Prilagodljiva metoda za predviđanje sportskih ishoda zasnovana na indeksu korisnosti i optimalnom vremenskom prozoru

Horvat, Tomislav

SVEUČILIŠTE J. J. STROSSMAYERA U OSIJEKU

Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek

Tomislav Horvat

Prilagodljiva metoda za predviđanje sportskih ishoda zasnovana na indeksu korisnosti i optimalnom vremenskom prozoru

Doktorska disertacija

Osijek, 2020.

Doktorska disertacija je izrađena na Zavodu za programsko inženjerstvo Fakulteta elektrotehnike, računarstva i informacijskih tehnologija Osijek Sveučilišta J. J. Strossmayera u Osijeku

Mentor: izv. prof. dr. sc. Josip Job

Doktorski rad ima 136 stranica.

Doktorski rad br.: 77

Zahvala

All dreams can come true, if we have courage to pursue them.

- Walt Disney

Ovim se putem zahvaljujem mentoru, izv.prof.dr.sc. Josipu Jobu, na podršci, strpljenju, pomoći i prijateljskom vođenju tijekom poslijediplomskog studija.

Zahvaljujem se i kolegama sa Sveučilišta Sjever koji su bili od iznimne pomoći u vidu konstruktivnih savjeta i podrške u raznim prilikama.

Najveća hvala mojoj obitelji, posebno suprugi Teni, sinu Jakovu i roditeljima, koji su bili puni razumijevanja i strpljenja te nepresušan izvor inspiracije.

Povjerenstvo za ocjenu doktorske diseracije

- dr.sc. Goran Martinović, redoviti profesor u trajnom zvanju, Sveučilište J.J. Strossmayera u Osijeku, Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek, predsjednik
- dr.sc. Josip Job, izvanredni profesor, Sveučilište J.J. Strossmayera u Osijeku, Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek, mentor
- dr.sc. Emil Dumić, izvanredni profesor, Sveučilište Sjever, Odjel za elektrotehniku, član

Povjerenstvo za obranu doktorske diseracije

- dr.sc. Goran Martinović, redoviti profesor u trajnom zvanju, Sveučilište J.J. Strossmayera u Osijeku, Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek, predsjednik
- dr.sc. Josip Job, izvanredni profesor, Sveučilište J.J. Strossmayera u Osijeku, Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek, mentor
- dr.sc. Emil Dumić, izvanredni profesor, Sveučilište Sjever, Odjel za elektrotehniku, član

Datum obrane doktorske disertacije: 24. rujan 2020.

Sadržaj

1. UVOD

Predviđanje ishoda vrlo je popularno područje istraživanja, a najčešće se vrši na temelju iskustva ili znanja o određenom procesu. Ljudska je potreba predviđati ishode pojedinih procesa, a razna predviđanja se događaju svakodnevno bez posebnog razmišljanja. Predviđanje je u znanosti stroga i točno određena metodologija koja predviđa što će se dogoditi u određenim uvjetima. Znanstvena metoda temelji se na testnim izjavama koje su logična posljedica znanstvenih teorija, a koje se postižu ponovljivim pokusima ili promatranjem. Primjena metoda strojnog učenja kao alata u rješavanju brojnih problema postaje sve učestalija, pa su tako metode strojnog učenja svoju primjenu našle i u predviđanju sportskih ishoda. Predviđanje ishoda u sportu postalo je zanimljivo i široj javnosti i to u vidu sportskog klađenja, dok eksperti, treneri i sportski menadžeri strojno učenje koriste u svrhu evaluacije učinka igrača i momčadi, odabira igrača, identifikacije sportskih talenata, definiranja novih strategija itd.

Evaluacija učinka igrača i momčadi vrlo je bitan segment analize sportskih događaja i sama po sebi predstavlja zanimljivo područje istraživanja. Velik problem predviđanja ishoda u sportu je nemogućnost definiranja univerzalnog modela predviđanja. Istraživači se tako posvećuju jednom sportu i njegovim posebnostima te posljedično i ispitivanju metoda u okviru analiziranog sporta. Rezultati predviđanja ovise o vrsti sporta, broju mogućih ishoda, ali i o konkurentnosti pojedine lige unutar određenog sporta. Istraživanja u području predviđanja ishoda sportskih rezultata najzastupljenija su kod najpopularnijih sportova kao što su nogomet, košarka, američki nogomet i bejzbol.

Dostupnost informacija je svakim danom sve veća. Suočeni bezgraničnim količinama podataka sve više ljudi je posvećeno istraživanju vrijednosti podataka koji se spremaju u baze podataka. Baze podataka služe isključivo za pohranu podataka te nisu dovoljne za razumijevanje podataka, analiziranje podataka ili pretvaranje podataka u znanje. U prošlosti su ljudi iskustvom eksperata znali kako da podatke filtriraju, uspoređuju, sintetiziraju te izvlače iz podataka određena pravila i znanje. Tradicionalne metode analiziranja podataka, odnosno ograničenost analize od strane eksperata, i pouzdanost dobivenih informacija se mogu smatrati upitnim kada se govori o velikim količinama podataka. U trenutku kada tradicionalne metode usvajanja znanja ne mogu analizirati velike količine podataka, kao logično rješenje postavlja se znanost o podacima (engl. *data science*). Znanost o podacima se bavi strukturiranim i nestrukturiranim podacima te uključuje postupke čišćenja, pripreme i konačne analize podataka, a uključuje programiranje, logičko zaključivanje, matematiku i statistiku. Znanost o podacima je interdisciplinarno područje koje

objedinjuje statistiku, analizu podataka, strojno učenje, dubinsku analizu podataka i ostala srodna područja. Svako od navedenih područja ima svoju svrhu, a poseban fokus ovog rada će biti na predviđanju ishoda, području kojim se bavi strojno učenje. Konačan produkt istraživanja će biti metoda predviđanja ishoda sportskih događaja, a cilj rada će biti izgraditi model koji će biti dobra i korisna aproksimacija podataka.

U radu će biti predložena prilagodljiva metoda za predviđanje sportskih ishoda korištenjem klasifikacijske metode nadziranog strojnog učenja zasnovana na indeksu korisnosti i optimalnom vremenskom prozoru. Cilj rada je predložiti metodu koja će biti lako prilagodljiva ostalim sportovima, ali i procesima koji se mogu podijeliti na komponente. Sukladno tome, u radu će se predložiti indeks korisnosti kao mjera uspješnosti procesa te optimalan vremenski prozor kao mehanizam određivanja relevantnosti statističkih podataka o prethodnim događajima. Također, predstavit će se i postupak optimiranja indeksa korisnosti, ali i određivanje specifičnih kategorija procesa kako bi se primjenom prilagodljive metode predviđanja dobili optimalni rezultati.

Istraživanje će uključiti i pripremu ulaznog skupa podataka te odabir korisnih značajki, tj. onih značajki koje će se koristiti za predviđanje ishoda sportskih događaja. Osim uobičajenih značajki podatkovnog skupa sportskog događaja, a u svrhu poboljšanja rezultata predviđanja, metoda će koristiti i dodatne značajke dobivene postupkom izlučivanja značajki. Dodatno će biti analiziran skup značajki prikupljen iz rezultata radova drugih istraživača i prema iskustvima eksperata. Predložena metoda će biti eksperimentalno ispitana na stvarnim podacima NBA lige, a rezultati će biti analizirani.

Predložena metoda upotrebe indeksa korisnosti u svrhu predviđanja ishoda, klasifikacija sportskih događaja prema izvjesnosti uspješnog predviđanja, upotreba optimalnog vremenskog prozora i postupak optimiranja parametara indeksa korisnosti svakako predstavljaju novinu u odnosu na uobičajeno korištene metode u području predviđanja ishoda sportskih događaja.

Znanstveni doprinosi ove disertacije su sljedeći:

- Metoda za predviđanje ishoda sportskih događaja na temelju indeksa korisnosti i optimalnog vremenskog prozora
- Postupak optimiranja sveobuhvatnog indeksa korisnosti
- Algoritam izračuna i prilagodbe optimalnog vremenskog prozora

Disertacija je podijeljena u sedam poglavlja, a kratak sadržaj svakog poglavlja disertacije je dan u nastavku.

Drugo poglavlje: *Predviđanje ishoda u sportu*. U ovom poglavlju dan je pregled područja predviđanja sportskih ishoda. U okviru istraživanja, napravljena je analiza trenutno dostupnih publikacija iz područja. Navedene su korištene metode i pristupi u rješavanju spomenute

problematike s naglaskom na analizi korištenih postupaka odabira i izlučivanja značajki. Pružen je uvid u indekse korisnosti koji predstavljaju postojeće metrike za postupke evaluacije korisnosti igrača i momčadi te je predložen sveobuhvatni indeks korisnosti uz pripadajući matematički opis. Dan je i kratki uvod u područje strojnog učenja te su objašnjeni postupci validacije modela strojnog učenja s ciljem razumijevanja primjene u predviđanju sportskih ishoda.

Treće poglavlje: *Početne pretpostavke i analiza ulaznog skupa sportskih događaja*. U ovom poglavlju su navedene početne pretpostavke za izradu modela te je napravljena analiza ulaznog skupa podataka. Provedena su ispitivanja početnih pretpostavki kako bi se identificiralo postupke koji će omogućiti izradu učinkovitog modela za predviđanje sportskih ishoda. U poglavlju je provjerena primjenjivost sveobuhvatnog indeksa korisnosti momčadi kao mjere uspješnosti pojedine momčadi, odnosno temelja predložene metode predviđanja sportskih ishoda. Također, u ovom poglavlju su ispitana dva načina validacije modela kako bi se odabrao primjeren pristup u analizi dobivenih rezultata istraživanja. Isto tako, provjerena je i opravdanost primjene ograničenog vremenskog perioda u predviđanju te je pokazano postojanje optimalnog vremenskog prozora podatkovnog skupa unutar kojega postoji dovoljna količina informacija za izradu učinkovitog modela predviđanja.

Četvrto poglavlje: *Model predviđanja sportskih ishoda zasnovan na indeksu korisnosti i optimalnom vremenskom prozoru.* U ovom poglavlju je opisana predložena metoda za predviđanje sportskih ishoda zasnovana na indeksu korisnosti i optimalnom vremenskom prozoru. Sam postupak izrade modela je detaljno opisan i matematički potkrijepljen. Tako je objašnjen postupak optimizacije indeksa korisnosti, predložen je način dodavanja skupa dodatnih značajki sa svrhom dodatnog vrednovanja učinka pojedinih analiziranih procesa te je definiran algoritam izračuna i prilagodbe optimalnog vremenskog prozora s ciljem pronalaska podskupa skupa za učenje koji najbolje opisuje trenutno stanje analiziranog procesa, a da pritom ne doprinosi posljedičnom smanjenju rezultata predviđanja. Uveden je i pojam događaja povećane neizvjesnosti, točnije događaja kod kojih je razlika projiciranih indeksa korisnosti dva suprotstavljena procesa unutar unaprijed definiranog raspona, a koristi ih se s ciljem dodatnog poboljšanja rezultata predviđanja.

Peto poglavlje: *Analiza rezultata ispitivanja i optimiranje predloženog modela na primjeru utakmica NBA lige.* U ovom poglavlju je prikazano vrednovanje predloženih metoda na konkretnom primjeru predviđanja ishoda u NBA ligi. Model je eksperimentalno ispitan korištenjem informacijskog sustava Basketball Coach Assistant te je napravljena analiza dobivenih rezultata koji su prikazani u numeričkom i grafičkom obliku.

Šesto poglavlje: *Rasprava.* U ovom je poglavlju dan kratak osvrt na cjelokupno istraživanje te su navedeni razlozi otežane usporedbe rezultata istraživanja različitih znanstvenih istraživanja.

Sedmo poglavlje: *Zaključak.* U zaključku je dan osvrt na cjelokupno istraživanje. Rezimirani su rezultati te su navedene smjernice budućih istraživanja vezana uz područje predviđanja sportskih ishoda.

2. PREDVIĐANJE ISHODA U SPORTU

Predviđanje sportskih ishoda popularno je područje kojim se bavi mnoštvo istraživača. Najčešća predviđanja vezana uz sport se tiču najpopularnijih sportova. Pojam popularnosti sporta nije jednostavno definirati, a fraza „najpopularniji sport“ može značiti najgledaniji, zatim može značiti sport kojim se bavi najviše ljudi ili pak sport koji donosi najveći profit. Svakako nije pravedno popularnost sporta ocjenjivati na temelju jednog kriterija. Postoji velik broj novinskih članaka koji se bave analizom popularnosti sportova, a najaktualnije i najopširnije istraživanje vezano uz pronalazak najpopularnijeg sporta dano je u [1]. Autor je koristio 15 kriterija da bi pronašao najpopularniji sport vezanih uz broj gledatelja, prisutnost na Internetu, broj profesionalnih/amaterskih igrača, konkurentnost, broj država u kojima je sport popularan itd. Najpopularnijim sportom tako je proglašen nogomet, kojeg slijede kriket, košarka, hokej, tenis, odbojka, stolni tenis, bejzbol, itd. Zanimljivo istraživanje vezano uz popularnost sporta na temelju broja obožavatelja dano je u radu [2] gdje je najpopularnijim sportom proglašen nogomet, a slijede ga kriket, hokej, tenis, odbojka, stolni tenis, košarka, itd. Zanimljivo istraživanje dano je i u radu [3] gdje je autor rangirao sportove prema regijama, državama, mjesecima, čak i prema velikim sportašima.

U vrijeme generiranja velikih količina strukturiranih i nestrukturiranih podataka (engl. *big data*) vezanih uz sportske događaje interes sportskih djelatnika, ali i publike postao je sve veći. Dostupni podaci se osim u analitičke svrhe koriste i za predviđanje ishoda budućih događaja. Predviđanja se mogu vršiti na temelju iskustva eksperta ili na temelju podatkovnih primjera, točnije korištenjem podataka što dovodi direktno do pojma strojnog učenja. Govoreći o sportu, strojno učenje može pomoći trenerima ili sportskim menadžerima ne samo u predviđanju ishoda već i kod predviđanja učinka igrača ili momčadi, detekciji mogućnosti ozljede, identifikaciji mladih talenata, ali i kao pomoć široj javnosti u svrhu donošenja odluka vezanih uz sportsko klađenje. Većina sportova nudi dva ili tri konačna ishoda. Gledajući matematički i ne ulazeći u složenost sporta, lako je zaključiti da je sportove s dva moguća ishoda lakše predvidjeti, točnije postoji veća vjerojatnost da će se pogoditi konačni ishod. Primjerice, u nogometu su vrlo vjerojatna tri moguća ishoda te je samim time predviđanje ishoda otežano. Također, nogomet je momčadski sport u kojem na konačan ishod utječe više parametara. U prilog nogometu ne ide i mali broj golova. Uzevši u obzir prethodno navedene razloge, matematički najveću vjerojatnost za točno predviđenim ishodom imaju pojedinačni sportovi s dva moguća ishoda kao što su tenis, šah, većina borilačkih sportova, ali ne i pojedinačni sportovi u kojima sudjeluje više natjecatelja kao što su primjerice gimnastika, atletika, plivanje i ostali slični sportovi. Kao što je već ranije rečeno,

gledajući striktnu definiciju pojedinog sporta, najlakše bi trebalo biti predvidjeti ishode individualnih sportova u kojima se nadmeću dva natjecatelja, a moguća su dva ishoda. U ovom slučaju se u obzir ne uzimaju posebnosti natjecanja, ograničenja u budžetu, konkurentnost lige, format natjecanja i ostali slični faktori.

Najveća iznenađenja u sportu je najlakše pronaći analizom kladioničarskih tipova koji su nudili najveće zarade. Od najpoznatijih događaja treba izdvojiti FC Leicester City, klub koji je 2016. godine neočekivano osvojio naslov engleske Premier Lige [4]. Na početku prvenstva nudio se omjer 5000/1. Veliko iznenađenje priredila je i reprezentacija Grčke koja je 2004. godine s omjerom 150/1 osvojila Europsko nogometno prvenstvo. Veliko iznenađenje priredio je 2001. godine i hrvatski tenisač Goran Ivanišević osvojivši Wimbledon, jedan od najprestižnijih teniskih turnira. Kao 125. igrač svijeta najprije je prošao kvalifikacije, ulaskom u glavni turnir pobjeđivao redom bolje plasirane igrače i na kraju nakon velike drame u finalu pobijedio favoriziranog Australca Patricka Raftera. U kategoriju iznenađenja spada i FC Liverpool osvajanjem nogometne Lige prvaka. Nakon vodstva AC Milan 3:0 na poluvremenu, FC Liverpool je okrenuo rezultat te na kraju pobijedio utakmicu. Kladionice su u trenutku vodstva AC Milan na poluvremenu nudile omjer 100/1 da će FC Liverpool pobijediti. Ovo su samo neki od najpoznatijih događaja koji govore u korist kompleksnosti predviđanja ishoda u sportu.

U ovom će se poglavlju dati pregled radova vezanih uz predviđanje ishoda u sportu. Prva istraživanja vezana uz predviđanje ishoda u sportu krenula su krajem 20. stojeća, a velik broj radova počeo se pojavljivati početkom 21. stoljeća. Generalno gledajući najčešće korištena metoda predviđanja ishoda u sportu svakako su neuronske mreže [5] i [6].

2.1. Kompleksnost predviđanja ishoda u sportu

U ovom potpoglavlju dat će se pregled rezultata ostalih istraživača vezan uz kompleksnost predviđanja ovisno o samom sportu, točnije, dat će se uvid u radove koji se prvenstveno bave analizom kompleksnosti predviđanja.

U radu [7] autori su statistički analizirali rezultate pet ligaških sportova u Engleskoj i Sjedinjenim Američkim Državama s ciljem pronalaska najkonkurentnije sportske lige. Ukupno je analizirano više od 300 000 utakmica. Autori su uveli koeficijent q koji predstavlja vjerojatnost pobjede lošije momčadi. Na kraju su dobiveni rezultati $q \approx 0{,}45$ (45 %) za slučaj nogometa i bejzbola te $q \approx 0{,}35$ (35 %) za slučaj košarke i američkog nogometa. Važno je napomenuti da su se u obzir, kao kriva predviđanja, uzimali i primjeri kada je vjerojatnost za pobjedu bila 51:49 % u korist favorizirane momčadi, a pobijedila je momčad s manjom vjerojatnošću. Ignorirane su

samo utakmice u kojima je razlika bila manja od 0,05 u postotku pobjeda. Kao polazna točka predviđanja korišten je omjer momčadi na kraju prethodne sezone.

U radu [8] statistički su obrađene utakmice 198 sportskih liga. Obrađeno je ukupno 1503 sezona iz 84 države u četiri različita sporta (košarka, nogomet, odbojka i rukomet). Ukupno je korišteno 270 713 utakmica od siječnja 2007. godine pa do srpnja 2016. godine. Autori su definirali odnos između konačnog rezultata i savršeno idealiziranog natjecanja u smislu vještina. U obzir je uzet i faktor sreće. Dobiveni rezultati su pokazali da u NBA (engl. *National Basketball Association*) ligi postoji 35 % vjerojatnosti da će favorizirana momčad izgubiti uz 9 % vjerojatnosti koja se daje domaćoj momčadi. Isto tako, pokazano je da je faktor sreće prisutan i u najkonkurentnijim ligama čime je objašnjeno zašto sofisticirani i složeni modeli zasnovani na skupu značajki rezultatski značajno ne odskaču od jednostavnih metoda predviđanja. Konačni rezultati su pokazali da od svih sportova na konačni ishod najviše utječe vještina. Košarku slijede odbojka, nogomet i tek na kraju rukomet. Autori su također napravili analizu po sportovima koja govori koliki postotak momčadi je potrebno isključiti iz lige da bi vjerojatnost za pobjedom suprotstavljenih momčadi bile gotovo identične. Tako je potrebno isključiti 50 % momčadi iz košarkaških liga, 40 % iz odbojkaških liga, 19 % iz nogometnih te 14 % iz rukometnih liga.

2.2. Pregled istraživanja vezanih uz predviđanje ishoda u sportu

Gotovo je nemoguće odrediti mogućnosti predloženih modela jer predviđanje znatno ovisi o faktorima kao što su broj mogućih ishoda, konkurentnost lige, format lige, promjene u sastavima momčadi, promjene trenera itd. Stoga je puno lakše predložiti metodologiju koja se uz sitne izmjene može prilagoditi bilo kojem sportu. U ovom radu će se predložiti metoda predviđanja ishoda u prvenstveno momčadskim sportovima, a konačan model će biti eksperimentalno ispitan na košarci, točnije na američkoj profesionalnoj ligi NBA. Međutim, moguće je izvesti zaključke vezane uz pojedine sportove i pojedine lige. U okviru istraživanja napisan je pregledni rad [9] koji pruža uvid u istraživanje u kojem je analiziran velik broj radova vezan uz predviđanje ishoda u momčadskim sportovima kao što su košarka, nogomet, američki nogomet, kriket i bejzbol te su izvučeni određeni zaključci. Autori su zaključili da istraživači uglavnom koriste klasifikacijske metode nadziranog strojnog učenja, a glavnim problemom navedena je nemogućnost usporedbe rezultata jer praktički ne postoje radovi koji koriste isti skup podataka. Kao metoda validacije u većini slučajeva se koristi metoda podjele skupa podataka. Manji broj istraživača koristi i metodu unakrsne provjere, međutim sportski događaji nisu u potpunosti neovisni te samim time unakrsna provjera nije pogodna za sportska predviđanja. Općenito najbolji rezultati su dobiveni korištenjem manjih skupova podataka. Autori su naveli i nekoliko prijedloga koji bi mogli dovesti do boljih

rezultata predviđanja, a to su poboljšanje metode učenja, korištenje većeg broja algoritama strojnog učenja u svrhu pronalaženja optimalnog, poboljšanje metode odabira značajki, optimizacija parametara strojnog učenja, upotreba relevantnih i optimalnih skupova značajki te pronalaženje pravilnosti među podacima. Istraživanje je pokazalo da upotreba novih, alternativnih algoritama strojnog učenja može postići dobre, u nekim slučajevima i bolje rezultate predviđanja. Također je navedeno da glavni izazov vezan uz predviđanje sportskih ishoda predstavlja predlaganje univerzalne metode koja će moći uspješno predviđati ishode više sportova. Pošto će se ispitivanje modela vršiti na košarci, poseban naglasak je dan radovima vezanim uz predviđanje ishoda u košarci. Autori članka [10] su zaključili da najbolje rezultate predviđanja ishoda u košarci, točnije NBA ligi, daje prilagodba modela na način da se tijekom faze učenja koriste i podaci skupa za ispitivanje za koje je predviđanje obavljeno, odnosno upotreba jedne sezone skupa za ispitivanje i jedna do tri sezone skupa za učenje.

Kao što je već napomenuto, svakodnevno se generiraju velike količine strukturiranih i nestrukturiranih podataka vezanih uz sportske događaje. Osim povećanja količine podataka, povećava se i broj relevantnih baza podataka koje sadrže različite sportske statistike. Analizirajući radove ostalih istraživača, lako je uočiti da se kao izvor informacija uglavnom koriste službene stranice sportskih organizacija. U radu [9] je analizirano preko stotinu radova vezanih uz sportska predviđanja ili izvlačenje korisnih informacija i pravilnosti vezanih uz sport. U obzir su uzeti radovi koji koriste barem jedan algoritam strojnog učenja u predviđanju sportskih ishoda ili radovi koji iz ulaznog skupa podataka izvlače korisne činjenice i pravilnosti. Mogući razlozi za isključenje radova bili su poteškoća u interpretaciji podataka, loša metodologija rada ili nedostatak jasnoće. Tablica 2.1 prema [9] prikazuje kronološki poredane radove te pripadajuće izvore podataka, ulazne skupove podataka i korištene algoritme strojnog učenja kod predviđanja ishoda.

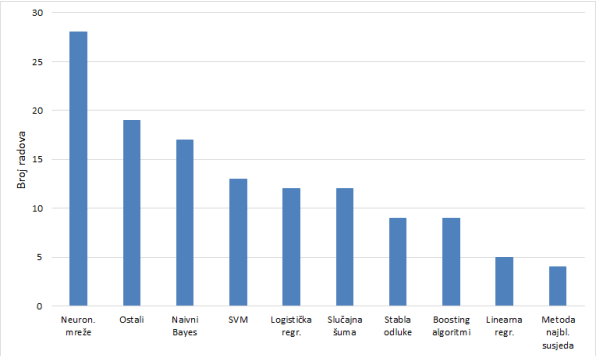
Tablica 2.1. Analizirani radovi poredani po godini objavljivanja.

#	Skup podataka	Izvor podataka	Korišteni algoritmi strojnog učenja
[11]	National Football League (NFL), sezona 1994, tjedni 11 - 16 (američki nogomet)	-	neuronska mreža
[12]	NFL, prvih 15 tjedana sezone 2013 (američki nogomet)	Službene stranice lige NFL	neuronska mreža
[13]	NFL, sezone 2003 – 2005 (američki nogomet)	Različiti izvori	logistička regresija, Stroj s potpornim vektorima (engl. *Support Vector Machine* - SVM)
[14]	English Premier League (EPL) i Australian Football League (AFL), sezone 2002 - 2007 (nogomet, američki nogomet)	Različiti izvori	neuronska mreža
[15]	Prvih 650 utakmica sezone 2007 (košarka)	ESPN	neuronska mreža
[16]	NBA, regularna sezona 2009 (košarka)	Službene stranice lige NBA, Yahoo Sports	Naivni Bayes, linearna regresija, metoda najbližih susjeda, stabla odluke, SVM
[17]	NBA, dvije uzastopne sezone (košarka)	Službene stranice lige NBA	37 algoritama strojnog učenja
[18]	Asociacion de Clubes de Baloncesto (ACB), sezona 2008 (košarka)	Službene stranice lige ACB	10 modela neizrazite logike

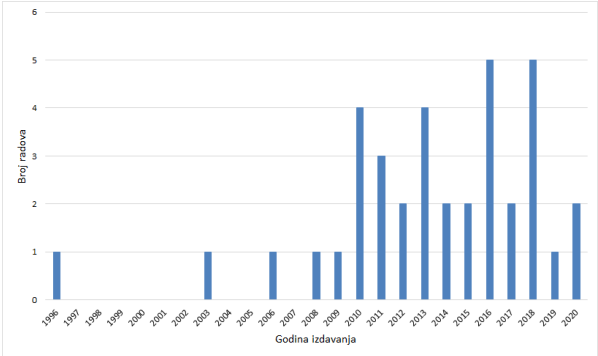
[19]	Srbija, Prva B liga, sezone 2005 – 2009 (košarka)	Košarkaški savez Srbije	neuronska mreža
[20]	Zadnjih 15 sezona nizozemskih nogometnih liga (nogomet)	www.football-data.co.uk	linearna/logistička regresija, stabla odluke, LogitBoost, Bayesova mreža, Naivni Bayes
[21]	NFL i sveučilišni američki nogomet, sezone 2003 – 2010 (američki nogomet)	ESPN, USA today	neuronska mreža
[22]	Liga prvaka (nogomet)	Različiti izvori	Naivni Bayes, Bayesova mreža, LogitBoost, metoda najbližih susjeda, slučajna šuma, neuronska mreža
[23]	NBA, sezone 2005 – 2010 (košarka)	Službene stranice lige NBA, Basketball Reference, Databasebasketball.com	logistička regresija, Naivni Bayes, SVM, neuronska mreža
[24]	Sveučilišni američki nogomet, sezone 2002 – 2010 (američki nogomet)	Različiti izvori	stabla odluke, neuronska mreža, SVM
[25]	NBA, sezone 2007 – 2009 (košarka)	Basketball Geek	SVM, logistička regresija
[26]	NBA, regularni dio sezona 2006 – 2012 (košarka)	Basketball Reference	linearna regresija, Maximum Likelihood klasifikator, Multilayer Perceptron – Back Propagation
[27]	National Collegiate Athletic Association Basketball (NCAAB), sezone 2008 – 2013 (košarka)	kenpom.com	stabla odluke, slučajna šuma, Naivni Bayes, neuronska mreža
[28]	Španjolska nogometna liga (Primera), sezona 2008 (nogomet)	Football356, Soccer-Spain, Stat-Football	Bayesova mreža
[29]	NBA, sezone 1991 – 1997 (košarka)	Službene stranice lige NBA	logistička regresija, Adaptive Boost, slučajna šuma, SVM, Naivni Bayes
[30]	EPL, sezona 2014 (nogomet)	-	neuronska mreža, logistička regresija
[31]	English Twenty over Country Cricket Cup, sezone 2009 – 2014 (kriket)	www.cricinfo.com	Naivni Bayes, logistička regresija, slučajna šuma, Gradient boosted stabla odluke
[32]	Dutch Eredivisie, sezone 2000 – 2012 (nogomet)	Različiti izvori	Naivni Bayes, LogitBoost, neuronska mreža, slučajna šuma, genetsko programiranje
[33]	NBA, sezona 2013/2014 (košarka)	Basketball Reference	Linearna regresija, Gaussova analiza, SVM, slučajna šuma, Adaptive Boost
[34]	EPL, sezone 2010 – 2015 (nogomet)	www.football-data.co.uk, sofifa.com/teams	logistička regresija
[35]	Regularni dio Major League Basketball (MLB), sezone 2005 – 2014 (bejzbol)	Retrosheet, Lahman Database	metoda najbližih susjeda, neuronska mreža, SVM, stabla odluke
[36]	NBA, sezone 2007 – 2014 (košarka)	Službene stranice lige NBA	Model maksimalne entropije
[37]	NBA, sezone 1985 – 2016 (košarka)	Službene stranice lige NBA, Basketball Reference	Faktorizacija matrica
[38]	NBA, sezone 2008 – 2010 (košarka)	Basketball Reference, Službene stranice lige NBA	SVM
[39]	Indian Premier League 2014/2015, Svjetsko prvenstvo 2015 (kriket)	Twitter, cricinfo.com, cricbuzz.com, ostale službene stranice momčadi	SVM, Naivni Bayes, linearna regresija
[40]	Euroliga (EL), sezone 2012 – 2017 (košarka)	Službene stranice Eurolige	metoda najbližih susjeda
[41]	MLB, sezone 1930 – 2016 (bejzbol)	Retrosheet	slučajna šuma, XGBoost, logistička regresija, GLM
[42]	NBA, sezone 2013 – 2014 (košarka)	Basketball Reference	Bayesova regresija
[43]	NBA, sezone 2002 – 2016 (košarka)	-	neuronska mreža
[44]	Španjolska nogometna liga, sezone 2012 – 2016 (nogomet)	Različiti izvori	logistička regresija, slučajna šuma, neuronska mreža, SVM, Naivni Bayes
[10]	NBA, sezone 2009 – 2017 (košarka)	Basketball Reference	naivni model strojnog učenja
[45]	NBA, sezone 2000 – 2014 (košarka)	Službene stranice lige NBA	neuronska mreža, logistička regresija
[46]	EPL, Ligue 1, Bundesliga, Seria A, Primera Division, sezone 2013 – 2017 (nogomet)	Sportal	slučajna šuma, gradient boosting algoritam, SVM, linearna regresija
[47]	5 europskih nogometnih liga i pripadajuće druge lige, sezone 2006 – 2017 (nogomet)	FIFA Index	slučajna šuma, boosting algoritam, SVM, linearna regresija

Ukupno je analizirano 39 radova iz razdoblja od 1996. do 2020. godine koji se tiču pet momčadskih sportova. Većina radova se odnosi na košarku, a slijede je nogomet, američki nogomet, bejzbol i kriket. Najpopularnija košarkaška liga je NBA (75,00 %), dok je najpopularnija nogometna liga Liga prvaka (55,56 %). Što se tiče američkog nogometa, najpopularnija liga je NFL, a uključuje ukupno 80 % pripadajućih radova. Ostale lige se pojavljuju dva puta ili manje od 10 %. Većina radova uključuje više algoritama strojnog učenja, tako da analizirani uzorak sadrži više od 100 rezultata predviđanja sportskih ishoda. Ukoliko autori koriste različite skupove značajki ili skupove podataka, analizirani su samo najbolji rezultati. Usporedba rezultata je vrlo teška ili gotovo nemoguća jer istraživači koriste različite skupove podataka i lige različite konkurentnosti.

Najčešće korištena metoda za predviđanje sportskih ishoda su neuronske mreže. Slika 2.1 prikazuje broj radova skupine algoritama strojnog učenja vezanih uz predviđanje sportskih ishoda nevezanih uz sport [9]. Algoritmi strojnog učenja grupirani su prema sličnostima. Varijante neuronskih mreža kategorizirane su pod nazivom „Neuronske mreže“. Grupiraju se i Naive Bayes i Bayesian mreže, kao i svi Boosting algoritmi (LogitBoost, AdaBoost, XGBoost itd.). Kada se u radu koristi više algoritama strojnog učenja, uključuju se samo najbolji rezultati po korištenom algoritmu. Slika 2.2 prema [9] prikazuje broj objavljenih radova po godinama.



Slika 2.1. Broj radova vezanih uz grupu algoritama strojnog učenja.



Slika 2.2. Broj radova po godinama.

Uz standardne algoritme strojnog učenja, istraživači koriste i druge metode kao što su modeli neizrazite logike (engl. *fuzzy rule-based models*), entropijski model, modeli faktorizacije matrica, itd. Slika 2.2 naglašava porast interesa istraživača za područje predviđanja sportskog ishoda. Pojačano zanimanje za predviđanje sportskih ishoda započelo je oko 2010. godine.

2.2.1. Odabir i izlučivanje značajki

Odabir i izlučivanje značajki predstavljaju vrlo važan preliminarni korak strojnog učenja jer značajno pridonosi uspjehu algoritma. Odabir značajki je postupak prepoznavanja i uklanjanja što

je više nebitnih i suvišnih značajki u svrhu smanjenja dimenzionalnosti problema i omogućavanja algoritmima da brže i učinkovitije djeluju [5], [48] i [49]. Izlučivanje značajki je proces smanjenja dimenzionalnosti kojim se početni skup neobrađenih podataka svodi na lakše upravljive i obradive skupove podataka [50]. Varijable su kombinirane u značajke koje učinkovito smanjuju količinu podataka koji se moraju obraditi, a još uvijek točno i potpuno opisuju izvorni skup podataka. Tehnike odabira i izlučivanja značajki povećavaju učinkovitost algoritma strojnog učenja i ubrzavaju rad samog algoritma, što je ključno za analizu velikih skupova podataka. Ulazni skup značajki obično sadrži značajke odabrane prema iskustvu eksperta koje su u pravilu kombinacija drugih poznatih, javno dostupnih značajki.

Većina istraživača koristi metode odabira i izlučivanja značajki tijekom faze pripreme podataka. Autori najčešće najprije na temelju iskustva definiraju skup ulaznih značajki, a nakon toga, ukoliko je potrebno, koristeći metodu izlučivanja značajki izračunavaju nedostajuće vrijednosti definiranih ulaznih značajki. Istraživanja koja uspoređuju učinak algoritma strojnog učenja prije i nakon korištenja odabira značajki vrlo su rijetka. Radove je moguće podijeliti na radove koji koriste odabir značajki temeljen samo na iskustvu istraživača i na radove koji koristi sofisticiranije metode odabira značajki. Relativno mali broj radova koristi dodatni odabir značajki temeljen na eksperimentima ili kombinacijom početnih skupova značajki, s ciljem pronalaženja optimalnog podskupa značajki koji će maksimizirati točnost predloženog modela. Sofisticiranije metode odabira značajki se koriste za povećanje relevantnosti i minimiziranje redundancije. Odabir značajki uključuje izračun podskupa značajki koji sadrži samo relevantne značajke. Metode odabira značajki obično se klasificiraju u filter metode (engl. *filter methods*), metode omotača (engl. *wrapper methods*), ugrađene metode (engl. *embedded methods*) i hibridne metode (engl. *hybrid methods*) [51], [52] i [53].

Teško je zaključiti koja metoda odabira značajki daje najbolje rezultate. Najčešće su korištene filter metode odabira značajki koje odabiru značajke na temelju mjere učinkovitosti bez obzira na algoritam modeliranja. Filter metode odabira značajki karakterizira smanjenje dimenzionalnosti problema prije korištenja samog algoritma strojnog učenja. Usporedba rezultata prije i nakon odabira značajki nije prikazana u većini članaka, ali pretpostavka je da su bolji rezultati dobiveni korištenjem metoda odabira značajki. Broj značajki se razlikuje od istraživanja do istraživanja. Većina autora koristi metode odabira i izlučivanja značajki, a prosječan broj korištenih značajki, uključujući i stršeće vrijednosti (engl. *outliers*), iznosi otprilike 57. Nakon korištenja metoda odabira i izlučivanja značajki, broj odabranih značajki, uključujući stršeće vrijednosti, iznosi približno 39. Istraživači uglavnom koriste elemente osnovne statistike analiziranog sporta te često uključuju i izlučene značajke vezane uz uspješnost momčadi. Ostali skupovi značajki, kao što su

npr. podaci o psihološkom stanju ili podaci s društvenih mreža i medija, rijetko se koriste. Mnogi istraživači ističu da bi povećanje skupa značajki moglo dovesti do boljih rezultata predviđanja. Dodatne značajke zasigurno mogu biti predloženi skupovi značajki poput psihološkog stanja ili podaci s društvenih mreža. Oba pristupa predstavljaju novost u području predviđanja sportskih ishoda. Analiza radova otkrila je da su raniji radovi vezani uz predviđanje sportskih ishoda često koristili odabir značajki na temelju znanja ili prethodnog iskustva. U posljednje vrijeme sve više autora koristi sofisticiranije metode odabira značajki. Generalni zaključak, a vidljiv iz analiziranih radova, jest da odabir i izlučivanje značajki doprinose povećanju učinkovitosti modela zasnovanog na strojnom učenju. Većina radova ne prikazuje usporedbu rezultata predviđanja korištenjem početnog skupa značajki i rezultata dobivenih uporabom odabranih značajki. Radovi s predstavljenim poboljšanjima rezultata predviđanja koristeći dodatan odabir značajki su [15], [18], [20], [38], [40] i [43]. Ostali istraživači samo zaključuju da dodatan odabir značajki doprinosi boljim rezultatima predviđanja. U idealnom slučaju, prilikom predlaganja modela predviđanja ishoda trebalo bi testirati nekoliko različitih skupova značajki i sukladno tome odabrati onaj koji donosi najbolje rezultate. Pretpostavka je da bi upotreba metode omotača ili ugrađenih metoda odabira značajki mogla dovesti do poboljšanih rezultata predviđanja.

2.2.2. Evaluacija rezultata ostalih istraživača

Neuronske mreže od početka su najčešće korištena metoda strojnog učenja u predviđanju sportskog ishoda ([5] i [50]). Iako svakim danom postoji sve više novih metoda strojnog učenja, trend korištenja neuronskih mreža nastavlja se i dalje. U ovom odjeljku će se dati uvid u rezultate korištenja algoritama strojnog učenja kod predviđanja sportskih ishoda. Tablica 2.1 daje popis analiziranih radova i korištenih algoritama strojnog učenja.

Tablica 2.2 prikazuje sažetak analiziranih radova [9]. Tablicom su prikazane informacije vezane uz najvišu postignutu točnost te algoritam kojim je točnost postignuta, broju korištenih sezona ulaznog skupa podataka, broju značajki, informaciju ukoliko su korištene metode odabira i izlučivanja značajki te informaciju o korištenoj metodi validacije. Istraživači u svojim radovima najčešće koriste dvije metode validacije, metodu podjele skupa podataka i metodu unakrsne provjere. Tablicom su prikazani rezultati predviđanja pet sportova. Vidljivo je da rezultati unutar istog sporta variraju. Najviše istraživanja je vezano uz košarku, a najbolji postignuti rezultat je 85,28 %, ali isto tako postoji i istraživanje u kojem je jedna od korištenih metoda postigla maksimalnu točnost od 60,23 %. Najveći nesrazmjer u točnosti predviđanja vezan je uz nogomet gdje najviša točnost iznosi 93,00 %, dok je najniža zabilježena točnost 54,03 %. Rezultati američkog nogometa (50,00 – 86,48 %) i kriketa (55,60 – 87,90 %) također predstavljaju veliki

nesrazmjer u rezultatima predviđanja. Najujednačeniji su rezultati predviđanja ishoda vezani uz bejzbol (53,75 – 58,92 %). Važno je napomenuti da je najmanji broj istraživanja vezan uz kriket i bejzbol te samim time i ne čudi veliki nesrazmjer ili pak velika ujednačenost rezultata.

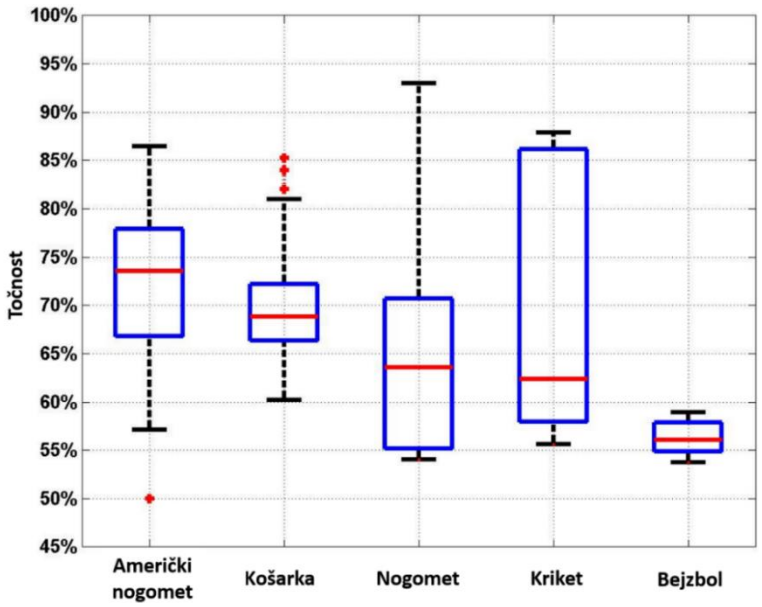
Tablica 2.2. Korišteni algoritam strojnog učenja i najveće točnosti po analiziranom radu.

#	Najveća točnost	Algoritam strojnog učenja	Broj sezona	Broj znač.	Odabir znač.	Izluč. Znač.	Metoda validacije
[30]	93,00 %	logistička regresija	1	9	✓	✓	Podj. skupa podat.
[28]	92,00 %	Bayesova mreža	1	13	✓	✓	Podj. skupa podat.
[39]	87,90 %	SVM	1	3	✓	✓	Unakr. validac.
[24]	86,48 %	stabla odluke	9	28	✓	✓	Unakr. validac.
[42]	85,28 %	Bayesova regresija	1	17	✓	x	Podj. skupa podat.
[38]	85,25 %	SVM	3	7/17	✓	x	Podj. skupa podat.
[40]	83,96 %	k-NN	3	15	✓	x	Podj. skupa podat.
[43]	82,00 %	neuronska mreža	15	352	✓	x	Podj. skupa podat.
[47]	81,77 %	integrirani model	12	40	✓	✓	Podj. skupa podat.
[19]	80,96 %	neuronska mreža	5	9	✓	x	Podj. skupa podat.
[11]	78,60 %	neuronska mreža	1	5	✓	x	Podj. skupa podat.
[46]	75,62 %	slučajna šuma	5	39	✓	✓	Podj. skupa podat.
[12]	75,00 %	neuronska mreža	1	5	✓	✓	Podj. skupa podat.
[27]	74,46 %	neuronska mreža	5	-	✓	x	Podj. skupa podat.
[36]	74,40 %	Model maksimalne entropije	8	28	✓	x	Podj. skupa podat.
[15]	74,33 %	neuronska mreža	1	4/22	✓	x	Podj. skupa podat.
[17]	72,80 %	logistička regresija	2	10	✓	✓	Podj. skupa podat.
[44]	71,63 %	logistička regresija	5	13	✓	✓	Unakr. validac.
[18]	71,50 %	Class-Fuzzy-Chi-RW algoritam	1	5/15	✓	✓	Unakr. validac.
[37]	70,95 %	Faktorizacija matrice	2	-	✓	x	Podj. skupa podat.
[25]	70,01 %	SVM	3	39	✓	✓	Unakr. validac.
[23]	69,67 %	logistička regresija	6	46	✓	✓	Podj. skupa podat.
[34]	69,50 %	logistička regresija	6	4	✓	✓	Podj. skupa podat.
[45]	68,83 %	neuronska mreža	15	-	✓	✓	Podj. skupa podat.
[22]	68,80 %	neuronska mreža, LogitBoost	1	20/30+	✓	✓	Podj. skupa podat.
[26]	68,44 %	neuronska mreža	7	8	✓	✓	Podj. skupa podat.
[13]	67,08 %	logistička regresija	3	30+	✓	✓	Pojed. unakr. validac.
[16]	67,00 %	Naivni Bayes	1	32	✓	✓	Unakr. validac.
[33]	65,53 %	Gaussova analiza	1	218	✓	✓	Unakr. validac.
[29]	65,15 %	slučajna šuma	7	7/17	✓	✓	Podj. skupa podat.
[31]	62,40 %	Naivni Bayes	6	31/500+	✓	✓	Podj. skupa podat.
[35]	58,92 %	SVM (klasifikacija)	10	3/60	✓	✓	Unakr. validac.
[14]	58,90 % (EPL), 67,20 % (AFL)	neuronska mreža	6	19	✓	✓	Podj. skupa podat.
[20]	57,00 %	linearna regresija	15	9/18	✓	✓	Unakr. validac.
[32]	56,10 %	LogitBoost	13	-	✓	x	Unakr. validac., podj. Skupa podat.
[41]	55,52 %	XGBoost	1	164	✓	✓	Unakr. validac.

Jedini analizirani sport koji se igra u dvorani je košarka i samim time vremenski uvjeti nemaju utjecaj na konačni ishod. S druge strane, košarka je sport u kojem se postiže mnogo više poena u odnosu na ostale sportove. Predviđanje sportskih ishoda složen je problem zbog niza neizvjesnosti koje mogu čak i tijekom igre utjecati na konačni ishod, poput ozljede igrača, vremenskih uvjeta, taktičkih promjena u igri, itd.

Većina istraživanja vezana je u predviđanje košarkaških i nogometnih ishoda, sportova s dva (košarka) i tri (nogomet) moguća ishoda. Važno je napomenuti da je neriješen ishod, koji ne postoji u košarci, vrlo vjerojatan u nogometu. Prosječna maksimalna točnost po analiziranom radovima i sportu je 73,92 % za košarku i 72,43 % za nogomet. Izuzevši dvije stršeće vrijednosti (maksimalna točnost od 92,00 % i 93,00 %) kod predviđanju nogometnih ishoda, prosječna maksimalna točnost pada na 67,42 %. Prosječna točnost za NBA ligu, najpopularniju košarkašku ligu, iznosi 72,83 %. Najveći problem, čak i za radove vezane uz NBA ligu gdje je dostupno najviše istraživanja, je uporaba različitih skupova podataka, što usporedbu rezultata čini vrlo teškom ili gotovo nemogućom. Najlošiji prosječni rezultati vezani su uz bejzbol, sport s dva moguća ishoda. Opći je zaključak da složenost predviđanja ishoda najviše ovisi o konkurentnosti lige i raste s brojem mogućih ishoda. Isto tako, rezultati predviđanja košarkaških ishoda su se pokazali dosljednijima od predviđanja nogometnih ishoda.

Najprikladniji način prikaza raspona rezultata predviđanja je kutijasti dijagram ili karakteristična petorka uzorka (engl. *box plot*). Kutijasti dijagram je metoda za grafički prikaz skupina numeričkih podataka kroz njihove kvartile, a može lako organizirati velike količine podataka te prikazati stršeće vrijednosti. Slika 2.3 prikazuje raspon rezultata kategorizan prema sportu [9].

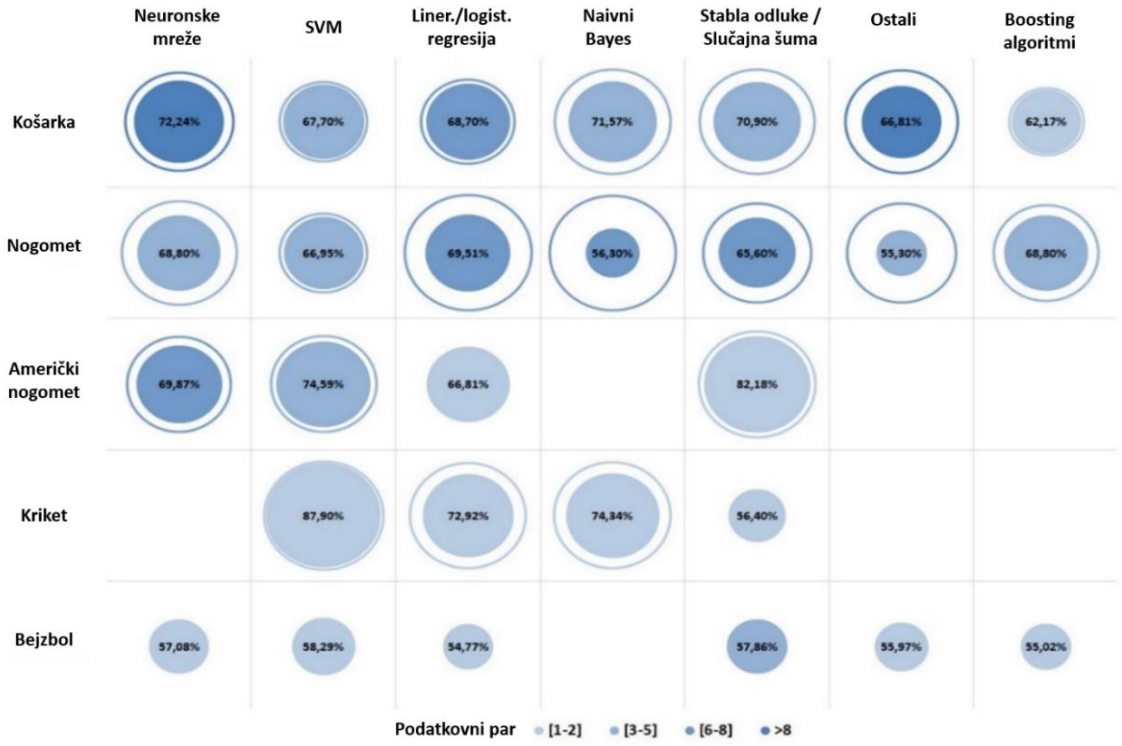


Slika 2.3. Kutijasti dijagram maksimalnih točnosti po sportu.

Na slici je jasno vidljivo da se rezultati vezani uz bejzbol ističu nižim rezultatima predviđanja, ali i da je okvir rezultata predviđanja relativno kratak. To ukazuje da istraživanja imaju visoku razinu uniformnosti. Za usporedbu, okvir predviđanja ishoda u nogometu je vrlo visok što upućuje na to da istraživanja imaju prilično nisku razinu uniformnosti. Okvir košarkaških predviđanja sadrži stršeće vrijednosti, ali je znatno kraći od rezultata predviđanja vezanih uz nogomet i kriket,

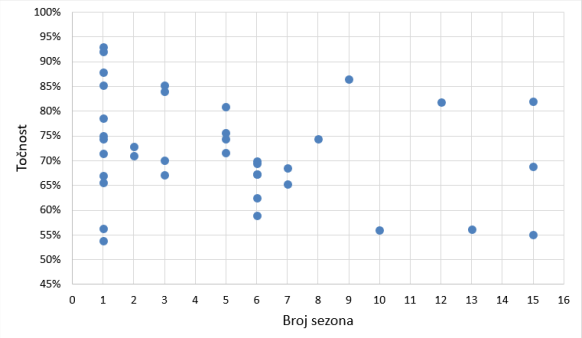
što sugerira da većina košarkaških istraživanja ima visoku razinu uniformnosti. Analizirajući rezultate predviđanja nogometnih ishoda jasno je da dva rezultata odskaču. Navedena istraživanja koriste skup podataka koji čini jednu sezonu, a podaci su podijeljeni na skup za učenje i skup za ispitavanje. U radu s najvišom točnosti koristi se isti skup za učenje i ispitivanje što je još jedan razlog visokih rezultata. Ukoliko se koristi skup podataka jedne sezone i ako se veći dio sezone koristi za učenje, a manji dio za ispitivanje, moguće je dobiti nerealno visoke rezultate predviđanja. Drugi problem vezan uz rezultate predviđanja za nogomet i kriket je položaj medijana koji je u ovom slučaju vrlo nizak (medijan je obično blizu prosjeka), što također ukazuje na nisku razinu ujednačenosti među radovima. Ostali sportovi, osim kriketa, također imaju veliki raspon rezultata, što se i očekuje s obzirom da radovi ne koriste isti skup podataka ili se ne odnose na isto natjecanje.

Zanimljiva, ali s obzirom na mali broj uzoraka neprikladna za prikaz kutijastim dijagramom, je analiza rezultata predviđanja po grupama algoritama strojnog učenja. Prikladan alat za prikaz rezultata predviđanja prema grupama strojnog učenja je matrični mjehurićasti graf (engl. *matrix bubble chart*), kategorijski grafički prikaz mjehurićima gdje su podaci prikazani elipsama, nijansom boje i veličinom. Slika 2.4 prema [9] prikazuje rezultate predviđanja grupe algoritama strojnog učenja i analiziranog sporta u kojem grupa algoritama strojnog učenja i sport formiraju par. Medijalna vrijednost rezultata predviđanja je prikazana veličinom elipse, dok je broj pojavljivanja podatkovnog para unutar skupa podataka prikazan intenzitetom boje elipse.

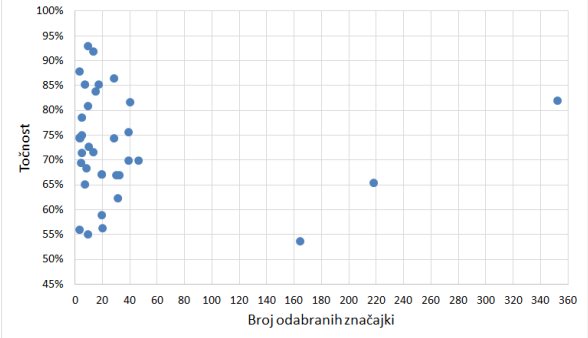


Slika 2.4. Medijalna vrijednost, maksimalna točnost i broj pojavljivanja predložene grupe algoritama strojnog učenja i analiziranog sporta u odnosu na cjelokupan skup podataka. Veličina vanjske elipse prikazuje maksimalnu točnost, medijalna vrijednost prikazana je veličinom elipse, a broj pojavljivanja podatkovnog para intezitetom boje.

