SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1577

DUBINSKA ANALIZA STATISTIČKIH KATEGORIJA PRAĆENJA IGRAČA U KOŠARKAŠKIM EKIPAMA

Igor Stančin

Zagreb, lipanj 2018.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA DIPLOMSKI RAD PROFILA

Zagreb, 6. ožujka 2018.

DIPLOMSKI ZADATAK br. 1577

Pristupnik: **Igor Stančin (0036466090)**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Dubinska analiza statističkih kategorija praćenja igrača u košarkaškim ekipama**

Opis zadatka:

Pravila pridruživanja i pravila prekrivanja dvije su poznate tehnike za pronalaženje skrivenih obrazaca u velikim skupovima podataka. Zadatak ovog diplomskog rada je provesti dubinsku analizu statističkih kategorija praćenja igrača u košarkaškim ekipama koristeći spomenute tehnike. Dva su cilja analize: pronalazak najboljeg modela za predviđanje koja košarkaška ekipa će biti pobjednička i pronalaženje značajnih pravila između različitih statističkih kategorija. Analiza će se temeljiti na prikupljenim statističkim kategorijama praćenja igrača tijekom nekoliko godina u američkoj košarkaškoj ligi National Basketball Association (NBA), koja je najpoznatija i najrasprostranjenija košarkaška liga na svijetu. Ovaj rad uključuje prikupljanje relevantnih podataka, temeljito razmatranje podataka, što uključuje njihovo predstavljanje postupcima deskriptivne statistike, prilagođavanje podataka za postupke dubinske analize, provođenje dubinske analize i tumačenje rezultata. Prilikom razmatranja skupa statističkih kategorija moguće je i predlaganje i izgradnja novih statističkih kategorija. U okviru diplomskog rada, očekuje se izrada fleksibilnog programskog rješenja za analizu podataka košarkaških ekipa koje će omogućiti jednostavnu pripremu podataka i provođenje analize, kako na do sada dostupnim podatcima tako i na podatcima koji će očekivano biti dostupni u budućnosti.

Zadatak uručen pristupniku: 16. ožujka 2018.
Rok za predaju rada: 29. lipnja 2018.

Mentor:

Doc. dr. sc. Alan Jović

Predsjednik odbora za diplomski rad profila:

Prof. dr. sc. Siniša Srbljić

Djelovođa:

Doc. dr. sc. Tomislav Hrkać

Zahvale

Zahvaljujem mentoru doc. dr. sc. Alanu Joviću na usmjeravanju prilikom izrade ovog diplomskog rada i na tome što je uvijek bio voljan odvojiti dio svog vremena kako bi mi pomogao. Također zahvaljujem Ani, prijateljima i svojoj obitelji, posebice roditeljima koji su mi bili podrška tijekom studija.

Sadržaj

1. Uvod

Košarka je jedan od najpopularnijih sportova na globalnoj razini. Svoju popularnost u velikoj mjeri može zahvaliti svojoj dinamičnosti i nepredvidljivosti. Akcije se brzo izmjenjuju te se rijetko ponavljaju isti događaji. Posjed lopte u košarci može započeti i završiti na nekoliko različitih načina. Primjerice, posjed može započeti dodavanjem iz auta, ukradenom loptom, defanzivnim skokom i slično, dok završiti može šutom na koš, izgubljenom loptom, istekom vremena za napad i slično. Između samog početka i kraja posjeda može se dogoditi puno različitih događaja, primjerice odbijena lopta u aut, tehnička pogreška, dodavanja, driblanje lopte i slično. S obzirom na mnogo različitih događaja u košarkaškoj utakmici, prikupljanje statistika zahtjevan je zadatak [16].

Kroz povijest se prikupljanje statistika razvijalo zajedno s košarkom. Tako se recimo prikupljanje broja skokova u NBA ligi (engl. *National Basketball Association*) uvelo tek u sezoni 1950./'51. U početnim fazama prikupljanja skokova, skokovi su se „namjerno“ brojali na krivi način. Današnji ekipni skok (npr. nakon promašenog prvog slobodnog bacanja) tada se pripisivao bilo kojem igraču, obično onome s najvećim brojem skokova u ekipi. Takvo prikupljanje statistika u startu nam stvara velike rupe u podatcima jer će određeni igrači imati nerealno visoke prosjeke skokova. Tek u sezoni 1973./'74. počeli su se brojati defanzivni skokovi zasebno od ofanzivnih. Od iste sezone počele su se prikupljati i statistike o ukradenim loptama i blokadama. Šut za tri poena je u NBA ligu uveden tek od sezone 1979./'80., tako da i podatci za njega postoje tek od te sezone.

U novije vrijeme prikupljanje statistika još se značajnije popravilo. Jedan od primjera je prikupljanje broja različitih vrsta šutova. Za svaki se šut zna je li bilo polaganje, zakucavanje, šut nakon driblinga ili nešto četvrto. Za svaki se šut zna s koje je pozicije na terenu upućen. Uz to, pojavile su se i mnoge napredne statistike koje na razne načine kombiniraju postojeće s ciljem što boljeg prikaza kvalitete ekipe i/ili igrača kroz samo jednu brojku. Tako recimo postoje napadački i obrambeni rejting, PIE (engl. *Player Impact Estimate*), PER (engl. *Player Efficiency Rating*) i mnoge druge. Statistike su se počele skalirati po minutama (ostvareni broj u nekoj statističkoj kategoriji skalira se tako da se vidi kakva bi statistika bila da je igrač igrao 40 minuta u utakmici) i po posjedima (isto kao za minute, samo ovoga puta se skalira na 100 posjeda lopte).

Prikupljanje kvalitetne i detaljne statistike od iznimne je važnosti kako za analitičare, tako i za same trenere. Kvalitetnom analizom statistike lakše je pripremiti ekipu za svaku pojedinu utakmicu. Lakše je predvidjeti što očekivati od protivnika. Mnoge stvari koje se ne vide samim gledanjem utakmica vide se u statistikama. Analitika je posebno uzela maha unazad nekoliko godina, kada su pojedine ekipe cijelu svoju logistiku počele zasnivati upravo na naprednim analizama statistika. Upravo je iz tog razloga NBA liga uložila velike novce u sustav računalnog vida, čija je primarna zadaća upravo to, prikupljanje što kvalitetnije i što detaljnije statistike.

Sustav računalnog vida prikuplja pozicije na terenu od svih igrača i lopte 25 puta u sekundi. Kasnije se obradom tih podataka automatski kreiraju mnoge statistike. Uz to što je prikupljanje statistika postalo bitno jednostavnije, mnoge su statističke kategorije unaprijeđene, a mnoge su se nove kategorije počele pratiti. Valja naglasiti kako je NBA liga u startu objavljivala javno podatke o pozicijama svih igrača i lopte u svakom trenutku. Danas, na žalost svih istraživača, ti podatci više nisu javni.

Cilj ovoga rada je analizirati jednu grupu novih statističkih kategorija koje se prikupljaju zadnjih nekoliko godina. Analizirana grupa zove se „praćenje igrača“ (engl. *player tracking)*. Cilj analize je identificirati koje od statističkih kategorija predstavljaju najveću razliku između pobjedničkih i gubitničkih ekipa. Ideja je dobiti odgovore na pitanja kao što su: „Jesu li bitniji branjeni ili nebranjeni šutovi?“, „Je li više pretrčanih kilometara uvijek bolje?“, „Je li više dodavanja uvijek bolje?“ i slično. Jedan dio analize proveli smo u našem nedavno objavljenom radu [16], ali u ovom radu provest ćemo dublju analizu koristeći neke od naprednih algoritama te na većem skupu podataka.

2. Srodna istraživanja

J. Kubatko i sur. [30] u svome radu daju pregled dostupnih i općeprihvaćenih naprednih statistika koje se koriste u košarci. T. Al Baghal i sur. [34] u svome radu analiziraju grupu takvih statistika pod nazivom „četiri faktora“ (engl. *four factor*), a u koju spadaju efektivni postotak šuta, učestalost slobodnih bacanja, broj izgubljenih lopti po posjedu i postotak napadačkih skokova. G. Ziv i sur. [28] u svome radu ispituju ima li uopće smisla koristiti statistike sa svake utakmice s ciljem predviđanja ranga ekipe na kraju sezone. Promatrali su sezone od 2003./'04. do 2009./'10. i pokazali su kako, ovisno o sezoni, postoje korelacije između ranga ekipa na kraju sezone i nekih od sljedećih statističkih kategorija: postotka šuta za dva i za tri poena, obrambenih i napadačkih skokova, asistencija i faula. Pokazali su i da su neke od statističkih kategorija visoko korelirane, primjerice, asistencije i broj zabijenih koševa. Time su pokazali kako takve analize imaju smisla jer postoje korelacije.

M. R. Summers [20] u svome radu statističkom analizom utvrđuje postotak šuta i obrambene skokove kao faktore koji najviše koreliraju s pobjedom utakmice. Do sličnog zaključka došli su R. Hofler i sur. [36] i S. Trninić i sur. [37]. G. Csataljay i sur. [24] provode sličnu statističku analizu kao i mi, ali razmatraju samo neizvjesne utakmice (relativno mala razlika u poenima na kraju utakmice) i grupu osnovnih statistika. Utvrdili su kako pobjedničke ekipe imaju manje pokušaja šutova za tri poena, viši postotak šuta, veći broj zabijenih slobodnih bacanja i postotak slobodnih bacanja te više obrambenih skokova. Na sličan način G. Csataljay i sur. [25] provode istraživanje nad neizvjesnim utakmicama, ali na razini četvrtina utakmice. Došli su do zaključka kako su broj zabijenih slobodnih bacanja i četiri kategorije povezane s brojem skokova ključne razlike u neizvjesnim četvrtinama. Slično istraživanje samo nad zadnjim četvrtinama neizvjesnih utakmica proveli su M. Ruano i sur. [29]. Istraživali su najznačajnije razlike među domaćim i gostujućim ekipama u tri različita „trenda“: 1) rezultatski izjednačenim četvrtim četvrtinama; 2) četvrtim četvrtinama u kojima je domaćin ostvario značajniju prednost; 3) četvrtim četvrtinama u kojima je gost ostvario značajniju prednost. Utvrdili su značaj rezultata na početku četvrtine za prvi i treći trend, kvaliteta protivnika također utječe u drugom i trećem trendu, koš za tri poena s središnjeg dijela terena u prvom i drugom

trendu i još nekoliko značajnih razlika kroz sva tri trenda. M. Gomez i sur. [35] utvrđuju da su asistencije i obrambeni skokovi ono što razlikuje pobjedničke ekipe u španjolskoj ligi.

Nešto drugačiji pristup analizi imao je M. U. Özmen [23] koji u svome radu promatra marginalni doprinos svakoj pojedinoj statistici na temelju podataka iz Eurolige. Najzanimljiviji rezultat je da, ukoliko ekipa ima jednu izgubljenu loptu više od protivničke, šanse za pobjedu im se smanjuju za 41%. G. Csátaljay i sur [27] proveli su zanimljivu analizu utjecaja obrambenog pritiska na uspješnost šutiranja. U radu navode kako su obrambeni pritisak na igrača koji upućuje udarac definirali u ovisnosti o udaljenosti najbližeg obrambenog igrača u trenutku ispuštanja lopte, ali ne navode koje su to konkretne udaljenosti. Pokazalo se kako tako definirani obrambeni pritisak uvelike utječe na smanjenje postotka šuta protivnika. M. Lopez i sur. [38] u svome radu opisuju model s kojim su pobijedili na Kaggle natjecanju u predviđanju NCAA (sveučilišna košarkaška liga) pobjednika. Uz to, pokazali su kako je sreća faktor koji utječe na ishod Kaggle natjecanja. Naime, čak i kada bi postavili vjerojatnosti koje je dao njihov model kao stvarne, i dalje su im šanse za pobjedu na natjecanju bile tek 20%.

S. J. Ibáñez i sur. [31] u svome radu koristili su diskriminativnu funkciju za određivanje diskriminativnih značajki između pobjedničkih i gubitničkih ekipa na razini sezone. Kao diskriminirajuće značajke pokazale su se asistencije, ukradene lopte i blokade te su uz pomoć navedene funkcije dobili točnost klasifikacije od 82,4%. A. Jadhav i sur. [18] u svome radu koriste stroj potpornih vektora (engl. *Support Vector Machines*) kako bi na temelju grupe osnovnih statistika predviđali pobjednika utakmica u doigravanju (engl. *playoff*). Na stvarnim podatcima su postigli točnost od 88%. G. Cheng i sur. [19] u svome radu također predviđaju pobjednika utakmica iz doigravanja s modelom maksimalne entropije i postižu točnost od 74,4% s prosjecima od zadnjih šest utakmica i 28 značajki osnovnih statističkih kategorija. S. B. Caudill [22] u svome radu koristi procjenitelj maksimalnog skora (engl. *maximum score estimator*) za predviđanje ishoda NCAA utakmica i postiže točnost od oko 75%.

Od istraživanja koja se zasnivaju na novoj tehnologiji prikupljanja podataka i novim statistikama, valja izdvojiti D. Cervone i sur. [21]. Oni su na temelju kretanja svih igrača i lopte kreirali novu mjeru koja u stvarnom vremenu predviđa očekivani broj poena iz svakog posjeda lopte. J. Sampaio i sur. [26] koristili su podatke o vrsti akcija kako bi

usporedili performanse all-star igrača i svih ostalih (svake godine se igra utakmica između dvije konferencije s najboljim pojedincima te sezone koje biraju navijači i treneri te svi koji igraju na toj utakmici su all-star igrači). A. Franks i sur. [33] u svome radu također koriste pozicije svih igrača i lopte u svakom trenutku utakmice kako bi pokušali kreirati novu obrambenu metriku, koja je tradicionalno zapostavljena u odnosu na napadačku. U svome radu računaju „kontra poene“ (engl. *counterpoints*) te oko njih grade tri zanimljive studije slučaja (engl. *case study*). U prvom slučaju kontra poene pripisuju igraču koji je inicijalno (u početku posjeda lopte) čuvao šutera, u drugom slučaju kontra poene pripisuju igraču koji je čuvao šutera neposredno prije šuta te u trećem slučaju kontra poene pripisuju proporcionalno svakome od igrača koji je tijekom posjeda čuvao šutera. Ovakvom metrikom pružaju novi zanimljivi uvid u kvalitetu obrane, no i sami zaključuju kako je bez točnog znanja o timskoj strategiji u obrani teško napraviti univerzalno kvalitetnu obrambenu metriku.

3. Korišteni materijali i metode

3.1. Korišteni podatci

NBA liga sastoji se od 30 ekipa. Svaka ekipa tijekom regularne sezone odigra 82 utakmice, što znači da se u regularnoj sezoni odigra ukupno 1230 utakmica. Skup podataka koji smo koristili u analizi sastoji se od statistika koje spadaju u grupu statistika pod imenom „praćenje igrača“ (engl. *player tracking),* i to iz pet sezona počevši od sezone 2013./'14. do sezone 2017./'18. Analizirani su samo podatci iz regularnih sezona. Nekoliko utakmica iz svake sezone nije imalo dostupne navedene statistike te su te utakmice uklonjene iz skupa podataka. Broj uklonjenih utakmica je mali, od dvije do osam, ovisno o sezoni, tako da je skup podataka i dalje dovoljno velik i reprezentativan za analizu.

Statističke kategorije koje se nalaze u navedenoj grupi su:

- DIST – (engl. *distance run*) pretrčana udaljenost (u miljama)
- SPD – (engl. *Prosjek speed*) prosječna brzina kretanja
- TCHS – (engl. *touches*) broj dodira s loptom
- PASS – (engl. *passes*) broj dodavanja
- AST – (engl. *assists*) broj asistencija. Asistencija je posljednje dodavanje suigraču koje je direktno dovelo do postizanja poena te se igrač koji je primio dodavanje odmah počeo kretati prema košu s ciljem postizanja poena.
- SAST – (engl. *secondary assists*) igrač dobiva sekundarnu asistenciju ako je dodao loptu igraču koji je ostvario asistenciju unutar jedne sekunde i bez driblinga s loptom
- DFGM – (engl. *field goals made by opponent while defending the rim*) broj zabijenih koševa protivnika kada je branjen u blizini obruča
- DFGA – (engl. *field goals attempted by opponent while defending the rim*) broj pokušaja zabijanja koša protivnika kada je branjen u blizini obruča
- DFG% - (engl. *field goal percentage by opponent while defending the rim*) postotak zabijenih koševa protivnika kada je branjen u blizini obruča
- ORBC – (engl. *offensive rebound chances*) broj prilika za hvatanje napadačkih skokova

- DRBC – (engl. *defensive rebound chances*) broj prilika za hvatanje obrambenog skoka
- RBC – (engl. *rebound chances*) ukupan broj prilika za hvatanje skoka
- FG% - (engl. *field goal percent*) postotak šuta
- CFGM – (engl. *contested field goals made*) broj zabijenih branjenih šutova
- CFGA – (engl. *contested field goals attempted*) broj pokušaja zabijanja branjenih šutova
- CFG% - (engl. *contested field goal percentage*) postotak zabijenih branjenih šutova
- UFGM - (engl. *uncontested field goals made*) broj zabijenih nebranjenih šutova
- UFGA - (engl. *uncontested field goals attemted*) broj pokušaja zabijanja nebranjenih šutova
- UFG% - (engl. *uncontested field goal percentage*) postotak zabijenih nebranjenih šutova
- FTAST – (engl. *free throw assist*) igrač je nagrađen asistencijom za slobodna bacanja ako je dodao loptu igraču koji je izborio slobodna bacanja unutar jednog driblinga od primanja lopte

Iako se grupa statistika koju smo proučavali u radu zove „praćenje igrača“, te statističke kategorije postoje i na razini ekipa. U ovome radu proučavane su upravo na razini ekipa. Nije poznat točan način na koji se statističke kategorije s razine igrača prebacuju na razinu ekipe, ali naša je pretpostavka da se kategorije za koje to ima smisla samo zbroje, dok se kategorije koje predstavljaju prosjeke onda računaju iz tih zbrojeva. Sve navedene statistike, kao i objašnjenja pojedinih kategorija, dostupne su na službenoj stranici NBA statistika [1]. Za pribavljanje podataka koristili smo nekoliko javno dostupnih programskih rješenja uz manje modifikacije nad njima. U ostatku rada koristit će se gore navedene kratice i navedeni opis za svaku pojedinu kraticu jer su moguće manje varijacije u opisu pojedinih kategorija na različitim NBA-ovim stranicama.

