UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Andrej Panjan

NAPOVEDOVANJE USPEŠNOSTI TENIŠKIH IGRALCEV Z METODAMI STROJNEGA UČENJA

DIPLOMSKO DELO NA UNIVERZITETNEM ŠTUDIJU

Mentor: doc.dr. Janez Demšar

Ljubljana, 2009

Univerza
v Ljubljani

Fakulteta za računalništvo
in informatiko

Tržaška 25
1000 Ljubljana, Slovenija
telefon: 01 476 84 11
faks: 01 426 46 47
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 01584/2009
Datum: 01.09.2009

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

Kandidat: **ANDREJ PANJAN**

Naslov: **NAPOVEDOVANJE USPEŠNOSTI TENIŠKIH IGRALCEV Z METODAMI STROJNEGA UČENJA**

PREDICTION OF SUCCESSFULNESS OF TENIS PLAYERS USING MACHINE LEARNING

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

Pri vzgoji teniških igralcev je pomembno čim prej identificirati perspektivne mlade igralce. Na uspešnost igralca vplivajo različni motorični in morfološki dejavniki. V okviru diplomskega dela z metodami strojnega učenja sestavite napovedne modele za uspešnost igralcev in preskusite njihovo natančnost. Posebno pozornost posvetite analizi modelov, v okviru katere identificirajte najpomembnejše dejavnike. Sestavite tudi modele, ki na podlagi izmerjenih podatkov napovedujejo uspešnost športnika čez nekaj let in ocenite njihovo uporabno vrednost v praksi.

Mentor:
doc. dr. Janez Demšar

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

Dekan:
prof. dr. Franc Solina

ZAHVALA

Za mentorstvo, vodenje in pomoč se iskreno zahvaljujem doc. dr. Janezu Demšarju.

Posebej bi se zahvalil doc. dr. Nejcu Šarabonu za izdatno pomoč pri zbiranju podatkov, nasvete, komentarje in ideje.

Zahvala gre tudi doc. dr. Alešu Filipčiču za pomoč pri zbiranju podatkov.

Nenazadnje hvala tudi družini, ki me je podpirala skozi vsa leta študija.

KAZALO

Povzetek

V tej nalogi je narejena raziskava napovedovanja tekmovalne uspešnosti slovenskih teniških igralcev v treh selekcijah v moški in ženski konkurenci z metodami strojnega učenja.

Strojno učenje postaja vse bolj pomembno orodje za odkrivanje znanja v podatkih. Razlog za to je v zelo intenzivnem večanju števila podatkov, ki smo mu priča v zadnjih letih. Ker se v zadnjih letih tudi v tenisu opravlja vse več meritev , smo se odločili, da naredimo raziskavo napovedovanja tekmovalne uspešnosti slovenskih teniških igralcev. Pridobljeni rezultati bodo lahko v pomoč teniški stroki pri vzgoji mladih teniških igralcev.

Na začetku je predstavljen problem napovedovanja tekmovalne uspešnosti teniškega igralca. Na kratko so predstavljeni vsi dejavniki, ki vplivajo na tekmovalno uspešnost. Bolj podrobno pa so predstavljeni dejavniki, ki so bili uporabljeni v tej raziskavi. To so motorični in morfološki dejavniki. Zatem so predstavljeni cilji te raziskave, teoretične osnove strojnega učenja, potrebne za rešitev problema, in metode dela, po katerih je bila narejena raziskava. Na koncu so podani rezultati in interpretacija napovedovanja tekmovalne uspešnosti teniških igralcev v posameznih trenutnih starostnih obdobjih in napovedovanja tekmovalne uspešnosti za bodoča starostna obdobja. Pri tem so bile uporabljene klasifikacijske in regresijske metode strojnega učenja z dvema pristopoma za izbiro optimalne podmnožice atributov.

Ključne besede:

tekmovalna uspešnost, strojno učenje, morfološki dejavniki, motorični dejavniki.

Abstract

In this thesis, a survey of prediction of successfulness of Slovenian tennis players using machine learning methods is conducted. Tests are performed for three different selections, in male and female competition.

Machine learning is becoming an increasingly important tool for knowledge discovery in data. The reason for this is a rapid increase of data that is collected in recent years. Since more and more measurements are performed in tennis in recent years, we decided to do a survey of prediction of successfulness of Slovenian tennis players. The results obtained will be of assistance to tennis profession in the education of young tennis players.

At the beginning a problem of prediction of tennis player successfulness is presented. Factors that affect the successfulness are presented briefly, while the factors that are used in the study are discussed in detail. The study focuses on the motor and morphological factors. The objectives of this survey, the machine learning theory used in the solution and working methodology used in the survey are presented next. The results and the interpretation of prediction of successfulness of tennis players for current and future age periods are given at the end. For prediction, classification and regression methods of machine learning with two different approaches for optimal attribute subset selection were used.

Keywords:

successfulness, machine learning, morphological factors, motor factors.

1. Uvod

Šport je za sodobnega človeka ena izmed najpomembnejših postranskih dejavnosti v današnjem času. Šport je lahko v človekovem vsakdanjiku prisoten v več različnih oblikah in intenzivnostih. Najpogosteje se človek ukvarja s športom kot obliko rekreativne vadbe, ki mu pomaga do boljšega počutja ali sprostitve po napornem delavniku. Potem poznamo še obliko, ko se človek resneje ukvarja s športom, vendar mu to ne predstavlja vira zaslužka, ampak gre za neko obliko odvisnosti v dobrem pomenu besede. Nenazadnje je lahko šport tudi osnovni vir zaslužka človeka. V tem primeru se človek popolnoma posveti treningom in ustreznemu načinu življenja, saj le tako lahko pride do vrhunskih rezultatov. V vseh teh oblikah se pojavlja tudi tenis, ki je postal ena najbolj priljubljenih športnih zvrsti pri nas in tudi po svetu.

Tenis je športna igra z loparjem, ki poteka med dvema nasprotnima igralcema (»posameznika«) ali dvema nasprotnima ekipama s po dvema igralcema (»dvojice«). Igra se na igrišču, razdeljenem na dva enako velika dela. Ta dva dela ločuje nizko postavljena mreža. Za igro se uporablja lopar in gumijasta žogica. Igralec mora žogico odbiti na nasprotnikovo polovico, preden se žogica dvakrat odbije na njegovi polovici. Igralec dobi točko, ko se žogica najmanj dvakrat odbije na nasprotnikovi polovici ali nasprotnik ne uspe odbiti žogice nazaj v polje igralca, ali nasprotnik ne udari žogice z loparjem ali pa nasprotnik udari žogico na nasprotni polovici igrišča.

Začetki tenisa segajo v drugo polovico 19. stoletja, ko so ga začeli igrati v Angliji. V naše kraje je igra prišla okoli leta 1880. Na začetku se je tenis najbolj razvijal v Ljubljani in Mariboru. Prvo pravo teniško igrišče v Sloveniji pa je dal zgraditi znani pisatelj Ivan Tavčar na Visokem v Poljanski dolini.

Bolj intenzivno širjenje teniške igre je pri Slovencih prišlo v začetku osemdesetih let dvajsetega stoletja, ko so se začela graditi številna teniška igrišča in dvorane. Hkrati se je začelo povečevati tudi število rekreativnih in profesionalnih igralcev.

Teniška zveza Slovenije, kot nacionalna in strokovna organizacija, združuje klube s področja Slovenije in je članica evropske teniške zveze in mednarodne teniške federacije. Teniška zveza Slovenije skrbi za organizacijo turnirjev, vodenje jakostnih lestvic po posameznih kategorijah, izobraževanje strokovnih delavcev, sodnikov, ažurnost teniških pravil itd.

Kljub majhnosti Slovenije in omejenosti finančnih in drugih sredstev v primerjavi z večjimi in bolj razvitimi državami so nekateri tekmovalci in tekmovalke že dosegli odmevnejše rezultate na mednarodnih tekmovanjih in s tem pokazali, da se lahko kosajo z najboljšimi svetovnimi igralci. Vendar se je potrebno zavedati, da bo brez razvoja strokovno-izobraževalnega in znanstveno-raziskovalnega dela vse težje držati stik z najboljšimi tekmovalci.

Za razliko od Slovenije imajo teniško razvite države številne ustanove, ki se ukvarjajo s športno in teniško znanostjo in skrbijo za iskanje novih dognanj o teniški igri. Visoko usposobljeni strokovni centri izvajajo načrtno treniranje mladih teniških igralcev ter hkrati v praksi preverjajo ugotovitve raziskav. Slovenija takšnih pogojev, kot jih imajo mnoge druge države, še nima, vendar se tudi v Sloveniji stvari premikajo na bolje.

V tej nalogi je predstavljena raziskava, ki naj bi s pomočjo metod umetne inteligence, poskušala pojasniti tekmovalno uspešnost teniških igralcev v moških in ženskih selekcijah v Sloveniji.

Strojno učenje je področje umetne inteligence, ki se ukvarja z odkrivanjem znanja v podatkih, analizo podatkov, avtomatskim generiranjem baz znanja za ekspertne sisteme, gradnjo numeričnih in kvalitativnih modelov, klasifikacijo, regresijo itd. Z zelo intenzivnim večanjem števila podatkov v digitalni obliki, ki smo mu priča v zadnjih letih, postaja strojno učenje vse bolj pomembno orodje za transformacijo teh podatkov v koristno informacijo, saj je postalo ročno obdelovanje take množice podatkov nemogoče. Uveljavljanje strojnega učenja se odraža tudi z vse več komercialnimi sistemi v industriji, medicini, ekonomiji, bančništvu itd. Osnovni princip strojnega učenja je avtomatsko opisovanje (modeliranje) pojavov iz podatkov. Naučeni modeli poskušajo razlagati podatke, iz katerih so bili modeli zgrajeni, in se lahko uporabljajo za odločanje pri opazovanju modeliranega procesa v bodočnosti (napovedovanje, diagnosticiranje, nadzor, preverjanje, simulacije itd.).

Raziskave strojnega učenja so začele pridobivati na veljavi od osemdesetih let prejšnjega stoletja. To se odraža z vse več mednarodnimi konferencami (npr. International Machine Learning Conference), specializiranimi znanstvenimi revijami (npr. Journal of Machine Learning Research) in nenazadnje z uspehi na realnih problemih v industriji, medicini, bančništvu itd.

V Ljubljani se s strojnim učenjem že od leta 1980 uspešno ukvarjajo v Laboratoriju za umetno inteligenco in Laboratoriju za kognitivno modeliranje na Fakulteti za računalništvo in informatiko ter Odsek za inteligentne sisteme in Odsek za tehnologije znanja na Inštitutu Jožef Stefan. Uspehi laboratorijev so vidni iz številnih publikacij, njihovega citiranja v svetovni literaturi in prenosu metod strojnega učenja v prakso.

2. Predmet in problem

2.1. Tekmovalna uspešnost

Tekmovalna uspešnost teniškega igralca je trenutna uspešnost igralca v primerjavi s trenutno aktivnimi igralci znotraj starostne kategorije, v kateri igralec nastopa. Tako ne moremo neposredno primerjati igralcev, ki so igrali v različnih časovnih obdobjih, zaradi naslednjih razlogov: konkurenca v nekem časovnem obdobju je lahko močnejša ali šibkejša v primerjavi z drugim časovnim obdobjem, skozi leta so se materiali, iz katerih so bili izdelani športni rekviziti, spreminjali, skozi leta so se spreminjala pravila itd. Prav tako ne moremo neposredno primerjati igralcev v različnih starostnih obdobjih ali igralcev različnega spola. Seveda se zmeraj delajo primerjave in razne lestvice skozi več časovnih obdobij, ki temeljijo na predpostavki, da je v povprečju konkurenca zmeraj enako močna.

2.2. Dejavniki tekmovalne uspešnosti

V profesionalnem tenisu nas zanima predvsem, kakšna bo tekmovalna uspešnost nekega igralca v članski konkurenci, saj le uspešni rezultati v članski konkurenci prinašajo tudi finančno nagrado. Napovedovanje tekmovalne uspešnosti mladih teniških igralcev v članski konkurenci se lahko izvaja na več različnih načinov. Tako poznamo v športu pogosto metodo, ko si strokovnjak z določenega področja ogleda nekaj nastopov določenega športnika in na podlagi svojih izkušenj in ekspertize napove uspešnost športnika. Pogosto se uporablja tudi primerjanje športnika z ostalimi športniki v določeni kategoriji. Tretja metoda je ocenjevanje tekmovalne uspešnosti na podlagi merljivih lastnosti športnika. Predvsem pri tej zadnji metodi je zanimivo preveriti, ali metode strojnega učenja uspešno napovedujejo tekmovalno uspešnost teniškega igralca znotraj določene kategorije in napovedovanje tekmovalne uspešnosti v kasnejših obdobjih na podlagi lastnosti pri določeni starosti.

2.3. Vrednotenje tekmovalne uspešnosti

Tekmovalna uspešnost igralca je odvisna od več notranjih in zunanjih dejavnikov. Notranji dejavniki določajo in opisujejo igralčevo zdravstveno stanje, morfološke značilnosti, funkcionalne, motorične in teniško motorične sposobnosti, tehnično in taktično znanje, kognitivne sposobnosti in psihosocialne lastnosti ter igralčev vrednostni sistem in socialni status. Zunanji dejavniki so sestavljeni iz objektivnih dejavnikov (igrišča, vremenski pogoji, sodnik, gledalci) in pogojev treniranja in tekmovanja (trener, rekviziti in pripomočki, finančna sredstva). Pomemben zunanji dejavnik, ki se kaže v času tekme, je nasprotnik s svojimi sposobnostmi, značilnostmi in lastnostmi. Določen delež k rezultatu prispeva tudi napaka oziroma prispevek nam nepoznanih dejavnikov.

