NBA

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Dimić Nikola, Krstić Dušica
dimic.nikola@gmail.com, dusicamkrstic@gmail.com

2. jun 2018.

Sažetak

Ovaj rad se bavi prikazom različitih metoda i tehnika korišćenih u procesu istraživanja podataka na primeru baze sa sezonskim podacima NBA lige. Obrađene su oblasti pretprocesiranja, klaster analize, klasifikacije kao i pravila pridruživanja.

Sadržaj

1 Uvod **2**

2 Analiza i pretprocesiranje podataka **2**
2.1 Analiza podataka . . . 2
2.2 Obrada podataka . . . 3
2.2.1 Analiza ekstremnih vrednosti . . . 3
2.3 Razdvajanje era NBA lige . . . 5

3 Pravila pridruživanja **6**

4 Klasifikacija **7**
4.1 Priprema podataka za klasifikaciju . . . 7
4.2 Klasifikacija korišćenjem stabla odlučivanja . . . 10
4.2.1 Ginijev indeks . . . 11
4.2.2 Odnos dobiti . . . 12
4.3 Klasifikacija metodom K najbližih suseda . . . 14
4.4 Klasifikacija naivnim Bajesovim algoritmom . . . 15
4.5 Klasifikacija metodom potpornih vektora (SVM) . . . 16
4.5.1 Polinomijalan kernel . . . 16
4.5.2 Sigmid kernel . . . 16
4.5.3 Gausov (RBF) kernel . . . 16

5 Klaster analiza **18**

6 Zaključak **20**

1 Uvod

Skup podataka "NBA Players stats since 1950" sadrži individualne statistike svakog igrača za svaku NBA sezonu od 1950. godine zaključno sa 2017. godinom.

Ovaj skup podataka je preuzet sa adrese https://www.kaggle.com/drgilermo/nba-players-stats.

2 Analiza i pretprocesiranje podataka

U ovom odeljku ćemo se baviti pripremom i obradom podataka kako bi bili pogodniji za dalju analizu i primenu algoritama za klasifikaciju i klasterovanje.

2.1 Analiza podataka

Skup podataka "NBA Players stats since 1950" se sastoji iz tri datoteke od kojih su za naš rad značajne dve datoteke: player_data.csv i Season_stats.csv.

Opisi datoteka su dati u tabelama 1 i 2.

name	ime i prezime
year_start	godina početka NBA karijere
year_end	kraj početka NBA karijere
position	pozicija
height	visina
weight	težina
birth_date	datum rođenja
college	koledž

Tabela 1: Opis tabele player_stats.csv

Player	ime i prezime
Year	sezona
Pos	pozicija
Age	starost
Tm	tim
3P	broj postignutih šuteva za 3 poena
3PA	broj pokušanih šuteva za 3 poena
2P	broj postignutih šuteva za 2 poena
2PA	broj pokušanih šuteva za 2 poena
FT	broj postignutih slobodnih bacanja
FTA	broj pokušanih slobodnih bacanja
ORB	broj ofanzivnih skokova
DRB	broj defanzivnih skokova
TRB	ukupan broj skokova
AST	broj asistencija
STL	broj ukradenih lopti
BLK	broj blokova
TOV	broj prodatih lopti
PF	broj faulova
PTS	ukupan broj postignutih poena

Tabela 2: Opis tabele Season_Stats.csv

Datoteke sadrže i dodatne podatke, od kojih su neki neupotrebljivi, loše definisani, a neki nisu imali poseban značaj za naše istraživanje.

2.2 Obrada podataka

Posmatrajući datoteku sa podacima o igračima uočili smo da su visina i težina u skladu sa američkim sistemom veličina, stoga smo ih konvertovali u SI sistem.

Atribut visine je kategoričkog tipa, string formata X-Y (gde je X broj stopa, a Y broj inča). Konverziju datog stringa u numerički atribut, konkretno u centimetre, postigli smo tako što smo razdvojili string u odnosu na poziciju crtice u tom stringu, vrednost ispred pomnožili sa 30.48, vrednost iza pomnožili sa 2.54, sabrali obe vrednosti i konvertovali tako da je njihov zbir celobrojnog tipa.

Expression

```
toInt(
     toDouble(substr($height$,0,1))* 30.48 +
     toDouble(substr($height$, indexOf($height$,"-")+1)) * 2.54
     )
```

Slika 1: Konverzija visine

Atribut težine je bio izražen u funtama tako da smo ga pomnožili sa 0.45359237 i zaokružili dobijenu vrednost tako da bude ceo broj. Na ovaj način smo dobili atribut u željenoj metrici.

Expression

```
round($weight$ * 0.45359237)
```

Slika 2: Konverzija težine

Nakon korigovanja atributa, bavili smo se nedostajućim vrednostima u tabeli na sledeći način:

- pozicija igrača - medijana obeležja
- visina igrača - srednja vrednost obeležja
- težina igrača - srednja vrednost obeležja
- datun rođenja - *Unknown*
- koledž - *International player* (uočili smo da ta informacija nedostaje za sve igrače koji nisu iz Sjedinjenih Američkih Država stoga smo im dodelili ovu vrednost)

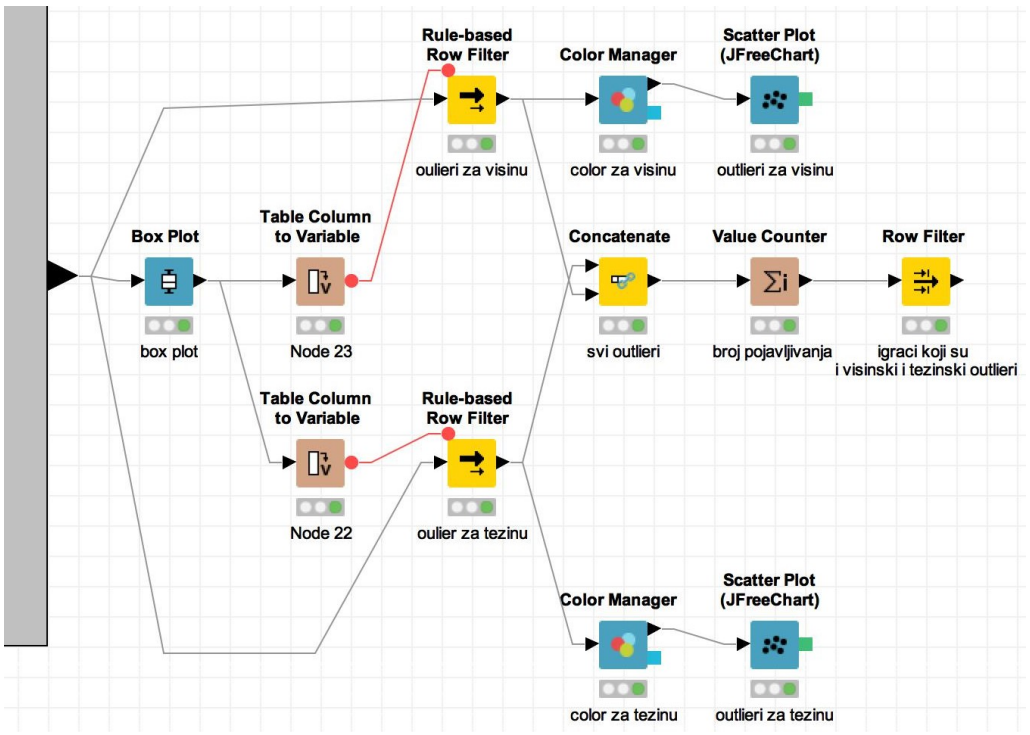
2.2.1 Analiza ekstremnih vrednosti

Na tabelu sa prethodno obrađenim atributima, primenili smo čvor *Box Plot* kako bismo dobili raspodelu vrednosti atributa.

Row ID	D year_start	D year_end	D height	D weight
Minimum	1,947	1,947	160	52
Smallest	1,947	1,947	170	62
Lower Quar...	1,969	1,973	190	86
Median	1,986	1,992	198	95
Upper Qua...	2,003	2,009	205	102
Largest	2,018	2,018	226	126
Maximum	2,018	2,018	231	163

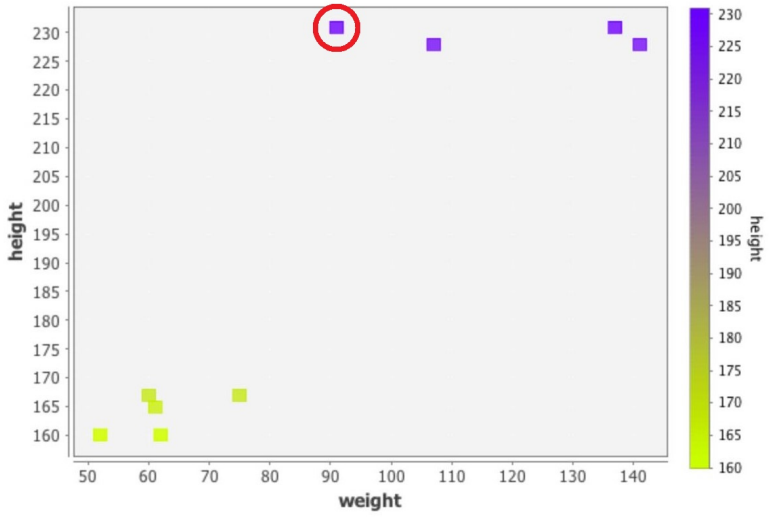
Slika 3: Raspodela atributa

Zbog nepoklapanja vrednosti atributa *Minimum* i *Smallest*, kao i vrednosti atributa *Maximum* i *Largest* zaključili smo da postoje autlajeri za visinu i težinu, koje smo dalje analizirali.