3.2. Statistička analiza

Prilikom statističke analize podijelit ćemo podatke iz svake sezone u dvije grupe te promatrati razliku među tim grupama. S obzirom da nam je cilj saznati koje od ranije navedenih novih statistika čine razliku između pobjednika i gubitnika u košarci, podjelu ćemo napraviti na dva načina. U prvom načinu ćemo pobjednika svake zasebne utakmice označiti kao pobjedničku ekipu, a gubitnika kao gubitničku, dok ćemo u drugom načinu podjele ekipe koje su na kraju sezone imale 50 ili više pobjeda označiti kao pobjedničke, a ekipe s 35 ili manje pobjeda ćemo označiti kao gubitničke. U nastavku rada će se izraz „na razini utakmice“ odnositi na razliku među ekipama u prvoj podjeli podataka, a „na razini sezone“ na razliku među ekipama u drugoj podjeli podataka.

Kako bismo uočili razlike među navedenim grupama, izračunat ćemo srednje vrijednosti i standardne devijacije za svaku pojedinu statističku kategoriju. Nakon toga, koristit ćemo Mann-Whitney U Test kako bismo odredili značajnost razlike među tim razdiobama. Kao stupanj značajnosti izabrali smo $\alpha_0 = 0{,}01$. S obzirom da test ponavljamo 20 puta, po jednom za svaku kategoriju, koristit ćemo i Bonferronievu korekciju kako bismo smanjili vjerojatnost lažno pozitivnih rezultata. Formula za Bonferronievu korekciju je

$$\alpha = \frac{\alpha_0}{m} \tag{1}$$

gdje nam je $m = 20$, tako da je korišteni stupanj značajnosti $\alpha = 5{,}0E - 4$.

3.3. Dubinska analiza

Primarni cilj dubinske analize u ovom radu je pokušati pronaći pravila o tome što je potrebno da bi ekipa bila pobjednička. U tu svrhu koristili smo dva algoritma za izgradnju pravila, Apriori i Ripper (engl. *repeated incremental pruning to produce error reduction*). Uz to, koristili smo i tri dodatna klasifikatora kako bismo određivali i uspoređivali točnosti klasifikacije na pojedinim značajkama te na taj način pokušali utvrditi „koliko informacije“ o pobjedniku se nalazi u pojedinoj značajci. Uz ranije spomenuti Ripper, korišteni klasifikatori su još slučajna šuma (engl. *Random Forest*), naivni Bayesov klasifikator (engl. *Naive Bayes Classifier*) i stroj s potpornim vektorima zasnovan na slijednoj minimalnoj optimizaciji (engl. *Sequential Minimal Optimization Based Support Vector Machine,* dalje:

SMO) s polinomijalnom jezgrom (engl. *Polynomial Kernel*). Za sve algoritme korištene su implementacije iz alata Weka [3].

Navedeni klasifikacijski algoritmi bit će iskorišteni i kako bismo pokušali izgraditi model za predikciju pobjednika utakmice. Predikcije će se raditi na temelju prosjeka zadnjih *n* utakmica. Parametar *n* odredit ćemo eksperimentalno.

3.3.1. Algoritam Apriori

Apriori je algoritam za izgradnju asocijativnih pravila. S obzirom da algoritam pronalazi skrivene obrasce u diskretnim podatcima, najčešće je korišten prilikom analize košarica u trgovačkim lancima (engl. *market basket analysis*) ili u jednostavnim sustavima za preporuke (engl. *recommendation systems*). Može se koristiti i u ovakvim slučajevima, gdje su podatci kompliciraniji, kako bi se dobio uvid u to koje grupe značajki su međuovisne. Prednost algoritma Apriori je ta što pronalazi skrivene obrasce među bilo kojim parovima i/ili grupama značajki, a ne samo u odnosu na ciljnu varijablu (oznaku klase). Veliki nedostatak algoritma Apriori je taj što radi isključivo s diskretnim vrijednostima. Ukoliko izvorni skup podataka nije diskretan, uvijek se postavlja pitanje na koji način provesti diskretizaciju kako bismo dobili smislene grupe među kojima bi algoritam pronašao smislena pravila. Implementacija algoritma Apriori u Weki zasniva se na radu R. Agrawal i sur. [4], te B. Liu i sur. [5]. P. Tanna i sur. u svome radu [2] prikazuju pseudokod algoritma Apriori. Navedeni pseudokod prikazan je u nastavku:

```
procedure Apriori (T, minSupport)
{ //T is the database and minSupport is the minimum support
        L1= {frequent items};
        for (k= 2; Lk-1 !=Ø; k++)
        {
                Ck= candidates generated from Lk-1
                //that is cartesian product Lk-1 x Lk-1 and
                //eliminating any k-1 size itemset that is not
                //frequent
                for each transaction t in database do
                {
                        #increment the count of all candidates in Ck
                        #that are contained in t
                        Lk = candidates in Ck with minSupport
                }//end for each
```

```
        }//end for
  return ;
}
```

3.3.2. Algoritam Ripper

Ripper je klasifikacijski algoritam za izgradnju indukcijskih pravila. Uz pomoć njega iz podataka dobivamo jasna pravila koja nam govore kakve moraju biti vrijednosti nekih od značajki iz podataka te koja će u tome slučaju biti oznaka klase. Za razliku od algoritma Apriori koji radi isključivo s diskretnim vrijednostima i pronalazi pravila među bilo kojim značajkama, Ripper radi i s numeričkim značajkama te uvijek gradi pravila u odnosu na ciljanu varijablu (oznaku klase). Algoritam Ripper gradi pravila slično kao i stabla odluke, po principu podijeli pa vladaj (engl. *divide and conquer*). Prilikom određivanja granica pojedinih značajki, algoritam isprobava sve vrijednosti koje ta značajka sadrži te traži granicu koja bi na najbolji način razdvojila primjere. Implementacija algoritma Ripper (JRip) u Weki temelji se na radu W. W. Cohena [6]. Pseudokod algoritma preuzet je iz knjige I. H. Witten i sur. [39] i prikazan je u nastavku:

```
Initialize E to the instance set
For each class C, from smallest to largest
        BUILD:
                Split E into Growing and Pruning sets in the ratio 2:1
                Repeat until (a) there are no more uncovered examples of C; or
                        (b) the description length (DL) of Pravilaet and examples is
                        64 bits greater than the smallest DL found so far, or (c)
                        the error rate exceeds 50%:
                GROW phase: Grow a rule by greedily adding conditions until the
                        rule is 100% accurate by testing every possible value of
                        each attribute and selecting the condition with greatest
                        information gain G
                PRUNE phase: Prune conditions in last-to-first order. Continue
                        as long as the worth W of the rule increases
        OPTIMIZE:
                GENERATE VARIANTS:
                For each rule R for class C,
                        Split E afresh into Growing and Pruning sets
                        Remove all instances from the Pruning set that are covered
                            by other Pravila for C
                        Use GROW and PRUNE to generate and prune two competing Pravila
                            from the newly split data:
                                R1 is a new rule, rebuilt from scratch;
```

R2 is generated by greedily adding antecedents to R.
Prune using the metric A (instead of W) on this reduced data
SELECT REPRESENTATIVE:
Replace R by whichever of R, R1 and R2 has the smallest DL.
MOP UP:
If there are residual uncovered instances of class C, return to
the BUILD stage to generate more Pravila based on these instances.
CLEAN UP:
Calculate DL for the whole Pravilaet and for the Pravilaet with each
rule in turn omitted; delete any rule that increases the DL
Remove instances covered by the Pravila just generated
Continue

3.3.3. Slučajna šuma

Algoritam slučajna šuma (engl. *random forest*) algoritam je koji koristi ansamble s više stabala odluke. S obzirom da se sastoji od mnogo stabala odluke, konačna odluka klase dobiva se većinskim glasanjem. Za razliku od običnih pojedinačnih stabala odluke, algoritam slučajnih šuma nije sklon prenaučenosti na skupu za učenje, budući da koristi dva izvora slučajnosti: pri generiranju skupova za učenje i pri grananju u čvorovima stabala. Implementacija algoritma slučajnih šuma u Weki temelji se na radu L. Breimana (2001) [7]. Pseudokod glavne funkcije algoritma slučajnih šuma nalazi se u nastavku. Pseudokod je preuzet s [8], gdje se mogu pronaći i pseudokodovi ostalih korištenih funkcija u algoritmu.

```
Require: Initially the tree has exactly one leaf (TreeRoot) which covers the whole space
Require: The dimensionality of the input, D. Parameters λ, m and τ .
  SelectCandidateSplitDimensions(TreeRoot, min(1 + Poisson(λ), D))
  for t = 1 . . . do
    Receive (Xt, Yt, It) from the environment
    At ← leaf containing Xt
    if It = estimation then
      UpdateEstimationStatistics(At, (Xt, Yt))
      for all S ∈ CandidateSplits(At) do
        for all A ∈ CandidateChildren(S) do
          if Xt ∈ A then
            UpdateEstimationStatistics(A, (Xt, Yt))
          end if
        end for
      end for
    else if It = structure then
```

```
        if At has fewer than m candidate split points then
          for all d ∈ CandidateSplitDimensions(At) do
            CreateCandidateSplit(At, d, πdXt)
          end for
        end if
        for all S ∈ CandidateSplits(At) do
          for all A ∈ CandidateChildren(S) do
            if Xt ∈ A then
              UpdateStructuralStatistics(A, (Xt, Yt))
            end if
          end for
        end for
        if CanSplit(At) then
          if ShouldSplit(At) then
            Split(At)
          else if MustSplit(At) then
            Split(At)
          end if
        end if
      end if
    end for
```

3.3.4. Naivni Bayesov klasifikator

Naivni Bayesov klasifikator (engl. *Naive Bayes*) spada u probabilističke klasifikatore. Temelji se na Bayesovom teoremu aposteriorne vjerojatnosti. Klasifikator je dobio naziv naivni jer pretpostavlja uvjetnu nezavisnost varijabli, što u praksi uglavnom ne vrijedi. Unatoč toj naivnoj pretpostavci, klasifikator često daje dobre rezultate. U području klasifikacije teksta naivni Bayesov klasifikator daje jako dobre rezultate i često je prvi izbor u tome području. Implementacija naivnog Bayesovog klasifikatora u Weki temelji se na poglavlju iz knjige G. H. Johna i P. Langleya [9]. U nastavku je prikazan pseudokod naivnog Bayesovog algoritma za klasifikaciju teksta, korišten u radu C. D. Manning i sur. [10].

```
TrainMultinomialNB(C,D)
V = extractVocabulary(D)
N = countDocs(D)
for each c ϵ C
do Nc = countDocsInClass(D,c)
        prior[c] = Nc/N
        textc = concatenateTextOfAllDocsInClass(D,c)
        for each t ϵ V
        do Tct = countTokensOfTerm(textc,t)
```

```
        for each t ϵ V
        do condprob[t][c] = (Tct+1)/(SUM by ti(Tcti + 1))
return V,prior,condprob

ApplyMultinomialNB(C,V,prior,condprob,d)
W = extractTokensFromDoc(V,d)
for each c ϵ C
do score[c] = log prior[c]
        for each t ϵ V
        do score[c] += log condprob[t][c]
return argmax score[c]
```

3.3.5. Stroj s potpornim vektorima zasnovan na slijednoj minimalnoj optimizaciji

SMO (engl. *Sequential Minimal Optimization*) koristi heuristiku prilikom učenja stroja s potpornim vektorima (SVM, engl. *Support Vector Machines*) kako bi učenje, koje je problem kvadratnog programiranja, bilo efikasnije. Sami SVM trenutno je smatran jednim od najboljih klasifikacijskih algoritama. Radi na način da traži maksimalnu moguću marginu među klasama. Potpornim vektorima smatraju se točke najbliže plohi razdvajanja. Korištenjem jezgrenog trika, koji primjere diže u prostor više dimenzije bez da se pritom povećava računska složenost, SVM odlično radi i prilikom nelinearnih klasifikacijskih problema. Implementacija SMO klasifikatora u Weki temelji se na radovima J. Platt i sur. [11], S.S. Keerthi i sur. [12] i T. Hastie i sur. [13]. Pseudokod prikazan u nastavku preuzet je iz J. C. Plattovog rada [14].

```
target = desired output vector
point = training point matrix
procedure takeStep(i1,i2)
        if (i1 == i2) return 0
        alph1 = Lagrange multiplier for i1
        y1 = target[i1]
        E1 = SVM output on point[i1] – y1 (check in error cache)
        s = y1*y2
        Compute L, H via equations (13) and (14)
        if (L == H)
                return 0
        k11 = kernel(point[i1],point[i1])
        k12 = kernel(point[i1],point[i2])
        k22 = kernel(point[i2],point[i2])
        eta = k11+k22-2*k12
        if (eta > 0)
        {
```

```
                a2 = alph2 + y2*(E1-E2)/eta
                if (a2 < L) a2 = L
                else if (a2 > H) a2 = H
        }
        else
        {
                Lobj = objective function at a2=L
                Hobj = objective function at a2=H
                if (Lobj < Hobj-eps)
                        a2 = L
                else if (Lobj > Hobj+eps)
                        a2 = H
                else
                        a2 = alph2
        }
        if (|a2-alph2| < eps*(a2+alph2+eps))
                return 0
        a1 = alph1+s*(alph2-a2)
        Update threshold to reflect change in Lagrange multipliers
        Update weight vector to reflect change in a1 & a2, if SVM is linear
        Update error cache using new Lagrange multipliers
        Store a1 in the alpha array
        Store a2 in the alpha array
        return 1
endprocedure

procedure examineExample(i2)
        y2 = target[i2]
        alph2 = Lagrange multiplier for i2
        E2 = SVM output on point[i2] – y2 (check in error cache)
        r2 = E2*y2
        if ((r2 < -tol && alph2 < C) || (r2 > tol && alph2 > 0))
        {
                if (number of non-zero & non-C alpha > 1)
                {
                        i1 = result of second choice heuristic (section 2.2)
                        if takeStep(i1,i2)
                                return 1
                }
                loop over all non-zero and non-C alpha, starting at a random point
                {
                        i1 = identity of current alpha
                        if takeStep(i1,i2)
                                return 1
                }
                loop over all possible i1, starting at a random point
                {
                        i1 = loop variable
                        if (takeStep(i1,i2)
                                return 1
```

```
                }
        }
        return 0
endprocedure

main routine:
        numChanged = 0;
        examineAll = 1;
        while (numChanged > 0 | examineAll)
        {
                numChanged = 0;
                if (examineAll)
                        loop I over all training examples
                                numChanged += examineExample(I)
                else
                        loop I over examples where alpha is not 0 & not C
                                numChanged += examineExample(I)
                        if (examineAll == 1)
                                examineAll = 0
                        else if (numChanged == 0)
                                examineAll = 1
        }
```

---

4. Programsko rješenje

Programsko rješenje napisano za potrebe ovog rada u potpunosti je napisano u Pythonu 3. Kod je napisan u stilu funkcijskog programiranja, bez korištenja klasa. Cijelo programsko rješenje napravljeno je kao konzolna aplikacija koju kontroliramo uz pomoć argumenata iz komandne linije. Cjelokupan kod dostupan je na Githubu, zajedno s detaljnim uputama za korištenje [15].

Algoritme koje smo koristili i opisali u prethodnom poglavlju nismo implementirali, već smo koristili gotove implementacije iz Weke [3]. S obzirom da je Weka pisana u Javi, koristili smo biblioteku (engl. *library*) *python-weka-wrapper3* kako bismo pozivali Wekine klase kroz Python. Kao što i samo ime kaže, ta biblioteka je Pythonova „omotnica“, koja preko biblioteke *javabridge* omogućava korištenje Wekinih klasa (Jave) u Pythonu. Navedena biblioteka također je slobodno dostupna na Githubu [17]. Na Githubu su i detaljne upute za instalaciju i uporabu te biblioteke.