Kot je razvidno iz prejšnjega odstavka, vpliva na tekmovalno uspešnost veliko dejavnikov. Nekatere od njih je zelo težko meriti, zato se jih navadno izpušča pri napovedovanju tekmovalne uspešnosti. V tej nalogi bomo uporabili dejavnike, ki jih je enostavno meriti in po

ocenah strokovnjakov s področja tenisa v večji meri vplivajo na tekmovalno uspešnost igralcev.

2.4. Opis uporabljenih dejavnikov

V tej nalogi so bili uporabljeni dejavniki morfoloških značilnosti, motorične in teniško motorične sposobnosti.

2.4.1 Morfološke razsežnosti

Morfološke razsežnosti predstavljajo telesno konstitucijo posameznika, to je njegove telesne razsežnosti, ki so specifična, strukturna in funkcionalna manifestacija posameznika. Temeljni usmerjevalec razvoja telesne konstitucije je genom. Morfološke razsežnosti lahko pozitivno ali negativno vplivajo na učinkovitost izvajanja gibanja. Pomembnost morfoloških razsežnosti na učinkovitost izvajanja določenih motoričnih nalog je zelo odvisna od same narave gibanja. Vzorec in opis spremenljivk morfoloških razsežnosti, ki so bile uporabljene v tej nalogi je opisan v tabeli 2.1.

Šifra	Ime testa
ATV	Telesna višina
ADZGO	Dolžina zgornjega uda
ADSPO	Dolžina spodnjega uda
ASR	Širina ramen
ASM	Širina medenice
APKOM	Premer komolca
APZ	Premer zapestja
APKOL	Premer kolena
APG	Premer gležnja
AON	Obseg nadlahti
AONMA	Obseg pokrčene nadlahti
AOP	Obseg podlahti
AOPR	Obseg prsi
AOPMA	Maksimalni obseg prsi
AOS	Obseg stegna
AOSLS	Srednji obseg stegna
AOG	Obseg goleni
AKGH	Kožna guba hrbta
AKGN	Kožna guba nadlahti (tricepsova)
AKGB	Kožna guba nadlahti (bicepsova)
AKGP	Kožna guba podlahti
AKGPR	Kožna guba prsi
AKGT	Kožna guba trebuha
AKGS	Kožna guba stegna
AKGSI	Kožna guba suprailiakalna
AKGG	Kožna guba goleni
ATT	Telesna teža

Tabela 2.1: Opis spremenljivk morfoloških razsežnosti

Večina morfoloških spremenljivk je takšnih, da je zelo malo verjetno, da bi bil njihov doprinos k tekmovalni uspešnosti neposreden. Zato smo večino spremenljivk združili v štiri nove spremenljivke, za katere predpostavljamo, da je njihov doprinos k tekmovalni uspešnosti neposreden. Novo ustvarjene spremenljivke so: indeks telesne teže (BMI), odstotek mišične mase (AMISP), odstotek kostne mase (AKOSP) in odstotek mase maščobnega tkiva(AMASP).

BMI je definiran kot:

$$BMI = \frac{ATT}{ATV^2}$$

AMISP je definiran kot:

$$AMISP = [AV * (0{,}0553 * (AOS - AKGS)\text{^}2 + 0{,}0987 * (AOP)\text{^}2 - 0{,}0331 * (AOM - AKGG)\text{^}2\ ) - 2445]/ATT * 100$$

AKOSP je definiran kot:

$$AKOSP = \frac{APKOM + APZ + APKOL + APG}{4 * ATT} * ATV * 120$$

AMASP je definiran kot:

$$AMASP = 13 * \frac{AKGB + AKGS + AKGPR + AKGT}{12 * ATT} * \sqrt{\frac{ATT + ATV}{3600}}$$

2.4.2 Motorične sposobnosti

Motorične sposobnosti predstavljajo posameznikovo sposobnost gibanja, moči, koordinacije in ravnotežja, zato jih naprej delimo na moč, hitrost, koordinacijo, gibljivost in ravnotežje. Vzorec in opis spremenljivk motoričnih razsežnosti, ki so bile uporabljene v tej nalogi je opisan v tabeli 2.2.

Šifra	**Ime testa**
MSARG	Sargent skok
MM2	Met medicinke (2 kg)
MSKOK	Četveroskok
MDT60	Dviganje trupa
MT20	Tek na 20 m
MT9X6	Tek 9 x 6 m
MREAK	Reakcijska palica
MTAPN	Taping z nogo
MTAPR	Taping z roko
MTPK	Predklon na klopici
MZVIN	Zvinek s palico
MIZPK	Izpadni korak
MPAH	Pahljača
MHEK	Heksagon
MHST	Hitrost stopanja
MPOL	Poligon nazaj
MOZL	Odbijanje žoge z loparjem
MOSMI	Osmica s pripogibanjem
MOBR	Obrati na gredi
MHOJA	Hoja po gredi in odbijanje
MPRIS	Prisunski koraki po gredi
M2400	Tek na 2400 m

Tabela 2.2: Vzorec in spremenljivk motoričnih razsežnosti

2.5. Napovedovanje tekmovalne uspešnosti na podlagi dejavnikov, ki so jih izbrali slovenski trenerji

Trenerji so tisti subjekti v tenisu, ki poskušajo posameznega igralca z različnimi metodami in pristopi pripeljati do čim večje uspešnosti. V ta namen se trenerji sproti izobražujejo in tako nadgrajujejo svoje znanje in sposobnosti. Sposobnost trenerja mora biti tudi pravilno načrtovati treninge tako, da igralec napreduje od svojega začetka ukvarjanja s tenisom in potem skozi celotno kariero profesionalnega igralca. S tega stališča je zanimivo preveriti, kako dobro se lahko napove tekmovalna uspešnost na podlagi dejavnikov, ki vplivajo na tekmovalno uspešnost in so jih izbrali slovenski trenerji.

3. Cilji

Z raziskavo smo želeli doseči naslednje cilje:

- ugotoviti smiselnost napovedovanja tekmovalne uspešnosti za različna starostna obdobja v moški in ženski konkurenci z metodami strojnega učenja na podlagi meritev,
- ugotoviti smiselnost napovedovanja tekmovalne uspešnosti za različna starostna obdobja v moški in ženski konkurenci z metodami strojnega učenja na podlagi meritev v predhodnih starostnih obdobjih,
- preizkusiti napovedovanje tekmovalne uspešnosti s klasifikacijskimi metodami; pri tem razdelimo igralce na zelo dobre in dobre,
- preizkusiti napovedovanje tekmovalne uspešnosti z regresijskimi metodami,
- preizkusiti pristope za izboljšanje učenja,
- najti morfološke in motorične dejavnike, ki največ prispevajo k tekmovalni uspešnosti,
- preveriti napovedovanje tekmovalne uspešnosti za različna starostna obdobja v moški in ženski konkurenci na podlagi 5 najpogostejših dejavnikov, ki so jih izbrali slovenski trenerji.

4. Teoretične osnove strojnega učenja potrebne za rešitev problema

4.1. Predstavitev učnih primerov

Zbrane podatke je potrebno predstaviti v strukturi, ki je primerna za učinkovito uporabo v učnih algoritmih. Različni algoritmi uporabljajo različne strukture za predstavitev predznanja, učnih primerov in prostora hipotez. Zato je potrebno za vsak algoritem pripraviti podatke v ustrezni strukturi.

Znanje lahko predstavimo na naslednje načine:
- izjavni račun,
- predikatni račun prvega reda,
- diskriminantne in regresijske funkcije,
- verjetnostne porazdelitve.

Ker smo pri tej nalogi uporabljali samo izjavni račun, oziroma atributno predstavitev učnih primerov, bomo v nadaljevanju opisali samo ta način predstavitve znanja.

4.1.1 Atributna predstavitev učnih primerov

Pri klasifikacijskih in regresijskih problemih se najpogosteje uporablja atributna predstavitev učnih primerov. Atribut je spremenljivka, ki je lahko diskretna ali zvezna in ima določeno množico možnih vrednosti. Tudi razred je podan kot atribut in je diskreten, če rešujemo klasifikacijski problem, oziroma zvezen, če rešujemo regresijski problem. Atributom pravimo tudi značilke (angl. feature). Vsak učni primer je opisan z vektorjem vrednosti atributov. Tako je množica učnih primerov množica vektorjev vrednosti atributov.

Definicija atributne predstavitve:
množica atributov $A = \{A_i, i = 0 \dots a\}$;
- za vsak diskretni atribut A_i je množica vrednosti $V_i = \{V_1, \dots, V_n\}$;
- za vsak zvezni atribut A_i je interval vrednosti $V_i = [Min_i, Max_i]$;
- razred je podan z diskretnim atributom R, če rešujemo klasifikacijski problem, oziroma z zveznim atributom R, če rešujemo regresijski problem;
- učni primer je predstavljen z vektorjem vrednosti atributov $u_j = \langle v^{(1,j)}, v^{(a,j)}, r^{(j)} \rangle$, kjer je razred označen z $r^{(j)}$ in vrednosti atributov z $v^{(x,j)}$;
- množica učnih primerov je $U = \{u_j, j = 1 \dots n\}$

Vsak atribut ima različne lastnosti, od katerih je odvisna predobdelava učnih primerov in tudi izbira in nastavitev klasifikacijskega ali regresijskega algoritma.

Atribut z manjkajočimi vrednostmi je takšen atribut, pri katerem manjka vrednost pri enem ali več učnih primerih. Nekateri klasifikacijski in regresijski algoritmi lahko uporabljajo atribute z manjkajočimi vrednosti, nekateri pa tega ne znajo. Zato je potrebno v teh primerih takšne atribute predhodno odstraniti ali pa manjkajoče vrednosti nadomesti z vrednostmi, ki se jih lahko izračuna z različnimi algoritmi.

Šumen atribut vsebuje napake v podatkih bodisi zaradi napak pri meritvah ali zaradi napačnega vnosa podatkov. Tako so skoraj vsi realni podatki bolj ali manj šumni. Podobno kot pri atributih z manjkajočimi vrednostmi tudi šumne atribute lahko nekateri klasifikacijski in regresijski algoritmi uporabljajo, nekateri pa v tem primeru odpovejo, oziroma postanejo nezanesljivi. Primer algoritma, ki uspešno deluje na šumnih podatkih, je odločitveno drevo. Z rezanjem spodnjih nivojev drevesa zmanjšuje nezanesljivost klasifikacije, ki je posledica šuma v podatkih.

Naključen atribut je nepovezan z ostalimi atributi in s ciljno spremenljivko. Tak atribut je nepomemben in ga je najbolje odstraniti, saj samo znižuje zanesljivost učnih algoritmov. Primer naključnega atributa je zaporedna številka učnega primera, ki nima nobene povezave z razredom.

Redundanten atribut je atribut, katerega informacijo vsebuje že nek drug atribut ali množica atributov. Tak atribut je najbolje odstraniti.

Koreliran atribut je atribut, katerega informacijo delno vsebuje že nek drug atribut ali množica atributov. Bolj kot so atributi korelirani med seboj, večje so njihove soodvisnosti in več je redundance.

Močno soodvisni atributi glede na razred. Če so atributi močno soodvisni glede na razred je ciljno funkcijo težko odkriti, saj se ti atributi šele v kontekstu z drugimi pokažejo za pomembne. Primer močne soodvisnosti glede na razred sta atributa, ki sama ne povesta ničesar o razredu, skupaj pa ga zelo uspešno napovesta. Ekskluzivni ali je primer funkcije med dvema močno soodvisnima atributoma. Algoritem ReliefF, ki je opisan v poglavju 4.5.5, je sposoben uspešnega reševanja te vrste problemov.

4.2. Diskretizacija

Veliko algoritmov za strojno učenje zna uporabljati samo diskretne atribute. V takšnem primeru je potrebno vse zvezne atribute diskretizirati. Diskretizacija pomeni, da se interval vseh vrednosti zveznega atributa razdeli na določeno število podintervalov. Vse vrednosti znotraj tako nastalega posameznega intervala se preslikajo v isto vrednost. Zato se pri diskretizaciji informacija kvečjemu izgubi, saj nov atribut ne ločuje med vrednostmi znotraj posameznih intervalov.

Pri diskretizaciji atributov je potrebno glede na problematiko določiti optimalno število intervalov in optimalne meje za vse intervale, tako da se ohrani čim več informacije za klasifikacijo ali regresijo.

Najbolj pomembna je izbira optimalnega števila intervalov. Število intervalov se lahko izbere avtomatsko, če imamo ustrezno ocenitveno funkcijo, ali pa ročno z nastavitvijo parametra. Večje število intervalov pomeni manjšo izgubo informacije, vendar je aproksimacija verjetnostnih porazdelitev znotraj intervalov manj zanesljiva. Pri majhnem številu intervalov je ravno obratno, torej je več informacije izgubljene, vendar pa je aproksimacija verjetnostnih porazdelitev znotraj intervalov bolj zanesljiva.

Poznamo dva osnovna pristopa k diskretizaciji: od spodaj navzgor (angl. bottom-up) in od zgoraj navzdol (top-down). Pri obeh pristopih je potrebno najprej vrednosti atributov sortirati po velikosti.

4.3. Klasifikacijske metode strojnega učenja

4.3.1 Naivni Bayes

Naivni Bayesov klasifikator (angl. naive Bayes classifikator) je preprost verjetnostni klasifikator, ki temelji na Bayesovem teoremu s predpostavko pogojne neodvisnosti vrednosti različnih atributov pri danem razredu. Prav zaradi te predpostavke se imenuje naivni. Preprosto povedano, naivni Bayesov klasifikator predpostavlja, da je prisotnost (ali odsotnosti) določene vrednosti atributa pri danem razredu neodvisna od prisotnosti (ali odsotnosti) katerekoli vrednosti drugega atributa.

