UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Romana Koprivec

# Napovedovanje bolezni ledvic z metodami strojnega učenja

DIPLOMSKO DELO
NA UNIVERZITETNEM ŠTUDIJU

Mentor: doc. dr. Zoran Bosnić

Ljubljana, 2012

*Besedilo je oblikovano z urejevalnikom besedil LaTeX.*

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Tržaška 25
1001 Ljubljana, Slovenija
telefon: 01 476 84 11
01 476 83 87
faks: 01 426 46 47
01 476 87 11
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 01794/2012
Datum: 03.01.2012

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **ROMANA KOPRIVEC**

Naslov: **NAPOVEDOVANJE BOLEZNI LEDVIC Z METODAMI STROJNEGA UČENJA**

**PREDICTION OF A KIDNEY DISEASE USING MACHINE LEARNING METHODS**

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

Kandidatka naj v diplomski nalogi obravnava medicinski problem napovedovanja obstruktivne nefropatije (bolezen neprehodnosti ledvičnih poti). Napovedovanje iz dane množice podatkov je težavno, saj je vzorec primerov majhen, atributi so močno korelirani, podatki prihajajo iz različnih virov, ki med seboj niso usklajeni in povezani. Kandidatka naj se problema loti z različnimi pristopi strojnega učenja, kot so gradnja napovednih modelov, uporaba metod za izbiro atributov itd. Uspešnost doseženega rezultata naj prikaže in ovrednoti z metodologijo mednarodnega tekmovanja, ki je obravnavalo ta problem.

Mentor:

doc. dr. Zoran Bosnić

Dekan:

prof. dr. Nikolaj Zimic

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

# IZJAVA O AVTORSTVU

diplomskega dela

Spodaj podpisani/-a Romana Koprivec,

z vpisno številko 63060129,

sem avtor/-ica diplomskega dela z naslovom:

Napovedovanje bolezni ledvic z metodami strojnega učenja

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal/-a samostojno pod mentorstvom
  doc. dr. Zorana Bosnića
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 14.03.2012 Podpis avtorja/-ice:

# Zahvala

*Na tem mestu se zahvaljujem docentu dr. Zoranu Bosniću za vse nasvete in strokovno pomoč pri izdelavi diplomske naloge.*

*Posebna zahvala gre tudi moji družini in Denisu, ki so me tekom študija podpirali in mi stali ob strani tudi ob mojih muhastih dnevih.*

# Seznam uporabljenih kratic in simbolov

| | |
|---|---|
| **DRF** | razlika ledvične funkcije (*angl. differential renal function*) |
| **miRNA** | mikro RNA (*angl. micro RNA*) |
| **MFMW** | kombinirana metoda izbire atributov, ki združuje filter metode z metodami notranje optimizacije (*angl. Multiple-Filter-Multiple-Wrapper*) |
| **ON** | obstruktivna nefropatija |
| **PD** | premer ledvičnega meha (*angl. pelvis diameter*) |
| **RRMSE** | koren relativne srednje kvadratne napake (*angl. relative root mean squared error*) |

# Kazalo

**Povzetek** **1**

**Abstract** **2**

**1 Uvod** **3**

**2 Izbor podmnožice atributov** **5**

**3 Metode strojnega učenja za regresijske probleme** **7**
3.1 Linearna regresija . . . . . 7
3.2 Metoda podpornih vektorjev . . . . . 8
3.3 Lokalne metode . . . . . 9
3.3.1 K-najbližjih sosedov . . . . . 9
3.3.2 Lokalno utežena regresija . . . . . 9
3.4 Večnivojski perceptron . . . . . 10
3.5 Regresijska drevesa . . . . . 12
3.6 Kombiniranje algoritmov strojnega učenja . . . . . 12
3.6.1 Bagging . . . . . 12
3.6.2 Metoda naključnih podprostorov . . . . . 13
3.6.3 Glasovanje . . . . . 13
3.6.4 Aditivna regresija . . . . . 13

**4 Opis problema in podatkov** **14**
4.1 Povezave med bazami podatkov . . . . . 15
4.2 Ciljne spremenljivke . . . . . 16

**5 Postopek testiranja** **18**
5.1 Orodja . . . . . 18
5.2 Uporabljene metode za izbiro atributov . . . . . 18
5.3 Uporabljene metode strojnega učenja . . . . . 19

5.4 Dodajanje proteinskih atributov . . . . . 21
5.5 Preverjanje uspešnosti modelov . . . . . 22

**6 Rezultati** **23**
6.1 Primerjava regresijskih modelov na posameznih bazah . . . . . 23
6.2 Primerjava kombinacij regresijskih modelov in metod za izbiro atributov . . . . . 25
6.3 Rezultati regresijskih metod po dodajanju povezanih proteinskih atributov . . . . . 27

**7 Zaključek** **30**

**Seznam slik** **32**
**Seznam tabel** **33**

**Literatura** **34**

# Povzetek

Skupina raziskovalcev je objavila tekmovanje za izdelavo diagnostičnih modelov, ki napovedujejo intenzivnost obolenja obstruktivne nefropatije. V diplomskem delu smo s pomočjo strojnega učenja razvili regresijski model, ki glede na molekularni profil pacienta napove stopnjo obolenja. Ta model napoveduje dve ciljni vrednosti: premer ledvičnega meha in razliko ledvične funkcije.

Napovedne modele smo poskušali izboljšati še z znanjem o potencialnih povezavah med biološkimi nivoji. S pomočjo objavljenih podatkov o povezavah med atributi in Mann-Whitneyevem testom smo implementirali postopek, ki poveže dve različni bazi podatkov na dveh različnih skupinah vzorcev. Po povezovanju baz je najnižji koren relativne srednje kvadratne napake (v nadaljevanju RRMSE) na učnih podatki dosegla lokalno utežena regresija. RRMSE te metode je bil na povezanih bazah nižji kot na osnovni bazah, prav tako se je izboljšal rezultat na testni množici.

Ocenjene najboljše regresijske metode na učnih in testnih podatkih so se zaradi izredno majhnega števila vzorcev razlikovale. Nekatere modeli so imeli višji RRMSE na učni množici, a so dosegli boljše rezultate na testnih podatkih. Kljub temu smo z najboljšim modelom glede na RRMSE na učni množici za 81% presegli najboljši objavljeni rezultat na tekmovanju.

## Ključne besede:

strojno učenje, regresija, bioinformatika, obstruktivna nefropatija.

# Abstract

A group of researchers announced a competition for making diagnostic models that predict intensity of obstructive nephropathy. In this thesis we have developed a regression model which can predict the level of the illness according to the molecular profile. This model is intended to predict two target values: pelvic diameter and differential renal function.

We wanted to improve the prediction models with the knowledge about potential connections between biological levels. We have implemented a procedure with the help of the published data about connections between attributes and Mann-Whitney test, which connects two different databases on two different groups of samples. The lowest relative root mean squared error (RRMSE) has been achieved using locally weighted regression. RRMSE of the method has been lower on connected databases than on the original databases. The result on test dataset has improved as well.

The best estimated regression methods on train and test dataset have differentiated because of extraordinary small number of samples. Some models had higher RRMSE on train dataset; however, they achieved better results on test dataset. Nevertheless, with the best model according to RRMSE on train dataset we have exceeded the best-published result on the competition for 81%.

## Key words:

machine learning, regression, bioinformatics, obstructive nephropathy.

# Poglavje 1

# Uvod

S hitrim razvojem visokopretočnih tehnologij v biologiji so postale pogoste tudi študije bioloških mehanizmov na podlagi profiliranja vzorcev populacije. Ti vzorci so pogosto pridobljeni z različnimi merilnimi tehnikami in se nanašajo na različne biološke nivoje, kot so genomika, proteomika, metabolomika itd. Tipični cilji teh študij so ekstrakcije modelov za napovedovanje določenih vrednosti, diagnostiko stanja ali pa omogočanje boljšega razumevanja bioloških mehanizmov. Te razmere so pripeljale do zanimivih izzivov na področju strojnega učenja in podatkovnega rudarjenja.

Enega takih izzivov je predstavila množica raziskovalcev z objavo tekmovanja za izdelavo modelov, ki napovedujejo intenzivnost obstruktivne nefropatije (v nadaljevanju ON).

Obstruktivna nefropatija pomeni motnjo v odtoku seča, ki povzroči akumulacijo urina v ledvici. To privede do zmanjšanja ledvične funkcije, poškodb ledvičnega tkiva in kasneje v odpoved ledvice. ON se pogosto pojavlja pri novorojenčkih in se zdravi z dializo ali transplantacijo. Ena glavnih težav je torej odločitev o vrsti zdravljenja. Cilj tekmovanja je tako konstrukcija diagnostičnega modela, ki bo glede na molekularni profil pacienta kar najbolj natančno napovedal stopnjo obolenja in tako omogočil podporo zdravljenju. Stopnja obolenja se določi glede na napovedan premer ledvičnega meha (PD) in razliko ledvične funkcije (DRF).

Podatki o meritvah so pridobljeni iz urina novorojenčkov, katerim je bila diagnosticirana ON ali pa obstaja sum nanjo. Meritve so pridobljene z različnimi tehnikami (npr. meritve izražanja genov s pomočjo mikromrež, mikromreže s protitelesi) ter se nanašajo na tri različne biološke nivoje: miRNA, proteine in metabolite. Poleg velike količine šuma je največji problem podatkov, ki jih pridobimo s temi meritvami, njihova visoka dimenzionalnost. Podatki

tipično vsebujejo več tisoč spremenljivk in majhno število vzorcev (pacientov).

Drugi izziv, ki se poraja ob teh podatkih, je visoka stopnja povezanosti in odvisnosti med atributi, ki so bili pridobljeni iz istega vira (npr. ko-regulirani geni) ali pa med atributi iz različnih virov (npr. geni nadzirajo nivo izražanja proteinov, ti pa določajo nivo izražanja številnih drugih proteinov in metabolitov).