Navedena biblioteka sa sobom povlači ovisnost o drugim bibliotekama. Te biblioteke su *numpy* i *javabridge* te, opcionalno, *matplotlib*, *pygraphviz* i *PIL*. Uz to, zahtjev je i *Oracle JDK 1,8+.* Uz te ovisnosti, naša aplikacija dodatno je ovisna o biblioteci *pandas* koju smo koristili za jednostavniju manipulaciju podatcima, biblioteci *liac-arff* koja nam je poslužila za lakšu konverziju iz *pandasa* u arff i obratno, te o biblioteci *openpyxl* ukoliko se želi koristiti skripta za automatsko kreiranje statistika u Excelu. S obzirom na velik broj ovisnosti programskog rješenja, preporučamo korištenje *Linuxa* kao operacijskog sustava radi jednostavnijeg namještanja okoline.

Prilikom pisanja izvornog koda puno je pažnje posvećeno izgradnji dobre strukture koda. Cilj je bio kreirati modularan kod koji će biti jednostavan za održavanje i gdje će biti jednostavno dodavanje novih funkcionalnosti. Kod je grupiran u 10 datoteka kako bi se postigla smislena logička odvojenost koda i kako bi se omogućilo lakše održavanje te lakše buduće izmjene. Zbog relativno dobre strukture koda, dodavanje novog klasifikacijskog algoritma svodi se na pisanje automatskog parsera koji čita argumente algoritma iz komandne linije te dodatne tri linije koda koje služe kao „preusmjeravanje“ na željeni algoritam.

Uz već opisane algoritme u poglavlju 3.3., iz Weke smo koristili još i klase za učitavanje i spremanje podataka, evaluaciju algoritama i diskretizaciju podataka. Aplikacija podržava učitavanje i spremanje podataka iz/u csv i arff format. Podržani načini evaluacije su unakrsna provjera (engl. *cross validation*), podjela na skup za učenje i skup za testiranje (engl. *train-test split*), gdje kao parametar šaljemo i postotak podjele te testiranje na cijelome skupu podataka. Ukoliko korisnik ništa ne specificira, automatski će biti provedena evaluacija unakrsnom provjerom. Za diskretizaciju se koristi nenadzirani (engl. *unsupprovised*) diskretizacijski filter iz Weke.

Dio koda za rukovanje podacima napisan je kompletno u Pythonu. Weka radi najbolje s arff-formatom podataka koji je kreiran upravo za Weku. Kako smo manipulacije podatcima radili kroz biblioteku *pandas* radi jednostavnosti, morali smo napraviti i konvertere koji su format podataka arff direktno pretvarali u *pandas DataFrame*, i obratno.

S obzirom na vrstu obrade podataka koju smo željeli provesti, morali smo napisati program s dobrim i jednostavnim mogućnostima manipulacije podataka. Primjerice, prilikom učitavanja podataka programu se može zadati koje značajke da učita ili koje značajke da izbaci iz podataka. Također, mogu se zadati uvjeti izbacivanja redaka iz skupa podataka. Broj uvjeta koje možemo zadati je neograničen.

Ukoliko u podatcima imamo više klasa, a mi želimo promatrati samo odnos jedne naprama svih ostalih, to također možemo jednostavno ostvariti. To se ostvaruje na način da kao argument pošaljemo ime značajke koja je oznaka klase, nakon toga stavimo dvotočku, napišemo vrijednost klase koju želimo zadržati te nakon toga zarez i „rest“. Primjerice, „class_label:50+,rest“ će među oznakama klase u značajci koja se zove class_label ostaviti klasu „50+“, dok će sve ostale zamijeniti s „rest“. Kao oznaku klase programski možemo postaviti bilo koju značajku, ona ne mora biti unaprijed definirana i postavljena kao posljednja značajka u podatcima.

Ukoliko želimo računati podatke za predikcije, kao argumente šaljemo broj utakmica koje želimo uzeti u računanje prosjeka. U ovom slučaju postoje dvije mogućnosti. Prva je da kao argument *n* pošaljemo prirodni broj, primjerice 8. U tom slučaju ekipa mora odigrati minimalno osam utakmica prije trenutne kako bi prosjek postojao, dok se u utakmicama

nakon toga uvijek gleda prosjek samo zadnjih osam utakmica. Za razliku od toga, imamo drugi slučaj u kojemu kao parametar možemo poslati prirodni broj s sufiksom „+“, primjerice 8+. U tom slučaju ekipa mora odigrati barem osam utakmica prije trenutne da bi prosjek postojao, ali nakon toga se u svakoj sljedećoj utakmici gleda prosjek svih do tada odigranih utakmica.

S obzirom na vrstu analize koju smo provodili, koja se zasniva na puno pokretanja algoritama s različitim podatcima i parametrima te kasnijem uspoređivanju dobivenih rezultata, napravili smo i skriptu za automatsko pokretanje sa svim željenim permutacijama parametara te automatsko generiranje Excel tablice s dobivenim točnostima. Imena stupaca i redaka, odnosno željeni format te automatski generirane Excel tablice, također se zadaju kroz parametre.

5. Rezultati

5.1. Statistička analiza

Provedenom statističkom analizom svih pet sezona, utvrdili smo kako u većini sezona mnoge od promatranih statistika ima manju *p-vrijednost* od α. Iz toga zaključujemo kako mnoge od promatranih kategorija predstavljaju značajnu razliku između pobjedničkih i gubitničkih ekipa. U tablicama 1 i 2 prikazane su sve kategorije po sezonama koje **ne** predstavljaju značajnu razliku među pobjedničkim i gubitničkim ekipama.

Tablica 1. Kategorije koje ne predstavljaju značajnu razliku na razini utakmice

Sezona	Kategorije
2013./'14.	DFGA, FTAST, DIST, UFGA, TCHS
2014./'15.	DFGA, TCHS, FTAST, PASS
2015./'16.	DFGA, TCHS, FTAST, PASS, DIST, UFGA
2016./'17.	DFGA, TCHS, PASS, DIST, UFGA
2017./'18.	DFGA, TCHS, PASS, DIST, UFGA, FTAST

Tablica 2. Kategorije koje ne predstavljaju značajnu razliku na razini sezone

Sezona	Kategorije
2013./'14.	FTAST, TCHS, RBC, PASS, CFGM, DFGA
2014./'15.	DFGA, FTAST, CFGM, RBC, DFGM
2015./'16.	TCHS, PASS, FTAST, ORBC, CFGM, RBC
2016./'17.	CFGM, FTAST, UFGA, RBC, DFGA, DFGM, ORBC, DRBC, CFGA, DFG_PCT
2017./'18.	UFGA, DFGM, CFGA, ORBC, FTAST, DFGA, RBC, CFGM

Iz tablica 1 i 2 vidimo kako se mnoge kategorije ponavljaju i na razini utakmice i na razini sezone. Primjerice, asistencije za slobodna bacanja (FTAST) pojavljuju se u četiri sezone na razini utakmice i u svih pet sezona na razini sezone. Na moguće značenje nekih od navedenih vrijednosti osvrnut ćemo se i kasnije u radu.

S obzirom da nam većina kategorija predstavlja značajnu razliku, u nastavku rada ćemo se usredotočiti samo na one koje su se nama činile najznačajnijima. U tablici 3 prikazane su srednje vrijednosti, standardne devijacije i *p-vrijednosti* na razini utakmice. U tablici se

nalazi i nekoliko kategorija koje ne predstavljaju značajnu razliku. U tablicu su stavljene jer ćemo se morati osvrnuti na njih u nastavku rada.

Tablica 3. Najznačajniji rezultati statističke analize na razini utakmice

Sezona	Kategorija	Pobjednička ekipa		Gubitnička ekipa		p-vrijednost	
		Srednja vrijednost	Standardna devijacija	Srednja vrijednost	Standardna devijacija		
2013./'14.	FG_PCT	0,482141	0,05101	0,429235	0,047654	3,02E-121	
	UFG_PCT	0,470235	0,06978	0,415133	0,06815	5,18E-76	
	DRBC	64,58483	10,24363	57,05954	10,51416	2,84E-67	
	AST	23,62398	4,89995	20,38091	4,557399	2,29E-57	
	UFGM	21,26264	4,41199	18,51387	4,070828	1,12E-52	
	CFG_PCT	0,496513	0,085129	0,445307	0,080898	4,88E-47	
	SAST	3,174551	1,981681	2,50571	1,74881	4,07E-18	
	CFGA	36,98532	6,823712	38,95677	7,48488	6,95E-11	
	CFGM	18,28467	4,235482	17,30016	4,306951	2,00E-08	
	PASS	295,3785	32,87041	291,2463	32,6657	4,13E-04	
	UFGA	45,26591	6,956032	44,64682	6,856909	1,59E-02	**
2014./'15.	FG_PCT	0,474821	0,050474	0,425398	0,047825	2,51E-112	
	UFG_PCT	0,461685	0,072516	0,409547	0,069543	1,44E-67	
	UFGM	20,88346	4,440502	17,92176	3,941092	1,68E-60	
	AST	23,66341	4,920078	20,40261	4,452313	7,93E-59	
	DRBC	62,40098	9,689668	56,07987	9,407495	2,66E-56	
	CFG_PCT	0,490319	0,080244	0,44262	0,075886	1,56E-46	
	SAST	3,223309	1,968515	2,444173	1,617332	1,69E-24	
	CFGA	37,96333	6,978489	40,07661	7,019944	2,10E-13	
	UFGA	45,22494	6,617363	43,81581	6,741097	1,07E-07	
	CFGM	18,51915	4,128671	17,68623	4,056915	8,19E-07	
	PASS	298,4425	34,40571	295,7506	37,67602	1,87E-03	**
2015./'16.	FG_PCT	0,477902	0,049402	0,428773	0,04805	1,84E-112	
	UFG_PCT	0,464504	0,071598	0,413855	0,068834	1,87E-62	
	AST	23,82166	5,040176	20,74511	4,438957	3,36E-54	
	DRBC	62,98697	10,04431	56,81026	9,721165	4,57E-53	
	UFGM	21,60749	4,522227	18,92427	4,119047	1,74E-48	
	CFG_PCT	0,49495	0,08296	0,446546	0,077044	1,34E-44	
	SAST	3,06759	1,902983	2,407166	1,624591	5,45E-18	
	CFGM	18,56352	4,266016	17,36482	4,091512	9,04E-13	
	CFGA	37,68974	7,137307	39,03665	7,031617	3,99E-07	
	UFGA	46,57166	7,051718	45,79479	6,982472	2,32E-03	**
	PASS	300,6344	33,27629	297,2533	33,42783	2,51E-03	**
2016./'17.	FG_PCT	0,482894	0,049402	0,433431	0,04624	1,79E-114	
	UFG_PCT	0,46385	0,074922	0,408059	0,073669	3,24E-67	
	AST	24,27406	5,260241	20,97961	4,654644	2,73E-54	
	UFGM	18,52202	4,068185	16,04568	3,885025	1,32E-49	
	CFG_PCT	0,500278	0,07341	0,455419	0,070468	2,11E-49	
	DRBC	59,51387	9,045295	53,99429	9,619069	4,94E-48	

	SAST	5,90783	2,796877	4,898858	2,338927	3,83E-19	
	CFGM	22,44698	4,430972	20,98777	4,312944	2,03E-15	
	CFGA	45,05057	7,245934	46,20065	7,125616	1,18E-05	
	UFGA	39,98042	6,460399	39,36297	6,712248	9,42E-03	**
	PASS	297,8997	30,97637	300,1705	31,9745	6,68E-02	**
2017./'18.	FG_PCT	0,485124	0,049986	0,437258	0,048784	2,09E-103	
	UFG_PCT	0,468612	0,066428	0,417376	0,066861	1,29E-70	
	DRBC	61,32242	9,072868	55,36743	8,978363	5,59E-55	
	UFGM	23,35352	4,46272	20,54992	4,232828	4,07E-51	
	AST	24,8257	5,208186	21,68331	4,72698	3,78E-48	
	CFG_PCT	0,509808	0,08797	0,464349	0,082493	3,39E-35	
	SAST	3,154664	1,948512	2,655483	1,725171	7,30E-11	
	CFGM	18,11293	4,289524	17,07447	4,263844	1,39E-09	
	CFGA	35,73159	6,989284	36,86416	7,105929	1,33E-04	
	UFGA	49,89444	6,817702	49,27823	6,727613	1,13E-02	**
	PASS	298,8707	33,07879	301,3633	33,3872	3,60E-02	**

Bilješke: α=5,0E-4

** p-vrijednost je veća od α– kategorija ne predstavlja značajnu razliku

U tablici 3 vidimo kako postotak šuta (FG_PCT) predstavlja razliku s najmanjom *p-vrijednosti* u svim promatranim sezonama. To je na neki način i očekivano jer je za pobjedu potrebno postići više poena od suparnika, tako da ekipa koja to radi efikasno automatski si značajno povećava šanse za pobjedu. Značajnost postotka šuta potvrđena je i u radu [24].

Postotak šuta (FG_PCT) nije nova statistika, ali je zanimljivo vidjeti odnos s postotkom nebranjenih šutova (UFG_PCT) i branjenih šutova (CFG_PCT). Odnos među njima je u svim sezonama isti, postotak šuta ima najmanju *p-vrijednost*, nakon toga ide postotak šuta nebranjenih udaraca sa osjetno većom *p-vrijednosti* i na kraju je postotak šuta branjenih udaraca sa najvećom *p-vrijednošću.* S obzirom na to, čini se da postotak nebranjenog šuta predstavlja bitniju razliku među ekipama na razini utakmice. Treba naglasiti kako takva tvrdnja ne može biti direktno potvrđena ovakvom statističkom analizom, ali mi ćemo je pokušati potvrditi u nastavku rada s ostalim provedenim analizama.

Pogledamo li vrijednosti u tablici 3 za broj zabijenih i pokušanih, branjenih i nebranjenih šutova, ponovno je odnos sličan. U svim sezonama broj zabijenih nebranjenih šutova (UFGM u odnosu na CFGM) ima manju *p-vrijednost* te se čini da predstavlja značajniju

razliku među pobjedničkim i gubitničkim ekipama. Ponovno, to se ne može tvrditi direktno iz ovakve statističke analize, ali ćemo mi pokušati potvrditi ostalim analizama.

Na prvu bi se moglo tvrditi kako je razlog tome taj što će bolje ekipe jednostavno stvoriti više prilika za zabiti nebranjeni šut. Pogledamo li broj pokušanih nebranjenih šutova (UFGA) kroz sezone, vidimo kako je njihova značajnost mala, u četiri od pet promatranih sezona se ne može smatrati značajnom kategorijom. Razlika u broju pokušanih nebranjenih šutova u svim sezonama je manja od jednog šuta po utakmici u prosjeku. Što se tiče broja pokušanih branjenih šutova (CFGA), njihova značajnost je nešto veća, ali je i dalje mala. Razlika u broju branjenih šutova po utakmici u prosjeku ide do maksimalno 2 šuta u svih pet sezona.

Prije ovako provedene analize mogle bi se pretpostaviti dvije stvari kao razlika među pobjedničkim i gubitničkim ekipama. Prva je kako će i jedni i drugi zabiti svoje lake, nebranjene šutove, dok će razliku činiti veći broj zabijenih teških, branjenih šutova. Druga je kako će bolje ekipe stvoriti više nebranjenih prilika te samim time i zabiti više nebranjenih šutova te zbog toga pobijediti. Pogledom na same prosječne vrijednosti navedenih kategorija već vidimo kako nijedna tvrdnja nije točna. Mali iznosi *p-vrijednosti* dodatni su razlog da sumnjamo u ispravnost te dvije tvrdnje. Najveću razliku čini broj zabijenih nebranjenih šutova iz sličnog broja pokušaja (što automatski znači i bolji postotak nebranjenih šutova UFG_PCT). Do istog zaključka došli smo u našem prethodnom radu koji se odnosio samo na sezonu 2016./'17. [16]. Ovakva značajnost nebranjenog postotka šuta u skladu je s rezultatima koje su dobili G. Csátaljay i sur [27]. Uz važnost nebranjenog postotka šuta, oni su još utvrdili i veću sposobnost pobjedničkih ekipa da zabijaju branjene šutove. Potvrda drugog njihovog zaključka vidi se i u tablicama 3 i 4 kroz značajnost razlike među razdiobama postotka branjenog šuta među dobrim i lošim ekipama.

Ako pogledamo trendove kroz sezone, vidimo kako su postotci relativno konstantni tijekom svih pet sezona, oko 0,495 za postotak branjenog šuta (CFG_PCT) te oko 0,465 za postotak nebranjenog šuta (UFG_PCT). Gledajući same brojke šutova, vidimo kako su u prve tri sezone relativno konstantne, dok u sezoni 2016./'17. broj pokušanih nebranjenih šutova (UFGA) padne za oko sedam šutova po utakmici u prosjeku, a broj pokušanih branjenih šutova (CFGA) poraste za oko sedam šutova. Nakon toga, u sezoni 2017./'18.

događa se ponovni veliki skok, ovoga puta broj pokušanih nebranjenih šutova poraste za oko deset šutova u prosjeku, dok broj pokušanih branjenih padne za oko deset šutova u prosjeku. Ako gledamo zbroj svih pokušaja šutova, vidimo kako se iz sezone u sezonu broj šutova blago povećava, s prosjeka od 82,1 u sezoni 2013./'14. došao je do prosjeka od 85,5 u sezoni 2017./'18. Ovo je pokazatelj toga da se igra kroz godine ubrzava.