Broj u sredini elipse i veličina unutarnje elipse predstavljaju medijalnu vrijednost po grupi algoritama strojnog učenja i sportu, tj. paru podataka, dok veličina vanjske elipse predstavlja postignutu maksimalnu točnost. Intenzitet boje predstavlja relevantnost podatkovnog para gdje tamnoplava elipsa predstavlja više pojava određenog podatkovnog para unutar skupa podataka nego ako je elipsa ispunjena svjetlijom plavom bojom. Parovi podataka bez jasno definirane razlike između medijalne vrijednosti i maksimalne točnosti su parovi gdje je medijalna vrijednost jednaka maksimalnoj točnosti (pojedinačna pojava para) ili su medijalna vrijednost i maksimalna točnost gotovo jednaki (dvije pojave). Sličnost veličine unutarnje i vanjske elipse predstavlja uniformnost rezultata predviđanja. Preklapanje elipsi predstavlja visoku ujednačenost rezultata, a uglavnom je vidljivo unutar parova s malim brojem rezultata kao što su kriket i bejzbol. Dvije iznimke vezane uz nisku uniformnost predviđanja ishoda u kriketu je uzrokovana analizom svega dva istraživanja koja se odnose na različite skupove podataka i analizu različitih liga. Općenito najniža razina uniformnosti vezana je uz predviđanje ishoda u nogometu, a prikazana je razlikom u veličini unutarnje i vanjske elipse. Takav rezultat je već sugerirao kutijasti dijagram (Slika 2.3) s razmjerno visokim okvirom i relativno niskim položajem medijalne vrijednosti. Razlog za to je već spomenut i odnosi se na korištenje različitih skupova podataka i liga različitih konkurentnosti. Isto tako, nogomet je jedini sport s tri vrlo vjerojatna ishoda. Najrelevantniji su rezultati predviđanja u košarci sa zadovoljavajućom ujednačenošću s obzirom da su korišteni različiti skupovi podataka, različite značajke i što je najvažnije, lige različitih konkurentnosti. Općenito najveću ujednačenost postižu grupe algoritama strojnog učenja i sportovi s manjim brojem pojava unutar skupa podataka. U nastavku odjeljka će se analizirati svojstva ulaznih skupova podataka (broj sezona, broj značajki, broj referenci, napredak točnosti predviđanja). Slikama u nastavku biti će prikazani samo najbolji rezultati predviđanja vezani uz pojedino istraživanje.



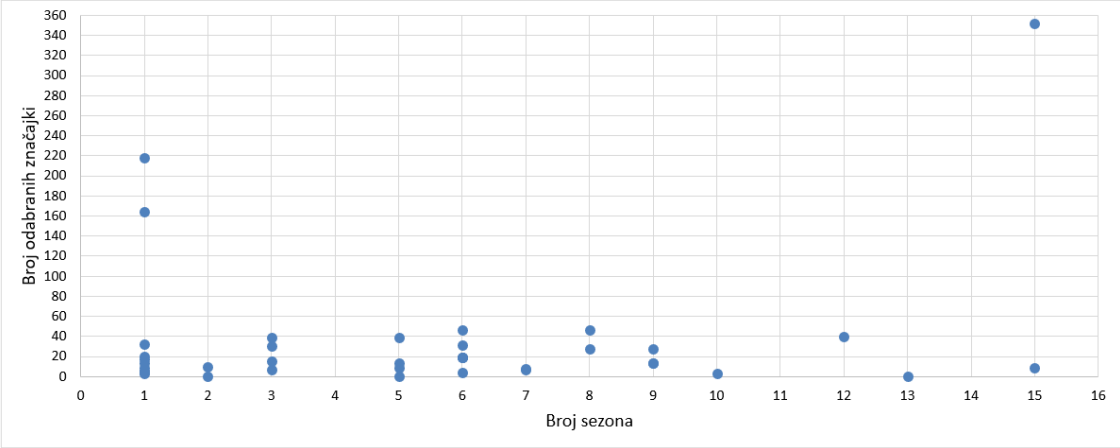
Slika 2.5. Ovisnost točnosti predviđanja o broju korištenih sezona.



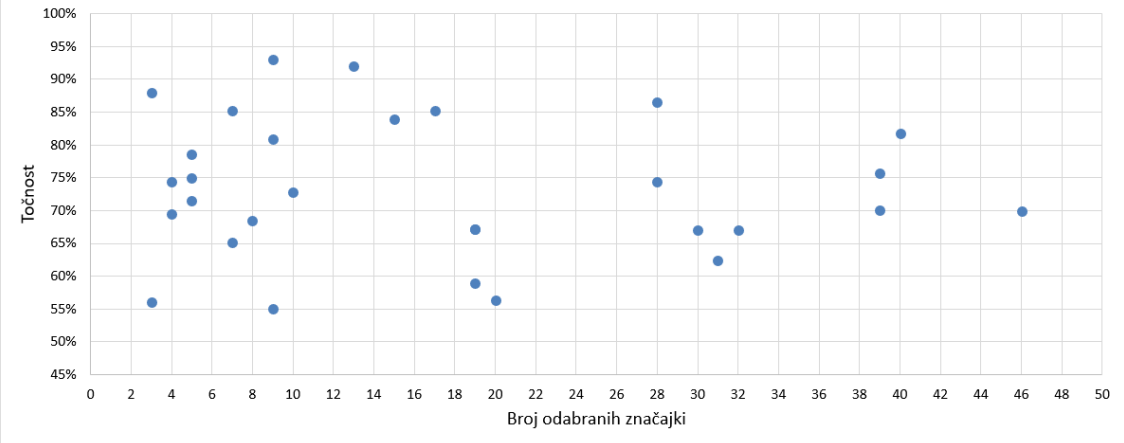
Slika 2.6. Ovisnost točnosti predviđanja o broju odabranih značajki.

Kao što je vidljivo sa slika, točnost algoritama predviđanja ishoda neovisna o analiziranom sportu u potpunosti ne ovisi o broju korištenih sezona i značajki. Veći skup podataka, konkretno

veći broj sezona ili značajki, ne znači nužno pad u točnosti rezultata, već samo sugerira da se gotovo jednaki ili čak bolji rezultati mogu postići korištenjem manjih skupova podataka. Slika 2.6 ima nekoliko stršećih vrijednosti [9]. Uklanjanjem stršećih vrijednosti (Slika 2.8 prema [9]) vidljivo je blago povećanje točnosti u slučajevima kada se koristi manje od 10 značajki. Kao što je prethodno rečeno, veći podskup značajki ne znači nužno lošije rezultate predviđanja, već sugerira da se analizirani proces može opisati manjim podskupom značajki bez da se ugrozi točnost predviđanja.

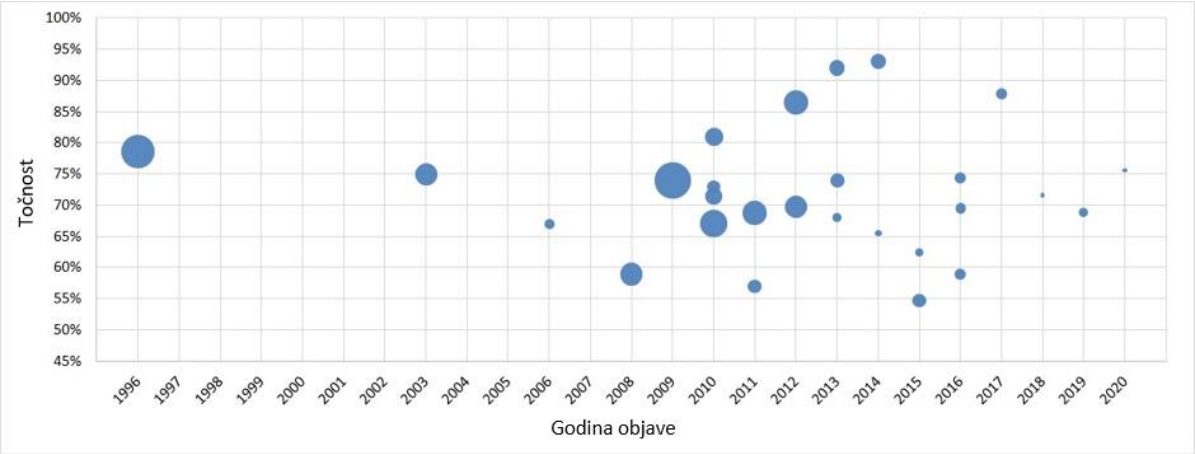


Slika 2.7. Ovisnost broja značajki i korištenih sezona.



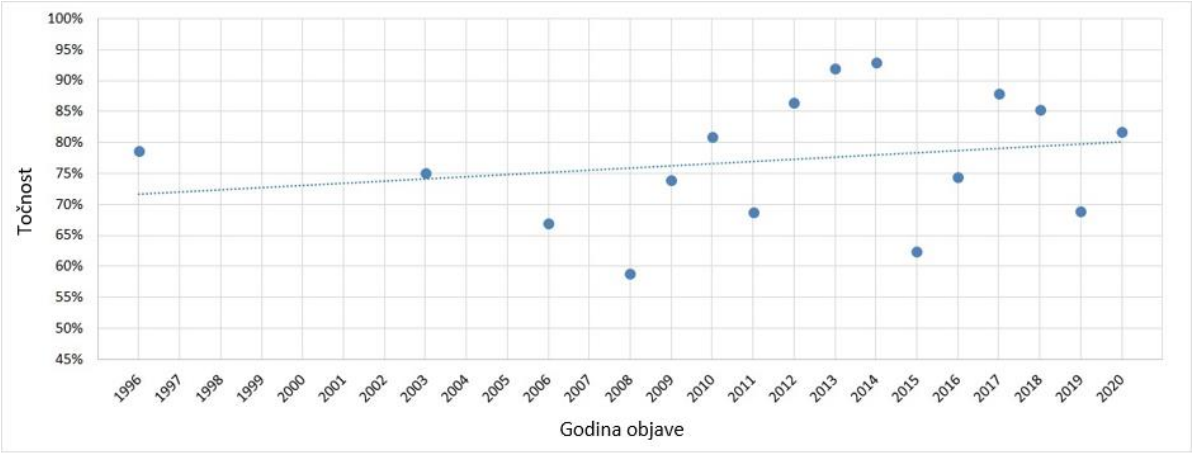
Slika 2.8. Ovisnost točnosti predviđanja o broju odabranih značajki (bez stršećih vrijednosti).

Slika 2.7 prema [9] prikazuje odnos broja značajki i korištenih sezona. U slučaju korištenja metode odabira značajki u analizi se koristi smanjeni skup značajki. Iz slike se jasno vidi da su broj značajki i broj korištenih sezona obično u obrnuto proporcionalnom odnosu. Postoji nekoliko radova s manjim brojem značajki i korištenih sezona i samo jedan rad s većim brojem značajki i korištenih sezona. Slikom je također pokazano kako istraživači uglavnom koriste manji broj sezona skupa za učenje i skupa za ispitivanje kako bi postigli bolje rezultate predviđanja.



Slika 2.9. Ovisnost rezultata predviđanja u odnosu na godinu objave i broj citata.

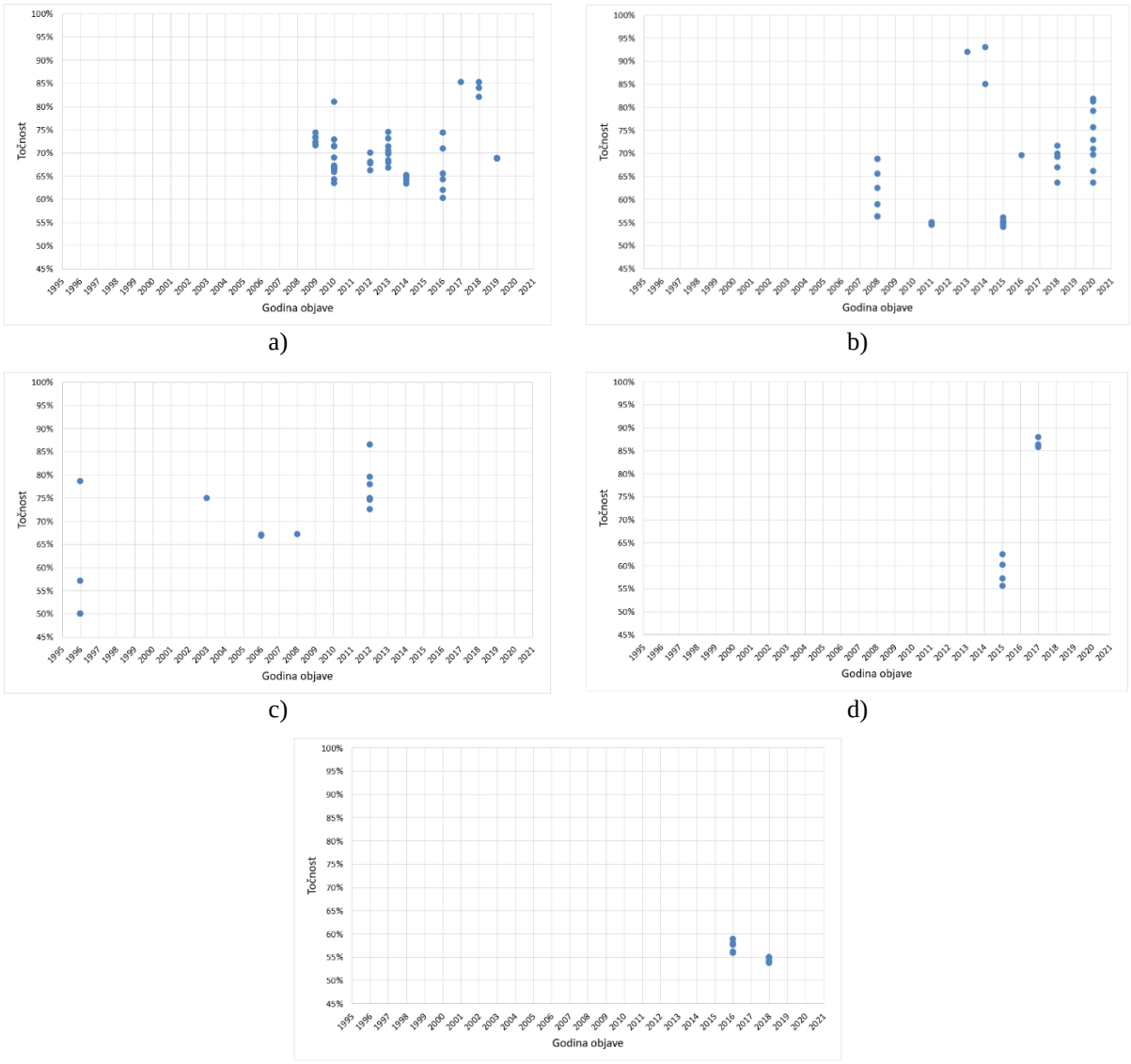
Slika 2.9 prikazuje ovisnost maksimalne točnosti u odnosu na godinu objave i broju citata prema Google znalcu (engl. *Google Scholar*) gdje veličina mjehurića predstavlja broj citata [9]. Najcitiraniji radovi su iz 1996. i 2009. godine, a najviše analiziranih radova je iz perioda 2010. do 2013. godine. Slika jasno pokazuje da broj citata najviše ovisi o godini objavljivanja, što ide u prilog najcitiranijim radovima koji se tiču predviđanja ishoda u američkom nogometu [11] i košarci [15]. Noviji radovi, bez obzira na njihovu kvalitetu, imaju manje citata od starijih, što je i očekivano zbog potrebnog vremenskog perioda do objave novih radova u kojima će ih se citirati.



Slika 2.10. Napredak algoritama strojnog učenja neovisno o analiziranom sportu.

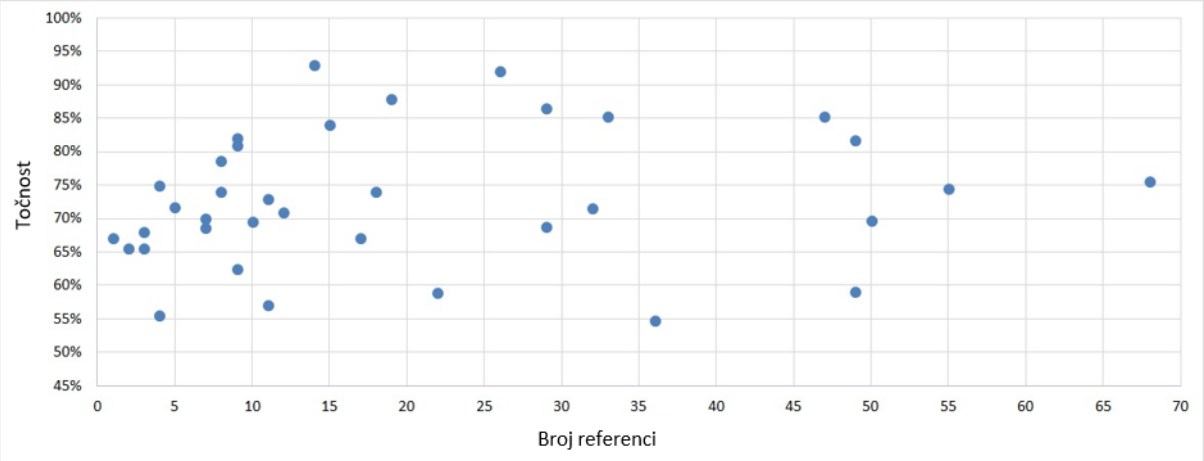
Slika 2.10 prikazuje napredak algoritama strojnog učenja neovisno o analiziranom sportu [9]. Linija trenda, izračunata metodom najmanjih kvadrata, sugerira napredak algoritama strojnog učenja. Napredak je razumljiv s obzirom da istraživači koriste rezultate ostalih istraživača uslijed čega dolaze do novih otkrića, što znači da će nove metode biti još točnije. Valja napomenuti kako linija trenda s godinama bilježi lagani porast točnosti neovisno o analiziranom sportu. Cilj analize bio je prikazati kako napredak u algoritmima predviđanja ishoda u sportu postoji, ali isto tako valja napomenuti kako je gotovo nemoguće očekivati daljnji znatni napredak. Mogućnosti daljnjeg

napretka algoritama predviđanja ishoda u sportu variraju od sporta do sporta, a najviše ovise o ljudskom faktoru koji se prvenstveno odnosi na same igrače, točnije njihovo trenutno raspoloženje i motivaciju, ali i vodstvo kluba koje svojim odlukama može znatno utjecati na konačan ishod, a radi se o stvarima koje algoritam predviđanja ne može znati niti predvidjeti. Slika 2.11 prikazuju napredak rezultata predviđanja po godinama i sportu [9]. Prikazani su najbolji rezultati metoda strojnog učenja u analiziranim radovima. Napredak je vidljiv u količini dostupnih radova. Općenito govoreći, napredak točnosti predloženih modela je evidentan, a najveći problem s usporedbom rezultata je uporaba različitih skupova podataka i liga različitih konkurentnosti. Uzorci za kriket i bejzbol su pak premali da bi se mogli izvući konkretni zaključci. Rezultati vezani uz kriket analiziraju rezultate predviđanja različitih liga što dodatno otežava donošenje zaključaka.



Slika 2.11. Napredak algoritama strojnog učenja vezan uz predviđanje ishoda u a) košarci, b) nogometu, c) američkom nogometu, d) kriketu i e) bejzbolu.

Analizirani radovi koriste različit broj referenci te je potrebno analizirati korelaciju broja referenci i najveće postignute točnosti. Slika 2.12 prikazuje korelaciju broja referenci i najveće postignute točnosti.



Slika 2.12. Ovisnost maksimalne točnosti u odnosu na broj referenci.

Slika 2.12 pokazuje da ne postoji korelacija između maksimalne točnosti korištenog algoritma strojnog učenja i broja referenci [9]. Također nema niti stršećih vrijednosti koje bi mogle dovesti do određenih zaključaka. Vidljivo je da većina radova ima manje od 20 referenci, što je i logično s obzirom da je većina tih radova objavljena u razdoblju od 1996. do 2013. godine.

Sportska predviđanja se obično klasificiraju kao problem u kojem se predviđa jedna klasa (pobjeda/poraz/neriješeno) [34], a rijetki su slučajevi u kojima se predviđa brojčana vrijednost kao što je točan učinak ili razlika u učincima. Točan učinak je točan broj poena momčadi, dok je razlika u učincima razlika učinaka suprotstavljenih momčadi. Predviđanje točnog učinka ili razlike u učincima je svakako izazovniji zadatak od predviđanja ishoda. Predviđanje koje uključuju dva moguća ishoda manje je izazovno od predviđanja koje uključuju tri moguća ishoda. Također, lakše je predvidjeti ishode pojedinačnih sportova u usporedbi s momčadskim sportovima u kojima učinak momčadi ovisi o većem broja igrača. Rezultati radova [24], [35] i [41] su pokazali kako klasifikacijski modeli predviđanja ishoda daju bolje rezultate u odnosu na regresijske modele. Odabir i izlučivanje značajki imaju važnu ulogu u točnosti algoritma strojnog učenja. Cilj oba pristupa je povećati točnost modela, olakšati razumijevanje i vizualizaciju podataka. Teoretski, cilj odabira i izlučivanja značajki je pronaći optimalni podskup značajki (onaj koji će maksimizirati točnost modela). U ovom radu daje se širi pregled literature vezane uz područje predviđanja sportskih ishoda koristeći klasifikaciju, analizu i usporedbu. U tu svrhu analizirano je nekoliko preglednih radova iz srodnih područja ([51], [54], [55], [56] i [57]). Autori rada [54] analizirali su znanstvene radove koji se odnose na predviđanja sportskih ishoda temeljenih na tehnikama dubinske analize podataka te istaknuli prednosti i nedostatke pojedinih modela predviđanja.

Otkrivena su dva glavna problema. Prvi problem vezan uz malu preciznost predviđanja ukazao je na potrebu daljnjih istraživanja kako bi se postigli pouzdani rezultati predviđanja. Drugi problem vezan je uz nemogućnost uspoređivanja rezultata uzrokovan korištenjem različitih skupova podataka. Rad [51] pružio je sveobuhvatan pregled radova koji se bave odabirom značajki. U radu su analizirane metode odabira značajki, uključujući filtar metode (engl. *filter*) kod kojih se odabir značajki vrši na temelju mjere izvedbe bez obzira na korišteni algoritam za modeliranje podataka, metode omotača (engl. *wrapper*) kod kojih se odabir značajki vrši na temelju kvalitete izvedbe algoritma modeliranja, ugrađene metode (engl. *embedded*) koje odabir značajki vrše tijekom izvođenja algoritma modeliranja te hibridne metode. Odabir značajki vrlo je važan korak strojnog učenja i njegova glavna svrha je postizanje boljih performansi smanjenjem skupa značajki. U radu [55] je izvršen sustavni pregled literature korištenja strojnog učenja u bejzbolu. Autori su analizirali 32 članka i otkrili da dva algoritma strojnog učenja dominiraju, a to su stroj s potpornim vektorima (SVM) i metoda najbližih susjeda (k-NN), nevezano radi li se o klasifikaciji ili regresiji. Autori su također zaključili da će neuronske mreže uskoro postati najčešće korišteni model u predviđanjima u bejzbolu. Radom [56] dan je pregled radova koji koriste strojno učenje za predviđanje ishoda u momčadskim sportovima. Krajnji cilj bio je odgovoriti na ključna pitanja vezana uz predviđanja u sportu. Autori su zaključili da većina studija (65 %) koristi umjetne neuronske mreže u svojim istraživanjima, ali i da umjetne neuronske mreže ne moraju nužno pružiti bolje rezultate u odnosu na ostale algoritme strojnog učenja. Također je zaključeno kako su ranija istraživanja odabir značajki radila ručno, obično na temelju znanja istraživača te da veći ulazni skup podataka ne mora nužno dovesti do visoke točnosti. Autori su zaključili i da sportovi s manjim brojem poena i koji uključuju veći broj mogućih ishoda općenito daju manju točnost. Iznesen je i problem uspoređivanja rezultata jer se istraživanja razlikuju u barem jednoj od slijedećih dimenzija: sportu, ulaznim podacima, modelu predviđanja ili razmatranom natjecanju.

Autori rada [57] su primijetili problem korištenja istih modela predviđanja za više sportova. Osim kritičkog pregleda literature vezanog uz upotrebu strojnog učenja za predviđanja u sportu, autori su zaključili da nije moguće izravno primijeniti model predviđanja na različite sportove. Stoga su predložili novu metodologiju i nazvali je „SRP-CRISM-DM“, kao proširenje CRISP-DM metodologije, pomoću koje se strojno učenje može koristiti kao strategija učenja. U trenutnom istraživanju naglasak je stavljen na predviđanje ishoda u momčadskim sportovima.

2.3. Indeksi korisnosti igrača

U današnje vrijeme se velika važnost daje evaluaciji učinka igrača ili momčadi. Analiza koja se bavi bilježenjem pokazatelja izvedbe i uspješnosti je notacijska analiza. Notacijska analiza je

metoda za označavanje događaja tijekom sportskog natjecanja te njihova statistička analiza. Koristi se u procesu sportske pripreme za poboljšanje sportske izvedbe, a temeljem zabilježenih događaja u igri statističkom se analizom dobivaju pokazatelji izvedbe koji ukazuju na tehničko-taktičku aktivnost, odnosno kvalitetu izvedbe igrača, a samim time i cijele momčadi. Notacijska analiza predstavlja objektivan način bilježenja pokazatelja izvedbe i uspješnosti [58].

Najčešći način evaluacije učinka igrača ili momčadi u košarci, osim konačnog rezultata, su indeksi korisnosti. Indeks korisnosti predstavlja metriku kojom se ocjenjuje učinak igrača ili momčadi na utakmici ili određenom vremenskom periodu, gdje viši indeks korisnosti predstavlja bolju izvedbu. Metrika se može definirati kao mjera kvantitativne procjene koja se obično koristi za ocjenjivanje, uspoređivanje ili praćenje uspješnosti procesa. Najkorišteniji, ujedno i najjednostavniji indeksi korisnosti u košarci su NBA i PIR (engl. *Performance Index Rating*) koji se prvenstveno koristi u europskim kupovima. Obično se računaju na temelju jedne utakmice, ali nije isključena upotreba na temelju većeg broja utakmica ili vremenskog perioda. Izračun indeksi NBA i PIR prikazan je formulama (2-1) i (2-2). Tablica 2.3 prikazuje popis kratica. Učinak igrača (N_e) ili momčadi ($N_{tm,e}$) određenog elementa uvijek je cijeli broj ($N_e, N_{tm,e} \geq 0$).

$$I_{NBA} = \left(N_{pts} + N_{rbs} + N_{asist} + N_{st} + N_{bl}\right) - \left(N_{miss\_fg} + N_{miss\_ft} + N_{to}\right) \quad (2\text{-}1)$$

$$I_{PIR} = \left(N_{pts} + N_{rbs} + N_{asist} + N_{st} + N_{bl} + N_{f\_dr}\right) - \left(N_{miss\_fg} + N_{miss\_ft} + N_{to} + N_{bl\_ag} + N_f\right) \quad (2\text{-}2)$$

Tablica 2.3. Popis elemenata i kratica osnovne košarkaške statistike.

Kratica (engleski)	**Značenje**
pts (engl. *points*)	Broj postignutih poena ($N_{pts} = 2 \times N_{2fgm} + 3 \times N_{3fgm} + N_{ftm}$)
$2fgm$ (engl. *2 field goals made*),	Broj pogođenih pokušaja za dva poena (N_{2fgm})
$2fga$ (engl. *2 field goals attempts*)	Broj pokušaja za dva poena (N_{2fga})
$2fgm$ (engl. *3 field goals made*)	Broj pogođenih pokušaja za tri poena (N_{3fgm})
$2fga$ (engl. *3 field goals attempts*)	Broj pokušaja za tri poena (N_{3fga})
ftm (engl. *free throws made*)	Broj pogođenih slobodnih bacanja (N_{ftm})
fta (engl. *free throws attempts*)	Broj pokušaja slobodnog bacanja (N_{fta})
rbs (engl. *rebounds*)	Broj skokova ($N_{rbs} = N_{def\_reb} + N_{of\_reb}$)
def_reb, of_reb (engl. *defensive / ofensive rebounds*)	Broj obrambenih / napadačkih skokova (N_{def_reb}, N_{of_reb})
$asist$ (engl. *assists*)	Broj asistencija (N_{asist})
st (engl. *steals*), to (engl. *turnovers*)	Broj osvojenih lopti (N_{st}), broj izgubljenih lopti (N_{to})
bl (engl. *blocks*), bl_ag (engl. *blocks against*)	Broj postignutih / pretrpljenih blokada (N_{bl}, N_{bl_ag})
f (engl. *fouls*), f_dr (engl. *fouls drawn*)	Broj počinjenih / pretrpljenih prekršaja (N_f, N_{f_dr})
$miss\_2fg$ (engl. *missed 2 field goals*)	Broj promašaja za dva poena ($N_{miss\_2fg} = N_{2fga} - N_{2fgm}$)
$miss\_3fg$ (engl. *missed 3 field goals*)	Broj promašaja za tri poena ($N_{miss\_3fg} = N_{3fga} - N_{3fgm}$)
$miss\_ft$ (engl. *missed free throws*)	Broj promašenih slobodnih bacanja ($N_{miss\_ft} = N_{fta} - N_{ftm}$)

Kompleksniji indeks korisnosti svakako je indeks PER (engl. *Player Efficiency Rating*). Autor indeksa PER je američki novinar John Hollinger. Kao i kod drugih indeksa korisnosti, cilj indeksa PER je prikazati učinak igrača koristeći jednu brojčanu vrijednost.

Navedeni indeksi korisnosti predstavljaju osnovni način evaluacije učinka igrača ili momčadi. Cilj ovog rada je predložiti novi indeks korisnosti koji će u još većoj mjeri moći predočiti stvaran rezultat utakmice, a koji neće biti ograničen brojem elemenata, koji neće imati unaprijed definirane elemente te koji će biti lako prilagodljiv ostalim sportovima ili sličnim procesima. Novo predloženi indeks korisnosti činit će osnovu za predlaganje metode (modela) za predviđanje ishoda sportskih događaja.

2.3.1. Sveobuhvatni indeks korisnosti

U ovom odjeljku će se dati uvod u novonastali indeks korisnosti nazvan sveobuhvatnim indeksom korisnosti (engl. *Comprehensive Efficiency Index* ili kraće CPE). Indeks korisnosti CPE je kumulativan indeks koji se sastoji od niza komponenti ponderiranih koeficijentom W_e, gdje svaka komponenta predstavlja element e promatranog procesa. Opća formula indeksa CPE prikazana je formulom (2-3).

$$I_{CPE} = \sum_{e \in E} W_e I_e, \quad E = skup\ elemenata\ promatranog\ e \tag{2-3}$$

Iako nije obavezno dobro je zadržati fiksni zbroj ponderiranih koeficijenata gdje je $card(E)$ broj elemenata (kardinalnost) skupa E.

$$\sum_{e \in E} W_e = const. = card(E) \tag{2-4}$$

Težinski faktori (W_e) uvedeni su radi općenitosti, tj. kako bi se omogućilo fino podešavanje pojedinih komponenti procesa. Podešavanje se može vršiti prema iskustvu eksperta ili pak korištenjem neke heurističke metode (dubinska analiza podataka, strojno učenje).

Svaki element procesa iz skupa E označen kao e može imati pozitivan i negativan doprinos. Osnovna je ideja indeksa CPE omogućiti korisniku podešavanje doprinosa. Pozitivan učinak komponente označen je oznakom N_e dok je negativni učinak označen oznakom N'_e. Da bi se postigla fleksibilnost, navedeni nenegativni brojevi (N_e i N'_e) množe se s težinskim faktorima v_e i v'_e. Početna formula doprinosa komponente e prikazana je formulom (2-5).

$$I_e = v_e N_e - v'_e N'_e, \quad v_e, v'_e \geq 0, N_e, N'_e \geq 0 \tag{2-5}$$

Kao što je vidljivo iz formule (2-5) indeks CPE omogućuje fino podešavanje doprinosa igre, a kao što je ranije navedeno, podešavanje se može vršiti prema iskustvu eksperta ili korištenjem neke od heurističkih metoda.

2.3.1.1. Koeficijent v (v')

Koeficijent $v_e(v'_e)$ omogućuje preciznije definiranje nagrađivanja pozitivnih i kažnjavanja negativnih doprinosa elemenata promatranog procesa. Tako će se funkcija izračuna koeficijenta $v_e(v'_e)$ prikazati jednostavnom nepadajućom funkcijom ovisnom o parametru $u_e \geq 0$ ($u'_e \geq 0$). Vrijednosti parametra u_e (u'_e) mogu biti odabrane na temelju iskustva korisnika ili korištenjem heurističkih metoda.

Funkcija izračuna koeficijenta $v_e(v'_e)$ je linearno ovisna o parametru $u_e(u'_e)$, te treba zadovoljiti četiri uvjeta koji se odnose na pozitivni i negativni doprinos:

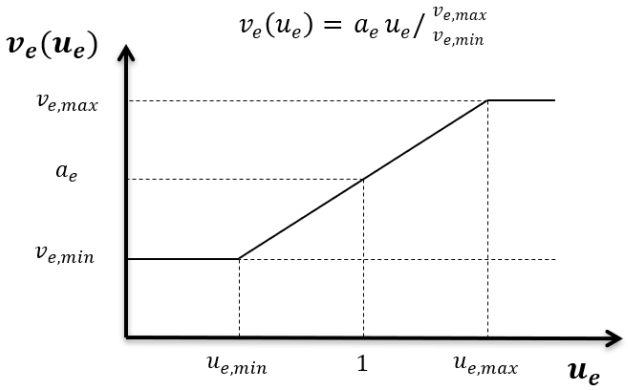
i. $v_e = v_e(u_e) \geq 0$
ii. funkcija v_e je nepadajuća funkcija
iii. kada je $u_e = 1$ potrebno je definirati parametar a_e koji će doprinijeti da je $v_e(1) = a_e \geq 0$
iv. funkcija $v_e = v_e(u_e)$ biti će limitirana parametrima $v_{e,min}$ i $v_{e,\max}$.

Funkcija koja zadovoljava gornje uvjete može se zapisati na dva načina:

$$v_e(u_e)\begin{cases} v_{e,min}, & u_e < u_{e,min} = v_{e,max}/v_e(1); \\ a_e u_e, & u_{e,min} \leq u_e \leq u_{e,max}; \\ v_{e,max}, & u_e > u_{e,max} = v_{e,max}/v_e(1). \end{cases}$$

$$v_e(u_e) = a_e u_e /{}^{v_{e,max}}_{v_{e,min}}$$

U ovom slučaju prikazan je primjer korištenja linearne funkcije izračuna koeficijenta $v_e(v'_e)$, a ovisno o vrsti promatranog procesa moguće je koristiti i sigmoidne funkcije. Slika 2.13. prikazuje linearnu ovisnost koeficijenta $v_e(v'_e)$ o parametru $u_e(u'_e)$ na doprinos komponente e. Kada je parametar $u_e = 1$, parametar v_e poprima vrijednost parametra a_e. Također, vidljiva je ovisnost koeficijenta $v_e(v'_e)$ u odnosu na parametre $v_{e,min}$ i $v_{e,max}$, a samim time i parametre $u_{e,min}$ i $u_{e,max}$.



Slika 2.13. Prikaz doprinosa komponente indeksa CPE.

Slika 2.13 prikazuje doprinos komponente indeksa CPE gdje je središnji dio funkcije segment pravca $v_e(u_e) = a_e u_e$. Funkcija je s gornje i donje strane ograničena horizontalnim pravcima $v_e(u_e) = v_{e,min}$ i $v_e(u_e) = v_{e,max}$.

Uzevši u obzir sva pravila i posebnosti indeksa CPE, doprinos komponente e je zapisan formulom (2-6).

$$I_e = a_e u_e / {}^{v_{e,max}}_{v_{e,min}} \times N_e - a'_e u'_e / {}^{v'_{e,max}}_{v'_{e,min}} \times N'_e \qquad (2\text{-}6)$$

Broj komponenti procesa ovisi o vrsti i složenosti samog procesa, a svaka pojedina komponenta može i ne mora imati pozitivni i/ili negativni doprinos.

2.3.1.2. Koeficijent u_e

Parametar u_e može poprimiti razne vrijednosti, dio predloženih od strane eksperta, proizvoljne vrijednosti definirane od strane korisnika ili pak vrijednosti dobivene heurističkom metodom. Tablica 2.4 prikazuje mogućnosti odabira parametra $u_e(u'_e)$ ovisnog o broju generatora komponenti.

Tablica 2.4. Mogućnosti odabira koeficijenta $u_e(u'_e)$ ovisan o broju generatora komponenti.

#	u_e (≥ 0)	
0	$c \geq 0$	*Korisnički definirana vrijednost*
1a	$\bar{N}_e/\bar{N}_{tm,e}, \bar{N}_{tm,e} > 0$	***Posebni slučajevi:*** Ako su $\bar{N}_e$ ili $\bar{N}_{tm,e}$ nedefinirani tada je $\bar{N}_e/\bar{N}_{tm,e} = 1$ inače ako je $\bar{N}_{tm,e} = 0 \Longrightarrow \bar{N}_e = 0$ tada je $\bar{N}_e/\bar{N}_{tm,e} = 1$
1b	$\bar{N}_{tm,e}/\bar{N}_e, \bar{N}_e > 0$	***Posebni slučajevi:*** Ako su $\bar{N}_e$ ili $\bar{N}_{tm,e}$ nedefinirani tada je $\bar{N}_e/\bar{N}_{tm,e} = 1$. inače ako je $\bar{N}_e = 0 \Longrightarrow u_e \to \infty$ tada je $\bar{N}_{tm,e}/\bar{N}_e = v_{e,max}$
2a	$N_e/\bar{N}_e, \bar{N}_e > 0$	***Posebni slučajevi:*** Ako su N_e ili $\bar{N}_e$ nedefinirani tada je $N_e/\bar{N}_e = 1$. inače ako je $\bar{N}_e = 0 \Longrightarrow u_e \to \infty$ tada je $N_e/\bar{N}_e = v_{e,max}$
2b	$\bar{N}_e/N_e, N_e > 0$	***Posebni slučajevi:*** Ako su $\bar{N}_e$ ili N_e nedefinirani tada je $\bar{N}_e/N_e = 1$. inače ako je $N_e = 0 \Longrightarrow u_e \to \infty$ tada je $\bar{N}_e/N_e = v_{e,max}$

Indeks korisnosti CPE je definiran kao kumulativan indeks koji se sastoji od niza komponenti ponderiranih koeficijentom W_e, gdje svaka komponenta predstavlja element e promatranog procesa. Konačan ishod procesa može definirati jedan ili više generatora komponenti. U slučaju da više generatora komponenti definira konačni ishod procesa moguće je koristiti razne mogućnosti koeficijenta u_e. U slučaju da konačni ishod definira jedan generator komponenti koeficijent u_e predlaže se slučaj 0. Slučaj 0 definira konstantnu vrijednost $c \geq 0$ definiranu na temelju iskustva eksperta ili neke od heurističkih metoda. Ostale varijante mogu se koristiti u slučaju kada više generatora komponenti definira konačni ishod procesa. Slučaj 1a definira omjer

prosječne vrijednosti elementa e u zadanom vremenskom periodu Δt ($\overline{N}_e$) i odgovarajućoj vrijednosti elementa e svih generatora komponenti tijekom istog vremenskog perioda ($\overline{N}_{tm,e}$). $\overline{N}_{tm,e}$ i $\overline{N}_e$ su nedefinirani ukoliko ne postoji poznati ishod procesa na temelju svih generatora komponenti ili točno određenog generatora komponenti u definiranom vremenskom periodu (Δt). Slučaj 2a definira učinak generatora komponente, točnije učinak generatora komponenti promatranog procesa i prosječnog učinka istog generatora komponenti u definiranom vremenskom periodu za element e promatranog procesa. Posebni slučajevi su definirani u tablici. Slučajevi 1b i 2b su recipročne vrijednosti slučajeva 1a i 2a.

Vremenski period definiranja prosječnih vrijednosti moguće je definirati na dva načina ovisno o duljini analiziranog vremenskog perioda. Prvi način definiranja prosječnih vrijednosti definira slučaj u kojem se ne uključuje trenutno promatrani proces. U tom slučaju prosječne vrijednosti parametara $\overline{N}_e$ i $\overline{N}_{tm,e}$ nisu poznate ukoliko ne postoji povijest izvođenja promatranog procesa. Drugi način definiranja prosječnih vrijednosti definira slučaj u kojem se uključuje i ishod promatranog procesa. U tom slučaju parametri $\overline{N}_e$ i $\overline{N}_{tm,e}$ su uvijek definirani, čak i u slučaju kada vremenski period (Δt) uključuje ishod samo promatranog procesa te će u tom slučaju $\overline{N}_e = N_e$.

Za predviđanje ishoda potrebno je koristiti prvi način, točnije način koji ne uključuje trenutni proces. Drugi način definiranja prosječnih vrijednosti se može prvenstveno koristiti za analizu.

2.4. NBA liga

Košarka spada među najpoznatije sportove na svijetu, a američka profesionalna košarkaška liga (NBA) najpopularnija je košarkaška liga na svijetu, ali i jedna od najpopularnijih svjetskih sportskih liga. Košarka je sport u kojem se dvije suprotstavljene momčadi s po pet igrača na terenu i najviše sedam igrača na klupi za rezerve bore postići što veći broj poena. Igra se na označenom pravokutnom terenu s dva koša koji se nalaze na suprotnim stranama terena. Cilj je postići više poena od protivničke momčadi, a moguća su dva konačna ishoda. U slučaju neriješenog rezultata nakon regularnog dijela, igraju se produžeci do trenutka kada jedna od momčadi na kraju produžetka ne postigne više poena. Pobjednikom se proglašava momčad koja je postigla više poena.

NBA liga broji 30 klubova, 29 iz Sjedinjenih Američkih Država i jedan iz Kanade, podijeljenih u dvije konferencije. Liga je osnovana 6. lipnja 1946. u New Yorku, a trenutno ime, National Basketball Association (NBA), dobila je 1949. godine spajanjem sa suparničkom ligom National Basketball League (NBL). Natjecateljska sezona se dijeli na regularni dio i doigravanje tijekom

kojih momčad odigra 82 utakmice regularnog dijela i eventualne utakmice doigravanje. Osim podjele na Istočnu i Zapadnu konferenciju koju čini po 15 momčadi, svaka konferencija se dijeli još i na divizije po pet klubova. Jedna momčad u regularnom dijelu sezone igra četiri puta sa svakim klubom iz divizije, tri do četiri utakmice s klubovima iz preostale dvije divizije iste konferencije te dvije utakmice s klubovima iz druge konferencije. Nakon regularnog dijela slijedi doigravanje (engl. *playoff*) u kojem sudjeluje osam najbolje plasiranih momčadi konferencija. Sistemom ispadanja, igrom na četiri pobjede u svakom krugu, proglašava se konačni pobjednik. Finale igraju najbolja momčad Istočne i najbolja momčad Zapadne konferencije.

Vrlo je važno iznijeti i činjenice vezane uz NBA ligu koje se tiču kompleksnosti predviđanja ishoda u odnosu na ostale sportske lige. Navedene činjenice će pokazati kako predviđanje ishoda u NBA ligi nije nimalo jednostavan proces te da je NBA liga specifična u odnosu na većinu neameričkih profesionalnih liga. Također, pojedine činjenice vezane uz kompleksnost predviđanja će biti produkt dugogodišnjeg iskustva praćenja, analiziranja i promišljanja, komunikacije sa sportskim djelatnicima i zaljubljenicima u sport te proučavanja prikladne literature, najčešće novinskih članaka. Iskustva eksperta vrlo su bitna kod predviđanja ishoda bilo kojeg procesa, a tiču se specifičnih znanja i spoznaja vezanih uz nemjerljive ili teško mjerljive elemente analiziranog procesa. U nastavku će se iznijeti činjenice vezane uz predviđanje u NBA ligi u odnosu na ostale sportove i pripadajuće profesionalne lige.

1. Momčadi u NBA ligi odigraju znatno veći broj utakmica u odnosu na većinu ostalih sportskih liga. Regularni dio NBA lige počinje krajem listopada, a završava krajem travnja. Tijekom regularnog dijela sezone svaka momčad odigra točno 82 utakmice. Nakon toga slijedi doigravanje koje završava početkom lipnja. Tablica 2.5 prikazuje broj utakmica NBA momčadi koje su ušle u finale po analiziranim sezonama.

Tablica 2.5. Broj utakmica finalisti po NBA sezonama.

Sezona	**Pobjednik**	**Broj utakmica**	**Finalist**	**Broj utakmica**
2009./2010.	LA Lakers	105	Boston Celtics	106
2010./2011.	Dallas Maverics	103	Miami Heat	103
2011./2012.*	Miami Heat	89	Oklahoma City Thunder	86
2012./2013.	Miami Heat	105	San Antonio Spurs	103
2013./2014.	San Antonio Spurs	105	Miami Heat	102
2014./2015.	Golden State Warriors	103	Cleveland Cavaliers	102
2015./2016.	Cleveland Cavaliers	103	Golden State Warriors	106
2016./2017.	Golden State Warriors	99	Cleveland Cavaliers	100
2017./2018.	Golden State Warriors	103	Cleveland Cavaliers	104
	Prosjek	103,25 (101,67*)	**Prosjek**	103,25 (101,33*)

Iz tablice je vidljivo da finalisti NBA lige odigraju u prosjeku 103 utakmice u nepunih osam mjeseci, od toga 82 utakmica do sredine travnja. Primjerice, u najpoznatijoj nacionalnoj nogometnoj ligi, engleskoj Premier liga, svaka momčadi odigra 38 utakmica, u pravilu jednu utakmicu tjedno. U većini ostalih liga, neovisno o formatu natjecanja, odigra se čak i manje utakmica, prvenstveno jer se lige sastoje od manjeg broja momčadi. Prvak europske Lige prvaka, najprestižnijeg nogometnog natjecanje na svijetu, odigra 17 utakmica plus eventualne dodatne utakmice kvalifikacija. U pravilu momčadi koje igraju finale europske Lige prvaka prošlogodišnjim plasmanom ulaze direktno u grupnu fazu natjecanja te samim time ne igraju kvalifikacije. Nacionalna nogometna prvenstva nude i Kup utakmice u kojima prvak nacionalnog Kupa odigra 10-12 utakmica, ovisno o broju prijavljenih momčadi i formatu natjecanja. Nogometne momčadi s najviše odigranih utakmica odigraju 70-ak utakmica raspoređenih u najmanje devet mjeseci. Analizirajući košarkaške lige, situacija je za većinu momčadi još i povoljnija, međutim najjače europske momčadi imaju ritam sličan NBA ligi te odigraju 20-ak utakmica manje. Njima u prilog ide početak natjecanja već polovicom rujna te završetak sredinom svibnja što znači da im službeni dio sezone traje minimalno mjesec dana duže. Također, konkurentnost liga u kojima oni nastupaju je znatno manja u odnosu na NBA ligu. Razlozi tome će biti objašnjeni u idućoj točki.

2. Većina sportskih liga ima otvoreno tržište i sukladno financijskoj moći kupuje/prodaje igrače. Profesionalne američke lige i nekolicina kanadskih neovisno o financijskoj moći izlaze na tzv. draft. Draft se održava jednom godišnje te predstavlja jedini način da mladi i perspektivni igrači igraju u NBA ligi. U slučaju da igrač napusti ligu u istu se može vratiti bez ponovnog drafta. Momčad s najlošijim omjerom prethodne sezone ima najveću vjerojatnost da na kraju sezone prva bira, a redoslijed odabira igrača se određuje lutrijom (izvlačenjem kuglica). Draft je podijeljen u dva kruga, a za pravo prvog izbora se natječu momčadi koje se nisu plasirali u doigravanje. Ovim načinom daje se prilika slabijim momčadi da ojačaju svoje momčadi i postanu konkurentnije.
3. NBA i ostale američke profesionalne lige (NFL, NHL, MLB, MLS…) uvele su tzv. *salary cap*, često nazvan i financijski fair-play, ograničenje u količini novca koji momčad može potrošiti na plaće igrača. Definirana je maksimalna plaća igrača i cjelokupni iznos koji momčad smije potrošiti na plaće igrača. Cilj navedenog ograničenja je povećanje konkurentnosti unutar lige. Ostale sportske lige nemaju ograničenja u iznosu plaća igrača te samim time ne postoje nikakve ograničenja.

4. U NBA ligi su uvedeni i izrazi kao što su „tanking“, „rebuilding“ ili pak ograničenje u minutama u igri. „Tanking“ je izraz koji se koristi u slučajevima kada momčad ne napravi sve da dobije utakmicu, a najčešće u svrhu dobivanja bolje pozicije na draftu. Uobičajeno je ovdje riječ o periodu kada momčad više matematički ne može osigurati doigravanje, a rjeđi su slučajevi kada momčad cijelu sezonu igra na navedeni način. U slučaju kada momčad jednu ili više sezona ne igra maksimalnim naporom je uglavnom riječ o „rebuildingu“. „Rebuilding“ ili obnova momčadi je pojam koji označava stvaranje nove momčadi uglavnom sastavljene od mladih igrača. Fazu obnove momčadi karakterizira velik broj mladih igrača i pojačani fond sati treninga. NBA momčadi zbog velikog broja utakmica koje znaju biti dan za danom i dugih putovanja avionom ne stignu trenirati kao momčadi u ligama s manje utakmica. Na taj način, točnije u fazi obnove, veći je fokus na trening i svjesno gubljenje utakmica koje donosi i bolju poziciju na draftu. Posljednji izraz „limited minutes“ ili ograničenje u minutama se odnosi na igrače koji su često i nositelji igre, ali zbog velikog broja utakmica i preventive od ozljede ili faze povratka nakon ozljede igraju po definiranom planu igre (engl. *playbook*). U pravilu se radi o igračima povratnicima nakon ozljede, ali vrijedi i za određene igrače u godinama koji svojom kvalitetom zavređuju više minuta na terenu. Manji broj utakmica i veći broj dana odmora igrači imaju tijekom doigravanja, međutim broj utakmica doigravanja je svega 6,5 %.

Navedene razloge bitno je izložiti iz razloga što su NBA liga i nekolicina profesionalnih sportskih liga SAD-a i Kanade po mnogočemu specifične u odnosu na ostale sportske lige. Fokus ovog rada će biti na upotrebi strojnog učenja u predviđanju sportskih ishoda, stoga je bilo potrebno navesti činjenice zbog kojih je predviđanje ishoda u NBA ligi kompleksnije od predviđanja ishoda u većini ostalih liga.

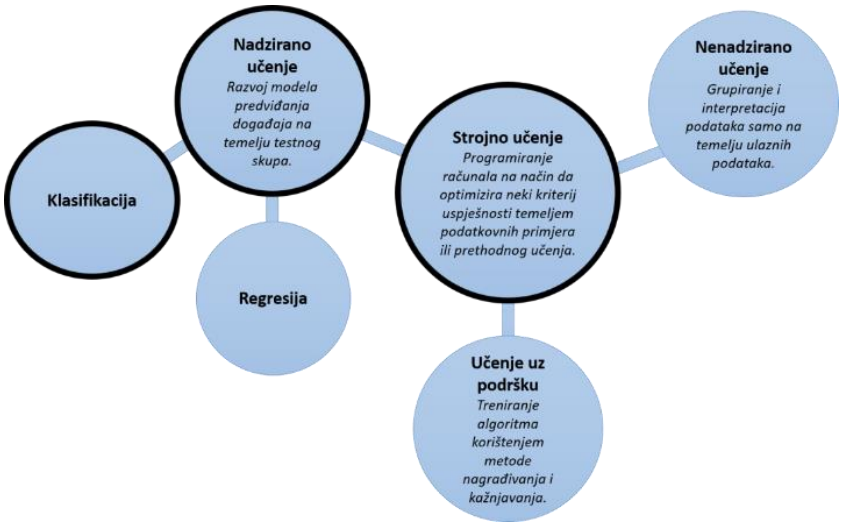
2.5. Strojno učenje

Zadatak strojnog učenja je predvidjeti ishod procesa na temelju dostupnih podataka. Jednu od prvih definicija strojnog učenja iznio je začetnik strojnog učenja Arthur Samuela (1959.) koji je strojno učenje definirao granom računalne znanosti koja daje računalu sposobnost učenja bez da je ponašanje računalnog procesa eksplicitno programirano[59]. Novija definicija strojnog učenja govori da je strojno učenje programiranje računala na način da optimizira kriterij uspješnosti temeljem podatkovnih primjera ili prethodnog učenja [60]. Ostale definicije strojnog učenja su i da se polje strojnog učenja bavi pitanjem kako konstruirati računalni program koji će se automatski unaprijediti s iskustvom [61]. Autori rada [62] pak strojno učenje definiraju sa statističke točke gledanja stavljajući podatke u centar pozornosti, definirajući strojno učenje kao postupak učenja

iz podataka. Vrlo zanimljivu misao izrekao je 1982. godine John Naisbitt u jednoj od svojih knjiga rekavši da se utapamo u informacijama, a gladujemo za znanjem. Cilj strojnog učenja je izgraditi model koji će biti dobra i korisna aproksimacija podataka, točnije izraditi model koji će na temelju poznatih podataka moći predvidjeti svojstva novih, još neviđenih podataka.

Strojno učenje se dijeli na nadzirano (engl. *supervised learning*), nenadzirano (engl. *unsupervised learning*) i učenje uz podršku (engl. *reinforcement learning*) [63], dok pojedini autori navode još i podjelu na polu-nadzirano učenje, transduktivno učenje, relacijsko učenje i genetsko programiranje [64]. Nadzirano učenje daje eksplicitnu informaciju o primjerima i vrijednostima ciljne varijable s konačnim ciljem izgradnje modela koji će na temelju poznatog skupa podataka vršiti predviđanja na još neviđenim podacima. Produkt nadziranog strojnog učenja je sustav koji generalizira odgovore na sve moguće ulazne podatke. Matematički rečeno, program koji uči dobiva skup ulaznih podataka $(x_1, x_2, x_3, \dots, x_n)$ i skup željenih vrijednosti takvih da za svaki ulazni podatak x_i vraća izlaz y_i. Zadatak nadziranog učenja je da „nauči“ kako da na novom, neobilježenom ulaznom skupu podataka predvidi točnu izlaznu vrijednost. Kod nenadziranog učenja postoje samo primjeri bez ikakve anotacije ili povratne informacije, a cilj je grupirati primjere te pronaći strukturnu pravilnost među podacima. Matematički rečeno, dobiva se samo skup ulaznih podataka $(x_1, x_2, x_3, \dots, x_n)$, a zadatak nenadziranog učenja otkriti je skrivene zakonitosti u podacima te na temelju njih definirati izlazne vrijednosti. Predviđanje korištenjem nenadziranog učenja vrši se grupiranjem (engl. *clustering*), procjenom gustoće (engl. *denstity estimation*) ili smanjenjem dimenzionalnosti (engl. *dimensionality reduction*) [65]. Učenje uz podršku je pak vrsta strojnog učenja koje trenira algoritam koristeći metode nagrađivanja i kažnjavanja, točnije dobivanje nagrade se odgađa do trenutka kada je konačni ishod poznat, a samim time i uspješnost algoritma. Učenje uz podršku je moguće definirati i kao učenje optimalne strategije na temelju pokušaja s odgođenom nagradom [65].

Područje strojnog učenja usko je povezano s računalnom statistikom koja je pak usko povezana s teorijom vjerojatnosti, linearnom algebrom, teorijom informacija te kognitivnim znanostima. Primjena strojnog učenja je široka. Strojno učenje se prvenstveno koristi za rješavanje složenih problema, točnije problema za koje ne postoji ljudsko znanje ili ljudi ne mogu objasniti proces. Strojno učenje se koristi i kod sustava koji se dinamički mijenjaju, kao i kod velikih količina podataka u svrhu otkrivanja znanja. Cilj strojnog učenja se može definirati i kao izgradnja modela koji objašnjavaju podatke, ali i omogućuju predviđanje ili zaključivanje. Slika 2.14 prikazuje podjelu na tri osnovne vrste strojnog učenja, a podebljanom linijom kružnica označena su područja koja će se koristiti u izradi ovog rada.



Slika 2.14. Podjela osnovnih metoda strojnog učenja.

