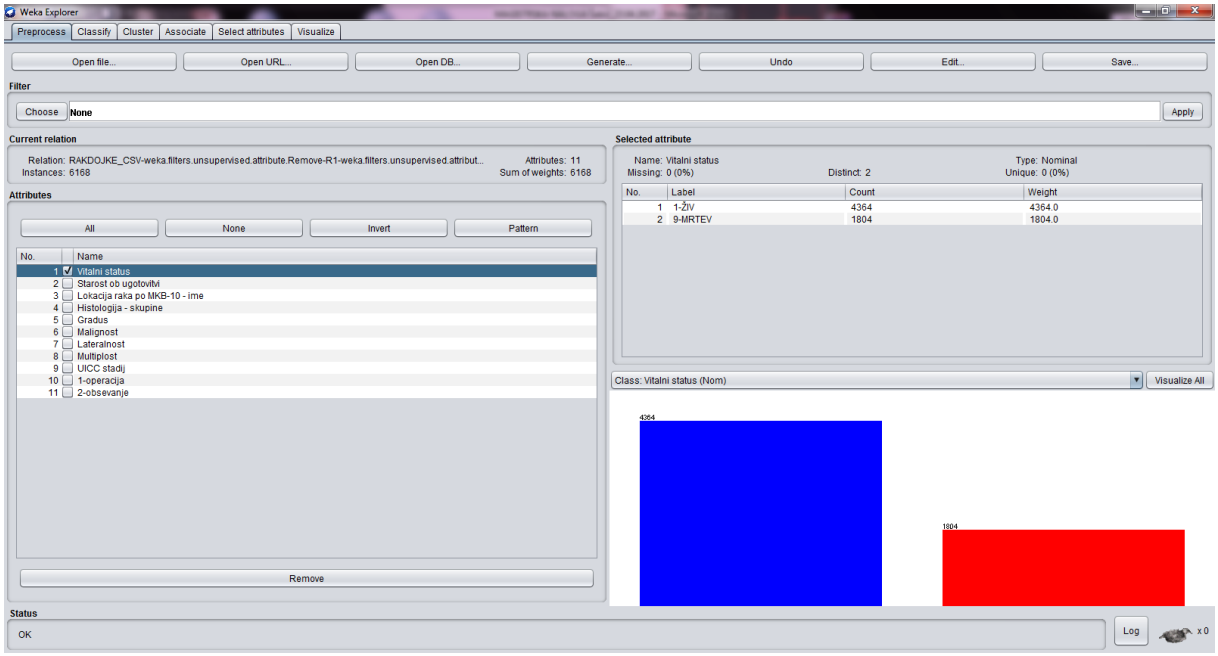
1) Raziskovalec: raziskovalni vmesnik vsebuje različne panele in sicer pred-obdelavo, klasifikacijo, razvrščanje v skupine, izbiro atributov in vizualizacijo.
2) Eksperimentator: ta vmesnik omogoča sistematično primerjavo različnih algoritmov na podlagi danih podatkov. Vsak algoritem se izvede 10-krat in nato dobimo rezultat natančnosti.
3) Pretok znanja: ta aplikacija je alternativa za raziskovalni vmesnik. Razlika je v tem, da tukaj uporabnik izbere komponento iz orodne vrstice orodja in to poveže, da bi oblikoval postavitev za izvajanje algoritmov.
4) Enostavni CLI: predstavlja enostavni vmesnik ukazne vrstice. Uporabnik izvaja ukaze prek ukazne vrstice vmesnika tako, da daje navodila operacijskemu sistemu. Ta vmesnik je v primerjavi z drugimi tremi najmanj priljubljen.

Slika 6.1: Uporabniški vmesnik orodja WEKA

Vir: WEKA (2017).

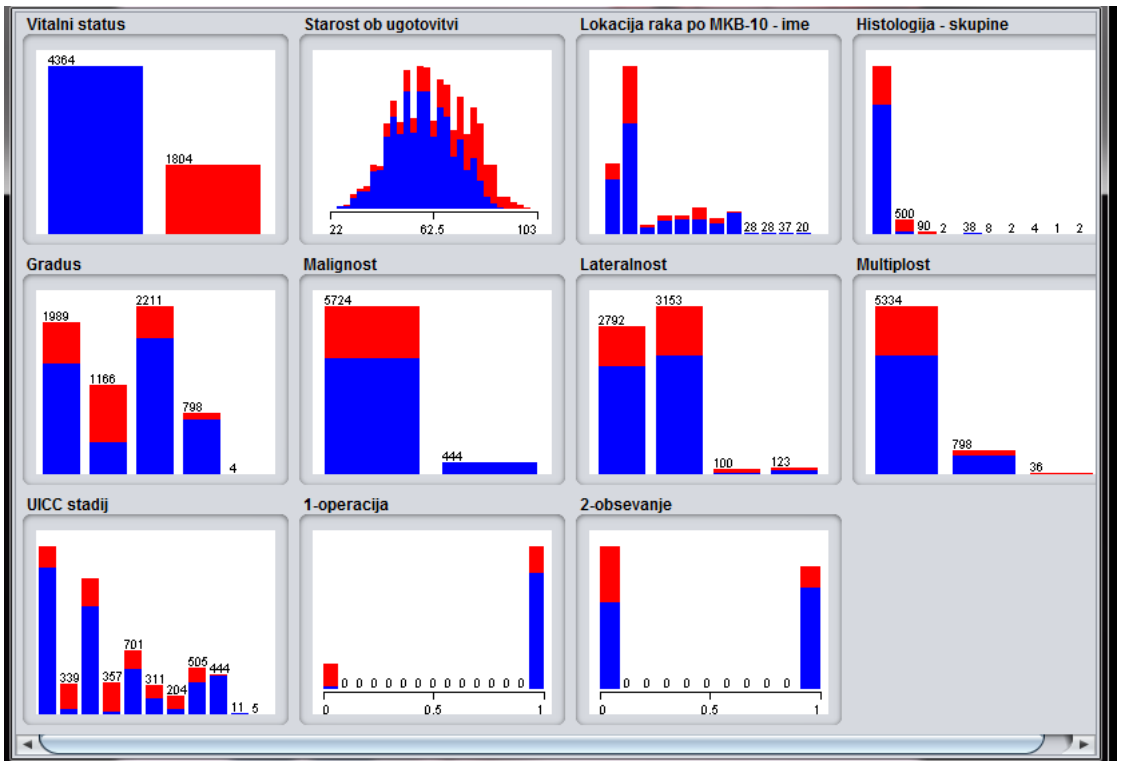
S pomočjo raziskovalnega vmesnika (Slika 6.1 in Slika 6.2) bomo uvozili našo bazo podatkov in s klasifikacijo identificirali problem opazovanih karakteristik raka dojke med bolniki, ter diagnosticirali oziroma predvideli, kateri algoritem je najustreznejši glede na statistične rezultate. Tukaj imamo tudi možnost vizualizacije vseh posameznih vhodnih spremenljivk (Slika 6.3).

