UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Miha Sedej

Analiza lastnosti pločevine z metodami podatkovnega rudarjenja

DIPLOMSKO DELO
NA UNIVERZITETNEM ŠTUDIJU

Mentor: izr. prof. dr. Uroš Lotrič

Ljubljana, 2012

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: 01768/2011
Datum: 02.09.2011

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

Kandidat: **MIHA SEDEJ**

Naslov: **ANALIZA LASTNOSTI PLOČEVINE Z METODAMI PODATKOVNEGA RUDARJENJA**

STEEL SHEET ANALYSIS WITH DATA MINING MODELS

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

Pri predelavi kovin na kvaliteto končnega izdelka močno vpliva kvaliteta same pločevine. Z metodami podatkovnega rudarjenja analizirajte mehanske in kemijske lastnosti pločevine in poiščite povezave med njimi. Ocenite smiselnost vzpostavitve laboratorijskega informacijskega sistema za avtomatizacijo nadaljnjih podobnih analiz.

Mentor:
prof. dr. Uroš Lotrič

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

Dekan:
prof. dr. Nikolaj Zimic

IZJAVA O AVTORSTVU

diplomskega dela

Spodaj podpisani Miha Sedej,

z vpisno številko 63060285,

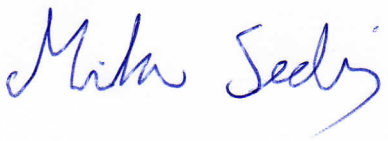
sem avtor diplomskega dela z naslovom:

Analiza lastnosti pločevine z metodami podatkovnega rudarjenja

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom izr. prof. dr. Uroša Lotriča
- 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 14.3.2012 Podpis avtorja:



Zahvala

Mentorju izr. prof. dr. Urošu Lotriču se iskreno zahvaljujem za strokovno pomoč in vse napotke pri pisanju diplomske naloge.

Zahvaljujem se staršem in sestri Kaji za podporo pri študiju ter dekletu Urški za potrpežljivost.

Zahvaljujem se tudi vsem zaposlenim v družbi Hidria, ki so na kakršenkoli način pripomogli k nastanku dela, še posebej Kristjanu Hladniku in Marku Kovaču za ves vložen trud in strokovno pomoč.

But if the technological Singularity can happen, it will.

(Vernor Vinge)

Kazalo

Povzetek **1**

Abstract **2**

1 Uvod **3**
1.1 Motivacija . . . 3
1.2 Predstavitev podjetja . . . 3
1.3 Predelava kovin . . . 4
1.4 Opis problematike, obdelane v diplomskem delu, in zgradba . . 6

2 Teorija **8**
2.1 Mere za oceno pomembnosti atributov . . . 9
2.1.1 RReliefF . . . 9
2.1.2 Metoda razlike standardnega odklona . . . 9
2.1.3 Metoda MARS . . . 10
2.1.4 Naključni gozdovi . . . 10
2.1.5 Permutacijski test . . . 11
2.2 Modeliranje . . . 12
2.2.1 Linearna regresija . . . 12
2.2.2 Metoda k najbližjih sosedov . . . 13
2.2.3 Metoda MARS . . . 13
2.2.4 Odločitvena drevesa . . . 15
2.2.5 Metoda podpornih vektorjev . . . 17
2.2.6 Naključni gozdovi . . . 18
2.2.7 Umetne nevronske mreže . . . 18
2.3 Vizualizacija podatkov . . . 20

3 Podatki **21**
3.1 Predprocesiranje . . . 23
3.2 Pregled . . . 24

3.3 Vizualizacija podatkov 24

4 Ocena pomembnosti atributov **29**
4.1 Rezultati 31
4.1.1 Podatkovna množica 1 31
4.1.2 Podatkovna množica 2 32
4.1.3 Podatkovna množica 3 33

5 Napovedovanje trdote **34**
5.1 Rezultati 36
5.1.1 Nastavitve prostih parametrov 36
5.1.2 Množica 1 37
5.1.3 Množica 2 39
5.1.4 Množica 3 43

6 Zaključek **45**
6.1 Ugotovitve 45
6.2 Nadaljnje delo 46

Literatura **47**

Povzetek

Dandanes vse več podatkov shranjujemo v elektronski obliki, ki nam omogoča enostaven dostop, iskanje in obdelavo shranjenih podatkov. Na vseh področjih našega življenja nastajajo različne zbirke podatkov, iz katerih je s sodobnimi pristopi podatkovnega rudarjenja mogoče izluščiti nove informacije. V podjetjih lahko pravilna uvedba takšnih metod prinese neposredno ali posredno ekonomsko korist, zato je interes za omenjeno področje velik.

V diplomski nalogi smo se v sodelovanju s Hidria Inštitutom za materiale in tehnologijo osredotočili na zbirko meritev trdote, ostalih mehanskih in kemijskih lastnosti pločevine. Zadali smo si cilj napovedati trdoto pločevine iz preostalih lastnosti. Naprej smo ocenili, katere mehanske in kemijske lastnosti najbolj prispevajo k napovedovanju trdote, nato pa jo z različnimi metodami poskušali napovedati. Primerjali smo razliko v natančnosti modelov, zgrajenih z vsemi znanimi lastnostmi pločevine in samo najbolje ocenjenimi. Uporabili smo metode, implementirane v programski paket Orange, ki vsebuje tudi močna orodja za vizualizacijo podatkov. Preizkusili smo, kako uspešna je metoda za samodejno iskanje vizualizacij podatkov VizRank v praksi, in prikazali nekaj najdenih zakonitosti v podatkih.

Metode za oceno najpomembnejših lastnosti so se izkazale kot učinkovite, saj med modeli z vsemi znanimi lastnostmi pločevine in samo najbolje ocenjenimi ni bilo bistvene razlike v natančnosti napovedi. Napovedovanje lastnosti trdote se je izkazalo kot nezanesljivo in tako neprimerno za kakršnokoli praktično uporabo. Metoda VizRank se je pokazala kot zelo uspešna, saj nam je praktično v trenutku prikazala zanimive povezave med podatki, ki bi jih človek zaradi velikega števila lastnosti pregledoval več ur.

Ključne besede:

podatkovno rudarjenje, pločevina, ocena pomembnosti atributov in vizualizacija, napovedovanje trdote

Abstract

There is more and more data being stored electronically nowadays to enable easy access, searching and processing of data. Various collections of data are being created for all aspects of our lives. These collections can provide us with new information if the modern techniques of data mining are applied. Companies can gain additional profit with these techniques, which is why this field of computer science is becoming more and more popular.

In this thesis, collaboration was done with the Hidria Institute of materials and technology. The focus was on chemical and mechanical properties of steel sheets. The aim was to predict the hardness of these steel sheets from the other properties. Firstly, mechanical and chemical properties were determined, to conclude which contribute most to the prediction of hardness. Using different data mining methods, this data was then used as training samples for further predictions. Models consisting of all properties and only the best determined were compared. The Orange software package was used for data mining, which also provides a set of tools for data visualization. The performance of the method for automatic data visualizations search VizRank was tested in practice. Some of the most interesting visualizations found in data were shown.

Methods for determining properties turned out to be useful as there was only a slight difference in accuracy between models built from all properties and only best scored ones. Prediction on the hardness was however, less successful. We detected a correlation between the chemical and mechanical properties and hardness, but accuracy was poor and thus not reliable enough for practical use. VizRank turned out to be very useful as it showed interesting correlations between data almost instantly, which would take a human many hours to find.

Key words:

data mining, steel sheets, attribute selection and visualization, hardness prediction

Poglavje 1

Uvod

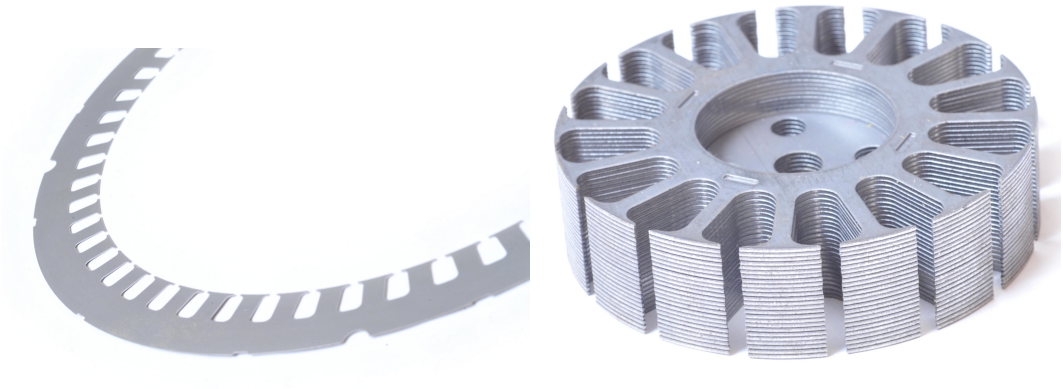
1.1 Motivacija

Dandanes računalniški sistemi, prisotni na vsakem našem koraku, zajemajo in shranjujejo najrazličnejše podatke. Spremljanje meteoroloških pojavov, nadzor prometa, telefonski klici, obisk spletnih strani so le nekateri izmed primerov, kjer se dnevno generirajo velike količine podatkov, ki jih je mogoče shraniti, analizirati in tako pridobiti nove koristne informacije. Tako na primer lahko meteorologi napovejo nevihto, prometna signalizacija se prilagaja prometu, telekomunikacijski ponudnik oblikuje najprimernejše pakete storitev in spletni oglasi oglašujejo stvari, ki obiskovalce dejansko zanimajo.

V industrijsko usmerjenih podjetjih se prav tako za zagotavljanje sledljivosti proizvodov beleži parametre tehnoloških postopkov, sestavne dele, vgrajene v proizvode, in podobno. Tako na primer lahko podjetje v primeru reklamacije proizvoda odkrije vzrok, zakaj natančno je do okvare prišlo, ostalim kupcem pa preventivno odpravi napako ali zamenja izdelek. To je zlasti pomembno na področju avtomobilske industrije, kjer so okvare zelo nezaželene, končni izdelki, avtomobili, pa imajo relativno visoko vrednost. Ni pa to edini možen način uporabe takšnih podatkov. Pravilno shranjeni parametri tehnoloških postopkov so ključni za enostavno in hitro izvajanje najrazličnejših analiz, s katerimi si lahko podjetja pridobijo pomembno konkurenčno prednost.

1.2 Predstavitev podjetja

Hidria Inštitut za materiale in tehnologijo, v sodelovanju s katerim je nastala diplomska naloga, je v lasti korporacije Hidria d.d. Skupaj s Hidria Inštitutom



(a) Lamela (b) Rotorsko jedro, sestavljeno iz več lamel

Slika 1.1: Izdelki, narejeni s postopkom hitrohodnega izsekavanja

za avtomobilsko industrijo in Hidria Inštitutom Klima spada v sklop Inovativnega centra korporacije. Sestavlja ga sedem laboratorijev, kjer se izvajajo najrazličnejše analize na vhodnih materialih in končnih produktih podjetij v skupini. Glavna dejavnost podjetij v skupini Hidria, poleg rešitev za klimatizacijo stavb, je razvoj in izdelava komponent za avtomobilsko in moto industrijo. Pomembno vlogo pri tem ima predelava kovin, saj ta predstavlja največji delež ustvarjenega prometa v korporaciji. Na podlagi tega smo se v nalogi osredotočili na analizo lastnosti pločevine kot vhodnega materiala, uporabljenega pri postopku izsekavanja izdelkov.

1.3 Predelava kovin

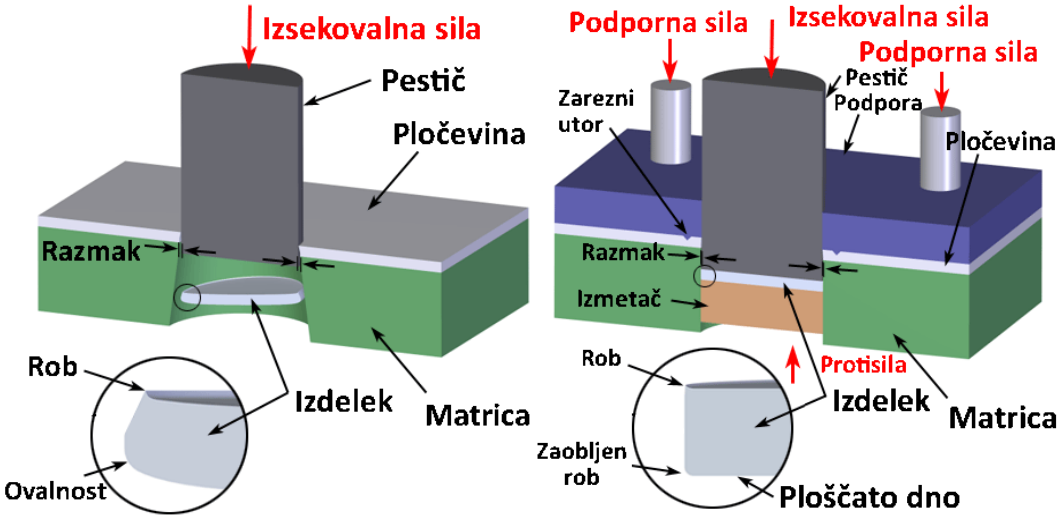
Pri izsekavanju gre trak pločevine v stiskalnico, kjer pestič oziroma nož pod pritiskom skozi matrico iz pločevine izseka izdelek glede na obliko le-te. Poleg pravilne zasnove orodja, ki ga sestavlja pestič, matrica in ostali predvsem vodilni deli, je za končni izdelek ključen tudi material, tako obdelovanca kot tudi orodja, saj ima zaradi drugačnih mehanskih lastnosti na izsekavanje velik vpliv. Orodje je običajno izdelano iz karbidne trdine ali kaljenega jekla. Poznamo tri metode izsekavanja: konvencionalno izsekavanje, hitrohodno izsekavanje rezin pločevine v obliki, podobni kolobarju, imenovanih lamele (slika 1.1a), in tako imenovano fino izsekavanje. Hidria se ukvarja predvsem s hitrohodnim izsekavanjem lamel iz elektropločevine na mehanskih stiskalnicah in finim iz-

(a) Omejilec ventila za klimatski kompresor (b) Prirobnica za izpušni sistem

Slika 1.2: Primera izdelkov, narejenih s postopkom finega izsekavanja

sekavanjem debelejših pločevin na hidravličnih stiskalnicah. Pri hitrohodnem izsekavanju lamel nastopa ena sama glavna sila, kjer pestič preko matrice deluje na material. Zaradi sile udarca se iz pločevine delno izreže in delno iztrže del v obliki posamezne rezalne stopnje orodja, ki na koncu predstavlja končni izdelek, preostanek pločevine pa je tehnološki odpadek, ki gre v reciklažo. Izdelek oziroma izsekanec zaradi deformacij materiala, ki nastanejo v postopku predelave, ni nikdar povsem ploščat in ima ostre robove. Fino izsekavanje je poseben, izpopolnjen in dražji postopek, ki delno odpravlja omenjene napake. Namesto ene tukaj nastopajo tri sile. Obdelovani material hidravlična stiskalnica stisne po celotni površini, ob mestih udarca pa je ta še dodatno stabiliziran s posebnim zareznim utorom. Prav tako pod mestom udarca pestiča ni praznina, ampak je znotraj matrice še dodatni podporni nož, ki hkrati deluje kot blažilnik in izmetač. Ta tudi podpira odrezani del in tako poskrbi, da se ob udarcu ne zvije. Rezultat so manj deformirani izdelki z bolj gladkim in manj natrganim stranskim robom. Mogoča je uporaba debelejše pločevine, a za ceno počasnejšega rezanja, ker je tak postopek v nekaterih primerih združen skupaj s tehnološkim postopkom hladnega kovanja ali iztiskavanja. Tipični predstavniki izdelkov finega izsekavanja so elementi za klimatske in izpušne sisteme (slika 1.2), za hitrohodno izsekavanje pa so lamele za statorska (slika 1.1b) in rotorska jedra elektromotorjev in generatorjev ter transformatorskih jeder.