Slika 4: Obrada ektremnih vrednosti

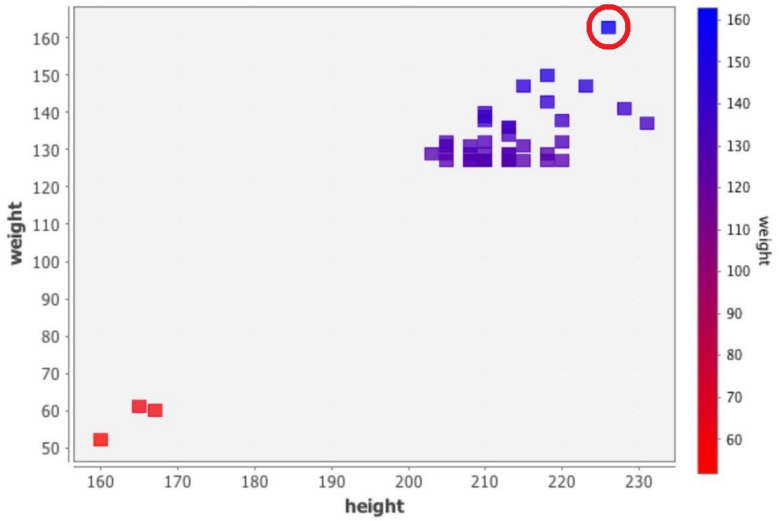
Posebno smo izdvojili ekstremne vrednosti za visinu i prikazali ih grafički korišćenjem *Scatter Plot* vizuelizacije.



Slika 5: Autlajeri za visinu

Kao što se može primetiti na slici 5, zaokruženi autlajer je posebno zanimljiv jer iako je autlajer za visinu, njegova težina je znatno manja od ostalih izuzetno visokih košarkaša. Ovaj košarkaš se zove Manute Bol, visok je 230cm i težak 91kg.

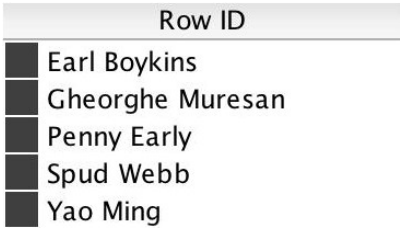
Istim postupkom smo prikazali i autlajere za težinu.



Slika 6: Autlajeri za težinu

Košarkaš koji se ovde najviše ističe čak i u odnosu na druge autlajere je košarkaš Sim Bhullar, težak 163kg.

Nakon ovoga smo ispitali da li postoje ekstremne vrednosti u obe kategorije, i izdvojili smo sledeće igrače:

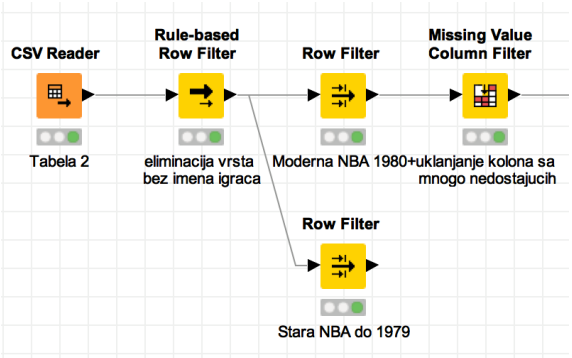


Slika 7: Autlajeri u obe kategorije

2.3 Razdvajanje era NBA lige

U datoteci koja sadrži sezonske statistike igrača su postojali unosi koji su imali nedefisane vrednosti za sve atribute i njih smo u startu eliminisali korišćenjem *Rule-based Row Filter* čvora.

Prolaskom kroz tabelu smo uočili veliku razliku u broju nedefinisanih vrednosti u podacima za period od 1950. do 1979. i od 1980. do 2017. zbog uvođenja linije za tri poena. Ovo nas je motivisalo da razdvojimo tabelu na dve zasebne koje predstavljaju dve ere NBA lige, što smo postigli koristeći *Row Filter*.



Slika 8: Razdvajanje dve NBA ere

3 Pravila pridruživanja

Zanimale su nas korelacije između srednjih vrednosti atributa ORB (en. *Offensive Rebound*), DRB (en. *Defensive Rebound*), STL (en. *Steals*) i BLK (en. *Blocks*) u odnosu na igrača i njegovu poziciju u toku cele karijere. Ovo nam je bilo interesantno jer smo želeli da pokažemo da postoji veza između ostvarenih defanzivnih rezultata košarkaša i pozicije koju igra.

Bitno je napomenuti da su analizirane samo sezone nove NBA ere, gde su eliminisani atributi sa više od 90% nedostajućih vrednosti. Rezultat ovoga je tabela sa 6 kolona i 2835 redova.

Normalizovanjem ovih podataka koristeći *Normalizer* čvor i njihovom kategorizacijom (korišćenjem *Auto-Binner* čvora) dobili smo tabelu u kojoj za svakog igrača i poziciju koju je igrao imamo rang svakog atributa na skali od 1 do 4:

- 1 - loše
- 2 - dobro
- 3 - vrlo dobro
- 4 - odlično

Player	First(Pos)	Mean(ORB)	Mean(DRB)	Mean(STL)	Mean(BLK)
Deng Gai	C	OR_ 1	DR_ 1	STL_ 1	BLK_ 1
Dennis Awtrey	C	OR_ 2	DR_ 2	STL_ 2	BLK_ 3
Dennis Hopson	SG	OR_ 4	DR_ 3	STL_ 4	BLK_ 4
Dennis Horner	PF	OR_ 1	DR_ 1	STL_ 1	BLK_ 1
Dennis Johnson*	SG	OR_ 4	DR_ 4	STL_ 4	BLK_ 4
Dennis Nutt	PG	OR_ 1	DR_ 1	STL_ 2	BLK_ 1
Dennis Rodman*	SF	OR_ 4	DR_ 4	STL_ 4	BLK_ 4
Dennis Schroder	PG	OR_ 2	DR_ 3	STL_ 4	BLK_ 2
Dennis Scott	SG	OR_ 3	DR_ 3	STL_ 4	BLK_ 3
Denzel Valentine	SG	OR_ 2	DR_ 3	STL_ 3	BLK_ 2
DerMarr Johnson	SF	OR_ 3	DR_ 3	STL_ 3	BLK_ 4
Derek Anderson	SF	OR_ 3	DR_ 3	STL_ 4	BLK_ 2
Derek Fisher	PG	OR_ 2	DR_ 3	STL_ 4	BLK_ 2

Slika 9: Kategorizacija

Korišćenjem *Create Collection Column* čvora pripremili smo podatke za primenu algoritma koji nalazi zavisnosti između gore pomenutih atributa. Postavljanjem minimalne podrške na 0.05 i minimalne pouzdanosti na 0.8 našli smo pravila koja su zanimljiva zbog toga što imaju *lift* meru veću od 1, što znači da se stavke zajedno pojavljuju više puta nego što je to očekivano.

Row ID	Support	Confid...	Lift	Conse...	implies	Items
rule11	0.06	0.805	3.195	OR_ 1	<---	[SG,BLK_ 1]
rule12	0.061	0.94	3.765	OR_ 4	<---	[BLK_ 4,DR_ 4,STL_ 3]
rule13	0.061	0.828	3.31	DR_ 4	<---	[BLK_ 4,STL_ 3,OR_ 4]
rule14	0.061	0.836	3.342	BLK_ 4	<---	[DR_ 4,STL_ 3,OR_ 4]
rule15	0.062	0.978	3.915	OR_ 4	<---	[BLK_ 4,DR_ 4,C]
rule16	0.062	0.863	3.45	DR_ 4	<---	[BLK_ 4,C,OR_ 4]
rule17	0.062	0.936	3.743	BLK_ 4	<---	[DR_ 4,C,OR_ 4]
rule18	0.063	0.84	3.365	OR_ 4	<---	[BLK_ 4,PF]
rule19	0.063	0.937	3.749	BLK_ 4	<---	[DR_ 4,C]
rule20	0.066	0.979	3.921	OR_ 4	<---	[DR_ 4,C]

Slika 10: Pravila

Iz ovog dela tabele možemo zaključiti:

- Ukoliko bek loše blokira ima mali broj ofanzivnih skokova
- Ukoliko krilni centar odlično blokira ima veliki broj ofanzivnih skokova
- Ukoliko centar odlično blokira i ima veliki broj ofanzivnih skokova onda ima i veliki broj defanzivnih skokova

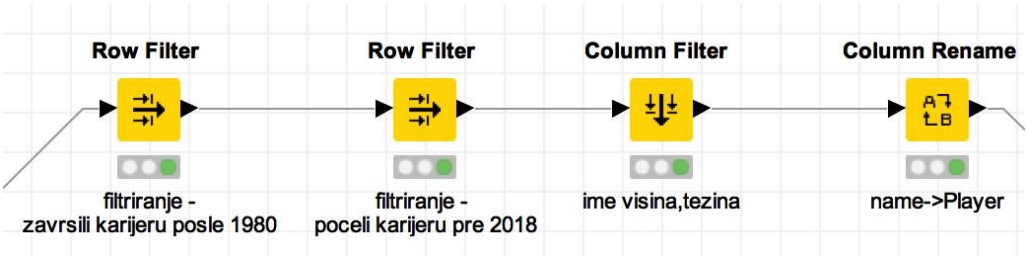
4 Klasifikacija

Opredelili smo se da nam pozicije igrača predstavljaju klase, a pripadnost određenoj klasi ćemo odrediti na osnovu karakteristika igrača poput visine i težine igrača i njihovih statistika, broja postignutih šuteva za 2 ili 3 poena, pokušaja za postizanje 3 poena, broja slobodnih bacanja, skokova, asistencija, blokova, prodatih lopti, faulova i ukupnog broja postignutih poena.