Slika 6.2: Pregled baze podatkov v raziskovalnem vmesniku Weke



Vir: Weka (2017).

Slika 6.3: Vizualizacija vhodnih spremenljivk



Vir: Weka (2017).

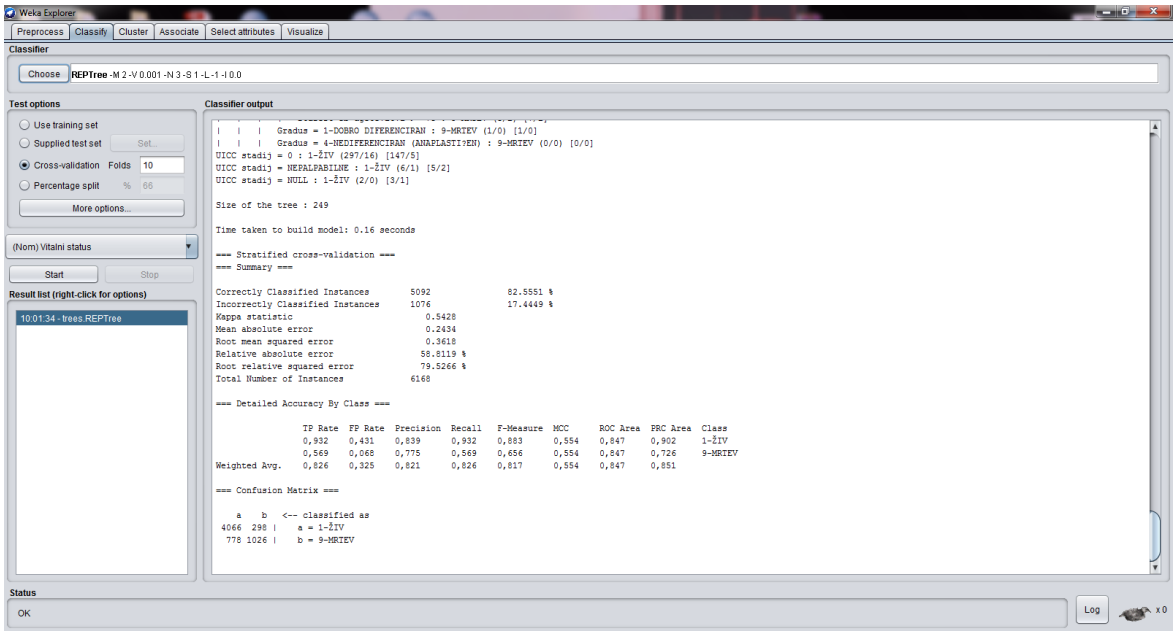
S pomočjo raziskovalnega vmesnika najprej analiziramo klasifikacijsko natančnost. Klasifikacijsko zaupanje in natančnost merimo z naslednjim (Shakil in drugi 2015):

- Stopnja pravilno razvrščenih primerov: pove nam odstotek natančnosti testa in koliko primerov je pravilno razvrščenih.
- Stopnja nepravilne klasifikacijske natančnosti: pove nam odstotno natančnost testa.
- Srednja absolutna napaka: prikazuje število napak pri analizi algoritma klasifikacijske natančnosti.
- Čas: koliko časa je potrebno za izgradnjo modela za predvidevanje.
- ROC območje: ang. »Receiver Operating Characteristic« predstavlja vodilo za izvajanje testa za klasifikacijsko natančnost, ki je diagnosticirana z ocenami odlično (0.90 - 1), dobro (0.80 – 0.90), slabo (0.60-0.70) in neuspelo (0.50 – 0.60).

V nadaljevanju bomo ocenili in primerjali učinkovitost različnih algoritmov za predvidevanje preživetja raka dojke. Test bomo izvedli z naslednjimi tehnikami podatkovnega rudarjenja:

- REP odločitveno drevo,
- logistična regresija in
- RBF umetna nevronska omrežja.