Eden pogostih problemov v biologiji je tudi ta, da meritev ni možno opraviti na vseh vzorcih. Učna množica je tako sestavljena iz dveh različnih skupin vzorcev, med katerima ni prekrivanja. Na dvajsetih vzorcih so narejene meritve miRNA in metabolitov, na desetih vzorcih pa meritve proteinov. Testna množica je sestavljena iz štirih vzorcev, na katerih so narejene le meritve miRNA in metabolitov. S standardnimi metodami strojnega učenja ni mogoče povezati teh dveh različnih baz podatkov, vendar pa obstajajo različne biološke baze podatkov, ki podajajo potencialne povezave med različnimi biološkimi sestavinami z istih ali različnih bioloških nivojev. To znanje se lahko uporabi za učne algoritme in tako izboljša kvaliteto napovednih modelov.

Cilj diplomskega dela je izdelava modela, ki bo čim bolj točno napovedal stopnjo ON. V diplomskem delu smo preverili, katere metode za izbiro atributov ublažijo posledice šuma, odvisnosti v podatkih ter problem visoke dimenzionalnosti. Na izbranih atributih smo nato preizkusili več različnih metod strojnega učenja. Poleg tega smo s pomočjo potencialnih povezav med različnimi bazami poskusili uporabiti tudi znanje iz proteinske baze ter z njim doseči večjo napovedno točnost.

V naslednjem poglavju opišemo izbiro podmnožice atributov. Sledi poglavje 3, kjer predstavimo uporabljene metode strojnega učenja za regresijske probleme. V poglavju 4 opišemo problem, ki ga bomo reševali v tej diplomski nalogi skupaj s podatki, ki so nam bili na voljo za reševanje problema. Sledi poglavje 5, kjer opišemo potek testiranja ter uporabljene ocene uspešnosti modelov. V poglavju 6 so predstavljeni rezultati na posameznih bazah podatkov, ocene regresijskih metod in metod za izbiro atributov. Podani so tudi rezultati, ki bi jih modeli dosegli na tekmovanju. Na koncu smo v poglavju 7 navedli še sklepne ugotovitve ter podali morebitne izboljšave.

# Poglavje 2

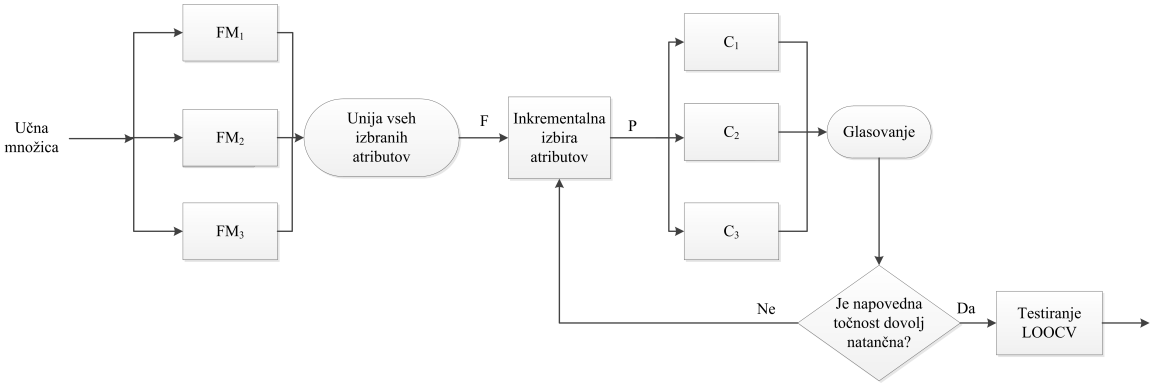
# Izbor podmnožice atributov

Metode pridobivanja podatkov, ki se v zadnjem času uporabljajo na področju bioinformatike (npr. pridobivanje podatkov o genskih izrazih z mikromrežami), hkrati merijo vrednosti več tisoč različnih molekul v biološkem vzorcu [3]. Pogosto so nekateri zajeti atributi šumni in nepomembni za napovedovanje ciljne spremenljivke. Ti atributi vplivajo na slabše učenje, zato je potrebno pred uporabo metod za strojno učenje izbrati le ustrezno podmnožico atributov.

Najpreprostejša in najhitrejša metoda je metoda filtriranja atributov[1]. Z izbrano ocenitveno funkcijo (npr. MDL ali ReliefF) ocenimo kvaliteto atributov. V novo učno množico nato vključimo le določeno število najboljših atributov ali pa vnaprej določimo prag kvalitete, ki odreže slabše atribute.

Bolj zanesljiva je metoda notranje optimizacije (*angl. wrapper*). Ta metoda izbira en atribut za drugim tako, da optimizira napovedno točnost izbrane metode za strojno učenje. Najprej z izbrano metodo strojnega učenja zgradimo model na vsakem posameznem atributu iz učne množice. Atribut, ki je imel največjo napovedno točnost, se nato doda v podmnožico najboljših atributov. Nato za vsako kombinacijo posameznega atributa iz množice preostalih atributov in podmnožico najboljših atributov naredimo nov model. Kombinacija atributov, ki ima največjo napovedno točnost, predstavlja novo podmnožico najboljših atributov. Postopek se ponavlja, dokler ni dosežena dovolj velika napovedna točnost. Slabost te metode je, da je ob velikem številu atributov izredno počasna.

Raziskovalci so predstavili več različnih metod, ki poskušajo odpraviti slabosti obeh metod. Ena takih metod je metoda, ki kombinira oba pristopa - MFMW [3] (*angl. Multiple-Filter-Multiple-Wrapper approach*). Po tej metodi se najprej izvede več filter metod $FM_i$ nad celotno učno množico, kar je




Slika 2.1: Model MFMW metode za izbiro atributov

prikazano na Sliki 2.1. Vsaka filter metoda ima svoje karakteristike, zato bodo nekateri atributi metod enaki, nekateri pa različni. Vsi izbrani atributi posamezne filter metode se nato združijo v novo množico $F$. Nad to novo množico se nato izvede metoda notranje optimizacije, ki je sestavljena iz večih klasifikatorjev. Rezultati različnih klasifikatorjev se združijo z metodo glasovanja. Novi atributi, pridobljeni z metodo notranje optimizacije, se dodajajo v množico atributov $P$, dokler ni dosežena dovolj visoka napovedna natančnost.

Postopki za preiskovanje prostora podmnožic atributov ponavadi uporabljajo kar požrešno iskanje, lahko pa uporabimo tudi kako bolj kompleksno metodo iskanja. Najpogosteje se uporabljata iskanje naprej in iskanje nazaj.

# Poglavje 3

# Metode strojnega učenja za regresijske probleme

## 3.1 Linearna regresija

Linearna regresija [1, 7] je preprosta metoda za numerično predikcijo. Ideja linearne regresije je, da izrazimo razred (**r**) kot linearno kombinacijo atributov s predefiniranimi utežmi (**w**):

$$\hat{r} = w_0 + \sum_{i=1}^{a} w_i v^{(i)} = \mathbf{w^T v} \tag{3.1}$$

pri čemer je $v^T = \langle 1, v^{(1)}, ..., v^{(a)} \rangle$ vektor vseh vrednost atributov $A_1$,...,$A_a$ z dodanim elementom 1.

Naloga je določiti vektor **w** uteži $w_i, i = 0...a$ tako, da minimiziramo vsoto kvadratov napake (SSE) napovedi razreda preko vseh učnih primerov:

$$SSE = \sum_{j=1}^{n} (r^{(j)} - w_0 - \sum_{i=1}^{a} w_i v^{(i,j)})^2 \tag{3.2}$$

Z **V** označimo matriko vseh učnih primerov:

$$\mathbf{V} = \begin{bmatrix} 1 & v^{(1,1)} & \dots & v^{(a,1)} \\ 1 & v^{(1,2)} & \dots & v^{(a,2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & v^{(1,n)} & \dots & v^{(a,n)} \end{bmatrix} \tag{3.3}$$

ter z $\mathbf{r}$ vektor razredov vseh učnih primerov :

$$r^T = \langle r^1, ..., r^n \rangle \tag{3.4}$$

Minimalni SSE dobimo, če velja:

$$\mathbf{w} = (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}\mathbf{r} \tag{3.5}$$

## 3.2 Metoda podpornih vektorjev

Metoda podpornih vektorjev [1, 7] je primerna za učenje na velikih množicah vzorcev, opisanih z velikim številom manj pomembnih atributov. Te metoda dosegajo visoko napovedno točnost, vendar pa je interpretacija naučenega težavna.

Metoda podpornih vektorjev za regresijske probleme poskuša z minimizacijo napake napovedi poiskati funkcijo, ki najbolje aproksimira učne primere. Pri tem pri računanju napake napovedi ne upoštevajo odstopanja primerov, ki so manjša od meje, določene s parametrom $\varepsilon$. Uporabniško določen parameter $\varepsilon$ določa območje okoli regresijske funkcije, v kateri se napake ne upoštevajo. Z njim lahko nadziramo, kako dobro se bo funkcija prilagajala učnim podatkom. Če je vrednost parametra prevelika, bo napovedni model neuporaben, saj bo vedno napovedoval povprečno vrednost razreda. V nasprotnem primeru pa premajhna vrednost parametra ne zajame vseh učnih primerov.

Za linearne primere lahko zapišemo funkcijo podpornih vektorjev za regresijo kot:

$$y = b + \sum_{\text{i je podporni vektor}} \alpha_i \mathbf{a_i} \bullet \mathbf{a} \tag{3.6}$$

kjer je $\mathbf{a_i}$ podporni vektor, $\mathbf{a}$ vektor testnega primera ter $b$ in $\alpha_i$ parametra. Za nelinearne primere se skalarni produkt $\mathbf{a_i} \bullet \mathbf{a}$ zamenja z jedrno funkcijo.

