ATP 2000-2017

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Marija Mijailović
mi14199@alas.matf.bg.ac.rs

Miroslav Mišljenović
mr12260@alas.matf.bg.ac.rs

jun 2018.

Sažetak

U ovom radu analizirali smo skup podataka ˝ATP - rezultati turnira od 2000-2017˝. Obradili smo pravila pridruživanja, klasterovanje, klasifikaciju i predstavili sve navedene metode odgovarajućom vizualizacijom. Skup podataka je preuzet sa https://www.kaggle.com/gmadevs/atp-matches-dataset.

Sadržaj

1 Uvod **2**

2 Analiza podataka **2**

3 Pravila pridruživanja **3**

4 Klasterovanje **5**
4.1 SPSS 6
4.2 KNIME 10

5 Klasifikacija **12**
5.1 SPSS 14
5.2 KNIME 18
5.2.1 Drveta odlučivanja 18
5.2.2 K najbližih suseda 22
5.2.3 SVM 24

1 Uvod

Skup podataka ATP mečeva podeljen je u 17 zasebnih .csv fajlova i svaki od njih prikazuje individualne statistike za svaki turnir u toku te godine.

2 Analiza podataka

U ovom poglavlju sledi kratak pregled najistaknutijih atributa ovog skupa podataka. Svaki red u skupu, označava jedan meč i sve informacije o tom meču.

U tabeli 1 prikazani su podaci o turniru.

Ime kolone	Objašnjenje
tourney_id	id turnira
tourney_name	ime turnira
surface	podloga(Grass, Clay, Hard)
tourney_level	nivo turnira(Grand Slam, Finals, Masters, Tour Series, Challenger)
round	runda(Round of 16, Quarterfinal...)
minutes	trajanje meča u minutima

Tabela 1: Podaci o turnirima

U tabeli 2 prikazani su podaci o pobedniku meča.

Ime kolone	Objašnjenje
winner_seed	nosilac na turniru
winner_entry	ulaznica(WildCard, Qualified, LuckyLoser, ProtectedRanking)
winner_name	ime pobednika
winner_ht	visina pobednika
winner_ioc	zemlja porekla pobednika
winner_age	godine pobednika
winner_rank	ATP rang pobednika
winner_rank_points	ATP poeni pobednika
w_ace	broj asova pobednika
w_df	broj duplih grešaka pobednika
w_svpt	broj poena dobijenih na servis pobednika
w_1stIn	broj ubačenih prvih servisa pobednika
w_1stWon	broj poena dobijenih nakon ubačenog prvog servisa pobednika
w_2ndWon	broj poena dobijenih nakon ubačenog drugog servisa pobednika
w_SvGms	broj gemova u kojima je servirao pobednik
w_bpSaved	broj spašenih brejk šansi pobednika
w_bpFaced	broj brejk šansi na servis pobednika

Tabela 2: Podaci o pobednicima

U tabeli 3 prikazani su podaci o gubitniku meča.

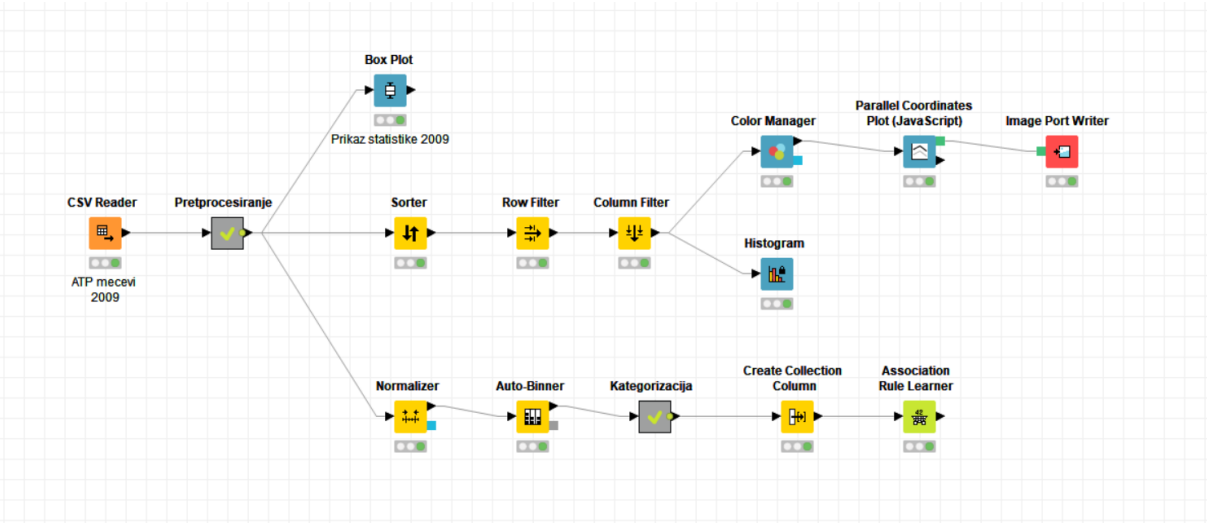
Ime kolone	Objašnjenje
loser_seed	nosilac na turniru
loser_entry	ulaznica(WildCard, Qualified, LuckyLoser, ProtectedRanking)
loser_name	ime gubitnika
loser_ht	visina gubitnika
loser_ioc	zemlja porekla gubitnika
loser_age	godine gubitnika
loser_rank	ATP rang gubitnika
loser_rank_points	ATP poeni gubitnika
l_ace	broj asova gubitnika
l_df	broj duplih grešaka gubitnika
l_svpt	broj poena dobijenih na servis gubitnika
l_1stIn	broj ubačenih prvih servisa gubitnika
l_1stWon	broj poena dobijenih nakon ubačenog prvog servisa gubitnika
l_2ndWon	broj poena dobijenih nakon ubačenog drugog servisa gubitnika
l_SvGms	broj gemova u kojima je servirao gubitnik
l_bpSaved	broj spašenih brejk šansi gubitnika
l_bpFaced	broj brejk šansi na servis gubitnika

Tabela 3: Podaci o gubitnicima

S obzirom na veliki broj raspoloživih godina, prvo smo se detaljno upoznali sa podacima, šta nam koja godina pruža i koji su najzanimljiviji atributi za svaku godinu. U zavisnosti od toga smo, po potrebama metoda, koristili različite godine, ali svuda smo se ograničili na četiri maksimalno.