Za proces same klasifikacije bilo je neophodno spremiti podatke iz obe tabele kako bi bili pogodni za spajanje tabela i dalju analizu i primenu algoritama.

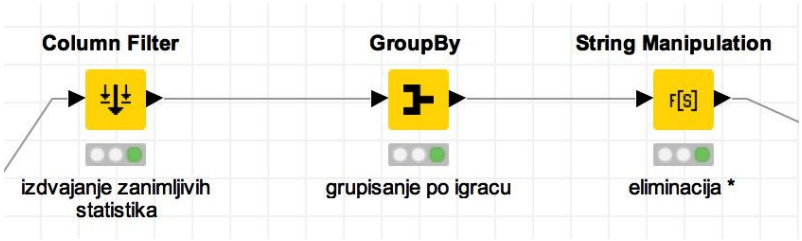
4.1 Priprema podataka za klasifikaciju

Iz tabele sa statistikama igrača smo eliminisali igrače koji su započeli karijeru pre 1979. nakon 2017. godine, pošto u tabeli sa sezonskim statistikama imamo podatke zaključno sa 2017. godinom. Preostalim igračima smo izdvojili ime, visinu i težinu. Dodatno, kolonu *name* smo preimenovali u *player* jer naziv atributa mora da se poklapa u svim tabelama koje učestvuju u spajanju.



Slika 11: Priprema tabele sa statistikama igrača

U koloni *player* postojali su podaci sa zvezdicom na kraju imena. Ovo je predstavljalo problem jer smo imali odvojene unose istog igrača. Korišćenjem čvora *String Manipulation* zvezdica je eliminisana, a nakon toga su unosi grupisani po igraču.



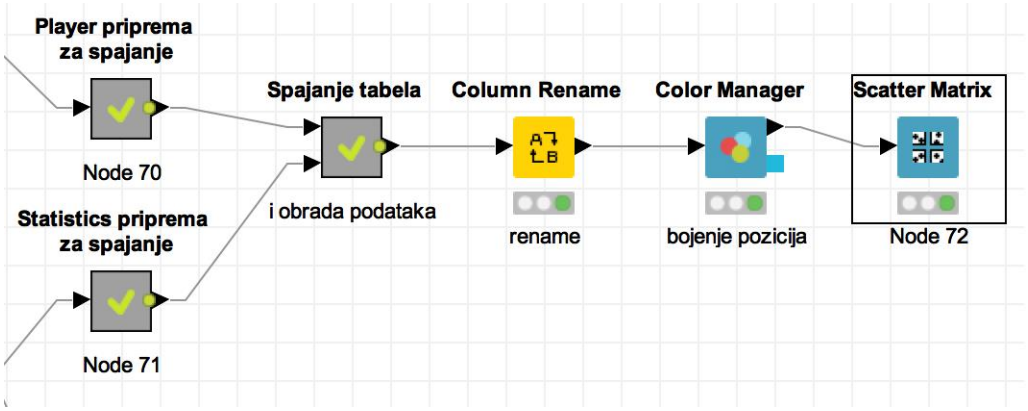
Slika 12: Priprema tabele sa sezonskm statistikama

Nakon priprema, spojili smo tabele sa *Concatenate* čvorom. Rezultujuća tabela ima 2812 redova i 15 kolona. Zatim smo eliminisali sve nedostajuce vrednosti i za igrače koji imaju duple pozicije (npr SG i PG) konvertovali poziciju tako da bude jedinstvena.

Slika 13: Spajanje tabela

Kada smo spojili tabele i dobili podatke u željenom formatu, preimenovali smo atribute tako da bolje oslikavaju karakteristike određenog igrača. Tako smo konačno dobili tabelu koja za svakog igrača izdvaja njegovo ime, tim, poziciju, kao i učinak u vidu postignutih poena, trojki i slično.

Korišćenjem *Color Manager* i *Scatter Matrix* čvorova grafički smo predstavili odnose između nekih atributa na osnovu čega smo započeli proces klasifikacije.



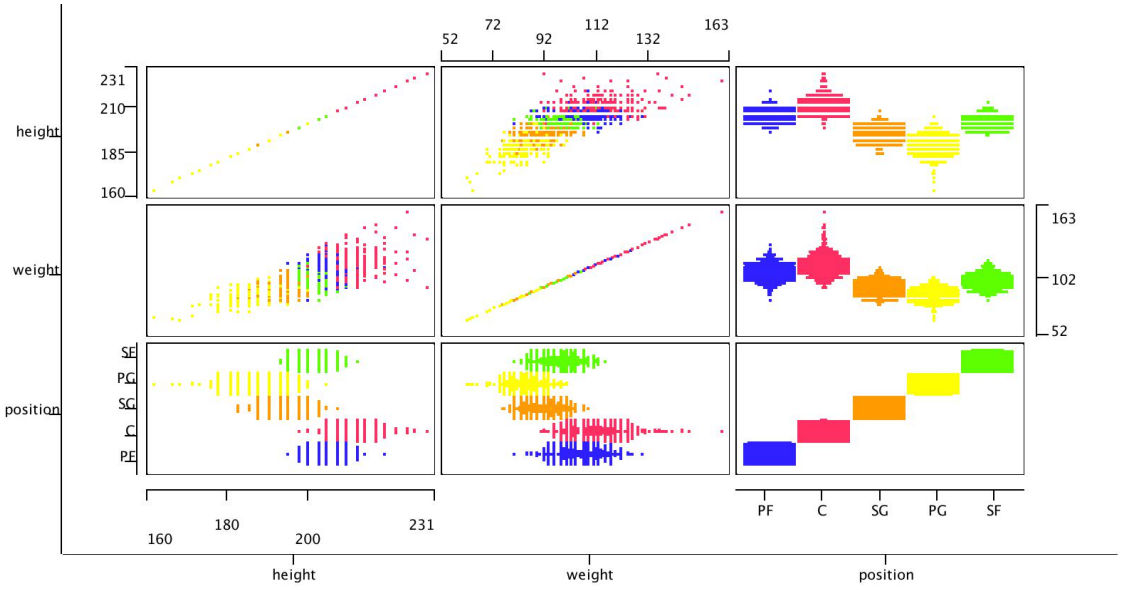
Slika 14: Priprema podataka za klasifikaciju

Table "default" – Rows: 2812 | Spec – Columns: 15 | Properties | Flow Variables

Player	height	weight	position	team	3P	3PA	2P	FT	TRB	AST	BLK	TOV	PF	PTS
Stefano Rusconi	205	109	PF	PHO	0	0	3	2	6	3	2	3	10	8
Steffond Johnson	203	109	PF	LAC	0	3	27	20	43	5	2	18	55	74
Stephane Lasme	203	98	PF	TOT	0	0	21.333	12.667	35.333	2	14.667	10	23.333	55.333
Stephen Curry	190	86	PG	GSW	239.625	547.375	334	249.125	312.75	489.5	15.125	227.25	180.125	1,636
Stephen Graham	198	98	SG	IND	6.111	20.222	38.222	19	50.111	13.778	2.333	18.667	42.556	113.778
Stephen Howard	205	102	SF	UTA	0	0	29.5	26.333	43	6	6.667	14.167	34.667	85.333
Stephen Jackson	203	99	SF	SAS	76.5	231.85	212.1	159.55	205.7	166.95	20.35	134.3	126.95	813.25
Stephen Thompson	193	84	SG	TOT	0	0.667	9.333	2	12.667	5.333	2	3.333	6	20.667
Stephen Zimmerman	213	109	C	ORL	0	0	10	3	35	4	5	3	17	23
Stephon Marbury	187	82	PG	NYK	67.294	207	329.471	255.647	171.824	448.647	8.118	174.118	135.941	1,116.471
Steve Alford	187	83	PG	DAL	9.167	27.167	64	29.5	36	44.667	2.167	22.333	28.333	185
Steve Bardo	195	86	SG	SAS	0.333	3.333	9	5.333	20	14.667	1.333	7.333	15	24.333
Steve Blake	190	78	PG	POR	68.368	180.053	69.526	32.368	122.684	239	4.474	83.211	85.053	376.526
Steve Bucknall	198	98	SG	LAL	0	1	9	5	7	10	1	11	10	23
Steve Burtt	187	84	SG	GSW	0.5	3.5	54	36.5	23	30.75	3.25	27.5	48.75	146
Steve Colter	190	75	PG	WSB	6.917	24.167	132	67.5	111	162.75	8.5	57.667	94	352.25
Steve Francis	190	88	PG	HOU	54.545	160.636	297.818	281.727	318.636	347.091	21.818	204	183.364	1,041
Steve Goodrich	208	100	PF	CHI	0.5	1.5	4	2.5	13	5.5	1.5	5.5	12	12
Steve Hamer	213	111	C	BOS	0	2	30	16	60	7	4	13	39	76
Steve Harris	195	88	SG	HOU	0.143	3.857	116.286	49.143	70.143	46.429	5.286	32.143	50.571	282.143
Steve Hawes	205	100	PF	ATL	2.714	9.714	176.571	95.286	331.286	106.143	19.857	90.714	167.429	456.571
Steve Hayes	213	93	C	TOT	0	0.143	41	15.857	91.857	16.143	17.429	16.857	72.714	97.857
Steve Henson	180	80	PG	MIL	15	34.714	21.286	18	25.143	66.429	0.286	25	40.429	105.571
Steve Johnson	208	107	C	KCK	0	0.286	227.929	132.071	280.071	62.5	43.143	119.429	198	587.929
Steve Kerr	190	79	PG	CHI	43.059	95.588	80.176	38.176	65	101.647	2.765	31.235	51.941	327.706