Podporni vektorji so v primeru linearne funkcije vse tiste točke, ki se nahajajo zunaj valja s polmerom $\varepsilon$ ali na njegovem robu. Vsi drugi koeficienti imajo vrednost 0 in jih lahko odstranimo iz učne množice. Ker je število podpornih vektorjev relativno majhno v primerjavi z vsemi učnimi primeri, je kompleksnost rešitve relativno majhna.

Metoda podpornih vektorjev poskuša poleg minimizacije napake minimizirati tudi naklon regresijske funkcije. Razmerje med napako rešitve in naklonom uravnavamo s parametrom $C$. Večja je vrednost parametra kompleksnosti, bolj se funkcija prilagaja učnim primerom.

## 3.3 Lokalne metode

Lokalne metode [1] sodijo med najpreprostejše metode strojnega učenja. Tem vrstam metod pravimo tudi leno učenje, saj učenja pri teh metodah skorajda ni. Klasifikator v fazi učenja ne zgradi modela na učni množici, ampak vsakič znova uporabi učne podatke za napovedovanje vrednosti novega primera. Časovna zahtevnost klasifikacije novega primerja je zato v primerjavi z drugimi metodami večja. Metoda shrani vse učne primere in ob napovedi novega primera poišče $k$ najbližjih učnih primerov glede na vnaprej definirano razdaljo.

### 3.3.1 K-najbližjih sosedov

Pri najpreprostejši različici algoritma $k$-najbližjih sosedov [1] shranimo vse učne primere. Pri napovedi razreda pri regresiji določimo vrednost razreda kot povprečje vrednosti razredov $k$ najbližjih sosedov.

Za optimalno izbiro parametra $k$, ki določa število najbližjih sosedov, je potrebno upoštevati, ali učni podatki vsebujejo šum. Če ga ni, se bo najbolje obnesel algoritem 1-NN. Kadar pa je šuma veliko, lahko s povečevanje parametra povprečimo napovedi več bližnjih primerov in s tem zmanjšamo šum. Po drugi strani pa s povečevanjem parametra $k$ povečujemo tudi možnost, da pri napovedovanju prispevajo tudi primeri, ki se precej razlikujejo od novega primera.

Za metriko pri računanju razdalje med novim in učnimi primeri se pogosto uporablja evklidska razdalja. Vsi zvezni atributi se normalizirajo na interval [0,1]. Razdalja med dvema vrednostma je enaka absolutni razliki med njima. Razdalja med dvema primeroma:

$$D(u_l, u_j) = \sqrt{\sum_{i=1}^{a} \left(v^{(i,l)} - v^{(i,j)}\right)^2} \tag{3.7}$$

### 3.3.2 Lokalno utežena regresija

Lokalno utežena regresija [1] je zelo podobna metodi k-najbližjih sosedov. Namesto uteževanja učnih primerov se uporabi poljubna regresijska funkcija skozi k najbližjih sosedov, npr. linearna funkcija, kvadratna funkcija, večnivojski perceptron itd. Lokalno utežena regresija sestavi lokalno aproksimacijo ciljne funkcije v okolici novega primera, ki se nato uporabi za napoved vrednosti funkcije za dani primer. Najpogosteje se uporablja linearna lokalno utežena

regresija. Preveč kompleksnih funkcij, ki bi natančno modelirale vseh k učnih primerov, ni priporočljivo uporabljati zaradi nevarnosti prevelikega prilagajanja.

## 3.4 Večnivojski perceptron

Večnivojski perceptrum [1, 2] je večnivojska usmerjena nevronska mreža. Z njim lahko rešujemo tudi nelinearne probleme.

Večnivojski perceptrum je sestavljen iz skupine vhodnih nevronov $X1, ..., X_{n_X}$, skupine izhodnih nevronov $Y1, ..., Y_{n_Y}$ in med njima še vsaj enega skritega nivoja. Število vmesnih skritih nivojev in število nevronov na posameznem skritem nivoju ni omejeno. Vezem med nevroni pravimo po analogiji z biološkimi nevronskimi mrežami tudi sinapse. Vsaka vez ima tudi določeno utež. Nevron je procesni element, ki izračuna uteženo vsoto vhodov nevronov in dobljeno vsoto preslika v izhodno vrednost. Uteženi vsoti pravimo funkcija aktivacije, funkciji, ki preslika aktivacijo v izhod, pa funkcija aktivacije.

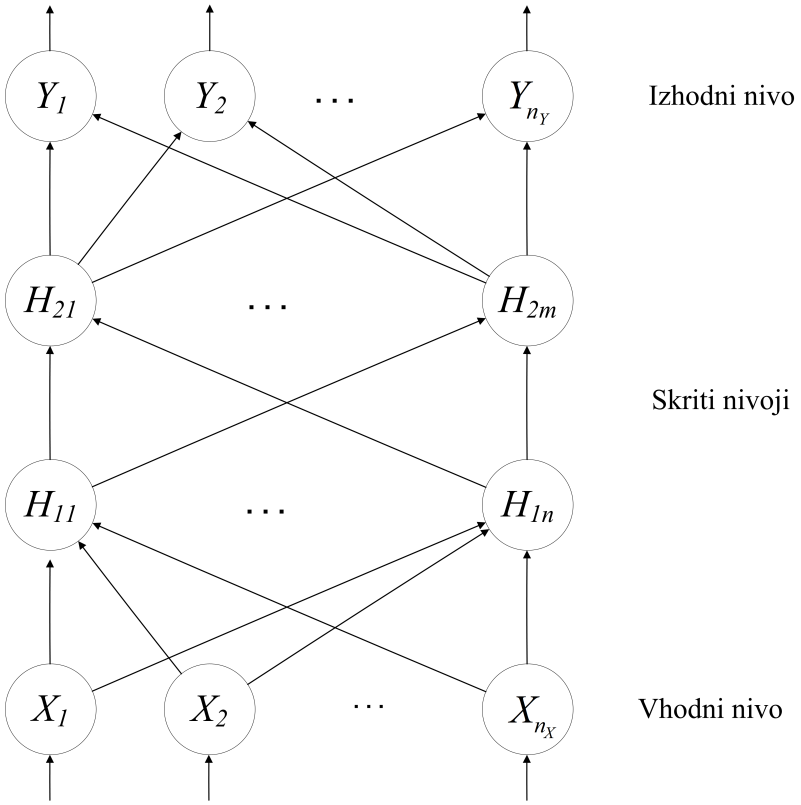
Model večnivojskega perceptruma z dvema skritima nivojema je prikazan na Sliki 3.1.

Učenje mreže, sestavljene iz več različnih skritih nivojev, omogoča *posplošeno pravilo delta*, ki mu pravimo tudi *pravilo vzvratnega razširjanja napake* (*angl. backpropagation of error*).

Izračun nevronske mreže poteka tako, da se najprej izbere število skritih nivojev, število nevronov na posameznem nivoju, aktivacijsko funkcijo ter določi naključne vrednosti utežem. Na vhodu dobi mreža vhodni vzorec in z razširjanjem po nivojih do izhodnega nivoja se najprej izračuna izhod mreže. Nato se za vsak izhodni nevron izračuna razlika med dejanskim in želenim izhodom. Glede na to razliko se spremenijo uteži na vezeh med zadnjim in predzadnjim nivojem. Zatem se izračunajo želene vrednosti nevronov na predzadnjem nivoju in rekurzivno se nadaljuje spreminjanje uteži vse do vhodnega nivoja nevronov.

Slabosti posplošenega delta pravila so:

- ne konvergira vedno k optimalni mreži,
- lahko zahteva zelo veliko število prehodov preko učnih primerov,
- potrebno je dodatno empirično testiranje za določitev pravega trenutka za ustavitev učenja, ki pa je časovno potratno,
- nima biološke analogije z možgani.




Slika 3.1: Večnivojski perceptron

Pri večnivojskem perceptrumu se mora izračunavati odvod napake tudi pri nevronih iz skritih nivojev in ne samo na izhodnih nivojih, zato je potrebno, da je izhodna funkcija zvezna in zvezno odvedljiva. Pogosto se uporablja sigmoidna funkcija:

$$f(X) = \frac{1}{1 + e^{-X}} \tag{3.8}$$

Za nevron na prvem skritem nivoju se izračuna njegov izhod pri učnem primeru $X(n)$ po enačbi:

$$H_{1i}(n) = f\left(\sum_{j=1}^{n_X} W_{ji}^1(n) X_j(n)\right) \tag{3.9}$$

kjer oznaka $W^k$ pomeni matriko uteži vhodnih sinaps, ki vodijo v k-ti skriti nivo.

## 3.5 Regresijska drevesa

Regresijsko drevo [1] je sestavljeno iz vozlišč, vej in listov. Vozlišča predstavljajo atribute, veje podmnožice atributov, listi pa funkcije, ki preslikajo vektor vrednosti atributov v zvezni razred. Vsaka pot od korena do lista predstavlja eno pravilo, pri čemer so pogoji, ki jih srečamo na poti, konjuktivno povezani. Najpreprostejša funkcija v listih je konstanta. Ta predstavlja povprečno vrednost razreda vseh učnih primerov v listu. Pogosto se uporabljajo tudi linearne funkcije na podmnožici zveznih atributov, pri čemer razred učnih primerov modeliramo z regresijsko premico.

Pri gradnji drevesa izberemo tisti atribut, ki množico učnih primerov najbolje razdeli na več podmnožic. Potem naredimo enako še za vsa poddrevesa. Postopek ponavljamo, dokler ne ostane premalo učnih primerov za zanesljivo nadaljevanje gradnje drevesa, zmanjka dobrih atributov ali pa funkcija v listu drevesa dovolj dobro ne modelira primerov, ki so še ostali. Za izbiro najboljšega atributa pri gradnji drevesa se uporabljata dve meri: razlika variance ali ReliefF.