U ovom istraživanju predstavit će se algoritam predviđanja ishoda košarkaške utakmice koji će se moći uz minimalne preinake koristiti za bilo koji drugi momčadski sport ili sličan proces. Osnovna je ideja algoritma koristiti osnovne elemente košarkaške statistike te ih uvesti u predloženi indeks korisnosti igrača. Budući da će se za treniranje sustava koristiti povijesni podaci poznatog ishoda, koristit će se nadzirano strojno učenje. Nadzirano strojno učenje se po vrsti dijeli na klasifikaciju i regresiju. Eksplicitna informacija kod klasifikacije je kategorijska (npr. ishod utakmice je pobjeda ili poraz) dok je kod regresije eksplicitna informacija numerička vrijednost (realan broj), kao što je predviđanje točnog rezultata ili razlike u poenima između dvije suprotstavljene momčadi. Predviđanje točnog rezultata svakako predstavlja puno izazovniji zadatak od predviđanja konačnog ishoda. Ukratko, ciljna varijabla kod klasifikacije je diskretna ili nominalna, a kod regresije kontinuirana. Sportska predviđanja obično se tretiraju kao klasifikacijski problemi u kojima se predviđa jedna klasa (pobjeda, poraz ili u nekim sportovima i neriješeno), a rijetki su slučajevi kad se predviđaju brojčane vrijednosti [34]. Rezultati radova [24], [35] i [41] su pokazali da klasifikacijski modeli predviđanja ishoda u sportu daju bolje rezultate od regresijskih modela. Matematički gledano cilj strojnog učenja postaviti je hipotezu h, odnosno funkciju koja primjerima dodjeljuje oznake, $h : ulazni\ skup \rightarrow izlazni\ skup$. U slučaju predviđanja košarkaške utakmice će se koristiti binarna klasifikacija, $h : ulazni\ skup \rightarrow \{0, 1\}$, gdje 0 označava pobjedu domaćina, a 1 pobjedu gosta. Neželjena pojava strojnog učenja je šum. Šum se može definirati kao neželjena anomalija među podacima, a uzrok šumu mogu biti nepreciznost pogreške u označavanju, nedostajuće vrijednosti ili pak subjektivnost.

Odabir značajki i izlučivanje značajki (u literaturi ponekad nazvano i transformacijom značajki) također doprinose povećanju točnosti modela predviđanja, a indirektno i lakšem razumijevanju

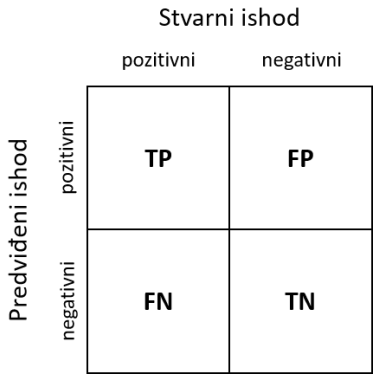
podataka te kao pomoć kod vizualizacije podataka. Glavni cilj odabira i izlučivanja značajki je pronaći optimalni podskup podataka koji maksimizira mogućnost predviđanja.

Strojno učenje se uglavnom koristi u situacijama kada je problem presložen da bi se riješio algoritamski, odnosno kada ne postoji ideja kako riješiti problem, u slučajevima kada se sustavi dinamički mijenjaju ili pak kada se koriste velike količine podataka.

Jedna od metrika definiranja uspješnosti modela strojnog učenja je točnost (engl. *accuracy*) prikazana formulom (2-7).

$$točnost = \frac{broj\ točnih\ predviđanja}{ukupan\ broj\ predviđanja} \tag{2-7}$$

Točnost se može definirati kao udio točno klasificiranih primjera u skupu svih primjera. Pošto će se u ovom istraživanju predviđati dva ishoda koristit će se binarna klasifikacija koju je moguće iskazati matricom zabune (engl. *confusion matrix*). Učenje binarne klasifikacije istovjetno je učenju Boolove funkcije. Matrica zabune prikazuje točnost klasifikatora usporedbom stvarnih i predviđenih klasa [66]. Slika 2.15 prikazuje matricu zabune za binarnu klasifikaciju koja se sastoji od četiri kombinacije stvarnih i previđenih ishoda.



Slika 2.15. Matrica zabune binarne klasifikacije.

Točnost predviđanja, prikazana formulom (2-8), se također može prikazati prema pozitivnim i negativnim previđenim i stvarnim ishodima.

$$točnost = \frac{TP + TN}{TP + TN + FP + FN} \tag{2-8}$$

U formuli (2-8) TP predstavlja točno pozitivnu (engl. *true positive*) klasu (ishod u kojem model predviđa pozitivnu klasu), TN predstavlja točno negativnu (engl. *true negative*) klasu (ishod u kojem model predstavlja točno predviđenu negativnu klasu), FP predstavlja netočno pozitivnu (engl. *false positive*) klasu (ishod gdje model predviđa pozitivnu klasu) te FN odnosno netočno negativnu (engl. *false negative*) klasu (ishod gdje model netočno predviđa negativnu klasu. Evaluacija predloženog modela će se vršiti na košarkaškim utakmica. Tablica 2.6 predstavlja matricu zabune vezanu uz problematiku istraživanja.

Tablica 2.6. Matrica zabune vezana uz problematiku predviđanja ishoda u sportu.

Točno pozitivno (TP): Stvarni ishod: pobjeda domaće momčadi Predviđeni ishod: pobjeda domaće momčadi	**Netočno pozitivno (FP):** Stvarni ishod: pobjeda gostujuće momčadi Predviđeni ishod: pobjeda domaće momčadi
Netočno negativno (FN): Stvarni ishod: pobjeda domaće momčadi Predviđeni ishod: pobjeda gostujuće momčadi	**Točno negativno (TN):** Stvarni ishod: pobjeda gostujuće momčadi Predviđeni ishod: pobjeda gostujuće momčadi

Osim točnosti, postoje i neke druge evaluacijske mjere kao što su preciznost (engl. *precision*) prikazana formulom (2-9), odziv (engl. *recall*) prikazan formulom (2-10) i specifičnost (engl. *specificity*) prikazana formulom (2-11). Preciznost predstavlja udio točno klasificiranih primjera u skupu pozitivno klasificiranih primjera, a još se naziva i *positive predictive value* ili skraćeno PPV.

$$preciznost = \frac{TP}{TP + FP} \tag{2-9}$$

Odziv predstavlja udio točno klasificiranih primjera u skupu svih pozitivnih primjera, a još se naziva i *true positive rate* ili skraćeno TPR.

$$odziv = \frac{TP}{TP + FN} \tag{2-10}$$

Specifičnost predstavlja udio točno klasificiranih primjera u skupu svih negativnih primjera, a još se naziva i *true negative rate* ili skraćeno TNR.

$$specifičnost = \frac{TN}{TN + FP} \tag{2-11}$$

Različite evaluacijske mjere se koriste u različitim područjima, a sve zbog raspodjele pozitivnih i negativnih primjera. U trenutnom istraživanju će se kao mjera evaluacije koristiti isključivo točnost.

2.6. Validacija modela

Cilj modela predviđanja je minimalan broj grešaka ($\min e(h,T)$) i ne prevelika složenost($C(h)$), a cilj validacije modela je procjena stvarne pogreške predloženog modela. Točnije, potrebno je definirati procedure za pretraživanje prostora mogućih modela ($h \in H$) koje vode računa o greškama ($e(h)$) i složenosti modela $C(h)$).

$$\min(e(h,T) + \alpha C(h) \tag{2-12}$$

Validacija modela je procjena stvarne pogreške modela. Važno je napomenuti da cilj modela učenja nije dobro klasificirati primjere za učenje jer je klasifikacija takvih primjera poznata, već što točnije klasificirati nepoznate primjere, tj. omogućiti generaliziranje. Poželjna svojstva modela svakako su da je model jednostavan što znači i lakše korištenje, ali i manju računalnu složenost.

Jednostavan model lakše je naučiti i tumačiti, a samim time i ekstrahirati znanje. Jednostavnost modela često se opisuje Occamovom britvom koja govori da ako jednu pojavu objašnjavaju dvije podjednako dobre teorije, vjerojatnije je istinita ona koja je jednostavnija. Teorija se primjenjuje kod razvoja teorijskih modela, ali se njome ne presuđuje pri odabiru različitih modela, već je potrebno pokusom odabrati bolji model. Kod učenja modela definirane su dvije krajnosti, prenaučenost ili pretreniranost (engl. *overfitting*) i podnaučenost ili podtreniranost (engl. *underfitting*).

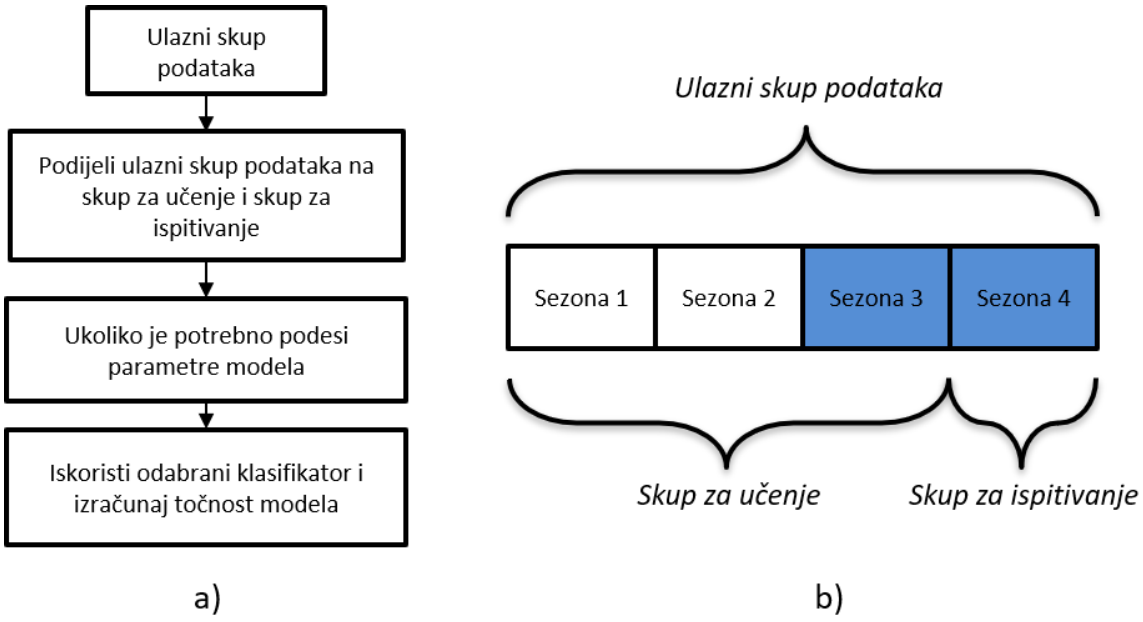
Postoje dvije osnovne tehnike probiranja podataka (engl. *sampling-resampling*), odnosno dvije tehnike validacije modela, a to su metoda podjele skupa podataka i unakrsna provjera (engl. *cross validation*).

2.6.1. Metoda podjele skupa podataka

Metoda podjele skupa podataka dijeli ulazni skup podataka ($\mathcal{D}_{skup\ podataka}$ ili kraće $\mathcal{D}$) na dva ili tri različita, ali ne nužno kronološki poredana skupa podataka, skup za učenje (engl. *training set*), skup za provjeru (engl. *validation set*) i skup za ispitivanje (engl. *test set*). Skup za provjeru ($\mathcal{D}_P$) nije uvijek korišten, a obično se koristi za završno podešavanje parametara modela. Kod predviđanja sportskih događaja preporuča se korištenje kronološki poredanih skupova podataka jer sportski događaji nisu u potpunosti neovisni događaji. Povijesni podaci mogu pružiti vrlo korisne informacije u predviđanju budućih događaja. Kod metode podjele skupa podataka model uči na skupu za učenje, a procjena modela se vrši testirajući model na skupu za učenje. Skup za učenje se sastoji od primjera i pripadnih oznaka, $\mathcal{D}_U = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, gdje N predstavlja ukupan broj primjera za učenje. Skup za ispitivanje, $\mathcal{D}_I$, ne sadrži poznate pripadajuće oznake, u ovom slučaju ishode procesa. Metodu podjele skupa podataka karakterizira disjunktna podjela skupova.

$$\mathcal{D} = \mathcal{D}_U \cup \mathcal{D}_I \tag{2-13}$$

Dobro svojstvo metode podjele skupa podataka je jednostavnost, odnosno odabir modela koji daje najmanju pogrešku na skupu za ispitivanje. Loša strana metode podjele skupa podataka je nepouzdana ocjena greške na malom skupu za učenje. Slika 2.16 prikazuje blok dijagram i grafički prikaz metode podjele skupa podataka kod koje omjer skupa za učenje i skupa za ispitivanje može biti proizvoljan.



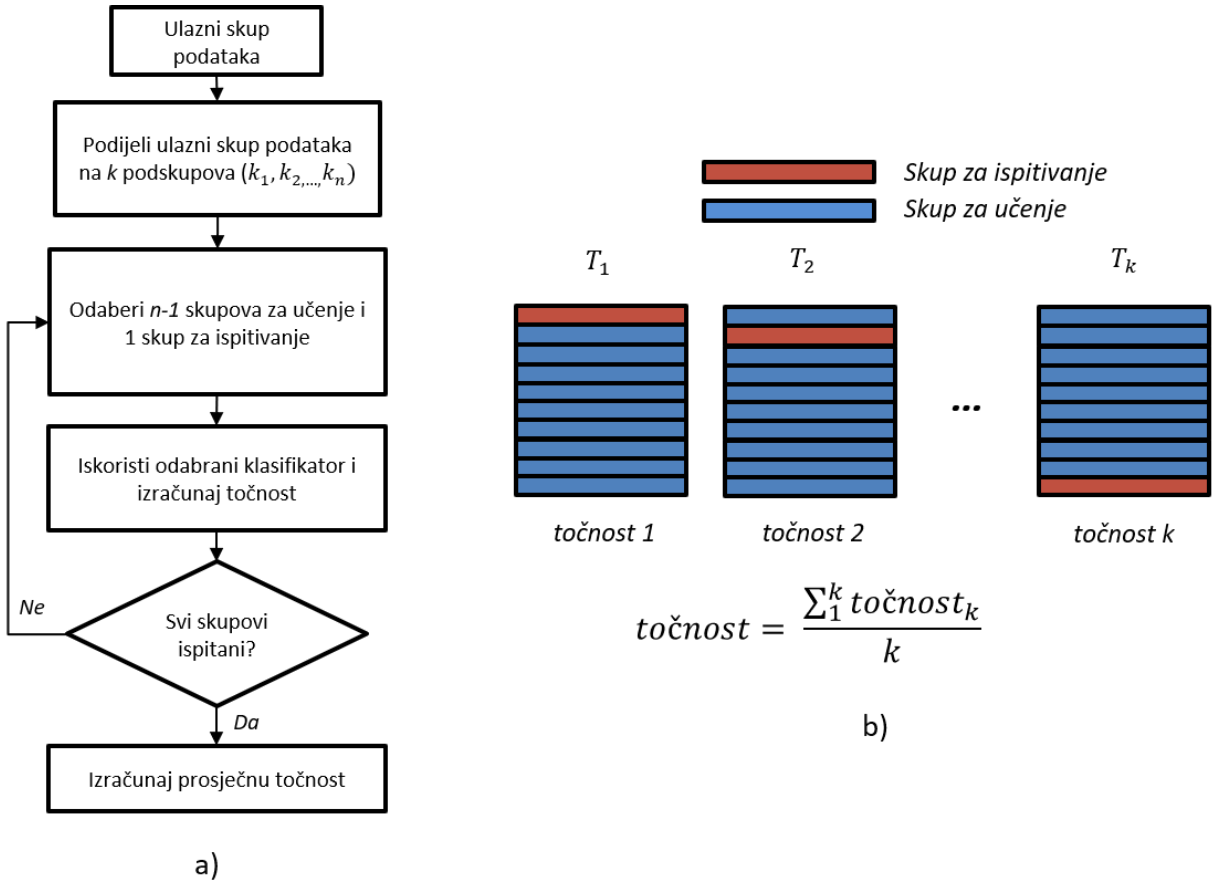
Slika 2.16. Metoda podjele skupa podataka. Slučaj a) prikazuje blok dijagram metode podjele skupa podataka, dok slučaj b) prikazuje grafički prikaz metode podjele skupa podataka.

Metoda podjele skupa podataka jednostavna je za razumijevanje i implementaciju. Jednostavnost razumijevanja i implementacije ne znači nužno i lošije sposobnosti predviđanja. Isto tako, korištenje skupa podataka za provjeru ne znači nužno i bolje rezultate predviđanja. Svaki proces je specifičan te je za odabir metode validacije potrebno dobro razumijevanje procesa.

2.6.2. Unakrsna provjera

Unakrsna provjera početni skup podataka dijeli na k odvojenih skupova T_i $(i = 1, 2, \dots, k)$. Broj odvojenih skupova k nije strogo definiran već ovisi o situaciji, a dobiva se na temelju procjene eksperta ili korištenjem heurističke metode. Jedan od k skupova definira se skupom za ispitivanje, a ostalih $n - 1$ odvojenih skupova skupom za učenje. Postupak se ponavlja k puta te se računa prosječna točnost modela. Metoda unakrsne provjere ima prednost u odnosu na metodu podjele skupa podataka zato što su svi odvojeni skupovi podataka T_i korišteni za učenje, ali i ispitivanje modela. Slika 2.17 prikazuje blok dijagram i grafički prikaz metode unakrsne provjere modela.

Loše strane u odnosu na metodu podjele skupa podataka su da je metoda unakrsne provjere k puta skuplja jer je potrebno učiti k modela, a dobra strana je da se u svakom trenutku gubi samo N/k odvojenih skupova te je procjena greške stabilnija. Unakrsna provjera s druge strane nije pogodna za predviđanje budućih sportskih događaja jer sportski događaji nisu u potpunosti neovisni događaji te ovise o poznatoj prošlosti [57].



Slika 2.17. Unakrsna provjera. Slučaj a) prikazuje blok dijagram metode unakrsne provjere, dok slučaj b) prikazuje grafički prikaz metode unakrsne provjere.

Odabir metode validacije modela predstavlja vrlo važan korak strojnog učenja što znači da je potrebno dobro poznavanje analiziranog procesa. Poseban naglasak potrebno je dati međusobnoj uzročnosti događaja, točnije kako protekli događaji utječu na buduće. U kasnijim poglavljima će se analizirati utjecaj dvije osnovne metode validacije i samim time donijeti odluka koja metoda validacije će biti korištena.

3. POČETNE PRETPOSTAVKE I ANALIZA PODATKOVNOG SKUPA SPORTSKIH DOGAĐAJA

Početak svakog istraživanja je definiranje hipoteza, odnosno pretpostavki koje se želi dokazati ili opovrgnuti. Obrada podataka započinje njihovim vizualnim prikazom, najčešće histogramima ili kutijastim dijagramima. Na taj se način može uočiti u kojim granicama se podaci nalaze te kako su raspoređeni. Cilj ovog poglavlja prikazati je na koji su način ulazni podaci strukturirani te kako će se koristiti u predviđanju ishoda, ali i postaviti početne pretpostavke vezane uz kasnije predloženi model predviđanja. Kada se govori o podatkom skupu potrebno je navesti dva pojma, reprezentativnost i relevantnost, koji uvelike utječu na kvalitetu modela pa posljedično i na mogućnosti primjene. Reprezentativnost govori koliko dobro korišteni skup podataka opisuje problem, dok relevantnost govori koliko je korišteni skup podataka značajan za opis analiziranog problema.

3.1. Podatkovni skupovi sportskih događaja

Osnovna pretpostavka modela predviđanja je postojanje podataka. Priprema podataka predstavlja možda i najteži korak strojnog učenja, a razlog tome je različitost i specifičnost skupa podataka, prvenstveno orijentiranih ka analiziranom problemu. Ipak postoje uobičajeni koraci tijekom faze pripreme podataka. Različiti autori predlažu i različit broj koraka pripreme podatkovnog skupa, međutim neovisno o broju koraka pripreme podatkovnog skupa konačan cilj je jednak. Pripremu podatkovnog skupa se tako može podijeli na nekoliko faza kao što su prikupljanje podataka, čišćenje podataka, identifikacija i izračun nedostajućih vrijednosti, transformacija podataka, podjela podataka na skup za učenje i skup za ispitivanje neovisno o korištenoj metodi validacije te na kraju upotreba samog skupa podataka.

Svaki sportski događaj je moguće opisati skupom statističkih podataka. Posebnost sporta je u tome što svaki sport karakterizira specifična statistika. Točnije, statističke parametre nogometa nije moguće upotrijebiti u košarci, ili pak statističke parametre košarke nije moguće upotrijebiti u bejzbolu itd. Ipak postoji jedan statistički parametar koji je jedinstven za sve sportove, a to je konačan rezultat, nebitno da li je definiran brojem golova, poena, setova… U konačnici, sportski događaji predstavljaju procese s jednakim ciljem, a to je biti uspješniji u onom statističkom parametru koji definira konačni ishod. Predviđanje ishoda na temelju jednog parametra je vrlo teško, stoga je potrebno sportski događaj opisati korištenjem većeg skupa parametara. Bez obzira što je navedeno da svaki sport karakterizira specifičan skup statističkih parametara, postoje i statistički parametri koji se koriste u više sportova, a to se odnosi prvenstveno na loptačke sportove

kao što su nogomet, košarka, rukomet, vaterpolo itd. Analizirani skup statističkih parametara sportskog događaja ne ovisi isključivo o sportu, već ovisi i o natjecanju, državi, ali i mogućnostima organizatora sportskog natjecanja. U ovom odjeljku je pokazano kako skup statističkih parametara varira ne samo od sporta do sporta, već i natjecanja unutar samog sporta.

Cilj ovog rada je predložiti metodu (model) predviđanja sportskih ishoda koji neće biti striktno fokusiran na jedan sport, već će se moći uz sitne preinake prilagoditi bilo kojem, prvenstveno momčadskom sportu. Rezultati modela će biti eksperimentalno ispitani i analizirani na rezultatima u košarci pa će se samim time i fokus biti na analizi košarkaške statistike.

3.1.1. Prikupljanje i predobrada ulaznih podataka

Ulazni skup podataka ovog istraživanja čini 9 uzastopnih NBA sezona, počevši od sezone 2009./2010. pa sve do sezone 2017./2018. Sezona 2011./2012. je specifična zbog četvrtog štrajka igrača (engl. *lockout*) u povijesti NBA lige te je trajala svega 8 mjeseci. Svaka momčad je tako tijekom regularnog dijela umjesto 82 utakmice odigrala 66 utakmica. Ulazni skup podataka čini ukupno 11578 utakmica od čega je 749 (6,5 %) utakmica doigravanja. Tablica 3.1 prikazuje ukupan broj utakmica po sezonama i broj utakmica doigravanja. Zvjezdicom je označena sezona štrajka igrača.

Tablica 3.1. Broj utakmica po analiziranoj sezoni NBA lige.

Sezona	Ukupan broj utakmica	Broj utakmica doigravanja
2009./2010.	1312	82 (6,25 %)
2010./2011.	1311	81 (6,18 %)
2011./2012.*	1074	84 (7,82 %)
2012./2013.	1314	84 (6,39 %)
2013./2014.	1319	89 (6,75 %)
2014./2015.	1311	81 (6,18 %)
2015./2016.	1316	86 (6,53 %)
2016./2017.	1309	79 (6,04 %)
2017./2018.	1312	82 (6,25 %)
Ukupno:	11578	749 (6,5 %)

Osnovna statistika koja će se koristiti u ovom radu se sastoji od 13 elemenata. Postoji još nekoliko elemenata koji se često koriste, kao što su pretrpljeni prekršaji (engl. *fouls drawn*) ili pretrpljene blokade (engl. *blocks against*), međutim u ovom slučaju nisu dostupni iz korištenog izvora podataka. Tablica 3.2 prikazuje popis elemenata osnovne košarkaške statistike koji će se koristiti u istraživanju.

Tablica 3.2. Elementi osnovne košarkaške statistike.

Kratica	Značenje
$2fgm$	Broj zabijenih pokušaja za dva poena (engl. *two field goals made*)
$2fga$	Broj pokušaja za dva poena (engl. *two field goals attempts*)
$3fgm$	Broj zabijenih pokušaja za tri poena (engl. (*three field goals made*)
$3fga$	Broj pokušaja za tri poena (engl. *three field goals attempts*)
ftm	Broj zabijenih slobodnih bacanja (engl. *free throws made*)
fta	Broj pokušaja slobodnog bacanja (engl. *free throws attempts*)
def_reb	Broj obrambenih skokova (engl. *defensive rebounds*)
of_reb	Broj napadačkih skokova (engl. *ofensive rebounds*)
$asist$	Broj asistencija (engl. *asists*)
st	Broj osvojenih lopti (engl. *steals*)
to	Broj izgubljenih lopti (engl. *turnovers*)
bl	Broj postignutih blokada (engl. *blocks*)
f	Broj počinjenih prekršaja (engl. *fouls*)

Osim 13 osnovnih elemenata košarkaške statistike koristit će se i tri izlučena elementa, a to su broj promašenih pokušaja za dva poena prikazan formulom (3-1), broj promašenih pokušaja za tri poena prikazan formulom (3-2) te broj promašenih slobodnih bacanja prikazan formulom (3-3). Izlučene značajke dobivene su na temelju 13 prethodno navedenih elemenata košarkaške statistike.

$$miss\_2fg = 2fga - 2fgm \tag{3-1}$$

$$miss\_3fg = 3fga - 3fgm \tag{3-2}$$

$$miss\_ft = fta - ftm \tag{3-3}$$

Pojedini elementi košarkaške igre se mogu prikazivati u obliku skupnih elemenata. Logične skupine elemenata čine elementi vezani uz pokušaje za dva poena, tri poena i slobodna bacanja, gdje svaki element ima pozitivan i negativan doprinos. Često se grupiraju i promašaji iz igre, prikazani formulom (3-4), koji čine sumu promašenih pokušaja za dva i tri poena.

$$miss\_fg = miss\_2fg + miss\_3fg \tag{3-4}$$

Od ostalih elemenata grupirati se mogu postignute i primljene blokade te načinjeni i pretrpljeni prekršaji. Nadalje, pogođeni pokušaji za dva i tri poena te pogođena slobodna bacanja mogu se grupirati u element nazvan poeni (engl. *points* ili kraće *pts*) prikazan formulom (3-5). Isto vrijedi i za element skokova (engl. *rebounds* ili kraće *rbs*) koji predstavlja sumu obrambenih i napadačkih skokova prikazan formulom (3-6).

$$pts = 2 \times 2fgm + 3 \times 3fgm + ftm \tag{3-5}$$

$$rbs = def\_reb + of\_reb \tag{3-6}$$

3.2. Identifikacija značajki

Univerzalna ili točna definicija značajke ne postoji, što znači da definicija značajke ovisi o samom problemu. Zadovoljavajuća definicija značajke (engl. *feature*) može biti da je značajka numerički prikaz neobrađenih podataka. Značajke će u ovom radu predstavljati elementi košarkaške statistike pobrojani u prethodnom potpoglavlju.

Važno je napomenuti kako se sve predstavljene izlučene značajke mogu prikazati i kao kombinacija osnovnih elemenata košarkaške statistike, ali ih je važno zbog lakšeg razumijevanja cjelokupne problematike podijeliti u logične skupine. Osim podjele značajki u logične skupove, moguće ih je podijeliti i na značajke pozitivnog i značajke negativnog doprinosa.

3.2.1. Pozitivni i negativni doprinosi

U odjeljku 2.3.1 uveden je sveobuhvatni indeks korisnosti. Učinak elemenata indeksa korisnosti je cijeli broj veći ili jednak 0 ($N_e \geq 0$). Indekse korisnosti se može definirati kao kumulativne indekse koji se sastoje od niza komponenti koje predstavlaju elemente e promatranog procesa p.

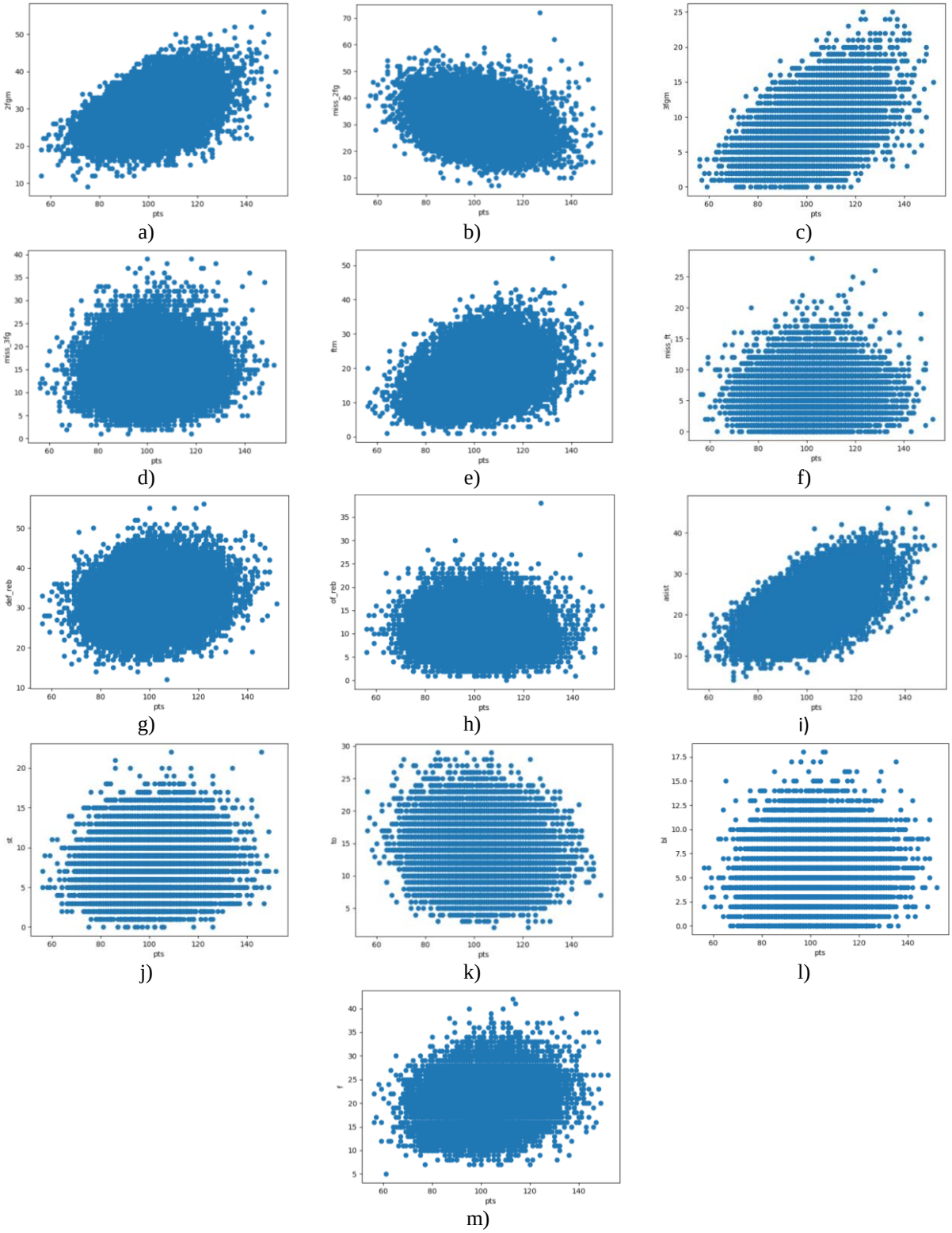
U odjeljku 2.3.1 je navedeno kako svaki element e skupa E procesa p može imati pozitivan i negativan doprinos, gdje je pozitivan učinak označen s N_e, a negativan učinak s N'_e. U svrhu postizanja fleksibilnosti navedeni nenegativni brojevi (N_e i N'_e) množe se s pripadajućim težinskim faktorima v_e i v'_e ($v_e, v'_e \geq 0$). Odabir težinskih faktora može se vršiti na temelju iskustvu eksperta ili korištenjem neke od heurističkih metoda. Tablica 3.3 prikazuje popis elemenata košarkaške statistike koji će se koristiti u ovom istraživanju.

Tablica 3.3. Prikaz pozitivnih i negativnih doprinosa elemenata košarkaške statistike.

e	**Naziv elementa**	$N_e \geq 0$ ($v_e \geq 0$)	$N'_e \geq 0$ ($v'_e \geq 0$)
$2fg$	šutevi za dva poena	N_{2fgm}	N_{miss_2fg}
$3fg$	šutevi za tri poena	N_{3fgm}	N_{miss_3fg}
ft	slobodna bacanja	N_{ftm}	N_{miss_ft}
rbs	skokovi	N_{def_reb}, N_{of_reb}	–
$asts$	asistencije	N_{asist}	–
$stls$	osvojene lopte	N_{st}	–
tos	izgubljene lopte	–	N_{to}
$blcks$	blokade	N_{bl}	–
fls	prekršaji	–	N_f *

Pripadajući faktori v_e i v'_e će u kasnijim poglavljima biti eksperimentalno ispitani, a kao polazna točka koristit će se skup doprinosa indeksa NBA. Konačan ishod događaja (u ovom slučaju košarkaške utakmice) definiran je izlučenom značajkom pts prikazanom formulom (3-5) gdje se

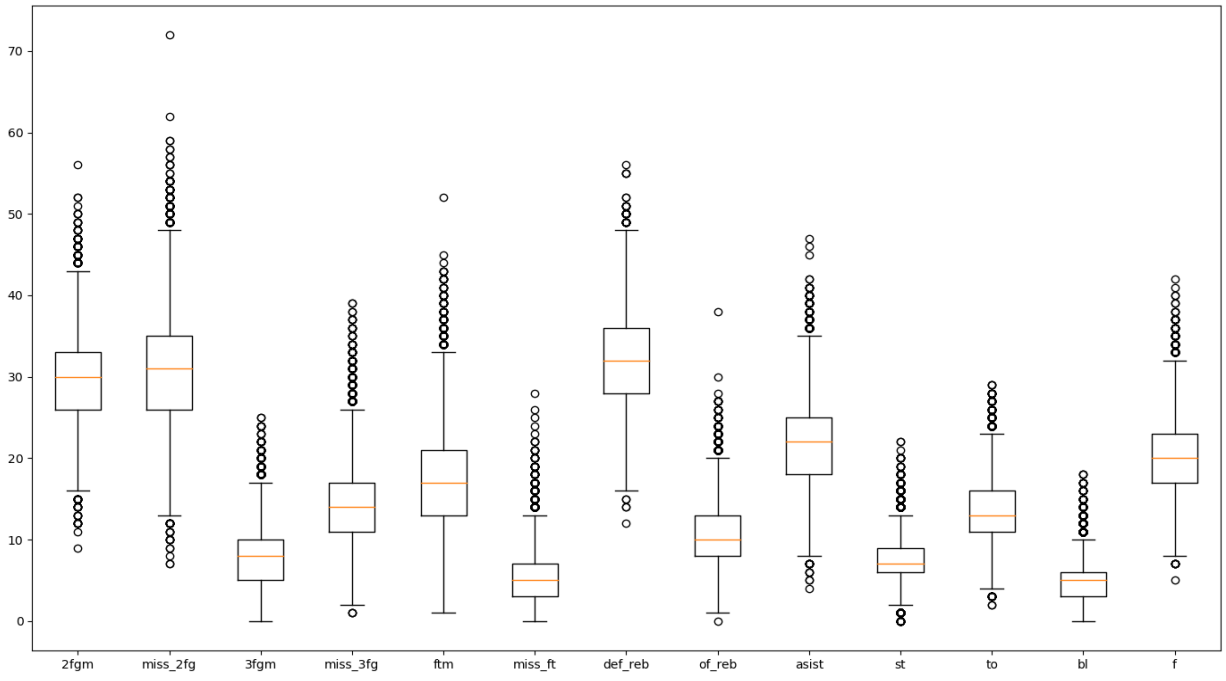
uspješnijim procesom proglašava onaj koji je postigao višu vrijednost. Izlučena značajka pts definira se kao suma umnožaka značajki $2fgm$, $3fgm$ i ftm i pripadajućih doprinosa. Važno je definirati odnos pojedinačnih značajki u odnosu na značajku pts. Slika 3.1 korištenjem raspršenog grafa (engl. *scatter plot*) prikazuje odnos izlučene značajke pts u odnosu na ostale značajke.



Slika 3.1. Odnos izlučene značajke pts u odnosu na a) $2fgm$, b) $miss\_2fg$, c) $3fgm$, d) $miss\_3fg$, e) ftm, f) $miss\_ft$, g) def_reb, h) of_reb, i) $asist$, j) st, k) to, l) bl i m) f.

Svrha raspršenih grafova je pronalaženje trendova među statističkim podacima. Formula (3-5) pokazuje kako značajke $2fgm$, $3fgm$ i ftm direktno utječu na izlučenu značajku pts, točnije kako postoji pozitivna korelacija. Analizom odnosa izlučene značajke pts u odnosu na ostale pojedinačne značajke jasno je vidljivo kako uzlazni trend (pozitivan odnos) prikazuju značajke koje direktno utječu na izlučenu značajku i značajka $asist$ koja je na neki način također vezana uz značajke koje direktno utječu na izlučenu značajku. Naime, značajka $asist$ posljedica je značajki $2fgm$, $3fgm$ i ftm, što znači da viša vrijednost značajke pts prosječno znači i višu vrijednost značajki $2fgm$, $3fgm$ i ftm, a posljedično i značajke $asist$. Ostale značajke, osim značajke $miss\_2fg$ koja pokazuje negativnu korelaciju, ne pokazuju povezanost u odnosu na izlučenu značajku pts, točnije ne postoji izražena korelacija. Posebno važan podatak statističke obrade podataka je linearnost kod kojeg odnos dviju značajki nalikuje pravcu, bilo uzlazno (pozitivan nagib pravca) ili silazno (negativan nagib pravca). Linearnost je vidljiva kod značajki $2fgm$, $3fgm$, ftm, $miss\_2fg$ i $asist$. Sve navedene značajke, osim značajke $miss\_2fg$, pokazuju pozitivan odnos. Pozitivni odnos znači proporcionalnost, dok negativan odnos znači obrnuto proporcijalnu vezu dvije značajke. Valja napomenuti kako raspršeni graf prikazuje moguću povezanost, točnije odnos dvije značajke, što ne znači da nužno postoji i uzročno-posljedična veza.

Osim raspršenih grafova, raspršenost ulaznih podataka može se prikazati i kutijastim dijagramom s ciljem prikaza raspona kvartila vrijednosti i identificiranja stršećih vrijednosti. Slika 3.2 prikazuje kutijasti dijagram raspršenosti značajki osnove košarkaške statistike.



Slika 3.2. Kutijasti dijagram raspodjele elemenata osnovne košarkaške statistike.

Na slici je jasno vidljivo kako stršeće vrijednosti postoje kod svake korištene značajke, ali i da je udio stršećih vrijednosti relativno nizak u odnosu na ukupan broj mjerenja. Također je jasno vidljivo kako su pojedini kutijasti dijagrami široki, a pojedini nešto uži. Širi kutijasti dijagrami se pojavljuju kod značajki kod kojih postoji viši raspon pojavljivanja, točnije značajki kod kojih je razlika između minimalne i maksimalne postignute vrijednosti veća. Značajke čiji je kutijasti dijagram razmjerno kratak, maksimalne stršeće vrijednosti u postotku su puno veće u odnosu na sam kutijasti dijagram.

Vrlo važnu ulogu u definiranju pozivnih i negativnih doprinosa predstavlja *a-priori* znanje o procesu tijekom postupka izgradnje modela. *A-priori* znanje ili znanje na temelju iskustva eksperta je usko povezano uz analizirani proces te ga je potrebno „ručno“ ugrađivati u pojedine korake, a može znatno skratiti sam proces izgradnje modela na temelju povijesnih podataka koji se u pravilu vrši iterativno. Negativna strana korištenja *a-priori* znanja je smanjenje univerzalnosti postupka izgradnje te je potrebno težiti automatizaciji navedenog postupka.

3.2.2. Nelinearni doprinosi

U prethodnom odjeljku definiran je način izračuna pozitivnih i negativnih doprinosa te su u obliku tablice prikazani učinci elemenata košarkaške statistike koji su temeljem *a-priori* znanja definirani kao pozitivni, odnosno negativni. U ovom odjeljku se želi skrenuti pozornost kako linearni doprinosi ponekad nisu rješenje za određenu vrstu problema. Tako će se u ovom istraživanju ispitati i nelinearan učinak težinskih faktora v_e i v'_e ($v_e, v'_e \geq 0$) na rezultate predviđanja. Cilj uvođenja nelinearnih doprinosa pojedinih elemenata svakako je poboljšati rezultate u odnosu na rezultate dobivene isključivo linearnim doprinosima.

Moguće korištene funkcije nelinearnog doprinosa su polinomne funkcije, logaritamske funkcije, eksponencijalne funkcije, trigonometrijske funkcije, racionalne funkcije itd. Analizom skupa podataka bit će važno odrediti granice u kojima se pojedini elementi analiziranog skupa pojavljuju u stvarnosti te sukladno tome definirati skup prikladnih matematičkih funkcija linearnih i nelinearnih doprinosa. Posebna pozornost trebat će se dati ekstremnim (odudarajućim ili stršećim) vrijednostima (engl. *outliers*) koje nisu pogodne za definiranje skupa potencijalnih funkcija doprinosa, ali i činjenici da davanje prevelikog doprinosa pojedinim elementima može u potpunosti zagušiti doprinos ostalih elemenata. Prikladnijom mjerom svakako se smatraju srednja vrijednost i medijalna vrijednost.

Valja napomenuti kako rješavanje problema korištenjem linearnih i nelinearnih doprinosa neće biti binarno, točnije neće se koristiti isključivo linearni ili nelinearni doprinosi, već će se postupkom optimizacije težiti što boljim rezultatima, a taj postupak će uključiti ispitivanje

korištenjem i linearnih i nelinearnih doprinosa te posljedično definiranjem optimalnog skupa pozitivnih i negativnih doprinosa.

3.3. Identifikacija specifičnih značajki

Procese je moguće kategorizirati u određene grupe i samim time definirati skup značajki koje je moguće primijeniti na određenu grupu procesa. Bez obzira o tome što je procese moguće grupirati, svaki proces je specifičan i samim time opisan sebi svojstvenim skupom elemenata (značajki). U ovom radu će biti predložena metoda predviđanja ishoda u sportu. Svaki sport sadrži specifičan skup značajki, ali postoje i značajke koje se mogu koristiti za većinu sportova. Tablica 2.3 prikazuje popis specifičnih značajki vezanih uz košarku. Značajke primjenjive na gotovo sve, prvenstveno momčadske sportove, su uglavnom značajke vezane uz povijest rezultata i uspjeh momčadi. U ovom potpoglavlju će se predložiti prednosti domaćeg terena. Opći oblik značajke prednosti domaćeg terena bit će dan u kasnijim poglavljima. U kasnijim poglavljima će se predložiti i skup izlučenih značajki koji se može koristiti u gotovo svakom sportu.

3.3.1. Prednost domaćeg terena

U momčadskih sportovima izraz prednosti domaćeg terena se opisuje kao prednost koju domaća momčad ima u odnosu na gostujuću. Prva istraživanja vezana uz prednost domaćeg terena krenula su 80-tih godina 20. stoljeća [67], [68] i [69]. Zbog činjenice da se u ovom radu predlaže metoda za predviđanje sportskih ishoda potrebno je analizirati prednost domaćeg terena u više sportova. Najpoznatiji momčadski sportovi su svakako nogomet, košarka i bejzbol. Predloženi model će biti eksperimentalno ispitan na podacima NBA lige, stoga će skup podataka vezan uz košarku biti kasnije detaljnije obrađen. U najpoznatijoj svjetskoj nogometnoj ligi, engleskoj Premier ligi, u sezoni 2018./2019. domaćin je pobijedio u 47 % utakmica, gost u 34 % utakmica, a 19 % utakmica završilo je neriješenim ishodom. Kako nogomet nudi tri moguća ishoda, ne čudi podatak da postotak pobjeda domaćina nije veći od 50 %, ali je svakako veći od postotaka pobjeda gosta i neriješenih ishoda. Najmanja razlika u postotku pobjeda domaćina u odnosu na gosta je u američkoj MLB ligi gdje je u sezoni 2018./2019. domaćin pobijedio 52,60 % utakmica, a gost 47,40 % utakmica. Prednost domaćeg terena varira od sporta do sporta, ali i države do države.

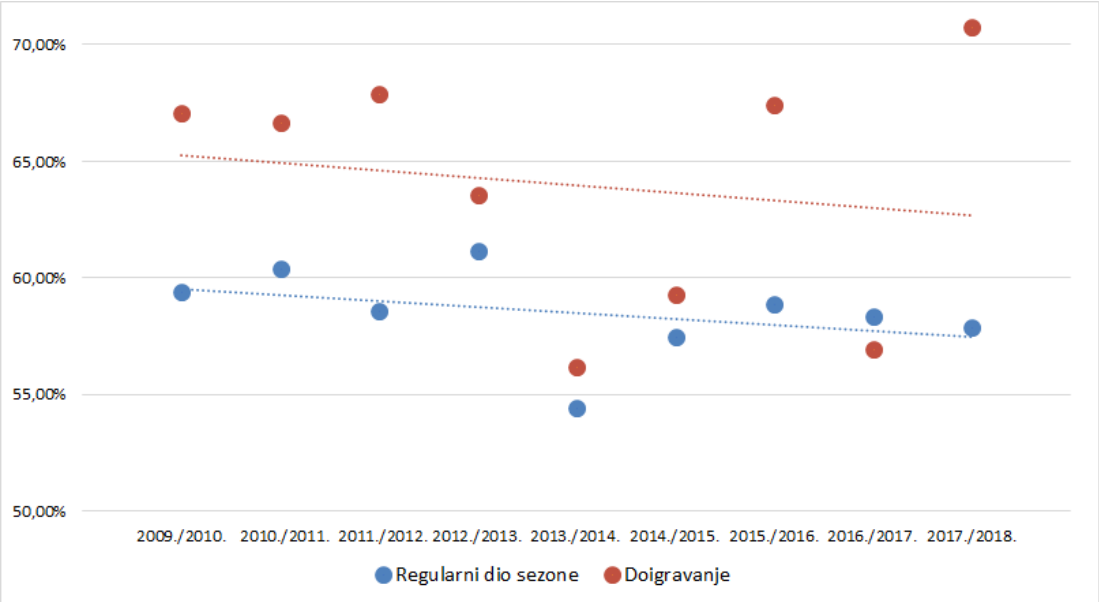
Odigravanjem više od polovice utakmica regularnog dijela NBA lige povećava se intenzitet i natjecateljski duh momčadi jer se polako dohvaćaju pozicije za doigravanje. Sve navedeno se radi s jednim ciljem, a to je osiguravanje prednosti domaćeg terena tijekom doigravanja. S druge strane, momčadi koje su izgubile matematičke mogućnosti ulaska u doigravanje mijenjaju način igre dajući više prilika mladim igračima čineći ih zanimljivijim za možebitne razmjene igrača, a ujedno

se i pripremaju za iduću sezonu. Detaljniji opis posebnosti NBA lige u odnosu na većinu ostalih profesionalnih liga je dana u potpoglavlju 2.4. Tablica 3.4 prikazuje postotak pobjeda domaćina tijekom regularnog dijela sezone i tijekom doigravanja.

Tablica 3.4. Omjer pobjeda domaćina i gosta tijekom regularnog dijela sezone i doigravanja.

Sezona	Regularni dio sezone	Doigravanje
2009./2010.	731/499 (59,43 %)	55/27 (67,07 %)
2010./2011.	743/487 (60,40 %)	54/27 (66,67 %)
2011./2012.	580/410 (58,58 %)	57/27 (67,86 %)
2012./2013.	752/477 (61,14 %)	54/31(63,53 %)
2013./2014.	714/516 (54,42 %)	50/39 (56,18 %)
2014./2015.	707/523 (57,48 %)	48/33 (59,26 %)
2015./2016.	724/506 (58,86 %)	58/28 (67,44 %)
2016./2017.	718/512 (58,37 %)	45/34 (56,96 %)
2017./2018.	712/518 (57,89 %)	58/24 (70,73 %)
Ukupno	58,51 %	63,98 %

Prednost domaćeg terena je očita, ali i izraženija tijekom doigravanja u odnosu na regularni dio sezone. Izraženija prednost tijekom doigravanja je očekivana s obzirom da prednost domaćeg terena stječu uspješnije momčadi tijekom regularnog dijela sezone. Pitanje koje se ovdje može postaviti je ukoliko bi se prednost domaćeg terena dala slabije rangiranoj momčadi hoće li brojevi ostati isti? Odgovor na navedeno pitanje teško je egzaktno odrediti jer ne postoji relevantan skup podataka, ali pretpostavka je da bi prednost bila na strani uspješnije momčadi tijekom regularnog dijela sezone. Slika 3.3 grafički prikazuje prednost domaćeg terena te trend promjene prednosti domaćeg terena tijekom regularnog dijela i doigravanja.



Slika 3.3. Postotak pobjeda domaćina tijekom regularnog dijela sezone i doigravanja.

Lagani pad trenda prednosti domaćeg terena je vidljiv što znači da je konkurentnost lige svakom sezonom sve veća, čime je i predviđanje ishoda dodatno otežano.

Ukoliko se analiziraju i sezone 1998./1999. – 2008./2009., dolazi se do brojke od 7021 pobjede domaćina i 4569 pobjeda gosta tijekom regularnog dijela. Postotak pobjeda domaćina u odnosu na gosta tako iznosi 60,60 %, što je više od 58,51 % koliko iznosi postotak pobjeda domaćina tijekom analiziranog perioda. Navedeni postotci predstavljaju još jedan pokazatelj povećanja konkurentnosti NBA lige. Tijekom istog vremenskog razdoblja, u fazi doigravanja domaćin je pobijedio u ukupno 513 utakmica, a gost u smo 278 utakmica čime se prethodno navedeni podatak povećao na 64,90 % (više od 4 % u odnosu na regularni dio sezone te gotovo 1 % u odnosu na sezone korištene u trenutnom istraživanju). Tablica 3.5 prikazuje usporedbu postotaka pobjeda domaćina u odnosu na gosta za vrijeme dva kronološki poredana natjecateljska perioda.

Tablica 3.5. Usporedba postotaka pobjeda domaćina u odnosu na gosta.

Sezone	**Regularni dio sezone**	**Doigravanje**
1998./1999. – 2008./2009.	60,60 %	64,90 %
2009./2010. – 2017./2018.	58,51 %	63,98 %
Razlika	2,09 %	0,92 %

Navedeno povećanje postotka vezano uz pobjede domaćih momčadi tijekom faze doigravanja se može potkrijepiti činjenicom da prednost domaćeg terena imaju bolje rangirane momčadi regularnog dijela. Svaka faza NBA doigravanja se igra na četiri pobjede, po principu 2-2-1-1-1 što znači da domaća momčad igra kod kuće prvu, drugu, eventualnu petu i sedmu utakmicu. NBA doigravanje se sastoji od 15 serija. U zadnjih 10 sezona je odigrano ukupno 150 serija, a momčad s prednošću domaćeg terena je pobijedila 111 serija (74,00 %). Tablica 3.6 prikazuje omjer pobjeda serija domaćina i gosta. Postotak pobjeda domaće momčadi u pravilu je veći od postotka pobjeda gosta, a rijetki su slučajevi kada je postotak jednak.

Tablica 3.6. Prikaz postotka pobjeda momčadi s prednošću domaćeg terena tijekom faze doigravanja.

	Prva runda	**Polufinale konferencije**	**Finale konferencije**	**Finale**
4 utakmice	14/2 (87,5 %)	8/2 (80,00 %)	2/1 (66,67 %)	1/0 (100,00 %)
5 utakmica	17/2 (89,47 %)	10/1 (90,90 %)	2/2 (50,00 %)	3/1 (75,00 %)
6 utakmica	16/12 (57,14 %)	9/3 (75,00 %)	4/4 (50,00 %)	1/1 (50,00 %)
7 utakmica	13/4 (76,47 %)	6/1 (85,71 %)	3/2 (60,00 %)	2/1 (66,67 %)

U radu [70] je analiziran učinak momčadi kod kuće i u gostima te je zaključeno da prosječna prednost domaćeg terena iznosi 3-5 poena, odnosno približno 1 poen po odigranoj četvrtini. Također je važno napomenuti da prednost domaćeg terena varira od momčadi do momčadi.

Ukoliko se zasebno analiziraju sezone 2009./2010. – 2017./2018., prosječna razlika u poenima između domaćina i gosta je 2,86 poena u korist domaćina, točnije 2,76 poena tijekom regularnog dijela sezone te 4,26 poena tijekom faze doigravanja. Broj utakmica regularnog dijela je puno veći od broja utakmica doigravanja pa samim time broj utakmica doigravanja slabo utječe na ukupnu razliku tijekom cijele sezone. Tablica 3.7 prikazuje prosječne razlike u poenima između domaće i gostujuće momčadi za svaku analiziranu sezonu.

Tablica 3.7. Prosječna razlika postignutih poena domaćina i gosta po sezonama.

Sezona	Regularni dio	Doigravanje	Ukupna prosječna razlika
2009./2010.	2,73	4,5	2,84
2010./2011.	3,17	3,64	3,20
2011./2012.	2,82	4,68	2,97
2012./2013.	3,27	4,00	3,28
2013./2014.	2,60	2,84	2,61
2014./2015.	2,41	2,14	2,39
2015./2016.	2,67	8,07	3,03
2016./2017.	3,15	2,06	3,08
2017./2018.	2,11	6,28	2,37
Ukupno	**2,76**	**4,26**	**2,86**

Iz rezultata je vidljivo da domaćin u prosjeku pobjeđuje više utakmica, s naglaskom da je prosječna razlika tijekom faze doigravanja još i veća. U prilog prednosti domaćeg terena ide i analiza portala SBNation.com koja je pokazala kako bi NBA momčadi pobijedile prosječno dodatnih 10,11 % utakmica da sve utakmice igraju kod kuće [71].

3.3.2. Definiranje značajke prednosti domaćeg terena

U odjeljku 3.3.1 je pokazano kako prednost domaćeg terena postoji u gotovo svim sportskim ligama, a poseban fokus dan je analizi skupa podataka koji će se koristiti u ovom istraživanju. Ukoliko se promatra samo skup podataka koji će se koristiti u ovom istraživanju, jasno je pokazano kako prednost domaćeg terena postoji, ali i da je prednost domaćeg terena izraženija tijekom faze doigravanja u odnosu na regularni dio sezone. Cilj ovog odjeljka je definirati značajku koja će definirati prednost domaće momčadi. Prednost domaćeg terena će se računati na temelju povijesnih podataka. Neka je u zadanom trenutku t broj pobjeda domaćina (tm_d) označen s N_{tm_d}, a broj pobjeda gosta (tm_g) s N_{tm_g}. Razlika omjera pobjeda (Δtm_N) domaćina u odnosu na gosta prikazana je formulom (3-9), dok formule (3-7) i (3-8) prikazuju izračun omjera pobjeda domaćina i gosta.

$$\Delta tm_d = \frac{N_{tm_d}}{N_{tm_d} + N_{tm_g}}, \Delta tm_d \in [0,1], N_{tm_d}, N_{tm_g} \geq 0 \quad (3\text{-}7)$$

$$\Delta tm_g = \frac{N_{tm_g}}{N_{tm_d} + N_{tm_g}} \; \Delta tm_g \in [0,1], N_{tm_d}, N_{tm_g} \geq 0 \quad (3\text{-}8)$$

$$\Delta tm_N = \Delta tm_d - \Delta tm_g \quad (3\text{-}9)$$

Prednost domaćeg terena će tako biti prikazana decimalnim brojem u intervalu $[0,1]$, a suma učinaka suprotstavljenih momčadi će u svakom trenutku biti jednaka 1 ($\Delta tm_d + \Delta tm_g = 1$). Valja napomenuti kako se uvedeni pojam prednosti domaćeg terena može prilagoditi gotovo svakom problemu u kojem postoje dva suprotstavljena procesa. Opći oblik značajke prednosti uspješnijeg procesa će biti dan u odjeljku 4.1.2.