Kljub svojii naivni naravi in očitno preveč enostavni predpostavki, naivni Bayesov klasifikator pogosto deluje boljše v veliko kompleksnih realnih situacijah kot bi to lahko pričakovali. Prednost naivnega Bayesovega klasifikatorja je tudi v tem da potrebuje majhno učno množico, da oceni potrebne parametre (srednje vrednosti in varianco spremenljivk) za klasifikacijo. Pri učenju lahko primere, ki nimajo vrednosti za dani atribut, preprosto izpustimo.

Naivni Bayesov klasifikator lahko neposredno uporabljamo pri diskretnih atributih, pri zveznih pa je potrebno atribute najprej diskretizirati. Pri tem se nam splača uporabiti mehko diskretizacijo, kar pomeni, da določen primer ne pripada samo enemu intervalu, ampak več intervalom, vsakemu z določeno verjetnostjo.

Izkaže se, da je pogojna neodvisnost v praksi pogosta sprejemljiva predpostavka, saj je naivni Bayesov klasifikator kljub naivnosti zelo uspešen. Zelo dobro se obnaša tudi, kadar predpostavka o pogojni neodvisnosti ne drži popolnoma. Kadar pa obstajajo močne odvisnosti med atributi, pa naivni Bayesov klasifikator odpove. V tem primeru si lahko pomagamo z delno naivnim Bayesovim klasifikatorjem.

Odločitve naivnega Bayesovaga klasifikatorja si lahko razlagamo kot vsote informacijskih prispevkov posameznih atributov za posamezne razrede, če je informacijski prispevek pozitiven, oziroma proti posameznim razredom, če je informacijski prispevek negativen.

Osnovna formula naivnega Bayesovega klasifikatorja, izpeljana s pomočjo Bayesovega pravila, je

$$P(r_k|V) = P(r_k)\prod_{i=1}^{a}\frac{P(r_k|v_i)}{P(r_k)},$$

kjer so $P(r_k), k = 1 \dots n_0$ apriorne verjetnosti razredov in $P(r_k|v_i), k = 1 \dots n_0$ pogojne verjetnosti razredov pri dani vrednosti v_i atributa $A_i, i = 1 \dots a$.

4.3.2 Odločitvena drevesa

Odločitvena drevesa so zelo pogosto uporabljena metoda v strojnem učenju. V takšnih drevesnih strukturah predstavljajo listi klasifikacije, medtem ko so vozlišča konjunkcije atributov, ki vodijo do klasifikacij. V odločitveni teoriji je odločitveno drevo graf ali model odločitev in možnih posledic. S takšnim modelom si lahko pomagamo narediti tudi plan, kako doseči zastavljen cilj. Interpretacija odločitev odločitvenih dreves je enostavna in lahko razumljiva.

Odločitveno drevo je sestavljeno iz listov, ki ustrezajo razredu, vozlišč, ki ustrezajo posameznim atributom, in vej, ki ustrezajo podmnožicam vrednosti atributov. Ena pot od korena do lista predstavlja eno odločitveno pravilo.

Algoritem učenja odločitvenega drevesa:
Če je izpolnjen vsaj eden ustavitveni pogoj
    potem postavi list z vsemi ustreznimi učnimi primeri;
    sicer
        izberi »najboljši« atribut A_i;
        označi naslednike z vrednostmi atributa A_i;
        za vsako vrednost V_j atributa A_i ponovi:
            rekurzivno zgradi poddrevo z ustrezno podmnožico učnih primerov;

Ustavitveni pogoj je lahko:
- premalo učnih primerov za zanesljivo nadaljevanje gradnje (prag se določi ročno),
- vsi ali večina primerov je iz istega razreda (prag se določi ročno),
- ni več *dobrih* atributov.

Pri gradnji odločitvenega drevesa je ključnega pomena izbira atributa za razvejitev. Za izbiro *najboljšega atributa* se večinoma uporabljajo mere: informacijski prispevek, razmerje informacijskega prispevka, Gini-indeks in ReliefF.

Ker je odločitveno drevo lahko vizualizirati, je tako zanimivo tudi za strokovnjake z določene problemske domene, saj lahko brez veliko predznanja razbere določene zakonitosti in strukture, ki se pojavljajo v samem modelu.

Učni primeri so opisani z vektorjem vrednosti atributov in razredom. Medtem ko mora biti razred diskreten, so lahko atributi diskretni ali zvezni. Klasifikacijo novega primera se naredi tako, da potujemo od korena proti listom in se v vsakem vozlišču odločimo za ustrezno vejo, dokler ne pridemo do lista. Novemu primeru priredimo vrednost, ki je prirejena listu.

Slaba lastnost odločitvenih dreves je, da so nižji nivoji drevesa nezanesljivi, ker vozliščem ustreza majhno število primerov in se ti lahko preveč prilagajajo učni množici. Z namenom preprečevanja nezanesljivosti se uporabljajo tudi ustavitveni pogoji, ki ustavijo gradnjo, ko ta postane nezanesljiva. Ker pa je nezanesljivost vnaprej težko oceniti, se gradnja drevesa nadaljuje in se nato drevo naknadno poreže. Za ocenjevanje napake pri rezanju se največkrat uporabljata m-ocena, ki ocenjuje klasifikacijsko napako, in MDL princip, ki ocenjuje dolžino kodiranja drevesa in porazdelitev razredov učnih primerov v listih. Metoda MDL ima pred m-oceno to prednost, da ne zahteva nastavitve nobenega parametra, medtem ko je to pri m-oceni potrebno.

Potek rezanja drevesa je sledeč:
- pojdi po drevesu od spodaj navzgor in pregleduj vozlišča
- za vsako notranje vozlišče primerjaj oceno, ki jo dobimo, če poddrevesa porežemo in vozlišče postane list, in če poddreves ne porežemo
- če je ocena v vozlišču porezanega poddrevesa boljša od ocene neporezanega poddrevesa, potem poddrevo porežemo in nadaljujemo po drevesu navzgor
- sicer se ustavimo

4.3.3 Algoritem C4.5

C4.5 je algoritem za gradnjo odločitvenega drevesa, ki ga je razvil Ross Quinlan in je nadgradnja ID3 algoritma. Tudi odločitveno drevo, ki ga generira algoritem C4.5, se uporablja za klasifikacijo tako kot odločitveno drevo, opisano v prejšnjem razdelku.

Gradnja drevesa poteka tako, da algoritem C4.5 izbere atribut, ki najbolje loči množico primerov v podmnožice glede na razred. Kriterij za izbiro atributa je razmerje informacijskega prispevka (angl. gain-ratio), tako da je izbran atribut z največjim razmerjem informacijskega prispevka, na podlagi katerega se učni primeri razdelijo na podmnožice.

Psevdokoda algoritma:
- Najdi atribut z največjim razmerjem informacijskega prispevka.
- Ustvari odločitveno vozlišče, ki se razdeli glede na izbrani atribut.
- Rekurzivno ponavljaj postopek na podmnožicah in dodaj nova vozlišča kot otroke.
- Ustavi se, ko zmanjka atributov ali pa je dosežen pogoj za zaustavitev (npr. minimalno število primerov v listu).

4.3.4 K-najbližjih sosedov

K-najbližjih sosedov je algoritem za klasifikacijo primerov na podlagi najbližjih primerov iz učne množice v prostoru atributov in je eden najbolj enostavnih algoritmov v strojnem učenju. Klasifikacija novega primera se naredi na podlagi glasov k-najbližjih sosedov, in sicer tako, da je izbran razred, kateremu pripada največ glasov. Ker učenja pri tej metodi skoraj da ni, pravimo tej vrsti učenja tudi *leno učenje*. Parameter k je naravno število in je tipično majhno. Poseben primer je za k = 1, takrat se objektu pripiše razred, kateremu pripada najbližji sosed.

Primeri učne množice so predstavljeni kot vektorji v večdimenzionalnem prostoru atributov. Največkrat se uporablja Evklidska razdalja, ki je najbolj primerna za zvezne atribute, čeprav so uporabne tudi druge razdalje kot na primer Manhattanska, ki je primerna za diskretne atribute. Algoritem k-najbližjih sosedov je občutljiv na lokalno strukturo podatkov, saj se lahko s povečanjem k-ja spremeni napoved algoritma. V fazi učenja se shranijo samo vektorji atributov in oznake razredov, pri dejanski klasifikaciji pa se računajo razdalje med novim vektorjem in vsemi shranjenimi vektorji, iz katerih se izbere k-najbližjih. Obstaja več načinov klasifikacije novega vektorja, med katerimi je največkrat uporabljena tehnika klasifikacije glede na najbolj zastopan razred med k-najbližjimi sosedi. Največja slabost te tehnike je, da razredi z največ primeri prevladujejo pri klasifikaciji. Ena izmed možnih izboljšav je otežitev vsakega glasu iz množice k-najbližjih sosedov z razdaljo do novega primera.

Izbira parametra k je odvisna od množice podatkov in problema, ki ga rešujemo. Velike vrednosti izboljšujejo učenje na šumnih podatkih, hkrati pa zabrišejo meje med različnimi razredi. Dober k se lahko izbere tudi s pomočjo hevrističnih metod, kot na primer navzkrižno preverjanje (angl. cross-validation). Velja še omeniti, da parameter k ne določa velikosti okolice novega primera, znotraj katere izbiramo učne primere, ampak se okolica dinamično spreminja v odvisnosti od gostote učnih primerov v danem podprostoru primerov. Tako se elegantno reši problem gostejših oziroma redkejših delov prostora.

Klasifikacijska natančnost algoritma je lahko zelo zmanjšana zaradi prisotnosti šuma v podatkih ali irelevantnih atributov, kar se da izboljšati s pravilno izbiro podmnožice atributov.

4.3.5 Metoda podpornih vektorjev

Metode podpornih vektorjev (angl. SVM – support vector machine) so med najbolj uspešnimi metodami za klasifikacijo in regresijo. Metode SVM so primerne za učenje na velikih množicah z velikim številom manj pomembnih atributov. Slaba stran metod SVM je v tem, da je interpretacija odločitev zelo težavna, podobno kot pri nevronskih mrežah.

V osnovi so klasifikacijski SVM namenjeni razločevanju dveh razredov med seboj. Pri več razredih ponovimo postopek za vsak razred, ki ga želimo ločevati od ostalih razredov. Pri klasifikaciji nov primer klasificiramo v razred z najvišjo vrednostjo odločitvene funkcije.

Osnovni princip metode SVM je v danem prostoru atributov ločiti podana razreda, kar metoda SVM naredi s postavitvijo optimalne hiperravnine. Optimalna hiperravnina je tista ravnina, ki je enako in hkrati najbolj oddaljena od najbližjih primerov obeh razredov. Najbližjim primerom hiperravnine pravimo podporni vektorji. Razdalji hiperravnine od podpornih vektorjev pa pravimo rob (angl. margin). Tako je optimalna hiperravnina tista, ki ima optimalni rob. Ker pa v originalnem prostoru pogosto linearna hiperravnina ne zadošča za zahtevano klasifikacijsko natančnost, je potrebno ta prostor nelinearno transformirati. Pri tem uporabnik vnaprej izbere transformacijo in lahko na ta način rešuje različne nelinearne probleme.

Vsako hiperravnino lahko zapišemo kot množico točk X, ki zadošča

$$w \cdot x - b = 0,$$

kjer $\cdot$ pomeni skalarni produkt, w je normala in je pravokotna na hiperravnino. Parameter $\frac{b}{\|w\|}$ pa določa odmik hiperravnine od izhodišča vzdolž normale. Izbrati je potrebno takšen w in b, da maksimizirata rob.

4.3.6 Logistična regresija

Logistična regresija je metoda, ki zgradi linearni model na podlagi transformirane ciljne spremenljivke. Transformirana spremenljivka je aproksimirana z uporabo linearne funkcije tako kot pri linearni regresiji. Tako dobimo naslednji model:

$$\Pr[1|a_1, a_2, \ldots, a_k] = \frac{1}{(1 + \exp(-w_0 - w_1 a_1 - \ldots - w_k a_k))},$$

kjer so w_i vrednosti uteži in a_i vrednosti atributov. Podobno kot pri linearni regresiji morajo biti uteži takšne, da se dobljeni model dobro prilega učni množici.

Posplošitev logistične regresije za uporabo na večrazrednih problemih lahko naredimo tako, da zgradimo klasifikator za vsak par razredov, pri tem pa uporabimo primere samo s teh dveh razredov. Nov primer je klasificiran tako, da mu pripišemo razred, ki je dobil največ glasov posameznih klasifikacij. S to metodo se pogosto dosegajo dobri rezultati.

4.4. Regresijske metode strojnega učenja

4.4.1 Regresijska drevesa

Regresijsko drevo je podobno odločitvenemu, vendar imamo pri regresijskem drevesu zvezen razred. Ostali atributi so lahko zvezni ali diskretni. Učni primeri so predstavljeni z atributno predstavitvijo.

Regresijsko drevo je sestavljeno iz vozlišč, vej in listov. Vozlišča predstavljajo atribute, veje predstavljajo podmnožice vrednosti atributov in listi predstavljajo funkcije, ki preslikajo vektor vrednosti atributov v zvezni razred. Funkcije v listih so lahko različne in se razlikujejo od lista do lista. Najpogostejša funkcija je konstanta, pogosto se uporablja še linearna funkcija na podmnožici zveznih atributov, lahko pa tudi poljubna druga funkcija. Vsaka pot od korena do lista predstavlja eno odločitveno pravilo, pri čemer so pogoji, ki jih srečamo na poti, konjunktivno povezani.

Podobno kot pri odločitvenih modelih je tudi tukaj ključnega pomena izbira najboljšega atributa. Za izbiro najboljšega atributa se večinoma uporablja razlika variance ali regresijski ReliefF.