Pri napovedovanje vrednosti odvisne spremenljivke novega primera potujemo od korena po ustreznih vejah do lista. Vrednost nove spremenljivke napovemo s funkcijo, ki se nahaja v listu.

Zaradi majhnega števila primerov v listih postanejo napovedi nezanesljive, zato se drevo ponavadi naknadno poreže.

## 3.6 Kombiniranje algoritmov strojnega učenja

### 3.6.1 Bagging

Bagging [1] (*angl. bootstrap aggregating*) je metaklasifikator za izboljšanje stabilnosti in napovedne točnosti metod za strojno učenje.

Pri baggingu generiramo serijo učnih množic. Posamezno generirano učno množico kreiramo tako, da iz učne množice z $n$ primeri vsakič $n$ krat naključno izberemo primer iz učne množice z vračanjem. Nad vsako tako generirano učno množico potem poženemo učni algoritem. Hipoteze se ponavadi razlikujejo, saj so zgrajene na različnih podatkovnih množicah. Za napovedovanje novega primera uporabimo nato kombinacijo generiranih hipotez.

### 3.6.2 Metoda naključnih podprostorov

Izbiri naključne podmnožice atributov pravimo metoda naključnih podprostorov [2, 5]. Za vsak klasifikator iz ansambla klasifikatorjev kreiramo novo učno množico. Iz učne množice z $n$ atributi brez vračanja naključno izberemo $d$ atributov, pri čemer velja, da je $d < n$. Nad vsako tako novo generirano učno množico potem zgradimo model z izbranim klasifikatorjem. Napovedovanje novega primera temelji na kombinaciji napovedi vseh klasifikatorjev iz ansambla (ponavadi kar na večinskem glasovanju). Metode naključnih prostorov delujejo dobro v primeru, ko nepomembne informacije niso osredotočene na neko podmnožico vseh atributov, ampak so razpršene po celotni množici atributov.

### 3.6.3 Glasovanje

Metoda glasovanja je ansambel klasifikatorjev [7], kjer vsak vključen klasifikator zgradi model na celotni učni množici in ločeno napove vrednost ciljne spremenljivke. Pri napovedovanju novega primera se nato upošteva kombinacija vseh napovedi, pri čemer se lahko uporabi npr. večinsko napoved, mediano, povprečje verjetnosti.

### 3.6.4 Aditivna regresija

Aditivni modeli [7] generirajo napovedi s seštevanjem prispevkov različnih modelov. Večina učnih algoritmov za aditivne modele ne gradi napovednih modelov ločeno, ampak poskušajo zgraditi ansambel klasifikatorjev tako, da se ti medsebojno dopolnjujejo in dosegajo boljšo napovedno točnost.

Pri aditivni regresiji v zaporedju zgradimo več regresijskih modelov. Najprej zgradimo osnovni regresijski model, ki napoveduje vrednost ciljne spremenljivke. Napake med dejanskimi in napovedanimi vrednostmi osnovnega modela na učni množici popravimo z dodajanjem novih modelov. Pred uporabo drugega modela zamenjamo vrednosti ciljnih spremenljivk z razliko med napovedano vrednostjo prvega modela in dejansko vrednostjo. Z dodajanjem napovedanih vrednosti drugega modela k prvemu modelu zmanjšamo napako na učni množici. Ker so nekatere napake po dodajanju drugega modela ponavadi še vedno prisotne, napako zmanjšamo z dodajanjem novih modelov, ki napovedujejo napake prejšnjih modelov. Posamezni modeli minimizirajo kvadratno napako posamezne napovedi, medtem ko algoritem minimizira kvadratno napako napovedi celotnega ansambla.

# Poglavje 4

# Opis problema in podatkov

Množica raziskovalcev je na spletni strani Tunedit [8] objavila tekmovanje za izdelavo diagnostičnega modela, ki bo glede na molekularni profil pacienta čim bolj natančno napovedal stopnjo obolenja obstruktivne nefropatije. ON je obolenje ledvic, ki nastane zaradi motnje v odtoku seča in se zdravi z dializo ali transplantacijo.

Neobdelane podatke je omogočil Inserm U858 iz Francije, predprocesirali pa sta jih Univerza Manchester iz Velike Britanije ter Univerza v Ženevi, Švica.

Objavljeni podatki predstavljajo številne izzive, saj:

- vsebujejo skupno več tisoč atributov in le 30 vzorcev v učni množici,
- so pridobljeni z različnimi merilnimi tehnikami na različnih skupinah vzorcev,
- obstajajo odvisnosti med atributi na istih in različnih bioloških nivojih,
- so nepopolni.

Meritve podatkov se nanašajo na tri različne biološke nivoje: miRNA (mikro RNA), proteine in metabolite. MiRNA podatki so bili pridobljeni z meritvami izražanja genov, metaboliti pa s pomočjo kapilarne elektroforeze, povezane z masno spektrometrijo. Nivo izražanja proteinov je izmerjen z dvema različnima tehnikama: mikromrežami s protitelesi in tekočo kromatografijo, povezano z masno spektrometrijo (LC-MS/MS metoda).

Celotna baza šteje skupaj 34 vzorcev. Razlogi, da so bile meritve opravljene na tako majhnem številu vzorcev, so: omejene količine urina, stroški in trajanje študije. Problem majhnega vzorca v bioinformatiki je zelo pogost in

je najbrž nekaj, s čimer se bomo srečevali tudi v prihodnosti. Kljub temu da je tehnološki razvoj in razvoj različnih tehnik testiranja omogočil cenejše pridobivanje vzorcev, namreč ne moremo mimo dejstva, da je količina nekaterih bioloških vzorcev omejena.

Učno množico sestavlja 30 vzorcev. Ker ni bilo možno izvesti vseh meritev na vseh vzorcih, je učna množica razdeljena na dve različni skupini vzorcev. Ti dve skupini vzorcev sta popolnoma neodvisni in izključujoči. Prva skupina vsebuje 20 vzorcev, na katerih so bile narejene meritve miRNA in meritve metabolitov. Druga skupina vsebuje 10 vzorcev, na katerih so bile narejene meritve proteinov s pomočjo dveh različnih tehnik.

Testna množica zajema 4 primere, na katerih so bile narejene le meritve miRNA in metabolitov. Razlog, da meritve proteinov niso bile narejene, izhaja iz dejstva, da je zanje potrebno 10-krat do 20-krat več bioloških vzorcev. Ker so biološki vzorci pridobljeni iz urina novorojenčkov, so ti precej omejeni. Vsak vzorec, pridobljen za meritve proteinov, je tako sestavljen iz enega do štirih vzorcev urina novorojenčkov.

Bolj podrobni podatki celotne baze se nahajajo v Tabeli 4.1.

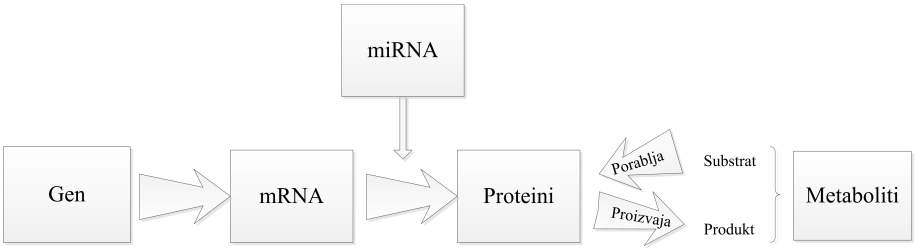
| **Biološki nivo** | | **miRNA** | **Proteini** | | **Metaboliti** |
|---|---|---|---|---|---|
| **Metoda meritev** | | Pan-miRNA mikromreže | Mikromreže s protitelesi | LC-MS/MS | CE-MS |
| **Št. atributov** | | 834 | 725 | 968 | 852 |
| **Tip atributov** | | numerični | numerični | numerični | numerični |
| **Učna množica - št. primerov** | **Skupina 1** | / | 10 | 10 | / |
| | **Skupina 2** | 20 | / | / | 20 |
| **Testna množica – št. primerov** | | 4 | / | / | 4 |

Tabela 4.1: Podatkovne baze

## 4.1 Povezave med bazami podatkov

Testna množica zaradi omejenih količin vzorcev ne vsebuje meritev proteinov. S standardnimi metodami strojenega učenja za napovedovanje ciljnih spremenljivk tako ne moremo uporabiti proteinskih podatkov iz učne množice. Vendar pa so raziskovalci objavili tudi potencialne povezave med atributi na različnih bioloških nivojih, s katerimi lahko povežemo podatke posameznih baz. Te povezave so bile pridobljene s pomočjo različnih bioloških baz, kot so: Target Scan, Mirbase, Uniprot, KEGG in HMDB.

Potencialne povezave so prikazane na Sliki 4.1. MiRNA nadzira nivo zatiranja mRNA, s čimer nadzira nivo izražanja proteinov. Nivo izražanja teh proteinov pa nato določa nivo izražanja posameznih metabolitov.




Slika 4.1: Prikaz odvisnosti med biološkimi nivoji

Povezave so podane kot seznami povezanih atributov med miRNA, proteini in metaboliti. Podani podatki o povezavah podajajo samo indekse atributov posameznih baz, ki so medsebojno povezani, ne pa tudi podatka o tem, kakšna je povezava oz. odvisnost med atributoma.

Tako so npr. povezave med miRNA in protitelesi podane v posebni datoteki v obliki:

1: 220;221;305;306;307;323;329;333;335;340;393
2:
3: 220;221;305;306;307;323;329;333;335;340;720
4:
5: 149;484 . . .

To pomeni, da je atribut št. 5 iz baze miRNA povezan z atributoma št. 149 in 484 iz baze Antibody, medtem ko atribut št. 4 nima povezanega atributa v bazi Anitbody.