U tablici 3 vidimo i kako prilike za hvatanje obrambenih skokova (DRBC) predstavljaju značajnu razliku među pobjedničkim i gubitničkim ekipama. Kako prilika za obrambeni skok nastaje na način da protivnička ekipa mora promašiti svoj šut, jasno je kako je ova kategorija usko povezana s postotkom šuta protivničke ekipe. Ova kategorija, i njezina značajnost, samo su još jedan pokazatelj važnosti postotka šuta ekipe (FG_PCT). Uz to, ovo možemo povezati i sa zaključkom G. Csataljay i sur. [24 - 25], koji su u svojim radovima pokazali izrazitu važnost samog broja obrambenih skokova.

Od ostalih kategorija iz tablice 3 vidimo kako veliku razliku čine asistencije (AST) i sekundarne asistencije (SAST). Za asistencije je to u neku ruku i logično jer one u velikoj većini slučajeva znače lake poene za ekipu. Sekundarna asistencija je po definiciji dodijeljena igraču ako je dodao loptu drugom igraču koji je upisao asistenciju unutar jedne sekunde i bez driblanja. Ako razmišljamo u smislu same igre, sekundarne asistencije najlakše je dobiti u situacijama kada je netko iz obrane „ispao“ iz obrambene sheme te je manjak igrača u obrani, a napadačka momčad samo prosljeđuje loptu dok se ne dođe do igrača koji je slobodan. To znači kako sekundarne asistencije isto u većini slučajeva donose lake poene.

U tablici 4 prikazane su najznačajnije razlike između pobjedničkih i gubitničkih ekipa na razini sezone.

Ako promatramo najznačajnije razlike na razini sezone (tablica 4), vidimo kako je postotak šuta (FG_PCT) i dalje među njima, ali ipak osjetno manje značajan nego na razini ekipe, odnosno ima osjetno veću *p-vrijednost*. Većina drugih značajki također imaju dosta veće *p-vrijednosti* nego na razini utakmice. Vidimo kako je postotak nebranjenog šuta (UFG_PCT) u tri od pet sezona ima manju *p-vrijednost* od postotka branjenog šuta (CFG_PCT). To je također razlika u odnosu na prethodnu podjelu gdje je u svim sezonama postotak nebranjenog šuta imao manju *p-vrijednost*.

Tablica 4. Najznačajniji rezultati na razini sezone

Sezona	Kategorija	Pobjednička ekipa		Gubitnička ekipa		p-vrijednost	
		Srednja vrijednost	Standardna devijacija	Srednja vrijednost	Standardna devijacija		
2013./'14.	FG_PCT	0,471245	0,05561	0,444789	0,054206	7,60E-21	
	SAST	3,299145	2,061911	2,400978	1,692549	5,19E-20	
	UFGM	20,89622	4,336429	18,95966	4,308721	1,70E-19	
	CFGA	36,72772	6,836931	39,80685	7,344927	5,61E-17	
	UFG_PCT	0,459205	0,071451	0,431082	0,074089	9,25E-15	
	AST	22,76923	5,032581	21,23472	4,772542	8,46E-11	
	CFG_PCT	0,486812	0,08925	0,459166	0,084867	3,37E-10	
	UFGA	45,60073	6,914642	44,02078	6,834881	5,87E-07	
	DRBC	62,45543	11,17679	59,72249	10,99015	9,04E-07	
	CFGM	17,80098	4,280587	18,21638	4,464288	4,49E-02	**
	PASS	292,1551	34,30767	290,9939	30,91016	1,19E-01	**
2014./'15.	UFGM	20,55	4,549658	18,20245	4,170652	4,38E-25	
	SAST	3,382222	1,967058	2,449728	1,683913	5,29E-24	
	UFGA	46,14667	6,845558	43,02853	6,46565	1,25E-21	
	CFGA	37,60556	7,060282	40,81794	6,84499	1,31E-19	
	FG_PCT	0,460566	0,053605	0,434586	0,053559	2,22E-19	
	PASS	302,7256	31,56238	291,4497	39,7072	1,06E-17	
	AST	23,20778	5,209317	20,98234	4,612431	1,15E-17	
	CFG_PCT	0,47942	0,082642	0,445708	0,079295	7,32E-16	
	DRBC	60,48778	10,16662	57,65082	10,0884	2,31E-09	
	UFG_PCT	0,445038	0,073377	0,4233	0,076466	7,39E-08	
	CFGM	17,93667	4,162093	18,14674	4,147969	1,05E-01	**
2015./'16.	UFGM	21,78208	4,870503	19,17216	4,381379	5,89E-20	
	CFGA	36,83096	7,550776	40,13797	6,596418	4,21E-16	
	FG_PCT	0,471919	0,053934	0,446663	0,055471	5,87E-16	
	DRBC	61,89817	10,27111	58,14286	10,13251	3,07E-12	
	UFGA	47,1833	7,123331	44,24542	7,120828	1,76E-11	
	UFG_PCT	0,461177	0,07286	0,43427	0,076766	2,10E-10	
	CFG_PCT	0,485633	0,084954	0,459647	0,082582	4,91E-08	
	AST	23,45418	5,574307	21,89255	4,578645	4,42E-07	
	SAST	3,069246	2,030203	2,521368	1,640851	5,99E-06	
	CFGM	17,78004	4,357551	18,42735	4,376313	6,04E-03	**
	PASS	296,6884	33,9304	295,5201	33,10952	9,54E-02	**
2016./'17.	FG_PCT	0,471077	0,05464	0,450077	0,05021	1,99E-14	
	UFG_PCT	0,454058	0,080014	0,425085	0,075547	5,11E-13	
	UFGM	18,36005	4,332693	17,04044	3,934068	1,47E-09	
	AST	23,61413	5,880952	21,94485	4,673187	2,29E-09	
	PASS	296,7609	33,35943	307,5515	31,49104	7,51E-09	
	SAST	5,888587	2,970802	5,093137	2,422016	3,90E-07	
	CFG_PCT	0,48775	0,075406	0,473502	0,072833	2,99E-05	
	CFGA	44,17663	7,033355	45,32353	7,294051	2,30E-03	**
	DRBC	56,71332	9,052418	55,84927	9,770901	1,83E-02	**

	UFGA	40,47147	6,734963	40,19485	6,74023	3,00E-01	**
	CFGM	21,44158	4,23357	21,38235	4,329483	3,99E-01	**
2017./'18.	PASS	295,4614	39,09476	311,6294	29,78296	2,02E-17	
	FG_PCT	0,470181	0,055582	0,447364	0,051405	1,91E-13	
	CFG_PCT	0,497339	0,090731	0,466786	0,084373	1,63E-09	
	DRBC	60,30702	9,43531	57,38011	9,672887	1,42E-07	
	UFG_PCT	0,452293	0,072821	0,434229	0,068697	8,82E-06	
	UFGM	22,61404	4,658663	21,6485	4,323752	5,90E-05	
	SAST	3,135088	2,008155	2,726158	1,721368	3,10E-04	
	AST	23,9579	5,828689	22,74387	4,638919	3,68E-04	
	CFGM	17,54737	4,138044	16,79973	4,256054	1,07E-03	**
	CFGA	35,55965	6,906159	36,12943	7,277833	0,123767	**
	UFGA	50,03158	6,862116	49,94687	6,80772	0,48967	**

Bilješke: α=5,0E-4

** p-vrijednost je veća od α– kategorija ne predstavlja značajnu razliku

Promatrajući same brojke branjenih i nebranjenih šutova, vidimo kako *p-vrijednost* broja zabijenih branjenih šutova (CFGM) u svih pet sezona ne prelazi zadani prag sigurnosti α. Broj pokušanih branjenih šutova (CFGA) značajan je u sezonama 2013./'14. do 2015./'16., dok u sezonama 2016./'17. i 2017./'18. nije značajan. Zanimljivo je primijetiti kako je prosjek nebranjenih pokušaja šuta (UFGA) veći u svih pet sezona, odnosno više šutova spada u nebranjene šutove, te kako je značajan u prve tri sezone, a nije u zadnje dvije. Isto tako, i broj pokušaja branjenih šutova predstavlja značajnu razliku u prve tri sezone, a ne predstavlja u zadnje dvije. Broj zabijenih nebranjenih šutova (UFGM) ponovno predstavlja značajnu razliku među pobjedničkim i gubitničkim ekipama.

Pogledamo li broj dodavanja (PASS) u tablici 4, vidimo kako se trend okrenuo, od toga da pobjedničke ekipe imaju više ili približno jednako dodavanja kroz prve tri sezone, do toga da pobjedničke ekipe imaju osjetno manje dodavanja u prosjeku. Razlog tomu može biti slučajnost jer broj dodavanja pobjedničkih ekipa uvelike ovisi o samom stilu igre pobjedničkih ekipa u pojedinoj sezoni, ali može biti i pokazatelj novog trenda u ligi i pokazatelj „novog načina“ kako biti pobjednička momčad. Svakako bi tu tvrdnju trebalo provjeriti i kroz sljedećih nekoliko sezona. Broj asistencija (AST) i sekundarnih asistencija (SAST) je ponovno značajan kroz svih pet sezona.

5.1.1. Nova statistička kategorija - postotak efektivnih dodavanja EPR

U našem prethodnom radu [16], u kojemu smo proučavali samo sezonu 2016./'17., predložili smo novu statističku kategoriju koju smo temeljili na značajnosti asistencija (AST), sekundarnih asistencija (SAST) i dodavanja (PASS) u navedenoj sezoni. Formula za navedenu novu statističku kategoriju koja je nazvana postotak efektivnih dodavanja (engl. *Effective Passing Ratio*, EPR) je:

$$EPR = \frac{AST + SAST + FTAST}{PASS} \quad (2)$$

Uz već navedene kategorije, u formulu smo dodali i asistencije za slobodna bacanja (FTAST). Iako ne predstavljaju značajnu razliku, ipak su i one dodavanja pa smo ih zbog toga uvrstili u formulu. Proučavajući samo sezonu 2016./'17., postotak efektivnih dodavanja (EPR) pokazao se kao dobra statistička kategorija [16].

Dodatna potvrda smislenosti ovakve kategorije može se pronaći u jednom od zaključaka koje su iznijeli G. Csataljay i sur. [25], gdje autori navode: „Veći broj izgubljenih lopti tijekom neizvjesnih četvrtina mogu spriječiti pobjedničke ekipe u povećanju vodstva. S druge strane, veće vodstvo postignuto je kada su pobjedničke ekipe igrale više kao ekipa (engl. *more collectively*) i kada su si igrači međusobno asistirali s uspješnijim dodavanjima."

U ovome radu pokušat ćemo provjeriti vrijedi li isto i na većem skupu podataka, odnosno na pet sezona koje proučavamo. Vrijednosti značajnosti razlike postotka efektivnih dodavanja (EPR-a) među pobjedničkim i gubitničkim ekipama prikazane su u tablicama 5 i 6.

Tablica 5. Značajnost EPR-a na razini utakmice

Sezona	Kategorija	Pobjednička ekipa		Gubitnička ekipa		p-vrijednost
		Srednja vrijednost	Standardna devijacija	Srednja vrijednost	Standardna devijacija	
2013./'14,	EPR	0,096549	0,020122	0,084074	0,018755	1,21E-50
2014./'15.	EPR	0,096003	0,020123	0,082919	0,018937	1,65E-55
2015./'16.	EPR	0,095297	0,020792	0,083575	0,018344	1,64E-45
2016./'17.	EPR	0,109173	0,023968	0,09325	0,021157	3,38E-63
2017./'18.	EPR	0,09759	0,020185	0,084762	0,018548	3,37E-53
Bilješka:		α=5,0E-4				

Tablica 6. Značajnost EPR-a na razini sezone

Sezona	Kategorija	Pobjednička ekipa		Gubitnička ekipa		p-vrijednost
		Srednja vrijednost	Standardna devijacija	Srednja vrijednost	Standardna devijacija	
2013./'14.	EPR	0,094966	0,021185	0,086855	0,01907	2,38E-15
2014./'15.	EPR	0,093027	0,019787	0,086651	0,021232	3,36E-11
2015./'16.	EPR	0,094669	0,022299	0,088694	0,019895	2,39E-07
2016./'17.	EPR	0,106989	0,027418	0,094896	0,020208	4,73E-20
2017./'18.	EPR	0,09549	0,021196	0,085324	0,017788	3,50E-18
Bilješka:		α=5,0E-4				

S obzirom na visoku koreliranost postotka efektivnih dodavanja (EPR) s asistencijama (AST), koja iznosi oko 0,7, usporedit ćemo *p-vrijednost* EPR-a s *p-vrijednošću* asistencija. Vidimo kako u prve tri sezone (izuzev sezone 2014./'15. u tablici 4) asistencije imaju manji iznos *p-vrijednosti*. U sezonama 2016./'17. i 2017./'18. postotak efektivnih dodavanja ima značajno manju *p-vrijednost* u odnosu na asistencije, posebice na razini sezone.

Vjerojatni razlog takve promjene ranije je opisana promjena trenda u broju dodavanja (PASS) jer se pokazalo kako bolje ekipe na razini sezone, u zadnje dvije sezone, imaju značajno manje dodavanja. Temeljem trenutnog trenda u statistikama, čini se kako bi u narednim sezonama EPR mogao predstavljati sve značajniju statističku kategoriju, ali to će svakako trebati provjeriti u narednim sezonama.

5.2. Dubinska analiza podataka

5.2.1. Analiza algoritmom Ripper na stvarnim podatcima

U tablici 7 prikazana su pravila dobivena algoritmom Ripper na skupu podataka s statistikama praćenja igrača.

Postotak šuta (FG_PCT) izbačen je iz podataka u tablici 7 jer, iako spada i u statistike praćenja igrača, on je zapravo osnovna statistika koja se mjeri već dugi niz godina. Iz generiranih pravila vidimo kako se branjeni (CFG_PCT) ili nebranjeni postotak šuta (UFG_PCT) spominje u svakom pravilu barem jednom. To je djelomično i očekivano s obzirom sve dosadašnje rezultate i to da je postizanje poena najbitnija stvar za pobjedu.

Iako su obje vrste postotka šuta zastupljene, vidimo kako se ipak češće pojavljuje postotak nebranjenih šutova (UFG_PCT). Pogledamo li primjerice sezonu 2017./'18., koja ima kratka i jasna pravila, vidimo kako ekipa s postotkom nebranjenog šuta većim od 0,453, dok protivnički ekipa ima niži od 0,456, pobjeđuje u 610 od 723 utakmica. Drugo pravilo je isto, s malo spuštenim granicama za oba postotka te ponovno s visokom vjerojatnošću u pobjedu ekipe koja ima viši postotak nebranjenog šuta. Slična pravila postoje i u ostalim sezonama. Ovakvi rezultati dodatna su potvrda za tvrdnje iz poglavlja 5.1 o važnosti branjenog (CFG_PCT) i nebranjenog postotka šuta, s naglaskom na to da je ipak nebranjeni postotak šuta važniji.

Vidimo kako se od ostalih statistika spominju asistencije (AST), prilike za obrambene skokove (DRBC) i postotak zabijenih šutova protivnika kada je branjen u blizini obruča (DFG_PCT). Važnost asistencija i prilika za obrambeni skok utvrdili smo i prilikom statističke analize.