Algoritem učenja regresijskega drevesa:

Če je izpolnjen vsaj eden ustavitveni pogoj
    potem postavi list z vsemi ustreznimi učnimi primeri;
    generiraj funkcijo, ki modelira učne primere v listu;
    sicer
        izberi »najboljši« atribut A_i;
        označi naslednike z vrednostmi atributa A_i;
        za vsako vrednost V_j atributa A_i ponovi:
            rekurzivno zgradi poddrevo z ustrezno podmnožico učnih primerov;

Napovedovanje vrednosti odvisne spremenljivke izvedemo tako, da potujemo od korena proti listom in se v vsakem vozlišču odločimo za ustrezno vejo, dokler ne pridemo do lista. Vrednost odvisne spremenljivke novega primera je enaka vrednosti funkcije v listu, ki se nahaja v končnem listu.

Ker se da regresijsko drevo enostavno vizualizirati, je tako zanimivo tudi za strokovnjake z določene problemske domene, saj lahko brez veliko predznanja razbere določene zakonitosti in strukture, ki se pojavljajo v samem modelu.

Slaba lastnost regresijskih dreves je podobno kot pri odločitvenih drevesih nezanesljivost nižjih nivojev drevesa. Nezanesljivost je pogosto posledica majhnega števila učnih primerov, šuma v podatkih ali prevelikega prilagajanja drevesa učnim primerom. Zato moramo tudi tukaj uporabljati ustavitvene pogoje, ko postane učenje nezanesljivo, ali pa drevo naknadno porezati. V praksi se praviloma uporabljata kar obe metodi za izboljšanje zanesljivosti napovedovanja.

Ustavitveni pogoj je lahko:
- premalo učnih primerov za zanesljivo nadaljevanje gradnje (prag se določi ročno),
- vsi ali večina primerov je iz istega razreda,
- ni več »dobrih« atributov.

Osnovni algoritem rezanja drevesa je enak kot pri odločitvenem drevesu. Za ocenjevanje napake v vozlišču pa lahko uporabimo prirejeno m-oceno. Druga možnost rezanja drevesa pa je uporaba principa MDL, pri tem namesto napake ocenjujemo dolžino kodiranja drevesa in dolžino kodiranja napake na učnih primerih v listih.

4.4.2 Linearna regresija

Linearna regresija je pristop k modeliranju odnosa med odvisnimi in neodvisnimi spremenljivkami, tako da dobimo linearen model. Linearna regresija je bila prva med regresijskimi analizami, ki je bila močno zastopana v praksi. Razlog za to je, da se modeli, ki so linearno odvisni od svojih neznanih parametrov, lažje izračunajo, kot pa modeli, ki so nelinearno odvisni od parametrov. Če linearno regresijo uporabljamo za napovedovanje, izračunamo linearni model na učni množici, potem pa ta model uporabimo za napovedovanje razreda. Funkcija, ki jo dobimo pri učenju, je linearna kombinacija vseh ali podmnožice atributov. Linearna regresija deluje le na zveznih atributih, če so atributi diskretni, jih obravnavamo kot zvezne. Linearni regresijski modeli se večinoma izračunajo s pristopom minimiziranja vsote kvadratov napake, čeprav se lahko uporabijo tudi drugi pristopi.

Model linearne regresije:

$$y = X\beta + \varepsilon\,,$$

kjer je

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

y je vektor vrednosti razreda, X je množica vektorjev vrednosti atributov, β je vektor uteži in ε je vektor šuma.

Slaba stran linearne regresije je v tem, da *trpi* zaradi svoje linearnosti. Če so podatki nelinearno odvisni, bo linearna regresija modelirala premico, ki se najbolje prilega. Vendar ta premica ne opisuje najbolje nelinearne odvisnosti.

So pa linearni modeli dobri sestavni deli bolj kompleksnih algoritmov. Primer uporabe linearne regresije kot sestavnega dela bolj kompleksnega algoritma je modeliranje regresijske premice v listu regresijskega drevesa.

4.5. Mere v klasifikacijskih in regresijskih problemih

4.5.1 Splošno

Pri klasifikacijskih in regresijskih problemih je osnovna naloga algoritma oceniti pomembnost atributa za podani učni problem, če pri tem uporabljamo atributni zapis. Ker je bila množica primerov za naš problem podana z atributno predstavitvijo, bomo v tem razdelku uporabljali atributno predstavitev pri opisu mer v klasifikacijskih in regresijskih problemih.

Opis atributne predstavitve:

- množica atributov $A = \{A_i, i = 0 \dots a\}$;
- za vsak diskretni atribut A_i je množica vrednosti $V_i = \{V_1, \dots, V_n\}$;
- za vsak zvezni atribut A_i je interval vrednosti $V_i = [Min_i, Max_i]$;
- razred je podan z diskretnim atributom R, če rešujemo klasifikacijski problem, oziroma z zveznim atributom R, če rešujemo regresijski problem;
- učni primer je predstavljen z vektorjem vrednosti atributov $u_j = \langle v^{(1,j)}, v^{(a,j)}, r^{(j)} \rangle$, kjer je razred označen z $r^{(j)}$ in vrednosti atributov z $v^{(x,j)}$;
- množica učnih primerov je $U = \{u_j, j = 1 \dots n\}$.

Mere, ki se pogosto uporabljajo za usmerjanje iskanja v klasifikacijskih in regresijskih problemih, so:

- informacijski prispevek (angl. information gain),
- razmerje informacijskega prispevka (angl. gain-ratio),
- Gini-index,
- ReliefF,
- razlika variance.

4.5.2 Informacijski prispevek

Informacijski prispevek je ena najosnovnejših mer pomembnosti atributa. Informacijski prispevek atributa je definiran kot prispevana informacija atributa za določitev njegove vrednosti. Da bi lahko izračunali informacijski prispevek, moramo najprej definirati še entropijo in pogojno entropijo pri dani vrednosti atributa.

H - entropija:

$$H = -\sum_{k} p(k) \log_2 p(k) ,$$

kjer je k število razredov.

$H_{res}(A)$ - pogojna entropija razreda pri dani vrednosti atributa:

$$H_{res}(A) = -\sum_{j} p(j) \sum_{k} p(k|j) \log_2 p(k|j) ,$$

kjer je A izbrani atribut, k število razredov in j število podmnožic pri delitvi atributa A.

Sedaj pa lahko definiramo informacijski prispevek kot $Gain(A)$:

$$Gain(A) = H - H_{res}(A) ,$$

kjer je A izbrani atribut.

Informacijski prispevek je navadno dobra mera pomembnosti atributa, ni pa idealna. Problem se pojavi pri atributih z velikim številom različnih vrednosti, saj informacijski prispevek atributa kvečjemu raste s številom vrednosti. Zato se velikokrat namesto informacijskega prispevka uporablja razmerje informacijskega prispevka, ki je opisano v naslednjem razdelku.

4.5.3 Razmerje informacijskega prispevka

Razmerje informacijskega prispevka je normalizacija informacijskega prispevka z entropijo vrednosti atributa in odpravlja problem precenjevanja atributov z velikim številom različnih vrednosti, ki je slaba lastnost informacijskega prispevka. Vendar ima tudi razmerje informacijskega prispevka slabo lastnost, da precenjuje atribute z zelo majhno entropijo vrednosti. Možna rešitev tega problema je, da se pri izbiri upoštevajo samo atributi, ki imajo informacijski prispevek večji od povprečnega.

Razmerje informacijskega prispevka:

$$GainR(A) = \frac{Gain(A)}{H(A)} ,$$

kjer je $Gain(A)$ informacijski prispevek in $H(A)$ entropija vrednosti atributa.

4.5.4 Gini-indeks

$$Gini = \sum_{i \neq j} p(i) p(j) ,$$

kjer sta i in j razreda.

Gini-indeks atributa:

$$Gini(A) = \sum_{v} p(v) \sum_{i \neq j} p(i|v)p(j|v) ,$$

kjer je A izbrani atribut, v vrednost atributa A, i in j pa sta razreda.

Gini-indeks lahko interpretiramo kot pričakovano stopnjo napake. Podobno kot informacijski prispevek tudi Gini-indeks kvečjemu raste z večanjem števila vrednosti atributa in tako precenjuje večvrednostne atribute.

4.5.5 ReliefF

ReliefF za razliko od do sedaj opisanih mer ne predpostavlja apriorne in pogojne neodvisnosti atributov pri danem razredu, zato učinkovito deluje tudi v primeru odvisnih atributov.

Osnovna ideja algoritma je, da za vsak učni primer poišče najbližji primer iz istega razreda (najbližji zadetek) in najbližji primer iz nasprotnega razreda (najbližji pogrešek). Tako lahko oceni kvaliteto atributov glede na lokalne značilnosti razločevanja razredov. Ker pa lokalnost vključuje v oceno tudi druge atribute, ReliefF implicitno ocenjuje atribute v odvisnosti od ostalih atributov.

ReliefF lahko pri ocenjevanju atributov uporablja tudi nepopolne podatke. Na šumnih podatkih se zanesljivost bistveno poveča z izbiro k najbližjih zadetkov in k najbližjih pogreškov ter poišče povprečje njihovih prispevkov. Tipična vrednost parametra $k = 5 \ldots 10$. Pri večrazrednih problemih ReliefF namesto k najbližjih pogreškov iz nasprotnega razreda poišče k najbližjih pogreškov iz vsakega razreda. Prispevki posameznih razredov so dodatno obteženi z apriornimi verjetnostmi razredov.

Psevdokoda algoritma:

inicializiraj polje ocen atributov
določi število najbližjih primerov k
določi število iteracij algoritma m
for $i = 0$ **to** m **do**
    iz množice primerov naključno izberi primer I
    primeru I poišči k najbližjih zadetkov
    primeru I poišči k najbližjih pogreškov za vse ostale razrede
    iz najbližjih zadetkov in najbližjih pogreškov izračunaj oceno
    oceno zapiši v ustrezno polje ocen atributov
end
vrni polje ocen atributov

4.5.6 Pričakovana razlika variance

Pričakovana razlika variance se uporablja za ocenjevanje pomembnosti atributov v regresijskih problemih. Če hočemo definirati pričakovano razliko variance, moramo najprej definirati še varianco zveznega razreda, ki je povprečni kvadrat napake.

Varianca zveznega razreda:

$$s^2 = \frac{1}{n}\sum_{k=1}^{n}(r(k) - \bar{r})^2 ,$$

kjer je n število vrednosti zveznega razreda in $\bar{r}$ povprečna vrednost zveznega razreda med n primeri.

Sedaj lahko definiramo pričakovano razliko variance:

$$ds^2(A_i) = \frac{1}{n}\sum_{k=1}^{n}(r(k) - \bar{r})^2 - \sum_{j=1}^{n_i}\left(p_j \frac{1}{n_j}\sum_{k=1}^{n_j}\left(r_j(k) - \overline{r_j}\right)^2\right),$$

kjer je A_i izbrani atribut, n število vrednosti zveznega razreda, $\bar{r}$ povprečna vrednost zveznega razreda med n primeri, n_i število vrednosti izbranega atributa, n_j število vrednosti zveznega razreda, ki imajo j-to vrednost atributa A_i, $r_j(k)$ vrednost zveznega razreda k-tega primera, ki imajo j-to vrednost atributa A_i in $\overline{r_j}$ povprečna vrednost zveznega razreda primerov z j-to vrednostjo atributa A_i.

4.6. Mere za ocenjevanje učenja

Algoritmi strojnega učenja so večinoma zgrajeni avtomatsko na podlagi nekega postopka, ki večinoma ni odvisen od narave problema, ki ga rešujemo. Zato nikoli ne moremo vnaprej vedeti, kako uspešno bo zgrajeni model reševal določen problem. Da lažje ugotovimo, kateri modeli so dobri in kateri so slabi, si pomagamo z merami za ocenjevanje učenja. Mere za ocenjevanje učenja so nam samo v pomoč, saj je dober model tisti, ki je uporaben v praksi. Tako ima lahko nek model za problem napovedovanja obolelosti za rakom veliko večjo zanesljivost napovedovanja od modela za napovedovanje zavarovalniških goljufij. Vendar je prvi v praski neuporaben, ker ne moremo tvegati, da ne bi poslali pacienta na dodatne preiskave. Drugi pa prinaša zavarovalnici večji zaslužek, ker odkrije vsaj nekaj potencialnih goljufov in je zato v praksi še kako uporaben.

Pri klasifikacijskih problemih želimo vedeti, kako uspešna bo klasifikacija na novih primerih. Pri regresijskih problemih pa nas zanima, kako natančne bodo napovedane vrednosti odvisne spremenljivke, oziroma s kakšnim zaupanjem lahko verjamemo vrednostim, ki jih vrne avtomatsko zgrajeni model.

Pri ocenjevanju uspešnosti avtomatsko zgrajenega modela praviloma ločimo podatke v dve množici. Prvo množico imenujemo učna množica. Učno množico podatkov uporablja

algoritem pri učenju. Drugo množico pa imenujemo testna množica, ki se uporablja za testiranje avtomatsko zgrajenega modela. S takšnim postopkom ocenjevanja uspešnosti se izognemo problemu, da se bi naučeni model preveč prilegal testnim primerom, in bi tako dobili napačno oceno uspešnosti napovedovanja. Vendar pa moramo paziti, da sta tako učna kot tudi testna množica reprezentativni podmnožici danega problema.

Za ocenjevanje uspešnosti klasifikacije lahko uporabljamo naslednje mere:
- klasifikacijska točnost
- tabela napačnih klasifikacij
- cena napačne klasifikacije
- krivulja ROC

Za analizo uspešnosti regresije pa se uporabljajo:
- srednja kvadratna napaka
- relativna kvadratna napaka
- srednja absolutna napaka
- relativna absolutna napaka

4.6.1 Klasifikacijska točnost

Rešitev vsakega primera klasifikacijskega problema je enolično določen razred iz množice možnih razredov. Tako lahko uspešnost reševanja klasifikacijskih problemov definiramo kot razmerje med pravilnimi rešitvami primerov problema na danem področju in vsemi možnimi primeri. To zapišemo kot:

$$CA = \frac{N_p}{N} * 100\% ,$$

kjer je N_p število pravilnih rešitev in N število vseh možnih primerov problema na danem področju. Na koncu razmerje še pomnožimo s 100%, da dobimo odstotek pravilnih rešitev. Klasifikacijsko točnost si lahko razlagamo tudi kot verjetnost, da bo naključno izbrani primer pravilno klasificiran.