Zavedati se je potrebno, da biološke baze, iz katerih so bile pridobljene povezave med atributi, izražajo trenutno znanje o povezavah in se stalno dopolnjujejo. Poleg tega so podatki o meritvah statistično obdelani, zato so lahko podatki o povezavah nepopolni in šumni.

## 4.2 Ciljne spremenljivke

Cilj tekmovanja je bil izdelati model, ki čimbolj natančno napove intenzivnost obolenja ON. Ta je podana s stopnjo obstrukcije, ki se meri na dva načina:

- premer ledvičnega meha izmerjen v milimetrih - PD (*angl. pelvis diameter*). Več urina se nabira v ledvicah, bolj se širi ledvični meh in s tem njegov premer.

- razlika ledvične funkcije, izmerjena v odstotkih - DRF (*angl. differential renal function*). DRF se izračuna iz izmerjene funkcije ovirane in neovirane ledvice. Večja je vrednost DRF, bolj je oslabljeno delovanje ledvic.

Obe vrednosti ciljnih spremenljivk sta normalizirani na interval med 0 in 1.

Rezultat na tekmovanju se je določal iz napovedane vrednosti PD, DRF in skupno PD in DRF na štirih testnih primerih. Rezultat se izračuna po formuli:

$$err_1 = 1/4 \sum_{i=1}^{4} [(PD_i^p - PD_i)^2] \tag{4.1}$$

$$err_2 = 1/4 \sum_{i=1}^{4} [(DRF_i^p - DRF_i)^2] \tag{4.2}$$

$$err_3 = 1/4 \sum_{i=1}^{4} [(PD_i^p - PD_i)^2 + (DRF_i^p - DRF_i)^2] \tag{4.3}$$

$$REZULTAT = BASELINE - (err_1 + err_2 + err_3) \tag{4.4}$$

pri čemer $PD_i^p$ predstavlja napovedano vrednost PD i-tega testnega primera, $PD_i$ pa dejansko vrednost i-tega testnega primera.

$BASELINE$ predstavlja rezultat, ki bi ga dobili, če bi za napovedane vrednosti vzeli kar povprečne vrednosti PD in DRF na učnih primerih. Ta znaša 0,116145. Izračunan $REZULTAT$ je izboljšava glede na $BASELINE$ vrednost.

Najboljši dosežen rezultat na tekmovanju je znašal 0,04146.

# Poglavje 5

# Postopek testiranja

## 5.1 Orodja

Celoten postopek za analizo podatkov in napovedovanje ciljnih spremenljivk je napisan v programskem jeziku Java. Za preizkušanje različnih metod strojnega učenja smo uporabili javanski odprto-kodni paket WEKA [9]. Paket WEKA je razvila Univerza Waikato z Nove Zelandije in podpira številne metode za predprocesiranje podatkov, izbiro atributov in regresijo. Pri statističnih izračunih pri povezovanju baz smo si pomagali s knjižnico za statistiko JSC (*angl. Java Statistical Classes*) [10].

## 5.2 Uporabljene metode za izbiro atributov

Preizkusili smo naslednje metode za izbiro atributov, ki so implementirane v paketu WEKA:

1. *CfsSubsetEval*- filter metoda izbire atributov, ki temelji na korelaciji. Metodo smo uporabili v kombinaciji s štirimi različnimi iskalnimi algoritmi:
    - *GreedyStepwise* (C-GS) - požrešno iskanje
    - *BestFirst* (C-BF) - metoda najboljši najprej
    - *LinearForwardSelection* (C-LFS) - iskanje naprej s povečevanjem števila atributov pri vsaki iteraciji
    - *SubsetSizeForwardSelection* (C-SSFS)- Nadgradnja metode *LinearForwardSelection*. Za določitev optimalnega števila atributov uporablja notranje prečno preverjanje.

2. *ReliefFAttributeEval* (RF) - filter metoda, ki temelji na ReliefF meri.

3. *MFMW* (*angl. Multiple-Filter-Multiple-Wrapper*) - implementirali smo kombinirano metodo izbire atributov, ki združuje filter metode z metodami notranje optimizacije. Vse izbrane atribute vseh prejšnjih filter metod smo združili in nato nad novo množico atributov izvedli metodo notranje optimizacije z metodo glasovanja dveh različnih regresorjev: linearne regresije in metode podpornih vektorjev.

V Tabeli 5.1 so predstavljene preizkušene metode za izbiro atributov, uporabljenim iskalnim algoritmom ter številom izbranih atributov na bazi miRNA, bazi CEMS (oz. bazi metabolitov) ter združeni bazi pri napovedovanju vrednosti PD in DRF. Združena baza (v nadaljevanju ZB) vsebuje vse atribute iz baze miRNA in baze CEMS.

| **Kratica** | **Št. atributov** | | | | | |
|---|---|---|---|---|---|---|
| | **MiRNA** | | **CEMS** | | **ZB** | |
| | **PD** | **DRF** | **PD** | **DRF** | **PD** | **DRF** |
| RF | 56 | 56 | 9 | 23 | 56 | 56 |
| C-GS | 11 | 11 | 8 | 12 | 13 | 24 |
| C-BF | 13 | 11 | 9 | 14 | 15 | 23 |
| C-LFS | 14 | 11 | 9 | 10 | 12 | 13 |
| C-SSFS | 13 | 9 | 9 | 11 | 16 | 12 |
| MFMW | | | 9 | | | 13 |

Tabela 5.1: Preizkušene metode za izbiro atributov

Prvih pet metod je skupno izbralo 68 različnih atributov na združeni bazi pri napovedovanju DRF in 18 različnih atributov na bazi CEMS pri napovedovanju PD. Vsi ti atributi so bili nato uporabljeni pri metodi notranje optimizacije z metodo MFMW za izbiro atributov.

## 5.3 Uporabljene metode strojnega učenja

Za izračun ciljnih spremenljivk smo preizkusili več regresijskih metod, ki so navedene v tabeli 5.3. Poleg naziva regresijske metode v paketu WEKA so navedeni tudi uporabljeni parametri in kratica, ki jo bomo uporabljali v nadaljevanju.

| **Naziv metode v paketu WEKA** | **Metoda strojnega učenja** | **Kratica** |
|---|---|---|
| LinearRegression | Linearna regresija | LR |
| | Parametri:<br>- metoda izbire atributov: M5<br>- odstranitev kolinearnih atributov | |
| LibSVM | Metoda podpornih vektorjev | LSVM |
| | Parametri:<br>- tip SVM: nu-SVR<br>- tip jedra: sigmoidna<br>- stopnja jedra: 3<br>- normalizacija: da<br>- ocenitev verjetnosti: da<br>- eps: 0.001 | |
| MultilayerPerceptron | Večnivojski perceptron | MP |
| | Parametri:<br>- stopnja učenja: 0.16<br>- momentum: 0.02 | |
| SMOreg | Metoda podpornih vektorjev | SMOR |
| | Parametri:<br>- parameter kompleksnosti c: 1.0<br>- tip filtra: normalizacija učne množice<br>- jedro: PolyKernel | |
| REPTree | Regresijsko drevo | REPT |
| | Parametri:<br>- obrezovanje: da | |
| M5P | Drevo regresijskih modelov | M5P |
| IBk | Metoda k-najbližjih sosedov | IBk |
| | Parametri:<br>- število najbližjih sosedov: 2<br>- algoritem iskanja najbližjih sosedov: LinearNNSearch z evklidsko razdaljo | |
| KStar | Lokalna metoda | K* |
| LWL | Lokalno utežena regresija | LWL |
| | Parametri:<br>- število kNN: vsi<br>- kvalifikator: SMOreg (metoda podpornih vektorjev)<br>- algoritem iskanja najbližjih sosedov: LinearNNSearch z evklidsko razdaljo | |
| AdditiveRegression | Aditivna regresija | AR |
| | Parametri:<br>- število iteracij: 10<br>- kvalifikator: SMOreg | |
| Bagging | Bagging metoda | B |
| | Parametri:<br>- število iteracij: 10<br>- kvalifikator: SMOreg | |
| RandomSubSpace | Metoda naključnih podprostorov | RSS |
| | Parametri:<br>- število iteracij: 10<br>- kvalifikator: SMOreg<br>- velikost pod-prostora v odstotkih št. atributov: 50 | |
| Vote | Glasovanje | V |
| | Parametri:<br>- kombinacija rezultatov regresijskih metod: povprečje<br>- regresijske metode: SMOReg, LWL in MultilayerPerceptron | |

Tabela 5.2: Uporabljene regresijske metode

## 5.4 Dodajanje proteinskih atributov

Podatki proteinske baze so bili pridobljeni na različni skupini vzorcev kot bazah miRNA in CEMS, zato proteinske baze nismo mogli neposredno uporabiti pri napovedovanju vrednosti PD in DRF. Odločili smo se, da bomo proteinske atribute vključili v primarni bazi miRNA in CEMS s pomočjo Mann-Whitneyevega testa.

Mann-Whitneyev test [4, 6] se uporablja za testiranje enakosti porazdelitev statističnih spremenljivk $X$ in $Y$ pri dveh neodvisnih vzorcih, pri čemer sta ničelna in alternativna hipoteza sledeči:

- $H_0$: oba vzorca imata enako mediano.
- $H_1$: mediana vzorca 1 je različna od mediane vzorca 2.

Naj bosta $X_1, X_2, \ldots, X_n$ in $Y_1, Y_2, \ldots, Y_m$ vzorca velikosti $n$ in $m$. Oba vzorca združimo in uredimo po velikosti v zaporedje $Z_1 <= Z_2 <= \ldots <= Z_{m+n}$ ter ga rangiramo.