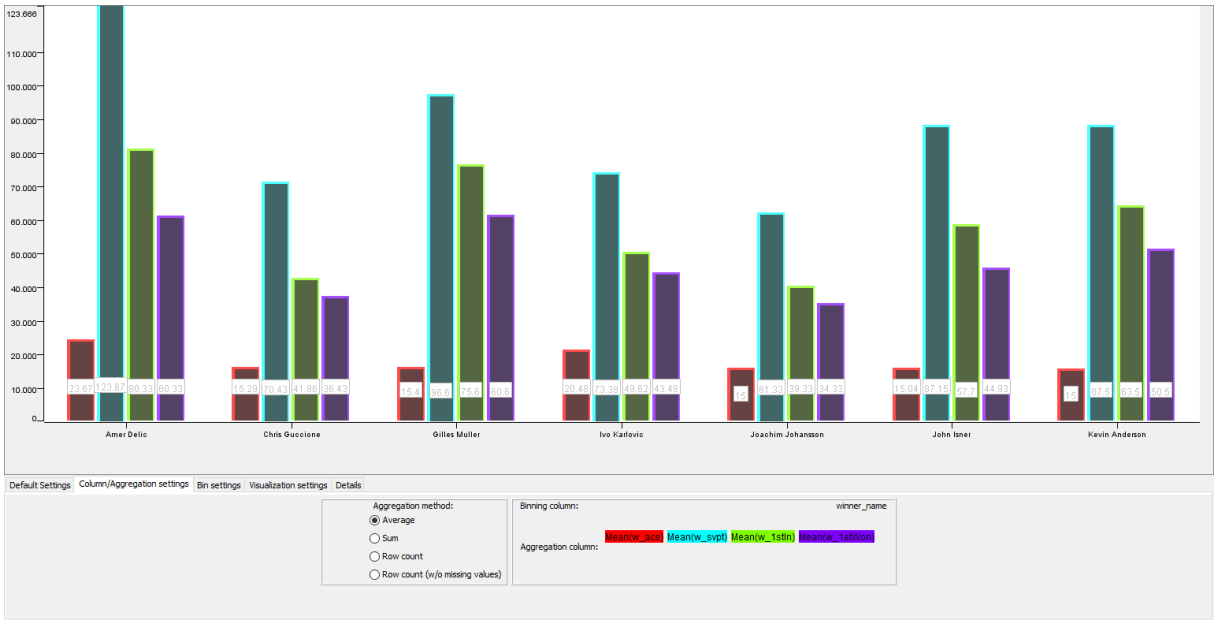
3 Pravila pridruživanja

Pravila pridruživanja smo obradili u programskom alatu KNIME (slika 1). Odlučili smo se za 2009. godinu, jer su rezultati reprezentativniji u odnosu na ostale godine.

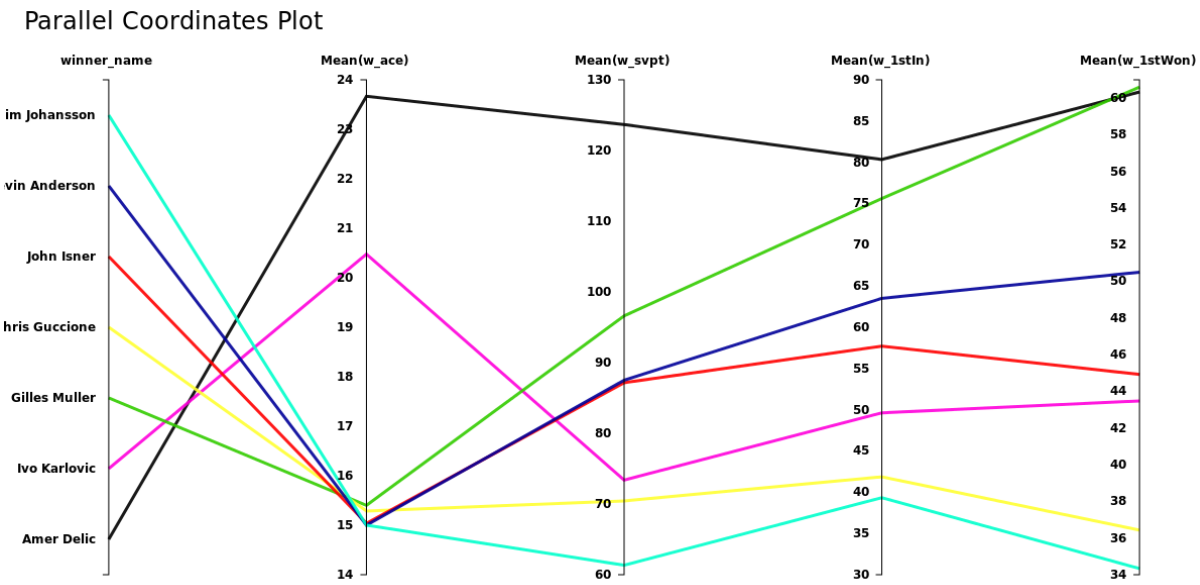


Slika 1: KNIME implementacija

Na slikama 2 i 3 grafički su prikazani rezultati za sedam tenisera koji su imali prosečno najviše asova po meču na kome su pobedili. Izabrali smo četiri parametra za svakog igrača: broj asova pobednika, broj dobijenih poena na servis pobednika, broj ubačenih prvih servisa pobednika i broj osvojenih poena nakon ubačenog prvog servisa pobednika. Na histogramu i grafiku paralelnih koordinata mogu se videti i uporediti rezultati.



Slika 2: Histogram



Slika 3: Paralelne koordinate

Iznenađenje je pojavljivanje Amera Delića u prvih sedam, jer je to autorima nepoznat igrač. Uvidom u podatke, utvrđeno je da je on te godine odigrao samo osam mečeva, a pobedio je samo tri puta, u mečevima u kojima je imao mnogo asova (što je kriterijum po kome je birano najboljih sedam).

U tri kategorije smo podelili sledeća četiri atributa: broj asova pobednika, broj duplih servis grešaka pobednika, broj osvojenih poena nakon ubačenog prvog servisa pobednika, broj spašenih brejk lopti pobednika. Na slici 4 se mogu videti pravila pridruživanja dobijena na osnovu te kategorizacije, sortirani po Lift meri. Za pouzdanost smo uzeli vrednost 0.4, a za minimalnu podršku vrednost 0.15. Analizirali smo podatke za sve godine i rezultati su prilično uniformni. Za 2009. godinu je dobijena druga najveća Lift mera (1.36) i odnosi se na pravilo [ACE 2, WON 2, DF 1] -> [BPS 1]. U 2004. godini smo dobili najveću vrednost Lift mere (1.441) za pravilo [BPS 1, ACE 1, DF 1] -> [WON 1].