V realnih problemih moramo biti pazljivi pri računanju klasifikacijske natančnosti, saj lahko hitro naredimo napako in tako dobimo napačno oceno napovedovanja. Paziti moramo predvsem na to, da testna in učna množica ne vsebujeta istih primerov, saj ni težko sestaviti modela, ki bo na učni množici dosegel 100% klasifikacijsko natančnost.

4.6.2 Tabela napačnih klasifikacij

Tabelo napačnih klasifikacij uporabljamo takrat, ko želimo vedeti kako dobro so klasificirani primeri iz posameznih razredov, česar nam pa klasifikacijska točnost ne pove, saj je povprečena preko vseh razredov. V praksi nas zanima uspešnost za vsak razred posebej.

Pravi razred	Klasificiran kot R1	R2	R3	Vsota
R1	58,7	5,5	2,1	66,3
R2	0,0	2,0	8,5	10,5
R3	2,4	8,5	12,3	23,2
vsota	61,1	16,0	22,9	100,0

Tabela 4.1: Primer tabele odstotkov napačnih klasifikacij za tri razrede

V tabeli 4.1 je podan problem za trirazredni klasifikacijski problem. Na diagonali (notranje 3×3 tabele) so podani odstotki pravilnih klasifikacij. Ostalo so odstotki nepravilnih klasifikacij. Vsota odstotkov vsake vrstice predstavlja apriorno verjetnost ustreznega razreda, vsota stolpcev pa odstotek primerov, klasificiranih v posamezen razred.

4.6.3 Cena napačne klasifikacije

V problemih, ko je napačna klasifikacija primerov iz nekaterih razredov hujša napaka kot napačna klasifikacija primerov iz drugih razredov, je smiselno uporabiti kot mero za uspešnost klasifikatorja ceno napačne klasifikacije. Če opredelimo bolnega človeka kot zdravega, je to ponavadi hujša napaka kot obratno.

Cena napačne klasifikacije je definirana kot:

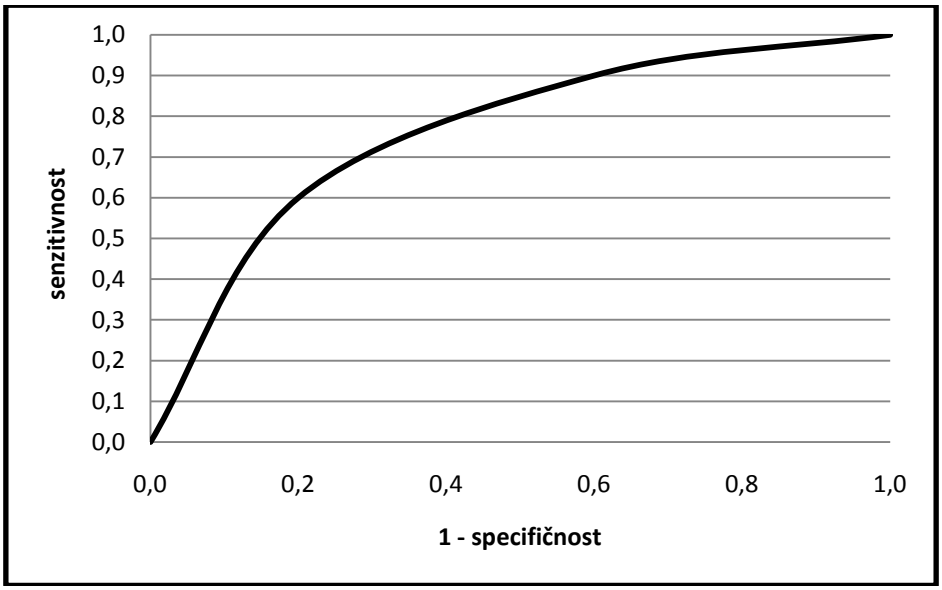
$$C = \frac{\sum_{i,j}(C_{ij} * N_{ij})}{N},$$

kjer je C_{ij} cena za napačno klasifikacijo primera, ki pripada i-temu razredu in je klasificiran v j-ti razred. N_{ij} je število testnih primerov, ki pripadajo i-temu razredu in so klasificirani v j-ti razred. N je število vseh primerov.

Cene za napačno klasifikacijo so navadno podane v kvadratni matriki. Na diagonali so vrednosti 0, saj so to klasifikacije v pravilni razred. Matrika ni nujno simetrična, saj je lahko cena klasifikacije primera iz i-tega razreda v j-ti razred različna od cene klasifikacije primera iz j-tega razreda v i-ti razred.

4.6.4 Krivulja ROC

Krivulja ROC (angl. Reciever Operating Characteristic) nam omogoča analizo razmerja med senzitivnostjo in specifičnostjo. Senzitivnost je definirana kot odstotek pravilno klasificiranih pozitivnih primerov, medtem ko je specifičnost odstotek pravilno klasificiranih negativnih primerov.



Slika 4.1: Primer krivulje ROC

Kot je razvidno iz slike 4.1 je na vodoravni osi prikazano relativno število napačno klasificiranih negativnih primerov (1 - specifičnost). Na navpični osi pa je prikazano relativno število pravilno klasificiranih pozitivnih primerov (senzitivnost).

Kvaliteto posameznega klasifikatorja določimo tako, da izračunamo ploščino pod krivuljo ROC (angl. Area Under the ROC Curve, AUC). Boljši klasifikatorji imajo večjo ploščino AUC. AUC nam pove verjetnost, da bo klasifikator pravilno razločil med pozitivnim in negativnim primerom.

4.6.5 Srednja kvadratna napaka in relativna srednja kvadratna napaka

Srednja kvadratna napaka (angl. mean squared error (MSE)) je mera uspešnosti avtomatsko zgrajenih zveznih modelov. Definirana je kot povprečni kvadrat razlike med napovedano in pravo vrednostjo:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f(i) - \hat{f}(i))^2 ,$$

kjer je N število vseh primerov danega problema, $\hat{f}(i)$ napovedana vrednost primera i in $f(i)$ prava vrednost primera i.

Ker je velikost MSE odvisna od dejanskega razpona možnih vrednosti funkcije, je smiselno uporabiti relativno srednjo kvadratno napako (RSE):

$$RSE = \frac{N * MSE}{\sum_{i=1}^{N}(f(i) - \bar{f})^2},$$

kjer je N število vseh primerov danega problema, MSE srednja kvadratna napaka, $\bar{f}$ povprečna vrednost pravih vrednosti in $f(i)$ prava vrednost primera i.

Relativna srednja kvadratna napaka je nenegativna in manjša od 1 za sprejemljive modele.

Pogosto se v praksi uporabljata tudi meri koren srednje kvadratne napake in koren relativne srednje kvadratne napake.

4.6.6 Srednja absolutna napaka in relativna srednja absolutna napaka

Srednja absolutna napaka (angl. mean absolute error (MAE) je mera uspešnosti avtomatsko zgrajenih zveznih modelov. Definirana je kot povprečna absolutna razlika med napovedano in pravo vrednostjo:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left|f(i) - \hat{f}(i)\right|,$$

kjer je N število vseh primerov danega problema, $\hat{f}(i)$ napovedana vrednost primera i in $f(i)$ prava vrednost primera i.

Ker je tudi velikost MAE odvisna od dejanskega razpona možnih vrednosti funkcije, je smiselno uporabiti relativno srednjo absolutno napako (RAE):

$$RAE = \frac{N * MAE}{\sum_{i=1}^{N}\left|f(i) - \bar{f}\right|},$$

kjer je N število vseh primerov danega problema, MAE srednja absolutna napaka, $\bar{f}$ povprečna vrednost pravih vrednosti in $f(i)$ prava vrednost primera i.

Tudi relativna srednja absolutna napaka je nenegativna in manjša od 1 za sprejemljive modele.

4.7. Ocenjevanje učenja

Kadar imamo dovolj veliko množico primerov z znano vrednostjo odvisne spremenljivke, lahko to množico razdelimo na testno in učno množico. Učno množico uporabimo za učenje algoritma in naučen model preverimo na testni množici. Za ocenjevanje kvalitete uporabimo eno izmed mer, opisanih v prejšnjem razdelku.

Če pa imamo na voljo malo vhodnih podatkov, si ne moremo privoščiti, da bi postopek učenja prikrajšali za primere testne množice. Vendar moramo tudi v tem primeru oceniti uspešnost avtomatsko zgrajenega modela, vendar tega ne smemo narediti na učni množici. Razlogi za to so bili podani v prejšnjem razdelku.

Najzanesljivejša metoda za ocenjevanje uspešnosti modela z malo primeri danega problema je *izloči enega* (angl. leave-one-out). Po tej metodi iz celotne množice izločimo en primer in ga uporabimo za testiranje uspešnosti modela, ki ga zgradimo iz preostalih primerov. Postopek ponovimo za vse učne primere. Tako zgradimo N modelov, ki jih testiramo na enem učnem primeru. Uspešnost modela dobimo tako, da izračunamo povprečje uspešnosti vseh N

modelov na izločenem primeru. Zavedati se moramo, da smo pri tem uporabili N različnih modelov, ki so si med seboj zelo podobni, saj so bili zgrajeni na zelo podobni učni množici in imajo zato zelo podobno uspešnost.

Metoda izloči enega je lahko velikokrat časovno nesprejemljiva, saj moramo zgraditi N modelov, če je N število primerov, namesto enega samega, ko imamo dovolj veliko množico primerov. Metodo izloči enega lahko posplošimo na *izloči N/K primerov*, ki ji pravimo *K-kratno prečno preverjanje* (angl. K-fold cross validation). K je število modelov, ki jih moramo zgraditi. Najprej množico razpoložljivih primerov razdelimo na K približno enako močnih podmnožic. Za vsako podmnožico zgradimo model na uniji preostalih podmnožic. Nato dobljeni model uporabimo za reševanje primerov na dani podmnožici. Uspešnost končnega modela, ki ga zgradimo iz vseh razpoložljivih primerov, ocenimo kot povprečno uspešnost vsek K zgrajenih modelov na celotni množici testnih primerov.

Zanesljivejša različica metode je *sorazmerno prečno preverjanje* (angl. stratified cross-validation). S to metodo ohranjamo približno enako distribucijo razredov v vseh podmnožicah.

4.8. Pristop z ovijanjem

Pristop z ovijanjem je metoda za izbor podmnožice atributov, na katerih naj bi klasifikacijski algoritem dosegel optimalno uspešnost. Večina metod za izbor podmnožice atributov deluje neodvisno od samega algoritma učenja in se zanaša na informacijo, pridobljeno samo iz atributov v povezavi z razredom. Pristop z ovijanjem pa deluje tako, da preiskuje prostor atributov z enim izmed preiskovalnih algoritmov in v vsaki iteraciji doda ali odstrani en ali več atributov. Vsaka iteracija vključuje tudi testiranje učnega algoritma na izbranih atributih in izračun njegove uspešnosti z eno izmed mer za ocenjevanje uspešnosti učenja.

Pri izbiri preiskovalnega algoritma moramo paziti, da izberemo takšen algoritem, da bo njegova časovna zahtevnost ustrezala našim željam. Velikost prostora stanj za n atributov je $O(2^n)$, zato ni smiselno, da izčrpno preiskujemo celoten prostor. Izberemo algoritem, ki bo to nalogo opravil učinkoviteje (npr. najprej najboljši). Pogosto se uporablja tudi požrešni algoritem.

4.9. Uporaba klasifikatorjev v regresiji

Včasih se izkaže, da je regresijski problem boljše reševati s klasifikacijskimi algoritmi. Najprej moramo zvezni razred diskretizirati, zato da lahko sploh uporabimo klasifikacijske algoritme. Razred diskretiziramo z eno od metod diskretizacije, ki najbolj ustreza izbranemu problemu. Nato poženemo izbran klasifikacijski algoritem, s katerim rešimo tako dobljeni klasifikacijski problem. Sedaj je potrebno še odgovore klasifikatorja preslikati nazaj v zvezni razred. To naredimo tako, da izračunamo uteženo vsoto srednjih vrednosti intervalov (diskretnih razredov), tako da so uteži enake verjetnostim, ki jih je klasifikator priredil posameznim intervalom (razredom).

Tak postopek diskretizacije je lahko koristen, saj izniči vpliv šuma odvisne spremenljivke v učnih primerih, kar lahko omogoči klasifikatorju, da doseže boljše napovedi, kot bi jih dosegli z regresijskim algoritmom.

5. Metode dela

Vzorec merjencev so predstavljali slovenski teniški igralci, ki so bili v posameznih obdobjih uvrščeni na jakostno lestvico Teniške zveze Slovenije in so v teh posameznih obdobjih opravili morfološke in motorične meritve. Skupno je vzorec merjencev predstavljalo 593 moških igralcev in 409 ženskih igralk.

Vzorec merjencev je bil razdeljen na skupine, ki so zanimive za analizo napovedovanja tekmovalne uspešnosti:

- Vzorec merjencev za napovedovanje tekmovalne uspešnosti v obdobju do 12 let na podlagi meritev, opravljenih v obdobju do 12 let. Vzorec je vseboval 170 moških igralcev in 157 ženskih igralk.

- Vzorec merjencev za napovedovanje tekmovalne uspešnosti v obdobju od 12 do16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let. Vzorec je vseboval 341 moških igralcev in 215 ženskih igralk.

- Vzorec merjencev za napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju nad 16 let. Vzorec je vseboval 82 moških igralcev in 37 ženskih igralk.