Poleg samih mehanskih lastnosti pločevine, od katerih je odvisna oblika izdelka, igrajo pomembno vlogo tudi kemične in elektromagnetne lastnosti pločevine. Vse lastnosti so med seboj povezane, odvisne druga od druge. Velik delež prodajnih izdelkov podjetja v skupini predstavljajo statorski in rotorski paketi kot sestavni del za najrazličnejše električne stroje. Postopek izdelave je



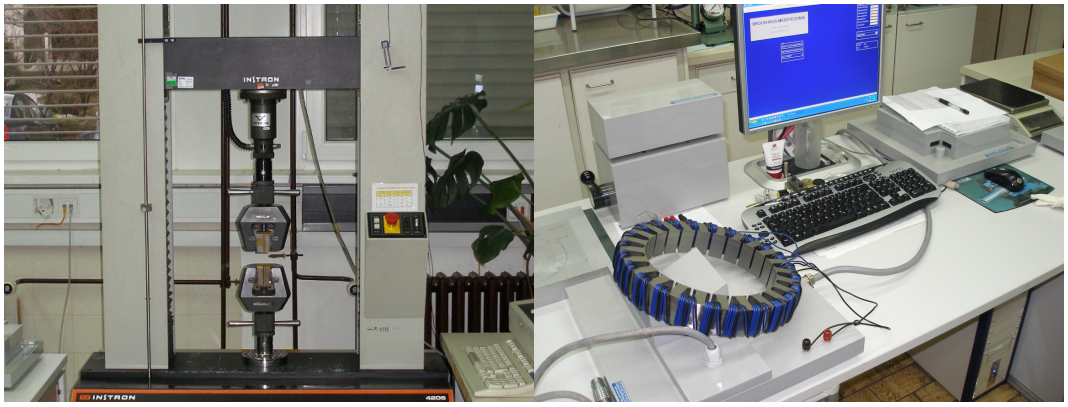
(a) Izskevanje (b) Fino izsekavanje

Slika 1.3: Primerjava postopkov izsekavanja [15]

tak, da se najprej iz pločevine izseka večje število lamel, te pa se potem zložijo ena vrh druge, v tako imenovan paket. Lamele v paketu je nato potrebno spojiti skupaj. Postopki spajanja so lahko mehanski, različni tipi varjenj ali posebno lepljenje. Mehanski postopek temelji na kovičenju. Danes je večina paketov izdelanih že kar v izsekovalnem orodju samem, kjer je sama kovica izdelana iz osnovnega materiala izdelka. Za izboljšanje elektromagnetnih lastnosti je lamele ali pakete iz nekaterih vrst pločevine potrebno še dodatno toplotno obdelati. Pri tem se nehote spremenijo tudi mehanske lastnosti, kar pa ni zaželeno. Elektromagnetne lastnosti pločevine in kvaliteta spoja lamel se nazadnje odražajo v karakteristikah električnih strojev.

1.4 Opis problematike, obdelane v diplomskem delu, in zgradba

Zaradi obsežnosti področja smo se v nalogi posvetili izključno analizi mehanskih in kemijskih lastnosti pločevine. Pri postopku izsekavanja se uporabljajo različne vrste pločevine, za vsako vrsto so predpisani točno določeni tehnično prevzemni pogoji, kjer so določene zgornje in spodnje meje fizikalnih lastnostih, ki jim mora dobavljena pločevina ustrezati. Tehnično prevzemni pogoji se lahko od dobavitelja do dobavitelja spreminjajo in so navadno predmet po-



(a) Trgalni stroj (b) Merjenje elektromagnetnih lastnosti

Slika 1.4: Merilne priprave v laboratoriju

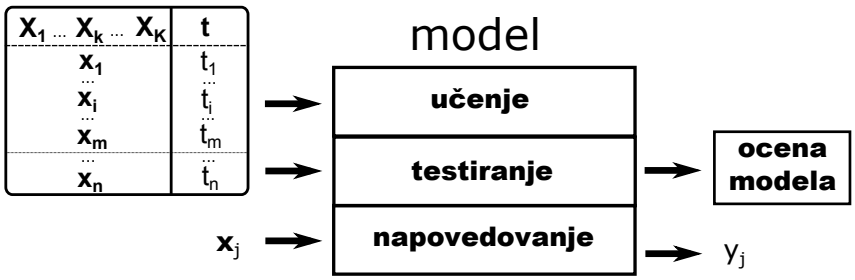
gajanj med dobaviteljem in kupcem, v tem primeru Hidrio. V nalogi smo se osredotočili na pločevino dveh vrst, M310-50A in M800-65K, ki ju podrobno opišemo v tretjem poglavju. Obe pločevini se uporabljata za izdelavo lamel po postopku hitrohodnega izsekavanja.

Drugo poglavje naloge posvetimo teoretičnim osnovam metod podatkovnega rudarjenja, uporabljenih v diplomski nalogi. V tretjem opišemo, katere izmerjene podatke o pločevini imamo na voljo, na kratko razložimo njihov pomen in na kakšen način v laboratorijih izmerijo podatke. Naštejemo predpisane tehnično prevzemne pogoje zanje in s pomočjo metode VizRank samodejno poiščemo nekaj zanimivih grafičnih prikazov. V četrtem poglavju z metodami za oceno pomembnosti atributov določimo podmnožico mehanskih in kemijskih lastnosti, ki v sebi skrivajo največ informacij za napoved lastnosti trdote pločevine. V petem poglavju modeliramo podatke z različnimi regresijskimi modeli in primerjamo natančnost modelov, zgrajenih z vsemi razpoložljivimi lastnostmi in samo podmnožico najbolje ocenjenih iz četrtega poglavja. Rezultate predstavimo kot povprečno korenjeno kvadratno napako med napovedano in testno množico ter s korelacijskim koeficientom med obema množicama. Vse metode podatkovnega rudarjenja, ki smo jih uporabili v diplomski nalogi, so bile del programskega paketa Orange [18].

Poglavje 2

Teorija

V diplomskem delu poskušamo iz nekaterih mehanskih in kemijskih lastnosti pločevine napovedati njeno trdoto. Lastnosti, iz katerih želimo napovedati drugo lastnost, imenujemo atributi, napovedano lastnost pa razred. Z omenjeno problematiko se ukvarja področje računalništva, imenovano strojno učenje. V našem primeru gre za nadzorovano atributno učenje, kjer imamo v učni množici določeno število vzorcev z znanimi atributi in razredi. Prav na podlagi teh znanih razredov naučimo naš model, da bo znal nato napovedati vrednosti tudi za nove, še nepoznane kombinacije vrednosti atributov. Običajno množico vzorcev razdelimo na učno in testno množico. Na podlagi vzorcev ocenimo, kako dobro naučen model napoveduje razred za še nepoznane vrednosti atributov. Testni vzorci ne smejo sodelovati pri učenju modela. Matriko z vrednostmi vseh atributov smo v nalogi označili z $\mathbf{X}$, vektor



Slika 2.1: Nadzorovano atributno učenje

vrednosti k-tega atributa z $\mathbf{X}_k$ in vektor vrednosti razreda s $\mathbf{t}$. Par $(\mathbf{x}_i,\ t_i)$ predstavlja i-ti vzorec, kjer je $\mathbf{x}_i$ vektor z vrednostmi vseh atributov v vzorcu, t_i pa razred v katerega je vzorec uvrščen. Na sliki 2.1 je prikazana shema nadzorovanega atributnega učenja modela, kjer imamo skupno n vzorcev, od

tega m v učni in $n - m$ v testni množici. Vektor $\boldsymbol{x}_j$ označuje nek nepoznan vhod v model, na podlagi katerega želimo napovedati vrednost razreda y_j.

2.1 Mere za oceno pomembnosti atributov

Metode za določanje pomembnosti atributov v grobem delimo v dve skupini:

1. kratkovidne, ki kot pomembne razpoznajo samo atribute, ki neposredno vplivajo na razred, in
2. nekratkovidne, ki prepoznajo tudi med seboj močne odvisne atribute.

Iz prve skupine smo uporabili metodo razlike standardnega odklona, iz druge pa metode RReliefF, MARS in naključne gozdove.

2.1.1 RReliefF

Mera za oceno kvalitete atributov ReliefF [1] je izboljšana različica mere Relief [3]. Atribute ocenjuje v odvisnosti od ostalih, zato spada v skupino nekratkovidnih mer. Celotna družina algoritmov Relief deluje po principu iskanja podobnih vzorcev z enakimi in različnimi vrednostmi razredov. Atributom, ki imajo za enake vrednosti razredov različne vrednosti, niža oceno, tistim, ki imajo za različne vrednosti razredov različne vrednosti, pa jo viša. Oceno ReliefF lahko opišemo tudi kot razliko verjetnosti [2]

$$\mathbf{w}[k] = P(\text{različen } \mathbf{x}_k \mid \text{različen } t_k) - P(\text{različen } \mathbf{x}_k \mid \text{enak } t_k) \ , \tag{2.1}$$

kjer je $\mathbf{w}$ vektor z ocenami atributov. Relief poišče po en najbližji vzorec z različnim in enakim razredom, izboljšana izpeljanka ReliefF poišče l vzorcev, kjer je l uporabniško določen parameter. Izpeljanka RReliefF je prilagojena na regresijske probleme in namesto različnih in enakih razredov išče razrede s podobnimi vrednostmi.

2.1.2 Metoda razlike standardnega odklona

Mera izvira iz algoritma za gradnjo regresijskih dreves. Če želimo zgraditi regresijsko drevo, se soočimo s problemom, kateri atribut izbrati za cepitev drevesa, da bomo ločili kar največ podatkov. Pri diskretnih atributih in razredih iščemo atribut z največjo medsebojno informacijo med atributom in

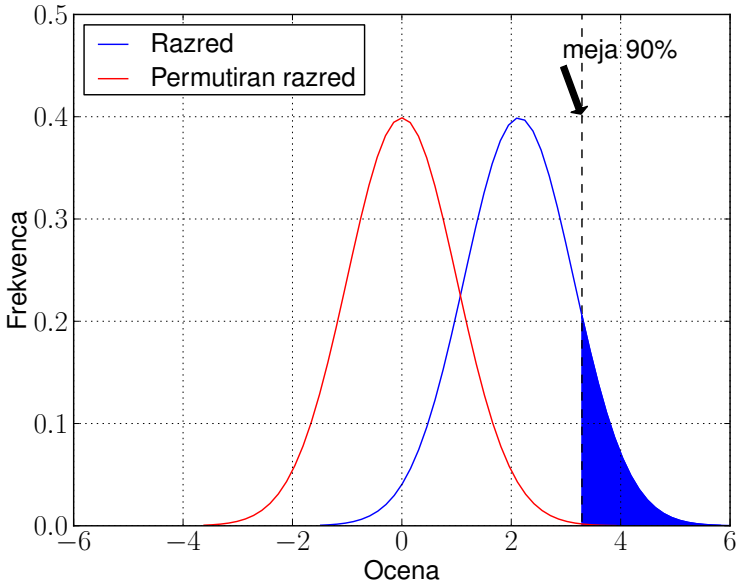
razredom. Pri zveznih vrednostih je mogoče uporabiti zvezno definicijo medsebojne informacije, večkrat pa se uporabi mero razlike standardnega odklona (ali razlike variance, ki je kvadrat standardnega odklona). Gre za eno izmed funkcij nečistoče (ang. impurity functions), kjer merimo, kako različni so uvrščeni vzorci. Primerjamo standardni odklon vrednosti v listih drevesa pred cepitvijo in po cepitvi za določen atribut. Atribut, ki najbolj zmanjša standardni odklon, je izbran kot najboljši. Ta postopek ponavljamo rekurzivno, atributi bližje korenu imajo večjo pomembnost kot tisti pri listih.

2.1.3 Metoda MARS

Pri metodi modeliranja z adaptivnimi multivariantnimi regresijskimi zlepki (ang. multivariate adaptive regression splines) na podlagi ocene najmanjše kvadratne napake v prvem delu izgradnje modela najprej dodajamo nove člene, v drugem pa odvzemamo najmanj pomembne člene. Za vsak člen ob dodajanju in odvzemanju izračunamo spremembo napake, ki jo prinese modelu, če ga v model dodamo ali pozneje odstranimo. Tako iz modela odstranimo najmanj pomembne člene in ga naredimo preprostejšega, obenem pa rešimo problem prekomernega prileganja učnim podatkom. Če stvar nekoliko obrnemo, je člen, brez katerega se najbolj zviša napaka modela, zelo pomemben. Ocena MARS tako na podlagi največjega prispevka kvadratne napake k modelu razvrsti atribute od najbolj do najmanj pomembnega.