Frequent itemsets/Association rules - 2:3 - Association Rule Learner

File Hilite Navigation View

Table "default" - Rows: 26 | Spec - Columns: 6 | Properties | Flow Variables

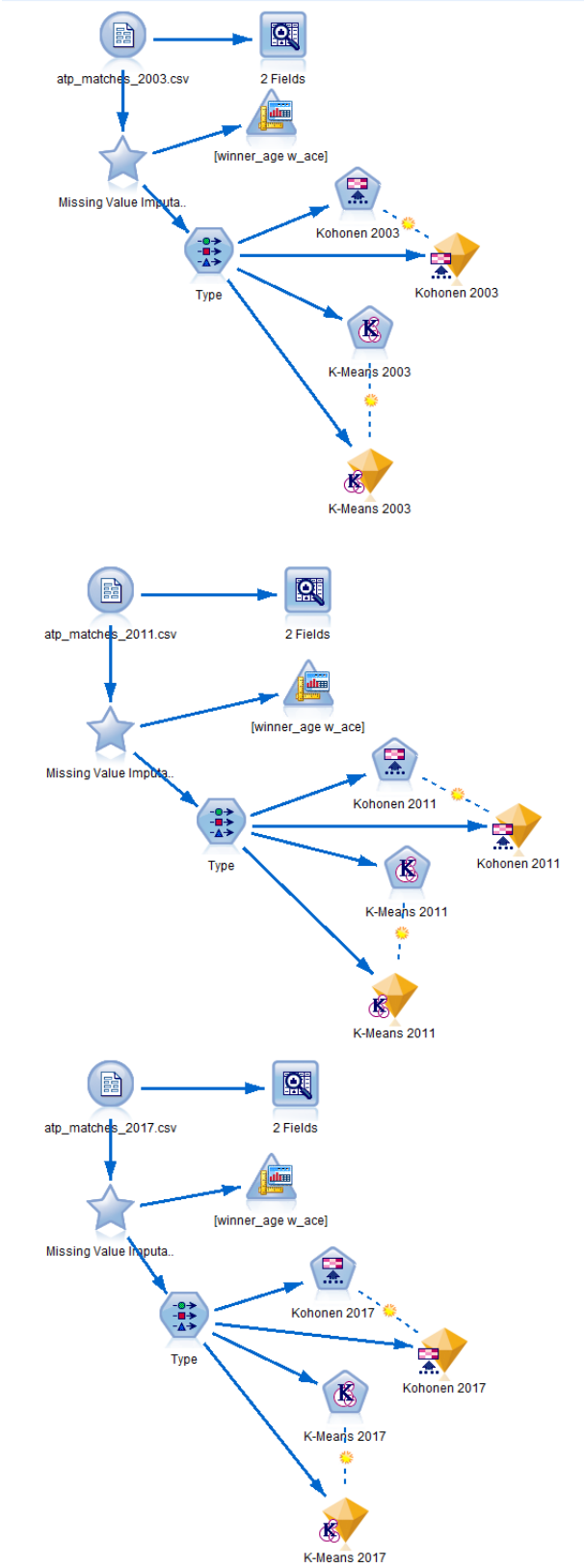
Row ID	Support	Confide...	Lift	Conseq...	implies	Items
rule2	0.177	0.837	1.36	BPS 1	<---	[ACE 2,WON 2,DF 1]
rule4	0.197	0.816	1.326	BPS 1	<---	[ACE 2,DF 1]
rule6	0.207	0.857	1.279	ACE 1	<---	[BPS 2,WON 2,DF 1]
rule10	0.217	0.83	1.239	ACE 1	<---	[BPS 2,DF 1]
rule7	0.207	0.955	1.174	WON 2	<---	[BPS 2,ACE 1,DF 1]
rule13	0.241	0.925	1.137	WON 2	<---	[BPS 2,DF 1]
rule14	0.246	0.909	1.118	WON 2	<---	[BPS 2,ACE 1]
rule0	0.177	0.9	1.107	WON 2	<---	[BPS 1,ACE 2,DF 1]
rule1	0.177	0.878	1.1	DF 1	<---	[BPS 1,ACE 2,WON 2]
rule17	0.325	0.868	1.088	DF 1	<---	[BPS 1,ACE 1]
rule18	0.419	0.867	1.087	DF 1	<---	[BPS 1,WON 2]
rule21	0.532	0.864	1.083	DF 1	<---	[BPS 1]
rule9	0.212	0.878	1.08	WON 2	<---	[ACE 2,DF 1]
rule16	0.31	0.875	1.077	WON 2	<---	[BPS 2]
rule12	0.236	0.857	1.074	DF 1	<---	[BPS 1,ACE 1,WON 2]
rule5	0.202	0.872	1.073	WON 2	<---	[BPS 1,ACE 2]
rule3	0.197	0.851	1.066	DF 1	<---	[BPS 1,ACE 2]
rule8	0.207	0.84	1.053	DF 1	<---	[BPS 2,ACE 1,WON 2]
rule20	0.448	0.827	1.037	DF 1	<---	[ACE 1,WON 2]
rule24	0.665	0.833	1.025	WON 2	<---	[DF 1]
rule25	0.665	0.818	1.025	DF 1	<---	[WON 2]
rule23	0.547	0.816	1.023	DF 1	<---	[ACE 1]
rule15	0.266	0.831	1.022	WON 2	<---	[ACE 2]
rule19	0.448	0.82	1.009	WON 2	<---	[ACE 1,DF 1]
rule11	0.217	0.8	1.002	DF 1	<---	[BPS 2,ACE 1]
rule22	0.542	0.809	0.995	WON 2	<---	[ACE 1]

Slika 4: Pravila pridruživanja za 2009. godinu

4 Klasterovanje

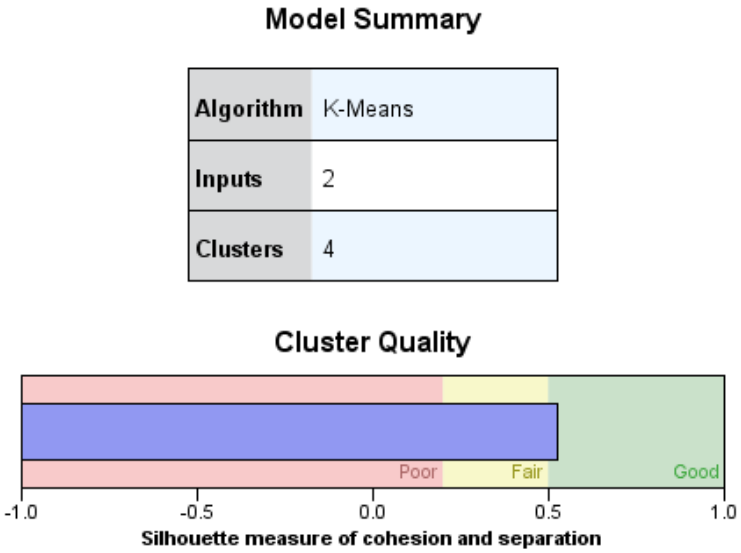
Što se tiče klasterovanja, s obzirom da podaci po godinama dosta osciliraju, odlučili smo da klasterovanje izvršimo za više godina. Izabrali smo 2003., 2011. i 2017. godinu. Pre svega, zanimala nas je zavisnost broja godina pobednika i broj asova pobednika. Prvo smo obradili nedostajuće vrednosti. Klasterovanje smo obradili u alatima SPSS i KNIME (slike 5 i 10).