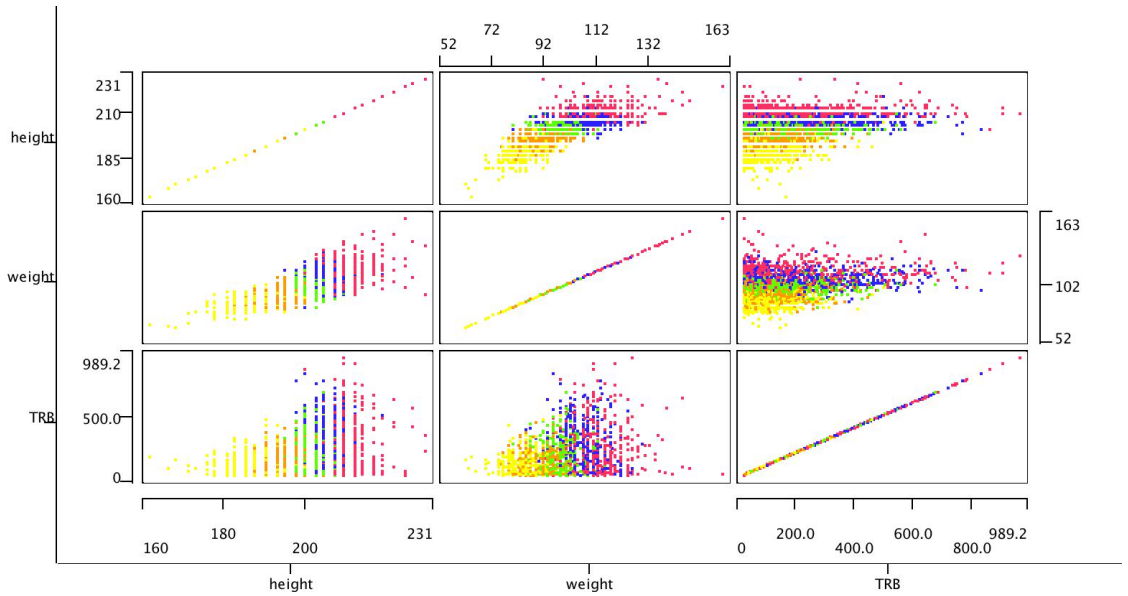
Slika 15: Tabela nad kojom je rađena klasifikacija

Uočavamo zanimljive korelacije između izdvojenih podataka.



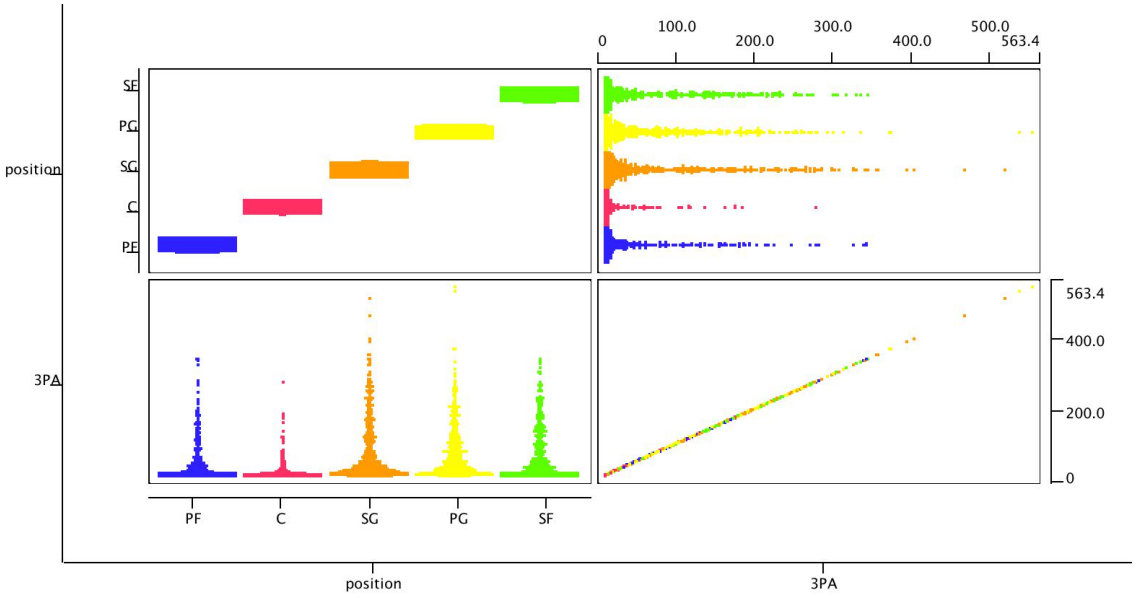
Slika 16: Korelacija visina-težina-pozicija

Ovde uočavamo da postoji jaka korelacija između visine i težine igrača i pozicije koju taj košarkaš igra. Na osnovu grafika jasno se vidi da su centri u najvećoj meri najviši i najteži igrači, a plejmejkeri najniži i najlakši, dok su igrači koji igraju pozicije krila i beka slične težine.



Slika 17: Korelacija visina-težina-skokovi

Sa najvećim brojem skokova prednjače centri, dok su tik iza njih krilni centri. Možemo videti da visina utiče na na to da li će neki igrač imati velik broj skokova, dok težina nema poseban uticaj na ovaj aspekt igre.



Slika 18: Korelacija pozicija-3 poena

Sa ovih grafika možemo zaključiti da najviše pokušaja za 3 poena imaju bekovi i plejmejkeri. Iako bekovi dosta šutiraju sa linije za tri poena, rekord u ovoj kategoriji drži nekolicina plejmejkera. Igrači koji igraju centra ubedljivo najmanje pokušavaju da postignu 3 poena.

4.2 Klasifikacija korišćenjem stabla odlučivanja

Kako ne bismo imali prepilagođen model, razdvojili smo skup na dva, test i trening skup. To smo učinili pomoću *Partitioning* čvora i to tako što smo 70% podataka iskoristili za trening a 30% za test podatke. Kako nam pozicije igrača predstavljaju tražene klase, uzimamo stratifikovane uzorke kako bismo najbolje klasifikovali podatke.

Klasifikaciju metodom stabla odlučivanja smo ostvarili primenom čvorova *Decision Tree Learner* i *Decision Tree Predictor*. Za dobijene rezultate smo posmatrali matricu konfuzije i statistiku preciznosti.

Row ID	PF	C	SG	PG	SF
PF	110	26	1	0	34
C	29	132	0	0	2
SG	2	0	122	35	23
PG	0	0	25	129	5
SF	29	0	36	0	104

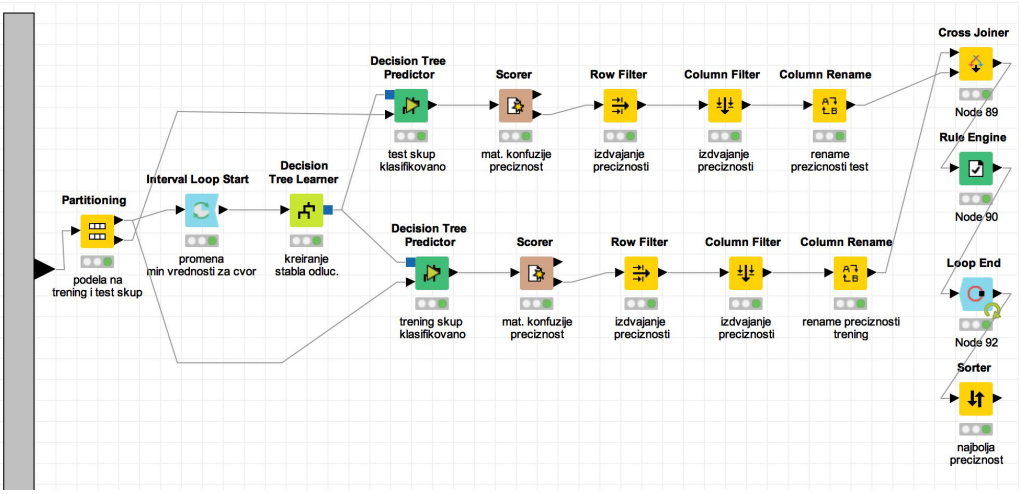
Slika 19: Matrica konfuzije za test skup

Analizom matrice konfuzije možemo primetiti kako se igrači koji igraju na poziciji centra odlično klasifikuju a kad se loše klasifikuju, pridružuje im se klasa krilnog centra. Iz ovog je odmah jasno da atribut visine ima veliku ulogu u klasifikaciji ovog skupa kako su igrači na ove dve pozicije relativno slične visine.