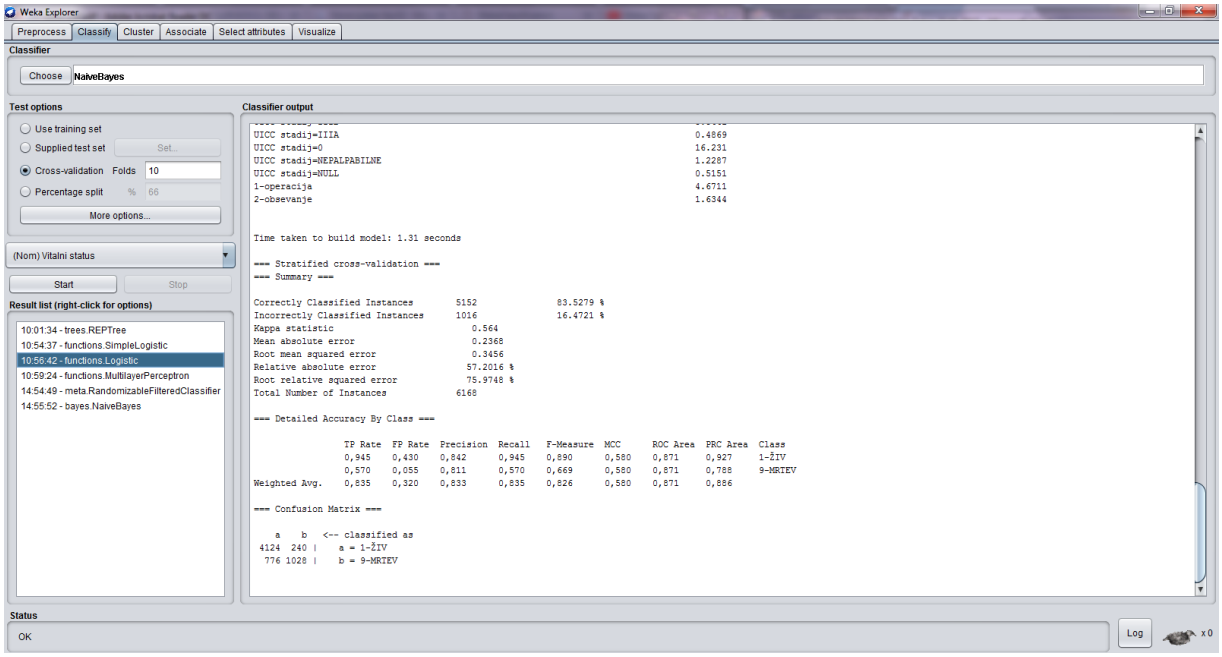
Slika 6.4: Klasifikacija z REP odločitvenim drevesom v orodju WEKA



Vir: Weka (2017).

REP odločitveno drevo smo uporabili, da bi zgradili odločitev in zmanjšali napake zaradi numeričnih atributov ter za razdelitev primerov na dele za določitev natančnosti. Na Sliki 6.4 vidimo doseženo klasifikacijsko natančnost, ki kaže 82.56 odstotkov pravilno razvrščenih primerov (5092 primerov) in 17.45 odstotkov nepravilno razvrščenih primerov (1076 primerov). Povprečna absolutna napaka je 0.24, izgradnja modela je trajala 0.16 sekunde in ROC območje je 0.847.

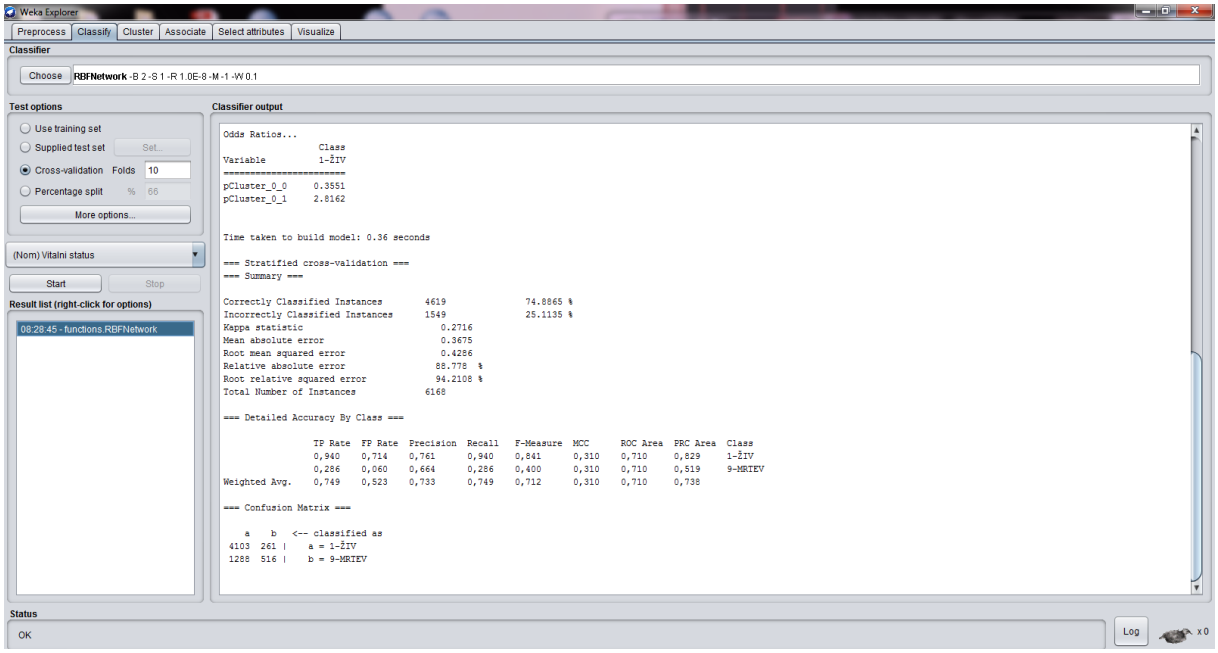
Slika 6.5: Klasifikacija z logistično regresijo v orodju WEKA



Vir: WEKA (2017).

Z metodo logistične regresije dosežemo 83.53-odstotno stopnjo natančnosti. 83.53 odstotkov ali 5152 primerov je pravilno razvrščenih (Slika 6.5). Povprečna absolutna napaka je 0.24, izgradnja modela pa je trajala 1.31 sekunde in ROC območje je 0.87.

Slika 6.6: Klasifikacija z RBF umetnim nevronskim omrežjem v orodju WEKA



Vir: WEKA (2017).

Klasifikacijski test, ki smo ga izvedli s tehniko RBF umetnega nevronskega omrežja, je porabil za izgradnjo modela 0.36 sekunde (Slika 6.6). 74.89-odstotna stopnja natančnosti kaže, da je 4619 primerov pravilno razvrščenih. Povprečna absolutna napaka je 0.37, ROC območje pa 0.71.

Tabela 6.2: Primerjava učinkovitosti tehnik podatkovnega rudarjenja za predvidevanje preživetja raka dojke

Ocenjevalni kriterij	**Klasifikator**		
	REP odločitveno drevo	**RBF nevronsko omrežje**	**Logistična regresija**
Čas zgraditve modela	0.16 s	0.36	1.31 s
Pravilno razvrščeni primeri	5092	4619	5152
Nepravilno razvrščeni modeli	1076	1549	1016
Natančnost	82.56 %	74.89 %	83.53 %

V Tabeli 6.2 je opravljena primerjava rezultatov klasifikacijskih testov. Kot najbolj natančen klasifikator se je izkazala metoda logistične regresije z 83.53-odstotno stopnjo natančnosti.
V naslednji Tabeli 6.3 so predstavljene tudi simulacijske napake, na podlagi katerih vidimo, da je najmanjša stopnja napake ravno pri logistični regresiji. Višja vrednost Kappa statistike ter čim nižja absolutna in relativna napaka pomenita boljšo izvedbo testa pri izgradnji modela za predvidevanje preživetja raka dojke s posameznimi algoritmi.