4.1 SPSS

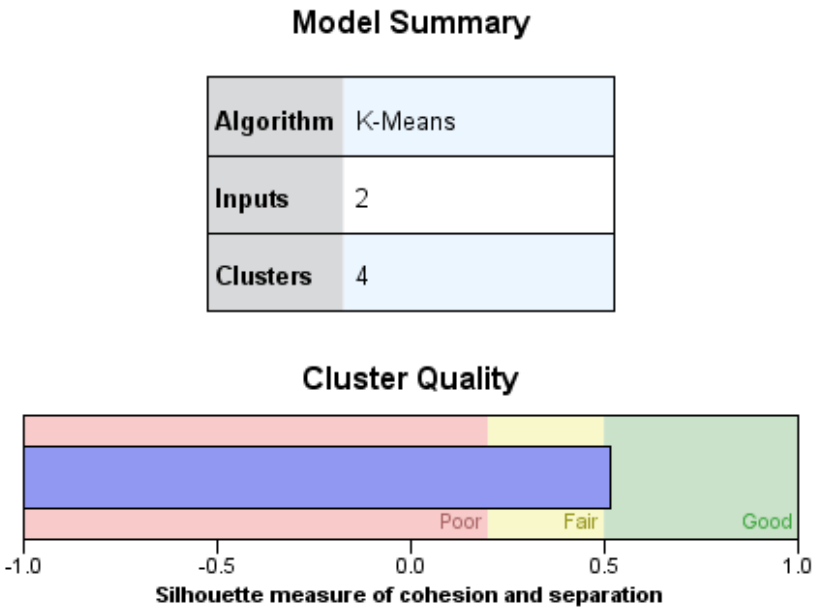


Slika 5: SPSS klasterovanje

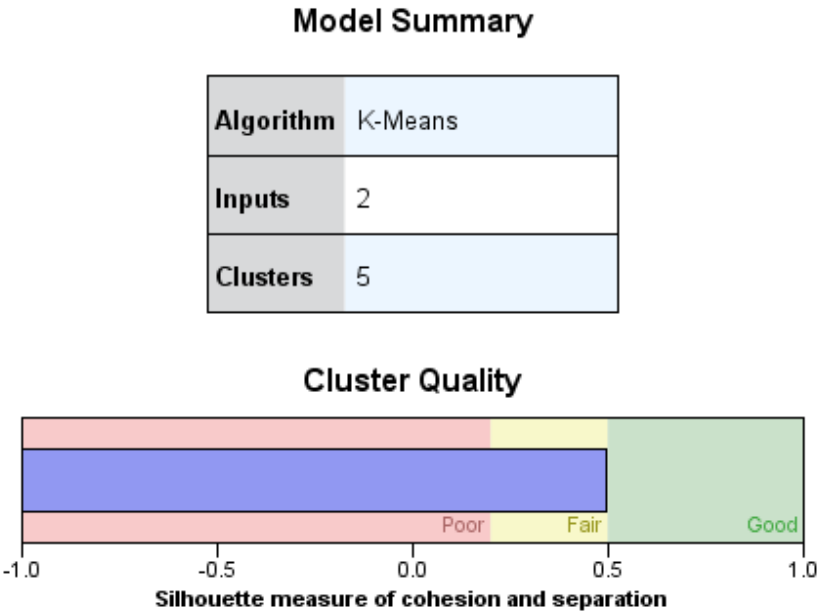
U alatu SPSS smo pomoću siluete pratili kako nam se kvalitet klasterovanja razlikuje u zavisnosti od broja klastera. *Kohonen* algoritam nam je za broj klastera između 3 i 5 davao ˝osrednji˝ kvalitet klasterovanja, pa ga nismo detaljno razmatrali. S druge strane, *K-Means* algoritam nam je davao dosta raznolike ocene klastera po godinama. 2003. godina nam je za 4 i 5 klastera pokazala kvalitet klasterovanja ˝dobar˝, uz važnost atributa $w\_age = 1$ i $w\_ace = 1$. U 2011. godini nam je za 5 klastera silueta pokazivala kvalitet ˝osrednji˝, promenivši broj klastera na 4 silueta je prešla u ˝dobar˝. Takođe, i sa 4 klastera i sa 5 klastera važnost atributa $w\_age = 1$ i $w\_ace = 1$. Rezultati za 2017. godinu za 4 klastera pokazuju kvalitet ˝osrednji˝; promenivši broj klastera na 5, silueta je na granici ˝osrednji˝-˝dobar˝, međutim, važnost atributa sa 4 klastera je $w\_age = 1$, $w\_ace = 0.77$, dok je sa 5 klastera w_ace opao na 0.46. Ipak smo odlučili da 2017. godinu odbradimo sa 5 klastera. Konačno, odlučili smo se za broj i kvalitet klastera koji su prikazani na slikama 6, 7 i 8.



Slika 6: Kvalitet klasterovanja - 2003. godina

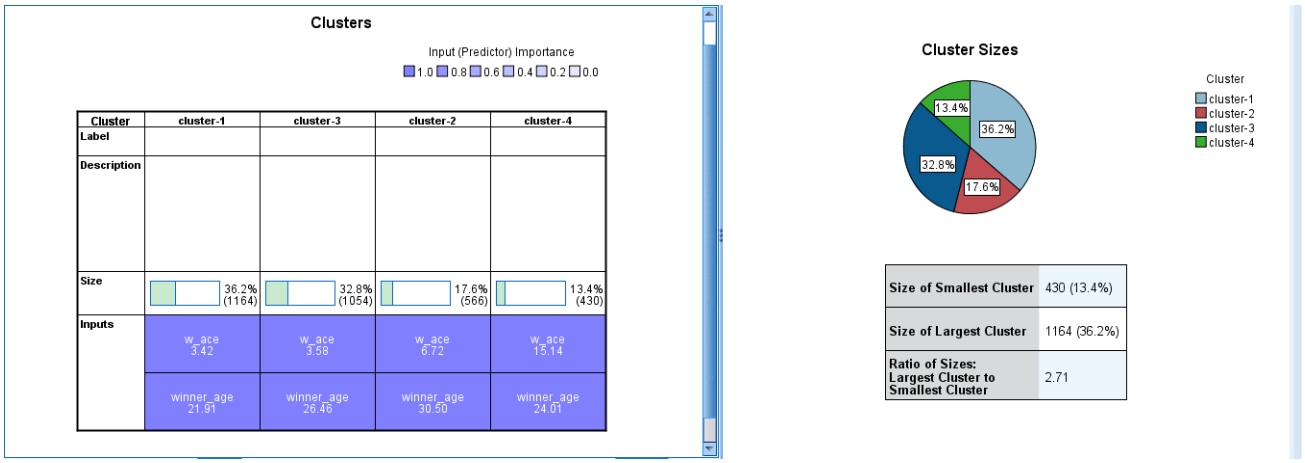


Slika 7: Kvalitet klasterovanja - 2011. godina



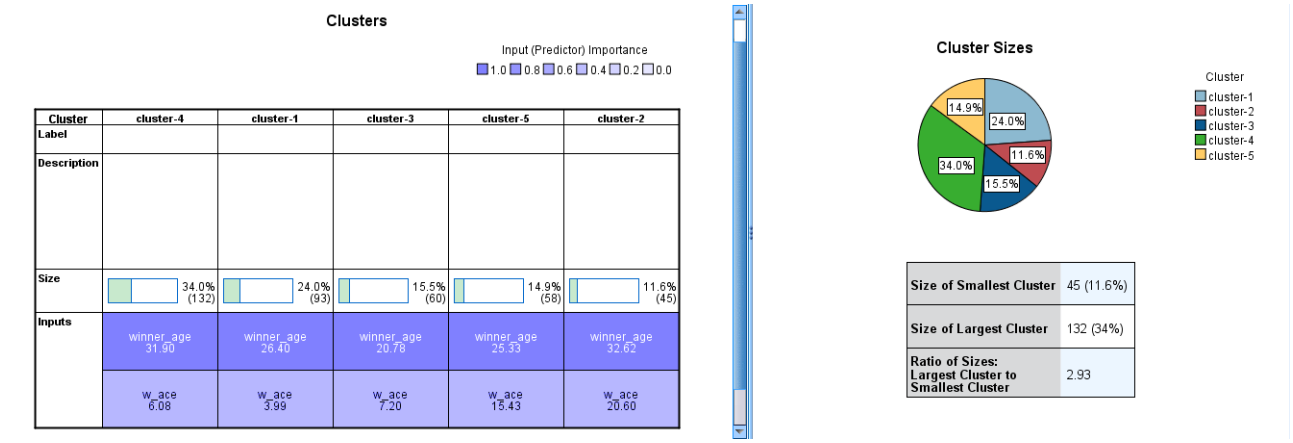
Slika 8: Kvalitet klasterovanja - 2017. godina

Kao što se vidi na slici 9, u sve tri godine su dobijeni interesantni podaci. Na primer, u 2003. godini najstariji igrači imaju slabiji prosek asova, dok u 2011. i 2017. godini imamo dva klastera sa prosekom godina oko 30; u jednom klasteru nam je broj asova mali, dok je u drugom najveći. Ovo nam je govorilo da možda imamo neki element van granica, koji je uticao na kreiranje dodatnog klastera. Odlučili smo da proverimo šta ćemo dobiti u KNIME-u.