2.1.4 Naključni gozdovi

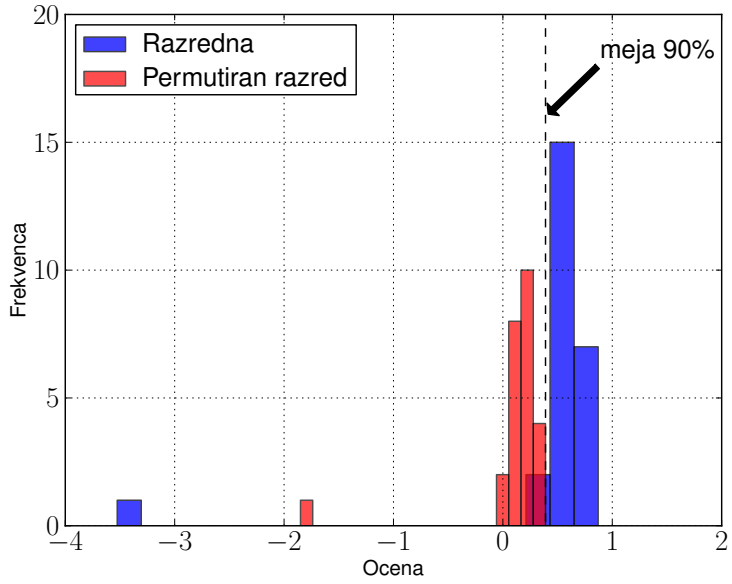
V delu [5] so predlagali možnost uporabe metode naključnih gozdov kot nekratkovidno mero za oceno kvalitete atributov. Postopek za izračun mere temelji na predpostavki, da naključno permutiranje vrednosti pomembnega atributa močno vpliva na pravilnost uvrstitve razreda, permutiranje manj pomembnega atributa pa malo. Za izračun tako najprej na podatkih zgradimo gozd naključnih dreves, tipično velikosti 100 dreves. Za testno množico vzamemo izpuščene vzorce (ang. out of the bag), ki jih pri učenju modela (gradnji drevesa) nismo upoštevali. Naučen model nato testiramo na testni množici dvakrat, prvič pustimo vrednosti opazovanega atributa nespremenjene, drugič pa jih naključno permutiramo. Iskana mera je tako razlika med številom pravilno uvrščenih vzorcev z izvornim nepermutiranim in permutiranim opazovanim atributom.



Slika 2.2: Permutacijski test: idealen primer

2.1.5 Permutacijski test

Vse zgoraj naštete mere nam za seznam atributov podajo oceno v obliki številčne vrednosti, ki nam pove, koliko je posamezen atribut pomemben pri napovedovanju razreda. Na ta način lahko ugotovimo razmerja med posameznimi atributi, kateri je boljši in kateri slabši, težko pa za neki atribut rečemo, da je v našem modelu nepomemben. Obstaja način [6], s katerim lahko določimo spodnjo mejo za mero, pod katero atributi niso več pomembni. Postopek se imenuje permutacijski test, pri katerem atribute ocenimo še na naključno premešani množici razredov. Ocena na permutirani množici bi v idealnem primeru morala biti enaka nič za vse atribute. Spodnjo mejo določimo tako, da pogledamo pogostost pojavitev vrednosti na permutirani množici in pri neki meji vseh pojavitev potegnemo črto. V našem primeru smo ocenili kot nepomembne tiste atribute, ki ne presegajo 90 % pojavitev vrednosti v permutirani množici. Na sliki 2.2 je grafično prikazan idealiziran primer, kjer bi imeli podatek o oceni za neskončno vrednosti atributa in bi bila ocena razporejena z normalno verjetnostno porazdelitvijo. Iz slike je razvidno, da se ocene pomembnih atributov pričnejo pri oceni okrog vrednosti 3,5. V praksi imamo končno število vrednosti in lahko za prikaz uporabimo histogram. Na sliki 2.3 vidimo, da so kot pomembni razpoznani atributi z oceno 45 in več.



Slika 2.3: Permutacijski test: histogram

2.2 Modeliranje

2.2.1 Linearna regresija

Model linearne regresije temelji na izbiri hiperravnine, ki se najbolje prilega učnim podatkom. Enačbo modela lahko zapišemo kot

$$\mathbf{y} = \mathbf{w} \cdot \mathbf{X} \, . \tag{2.2}$$

Vektor napake

$$\boldsymbol{\epsilon} = \boldsymbol{t} - \boldsymbol{y} \tag{2.3}$$

je razlika med izmerjenimi podatki in prilegajočo se hiperravnino. Naš cilj je določiti uteži $\mathbf{w}$ tako, da minimizirajo vektor napake in tako določimo najboljšo prilegajočo hiperravnino. Uteži $\mathbf{w}$ izračunamo po enačbi

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t} \, . \tag{2.4}$$

V primeru, ko napovedujemo razred samo iz enega (dveh) atributov, pravzaprav aproksimiramo podatke s premico (ravnino) v prostoru.

2.2.2 Metoda k najbližjih sosedov

Metoda k najbližjih sosedov [12] je ena izmed preprostejših metod podatkovnega rudarjenja. Deluje po principu iskanja podobnih vrednosti atributov z znanimi razredi iz učne množice. Za iskanje podobnih vrednosti izračunamo razdaljo med atributi za neznan vzorec j in ostalimi vzorci v učni množici. Za izračun razdalje lahko uporabimo katerokoli veljavno metriko, najpogosteje pa se uporablja razdalja Manhattan

$$d_1(\mathbf{x}_i, \mathbf{x}_j) = ||\mathbf{x}_i - \mathbf{x}_j||_1 = \sum_k |x_{ik} - x_{jk}| \ , \tag{2.5}$$

ali evklidska razdalja

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = ||\mathbf{x}_i - \mathbf{x}_j||_2 = \sum_k (x_{ik} - x_{jk})^2 \ . \tag{2.6}$$

x_{ik} predstavlja vrednost k-tega atributa v i-tem vzorcu. Pri diskretnih atributih za določanje razdalje uporabimo kar Hammingovo razdaljo, ki je definirana kot število različnih si elementov v vektorju. Ker želimo, da vsak atribut enakovredno prispeva k skupni razdalji, je potrebno vrednosti vseh atributov preslikati na enak interval. Za preslikavo na interval [0, 1] oziroma standardizacijo uporabimo enačbo

$$\mathbf{X}_k^s = \frac{\mathbf{X}_k - \min(\mathbf{X}_k)}{\max(\mathbf{X}_k) - \min(\mathbf{X}_k)} \ .$$

Vrednost razreda izračunamo tako, da povprečimo k vrednosti razredov iz učne množice z najmanjšo izračunano razdaljo, kjer je k uporabniško določen parameter.

2.2.3 Metoda MARS

Metoda adaptivnih multivariantnih regresijskih zlepkov je nelinearna regresijska metoda, kjer namesto izbire koeficientov določene matematične funkcije zgradimo model kot vsoto koeficientov pomnoženih z eno ali več odsekovno linearnih funkcij

$$y_i = \beta_0 + \sum_m \beta_m h_m(\mathbf{x}_k) \ , \tag{2.7}$$

kjer so β_0 in β_m parametri, ki se nastavijo z učenjem modela, h_m pa ena ali produkt več odsekovno linearnih funkcij iz podanega nabora. Osnovna

odsekovno linearna funkcija $h(u)$, uporabljena pri modeliranju, je definirana kot

$$h(u) = \max(0, u \pm t) \ , \tag{2.8}$$

kjer t predstavlja koleno (vozlišče) funkcije, odvisna spremenljivka pa je eden izmed atributov. Oba se izbereta med postopkom učenja. Omenjeni model se zelo dobro izkaže na nelinearnih večdimenzijskih problemih (torej primerih, kjer imamo veliko število atributov) in predstavlja alternativo nevronskim mrežam. Postopek gradnje modela je sestavljen iz dveh korakov:

1. dodajanja (ang. forward selection) in
2. odvzemanja (ang. backward elimination).

V prvem delu dodajamo člene, ki najbolj zmanjšajo kvadratno napako (ang. residual square error). Gre za tako imenovani požrešen algoritem, za katerega je značilno, da v vsakem koraku poišče lokalno najboljšo možno rešitev, kar pa ne vodi nujno do globalno najboljše. Dodani členi so lahko:

- konstanta 1,
- ena izmed osnovnih odsekovno linearnih funkcij ali
- produkt dveh ali več osnovnih odsekovno linearnih funkcij.

Člene dodajamo, dokler ni preseženo njihovo vnaprej določeno največje število. Med dodajanjem pravzaprav pregledujemo celoten prostor možnih členov. Pregledati je potrebno:

- vse atribute, da izberemo pravega za odvisno spremenljivko osnovne funkcije,
- vse vrednosti vseh atributov, da izberemo koleno osnovne funkcije.

Namen drugega dela postopka je znižanje prekomernega prileganja modela učnim podatkom. V vsakem koraku tega postopka preverimo, kateri člen najmanj zviša napako v modelu, in ga odstranimo. Člene odstranjujemo, dokler ta ne doseže optimalnega števila členov glede na mero GCV (ang. generalized cross validation)

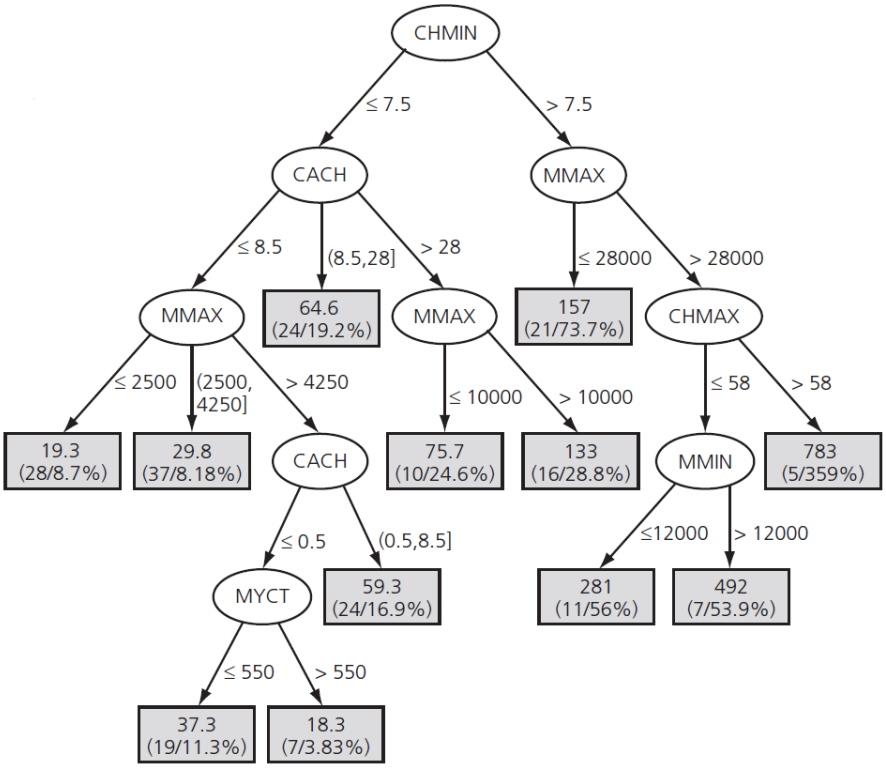
$$\mathrm{GCV} = \frac{\sum_{i=1}^{m}(t_i - y_i)^2}{\left(1 - \frac{C}{m}\right)^2} \ . \tag{2.9}$$

m v enačbi predstavlja število vseh primerov v učni množici, C, ki nastopa v imenovalcu, je definiran kot $C = 1 + cd$, kjer je d število različnih osnovnih

funkcij v modelu, c pa je uporabniško izbran parameter in predstavlja kazen za veliko število členov v modelu. Empirični rezultati kažejo, da se za c kot najboljša izkaže vrednost na intervalu $2 < c < 3$ [4].

2.2.4 Odločitvena drevesa

Odločitvena drevesa delimo na uvrstitvena in regresijska. Pri obojih podatke modeliramo tako, da na podlagi učnih podatkov izgradimo drevo, ki ima v listih vrednosti diskretnega razreda (uvrstitvena drevesa) ali pa povprečno vrednost zveznega razreda (regresijska drevesa). V vozliščih imamo postavljene pogoje na podlagi vrednosti atributov, preko katerih potujemo do končnih vrednosti razredov. Na sliki 2.4 so v vozliščih v elipsah imena atributov, na vejah



Slika 2.4: Primer regresijskega drevesa [10]

meje oziroma intervali vrednosti atributov in v listih, označenimi s kvadrati,

napovedane vrednosti razreda skupaj z deležem pripadajočih vzorcev v učni množici.

Izgradnjo odločitvenega drevesa navadno predstavimo rekurzivno, kjer učno množico podatkov rekurzivno cepimo v vedno manjše podmnožice na podlagi izbranega atributa. Ključna je izbira pravega atributa, da ločimo podatke tako, da je napoved modela kar se da uspešna [7]. Atributi, uporabljeni v vozliščih bližje korenu, so bolj pomembni pri modeliranju, saj ločijo večji delež podatkov. Pri odločitvenih drevesih, kjer napovedujemo diskretne vrednosti razreda, najpomembnejši atribut določimo tako, da izračunamo razmerje informacijskega prispevka I_R med razredom in vsakim izmed atributov

$$I_R(\mathbf{t}; \mathbf{X}_k) = \frac{I(\mathbf{t}; \mathbf{X}_k)}{H(\mathbf{X}_k)} \ , \tag{2.10}$$

kjer je $I(\mathbf{t}; \mathbf{X}_k)$ [11] povprečna medsebojna informacija med razredom in izbranim atributom

$$I(\mathbf{t}; \mathbf{X}_k) = H(\mathbf{t}) – H(\mathbf{t}|\mathbf{X}_k) \ , \tag{2.11}$$

$H(\mathbf{X}_k)$ in $H(\mathbf{t})$ entropija atributa $\mathbf{X}_k$ oziroma razreda $\mathbf{t}$

$$H(\mathbf{t}) = -\sum_{i \in \mathbf{t}} p_i \cdot \log(p_i) \ , \tag{2.12}$$

$$H(\mathbf{X}_k) = -\sum_{j \in \mathbf{X}_k} p'_j \cdot \log(p'_j) \ , \tag{2.13}$$

$H(\mathbf{t}|\mathbf{X}_k)$ pa pogojna entropija razreda $\mathbf{t}$ pri dani vrednosti izbranega atributa $\mathbf{X}_k$

$$H(\mathbf{t}|\mathbf{X}_k) = \sum_i \sum_j p_{ij} \cdot \log(p_{i|j}) \ . \tag{2.14}$$

Pri tem je p_i verjetnost vrednosti t_i razreda $\mathbf{t}$, p'_j pa verjetnost vrednosti x_{kj} atributa $\mathbf{X}_k$. p_{ij} je verjetnost vrednosti t_i razreda $\mathbf{t}$ pri vrednosti x_{kj} atributa $\mathbf{x}_k$, $p_{i|j}$ pa je verjetnost definirana kot

$$p_{i|j} = \frac{p_{ij}}{p'_j} \ . \tag{2.15}$$

Za cepitev tako vedno izberemo atribut z največjim razmerjem informacijskega prispevka. Postopek ponavljamo rekurzivno, dokler ni informacijski prispevek ustrezno majhen. Pri regresijskih razredih se namesto zvezno definirane povprečne medsebojne informacije za mero uporablja razlika standardnega odklona

$$\Delta\sigma(\mathbf{t};\mathbf{X}_k) = \sigma(\mathbf{t})–\sigma(\mathbf{t},\mathbf{X}_k)\ , \tag{2.16}$$

kjer sta $\sigma(\mathbf{t})$ in $\sigma(\mathbf{t},\ \mathbf{X}_k)$ standardni odklon ene in dveh spremenljivk definirana kot:

$$\sigma(\mathbf{t}) = \sqrt{\frac{\sum_i (t_i - \bar{t})^2}{m}}\ , \tag{2.17}$$

$$\sigma(\mathbf{t},\mathbf{X}_k) = \sum_{i\in\mathbf{X}_k} (p_i \cdot \sigma_i)\ , \tag{2.18}$$

kjer je m število vzorcev v učni množici, $\bar{t}$ povprečna vrednost vektorja $\mathbf{t}$, p_i verjetnost za vrednost atributa x_{ki}, σ_i pa standardni odklon razreda pri vrednosti atributa x_{ki}. Omenjeno mero uporabljajo znani algoritmi za izgradnjo dreves ID3, C4.5 in C5.