Tablica 7. Pravila dobivena algoritmom Ripper nad statistikama praćenja igrača

Sezona	Pravila		Točnost
2013./'14.	(UFG_PCT >= 0,456) and (DRBC >= 60) and (CFG_PCT >= 0,471)	WINNER=1 (303,0/24,0)	71,73%
	(UFG_PCT_OPP <= 0,424) and (UFG_PCT >= 0,436) and (CFG_PCT_OPP <= 0,517)	WINNER=1 (244,0/35,0)	
	(UFG_PCT_OPP <= 0,46) and (DRBC_OPP <= 60)	WINNER=1 (388,0/132,0)	
	(CFG_PCT_OPP <= 0,484) and (AST >= 23)	WINNER=1 (311,0/128,0)	
		WINNER=0 (1206,0/299,0)	
2014./'15.	UFG_PCT_OPP <= 0,425) and (DRBC_OPP <= 58) and (AST >= 21) and (CFG_PCT_OPP <= 0,51) and (UFGM >= 19)	WINNER=1 (213,0/8,0)	70,57%
	(UFG_PCT_OPP <= 0,45) and (UFG_PCT >= 0,426) and (AST >= 23) and (CFG_PCT_OPP <= 0,559)	WINNER=1 (245,0/39,0)	
	(AST_OPP <= 20) and (AST >= 20) and (CFG_PCT_OPP <= 0,469)	WINNER=1 (261,0/61,0)	
	(UFG_PCT >= 0,451) and (CFG_PCT >= 0,487)	WINNER=1 (210,0/59,0)	
	(UFG_PCT_OPP <= 0,396)	WINNER=1 (355,0/173,0)	
		WINNER=0 (1170,0/283,0)	
2015./'16.	(UFG_PCT <= 0,443) and (UFG_PCT_OPP >= 0,417) and (CFG_PCT <= 0,519)	WINNER=0 (548,0/85,0)	72,52%
	(CFG_PCT_OPP >= 0,514) and (AST <= 23)	WINNER=0 (324,0/80,0)	
	(AST_OPP >= 23) and (UFG_PCT <= 0,476)	WINNER=0 (337,0/139,0)	
		WINNER=1 (1247,0/323,0)	

2016./'17.	(UFG_PCT >= 0,423) and (UFG_PCT_OPP <= 0,449) and (CFG_PCT >= 0,49)	WINNER=1 (344,0/33,0)	72,47%
	(AST_OPP <= 22) and (UFG_PCT >= 0,451)	WINNER=1 (352,0/85,0)	
	(CFG_PCT >= 0,456) and (CFG_PCT_OPP <= 0,467) and (UFG_PCT >= 0,38)	WINNER=1 (242,0/62,0)	
	(UFG_PCT_OPP <= 0,419) and (CFG_PCT >= 0,462)	WINNER=1 (268,0/111,0)	
		WINNER=0 (1246,0/311,0)	
2017./'18.	(UFG_PCT >= 0,453) and (UFG_PCT_OPP <= 0,456)	WINNER=1 (610,0/113,0)	70,54%
	(UFG_PCT_OPP <= 0,415) and (UFG_PCT >= 0,429)	WINNER=1 (125,0/25,0)	
	(CFG_PCT >= 0,529) and (DFG_PCT <= 0,679)	WINNER=1 (326,0/112,0)	
		WINNER=0 (1383,0/411,0)	

Bilješka: Podatci ne uključuju nove statističke kategorije (AFG_PCT I EPR) kao ni postotak šuta (FG_PCT)

Kako su postotci šuta, odnosno što efikasnije postizanje poena, očekivano bitne statistike, pokušali smo pokrenuti algoritam Ripper i bez postotka branjenog i nebranjenog šuta. Cilj ovoga nam je da vidimo koje od ostalih statističkih kategorija su bitne te hoće li i koliko pasti točnost algoritma. Radi uštede prostora, ovoga puta nećemo prikazati cijela pravila, već samo broj pojavljivanja pojedinih statističkih kategorija. Rezultati su prikazani u tablici 8.

Tablica 8. Broj pojavljivanja pojedinih kategorija u pravilima algoritma Ripper

Sezona	Broj pojavljivanja kategorija	Broj pravila	Točnost
2013./'14.	UFGM - 2, AST - 2 and DRBC – 2	3	72,15%
2014./'15.	UFGM - 6, DRBC - 6, AST - 3 and DFG_PCT - 1	6	69,89%
2015./'16.	UFGM - 4, AST - 4, DRBC - 4 and CFGM - 1	5	69,67%
2016./'17.	UFGM - 3, AST - 3 and DRBC – 3	4	71,61%
2017./'18.	DRBC - 5, AST - 3, UFGM - 2 and DFG_PCT - 2	4	70,09%

Bilješke: Statističke kategorije protivničkih ekipa nisu zasebno razdvajane. Primjerice, AST_OPP je ubrojen pod AST
Postotci šuta i nove kategorije izbačeni su iz podataka

U rezultatima tablice 8 vidimo kako se najčešće pojavljuju broj zabijenih nebranjenih šutova (UFGM), broj asistencija (AST) i broj prilika za obrambeni skok (DRBC). Sve tri kategorije okarakterizirane su kao bitna razlika i u statističkoj analizi. Ako pogledamo razlike u točnosti između tablica 7 i 8, vidimo kako je u tablici 8 došlo do blagog pada točnosti, ali je ona i dalje relativno visoka. Broj nebranjenih šutova se pojavljuje puno

češće od broja branjenog šuta što ponovno ukazuje na to da ima veću važnost pri određivanju razlike između pobjedničkih i gubitničkih ekipa.

U tablici 9 prikazani su isti rezultati, no ovoga puta s dodanim EPR statistikama u podatke.

Tablica 9. Broj pojavljivanja pojedinih kategorija u pravilima algoritma Ripper s dodanim EPR-om u podatke

Sezona	Broj pojavljivanja kategorija	Broj pravila	Točnost
2013./'14.	DRBC - 4, EPR - 2, AST - 2, UFGM - 1 and DFG_PCT - 1	4	71,45%
2014./'15.	EPR - 4, DRBC - 3, UFGM - 2 and AST - 1	4	70,04%
2015./'16.	DRBC - 4, AST - 3, UFGM - 1, CFGM - 1 and DFG_PCT - 1	4	70,27%
2016./'17.	EPR - 5, AST - 3, UFGM - 2, DRBC - 2, CFGM - 1 and DFG_PCT - 1	5	70,76%
2017./'18.	DRBC - 6, EPR - 4, UFGM - 4, CFGM - 1 and DFG_PCT - 1	5	70,09%

Bilješke: Statističke kategorije protivničkih ekipa nisu zasebno razdvajane. Primjerice, AST_OPP je ubrojen pod AST
Postotci šuta

Vidimo kako se postotak efektivnih dodavanja (EPR) nalazi u značajnom broju pravila kroz sve sezone, izuzev sezone 2015./'16. gdje se ne pojavljuje. Iako su se sama pravila promijenila, vidimo kako su točnosti ostale slične. Ako u podatke ponovno vratimo postotke šuta i provedemo istu analizu, postotci šuta će ponovno biti glavne diskriminirajuće varijable, ali će se postotak efektivnih dodavanja svejedno pojaviti barem jednom u pravilima za svaku sezonu, izuzev 2015./'16. Ovakvi rezultati ponovno sugeriraju kako EPR predstavlja mjeru koja čini bitnu razliku među pobjedničkim i gubitničkim ekipama. Postotak efektivnih dodavanja pojavljuje se češće od asistencija u pravilima u četiri od pet promatranih sezona. To nas navodi na zaključak kako prema ovoj analizi postotak efektivnih dodavanja predstavlja bolju mjeru od asistencija.

Ukoliko pokušamo dobiti pravila algoritmom Ripper na razini sezone, jedino za sezonu 2013./'14. dobijemo smisleno pravilo. U toj sezoni pravilo je da, ukoliko ekipa ima manje od 18 nebranjenih zabijenih šutova (UFGM) po utakmici te njen protivnik ima više od 23 asistencije (AST), ekipa je gubitnička, odnosno ima na kraju sezone između 0 i 35 pobjeda. Ovo pravilo točno klasificira u 442 slučaja od njih ukupno 639. Za sve ostale utakmice u toj sezoni algoritam predviđa da je ekipa pobjednička, odnosno da ima 50 ili više pobjeda u sezoni.

Prilikom statističke analize na razini sezone (tablica 4) utvrdili smo značajnost razlika nekih kategorija. Sada vidimo kako nam algoritam Ripper ne pronalazi značajna pravila na razini

sezone. Usporedimo li *p-vrijednosti* na razini sezone (tablica 4) s onima na razini utakmice (tablica 3), vidimo kako su one dosta veće. Razlikuju se čak i do reda veličine 10^{60}. To što su značajnosti razlika među pobjednicima i gubitnicima toliko niže na razini sezone mogao bi biti razlog zašto s algoritmom Ripper na razini sezone ne dobivamo smislena pravila.

5.2.2. Analiza algoritmom Apriori na stvarnim podatcima

Kao što smo već naveli u poglavlju 3.3, algoritam Apriori radi isključivo s diskretnim vrijednostima. Zbog toga je provedena diskretizacija nad podatcima na način da se raspon vrijednosti podijeli na jednake intervale. Teško je utvrditi je li takva diskretizacija ispravna za ovakav skup podataka ili postoji bolji način. Mi ćemo proučiti dobivena pravila i vidjeti što nam govore, bez obzira na problem ispravnosti diskretizacije.

U tablici 10 prikazana su za nas najzanimljivija dobivena pravila.

Tablica 10. Najznačajnija pravila dobivena algoritmom Apriori

Sezona	Pravila	Pouzdanost
2013./'14.	FG_PCT='(0,5106-0,5507]' ==> WINNER=1	0,84
	AFG_PCT='(0,579505-0,625089]' ==> WINNER=1	0,79
	UFG_PCT='(0,4982-0,5499]' ==> WINNER=1	0,73
	UFGM='(13,2-16,8]' ==> WINNER=0	0,68
	DRBC='(66-74,6]' ==> WINNER=1	0,67
	EPR='(0,058916-0,074388]' ==> WINNER=0	0,67
	UFG_PCT='(0,3431-0,3948]' ==> WINNER=0	0,67
	AST='(24,5-28]' ==> WINNER=1	0,67
2014./'15.	AFG_PCT='(0,410905-0,457607]' ==> WINNER=0	0,81
	FG_PCT='(0,363-0,401]' ==> WINNER=0	0,78
	EPR='(0,101155-0,11477]' ==> WINNER=1	0,66
	AST='(24-27,6]' ==> WINNER=1	0,66
	EPR='(0,060309-0,073925]' ==> WINNER=0	0,65
	UFG_PCT='(0,4728-0,5266]' ==> WINNER=1	0,65
	DRBC='(47,7-54,6]' ==> WINNER=0	0,64
	UFGM='(14,3-17,4]' ==> WINNER=0	0,63
2015./'16.	AFG_PCT='(0,414561-0,46455]' ==> WINNER=0	0,82
	FG_PCT='(0,4775-0,5188]' ==> WINNER=1	0,72
	. AST='(25-28,2]' ==> WINNER=1	0,7
	DRBC='(44-53]' ==> WINNER=0	0,68
	EPR='(0,103926-0,118828]' ==> WINNER=1	0,68
	EPR='(0,059217-0,07412]' ==> WINNER=0	0,67
	DRBC='(62-71]' ==> WINNER=1	0,65
	UFG_PCT='(0,3621-0,4108]' ==> WINNER=0	0,64

	UFGM='(15-18]' ==> WINNER=0	0,61
2016./’17.	AST='(19-23]' WINNER=1 ==> EPR='(0,084294-0,112973]'	0,78
	FG_PCT='(0,4992-0,5344]' ==> WINNER=1	0,75
	AST='(19-23]' ==> EPR='(0,084294-0,112973]'	0,75
	EPR='(0,055615-0,084294]' ==> WINNER=0	0,69
	EPR='(0,112973-0,141652]' ==> WINNER=1	0,68
	AFG_PCT='(0,535988-0,579077]' ==> WINNER=1	0,66
2017./’18.	UFG_PCT='(0,3406-0,3908]' ==> WINNER=0	0,73
	UFG_PCT='(0,4912-0,5414]' ==> WINNER=1	0,7
	AFG_PCT='(0.567354-0.613645]' ==> WINNER=1	0,69
	DRBC='(43,6-50,8]' ==> WINNER=0	0,69
	EPR='(0,064127-0,07873]' ==> WINNER=0	0,66
	AFG_PCT='(0.474771-0.521062]' 576 ==> WINNER=0	0,65
	FG_PCT='(0,4745-0,513]' ==> WINNER=1	0,64
	UFGM='(16,6-19,8]' ==> WINNER=0	0,6

Bilješke: AFG_PCT će u jednom od kasnijih poglavlja biti predstavljen i objašnjen.
Prag za pouzdanost (engl. *confidence*) je postavljen na 0,6

Iako algoritam Apriori određuje pravila među bilo kojim varijablama, a ne nužno samo u odnosu na ciljanu varijablu (oznaku klase), u našem slučaju nismo pronašli takva pravila koja bi nam otkrivala nešto novo. Takva pravila su uglavnom govorila o vezi postotka efektivnih dodavanja (EPR) s asistencijama (AST) i o vezi prilagođenog postotka šuta (AFG_PCT) s običnim postotkom šuta (FG_PCT). Iz samih formula su ta pravila očita i samo govore o relativno visokoj koreliranosti tih kategorija.

Nedostatak takvih pravila koja se ne odnose na ciljnu varijablu je očekivan i zbog toga što korelacije među varijablama nisu prikazale nikakve bitne ovisnosti. Jedine jake korelacije u podatcima bile su one za koje se to i očekivalo, kao što su recimo broj dodavanja (PASS) i broj dodira s loptom (TCHS).

Pogledamo li rezultate kroz sezone, vidimo kako u svakoj sezoni postoji barem jedno pravilo povezano s postotkom efektivnih dodavanja (EPR). U nekim sezonama odnosi se na to da visoki postotak efektivnih dodavanja za sobom povlači pobjedu, u nekima da niski postotak efektivnih dodavanja povlači poraz, dok u tri sezone vidimo oba slučaja među značajnim pravilima. Asistencije (AST) se pojavljuju u četiri od pet sezona, i to uvijek samo jedno pravilo. U dvije sezone imaju nešto veću pouzdanost (engl. *confidence*) od postotka efektivnih dodavanja, dok u dvije sezone imaju istu pouzdanost. Iako je korelacija među asistencijama i postotkom efektivnih dodavanja relativno visoka (oko

0,7), temeljem ovih pravila se čini kako postotak efektivnih dodavanja ipak predstavlja nešto bolju mjeru od asistencija.

Pogledamo li ostala pravila, vidimo kako se često pojavljuje postotak nebranjenih šutova (UFG_PCT) i broj zabijenih nebranjenih šutova (UFGM). Značaj ovih varijabli smo utvrdili i u prethodnim analizama. Razina pouzdanosti pravila je viša za postotak zabijenih nebranjenih šutova.

Od ostalih pravila ponovno možemo utvrditi da broj prilika za obrambeni skok (DRBC) predstavlja bitnu značajku. Kao što smo već ranije spomenuli, postoji velika mogućnost da je broj prilika za obrambeni skok izravno povezan s protivničkim postotkom šuta tako da značajnost ove kategorije ne treba čuditi.

5.2.3. Analiza korištenjem klasifikatora na stvarnim podatcima

U ovom ćemo dijelu rada uz pomoć četiri opisana klasifikatora pokušati identificirati statističke kategorije koje u sebi sadrže „puno informacije“ o pobjedniku, odnosno gubitniku. U tablici 11 prikazane su točnosti klasifikacije po sezonama za svaki od pojedinih klasifikatora. Korišteno je svih 20 kategorija koje spadaju u grupu statistika praćenja igrača.

Tablica 11. Točnosti klasifikacije sa stvarnim podatcima

Sezona	JRip		RandomForest		NaiveBayes		SMO		Prosjek	
	Sr. vr.	Std	Sr. vr.	Std	Sr. vr.	Std	Sr. vr.	Std	Sr. vr.	Std
2013./'14.	77,7732	0	80,8401	0,2086	81,5253	0	82,4225	0	80,64029	1,7478
2014./'15.	77,3431	0	80,978	0,2491	81,6218	0	83,3333	0	80,81907	2,1837
2015./'16.	77,158	0	81,1808	0,2656	81,0261	0	82,9805	0	80,58632	2,1232
2016./'17.	78,2219	0	80,9299	0,3102	80,3018	0	83,7276	0	80,79527	1,9674
2017./'18.	74,9182	0	80,8838	0,0982	79,9918	0	84,2062	0	80	3,3279

Bilješke: Uključene su samo statistike koje spadaju u grupu praćenje igrača.
Sr. vr. označava srednju vrijednost, a Std standardnu devijaciju.

Vidimo kako najveću točnost, čak do 84%, dobivamo uz pomoć SMO klasifikatora. Ripper (JRip) daje najmanju točnost. Prosječna točnost sva četiri klasifikatora iznosi nešto više od 80%. Ove rezultate koristit ćemo kao referentnu točku za usporedbu rezultata koristeći pojedinačne statističke kategorije.

U tablici 12 prikazane su srednje vrijednosti točnosti klasifikacije kroz svih pet sezona za svaku od pojedinačnih kategorija iz promatrane grupe statistika.

U tablici 12 vidimo kako najveću prosječnu točnost dobivamo za postotak šuta (FG_PCT). Točnost je 78,677 ± 2,073%, što je blizu maksimalne točnosti koju smo utvrdili u tablici 11. Ovakav rezultat je očekivan jer su sve prethodne analize također pokazale iznimnu značajnost postotka šuta.

Pogledamo li postotke branjenih (CFG_PCT) i nebranjenih šutova (UFG_PCT), ponovno vidimo istu stvar kao i u prethodnim analizama, nebranjeni šutovi imaju veću točnost (70,037 ± 2,454%), što se može tumačiti kao da sadrže više informacije o pobjedniku od branjenih šutova (65,06 ± 2,407%). S obzirom da se radi o klasifikaciji na temelju samo jedne značajke, točnost klasifikacije uz pomoć postotka nebranjenih šutova također je na visokoj razini u odnosu točnost na temelju svih značajki. Zanimljivo je vidjeti kako broj zabijenih šutova (UFGM) također ima veću prosječnu točnost (66,381 ± 2,387%) od postotka branjenog šuta. Nakon ovakve analize možemo potvrditi kako su nebranjeni šutovi značajniji od branjenih šutova.