$R$ je vsota vseh rangov, ki jih zavzame spremenljivka $X$. Kadar se v zgornjem zaporedju pojavi vrednost $Y_j$ pred $X_i$ pravimo, da je nastopila inverzija. Število inverzij je

$$U = n * m + \frac{n(n+1)}{2} - R \tag{5.1}$$

Izračunamo statistiko z:

$$z = \frac{U - \frac{n*m}{2}}{\sqrt{\frac{n*m(n+m+1)}{12}}} \tag{5.2}$$

Če je vrednost $|z|$ večja ali enaka kritični vrednosti $\alpha$, $H_0$ zavrnemo. Iz statistike $z$ lahko izračunamo verjetnost $p$, da ničelna hipoteza drži.

Postopek dodajanja proteinskih atributov v primarni bazi miRNA in CEMS je sledeč:

1. Na obeh proteinskih bazah (Antibody in LCMS) naredimo izbiro atributov po metodi C-GS, s čimer dobimo ustrezno podmnožico domnevno najboljših proteinskih atributov.
2. Nato za vsako primarno bazo miRNA in CEMS naredimo novo proteinsko bazo, ki bo vsebovala najboljše povezane proteinske atribute po naslednjem postopku:

- za vsak izbran atribut iz podmnožice proteinskih atributov poiščemo vse povezane atribute iz izbrane primarne baze,
- izmed vseh povezanih atributov iz primarne baze izberemo tistega, ki ima največjo vrednost $p$ po Mann-Whitneyevem testu za izbrani atribut iz normalizirane proteinske in normalizirane primarne baze,
- nato za vsako normalizirano vrednost vzorca iz izbrane primarne baze poiščemo $k$ najbližjih normaliziranih vrednosti iz proteinskega atributa ter povprečimo dejanske vrednosti proteinskega atributa,
- povprečen proteinski atribut dodamo v novo proteinsko bazo.

3. Na primarni bazi naredimo izbiro atributov z izbrano metodo.
4. Združimo primarno bazo in novo proteinsko bazo.

## 5.5 Preverjanje uspešnosti modelov

Za preverjanje uspešnosti smo uporabili metodo izloči enega [7] (*angl. leave-one-out*), ki se pogosto uporablja v primerih, ko imamo majhno število vzorcev. Pri tej metodi vsak vzorec izločimo iz učne množice, model naučimo na vseh preostalih vzorcih ter ga uporabimo za reševanje izločenega vzorca. Uspešnost hipoteze, zgrajene na vseh vzorcih, ocenimo kot povprečno uspešnost vseh zgrajenih hipotez na izločenem vzorcu.

Uspešnost regresijskih modelov smo ocenili s korenom relativne srednje kvadratne napake:

$$RRMSE = \sqrt{\frac{(p_1 - a_1)^2 + ... + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + ... + (a_n - \bar{a})^2}} \tag{5.3}$$

kjer so $p_1, p_2, ..., p_n$ napovedane vrednosti na testnih primerih, $a_1, a_2, ..., a_n$ dejanske vrednosti na testnih primerih, $\bar{a}$ pa povprečna vrednost na učni množici.

# Poglavje 6

# Rezultati

## 6.1 Primerjava regresijskih modelov na posameznih bazah

Vsako regresijsko metodo smo najprej preizkusili pri napovedovanju ciljnih vrednosti na bazi miRNA, bazi CEMS (bazi metabolitov) ter združeni bazi (ZB). RRMSE je izračunan po metodi izloči-enega na učni množici.

Tabela 6.1 in Slika 6.1 prikazujeta RRMSE regresijskih metod pri ločenem napovedovanju vrednosti PD s predhodno izbiro atributov po metodi C-LFS na bazi miRNA, bazi CEMS in ZB.

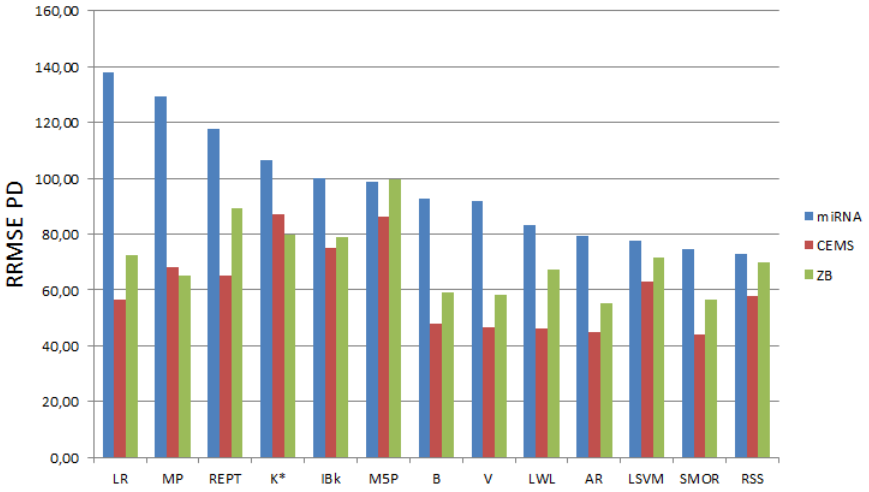
| | **LR** | **LSVM** | **MP** | **SMOR** | **REPT** | **M5P** | **IBk** | **K*** | **LWL** | **AR** | **B** | **RSS** | **V** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **miRNA** | 137,93 | 77,53 | 129,47 | 74,68 | 117,75 | 98,67 | 99,81 | 106,60 | 83,33 | 79,15 | 92,76 | 72,99 | 91,73 | **97,11** |
| **CEMS** | 56,38 | 63,08 | 68,02 | 44,04 | 65,14 | 86,11 | 74,83 | 86,93 | 46,26 | 44,67 | 47,73 | 57,68 | 46,46 | **60,56** |
| **ZB** | 72,27 | 71,60 | 65,07 | 56,56 | 89,08 | 99,71 | 79,03 | 79,68 | 67,28 | 55,38 | 59,26 | 69,74 | 58,13 | **70,98** |

Tabela 6.1: RRMSE pri napovedovanju vrednosti PD. Z zeleno barvo so označeni trije najboljši modeli, z rdečo pa trije najslabši modeli.

Opazimo lahko, da pri napovedovanje vrednosti PD vse metode razen K* in MP dosegajo nižji RRMSE na bazi CEMS. V povprečju je RRMSE na bazi CEMS znašal 60,56%, na ZB 70,98%, na bazi miRNA pa kar 97,11%, zato smo v nadaljevanju pri napovedovanju vrednosti PD uporabili le bazo CEMS.

Tabela 6.1 in Slika 6.2 prikazujeta RRMSE regresijskih metod pri ločenem napovedovanju vrednosti DRF s predhodno izbiro atributov po metodi C-LFS na bazi miRNA, bazi CEMS in ZB.

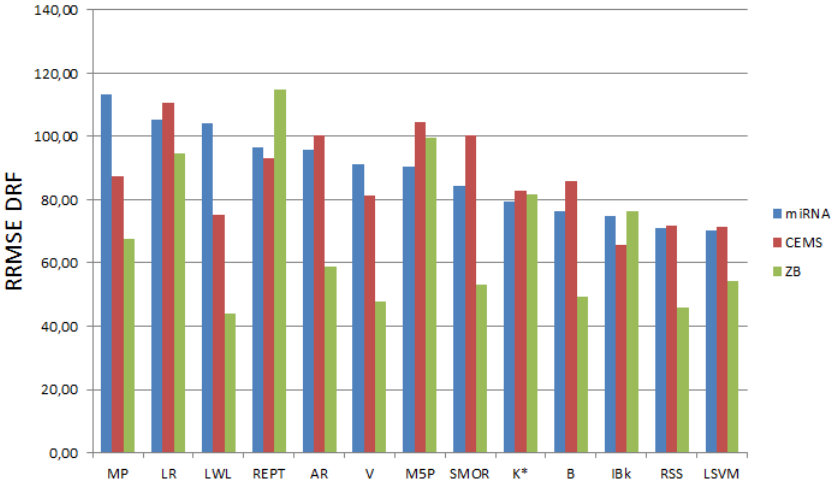
Ker so regresijske metode dosegle najnižji povprečni RRMSE pri napovedovanju vrednosti DRF na ZB, smo v nadaljevnaju pri napovedovanju vred-




Slika 6.1: Graf RRMSE pri napovedovanju vrednosti PD posameznih baz.

| | LR | LSVM | MP | SMOR | REPT | M5P | IBk | K* | LWL | AR | B | RSS | V | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **miRNA** | 105,42 | 70,10 | 113,34 | 84,52 | 96,41 | 90,41 | 74,83 | 79,38 | 104,24 | 95,91 | 76,41 | 71,06 | 91,25 | **88,71** |
| **CEMS** | 110,59 | 71,46 | 87,52 | 100,32 | 92,94 | 104,70 | 65,60 | 82,75 | 75,33 | 100,30 | 86,02 | 71,76 | 81,12 | **86,95** |
| **ZB** | 94,76 | 54,18 | 67,60 | 53,23 | 114,98 | 99,57 | 76,18 | 81,63 | 44,06 | 58,71 | 49,39 | 45,86 | 47,94 | **68,32** |

Tabela 6.2: RRMSE pri napovedovanju vrednosti DRF. Z zeleno barvo so označeni trije najboljši modeli, z rdečo pa trije najslabši modeli.




Slika 6.2: RRMSE pri napovedovanju vrednosti DRF posameznih baz

nosti DRF uporabili združeno bazo.

## 6.2 Primerjava kombinacij regresijskih modelov in metod za izbiro atributov

Sledil je preizkus vseh regresijskih metod v kombinaciji z različnimi metodami za izbiro atributov.