2.2.5 Metoda podpornih vektorjev

Metoda podpornih vektorjev je metoda strojnega učenja za uvrščanje diskretnih razredov. Bistvo algoritma je, da med dvema razredoma potegne hiperravnino, ki ju ločuje in je od njiju maksimalno oddaljena. Pravimo, da takšna hiperravnina najbolje ločuje razreda. Za definicijo hiperravnine so pomembne le točke, ki so najbližje ravnini. Imenujemo jih podporni vektorji in z njimi lahko zapišemo enačbo modela

$$y_j = \beta + \sum_{i\in V} \alpha_i \mathrm{k}(\mathbf{x}_i,\mathbf{x}_j)\ , \tag{2.19}$$

kjer je V množica podpornih vektorjev, β in α_i pa so parametri, ki se nastavijo med učenjem. V enačbi nastopa tudi jedrna funkcija $\mathrm{k}(\mathbf{x}_i,\mathbf{x}_j)$, ki pride v poštev pri linearno neločljivih problemih. Takrat med dvema razredoma ne moremo potegniti ločujoče hiperravnine. V praksi se srečamo z veliko nelinearno ločljivimi razredi, katere lahko ločimo z uporabo nelinearne jedrne funkcije in tako preslikamo podatke v višjo dimenzijo, kjer jih linearno ločimo. V ta namen se največkrat uporablja polinomsko in radialno bazno jedrno funkcijo. Za algoritem so izmed vseh vzorcev v učni množici pomembni samo podporni vektorji, vse ostale vzorce lahko odmislimo.

Ker ne vemo, katera stopnja polinoma n je optimalna za naše podatke, je smiselno začeti s stopnjo $n = 1$ (linearno) in jo nato povečevati do neke določene vrednosti. V večini primerov se do določenega n prileganje izboljšuje, nato pa poslabša. Velikokrat se to zgodi že pri stopnji, manjši od pet. Regresijske probleme z metodo podpornih vektorjev rešujemo tako, da poskušamo

vrednosti razreda zajeti v pas širine ϵ. Parameter ϵ tako definira debelino pasu okrog ločitvene ravnine, znotraj katerega ignoriramo napako. Nato pa poskušamo ta pas postaviti tako, da najbolj minimiziramo napako.

2.2.6 Naključni gozdovi

Gre za eno izmed skupinskih metod, kjer je ideja uporabiti več različnih osnovnih uvrstitvenih metod in tako z vsako bolje modelirati določen del podatkov [8]. Omenjene metode se navadno uporabljajo na velikih podatkovnih množicah. Po metodi naključnih dreves zgradimo množico, veliko navadno 100 odločitvenih oziroma regresijskih dreves. Pri gradnji vsakega drevesa naključno izbiramo vzorce iz učne množice po principu izbiranja z vračanjem. Iz kombinatorike sledi, da v limiti ostane $1/\mathrm{e} = 37$ % vzorcev neizbranih (ang. out of the bag). Te lahko nato uporabimo za nepristransko oceno modela. Sam model postane še bolj nedeterminističen, ko za cepitev drevesa namesto celotne množice atributov naključno izberemo le podmnožico velikosti m, izmed katerih nato ugotovimo, kateri je najboljši. Parameter m je navadno kvadratni koren števila vseh atributov. Prav ta naključnost naredi model odporen na šum in na prekomerno prileganje učni množici.

2.2.7 Umetne nevronske mreže

Umetne nevronske mreže se zgledujejo po nevronskih strukturah iz narave in predstavljajo omrežje več med seboj povezanih enostavnih elementov, nevronov. Vsak nevron vrednosti iz množice atributov $\mathbf{x}_i$ pomnoži z utežmi $\boldsymbol{w}$, te vrednosti sešteje in jim prišteje prag b, to vrednost pa potem pošlje naprej v prenosno funkcijo f. V našem primeru smo uporabili sigmoidno prenosno funkcijo

$$f(u) = \frac{1}{1 + e^{-\beta u}} \,, \tag{2.20}$$

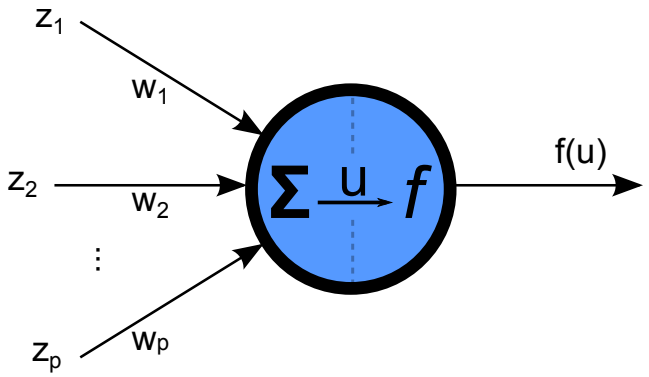
$$u = \boldsymbol{w}^T \cdot \boldsymbol{z} + b \,. \tag{2.21}$$

Glede na način povezovanja nevronov med seboj poznamo več vrst nevronskih mrež. V našem primeru smo uporabili večnivojski perceptron z vzvratnim učenjem. Nevroni so razporejeni v dve plasti, skrito in izhodno. V skriti plasti smo imeli q nevronov in enega v izhodni, ker imamo en sam razred. Pri algoritmu vzvratnega učenja za vsak učni vzorec izračunamo razliko med napovedano in učno vrednostjo vzorca. To razliko z verižnim pravilom vzvratno prenesemo do uteži med skrito in izhodno plastjo, izračunamo nove uteži ter

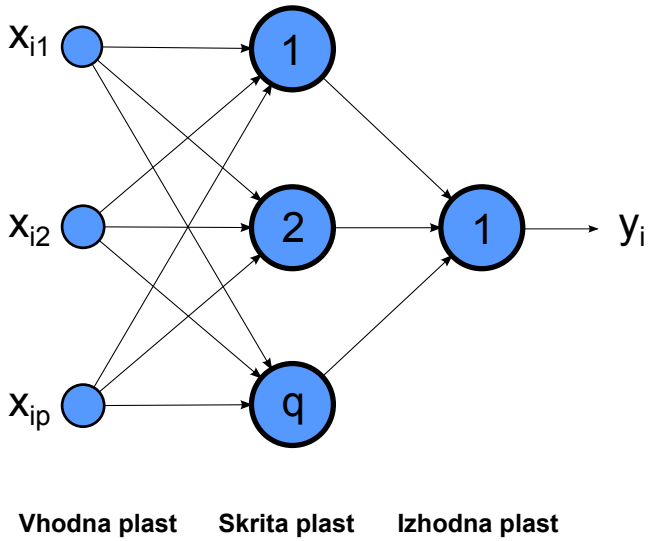
jih popravimo. Enako storimo za uteži med vhodno in skrito plastjo. Model uporabljenega večnivojskega perceptrona popišemo z enačbo [9]

$$y_i = f(b_I + \mathbf{w}_I^T \boldsymbol{f}(\mathbf{W}_S \mathbf{x}_i + \mathbf{b}_S)) \,. \tag{2.22}$$

Uteži in pragovi v skriti plasti, $\mathbf{W}_S$ in $\mathbf{b}_S$, ter izhodni plasti, $\mathbf{w}_I$ in b_I, se na



Slika 2.5: Perceptron s p vhodnimi spremenljivkami

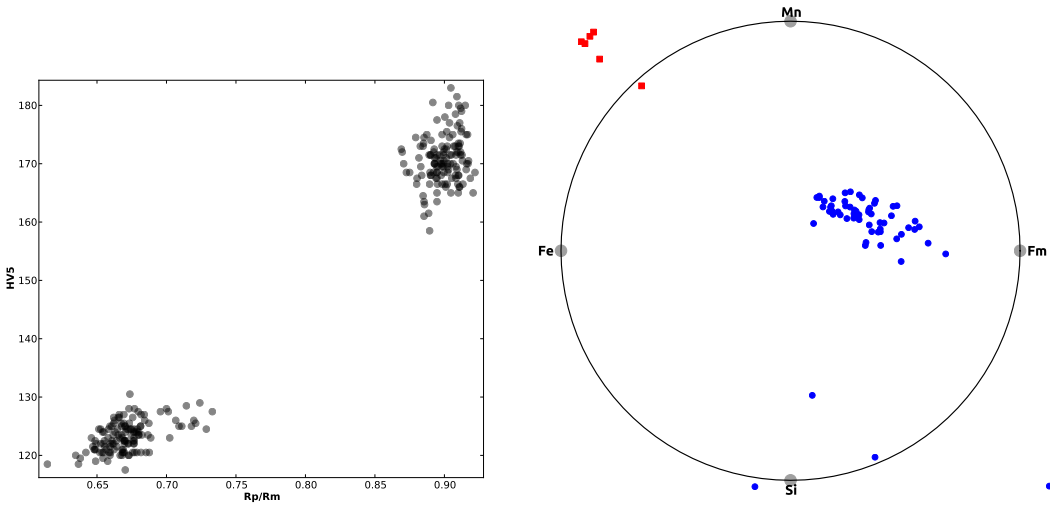


Slika 2.6: Nevronska mreža s p vhodi, q nevroni v skriti in enim v izhodni plasti

začetku naključno nastavijo na neko vrednost, nato pa se skozi postopek učenja prilagodijo podatkom, da je napaka modela najmanjša. Sigmoidna vektorska funkcija $\boldsymbol{f}$ za vsak element vhodnega vektorja izračuna vrednost sigmoidne funkcije f.

2.3 Vizualizacija podatkov

Zakonitosti v podatkih ljudje najlažje zaznamo z grafičnim prikazom le-teh. Težava nastane pri velikih zbirkah podatkov, kjer imamo veliko število atributov in je tako pregledovanje vseh mogočih prikazov dolgotrajno. VizRank [13] je metoda, ki na podlagi strojnega učenja iz vseh mogočih prikazov podatkov prepozna najboljše. Vsaka kombinacija atributov je ocenjena na podlagi tega, kako dobro ločuje razred pregledovane množice podatkov. Metoda deluje le za diskretne razrede, možne so implementacije z različnimi uvrstitvenimi algoritmi strojnega učenja, niso pa vsi enako učinkoviti. V orodju Orange je uporabljena metoda k najbližjih sosedov. VizRank je mogoče uporabiti z različnimi metodami za vizualizacijo, najpogosteje pa se uporablja razsevni diagram (ang. scatter plot), kjer naenkrat prikažemo odvisnost dveh spremenljivk (slika 2.7a), ali diagram Radviz [14] (ang. radial coordnate visualization), kjer lahko naenkrat prikažemo odvisnosti več kot dveh spremenljivk (slika 2.7b). Diagram Radviz prikaže podatke bližje sprejemljivki, ki ima nanje močnejši vpliv. Te so enakomerno razporejene po obodu krožnice.



(a) Razsevni diagram (b) Radviz diagram

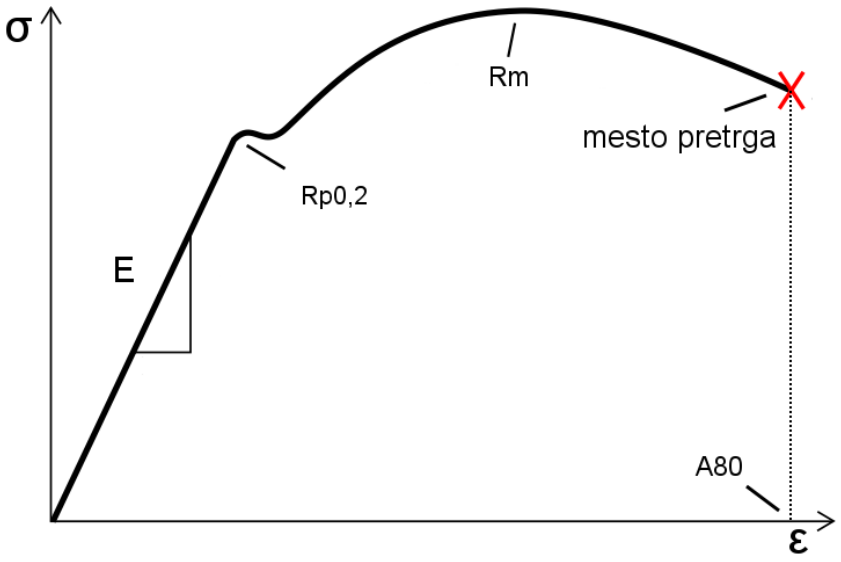
Slika 2.7: Razsevni in Radviz diagram

Poglavje 3

Podatki

Od Inštituta za materiale in tehnologijo smo prejeli tri množice podatkov. Ker je za uspešno modeliranje potrebna kar se da velika množica vzorcev, smo se osredotočili le na take z več kot 50 vzorci. Na koncu smo uspeli izbrskati tri primerne množice, kjer je prva imela 132 vzorcev, druga 70 in tretja 67. Gre za pločevini dveh različnih kvalitet z oznakama M310-50A in M800-65K. Oznaka M nakazuje, da gre za elektropločevino, torej bodo končni izdelki sestavni deli električnih strojev. Sledeča številka 310 oziroma 800 določa magnetno lastnost vatne izgube, ki mora pri gostoti magnetnega polja 1,5T znašati 3,1 W/kg (M310) oziroma 8 W/kg (M800). Številki 50 oziroma 65 določata debelino pločevine v stotinkah milimetra, zadnji črki A ali K pa določata gotovo ali polgotovo pločevino. Slednjo je potrebno za doseganje ustreznih lastnosti še dodatno toplotno obdelati.