Tabela 6.3 Primerjava simulacijskih napak za klasifikacijske teste

Ocenjevalni kriterij	**Klasifikator**		
	REP odločitveno drevo	**RBF nevronsko omrežje**	**Logistična regresija**
Kappa statistika	0.54	0.27	0.56
Povprečna absolutna napaka	0.24	0.37	0.24
Relativna absolutna napaka	58.81 %	88.78 %	57.20 %

Tabela 6.4: Primerjava meritev natančnosti za izbrane tehnike

Klasifikator	**TP**	**FP**	**Preciznost**	**Priklic**
REP odločitveno drevo	0.932	0.431	0.839	0.932
	0.569	0.068	0.775	0.569
RBF umetno nevronsko omrežje	0.940	0.714	0.761	0.940
	0.286	0.060	0.664	0.286
Logistična regresija	0.945	0.430	0.842	0.945
	0.570	0.055	0.811	0.570

Po izgradnji modelov predvidevanja je potrebno preveriti tudi, kako natančni so, ta natančnost pa je izračunana glede na preciznost in priklicane vrednosti klasifikacijske matrike. Zgornja

Tabela 6.4 prikazuje vrednosti TP, FP, preciznosti in priklica za REP odločitveno drevo, RBF umetno omrežje in logistično regresijo.

Za boljše razumevanje pomembnosti vhodnih spremenljivk je pomembno analizirati njihov vpliv na predvidevanje preživetja raka dojke. Analizirali bomo, v kolikšni meri katera izmed 10 vhodnih napovednih spremenljivk vpliva na to, in sicer s pomočjo treh različnih testov: Hi-kvadrat test, test informacijskega prispevka (ang. »Info Gain test«) in test razmerja informacijskega prispevka (ang. »Gain Ratio test«).

Hi kvadrat test (ang. »Chi square«) je najpogosteje uporabljana statistika za testiranje povezanosti kategoričnih spremenljivk. Pri tem postavimo ničelno hipotezo – da ne obstaja povezanost med spremenljivkami, in hipotezo – da sta spremenljivki povezani. Hi kvadrat test temelji na primerjavi dejanskih frekvenc s frekvencami, kakršne bi bile v primeru nepovezanosti tj. s teoretičnimi. Formula se glasi $x^2 = \sum_{i=1}^{k} \frac{(f_i - f_j')^2}{f_{i'}}$, kjer je i oznaka za i-to celico med k celicami kontingenčne tabele (Ferligoj in drugi 2011).

Test informacijskega prispevka (ang. »Info Gain«) izbere najboljši atribut v vsakem koraku razvoja odločitvenega drevesa in uporablja entropijo. Je redukcija entropije vrednosti razreda, ko dobimo vrednost atributa. Z njegovo pomočjo izračunamo kako dobro posamezni atribut določa primere v skladu s klasifikacijo. *Test razmerja informacijskega prispevka* (ang. »Gain Ratio«) odpravlja precenjenost večvrednostnih atributov (Mitchell 1997). Natančneje, informacijski prispevek atributa A v povezavi z zbirko primerov S je definiran kot

$$Gain\ (S,A) = Entropy(S) - \sum_{v \in Vrednosti\,(A)} \frac{|S_v|}{|S|}\ Entropy(S_v)$$

kjer $Vrednosti(A)$ predstavlja zbirko vseh možnih vrednosti za atribut A, in S_v je pod-zbirka vrednosti S, za katero ima atribut A vrednost v (tj. $S_v = \{s \in S | A(s) = v\}$) (Mitchell 1997).

Test razmerja informacijskega prispevka izračunamo s pomočjo dodatne razdelitvene informacije

$$\text{SplitInfo}\ (S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

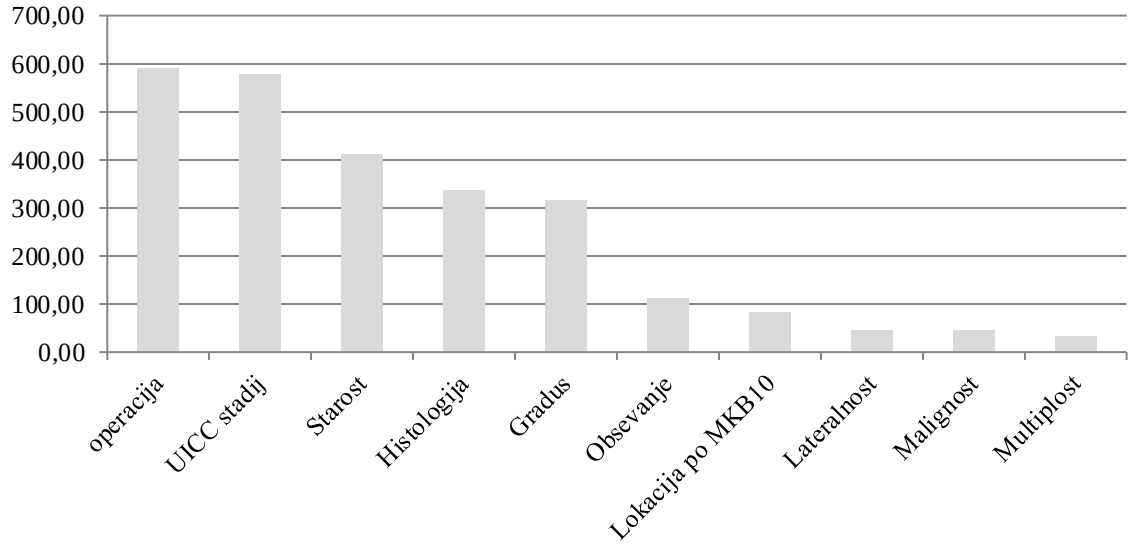
kjer so zbirke c, od S_1 do S podmnožice, primeri particioniranja S s c-vrednotenim atributom A. *SplitInfo* je dejansko entropija vrednosti S glede na vrednosti atributa A. Razmerje informacijskega dobička je tako rezultat naslednje formule (Mitchell 1997):