Tabela 6.2 prikazuje RRMSE preizkušenih parov metod za izbiro atributov in metod strojnega učenja pri napovedovanju PD in DRF. Vsak model je ločeno napovedal vrednosti PD in DRF, RRMSE obeh napovedi pa smo povprečili.

| | LR | LSVM | MP | SMOR | REPT | M5P | IBk | K* | LWL | AR | B | RSS | V | Povpr. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **BREZ** | 93,19 | 95,62 | | 92,94 | 135,48 | 101,39 | 104,68 | 137,58 | 89,50 | 92,99 | 88,96 | 96,11 | | **102,58** |
| **RF** | 73,46 | 76,69 | 91,36 | 73,83 | 110,63 | 87,07 | 75,94 | 90,17 | 75,82 | 75,13 | 76,19 | 74,78 | 76,96 | **81,39** |
| **C-GS** | 74,30 | 69,60 | 77,08 | 69,07 | 117,33 | 94,48 | 72,80 | 88,72 | 71,53 | 72,59 | 70,49 | 71,67 | 70,17 | **78,45** |
| **C-BF** | 90,20 | 66,17 | 68,29 | 56,62 | 97,54 | 91,01 | 69,50 | 82,54 | 65,47 | 54,58 | 64,56 | 68,81 | 61,32 | **72,05** |
| **C-LFS** | 75,57 | 58,63 | 67,81 | 48,63 | 90,06 | 92,84 | 75,51 | 84,28 | 45,16 | 51,69 | 48,56 | 51,77 | 47,20 | **64,44** |
| **C-SSFS** | 79,02 | 62,50 | 57,58 | 53,61 | 88,44 | 90,82 | 71,66 | 86,90 | 47,90 | 50,78 | 53,86 | 60,28 | 49,97 | **65,64** |
| **MFMW** | 84,85 | 64,25 | 67,17 | 60,10 | 85,40 | 88,39 | 68,35 | 77,39 | 63,56 | 60,45 | 59,03 | 63,84 | 59,65 | **69,42** |
| | **81,51** | **70,49** | **71,55** | **64,97** | **103,55** | **92,28** | **76,92** | **92,51** | **65,56** | **65,46** | **65,95** | **69,61** | **60,88** | |

Tabela 6.3: RRMSE preizkušenih kombinacij metod za izbiro atributov in regresijskih metod. Z zeleno barvo so označeni trije najboljši modeli, z rdečo pa trije najslabši modeli.

Opazimo lahko, da je veliko atributov v bazi miRNA in CEMS nepotrebnih ali šumnih, saj vse regresijske metode na celotnih podatkih dosegale slabši RRMSE, kot pa na podatkih po izvedeni izbiri atributov.

Pri metodah za izbiro atributov se je najbolje obnesla metoda C-LFS, le malo slabša pa je bila metoda C-SSFS. Pričakovali smo nekoliko boljše rezultate pri metodi MFMW, saj naj bi bile metode notranje optimizacije zanesljivejše kot filter metode izbire atributov. Je pa ta metoda izbire atributov dosegla najmanjši RRMSE v primerjavi z drugimi metodami izbire atributov pri regresijskih metodah, ki so imele v povprečju nekoliko višji RRMSE - torej pri regresijskih drevesih in lokalnih metodah.

Pri regresijskih metodah je imela najmanjši povprečni RRMSE metoda glasovanja, ki vzame povprečje treh regresijskih metod: SMOR, LWL in MP. Sledijo metode SMOR, AR in LWL. Najvišji povprečni RRMSE je imela metoda regresijskih dreves.

Tabela 6.4 prikazuje 10 najboljših modelov glede na izračunani RRMSE na učni množici s pripadajočim rezultatom na tekmovanju (4.4) ter njegovim

odstopanjem od najboljšega rezultata na tekmovanju, ki je znašal 0,04146. RRMSE je izračunan po metodi izloči-enega na učni množici, rezultat pa na štirih testnih primerih. Podatki so urejeni po naraščajočem RRMSE.

| | **Regresijska metoda** | **Metoda izbire atributov** | **RRMSE** | **Rezultat** | **Odstopanje od najboljšega rezultata** |
|---|---|---|---|---|---|
| 1 | LWL | C-LFS | 45,1612 | 0,07140 | 72,20% |
| 2 | V | C-LFS | 47,2002 | 0,07515 | 81,30% |
| 3 | LWL | C-SSFS | 47,9012 | 0,08289 | 99,90% |
| 4 | B | C-LFS | 48,5613 | 0,05481 | 32,20% |
| 5 | SMOR | C-LFS | 48,6335 | 0,06241 | 50,50% |
| 6 | V | C-SSFS | 49,9667 | 0,07394 | 78,30% |
| 7 | AR | C-SSFS | 50,7845 | 0,07464 | 80,00% |
| 8 | AR | C-LFS | 51,6892 | 0,06611 | 59,50% |
| 9 | RSS | C-LFS | 51,7695 | 0,05100 | 23,00% |
| 10 | SMOR | C-SSFS | 53,6121 | 0,06372 | 53,70% |

Tabela 6.4: Rezultati na osnovnih bazah

Opazimo lahko, da najmanjši RRMSE dosega metoda LWL in SMOR ter metode, ki kombinirajo različne regresijske metode. Metoda LWL je v kombinaciji z C-LSF metodo izbire atributov dosegla najnižji RRMSE in rezultat, ki je za 72% boljši od najboljšega objavljenega rezultata na tekmovanju. Iz rezultatov je razvidno tudi, da manjši RRMSE na učni množici še ne pomeni boljšega rezultata na testnih podatkih. Tako so imele metode glasovanja in metoda LWL v kombinaciji z metodo C-SSFS izbire atributov boljši rezultat na tekmovanju, a višji RRMSE. Razlog za odstopanje je v izbiri metode ocenjevanja uspešnosti regresijskih modelov. Za preverjanje uspešnosti modelov smo izbrali metodo izloči-enega, ki se uporablja v primerih, ko je na voljo malo podatkov. A tudi ta metoda postane nezanesljiva, ko imamo za učenje le 20 primerov.

Tabela 6.5 prikazuje napovedane in dejanske vrednosti PD in DRF najboljšega modela iz tabele 6.4 na štirih testnih vzorcih.

| | **Dejanski PD** | **Napovedani PD** | **Dejanski DRF** | **Napovedani DRF** |
|---|---|---|---|---|
| Vzorec 1 | 0,23529 | 0,41061 | 0,32000 | 0,31356 |
| Vzorec 2 | 1,00000 | 0,70690 | 0,04000 | 0,17639 |
| Vzorec 3 | 0,17647 | 0,24683 | 0,04000 | 0,11974 |
| Vzorec 4 | 0,00000 | 0,25012 | 0,00000 | 0,24353 |

Tabela 6.5: Napovedane in dejanske vrednosti ciljnih spremenljivk

## 6.3 Rezultati regresijskih metod po dodajanju povezanih proteinskih atributov

Na koncu smo poskušali rezultat izboljšati še z dodajanjem povezanih proteinskih atributov. Na obeh proteinskih bazah Antibody in LCMS smo naredili izbiro atributov z metodo C-GS, s čimer smo dobili 9 potencialno najboljših atributov iz baze Antibody in 7 iz baze LCMS. Za te atribute smo poiskali vse povezane atribute iz baz miRNA in CEMS. Zanje smo z Mann-Whitneyevem testom izračunali verjetnost $p$, da imajo normalizirane vrednosti atributov enako mediano s čimer smo ocenili podobnost atributov. Bolj podroben postopek je opisan v Poglavju 5.4.

V tabeli 6.6 so navedeni najboljši proteinski atributi izbrani z metodo C-GS, indeksi vseh povezanih atributov iz baz miRNA in CEMS ter v oglatih oklepajih še izračunana vrednost $p$.

| **Proteinski atributi** | | **Povezani miRNA atributi** | **Povezani CEMS atributi** |
|---|---|---|---|
| **Antibody** | C9603 | | |
| | D2178 | 365 [0,243] | |
| | G9269 | **164[0,803]**, 407[0,1928] | |
| | H4163 | 347[0,2257] | |
| | L7391 | | |
| | M7693 | | |
| | P1870 | 148[0,301], 487[0,0813], 514[0,4957], 538[0,0032], 580[0,0476] | |
| | R6028 | | 803[0,3675], 817[0,0034], 819[0,2455] |
| | R6878 | 369[0,0157] | |
| **LCMS** | IPI00305286 | 181[0,0171], 213[0,0171] | |
| | IPI00913924 | 536[0,0091], 588[0,4399], 823[0,0428] | |
| | IPI00025840 | 602[0,209] | |
| | IPI00296608 | | |
| | IPI00013976 | 822[0,1072] | |
| | IPI00439344 | 823[0,3883] | |
| | IPI00220642 | 507[0,0058] | **501[0,5222]** |

Tabela 6.6: Potencialno najboljši proteinski atributi s povezanimi atributi iz baz miRNA in CEMS. Odebeljena atributa sta dosegla vrednost $p$ večjo od 0,5

Pri testiranju smo preizkusili več različnih vrednosti parametrov $k$ in $p$. Za število najbližjih sosedov $k$ pri povprečenju vrednosti proteinskega atributa smo preizkusili vrednosti od 1 do 4. K osnovnim bazam smo nato dodali le atribute, ki imajo vrednost $p$ večjo od določene meje. Za mejo smo preizkusili vrednosti od 0,5 do 0,9. Regresijske metode so imele najnižji RRMSE pri vrednosti parametra $k = 2$ in $p = 0,5$. Tem kriterijem sta ustrezala le dva povezana atributa iz baz CEMS in miRNA, ki sta v Tabeli 6.6 odebeljena. Za

proteina IPI00220642 iz baze LCMS in G9269 iz baze Antibody smo izračunali vrednosti po postopku, opisanem v Poglavju 5.4.

Nato smo preverili kako dodajanje izbranih proteinskih atributov vpliva na napovedovanje vrednosti PD in DRF. Izbrana proteinska atributa smo dodali k ZB pri napovedovanju vrednosti DRF in k bazi CEMS pri napovedovanju vrednosti PD. Obe bazi sta imeli že opravljeno izbiro atributov po metodi L-LFS. Povprečni RRMSE vseh regresijskih metod se je pri napovedovanju vrednosti DRF znižal z 68,32 % na 65,55 %, pri napovedovanju vrednosti PD pa povečal z 60,56% na 62,67 %. Proteinske atribute smo zato v nadaljevanju uporabili le za izboljšanje napovedovanja vrednosti DRF.