3.4. Primjena sveobuhvatnog indeksa korisnosti u predviđanju ishoda sportskih događaja

U odjeljku 2.3.1 opisan je sveobuhvatni indeks korisnosti. Sveobuhvatni indeks korisnosti definiran je kao kumulativan indeks koji se sastoji od niza komponenti ponderiranih koeficijentom W_e, gdje svaka komponenta predstavlja element e promatranog procesa. Osnovna ideja indeksa CPE je da bude neograničen brojem elemenata, da ne postoje unaprijed definirani elementi te da je lako prilagodljiv svim procesima koji se mogu podijeliti na komponente. U ovom potpoglavlju bit će objašnjeno na koji način je sveobuhvatni indeks korisnosti moguće koristiti u predviđanju ishoda sportskih događaja.

3.4.1. Indeks korisnosti kao pokazatelj ishoda sportskog događaja

Indeks korisnosti je relativan indikator kvalitete igrača ili momčadi. Ukoliko ga se promatra kao relativan indikator kvalitete momčadi, može ga se definira kao sumu indeksi korisnosti igrača koji su participirali na utakmici za pojedinu momčad. Indeks korisnosti momčadi za utakmicu gm promatrane momčadi tm u kojoj su nastupili igrači pl prikazan je formulom (3-10). Učinak igrača koji nisu aktivno sudjelovali na utakmici gm jednak je nuli te ne utječe na indeks momčadi.

$$I(tm, gm) = \sum_p I(pl, gm) \quad (3\text{-}10)$$

Jedini indikator definiranja ishoda utakmice su, ovisno o analiziranom sportu, postignuti poeni ili golovi. Točan broj poena ili golova gotovo je nemoguće predvidjeti. Uvođenjem većeg broja značajki olakšava se postupak predviđanja. Posebnu pozornost važno je dati odabiru ulaznih značajki, točnije smanjenjem ili povećanjem dimenzionalnosti problema pridonijeti bržem

izvođenju algoritma i povećanju efikasnosti. Točnije, potrebno je pronaći optimalan skup značajki koji će doprinijeti maksimiziranju točnosti modela predviđanja.

Točan broj elemenata, u slučaju strojnog učenja broj značajki, ovisi o samom procesu. Govoreći o sportu, skup značajki najčešće čini specifična statistika vezana uz analizirani sport te značajke vezane uz uspješnost momčadi kao cjeline. Indeksi korisnosti se najčešće koriste kao metrika kojom se evaluira učinak igrača, a kumulativno i učinak momčadi.

Vrlo je važno izračunati koliko je indeks korisnosti relevantan pokazatelj ishoda sportskog događaja ili bilo kojeg drugog procesa. Pojam pokazatelja ishoda sportskog događaja pokazuje koliko informacije indeks korisnosti donosi o događaju, tj. koliko je omjer indeksa korisnosti dobra mjera za procjenu ishoda događaja.

3.4.2. NBA indeks kao pokazatelj ishoda košarkaške utakmice

NBA indeks je definiran u potpoglavlju 2.3. Cilj ovog odjeljka je izračunati koliko je NBA indeks relevantan pokazatelj ishoda košarkaške utakmice. Za izračun pokazatelja ishoda korišten je stvarni NBA indeks momčadi analizirane utakmice, a pobjednikom je proglašena momčad čiji je indeks korisnosti viši. U slučaju kada su indeksi momčadi jednaki pobjednikom se proglašava domaća momčad, a razlog tome je naveden u odjeljku 3.3.1 u kojem je pokazano da u NBA ligi postoji izražena prednost domaćeg terena. Formula (3-11) prikazuje matematičku funkciju $(winn(I_{NBA}(tm_d), I_{NBA}(tm_g))$ izračuna pobjedničke momčadi utakmice gm na temelju stvarnih NBA indeksa suprotstavljenih momčadi. Tablica 3.8 prikazuje rezultate NBA indeksa kao pokazatelja ishoda košarkaške utakmice.

$$winn(I_{NBA}(tm_d), I_{NBA}(tm_g)) = \begin{cases} tm_d, \frac{I_{NBA}(tm_d, gm)}{I_{NBA}(tm_g, gm)} \geq 1; \\ tm_g, \frac{I_{NBA}(tm_d, gm)}{I_{NBA}(tm_g, gm)} < 1; \end{cases} \quad (3\text{-}11)$$

Tablica 3.8. NBA indeks kao pokazatelj ishoda utakmice.

Sezona	NBA	NBA (prednost domaćeg terena)
2009./2010.	92,30 %	92,76 %
2010./2011.	92,30 %	92,60 %
2011./2012.	91,06 %	91,71 %
2012./2013.	91,55 %	92,24 %
2013./2014.	92,04 %	92,65 %
2014./2015.	91,69 %	92,30 %
2015./2016.	91,95 %	92,55 %
2016./2017.	91,98 %	92,13 %
2017./2018.	91,46 %	91,77 %
Ukupno:	**91,83 % (10632/11578)**	**92,31 % (10688/11578)**

NBA indeks definira pobjednika u prosječno 92,31 % slučajeva kada se u slučaju jednakih NBA indeksa domaćina i gosta pobjednikom proglasi domaća momčad, te prosječno 91,83 % u slučaju kada se predviđanje u slučaju kada su NBA indeks domaćina i gosta jednaki smatra neuspješnim. S obzirom da je košarka sport u kojem se u odnosu na većinu momčadskih sportova postiže relativno velik broj poena, promjena od nekoliko postotnih bodova znači i vrlo vjerojatnu promjenu ishoda, prosječna točnost od 92,31 % je zadovoljavajuća.

Osnovna košarkaška statistika bilježi i element broja počinjenih prekršaja (f) koji NBA indeks ne koristi. Ukoliko NBA indeksu dodamo i negativni učinak počinjenih prekršaja, točnost predviđanjem modificiranog NBA indeksa, prikazanog formulom (3-12), se povećava na 93,95 %. Zbroj težinskih faktora (W_e) NBA indeksa i modificiranog NBA indeksa jednak je kardinalnosti skupa E, što znači da su težinski faktori (W_e) svih elemenata $W_e = 1$, gdje skup E predstavlja skup elemenata osnove košarkaške statistike. Valja napomenuti kako je indeks korisnosti relativan indikator kvalitete momčadi te veća točnost definiranja pobjednika korištenjem stvarnih podataka ne znači nužno i bolje sposobnosti predviđanja na temelju povijesnih podataka. Tablica 3.9 prikazuje usporedbu rezultata predviđanja ishoda korištenjem NBA indeksa u odnosu na modificirani NBA indeks. Valja napomenuti kako je u oba slučaja korištena prednost domaćeg terena u slučaju istih vrijednosti.

$$I'_{NBA} = \left(N_{pts} + N_{rbs} + N_{as} + N_{st} + N_{bl}\right) - \left(N_{miss\_fg} + N_{miss\_ft} + N_{to} + N_f\right) \qquad (3\text{-}12)$$

Tablica 3.9. NBA indeks i modificirani NBA indeks kao pokazatelji ishoda utakmice.

Sezona	NBA indeks	Modificirani NBA indeks
2009./2010.	92,76 %	93,75 %
2010./2011.	92,60 %	93,97 %
2011./2012.	91,71 %	93,58 %
2012./2013.	92,24 %	93,99 %
2013./2014.	92,65 %	94,62 %
2014./2015.	92,30 %	94,81 %
2015./2016.	92,55 %	94,07 %
2016./2017.	92,13 %	92,97 %
2017./2018.	91,77 %	93,67 %
Ukupno:	**92,31 % (10688/11578)**	**93,95 % (10877/11578)**

3.4.3. Primjena indeksa CPE u predviđanju ishoda sportskih događaja

Indeksi korisnosti su se pokazali uspješnim indikatorima predviđanja ishoda korištenjem stvarnih podataka analiziranih utakmica. Kod predviđanja budućih ishoda korištenje stvarnih podataka nije moguće, već je predviđanje potrebno vršiti na temelju povijesnih podataka. Svaki

sport karakterizira specifična statistika te je indeks korisnosti potrebno prilagoditi što sveobuhvatni indeks korisnosti čini posebno pogodnim. Sveobuhvatni indeks korisnosti neograničen je brojem elemenata i ne postoje unaprijed definirani elementi te je samim time pogodan za predviđanje ishoda sportskih događaja.

Ukoliko se s E prikaže skup elemenata analiziranog sporta, moguće je definirati i sveobuhvatni indeks korisnosti. Formula (3-13) prikazuje izračun CPE indeksa korisnosti na temelju skupa elemenata E.

$$I_{CPE} = \sum_{e \in E} W_e I_e, \quad E = skup\ elemenata\ promatranog\ e \tag{3-13}$$

U sportu postoje dvije suprotstavljene momčadi, tako da je za predviđanje ishoda utakmice potrebno definirati indeks korisnosti svake momčadi. Formula (3-14) prikazuje način izračuna ishoda sportskog događaja na temelju povijesnih podataka.

$$winn(I_{CPE}(tm_A), I_{CPE}(tm_B)) \begin{cases} tm_A, \dfrac{I_{CPE}(tm_A)}{I_{CPE}(tm_B)} > 1; \\ tm_x, \dfrac{I_{CPE}(tm_A)}{I_{CPE}(tm_B)} = 1\ ; \\ tm_B, \dfrac{I_{CPE}(tm_A)}{I_{CPE}(tm_B)} < 1. \end{cases} \tag{3-14}$$

Kao što je vidljivo iz formule (3-14) pobjedničkom momčadi proglašava se momčad čiji je projicirani CPE indeks viši. U slučaju kada je CPE indeks obje momčad jednak pobjedničku momčad nije moguće definirati, a taj je slučaj označen oznakom tm_x. Problem nemogućnosti definiranja pobjedničke momčad će biti objašnjen na konkretnim primjerima u kasnijim poglavljima.

3.4.4. Primjena indeksa CPE u predviđanju ishoda košarkaških utakmica

Opća formula indeksa CPE vezana uz proizvoljan proces je dana u odjeljku 2.3.1. Primjena novonastalog indeksa CPE će u ovom radu biti na predviđanju ishoda košarkaških utakmica. Generatori komponenti će biti igrači, a analizirani proces košarkaška utakmica. Na košarkaškoj utakmici sudjeluje više igrača (generatora komponenti), a učinak svakog igrača opisan je elementima košarkaške igre. Tablica 3.2 prikazuje elemente osnovne košarkaške statistike.

U košarci se koriste još tri izlučena elementa, a to su broj promašenih pokušaja za dva poena ($miss\_2fg$) prikazan formulom (3-1), broj promašenih pokušaja za tri poena ($miss\_3fg$) prikazan formulom (3-2) te broj promašenih slobodnih bacanja ($miss\_ft$) prikazan formulom(3-3).

Govoreći o konkretnom primjeru predviđanja ishoda u košarci, način definiranja ishoda košarkaške utakmice gm korištenjem povijesnih podataka je prikazan formulom (3-15).

$$winn(I_{CPE}(tm_d), I_{CPE}(tm_g)) \begin{cases} tm_d, \frac{I_{CPE}(tm_d)}{I_{CPE}(tm_g)} > 1; \\ tm_x, \frac{I_{CPE}(tm_d)}{I_{CPE}(tm_g)} = 1\,; \\ tm_g, \frac{I_{CPE}(tm_d)}{I_{CPE}(tm_g)} < 1. \end{cases} \quad (3\text{-}15)$$

Kao i u prethodnom slučaju, nemogućnost definiranja pobjednika javlja se u slučaju kada omjer projiciranih CPE indeksa domaće i gostujuće momčadi iznosi jedan. U tom slučaju je potrebno predviđanju pristupiti na poseban način koji će biti definiran u kasnijim poglavljima.

3.4.5. NBA indeks kao poseban slučaj CPE indeksa

Novonastali indeks CPE moguće je prilagoditi indeksima NBA i PIR. Za postavljenje $v_e = v_e(u_e)$ na željenu nenegativnu vrijednost c_e dovoljno je:

i. postaviti u_e na konstantu vrijednost $u_e = const. = 1$, kao što je prethodno i predloženo, te odabrati $a_e = c_e$ što dovodi do željenog rezultata $v_e(u_e) = a_e = c$.

ii. ograničiti koeficijente $v_{e,min}$ i $v_{e,max}$, što znači da je $v_{e,min} = v_{e,max} = a_e = c_e$ te da je $u_{e,min} = u_{e,max} = 1$ što daje isti rezultat kao i prethodni uvjet, odnosno $v_e(1) = a_e = c$.

Tablica 3.10 prikazuje doprinos elemenata vezanih uz indekse korisnosti NBA i PIR, točnije prilagodbu indeksa CPE vrijednostima indeksa NBA i PIR. Suma težinskih faktora W_e indeksa NBA i PIR fiksirana je na kardinalnost igre, gdje je težinski faktor svakog elementa košarkaške igre $W_e = 1$, a isto pravilo vrijedi i za indeks CPE. Novonastali indeks CPE (I_{CPE}) može koristiti razrađeniji i precizniji prikaz doprinosa pojedinog igrača ili momčadi s ciljem stvaranja ravnoteže između nagrađivanja pozitivnih i kažnjavanja negativnih elemenata košarkaške igre. Zvjezdicom (*) su označeni elementi koji se ne koriste kod indeksa NBA, a koriste se kod indeksa PIR.

Tablica 3.10. Tablica doprinosa elemenata košarkaške statistike.

e	**Naziv elementa**	v_e	N_e	v'_e	N'_e
$2fg$	šutevi za dva poena	2	N_{2fgm}	1	N_{miss_2fg}
$3fg$	šutevi za tri poena	3	N_{3fgm}	1	N_{miss_3fg}
ft	slobodna bacanja	1	N_{ftm}	1	N_{miss_ft}
rbs	skokovi (def_reb, of_reb)	1	$N_{rbs} = \sum_{i=1}^{2} N_{rbs,i}$	–	–
$asts$	asistencije	1	N_{asist}	–	–
$stls$	osvojene lopte	1	N_{st}	–	–
tos	izgubljene lopte	–	–	1	N_{to}
$blcks$	blokade	1	N_{bl}	1*	N_{bl_ag} *
fls	prekršaji	1*	N_{f_dr} *	1*	N_f *

Iz prethodnih razmatranja, prikazano formulom (3-16), se može definirati i CPE indeks, sveden na opći oblik NBA ili PIR indeksa na nivou igrača ili momčadi.

$$I_{CPE} = \sum_{e \in E} W_e I_e = I_{pts} + I_{rbs} + I_{asts} + I_{stls} + I_{tos} + I_{blcks} + I_{fls} \tag{3-16}$$

Također, moguće je definirati i CPE indeks po minutama ($I_{CPE/min}$) prikazan formulom (3-17) te normaliziran indeks na cijelu utakmicu ($I_{CPE/N_{reg}}$) prikazan formulom (3-18) gdje N_{min} predstavlja broj minuta igrača provedenih u igri, dok N_{reg} predstavlja ukupan broj minuta regularnog dijela utakmice.

$$I_{CPE/min} = \frac{1}{N_{min}} I_{CPE} \tag{3-17}$$

$$I_{CPE/N_{reg}} = \frac{N_{reg}}{N_{min}} I_{CPE} \tag{3-18}$$

CPE indeks po minutama i normaliziran indeks na cijelu utakmicu se u pravilu ne koriste, ali se mogu koristiti u analitičke svrhe vezane uz evaluaciju učinka igrača ili momčadi.

3.4.6. Koeficijent $u_e(u'_e)$

Detaljniji opis koeficijenta u_e dan je u odjeljku 2.3.1.2. Koeficijent $u_e(u'_e)$ može biti predložen od strane eksperta, definiran od strane korisnika ili pak izračunat nekom od heurističkih metoda. Tablica 3.11 prikazuje mogućnosti odabira koeficijenta $u_e(u'_e)$ predloženih na temelju iskustva eksperta vezan uz konkretan problem predviđanja košarkaških ishoda.

Tablica 3.11. Mogućnosti odabira koeficijenta u_e (u'_e).

#	u_e (≥ 0)	
0	$c \geq 0$	*Korisnički definirana vrijednost*
1a	$\bar{N}_e/\bar{N}_{tm,e}$, $\bar{N}_{tm,e} > 0$	*Posebni slučajevi:* Ako su $\bar{N}_e$ ili $\bar{N}_{tm,e}$ nedefinirani tada je $\bar{N}_e/\bar{N}_{tm,e} = 1$ inače ako je $\bar{N}_{tm,e} = 0 \Longrightarrow \bar{N}_e = 0$ tada je $\bar{N}_e/\bar{N}_{tm,e} = 1$
1b	$\bar{N}_{tm,e}/\bar{N}_e$, $\bar{N}_e > 0$	*Posebni slučajevi:* Ako su $\bar{N}_e$ ili $\bar{N}_{tm,e}$ nedefinirani tada je $\bar{N}_e/\bar{N}_{tm,e} = 1$. inače ako je $\bar{N}_e = 0 \Longrightarrow u_e \to \infty$ tada je $\bar{N}_{tm,e}/\bar{N}_e = v_{e,max}$
2a	$N_e/\bar{N}_{\mathrm{e}}$, $\bar{N}_{\mathrm{e}} > 0$	*Posebni slučajevi:* Ako su N_e ili $\bar{N}_e$ nedefinirani tada je $N_e/\bar{N}_e = 1$. inače ako je $\bar{N}_e = 0 \Longrightarrow u_e \to \infty$ tada je $N_e/\bar{N}_e = v_{e,max}$
2b	$\bar{N}_{\mathrm{e}}/N_e$, $N_e > 0$	*Posebni slučajevi:* Ako su $\bar{N}_e$ ili N_e nedefinirani tada je $\bar{N}_{\mathrm{e}}/N_e = 1$. inače ako je $N_e = 0 \Longrightarrow u_e \to \infty$ tada je $\bar{N}_{\mathrm{e}}/N_e = v_{e,max}$

Primjer 0 definira konstantnu vrijednost $c \geq 0$ definiranu na temelju iskustva ili neke heurističke metode. Primjer 1a definira omjer prosječnog učinka igrača elementa e u zadanom vremenskom periodu Δt ($\bar{N}_{\mathrm{e}}$) i odgovarajućeg učinka momčadi tijekom istog vremenskog perioda

($\overline{N}_{tm,e}$). $\overline{N}_{tm,e}$ i $\overline{N}_{\mathrm{e}}$ su nedefinirani ukoliko ne postoji utakmica momčadi ili igrača u definiranom vremenskom periodu (Δt). Primjer 2a definira učinak igrača, točnije učinak igrača na utakmici i prosječnog učinka igrača u definiranom vremenskom periodu za element e košarkaške igre. Posebni slučajevi su također definirani u tablici, a slučajevi 1b i 2b su recipročne vrijednosti slučajeva 1a i 2a.

3.4.7. Sveobuhvatni indeks korisnosti momčadi

U ovom odjeljku će se uvesti pojam sveobuhvatnog indeksa korisnosti momčadi (engl. *Comprehensive Team Efficiency Indeks* ili kraće CTE). CTE indeks, prikazan formulom (3-19), predstavlja sumu CPE indeksa korisnosti igrača koji su igrali na utakmici gm za momčad tm i postigli određeni doprinos, a CPE indeks igrača pl na utakmici gm zapisuje se kao $I_{CPE}(pl, gm)$.

$$I_{CTE}(tm, gm) = \sum_{p} I_{CPE}(pl, gm)\,;\ \ \forall\, pl\ momčadi\ tm\ s\ definiranim\ I_{CPE}(pl, gm) \tag{3-19}$$

Kao što je već ranije napomenuto, predviđanje ishoda se vrši isključivo na temelju povijesnih podataka te je potrebno uvesti pojam projiciranog indeksa korisnosti. Projicirani indeks korisnosti se računa za svaku momčad na temelju poznatih povijesnih podataka tijekom definiranog vremenskog perioda. Formula (3-20) prikazuje opći oblik izračuna projiciranog indeksa korisnosti u trenutku t neposredno prije predviđanja utakmice n. Projicirani indeks korisnost momčadi tm u trenutku t će se tako računati na temelju $n-1$ utakmica skupa za učenje.

$$I(tm, t) = \frac{1}{n-1}\sum_{i=1}^{n-1} I(tm, i) \tag{3-20}$$

Valja napomenuti da će u budućim zapisima oznaka I predstavljati projicirani indeks korisnosti, osim u slučaju kada će biti napomenuto drugačije. U slučaju kada ne postoji poznata povijest projicirani indeks korisnosti nije moguće izračunati.

3.4.8. Korelacija relativnog rezultata (učinka) i relativnog indeksa korisnosti

Dva važna pojma vezana uz predviđanje ishoda su relativni rezultat (učinak) i relativni indeks korisnosti. Ukoliko je promatrana momčad A na utakmici gm postigla $N_A(gm)$ poena, a momčad B također na utakmici gm postigla $N_B(gm)$ poena, relativni rezultat utakmice gm normaliziran u odnosu na promatranu momčad može se zapisati kako slijedi u formuli (3-21).

$$R_{N_A/N_B}(gm) = \frac{N_A(gm)}{N_B(gm)}\,,\ \ (N_A(gm) > 0, N_B(gm) > 0) \tag{3-21}$$

Relativni rezultat (učinak) utakmice može poprimiti vrijednost veću od 1 u slučaju pobjede promatrane momčadi ili manju od 1 u slučaju pobjede protivničke momčadi. Na temelju relativnog rezultata moguće je definirati i relativni indeks korisnosti. Ukoliko su indeksi korisnosti momčadi označeni s $I(A, gm)$ i $I(B, gm)$, relativni indeks utakmice gm prikazan je formulom (3-22).

$$I_{A/B}(gm) = \frac{I(A, gm)}{I(B, gm)} \tag{3-22}$$

Kao i u općem slučaju, relativni indeks utakmice se ne može vrijednosno strogo definirati što znači da indeks korisnosti poražene momčadi može biti veći ili jednak indeksu korisnosti pobjedničke momčadi. Bez obzira na to, relativni indeks korisnosti treba biti linearno koreliran s relativnim rezultatom (učinkom) utakmice. Ista tvrdnja vrijedi u slučaju kada se koristi projicirani indeks korisnosti ili stvarni indeks korisnosti pojedine utakmice.

3.4.9. Prilagodljivost sveobuhvatnog indeksa korisnosti

Sveobuhvatni indeks korisnosti je neograničen brojem elemenata te je samim time lako prilagodljiv ostalim sportovima ili sličnim procesima. Prilagodljiv je u vidu definiranja doprinosa elemenata te je jednostavan za vizualizaciju i razumijevanje. Konačan produkt je metrika, pojam definiran u potpoglavlju 2.3. U ovom radu će se sveobuhvatni indeks korisnosti koristiti za procjenu kvalitete i evaluaciju učinka košarkaške momčadi, a zbog svojstva fleksibilnosti moguće ga je koristiti i za ostale procese koji se mogu podijeliti na komponente. Spomenuti indeks korisnosti je vremenski ograničen te lako prilagodljiv i usporediv s ostalim indeksima korisnosti.

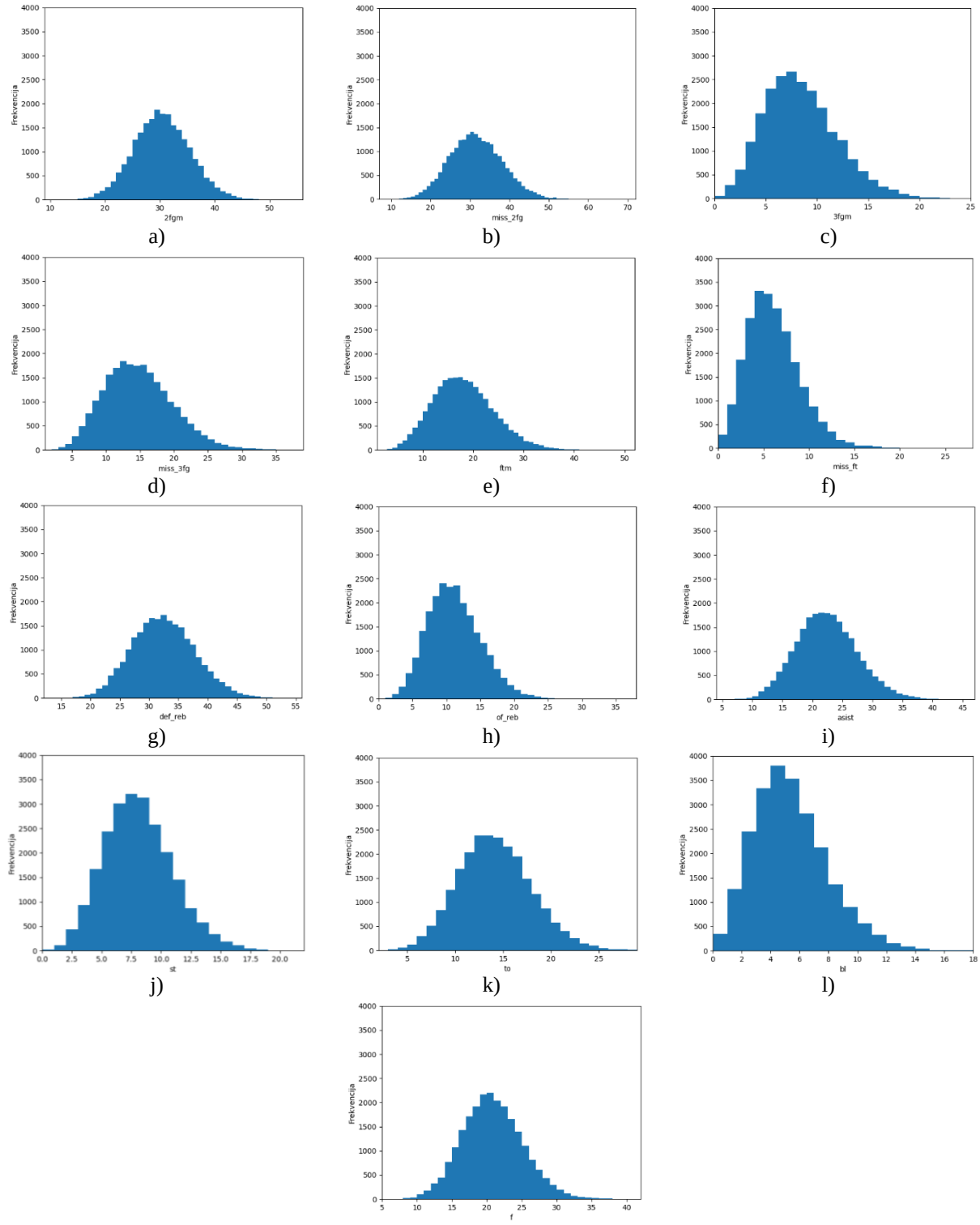
3.5. Analiza osjetljivosti indeksa na doprinos pojedinačnih značajki

Analiza osjetljivosti služi za ispitivanje na koje je čimbenike, u ovom slučaju značajke, podložan proces. Točnije cilj analize osjetljivosti definirati je ključne čimbenike te identificirati čimbenike koji će dati najveću vjerojatnost za uspjehom. Najpogodniji alat za grafički prikaz osjetljivosti značajki je tornado dijagram. Tornado dijagrami prikazuju utjecaj varijabilnosti ulaznih varijabli, u ovom slučaju značajki indeksa korisnosti, na varijabilnost konačnog rezultata. Varijable tornado dijagrama su poredane odozgo prema dolje prema intenzitetu utjecaja na konačan rezultat.

3.5.1. Analiza osjetljivosti NBA indeksa

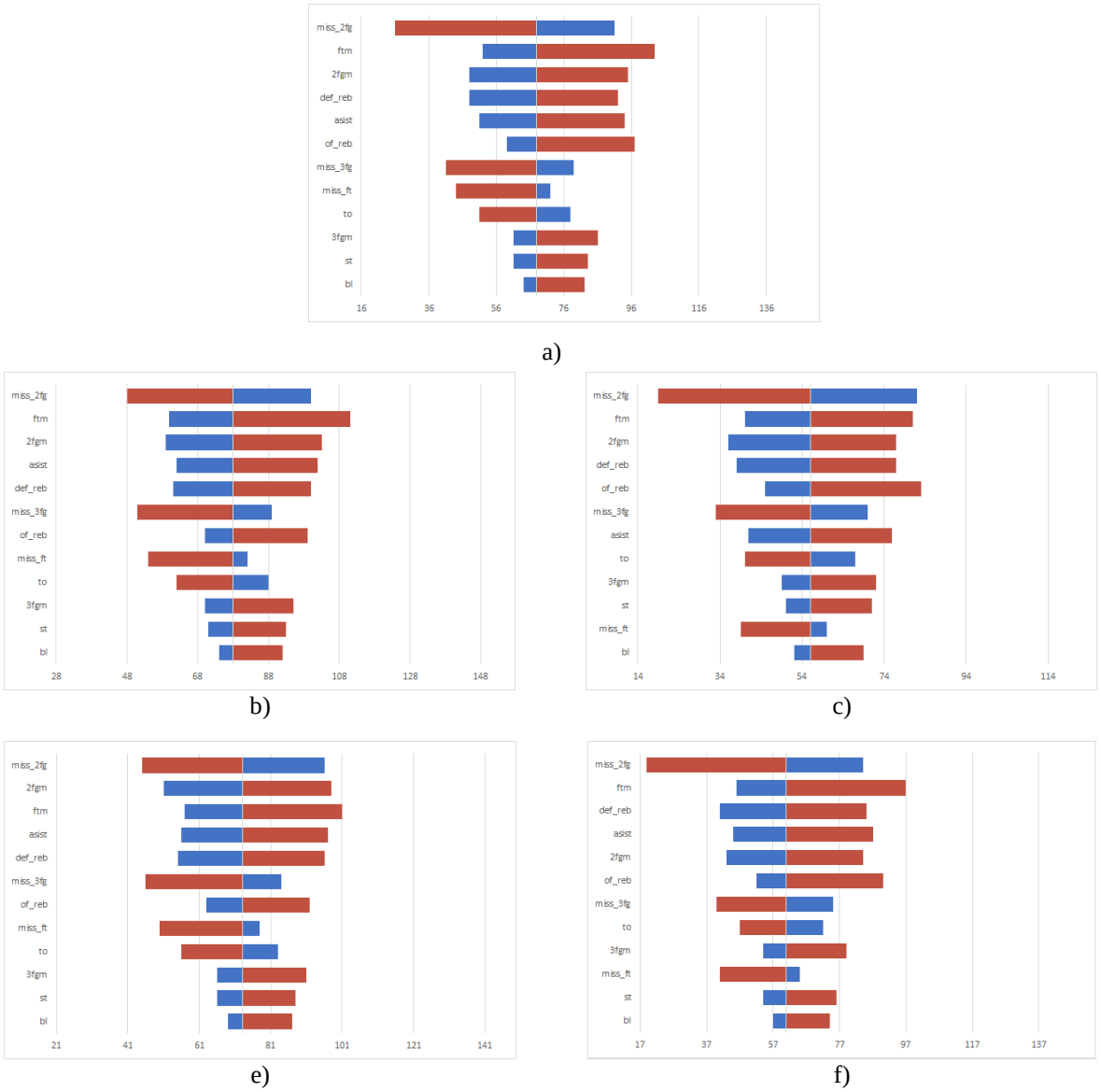
Važan segment upotrebe skupa podataka je analiza podataka, u ovom slučaju analiza osjetljivosti NBA indeksa na doprinos pojedinačnih značajki. Već je navedeno kako osnovna košarkaška statistika sadrži 13 elemenata igre, a da NBA indeks koristi 12 elemenata, točnije ne koristi element počinjenih prekršaja. Ulazne vrijednosti su diskretne što ih čini pogodnim za prikaz histogramom

kojim je moguće vizualno prikazati frekvenciju pojavljivanja određenih grupa ili kategorija numeričkih podataka, gdje os apscise prikazuje vrijednosti mjerenja, a os ordinata frekvenciju pojavljivanja određenog intervala. Slika 3.4 prikazuje histograme razdiobe osnovnih elemenata košarkaške statistike.



Slika 3.4. Raspodjela mjerenja značajke a) $2fgm$, b) $miss\_2fg$, c) $3fgm$, d) $miss\_3fg$, e) ftm, f) $miss\_ft$, g) def_reb, h) of_reb, i) $asist$, j) st, k) to, l) bl i m) f.

Korišteni ulazni skup podataka se sastoji od ukupno 11 578 utakmica, točnije svaku utakmicu igraju dvije suprotstavljene momčadi što čini ukupno 23 156 mjerenja. Raspodjela uzoraka u pravilu je normalna (Gaussova) s nekoliko izuzetaka kod kojih je razdioba iskrivljena u desno. Važan pojam vezan uz statističku analizu podataka je i osjetljivost (engl. *sensitivity*). Osjetljivost značajki prikazuje stupanj osjetljivosti konačnog rezultata u odnosu na nezavisne varijable. Najpogodniji alat za grafički prikaz osjetljivosti značajki je tornado dijagram. Slika 3.5 prikazuje osjetljivost NBA indeksa na doprinos pojedinačnih značajki korištenjem cijelog ulaznog skupa podataka te podjelom na pobjedničku i poraženu momčad, odnosno momčad domaćina i momčad gosta.



Slika 3.5. Osjetljivost NBA indeksa na doprinos pojedinačnih značajki vezana uz a) sva mjerenja b) pobjedničke momčadi c) poražene momčadi d) domaćina e) gosta.

NBA indeks je najosjetljiviji na značajku promašenih pokušaja za dva poena ($miss\_2fg$), a najmanje osjetljiv na značajku načinjenih blokada (bl). Tablica 3.12 potvrđuje prethodnu tvrdnju te prikazuje redoslijed osjetljivosti NBA indeksa na pojedinačne značajke za sva tri korištena pristupa.

Tablica 3.12. Osjetljivost NBA indeksa prema pojedinačnim značajkama.

#	Sve utakmice	Podjela pobjednik/poraženi		Podjela domaćin/gost	
		Pobjednik	Poraženi	Domaćin	Gost
1.	$miss\_2fg$	$miss\_2fg$	$miss\_2fg$	$miss\_2fg$	$miss\_2fg$
2.	ftm	ftm	ftm	$2fgm$	ftm
3.	$2fgm$	$2fgm$	$2fgm$	ftm	def_reb
4.	def_reb	$asist$	def_reb	$asist$	$asist$
5.	$asist$	def_reb	of_reb	def_reb	$2fgm$
6.	of_reb	$miss\_3fg$	$miss\_3fg$	$miss\_3fg$	of_reb
7.	$miss\_3fg$	of_reb	$asist$	of_reb	$miss\_3fg$
8.	$miss\_ft$	$miss\_ft$	to	$miss\_ft$	to
9.	to	to	$3fgm$	to	$3fgm$
10.	$3fgm$	$3fgm$	st	$3fgm$	$miss\_ft$
11.	st	st	$miss\_ft$	st	st
12.	bl	bl	bl	bl	bl

Pojedine pravilnosti lako su uočljive. NBA indeks je najosjetljiviji na značajku promašenih pokušaja za dva poena ($miss\_2fg$), a najmanje osjetljiv na značajku broja načinjenih blokada (bl). Također je jasno vidljivo kako u značajnije značajke spadaju i broj pogođenih slobodnih bacanja (ftm) i broj pogođenih pokušaja za dva poena ($2fgm$), dok značajka broja osvojenih lopti (st) uz već navedenu značajku broja načinjenih blokada (bl) najmanje utječe na NBA indeks. Najznačajnija je svakako analiza u kojoj su uključena sva mjerenja.

3.6. Usporedba metoda validacije i načina korištenja podataka

U ovom potpoglavlju analizirat će se rezultati predviđanja ishoda korištenjem osnovnih metoda nadziranog strojnog učenja kao što su logistička regresija (engl. *logistic regression*), Naivni Bayes (engl. *Naive Bayes*), stabla odluke (engl. *decision trees*, varijanta J-48), vrsta neuronske mreže naziva višeslojni perceptron (engl. *multilayer perceptron*), slučajna šuma (engl. *random forest*), metoda najbližih susjeda (engl. *k-nearest neighbours*) i LogitBoost. Osim analize rezultata, cilj je definirati i metodu validacije koja će dati najbolje rezultate predviđanja. Istraživanje će biti provedeno u programskom alatu Waikanato Environment for Knowledge Analysis (WEKA), a usporedit će se rezultati predviđanja korištenjem metoda podjele skupa podataka i unakrsne provjere. Najprije će se usporediti rezultati korištenjem metode podjele skupa podataka i metode unakrsne provjere gdje će se koristiti fiksno definirani skupovi podataka za učenje i ispitivanje. Problem metode unakrsne provjere je što se kod učenja sustava ponekad koriste i stvarni podaci, u tom trenutku, budućih događaja, a sportski događaji, samim time i košarkaške utakmice, nisu u

potpunosti nezavisni događaji stoga je nemoguće ili vrlo teško pripremiti podatke za fazu učenja. Neki od razloga su ozljede ili odmaranje igrača, format natjecanja (prvenstveno se odnosi na format natjecanja koji uključuje eliminaciju), taktičke promjene (promjene trenera, promjene načina igre) itd. Uvest će se i pojam aktualnih podataka (engl. *up-to-date*), što znači da će se tijekom faze ispitivanja vršiti prilagodba modela na način da će se podaci faze ispitivanja nad kojima je predviđanje već obavljeno koristi za učenje. Drugim riječima, poznati podaci faze ispitivanja koristit će se u kasnijim iteracijama faze učenja. Konačan cilj potpoglavlja prikazati je koliku točnost predviđanja je moguće dobiti korištenjem osnovnih metoda nadziranog strojnog učenja, a samim time i definirati metodu validacije koju će koristiti kasnije predložena metoda (model) predviđanja ishoda.

Waikato Environment for Knowledge Analysis (Weka)

Weka je program otvorenog koda pod GNU općom javnom licencom [72]. Osmišljen je na Sveučilištu Waikato na Novom Zelandu. U potpunosti je napisan u programskom jeziku Java te je samim time dostupan na svim platformama. Weka je zapravo kolekcija algoritama strojnog učenja koji se koriste za dubinsku analizu podataka, a sadrži alate za obradu, klasifikaciju, regresiju i vizualizaciju podataka te podržava duboko učenje.

Cilj potpoglavlja je kroz usporedbu rezultata klasifikacijskih metoda strojnog učenja definirati metodu validacije i način korištenja podataka koji će dati najbolje rezultate predviđanja korištenjem predložene metode (modela). Koristit će se jedna do tri sezone skupa za učenje i jedna do dvije sezone skupa za ispitivanje. Razlog nekorištenja većeg skupa podataka su rezultati rada [10] koji su pokazali da najbolje rezultate predviđanja daje korištenje maksimalno tri sezone skupa za učenje i jedne sezone skupa za ispitivanje. Tako će se usporediti dva načina korištenja podataka, fiksno definirani skupovi podataka i aktualni podaci te dvije metode validacije, metoda podjele skupa podataka i unakrsna provjera, definirane u potpoglavlju 2.6. Pošto košarkaška utakmica nudi dva moguća ishoda koristit će se binarna klasifikacija. U ovom primjeru i ostalim primjerima vezanim uz osnovne metode strojnog učenja će se koristiti prosječni podaci faze učenja, skup od 13 osnovnih elemenata košarkaške statistiske te klasifikacijska značajka. Vektor značajki (engl. *feature vector*) koji se sastoji od statistike domaće i gostujuće momčadi te klasifikacijske značajke prikazan je formulom (3-23) gdje f_{tm_A}, f_{tm_B} i f_C označuju redom: vektor značajki domaće momčadi, vektor značajki gostujuće momčadi i klasifikacijsku značajku.

$$f = f_{tm_A} + f_{tm_B} + f_C \tag{3-23}$$

Statistika momčadi predstavljena vektorom $\overrightarrow{tm}$ se sastoji od 13 značajki (elemenata osnovne košarkaške statistike). Vektor značajki momčadi prikazan je formulom (3-24).

$$\overrightarrow{tm} = \begin{bmatrix} tm_{2fgm}, tm_{miss\_2fg}, tm_{3fgm}, tm_{miss\_3fg}, tm_{ftm}, tm_{miss\_ft}, \\ tm_{def\_reb}, tm_{of\_reb}, tm_{asist}, tm_{st}, tm_{to}, tm_{bl}, tm_{f} \end{bmatrix} \quad (3\text{-}24)$$

Podaci koje pojedina komponenta vektora značajki sadrži ovise o fazi strojnog učenja. Komponenta vezana uz učenje modela sadrži stvarne podatke odigranih utakmica, dok komponenta vezana uz ispitivanje modela sadrži prosječan učinak momčadi tijekom faze učenja. Komponenta vektora značajki vezana uz fazu ispitivanja može sadržavati i bilo koji drugi poznati podatak na temelju poznate povijesti ili iskustva eksperta.

3.6.1. Predviđanja korištenjem metode podjele skupa podataka

Prvi rezultati predviđanja vezani su uz korištenje fiksno definiranih skupova za učenje i ispitivanje te metode podjele skupa podataka. Ulazni skup podataka će se podijeliti na dva kronološki poredana skupa podataka, skup za učenje i skup za ispitivanje, gdje skup za učenje prethodi skupu za ispitivanje. U svakom trenutku će se koristiti međusobno disjunktni skupovi. Ukoliko cijeli skup podataka označimo kao $\mathcal{D}$, skup za učenje oznakom $\mathcal{D}_U$, a skup za ispitivanje $\mathcal{D}_I$, navedene skupove podataka može se označiti kao $\mathcal{D} = \mathcal{D}_U \cup \mathcal{D}_I$.

Skup za učenje sadrži stvarne statističke podatke utakmica, dok skup za ispitivanje sadrži prosječnu statistiku momčadi tijekom definiranog vremenskog perioda. Problem ovog pristupa je korištenje istog skupa za učenje, definiranog neposredno prije predviđanja kronološki prvog događaja, tijekom cijele faze ispitivanja te se samim time očekuju lošiji rezultati u odnosu na pristup koji će biti korišten u kasnijim odjeljcima. Pošto su utakmice u pravilu kronološki poredane, moguća je prilagodba skupa za učenje neposredno prije predviđanja svakog sportskog događaja. Tablica 3.13 prikazuje rezultate predviđanja ishoda košarkaških utakmica korištenjem fiksno definiranih skupova za učenje i ispitivanje te metode podjele skupa podataka.

Tablica 3.13. Rezultati predviđanja korištenjem fiksno definiranih skupova i metode podjele skupa podataka.

Br. sezona skupa za učenje	Br. sezona skupa za ispit.	Logist. regr.	Naivni Bayes	Stabla odluke	Višesl. perc.	Metoda najbl. susjeda	Slučajna šuma	LogitBoost
1	1	57,09 %	57,40 %	55,03 %	57,13 %	58,94 %	57,96 %	56,46 %
1	2	55,53 %	55,76 %	53,75 %	55,64 %	57,76 %	56,94 %	55,31 %
2	1	56,47 %	57,20 %	55,16 %	56,32 %	59,04 %	56,94 %	54,48 %
2	2	56,01 %	54,97 %	53,66 %	55,86 %	57,33 %	55,39 %	53,56 %
3	2	55,62 %	53,65 %	49,87 %	55,58 %	56,42 %	54,14 %	52,84 %

Prosječno najlošije rezultate predviđanja daju stabla odluke, a prosječno najbolje rezultate daje metoda najbližih susjeda. Važno je napomenuti da se istraživanje koristi samo kao polazna točka koja prikazuje koliko uspješno mogu osnovne klasifikacijske metode nadziranog strojnog učenja predviđati ishode utakmica korištenjem metode podjele skupa podataka te prosječnog NBA

indeksa momčadi. Sukladno tome moći će se usporediti rezultati predložene metode predviđanja ishoda u odnosu na osnovne metode strojnog učenja.

Podrobnijom analizom rezultata je vidljivo da najbolji pojedinačni rezultat predviđanja koji iznosi 58,94 % daje korištenje metode najbližih susjeda korištenjem jedne sezone skupa za učenje i jedne sezone skupa za ispitivanje. Od ostalih korištenih metoda, najbolji pojedinačni rezultati su također postignuti korištenjem jedne sezone skupa za ispitivanje. Time su rezultati rada [10], koji pokazuju da se najbolji rezultati predviđanja košarkaških utakmica dobivaju korištenjem jedne do tri sezone skupa za učenje i jedne sezone skupa za ispitivanje, potvrđeni.

3.6.2. Predviđanje korištenjem unakrsne provjere

U ovom odjeljku, u odnosu na odjeljak 3.6.1, koristit će se metoda unakrsne provjere. Cilj odjeljka je usporediti rezultate upotrebe osnovnih metoda nadziranog strojnog učenja korištenjem metode podjele skupa podataka i metode unakrsne provjere na fiksno definiranim skupovima za učenje i ispitivanje. Košarkaške utakmice, kao i utakmice u ostalim sportovima, nisu u potpunosti neovisni događaji te ishodi prethodnih utakmica u velikoj mjeri utječu na buduće utakmice. Metoda unakrsne provjere u fazi učenja koristi i buduće događaje. Pretpostavka je da će rezultati korištenja metode unakrsne provjere biti bolji u odnosu na metodu podjele skupa podataka. Razlog tome je korištenje budućih događaja u predviđanju prethodnih koji u velikoj mjeri utječu na buduće događaje.

Tablica 3.14 prikazuje rezultate predviđanja korištenjem fiksno definiranih skupova i metode unakrsne provjere.

Tablica 3.14. Rezultati predviđanja korištenjem fiksno definiranih skupova i metode unakrsne provjere.

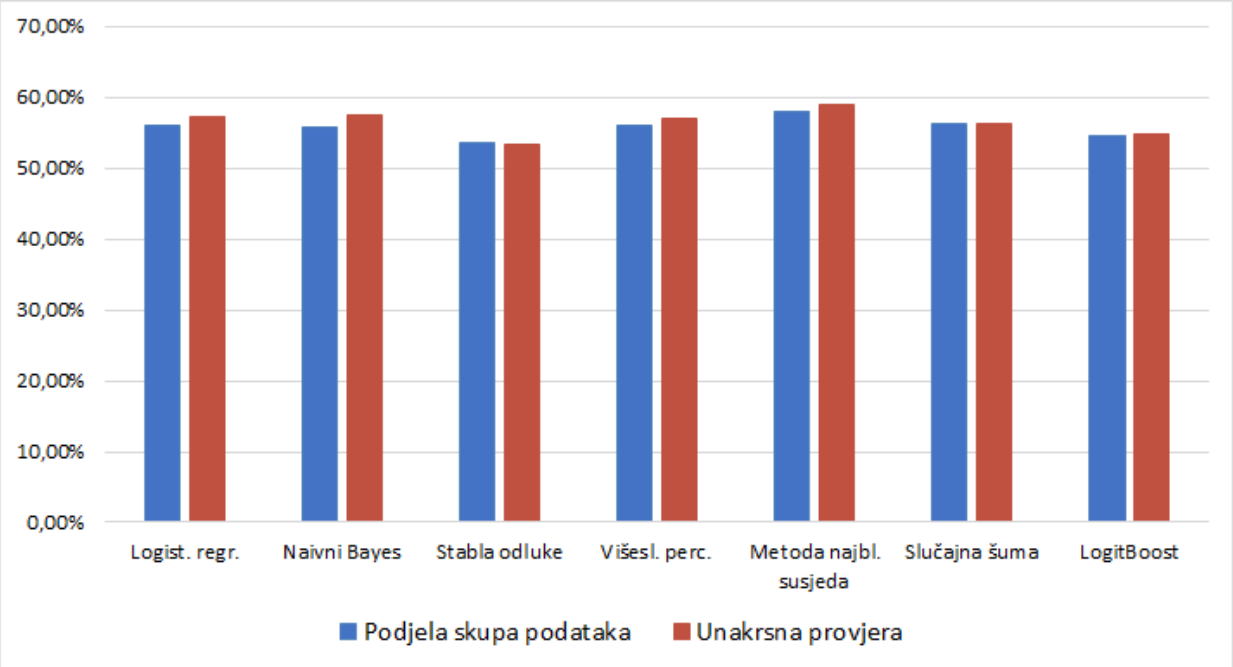
Ulazni skup podataka (broj sezona)	Logist. regr.	Naivni Bayes	Stabla odluke	Višesl. perc.	Metoda najbl. susjeda	Slučajna šuma	LogitBoost
2 sezone	58,14 %	58,67 %	55,07 %	57,89 %	60,12 %	58,74 %	55,78 %
3 sezone	57,50 %	58,08 %	53,61 %	57,38 %	59,46 %	57,23 %	55,83 %
4 sezone	57,01 %	57,54 %	52,95 %	56,95 %	58,53 %	55,85 %	54,80 %
5 sezona	56,02 %	56,00 %	51,85 %	56,00 %	57,69 %	53,59 %	52,50 %

Kao i kod metode podjele skupa podataka prosječno najlošije rezultate daju stabla odluke, dok su najbolji rezultati postignuti metodom najbližih susjeda. Najbolji pojedinačni prosječni rezultat dobiven je metodom najbližih susjeda korištenjem dvije sezone. Ostale korištene metode strojnog učenja također su najbolje rezultate polučile korištenjem dvije sezone.

3.6.3. Usporedba rezultata metode podjele skupa podataka i metode unakrsne provjere

U ovom odjeljku će se usporediti rezultati predviđanja korištenjem metode podjele skupa podataka i metode unakrsne provjere korištenjem fiksno definiranih skupova za učenja i ispitivanje. U analizi će se koristiti prosječne vrijednosti dobivene za pojedinu metodu strojnog

učenja. Slika 3.6 daje usporedbu prosječnih točnosti algoritma korištenjem metode podjele skupa podataka i metode unakrsne provjere. Slikom su prikazani prosječni rezultati korištenjem metode podjele skupa podataka iz odjeljka 3.6.1 i prosječnih podataka korištenjem metode unakrsne provjere iz odjeljka 3.6.2. Metoda podjele skupa podataka koristi najviše tri sezone skupa za učenje i 2 sezone skupa za ispitivanje, dok metoda unakrsne provjere dijeli ulazni skup na pet odvojenih skupova ($k = 5$).



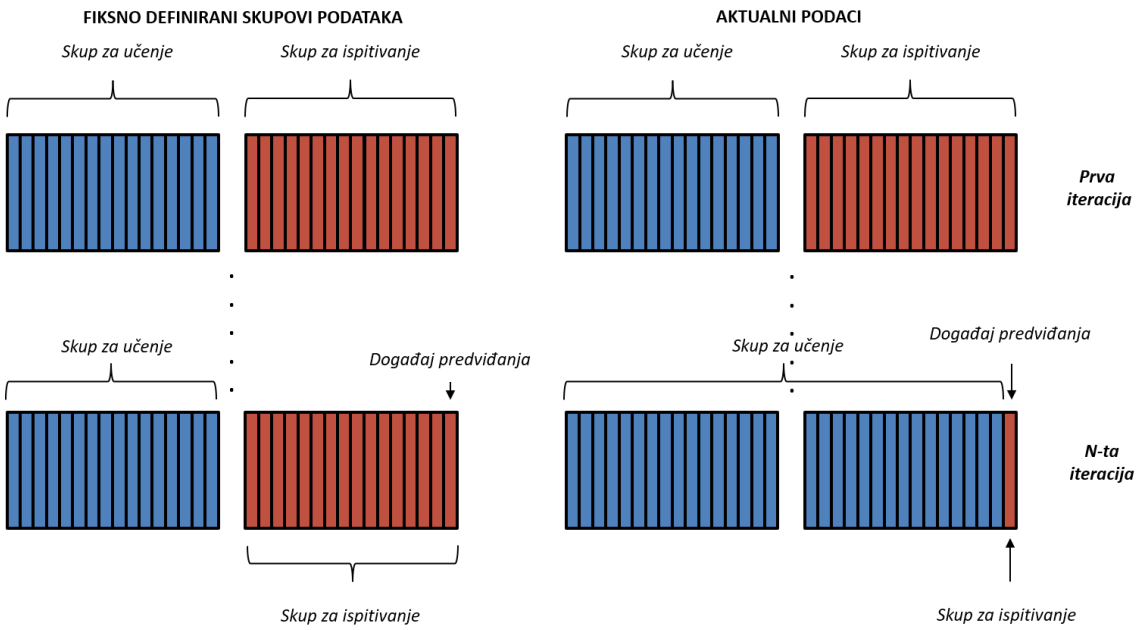
Slika 3.6. Usporedba rezultata korištenjem metode podjele skupa podataka i unakrsne provjere.

Pretpostavka da će rezultati korištenjem metode podjele skupa podataka biti lošiji u odnosu na metodu unakrsne provjere se pokazala točnom. Slika jasno pokazuje kako sve korištene metode strojnog učenja, osim stabla odluke, bolje rezultate predviđanja postižu korištenjem metode unakrsne provjere.

Svrha usporedbe metoda validacija je provjeriti daju li različiti pristupi u metodama validacije različite rezultate, a što se pokazalo točnim. Bolji prosječni rezultati su dobiveni korištenjem metode unakrsne provjere. Metoda unakrsne provjere koristi buduće podatke, a pošto sportski događaji nisu u potpunosti neovisni, upotreba unakrsne provjere nije poželjna jer, zbog upotrebe budućih događaja u predviđanju prošlih, ne oslikava stvarne mogućnosti modela predviđanja, tj. u stvarnosti, tijekom primjene modela predviđanja, budući podaci nisu dostupni pa bi samim time i točnost takvog modela vjerojatno bila manja u odnosu na vrijednosti dobivene tijekom faze ispitivanja. Zbog svega gore obrazloženog, u ovom istraživanju koristit će se metoda podjele skupa podataka.

3.6.4. Predviđanje korištenjem aktualnih podataka i metode podjele skupa podataka

U prethodnom odjeljku su uspoređeni rezultati korištenjem dvije metode validacije nad fiksno definiranim skupovima podataka. Rezultati korištenja fiksno definiranih skupova pokazali su da bolje rezultate predviđanja daje metoda unakrsne provjere, ali je i navedeno da metodu unakrsne provjere nije poželjno koristiti iz objektivnog razloga, a to je da metoda unakrsne provjere koristi buduće podatke u predviđanju prošlih. U ovom odjeljku će se koristiti aktualni podaci, a dobiveni rezultati će se usporediti s rezultatima dobivenim korištenjem disjunktnih skupova. Pojam aktualnih podataka smatra prilagodbu modela na način da će se u fazi učenja koristiti i podaci skupa za ispitivanje za koje je predviđanje već obavljeno. Tako će se prilikom svake iteracije skupu podataka za učenje, koji sadrži isključivo stvarne podatke poznatih događaja, dodavati i stvarni podaci događaja faze ispitivanja. Skup podataka za učenje će se tako proširivati te samim time nuditi dulji pogled u prošlost. Učenje tako više neće biti strogo nadzirano, već će poprimiti i elemente učenja uz podršku. Točnije, skupu za učenje bit će priključeni i poznati podaci faze ispitivanja te će sustav moći učiti na temelju metode pogodaka i pogrešaka. Pretpostavka je da će rezultati korištenja aktualnih podataka dati bolje rezultate u odnosu kada se koriste fiksno definirani skupovi za učenje i ispitivanje. Pogled u prošlost se neće ograničavati u niti jednom trenutku već će se skup podataka za učenje svakom iteracijom povećavati. Ograničavanje skupa za učenje te sam utjecaj ograničavanja pogleda u prošlost će se razmotriti u kasnijim poglavljima. Slika 3.7 grafički prikazuje metode podjele skupa podataka i dva načina pripreme podataka.