$$GainRatio\ (S,A) = \frac{Gain\ (S,A)}{SplitInfo\ (S,A)}.$$

Tabela 6.5: Rezultati testov in povprečna uvrstitev napovednih spremenljivk

Spremenljivka	**Hi-kvadrat test**	**Gain Ratio**	**Info Gain**	**Povprečna uvrstitev**
operacija	1777,20	0,316	0,194	592,57
UICC stadij	1738,99	0,071	0,200	579,75
Starost	1239,77	0,069	0,140	413,33
Histologija	1014,71	0,183	0,110	338,33
Gradus	949,63	0,056	0,106	316,60
Obsevanje	335,24	0,041	0,040	111,77
Lokacija po MKB10	253,38	0,014	0,035	84,48
Lateralnost	140,79	0,012	0,015	46,94
Malignost	138,98	0,058	0,022	46,35
Multiplost	98,78	0,019	0,012	32,94

Graf 6.1: Primerjava pomembnosti vpliva napovednih spremenljivk za predvidevanje preživetje raka dojke



S pomočjo testov izračunamo povprečno uvrstitev vsake posamezne napovedne spremenljivke. Različni algoritmi zagotavljajo zelo različne rezultate. Namreč, vsak od njih predstavlja pomembnost spremenljivk na drugačen način. Kot končni rezultat za razvrstitev spremenljivk bomo upoštevali povprečno vrednost vseh uporabljenih algoritmov. Tabela 6.5 kaže, da ima največji vpliv napovedna spremenljivka Operacija, ki se je tudi najbolje izkazala v vseh treh posameznih testih. Pomembno vplivajo tudi naslednje spremenljivke: UICC stadij, Starost, Histologija in Gradus. Graf 6.1 prikazuje pomembnost vseh posameznih napovednih spremenljivk pri predvidevanju preživetja raka dojke.

7 Zaključek

V magistrskem delu smo se ukvarjali z odkrivanjem zakonitosti in podatkovnim rudarjenjem na področju zdravstva. Natančno smo predstavili proces odkrivanja zakonitosti in podatkovno rudarjenje. Ugotovili smo, kako lahko s pomočjo tega procesa in s pomočjo različnih tehnik podatkovnega rudarjenja koristno analiziramo ogromne količine podatkov, ki se vsakodnevno zbirajo v naših podatkovnih bazah. Iz tega pridobimo znanje, ki pomembno vpliva na ključne odločitve na zdravstvenem področju. Podatkovno rudarjenje se zelo pogosto enači s procesom odkrivanja zakonitosti, a to ne velja, saj vemo, da je podatkovno rudarjenje eden izmed najpomembnejših korakov tega procesa. Pri procesu odkrivanja zakonitosti je izredno pomembno v prvi fazi podatkovno bazo primerno pripraviti za analizo, saj nam lahko na primer šumni podatki, manjkajoče vrednosti in odvečni podatki povzročijo težave pri podatkovnem rudarjenju. Pomembno je izbrati tehniko podatkovnega rudarjenja, ki je najprimernejša za posamezno raziskavo. Za modele predvidevanja se tako uporabljajo klasifikacijske metode in regresija.

Pred samo izbiro metod podatkovnega rudarjenja se moramo zavedati, da so zdravstveni podatki izredno heterogeni in občutljivi. Pri izvedbi raziskave moramo upoštevati, da so zdravstveni podatki heterogeni (slike, interpretacije, diagnoze, in tako naprej), da imamo veliko količino podatkov, ter je tako potrebno izbrati primerno kombinacijo. Prav tako se moramo zavedati problema občutljivosti, zasebnosti in pravnega vidika zdravstvenih podatkov. V naši raziskavi smo za to poskrbeli tako, da smo dobili dovoljenje za uporabo zdravstvenih podatkov v zdravstvene namene in se zavezali, da bomo po končani raziskavi podatke uničili in jih dokončno izbrisali z vseh naprav. Pri sekundarni analizi virov na temo zdravstvenih podatkov smo ugotovili, da so le-ti nekoliko drugačni od ostalih predvsem po svoji značilnosti slabe matematične karakterizacije in nezmožnosti za doseganje kanonične oblike podatkov, kar prinaša težave pri statistični analizi podatkov. Vse to vodi k povečanju vloge in koristnosti uporabe tehnik podatkovnega rudarjenja in za to prilagojenih orodij, ki omogočajo v primerjavi s klasično statistiko boljše načine analize tovrstnih podatkov.

V magistrski nalogi smo prav tako podrobno predstavili bolezen raka dojke, njen nastanek, dejavnike za tveganje razvoja ter prognostične dejavnike. Ugotovili smo, da je med vsemi raki ravno rak dojke najbolj pogost pri ženski populaciji tako v Sloveniji kot po svetu, a tudi moški

zbolevajo za to boleznijo. Pri pregledu literature na temo podatkovnega rudarjenja in odkrivanja zakonitosti v zdravstvu smo ugotovili, da je eno izmed najpomembnejših področij ravno napovedovanje in predvidevanje raka pri bolnikih. Na področju raka dojke se tehnike podatkovnega rudarjenja uporabljajo tako za diagnosticiranje kot tudi za napovedovanje oziroma predvidevanje preživetja raka dojke. Za predvidevanje preživetja raka dojke so raziskovalci primerjali različne metode. Najpogosteje uporabljane metode so logistična regresija, MLP in RBF nevronska omrežja, metoda podpornih vektorjev, C4.5 odločitveno drevo, Bayesianovo klasifikacijo. Pri tem delu smo podrobno predstavili primer predvidevanja preživetja raka dojke s pomočjo umetnih nevronskih omrežij, odločitvenega drevesa in logistične regresije.