(a) 2003. godina

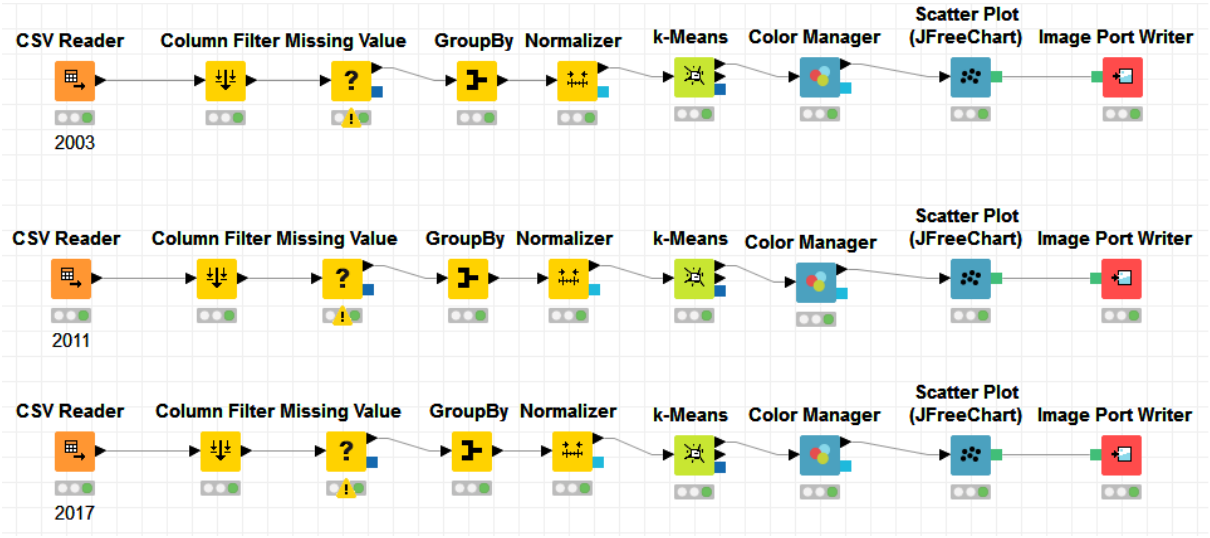
(b) 2011. godina



(c) 2017. godina

Slika 9: Modeli klastera

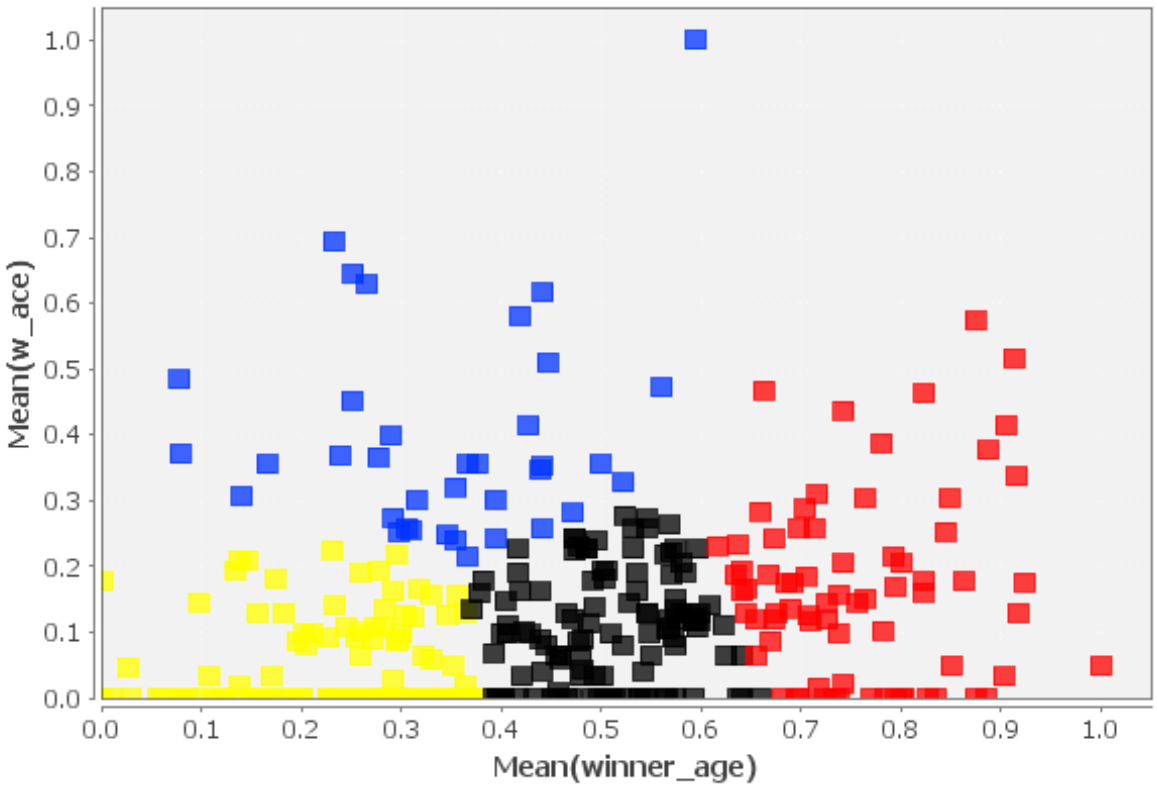
4.2 KNIME



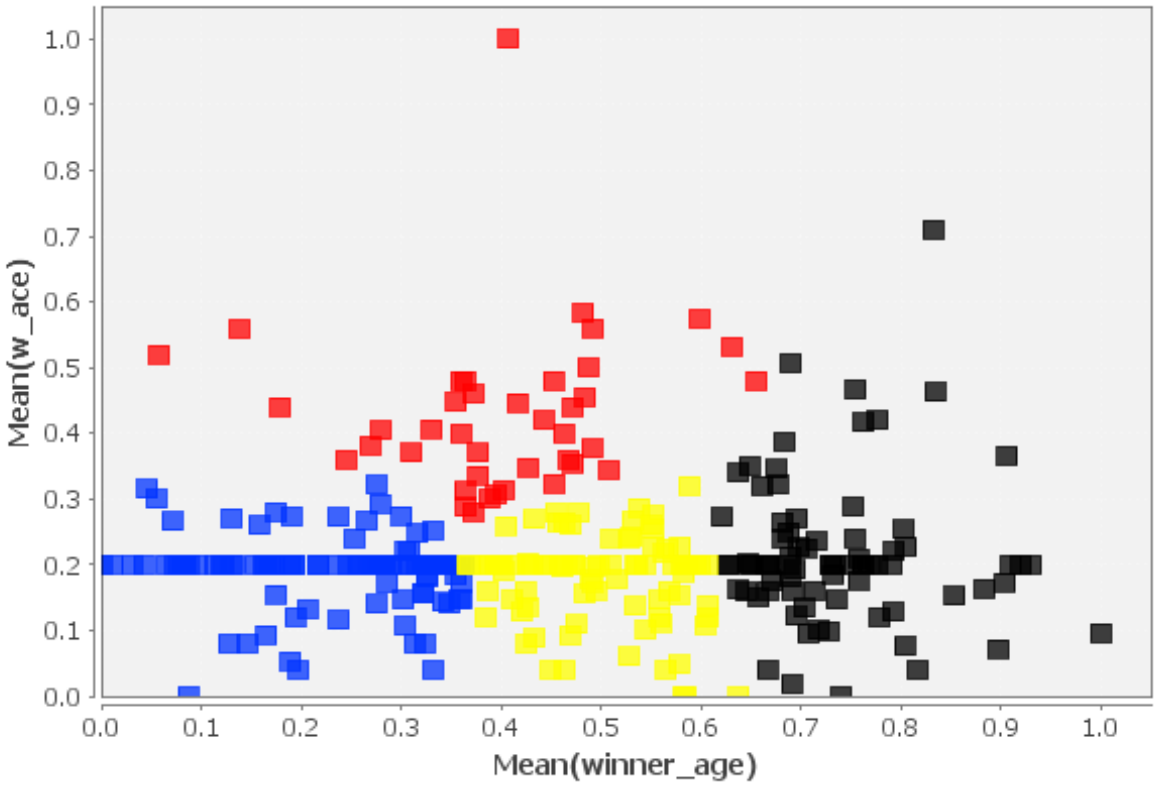
Slika 10: KNIME klasterovanje

U fazi pretprocesiranja podataka, otkrili smo jednog igrača sa nepoznatim brojem godina i taj red smo obrisali. U situaciji kada je broj asova bio nepoznat, stavljali smo vrednost nula. Nakon toga smo grupisali podatke po igračima, kako bi za svakog igrača dobili prosek koliko je imao asova tokom godine. Da bi iskoristili *K-Means* algoritam, normalizovali smo podatke kako bi broj godina i broj asova imali isti uticaj na računanje rastojanja među instancama.

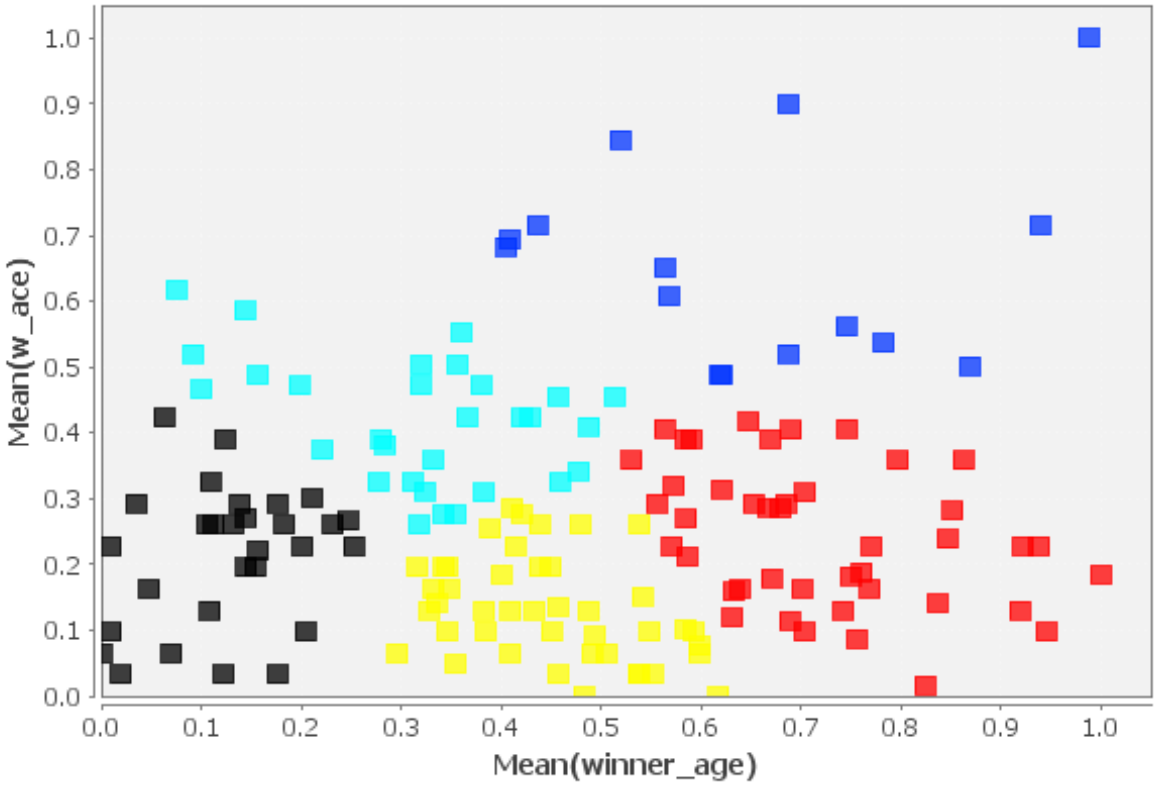
U KNIME-u smo se opredelili da klasterovanje vršimo sa istim brojem klastera kao što smo činili u SPSS-u. Dobijeni klasteri su prikazani na slikama 11, 12 i 13.



Slika 11: Klasteri - 2003. godina



Slika 12: Klasteri - 2011. godina



Slika 13: Klasteri - 2017. godina