Slika 3.7. Načini pripreme podataka.

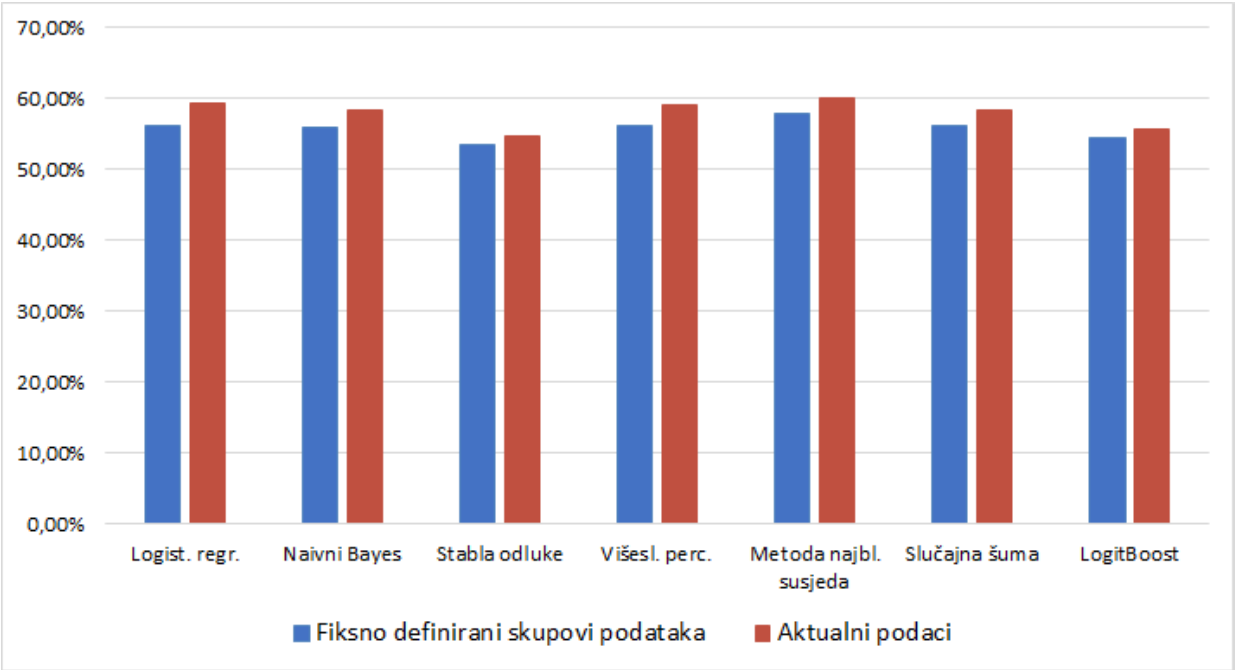
Način korištenja aktualnih podataka je složeniji i računalno skuplji u odnosu na fiksno definirane skupove za učenje i ispitivanje. Pretpostavka je da će način pripreme podataka korištenjem aktualnih podataka dati bolje rezultate u odnosu na fiksno definirane skupove. Tablica 3.15 prikazuju rezultate dobivene korištenjem aktualnih podataka i metode podjele skupa podataka.

Tablica 3.15. Rezultati predviđanja korištenjem aktualnih podataka i metode podjele skupa podataka.

Br. sezona skupa za učenje	Br. sezona skupa za ispit.	Logist. regr.	Naivni Bayes	Stabla odluke	Višesl. perc.	Metoda najbl. susjeda	Slučajna šuma	LogitBoost
1	1	59,29 %	59,22 %	54,97 %	58,23 %	60,06 %	59,56 %	57,55 %
1	2	58,97 %	58,03 %	55,10 %	58,70 %	59,23 %	58,63 %	56,24 %
2	1	59,97 %	58,58 %	54,85 %	59,97 %	60,82 %	58,92 %	55,82 %
2	2	59,47 %	57,77 %	54,20 %	59,46 %	59,87 %	57,50 %	54,57 %
3	2	59,44 %	57,58 %	54,18 %	59,50 %	60,06 %	56,60 %	54,39 %

Najbolji prosječni rezultati dobiveni su korištenjem metode najbližih susjeda. Ostale metode, osim stabla odluke, najbolje rezultate postigle su korištenjem jedne sezone skupa za ispitivanje i jedne ili dvije sezone skupa za učenje. Time su rezultati rada [10], koji pokazuju da se najbolji rezultati predviđanja košarkaških utakmica dobivaju korištenjem jedne do tri sezone skupa za učenje i jedne sezone skupa za ispitivanje, ponovo potvrđeni.

Za predlaganje budućeg modela predviđanja vrlo je važno usporediti i rezultate dobivene korištenjem različitih duljina fiksno definiranih skupova za učenje i ispitivanje te korištenjem poznatih podataka skupa za ispitivanje tijekom faze učenja. Slika 3.8 prikazuje dobivene rezultate.



Slika 3.8. Usporedba prosječnih rezultata predviđanja korištenjem različitih duljina fiksno definiranih skupova za učenje i ispitivanje i aktualnih podataka na temelju metode podjele skupa podataka.

Bolji rezultati predviđanja su očekivano dobiveni korištenjem poznatih podataka faze ispitivanja tijekom faze učenja. Pretpostavka ostvarivanja boljih rezultata korištenjem podataka faze ispitivanja za koje je predviđanje obavljeno u prilagodbi modela pokazala se točnom.

3.7. Predviđanje ishoda na temelju prosječnih učinaka

Jedini indikator proglašenja pobjednika u sportu je konačan rezultat. Svaki sport ima jedinstven način bodovanja i praćenja rezultata. Konačan rezultat ne treba nužno biti u vidu golova ili poena, već može biti primjerice i u obliku vremena ili udaljenosti. Analizirani skup podataka koristi broj poena pa je istraživanje prilagođeno specifičnostima takve vrste problema.

U prethodnim potpoglavljima se za predviđanje koristila cijela poznata povijest. Cilj ovog potpoglavlja pokazati je kako duljina povijesnih podataka utječe na predviđanje ishoda korištenjem različitog skupa značajki. Tijekom postupka ispitivanja vršit će se prilagodba modela na način da se skupu za učenje dodaju i podaci skupa za ispitivanje za koje je predviđanje već obavljeno. Predviđanja će se vršiti na temelju prosječnog broja poena i prosječnog NBA indeksa definiranog u potpoglavlju 2.3.

3.7.1. Predviđanje ishoda korištenjem prosječnog broja postignutih poena

Pretpostavka je kako će rezultati korištenja prosječnog broja poena biti lošiji od rezultata korištenja prosječnog indeksa korisnosti. U prethodnim potpoglavljima je pokazano kako duljina skupa za učenje utječe na rezultate predviđanja, ali i da postoji prednost domaćeg terena koja je još izraženija tijekom faze doigravanja. Broj utakmica faze doigravanja prosječno iznosi 6,47 % od ukupnog broja utakmica, što čini vrlo mali udio u skupu svih utakmica. Tablica 3.1 prikazuje omjer broja utakmica doigravanja u odnosu na broj utakmica po analiziranim sezonama. Prilikom predviđanja će se koristiti cijela poznata povijest, a kao logične cjeline su odabrane natjecateljske sezone. Cilj je pokazati kako segmentiranje sezone može dovesti do boljih rezultata predviđanja, odnosno pokazati kako postoji vremenski period koji najbolje opisuje momčad, a samim time i daje najbolje rezultate predviđanja. Skup za učenje činit će n odigranih utakmica pojedine momčadi, gdje je $n \in [1, N]$, a N ukupan broj utakmica skupa za učenje. NBA momčad tijekom jedne sezone odigra 82 utakmice regularnog dijela i eventualne utakmice doigravanja. Ukoliko se u obzir uzme jedna sezona skupa za učenje prilikom predviđanja ishoda prve utakmice n će biti broj u intervalu $n = \{1, \dots 82\}$ u slučaju kada momčad nije izborila doigravanje, odnosno $n = \{1, \dots 82, \dots, 82 + m_{tm}\}$ u slučaju kada je momčad izborila doigravanje, gdje m_{tm} predstavlja broj utakmica momčadi tm tijekom faze doigravanja. Broj utakmica faze doigravanje nije jednak za sve momčadi, već ovisi o uspješnosti same momčadi. Skupu za učenje će se dodavati i poznati

podaci utakmica skupa za ispitivanje nad kojima je predviđanje izvršeno. Ukoliko je k_{tm} broj utakmica skupa za ispitivanje momčadi tm nad kojima je predviđanje već izvršeno, skup za učenje sadrži $n = 82 + m_{tm} + k_{tm}$ utakmica ukoliko se početni skup za učenje sastoji od jedne natjecateljske sezone. Pretpostavka je da će točnost predviđanja rasti povećanjem duljine skupa za učenje, dostići maksimum te nakon toga početi lagano opadati kako će se duljina skupa za učenje povećavati. Skup za učenje i skup za ispitivanje moraju biti kronološki poredani. Predviđanje korištenjem broja postignutih poena koristi jednu izlučenu značajka nazvanu poeni (pts). Izlučena značajka je suma umnožaka pripadajućeg doprinosa značajke i prosječnog učinka tri elementa osnovne košarkaške statistike. Formula (3-25) prikazuje formulu izračuna značajke pts, gdje $2fgm$ predstavlja broj pogođenih pokušaja za dva poena, $3fgm$ broj pogođenih pokušaja za tri poena i ftm broj pogođenih slobodnih bacanja.

$$N_{pts} = 2 \times N_{2fgm} + 3 \times N_{3fgm} + N_{ftm} \tag{3-25}$$

Formula (3-26) prikazuje način izračuna ishoda događaja temeljen na prosječnom broju poena. Valja napomenuti kako je prosječan broj poena u pravilu realan broj, a u slučaju jednakih vrijednosti za domaćina i gosta pobjednikom se proglašava domaća momčad. Razlog proglašenja domaće momčadi pobjedničkom momčadi objašnjen je u odjeljku 3.3.1.

$$winn(pts(tm_d), pts(tm_g)) \begin{cases} tm_d, \dfrac{\frac{1}{n}\sum_{i=1}^{n} pts(tm_d, i)}{\frac{1}{n}\sum_{i=1}^{n} pts(tm_g, i)} \geq 1; \\ tm_g, \dfrac{\frac{1}{n}\sum_{i=1}^{n} pts(tm_d, i)}{\frac{1}{n}\sum_{i=1}^{n} pts(tm_g, i)} < 1\,; \end{cases} \tag{3-26}$$

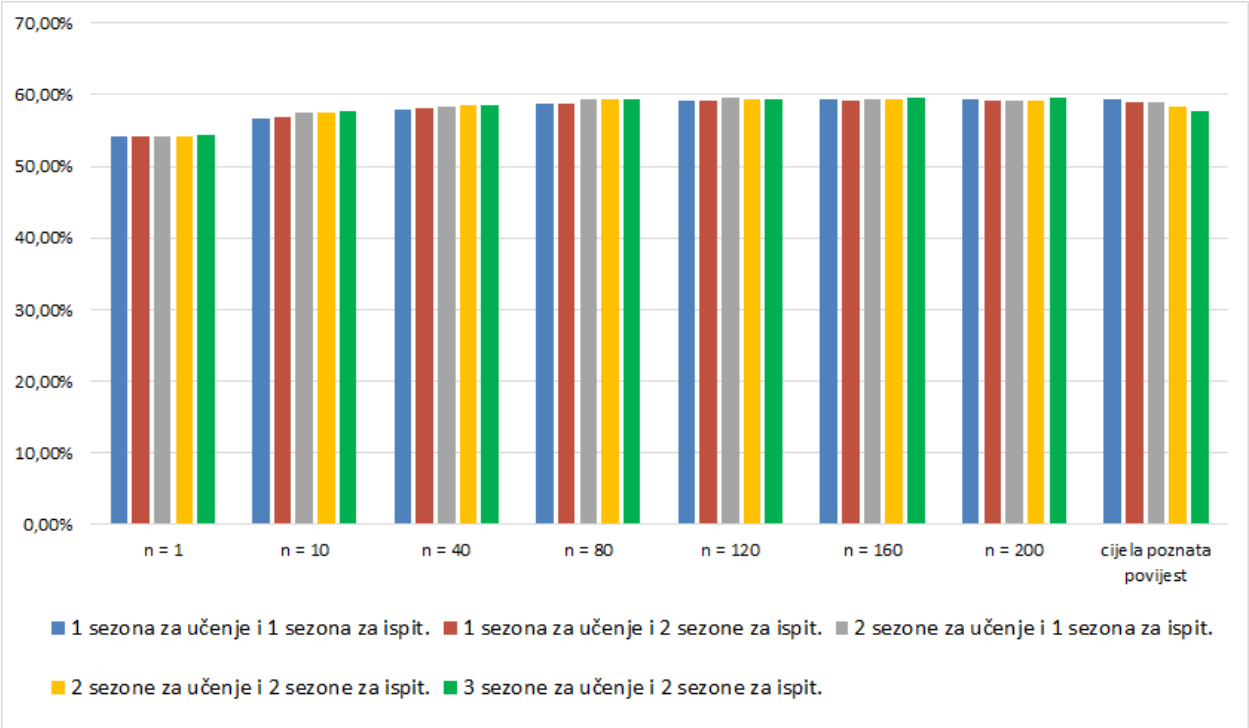
Tablica 3.16 prikazuje rezultate predviđanja na temelju definiranog broja utakmica skupa za učenje ($n \leq N$).

Tablica 3.16. Točnost predviđanja ovisna o duljini skupa za učenje i prosječno zabijenih poena.

Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	$n = 1$	$n = 10$	$n = 40$	$n = 80$	$n = 120$	$n = 160$	$n = 200$	**cijela povijest**
1	1	54,20 %	56,75 %	58,00 %	58,73 %	59,08 %	59,36 %	59,36 %	59,36 %
1	2	54,12 %	56,94 %	58,12 %	58,83 %	59,12 %	59,25 %	59,11 %	58,99 %
2	1	54,19 %	57,46 %	58,35 %	59,31 %	59,51 %	59,41 %	59,17 %	58,92 %
2	2	54,19 %	57,51 %	58,46 %	59,38 %	59,44 %	59,46 %	59,24 %	58,34 %
3	2	54,28 %	57,76 %	58,63 %	59,36 %	58,46 %	59,59 %	59,51 %	57,64 %

Prosječna točnost predviđanja raste povećanjem broja utakmica skupa za učenje, dostiže svoj maksimum i počinje lagano opadati. Važno je napomenuti da korištenje jedne sezone skupa za učenje i jedne sezone skupa za ispitivanje u nekim kombinacijama ne sadrži dovoljan broj utakmica skupa za učenje pa se samim time rezultati ponavljaju u slučaju kada je $n > N$, tj. kada je analizirani broj utakmica veći od ukupnog broja utakmica skupa za učenje i skupa za ispitivanje.

Slika 3.9 prikazuje trend kretanja točnosti predviđanja korištenjem prosječnog broja postignutih poena.



Slika 3.9. Trend kretanja točnosti predviđanja korištenjem prosječnog broja poena i različitih duljina povijesti.

Slika 3.9 potvrđuje pretpostavke, što znači da prosječna točnost povećanjem broja utakmica skupa za učenje raste, dostiže svoj maksimum te počinje lagano opadati. Pouzdaniji opis procesa, u ovom slučaju košarkaške utakmice, će se u nastavku pokušati dobiti korištenjem većeg skupa značajki.

3.7.2. Predviđanje na temelju prosječnog NBA indeksa

U odjeljku 3.4.2 je prikazano kako NBA indeks s definiranom prednošću domaćeg terena prosječno daje 92,31 % informacija o ishodu utakmica. Cilj ovog odjeljka pokazati je kako duljina povijesnih podataka utječe na predviđanje ishoda korištenjem prosječnog NBA indeksa. Priprema podataka će se vršiti na isti način kao i u odjeljku 3.7.1. Matematički opis izračuna pobjednika prikazan je formulom (3-27).

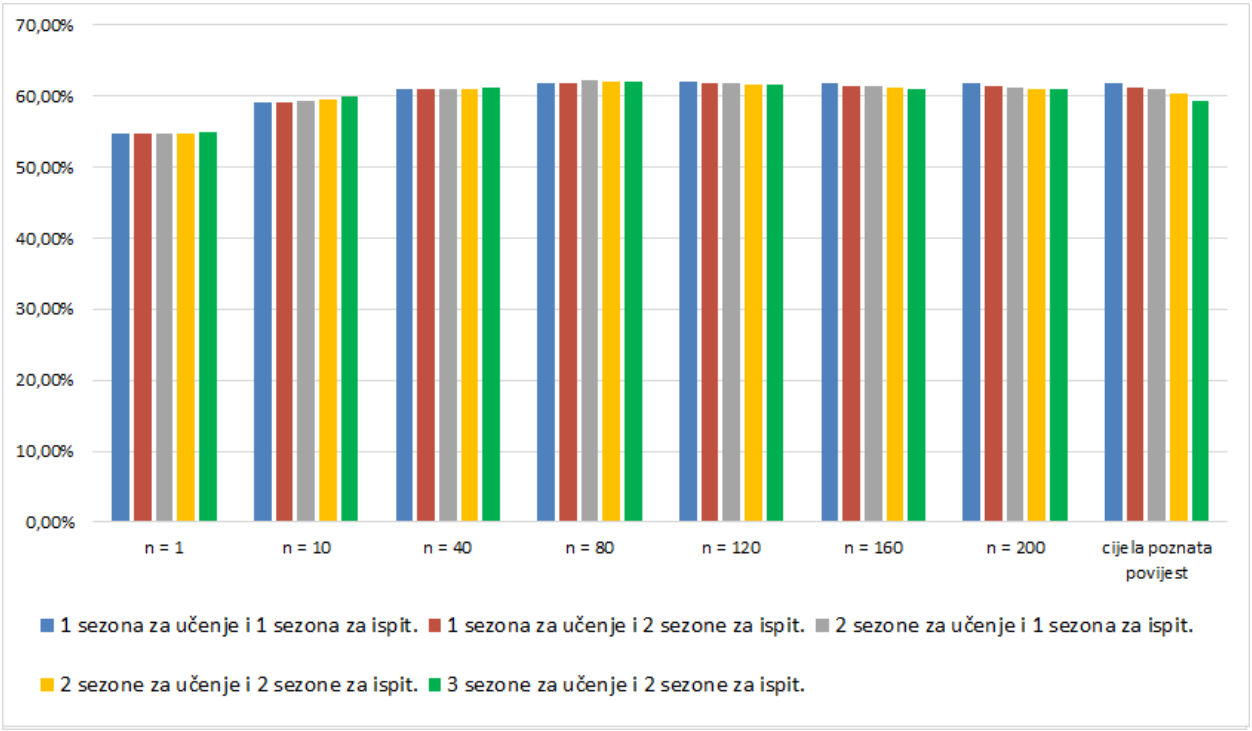
$$winn(I_{NBA}(tm_d), I_{NBA}(tm_g)) \begin{cases} tm_d, \dfrac{\frac{1}{n}\sum_{i=1}^{n} I_{NBA}(tm_d, i)}{\frac{1}{n}\sum_{i=1}^{n} I_{NBA}(tm_g, i)} \geq 1; \\ tm_g, \dfrac{\frac{1}{n}\sum_{i=1}^{n} I_{NBA}(tm_d, i)}{\frac{1}{n}\sum_{i=1}^{n} I_{NBA}(tm_g, i)} < 1; \end{cases} \quad (3\text{-}27)$$

Tablica 3.17 prikazuje rezultate predviđanja korištenjem prosječnog NBA indeksa. Kao i u varijanti korištenja samo prosječnog broja postignutih poena, pretpostavka je da će prosječna točnost korištenjem NBA indeksa s povećanjem duljine poznate povijesti rasti, dostići maksimum te početi lagano opadati. Isto tako, pretpostavka je da će duži pogled u povijest (dvije ili više sezona) dati lošije rezultate od varijante koja koristi jednu sezonu skupa za učenje i aktualne podatke skupa za ispitivanje.

Tablica 3.17. Točnost predviđanja ovisna o duljini skupa za učenje i prosječnog NBA indeksa.

Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	$n = 1$	$n = 10$	$n = 40$	$n = 80$	$n = 120$	$n = 160$	$n = 200$	**cijela povijest**
1	1	54,81 %	59,09 %	60,98 %	61,90 %	61,95 %	61,87 %	61,80 %	61,80 %
1	2	54,75 %	59,18 %	61,00 %	61,90 %	61,76 %	61,43 %	61,40 %	61,28 %
2	1	54,81 %	59,40 %	60,97 %	62,17 %	61,83 %	61,32 %	61,28 %	61,04 %
2	2	54,79 %	59,57 %	61,05 %	62,01 %	61,60 %	61,12 %	61,05 %	60,31 %
3	2	54,99 %	59,90 %	61,12 %	62,01 %	61,53 %	61,07 %	60,89 %	59,21 %

Prosječna točnost predviđanja raste povećanjem duljine skupa za učenje, dostiže svoj maksimum i počinje lagano opadati. Kao i u odjeljku 3.7.1, korištenje jedne sezone skupa za učenje i jedne sezone skupa za ispitivanje u pojedinim varijantama korištenja parametra n ne sadrži dovoljan broj utakmica pa se samim time rezultati ponavljaju u slučaju kada je $n \geq N$, tj. kada je analizirani broj utakmica veći od ukupnog broja utakmica skupa za učenje i skupa za ispitivanje. Slika 3.10 prikazuje trend kretanja točnosti predviđanja korištenjem prosječnog NBA indeksa ovisnog o duljini ulaznog skupa podataka.



Slika 3.10. Trend kretanja točnosti predviđanja korištenjem prosječnog NBA indeksa i različitih duljina povijesti.

Slika 3.10 samo potvrđuje postavljene pretpostavke. Rezultati predviđanja korištenjem prosječnog NBA indeksa su u prosjeku dali bolje rezultate od korištenja izlučene značajke pts, što znači da je veći skup značajki dao bolje rezultate.

3.7.3. Optimizacija doprinosa sveobuhvatnog indeksa korisnosti

Problem optimizacije, odnosno problem pronalaženja najboljeg rješenja u matematičkom skupu mogućih rješenja u skupini je problema čije se rješavanje omogućilo pojavom računala, a dodatno olakšalo povećanjem računalne snage. Optimizacija se vrši iterativno, slijedeći zadane korake kojima se nastoji poboljšati prethodno rješenje. Bez obzira što je pojava računala i svakodnevno povećanje računalne snage olakšalo problem optimizacije, metode iscrpnog pretraživanja (engl. *exhaustive search*) prostora svih mogućih rješenja za određenu vrstu problema i dalje predstavljaju veliki problem. Razvijeni su različiti pristupi definiranja strategije pretraživanja prostora, a mogu se podijeliti na heurističke i metaheurističke metode (kraće heuristike i metaheuristike). Navedene metode optimizacije ne mogu garantirati optimalno rješenje, ali omogućuju pronalazak zadovoljavajućeg rješenja u razumnom vremenskom roku. Heuristike se koriste za određene probleme, dok se metaheuristike koriste za širok spektar različitih problema.

Kao što je već navedeno, početna točka predviđanja ishoda sportskih događaja biti će CPE indeks. Pošto u momčadskim sportovima više sudionika utječe na konačan ishod, potrebno je koristiti CTE indeks predstavljen u odjeljku 3.4.7.

Jedan od ciljeva rada je predložiti postupak optimizacije doprinosa sveobuhvatnog indeksa korisnosti, točnije optimizirati doprinose koeficijenta $v_e(v'_e)$ skupa elemenata E analiziranog procesa. Optimizacija će se vršiti iterativno, a uspješnost procesa optimizacije procijenit će se na temelju rezultata predviđanja.

3.7.4. Predviđanje na temelju optimiziranog indeksa CTE

Rezultati predviđanja korištenjem osnovnog NBA indeksa su dani u odjeljku 3.7.2. Najbolji rezultati su dobiveni korištenjem jedne sezone skupa za učenje i jedne sezone skupa za ispitivanje. Povećanjem ulaznog skupa podataka, točnije povećanjem broja utakmica skupa za učenje, točnost je rasla, dostigla svoj maksimum te počela opadati. Navedena činjenica je pokazala da postoji optimalni vremenski period koji daje najbolje rezultate predviđanja. Jedan od ciljeva rada je koristiti optimizirani indeks CTE i na taj način dobiti bolje rezultate u odnosu na indeks NBA.

Istraživanje će se provesti kronološkom podjelom ulaznog skupa podataka na skup za učenje i skup za ispitivanje, a samo predviđanja će se vršiti na temelju prosječnih podataka skupa za učenje. Skup za učenje će se dodatno proširivati podacima skupa za ispitivanje nad kojima je predviđanje

izvršeno. Pretpostavka je da će rezultati optimiziranog indeksa CTE biti bolji u odnosu na osnovni NBA indeks i modificirani NBA indeks. Formula (3-29) prikazuje način izračuna ishoda gdje n predstavlja broj utakmica skupa za učenje.

$$winn(I_{CTE}(tm_d), I_{CTE}(tm_g)) \begin{cases} tm_d, \dfrac{\frac{1}{n}\sum_{i=1}^{n} I_{CTE}(tm_d, i)}{\frac{1}{n}\sum_{i=1}^{n} I_{CTE}(tm_g, i)} \geq 1; \\ tm_d, \dfrac{\frac{1}{n}\sum_{i=1}^{n} I_{CTE}(tm_g, i)}{\frac{1}{n}\sum_{i=1}^{n} I_{CTE}(tm_g, i)} < 1\ ; \end{cases} \quad (3\text{-}28)$$

Pretpostavka je da će najbolji rezultati biti dobiveni korištenjem jedne sezone skupa za učenje i jedne sezone skupa za ispitivanje. Također će biti važno i pokazati kako bolji rezultati korištenjem stvarnih podataka za definiranje pobjednika ne znače nužno i bolje rezultate predviđanja ishoda nad povijesnim podacima.

3.7.5. Uvođenje značajke prednosti domaćeg terena

U odjeljku 3.3.1 je pokazano kako postoji prednost domaćeg terena te je prikazan način izračuna. Cilj ovog odjeljka je predložiti uvođenje dodatne značajke koja će predstavljati prednost domaćeg terena. Kao referentna točka definiranja prednosti domaćeg terena koristit će se razlika postotka pobjeda domaćih momčadi u odnosu na gostujuće, a izračunavat će se na temelju poznate povijesti. U kasnijim iteracijama skupu za učenje će se dodavati i podaci skupa za ispitivanje nad kojima je predviđanje izvršeno. Značajka prednosti domaćeg terena će se ponovno izračunavati prilikom svake sljedeće iteracije algoritma predviđanja. Formula (3-29) prikazuje način izračuna značajke prednosti domaćeg terena.

$$\Delta tm_N = \frac{N_{tm_d}}{N_{tm_d} + N_{tm_g}} - \frac{N_{tm_g}}{N_{tm_d} + N_{tm_g}} \quad (3\text{-}29)$$

Osnovna ideja korištenja značajke prednosti domaćeg terena je povećanje projiciranog CTE indeksa domaće momčadi. Ukoliko projicirani CTE indeks domaće momčadi označimo kao $I_{CTE}(tm_d)$, a projicirani CTE indeks gostujuće momčadi kao $I_{CTE}(tm_g)$, projicirani CTE indeks domaćina ćemo povećati za razliku umnoška postotka pobjede i pripadnog doprinosa zapisanog $I_{CTE}(tm_d(dt))$. Pripadni koeficijent doprinosa razlike postotka, točnije korektivni faktor, će se zapisati kao kf. Formulom (3-30) je prikazan projicirani indeks domaće momčadi pomnožen s razlikom postotka pobjeda domaćina u odnosu na gosta ($I_{CPE}(tm_d(dt))$. Projicirani indeks gostujuće momčadi ostaje nepromijenjen.

$$I_{CTE}(tm_d(dt)) = I_{CTE}(tm_d) + I_{CTE}(tm_d) \times kf \times \Delta tm_N, kf \geq 0 \quad (3\text{-}30)$$

Kao što je vidljivo iz formule (3-30), korektivni faktor (kf) treba biti broj veći ili jednak 0. Optimalni korektivni faktor će se izračunati korištenjem skupa linearnih i nelinearnih funkcija. Kao i u prethodnim istraživanjima pobjednikom će se proglasiti ekipa s većim projiciranim indeksom korisnosti. Formulom (3-31) prikazan je način definiranja pobjedničke momčadi. Pobjeda domaćina definirana je u slučaju kada navedeni omjer daje broj veći ili jednak jedan, dok se pobjedom gosta smatra rezultat manji od jedan.

$$winn(I_{CTE}(tm_d(dt)), I_{CTE}(tm_g(dt))) \begin{cases} tm_d, \dfrac{\frac{1}{n}\sum_{i=1}^{n} I_{CTE}(tm_d(dt), i)}{\frac{1}{n}\sum_{i=1}^{n} I_{CTE}(tm_g, i)} \geq 1; \\ tm_g, \dfrac{\frac{1}{n}\sum_{i=1}^{n} I_{CTE}(tm_d(dt), i)}{\frac{1}{n}\sum_{i=1}^{n} I_{CTE}(tm_g, i)} < 1\ ; \end{cases} \tag{3-31}$$

Najbolji rezultati predviđanja su dobiveni korištenjem jedne sezone skupa za učenje i jedne sezone skupa za ispitivanje, ali je isto tako pokazano kako postoji vremenski period koji daje najbolje rezultate predviđanja. Sukladno istraživanju iz odjeljka 3.7.2, izvršit će se ispitivanje indeksa CTE s uključenom značajkom prednosti domaćeg terena. Rezultati samog istraživanja će se prikazati u kasnijim poglavljima.

3.8. Optimalni vremenski prozor

Prethodna potpoglavlja, pogotovo odjeljak 3.7.2 u kojem se predviđanje vršilo na temelju prosječnog NBA indeksa, su pokazala kako postoji vremenski period koji najbolje opisuje stanje momčadi. Cilj vremenskog perioda, nazvanog optimalnim vremenskim prozorom (oP), bit će odabrati podskup skupa za učenje koji će dati bolje rezultate predviđanja, ali ujedno i smanjiti ulazni skup podataka bez posljedičnog smanjenja rezultata predviđanja. U prilog uvođenja optimalnog vremenskog prozora ide i rad [73] gdje je autor predstavio metodologiju korištenja pomičnog prozora (engl. *moving window*) za estimaciju teško-mjerljivih procesnih veličina korištenjem lako-mjerljivih procesnih veličina koje su u korelaciji s teško-mjerljivom veličinom. Autor rada koristi povijesne podatke koje naziva pogonskim podacima (engl. *plant data*). Kada pristigne novi mjerni uzorak najstariji podaci se izbacuju, a novi ubacuju u okvir ili prozor. Vrlo važan pojam vezan uz korištenje pomičnog prozora je veličina samog pomičnog prozora, odnosno pronalaženje kompromisa između „zaboravljanja“ starih informacija i „učenja“ novih. Osim navedenog rada, postoje i radovi koji za preračunavanje parametara modela koriste pomični prozor [74], [75], [76] i [77]. Na tragu navedenih istraživanja će se predstaviti algoritam izračuna i prilagodbe vremenskog prozora.

Cilj optimalnog vremenskog prozora je ograničiti pogled u prošlost te pronaći optimalni skup podataka koji će u zadanom trenutku najbolje opisati pojedinu momčad.

Ulazni skup podataka ($\mathcal{D}$) se sastoji od skupa za učenje ($\mathcal{D}_U$) i skupa za ispitivanje ($\mathcal{D}_I$), gdje su skup za učenje i skup za ispitivanje u svakom trenutku međusobno disjunktni ($\mathcal{D}_U \cap \mathcal{D}_I = \emptyset$). Nadzirano strojno učenje podrazumijeva postojanje skupa za učenje i skupa za ispitivanje.

U kasnijim odjeljcima će se predstaviti dva načina izračuna optimalnog vremenskog prozora, a uključivat će dva pojma predstavljena u odjeljku 3.4.8. Prvi način će koristi relativni rezultat, a drugi način relativni indeks korisnosti. Optimalni vremenski prozor će se računati na temelju skupa za učenje te proširivati podacima skupa za ispitivanje nad kojima je predviđanje obavljeno.

Računanje vremenskog prozora će se podijeliti u dva koraka:

1. Računanje optimalnog vremenskog prozora prije početka predviđanja – tzv. početni vremenski prozor (P_0)
2. Prilagodba optimalnog vremenskog prozora

3.8.1. Računanje i prilagodba optimalnog vremenskog prozora

Početni vremenski prozor će se računati neposredno prije predviđanja ishoda prvog događaja skupa za ispitivanje. Prilikom računanja početnog vremenskog prozora, ali i svih kasnijih optimalnih vremenskih prozora, skup za učenje i skup za ispitivanje su disjunktni. Za računanje početnog vremenskog prozora uzima se cijeli poznati skup za učenje. Početni vremenski prozor podskup je skupa za učenje ($P_0 \subseteq \mathcal{D}_U$).

Početni vremenski prozor, kao i svaki drugi optimalni vremenski prozor, se može izračunati na temelju prosječnog relativnog učinka ili prosječnog relativnog indeksa korisnosti. Konkretan način odabira početnog vremenskog prozora vezan uz trenutno istraživanje će biti objašnjen kasnije.

Vrlo važan korak algoritma izračuna vremenskog prozora bit će i prilagodba optimalnog vremenskog prozora. Prilagodba optimalnog vremenskog prozora podrazumijeva ponovni postupak izračuna optimalnog vremenskog prozora, a vrši se u slučaju kada je dostignut dozvoljen broj krivo predviđenih ishoda, odnosno neposredno prije predviđanja ishoda svakog. Pretpostavka je da će se prilagodbom optimalnog vremenskog prozora dodatno poboljšati rezultati predviđanja neposrednim odabirom relevantnijih vremenskih prozora. Broj dozvoljenih krivih predviđanja se može izračunati korištenjem neke od heurističkih metoda ili na temelju iskustva eksperta. Razlika u odnosu na računanje početnog vremenskog prozora je u tome da se skupu za učenje dodaju i utakmice skupa za ispitivanje za koje je predviđanje već obavljeno. Navedenim postupkom skup za učenje će se povećavati, a skup za ispitivanje smanjivati.

U kasnijim potpoglavljima će se najprije prikazati opći slučaj izračuna i prilagodbe vremenskog prozora, a kasnije i za konkretan problem vezan uz predviđanje ishoda košarkaških utakmica. Smisao upotrebe optimalnog vremenskog prozora je pronaći optimalni podskup skupa za učenje koji će u promatranom trenutku najbolje opisivati analiziranu momčad.

3.9. Odabir i izlučivanje značajki

Osim specifičnog skupa značajki vezanog uz analizirani proces, moguće je koristiti i skup univerzalnih izlučenih značajki. Skup izlučenih značajki može se dobiti analizom radova drugih istraživača koji se bave istim ili sličnim područjem, prema iskustvu autora ili ispitivanjem na konkretnom skupu podataka. Svaku potencijalnu izlučenu značajku potrebno je eksperimentalno ispitati i na taj odlučiti hoće li se koristiti u daljnjem istraživanju. Izlučene značajke u pravilu su značajke vezane uz uspješnost procesa, u ovom slučaju uspješnost momčadi, te samim time mogu poprimiti vrijednosti iz skupa prirodnih brojeva $N_{zn} \in N$. Formula (3-32) prikazuje način izračuna učinka analiziranog procesa p gdje n predstavlja broj izlučenih značajki.

$$I_p = \sum_{i=1}^{n} N_{p,zn_i} \tag{3-32}$$

Konkretne izlučene značajke vezane uz problem predviđanja ishoda u sportu će se dati u kasnijim poglavljima, a prijedlog je da se postupak predviđanja na temelju izlučenih značajki vrši na isti način kao i korištenjem indeksa korisnosti. Valja napomenuti kako korištenje dodatnih izlučenih značajki ne treba nužno značiti i bolje rezultate predviđanja, ali isto tako da i korištenje dodatnog skupa izlučenih značajki može dati bolje rezultate od korištenja indeksa korisnosti i optimalnog vremenskog prozora. Pretpostavka vezana uz trenutno istraživanje je da bi kombinacija korištenja indeksa korisnosti i optimalnog vremenskog prozora u kombinaciji s dodatnim izlučenim značajkama mogla dati bolje rezultate od korištenja indeksa korisnosti ili korištenja dodatnih izlučenih značajki.

U radu će se razmotriti dodavanje podskupa dodatnih izlučenih značajki indeksu korisnosti i optimalnom vremenskom prozoru te na taj način pronaći optimalni podskup dodatnih izlučenih značajki koji će polučiti najbolje rezultate predviđanja.

3.9.1. Događaji povećane neizvjesnosti

U ovom odjeljku će se uvesti pojam događaja povećane neizvjesnosti. Događajem povećane neizvjesnosti se smatra događaj u kojem je razlika projiciranih indeksa korisnosti dva suprotstavljena procesa unutar unaprijed definiranog raspona. Raspon identifikacije događaja

povećane neizvjesnosti se može izračunati korištenjem neke od heurističkih metoda ili na temelju iskustva eksperta.

Pretpostavka je da se bolji rezultati predviđanja mogu dobiti identifikacijom događaja povećanje neizvjesnosti, a prijedlog je da se za predviđanje događaja povećane neizvjesnosti, kao i za predviđanje događaja kojima nije potrebna analiza, koristiti projicirani indeks korisnosti, optimalni vremenski prozor te podskup dodatnih izlučenih značajki. Na taj bi se način koristio isti skup značajki, ali bi se pokušalo dodatno poboljšati rezultate predviđanja događaja povećane neizvjesnosti. Konkretni rezultati istraživanja će biti prikazani u kasnijim poglavljima.

Također valja napomenuti kako identifikacija i kasnije predloženi postupak predviđanja događaja povećane neizvjesnost ne mora nužno značiti i bolje rezultate predviđanja.

4. MODEL PREDVIĐANJA SPORTSKIH ISHODA ZASNOVAN NA INDEKSU KORISNOSTI I OPTIMALNOM VREMENSKOM PROZORU

U ovom poglavlju dana je teorijska pozadina predložene metode za predviđanje sportskih ishoda zasnovana na indeksu korisnosti i optimalnom vremenskom prozoru. Izgradnja modela odvija se u tri koraka. U prvom koraku se vrši prilagodba i optimizacija indeksa korisnosti, nakon toga i predviđanje na temelju optimiziranog indeksa korisnosti te se uvodi nova značajka kojoj je svrha dodatno vrednovati učinak pojedinih analiziranih procesa. Nakon toga se u drugom koraku uvodi pojam optimalnog vremenskog prozora sa svrhom definiranja vremenskog perioda koji će najbolje opisati suprotstavljene procese. U posljednjem koraku se u svrhu poboljšanja uspješnosti predviđanja identificiraju događaji povećane neizvjesnosti s ciljem definiranja načina predviđanja.

4.1. Predviđanje ishoda na osnovu indeksa korisnosti

Indeks korisnosti relativni je indikator kvalitete analiziranog procesa te ga se može definirati kao kumulativan indeks koji se sastoji od niza komponenti ponderiranih koeficijentom W_e, gdje svaka komponenta predstavlja element e promatranog procesa p. Opća formula indeksa korisnosti prikazana je formulom (4-1), dok je detaljan opis predloženog sveobuhvatnog indeksa korisnosti objašnjen u odjeljku 2.3.1.

$$I = \sum_{e \in E} W_e I_e \tag{4-1}$$

U formuli (4-1) E predstavlja skup elemenata e promatranog procesa p. Konačan indeks korisnosti procesa može definirati jedan ili više generatora komponenti. Ukoliko konačan indeks korisnosti procesa definira više generatora komponenti, kumulativni indeks korisnosti predstavlja sumu indeksa generatora komponenti koji su dio procesa p. Konačan indeks korisnosti $I(p)$ procesa p koji se sastoji od n generatora komponenti gn zapisan je fomulom (4-2).

$$I(p) = \sum_{n} I(gn)\,;\ \forall\ gn\ procesa\ p\ s\ definiranim\ I(gn) \tag{4-2}$$

Predviđanje ishoda se vrši na temelju povijesnih podataka te je potrebno uvesti pojam projiciranog indeksa korisnosti. Projicirani indeks korisnosti računa se isključivo na temelju povijesnih podataka. Cilj projiciranih indeksa korisnosti je predvidjeti ishod dva suprotstavljena procesa.

U odjeljku 3.4.8 uvedena su dva važna pojma vezana uz predviđanje ishoda korištenjem indeksa korisnosti, a to su pojmovi relativnog učinka i relativnog indeksa korisnosti. Ukoliko se promatrani

proces označi kao p_A, a suprotstavljeni proces kao p_B te ukoliko su promatrani procesi postigli učinak N_{p_A} i N_{p_B}, relativni učinak suprotstavljenih procesa, prikazan formulom (4-3), može se zapisati kao omjer učinka procesa A u odnosu na proces B.

$$R_{p_A/p_B} = \frac{N_{p_A}}{N_{p_B}} \tag{4-3}$$

U slučaju boljeg učinka procesa A, relativni učinak uvijek je veći je od 1, dok je u slučaju slabijeg učinka procesa A u odnosu na proces B relativni učinak uvijek manji od 1. U slučaju jednakih učinaka suprotstavljenih procesa nije moguće definirati uspješniji proces. Na temelju relativnog učinka moguće je definirati i relativni indeks korisnosti. Ukoliko se indeks korisnosti procesa A označi kao $I(p_A)$, a indeks korisnosti procesa B kao $I(p_B)$, relativni indeks korisnosti se može zapisati kako slijedi u formuli (4-4).

$$I_{p_A/p_B} = \frac{I(p_A)}{I(p_B)} \tag{4-4}$$

Nakon što su pojmovi relativnog učinka i relativnog indeksa korisnosti uvedeni, moguće je predložiti postupak predviđanja ishoda na temelju povijesnih podataka. Neka su s I_{p_A} i I_{p_B} označeni indeksi korisnosti dva suprotstavljena procesa analiziranog događaja te neka je funkcija izračuna pobjednika ovisna o indeksima suprotstavljenih procesa prikazana s $winn(I(p_A), I(p_B))$. Uspješnijim procesom će se smatrati proces čiji je indeks korisnosti veći. U slučaju istih vrijednosti nije moguće definirati uspješniji proces (p_X). Izračun ishoda procesa dan je formulom (4-5).

$$winn(I(p_A), I(p_B)) \begin{cases} p_A, \frac{I(p_A)}{I(p_B)} > 1; \\ p_X, \frac{I(p_A)}{I(p_B)} = 1\,; \\ p_B, \frac{I(p_A)}{I(p_B)} < 1. \end{cases} \tag{4-5}$$

Važno je napomenuti kako se u formuli (4-5) prikazuje način izračuna ishoda jednog događaja u kojem postoje dva suprotstavljena procesa. Projicirani indeks korisnosti računa se isključivo na temelju poznatih, povijesnih podataka, točnije projicirani indeks korisnosti prosječan je indeks korisnosti procesa do trenutka t tijekom kojeg postoji zapis o n događaja. Prosječni indeks korisnosti procesa u vremenskom periodu t tijekom kojeg postoji zapis o n događaja prikazan je formulom (4-6).

$$I(p) = \frac{1}{n}\sum_{i=1}^{n} I(p, i) \tag{4-6}$$

Na temelju formule (4-6), sukladno formuli (4-5), moguće je predložiti formulu predviđanja ishoda događaja dva suprotstavljena procesa. Formula (4-7) prikazuje matematički opis izračuna ishoda događaja dva suprotstavljena procesa na temelju projiciranog indeksa korisnosti u trenutku $t+1$ kada za vremenski period t postoji zapis za o n_{p_A} i n_{p_B} događaja vezanih uz proces A i proces B.

$$winn(I(p_A), I(p_B)) \begin{cases} p_A, \dfrac{\frac{1}{n_{p_A}} \sum_{i=1}^{n_{p_A}} I(p_A, i)}{\frac{1}{n_{p_B}} \sum_{i=1}^{n_{p_B}} I(p_B, i)} > 1; \\ p_X, \dfrac{\frac{1}{n_{p_A}} \sum_{i=1}^{n_{p_A}} I(p_A, i)}{\frac{1}{n_{p_B}} \sum_{i=1}^{n_{p_B}} I(p_B, i)} = 1; \\ p_B, \dfrac{\frac{1}{n_{p_A}} \sum_{i=1}^{n_{p_A}} I(p_A, i)}{\frac{1}{n_{p_B}} \sum_{i=1}^{n_{p_B}} I(p_B, i)} < 1. \end{cases} \quad (4\text{-}7)$$

Kao i u prethodna dva primjera, u slučaju jednakih vrijednosti projiciranih indeksa korisnosti nije moguće definirati ishod događaja, već se predviđanju ishoda događaja mora pristupiti na drugi način. Formulu (4-7) je moguće zapisati i kraće.

$$winn(I(p_A), I(p_B)) \begin{cases} p_A, \dfrac{I(p_A)}{I(p_B)} > 1; \\ p_X, \dfrac{I(p_A)}{I(p_B)} = 1; \\ p_B, \dfrac{I(p_A)}{I(p_B)} < 1. \end{cases} \quad (4\text{-}8)$$

Valja napomenuti kako će u nastavku pojam $I(p)$ predstavljati projicirani indeks korisnosti procesa. Iz formule je jasno vidljivo da ishod događaja nije moguće definirati u slučaju kada su projicirani indeksi korisnosti suprotstavljenih procesa jednaki. U tom slučaju se predviđanju ishoda događaja mora pristupiti na način prilagođen specifičnosti problema.

4.1.1. Optimizacija doprinosa elemenata indeksa korisnosti

Problem optimizacije, odnosno problem pronalaženja najboljeg rješenja u matematičkom skupu mogućih rješenja u skupini je problema koji se počinju učinkovito rješavati pojavom računala, a dodatni razvoj je postignut povećanjem računalne snage. Optimizacija se vrši iterativno, slijedeći zadane korake kojima se nastoji poboljšati prethodno rješenje. Bez obzira na to što je pojava računala i svakodnevno povećanje računalne snage olakšalo problem optimizacije, metode iscrpnog pretraživanja prostora svih mogućih rješenja za određenu vrstu problema i dalje predstavljaju veliki problem. Razvijeni su različiti pristupi definiranja strategije pretraživanja prostora, a mogu se podijeliti na heurističke i metaheurističke metode (kraće heuristike i

metaheuristike). Navedene metode optimizacije ne mogu garantirati optimalno rješenje, ali omogućuju pronalazak zadovoljavajućeg rješenja u razumnom vremenskom roku. Heuristike se koriste za određene probleme, dok se metaheuristike koriste za širok spektar različitih problema.

Sveobuhvatni indeks korisnosti je definiran u odjeljku 2.3.1 kao kumulativan indeks koji se sastoji od niza komponenti ponderiranih koeficijentom W_e, gdje svaka komponenta predstavlja element e promatranog procesa. Opća formula indeksa CPE prikazana je formulom (2-3). Zbroj ponderiranih koeficijenti W_e jednak je kardinalnosti skupa $E = (e_1, e_2, \dots, e_n)$ gdje n predstavlja broj elemenata skupa E. Svaki element e procesa p može imati pozitivan i/ili negativan doprinos, a početna formula doprinosa komponente e prikazana je formulom (4-9).

$$I_e = v_e N_e - v'_e N'_e, \;\; v_e, v'_e \geq 0, N_e, N'_e \geq 0 \tag{4-9}$$

Cilj predloženog postupka optimizacije prilagoditi je koeficijente $v_e(v'_e)$ indeksu korisnosti, točnije pronaći skup neovisnih varijabli koje će minimizirati vrijednost zadane funkcije. U slučaju dva suprotstavljena procesa p_A i p_B potrebno je na temelju korištenog indeksa korisnosti definirati uspješniji proces. Opća formula funkcije izračuna uspješnijeg procesa ($winn(I(p_A), I(p_B))$) dana je formulom (4-5). Funkcija izračuna uspješnosti suprotstavljenih procesa prima dva argumenta, indeks korisnosti procesa A ($I(p_A)$) i indeks korisnosti procesa B ($I(p_B)$). Cilj optimizacije koeficijenata $v_e(v'_e)$ minimizirati je funkciju izračuna uspješnosti procesa koji će na temelju povijesnih podataka izračunavati uspješnost suprotstavljenih procesa. Jedini pravi indikator uspješnosti procesa je relativni učinak prikazan formulom (4-3). Na temelju prosječnih povijesnih podataka potrebno je prilagoditi koeficijente $v_e(v'_e)$ elementima e skupa E na način da razlika omjera prosječnih indeksa korisnosti i relativnog učinka bude minimalna. Funkcija $opt(x)$, kojoj je cilj težiti globalnom minimumu, u trenutku t u kojem je za pojedini proces poznato n ishoda je prikazana formulom (4-10). Valja napomenuti kako u trenutku t broj poznatih ishoda procesa ne treba biti jednak, a oznakom N označen je stvaran učinak procesa.

$$opt(x) = \frac{\frac{1}{n_{p_A}} \sum_{i=1}^{n_{p_A}} I(p_A, i)}{\frac{1}{n_{p_B}} \sum_{i=1}^{n_{p_B}} I(p_B, i)} - \frac{N_{p_A}(t)}{N_{p_B}(t)} \tag{4-10}$$

Nakon što je definirana funkcija optimizacije moguće je pristupiti i definiranju samog postupka optimizacije. Neka su elementi skupa E koji opisuje skup značajki definirani kao $E = (e_1, e_2, \dots, e_n)$ te skup pripadajućh koeficijenata koje je potrebno prilagoditi s $V = (v_{e_1}, v_{e_2}, \dots, v_{e_n})$. Kao što je definirano u odjeljku 2.3.1.1, koeficijent $v_e(v'_e)$ limitiran je minimalnim i maksimalnim vrijednostima ($v_e(v'_e) = \left[v^{(\prime)}_{e,min}, v^{(\prime)}_{e,max}\right]$). Gornja i donja granica se

mogu definirati na temelju neke od heurističkih metoda ili empirijski, a glavni cilj limitiranja granica koeficijenta $v_e(v_e')$ je sprječavanje da pojedini elementi zaguše doprinos ostalih elemenata. Pod pojmom zagušenja smatra se preveliki udio pojedinih značajki u odnosu na druge značajke, čime preostale značajke u potpunosti gube značaj.

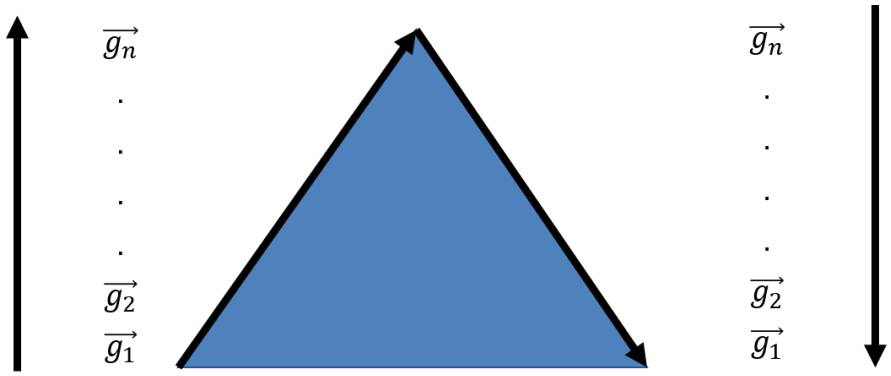
Proces optimizacije se vrši iterativno. Skup E predstavlja skup značajki, a skup V predstavlja skup pripadajućih koeficijenata gdje je svaki koeficijent $v_e(v_e')$ limitiran minimalnom i maksimalnom vrijednošću. Oznakom $G = (g_1, g_2, \dots g_n)$ je označen skup funkcija linearnog i nelinearnog doprinosa. Valja napomenuti kako gornje i donje vrijednosti elemenata skupa E ne trebaju biti jednake, stoga je potrebno definirati podskup funkcija linearnog i nelinearnog doprinosa za svaki element (značajku), pri čemu je skup funkcija linearnog i nelinearnog doprinosa elementa podskup skupa G ($G_e \subseteq G$). Nakon što su definirani svi potrebni pojmovi, točnije skup elemenata (značajki) E, skup koeficijenata V i pripadnih minimalnih i maksimalnih vrijednosti $V_{e,min}$ i $V_{e,max}$ te skup funkcija linearnog i nelinearnog doprinosa G, moguće je pristupiti optimizaciji.

Postupak optimizacije se vrši u dva koraka:

1. Izračun redoslijeda značajki (predkorak postupka optimizacije)
2. Optimizacija doprinosa pojedinačnih značajki

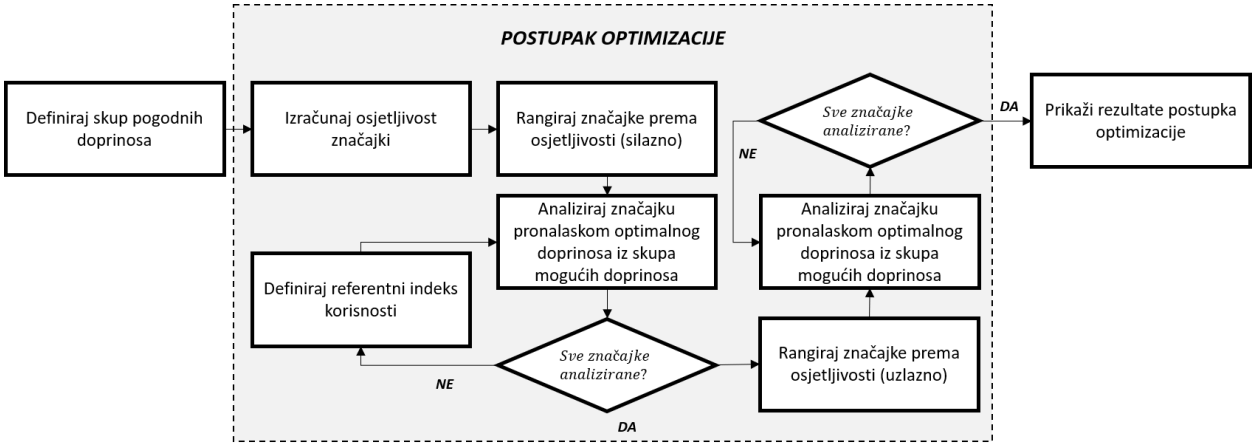
U prvom koraku je potrebno na temelju odabranog indikatora definirati redoslijed optimizacije značajki te definirati početni skup doprinosa skupa V. Redoslijed značajki se može definirati empirijski ili korištenjem neke od heurističih metoda. Neke od metoda koje se mogu koristiti za definiranje redoslijeda značajki su informacijska dobit i osjetljivost (engl. *sensitivity*), a sam odabir metode definiranja redoslijeda ponekad ovisi i o samom procesu. Skup poredanih značajki se označuje oznakom $\vec{G} = (\overrightarrow{g_1}, \overrightarrow{g_2}, \dots \overrightarrow{g_n})$ u kojem su značajke poredane od najznačajnije prema manje značajnima. U drugom koraku slijedi optimizacija pojedinačnih značajki eksperimentalno ispitivajući učinak definiranog doprinosa iz skupa mogućih doprinosa G_e na rezultat predviđanja. Najbolji rezultat trenutne iteracije koristi se kao referentna točka sljedeće iteracije.