Glede na analizo sekundarnih virov smo se odločili, da bomo v empiričnem delu uporabili metode, ki so se v preteklosti izkazale kot najbolj učinkovite. S pomočjo prosto dostopnega orodja WEKA za podatkovno rudarjenje smo opravili analizo podatkov o raku dojke za obdobje od leta 2005 do 2009. Naša odvisna spremenljivka je preživetje raka dojke, uporabili pa smo še 10 napovednih neodvisnih spremenljivk. Pri izvedbi testov z različnimi algoritmi za izgradnjo modela predvidevanja preživetja raka dojke smo tudi sami ugotovili, kako pomemben je proces pred-obdelave podatkov. Dokler podatki niso bili urejeni v skladu s pravili za podatkovno rudarjenje in v skladu z zahtevami orodja WEKA, analize ni bilo mogoče izvesti. Po ustrezni pripravi podatkov smo le-te uvozili v orodje in opravili testiranje s pomočjo treh tehnik podatkovnega rudarjenja: REP odločitveno drevo, RBF nevronsko omrežje in logistična analiza. Podobno kot ostali raziskovalci smo tudi mi ugotovili, da se je za oblikovanje modelov predvidevanja preživetja raka dojke najbolje izkazala metoda logistične regresije. Po uspešnosti ji sledi REP odločitveno drevo in nazadnje RBF nevronsko omrežje. Poleg tega nas je zanimalo tudi, katere napovedne spremenljivke v modelu predvidevanja preživetja raka dojke imajo največji vpliv. S pomočjo več različnih testov smo ugotovili, da je v našem primeru najboljši napovednik to, ali je bila pri pacientu izveden operativni poseg ali ne. Pomemben vpliv imajo tudi UICC stadij raka dojke, starost bolnika pri diagnozi bolezni, histološka skupina raka in gradus.

Glede na potek podatkovnega rudarjenja in pregled literature lahko kot sklep izpostavimo in potrdimo, da je izredno pomembno pravilno izvesti vse korake procesa odkrivanja zakonitosti

– od samega začetnega razumevanja področja uporabe in ciljev procesa do uporabe pridobljenega znanja. Pomembno je, da razumemo potrebe raziskovanja in področje, na katerem poteka, saj lahko le na ta način razumemo ugotovitve in pridobljeno znanje tudi apliciramo, ter s tem pripomoremo k napredku družbe.

8 Literatura

- Abdelaal, Ahmed Mohamed Medhat in Farouq Wael Muhamed. 2010. Using data mining for assessing diagnosis of breast cancer. *Proc. International multiconfrence on computer science and information Technology:* 11–17.
- Acir, Nurettin in Güzeliş Cüneyt. 2004. Automatic Spike Detection in EEG by a Two-stage Procedure Based on Support Vector Machines. *Computers in Biology and Medicine* 34 (7): 561–575.
- Ahmad, L.G., Eshlaghy Abbas Toloie., Poorebrahimi A., Ebrahimi M. and Razavi A.R. 2013. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *Health & Medical Informatics* 4 (2).
- Anunciacao, Orlando, Gomes C. Bruno, Vinga Susana, Gaspar Jorge, Oliveira L. Arlindo in Rueff Jose. 2010. A Data Mining approach for detection of high-risk Breast Cancer groups. *Advances in Soft Computing* (74): 43–51.
- Bellaachia, Abdelghani in Erhan Guven. 2006. Predicting Breast Cancer Survivability Using Data Mining Techniques. *Age* 58(13): 10–110.
- Bellazzi, Riccardo in Blaz Zupan. 2008. Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics* 77: 81–97.
- Berman, Jules J., G. William Moore in Grover M. 1996. Hutchins. Maintaining patient confidentiality in the public domain Internet autopsy database (IAD). *Proc AMIA Annu Fall Symp:* 328–332.
- Borštnar, Simona, Matej Bračko, Tanja Čufer, Kristjana Hertl, Marko Hočevar, Marija Us Krašovec, Elga Majdič, Bojana Pajk, Maja Primic Žakelj, Mileva Rener, Andreja Škufca Smrdel, Marjetka Uršič Vrščaj, Marija Vegelj Pirc in Janez Žgajnar. 2006. *Rak dojke: Kaj morate vedeti.* Ljubljana: Onkološki inštitut. Dostopno prek: https://www.onko-i.si/fileadmin/onko/datoteke/dokumenti/Rak-dojke.pdf (13. maj 2017).
- Breault, Joseph L., Colin R. Goodall, Peter J. Fos. 2002. Data Mining a Diabetic Data Warehouse. *Artificial Intelligence in Medicine* 26: 37–54.
- Bundred NJ. 2001. Prognostic and predictive factors in breast cancer. *Cancer Treatment* 27: 137–142.

- Burke, Harry B., David Rosen, Philip Goodman. 1995. Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. *Advances in neural information processing systems* 7: 1063–1067.
- Chang, Pin Wei in Liou Ming Der. 2008. Comparision of three Data Mining techniques with Ginetic Algorithm in analysis of Breast Cancer data. *Journal of Telemedicine and Telecare* 9.
- Ceusters, Werner. 2001. Medical Natural Language Understanding as a Supporting Technology for Data Mining in Healthcare. *Medical Data Mining and Knowledge Discovery:* 41–67.
- Chaea, M. Young, Hye S. Kima, Kwan C. Tarkb, Hyun J. Parkb, Seung H. Hoa. 2003. Analysis of healthcare quality indicator using data mining and decision support system. *Expert Systems with Applications* 24: 167–172.
- Chen, Hsinchun, Sherrilynne S. Fuller, Carol Friedman, & William Hersh. 2005. Knowledge management, data mining, and text mining in medical informatics. *Medical Informatics*: 3–33.
- Cios J.Krzysztof in William G. Moore. 2002. Uniqueness of medical data mining. *Artificial Intelligence in Medicine* 26: 1–24.
- Cios, Krzysztof J. in Lukasz A. Kurgan. 2005. Trends in data mining and knowledge discovery. *Advanced techniques in knowledge discovery and data mining*: 1-26.
- Cios, Krzysztof J., Witold Pedrycz in Roman W. Swiniarski. 2012. *Data mining methods for knowledge discovery*. New York: Springer Science & Business Media. Dostopno prek: Google Books.
- Chi, Chih-Lin, W. Nick Street in William H. Wolberg. 2007. Application of Artifical Neural Network- based Survival Analysis on Two Breast Cancer Datasets. *Annual Symposium Proceedings / AMIA Symposium*.
- Choi Jong Pill, Tae Hwa Han and Rea Woong Park. 2009. A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis. *Journal of Korean Society of Medical Informatics*: 49–57.
- Cooper, Ted in Jeff Collman. 2005. Managing Information Security and Privacy. *Medical Informatics*: 95–137.