Podatki so sestavljeni iz dveh glavnih skupin meritev. Parametri iz prve skupine, mehanske meritve, se merijo na treh merilnih instrumentih - trgalnem stroju Instron, merilniku trdote Vickers in merilniku hrapavosti Surtronic. Na trgalniku se opravi natezni preizkus, tako da vanj vstavimo podolgovat vzorec pločevine, imenovan tudi epruveta, katerega obremenimo z naraščajočo enoosno napetostjo, vse dokler se epruveta ne pretrže. Potek nateznega poizkusa najlažje razložimo z napetostno deformacijskim diagramom, ki prikazuje napetost materiala $\sigma = \frac{F}{S}$ v odvisnosti od specifične deformacije $\epsilon = \frac{\Delta x}{x}$ (slika 3.1). F je sila, s katero obremenimo vzorčno epruveto, S presek, x dolžina in Δx raztezek epruvete. Napetostno deformacijski diagram lahko poenostavljeno razdelimo v dve prevladujoči območji: elastični in plastični del. Za elastični del praviloma predpostavimo linearni model, ki ga popišemo s Hookovim zakonom. Za potrebe testiranj se meja med elastičnim in plastičnim območjem poenostavljeno določi pri vrednosti specifične deformacije 0,2 %. Naklon pre-



Slika 3.1: Napetostno deformacijski diagram [16]

mice predstavlja Youngov modul E. V območju plastičnosti natezna napetost narašča počasneje vse do točke, kjer začne padati. Točko največje napetosti materiala označimo z Rm. Natezna napetost nato pada do mesta, kjer se epruveta pretrga in je poizkusa konec. Na podlagi relativnega raztezka na trgalnem stroju določimo še parameter A80, ki predstavlja dolžino 80 mm dolge epruvete po pretrganju in na podlagi natezne trdnosti največjo silo, ki jo material še prenese.

Merilnik trdote po metodi Vickers deluje tako, da se na material pritisne diamantni tetraeder s kotom konice 136°. Zaradi delovanja sile tetraeder pusti v materialu zarezo, velikost pa je odvisna od trdote materiala. Trdoto po Vickersu HV nato določimo po enačbi $\frac{F}{S}$, kjer F sila pritiska tetraedra in S izmerjena površina zaznamka. V našem primeru smo tetraeder obremenili s težo 5 kg, zato trdoto označimo z oznako HV5. Merilnik hrapavosti deluje tako, da z gramofonski igli podobnim tipalom drsimo po površini materiala in beležimo vertikalno gibanje igle. V našem primeru je rezultat za vsak vzorec predstavljen s parametrom hrapavosti površine Ra, ki je določen kot aritmetično povprečje absolutnih vertikalnih pomikov igle. Za vsak vzorec smo pomerili hrapavost v vzdolžni in prečni smeri na obeh straneh pločevine. Vsebnosti 16 kemijskih elementov so izmerjene z analizatorjem kovine spektrometrom Thermo Scientific ARL.

Ker se na trgalniku izmeri več parametrov, na merilniku trdote pa le enega,

je smiselno, da poskusimo iz izmerjenih parametrov prve naprave napovedati vrednosti parametrov druge. V kolikor natančnost napovedane vrednosti ne bo zadostila želenim, bi bilo smiselno poskusiti določiti model, ki ne napove vrednosti parametra HV5, ampak le poskuša napovedati, kdaj je vrednost znotraj ali zunaj predpisanih tehnično prevzemnih pogojev. Za omenjeni pločevini M310-50A in M800-65K so tehnično prevzemni pogoji za mehanske lastnosti predstavljeni v tabelah 3.1 in 3.2.

Tabela 3.1: Predpisane mehanske lastnosti za pločevino M310-50A [17]

	Enota	Min.	Max.
Napetost pri 0,2 % Rp	N/mm^2	280	400
Največja napetost Rm	N/mm^2	420	550
Razmerje Rp/Rm	-	0,62	-
Raztezek A80	%	-	-
Trdota HV5	-	140	165

Tabela 3.2: Predpisane mehanske lastnosti za pločevino M800-65K [17]

	Enota	Min.	Max.
Napetost pri 0,2 % Rp	N/mm^2	370	500
Največja napetost Rm	N/mm^2	400	560
Razmerje Rp/Rm	-	0,8	-
Raztezek A80	%	-	-
Trdota HV5	-	150	175

3.1 Predprocesiranje

Ker v podjetju še nimamo delujočega sistema za zajem in shranjevanje podatkov v računalniško podatkovno bazo, je bilo najprej potrebno rezultate v obliki množice datotek v obliki zapisa Microsoft Excel in besedilnih datotek pretvoriti v uporabnejšo obliko. To smo storili s pomočjo skriptnega jezika Python. Končni rezultat je bil za vsako podatkovno množico svoja besedilna datoteka s tabulatorjem ločenih vrednosti, kjer stolpci predstavljajo atribute

in razrede, vrstice pa posamezne vzorce. Takšen zapis zna prebrati praktično vsak programski paket za obdelavo podatkov.

3.2 Pregled

Seznam vseh parametrov z oznakami in opisi je podan v tabeli 3.3. Stolpec Tip v tabeli označuje ali gre za zvezen (Z) ali diskreten (D) tip podatkov, stolpec Množica pa v katerih izmed treh pridobljenih množicah podatkov je parameter vsebovan.

Množica 1

Prva podatkovna množica vsebuje 132 vzorcev pločevine vrste M310-50A, sestavljenih iz sedmih mehanskih parametrov, opravljenih na trgalnem stroju Instron in merilniku trdote Vickers. Vzorci so bili vzeti pred toplotno obdelavo in po njej, tako da je v množici dejansko 264 vzorcev. Odvzeti so bili iz 44 kolutov pločevine, iz vsakega na treh različnih pozicijah – začetku, koncu in sredini koluta. Prav tako so vzorci opremljeni še s številko sarže, iz katere je bil kolut pločevine izdelan v železarni, z debelino in s širino.

Množica 2

V drugi množici je bilo izmerjenih 70 vzorcev pločevine vrste M800-65K. Poleg mehanskih meritev imamo v tej množici meritev tudi vsebnost kemijskih elementov. Pločevina, iz katere so vzorci vzeti, ni bila dodatno toplotno obdelana.

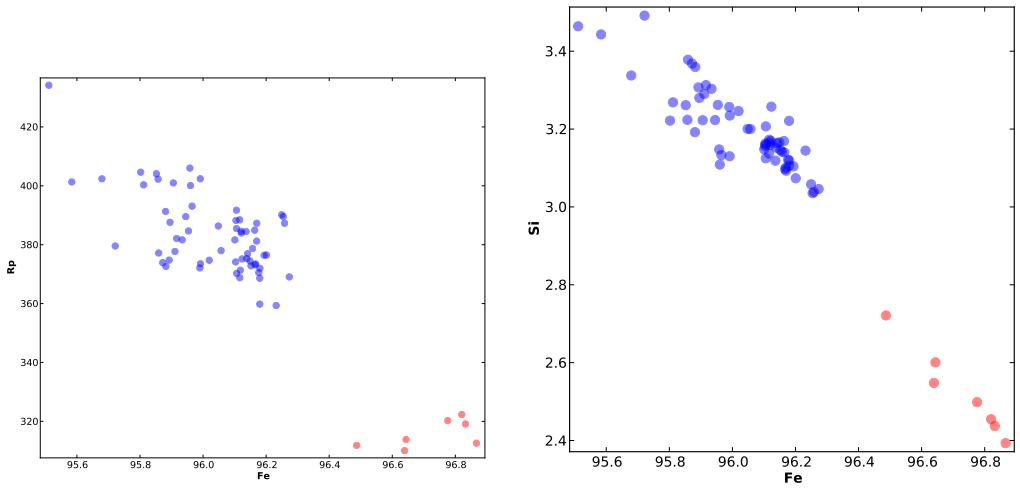
Množica 3

Tretja množica vsebuje 67 vzorcev pločevine M310-50K, kjer smo imeli od mehanskih meritev izmerjene samo podatke o Rm, Rp, Rp/Rm in trdoti. Poznamo tudi vsebnosti kemijskih elementov, dodatno pa imamo še podatek o hrapavosti in datumu, kdaj so bili vzorci odvzeti.

3.3 Vizualizacija podatkov

Nekaj zanimivih povezav med podatki smo želeli prikazati grafično. Nabor gradnikov za vizualizacijo v programskem paketu Orange je velik, za naše podatke

sta se kot najboljša izkazala osnovni razsevni diagram in diagram Radviz. Pustili smo, da nam program s pomočjo metode VizRank sam poišče najboljše prikaze podatkov. Žal metoda deluje samo na diskretnih razredih, zato smo se odločili podatke na podlagi ustreznosti mehanskih lastnosti tehnično prevzemnim pogojem razdeliti v dve skupini. Neustrezni vzorci so bili prisotni samo v drugi množici, tako da smo na ta način vizualizirali le omenjeno množico. Program nam je po slabi minuti prikazal zanimive vizualizacije. Na slikah so vzorci neustrezne kvalitete prikazani z rdečo barvo.

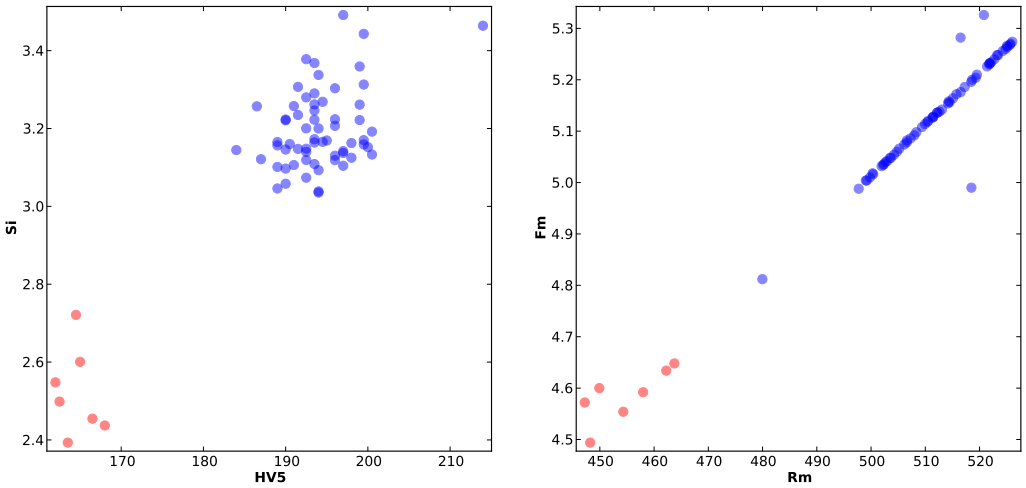


(a) Rp v odvisnosti od Fe

(b) Linearna povezava vsebnosti silicija od vsebnosti železa

Slika 3.2: Razsevna diagrama

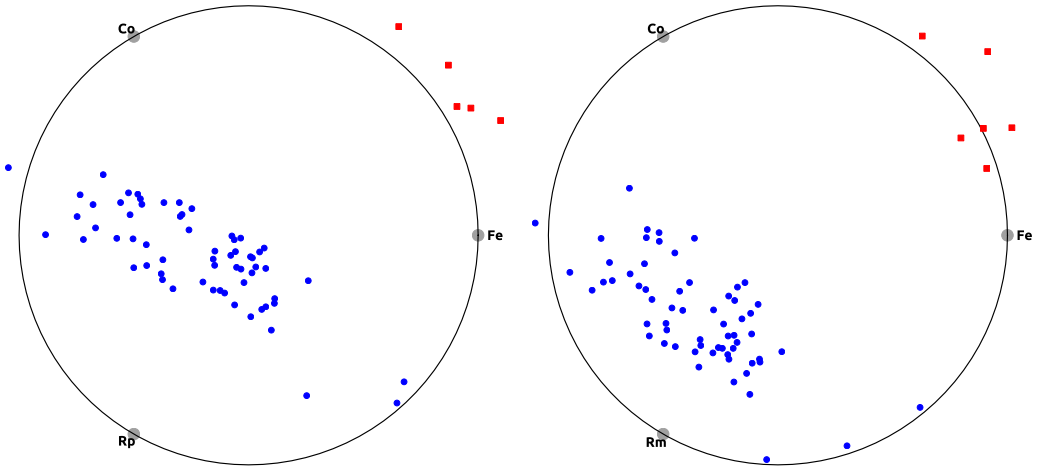
Na sliki 3.2a je vidna negativna linearna povezava med napetostjo pri raztezku 0,2 % (Rp) in vsebnostjo železa. Neustrezni vzorci so izrazito pomaknjeni v desni spodnji kot diagrama, kjer so vsebnosti železa velike, Rp pa majhen. Še bolj očitna je na sliki 3.2b negativna linearna povezava vsebnosti silicija v odvisnosti od vsebnosti železa. Neustrezni vzorci so tudi tukaj postavljeni v desni spodnji kot diagrama, kjer so vsebnosti železa visoke, silicija pa majhne. Na sliki 3.3a, kjer je prikazana odvisnost vsebnosti železa od trdote, sta se izoblikovali dve gruči ustreznih in neustreznih vzorcev. Neustrezni vzorci se tokrat nahajajo v levem spodnjem kotu, kjer so nizke vsebnosti silicija in majhna trdota. Prav tako je zaznati linearno povezavo med količinama. Na sliki 3.3b smo odkrili že znano linearno povezavo med največjo silo in napetostjo. Na



(a) Vsebnost silicija v odvisnosti od trdote (b) Očitna linearna povezava Fm in Rm

Slika 3.3: Razsevna diagrama

slikah 3.4a in 3.4b sta prikazana diagrama Radviz, ki prikazujeta odvisnost treh spremenljivk. Na obeh je vidno, da vsebnost kobalta in železa močno vpliva na to, ali je vzorec razpoznan kot ustrezen ali neustrezen.



(a) Povezave med vsebnostjo kobalta, železa in Rp

(b) Povezave med vsebnostjo kobalta, železa in Rm

Slika 3.4: diagrama Radviz

Tabela 3.3: Seznam pridobljenih parametrov po množicah podatkov.