Tabela 6.7 prikazuje 10 najboljših regresijskih modelov po dodajanju povprečenih proteinskih atributov glede na izračunani RRMSE s pripadajočim rezultatom na tekmovanju ter njegovim odstopanjem od najboljšega rezultata na tekmovanju. RRMSE je izračunan po metodi izloči-enega na učni množici, rezultat pa na štirih testnih primerih. Podatki so urejeni po naraščajočem RRMSE. Poleg tega so za primerjavo prikazani tudi rezultati na bazah brez dodajanja proteinskih atributov ter izboljšava rezultata po dodajanju proteinskih atributov v odstotkih.

| **Regresijska metoda** | | **Metoda izbire atributov** | **Baze z dodanimi proteinskimi atributi** | | | **Baze brez dodanih atributov** | | **Izboljšava rezultata** |
|---|---|---|---|---|---|---|---|---|
| | | | **RRMSE** | **Rezultat** | **Odstopanje od najboljšega rezultata** | **RRMSE** | **Rezultat** | |
| 1 | LWL | C-LFS | 40,3245 | 0,07540 | 81,90% | 45,1612 | 0,07140 | 5,60% |
| 2 | V | C-LFS | 42,9600 | 0,08180 | 97,30% | 47,2002 | 0,07515 | 8,86% |
| 3 | SMOR | C-LFS | 46,3289 | 0,07090 | 71,00% | 48,6335 | 0,06241 | 13,61% |
| 4 | B | C-LFS | 47,0314 | 0,06285 | 51,60% | 48,5613 | 0,05481 | 14,67% |
| 5 | B | C-SSFS | 49,4438 | 0,06824 | 64,60% | 53,8617 | 0,05943 | 14,83% |
| 6 | LWL | C-SSFS | 49,9992 | 0,08456 | 104,00% | 47,9012 | 0,08289 | 2,01% |
| 7 | V | C-SSFS | 50,2557 | 0,07999 | 92,90% | 49,9667 | 0,07394 | 8,19% |
| 8 | AR | C-LFS | 50,4313 | 0,07326 | 76,70% | 51,6892 | 0,06611 | 10,82% |
| 9 | SMOR | C-SSFS | 51,0716 | 0,07671 | 85,00% | 53,6121 | 0,06372 | 20,39% |
| 10 | AR | C-SSFS | 52,6958 | 0,07974 | 92,30% | 51,6892 | 0,06611 | 20,61% |

Tabela 6.7: Rezultati po dodajanju proteinskih atributov

Vseh deset najboljših metod glede na dosežen RRMSE na učni množici je v primerjavi z rezultati na osnovnih bazah po dodajanju doseglo boljši rezultat na testnih podatkih. Najboljša metoda na učnih podatkih glede na RRMSE je še vedno lokalno utežena regresija v kombinaciji z metodo C-LFS. Ta metoda je v primerjavi z modelom na osnovnih bazah dosegla za 5,6 % boljši rezultat na testni množici, nižji pa ima tudi RRMSE. V nasprotju z najboljšim modelom pa se je pri nekaterih modelih RRMSE na učnih množicah po dodajanju proteinskih atributov povišal (npr. pri metodi glasovanja). Podobno kot na

osnovnih bazah je tudi tu prihajalo do neskladnosti med rezultati na učnih množicah in rezultatih na testnih množicah.

Tabela 6.8 prikazuje napovedane in dejanske vrednosti PD in DRF najboljšega modela iz Tabele 6.7 na štirih testnih vzorcih po dodajanju proteinskih atributov.

| | **Dejanski PD** | **Napovedani PD** | **Dejanski DRF** | **Napovedani DRF** |
|---|---|---|---|---|
| Vzorec 1 | 0,23529 | 0,41061 | 0,32000 | 0,32374 |
| Vzorec 2 | 1,00000 | 0,70690 | 0,04000 | 0,13505 |
| Vzorec 3 | 0,17647 | 0,24683 | 0,04000 | 0,07769 |
| Vzorec 4 | 0,00000 | 0,25012 | 0,00000 | 0,22325 |

Tabela 6.8: Napovedane in dejanske vrednosti ciljnih spremenljivk po dodajanju proteinskih atributov

# Poglavje 7

# Zaključek

V diplomskem delu smo poskušali izdelati regresijski model, ki bo kar najbolj natančno določal stopnjo obolenja obstruktivne nefropatije. V ta namen smo preizkusili različne obstoječe metode za izbiro atributov ter implementirali metodo MFMW, ki kombinira filter metode z metodami notranje optimizacije. Posamezne metode za izbiro atributov smo ocenili v kombinaciji z različnimi regresijskimi metodami, ki jih ponuja WEKA. Izdelali smo tudi postopek, ki s pomočjo bioloških baz poišče najboljše povezane proteinske atribute in jih doda v učno in testno množico.

Na podatkovni bazi, ki ni vsebovala proteinskih atributov, se je najbolje obnesla lokalno utežena regresija (LWL). V kombinaciji s metodo C-LFS izbire atributov je dosegla najnižji RRMSE, ki je znašal 45,16 %. Rezultat na testni množici znaša 0,07140 in za 72,20 % presega najboljši objavljen rezultat na tekmovanju. Po dodajanju proteinskih atributov v podatkovno bazo je lokalno utežena regresija (LWL) v kombinaciji z metodo C-LFS izbire atributov še vedno dosegala najnižji RRMSE. RRMSE se je v primerjavi s podatkovno bazo brez proteinskih atributov znižal na 40,32 %. Rezultat se prav tako izboljšal. Z 0,07540 za 81,90 % presega najboljši rezultat na tekmovanju. Čeprav so podatki o potencialnih povezavah med atributi šumni in nepopolni, nam je uspelo z izbiro najboljših proteinskih atributov in postopkom povezovanja diagnostični model še izboljšati.

Z implementiranim modelom smo presegli rezultate na tekmovanju, vendar bi bilo potrebno v nadaljnjem raziskovanju preizkusiti še kako bolj zanesljivo metodo ocenjevanja uspešnosti. V postopku smo uporabili metodo izločienega, ki se pogosto uporablja, ko imamo na voljo majhno število vzorcev, a je tudi ta nezanesljiva, ko je teh le 20. Tako so nekateri modeli dosegali boljši rezultat, a višji RRMSE in smo jih tako ocenili kot slabše. Metoda LWL

je imela po dodajanju proteinskih atributov v kombinaciji s metodo C-SSFS izbire atributov rezultat na testnih podatkih 0,08456, a kar za 10 % slabši RRMSE glede na najboljši model na učni množici. Ena takih metod, ki se izkaže ravno v primerih, ko je vzorev izredno malo, je metoda *razmnoževanja učnih primerov* (*angl. bootstrapping*).

# Slike

2.1 Model MFMW metode za izbiro atributov . . . . . 6

3.1 Večnivojski perceptron . . . . . 11

4.1 Prikaz odvisnosti med biološkimi nivoji . . . . . 16

6.1 Graf RRMSE pri napovedovanju vrednosti PD posameznih baz. 24
6.2 RRMSE pri napovedovanju vrednosti DRF posameznih baz . . . 24

# Tabele

4.1 Podatkovne baze . . . . . 15

5.1 Preizkušene metode za izbiro atributov . . . . . 19
5.2 Uporabljene regresijske metode . . . . . 20

6.1 RRMSE pri napovedovanju vrednosti PD. Z zeleno barvo so označeni trije najboljši modeli, z rdečo pa trije najslabši modeli. 23
6.2 RRMSE pri napovedovanju vrednosti DRF. Z zeleno barvo so označeni trije najboljši modeli, z rdečo pa trije najslabši modeli. 24
6.3 RRMSE preizkušenih kombinacij metod za izbiro atributov in regresijskih metod. Z zeleno barvo so označeni trije najboljši modeli, z rdečo pa trije najslabši modeli. . . . . 25
6.4 Rezultati na osnovnih bazah . . . . . 26
6.5 Napovedane in dejanske vrednosti ciljnih spremenljivk . . . . . 26
6.6 Potencialno najboljši proteinski atributi s povezanimi atributi iz baz miRNA in CEMS. Odebeljena atributa sta dosegla vrednost $p$ večjo od 0,5 . . . . . 27
6.7 Rezultati po dodajanju proteinskih atributov . . . . . 28
6.8 Napovedane in dejanske vrednosti ciljnih spremenljivk po dodajanju proteinskih atributov . . . . . 29

# Literatura


[1] I. Kononenko, *Strojno učenje*, Ljubljana, Založba fakultete za elektrotehniko in fakultete za računalništvo in informatiko, 2005

[2] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, New Jersey: Wiley-Interscience, 2004, pogl. 2.5

[3] Y. Leung, Y. Hung, *„A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification“*, Bioinformatics, št. 7, str. 108-117, 2010

[4] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, 3. izdaja, Boca Raton: Chapman and Hall/CRC, 2004

[5] M. Skurichina, R. P. W. Duin, *"Bagging, Boosting and the Random Subspace Method for Linear Classifiers"*, Pattern Analysis Applications, št.5, str. 121-135, 2002

[6] M. F. Triola, *Elementary statistic*, 11. izdaja, Boston: Addison Wesley, 2009, pogl. 13

[7] I. H. Witten, *Data Mining: Practical machine learning tools and techniques*, 3. izdaja, Burlington: Morgan Kaufmann, 2011

[8] (2011) Spletna stran tekmovanja. Dostopno na: http://tunedit.org/challenge/ON

[9] (2011) Programski paket WEKA. Dostopno na: http://www.cs.waikato.ac.nz/ml/weka/

[10] (2012) Javanska statistična knjižnica JSC. Dostopno na: http://www.jsc.nildram.co.uk/