- Delen, Dursun, Glenn Walker in Amit Kadam. 2005. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial Intelligence in Medicine* 34 (2): 113–127.
- Dreiseitl, Stephan, Lucila Ohno-Machado, Harald Kittler, Staal Vinterbo, Holger Billhardt, Michael Binder. 2001. A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions. *Journal of Biomedical Informatics* 34: 28–36.
- Fayyad, Usama, Gregory Piatetsky-Shapiro in Padhraic Smyth. 1996. From data mining to knowledge discovery in databases. *AI magazine* 17 (3): 37–54.
- Ferligoj, Anuška, Katja Lozar Manfreda in Aleš Žiberna. 2011. *Osnove statistike na prosojnicah*. Študijsko gradivo pri predmetu statistika. Ljubljana: Fakulteta za družbene vede. Dostopno prek: https://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0ahUKEwi9LXmqdHVAhXL1xQKHab5D1YQFggtMAE&url=http%3A%2F%2Fstudentski.net%2Fget%2Fulj_fdv_nv1_sta_sno_bivarna_analiza_01.pdf&usg=AFQjCNGH6jiGmc0EI7-9pLfTA-bRc8hJhA (11. avgust 2017).
- Freitas, A. Alex. 2013. *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer Science & Business Media. Dostopno prek: Google Books.
- Gandhi, Rajiv K., Karnan Marcus in Kannan S. 2010. Classification rule construction using particle swarm optimization algorithm for breast cancer datasets. *Signal Acquisition and Processing, ICSAP, International Conference:* 233–237.
- Goldner, Martin G., Knatterud Genell L., Prout Thaddeus E. Effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes 3. Clinical implications of UGDP results. *JAMA* 218 (9): 1400–10.
- Gupta, Shelly, Dharminder Kumar in Anand Sharma. 2011. Data mining classifications techniques applied for breast cancer diagnosis and prognosis. *Indian Journal of Computer Science and Engineering* 2 (2): 188–195.
- Han, Jiawei in Micheline Kamber. 2006. Data mining: Concepts and techniques. Elsevier Science. Dostopno prek: https://cs.wmich.edu/~yang/teach/cs595/han/ch01.pdf (17. junij 2017).

- Hripcsak, George, J. H. Austin, P. O. Alderson in C. Friedman. 2002. Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports. *Radiology* 224 (1): 157–163.
- Jamarani, S. M., H. Behnam in G. A. Rezairad. 2005. Multiwavelet Based Neural Network for Breast Cancer Diagnosis. *GVIP 05 Conference:* 19–21.
- Jerez, J. M., L. Franco, E. Alba, A. Llombart-Cussac, A. Lluch, N. Ribelles, , B. Munarriz in M. Martín. 2005. Improvement of breast cancer relapse prediction in high risk intervals using artificial neural networks. *Breast Cancer Research and Treatment, 94* (3): 265–272.
- Jimenez-Lee, Ricardo, Bruce Ham, John Vetto, Rodney Pommier. 2003. Breast cancer severity score is an innovative system for prognosis. *The American journal of surgery* 186 (4): 404–408.
- Jonsdottir, Thora, Ebba Thora Hvannberg, Helgi Sigurdsson in Sven Sigurdsson. 2008. The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining. *Expert Systems with Applications* 34: 108–118.
- Kandaswamy, A., C. S. Kumar, R. P. Ramanathan, R. Jayaraman in N. Malmurugan. 2004. Neural Classification of Lung Sounds Using Wavelet Coefficients. *Computers in Biology and Medicine* 34: 523–537.
- Kaur, Harleen in Siri Krishan Wasan. 2006. Empirical Study on Applications of Data Mining Techniques in Healthcare. *Journal of Computer Science* 2 (2): 194–200.
- Kharya, Shweta. 2012. Using data mining techniques for diagnosis and prodnosis of cancer disease. *International Journal of Computer Science, Engineering and Information Technology* 2 (2): 55–65.
- Kovalerchuk, Boris, Evgenii Vityaev in James F. Ruiz. 2001. Consistent and Complete Data and 'Expert' Mining in Medicine. *Studies in Fuzziness and Soft Computing* 60: 238–281.
- Kulkarni, Siddhivinayak. 2012. Machine *Learning Algorithms for problem solving in computer applications: Intelligent Techniques*. Hershey: IGI Global.