Vrsta	Parameter	Ime	Tip	Instrument	Množica
ostalo	d	dolžina	Z	mikrometer	1, 2, 3
	š	širina	Z		1, 2, 3
	datum	datum	D		3
	pozicija	mesto vzorca	D		1
	s	številka sarže	D		1
	T	temp. obdelava	D		1
mehanske	Rm	največja napetost	Z	trgalni stroj	1, 2, 3
	Rp	napetost pri 0,2 %			
	Rp/Rm	razmerje			
	HV5	trdota			
mehanske	Fm	največja sila	Z	trgalni stroj	1, 2
	A80	dolžina po pretrgu			
	E	Youngov modul			
mehanske		hrapavost	Z	merilnik trdote	3
	Ra 1-v	zgoraj, vzdolžno			
	Ra 1-p	zgoraj, prečno			
	Ra 2-v	spodaj, vzdolžno			
	Ra 2-p	spodaj, prečno			
kemijske	Fe	železo	Z	spektrometer	2, 3
	Co	kobalt			
	Cr	krom			
	Mo	molibden			
	Mn	mangan			
	V	vanadij			
	Nb	niobij			
	Cu	baker			
	P	fosfor			
	S	žveplo			
	C	ogljik			
	Ca	kalcij			
	Al	aluminij			
	Si	silicij			
	W	volfram			
	Ni	nikelj			

Poglavje 4

Ocena pomembnosti atributov

Oceno pomembnosti atributov smo poskušali izvesti z več različnimi orodji, od preprostih izračunov korelacijskega koeficienta s pomočjo programskega jezika Python, preko orodij za podatkovno rudarjenje Weka in Orange. Slednji se je izkazal za najbolj učinkovitega, zato smo uporabo ostalih orodij tekom izdelave diplomskega dela opustili. Za izračune ocen smo uporabili kar njegov intuitivni grafični vmesnik, kjer s pomočjo enostavnih gradnikov sestavimo shemo, ki predstavlja uvoz podatkov, morebitno predprocesiranje in ločitev na podmnožice, končno oceno atributov z različnimi metodami in prikaz rezultatov. Za izvedbo permutacijskega testa je bilo potrebno rahlo spremeniti gradnik za uvoz podatkov iz datoteke, saj v grafičnem načinu ni bilo mogoče premešati vrednosti razreda ob nespremenjenem vrstnem redu vrednosti atributov. Zaradi narave programskega paketa Orange, kjer je večji del implementiran s pomočjo skriptnega programskega jezika Python, je poseg vzel malo časa. Enako v času pisanja diplomske naloge še ni bilo implementirano ocenjevanje zveznih razredov pri metodi naključnih gozdov. Po pregledu izvorne kode smo ugotovili, da omejitev izvira iz primerjave pravilnosti napovedanih razredov pred in po permutaciji izbranega atributa. Z manjšim dodatkom v izvorni kodi, ki namesto enakosti razredov vrača razliko, smo implementirali metodo tudi za zvezne vrednosti. Opisano v psevdokodi, smo v datoteki */orange/Orange/ensemble/forest.py* del programa

```
if oob[i].getclass() == classifier(shuffle_ex(i)):
    return 1
else
    return 0
```

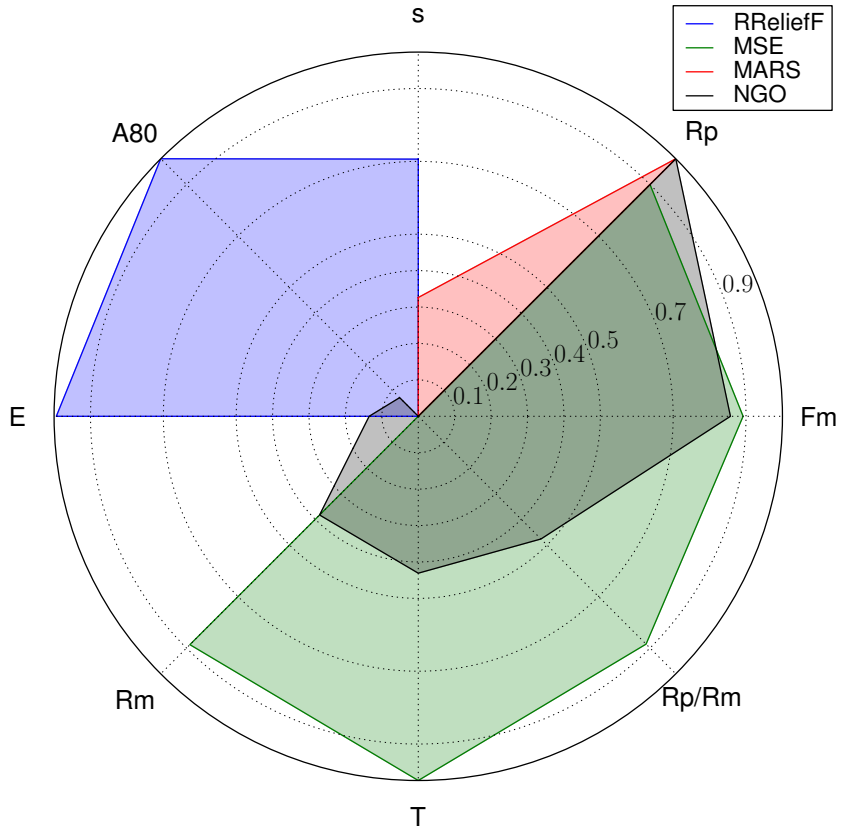
zamenjali z

```
return abs(oob[i].getclass() - classifier(shuffle_ex(i))).
```

Pri oceni pomembnosti atributov smo za vsako podatkovno množico določili mehansko lastnost HV5 kot razred, ostale lastnosti pa so nastopale kot atributi. Za vsako podatkovno množico smo izvedli permutacijski test, tako da smo naključno premešali vrednosti razredov, atribute pa pustili nedotaknjene, nato pa še enkrat ocenili pomembnost atributov. S permutacijskim testom smo nato določili mejo, pod katero atributi niso več relevantni, in jih izločili. Pomembnost smo ocenjevali s štirimi različnimi metodami, ki so atributom dodelile različne ocene. Te metode so RReliefF, metoda najmanjšega standardnega odklona, metoda MARS in metoda naključnih gozdov. Za prikaz rezultatov smo uporabili mrežni diagram. Vse ocene smo normalizirali, da jih lahko za vsako množico podatkov prikažemo na istem diagramu. Iz vsake metode smo izbrali do največ pet najboljših ocen, na diagramu tako prikažemo unijo najboljših atributov vseh metod. Za ocenjevanje smo uporabili privzete parametre metod. Zaradi bolj preglednega označevanja smo v legendi na mrežnih diagramih metodo razlike standardnega odklona označili z angleško okrajšavo MSE, naključne gozdove pa z okrajšavo NGO.

4.1 Rezultati

4.1.1 Podatkovna množica 1

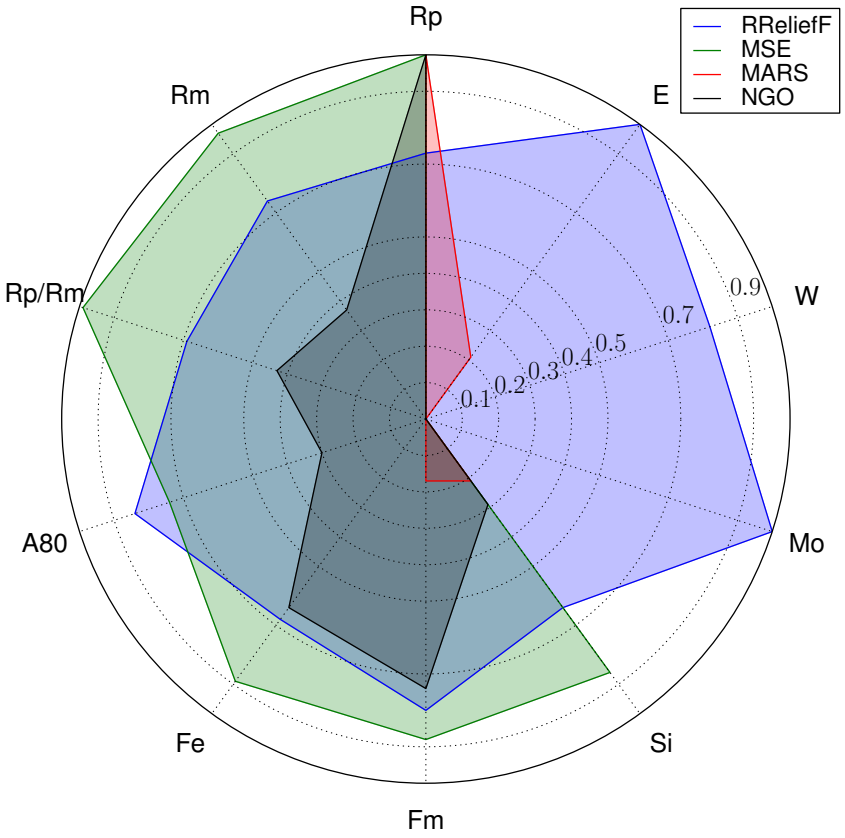


Slika 4.1: Najbolje ocenjeni atributi, množica 1

V prvi podatkovni množici (slika 4.1) je metoda RReliefF kot pomembne razpoznala samo Youngov modul, raztezek A80 in številko sarže, iz katere so bili posamezni koluti pločevine. Metoda najmanjšega standardnega odklona (MSE) je za pomembne razpoznala večino mehanskih lastnosti (največjo napetost Rm, največjo silo Fm, napetost pri raztezku 0,2 % Rp in razmerje Rp/Rm) in podatek o tem, ali je bil vzorec že toplotno obdelan. Ocena po metodi MARS je za pomembne razpoznala le natezno trdnost in številko sarže, metoda naključnih gozdov pa se je na tej množici podatkov odrezala podobno kot metoda najmanjšega standardnega odklona.

4.1.2 Podatkovna množica 2

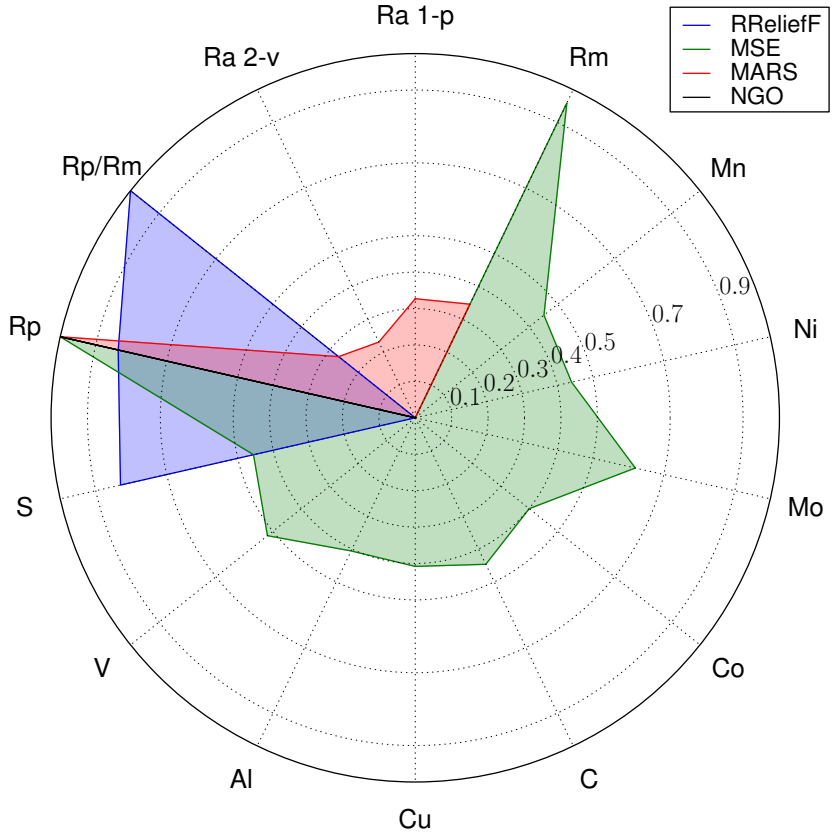
Na sliki 4.2 vidimo rezultate za drugo podatkovno množico. Jasno je, da so različne metode različno ocenile posamezne lastnosti. Metoda RRelifF je vse lastnosti, ki niso bile izločene s permutacijskim testom, ocenila relativno visoko. Metodi najmanjšega standardnega odklona in MARS sta v nasprotju z metodo RReliefF Youngov modul, vsebnost kemijskega elementa volframa in molibdena ocenila kot nepomembne. Metoda naključnih gozdov je dala nekoliko drugačne rezultate, saj je kot pomembne razpoznala le napetost pri raztezku 0,2 % (Rp), največjo silo Fm, vsebnost železa in silicija.



Slika 4.2: Najbolje ocenjeni atributi, množica 2

4.1.3 Podatkovna množica 3

V tretji podatkovni množici (slika 4.3) je metoda RReliefF kot pomembne razpoznala samo napetost pri raztezku 0,2 %, razmerje Rp/Rm in vsebnost žvepla, metoda najmanjšega standardnega odklona pa napetost pri raztezku 0,2 %, največjo napetost in vsebnost kemijskih elementov žvepla, vanadija, aluminija, bakra, ogljika, kobalta, molibdena in niklja. Metoda MARS je v tem primeru kot pomembne razpoznala napetost pri raztezku 0,2 %, natezno trdnost in razmerje obeh količin ter vzdolžno hrapavost na spodnji strani vzorca pločevine in prečno na zgornji. Metoda naključnih gozdov je v tem primeru za pomembno razpoznala samo mejo tečenja.



Slika 4.3: Najbolje ocenjeni atributi, množica 3

Poglavje 5

Napovedovanje trdote

Z regresijskimi metodami strojnega učenja smo se odločili oceniti možnost napovedovanja zvezne mehanske lastnosti trdote, izmerjene po metodi Vickers. Tudi za preizkus teh metod smo uporabili programski paket Orange in njegov derivat AZOrange. AZOrange je specializiran za delo na področju bioinformatike, mi pa smo ga uporabili zaradi vgrajene podpore nevronskim mrežam, ki jih izvorni Orange nima. Za implementacijo metode nevronskih mrež uporablja odprtokodno programsko knjižnico OpenCV (Open computer vision), ki ima implementiranih tudi nekaj v Orange že obstoječih metod, zato smo za primerjavo teste pognali tudi na njih. Tako smo iz izvornega Orange uporabili metode linearne regresije, regresijskih odločitvenih dreves, naključnih gozdov, k najbližjih sosedov, metode podpornih vektorjev in metode MARS, iz paketa AZOrange pa metode umetnih nevronskih mrež, metode podpornih vektorjev in metode naključnih gozdov. Pri metodi podpornih vektorjev smo preizkusili, kako se obnesejo različne jedrne funkcije na podatkovnih množicah, pri metodi umetnih nevronskih mrež pa število nevronov. V primerjavi z ostalimi metodami smo nato za vsako množico izbrali nastavitve, ki dajo najboljše rezultate. Pri ostalih metodah privzetih parametrov nismo spreminjali. Vsako podatkovno množico smo razdelili na učno in testno podmnožico v razmerju $9:1$. Pri oceni modelov smo uporabili postopek prečnega preverjanja z delitvijo na 10 podmnožic. Vsakega od šestih modelov smo tako 10-krat naučili, vsakič z drugačno učno podmnožico, in preverili na drugačni testni podmnožici. Na podlagi odstopanj med izračunano (y_i) in dejansko vrednostjo (t_i) smo izračunali povprečno korenjeno kvadratno napako (ang. root mean squared error - RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n-m}\sum_{i=m}^{n}(y_i - t_i)^2}\,, \tag{5.1}$$

in s povprečenjem dobili oceno za model. Prav tako smo med napovedano in testno množico izračunali Pearsonov korelacijski koeficient r

$$r = \frac{\text{Covar}(\mathbf{u}, \mathbf{v})}{\sqrt{\text{Var}(\mathbf{u})\text{Var}(\mathbf{v})}}\,, \tag{5.2}$$

kjer smo s $\text{Covar}(\mathbf{u}, \mathbf{v})$ označili kovarianco dveh spremenljivk in z $\text{Var}(\mathbf{u})$ varianco spremenljivke $\mathbf{u}$

$$\text{Covar}(\mathbf{u}, \mathbf{v}) = \frac{\sum_i (u_i - \bar{u})(v_i - \bar{v})}{n-m}\,, \tag{5.3}$$