- Vzorec merjencev za napovedovanje tekmovalne uspešnosti v obdobju od 12 do 16 let na podlagi meritev, opravljenih v obdobju do 12 let. Vzorec je vseboval 89 moških igralcev in 84 ženskih igralk.

- Vzorec merjencev za napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju do 12 let. Vzorec je vseboval 47 moških igralcev in 35 ženskih igralk.

- Vzorec merjencev za napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let. Vzorec je vseboval 125 moških igralcev in 79 ženskih igralk.

Vsak od naštetih vzorcev je bil naprej razdeljen še glede na spol.

Vzorec trenerjev je sestavljalo 5 slovenskih trenerjev. Trenerji so bili stari od 24 do 38 let, s trenerskim stažem od 2 do 18 let. Prav tako so bili vsi trenerji nekdanji tekmovalci.

Ker smo imeli opravka z realnimi podatki, se je po pričakovanjih v podatkih pojavilo več primerov z manjkajočimi vrednostmi. Zaradi želje po čim boljši oceni napovedovanja tekmovalne uspešnosti so bili primeri z več kot tremi manjkajočimi vrednostmi odstranjeni.

V realnih podatkih je zmeraj prisoten tudi šum, ki je lahko posledica napačne meritve ali napačnega vnosa podatka. Obravnavanje šuma je bilo prepuščeno kar posameznim algoritmom strojnega učenja, ki so sposobni obravnavati šum.

Kriterij tekmovalne uspešnosti je bil uvrstitev na jakostni lestvici Teniške zveze Slovenije za posamezno leto. Jakostna lestvica upošteva dosežene rezultate v zadnjem tekmovalnem letu.

Mesto na lestvici se določi na osnovi koeficienta, ki predstavlja skupno število točk, deljeno s številom odigranih turnirjev. Na osnovi koeficienta, ki pomeni razmerje med skupnim številom osvojenih točk in številom odigranih tekmovanj, so igralci razvrščeni na jakostni lestvici Teniške zveze Slovenije.

Podatki meritev morfoloških in motoričnih testov ter jakostne lestvice za posamezna leta so bile pridobljene s strani Teniške zveze Slovenije. Slednja poleg vodenja jakostnih lestvic enkrat letno izvede tudi celostna testiranja motoričnih, funkcionalnih in morfoloških sposobnosti/lastnosti v sodelovanju z Inštitutom za šport na Fakulteti za šport Univerze v Ljubljani.

Za analizo podatkov smo uporabili naslednja orodja:

- odprtokodni programski paket za strojno učenje Orange
- programski jezik Python
- programski paket za statistično analizo SPSS
- Microsoft Excel

Najprej je bilo potrebno vzorec teniških igralcev razdeliti v skupine, opisane v poglavju 5.2. Pri analizi napovedovanja tekmovalne uspešnosti s klasifikacijskimi algoritmi je bilo potrebno razred diskretizirati, ker klasifikacijski algoritmi ne delujejo na zveznih razredih. Razred je bil predstavljen z uvrstitvijo na jakostni lestvici Teniške zveze Slovenije. Razred je bil razdeljen na dva dela. Na igralce, ki so uvrščeni v prvo deseterico in ostale. Razlog za to je, da želimo ločiti najboljše igralce od ostalih, saj lahko le najboljši uspejo tudi na mednarodni ravni. Klasifikacija je bila izvedena z naslednjimi metodami: naivni Bayes, odločitveno drevo, algoritem C4.5, k-najbližjih sosedov, SVM in logistično regresijo. Ocenjevanje uspešnosti klasifikatorjev je bilo narejeno s klasifikacijsko točnostjo in površino pod krivuljo ROC ter uporabo metode k-kratnega prečnega preverjanja, pri čemer je bil $k = 10$.

Regresijska analiza deluje z zveznim razredom, zato ni bilo potrebne dodatne predobdelave podatkov. Tudi tukaj je bil razred predstavljen z uvrstitvijo na jakostni lestvici Teniške zveze Slovenije. Regresijska analiza je bila izvedena z naslednjima metodama: linearna regresija in regresijska drevesa. Ocenjevanje uspešnosti je bilo narejeno s korenom srednje kvadratne napake, srednjo absolutno napako in relativno srednjo absolutno napako.

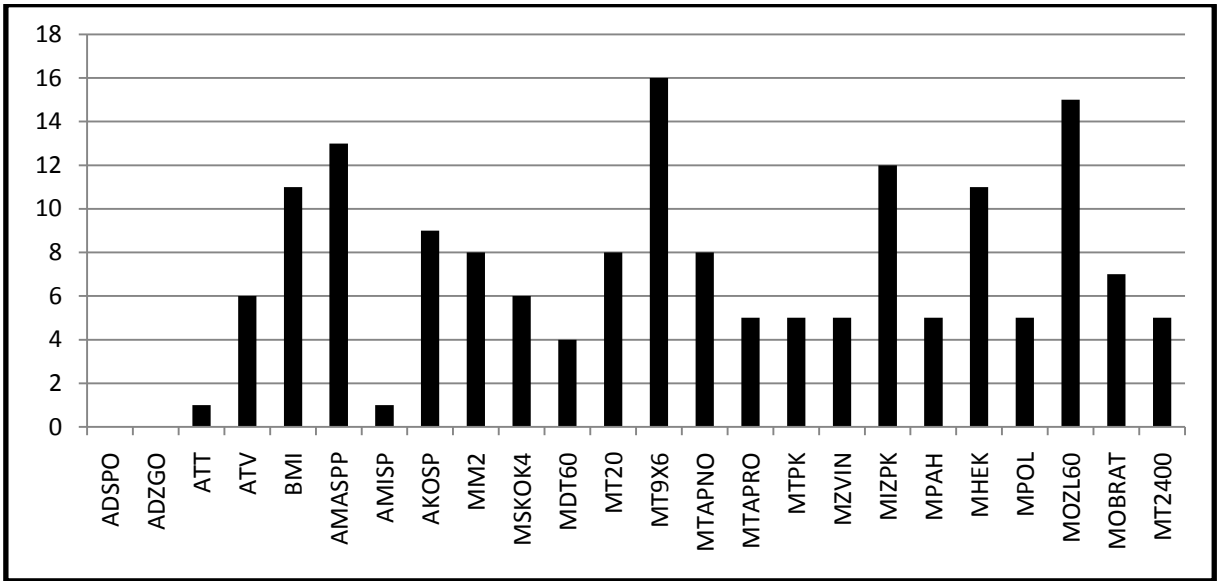
Test napovedovanja tekmovalne uspešnosti je bil izveden s klasifikacijskimi in regresijskimi metodami za vse skupine, opisane v poglavju 5.2., pri tem pa so bili uporabljeni vsi morfološki in motorični atributi. Pri tem testu je bil pri metodah naivni Bayes, odločitveno drevo, k-najbližjih sosedov, SVM in logistična regresija uporabljen tudi pristop za izbiro podmnožice atributov, imenovan pristop z ovijanjem. Nato je bilo z metodo ReliefF izbranih pet najobetavnejših morfoloških in motoričnih atributov za vse skupine. Metoda ReliefF je bila izbrana zato, ker je veliko atributov med seboj odvisnih. Metoda ReliefF deluje dobro tudi na odvisnih atributih, za razliko od ostalih metod za ocenjevanje pomembnosti atributov (npr. informacijski prispevek). Spet je bil opravljen test napovedovanja tekmovalne uspešnosti s klasifikacijskimi metodami za vse skupine. Nazadnje je bil narejen test napovedovanja tekmovalne uspešnosti na pet najpogosteje izbranih morfoloških in motoričnih atributih s strani trenerjev s klasifikacijskimi in regresijskimi metodami za vse skupine.

6. Rezultati in interpretacija

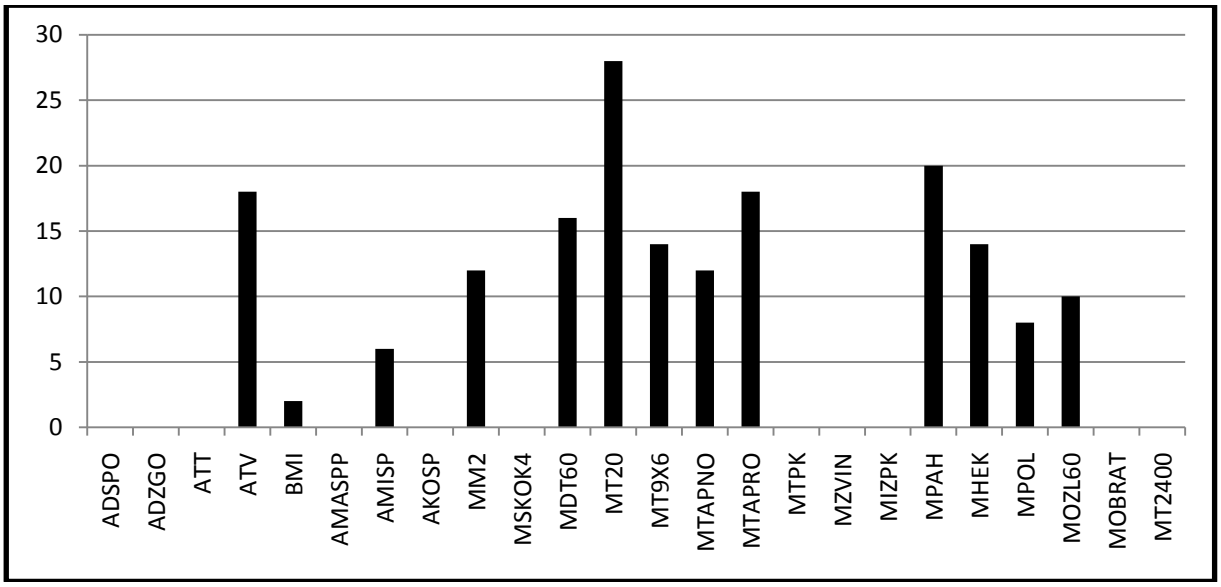
V tem poglavju so predstavljeni rezultati, ki smo jih dobili pri raziskavi. Poleg rezultatov je vključena tudi interpretacija rezultatov, ki bralcu pomaga razumeti vsebino in uporabnost rezultatov ter podaja odgovore na zastavljene cilje.

6.1. Pogostost najobetavnejših atributov

Na napovedovanje tekmovalne uspešnosti teniškega igralca vplivajo morfološki in motorični atributi. Vpliv posameznega atributa je lahko večji ali manjši, zato smo se odločili, da povprašamo slovenske teniške trenerje za njihovo mnenje, kateri atributi najbolj vplivajo na napovedovanje tekmovalne uspešnosti. Pogostost izbere posameznega atributa je predstavljena v grafu 6.2. Pogostost izbere atributov s strani trenerjev smo nato primerjali s pogostostjo izbere atributov z algoritmov ReliefF, ki je predstavljena v grafu 6.1.



Graf 6.1: Pogostost izbire najobetavnejših morfoloških in motoričnih atributov za vse skupine z algoritmom ReliefF.



Graf 6.1: Pogostost izbire najobetavnejših morfoloških in motoričnih atributov za vse skupine s strani trenerjev.

Pri izbiri najobetavnejših atributov z metodo ReliefF izstopata tek 9 x 6 metrov (M9x6) in odbijanje žogice z loparjem 60 sekund (MOZL60). Izbira teh dveh atributov je smiselna, saj so kratki teki in teniška koordinacija pri tenisu pomembna dejavnika. Gledano s strani trenerjev pa sta najpomembnejša atributa tek na 20 metrov (MT20) in hitrost kratkih tekov s spremembami (MPAH). Tudi izbira teh dveh atributov je smiselna, saj sta prav tako pomembna dejavnika pri teniški igri. Na tem mestu velja omeniti, da so tek na 9 x 6 metrov, tek na 20 metrov in hitrost kratkih tekov s spremembami med seboj zelo korelirani, tako da je med njimi zelo težko izbrati atribut, ki najbolj vpliva na tekmovalno uspešnost. Vsi do sedaj omenjeni atributi so motorični dejavniki. Od morfoloških dejavnikov je po izboru trenerjev najobetavnejši telesna višina ATV, medtem ko po izbiri metode ReliefF največ obetata odstotek maščobne mase (AMASPP) in indeks telesne teže (BMI).

6.2. Rezultati in interpretacija napovedovanja tekmovalne uspešnosti s klasifikacijskimi metodami

V tem poglavju so predstavljeni rezultati in interpretacija rezultatov napovedovanja tekmovalne uspešnosti s klasifikacijskimi metodami. Najprej so v tabelah 6.2, 6.3, 6.4, 6.5, 6.6 in 6.7 podani rezultati napovedovanja tekmovalne uspešnosti z uporabo vseh morfoloških in motoričnih atributov, nato so v tabelah 6.8, 6.9, 6.10, 6.11, 6.12 in 6.13 podani rezultati napovedovanja tekmovalne uspešnosti z uporabo petih najobetavnejših morfoloških in motoričnih atributov izbranih z metodo ReliefF, na koncu so v tabelah 6.14, 6.15, 6.16, 6.17, 6.18 in 6.19 podani rezultati napovedovanja tekmovalne uspešnosti z uporabo petih najobetavnejših morfoloških in motoričnih atributov izbranih s strani trenerjev.

Zaradi preglednosti in poenostavljenega prikaza rezultatov v naslednjih podpoglavjih, smo uporabili oznake, katerih pomen je opisan v tabeli 6.1.