Takođe igrači na poziciji plejmejkera se dobro klasifikuju i slično važi kao u prethodnom primeru, a to je da ako se ne klasifikuju kako treba, dodeliće im se klasa bekova. Iz ovog izvlačimo identičan zaključak.

Ono što nam predstavlja problem jeste klasifikacija igrača na poziciji krila gde imamo veliki broj igrača (procentualno od ukupnog broja) koji je pogrešno klasifikovan na pozicije beka i krilnog centra.

Nakon detaljne analize smo izvukli preciznost pogađanja i spojili skupove u jedan čvor kako bismo bolje analizirali podatke.



Slika 20: Shema za stablo odlučivanja

Da bismo dobili što bolju preciznost ovaj proces smo ponavljali u petlji, tako što smo menjali minimalan broj igrača koji će se naći u čvoru stabla odlučivanja, kako se stablo ne bi granalo dalje tj. menjali smo kriterijum zaustavljanja grananja stabla od 5 do 100 sa korakom 5.

Mere koje su razmatrane su Ginijev indeks i odnos dobiti.

4.2.1 Ginijev indeks

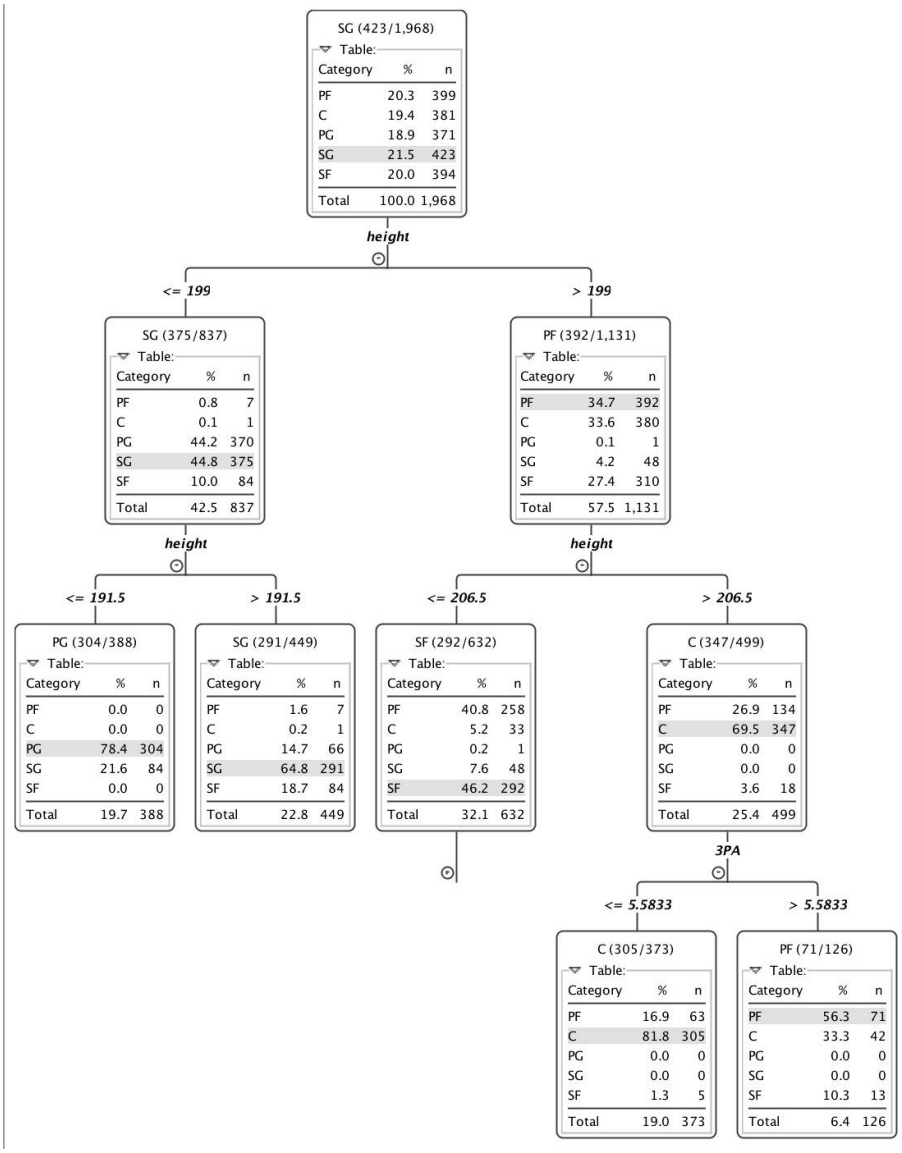
Ginijev indeks predstavlja meru nečistoće koja se koristi za izračunavanje dobiti nekog čvora. Izračunava se sledećom formulom:

$$Gini(t) = 1 - \sum_j [p(j|t)]^2$$

Korišćenjem ove mere dobili smo drvo odlučivanja predstavljeno na slici 21.

Ono što smo intuitivno i preko grafika u prethodnoj glavi zaključili samo je potvrđeno ovim stablom odlučivanja, a to je činjenica da visina najviše utiče na poziciju igrača. Zbog ovoga je prvih par nivoa stabla upravo podela igrača po visini.

Na nižim nivoima, bitne podele stabla dešavaju se na osnovu broja pokušaja za tri poena, gde se vidi da igrači na poziciji krilnog centra više šutiraju za tri poena nego centri. Analogno ovome, podela između pozicija plejmejkera i beka se vrši na osnovu broja asistencija igrača.



Slika 21: Stablo dobijeno korišćenjem Ginijevog indeksa

Kako možemo primetiti na slici 23 nakon 20 iteracija izvršavanja algoritma sa različitim vrednostima za minimalan broj igrača u listu stabla odlučivanja, optimalan broj za granicu grananja je između 20 i 30 instanci, kako za taj broj preciznost dostiže svoj maksimum.

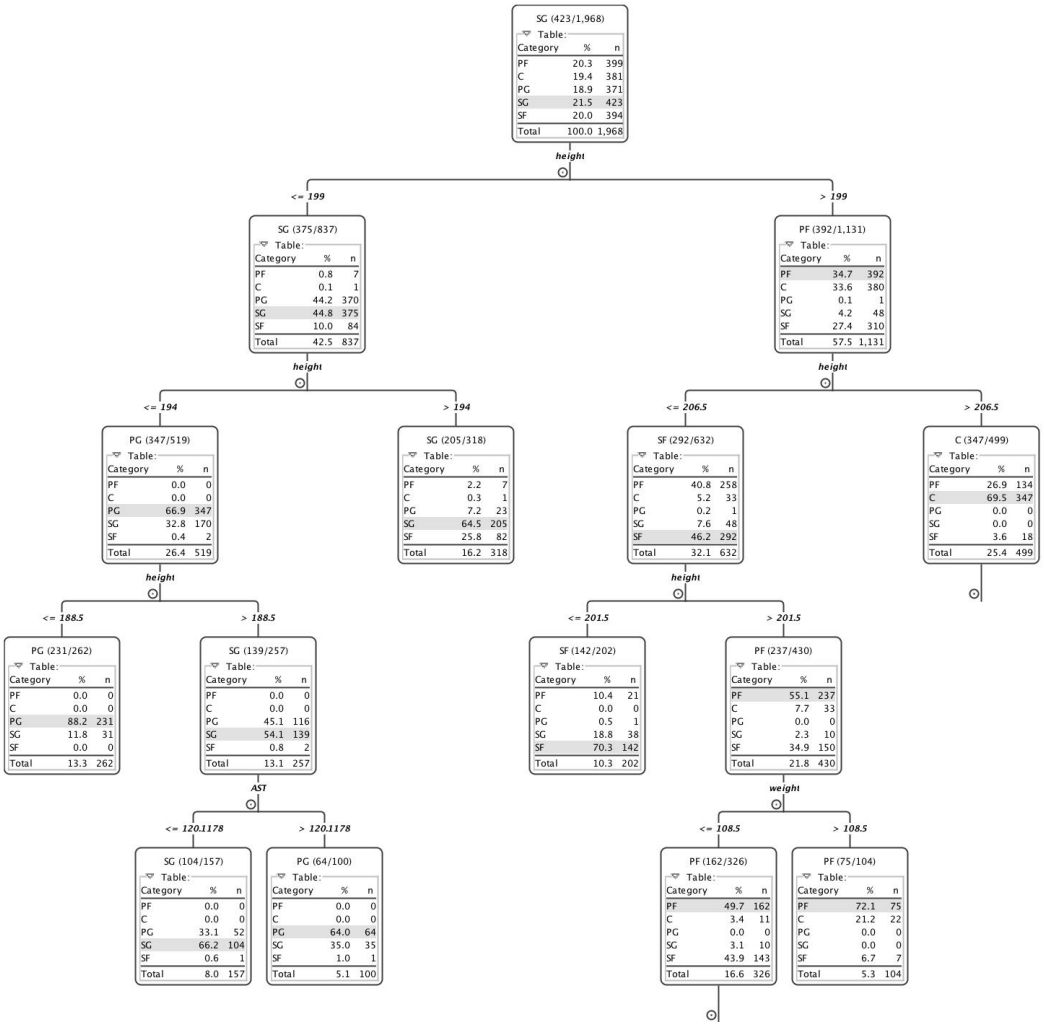
4.2.2 Odnos dobiti

Odnos dobiti je mera koja se koristi za određivanje valjanosti podele stabla odlučivanja. Definisana je sledećom formulom:

$$GainRatio_{split} = \frac{Gain_{split}}{SplitInfo}$$

$$SplitInfo = -\sum_{i=1}^{k} \frac{n_i}{n} log \frac{n_i}{n}$$

Stablo odlučivanja je jako slično slablu odlučivanja koje se generisalo prilikom primene mere Ginijevog indeksa, pa nismo ulazili u detaljnu analizu ovog stabla.



Slika 22: Stablo dobijeno korišćenjem dobiti

Analiza dobijenih rezultata preciznosti prikazane na slici 24, rađena je na sličan način kao za Ginijev indeks. Nakon 20 iteracija izvršavanja algoritma sa različitim vrednostima za minimalan broj instanci u listovima, zaključili smo da je u ovom slučaju optimalna granica za broj čvorova 40 iako se algoritam dobro ponaša za vrednosti 30 ili 15.

Upoređivanjem najboljih vrednosti iz jedne i druge tabele preciznosti zaključili smo da je Ginijev indeks za nijansu bolji i da maksimalna preciznost koju na našem skupu metoda stabla odlučivanja može da dostigne iznosi oko 0.7.

Row ID	AccuracyTrening	AccuracyTest	minPvalue	Iteration
Overall_Ov...	0.755	0.707	30	5
Overall_Ov...	0.759	0.706	25	4
Overall_Ov...	0.768	0.704	20	3
Overall_Ov...	0.748	0.699	35	6
Overall_Ov...	0.729	0.698	55	10
Overall_Ov...	0.735	0.694	40	7
Overall_Ov...	0.733	0.691	45	8
Overall_Ov...	0.731	0.691	50	9
Overall_Ov...	0.706	0.69	85	16
Overall_Ov...	0.715	0.687	70	13
Overall_Ov...	0.715	0.687	75	14
Overall_Ov...	0.715	0.685	80	15
Overall_Ov...	0.707	0.685	90	17
Overall_Ov...	0.707	0.685	95	18
Overall_Ov...	0.707	0.685	100	19
Overall_Ov...	0.725	0.684	60	11
Overall_Ov...	0.725	0.684	65	12
Overall_Ov...	0.721	0.667	10	1
Overall_Ov...	0.726	0.666	15	2
Overall_Ov...	0.686	0.665	5	0

Slika 23: Statistike iteracija dobijenih korišćenjem Ginijevog indeksa

Row ID	Accur...	Accur...	minPv...	Iteration
Overall_Ov...	0.752	0.705	40	7
Overall_Ov...	0.753	0.698	30	5
Overall_Ov...	0.78	0.695	15	2
Overall_Ov...	0.778	0.695	10	1
Overall_Ov...	0.798	0.694	5	0
Overall_Ov...	0.758	0.694	25	4
Overall_Ov...	0.752	0.694	35	6
Overall_Ov...	0.76	0.691	20	3
Overall_Ov...	0.713	0.69	90	17
Overall_Ov...	0.713	0.688	85	16
Overall_Ov...	0.711	0.688	100	19
Overall_Ov...	0.713	0.687	95	18
Overall_Ov...	0.722	0.684	80	15
Overall_Ov...	0.73	0.674	75	14
Overall_Ov...	0.734	0.673	70	13
Overall_Ov...	0.741	0.661	50	9
Overall_Ov...	0.738	0.66	45	8
Overall_Ov...	0.734	0.66	65	12
Overall_Ov...	0.741	0.659	55	10
Overall_Ov...	0.734	0.652	60	11

Slika 24: Statistike iteracija dobijenih korišćenjem odnosa dobiti

4.3 Klasifikacija metodom K najbližih suseda

Algoritam K najbližih suseda zasniva se na ideji da se isto klasifikuju k vrednosti koje su bliske po nekom rastojanju npr. Euklidskom.

Ovaj algoritam ne radi sa kategoričkim vrednostima, a kako nema smisla pretvarati kategorički atribut *tim* u numerički, eliminisali smo ga. Zatim smo particionisali podatke na test i trening skup na isti način kao i pri korišćenju stabla odlučivanja.

U petlji smo menjali broj suseda od 4 do 100 sa korakom 2 i računali preciznost klasifikacije korišćenjem ove metode. Ovo smo uradili kako bismo dobili najbolju moguću preciznost tj. da bismo pogodili optimalan broj suseda za naše podatke. Za primenu algoritma koristili smo čvor *K Nearest Neighbor*.

Nakon 49 iteracija zaključili smo da je najbolja preciznost ukoliko se za broj suseda izabere između 10 i 14 suseda kao što se vidi na slici 26. Maksimalna preciznost koja se dostiže u ovom skupu je 0.666, što je lošije od preciznosti dobijene metodom stabla odlučivanja.

Row ID	PF	C	SG	PG	SF
PF	102	40	8	0	21
C	71	87	1	0	4
SG	7	1	129	19	26
PG	0	0	34	124	1
SF	41	3	41	1	83

Slika 25: Matrica konfuzije za metodu k najbližih suseda

Posmatrajući matricu konfuzije na test skupu možemo primetiti da se od 169 igrača koji igraju na poziciji krila samo 83 dobro klasifikovalo dok je 41 igrač pogrešno klasifikovan na pozicije beka ili krilnog centra. To nam govori da su igrači koji igraju na poziciji krila po stilu igre i po naprednim statistikama često vrlo slični igračima na pozicijama beka ili krilnog centra.

Takođe možemo primetiti da se slična situacija dešava u drugom redu gde je od ukupnog broja centara skoro pola klasifikovano pogrešno na poziciju krilnog centra. Analogno centri često mogu da igraju slično kao krilni centri.

Row ID	Accur...	k	Iteration
Overall#4	0.666	12	4
Overall#5	0.655	14	5
Overall#3	0.649	10	3
Overall#23	0.649	50	23
Overall#2	0.647	8	2
Overall#9	0.645	22	9
Overall#10	0.643	24	10
Overall#24	0.637	52	24
Overall#21	0.636	46	21
Overall#25	0.636	54	25
Overall#22	0.635	48	22
Overall#1	0.634	6	1
Overall#7	0.634	18	7
Overall#26	0.634	56	26
Overall#20	0.633	44	20
Overall#8	0.63	20	8
Overall#19	0.63	42	19
Overall#11	0.628	26	11
Overall#27	0.628	58	27
Overall#28	0.627	60	28
Overall#33	0.626	70	33
Overall#13	0.624	30	13
Overall#18	0.624	40	18
Overall#6	0.623	16	6
Overall#17	0.623	38	17
Overall#0	0.622	4	0
Overall#14	0.622	32	14
Overall#32	0.621	68	32
Overall#12	0.62	28	12
Overall#31	0.62	66	31

Slika 26: Statistike iteracija dobijenih metodom k najbližih suseda

4.4 Klasifikacija naivnim Bajesovim algoritmom

Bajesovska klasifikacija zasniva se na rešavanju problema klasifikacije primenom verovatnosnih i statističkih metoda i teorema kao što su Bajesova teorema i teorema o uslovnoj verovatnoći.

Bajesova teorema:

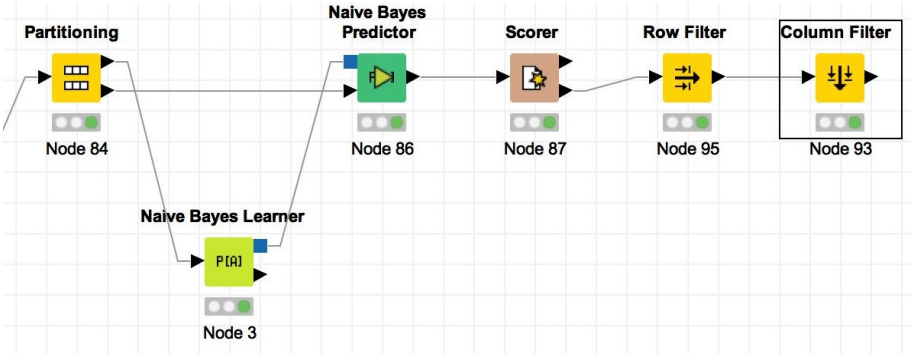
$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Particionisanjem skupa na test i trening skup, a zatim primenom čvorova *Naive Bayes Learner* za kreiranje modela odnosno *Naive Bayes Predictor* za klasifikovanje podataka, dobijamo sličnu matricu konfuzije kao i preciznost klasifikacije test skupa kao što smo dobili primenom algoritma k najblizih suseda.