Od ostalih statističkih kategorija valja izdvojiti asistencije (AST) s visokom točnošću od 68,619 ± 2,112% te broj prilika za obrambeni skok (DRBC) s točnošću od 67,597 ± 2,989%. Obje su se kategorije i u prijašnjim analizama pokazale kao značajne.

Tablica 12. Prikaz srednje vrijednosti i standardne devijacije točnosti algoritama kroz svih pet sezona

Kat.	JRip		RandomForest		NaiveBayes		SMO		Prosjek	
	Sr. vr.	Std.	Sr. vr.	Std.	Sr. vr.	Std.	Sr. vr.	Std.	Sr. vr.	Std.
AST	68,176	0,849	65,348	1,111	70,322	1,115	70,631	0,924	68,619	2,112
DIST	52,408	2,137	51,108	0,904	53,468	2,055	54,528	2,957	52,878	1,267
SPD	49,878	0,036	49,985	0,023	49,878	0,036	49,878	0,036	49,905	0,047
TCHS	50,628	0,684	49,821	1,309	51,68	0,878	52,268	0,35	51,099	0,943
PASS	50,952	1,914	49,763	2,214	52,349	1,427	52,676	0,997	51,435	1,162
SAST	57,764	1,626	56,563	3,164	59,126	1,767	59,991	1,602	58,361	1,307
DFGA	49,878	0,814	49,747	1,69	49,724	1,087	50,042	1,546	49,848	0,127
DFGM	56,404	1,015	51,196	2,217	58,109	1,498	58,468	1,202	56,044	2,906
DFG_PCT	60,792	2,631	55,468	2,115	63,729	2,192	63,696	2,166	60,921	3,367
ORBC	55,628	1,381	52,485	2,446	57,553	1,678	57,545	1,669	55,803	2,07
DRBC	67,728	0,942	62,675	1,677	69,898	1,093	70,086	1,005	67,597	2,989
RBC	59,055	1,121	54,211	0,719	61,829	0,376	62,041	0,411	59,284	3,157
FG_PCT	78,258	0,74	75,454	0,375	80,526	0,586	80,469	0,633	78,677	2,073

CFGM	54,895	2,249	51,117	2,283	57,562	2,118	57,782	1,958	55,339	2,69
CFGA	54,314	1,351	50,699	1,273	55,709	1,477	55,995	1,484	54,179	2,107
CFG_PCT	64,862	1,531	61,199	0,436	67,089	1,499	67,089	1,62	65,06	2,407
UFGM	66,056	0,785	62,607	1,606	68,267	1,205	68,593	1,218	66,381	2,387
UFGA	51,451	0,711	48,668	2,375	52,903	1,725	53,565	1,07	51,647	1,882
UFG_PCT	70,298	0,777	65,95	1,082	71,921	0,669	71,978	0,665	70,037	2,454
FTAST	51,1	1,58	48,675	2,294	51,099	2,003	51,401	1,781	50,569	1,1
EPR	67,491	1,186	64,387	0,545	68,861	0,981	68,853	0,942	67,398	1,826

Vidimo kako postotak efektivnih dodavanja (EPR) ima prosječnu točnost od 67,398 ± 1,826%. Točnost mu je nešto niža od točnosti asistencija (AST). Sekundarne asistencije (SAST) imaju točnost 58,361 ± 1,307%, dodavanja (PASS) su na razini slučajnog odabira, isto kao i asistencije za slobodna bacanja (FTAST). Kako samo asistencije i donekle sekundarne asistencije predstavljaju bitnu razliku u ovoj analizi, a broj dodavanja i asistencije za slobodna bacanja su na razini slučajnosti, možemo pretpostaviti kako te dvije značajke „guše“ značajnost EPR-a te je zbog toga nešto lošiji od samih asistencija. U tablici 13 prikazani su iznosi točnosti klasifikacije za postotak efektivnih dodavanja i asistencije po svim sezonama kako bismo ih mogli usporediti.

Tablica 13. Usporedba točnosti klasifikacije među asistencijama i EPR-om na razini utakmice

Sezona	JRip		RandomForest		NaiveBayes		SMO	
	AST	EPR	AST	EPR	AST	EPR	AST	EPR
2013./’14.	68,883	67,007	66,786 (0,164)	63,972 (0,215)	71,533	67,659	71,860	67,741
2014./’15.	68,378	68,011	66,536 (0,238)	64,759 (0,183)	70,293	69,764	70,660	69,804
2015./’16.	66,857	65,676	64,128 (0,323)	63,982 (0,181)	68,445	67,915	69,300	67,956
2016./’17.	69,168	69,290	64,967 (0,186)	65,285 (0,224)	71,370	70,147	71,370	70,065
2017./’18.	67,594	67,471	64,320 (0,397)	63,936 (0,169)	69,967	68,822	69,967	68,699

Bilješka: U zagradama su prikazane standardne devijacije

U tablici 13 vidimo kako postotak efektivnih dodavanja (EPR) ima veću točnost klasifikacije samo u dva slučaja, i to upravo u sezoni 2016./’17., koju smo proučavali u našem prethodnom radu kada smo predložili postotak efektivnih dodavanja kao novu kategoriju. Temeljem ove tablice čini se kako je to izolirani slučaj, odnosno da je EPR eventualno u toj sezoni predstavljao bitniju kategoriju od asistencija (AST) u smislu

klasifikacijske točnosti. Iako je očekivano da će EPR i asistencije imati sličnu točnost zbog njihove relativno visoke korelacije (oko 0,7), ovakav rezultat pomalo iznenađuje s obzirom da su prijašnje analize davale nešto veću značajnost EPR-u.

5.2.4. Nova statistička kategorija – prilagođeni postotak šuta

S obzirom na značajne razlike među branjenim i nebranjenim šutovima koje su opisane u poglavlju 5.1., važnosti nebranjenog postotka šuta (UFG_PCT) u tablici 7 i važnosti broja zabijenih nebranjenih koševa (UFGM) u tablici 8, u ovom radu predlažemo još jednu novu statističku kategoriju. Novu kategoriju nazvali smo prilagođeni postotak šuta (engl. *adjusted field goal percentage*) (AFG_PCT) i formula za njeno računanje je:

$$AFG\_PCT = \frac{k * UFGM + CFGM}{UFGA + CFGA} \quad (3)$$

Ovakvim računanjem nove statističke kategorije želimo dati veću važnost broju nebranjenih zabijenih šutova te na taj način modificirati postotak šuta u nadi da će predstavljati još značajniju razliku od običnog postotka šuta. Razmatrali smo nekoliko različitih formula za računanje nove kategorije a odlučili smo se na ovu radi njene jednostavnosti i postotne interpretacije. Postotna interpretacija nam omogućava jednostavniju usporedbu s ostalim postotcima šuta, kako običnim, tako i „naprednim“ postotcima šuta opisanim u sljedećoj crtici. Matematički gledano, ovakva formula ne daje postotak, no ona u praksi nikada neće biti veća od 100%. Isto vrijedi i za formule za ostale napredne postotke šuta. Prethodnim analizama utvrđena je velika važnost i postotka nebranjenog šuta (UFG_PCT). Iako ta kategorija nije izravno uključena u računanje nove statističke kategorije, ona ima velik neposredni učinak na njen rezultat. Naime, u formuli su i broj zabijenih nebranjenih šutova (UFGM) i broj pokušanih nebranjenih šutova (UFGA), znači da i postotak nebranjenog šuta neposredno utječe na novu statističku kategoriju.

Kako smo već ranije spomenuli, slične statističke kategorije koje modificiraju postotak šuta već postoje. Efektivni postotak šuta (engl. *effective field goal percent*) modificira obični postotak na način da daje veći značaj šutovima za tri poena. Uz to, postoji i stvarni postotak šuta (engl. *true shooting percentage*) koji direktno kroz broj poena i broj

pokušanih šutova i slobodnih bacanja modificira sami postotak šuta i uzima u obzir i šutove za tri poena i slobodna bacanja. J. Kubatko i sur. [30] u svome radu daju jednostavan uvid u sve napredne statističke kategorije, pa tako i u efektivni postotak šuta i stvarni postotak šuta. Mi našom modifikacijom postotka šuta želimo „favorizirati“ ekipe s kvalitetnom selekcijom šuta i dati jedinstveni uvid u kvalitetu selekcije šuta i efikasnost šuta kroz jednu brojku. Kada kažemo kvaliteta selekcije šuta, mislimo na to da ekipa zna prepoznati i kreirati situacije za dobar (nebranjeni) šut i iskoristiti ih.

Parametar *k* odredili smo eksperimentalno uz pomoć SMO i Naive Bayes klasifikatora. S obzirom na relativno malu očekivanu domenu mogućih vrijednosti parametra *k,* koristili smo pristup brutalnom silom (engl. *brute force*), odnosno proveli smo lokalno pretraživanje „svih“ smislenih vrijednosti parametra *k*. Testiranje je provedeno na svih 5 sezona, s dva navedena klasifikatora. Prosjek točnosti obaju klasifikatora na svih pet sezona uzet je kao mjera dobrote pojedinog parametra *k*. Takvim eksperimentom dobili smo najbolje prosječne rezultate klasifikacije za *k=1,3*, tako da nam je konačna verzija formule za novu statističku kategoriju jednaka:

$$AFG\_PCT = \frac{1{,}3 * UFGM + CFGM}{UFGA + CFGA} \tag{4}$$

U tablicama 14 i 15 prikazane su razine statističke značajnosti razlike među dobrim i lošim ekipama navedene nove kategorije.

Tablica 14. Značajnost AFG_PCT na razini utakmice

Sezona	Kategorija	Pobjednička ekipa		Gubitnička ekipa		p-vrijednost
		Srednja vrijednost	Standardna devijacija	Srednja vrijednost	Standardna devijacija	
2013./'14.	AFG_PCT	0,559987	0,061096	0,495968	0,056429	6,27E-125
2014./'15.	AFG_PCT	0,550377	0,061249	0,489636	0,056284	5,74E-117
2015./'16.	AFG_PCT	0,555072	0,05979	0,495867	0,056979	4,62E-113
2016./'17.	AFG_PCT	0,548596	0,058023	0,489871	0,053432	9,84E-119
2017./'18.	AFG_PCT	0,567319	0,060192	0,509181	0,058109	1,59e-106
Bilješka:	α=5,0E-4					

Tablica 15. Značajnost AFG_PCT na razini sezone

Sezona	Kategorija	Pobjednička ekipa		Gubitnička ekipa		p-vrijednost
		Srednja vrijednost	Standardna devijacija	Srednja vrijednost	Standardna devijacija	
2013./'14.	AFG_PCT	0,547734	0,065751	0,512882	0,063977	3,34E-25
2014./'15.	AFG_PCT	0,534375	0,064674	0,49988	0,063221	1,67E-23
2015./'16.	AFG_PCT	0,549958	0,065845	0,51494	0,064959	3,14E-20
2016./'17.	AFG_PCT	0,536429	0,064842	0,510222	0,058425	1,04E-15
2017./'18.	AFG_PCT	0,549655	0,067465	0,523558	0,061619	2,54e-12
Bilješka:	α=5,0E-4					

U tablicama 14 i 15 vidimo kako nova statistička kategorija predstavlja značajnu razliku među pobjedničkim i gubitničkim ekipama. Usporedimo li *p-vrijednost* prilagođenog postotka šuta s običnim postotkom šuta (FG_PCT) u svim sezonama, izuzev sezone 2017./'18. s podjelom na razini sezone, prilagođeni postotak šuta ima manju *p-vrijednost*.

Pogledamo li rezultate iz tablice 15, vidimo kako značajnost razlike između dobrih i loših ekipa, prilagođenog postotka šuta (AFG_PCT) na razini sezone, opada iz sezone u sezonu. To bi mogao biti pokazatelj novog trenda u ligi i svakako bi ga trebalo provjeriti kroz naredne sezone.

U tablici 10 prikazana su pravila dobivena algoritmom Apriori. Uspoređujemo li prilagođeni postotak šuta (AFG_PCT) s običnim postotkom šuta, vidimo kako u tri od pet sezona prilagođeni postotak šuta ima veći stupanj pouzdanosti. Još je zanimljivo primijetiti kako u sezoni 2017./'18. postotak nebranjenog šuta (UFG_PCT) ima veću značajnost od prilagođenog postotka šuta. Iako u formuli za prilagođeni postotak šuta direktno povećavamo zabijene nebranjene šutove za 30%, postotak nebranjenog šuta je i dalje značajniji.

U tablici 16 prikazana su pravila dobivena algoritmom Ripper kada uključimo sve promatrane statističke kategorije u podatke, zajedno s postotkom efektivnih dodavanja (EPR) i prilagođenim postotkom šuta (AFG_PCT).

Ono što odmah primjećujemo jest činjenica da su sada točnosti osjetno veće nego u istoj analizi, ali bez postotka šuta (FG_PCT), postotka efektivnih dodavanja (EPR) i prilagođenog postotka šuta (AFG_PCT), prikazanoj u tablici 7. Točnosti su porasle za oko 5-6%, ovisno od sezone do sezone. Kako ne postoji niti jedno pravilo tijekom svih pet sezona u kojemu

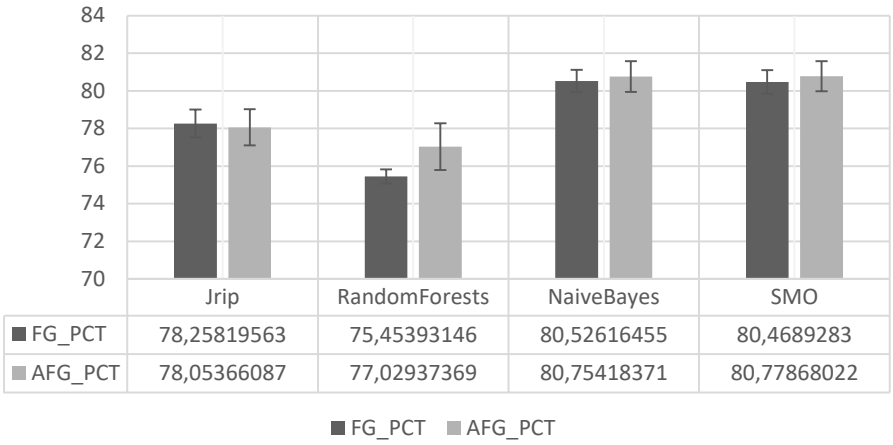
prilagođeni postotak šuta nije sadržan, možemo zaključiti da ta kategorija ima velik utjecaj na ishod klasifikacije. Samim time ima i velik utjecaj na pobjedničke ekipe i predstavlja značajnu razliku pobjedničkim i gubitničkim ekipama

Tablica 16. Pravila dobivena algoritmom Ripper s uključenim svim statističkim kategorijama

Sezona	Pravila		Točnost
2013./'14.	(AFG_PCT_OPP <= 0,526136) and (AFG_PCT >= 0,526582)	=> WINNER=1 (623,0/60,0)	76,99%
	(AFG_PCT_OPP <= 0,546429) and (AFG_PCT >= 0,507792)	=> WINNER=1 (309,0/98,0)	
	(AFG_PCT_OPP <= 0,526136) and (AFG_PCT >= 0,472816)	=> WINNER=1 (207,0/79,0)	
	(AFG_PCT >= 0,527848) and (FG_PCT_OPP <= 0,5)	=> WINNER=1 (223,0/93,0)	
		=> WINNER=0 (1090,0/194,0)	
2014./'15.	(AFG_PCT_OPP <= 0,516456) and (AFG_PCT >= 0,511628)	=> WINNER=1 (660,0/76,0)	77,02%
	(AFG_PCT >= 0,542553) and (AFG_PCT_OPP <= 0,564634)	=> WINNER=1 (241,0/52,0)	
	(AFG_PCT_OPP <= 0,515854) and (AFG_PCT >= 0,468041) and (FG_PCT_OPP <= 0,397)	=> WINNER=1 (104,0/18,0)	
		=> WINNER=0 (1449,0/368,0)	
2015./'16.	(AFG_PCT_OPP >= 0,536364) and (AFG_PCT <= 0,538202)	=> WINNER=0 (595,0/54,0)	77,32%
	(AFG_PCT <= 0,502247) and (FG_PCT_OPP >= 0,42)	=> WINNER=0 (281,0/57,0)	
	(AFG_PCT_OPP >= 0,534831) and (FG_PCT <= 0,493)	=> WINNER=0 (205,0/68,0)	
		=> WINNER=1 (1375,0/326,0)	
2016./'17.	(AFG_PCT <= 0,522093) and (AFG_PCT_OPP >= 0,501064)	=> WINNER=0 (787,0/115,0)	78,06%
	(FG_PCT <= 0,468) and (AFG_PCT_OPP >= 0,472527)	=> WINNER=0 (320,0/119,0)	
	(AFG_PCT_OPP >= 0,560577)	=> WINNER=0 (243,0/90,0)	
		=> WINNER=1 (1102,0/200,0)	
2017./'18.	(AFG_PCT_OPP <= 0.524691) and (AFG_PCT >= 0.505682)	=> WINNER=1 (691.0/91.0)	75,04%
	(FG_PCT >= 0.467) and (AFG_PCT_OPP <= 0.57931)	=> WINNER=1 (362.0/110.0)	
	(FG_PCT >= 0.466) and (AFG_PCT_OPP <= 0.583784)	=> WINNER=1 (37.0/14.0)	
		=> WINNER=0 (1354.0/347.0)	

Na grafu 1 prikazane su srednje vrijednosti točnosti klasifikacije kroz svih pet sezona. Vidimo kako su točnosti za postotak šuta (FG_PCT) i prilagođeni postotak šuta (AFG_PCT) u slučaju naivnog Bayesovog klasifikatora i SMO-a minimalno veće za prilagođeni postotak šuta, dok je kod Rippera minimalno veća točnost za obični postotak šuta, a kod slučajne šume je nešto veća točnost za prilagođeni postotak šuta.