Oznaka	Pomen
Naivni Bayes	Metoda naivni Bayes
Naivni Bayes WA	Metoda naivni Bayes s pristopom ovijanja
Odločitveno drevo	Metoda odločitveno drevo
C4.5	Metoda C4.5
k-najbližjih sosedov	Metoda k-najbližjih sosedov
k-najbližjih sosedov WA	Metoda k-najbližjih sosedov s pristopom ovijanja
SVM	Metoda podpornih vektorjev
SVM WA	Metoda podpornih vektorjev s pristopom ovijanja
Logistična regresija	Metoda logistična regresija
Logistična regresija WA	Metoda logistična regresija s pristopom ovijanja
Regresijsko drevo	Metoda regresijsko drevo
Linearna regresija	Metoda linearna regresija
CA	Klasifikacijska točnost
AUC	Površina pod krivuljo ROC
RMSE	Koren srednje kvadratne napake
MAE	Srednja absolutna napaka
RAE	Relativna srednja absolutna napaka

Tabela 6.1: Pomen oznak v poglavju 6.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.665	0.728	0.656	0.728
Naivni Bayes WA	0.624	0.666	0.695	0.778
Odločitveno drevo	0.588	0.535	0.620	0.603
C4.5	0.576	0.592	0.617	0.613
k-najbližjih sosedov	0.641	0.635	0.581	0.603
k-najbližjih sosedov WA	0.529	0.539	0.625	0.667
SVM	0.571	0.456	0.638	0.694
SVM WA	0.653	0.611	0.611	0.636
Logistična regresija	0.600	0.627	0.555	0.585
Logistična regresija WA	0.688	0.773	0.668	0.720

Tabela 6.2: Napovedovanje tekmovalne uspešnosti v obdobju do 12 let na podlagi meritev, opravljenih v obdobju do 12 let.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.674	0.730	0.623	0.654
Naivni Bayes WA	0.680	0.737	0.606	0.678
Odločitveno drevo	0.648	0.627	0.559	0.545
C4.5	0.601	0.623	0.576	0.582
k-najbližjih sosedov	0.545	0.522	0.609	0.592
k-najbližjih sosedov WA	0.619	0.653	0.552	0.561
SVM	0.610	0.650	0.549	0.549
SVM WA	0.552	0.465	0.540	0.567
Logistična regresija	0.578	0.600	0.596	0.632
Logistična regresija WA	0.671	0.735	0.616	0.669

Tabela 6.3: Napovedovanje tekmovalne uspešnosti v obdobju od 12 do16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.475	0.453	0.567	0.700
Naivni Bayes WA	0.560	0.547	0.500	0.625
Odločitveno drevo	0.453	0.445	0.633	0.662
C4.5	0.537	0.490	0.450	0.537
k-najbližjih sosedov	0.535	0.529	0.500	0.525
k-najbližjih sosedov WA	0.451	0.343	0.542	0.475
SVM	0.536	0.541	0.542	0.600
SVM WA	0.574	0.484	0.500	0.425
Logistična regresija	0.414	0.351	0.557	0.479
Logistična regresija WA	0.429	0.359	0.608	0.638

Tabela 6.4: Napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju nad 16 let.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.554	0.496	0.544	0.608
Naivni Bayes WA	0.531	0.670	0.522	0.533
Odločitveno drevo	0.461	0.448	0.469	0.472
C4.5	0.478	0.511	0.547	0.507
k-najbližjih sosedov	0.557	0.603	0.574	0.551
k-najbližjih sosedov WA	0.614	0.682	0.436	0.503
SVM	0.411	0.398	0.522	0.561
SVM WA	0.478	0.421	0.533	0.551
Logistična regresija WA	0.508	0.521	0.613	0.630

Tabela 6.5: Napovedovanje tekmovalne uspešnosti v obdobju od 12 do 16 let na podlagi meritev, opravljenih v obdobju do 12 let.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.690	0.725	0.433	0.400
Naivni Bayes WA	0.645	0.700	0.450	0.525
Odločitveno drevo	0.485	0.483	0.542	0.500
C4.5	0.540	0.542	0.375	0.313
k-najbližjih sosedov	0.430	0.442	0.442	0.450
k-najbližjih sosedov WA	0.450	0.542	0.500	0.500
SVM	0.420	0.408	0.383	0.375
SVM WA	0.595	0.725	0.517	0.500
Logistična regresija WA	0.485	0.450	0.550	0.650

Tabela 6.6: Napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju do 12 let.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.608	0.644	0.605	0.613
Naivni Bayes WA	0.578	0.658	0.495	0.489
Odločitveno drevo	0.535	0.542	0.520	0.540
C4.5	0.591	0.619	0.455	0.402
k-najbližjih sosedov	0.531	0.543	0.493	0.489
k-najbližjih sosedov WA	0.549	0.585	0.518	0.496
SVM	0.377	0.353	0.534	0.571
SVM WA	0.576	0.573	0.430	0.425
Logistična regresija WA	0.576	0.619	0.570	0.625

Tabela 6.7: Napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.665	0.726	0.725	0.788
Odločitveno drevo	0.659	0.591	0.630	0.598
C4.5	0.653	0.641	0.611	0.628
k-najbližjih sosedov	0.618	0.694	0.561	0.586
SVM	0.600	0.386	0.586	0.543
Logistična regresija	0.671	0.715	0.650	0.713

Tabela 6.8: Napovedovanje tekmovalne uspešnosti v obdobju do 12 let na podlagi meritev, opravljenih v obdobju do 12 let. Pri moških so bili izbrani atributi MTAPR, MM2, MOZL60, ATV in MIZPK, pri ženskah pa MOZL60, MTAPNO, MT2400, SKOK4 in MHEK.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.657	0.671	0.670	0.752
Odločitveno drevo	0.528	0.501	0.655	0.621
C4.5	0.598	0.589	0.655	0.632
k-najbližjih sosedov	0.551	0.573	0.629	0.697
SVM	0.595	0.522	0.577	0.583
Logistična regresija	0.660	0.677	0.713	0.754

Tabela 6.9: Napovedovanje tekmovalne uspešnosti v obdobju do 12 do 16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let. Pri moških so bili izbrani atributi MTPK, MIZPK, MT9X6, MOBRAT in MPAH, pri ženskah pa MHEK, MIZPK, MZOL60, MM2 in MT9X6.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.524	0.475	0.725	0.775
Odločitveno drevo	0.450	0.435	0.567	0.538
C4.5	0.574	0.500	0.567	0.588
k-najbližjih sosedov	0.524	0.511	0.633	0.575
SVM	0.535	0.500	0.483	0.625
Logistična regresija	0.482	0.495	0.600	0.725

Tabela 6.10: Napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju nad 16 let. Pri moških so bili izbrani atributi AMASPP, ATV, MHEK, MT2400 in MM2, pri ženskah pa MABAL, MT20, MSKOK4, MOBRAT in MHEK.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.589	0.640	0.629	0.664
Odločitveno drevo	0.625	0.646	0.617	0.627
C4.5	0.581	0.624	0.522	0.568
k-najbližjih sosedov	0.596	0.642	0.597	0.661
SVM	0.514	0.424	0.521	0.437
Logistična regresija	0.569	0.616	0.664	0.690

Tabela 6.11: Napovedovanje tekmovalne uspešnosti v obdobju od 12 do 16 let na podlagi meritev, opravljenih v obdobju do 12 let. Pri moških so bili izbrani atributi MT9X6, MPAH, MPOL, BMI in MT20, pri ženskah pa AMASPP, MTAPNO, MOZL60, MDT60 in MZVIN.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.665	0.750	0.542	0.650
Odločitveno drevo	0.620	0.642	0.483	0.475
C4.5	0.455	0.475	0.525	0.538
k-najbližjih sosedov	0.540	0.517	0.533	0.550
SVM	0.490	0.425	0.433	0.525
Logistična regresija	0.520	0.475	0.567	0.550

Tabela 6.12: Napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju do 12 let. Pri moških so bili izbrani atributi MT9X6, MABAL, MT20, MPOL in MSKOK4, pri ženskah pa BMI, AKOST, MOBRAT, MDT60 in ATT.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.575	0.584	0.580	0.637
Odločitveno drevo	0.579	0.602	0.455	0.445
C4.5	0.480	0.492	0.480	0.487
k-najbližjih sosedov	0.544	0.587	0.480	0.548
SVM	0.478	0.457	0.495	0.488
Logistična regresija	0.543	0.596	0.570	0.575

Tabela 6.13: Napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let. Pri moških so bili izbrani atributi MHEK, MZVIN, MIZPK, MPOL in MOZL60, pri ženskah pa AKOSP, BMI, AMASPP, MT9X6 in AMISP.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.653	0.729	0.593	0.639
Odločitveno drevo	0.547	0.520	0.458	0.442
C4.5	0.594	0.603	0.503	0.523
k-najbližjih sosedov	0.594	0.563	0.492	0.493
SVM	0.600	0.577	0.605	0.543
Logistična regresija	0.700	0.773	0.617	0.628

Tabela 6.14: Napovedovanje tekmovalne uspešnosti v obdobju do 12 let na podlagi meritev, opravljenih v obdobju do 12 let. Izbrani atributi so bili ATV, MPOL, MDT60, MPAH in BMI.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.642	0.685	0.549	0.573
Odločitveno drevo	0.607	0.555	0.494	0.523
C4.5	0.569	0.555	0.470	0.429
k-najbližjih sosedov	0.593	0.614	0.437	0.414
SVM	0.601	0.536	0.483	0.477
Logistična regresija	0.616	0.650	0.545	0.563

Tabela 6.15: Napovedovanje tekmovalne uspešnosti v obdobju od 12 do16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let. Izbrani atributi so bili MT20, ATV, MSARG, MPAH in MT9X6.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.642	0.685	0.549	0.573
Odločitveno drevo	0.607	0.555	0.494	0.523
C4.5	0.569	0.555	0.470	0.429
k-najbližjih sosedov	0.593	0.614	0.437	0.414
SVM	0.601	0.536	0.483	0.477
Logistična regresija	0.616	0.650	0.545	0.563

Tabela 6.16: Napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju nad 16 let. Izbrani atributi so bili MPAH, MSARG, MT9X6, MHEK in AMISP.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.500	0.486	0.593	0.660
Odločitveno drevo	0.536	0.531	0.588	0.602
C4.5	0.443	0.452	0.536	0.547
k-najbližjih sosedov	0.479	0.500	0.561	0.591
SVM	0.429	0.395	0.512	0.477
Logistična regresija	0.546	0.540	0.557	0.646

Tabela 6.17: Napovedovanje tekmovalne uspešnosti v obdobju od 12 do 16 let na podlagi meritev, opravljenih v obdobju do 12 let. Izbrani atributi so bili MTAPN, MTAPR, MHOJA, MPAH in MOZL60.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.500	0.486	0.575	0.350
Odločitveno drevo	0.536	0.531	0.492	0.525
C4.5	0.443	0.452	0.458	0.388
k-najbližjih sosedov	0.479	0.500	0.533	0.475
SVM	0.429	0.395	0.533	0.450
Logistična regresija	0.546	0.540	0.550	0.325

Tabela 6.18: Napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju do 12 let. Izbrani atributi so bili MT20, MTAPR, MHST, MOZL in MOBR.

	Moški		Ženske	
Metoda	**CA**	**AUC**	**CA**	**AUC**
Naivni Bayes	0.606	0.583	0.443	0.585
Odločitveno drevo	0.560	0.547	0.441	0.443
C4.5	0.545	0.585	0.495	0.502
k-najbližjih sosedov	0.531	0.542	0.566	0.608
SVM	0.469	0.445	0.593	0.521
Logistična regresija	0.505	0.521	0.504	0.590

Tabela 6.19: Napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let. Izbrani atributi so bili MM2, MT20, MDT60, MHEK in MT9X6.

Metodi naivni Bayes in logistična regresija s pristopom ovijanja dobro napovedujeta tekmovalno uspešnost v obdobju do 12 let na podlagi meritev, opravljenih v obdobju do 12 let, in v obdobju od 12 do16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let, kar je razvidno iz tabel 6.2 in 6.3. Iz rezultatov v tabeli 6.4 sledi, da v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju nad 16 let, odpovejo vse metode za napovedovanje tekmovalne uspešnosti. Ti rezultati so pričakovani, saj na tekmovalno uspešnost v obdobju do 16 let vplivajo morfološki in motorični dejavniki v precejšnji meri, medtem ko v obdobju po 16 letu prideta do vse večjega izraza tehnično in taktično znanje igralca. Najpogosteje izbrani atributi pri metodi logistična regresija s pristopom ovijanja za napovedovanje tekmovalne uspešnosti v obdobju do 12 let na podlagi meritev, opravljenih v obdobju do 12 let, in v obdobju od 12 do16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let, sta bila odbijanje žogice z loparjem 60 sekund (MOZL60), kratki teki s spremembami smeri (MPAH). To je primerljivo z izbiro najobetavnejših atributov z metodo ReliefF, saj se izbira odbijanje žogice z loparjem 60 sekund ujema, medtem ko atributa kratki teki s spremembami smeri in tek 9 x 6 temeljita na zelo podobnem testu in sta tudi zelo korelirana.

Pri napovedovanju tekmovalne uspešnosti v obdobju od 12 do 16 let na podlagi meritev, opravljenih v obdobju do 12 let, v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju do 12 let, in v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let, izstopa le napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju do 12 let pri moških z metodo naivni Bayes, kar je razvidno iz tabele 6.6, vendar temu le težko pripišemo večjo težo, saj ostale metode popolnoma odpovejo. Iz rezultatov sledi, da je napovedovanje tekmovalne uspešnosti za bodoča obdobja na podlagi morfoloških in motoričnih dejavnikov nezanesljivo.

Z izbiro najobetavnejših atributov z metodo ReliefF praviloma ne pridobimo ničesar v primerjavi z napovedovanjem tekmovalne uspešnosti z uporabo vseh morfoloških in motoričnih atributov. To lahko pomeni, da metoda ReliefF ni primerna za uporabo v tej problematiki ali pa da zaradi velike odvisnosti med atributi ne moremo doseči večje uspešnosti napovedovanja tekmovalne uspešnosti na podlagi morfoloških in motoričnih testov. Znatnejši popravek je opazen le pri napovedovanju tekmovalne uspešnosti z metodo naivni Bayes v obdobju do 12 let na podlagi meritev, opravljenih v obdobju do 12 let (tabela 6.8), v obdobju od 12 do16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let (tabela 6.9), in v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju nad 16 let pri ženskah (tabela 6.10). S primerjavo klasifikacijskih točnosti pri testih z vsemi atributi in testih z izbranimi petimi najobetavnejšimi atributi lahko sklepamo, da je pristop ovijanja v kombinaciji z logistično regresijo dober pristop k učenju za to problematiko.