Row ID	Accuracy
Overall	0.678

Slika 27: preciznost

Naivni Bajesov algoritam primenjen nad podacima iz našeg skupa u alatu KNIME može se videti na slici 28.



Slika 28: Čvorovi klasifikacije Bajesovom metodom

4.5 Klasifikacija metodom potpornih vektora (SVM)

Metod potpornih vektora zasniva se na ideji nalaženja razdvajajuće hiper ravni tako da su svi podaci date klase sa iste strane ravni. Kako i u našem skupu podaci nisu linearno separabilni ideja je da se osnovni vektorski prostor preslika u neki višedimenzioni prostor u kome je skup podataka za trening linearno razdvojiv. To se postiže uvođenjem Kernel funkcije koja odgovara skalarnom proizvodu u nekom višedimenzionom prostoru. Pošto se ovaj algoritam u teoriji dobro ponaša za numeričke podatke pretpostavili smo da će nam on dati preciznije rezultate.

Pre particionisanja našeg skupa na test i trening skup, podaci se moraju normalizovati, jer su normalizovani podaci jedan od uslova za primenu ovog metoda klasifikacije. To smo uradili primenom čvora *Normalizer*.

Nakon particionisanja, primenom *SVM Learner* i *SVM Predictor* čvorova dobili smo klasifikovane podatke kako za trening tako i za test skup. Da bismo dobili podatke koje su najbolje moguće klasifikovani koristili smo različite kernele i upoređivali rezultate.

4.5.1 Polinomijalan kernel

Polinomijalan kernel definisan je sledećom formulom:

$$K(x, y) = (x^{\mathsf{T}} y + c)^d$$

Primenom ovog kernela preciznost kojom se trening i test skup klasifikuju je jako loša, kao što se može videti na slici 29, stoga ovaj kernel nećemo koristiti.

Row ID	AccuracyTrening	AccuracyTest
Overall_Ov...	0.551	0.54

Slika 29: Preciznost primenom polinomijalnog kernela

4.5.2 Sigmid kernel

HiperTangent kernel odnosno Sigmid kernel definisan je sledećom formulom:

$$K(x, y) = tanh(\gamma x^{\mathsf{T}} y + r)$$

Primenom ovog kernela preciznost kojom se trening i test skup klasifikuju je jako loša, kao što se može videti na slici 30, stoga ni ovaj kernel nećemo koristiti.

Row ID	AccuracyTrening	AccuracyTest
Overall_Ov...	0.466	0.462

Slika 30: Preciznost primenom sigmid kernela

4.5.3 Gausov (RBF) kernel

Gausov kernel, odnosno RBF (en. *Radial basis function*) kernel, definisan je sledećom formulom:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

Primenom ovog kernela preciznost kojom se test odnosno trening skup klasifikuju je jako velika, znatno veća od SVM algoritma sa prethodna dva kernela, ali i znatno veća od one koja se dobija primenom metode k najbližih suseda, metode stabla odlučivanja kao i metode Bajesovske klasifikacije što se ogleda i na slici 31.

Row ID	AccuracyTrening	AccuracyTest
Overall_Ov...	0.874	0.72

Slika 31: Preciznost primenom RBF kernela

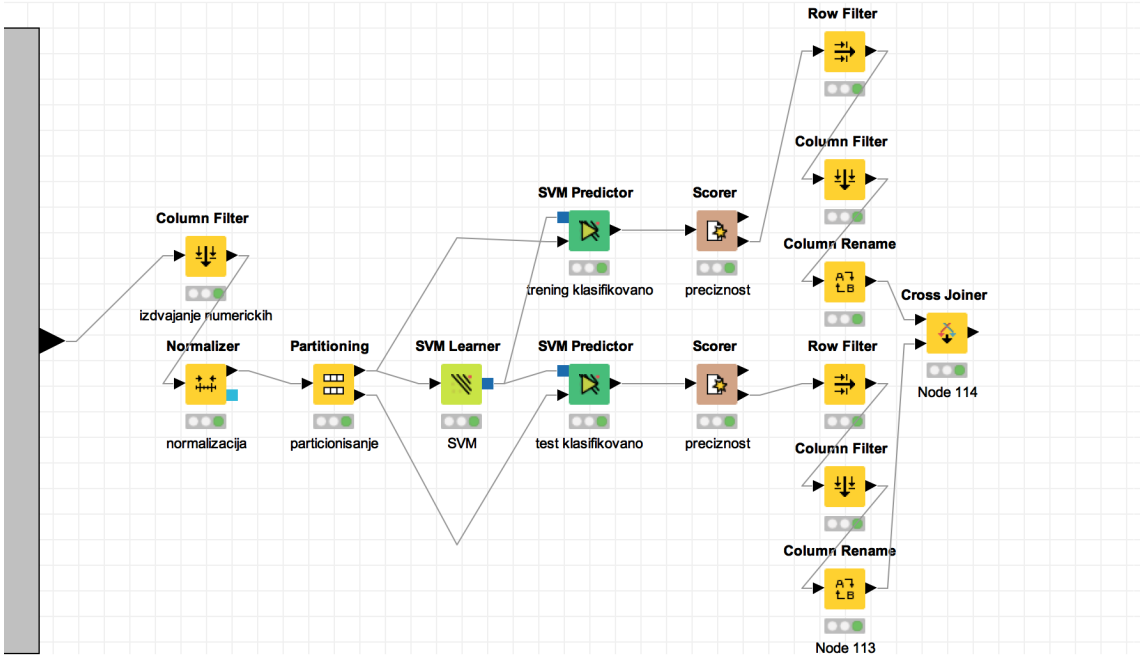
Analizom matrice konfuzije zaključujemo da se pri primeni SVM klasifikacije osetno bolje klasifikuju svi podaci, a posebno igrači koji igraju krila. U ovom slučaju se krila takođe nekad zamene za krilne centre ili bekove ali u znatno manjem broju. Taj broj je sada 20 za bekove odnosno 27 za krilne centre, što implicira da se znatno manje igrača na poziciji krila loše klasifikovalo nego npr. prilikom primene k najbližih suseda.

Ne samo što se poboljšanje vidi na poziciji krila nego je i pozicija centra znatno bolje klasifikovana. Loše se klasifikuje samo 36 igrača koji su centri za razliku od 71 igrača koliko se loše klasifikovalo na ovoj poziciji korišćenjem algoritma k najbližih suseda.

Row ID	PF	C	SG	PG	SF
PF	109	31	1	4	26
C	31	127	0	4	1
SG	2	0	123	30	27
PG	0	0	23	135	1
SF	27	2	20	6	114

Slika 32: Matrica konfuzije nad test skupom primenom SVM metode (Gausov kernel)

Kako bismo dobili podatke o preciznosti u formatu koji je prikazan na prethodnim stranama koristili smo kombinaciju *Row Filter*, *Column Filter* i *Column Rename* čvorova. Ukupna SVM klasifikacija implementirana koristeći KNIME može se videti na slici 33.



Slika 33: SVM čvorovi

5 Klaster analiza

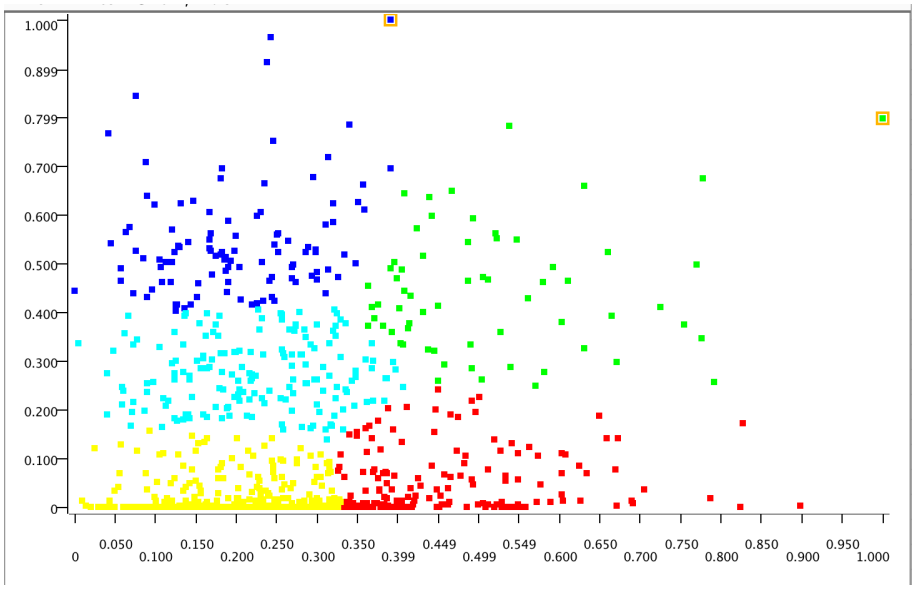
Klasterovanje predstavlja pronalaženje grupa objekata takvih da su objekti u grupi međusobno slični, odnosno da su objekti u razlicitim grupama medjusobno različiti.