Možemo primetiti da stvarno postoje elementi van granica koji su uticali na to da se formiraju novi klasteri. Igrači koji predstavljaju elemente van granica su dati na slici 14.

Row ID	winner_name	Mean(winner_age)	Mean(w_ace)
Row91	Frederic Niemeyer	27.162	31

(a) 2003. godina

Row ID	winner_name	Mean(winner_age)	Mean(w_ace)
Row97	Fritz Wolmarans	24.903	25

(b) 2011. godina

Row ID	winner_name	Mean(winner_age)	Mean(w_ace)
Row68	Ivo Karlovic	37.864	30.75

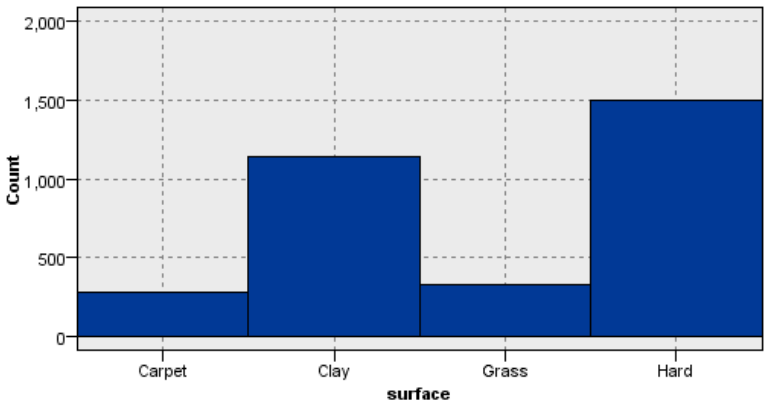
(c) 2017. godina

Slika 14: Elementi van granica

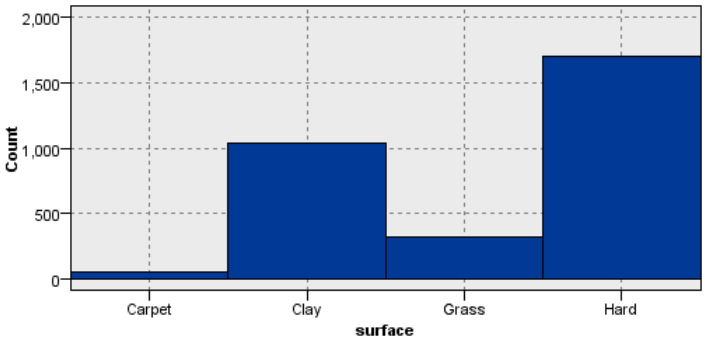
Najinteresantiji je definitivno Ivo Karlović, koji je na meču protiv Orasia Zebaljosa na Australian Open-u postigao 75 asova. Treba imati na umu da je Karlović odigrao četiri meča na ovom turniru. Moramo napomenuti, da je skup podataka o 2017. godini nepotpun, jer je u toku te godine napravljen skup i da dosta turnira još uvek nije upisano. Trenutni skup ima samo podatke sa turnira odigranih u januaru i februaru.