Pri klasifikaciji na podlagi izbranih petih najobetavnejši atributov s strani trenerjev je opazno poslabšanje uspešnosti napovedovanja tekmovalne uspešnosti. V primerjavi z napovedovanjem tekmovalne uspešnosti na podlagi vseh atributov je primerljivo le v obdobju do 12 let na podlagi meritev, opravljenih v obdobju do 12 let pri ženskah in moških (primerjava tabel 6.2 in 6.14) ter v obdobju od 12 do16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let pri moških (primerjava tabel 6.3 in 6.15). Na podlagi tega lahko sklepamo, da je izbira najobetavnejših morfoloških in motoričnih atributov zelo težek problem, saj je tenis kompleksna igra in na uspešnost igralcev vpliva veliko različnih dejavnikov. Po drugi strani pa lahko sklepamo, da trenerji posvečajo več časa dejavnikom, ki v večji meri vplivajo na tekmovalno uspešnost v obdobju po 16-tem letu in zato ne posvečajo večje pozornosti preučevanju morfoloških in motoričnih dejavnikov ter odvisnostim med njimi in tekmovalno uspešnostjo.

6.3. Rezultati in interpretacija napovedovanja tekmovalne uspešnosti z regresijskimi metodami

V tem poglavju so predstavljeni rezultati in interpretacija rezultatov napovedovanja tekmovalne uspešnosti z regresijskimi metodami. Najprej so v tabelah 6.20, 6.21, 6.22 in 6.23 podani rezultati napovedovanja tekmovalne uspešnosti z uporabo vseh morfoloških in motoričnih atributov, nato pa so v tabelah 6.24, 6.25, 6.26 in 6.27 podani rezultati napovedovanja tekmovalne uspešnosti z uporabo petih najobetavnejših morfoloških in motoričnih atributov izbranih s strani trenerjev.

Testiranje napovedovanja tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju nad 16 let pri moških in ženskah, v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju do 12 let pri moških in ženskah ter v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let pri ženskah, niso bili narejeni zaradi prevelikega števila primerov z manjkajočimi vrednostmi.

	Moški			Ženske		
Metoda	**RMSE**	**MAE**	**RAE**	**RMSE**	**MAE**	**RAE**
Regresijsko drevo	28.690	20.335	1.018	23.475	17.024	0.983
Linearna regresija	14.934	11.314	0.591	15.435	12.23	0.634

Tabela 6.20: Napovedovanje tekmovalne uspešnosti v obdobju do 12 let na podlagi meritev, opravljenih v obdobju do 12 let.

	Moški			Ženske		
Metoda	**RMSE**	**MAE**	**RAE**	**RMSE**	**MAE**	**RAE**
Regresijsko drevo	36.485	25.309	1.162	15.075	10.476	1.060
Linearna regresija	18.875	14.704	0.759	10.639	8.364	0.773

Tabela 6.21: Napovedovanje tekmovalne uspešnosti v obdobju od 12 do16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let.

	Moški			Ženske		
Metoda	**RMSE**	**MAE**	**RAE**	**RMSE**	**MAE**	**RAE**
Regresijsko drevo	24.161	15.677	0.782	27.800	20.898	1.155
Linearna regresija	3.777	2.759	0.136	0.925	0.661	0.032

Tabela 6.22: Napovedovanje tekmovalne uspešnosti v obdobju od 12 do 16 let na podlagi meritev, opravljenih v obdobju do 12 let.

	Moški		
Metoda	**RMSE**	**MAE**	**RAE**
Regresijsko drevo	46.316	38.279	1.075
Linearna regresija	15.906	12.464	0.421

Tabela 6.23: Napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let

	Moški			Ženske		
Metoda	**RMSE**	**MAE**	**RAE**	**RMSE**	**MAE**	**RAE**
Linearna regresija	19.840	15.398	0.805	21.510	17.097	0.887

Tabela 6.24: Napovedovanje tekmovalne uspešnosti v obdobju do 12 let na podlagi meritev, opravljenih v obdobju do 12 let. Izbrani atributi so bili ATV, MPOL, MDT60, MPAH in BMI.

	Moški			Ženske		
Metoda	**RMSE**	**MAE**	**RAE**	**RMSE**	**MAE**	**RAE**
Linearna regresija	23.928	18.787	0.970	14.449	10.666	0.985

Tabela 6.25: Napovedovanje tekmovalne uspešnosti v obdobju od 12 do16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let. Izbrani atributi so bili MT20, ATV, MSARG, MPAH in MT9X6.

	Moški			Ženske		
Metoda	**RMSE**	**MAE**	**RAE**	**RMSE**	**MAE**	**RAE**
Linearna regresija	26.117	20.344	1.005	18.413	15.059	0.740

Tabela 6.26: Napovedovanje tekmovalne uspešnosti v obdobju od 12 do 16 let na podlagi meritev, opravljenih v obdobju do 12 let. Izbrani atributi so bili MTAPN, MTAPR, MHOJA, MPAH in MOZL60.

	Moški		
Metoda	**RMSE**	**MAE**	**RAE**
Linearna regresija	32.955	26.942	0.910

Tabela 6.27: Napovedovanje tekmovalne uspešnosti v obdobju nad 16 let na podlagi meritev, opravljenih v obdobju od 12 do 16 let. Izbrani atributi so bili MM2, MT20, MDT60, MHEK in MT9X6.

Najprej opazimo, da je relativna srednja absolutna ocena napovedovanja regresijskega drevesa pri petih od sedmih testov nad 1 (tabele 6.20, 6.21, 6.22 in 6.23), kar pomeni, da je napovedovanje tekmovalne uspešnosti teniških igralcev v praksi s to metodo popolnoma neuporabno. Zato je bila ta metoda tudi izpuščena pri napovedovanju tekmovalne uspešnosti z uporabo petih najobetavnejših morfoloških in motoričnih atributov, izbranih s strani trenerjev.

Linearna regresija se je izkazala za veliko boljšo metodo, saj je relativna absolutna napaka zmeraj pod 1. V dveh primerih (napovedovanje tekmovalne uspešnosti v obdobju od 12 do 16 let na podlagi meritev, opravljenih v obdobju do 12 let pri moških in ženskah, ki je prikazano v tabeli 6.22) je tudi blizu 0, kar pomeni skoraj idealno regresijsko funkcijo, vendar bi glede na ostale teste to lahko pripisali nereprezentativnemu vzorcu primerov. V ostalih primerih daje linearna regresija zadovoljive rezultate, saj je srednja absolutna napaka približno v območju med 8 in 15. To pomeni, da lahko tekmovalno uspešnost teniškega igralca na podlagi morfoloških in motoričnih dejavnikov napovemo v povprečju na vsaj 15 mest točno.

Na podlagi izbranih petih najobetavnejših atributov s strani trenerjev je opazno poslabšanje napovedovanja tekmovalne uspešnosti z metodo linearne regresije, kar je razvidno iz tabel 6.24, 6.25, 6.26 in 6.27. Spet lahko sklepamo podobno kot pri klasifikaciji na podlagi izbranih petih najobetavnejših atributov s strani trenerjev. Torej, da je izbira najobetavnejših morfoloških in motoričnih atributov zelo težek problem, saj je tenis kompleksna igra in na uspešnost igralcev vpliva veliko različnih dejavnikov.

7. Zaključek

Namen raziskave je bil preveriti zanesljivost napovedovanja tekmovalne uspešnosti slovenskih teniških igralcev v različnih obdobjih tekmovanja z metodami strojnega učenja. Pri tem smo uporabili klasifikacijske in regresijske metode. Za izboljšanje učenja smo uporabili metodo ReliefF in pristop ovijanja. Na koncu smo še preverili uspešnost napovedovanja tekmovalne uspešnosti na podlagi petih najpogosteje izbranih morfoloških in motoričnih dejavnikov s strani trenerjev.

Ugotovili smo, da s klasifikacijskimi metodami lahko ločimo najboljše igralce od ostalih do 16 leta starosti na podlagi morfoloških in motoričnih dejavnikov s točnostjo približno 0,65, medtem ko po 16. letu starosti klasifikacija na podlagi morfoloških in motoričnih dejavnikov odpove. Za nezanesljivo se je pokazala tudi klasifikacija na podlagi meritev, opravljenih v predhodnih starostnih obdobjih. Pokazali smo tudi, da lahko avtomatski algoritmi izberejo boljšo podmnožico atributov kot pa strokovnjaki na področju tenisa. Najuspešnejši metodi za napovedovanje tekmovalne uspešnosti sta bili naivni Bayes in logistična regresija s pristopom ovijanja. Pri regresijskih metodah se je za uspešnejšo izkazala linearna regresija, ki lahko napove tekmovalo uspešnost v povprečju na 8 do 15 mest točno, medtem ko je metoda regresijskega drevesa za reševanje tega problema popolnoma neuporabna.

V prihodnje bi bilo smiselno bolj organizirano opravljanje meritev in zbiranje podatkov, saj se je v nekaterih primerih izkazalo, da je bilo pri veliko meritvah precejšnje število manjkajočih vrednosti. Seveda je ta segment povezan z razpoložljivimi finančnimi sredstvi, vendar se naj bi z razvojem tudi to popravilo. Smiselno bi bilo narediti raziskavo napovedovanja tekmovalne uspešnosti s čim več dejavniki, ki vplivajo na tekmovalno uspešnost teniškega igralca. Tako bi verjetno dobili še nekaj boljše rezultate. Predvsem zanimivo bi bilo videti izboljšanje pri napovedovanju tekmovalne uspešnosti igralca za več let vnaprej.

V tej raziskavi se je izkazalo, da je napovedovanje tekmovalne uspešnosti teniških igralcev zelo kompleksen problem, saj je bila točnost napovedanih modelov, ki so temeljili na morfoloških in motoričnih dejavnikih relativno slaba.

8. Priloge

8.1. Vprašalnik za trenerje

PREDIKCIJA USPEŠNOSTI V TENISU

Izberite 10 testov (iz spodnjega nabora), ki po vašem mnenju najbolj vplivajo na tekmovalno uspešnost igralca za posamezno starostno obdobje. Testi si naj sledijo od najpomembnejšega navzdol.

Šifra	**Ime testa**
ATV	Telesna višina (vzdolžna razsežnost)
ATT	Telesna teža (telesna teža)
BMI	Indeks telesne mase (grob indeks telesa, vezan le na težo in višino)
AMASPP	% telesne maščobe
AMISP	% mišične mase
AKOSP	% kostne mase
MSARG	Sargent skok (hitra moč)
MM2	Met medicinke (2 kg) (hitra moč)
MSKOK	Četveroskok (hitra moč)
MDT60	Dviganje trupa (repetitivna/vzdržljivostna moč)
MT20	Tek na 20 m (hitrost)
MT5	Tek na 5 m (hitrost)
MT9X6	Tek 9 x 6 m (hitrost)
MREAK	Reakcijska palica (hitrost reakcije)
MTAPN	Taping z nogo (hitrost izmeničnih gibov)
MTAPR	Taping z roko (hitrost izmeničnih gibov)
MTPK	Predklon na klopici (gibljivost nog in hrbta)
MZVIN	Zvinek s palico (gibljivost rok in ramen)
MIZPK	Izpadni korak (gibljivost nog in kolkov)
MPAH	Pahljača (agilnost / hitrost kratkih tekov s spremembami smeri)
MHEK	Heksagon (agilnost / hitrost kratkih tekov s spremembami smeri)
MHST	Hitrost stopanja (agilnost / hitrost kratkih tekov s spremembami smeri)
MPOL	Poligon nazaj (koordinacija celega telesa)
MOZL	Odbijanje žoge z loparjem (koordinacija - teniška)
MOSMI	Osmica s pripogibanjem (koordinacija – v povezavi s tekom)
MOBR	Obrati na gredi (ravnotežje)
MHOJA	Hoja po gredi in odbijanje (ravnotežje)
MPRIS	Prisunski koraki po gredi (ravnotežje)
M2400	Tek na 2400 m (aerobna vzdržljivost)

V obdobju do 12 let za obdobje do 12 let:

V obdobju od 12 do 16 let za obdobje od 12 do 16 let:

V obdobju nad 16 let za obdobje nad 16 let:

V obdobju do 12 let za obdobje od 12 do 16 let:

V obdobju do 12 let za obdobje nad 16 let:

V obdobju od 12 do 16 let za obdobje nad 16 let:

PODATKI O IZPOLNJEVALCU:

Priimek in ime:

Starost:

Trenerski staž:

Nekdanji tekmovalec (DA/NE):

Tekmovalski staž:

Klub:

9. Literatura

[1] I. Bratko, *Prolog programming for artificial intelligence*, Harlow: Addison-Wesley, 2001

[2] A. Filipčič, *Evaluacija tekmovalne in potencialne uspešnosti mladih teniških igralcev*, doktorsko delo na Fakulteti za šport UL, Ljubljana, 1996

[3] D. Hand, H. Mannila, P. Smyth, *Priciples of Data Mining*, Cambridge: A Bradford Book The MIT Press, 2001

[4] I. Kononenko, M. Kukar, *Machine Learning and Data Mining*, Chichester: Horwood, Publishing, 2007

[5] I. H. Witten, E. Frank, *Data mining: practical machine learning tools and techniques*, San Francisco: Morgan Kaumann, 2005

[6] (2009) Orange. Dostopno na: http://www.ailab.si/orange/