Graf 1. Prikaz prosječne točnosti kroz svih pet sezona za FG_PCT i AFG_PCT.



Usporedimo li dobivene vrijednosti za prosječnu točnost klasifikacije prilagođenog postotka šuta (AFG_PCT) s prosječnim vrijednostima ostalih statističkih kategorija prikazanih u tablici 12, vidimo kako nijedna druga statistička kategorija osim postotka šuta (FG_PCT) nije blizu točnosti prilagođenog postotka šuta.

Prethodne analize algoritmima za izgradnju pravila pokazale su veću značajnost prilagođenog postotka šuta (AFG_PCT) u odnosu na obični postotak šuta (FG_PCT). S druge strane, rezultati prikazani na grafu 1 odaju dojam kako prilagođeni postotak šuta ne predstavlja bitniji napredak u odnosu na obični postotak šuta u smislu točnosti klasifikacije. Unatoč tome, ovakva kategorija kroz jednu brojku daje dobar uvid u to koliko je ekipa efikasna i koliko ima dobru selekciju šuta. Sigurno je kako svaka ekipa želi imati što više nebranjenih šutova te im ovakva kategorija može dati informaciju o kvaliteti selekcije šuta zajedno s time koliko su bili efikasni prilikom šutiranja.

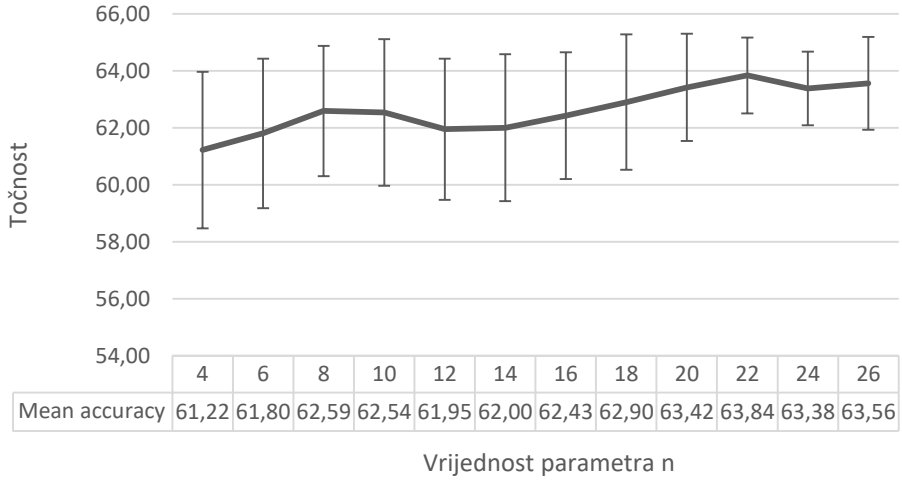
5.3. Predikcije na temelju prosjeka *n* utakmica

Kao zadnji korak naše analize napravit ćemo nekoliko testova u kojima ćemo predviđati ishode utakmica unaprijed. Predviđanje će se napraviti na temelju prosjeka zadnjih *n* odigranih utakmica. Prvi korak analize nam je određivanje parametra *n.*

5.3.1. Određivanje parametra *n*

Kako bismo odredili parametar *n*, ponovno smo koristili pristup brutalne sile (engl. *brute force*). Pokrenuli smo sva četiri algoritma, na podatcima iz tri sezone (13/14, 14/15 i 15/16) s različitim vrijednostima *n,* odnosno s uzimanjem različitog broja utakmica u računanje prosjeka na temelju kojega se radi predikcija. Nakon toga, izračunali smo srednje vrijednosti točnosti predikcije za svaki algoritam i svaku od navedete tri sezone. Prosječnu točnost klasifikacije svih algoritama koristili smo kao mjeru dobrote parametra n. Rezultati su prikazani na grafu 2.

Graf 2. Analiza za izbor parametra n

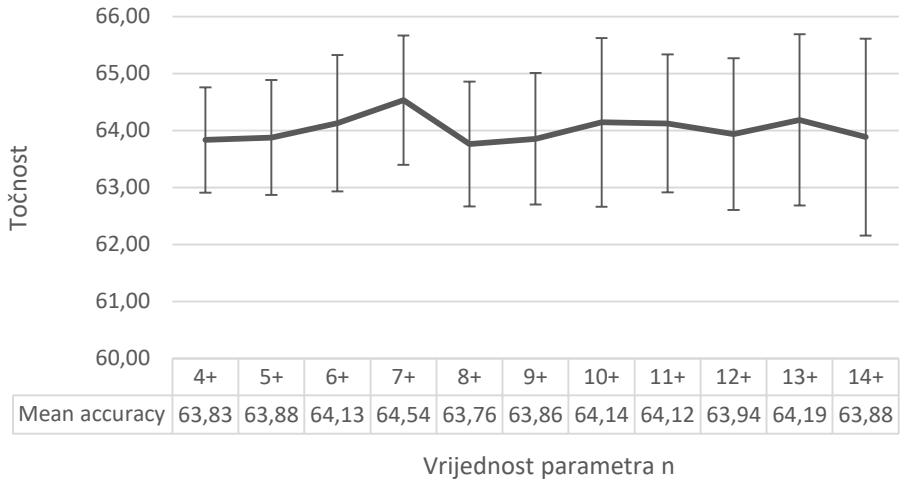


	4	6	8	10	12	14	16	18	20	22	24	26
Mean accuracy	61,22	61,80	62,59	62,54	61,95	62,00	62,43	62,90	63,42	63,84	63,38	63,56

Vidimo kako s porastom parametra *n* raste i točnost predikcija. Problem s izborom velikog parametra *n* predstavlja činjenica da se u tome slučaju gubi veliki dio skupa podataka jer ukoliko barem jedna ekipa nije odigrala 26 utakmica, ta utakmica se izbacuje iz skupa podataka. Naime, izaberemo li primjerice *n=26*, to znači da će nam iz skupa podataka otpasti minimalno (*26*30)/2=390* utakmica (30 je broj ekipa u ligi). U realnom slučaju je ta brojka i nešto veća jer zbog zgusnutog rasporeda ne mogu sve ekipe doći u isto vrijeme do 26 utakmice.

S obzirom na problem gubitka velikog broja utakmica izborom velikog parametra *n*, odlučili smo se pokušati uzeti manji broj utakmica kao početnu točku, a nakon toga uvijek gledati prosjek svih do tada odigranih utakmica u tekućoj sezoni. Ovime bismo izbjegli gubitak velikog broja utakmica, a kako bi liga odmicala, tako bi nam se i točnost predikcija povećavala jer se parametar *n* u tome slučaju povećava iz kola u kolo. S obzirom da u ovom slučaju *n* samo predstavlja koliko je utakmica potrebno odigrati da bi prosjek postojao, a nakon toga se uvijek računa prosjek svih do tada odigranih utakmica, ovdje ćemo taj parametar nazvati *n+.* Dobiveni rezultati prikazani su na grafu 3.

Graf 3. Analiza za izbor parametra n+



	4+	5+	6+	7+	8+	9+	10+	11+	12+	13+	14+
Mean accuracy	63,83	63,88	64,13	64,54	63,76	63,86	64,14	64,12	63,94	64,19	63,88

Vidimo kako su razlike među prosječnim točnostima male. To je i očekivano s obzirom kako je zapravo velika većina skupa podataka ista. Primjerice, razlika između 7+ i 8+ je samo u onim utakmicama gdje jedna ekipa ima odigranih točno 7 utakmica prije trenutne. Sve utakmice gdje su obje ekipe odigrale 8 ili više utakmica imat će iste podatke jer se računaju prosjeci svih do tada odigranih utakmica.

Vidimo na grafu 2 kako su točnosti relativno visoke za parametar *n* između osam i deset. S obzirom na to da su oko tog raspona brojki na oba grafa (2 i 3) relativno visoke točnosti, mi ćemo se odlučiti za parametar s najvišom točnošću na grafu 3, odnosno za *n=7+.*

Također na grafu 3 vidimo kako prosječna točnost svih algoritama kroz sve sezone za *n=7+* iznosi 64,54 ± 1,13%. Usporedimo li to s točnostima na grafu 2, gdje niti jedna vrijednost ne prelazi granicu od 64%, možemo potvrditi ranije iznesenu tvrdnju kako veći broj utakmica uračunat u prosjek donosi i veću točnost predikcije. Do istog zaključka

došao je D. Buursma [32] koji je radio predikcijski model za nogometne utakmice te došao do zaključka kako se točnost predikcija povećava s porastom broja utakmica uzetih u računanje prosjeka. U svome radu isprobao je računanje prosjeka čak do 75 utakmica unazad kada je došao do navedenog zaključka.

Sve buduće predikcije u ovome radu radit će se s parametrom *n=7+,* odnosno potrebno je minimalno 7 utakmica da bi prosjek postojao, a nakon toga računaju se prosjeci svih do tada odigranih utakmica kako bismo si povećali točnosti predikcija.

5.3.2. Predikcije s dobivenim parametrom *n=7+*

Kako smo *n* odredili na skupu podataka iz tri sezone, model ćemo testirati na preostale dvije sezone (14/15 i 17/18). U tablici 17 prikazane su točnosti predikcija na temelju svih statističkih kategorija iz grupe praćenje igrača, uključujući kategorije koje smo mi predložili. Vidimo kako se točnosti kreću oko 65%. To je dosta dobar rezultat s obzirom da su predikcije rađene s relativno malim skupom statistika, odnosno samo s grupom statistika praćenje igrača. U tablici 18 prikazane su točnosti predikcija na temelju svih statističkih kategorija koje su javno dostupne na službenoj NBA stranici [1].

Tablica 17. Točnosti predikcija na temelju svih kategorija iz grupe praćenja igrača i dvije nove kategorije

Sezona	JRip		RandomForest		NaiveBayes		SMO	
	Sr. vr.	Std	Sr. vr.	Std	Sr. vr.	Std	Sr. vr.	Std
2014./'15.	64,10	0	64,49	0,25	66,69	0	65,17	0
2017./'18.	65,25	0	65,20	0,17	64,98	0	65,52	0

Tablica 18. Točnosti predikcija na temelju svih javno dostupnih statističkih kategorija

Sezona	JRip		RandomForest		NaiveBayes		SMO	
	Sr. vr.	Std	Sr. vr.	Std	Sr. vr.	Std	Sr. vr.	Std
2014./'15.	63.47	0	65.46	0.34	67.68	0	65.35	0
2017./'18.	64.71	0	66.17	0.14	65.97	0	65.79	0

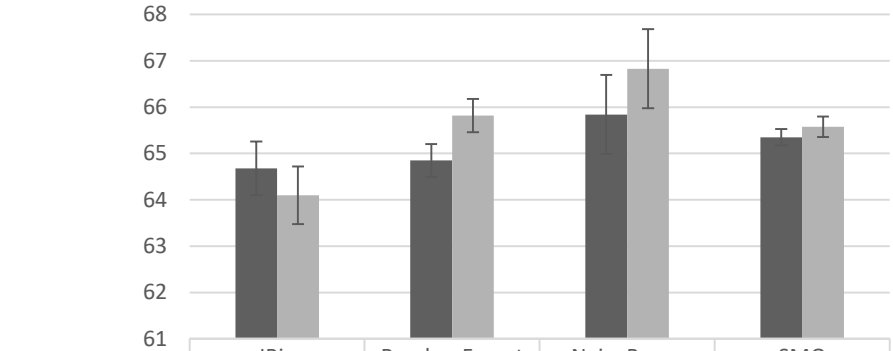
Bilješka: Podatci sadrže 96 statističkih kategorija

Na grafu 4 prikazana je usporedba točnosti klasifikacije na temelju statističkih kategorija iz grupe praćenja igrača sa svim javno dostupnim statističkim kategorijama. Vidimo kako je

u prosjeku točnost temeljem svih značajki nešto malo veća. Svih značajki ima 96, dok značajki iz grupe praćenje igrača ima 20, uz dodatne dvije koje smo mi predložili. Mjerenjem vremena učenja klasifikatora utvrdili smo da se vrijeme učenja za sve značajke poveća za oko tri puta kod algoritma Ripper (s šest sekundi na 18 sekundi), za oko dva i pol puta za naivni Bayesov klasifikator (s pet sekundi na 12 sekundi) te oko dva puta za slučajne šume i SMO (s 1:30 minute na 3:00 minute). Prilikom navedenog mjerenja vremena mjerimo samo vrijeme učenja klasifikatora, ne i vrijeme pripreme podataka. Vrijeme pripreme podataka produži se za oko 3 puta u slučaju svih značajki (s 1:20 minute na 4:00 minute). Vidimo kako korištenjem samo statistika iz grupe praćenje igrača štedimo puno vremena, a da pritom ne gubimo previše na točnosti.

Navedeni rezultati nam potvrđuju da su statistike sadržane u grupi praćenja igrača kvalitetne statistike koje čine bitnu razliku između pobjedničkih i gubitničkih ekipa. Valja naglasiti i kako bi se nekom selekcijom značajki vjerojatno dobila nešto veća točnost u tablici 18, ali naša je pretpostavka kako razlika ne bi bila velika.

Graf 4. Usporedba prosječnih točnosti na temelju kategorija praćenja igrača i na temelju svih kategorija



	JRip	RandomForest	NaiveBayes	SMO
player_tracking	64,67839711	64,84835055	65,84150358	65,3505637
all_stats	64,09503081	65,81783693	66,82894341	65,57510272

U tablici 19 prikazane su prosječne točnosti predikcija na temelju svake pojedine kategorije. Zanimljivo je vidjeti kako je minimalna prosječna točnost za kategoriju asistencije za slobodna bacanja (FTAST) i iznosi 56,386 ± 1,725%. Nema niti jedne kategorije za koju bi točnost bila na razini slučajnosti. Objašnjenje za to moglo bi biti to da, bez obzira koliko statistička kategorija bila beznačajna prema našoj prijašnjoj analizi,

bolje ekipe će ipak imati nešto bolje prosjeke i u tim kategorijama. Kako se prosjeci za predikcije računaju na temelju svih do tada odigranih utakmica, u drugom dijelu sezone će prosjeci već biti uvelike formirani i jako će se sporo i teško mijenjati. S obzirom da bolje ekipe imaju više pobjeda na razini sezone, upravo zbog te tromosti promjene prosjeka i činjenice da će bolje ekipe u globalu imati nešto bolje prosjeke u svim kategorijama, klasifikatori će uvijek imati točnost iznad teoretskog slučajnog odabira.

Vidimo kako prilagođeni postotak šuta (AFG_PCT) daje nešto veću točnost od običnog postotka šuta (FG_PCT), te također od postotka nebranjenog šuta (UFG_PCT). Točnosti tih triju kategorija ističu se s nešto većom točnošću od ostalih što ponovno govori o njihovoj važnosti. Od ostalih kategorija još valja izdvojiti asistencije (AST) i sekundarne asistencije (SAST) kao kategorije koje se malo izdvajaju točnošću.