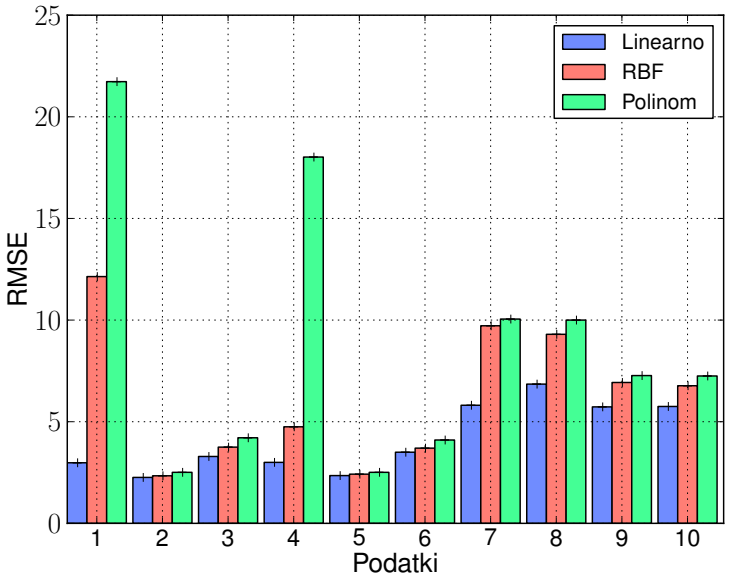
$$\text{Var}(\mathbf{u}) = \frac{\sum_i (u_i - \bar{u})}{n-m}\,. \tag{5.4}$$

Z $\bar{u}$ in $\bar{v}$ smo označili povprečni vrednost vektorja $\mathbf{u}$ in $\mathbf{v}$. Na enak način smo s povprečenjem določili korelacijski koeficient za celoten model. Vsak model smo zgradili dvakrat, prvič smo uporabili čisto vse razpoložljive atribute, drugič pa le najbolje ocenjene v četrtem poglavju. Poleg regresijskih modelov, ki smo jih želeli preskusiti za napovedovanje razreda, smo uporabili tudi naivni model, ki vedno vrne povprečno vrednost razreda iz učne množice. Ta bo služil kot testni primer, koliko so ostali preskušeni modeli zares relevantni. Za vsak model smo celoten postopek testiranja ponovili 100-krat in kot končni rezultat vzeli povprečje vseh rezultatov. Prav tako smo za povprečno korenjeno kvadratno napako in korelacijski koeficient izračunali standardni odklon. Rezultate smo prikazali s stolpičnimi diagrami, na katerih je prikazan tudi standardni odklon posameznih količin. Zaradi preglednejšega označevanja smo na grafih uporabili okrajšave imen preizkušenih metod: PV označuje povprečno vrednost razreda oziroma naivni model, RDR regresijska drevesa, KNS metodo k najbližjih sosedov, NGO metodo naključnih gozdov, MPV metodo podpornih vektorjev, UNM mtodo umetnih nevronskih mrež, cMPV in cNGO pa implementacijo openCV metode podpornih vektorjev oziroma naključnih gozdov.

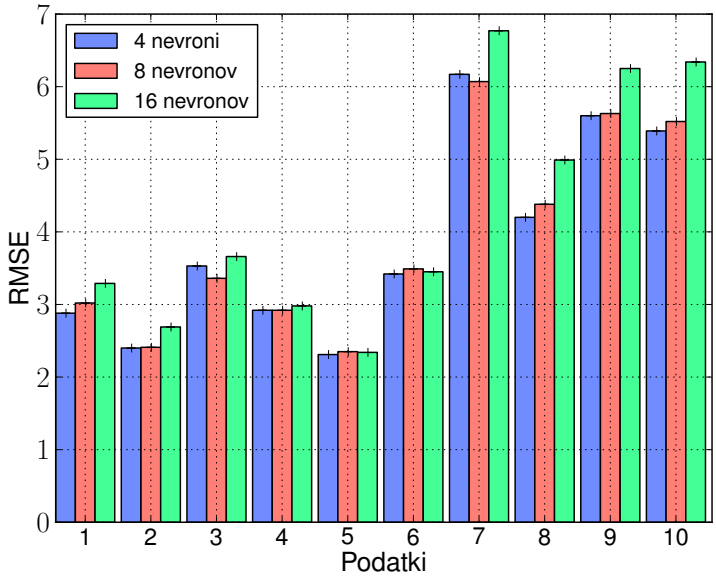
5.1 Rezultati

5.1.1 Nastavitve prostih parametrov

Na testnih množicah smo preizkusili različne parametre metode podpornih vektorjev in umetnih nevronskih mrež. Pri prvi smo preizkusili tri različne jedrne funkcije (linearno, radialno bazno ter polinomsko). Iz slike 5.1 je očitno, da se v vseh primerih najbolje odreže kar privzeta, linearna jedrna funkcija. Pri umetnih nevronskih mrežah smo naredili primerjavo za različno število nevronov v skriti plasti (4, 8 in 16). Na sliki 5.2 pa vidimo, da v večini primerov dajo najboljše rezultate že samo štirje nevroni v skriti plasti.



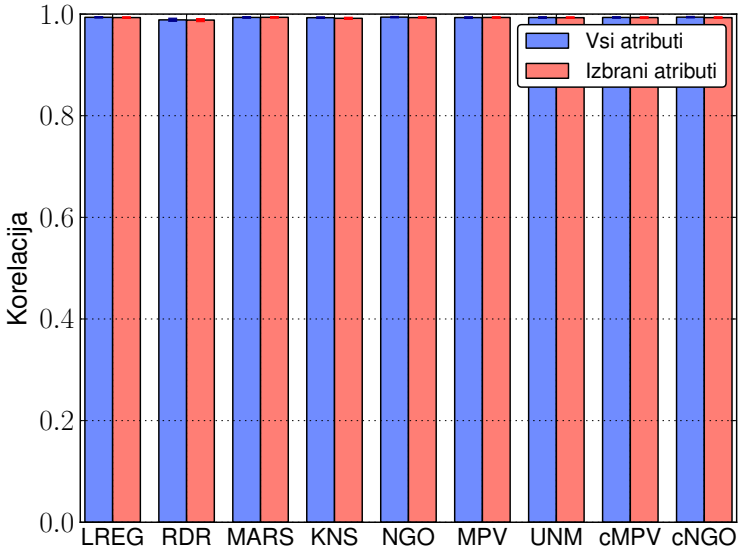
Slika 5.1: Različne jedrne funkcije pri metodi podpornih vektorjev



Slika 5.2: Različno število nevronov v skriti plasti pri metodi umetnih nevronskih mrež

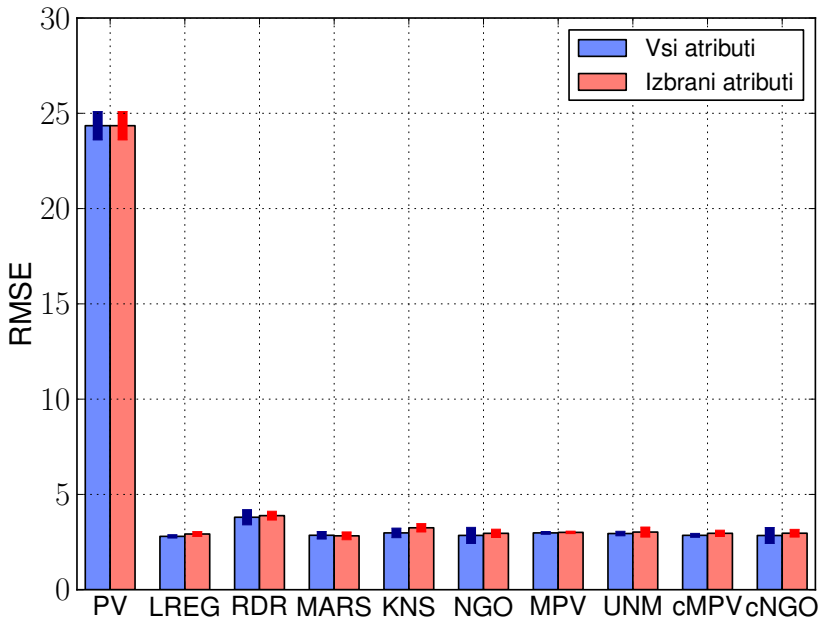
5.1.2 Množica 1

V prvi množici podatkov smo imeli polovico vzorcev izmerjenih pred toplotno obdelavo, polovico pa po njej. Postopek toplotne obdelave močno spremeni mehanske lastnosti pločevine, zaradi tega se nam znotraj množice oblikujeta dve skupini vzorcev, ki sta na sliki 2.7a lepo vidni. To neugodno upliva na izračun korelacijskega koeficienta, saj je ta pri vseh metodah praktično enak ena (slika 5.3). Prav tako naivni model vrača povprečje obeh skupin, zaradi česar je povprečna korenjena kvadratna napaka naivnega modela, za razliko od ostalih, zelo visoka (slika 5.4).



Slika 5.3: Vsi vzorci, korelacijski koeficient

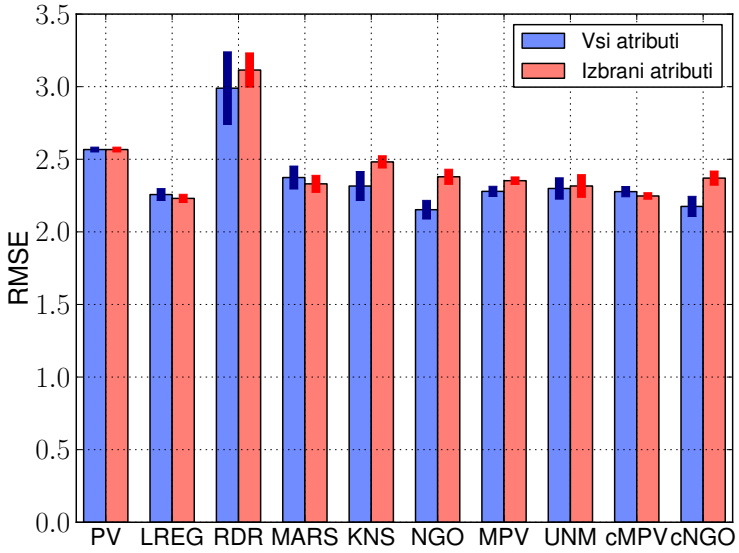
Zaradi tega smo se odločili to množico še dodatno razdeliti na dve podmnožici, kjer so zajeti le vzorci pred toplotno obdelavo ali po njej. V tem primeru so rezultati precej drugačni, povprečna korenjena kvadratna napaka med naivnim modelom in ostalimi se komaj opazno razlikuje, pri regresijskih drevesih je ta celo višja (sliki 5.5 in 5.7). Prav tako so zelo majhne razlike v korelaciji, izstopajo ponovno regresijska drevesa, ki se odrežejo najslabše (sliki 5.6 in 5.8). Pri korelaciji je potrebno omeniti zelo velik standardni odklon med večkratnim ponavljanjem testa, saj so vrednosti nihale tudi za 40 %, česar pri povprečni korenjeni kvadratni napaki ni bilo. Razlike med modeli, naučenimi z vsemi razpoložljivimi atributi, in samo najbolje ocenjenimi praktično ni.



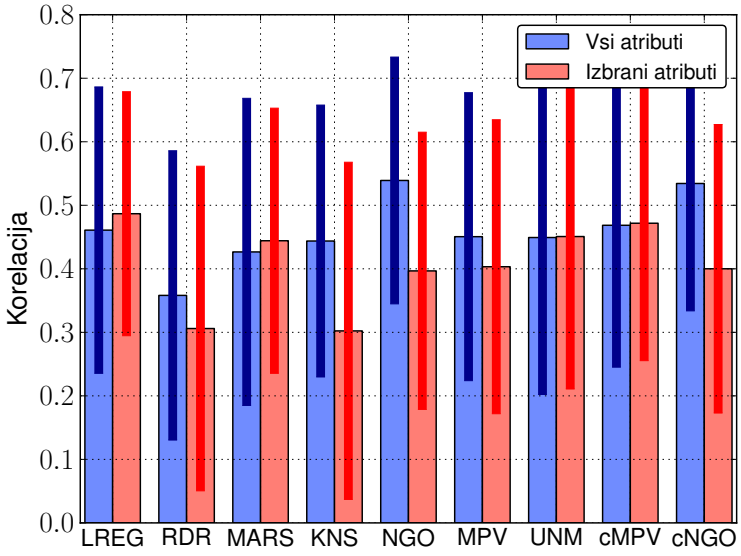
Slika 5.4: Vsi vzorci, povprečna korenjena kvadratna napaka

5.1.3 Množica 2

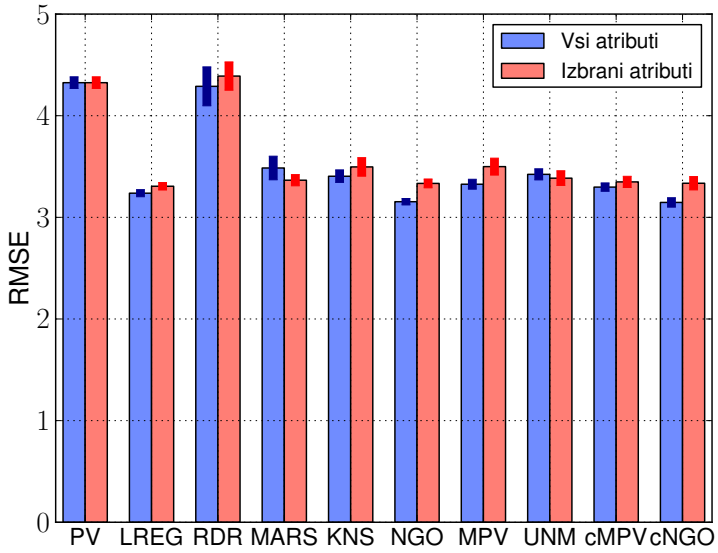
V drugi množici so rezultati nekoliko boljši, povprečna korenjena kvadratna napaka naivnega modela je blizu 10, pri večini ostalih modelov pa okrog 4. Nekoliko slabše se odreže metoda umetnih nevronskih mrež in implementacija metode podpornih vektorjev v paketu Orange. Razlike med modeli, naučenimi z vsemi razpoložljivimi atributi, in samo najbolje ocenjenimi ponovno skoraj ni, omembe vredna je le pri modelu umetnih nevronskih mrež (RMSE 7 proti 5). Razlik v korelaciji tudi v tem primeru praktično ni, so pa ponovno velika odstopanja med ponovitvami testa. Standardni odklon znaša tudi 50 %.