5 Klasifikacija

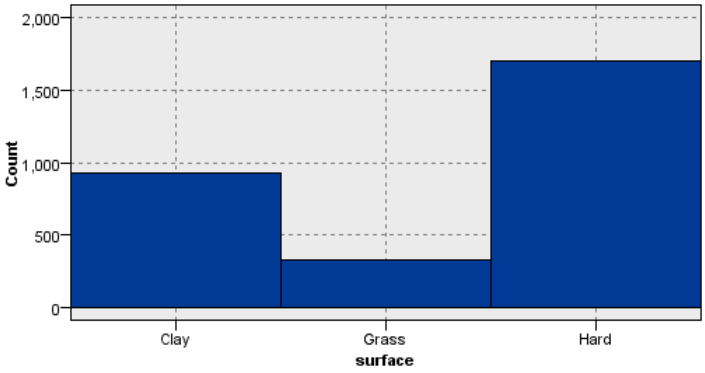
Za klasifikaciju smo odlučili da nam klase budu podloge terena, a pripadnost svakoj klasi se određuje na osnovu karakteristika četiri gubitnikova atributa (broj asova gubitnika, broj duplih servis grešaka gubitnika, broj ubačenih prvih servisa gubitnika, broj brejk šansi na servis gubitnika). Kao i kod klasterovanja i ovde smo želeli da vidimo šta se dešava u više godina. S obzirom na to da su neke podloge prisutnije u odnosu na ostale. Nakon detaljne analize raspodele podloga, odlučili smo se za 2005., 2008. i 2015. godinu. Razlog što smo odabrali baš ove godine jeste polako gubljenje ˝tepiha˝ kao podloge (slike 15, 16 i 17), pa nas je zanimalo kako će ova činjenica uticati na sam proces klasifikacije.



Slika 15: Podloga - 2005. godina



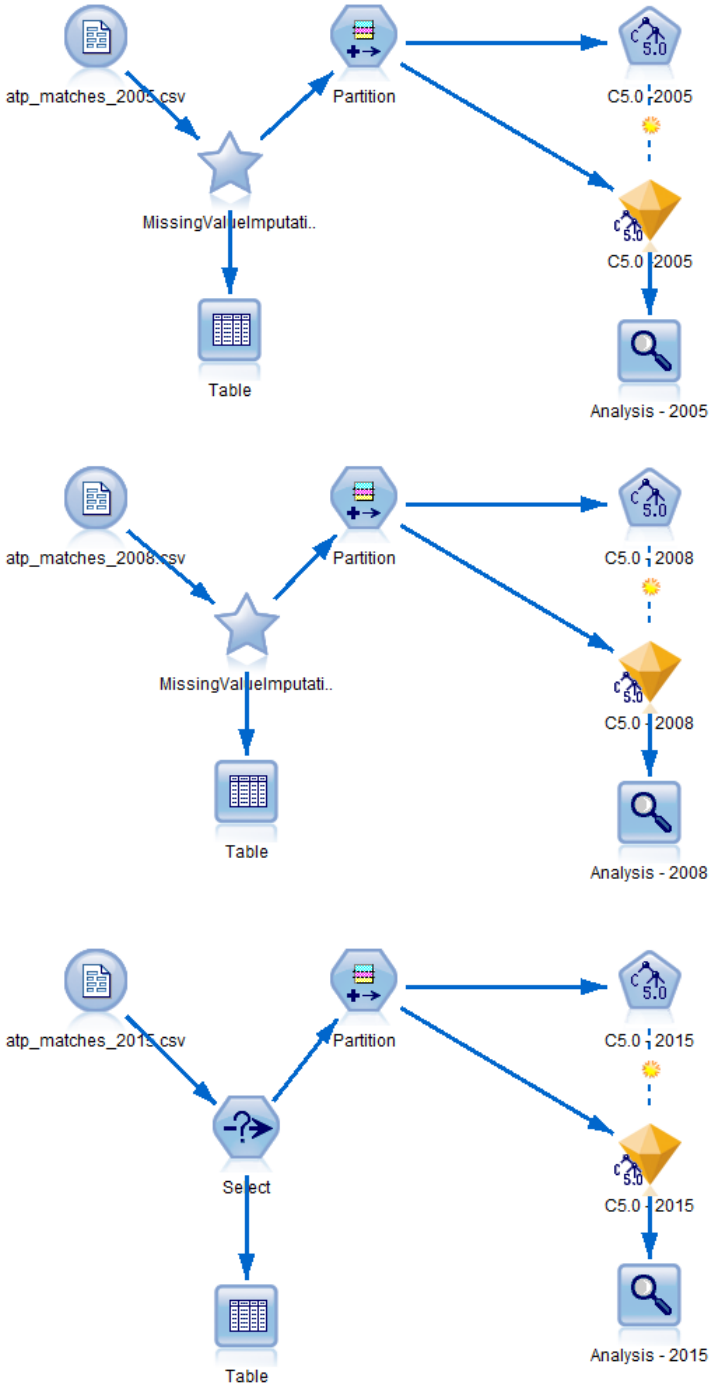
Slika 16: Podloga - 2008. godina



Slika 17: Podloga - 2015. godina

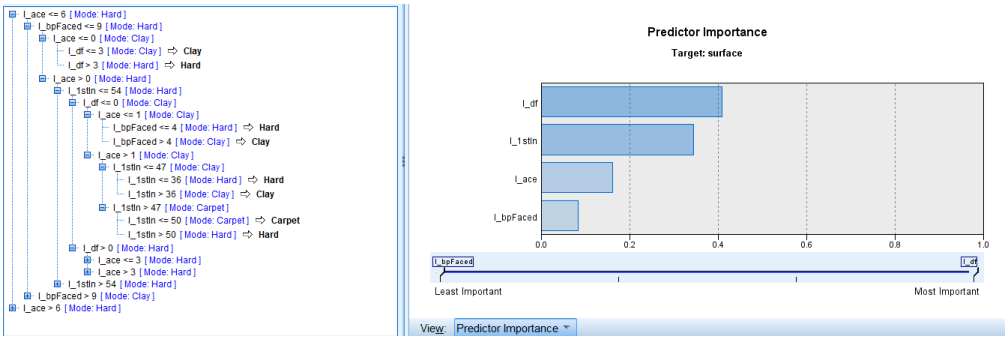
Klasifikaciju smo, također, obradili u SPSS-u i u KNIME-u.