Tablica 19. Prikaz srednje vrijednosti i standardne devijacije točnosti predikcija kroz svih pet sezona

Kat.	JRip		RandomForests		NaiveBayes		SMO		Prosjek	
	Sr. vr.	Std	Sr. vr.	Std	Sr. vr.	Std	Sr. vr.	Std	Sr. vr.	Std
AST	59,873	1,406	56,46	1,067	60,196	1,071	58,259	0,92	58,697	1,486
DIST	57,507	1,63	55,744	1,799	58,35	1,3	58,17	0,973	57,443	1,03
SPD	58,17	0,973	58,17	0,973	58,17	0,973	58,17	0,973	58,17	0
TCHS	57,201	1,349	55,776	1,264	58,385	1,464	58,17	0,973	57,383	1,03
PASS	57,237	1,317	55,857	1,132	57,866	1,745	58,17	0,973	57,282	0,889
SAST	59,602	2,45	55,44	2,675	60,231	2,07	58,421	1,436	58,424	1,841
DFGA	57,686	0,781	54,833	1,485	57,775	0,973	58,17	0,973	57,116	1,331
DFGM	58,189	1,624	56,015	1,759	58,17	1,28	58,17	0,973	57,636	0,936
DFG_PCT	58,958	1,848	57,082	1,919	59,461	1,693	58,17	0,973	58,418	0,898
ORBC	57,937	1,066	56,396	1,24	58,636	0,619	58,17	0,973	57,785	0,84
DRBC	58,368	1,888	55,803	0,825	59,802	1,175	58,17	0,973	58,036	1,435
RBC	57,093	1,527	55,624	1,384	58,35	0,592	58,17	0,973	57,309	1,085
FG_PCT	59,065	2,482	55,931	1,438	60,177	1,504	58,42	1,638	58,399	1,557
CFGM	57,937	1,045	54,393	1,407	58,08	1,317	58,17	0,973	57,145	1,591
CFGA	57,866	2,232	55,803	0,762	58,098	0,951	58,17	0,973	57,484	0,977
CFG_PCT	59,174	1,862	56,722	0,946	60,143	2,049	58,277	1,168	58,579	1,259
UFGM	58,457	1,145	55,244	1,345	59,999	1,702	58,17	0,973	57,967	1,719
UFGA	57,166	1,157	55,812	1,747	57,579	1,385	58,17	0,973	57,182	0,868
UFG_PCT	58,493	1,348	56,168	0,739	59,516	1,341	58,17	0,973	58,087	1,214
FTAST	56,18	0,607	53,668	1,092	57,524	1,065	58,17	0,973	56,386	1,725
EPR	58,404	1,737	54,314	1,18	59,713	1,757	58,17	0,973	57,65	2,014
AFG_PCT	59,659	1,274	55,58	0,608	60,142	1,385	58,599	1,942	58,495	1,773

S obzirom na tromost promjene prosjeka i na činjenicu kako imamo samo jednu kategoriju na temelju koje moramo odrediti pobjednika, ovako niske točnosti su djelomično očekivane. Ono što nas je ipak iznenadilo je da ne postoji niti jedna kategorija koja bi davala točnosti na razini slučajnosti.

6. Rasprava i zaključak

Nekoliko vrsta analiza provedenih u ovome radu pokazale su kako unutar promatrane grupe statistika postoji nekoliko statističkih kategorije koje predstavljaju veliku razliku između pobjedničkih i gubitničkih ekipa. Statističkom analizom pokazali smo značajnu razliku za nekoliko kategorija: 1) postotci šuta (FG_PCT), posebice je značajan postotak nebranjenog šuta (UFG_PCT), ali i postotak branjenog šuta (CFG_PCT); 2) broj postignutih nebranjenih šutova (UFGM) također je od velikog značaja; 3) broj asistencija (AST) i sekundarnih asistencija (SAST) je u svim sezonama veći kod pobjedničkih ekipa; 4) broj prilika za obrambeni skok (DRBC) koji može izravno biti povezan s protivničkim postotcima šuta.

Dubinskom analizom navedene grupe statistika potvrdili smo sve zaključke iz statističke analize. Pokazalo se kako su nebranjeni šutovi bitniji od branjenih. Algoritmi koji generiraju pravila u velikoj su mjeri davali važnost postotcima šuta. Kada smo uklonili postotke šuta iz podataka, glavnu ulogu u pravilima preuzele su ostale kategorije koje su se pokazale kao značajne prilikom statističke analize.

U našem prethodnom radu [16] predložili smo novu kategoriju, postotak efektivnih dodavanja (EPR). U ovome radu pokazali smo kako jedan od razloga zbog kojega smo tu kategoriju predstavili ne vrijedi na većem skupu podataka, odnosno pobjedničke ekipe nemaju manje dodavanja tijekom svih pet sezona. Unatoč tome, postotak efektivnih dodavanja se u većini provedenih analiza pokazao kao bitna kategorija.

Ovdje smo predložili još jednu novu statistiku, prilagođeni postotak šuta (AFG_PCT). S obzirom da smo svim analizama utvrdili važnost nebranjenih šutova, odlučili smo kreirati novu kategoriju koja predstavlja postotak šuta koji daje veći značaj nebranjenim šutovima. Kategorija se kroz većinu analiza pokazala kao značajna, jedino prilikom analize klasifikacijom na stvarnim podatcima daje približno jednake rezultate kao i obični postotak šuta (FG_PCT).

Postotak efektivnih dodavanja (EPR) i prilagođeni postotak šuta (AFG_PCT) u analizi klasifikacijom daju rezultate koji su slični rezultatima s postojećim kategorijama, što se može tumačiti kao da ne unose neko poboljšanje u odnosu na postojeće kategorije.

Unatoč tome, mi preporučamo korištenje obiju novih statistika. Smatramo kako prilagođeni postotak šuta predstavlja jedinstvenu kategoriju koju tumačimo kao postotak šuta koji uzima u obzir i kvalitetu selekcije šuta i samu efikasnost prilikom šutiranja. Ekipe koje imaju dobru selekciju šuta na utakmici, odnosno puno nebranjenih zabijenih šutova te su uz to efikasne, imat će i visok prilagođeni postotak šuta. Isto tako, postotak efektivnih dodavanja daje uvid u odnos kvalitete i kvantitete dodavanja. Iako sam broj dodavanja dosta ovisi o stilu igre pojedine ekipe, svakako da svaka ekipa teži tome da što lakše postiže poene. Smatramo kako nema lakšeg načina od toga da se u što kraćem roku i sa što manje dodavanja upiše asistencija, a samim time i poeni. Svjesni smo kako to u realnosti nije tako jednostavno ostvariti, ali to je neki ideal za koji vjerujemo da mu sve ekipe teže. Zbog toga preporučamo korištenje ove kategorije kao relevantne mjere.

Prilikom predikcija na temelju prosječnih podataka, utvrdili smo kako s porastom broja utakmica koje se uzimaju u prosjek raste i točnost predikcija. Mi smo se kroz analizu odlučili da minimalni broj utakmica da bi prosjek postojao bude 7, a nakon toga smo računali prosjeke na temelju svih do tada odigranih utakmica. Točnosti predikcija dobivene na grupi statistika praćenja igrača, zajedno s dvije novopredložene kategorije, kreće se oko 65%, što je značajno bolje od slučajnog predviđanja (oko 50%). G. Cheng i sur. [19] u svome radu postižu točnost predikcija od 74,4% na skupu podataka s 28 osnovnih značajki. Predikcije su radili na temelju prosjeka zadnjih šest utakmica. A. Jadhav i sur. [18] postižu točnost od 88% uz pomoć SVM-a, ali na stvarnim podatcima osnovnih statističkih kategorija, bez računanja prosjeka. Mi smo takvim načinom predikcija dobili prosječne točnosti oko 80%, dok se kao najbolji klasifikator pokazao SMO s točnostima između 82% i 84%.

U navedenim radovima su postignute veće točnosti od naših, međutim valja naglasiti kako je naš skup podataka manji, odnosno imamo manje značajki. Uz to, dosta značajki koje smo mi promatrali ne predstavljaju bitnu razliku među pobjedničkim i gubitničkim ekipama. Kako smo u radu pokazali značajnost nekih statističkih kategorija i dobili relativno visoke postotke točnosti klasifikacije s tim kategorijama, vjerujemo kako bi neke od tih kategorija poboljšale točnost klasifikacije u navedenim radovima.

Moguća nadogradnja našeg istraživanja može biti proučavanje dvije predložene kategorije u narednim sezonama kako bi se dodatno utvrdio stupanj njihove značajnosti. Ukoliko

uđu u uporabu smatramo da će njihova značajnost rasti jer će ekipe početi „voditi računa“ o njima. Uz to, moguće je i proširenje istraživanja na još više prijašnjih i budućih sezona radi praćenja nekih od u radu navedenih trendova u košarci. Detaljna usporedba predloženog prilagođenog postotka šuta s već postojećim efektivnim postotkom šuta i stvarnim postotkom šuta bi se također moglo pokazati kao zanimljivo istraživanje.

_____

Igor Stančin

7. Literatura

[1] National Basketball Association, *Official NBA stats API* https://stats.nba.com/

[2] P. Tanna i Y. Ghodasara, „Using Apriori with WEKA for Frequent Pattern Mining“, International Journal of Engineering Trends and Technology,vol. 12, no. 3, pp. 127-131, 2014, doi: 10.14445/22315381/IJETT-V12P223.

[3] Machine Learning Group at the University of Waikato, *Weka 3: Data Mining Software in Java*, https://www.cs.waikato.ac.nz/ml/weka/

[4] R. Agrawal i R. Srikant,„Fast Algorithms for Mining Association Pravila“, Proc. 20th Int. Pouzdanost Very Large Data Bases VLDB,San Jose, SAD, 2000, pp. 478-499.

[5] B. Liu, W. Hsu i Y. Ma, „Integrating Classification and Association Rule Mining“, proceedings of Fourth International Conference on Knowledge Discovery and Data Mining, New York, SAD, 1998, pp. 80-86.

[6] W. W. Cohen, „Fast Effective Rule Induction“, Twelfth International Conference on Machine Learning, Tahoe City, SAD, 1995, pp. 115-123.

[7] L. Breiman, „Random Forests“, Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[8] M. Denil, D. Matheson i N. de Freitas, „Consistency of Online Random Forests“, Proceedings of the 30th International Conference on Machine Learning, Atlanta, SAD, 2013, pp. 1256-1264

[9] G. H. John i P. Langley, „Estimating Continuous Distributions in Bayesian Classifiers“, Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, Canada, 1995, pp. 338-345

[10] C. D. Manning, P. Raghavan i H. Schütze, „Introduction to Information Retrieval“, Cambridge University Press, 2008.

[11] J. Platt, „Advances in Kernel Methods - Support Vector Learning“, 1st ed., MIT Press, 1998

[12] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya i K.R.K. Murthy, „Improvements to Platt's SMO Algorithm for SVM Classifier Design“, Neural Computation, vol. 13, no. 3, pp. 637-649, 2001.

[13] T. Hastie i R. Tibshirani,„Advances in Neural Information Processing Systems“, vol. 10MIT Press, 1998.

[14] J. Platt,, „Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines“, Advances in Kernel Methods-Support Vector Learning, Microsoft Research, MIT Press, 1998

[15] I. Stančin, *Master thesis*, https://github.com/istancin/master-thesis

[16] I. Stančin i A. Jović, „Analyzing the Influence of Player Tracking Statistics on Winning Basketball Teams“, MIPRO 2018 41st International Convention, Opatija 2018, pp. 1779-1784

[17] *python-weka-wrapper3*, https://github.com/fracpete/python-weka-wrapper3

[18] A. Jadhav, S. Das, S. Khatode i R. Degaonkar, „Predicting the NBA Playoff Using SVM“, Journal of Web Development and Web Designing, vol. 1, no. 1, pp. 1-6, 2016.

[19] G. Cheng, Z. Zhang, M. N. Kyebambe i N. Kimbugwe, „Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle“, Entropy, vol. 18, no. 12, pp. 450-465.

[20] M. R. Summers, (2013) "How to Win in the NBA Playoffs: A Statistical Analysis", American Journal of Management, Vol. 13, Iss. 3, pp. 11 - 24, 2013.

[21] D. Cervone, A. D'Amour, L. Bornn i K. Goldsberry, „A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes“, Journal Of The American Statistical Association, vol. 111, iss. 514, pp. 585-599, 2016.

[22] S. B. Caudill, „Predicting discrete outcomes with the maximum score estimator: The case of the NCAA men's basketball tournament“, International Journal of Forecasting, vol. 19, iss. 2, pp. 313-317, 2003.

[23] M. Utku Özmen, „Marginal contribution of game statistics to probability of winning at different levels of competition in basketball: Evidence from the Euroleague“, International Journal of Sports Science & Coaching, vol. 11, iss. 1, pp. 98–107, 2016.

[24] G. Csátaljay, P. O'Donoghue, M. Hughes i H. Dancs, „Performance indicators that distinguish winning and Gubitnička ekipas in basketball“, International Journal of Performance Analysis in Sport, vol. 9. pp. 60-66, 2009.

[25] G. Csátaljay, N. James, M. Hughes i H. Dancs, „Performance differences between winning and losing basketball teams during close, balanced and unbalanced quarters“, Journal of Human Sport and Exercise, vol. 7, no. 2, pp. 356-364, 2012.

[26] J. Sampaio, T. McGarry, J. Calleja-González, S. Jiménez Sáiz, Schelling, X. del Alcázar i M. Balciunas, „Exploring Game Performance in the National Basketball Association Using Player Tracking Data" PLoS ONE, vol. 10, no. 7, e0132894, 2015.

[27] G. Csátaljay, N. James, M. Hughes i H. Dancs, „Effects of defensive pressure on basketball shooting performance“, International Journal of Performance Analysis in Sport, vol. 13, no. 3, pp. 594-601, 2013.

[28] G. Ziv, R. Lidor i M. Arnon, „Predicting team rankings in basketball: The questionable use of on-court performance statistics“, International Journal of Performance Analysis in Sport, vol. 10, no. 2, pp. 103-114, 2010.

[29] M. Ruano, L. Gasperi i C. Lupo, „Performance analysis of game dynamics during the 4th game quarter of NBA close games“, International Journal of Performance Analysis in Sport, vol. 15 no. 1, pp. 249-263, 2016.

[30] J. Kubatko, D. Oliver, K. Pelton i D.T. Rosenbaum, „A Starting Point for Analyzing Basketball Statistics“, Journal of Quantitative Analysis in Sports, vol. 3, no. 3, 1559-0410.1070, 2007.

[31] S. Ibáñez, J. Sampaio, S. Feu, A.L. Calvo,M. Ruano i E. Ortega, „Basketball game-related statistics that discriminate between teams Sezona-long success“, European Journal of Sport Science, vol. 8,no. 6, pp. 369-372, 2008.

[32] D. Buursma, „Predicting sports events from past results“, 14th Twente Student Conference on IT, Enschede, Nizozemska, 2011

[33] A. Franks, A. Miller, L. Bornn i K. Goldsberry, “Counterpoints: Advanced Defensive Metrics for NBA Basketball.” MIT Sloan, 9th Annual Sports Analytics Conference (SSAC15), Boston, MA, USA, Feb. 27-28, 2015, pp. 1–8.

[34] T. Al Baghal, „Are the Four Factors Indicators of One Factor? An Application of Structural Equation Modeling Methodology to NBA Data in Prediction of Winning Percentage“, Journal of Quantitative Analysis in Sports, vol. 8, no. 1, 1559-0410.1355, 2012.

[35] M. Ruano, A. L. Calvo, J. Sampaio, S. Ibáñez i E. Ortega, „Game-related statistics that discriminated winning and Gubitnička ekipas from the Spanish Men's Professional Basketball Teams“, Collegium antropologicum, vol. 32, no. 2, pp. 451-6, 2008.

[36] R. Hofler i J. Payne, „Efficiency in the National Basketball Association: A Stochastic Frontier Approach With Panel “ana", Managerial and Decision Economics, vol. 27, no. 4, pp. 279-285, 2006.

[37] S. Trninić, D. Dizdar i E. Luksić, „Differences Between Winning and Defeated Top Quality Basketball Teams in Final Tournaments of European ClubChampionship“, Collegium antropologicum, vol. 26, no. 2, pp. 521-31, 2003.

[38] M. Lopez i G. Matthews, „Building an NCAA mens basketball predictive model and quantifying its success“, Journal of Quantitative Analysis in Sports, vol. 11, no. 1, pp. 5-12, 2014.

[39] I. H. Witten, E. Frank i M. A. Hall, „Data Mining: Practical Machine Learning Tools and Techniques“, 3rd ed., Morgan Kaufmann, 2011

Dubinska analiza statističkih kategorija praćenja igrača u košarkaškim ekipama

Sažetak: U ovom radu proučavana je značajnost statističkih kategorija unutar grupe praćenja igrača, ali na razini ekipa. Provedena je statistička analiza, dubinska analiza algoritmima za generiranje pravila, analiza klasifikacijom te su napravljene predikcije na temelju prosjeka. Najznačajnije utvrđene značajnosti su: 1) nebranjeni postotka šuta; 2) broja postignutih nebranjenih koševa; 3) postotak branjenog šuta; 4) asistencije i sekundarne asistencije; i 5) prilika za obrambeni skok. U radu predlažemo dvije nove statističke kategorije: postotak efektivnih dodavanja i prilagođeni postotak šuta. Predložene kategorije su također pokazale visok stupanj značajnosti u analizama. Prilikom predikcija na temelju prosjeka zadnjih n utakmica, utvrdili smo kako što je *n* veći, to su i točnosti veće. Dobivene točnosti predikcija samo na skupu podataka praćenja igrača kreću se oko 65%.

Ključne riječi: košarka, statistika, praćenje igrača, klasifikacija, predikcija, dubinska analiza

Data mining analysis of player tracking statistics in basketball teams

Summery: Main goal of this thesis was to identify the most significant statistical categories in group of player tracking statistics, but on a team level. We have done statistical analysis on data, analysis with data mining algorithms for rule inducting, analysis with classification, and predictions based on Prosjek values of categories. Results show that most significant categories are: 1) uncontested field goal percentage; 2) uncontested field goals made; 3) contested field goal percentage; 4) assists and secondary assists; and 5) defensive rebound chances. In the thesis, we suggest two novel statistical categories: effective passing ratio and adjusted field goal percentage. With our model for predictions, we obtain Prosjek Točnost around 65% based only on player tracking data. Also, we concluded that if we take more games in calculating the Prosjek, we get higher Točnost of predictions.

Key words: basketball, statistics, player tracking, classification, prediction, data mining