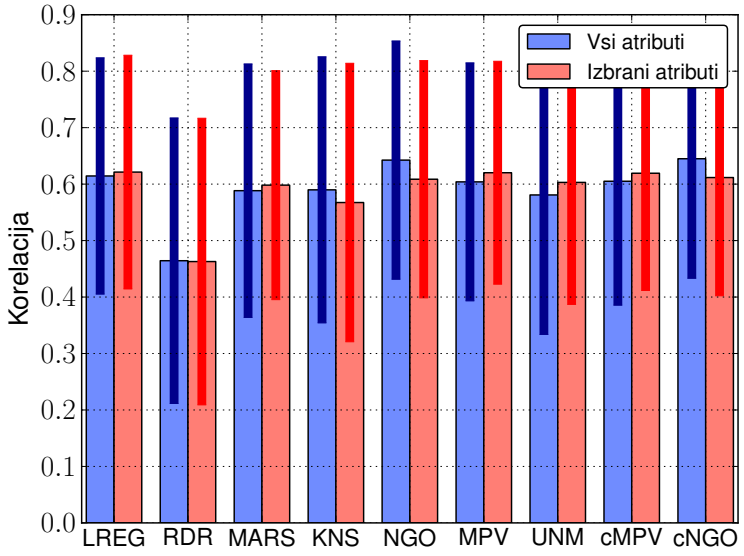
Slika 5.5: Toplotno obdelani vzorci, povprečna korenjena kvadratna napaka



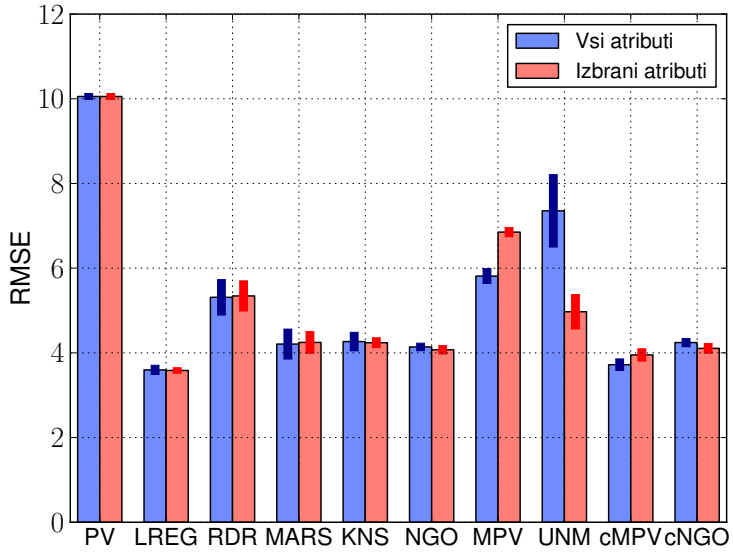
Slika 5.6: Toplotno obdelani vzorci, korelacijski koeficient



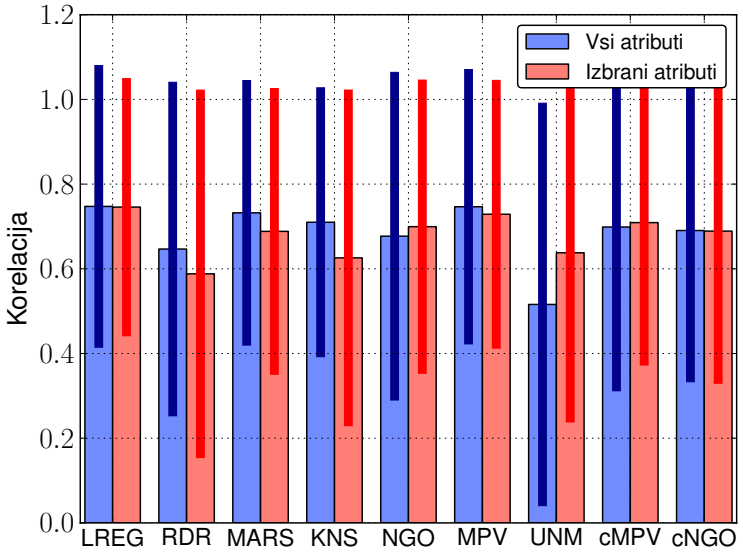
Slika 5.7: Toplotno neobdelani vzorci, povprečna korenjena kvadratna napaka



Slika 5.8: Toplotno neobdelani vzorci, korelacijski koeficient



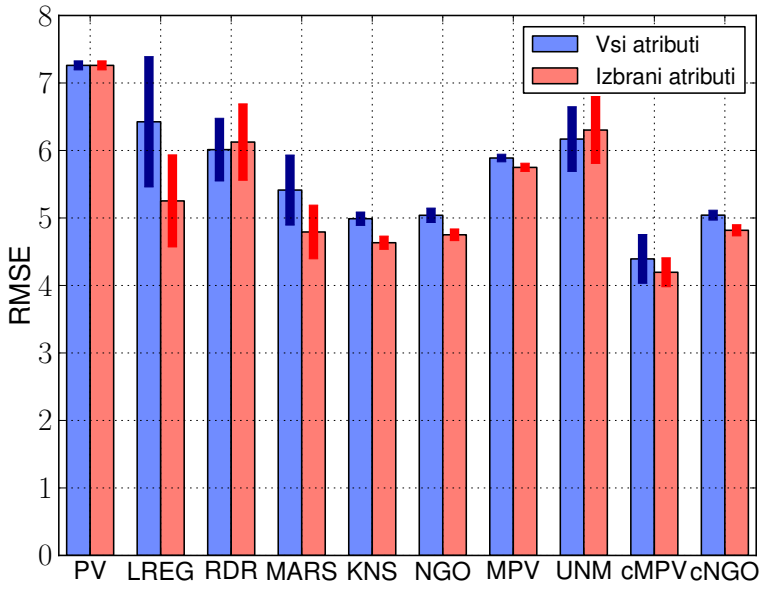
Slika 5.9: Povprečna korenjena kvadratna napaka



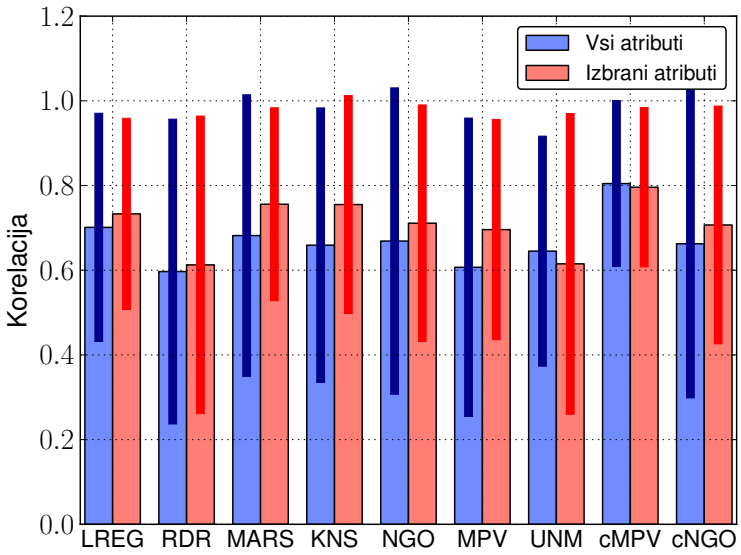
Slika 5.10: Korelacijski koeficient

5.1.4 Množica 3

Zelo podobno so se obnesli modeli tudi pri tretji množici podatkov, razlika v RMSE med naivnim modelom in najboljšim, z implementacijo AZOrange nevronskih mrež, znaša 3. Razlike med modeli, naučenimi z vsemi razpoložljivimi atributi, in samo najboljšimi ponovno praktično ni, največja razlika RMSE je pri linearni regresiji (približno 6 proti 5). Korelacija nam ponovno ne pove dosti, saj je za vse modele približno enaka ob zelo majhnem odstopanju med posameznimi ponovitvami (do 50 %).



Slika 5.11: Povprečna korenjena kvadratna napaka



Slika 5.12: Korelacijski koeficient

Poglavje 6

Zaključek

V diplomski nalogi smo poskušali iz mehanskih in kemijskih lastnosti pločevine napovedati njeno trdoto. Z metodami za oceno pomembnosti atributov smo ocenili, katere mehanske in kemijske lastnosti so najpomembnejše pri napovedovanju trdote pločevine. Uporabili smo več različnih regresijskih modelov, jih primerjali med seboj in primerjali razliko v natančnosti napovedi modelov, sestavljenih iz vseh razpoložljivih in iz samo najbolje ocenjenih atributov. Vse omenjeno smo storili z uporabo programskega paketa Orange, ki je tudi močno orodje za vizualizacijo podatkov. S pomočjo metode VizRank smo poiskali nekaj najbolj zanimivih povezav v podatkih in jih grafično prikazali.

6.1 Ugotovitve

Pridobljene ocene pomembnosti atributov pri napovedovanju trdote pločevine očitno veljajo, saj modeli zgrajeni samo iz najbolje ocenjenih atributov, dajo praktično enake rezultate, kot če uporabimo vse razpoložljive. Presek najbolje ocenjenih atributov vseh štirih metod je relativno velik in je bil tudi potrjen kot logičen pri osebju z znanji s področja strojništva in metalurgije. Napoved lastnosti trdote iz preostalih lastnosti pločevine ni dala uporabnih rezultatov. Povezava med količinami vsekakor obstaja, zaznali smo korelacijo med napovedanimi in testnimi množicami, čeprav majhno. Med ponovitvami poskusov učenja smo izmerili precejšnja nihanja izračunanih vrednosti. Standardni odklon je segal tudi to 50 % povprečne napovedane vrednosti. Izmerjena povprečna korenjena kvadratna napaka med napovedano in testno množico je znašala do 10 % absolutne napovedane vrednosti, kar ni slab rezultat. Vendar za opustitev izvajanja merjenja trdote v laboratoriju to ni dovolj. Pri učenju na prvi in tretji množici podatkov smo izmerili zelo majhno razliko v napaki

med naivnim modelom in ostalimi. To nakazuje na neustreznost modela za zanesljivo napovedovanje razreda. K slabim rezultatom in predvsem velikemu nihanju rezultatov so najverjetneje precej pripomogle relativno majhne množice podatkov. Zelo dobri so bili rezultati uporabe metode VizRank za iskanje najboljših grafičnih prikazov podatkov, kjer je metoda samodejno odkrila precej nazornih prikazov povezav med podatki.

6.2 Nadaljnje delo

Napovedovanje lastnosti trdote pločevine z regresijskimi metodami nadzorovanega atributnega učenja se ni izkazalo kot uporabno, bi pa bilo vredno namesto napovedovanja zveznih vrednosti poskusiti napovedati, ali vrednost trdote ustreza predpisanim tehnično prevzemnim pogojem. Tukaj zaradi narave dvorazrednega diskretnega razreda pridejo v poštev povsem druge metode strojnega učenja, imenovane uvrstitvene metode. Prav tako je zaradi majhnega števila trenutno dostopnih vzorcev bistveno, da se v podjetju čim prej vzpostavi podatkovni sistem za shranjevanje podatkov o meritvah. V podjetju že nekaj časa iščemo rešitev, ki bi neračunalnikarjem omogočala delo s podatkovno bazo izmerjenih fizikalnih lastnosti pločevine in morda še drugih vrst kovine. V diplomski nalogi smo spoznali drobovje programskega paketa Orange, ki se je izkazal kot zelo fleksibilen, a obenem za uporabo enostaven program. Z nekaj dopolnjevanja gradnikov in razširitvijo na neposredno interakcijo s podatkovno bazo bi Orange prestavljal idealno rešitev za naše potrebe.

Literatura

[1] M. Robnik-Šikonja, I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning Journal*, št. 53, str. 23-69, 2003.

[2] I. Kononenko, "Estimating attributes: analysis and extensions of Relief," v zborniku *European Conference on Machine Learning*, Catania, Italy, apr. 1994, str. 171-182.

[3] K. Kira, L. A. Rendell, "The feature selection problem: traditional methods and new algorithm," v zborniku *10th National Conference on Artifical Intelligence*, San Jose, California, jul. 1992, str. 129-134.

[4] (2012) Multivariate Adaptive Regression Splines. Dostopno na: http://www.statsoft.com/textbook/multivariate-adaptive-regression-splines

[5] L. Breiman, "Bagging Predictors," *Technical report*, Department of Statistics, University of California, št. 421, 1994.

[6] A. Starič, "Pristopi strojnega učenja za tekmovanje UCSD Data Mining Contest," Diplomsko delo, Ljubljana, 2010.

[7] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, New York: Springer-Verlag, 2001, pogl. 9.

[8] M. Robnik-Šikonja, "Improving Random Forests," v zborniku *European Conference on Machine Learning*, Pisa, Italy, sept. 2004, str. 359-371.

[9] M. Bratina, A. Dobnikar, U. Lotrič, "Modeliranje časovnih vrst z metodami teorije informacij," *Elektrotehniški vestnik*, letn. 76, št. 4, str. 240-245, 2009.

[10] I. H. Witten, E. Frank, *Data Mining Practical Machine Learning Tools and Techniques*, San Francisco: Morgan Kaufmann, 2005, 2. izdaja, pogl. 4, 5, 6.

[11] A. Dobnikar, *Osnove teorije informacij*, Ljubljana, 2005, pogl. 2, 3.

[12] (2012) An Introduction to Data Mining. Dostopno na: http://chem-eng.utoronto.ca/ datamining/dmc

[13] G. Leban, B. Zupan in drugi, "VizRank: Data Visualization Guided by Machine Learning," *Data Mining and Knowledge Discovery*, št. 13, str. 119-136, 2006.

[14] P. Hoffmann, "A Survey of Visualizations for High-Dimensional Data Mining," *Information Visualization in Data Mining and Knowledge Discovery*, str. 47-82, 2002.

[15] (2012) Sheet metal cutting. Dostopno na: http://www.custompartnet.com/wu/sheet-metal-shearing

[16] (2012) Elastic deformation. Dostopno na: http://en.wikipedia.org/wiki/Deformation_(engineering)

[17] R. Mohorič, "Dodatek k tehnično prevzemnim pogojem za elektropločevini M310-50A in M800-65K," interna dokumentacija družbe Hidria, 2011.

[18] (2012) Programski paket Orange: dokumentacija. Dostopno na: http://orange.biolab.si/doc