Sam tijek optimizacije može biti proizvoljan, a u ovom slučaju će se objasniti tzv. piramidalni način optimizacije. Piramidalni način optimizacije se sastoji od dvije faze. U prvoj fazi se optimizacija vrši počevši s najznačajnijom značajkom pa prema manje značajnim, a u drugoj fazi se kreće obrnutim redoslijedom, točnije od najmanje značajne značajke prema najznačajnijoj. Valja napomenuti kako predloženi postupak optimizacije rješava problem optimizacije suprotstavljenih procesa korištenjem funkcija linearnog i nelinearnog doprinosa. Slika 4.1 grafički prikazuje opći oblik piramidalne optimizacije.



Slika 4.1. Opći oblik piramidalne optimizacije.

Slika 4.2 prikazuje blok dijagram postupka optimizacije doprinosa koeficijenata indeksa korisnosti.



Slika 4.2. Blok dijagram predloženog postupka optimizacije doprinosa značajki.

4.1.2. Definiranje značajke prednosti uspješnijeg procesa

Nakon što je definiran način predviđanja ishoda događaja dva suprotstavljena procesa, potrebno je uvesti pojmove vezane uz dodatno nagrađivanje procesa. Neka su s p_A i p_B označena dva suprotstavljena procesa i neka je u zadanom trenutku t broj uspješnijih ishoda procesa p_A označen s N_{p_A}, a broj uspješnijih ishoda procesa p_B s N_{p_B}. Postotci uspješnosti procesa A u odnosu na proces B te procesa B u odnosu na proces A su zapisani formulama (4-11) i (4-12).

$$\Delta p_A = \frac{N_{p_A}}{N_{p_B} + N_{p_B}}, \Delta p_A \in [0,1], N_{p_A}, N_{p_B} \geq 0 \tag{4-11}$$

$$\Delta p_B = \frac{N_{p_B}}{N_{p_A} + N_{p_B}}, \Delta p_B \in [0,1], N_{p_A}, N_{p_B} \geq 0 \tag{4-12}$$

Suma uspješnosti procesa A i procesa B mora u svakom trenutku iznositi jedan što je i prikazano formulom (4-13), dok je razlika uspješnosti procesa prikazana formulom (4-14).

$$\Delta p_A + \Delta p_B = 1 \tag{4-13}$$

$$\Delta p = \Delta p_A - \Delta p_B \tag{4-14}$$

Neka su koeficijenti nagrađivanja suprotstavljenih procesa prikazani kao χ_{p_A} i χ_{p_B}. Konačan indeks korisnosti uspješnijeg procesa tako se uvećava za razliku postotka uspješnosti učinka suprotstavljenih procesa. Konačan učinak neuspješnijeg procesa ostaje nepromijenjen. Formule (4-15) i (4-16) prikazuju izračun koeficijenata suprotstavljenih procesa.

$$\chi_{p_A} = \begin{cases} 1, \Delta p_A > \Delta p_B \\ 0, \Delta p_A \leq \Delta p_B \end{cases} \tag{4-15}$$

$$\chi_{p_B} = \begin{cases} 1, \Delta p_B > \Delta p_A \\ 0, \Delta p_B \leq \Delta p_A \end{cases} \tag{4-16}$$

Nakon što je definiran koeficijent nagrađivanja suprotstavljenih procesa, moguće je definirati i konačnu formulu izračuna indeksa korisnosti. Konačan učinak suprotstavljenih procesa prikazan je formulama (4-17) i (4-18). Kao što je već napomenuto, $I(p)$ predstavlja projicirani indeks korisnosti procesa na temelju poznate povijesti.

$$I(p_A, \chi) = I(p_A, \chi) + \chi_{p_A} \times I(p_A, \chi) \times \Delta p \tag{4-17}$$

$$I(p_B, \chi) = I(p_B, \chi) + \chi_{p_B} \times I(p_B, \chi) \times \Delta p \tag{4-18}$$

Nakon što je definiran koeficijent nagrađivanja suprotstavljenih procesa, potrebno je optimizirati doprinos nagrađivanja uvođenjem korektivnog faktora (kf). Korektivni faktor limitiran je parametrima kf_{min} *i* kf_{max}, a sam postupak optimizacije se vrši na isti način kao i optimizacija doprinosa elemenata indeksa korisnosti prikazana u odjeljku 4.1.1. Formulama (4-19) i (4-20) matematički je opisan način korištenja korektivnog faktora.

$$I(p_A, \chi) = I(p_A, \chi) + \chi_{p_A} \times I(p_A, \chi) \times kf \times \Delta p \tag{4-19}$$

$$I(p_B, \chi) = I(p_B, \chi) + \chi_{p_B} \times I(p_B, \chi) \times kf \times \Delta p \tag{4-20}$$

4.2. Predviđanje na temelju optimalnog vremenskog prozora

Metoda nadziranog strojnog učenja podrazumijeva postojanje skupa za učenje ($\mathcal{D}_U$) i skupa za ispitivanje ($\mathcal{D}_I$), gdje su skup za učenje i skup za ispitivanje u svakom trenutku disjunktni ($\mathcal{D}_U \cap \mathcal{D}_I = \emptyset$).

Optimalni vremenski prozor se računa na temelju skupa za učenje te se u svakoj iteraciji proširuje podacima skupa za ispitivanje nad kojima je predviđanje već obavljeno. Drugi način izračuna optimalnog vremenskog prozora može biti fiksno definirana duljina vremenskog prozora, gdje se najstariji podaci prilikom pristizanja novih izbacuju, a novi ubacuju u vremenski prozor. Dulji vremenski prozor sadrži veću količinu kronološki poredanih podataka te je samim time utjecaj novijih podataka u prilagodbi modela smanjen, što je posebno važno kod međusobno

ovisnih događaja. S druge strane, u kraćem vremenskom prozoru informacije iz bliske prošlosti dominiraju. Ukoliko je vremenski prozor prevelik model će se sporo prilagoditi na promjene procesa što može rezultirati smanjenim sposobnostima predviđanja. S druge strane, premali prozor može proizvesti model s lošim sposobnostima predviđanja jer u takvim prozorima često nema dovoljno podataka za procjenu parametara modela, a dostupni podaci mogu sadržavati pogreške. Stoga je vrlo važno pronaći kompromis koji treba težiti pronalasku optimalne duljine vremenskog prozora. U ovom slučaju će se predstaviti način proširenja optimalnog vremenskog prozora.

Neka se skup za učenje sastoji od n kronološki poredanih događaja, $\mathcal{D}_U = \{1, 2, 3, \dots, n\}$, gdje je kronološki zadnji događaj označena s n. Kronološki zadnji događaj skupa za učenje prethodi kronološki prvom događaju skupa za ispitivanje. Na temelju n događaja skupa za učenje predlaže se isto toliko potencijalnih vremenskih prozora (poP_n). Prvi potencijalni vremenski prozor uključuje samo n-ti događaj skupa za učenje, a svakom slijedećem potencijalnom vremenskom prozoru se dodaje prethodni događaj, samim time povećavajući svaki sljedeći potencijalni vremenski prozor za jedan događaj pri čemu valja napomenuti kako je potencijalni vremenski prozor kontinuirani niz događaja. Tako će prvi potencijalni vremenski prozor uključivati samo jedan događaj ($poP_1 = \{n\}$), a zadnji potencijalni vremenski prozor sve događaje skupa za učenje ($poP_n = \{1, 2, \dots, n\}$). Optimalni vremenski prozor se može računati na temelju relativnog učinka ili relativnog indeksa korisnosti, a optimalnim vremenskim prozorom (oP) će se proglasiti podskup skupa za učenje koji zadovoljava određene uvjete. Formulama (4-21) i (4-22) je prikazan način izračuna relativnog učinka i relativnost indeksa dva suprotstavljena procesa.

$$R_{p_A/p_B} = \frac{N_{p_A}}{N_{p_B}} \tag{4-21}$$

$$I_{p_A/p_B} = \frac{I(p_A)}{I(p_B)} \tag{4-22}$$

Relativni učinak i relativni indeks korisnosti potencijalnog vremenskog prozora se računa kao prosječna vrijednost relativnih učinaka ili relativnih indeksa korisnosti kontinuiranih događaja. Formulama (4-23) i (4-24) prikazani su načini izračuna prosječnog rezultata i prosječnog indeksa korisnosti potencijalnog vremenskog prozora na temelju n događaja.

$$R_{poP_n} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_{p_A,i}}{N_{p_B,i}} \tag{4-23}$$

$$I_{poP_n} = \frac{1}{n}\sum_{i=1}^{n}\frac{I(p_A, i)}{I(p_B, i)} \tag{4-24}$$

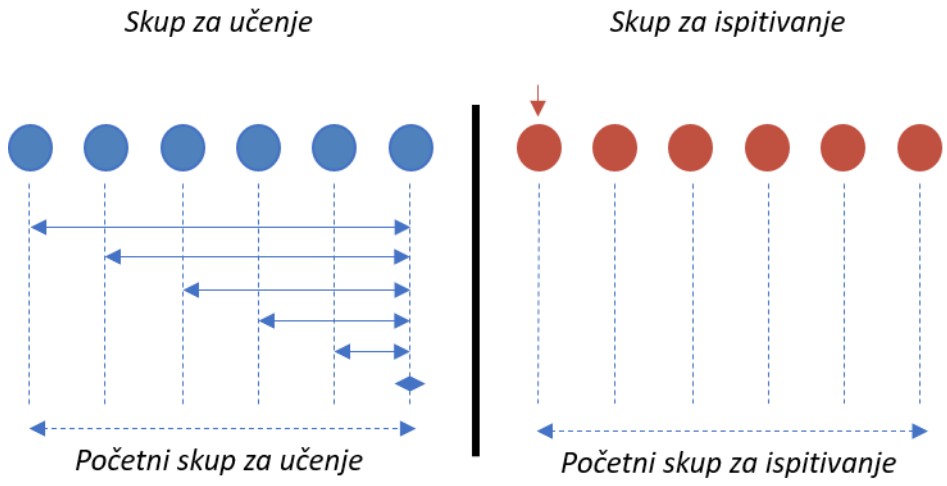
Računanje vremenskog prozora dijeli se u dva koraka:

1. Računanje početnog vremenskog prozora (P_0)
2. Prilagodba optimalnog vremenskog prozora

4.2.1. Računanje optimalnog vremenskog prozora

Početni vremenski prozor računa se neposredno prije predviđanja ishoda kronološki prvog događaja skupa za ispitivanje. Prilikom računanja početnog vremenskog prozora, ali i svih kasnijih optimalnih vremenskih prozora, skup za učenje i skup za ispitivanje su disjunktni. Za računanje početnog vremenskog prozora se uzima cijeli poznati skup za učenje. Početni vremenski prozor podskup je skupa za učenje ($P_0 \subseteq \mathcal{D}_U$).

Početni vremenski prozor, kao i svaki drugi optimalni vremenski prozor, se može izračunati na temelju prosječnog relativnog učinka ili prosječnog relativnog indeksa korisnosti. Konkretan način odabira početnog vremenskog prozora potrebno je prilagoditi analiziranom događaju. Slika 4.3 grafički prikazuje izračun početnog vremenskog prozora (P_0). Plavim strelicama s punim linijama označeni su potencijalni vremenski prozori u kojima zadnji, šesti po redu, označava potencijalni vremenski prozor koji uključuje samo jednu utakmicu.

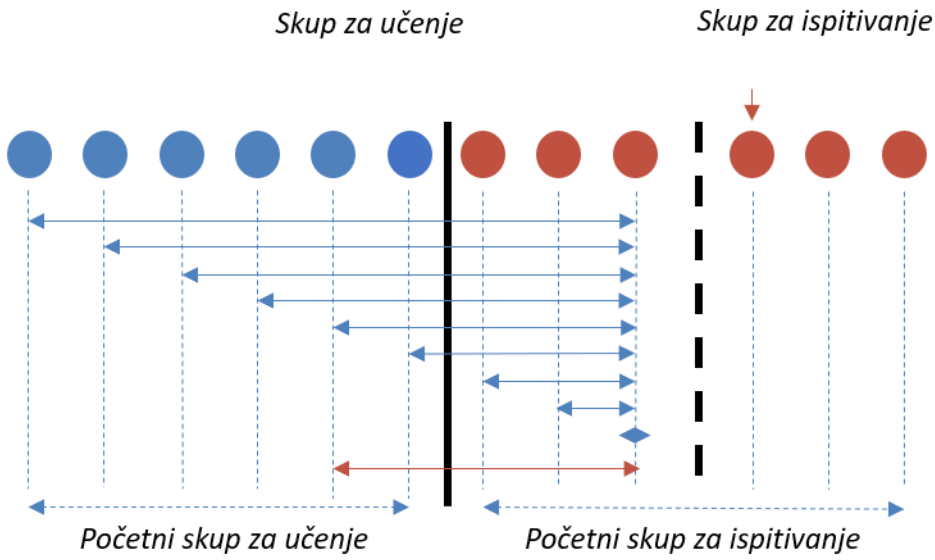


Slika 4.3. Grafički prikaz izračuna početnog vremenskog prozora.

Ishod događaja kojeg je potrebno prvog predvidjeti te ostali događaji skupa za ispitivanje označeni su crvenim krugovima, dok su plavim krugovima označeni događaji skupa za učenje. Crvenom strelicom označen je događaj predviđanja. Kao što je već navedeno, optimalni vremenski prozor računa se na temelju skupa za učenje te sadrži najmanje jedan događaj skupa za učenje. Važno je napomenuti da svaki potencijalni vremenski prozor, a samim time i početni vremenski prozor, kao i svi kasnije izračunati optimalni vremenski prozori, predstavljaju kontinuirani skup događaja.

4.2.2. Prilagodba optimalnog vremenskog prozora

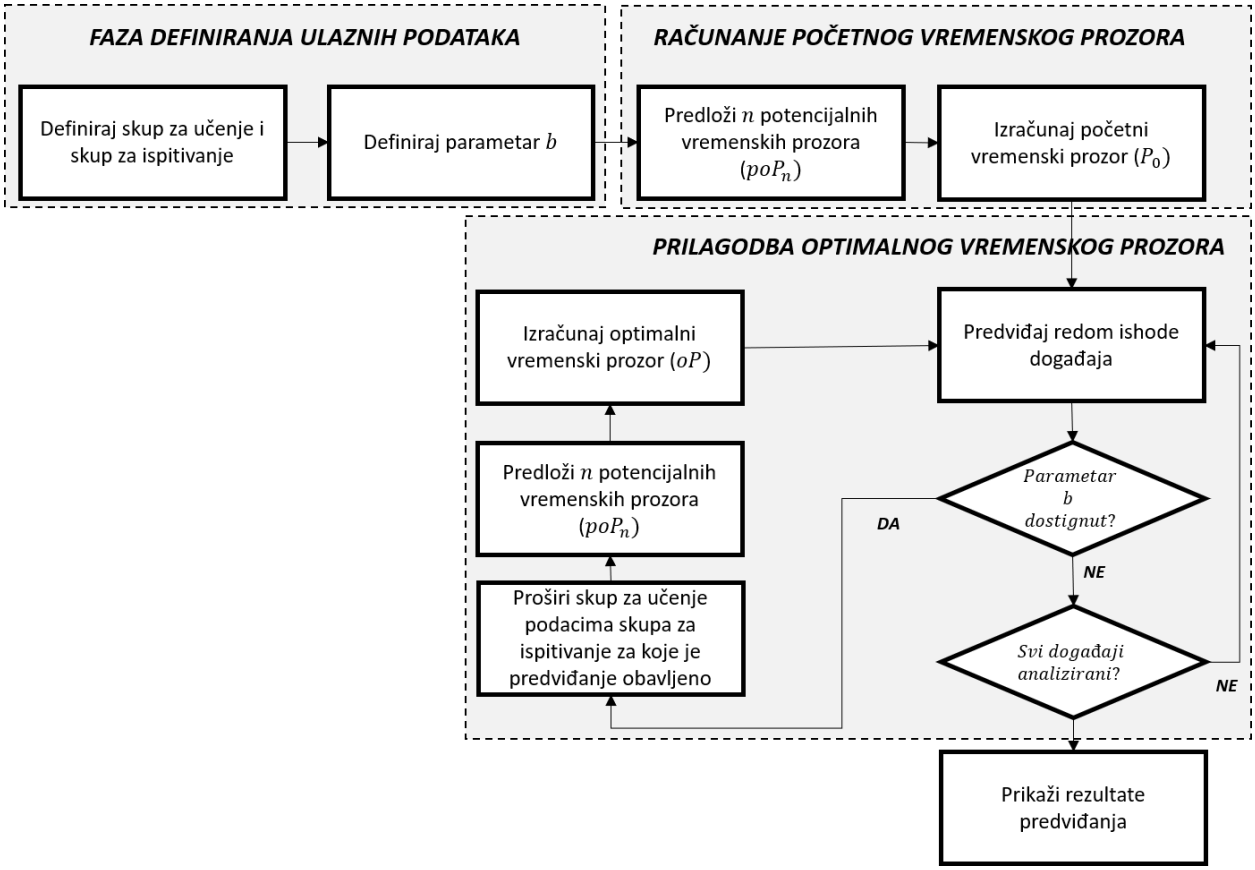
Vrlo važan korak algoritma izračuna vremenskog prozora je prilagodba optimalnog vremenskog prozora. Prilagodba optimalnog vremenskog prozora podrazumijeva ponovni postupak izračuna optimalnog vremenskog prozora, a vrši se u slučaju kada je dostignut dozvoljen broj krivo predviđenih ishoda ($b \in \{0, 1, \dots l\}$). Prilagodba optimalnog vremenskog prozora može se vršiti i prilikom predviđanja ishoda svakog događaja ($b = 0$). Broj dozvoljenih krivih predviđanja može se izračunati korištenjem neke od heurističkih metoda ili na temelju iskustva eksperta. Kao i kod izračuna početnog optimalnog vremenskog prozora, prilagodba vremenskog prozora skup podataka dijeli na skup za učenje i skup za ispitivanje. Razlika u odnosu na računanje početnog vremenskog prozora je u tome što se skupu za učenje dodaju i utakmice skupa za ispitivanje nad kojima je predviđanje već obavljeno. Navedenim postupkom se skup za učenje povećava, skup za ispitivanje smanjuje, ali skupovi i dalje ostaju disjunktni. Slika 4.4 grafički prikazuje prilagodbu optimalnog vremenskog prozora. Plavim strelicama s punim linijima označeni su potencijalni vremenski prozori u kojima zadnji, deveti po redu, označava potencijalni vremenski prozor koji uključuje samo jednu utakmicu.



Slika 4.4. Grafički prikaz prilagodbe optimalnog vremenskog prozora.

Ishod događaja kojeg je potrebno predvidjeti označen je crvenom strelicom. Broj dozvoljenih krivih predviđanja je dostignut te je potrebna prilagodba optimalnog vremenskog prozora. Kao i kod računanja početnog optimalnog vremenskog prozora potrebno je na temelju poznate povijesti i načina izračuna prosječne vrijednosti izračunati potencijalne vremenske prozore te odabrati najpogodniji. Prilagodba optimalnog vremenskog prozora se ponavlja ukoliko je ponovno dostignut dozvoljen broj krivih predviđanja. Slika 4.5 prikazuje dijagram toka algoritma izračuna

i prilagodbe optimalnog vremenskog prozora. Na slici su prvenstveno zbog lakše vizualizacije označeni koraci računanja vremenskog prozora, točnije korak računanja početnog vremenskog prozora te korak prilagodbe optimalnog vremenskog prozora koji su detaljno objašnjeni ranije u potpoglavlju.



Slika 4.5. Dijagram toka algoritma izračuna i prilagodbe optimalnog vremenskog prozora.

Valja napomenuti kako se optimalni vremenski prozor računa za svaki analizirani proces, a da se kao mogućnost korištenja optimalnog vremenskog prozora dva suprotstavljena procesa nudi više mogućnosti koje je potrebno eksperimentalno ispitati i na taj odrediti onu mogućnost koja će dati najbolji rezultat. Ukoliko se optimalni vremenski prozori dva suprotstavljena procesa označe s oP_{p_A} i oP_{p_B}, konačan optimalni vremenski prozor (oP) može se izračunati kao unija, presjek ili razlika optimalnih vremenskih prozora suprotstavljenih procesa ili koristiti optimalni vremenski prozori specifični svakom procesu.

Formula izračuna ishoda događaja dva suprotstavljena procesa korištenjem indeksa korisnosti i optimalnog vremenskog prozora dana je formulom (4-25) gdje n predstavlja broj utakmica optimalnog vremenskog perioda pojedinog procesa. U navedenom općem obliku koristi se vremenski prozor specifičan za svaki analizirani proces.

$$winn(I(p_A), I(p_B)) \begin{cases} p_A, \dfrac{\frac{1}{n_{p_A}} \sum_{i=1}^{n_{p_A}} I(p_A, i)}{\frac{1}{n_{p_B}} \sum_{i=1}^{n_{p_B}} I(p_B, i)} > 1 \\ p_X, \dfrac{\frac{1}{n_{p_A}} \sum_{i=1}^{n_{p_A}} I(p_A, i)}{\frac{1}{n_{p_B}} \sum_{i=1}^{n_{p_B}} I(p_B, i)} = 1 \\ p_B, \dfrac{\frac{1}{n_{p_A}} \sum_{i=1}^{n_{p_A}} I(p_A, i)}{\frac{1}{n_{p_B}} \sum_{i=1}^{n_{p_B}} I(p_B, i)} < 1 \end{cases} \quad (4\text{-}25)$$

4.3. Predviđanje na temelju dodatnih značajki

Osim osnovnih elemenata korištenog indeksa korisnosti prilagođenih analiziranom događaju, model predviđanja može koristiti i dodatan skup značajki. Odabir dodatnih značajki se može vršiti na temelju iskustva eksperta, a njihov učinak na rezultate predviđanja potrebno je eksperimentalno ispitati. Važno je napomenuti kako skup dodatnih značajki može biti univerzalan za sve vrste procesa. Formula izračuna ishoda dva suprotstavljena procesa određenog događaja na temelju proizvoljnog skupa dodatnih značajki dana je formulom (4-26) gdje n_{zn} predstavlja broj značajki dodatnog izlučenog skupa značajki.

$$winn(N_{p_A}, N_{p_B}) \begin{cases} p_A, \dfrac{\sum_{i=1}^{n_{zn}} N_{p_A, zn_i}}{\sum_{i=1}^{n_{zn}} N_{p_B, zn_i}} > 1; \\ p_X, \dfrac{\sum_{i=1}^{n_{zn}} N_{p_A, zn_i}}{\sum_{i=1}^{n_{zn}} N_{p_B, zn_i}} = 1\,; \\ p_A, \dfrac{\sum_{i=1}^{n_{zn}} N_{p_A, zn_i}}{\sum_{i=1}^{n_{zn}} N_{p_B, zn_i}} < 1. \end{cases} \quad (4\text{-}26)$$

Kao što je već utvrđeno, uspješnijim se proglašava proces s većom projiciranom sumom učinaka izlučenih značajki. U slučaju kada je projicirana suma učinaka suprotstavljenih procesa jednaka, predviđanju ishoda potrebno je pristupiti na neki drugi način.

4.3.1. Predviđanje na temelju indeksa korisnosti i izlučenih značajki

Upotreba dodatnog skupa značajki je opcionalna. Ukoliko se kod predviđanja ishoda događaja koristi i dodatan skup značajki, potrebno je definirati i način predviđanja. Ideja korištenja kombinacije osnovnih značajki indeksa korisnosti i dodatnih značajki istovjetna je predviđanju ishoda korištenjem indeksa korisnosti. Formula (4-27) matematički opisuje predviđanje ishoda događaja dva suprotstavljena procesa korištenjem kombinacije indeksa korisnosti i dodatnog skupa izlučenih značajki označenih s $I(p)$ gdje n_{zn} predstavlja broj značajki dodatnog skupa izlučenih značajki, a n_p broj događaja optimalnog vremenskog prozora pojedinog procesa.

$$winn(I(p_A), I(p_B)) \begin{cases} p_A, \dfrac{\frac{1}{n_{p_A}} \sum_{i=1}^{n_{p_A}} I(p_A, i) + \sum_{i=1}^{n_{zn}} N_{p_A, zn_i}}{\frac{1}{n_{p_B}} \sum_{i=1}^{n_{p_B}} I(p_B, i) + \sum_{i=1}^{n_{zn}} N_{p_B, zn_i}} > 1 \\ p_X, \dfrac{\frac{1}{n_{p_A}} \sum_{i=1}^{n_{p_A}} I(p_A, i) + \sum_{i=1}^{n_{zn}} N_{p_A, zn_i}}{\frac{1}{n_{p_B}} \sum_{i=1}^{n_{p_B}} I(p_B, i) + \sum_{i=1}^{n_{zn}} N_{p_B, zn_i}} = 1 \\ p_B, \dfrac{\frac{1}{n_{p_A}} \sum_{i=1}^{n_{p_A}} I(p_A, i) + \sum_{i=1}^{n_{zn}} N_{p_A, zn_i}}{\frac{1}{n_{p_B}} \sum_{i=1}^{n_{p_B}} I(p_B, i) + \sum_{i=1}^{n_{zn}} N_{p_B, zn_i}} < 1 \end{cases} \quad (4\text{-}27)$$

4.4. Događaji povećane neizvjesnosti

U ovom potpoglavlju će se uvesti pojam događaja povećane neizvjesnosti. Događajem povećane neizvjesnosti se smatra događaj u kojem je razlika projiciranih indeksa korisnosti dva suprotstavljena procesa unutar unaprijed definiranog raspona. Ukoliko su s p_A i p_B označena dva suprotstavljena procesa, samim time i pripadajući projicirani indeksi korisnosti, $I(p_A)$ i $I(p_B)$, postotnu razliku ($pr_{\%}$) projiciranih indeksa korisnosti suprotstavljenih procesa može se definirati kako je zapisano u formuli (4-28).

$$pr_{\%} = \frac{|I(p_A) - I(p_B)|}{I(p_A)} \times 100 \quad (4\text{-}28)$$

Definiranjem pojma postotne razlike potrebno je definirati i pojam raspona pojavljivanja događaja povećane neizvjesnosti (r). Raspon događaja povećane neizvjesnosti ograničen je minimalnom i maksimalnom vrijednosti, točnije može poprimiti vrijednost iz intervala $[r_{min}, r_{max}]$. Formula (4-29) matematički opisuje predviđanje korištenjem događaja povećane neizvjesnosti.

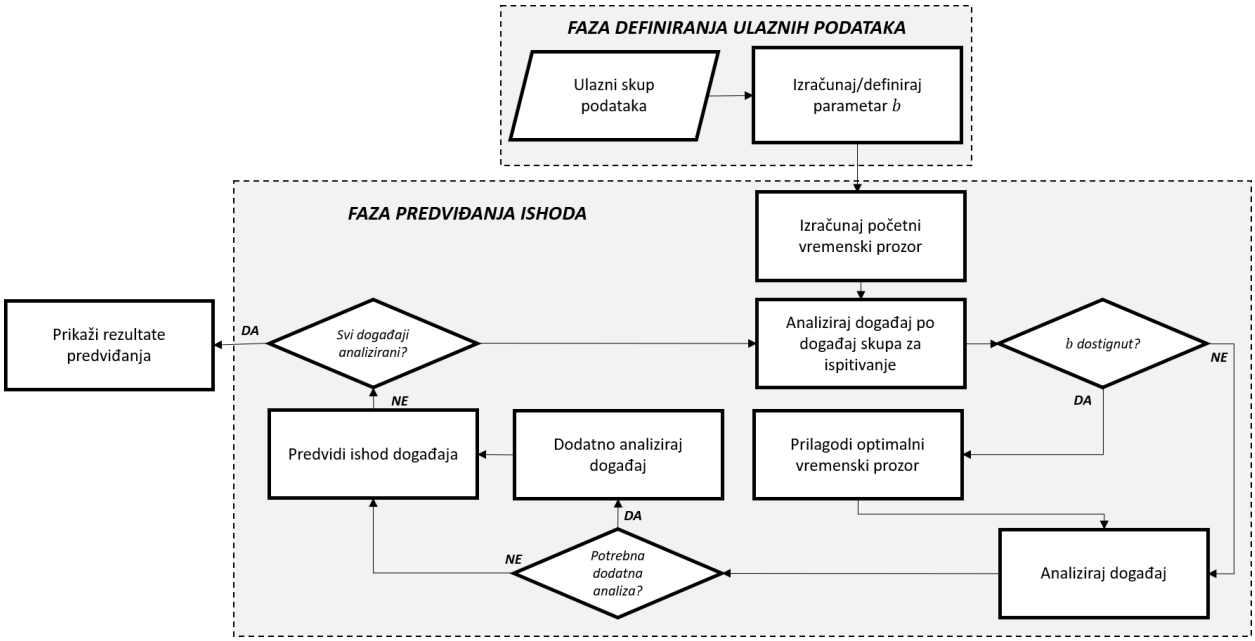
$$predviđanje \begin{cases} događaj\ povećane\ neizvjesnosti, r_{min} \leq pr_{\%} \leq r_{max} \\ predviđanje\ ishoda\ opisano\ u\ poglavljima\ 4.1 - 4.3, ostalo \end{cases} \quad (4\text{-}29)$$

Kao što je vidljivo iz formule (4-29), ukoliko se parametar postotne razlike nalazi u definiranom rasponu, točnije rasponu pojavljivanja događaja povećane neizvjesnosti, događaj se smatra događajem povećane neizvjesnosti. U ostalim slučajevima kada se parametar postotne razlike ne nalazi u rasponu pojavljivanja događaja povećane neizvjesnosti, predviđanje ishoda se vrši na način opisan u poglavljima 4.1 - 4.3.

Predviđanje ishoda događaja povećane neizvjesnosti, uz predloženi indeks korisnosti koristi i skup izlučenih značajki koje je potrebno eksperimentalno ispitati i na taj način identificirati one koje daju najbolje rezultate predviđanja.

4.5. Model predviđanja

Nakon što su definirani i matematički opisani svi ključni pojmovi, moguće je blok dijagramom opisati predloženi model predviđanja. Blok dijagram je grafički prikaz algoritma koji se sastoji od niza simbola povezanih strelicama koje definiraju tok i smjer realizacije. Slika 4.6 prikazuje blok dijagram modela predviđanja sportskih ishoda zasnovanog na indeksu korisnosti i optimalnom vremenskom prozoru.



Slika 4.6. Model predviđanja sportskih ishoda zasnovan na indeksu korisnosti i optimalnom vremenskom prozoru.

Zbog lakše vizualizacije predloženi model je podijeljen u dvije faze. Faze predstavljaju logičke cjeline vezane uz izradu samog modela. Prvu fazu karakterizira definiranje ulaznih podataka i parametra b. Ulazni skup podataka je potrebno podijeliti na skup za učenje i skup za ispitivanje. Kvalitetan i reprezentativan ulazni skup podataka osnovni je preduvjet uspješnog predviđanja ishoda analiziranog procesa. Vrijednost parametra b moguće je definirati empirijski ili izračunati na temelju skupa za učenje. Definiranjem parametra b priprema ulaznih podataka je završena te su stvoreni su svi potrebni preduvjeti za pristupanje fazi predviđanja ishoda. Predviđanje ishoda započinje izračunom početnih vremenskih prozora analiziranih procesa. Izračun početnih vremenskih prozora, ujedno i optimalnih vremenskih prozora, vrši se isključivo nad podacima skupa za učenje. Nakon što su početni vremenski prozori definirani kreće analiza i predviđanje ishoda događaja skupa za ispitivanje. Predviđanje se vrši iterativno, što znači da se u jednom trenutku predviđa ishod jednog događaja. Tijekom faze predviđanja potrebno je vršiti prilagodbu optimalnog vremenskog prozora u slučaju kada je parametar b dostignut. Pojedinačni događaj uključuje dva suprotstavljena procesa za koje je potrebno izračunati projicirane indekse korisnosti.

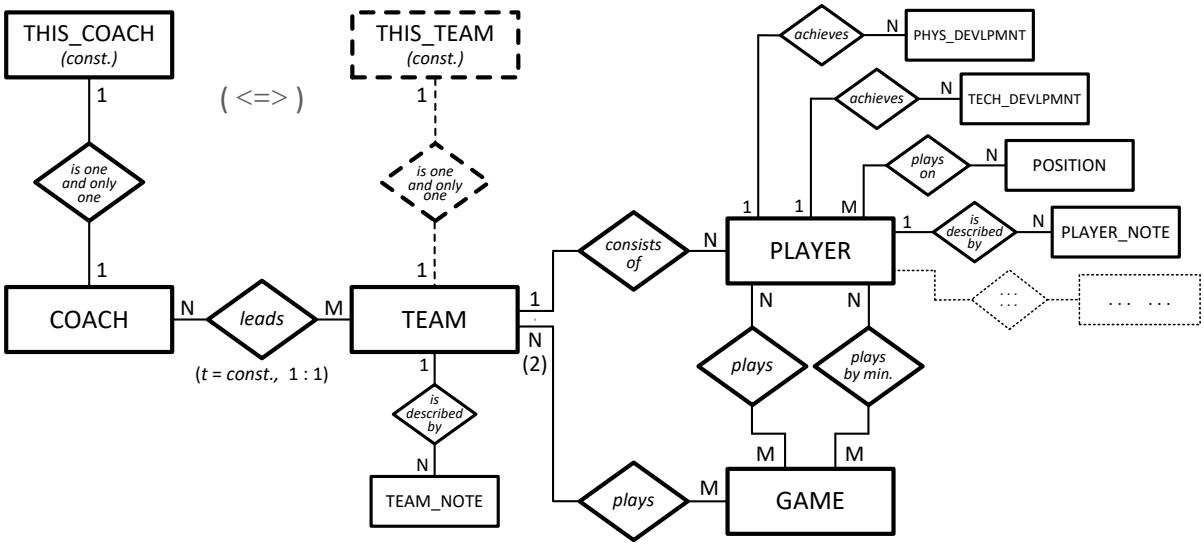
U svrhu poboljšanja uspješnosti predviđanja ishoda sportski događaji su podijeljeni u dvije kategorije događaja čime je omogućena primjena prilagodljovog postupka. Podjela događaja vrši se korištenjem projiciranog indeksa korisnosti suprotstavljenih događaja te se ovisno o definiranom ili izračunatom intervalu odlučuje da li je potrebna dodatna analiza trenutno analiziranog događaja. Konačni rezultati prikazuju se kada su ishodi svih događaja skupa za ispitivanje predviđeni.

5. ANALIZA REZULTATA ISPITIVANJA I OPTIMIRANJE PREDLOŽENOG MODELA NA PRIMJERU UTAKMICA NBA LIGE

Svojstva predložene metode za predviđanje sportskih ishoda zasnovane na indeksu korisnosti i optimalnom vremenskom prozoru ispitana su na modelu predviđanja utakmica NBA lige. Mjerne podatke čini devet uzastopnih NBA sezona, što čini skup od 11 578 utakmica. Za predviđanje je korišten matematički model predstavljen u poglavlju 4.

5.1. Ispitivanje modela

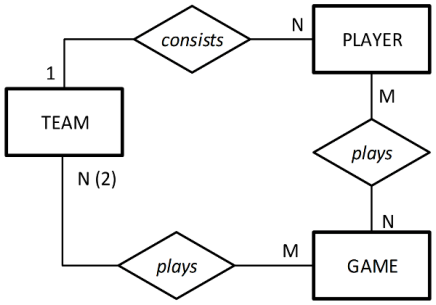
Za potrebe ispitivanja modela izgrađen je informacijski sustav Basketball Coach Assistant (BCA). BCA je informacijski sustav prvenstveno predviđen za upotrebu od strane košarkaških trenera, izgrađen korištenjem skripnog jezika PHP i relacijske baze MySQL. Aplikacija je na klijentskoj strani potpomognuta JavaScriptom i jQuery-em. Prva verzija informacijskog sustava prezentirana je na međunarodnoj konferenciji icSports 2015. godine u Lisabonu [79], a kasnije i na međunarodnoj konferenciju icSports 2019. u Beču [80]. Mogućnosti informacijskog sustava BCA su redovito nadograđivane u skladu s potrebama košarkaškog trenera i aktualnog istraživanja. Slika 5.1 prikazuje kompletan konceptualni model informacijskog sustava BCA.



Slika 5.1. Konceptualni model informacijskog sustava BCA.

Za potrebe istraživanja veći dio konceptualnog modela informacijskog sustava nije potreban pa će se naglasak dati na dio potreban za istraživanje. Slika 5.2 prikazuje dio konceptualnog modela potrebnog za istraživanje. Za istraživanje su dovoljna svega tri entiteta, a to su entiteti popisa momčadi, igrača i utakmica. Veze među entitetima su binarne, a iz kardinalnosti veza je vidljivo

da igrač može igrati više utakmica te da igrač u jednom trenutku može igrati za jednu momčad. Utakmicu igraju dvije momčadi, a momčad čini više igrača.



Slika 5.2. Dio konceptualnog modela informacijskog sustava BCA.

Standardnim pravilima pretvorbe konceptualni model je pretvoren u relacijski te je implementiran u MySQL relacijsku bazu podataka koja se koristi za pohranu prikupljenih podataka. Tablica 5.1 prikazuje dio relacijskog modela potreban za realizaciju modela predviđanja.

Tablica 5.1. Dio relacijskog modela informacijskog sustava BCA.

Relacija	**Atributi** (primarni ključ podcrtan, strani ključevi *ukošeni*)
team	id, name
player	id, name, surname, *team_id*
game	id, *home_team_id*, *guest_team_id*, result, 1st_quarter, 2nd_quarter, 3rd_quarter, 4th_quarter, overtime, date, playoff
statistics	*player_id*, *game_id*, *team_id*, 2fgm, 2fga, 3fgm, 3fga, ftm, fta, def_reb, of_reb, asist, st, to, bl, f

Ostatak relacijskog modela nije potrebno dodatno analizirati jer se ne koristi u istraživanju te nije od primarne važnosti za ovo istraživanje.

5.1.1. Skup odabranih značajki

Predloženi model koristi dva skupa ulaznih značajki. Prvi skup značajki čini 13 značajki osnovne košarkaške statistike i tri izlučene značajke izračunate na temelju osnovnih značajki košarkaške statistike. Skup od tri izlučene značajke su značajka broja promašaja za dva poena ($miss\_2fg$), značajka broja promašaja za tri poena ($miss\_3fg$) i značajka broja promašenih slobodnih bacanja ($miss\_ft$), prikazane redom formulama (3-1), (3-2) i (3-3). Značajke $2fga$, $3fga$ i fta se koriste isključivo za izlučivanje značajki $miss\_2fg$, $miss\_3fg$ i $miss\_ft$, te se u izvornom obliku neće koristiti tijekom ispitivanja predloženog modela predviđanja. Tablica 5.2 prikazuje podskup značajki osnovne košarkaške statistike i tri izlučene značajke. Drugi skup značajki čini osam izlučenih značajki, sedam vezanih uz uspješnost momčadi temeljenu na prethodnim rezultatima i značajka prednosti domaćeg terena definirana u odjeljku 3.3.2. Tablica 5.3 prikazuje skup izlučenih značajki vezanih uz prethodne rezultate te značajku prednosti domaćeg terena.

Tablica 5.2. Skup značajki košarkaške statistike.

Kratica	**Opis**
$2fgm$	Broj zabijenih pokušaja za dva poena
$2fgm$	Broj zabijenih pokušaja za tri poena
ftm	Broj zabijenih slobodnih bacanja
def_reb	Broj obrambenih skokova
of_reb	Broj napadačkih skokova
$asist$	Broj asistencija
st	Broj osvojenih lopti
to	Broj izgubljenih lopti
bl	Broj postignutih blokada
f	Broj počinjenih prekršaja
$miss\_2fg$	Broj promašaja za dva poena
$miss\_3fg$	Broj promašaja za tri poena
$miss\_ft$	Broj promašenih slobodnih bacanja

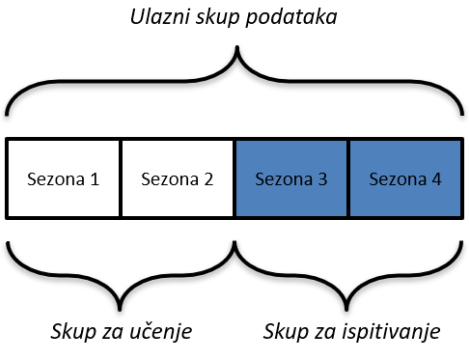
Tablica 5.3. Skup izlučenih značajki.

Naziv	**Kratica**		**Izračun na temelju**		**Izračun za**	
			Skup za učenje	**Skup za ispit.**	**Domaćin**	**Gost**
Omjer zadnjih 10 utakmica	om_{10}		✓	✓	✓	✓
Omjer domaćina u zadnjih 10 domaćih utakmica	om_{10d}	om_{10dg}	✓	✓	✓	
Omjer gosta u zadnjih 10 gostujućih utakmica	om_{10g}		✓	✓		✓
Međusobni omjer	om_{med}		✓	✓	✓	
Pobjednički niz	om_{pNiz}		✓	✓	✓	✓
Omjer faze ispitivanja	om_l			✓	✓	✓
Broj utakmica u zadnjih 10 dana	br_{10}		✓	✓	✓	✓
Prednost domaćeg terena	dt		✓	✓	✓	

5.1.2. Validacija modela

Validacija predloženog modela vršit će se metodom podjele skupa podataka. Ulazni podaci su podijeljeni u dva disjunktna, kronološki poredana skupa podataka u kojem prvi skup podataka predstavlja skup za učenje, a drugi skup za ispitivanje. Kao što je već navedeno, sportski događaji nisu u potpunosti neovisni događaji te samim time metoda unakrsne provjere nije pogodna.

Pitanje koje se ovdje postavlja je kako podijeliti skup podataka na skup za učenje i skup za ispitivanje, odnosno da li ih podijeliti na temelju broja utakmica, mjeseci, sezona ili nečeg sličnog. Logično izbor je podjela prema sezonama ili broju utakmica. Slika 5.3 prikazuje primjer odabira skupa za učenje i skupa za ispitivanje na temelju četiri sezone gdje prve dvije sezone predstavljaju skup za učenje, a druge dvije sezone skup za ispitivanje.



Slika 5.3. Metoda podjele skupa podataka.

Skup podataka će u ovom slučaju biti podijeljen na temelju sezona. Kao što je već ranije napomenuto, ulazni skup podataka će koristiti najviše tri sezone skupa za učenje i najviše dvije sezone skupa za ispitivanje. Također, skupu za učenje će se dodavati podaci skupa za ispitivanje nad kojima je predviđanje izvršeno.

5.2. Postupak optimiranja sveobuhvatnog indeksa korisnosti

Kao što je sugerirano i u naslovu rada, predviđanje ishoda će se vršiti korištenjem indeksa korisnosti i optimalnog vremenskog prozora. U odjeljku 2.3.1 je predložen sveobuhvatni indeks korisnosti. Valja napomenuti kako je indeks korisnosti relativni indikator kvalitete momčadi. Sveobuhvatni indeks korisnosti odlikuje svojstvo fleksibilnosti i lake prilagodljivosti ostalim indeksima. Početna točka predviđanja ishoda sportskih događaja bit će CTE indeks prilagođen indeksu NBA. Pojam sveobuhvatnog indeksa momčadi (CTE) definiran je u odjeljku 3.4.7.

Tablica 5.4. Početna točka definiranja doprinosa indeksa CTE prilagođenog indeksu NBA.

e	**Naziv elementa**	v_e	N_e	v'_e	N'_e
$2fg$	šutevi za dva poena	2	N_{2fgm}	1	N_{miss_2fg}
$3fg$	šutevi za tri poena	3	N_{3fgm}	1	N_{miss_3fg}
ft	slobodna bacanja	1	N_{ftm}	1	N_{miss_ft}
rbs	skokovi (def_reb, of_reb)	1	N_{rbs}= $\sum_{i=1}^{2} N_{rbs,i}$	–	–
$asts$	asistencije	1	N_{asist}	–	–
$stls$	osvojene lopte	1	N_{st}	–	–
tos	izgubljene lopte	–	–	1	N_{to}
$blcks$	blokade	1	N_{bl}	–	–
fls	prekršaji	–	–	–	–

Tablica 5.4 prikazuje doprinose elemenata indeksa CTE prilagođene indeksu NBA koji će predstavljati početnu točku predviđanja ishoda. Suma težinskih faktora W_e indeksa ostaje fiksirana

na kardinalnost samog skupa elemenata igre, gdje je težinski faktor svakog elementa košarkaške igre $W_e = 1$.

U odjeljku 3.4.2 je pokazano kako NBA indeks kao pokazatelj ishoda košarkaške utakmice korištenjem stvarnih podataka postiže točnost predviđanja ishoda od 92,31 % u slučaju davanja prednosti domaćoj momčadi, odnosno 91,83 % kada se u slučaju istog učinka NBA indeksa suprotstavljenih momčadi predviđanje ishoda smatra netočnim. Cilj odjeljka je optimizirati doprinose značajki indeksa CTE, točnije optimizirati doprinose elemenata osnovne košarkaške statistike (koeficijent $v_e(v'_e)$). Optimizacija će se vršiti iterativno, a uspješnost procesa optimizacije procijenit će se na temelju rezultata predviđanja.

Koristit će se period koji se pokazao najboljim, a to su jedna sezona skupa za učenje i jedna sezona skupa za ispitivanje. Rezultati optimizacije korištenjem isključivo linearne optimizacije, optimizacija rojem čestica (engl. *Particle Swarm Optimization)* i informacijska dobit (engl. *information gain*), nisu polučili bolje rezultate stoga je korišten postupak optimizacije korištenjem kombinacije nelinearnih i linearnih doprinosa predloženih u odjeljku 4.1.1.

Učinak igrača ili momčadi elemenata osnovne košarkaške statistike pozitivan je cijeli broj veći ili jednak nuli ($N_e, N_{tm,e} \geq 0$). Analizom skupa podataka važno je odrediti granice u kojima se pojedini elementi osnovne košarkaške statistike pojavljuju u stvarnosti te sukladno tome definirati skup prikladnih matematičkih funkcija.

Tablica 5.5. Minimalne, maksimalne, srednje i medijalne vrijednosti osnovne košarkaške statistike.

e	Minimalna vrijednost	Maksimalna vrijednost	Srednja vrijednost	Medijalna vrijednost
$2fgm$	9	56	29,87	30
$2fga$	28	113	60,91	61
$3fgm$	0	25	7,90	8
$3fga$	3	61	22,20	22
ftm	1	52	17,62	17
fta	1	64	23,23	23
def_reb	12	56	31,97	32
of_reb	0	38	10,68	10
$asist$	4	47	21,90	22
st	0	22	7,61	7
to	2	29	13,66	13
bl	0	18	4,89	5
f	5	42	20,32	20

Tablica 5.5 prikazuje stvarne minimalne, maksimalne, srednje i medijalne vrijednosti elemenata osnovne košarkaške statistike. Analizirajući rezultate vidljivo je da su rasponi minimalnih i

maksimalnih vrijednosti analiziranih značajki poprilično veliki. Rijetke su utakmice u kojoj momčad zabije svega devet pokušaja za dva poena, iskoristi 113 lopti kao pokušaj za dva poena, nema niti jedan pogodak za tri poena ili pak ima svega jedno slobodno bacanje. Takve vrijednosti se u statistici nazivaju ekstremnim (stršećim) vrijednostima, a događaju se vrlo rijetko te nisu pogodne za definiranje optimalne funkcije doprinosa. Prikladnijom mjerom smatraju se srednja i medijalna vrijednost. Razlika između srednje i medijalne vrijednosti je u prosjeku manja od 10 % što predstavlja zanemarivu razliku te će se kao polazna točka definiranja potencijalnih funkcija doprinosa koristiti srednja vrijednost. Analizirajući srednje vrijednosti elemenata košarkaške igre vidljivo je da doprinos kvadratne funkcije ili funkcija s još većom potencijom pridaju preveliki doprinos pojedinoj komponenti čime se doprinos ostalih elemenata košarkaške statistike može u potpunosti zagušiti. Ista stvar vrijedi i za eksponencijalne funkcije gdje povećanje doprinosa može biti još i veće. Tako će se u obzir uzeti nelinearne funkcije drugog korijena ($\sqrt{e}$) i trećeg korijena $\sqrt[3]{e}$ te skup odabranih linearnih funkcija. U ovom slučaju će se kao funkcija povećanja doprinosa uzeti i razlika učinaka. Razlika učinaka je razlika u učinku momčadi na zadanoj utakmici ili skupu utakmica. Doprinos se povećava momčadi koja je postigla veći učinak analiziranog elementa. Neka je učinak elementa e momčadi A prikazan kao $N_{e,A}$, a učinak momčadi B kao $N_{e,B}$ te neka su pripadajući koeficijenti povećanja doprinosa χ_A i χ_B. Formule (5-1) i (5-2) prikazuju izračun koeficijenata povećavanja doprinosa elementa e suprotstavljenih momčadi.

$$\chi_{e,A} = \begin{cases} 1, N_{e,A} > N_{e,B} \\ 0, N_{e,A} \leq N_{e,B} \end{cases} \tag{5-1}$$

$$\chi_{e,B} = \begin{cases} 1, N_{e,B} > N_{e,A} \\ 0, N_{e,B} \leq N_{e,A} \end{cases} \tag{5-2}$$

Nakon što je definiran koeficijent povećanja doprinosa suprotstavljenih momčadi, moguće je definirati i konačnu formulu izračuna doprinosa elementa e. Konačni doprinosi elementa e momčadi A i momčadi B prikazani su formulama (5-3) i (5-4) gdje $h(x)$ predstavlja funkciju izračuna povećanja doprinosa.

$$N_{e,A}(\chi_{e,A}) = N_{e,A} + \chi_A \times h(N_{e,A} - N_{e,B}) \tag{5-3}$$

$$N_{e,B}(\chi_{e,B}) = N_{e,B} + \chi_B \times h(N_{e,B} - N_{e,A}) \tag{5-4}$$

Momčadi koja ostvari veći učinak elementa e dodaje se razlika učinaka suprotstavljenih momčadi izračunata na temelju definirane funkcije izračuna povećanja doprinosa. Element koji je okarakteriziran kao pozitivan poveća te dodatno nagrađuje momčad koja je ostvarila veći učinak, dok element košarkaške igre koji je okarakteriziran kao negativan smanjuje te dodatno kažnjava momčad koja je ostvarila veći učinak negativnog elementa igre. Na taj se način pozitivni učinci

momčadi dodatno nagrađuju, a negativni učinci dodatno kažnjavaju. Vrlo je bitno i definirati granice koeficijenta $v_e(v'_e)$ čime se dodatno ograničava doprinos pojedine značajke. Granice koeficijenata $v_e(v'_e)$ se najčešće definiraju na temelju iskustva eksperta. Donja granica ($v_{e,min}(v'_{e,min})$) će biti 0, dok će gornja granica ($v_{e,max}(v'_{e,ax})$) biti definirana kao dvostruka vrijednost doprinosa definiranog NBA indeksom, što konkretno znači da će doprinos pojedinog elementa košarkaške igre biti u intervalu $[0, 2 \times v_{e,NBA}(v'_{e,NBA})]$.

U odjeljku 3.5.1 napravljena je analiza osjetljivosti NBA indeksa na skup korištenih značajki te je pokazano kako je NBA indeks najosjetljiviji na značajku promašenih pokušaja za dva poena ($miss\_2fg$), a najmanje osjetljiv na značajku načinjenih blokada (bl).

Postupak optimizacije doprinosa vršit će se iterativno definiranim redoslijedom na temelju osjetljivosti u kojem su u obzir uzeta sva mjerenja (Slika 3.5 a)), što znači da će se najprije analizirati učinak promjene doprinosa značajki s većom osjetljivošću. Vrijednost koeficijenta $v_e(v'_e)$ s najboljim rezultatom optimizacije će se koristiti kao referentna vrijednost iduće iteracije. Postupak optimizacije se vrši piramidalno u dvije faze. U prvoj fazi se koristi silazni redoslijed značajki, dok se u drugoj fazi koristi uzlazni redoslijedom definiranja doprinosa elemenata.

Tablica A.1 prikazuje rezultate optimizacije doprinosa koeficijenta $v_e(v'_e)$ korištenjem predloženog postupka optimizacije. Redoslijed značajki određen je osjetljivošću NBA indeksa prema skupu korištenih značajki. Element počinjenih prekršaja (f) koji nije dio osnovnog NBA indeksa je analiziran posljednji, a referentni doprinos navedene značajke iznosi v'_f= 1. Tablica 5.6 prikazuje rezultate prve faze optimizacije indeksa CTE gdje je redoslijed optimizacije komponenti označenim brojem u indeksu prefiksa učinka pojedine komponente.

Tablica 5.6. Tablica doprinosa elemenata košarkaške igre dobivena optimizacijom (prva faza).

$\boldsymbol{e}$	$[\boldsymbol{v_{min}}, \boldsymbol{v_{max}}]$	$\boldsymbol{v_e}$	${}^{\#}\boldsymbol{N_e}$	$[\boldsymbol{v_{min}}, \boldsymbol{v_{max}}]$	$\boldsymbol{v'_e}$	${}^{\#}\boldsymbol{N'_e}$
$2fg$	[0, 2]	2	${}^{3}N_{2fgm}$	[0, 1]	$miss\_2fg + \sqrt[3]{miss\_2fg\_diff}$	${}^{1}N_{miss\_2fg}$
$3fg$	[0, 2]	$3fgm - \sqrt[3]{3fgm}$	${}^{10}N_{3fgm}$	[0, 1]	1	${}^{7}N_{miss\_3fg}$
ft	[0, 2]	$ftm + \sqrt[3]{ftm\_diff}$	${}^{2}N_{ftm}$	[0, 1]	1	${}^{8}N_{miss\_ft}$
rbs	[0, 2]	$\frac{def\_reb}{2}$	${}^{4}N_{\text{def\_reb}}$	–	–	–
		1	${}^{6}N_{of\_reb}$	–	–	–
$asts$	[0, 2]	$asist - \sqrt[3]{asist}$	${}^{5}N_{asist}$	–	–	–
$stls$	[0, 2]	$\frac{st}{2}$	${}^{11}N_{2fgmst}$	–	–	–
tos	[0, 2]	–	–	[0, 1]	$to + \sqrt[3]{to\_diff}$	${}^{9}N_{to}$
$blcks$	[0, 2]	bl	${}^{12}N_{bl}$	–	–	–
fls	[0, 2]	–	–	[0, 1]	–	${}^{13}N_{f}$

Prvom fazom optimizacije točnost predviđanja je s 61,80 % povećana na 62,62 %, zabilježivši napredak od 0,82 %. Najveći porast točnosti dobiven je najosjetljivijom značajkom ($miss\_2fg$). Druga faza optimizacije se vrši obrnutim redoslijedom optimizacije značajki. Postupak se ponavlja krenuvši od značajke s najmanjom osjetljivošću. Logično, značajku koja je analizirana posljednja (f) nije potrebno ponovo analizirati. Postupak optimizacije se tako može opisati oblikom piramide gdje se u podnožju nalazi značajka s najvećom osjetljivošću, a na vrhu značajka s najmanjom osjetljivošću. U ovom slučaju na vrhu se nalazi dodana značajka (f) koja nije dio referentnog indeksa korisnosti. Slika 5.4 grafički prikazuje tijek optimizacije konkretnog problema.

Slika 5.4. Piramidalni prikaz postupka optimizacije.