- Kumar, G. Ravi, G. A. Ramachandra in K. Nagamani. 2013. An Efficient Prediction of Breast Cancer Data using Data Mining Techniques. *International Journal of Innovations in Engineering and Technology* 2 (4): 139–144.
- Lai, Hui-Chuan, Stacey C. FitzSimmons, David B. Allen, Michael R. Kosorok, Michael R. Kosorok, Preston W. Campbell in Philip M. Farrell. 2000. Risk of persistent growth impairment after alternate-day prednisone treatment in children with cystic fibrosis. *New England Journal of Medicine* 342 (12): 851–859.
- Leskovec, Jure, Anand Rajarman, Jeffrey D. Ullman. 2014. *Mining of Massive Datasets*. Cambridge: Cambridge University Press.
- Lundin, Mikael, Johan Lundin, B.H Burke., S. Toikkanen, L. Pylkkänen in H. Joensuu. 1999. Artificial Neural Networks Applied to Survival Prediction in Breast Cancer. *Oncology International Journal for Cancer Resaerch and Treatment* 57: 281–286.
- Maimon, Oded in Lior Rokach. 2005. *The Data Mining and Knowledge Discovery Handbook*. New York: Springer.
- Mansour, Edward G., Robert Gray, Ahmad H. Shatila, C.K. Osborne, Tormey Douglass C., Gilchrist Kennedy W., Robert M. Cooper, Falkson Geoffrey. 1989. Efficacy of adjuvant chemotherapy in high-risk node-negative breast cancer: an intergroup study. *N Engl J Med* 320 (8): 485–90.
- Mitchell, M. Tom. 1997. *Machine Learning*. McGraw-Hill Science/Engineering/Math. Dostopno prek: http://www.cs.ubbcluj.ro/~gabis/ml/ml-books/McGrawHill%20-%20Machine%20Learning%20-Tom%20Mitchell.pdf (11. avgust 2017).
- Ngan, Po Shun, Man Leung Wong, Wai Lam, Kwong Sak Leung, Jack C. Y. Cheng. 1999. Medical data mining using evolutionary computation. *Artificial Intelligence in Medicine* 16: 73–96.
- Ohno-Machado, L. 2001. Modeling medical prognosis: Survival analysis techniques. *Journal of Biomedical Informatics* 34: 428–439.
- Onkološki inštitut Ljubljana, Epidemiologija in register raka, Register raka. 2016. *Rak v Sloveniji 2013*. Ljubljana: Onkološki inštitut Ljubljana. Dostopno prek:

https://www.onko-i.si/fileadmin/onko/datoteke/dokumenti/RRS/LP_2013.pdf (13. maj 2017).

- Pal Nikhil R. in Lakhmi Jain, ur. 2005. *Advanced Techniques in Knowledge Discoveryand Data Mining*. London:Springer.
- Padmavati J. 2011. A Comparative study on Breast Cancer Prediction Using RBF and MLP. *International Journal of Scientific & Engineering Research* (2): 1.
- Prather, Jonathan C., Lobach D. F., Goodwin L. K., Hales J. W., Hage M. L. in Hammond W. E. 1997. Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse. *Proceedings of the AMIA Annual Symposium Fall:* 101–105.
- Rampaul, R. S., S. E. Pinder, C. W. Elston, I. O. Ellis. 2001. Prognostic and predictive factors in primary breast cancer and their role in patient management: the Nottingham breast team. *European Journal of Surgical Oncology* 27: 229–238.
- Razavi, Amir, Hans Gill, Hans Åhlfeldt in Nosrat Shahsavar. 2007. Predicting Metastasis in Breast Cancer: Comparing a Decision Tree with Domain Experts. *Journal of Medical Systems* 31 (4): 263–273.
- Sammut, Claude in Geofferey I. Webb. 2011. *Encyclopedia of Machine Learning*. New York: Springer.
- Sarvestani, Soltani A., A. A. Safavi, M. N. Parandeh in M. Salehi. 2010. Predicting Breast Cancer Survivability using data mining techniques. *Software Technology and Engineering (ICSTE)* (2): 227–231.
- Sawarkar, Sudhir D., Ghatol Ashok A. in Pande Amol P. 2006. Neural Network aided Breast Cancer Detection and Diagnosis. *Proceedings of the 7th WSEAS International Conference on Neural Networks:* 158–163.
- Slora. 2016. Osnovni epidemiološki podatki o raku: Dojka (C50). Dostopno prek: http://www.slora.si/c/document_library/get_file?uuid=4e2f0c00-dfd0-4400-b2cd-f261c9b181c9&groupId=11561 (6. maj 2017).
- Schneier, Bruce. 1996. *Applied cryptography: protocols, algorithms, and source code in* C. New York: John Wiley & Sons, Inc.
- Shakil, Kashish Ara, Shadma Anis in Mansaf Alam. 2015. Dengue disease prediction using weka data mining tool. Dostopno prek:

https://www.researchgate.net/profile/Mansaf_Alam/publication/272521958_Den/links/55952f8d08ae793d1379d0df.pdf (28. avgust 2017).

- Shrivastava, Shiv Shakti, Anjali Sant in Ramesh Prasad Aharwal. 2013. An Overview on Data Mining Approach on Breast Cancer data. *International Journal of Advanced Computer Research* 3 (4): 256.
- Street, W. Nick. 1998. A Neural Network Model for Prognostic Prediction. Fifteenth International Conference on Machine Learning, ICML: 540–546.
- Sweeney Latanya. 2001. *Computational disclosure control: a primer on data privacy protection*. Spring: Massachusetts Institute of Technology.
- Thongkam, Jaree, Guandong Xu, Yanchun Zhang in Fuchun Huang. 2009. Toward breast cancer survivability prediction models through improving training space. *Expert Systems with Applications* 36 (10): 12200–12209.
- Trifiro, Gianluca, Antoine Pariente, Preciosa M. Coloma, Jan A. Kors, Giovanni Polimeni, Ghada Miremont-Salame, Maria Antonietta Catania, Francesco Salvo, Anaelle David, Nicholas Moore, Achille Patrizio Caputi, Miriam Sturkenboom, Mariam Molokhia, Julia Hippisley-Cox, Carlos Diaz Acedo,Johan van der Lei in Annie Fourrier-Reglat. 2009. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? *Pharmacoepidemiology and Drug Safety* 18: 1176–1184.
- US Veterans Administration Co-operative Urological Research Group. 1967. Treatment and survival of patients with cancer of the prostate. *Surg Gynecol Obstet* 124(5): 1011–1017.
- Viera, J. Anthony in Joanne M. Garrett. 2005. Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine* 37 (5): 360–363. Dostopno prek: http://www1.cs.columbia.edu/~julia/courses/CS6998/Interrater_agreement. Kappa_statistic.pdf (11. avgust 2017).
- Zupan, Blaž, Janez Demšar, Michael W. Kattan, Robert J. Beck in Ivan Bratko. 2000. Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artificial intelligence in medicine 1 (20): 59–75.
- Xiong, Xiangchun, Kim Yangon, Yuncheol Baek, Dae Wong Rhee in Soo-Hong Kim. 2005. Analysis of breast cancer using data mining and statistical techniques.

The sixth international conference on software engineering, artificial intelligence, networking and parallel: 82–87.