Kako bismo posmatrali grupe igrača u odnosu na to koliko šutiraju i pogađaju šuteve za 3 poena, kao i u odnosu na to koliko šutiraju sa linije za slobodna bacanja morali smo prethodno pripremiti podatke na sličan način kao i za klasifikaciju korišćenjem čvorova *Row filter*, *Column filter* i *String manipulation*.

Kako bismo mogli da koristimo metodu k sredina za klasterovanje moramo prvo normalizovati podatke da bi broj postignutih trojki i broj postignutih slobodnih bacanja imali isti uticaj na distancu između instanci.

Nakon izvršavanja metode k sredina u 100 iteracija primenom čvora *k-Means* dobili smo klastere koji se vide na slici 34.

Pošto se kvalitet dobijenih rezultata razlikuje u odnosu na broj klastera, u petlji smo menjali broj klastera od 2 do 10 i prikazivali odgovarajuće grafike. U slučaju vrednosti broja klastera jednakih 5 dobijaju se najsmisleniji rezultati za tumačenje.



Slika 34: Klasterovanje

Igrači koji su pridruženi klasteru žute boje predstavljaju igrače koji imaju malo pogodaka za 3 poena i malo do umereno pogođenih slobodnih bacanja. U ovoj grupaciji se većinski nalaze igrači koji igraju na poziciji centra.

Igrači koji su pridruženi u crvenom klasteru predstavljaju igrače koji retko pogađaju šuteve za 3 poena ali su često faulirani i imaju veliki procenat pogođenih slobodnih bacanja.

Igrači koji su pridruženi klasteru plave boje predstavljaju one koji idu slabo do umereno na liniju za slobodna bacanja ali pogađaju mnogo šuteva za 3 poena. Košarkaš iz ovog skupa koji se izdvaja je Stephen Curry koji drži rekord za broj pogođenih trojki u istoriji NBA lige.

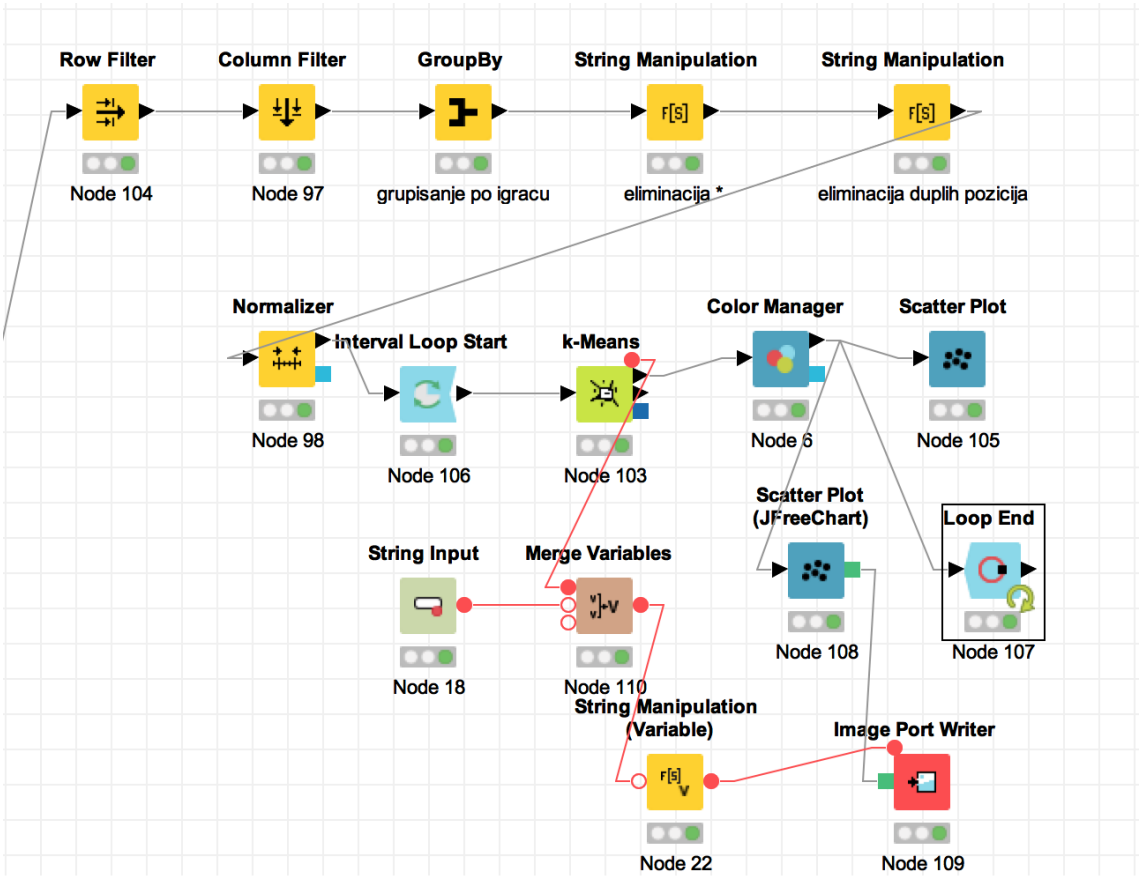
Zeleni klaster predstavlja igrače koji često idu na liniju za slobodna bacanja ali odlično šutiraju i za 3 poena. Košarkaš koji se izdvaja iz ovog klastera je James Harden koji je poznat kao odličan šuter ali i igrač koji najčešće iznudi prekršaj i ide na liniju za slobodna bacanja.

Row ID	S Player	D Mean(3P)	D Mean(3PA)	D Mean(FT)	D Mean(FTA)
Row730	Stephen Curry	1	0.897	0.392	0.369

Slika 35: Stephen Curry

Row ID	S Player	D Mean(3P)	D Mean(3PA)	D Mean(FT)	D Mean(FTA)
Row356	James Harden	0.798	0.866	1	1

Slika 36: James Harden



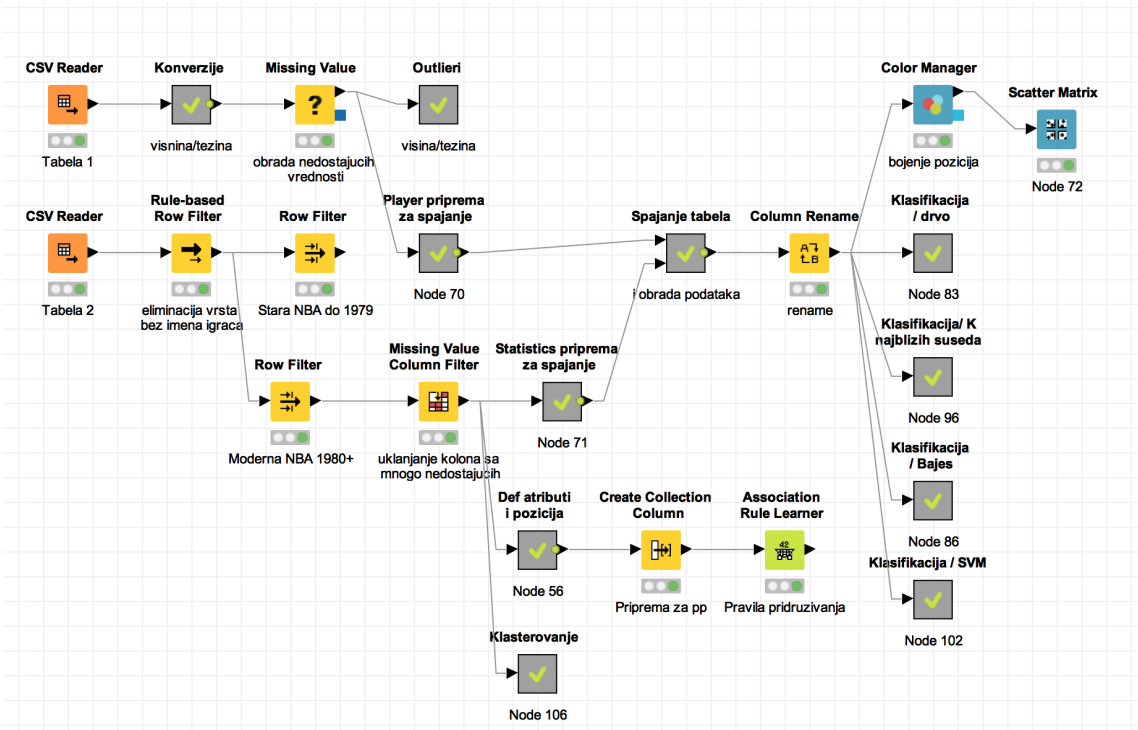
Slika 37: Klasterovanje čvorovi

6 Zaključak

U ovom radu smo prikazali osnovne metode i tehnike za istraživanje podataka, od pripreme, obrade i analize podataka, do implementacije konkretnih algoritama za klasifikaciju poput stabla odlučivanja, metode k najbližih suseda, naivnog Bajesovog algoritma, potpornih vektora, i klasterovanje nad tim podacima.

Kako je naša baza pored osnovnih karakteristika igrača sadržala i statistike o performansama tokom sezona, fokus u ovom radu je bio uočiti da li i kako te performanse utiču na poziciju koju neki košarkaš igra i obrnuto.

Na slici 38 se mogu videti sve nabrojane metode implementirane u programu KNIME.



Slika 38: Svi čvorovi