Tablica A.2 prikazuje rezultate druge faze optimizacije doprinosa koeficijenta $v_e(v'_e)$. Ulaz u drugu fazu je konačan skup doprinosa prve faze. Doprinosi druge faze nadovezuju se na doprinose prve faze optimizacije, a koristi se isti skup mogućih doprinosa kao i za prvu fazu optimizacije. Tablica 5.7 prikazuje rezultate druge faze optimizacije gdje je redoslijed optimizacije komponenti označenim brojem u indeksu prefiksa učinka pojedine komponente.

Tablica 5.7. Tablica doprinosa elemenata košarkaške igre dobivena optimizacijom (druga faza).

e	$[v_{min}, v_{max}]$	v_e	${}^{\#}N_e$	$[v_{min}, v_{max}]$	v'_e	${}^{\#}N'_e$
$2fg$	[0, 2]	2	${}^{11}N_{2fgm}$	[0, 2]	1	${}^{13}N_{miss\_2fg}$
$3fg$	[0, 2]	$3fgm - \sqrt{3fgm}$	${}^{4}N_{3fgm}$	[0, 2]	1	${}^{7}N_{miss\_3fg}$
ft	[0, 2]	1	${}^{12}N_{ftm}$	[0, 2]	1	${}^{6}N_{miss\_ft}$
rbs	[0, 2]	$def\_reb + \sqrt{def\_reb}$	${}^{10}N_{\mathrm{def\_reb}}$	[0, 2]	–	–
		$of\_reb + \sqrt{of\_reb}$	${}^{8}N_{of\_reb}$	–	–	–
$asts$	[0, 2]	$asist + \sqrt{asist}$	${}^{9}N_{asist}$	[0, 2]	–	–
$stls$	[0, 2]	$\sqrt{st}$	${}^{3}N_{2fgmst}$	[0, 2]	–	–
tos	[0, 2]	–	–	[0, 2]	$to + \sqrt[3]{to}$	${}^{5}N_{to}$
$blcks$	[0, 2]	bl	${}^{2}N_{bl}$	[0, 2]	–	–
fls	[0, 2]	–	–	[0, 2]	–	${}^{1}N_f$

Druga faza optimizacije dala je bolje rezultate u odnosu na prvu fazu, a relativni napredak je očekivano slabiji u odnosu na prvu fazu. Točnije, rezultat predviđanja je porastao s 62,62 % iz prve faze na 62,85 % u drugoj fazi optimizacije što čini porast od 0,23 %. Ukupno povećanje optimizacija iznosi 1,15 %.

Bolje rezultate predviđanja može sugerirati i postotak kojim stvarni CTE indeks predviđa ishod utakmica objašnjen u odjeljku 3.4.1. Valja ponovo napomenuti kako je indeks korisnosti relativni indikator kvalitete momčadi te da viša uspješnost definiranja pobjednika korištenjem stvarnih podataka ne znači nužno i bolje sposobnosti predviđanja na temelju povijesnih podataka. Tablica 5.8 prikazuje u kojem je postotku indeks CTE, izračunat formulom (5-5) na temelju stvarnih podataka utakmice gm, pokazatelj ishoda utakmica u odnosu na osnovni NBA indeks i modificirani NBA indeks momčadi.

$$winn(I_{CTE}(tm_d), I_{CTE}(tm_g)) \begin{cases} tm_d, \dfrac{I_{CTE}(tm_d, gm)}{I_{CTE}(tm_g, gm)} \geq 1; \\ tm_d, \dfrac{I_{CTE}(tm_d, gm)}{I_{CTE}(tm_g, gm)} < 1; \end{cases} \tag{5-5}$$

Tablica 5.8. CTE indeks kao pokazatelj ishoda utakmice.

Sezona	NBA indeks	Modificirani NBA indeks	CTE indeks
2009./2010.	92,76 %	93,75 %	92,76 %
2010./2011.	92,60 %	93,97 %	92,98 %
2011./2012.	91,71 %	93,58 %	92,92 %
2012./2013.	92,24 %	93,99 %	92,09 %
2013./2014.	92,65 %	94,62 %	92,80 %
2014./2015.	92,30 %	94,81 %	93,44 %
2015./2016.	92,55 %	94,07 %	92,86 %
2016./2017.	92,13 %	92,97 %	91,52 %
2017./2018.	91,77 %	93,67 %	91,92 %
Prosjek:	**92,31 % (10688/11578)**	**93,95 % (10877/11578)**	**92,58 % (10719/11578)**

Najbolji rezultat predviđanja ishoda na temelju stvarnih podataka dobiven je korištenjem modificiranog NBA indeksa, dok najlošije rezultate predviđanja daje osnovni NBA indeks. Također, vidljivo je da se značajka počinjenih prekršaja (f), koja je dio modificiranog NBA indeksa, optimizacijom indeksa CTE pokazala suvišnom, bez obzira što kao dio modificiranog NBA indeksa daje najbolje rezultate definiranja pobjednika na temelju stvarnih podataka.

5.2.1. Predviđanje na temelju optimiziranog indeksa CTE

Rezultati predviđanja korištenjem osnovnog NBA indeksa su dani u odjeljku 3.7.2. Najbolji rezultati su dobiveni korištenjem jedne sezone skupa za učenje i jedne sezone skupa za ispitivanje.

Povećanjem ulaznog skupa podataka, u ovom slučaju povećanjem broja utakmica skupa za učenje, točnost je rasla, dostigla svoj maksimum te počela padati. Navedena činjenica je pokazala da postoji optimalni vremenski period koji daje najbolje rezultate predviđanja. U ovom odjeljku će se prikazati rezultati predviđanja korištenjem optimiziranog indeksa CTE na cijelom ulaznom skupu podataka. Cilj je dobiti bolje rezultate predviđanja u odnosu na osnovni NBA indeks i modificirani NBA indeks. Predviđanje će se vršiti na temelju prosječnih podataka skupa za učenje, a uz početni skup za učenje koristit će se i podaci skupa za ispitivanje nad kojima je predviđanje izvršeno. Pretpostavka je da će rezultati optimiziranog indeksa CTE biti bolji u odnosu na osnovni NBA indeks i modificirani NBA indeks. Tablica 5.9 prikazuje rezultate predviđanja korištenjem indeksa CTE u odnosu na NBA indeks i modificirani NBA indeks. Formula (5-6) prikazuje način izračuna ishoda, gdje n predstavlja broj utakmica skupa za učenje.

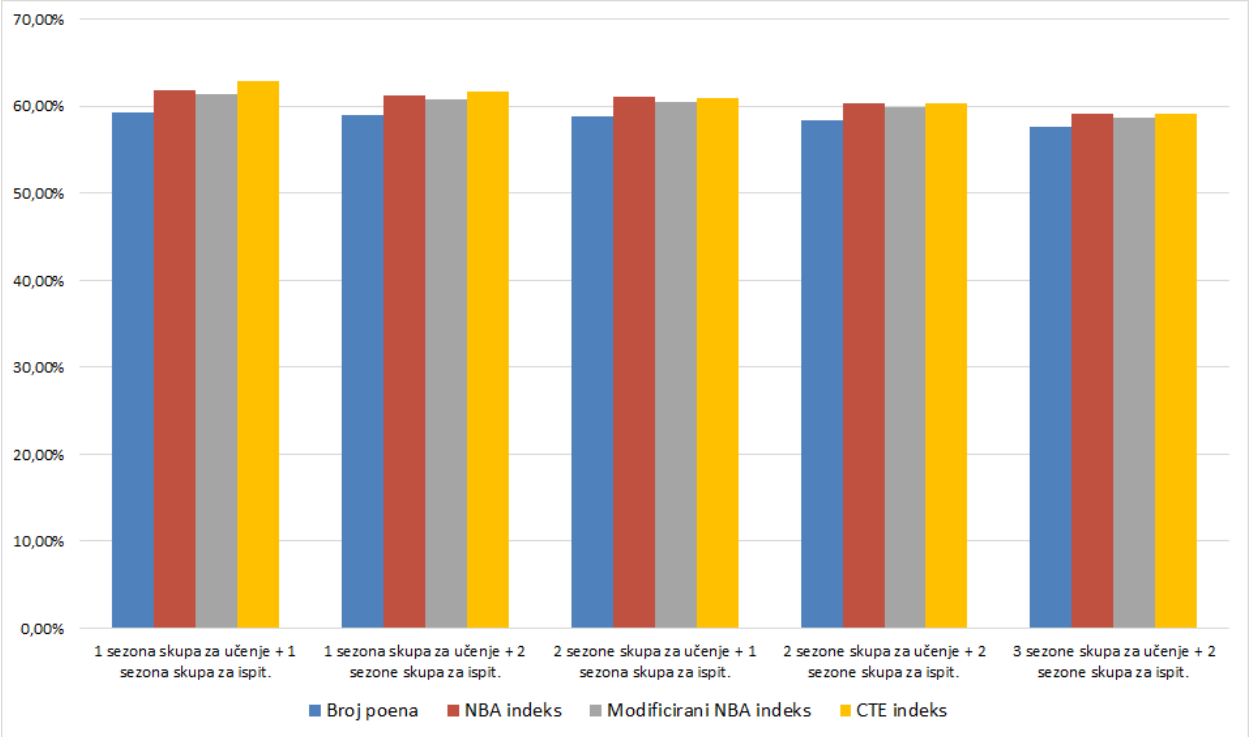
$$winn(I_{CPE}(tm_d), I_{CPE}(tm_g)) \begin{cases} tm_d, \dfrac{\frac{1}{n}\sum_{i=1}^{n} I_{CPE}(tm_d, i)}{\frac{1}{n}\sum_{i=1}^{n} I_{CPE}(tm_g, i)} \geq 1; \\ tm_g, \dfrac{\frac{1}{n}\sum_{i=0}^{n} I_{CPE}(tm_d, i)}{\frac{1}{n}\sum_{i=1}^{n} I_{CPE}(tm_g, i)} < 1; \end{cases} \quad (5\text{-}6)$$

Tablica 5.9. Rezultati predviđanja varijanti NBA indeksa u odnosu na optimizirani indeks CTE.

		NBA indeks		Modificirani NBA indeks		Optimizirani indeks CTE	
Skup za učenje	**Skup za ispitivanje**	**Točnost**	**Prosjek**	**Točnost**	**Prosjek**	**Točnost**	**Prosjek**
2016./2017.	2017./2018.	63,72 %	61,80 % (6344/10266)	62,73 %	61,40 % (6303/10266)	64,18 %	62,85 % (6452/10266)
2015./2016.	2016./2017.	60,35 %		59,97 %		63,10 %	
2014./2015.	2015./2016.	63,45 %		63,30 %		64,44 %	
2013./2014.	2014./2015.	61,94 %		61,48 %		62,32 %	
2012./2013.	2013./2014.	60,05 %		58,38 %		60,88 %	
2011./2012.	2012./2013.	61,72 %		60,96 %		63,39 %	
2010./2011.	2011./2012.	64,71 %		64,80 %		64,25 %	
2009./2010.	2010./2011.	58,96 %		60,18 %		60,49 %	
2015./2016.	2016./2017. – 2017./2018.	61,24 %	61,28 % 10974/17909	59,94 %	60,87 % (10902/17909)	62,65 %	61,79 % (11066/17909)
2014./2015.	2015./2016. – 2016./2017.	62,97 %		62,32 %		63,43 %	
2013./2014.	2014./2015. – 2015./2016.	61,97 %		62,43 %		62,39 %	
2012./2013.	2013./2014. – 2014./2015.	59,62 %		59,70 %		59,43 %	
2011./2012.	2012./2013. – 2013./2014.	60,16 %		58,79 %		61,45 %	
2010./2011.	2011./2012. – 2012./2013.	63,23 %		62,69 %		63,11 %	
2009./2010.	2010./2011. – 2011./2012.	59,79 %		60,38 %		60,04 %	
2015./2016. – 2016./2017.	2017./2018.	62,12 %	61,04 % 5466/8955	59,91 %	60,55 % (5422/8955)	62,20 %	60,93 % (5456/8955)
014./2015. – 2015./2016.	2016./2017.	62,49 %		61,34 %		62,41 %	
2013./2014. – 2014./2015.	2015./2016.	62,01 %		63,37 %		62,46 %	
2012./2013. – 2013./2014.	2014./2015.	59,19 %		61,02 %		57,97 %	
2011./2012. – 2012./2013.	2013./2014.	58,61 %		56,63 %		59,51 %	
2010./2011. – 2011./2012.	2012./2013.	62,02 %		60,96 %		62,18 %	
2009./2010. – 2010./2011.	2011./2012.	60,80 %		60,61 %		59,50 %	
2014./2015. – 2015./2016.	2016./2017. – 2017./2018.	61,08 %	60,31 % 9362/15524	60,90 %	59,92 % (9302/15524)	60,82 %	60,24 % (9351/15524)
2013./2014. – 2014./2015.	2015./2016. – 2016./2017.	61,45 %		61,75 %		61,90 %	
2012./2013. – 2013./2014.	2014./2015. – 2015./2016.	60,07 %		61,52 %		59,76 %	
2011./2012. – 2012./2013.	2013./2014. – 2014./2015.	58,82 %		57,49 %		58,78 %	
2010./2011. – 2011./2012.	2012./2013. – 2013./2014.	60,31 %		58,75 %		60,43 %	
2009./2010. – 2010./2011.	2011./2012. – 2012./2013.	60,09 %		59,92 %		59,67 %	
2013./2014. – 2015./2016.	2016./2017. – 2017./2018.	59,86 %	59,21 % 7778/13136	59,60 %	58,72 % (7713/13136)	59,86 %	59,31 % (7791/13136)
2012./2013. – 2014./2015.	2015./2016. – 2016./2017.	61,10 %		61,75 %		61,41 %	
2011./2012. – 2013./2014.	2014./2015. – 2015./2016.	59,27 %		58,96 %		58,96 %	
2010./2011. – 2012./2013.	2013./2014. – 2014./2015.	57,45 %		55,86 %		57,45 %	
2009./2010. – 2011./2012.	2012./2013. – 2013./2014.	58,37 %		57,42 %		58,87 %	

Najbolji rezultati dobiveni su korištenjem optimiziranog indeks CTE na temelju jedne sezone skupa za učenje i jedne sezone skupa za ispitivanje. Rezultati predviđanja su pokazali kako bolji

rezultati korištenjem stvarnih podataka za definiranje pobjednika ne znače nužno i bolje rezultate predviđanja ishoda nad povijesnim podacima. Navedena spoznaja je vidljiva ukoliko se usporede rezultati modificiranog NBA indeksa i indeksa CTE. Najbolji rezultat definiranja pobjednika na temelju stvarnih podataka dobiven je korištenjem modificiranog NBA indeksa, a najbolji rezultati predviđanja su dobiveni korištenjem indeksa CTE. Slika 5.5 prikazuje usporedbu rezultata predviđanja korištenjem prosječnih učinaka.



Slika 5.5. Usporedba rezultata previđanja korištenjem prosječnih učinaka za različite modele predviđanja u odnosu na veličinu skupova za učenje i ispitivanje.

Najbolji rezultati predviđanja dobiveni su korištenjem optimiziranog indeksa CTE, a prosječna točnost predviđanja opada povećanjem broja sezona skupa za učenje i skupa za ispitivanje, što je još jedan pokazatelj postojanja vremenskog perioda koji daje optimalne rezultate predviđanja.

5.2.2. Uvođenje značajke prednosti domaćeg terena

U odjeljku 3.3.1 je pokazano kako u NBA ligi postoji prednost domaćeg terena, a matematički opis definiranja značajki prednosti uspješnijeg procesa na temelju kojeg će se definirati i značajka prednosti domaćeg terena dan je u odjeljku 4.1.2. Cilj ovog odjeljka je pokazati kako uvođenje značajke prednosti domaćeg terena u kombinaciji s indeksom korisnosti može dati bolje rezultate predviđanja. Kao referentna točka definiranja prednosti domaćeg terena uzet će se razlika postotka pobjeda domaćih momčadi u odnosu na gostujuće, a izračunavat će se na temelju poznate povijesti. U kasnijim iteracijama uz skup za učenje koriste se i podaci skupa za ispitivanje nad kojima je

predviđanje izvršeno. Značajka prednosti domaćeg terena se ponovno izračunava prilikom svake sljedeće iteracije algoritma predviđanja. Formula (5-7) prikazuje način izračuna značajke prednosti domaćeg terena.

$$\Delta tm = \frac{N_{tm_d}}{N_{tm_d} + N_{tm_g}} - \frac{N_{tm_g}}{N_{tm_d} + N_{tm_g}} \tag{5-7}$$

Osnovna ideja korištenja značajke prednosti domaćeg terena je povećanje projiciranog CTE indeksa domaće momčadi. Ukoliko projicirani CTE indeks domaće momčadi označimo kao $I_{CTE}(tm_d)$, a projicirani CTE indeks gostujuće momčadi kao $I_{CTE}(tm_g)$, projicirani CTE indeks domaćina ćemo povećati za umnožak razlike postotka pobjede domaćina u odnosu na gosta (Δtm) i projiciranog indeksa domaćina. Projicirani CTE indeks domaćina ćemo zapisati kao $I_{CTE}(tm_d(dt))$. U odjeljku 4.1.2 je uveden i pripadajući koeficijent doprinosa razlike postotka nazvan korektivni faktor (kf). Formula (5-8) prikazuje izračun konačnog projiciranog indeksa korisnosti domaće momčadi. Projicirani indeks gostujuće momčadi ostaje nepromijenjen.

$$I_{CPE}(tm_d(dt)) = I_{CPE}(tm_d) + I_{CPE}(tm_d) \times kf \times \Delta tm, kf \geq 0 \tag{5-8}$$

Optimalni korektivni faktor (kf) će se izračunati korištenjem skupa linearnih i nelinearnih funkcija doprinosa. Kao i u prethodnim istraživanjima pobjednikom će se proglasiti ekipa s većim projiciranim indeksom korisnosti. Formulom (5-9) prikazan je način definiranja pobjedničke momčadi. Pobjeda domaćina definirana je u slučaju kada navedeni omjer daje broj veći ili jednak jedan, dok se pobjedom gosta smatra rezultat manji od jedan.

$$winn(I_{CTE}(tm_d(dt)), I_{CTE}(tm_g)) \begin{cases} tm_d, \dfrac{\frac{1}{n}\sum_{i=1}^{n} I_{CTE}(tm_d(dt), i)}{\frac{1}{n}\sum_{i=1}^{n} I_{CTE}(tm_g, i)} \geq 1; \\ tm_g, \dfrac{\frac{1}{n}\sum_{i=1}^{n} I_{CTE}(tm_d(dt), i)}{\frac{1}{n}\sum_{i=1}^{n} I_{CTE}(tm_g, i)} < 1; \end{cases} \tag{5-9}$$

Tablica 5.10 prikazuje rezultate optimizacije parametra kf gdje je parametar kf ograničen vrijednošću iz intervala $[0, 2]$. Referentna vrijednost dobivena je optimizacijom indeksa CTE te iznosi 62,85 % te ne uključuje korištenje značajke prednosti domaćeg terena, a samim time i parametra kf.

Tablica 5.10. Ovisnost rezultata predviđanja o parametru kf.

Smanjeni doprinos parametra kf						**kf**	**Pojačani doprinos parametra kf**			
$\frac{kf}{8}$	$\frac{kf}{4}$	$\frac{kf}{3}$	$\frac{kf}{2}$	$\frac{2kf}{3}$	$\frac{3kf}{4}$		$\frac{3}{2} \times kf$	$2 \times kf$	$\sqrt{kf}$	$\sqrt[3]{kf}$
63,68 %	64,11 %	64,52 %	63,65 %	62,60 %	62,26 %	60,84 %	59,71 %	59,37 %	59,19 %	59,17 %

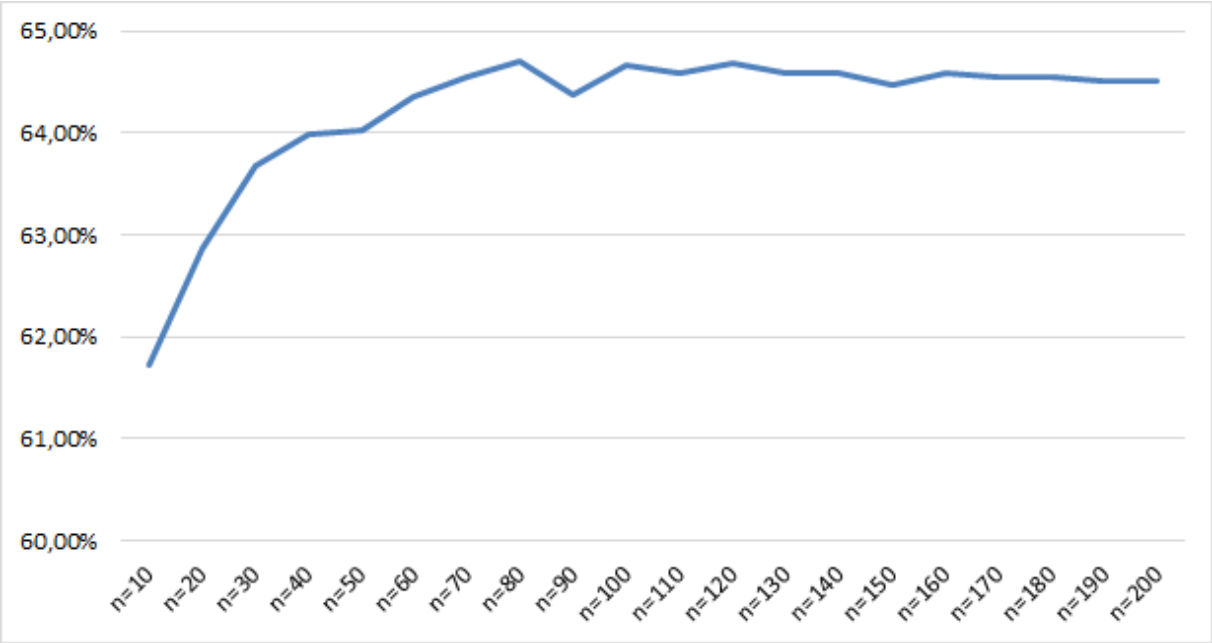
Rezultati predviđanja su pokazali da prednost domaćeg terena izražena kao razlika postotka pobjeda doprinosi boljim rezultatima. Najbolji rezultat je dobiven korištenjem smanjenog doprinosa razlike postotka pobjede domaćih momčadi u odnosu na gostujuće, točnije u slučaju kada je $kf = \frac{1}{3}$. Tako je zabilježen porast od 1,67 % u odnosu na indeks CTE te 2,72 % u odnosu na NBA indeks.

Najbolji rezultati predviđanja su dobiveni korištenjem jedne sezone skupa za učenje i jedne sezone skupa za ispitivanje. Rezultati odjeljka 3.7.2, u kojem je za predviđanje ishoda korišten prosječan NBA indeks momčadi, pokazali su kako postoji vremenski period koji daje najbolje rezultate predviđanja. Najbolji rezultati predviđanja dobiveni su korištenjem zadnjih $[80, 160]$ utakmica pojedine momčadi. Sukladno istraživanju iz odjeljka 3.7.2, izvršit će se ispitivanje indeksa CTE s uključenom značajkom prednosti domaćeg terena. Broj utakmica skupa (n) za učenje će biti prikazan deseticama, $n \in [10, 200]$. U slučaju kada broj skupa za učenje ne sadrži dovoljan broj utakmica, ispitivanje će se izvršiti na najvećem mogućem broju utakmica koje nudi skup za učenje. Ulazni skup podataka čini jedna sezona skupa za učenje i jedna sezona skupa za ispitivanje. Tablica 5.11 prikazuje rezultate predviđanja korištenjem indeksa CTE s uključenom značajkom prednosti domaćeg terena na unaprijed definiranom broju utakmica skupa za učenje.

Tablica 5.11. Rezultati predviđanja upotrebom indeksa CTE i prednosti domaćeg terena na definiranom broju utakmica za učenje.

$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 60$	$n = 70$	$n = 80$	$n = 90$	$n = 100$
61,73 %	62,86 %	63,68 %	63,99 %	64,03 %	64,35 %	64,56 %	64,70 %	64,37 %	64,66 %
$n = 110$	$n = 120$	$n = 130$	$n = 140$	$n = 150$	$n = 160$	$n = 170$	$n = 180$	$n = 190$	$n = 200$
64,58 %	64,69 %	64,59 %	64,59 %	64,47 %	64,58 %	64,56 %	64,56 %	64,52 %	64,52 %

Rezultati pokazuju kako prosječna točnost povećanjem broja utakmica skupa za učenje raste, da svoj maksimum postiže oko $n = 80$, nakon toga bilježi lagani pad te na kraju stagnira. Stagnacija rezultata je očekivana s obzirom da većina momčadi odigra 82 utakmice u jednoj natjecateljskoj sezoni te da samo nekolicina dostigne brojku od 100 utakmica po sezoni.



Slika 5.6. Rezultati predviđanja korištenjem indeksa CTE i značajke prednosti domaćeg terena.

Slika 5.6 samo potvrđuje prethodno navedene zaključke. Vremenski period koji daje najbolje rezultate predviđanja je $n \in [60, 120]$.

Tablica 5.12 prikazuje usporedbu rezultata predviđanja korištenjem NBA indeksa, indeksa CTE i indeksa CTE sa značajkom prednosti domaćeg terena.

Tablica 5.12. Usporedba rezultata predviđanja na temelju cijele poznate povijesti.

Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	**NBA indeks**	**CTE indeks**	**CTE indeks s značajkom prednosti domaćeg terena $kf = \frac{1}{3}$**
1	1	61,80 %	62,85 %	64,52 %
1	2	61,28 %	61,79 %	64,63 %
2	1	61,04 %	60,93 %	64,66 %
2	2	60,31 %	60,24 %	64,68 %
3	2	59,21 %	59,31 %	64,46 %

Rezultati korištenjem značajke prednosti domaćeg terena dali su najbolje rezultate, te za razliku od ostalih varijanti pokazuju i lagani porast ili stagnaciju rezultata povećanjem ulaznog skupa podataka što nije slučaj vezan uz indekse NBA i CTE.

5.3. Predviđanje na temelju optimalnog vremenskog prozora

Dosadašnja istraživanja su pokazala kako postoji vremenski period koji najbolje opisuje stanje momčadi. Također je pokazano kako značajka prednosti domaćeg terena s pripadnim korektivnim faktorom može dodatno poboljšati rezultate predviđanja. Cilj vremenskog perioda, u ovom slučaju nazvanog optimalnim vremenskim prozorom, je odabrati podskup skupa za učenje koji će dati najbolje rezultate predviđanja te samim time smanjiti ulazni skup podataka bez posljedičnog smanjenja rezultata predviđanja.

U potpoglavlju 4.1 su predložena dva moguća načina izračuna optimalnog vremenskog prozora. Prvi način uključuje korištenje relativnog učinka (u ovom slučaju relativnog rezultata), a drugi način uključuje korištenje relativnog indeksa korisnosti. Oba pojma i sam algoritam izračuna i prilagodbe optimalnog vremenskog prozora su opisani u potpoglavlju 4.2.

Za rješavanje trenutnog problema će se koristi relativan rezultat i relativan indeks CTE. Dobiveni rezultati korištenjem relativnog indeksa lošiji su u odnosu na način korištenja relativnog rezultata. Što se tiče relativnog rezultata, korištene su dvije varijante. Prva varijanta optimalnim vremenskim prozorom proglašava potencijalni vremenski prozor čiji je prosječni relativni rezultat najbliži broju jedan. Logika vezana uz broj jedan i relativni rezultat je već objašnjena. Druga varijanta, odnosno varijanta koja je dala bolje rezultate, optimalnim vremenskim prozorom je proglasila najduži potencijalni vremenski prozor s prosječnim relativnim rezultatom u intervalu [0,95 – 1,05]. U slučaju kada ne postoji niti jedan potencijalni vremenski prozor s prosječnim rezultatom u intervalu [0,95 – 1,05], optimalnim vremenskim prozorom se proglašava potencijalni vremenski prozor s prosječnim relativnim rezultatom najbližim broju jedan.

5.3.1. Rezultati predviđanja korištenjem optimalnog vremenskog prozora

U ovom odjeljku će se prikazati rezultati predviđanja korištenjem optimalnog vremenskog prozora i indeksa CTE. Svaka momčad ima svoje posebnosti, a samim time vrlo vjerojatno i različit vremenski prozor koji najbolje oslikava njeno stanje. Stoga je vrlo važno izračunati optimalni vremenski prozor domaće momčadi (oP_{tm_d}) te optimalni vremenski prozor gostujuće momčadi (oP_{tm_g}). Konačan optimalni vremenski prozor za predviđanje ishoda košarkaških utakmica definiran je unijom optimalnih vremenskih prozora suprotstavljenih momčadi ($oP = oP_{tm_d} \cup oP_{tm_g}$).

Važan segment prilagodbe optimalnog vremenskog prozora je parametar b koji predstavlja broj dozvoljenih krivih predviđanja za pojedinu momčad nakon čega je potrebno vršiti prilagodbu optimalnog vremenskog prozora te rezultati iz odjeljka 3.7.5 koji su pokazali da najbolje rezultate predviđanja daje korištenje $[60, 120]$ utakmica skupa za učenje. U obzir će se uz indeks CTE uzeti i značajka prednosti domaćeg terena. Formula (5-10) prikazuje izračun CTE indeksa momčadi domaćina s definiranom prednošću domaćeg terena na temelju izračunatog optimalnog vremenskog prozora gdje $n_{\Delta t}$ predstavlja broj utakmica optimalnog vremenskog prozora.

$$winn(I_{CTE}(tm_d(dt)), I_{CTE}(tm_g)) \begin{cases} tm_d, \dfrac{\frac{1}{n_{\Delta t}}\sum_{i=1}^{n_{\Delta t}} I_{CTE}(tm_d(dt), i)}{\frac{1}{n_{\Delta t}}\sum_{i=1}^{n_{\Delta t}} I_{CTE}(tm_g, i)} \geq 1; \\ tm_g, \dfrac{\frac{1}{n_{\Delta t}}\sum_{i=1}^{n_{\Delta t}} I_{CTE}(tm_d(dt), i)}{\frac{1}{n_{\Delta t}}\sum_{i=1}^{n_{\Delta t}} I_{CTE}(tm_g, i)} < 1; \end{cases} \quad (5\text{-}10)$$

Tablica 5.13 prikazuje rezultate predviđanja korištenjem indeksa CTE, optimalnog vremenskog prozora i značajke prednosti domaćeg terena.

Tablica 5.13. Rezultati predviđanje korištenjem indeksa CTE, optimalnog vremenskog prozora i značajke prednosti domaćeg terena.

Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	$\boldsymbol{b}$ =							
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	1	64,55 %	64,51 %	64,59 %	64,68 %	64,76 %	64,63 %	64,57 %	64,65 %
1	2	64,50 %	64,43 %	64,48 %	64,50 %	64,51 %	64,41 %	64,42 %	64,42 %
2	1	64,53 %	64,48 %	64,57 %	64,42 %	64,32 %	64,34 %	64,29 %	64,28 %
2	2	64,58 %	64,43 %	64,46 %	64,50 %	64,40 %	64,37 %	64,38 %	64,29 %
3	2	64,47 %	64,25 %	64,33 %	64,34 %	64,32 %	64,18 %	64,26 %	64,30 %
Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	$\boldsymbol{b}$ =							
		8	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	1	64,69 %	64,58 %	64,56 %	64,59 %	64,55 %	64,50 %	64,58 %	64,67 %
1	2	64,41 %	64,36 %	64,35 %	64,34 %	64,27 %	64,35 %	64,39 %	64,25 %
2	1	64,17 %	64,34 %	64,13 %	64,25 %	64,18 %	64,14 %	64,15 %	64,12 %
2	2	64,31 %	64,35 %	64,31 %	64,37 %	64,31 %	64,22 %	64,19 %	64,11 %
3	2	64,27 %	64,17 %	64,05 %	64,30 %	64,21 %	64,11 %	64,21 %	64,02 %
Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	$\boldsymbol{b}$ =							
		16	**17**	**18**	**19**	**20**	**21**	**22**	**23**
1	1	64,60 %	64,54 %	64,60 %	64,63 %	64,59 %	64,55 %	64,47 %	64,47 %
1	2	64,24 %	64,05 %	64,29 %	64,31 %	64,36 %	64,26 %	64,15 %	64,21 %
2	1	64,19 %	64,18 %	64,14 %	64,06 %	63,95 %	64,05 %	63,94 %	63,96 %
2	2	64,24 %	64,20 %	64,09 %	64,13 %	64,10 %	64,03 %	63,95 %	64,11 %
3	2	63,94 %	64,00 %	64,12 %	64,01 %	64,00 %	63,89 %	63,84 %	63,98 %

Rezultati korištenjem optimalnog vremenskog prozora i pripadajućeg parametra b daju bolje rezultate u odnosu na korištenje cijelog poznatog skupa ulaznih podataka. Povećanjem parametra b točnost raste, u jednom trenutku dostiže svoj maksimum te počinje stagnirati ili u prosjeku lagano padati. Najbolji rezultati su dobiveni korištenjem jedne sezone skupa za učenje i jedne sezone skupa za ispitivanje te parametrom $b = 4$. Povećanje iznosi 0,24 % u odnosu na korištenje indeksa CTE na temelju cijelog poznatog ulaznog skupa i značajke prednosti domaćeg terena. Također je važno napomenuti da korištenjem optimalnog vremenskog prozora pogled u prošlost nije od primarne važnosti. Razlog tome je korištenje vremenskog perioda koji najbolje opisuje stanje momčadi, a koji se računa na temelju zadnjih $[60, 120]$ utakmica pojedine momčadi, što se pokazalo najboljim pristupom.

5.4. Predviđanje na temelju izlučenih značajki

Osim osnovnih elemenata košarkaške statistike i značajke prednosti domaćeg terena, predloženi model koristi i skup izlučenih značajki. Tablica 5.3 prikazuje popis izlučenih značajki i pripadajućih kratica koji će se koristiti u istraživanju. Za potrebe istraživanja analiziran je veći skup izlučenih značajki, a konačan skup izlučenih značajki rezultat je analize radova drugih istraživača, dobiven prema iskustvu autora ili eksperimentalnim ispitivanjem. Sve izvedene značajke dobivene su transformacijama nad ulaznim skupom podataka.

5.4.1. Rezultati predviđanja korištenjem skupa izlučenih značajki

U ovom odjeljku će se prikazati rezultati predviđanja korištenjem pojedinačnih izlučenih značajki. Ispitivanja će se vršiti na temelju cijele poznate povijesti. Tablica 5.14 prikazuje rezultate predviđanja korištenjem pojedinačnih izlučenih značajki. U slučaju istog rezultata za obje momčadi predviđanje se smatra neuspješnim. Vrijednosti koje mogu poprimiti izlučene značajke isključivo su diskretne, odnosno iz skupa prirodnih brojeva $N_{zn} \in N$. Formula (5-11) prikazuje način izračuna konačnog ishoda korištenjem skupa izlučenih značajki gdje n_{zn} predstavlja broj izlučenih značajki, a $N_{zn,i}$ vrijednost izlučene značajke. Korištene izlučene značajke, točnije kratice izlučenih značajki, prikazane u nastavku poglavlja koje će se koristiti za eksperimentalno ispitivanje prikazane su tablicom 5.3.

$$winn(N_{zn}(tm_d), N_{zd}(tm_g)) \begin{cases} tm_d, \dfrac{\sum_{i=1}^{n_{zn}} N_{zn,i}(tm_d)}{\sum_{i=1}^{n_{zn}} N_{zn,i}(tm_g)} > 1; \\ tm_x, \dfrac{\sum_{i=1}^{n_{zn}} N_{zn,i}(tm_d)}{\sum_{i=1}^{n_{zn}} N_{zn,i}(tm_g)} = 0; \\ tm_g, \dfrac{\sum_{i=1}^{n_{zn}} N_{zn,i}(tm_d)}{\sum_{i=1}^{n_{zn}} N_{zn,i}(tm_g)} < 1; \end{cases} \quad (5\text{-}11)$$

Tablica 5.14. Rezultati predviđanja korištenjem pojedinačnih izlučenih značajki.

Br. sezona skupa za učenje	Br. sezona skupa za ispit.	Izlučena značajka (bez davanja prednosti domaćeg terena)					
		om_1	om_{10}	om_{10dg}	om_{med}	om_{pNiz}	br_{10}
1	1	61,98 %	54,49 %	58,69 %	50,88 %	50,31 %	28,50 %
1	2	62,78 %	54,45 %	58,68 %	52,28 %	50,21 %	28,53 %
2	1	61,79 %	54,41 %	58,48 %	53,47 %	50,09 %	28,39 %
2	2	62,63 %	54,61 %	58,50 %	54,03 %	50,16 %	28,56 %
3	2	62,62 %	54,89 %	58,37 %	54,42 %	50,21 %	28,46 %

Najveću prosječnu točnost daje značajka *Omjer tijekom faze ispitivanja*, dok najlošije rezultate predviđanja daje značajka *Broj utakmica u zadnjih 10 dana*. Lošiji rezultati vezani su uz značajke koje mogu poprimiti mali broj vrijednosti kao što su primjerice značajke *Broj utakmica u zadnjih 10 dana* ili *Pobjednički niz*, pa čak i značajka *Međusobni omjer*.

U odjeljku 3.3.1 je pokazano kako postoji prednost domaćeg terena u NBA ligi te da je prednost domaćeg terena izraženija tokom faze doigravanja. Tablica 5.15 prikazuje rezultate pojedinačnih izvedenih značajki, a u slučaju istog rezultata za obje momčadi pobjednikom se proglašava domaća momčad. Pretpostavka je da će rezultati korištenjem prednosti domaćeg terena dati bolje rezultate. Rezultati predviđanja poredani su prema uspješnosti predviđanja, dok je formulom (5-12) prikazan način izračuna konačnog ishoda.

$$winn(N_{zn}(tm_d), N_{zn}(tm_g)) \begin{cases} tm_d, \dfrac{\sum_{i=1}^{n_{zn}} N_{zn,i}(tm_d)}{\sum_{i=1}^{n_{zn}} N_{zn,i}(tm_g)} \geq 1; \\ tm_d, \dfrac{\sum_{i=1}^{n_{zn}} N_{zn,i}(tm_d)}{\sum_{i=1}^{n_{zn}} N_{zn,i}(tm_g)} < 1; \end{cases} \tag{5-12}$$

Tablica 5.15. Rezultati predviđanja korištenjem pojedinačnih izlučenih značajki i značajke prednosti domaćeg terena.

Br. sezona skupa za učenje	Br. sezona skupa za ispit.	Izlučena značajka (davanje prednosti domaćeg terena)					
		om_l	om_{10}	om_{10dg}	om_{med}	om_{pNiz}	br_{10}
1	1	64,84 %	62,84 %	62,75 %	60,35 %	57,84 %	54,47 %
1	2	64,44 %	62,77 %	62,43 %	60,01 %	57,73 %	54,59 %
2	1	64,68 %	62,67 %	62,41 %	59,46 %	57,61 %	54,61 %
2	2	64,30 %	62,83 %	62,18 %	59,25 %	57,66 %	54,77 %
3	2	64,26 %	63,03 %	62,00 %	58,69 %	57,68 %	54,65 %

Rezultati predviđanja na temelju pojedinačnih značajki jasno pokazuju kako se bolji rezultati u slučaju istog učinka domaće i gostujuće momčadi dobivaju davanjem prednosti domaćoj momčadi. Važan dio istraživanja svakako je i ispitivanje mogućnosti cijelog skupa izlučenih značajki. Tablica 5.16 prikazuje rezultate predviđanja korištenjem cijelog skupa izlučenih značajki.

Tablica 5.16. Rezultati predviđanja korištenjem cijelog skupa izlučenih značajki.

Br. sezona skupa za učenje	Br. sezona skupa za ispit.	Bez prednosti domaćeg terena	Davanje prednosti domaćeg terena
1	1	63,65 %	64,88 %
1	2	63,92 %	64,67 %
2	1	63,90 %	64,79 %
2	2	64,03 %	64,62 %
3	2	63,94 %	64,46 %

Rezultati korištenja cijelog skupa izlučenih značajki bolji su u odnosu na rezultate pojedinačnih značajki. Bolje rezultate u oba slučaja dobiva se korištenjem prednosti domaćeg terena. Predviđanje korištenjem prednosti domaćeg terena na temelju cijelog skupa ulaznih značajki nije toliko značajno kao kod predviđanja korištenjem pojedinačnih značajki. Razlog tome je veća domena mogućih vrijednosti.

5.5. Predviđanje na temelju indeksa korisnosti i izlučenih značajki

Prethodna potpoglavlja su pokazala kako predviđanja na temelju indeksa korisnosti i optimalnog vremenskog prozoru daju vrlo slične rezultate kao i predviđanje na temelju skupa izlučenih značajki. Pretpostavka je da će kombinacija prethodno navedenih rezultata dovesti do još boljih rezultata predviđanja pa su ispitane mogućnosti predviđanja kombinacijom CTE indeksa korisnosti i skupa izlučenih značajki. Referentna točka će biti CTE indeks korisnosti s uključenom značajkom prednosti domaćeg terena, kojem će se dodavati pojedinačne značajke iz skupa izlučenih značajki. Opća formula predviđanja korištenjem indeksa CTE i izlučenih značajki, gdje je n_{zn} broj izlučenih značajki, a $n_{\Delta t}$ broj utakmica optimalnog vremenskog prozora, prikazana je formulom (5-13).

$$winn(I_{CTE}(tm_d(dt)), I_{CTE}(tm_g)) \begin{cases} tm_d, \dfrac{\frac{1}{n_{\Delta t}}\sum_{i=1}^{n_{\Delta t}} I_{CTE}(tm_d(dt), i) + \sum_{j=1}^{n_{zn}} N_{zn}(tm_d, j)}{\frac{1}{n_{\Delta t}}\sum_{i=1}^{n_{\Delta t}} I_{CTE}(tm_{dg}, i) + \sum_{j=1}^{n_{zn}} N_{zn}(tm_g, j)} \geq 1; \\ tm_g, \dfrac{\frac{1}{n_{\Delta t}}\sum_{i=1}^{n_{\Delta t}} I_{CTE}(tm_d(dt), i) + \sum_{j=1}^{n_{zn}} N_{zn}(tm_d, j)}{\frac{1}{n_{\Delta t}}\sum_{i=1}^{n_{\Delta t}} I_{CTE}(tm_g, i) + \sum_{j=1}^{n_{zn}} N_{zn}(tm_g, j)} < 1; \end{cases} \quad (5\text{-}13)$$

Od ostalih parametara valja napomenuti da će raspon utakmica na temelju kojih će se računati optimalni vremenski prozor biti $[60,\ 120]$, dok će parametar b poprimiti vrijednost koja je dala najbolje rezultate predviđanja ($b = 4$). Tablica 5.17 prikazuje rezultate predviđanja zasnovane na CTE indeksu korisnosti, optimalnom vremenskom prozoru i pojedinačnim izlučenim značajkama.

Tablica 5.17. Rezultati predviđanja zasnovani na indeksu korisnosti, optimalnom vremenskom prozoru i pojedinačnoj izlučenoj značajci.

Br. sezona skupa za učenje	Br. sezona skupa za ispit.	Izlučena značajka					
		om_I	om_{10}	om_{10dg}	om_{med}	om_{pNiz}	br_{10}
1	1	66,14 %	65,59 %	65,33 %	64,91 %	64,79 %	64,41 %
1	2	65,44 %	65,44 %	65,26 %	64,46 %	64,78 %	64,20 %
2	1	66,34 %	65,48 %	65,14 %	64,34 %	64,72 %	64,31 %
2	2	65,48 %	65,34 %	65,16 %	64,15 %	64,71 %	64,23 %
3	2	65,16 %	65,04 %	64,88 %	63,68 %	64,63 %	63,98 %

Dodavanje pojedinačnih značajki daje bolje rezultate u odnosu na indeks korisnosti i optimalni vremenski prozor, osim u kombinaciji sa značajkom *Broj utakmica u zadnjih 10 dana*. Kombinacija sa značajkom *Pobjednički niz* daje slične, ali nešto slabije rezultate. Najbolji rezultati su dobiveni korištenjem jedne sezone skupa za učenje i jedne sezona skupa za ispitivanje. Točnost u pravilu opada povećanjem ulaznog skupa podataka. U ovom slučaju, kao i u slučaju korištenja indeksa CTE u kombinaciji s optimalnim vremenskim prozorom, različite duljine skupa za učenje i skupa za ispitivanje daju slične rezultate, što nije bio slučaj korištenjem cijele povijesti.

U nastavku će se analizirati učinak dodavanja podskupa od dvije izlučene značajke. Bolji rezultati u odnosu na referentnu točnost dobiveni su korištenjem četiri pojedinačne izlučene značajke, što znači da je potrebno analizirati učinak $\binom{4}{2} = 6$ kombinacija podskupa od dvije izlučene značajke. Formulom (5-13) je prikazan način predviđanja ishoda ovisan o broju izlučenih značajki. Tablica 5.18 prikazuje rezultate predviđanja zasnovane na indeksu korisnosti, optimalnom vremenskom prozoru i podskupu od dvije izlučene značajke.

Tablica 5.18. Rezultati predviđanja zasnovani na indeksu korisnosti, optimalnom vremenskom prozoru i podskupu od dvije izlučene značajke.

Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	**Podskup izlučenih značajki**		
		om_I, om_{10}	om_I, om_{10dg}	om_I, om_{med}
1	1	66,27 %	66,42 %	65,91 %
1	2	65,69 %	65,87 %	65,40 %
2	1	66,29 %	66,42 %	66,14 %
2	2	65,60 %	65,82 %	65,48 %
3	2	65,23 %	65,48 %	65,04 %
Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	**Podskup izlučenih značajki**		
		om_{10}, om_{10dg}	om_{10}, om_{med}	om_{10dg}, om_{med}
1	1	65,80 %	65,54 %	65,54 %
1	2	65,77 %	65,45 %	65,43 %
2	1	65,62 %	65,47 %	65,47 %
2	2	65,58 %	65,08 %	65,27 %
3	2	65,46 %	64,76 %	64,84 %

Rezultati korištenja podskupa dvije izlučene značajke u pojedinim kombinacijama daju bolje rezultate u odnosu na pojedinačne značajke. U daljnje razmatranje uzet će se podskup tri izlučene značajke, a razmatrat će se učinak podskupa tri izlučene značajke koje su u podskupu dvije značajke dale bolje rezultate u odnosu na korištenje pojedinačne značajke. Konkretno, radi se o slučajevima u kojima su korištene kombinacije značajki *Omjer faze ispitivanja, Omjer u zadnjih 10 utakmica* te značajke *Omjer domaćina u zadnjih 10 domaćih utakmica* i *Omjer gosta u zadnjih 10 gostujućih utakmica* koje dolaze u paru. Ukupan broj kombinacija koje nudi podskup tri značajke u kojima se biraju tri značajke iznosi $\binom{3}{3} = 1$. Formulom (5-13) prikazan je način predviđanja ishoda ovisan o broju izlučenih značajki. Tablica 5.19 prikazuje rezultate predviđanja zasnovane na indeksu korisnosti, optimalnom vremenskom prozoru i podskupu tri izlučene značajke.

Tablica 5.19. Rezultati predviđanja zasnovani na indeksu korisnosti, optimalnom vremenskom prozoru i podskupu od tri izlučene značajke.

Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	**Podskup izlučenih značajki**
		$om_I + om_{10} + om_{10dg}$
1	1	66,34 %
1	2	65,87 %
2	1	66,40 %
2	2	65,79 %
3	2	65,52 %

Rezultati korištenja podskupa tri izlučene značajke dali su lošije rezultate u odnosu na kombinacije korištenja podskupa dvije značajke. Najbolji rezultat od 66,42 % tako je dobiven korištenjem dvije izlučene značajke, *Omjer faze ispitivanja* te značajki *Omjer domaćina u zadnjih 10 domaćih utakmica* i *Omjer gosta u zadnjih 10 gostujućih utakmica* koje dolaze u paru.

5.6. Događaji povećane neizvjesnosti

Pojam događaja povećane neizvjesnosti je uveden u potpoglavlju 4.4. Događajem povećane neizvjesnosti se smatra događaj u kojem je razlika projiciranih indeksa korisnosti dva suprotstavljena procesa unutar unaprijed definiranog raspona, tzv. raspona neizvjesnosti ($r \in [r_{min}, r_{max}]$).

U konkretnom primjeru, procesi su predstavljeni suprotstavljenim momčadima košarkaške utakmice u kojem proces A predstavlja domaću momčad, a proces B gostujuću momčad. Formula (5-14) prikazuje prilagodbu formule izračuna postotne razlike ($pr_{\%}$) za problem predviđanja ishoda košarkaške utakmice temeljen na projiciranom indeksu CTE. Indeks CTE se računa na temelju optimalnog vremenskog prozora izračunatog u intervalu $[60, 120]$, a koristi se parametar prednosti domaćeg terena, parametar $b = 4$ te podskup izlučenih značajki.

$$pr_{\%} = \frac{\left|I_{CPE}(tm_d(dt)) - I_{CPE}(tm_g)\right|}{I_{CPE}(tm_d(dt))} \times 100 \qquad (5\text{-}14)$$

Prosječno najbolje rezultate predviđanja u pravilu daje upotreba jedne sezone skupa za učenje i jedne sezone skupa za ispitivanje pa će se shodno tome koristiti navedeni ulazni skup podataka s ciljem pronalaženja optimalne maksimalne vrijednosti (r_{max}) raspona identificiranja utakmica povećane neizvjesnosti. Tablica 5.20 prikazuje utjecaj raspona utakmica povećane neizvjesnosti na točnost predviđanja na temelju jedne sezone skupa za učenje i jedne sezone skupa za ispitivanje. Cilj ispitivanja je odrediti optimalni raspon identificiranja utakmica povećane neizvjesnosti, ali i analizirati točnost predviđanja utakmica kojima nije potrebna dodatna analiza u odnosu na utakmice povećane neizvjesnosti kojima je potrebna dodatna analiza.

Rezultati jasno sugeriraju kako se povećanjem raspona događaja povećane neizvjesnosti udio utakmica povećane neizvjesnosti povećava. Ujedno rastu i točnosti predviđanja, a udio utakmica kojima nije potrebna dodatna analiza opada. Svaka promjena raspona, točnije promjena gornje granice identificiranja utakmice povećane neizvjesnosti, doprinijela je povećanju točnosti predviđanja, osim u slučaju u kojem je gornja granica identificiranja utakmica povećane neizvjesnosti u intervalu $\langle 0, 10]$. U tom je slučaju 28,31 % utakmica identificirano utakmicama povećane neizvjesnosti.

Tablica 5.20. Utjecaj raspona utakmica povećane neizvjesnosti na točnost predviđanja.

Raspon utakmica povećane neizvjesnosti (%)	Predviđanje na temelju indeksa CTE		Razlika	Utakmice povećane neizvjesnosti		Razlika	Točnost modela
	Udio	Točnost		Udio	Točnost		
$\langle 0, 1]$	97,02 %	66,96 %	–	2,98 %	49,02 %	–	66,42 %
$\langle 0, 2]$	94,17 %	67,42 %	+0,46 %	5,83 %	50,33 %	+1,31 %	
$\langle 0, 3]$	91,70 %	67,88 %	+0,46 %	8,30 %	50,35 %	+0,02 %	
$\langle 0, 4]$	88,78 %	68,22 %	+0,34 %	11,22 %	52,17 %	+1,82 %	
$\langle 0, 5]$	85,81 %	68,71 %	+0,49 %	14,19 %	52,57 %	+0,40 %	
$\langle 0, 6]$	82,95 %	69,07 %	+0,36 %	17,05 %	53,54 %	+0,97 %	
$\langle 0, 7]$	80,12 %	69,56 %	+0,49 %	19,88 %	53,80 %	+0,26 %	
$\langle 0, 8]$	77,41 %	70,00 %	+0,44 %	22,59 %	54,16 %	+0,36 %	
$\langle 0, 9]$	74,77 %	70,26 %	+0,26 %	25,23 %	55,06 %	+0,90 %	
$\langle 0, 10]$	71,63 %	70,94 %	+0,68 %	28,37 %	55,01 %	-0,04 %	
$\langle 0, 11]$	68,94 %	71,53 %	+0,59 %	31,03 %	55,10 %	+0,09 %	
$\langle 0, 12]$	66,08 %	71,98 %	+0,45 %	33,92 %	55,60 %	+0,40 %	
$\langle 0, 13]$	63,62 %	72,35 %	+0,37 %	36,38 %	56,06 %	+0,46 %	
$\langle 0, 14]$	60,85 %	72,90 %	+0,55 %	39,15 %	56,36 %	+0,30 %	
$\langle 0, 15]$	58,16 %	73,19 %	+0,39 %	41,84 %	57,02 %	+0,66 %	
$\langle 0, 16]$	55,69 %	73,34 %	+0,15 %	44,31 %	57,73 %	+0,71 %	
$\langle 0, 17]$	53,23 %	73,67 %	+0,33 %	46,77 %	58,18 %	+0,35 %	
$\langle 0, 18]$	51,00 %	74,31 %	+0,64 %	49,00 %	58,21 %	+0,03 %	
$\langle 0, 19]$	48,85 %	74,62 %	+0,31 %	51,15 %	58,60 %	+0,39 %	
$\langle 0, 20]$	46,67 %	75,08 %	+0,46 %	53,33 %	58,85 %	+0,25 %	
	Apsolutna razlika		+8,12 %	**Apsolutna razlika**		+9,83 %	

5.6.1. Predviđanje utakmica povećane neizvjesnosti

Predviđanje utakmica povećane neizvjesnosti će, kao i za predviđanje utakmica kojima nije potrebna analiza, koristiti projicirani CTE indeks korisnosti, optimalni vremenski prozor te podskup izlučenih značajki. Kao točka identificiranja utakmice povećane neizvjesnosti će se koristiti interval $\langle 0, 10]$. Na rezultate dobivene indeksom korisnosti će se redom najprije dodavati pojedinačne značajke. Tablica 5.21 prikazuje rezultate predviđanja ishoda s uključenom identifikacijom utakmica povećane neizvjesnosti korištenjem projiciranog CTE indeksa i pojedinačne izlučene značajke. Pretpostavka je da rezultati korištenjem pojedinačne izlučene značajke neće znatno odstupati od rezultata dobivenih bez postupka identificiranje utakmica povećane neizvjesnosti.

Tablica 5.21. Rezultati predviđanja korištenjem pojedinačne izlučene značajke i utakmica povećane neizvjesnosti.

Br. sezona skupa za učenje	Br. sezona skupa za ispit.	Izlučena značajka					
		om_I	om_{10}	om_{10dg}	om_{med}	om_{pNiz}	br_{10}
1	1	66,05 %	66,03 %	67,49 %	66,18 %	65,60 %	65,74 %
1	2	65,43 %	65,86 %	66,84 %	67,50 %	65,70 %	65,65 %
2	1	66,15 %	66,04 %	67,49 %	66,75 %	65,86 %	65,91 %
2	2	65,42 %	65,76 %	66,71 %	67,44 %	65,72 %	65,60 %
3	2	65,19 %	65,42 %	66,44 %	66,63 %	65,42 %	65,29 %

Iz rezultata je jasno vidljivo kako je podskup značajki kojima će se predviđati utakmice povećane neizvjesnosti potrebno dodatno proširiti korištenjem većeg podskupa izlučenih značajki. Skup izlučenih značajki sadrži sedam značajki pri čemu dvije značajke, *Omjer domaćina u zadnjih 10 domaćih utakmica* i *Omjer gosta u zadnjih 10 gostujućih utakmica*, čine cjelinu. Samim time

potrebno je analizirati učinak $\binom{6}{2} = 15$ kombinacija odabira izlučenih značajki. Obećavajuće rezultate daju sve pojedinačne izlučene značajke osim značajke *Pobjednički niz* i *Broj utakmica u zadnjih 10 dana*. Tablica 5.22 prikazuje rezultate predviđanja ishoda s uključenom identifikacijom utakmica povećane neizvjesnosti korištenjem projiciranog CTE indeksa i podskupa dvije izlučene značajke.

Tablica 5.22. Rezultati predviđanja korištenjem podskupa dvije izlučene značajke i utakmica povećane neizvjesnosti.

Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	**Podskup dvije izlučene značajke**			
		$\boldsymbol{om_I}$ $\boldsymbol{om_{10}}$	$\boldsymbol{om_I}$ $\boldsymbol{om_{10dg}}$	$\boldsymbol{om_I}$ $\boldsymbol{om_{med}}$	$\boldsymbol{om_I}$ $\boldsymbol{om_{pNiz}}$
1	1	66,10 %	67,86 %	67,40 %	65,95 %
1	2	65,51 %	66,82 %	66,36 %	65,54 %
2	1	66,10 %	67,85 %	67,66 %	65,76 %
2	2	65,43 %	66,74 %	66,46 %	65,31 %
3	2	65,10 %	66,45 %	66,08 %	65,04 %
Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	**Podskup dvije izlučene značajke**			
		$\boldsymbol{om_I}$ $\boldsymbol{br10}$	$\boldsymbol{om_{10}}$ $\boldsymbol{om_{10dg}}$	$\boldsymbol{om_{10}}$ $\boldsymbol{om_{med}}$	$\boldsymbol{om_{10}}$ $\boldsymbol{om_{pNiz}}$
1	1	65,99 %	67,52 %	67,55 %	65,87 %
1	2	65,40 %	66,90 %	66,87 %	65,86 %
2	1	66,03 %	67,68 %	67,53 %	65,94 %
2	2	65,26 %	66,90 %	66,66 %	65,81 %
3	2	65,06 %	66,53 %	66,43 %	65,51 %
Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	**Podskup dvije izlučene značajke**			
		$\boldsymbol{om_{10}}$ $\boldsymbol{br_{10}}$	$\boldsymbol{om_{10dg}}$ $\boldsymbol{om_{med}}$	$\boldsymbol{om_{10dg}}$ $\boldsymbol{om_{pNiz}}$	$\boldsymbol{om_{10dg}}$ $\boldsymbol{br_{10}}$
1	1	65,96 %	68,75 %	67,27 %	67,46 %
1	2	65,74 %	67,68 %	66,73 %	66,77 %
2	1	66,05 %	68,83 %	67,49 %	67,71 %
2	2	65,73 %	67,67 %	66,73 %	66,81 %
3	2	65,35 %	67,36 %	66,38 %	66,47 %
Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	**Podskup dvije izlučene značajke**			
		$\boldsymbol{om_{med}}$ $\boldsymbol{om_{pNiz}}$	$\boldsymbol{om_{med}}$ $\boldsymbol{br_{10}}$	$\boldsymbol{om_{pNiz}}$ $\boldsymbol{br_{10}}$	
1	1	67,23 %	67,45 %	65,73 %	
1	2	66,61 %	66,72 %	65,75 %	
2	1	67,33 %	67,38 %	65,85 %	
2	2	66,52 %	66,58 %	65,68 %	
3	2	66,10 %	66,18 %	65,36 %	

Iz rezultata je jasno vidljivo kako korištenje podskupa dvije izlučene značajke daje bolje rezultate u odnosu na korištenje pojedinačnih značajki, ali i napredak u odnosu na predviđanje bez identifikacije utakmice povećane neizvjesnosti. Za gotove sve podskupove od dvije izlučene značajke je vidljiv napredak, međutim kombinacije s četiri izlučene značajke posebno odskaču. To su značajka *Omjer tijekom faze ispitivanje*, značajki *Omjer domaćina u zadnjih 10 domaćih utakmica* i *Omjer gosta u zadnjih 10 gostujućih utakmica* koje dolaze u paru, značajke *Omjer zadnjih 10 utakmica* i *Međusobni omjer*. U nastavku će se ispitati mogućnost predviđanja korištenjem podskupa tri izlučene značajke. Skup izlučenih značajki tako će se sastojati od četiri

značajki pri čemu će tri značajke činit podskup. Samim time potrebno je analizirati učinak $\binom{4}{3} = 4$ kombinacija. Tablica 5.23 prikazuje rezultate predviđanja ishoda s uključenom identifikacijom utakmica povećane neizvjesnosti korištenjem projiciranog CTE indeksa i podskupa tri izlučene značajke.

Tablica 5.23. Rezultati predviđanja korištenjem podskupa tri izlučene značajke i utakmica povećane neizvjesnosti.

Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	**Podskup tri izlučene značajke**			
		om_l om_{10} om_{10dg}	om_l om_{10} om_{med}	om_l om_{10dg} om_{med}	om_{10} om_{10dg} om_{med}
1	1	67,17 %	67,61 %	68,81 %	68,94 %
1	2	66,54 %	66,51 %	67,49 %	67,75 %
2	1	67,16 %	67,41 %	68,97 %	68,93 %
2	2	66,44 %	66,36 %	67,43 %	67,73 %
3	2	66,17 %	65,92 %	67,17 %	67,33 %

Rezultati korištenjem podskupa tri izlučene značajke su u pravilu bolji od kombinacije korištenja podskupa dvije izlučene značajke. Najbolji rezultat dobiven je korištenjem podskupa tri izlučene značajke, *Omjer zadnjih 10 utakmica*, značajki *Omjer domaćina u zadnjih 10 domaćih utakmica* i *Omjer gosta u zadnjih 10 gostujućih utakmica* koje dolaze u paru i značajke *Međusobni omjer*. Posljednje što valja provjeriti su rezultati korištenja podskupa sve četiri izlučene značajke. Tablica 5.24 prikazuje rezultate predviđanja ishoda s identifikacijom utakmica povećane neizvjesnosti korištenjem projiciranog CTE indeksa i podskupa četiri izlučene značajke.

Tablica 5.24. Rezultati predviđanja korištenjem podskupa četiri izlučene značajke i utakmica povećane neizvjesnosti.

Br. sezona skupa za učenje	**Br. sezona skupa za ispit.**	**Podskup četiri izlučene značajke**
		$om_l, om_{10}, om_{10dg}, om_{med}$
1	1	68,89 %
1	2	67,55 %
2	1	68,69 %
2	2	67,40 %
3	2	66,89 %

Rezultati korištenja četiri izlučene značajke su u prosjeku nešto slabiji od najbolje kombinacije korištenja tri izlučene značajke. Generalno najbolji rezultat dobiven je korištenjem podskupa tri izlučene značajke, točnije korištenjem podskupa značajki *Omjer zadnjih 10 utakmica* , *Omjer domaćina u zadnjih 10 domaćih utakmica* i *Omjer gosta u zadnjih 10 gostujućih utakmica* koje dolaze u paru i značajke *Međusobni omjer*.

5.7. Analiza rezultata

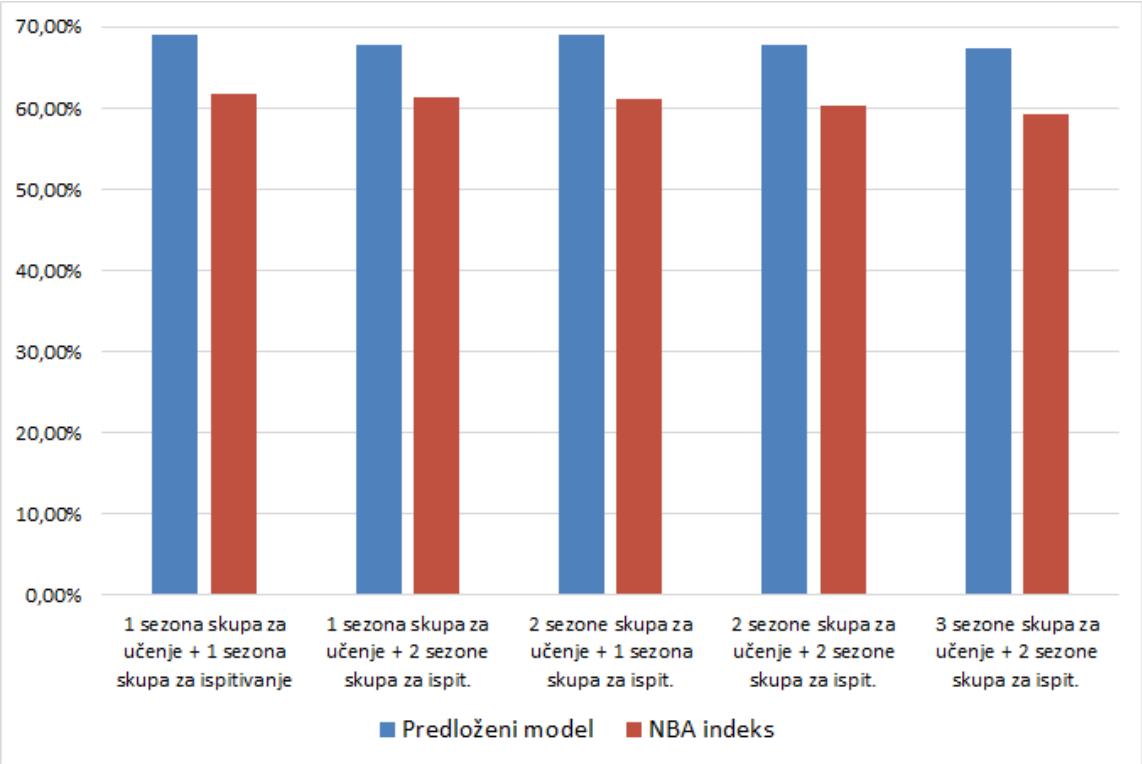
U ovom potpoglavlju će se prikazati rezultati konačnog modela predviđanja ishoda košarkaških utakmica zasnovanog na sveobuhvatnom indeksu korisnosti i optimalnom vremenskom prozoru. Osim prosječnih rezultata kombinacija korištenja skupa za učenje i skupa za ispitivanje, prikazani

će biti i rezultati pojedinačnih parova skupa za učenje i skupa za ispitivanje. Tablica 5.25 prikazuje konačne rezultate predviđanja ishoda.

Tablica 5.25. Konačni rezultati predloženog modela predviđanja ishoda košarkaških utakmica.

Ulazni skup podataka		**Točnost**	**Prosjek**
Skup za učenje	**Skup za ispitivanje**		
2016./2017.	2017./2018.	69,21 %	68,94 %
2015./2016.	2016./2017.	67,76 %	
2014./2015.	2015./2016.	70,67 %	
2013./2014.	2014./2015.	69,41 %	
2012./2013.	2013./2014.	66,49 %	
2011./2012.	2012./2013.	70,40 %	
2010./2011.	2011./2012.	69,65 %	
2009./2010.	2010./2011.	68,04 %	
2015./2016.	2016./2017. – 2017./2018.	66,84 %	67,75 %
2014./2015.	2015./2016. – 2016./2017.	68,61 %	
2013./2014.	2014./2015. – 2015./2016.	68,67 %	
2012./2013.	2013./2014. – 2014./2015.	66,31 %	
2011./2012.	2012./2013. – 2013./2014.	67,00 %	
2010./2011.	2011./2012. – 2012./2013.	69,14 %	
2009./2010.	2010./2011. – 2011./2012.	67,84 %	
2015./2016. – 2016./2017.	2017./2018.	68,45 %	68,93 %
014./2015. – 2015./2016.	2016./2017.	68,14 %	
2013./2014. – 2014./2015.	2015./2016.	70,59 %	
2012./2013. – 2013./2014.	2014./2015.	69,03 %	
2011./2012. – 2012./2013.	2013./2014.	66,34 %	
2010./2011. – 2011./2012.	2012./2013.	70,47 %	
2009./2010. – 2010./2011.	2011./2012.	69,65 %	
2014./2015. – 2015./2016.	2016./2017. – 2017./2018.	66,81 %	67,73 %
2013./2014. – 2014./2015.	2015./2016. – 2016./2017.	68,61 %	
2012./2013. – 2013./2014.	2014./2015. – 2015./2016.	68,56 %	
2011./2012. – 2012./2013.	2013./2014. – 2014./2015.	66,31 %	
2010./2011. – 2011./2012.	2012./2013. – 2013./2014.	67,00 %	
2009./2010. – 2010./2011.	2011./2012. – 2012./2013.	69,26 %	
2013./2014. – 2015./2016.	2016./2017. – 2017./2018.	66,58 %	67,33 %
2012./2013. – 2014./2015.	2015./2016. – 2016./2017.	68,91 %	
2011./2012. – 2013./2014.	2014./2015. – 2015./2016.	68,33 %	
2010./2011. – 2012./2013.	2013./2014. – 2014./2015.	66,01 %	
2009./2010. – 2011./2012.	2012./2013. – 2013./2014.	66,81 %	

Najbolji prosječni rezultat od 68,94 % dobiven je korištenjem jedne sezone skupa za učenje i jedne sezone skupa za ispitivanje. Također, gotovo identičan rezultat dobiven je korištenjem obje kombinacije jedne sezone skupa za ispitivanje. Gotovo identični rezultati vezani uz isti broj sezona skupa za ispitivanje su logični s obzirom da se optimalni vremenski prozor računa u intervalu $[60, 120]$ utakmica skupa za učenje, što je brojka kojom skup za učenje najčešće obuhvaća jednu sezonu unatrag. Najbolji pojedinačni rezultat je također dobiven jednom sezonom skupa za učenje i jednom sezonom skupa za ispitivanje u kombinaciji kada je natjecateljska sezona 2014./2015. korištena za učenje, a sezona 2016./2017. za ispitivanje. Slika 5.7 prikazuje konačne rezultate predloženog modela predviđanja u odnosu na početne rezultate dobivene korištenjem NBA indeksa.



Slika 5.7. Usporedba konačnih rezultata predloženog modela u odnosu na NBA indeks.

6. RASPRAVA

Predložena metoda upotrebe indeksa korisnosti, klasifikacija sportskih događaja prema izvjesnosti uspješnog predviđanja, upotreba optimalnog vremenskog prozora i postupak optimizacije parametara indeksa korisnosti svakako predstavljaju novinu u odnosu na uobičajene metode u području predviđanja ishoda sportskih događaja. Navedena tvrdnja se posebno odnosi na upotrebu indeksa korisnosti kao polazne točke cjelokupnog istraživanja i mjere uspješnosti analiziranog procesa, izračuna i prilagodbe optimalnog vremenskog prozora kojim se određuje relevantnost statističkih podataka o prethodnim događajima te određivanje specifičnih kategorija sportskih događaja kako bi se primjenom prilagodljive metode predviđanja dobili optimalni rezultati.

Najveći problem predviđanja ishoda u sportu svakako predstavlja vrlo teška ili gotovo nemoguća usporedba rezultata istraživanja drugih istraživača koji u većini slučajeva koriste različite skupove podataka i lige različitih konkurentnosti. Drugi problem svakako predstavlja vrlo visoka točnost pojedinih rezultata istraživanja potkrijepljena otežanom interpretacijom, lošom metodologijom istraživanja ili nedostatkom jasnoće.

Velik problem predviđanja ishoda u sportu je i definiranje univerzalnog modela predviđanja zbog čega se istraživači posvećuju pretežno jednom sportu i njegovim posebnostima te posljedično i ispitivanju metoda u okviru tog jednog sporta. Prednost predložene metode predviđanja u odnosu na metode korištene od strane ostalih istraživača je prilagodljivost ostalim, prvenstveno momčadskim sportovima. Predložene metode ostalih istraživača u pravilu su orijentirane ka rješavanju problema predviđanja vezanog uz određeni sport. Tako fokus na određeni sport, ili ako se problemu pristupa općenito, na određeni proces, može uključivati i empirijsko znanje vezano uz analizirani problem kojim se u pravilu može dodatno poboljšati rezultate predviđanja. Također, istraživanja ostalih istraživača uglavnom koriste manje skupove podataka čime je prilagodba modela ka specifičnom problemu znatno olakšana. I posljednji, možda i najvažniji razlog zbog kojeg su rezultati trenutnog istraživanja u prosjeku nešto lošiji od najboljih rezultata vezanih uz analizirani sport je prezentacija najboljih pojedinačnih rezultata. Pod pojmom najboljih pojedinačnih rezultata se smatra najbolji rezultat dobiven jednom od analiziranih kombinacija korištenja ulaznih skupova podataka. Najbolji pojedinačni rezultat tako može predstavljati ekstremnu vrijednost, točnije vrijednost koja znatno odskače od ostalih rezultata. Relevantnijim pokazateljem svakako treba smatrati srednju vrijednost. Bitnu ulogu u predviđanju ishoda sportskih događaja ima i odabir metode validacije. Pošto sportski događaji nisu u potpunosti neovisni događaji, korištenje unakrsne provjere u njenom općem obliku nije moguća iz razloga što

unakrsna provjera u pojedinim fazama koristi i buduće podatke čime je moguće uočiti trendove koji u trenutku predviđanja još nisu poznati. Najprimjerenija metoda validacije je metoda podjele skupa podataka u kojoj skup za učenje i skup za ispitivanje trebaju biti kronološki poredani. Postojanje skupa za provjeru nije nužno, što ne znači da korištenje skupa za provjeru neće dovesti do boljih rezultata predviđanja. U ovom istraživanju skup za provjeru nije korišten. Relevantnost statističkih podataka o prethodnim događajima definirana je računanjem početnog optimalnog vremenskog prozora te prilagodbom optimalnih vremenskih prozora u kasnijim iteracijama algoritma predviđanja.

Cilj istraživanja svakako je bio predložiti prilagodljivu metodu predviđanja ishoda sportskih događaja zasnovanu na indeksu korisnosti i optimalnom vremenskom prozoru, na način da je predloženu metodu moguće lako prilagoditi i drugim, prvenstveno momčadskim sportovima, ali i brojnim drugim procesima koje je moguće podijeliti na komponente. Predložena metoda se može koristiti i u analitičke svrhe. Način korištenja predloženog indeksa korisnosti i optimalnog vremenskog prozora ovisi o načinu definiranja analiziranog vremenskog perioda. Prvi način definira korištenje prosječnih vrijednosti koje ne uključuje trenutno promatrani proces. Drugi način definiranja korištenja predloženog indeksa korisnosti uključuje i promatrani proces. Za predviđanje ishoda potrebno je koristiti prvi način, točnije način koji ne uključuje trenutni proces, dok je u analitičke svrhe potrebno koristiti drugi način koji uključuje i promatrani proces. Valja napomenuti kako upotreba indeksa korisnosti i vremenskog prozora može analizirati proces na temelju izračunatog optimalnog vremenskog prozora te na taj način dati ocjenu samog procesa.

Predloženom metodom predviđanja, ali i analizom radova drugih istraživača uočeno je kako vrlo važnu ulogu kod predviđanja sportskih ishoda ima korištenje skupa izlučenih značajki vezanih uz uspješnost momčadi. Rezultati istraživanja su pokazali kako kombinacija korištenja osnovne statistike analiziranog sporta i skupa izlučenih značajki vezanih uz uspješnost momčadi mogu doprinijeti boljim rezultatima predviđanja. U radu je eksperimentalno ispitan velik broj značajki vezanih uz uspješnost momčadi. Valja napomenuti kako je sve izlučene značajke moguće koristiti u gotovo svim momčadskim, ali i većini individualnih sportova. Izlučene značajke su isključivo vezane uz uspješnost momčadi te ne uključuju osnovne statističke elemente analiziranog sporta. Izlučene značajke su se pokazale posebno pogodnim za predviđanje specifičnih kategorija sportskih događaja.

7. ZAKLJUČAK

U ovom radu je predložen model predviđanja sportskih ishoda zasnovan na indeksu korisnosti i optimalnom vremenskom prozoru. Govoreći općenito o predviđanju ishoda u sportu, valja napomenuti kako se radi o izazovnom problemu. Glavni cilj istraživanja bio je predložiti metodu predviđanja ishoda, ne samo u košarci koja je poslužila isključivo kao skup za ispitivanje, već predloženu metodu učiniti lako prilagodljivom i ostalim, prvenstveno momčadskim sportovima, ali i procesima koju se mogu podijeliti na komponente. Predviđanje ishoda u sportu specifičan je problem u kojem postoje dvije suprotstavljene ekipe opisane istim skupom značajki, čime je predviđanje dodatno otežano. Rezultati istraživanja ukazali su i na dodatne probleme vezene uz predviđanje ishoda u sportu.

U početnoj fazi istraživanja proučen je niz tema vezanih uz analizirano i srodna područja s ciljem stjecanja potrebnog znanja i uvida u samu problematiku. Samim time, analizirana je dostupna literatura vezana uz područje strojnog učenja, točnije klasifikacijske metode nadziranog strojnog učenja. Proučena je i literatura vezana uz predviđanje ishoda u sportu, metode evaluacije učinka igrača ili momčadi, ali i dostupna literatura koja se bavi upotrebom informacijskih sustava vezanih uz podršku u donošenju odluka vezanih uz sport.

Predloženi model je evaluiran nad podacima vezanim uz NBA ligu. Razvoj novih metoda istraživanja može se podijeliti u nekoliko faza. U prvoj fazi je predložen sveobuhvatni indeks korisnosti kojim se može, ovisno o konkretnom problemu, evaluirati učinak igrača ili momčadi, a koji će u kasnijim fazama služiti kao početna točka predviđanja ishoda. Svrha sveobuhvatnog indeksa korisnosti je evaluirati učinak procesa koristeći skup specifičnih značajki vezanih uz analizirani proces. Predložen je i postupak optimiranja parametara sveobuhvatnog indeksa korisnosti korištenjem kombinacije linearnih i nelinearnih doprinosa. Predloženim postupkom optimiranja parametara sveobuhvatnog indeksa korisnosti rezultati predviđanja ishoda dodatno su poboljšani. Predložen je i algoritam izračuna i prilagodbe optimalnog vremenskog prozora. Svrha optimalnog vremenskog prozora ograničiti je pogled u prošlost te samim time odrediti relevantnost statističkih podataka o prethodnim događajima. Uvođenjem optimalnog vremenskog prozora rezultati predviđanja su dodatno poboljšani. U svrhu poboljšanja uspješnosti predviđanja ishoda sportski događaji su podijeljeni u različite kategorije događaja čime je omogućena primjena prilagodljivog postupka. Na temelju poznatih podataka potrebno je klasificirati sportske događaje prema složenosti predviđanja te sukladno tome pristupiti predviđanju ishoda analiziranog događaja. Tako je uveden pojam utakmica povećane neizvjesnosti te je sukladno tome predložen

algoritam predviđanja takvih utakmica. Predloženi model je naposljetku eksperimentalno evaluiran, a rezultati su analizirani.

Konkretan model koristi dva skupa ulaznih značajki. Prvi skup značajki tiče se osnovne statistike vezane uz analizirani sport, u ovom slučaju košarke, a drugi skup značajki čine značajke vezane uz uspješnost momčadi tijekom faze učenja. Iz prethodnih objašnjenja moguće je zaključiti kako eksperimentalna evaluacija implementiranih metoda uključuje pripremu ulaznog skupa podataka, korištenje predloženih metoda predviđanja, prilagodbu parametara predloženog indeksa korisnosti, primjenu postupka izlučivanja značajki u svrhu poboljšanja rezultata predviđanja, izračun i prilagodbu optimalnog vremenskog prozora, klasifikaciju događaja prema izvjesnosti uspješnog predviđanja te naposljetku analizu dobivenih rezultata.

Iako je disertacijom obuhvaćen i razriješen dio problematike vezane uz predviđanje ishoda u sportu, još uvijek postoji mnogo neodgovorenih pitanja vezanih uz ovo zanimljivo znanstveno područje. Najveći problem svakako predstavlja nemogućnost usporedbe rezultata jer istraživači koriste različite skupove podataka i lige različite konkurentnosti. Drugi problem vezan je uz nemogućnost predlaganja univerzalnog modela predviđanja, te je samim time nužan fokus na određeni sport. Stoga se umjesto predlaganja modela predlaže predlaganje metode predviđanja koja će se uz sitne preinake moći prilagoditi ostalim sportovima.

Predložena metoda upotrebe indeksa korisnosti, klasifikacija sportskih događaja prema izvjesnosti uspješnog predviđanja, upotreba optimalnog vremenskog prozora i postupak optimizacije parametara indeksa korisnosti svakako predstavljaju novinu u odnosu na uobičajeno korištene metode u području predviđanja ishoda sportskih događaja. Buduća istraživanja svakako uključuju uvođenje novih značajki koje će još podrobnije opisati sam proces, samim time povećati točnost modela te prilagoditi model predviđanja sukladno posebnostima, najprije momčadskih sportova, a zatim i ostalih sličnih procesa.

Literatura

[1] Sourav Das, „Top 10 Most Popular Sports in The World [Updated 2020]“, https://sportsshow.net/top-10-most-popular-sports-in-the-world/, pristup 15.7.2020.

[2] Benjamin Elisha Sawe, „The Most Popular Sports in the World“, https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html, pristup 15.7.2020.

[3] Michael Brown, „Biggest Global Sports“, http://www.biggestglobalsports.com/calculation-method/4581015470, pristup 15.7.2020.

[4] H. Ruiz, P. Power, X. Wei, P. Lucey, ““The Leicester City Fairytale?“: Utilizing New Soccer Analytics Tools to Compare Performance in the 15/16 & 16/17 EPL Seasons”, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’17, str. 1991-2000, 2017.

[5] S. Grossberg, „Nonlinear Neural network: principles, mechanisms and architectures“, Neural network, sv. 1, br 1, str. 17-617, 1988.

[6] G.P. Zhang, „Neural network for Classification: A Survey“, IEEE Transactions on Systems, Man and Cybernetics (Applications and Reviews), sv. 30, br. 4, str. 451-462, 2000.

[7] E. Ben-Naim, F. Vazquez, S. Redner, „What is the most Competitive Sport?“, Journal of the Korean Physics Society, sv. 50, br. 1, str. 124–126, siječanj 2007.

[8] R.Y.S. Aoki, R.M. Assuncao, P.O.S. Vaz de Melo, „Luck is Hard to Beat“, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’17, str. 1367-1376, 2017.

[9] T. Horvat, J. Job, „The Use of Machine Learning in Sport Outcome Prediction: A Review”, Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, lipanj 2020.

[10] T. Horvat, J. Job, „Importance of the training dataset length in basketball game outcome prediction by using naive classification machine learning methods”, Elektrotehniški vestnik - Journal of Electrical Engineering and Computer Science, sv. 86, br. 4, str. 197-202, 2019.

[11] M.C. Purucker, „Neural network Quarterbacking“, IEEE Potentials, sv. 15, br., str. 9 – 15, 1996.

[12] J. Kahn, „Neural Network Prediction of NFL Football Games“, IEEE Computer Society Washington, str. 1194-1197, siječanj 2003.

[13] Babak Hamadani, „Predicting the outcome of NFL games using machine learning“, http://cs229.stanford.edu/proj2006/BabakHamadani-PredictingNFLGames.pdf, pristup 6. srpnja 2020.

[14] A. McCabe, J. Travathan, „Artificial Intelligence in Sports Prediction“, Fifth International Conference on Information Technology: New Generations, str. 1994-1997, 2008.

[15] B. Loeffelholz, E. Bednar, K.W. Bauer, „Predicting NBA Games Using Neural Networks“, Journal of Quantitative Analysis in Sports, sv. 5, br. 1, 2009.

[16] D. Miljković, Lj. Gajić, A. Kovačević, Z. Konjović, „The use of data mining for basketball matches outcomes prediction“, IEEE 8th International Symposium on intelligent and informatics, str. 309-312, 2010.

[17] E. Zdravevski, A. Kulakov, „System for Prediction of the Winner in a Sports Game“, ICT Innovations 2009, str. 55-63, 2010.

[18] K. Trawinski, „A fuzzy classification system for prediction of the results of the basketball games“, International Conference on Fuzzy Systems, 2010.

[19] Z. Ivanković, M. Racković, B. Markoviski, „Analysis of basketball games using neural networks“, 11th International Symposium on Computational Intelligence and Informatics (CINTI), str. 251-256, 2010.

[20] D. Buursma, „Predicting sports events from past results: Towards effective betting on football matches“, 14th Twente Student Conference on IT, 2011.

[21] A.D. Blaikie, J.A. David, G.J. Abud, R.D. Pasteur, „NFL & NCAA Football Prediction using Artificial Neural network“, Proceedings of the 2011 Midstates Conference on Undergraduate Research in Computer Science and Mathematics, 2011.

[22] J. Hucaljuk, A. Rakipović, „Predicting football scores using machine learning techniques“, Proceedings of the 34th International Convention MIPRO, str. 23-27, 2011.

[23] Chenjie Cao, „Sports Data Mining Technology Used in Basketball Outcome Prediction“, Dublin Institute of Technology, Ireland, 2012.

[24] D. Delen, D. Cogdell, N. Kasap, N, „A comparative analysis of data mining methods in predicting NCAA bowl outcomes“, International Journal of Forecasting, sv. 28, br. 2, str. 543 – 552, 2012.

[25] Anže Kravanja, „Napovedanje zmagovalcev košarkaških tekem“, Fakultet za računalništvo in informatiko, Sveučilište u Ljubljani, 2013.

[26] Renato Amorim Torres, „Prediction of NBA games based on Machine Learning Methods“, Computer-Aided Engineering, Sveučilište Winconsin Madison, 2013.

[27] Albrecht Zimmermann, Sruthi Moorthy, Zifan Shi, „Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned“, https://arxiv.org/pdf/1310.3607.pdf, pristup 6.7.2020.

[28] F. Owramipur, P. Eskandarian, F. Mozneb Sadat, „Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team“, International Journal of Computer Theory and Engineering, sv. 5, br. 5, pp. 812-815, 2013.

[29] Jasper Lin, Logan Short, Vishnu Sundaresan, „Predicting National Basketball Association Winners“, završni projekt, 2014.

[30] C.P. Igiri, E.O. Nwachukwu, „An Improved Prediction System for Football a Match Result“, IOSR Journal of Engineering, sv. 4, br.12, str. 12-20, 2014.

[31] Stylianos Kampakis, William Thomas, „Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches“, https://arxiv.org/ftp/arxiv/papers/1511/1511.05837.pdf, pristup 6.7.2020.

[32] N. Tax, Y. Joustra, „Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach“, Transactions on Knowledge and Data Engineering, sv. 10, br.10, str. 1–13, 2015.

[33] Grant Avalon, Batuhan Balci, Jesus Guzman, „Various Machine Learning Approaches to Predicting NBA Score Margins“, http://cs229.stanford.edu/proj2016/report/Avalon_balci_guzman_various_ml_approaches_NBA_Scores_report.pdf, pristupljeno 6.7.2020.

[34] D. Prasetio, D. Harlili, „Predicting football match results with logistic regression“, International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016.

[35] C. Soto Valero, „Predicting Win-Loss outcomes in MLB regular season games – A comparative study using data mining methods“, International Journal of Computer Science in Sport, sv. 15, br.2, str. 91 – 112, 2016,

[36] G. Cheng, Z. Zhang, M.N. Kyebambe, N. Kimbugwe, „Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle“, Entropy, sv. 18, br. 12, str. 450, 2016.

[37] Tuan Tran, „Predicting NBA Games with Matrix Factorization“, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2016.

[38] P. Ping-Feng, C. Lan-Hung, L. Kuo-Ping, „Analyzing basketball games by a support vector machines with decision tree model“, Neural Computing & Applications, sv. 28, br. 12, str. 4159-4167, 2017.

[39] R.U. Mustafa, S.M. Nawaz, M. Ikram Ullah Lali, T. Zia, W. Mehmood, „Predicting The Cricket Match Outcome Using Crowd Opinions On Social Networks: A Comparative Study Of Machine Learning Method“, Malaysian Journal of Computer Science, sv. 30, br. 1, str. 63-76, 2017.

[40] T. Horvat, J. Job, V. Medved, „Prediction of Euroleague games based on supervised classification algorithm k-nearest neighbours“, Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support : K-BioS, str. 203 – 207, 2018.

[41] Tim Elfrink, „Predicting the outcomes of MLB games with a machine learning approach“, Vrije Universiteit Amsterdam, 2018.

[42] M.W.Y Lam, „One-Match-Ahead Forecasting in Two-Team Sports with Stacked Bayesian Regressions“, Journal of Artificial Intelligence and Soft Computing Research, sv. 8, br. 3, str. 159-171, 2018.

[43] Sujoy Ganguly, Nathan Frank, „The Problem with Win Probability“, http://www.sloansportsconference.com/wp-content/uploads/2018/02/2011.pdf, pristupljeno 6.7.2020.

[44] N. Zaveri, U. Shah, S. Tiwari, P. Shinde, L. Kumar Teli, „Prediction of Football Match Score and Decision Making Process“; International Journal on Recent and Innovation Trends in Computing and Communication, sv. 6, br. 2, str. 162-165, 2018.

[45] O. Hubáček, G. Šourek, F. Železný, “Exploiting sports-betting market using machine learning“, International Journal of Forecasting, sv. 35, br. 2, str.783-796, 2019.

[46] J. Knoll, J. Stübinger, „Machine-learning-based statistical arbitrage football betting“, KI-Künstliche Intelligenz, sv. 34, br. 1, str. 69-80, 2020.

[47] J. Stübinger, B. Mangold, J. Knoll, „Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteritics“, Applied Sciences, sv. 10, br. 1, str. 46, 2020.

[48] L. Yu, H. Liu, „Efficient Feature Selection via Analysis of Relevance and Redundancy“, Journal of Machine Learning Research, sv. 5, str. 1205 – 1224, 2004.

[49] Jiawei Han, Micheline Kamber, „Data Mining Concepts and Techniques (2nd ed.)“, Morgan Kaufmann Publishers, 2006.

[50] P.Z. Zhang, „Neural network for Classification: A Survey“, IEEE Transactions on Systems, Man and Cybernetics (Applications and Reviews), sv. 30, br. 4, str. 451-462, 2000.

[51] A. Jović, K. Brkić, N. Bogunović, „A review of feature selection methods with applications“, 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Hrvatska, 2015.

[52] R. Kohavi, G. John, “Wrappers for feature subset selection”, Artificial intelligence, sv. 97, br. 1-2, str. 273–324, 1997.

[53] I. Guyon, A. Elisseeff, “An introduction to variable and feature selection”, The Journal of Machine Learning Research, sv. 3, str. 1157–1182, 2003.

[54] M. Haghighat, H. Rastegari, N. Nourafza, „A Review of Data Mining Techniques for Result Prediction in Sports“, Advances in Computer Science, sv. 2, br. 5, 2013.

[55] K. Koseler, M. Stephan, „Machine Learning Applications in Baseball: A Systematic Literature Review“, Intelligence, sv. 31, br. 9-10, str. 745-763, 2018.

[56] Rory Bunker, Teo Susnjak, T., The Application of Machine Learning Techniques for Predicting Results in Team Sport: A Review, https://arxiv.org/pdf/1912.11762.pdf , pristupljeno 6.7.2020.

[57] R.P. Bunker, F. Thabtah, „A machine learning framework for sport result prediction“, Applied Computing and Informatics, sv. 15, br. 1, str. 27-33, 2019.

[58] Dario Bašić, Valentin Barišić, Romeo Jozak, Dražan Dizdar, „Notacijska analiza nogometnih utakmica“, Leonardo Media, Zagreb, 2015.

[59] A. Samuel, „Some Studies on Machine Learning Using the Game of Checkers“, IBM Journal of Research and Development, sv. 3, br. 3., str. 210 – 229, 1959.

[60] Ethem Alpydin, „Introduction to Machine Learning: Second Edition“, Cambridge, Massachusetts, London, England, 2010.

[61] Tom M. Mitchell, „Machine learning“, McGraw-Hill Science/Engineering/Math, 1997.

[62] Trevor Hastie, Robert Tibshirani, Jerome Friedman, „The Elements of Statistical Learnig: Data Mining, Inference, and Prediction: Second Edition“, 2009.

[63] S.B. Kotsiantis, „Supervised Machine Learning: A Review of Classification Techniques“, Informatica, sv. 31, str. 249-268, 2007.

[64] Darija Marković, „Osnove umjetne inteligencije“, http://www.mathos.unios.hr/oui/p11.pdf, pristupljeno 7.6.2020.

[65] Bojana Dalbelo Bašić, Jan Šnajder, „Uvod u umjetnu inteligenciju“, Fakultet elektrotehnike i računarstva, Sveučilište u Zagreb, https://www.fer.unizg.hr/_download/repository/UI-1-Uvod.pdf , pristupljeno 7.6.2020.

[66] Nathalie Japkowicz, Mohak Shah, „Evaluating Learning Algorithms: A Classification Perspective“, Cambridge University Press, 2011.

[67] Desmond Morris, „The Soccer Tribe“, London:Cape, 1981.

[68] J. Dowie, „Why Spain Should Win in the World Cup“, New Sci, sv. 94, br. 1309, str. 693-695, 1982.

[69] R. Pollard, „Home advantage in soccer: a retrospective analysis“, J Sports Sci, sv. 4, str. 237-248, 1986.

[70] „Home Court Advantage (HCA)“, https://www.nbastuffer.com/analytics101/home-court-advantage/, pristupljeno 6.7.2020.

[71] Jon Bois, https://www.sbnation.com/2011/1/19/1940438/home-field-advantage-sports-stats-data, pristupljeno 6.7.2020.

[72] „Weka 3: Data Mining Software in Java“, https://www.cs.waikato.ac.nz/ml/weka/, pristupljeno 6.7.2020.

[73] Ratko Grbić, „Estimacija teško mjerljivih procesnih veličina zasnovana na mješavini Gaussovih regresijskih modela“, Elektrotehnički fakultet Osijek, 2013.

[74] A. AlGhazzawi, B. Lennox, „Monitoring a complex refining process using multivariate statistics“, Control Engineering Practice, sv. 16, br. 3, str. 294–307, 2008.

[75] X. Wang, U. Kruger, and G. W. Irwin, „Process Monitoring Approach Using Fast Moving Window PCA“, Industrial & Engineering Chemistry Research, sv. 44, br. 15, str. 5691–5702, 2005

[76] J.-C. Jeng, „Adaptive process monitoring using efficient recursive PCA and moving window PCA algorithms“, Journal of the Taiwan Institute of Chemical Engineers, str. 41, br. 4, str. 475–481, 2010.

[77] M. Shu, J.J. Jiang, M. Willey, „The Effect of Moving Window on Acoustic Analysis“, Journal of Voice, sv. 30, br. 1, str. 5–10, 2016.

[78] Tom Mitchell, „Machine learning“, McGraw – Hill, New York, 1. izdanje, 1997.

[79] T. Horvat, L. Havaš, V. Medved, “Web Application for Support in Basketball Game Analysis”, Proceedings of the 3rd International Congress on Sport Sciences Research and Technology Support, str. 225–231, Lisabon, Portugal, 2015.

[80] T. Horvat, L. Havaš, D. Srpak, V. Medved, “Data-driven Basketball Web Application for Support in Making Decisions:,”, Proceedings of the 7th International Conference on Sport Sciences Research and Technology Support, str. 239–244, Beč, Austrija

Popis slika

Popis tablica

Sažetak

Predviđanje sportskih procesa, vršeno na temelju iskustva ili znanja o određenom procesu i korištenjem informacija o događaju, zanimljivo je široj javnosti u vidu sportskog klađenja, a dostupnošću velikih količina podataka sve češće postaje i tema znanstvenih istraživanja. Osim uobičajenih statističkih metoda za analizu podataka, koristi se i strojno učenje kako bi se ostvarili što bolji rezultati predviđanja sportskih ishoda. Rezultati znanstvenih istraživanja o sportskim događajima posebno su zanimljivi ekspertima, trenerima, sportskim menadžerima i upravama sportskih klubova koji ih koriste u svrhu vrednovanja učinka igrača i momčadi, kod odabira igrača, identifikacije sportskih talenata, definiranja novih strategija, itd.

U disertaciji je tema istraživanja primjena metoda strojnog učenja u predviđanju sportskih ishoda. Osim pruženog detaljnog uvida u dostupnu literaturu i trenutna postignuća u području, napravljena je i analiza znanstvenih radova i ostvarenih rezultata istraživanja te analiza dostupnih podataka o sportskim događajima kako bi se identificiralo značajke od interesa za izradu modela predviđanja sportskih ishoda. U radu su opisani postojeći indeksi korisnosti te je predložen sveobuhvatni indeks korisnosti prilagodljiv različitim sportovima koji predstavlja temelj predložene metode. Ispitana je hipoteza da za učinkovit model nije potrebno poznavanje cijele povijesti, već je dovoljno pronaći reprezentativni dio povijesti koji se u ovom slučaju naziva optimalnim vremenskim prozorom. Također, zaključeno je kako je događaje moguće klasificirati u različite kategorije kako bi se primjenom prilagodljive metode predviđanja dobili još bolji rezultati.

U radu je predstavljen sveobuhvatni indeks korisnosti kojim se može, ovisno o konkretnom problemu, vrednovati učinak igrača ili momčadi, a koji se u kasnijim fazama koristi kao početna točka predviđanja ishoda. Predložen je i postupak optimiranja parametara sveobuhvatnog indeksa korisnosti korištenjem kombinacije linearnih i nelinearnih doprinosa, a predstavljen je i algoritam izračuna i prilagodbe optimalnog vremenskog prozora sa svrhom ograničavanja doprinosa događaja iz daleke prošlosti. Sam postupak optimiranja je iterativan, a osim pronalaska optimalnog doprinosa, uključuje i postupak definiranja redoslijeda optimizacije skupa korištenih značajki. Optimalni vremenski prozor predstavlja kontinuirani vremenski period koji se koristi u svrhu određivanja relevantnosti statističkih podataka o prethodnim događajima s ciljem pronalaska podskupa skupa za učenje koji najbolje opisuje trenutno stanje analiziranog procesa, a da pritom ne doprinosi posljedičnom smanjenju rezultata predviđanja. U radu je predložen način izračuna optimalnog vremenskog prozora na temelju prosječnog učinka ili prosječnog indeksa korisnosti. Proces izračuna i prilagodbe optimalnog vremenskog prozora uključuje dva koraka. Jedan je

izračun početnog vremenskog prozora, a drugi je prilagodba vremenskog prozora na eventualne promjene. U svrhu poboljšanja rezultata predviđanja predložen je i način identifikacije događaja povećane neizvjesnosti. Događaji povećane neizvjesnosti omogućuju primjenu prilagodljivog postupka u vidu određivanja razine složenosti i načina predviđanja s konačnim ciljem poboljšanja rezultata predviđanja. Predloženi model ispitan je korištenjem skupa podataka o utakmicama NBA lige.

Ključne riječi: optimalni vremenski prozor, optimizacija, predviđanje, sportski ishodi, strojno učenje, sveobuhvatni indeks korisnosti

Abstract

An adaptive method for predicting sport outcomes based on the efficiency index and optimal time window

Although the sporting process prediction, which is based on the experience or knowledge of a particular process and the use of information about the event, is interesting to the general public mainly in the form of sports betting, the availability of large data amounts is increasingly becoming a topic of scientific research. Currently, in addition to the application of common statistical methods to analyze the available data, machine learning is used to achieve the best possible results in predicting sports outcomes. Scientific research results on sporting events are of particular interest to experts, coaches, sports managers and the management of sports clubs who use them to evaluate the players' and team's performance, to select players, to identify sporting talents, to define new strategies, etc.

The research topic of the dissertation pertains to the application of machine learning methods in predicting sports outcomes. In addition to providing a detailed insight into the available literature and the current achievements in the field, an analysis of scientific papers and research results was carried out, together with an analysis of the available data on sporting events, all in order to identify the features of interest for the development of models for predicting sports outcomes. The dissertation describes the existing efficiency indexes, specific to individual sports, which are used to assess the players' and team's performance; and it proposes a comprehensive efficiency index which can be used to assess performance in different sports. The proposed comprehensive efficiency index is the basis of the proposed predicting method. The hypothesis that was tested states that an effective model does not require the knowledge of the entire history, but that it is enough to find a representative part of history and use it to make predictions with satisfactory accuracy, which is achieved by applying the optimal time window. In the proposed method, the time window, in addition to being determined before the beginning of the application, is adjusted on the basis of the data present during the application itself, i.e. during the testing of the model.

The dissertation presents a comprehensive efficiency index which can, depending on the specific problem, evaluate the players' or team's performance, and which is to be used in the later stages as a starting point for predicting the outcome. A procedure of optimizing the parameters of the comprehensive efficiency index by using a combination of linear and nonlinear contributions is also proposed, and an algorithm for calculating and adjusting the optimal time window is presented in order to limit the event contribution from the distant past. The goal of the optimization process

is to reduce the dimensionality problem by identifying irrelevant and redundant features for the purpose of faster and more efficient execution of the prediction algorithm. The optimization process is iterative, and in addition to finding the optimal contribution, it also includes the process of defining the order of the optimization of the used set of features. The optimal time window is a continuous period used to determine the relevance of the statistical data of previous events. The goal of the optimal time window is to find a training subset that best describes the current state of the analyzed process, without consequently reducing the prediction results. The dissertation proposes a method of calculating the optimal time window based on the average process performance or average process efficiency index. The process of calculating and adjusting the optimal time window involves two steps – one is to define the initial time window, and the other is to adjust the time window in case of any changes. The proposed model was evaluated by using the NBA league game dataset.

Keywords: comprehensive efficiency index, machine learning, optimal time window, optimization, prediction, sports outcomes

Životopis

Tomislav Horvat rođen je 2. studenog 1986. godine u Varaždinu. Osnovnu školu završava u Ludbregu. Obrazovanje nastavlja u Varaždinu gdje s odličnim uspjehom završava prirodoslovno-matematički smjer Prve gimnazije Varaždin. Nakon završene srednje škole u Zagrebu upisuje Fakultet elektrotehnike i računarstva. Preddiplomski studij Računarstva završava 2008. godine te stječe titulu sveučilišnog prvostupnika inženjera računarstva. Diplomski sveučilišni studij Telekomunikacije i informatika završava u siječnju 2011. godine pod mentorstvom prof.dr.sc. Zorana Skočira. Krajem 2011. godina završava tečaj za Web dizajnera, a sredinom 2012. godine upisuje izvanredni studij pedagoško-psihološke i metodičko-didaktičke izobrazbe kojeg uspješno završava u veljači 2013. godine. U studenom 2017. godine završava program osposobljavanja za obavljanje poslova trenera košarke.

Nakon završetka studija godinu dana je radio kao pripravnik u informatičkom odjelu Hrvatskih šuma te 6 mjeseca kao programer SMS servisa nakon čega se najprije kao vanjski suradnik, a zatim i kao asistent na kojem mjestu radi i danas, zapošljava na Sveučilištu Sjever. Kao član odjela za Elektrotehniku sudjeluje u izvođenju nastave na kolegijima Baze podataka i SQL, Programski alati 1 i Stručna praksa, a do nedavno je sudjelovao i na izvođenju nastave iz kolegija Programski jezici i algoritmi i Programski alati 2. Uže područje interesa odnosi se na strojno učenje i upotrebu ICT u sportu, a posebno na predviđanje sportskih ishoda.

Do prije nekoliko godina aktivno se bavio igranjem košarke na amaterskoj razini, a već devet godina trenira mlađedobne kategorije. Godinu dana je bio trener seniorske momčadi te tri godine pomoćni trener seniorske momčadi. Kao autor ili koautor objavio je deset znanstvenih radova u inozemnim znanstvenim časopisima i zbornicima. Govori engleski jezik.

Trenutno živi u Ludbregu sa suprugom Tenom i sinom Jakovom.

Prilog

A.1. Rezultati optimizacije sveobuhvatnog indeksa korisnosti

U ovom poglavlju su prikazani rezultati optimizacije koeficijenata $v_e(v_e')$. Tablica A.1 prikazuje rezultate prve faze piramidalne optimizacije, dok Tablica A.2 prikazuje rezultate druge faze piramidalne optimizacije problema predviđanja ishoda košarkaških utakmica.

Tablica A.1. Doprinos koeficijenta $v(v')$ značajki – prva faza.

Značajka (x)	**Bez x**	**Smanjeni doprinos komponente**					**Referentni indeks (x)**	**Pojačani doprinos komponente**				**$y = razlika\ učinaka$**				**Razlika**
		$\sqrt[3]{x}$	$\sqrt{x}$	$\frac{x}{2}$	$x-\sqrt{x}$	$x-\sqrt[3]{x}$		$x+\sqrt[3]{x}$	$x+\sqrt{x}$	$x+\frac{x}{2}$	$x+x$	$x+\sqrt[3]{y}$	$x+\sqrt{y}$	$x+y$	$x+y^2$	
$miss\_2fg$	58,73 %	58,95 %	59,13 %	60,50 %	61,59 %	61,81 %	61,80 %	61,85 %	62,12 %	62,21 %	62,09 %	62,27 %	62,23 %	62,09 %	61,50 %	+ 0,47 %
ftm	60,73 %	60,87 %	60,86 %	61,74 %	62,23 %	62,25 %	62,27 %	62,24 %	62,22 %	62,06 %	61,85 %	62,28 %	62,19 %	61,85 %	61,02 %	+ 0,01 %
$2fgm$	61,61 %	61,67 %	61,74 %	62,10 %	62,14 %	62,19 %	62,28 %	62,25 %	62,18 %	61,16 %	59,96 %	61,71 %	61,41 %	59,96 %	56,51 %	-
def_reb	61,81 %	61,93 %	62,04 %	62,38 %	62,22 %	62,27 %	62,28 %	62,22 %	62,18 %	62,02 %	61,89 %	62,09 %	62,07 %	61,89 %	61,75 %	+ 0,10 %
$asist$	62,26 %	62,34 %	62,26 %	62,15 %	62,32 %	62,41 %	62,38 %	62,33 %	62,33 %	61,94 %	61,46 %	61,88 %	61,79 %	61,46 %	60,80 %	+ 0,03 %
of_reb	61,83 %	61,82 %	61,90 %	62,22 %	62,23 %	62,32 %	62,41 %	62,30 %	62,25 %	62,18 %	62,05 %	62,03 %	62,16 %	62,05 %	61,05 %	-
$miss\_3fg$	61,40 %	61,41 %	61,53 %	61,73 %	62,22 %	62,25 %	62,41 %	62,24 %	62,23 %	61,88 %	61,22 %	61,99 %	61,86 %	61,22 %	57,07 %	-
$miss\_ft$	62,12 %	62,09 %	62,08 %	62,28 %	62,22 %	62,28 %	62,41 %	62,28 %	62,28 %	62,24 %	62,20 %	62,13 %	62,22 %	62,20 %	61,80 %	-
to	61,49 %	61,50 %	61,59 %	61,84 %	62,22 %	62,29 %	62,41 %	62,42 %	62,38 %	62,36 %	62,39 %	62,45 %	62,36 %	62,39 %	62,06 %	+ 0,04 %
$3fgm$	59,09 %	59,47 %	59,73 %	61,39 %	62,49 %	62,49 %	62,45 %	62,28 %	62,48 %	62,06 %	61,36 %	61,57 %	61,62 %	61,86 %	62,20 %	+ 0,04 %
st	62,53 %	62,56 %	62,58 %	62,62 %	62,56 %	62,52 %	62,49 %	62,42 %	62,45 %	62,24 %	61,98 %	61,97 %	61,94 %	61,98 %	62,04 %	+ 0,13 %
bl	62,32 %	62,34 %	62,37 %	62,53 %	62,59 %	62,54 %	62,62 %	62,60 %	62,52 %	62,48 %	62,49 %	62,52 %	62,57 %	62,49 %	62,43 %	-
f	62,62 %	62,61 %	62,61 %	62,46 %	62,43 %	62,33 %	62,26 %	62,24 %	62,25 %	62,05 %	61,54 %	61,87 %	61,73 %	61,54 %	60,77 %	-
															Ukupno:	+ 0,82 %

Tablica A.2. Doprinos koeficijenta $v(v')$ značajki – druga faza.

Značajka (x)	**Bez x**	**Smanjeni doprinos komponente**					**Referentni indeks (x)**	**Pojačani doprinos komponente**				**$y = razlika\ učinaka$**				**Razlika**
		$\sqrt[3]{x}$	$\sqrt{x}$	$\frac{x}{2}$	$x-\sqrt{x}$	$x-\sqrt[3]{x}$		$x+\sqrt[3]{x}$	$x+\sqrt{x}$	$x+\frac{x}{2}$	$x+x$	$x+\sqrt[3]{y}$	$x+\sqrt{y}$	$x+y$	$x+y^2$	
f	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
bl	62,32 %	62,34 %	62,37 %	62,53 %	62,59 %	62,54 %	62,62 %	62,60 %	62,52 %	62,48 %	62,49 %	62,52 %	62,57 %	62,49 %	62,43 %	-
st	62,53 %	62,54 %	62,62 %	62,56 %	62,55 %	62,59 %	62,62 %	62,60 %	62,59 %	62,56 %	62,49 %	62,22 %	62,35 %	62,49 %	62,52 %	-
$3fgm$	59,00 %	59,08 %	59,47 %	61,19 %	62,69 %	62,59 %	62,62 %	62,57 %	62,59 %	62,05 %	61,37 %	62,26 %	62,18 %	61,37 %	59,45 %	+ 0,07 %
to	61,52 %	61,62 %	61,73 %	62,17 %	62,41 %	62,59 %	62,69 %	62,75 %	62,61 %	62,44 %	61,77 %	62,17 %	62,19 %	61,77 %	60,78 %	+ 0,06 %
$miss\_ft$	62,46 %	62,47 %	62,47 %	62,53 %	62,62 %	62,64 %	62,75 %	62,65 %	62,66 %	62,65 %	62,18 %	62,19 %	62,25 %	62,18 %	61,95 %	-
$miss\_3fg$	61,84 %	61,88 %	61,99 %	62,53 %	62,58 %	62,59 %	62,75 %	62,70 %	62,62 %	61,97 %	61,06 %	61,85 %	61,68 %	61,06 %	56,84 %	-
of_reb	61,81 %	61,84 %	61,89 %	62,16 %	62,58 %	62,63 %	62,75 %	62,75 %	62,77 %	62,40 %	62,22 %	62,45 %	62,39 %	62,22 %	61,12 %	+ 0,02 %
$asist$	62,52 %	62,43 %	62,39 %	62,33 %	62,62 %	62,68 %	62,77 %	62,73 %	62,79 %	62,22 %	61,79 %	61,95 %	61,90 %	61,79 %	61,18 %	+ 0,02 %
def_reb	62,17 %	62,19 %	62,25 %	62,43 %	62,67 %	62,78 %	62,79 %	62,79 %	62,85 %	62,72 %	62,66 %	62,48 %	62,56 %	62,66 %	62,81 %	+ 0,06 %
$2fgm$	61,87 %	61,95 %	61,98 %	62,67 %	62,77 %	62,74 %	62,85 %	62,76 %	62,61 %	61,00 %	59,69 %	61,63 %	61,22 %	59,69 %	56,43 %	-
ftm	60,73 %	60,90 %	61,06 %	61,74 %	62,54 %	62,73 %	62,85 %	62,82 %	62,53 %	62,23 %	61,42 %	62,15 %	61,92 %	61,42 %	58,80 %	-
$miss\_2fg$	58,45 %	58,60 %	58,83 %	60,87 %	62,65 %	62,65 %	62,85 %	62,70 %	62,47 %	62,53 %	62,11 %	62,44 %	62,52 %	62,11 %	60,81 %	-
															Ukupno:	+ 0,23 %