5.1 SPSS

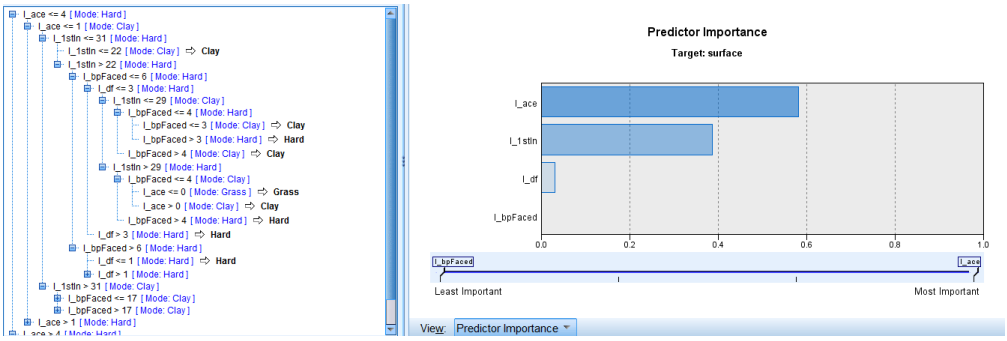


Slika 18: SPSS klastifikacija

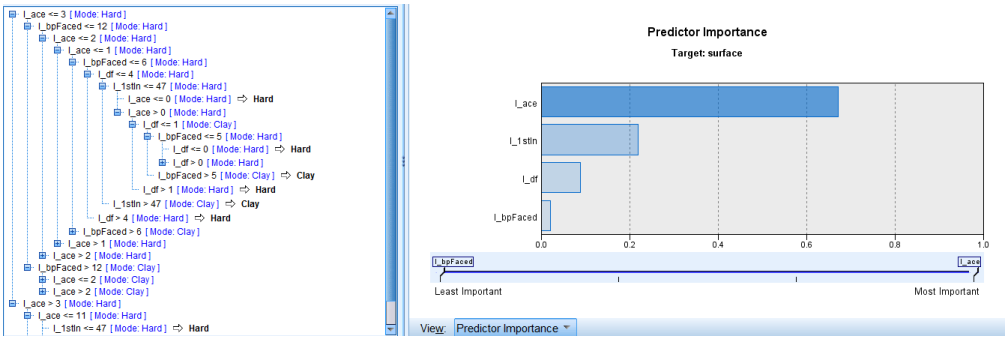
Primenili smo *C5.0* algoritam sa podelom na trening i test skup u odnosu 70-30. Na slici 19 možemo videti analizu najvažnijjih atributa. Najvažniji atribut u 2008. i 2015. godini je broj asova, dok u 2005. godini broj asova zauzima treće mesto po važnosti. Ono što je interesantno jeste da se u 2005. godini broj duplih servis grešaka smatra najvažnijim, dok se u 2008. i 2015. godini može videti da je važnost ovog atributa skoro nula, 0.03 i 0.09 respektivno.



(a) 2005. godina



(b) 2008. godina

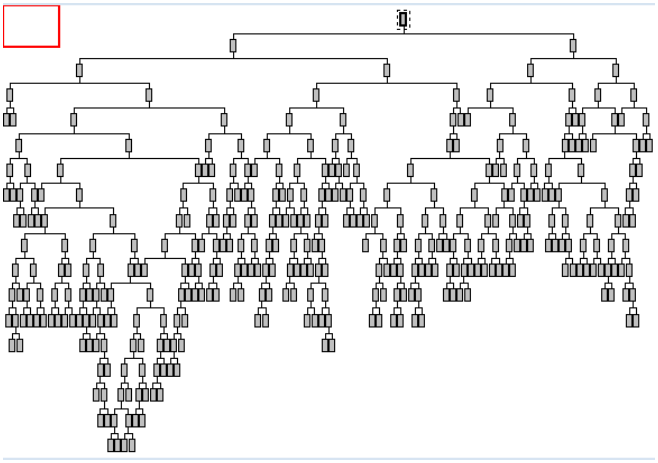


(c) 2015. godina

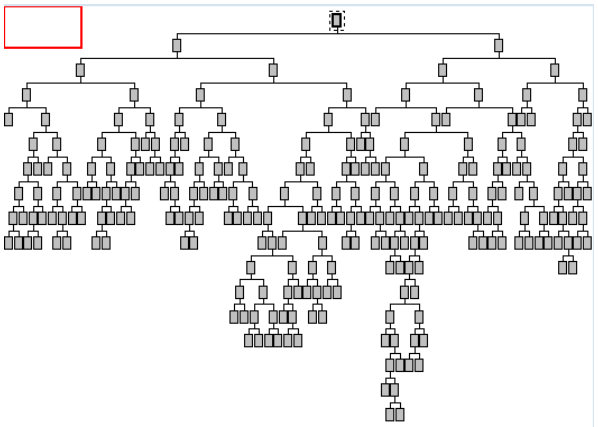
Slika 19: Model klasifikacije

Neke segmente drveta odlučivanja smo mogli da vidimo i na prethodnoj slici, a detaljniji prikaz se može pogledati na sledećim linkovima: 2005, 2008, 2015.

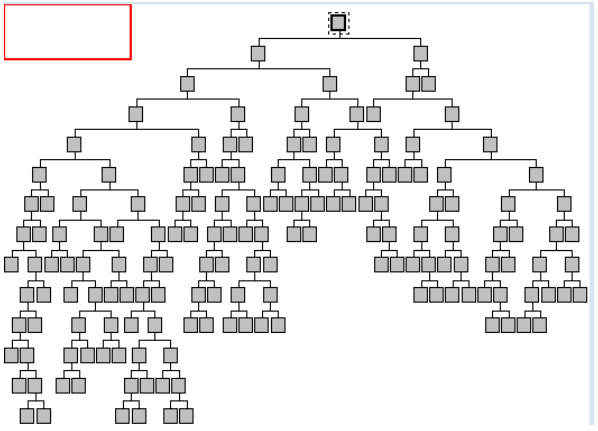
Intuitivan prikaz drveta odlučivanja za sve godine se može videti na slici 20.



(a) 2005. godina



(b) 2008. godina



(c) 2015. godina

Slika 20: Drveta odlučivanja dobijena algoritmom C-5.0

Kao što možemo da vidimo, drveta su veoma duboka i razgranata. Na slici 21 se može videti da je dobijena preciznost na trening skupu manja od 70%, dok je na test skupu ta preciznost oko 50%.

Results for output field surface

Comparing $C-surface with surface

'Partition'	Testing		Training	
Correct	443	47.63%	1,315	66.35%
Wrong	487	52.37%	667	33.65%
Total	930		1,982	

Coincidence Matrix for $C-surface (rows show actuals)

'Partition' = Testing	Carpet	Clay	Grass	Hard
Carpet	3	27	1	53
Clay	5	143	16	160
Grass	2	24	8	61
Hard	13	111	14	289

'Partition' = Training	Carpet	Clay	Grass	Hard
Carpet	39	32	3	89
Clay	5	431	8	239
Grass	6	34	55	115
Hard	6	122	8	790

(a) 2005. godina

Results for output field surface

Comparing $C-surface with surface

'Partition'	Testing		Training	
Correct	453	51.54%	1,311	69.55%
Wrong	426	48.46%	574	30.45%
Total	879		1,885	

Coincidence Matrix for $C-surface (rows show actuals)

'Partition' = Testing	Clay	Grass	Hard
Carpet	1	0	10
Clay	115	14	163
Grass	24	5	65
Hard	126	23	333

'Partition' = Training	Clay	Grass	Hard
Carpet	4	1	15
Clay	369	4	236
Grass	36	44	124
Hard	146	8	898

(b) 2008. godina

Results for output field surface

Comparing $C-surface with surface

'Partition'	Testing		Training	
Correct	455	54.23%	1,228	68.76%
Wrong	384	45.77%	558	31.24%
Total	839		1,786	

Coincidence Matrix for $C-surface (rows show actuals)

'Partition' = Testing	Clay	Grass	Hard
Clay	71	16	182
Grass	8	8	95
Hard	70	13	376

'Partition' = Training	Clay	Grass	Hard
Clay	242	7	285
Grass	30	39	141
Hard	81	14	947

(c) 2015. godina

Slika 21: Matrice konfuzije - C5.0

Možemo da primetimo da se ˝tepih˝ klasifikuje kao ˝beton˝, što je i očekivano s obzirom da se najviše turnira igra na betonu.

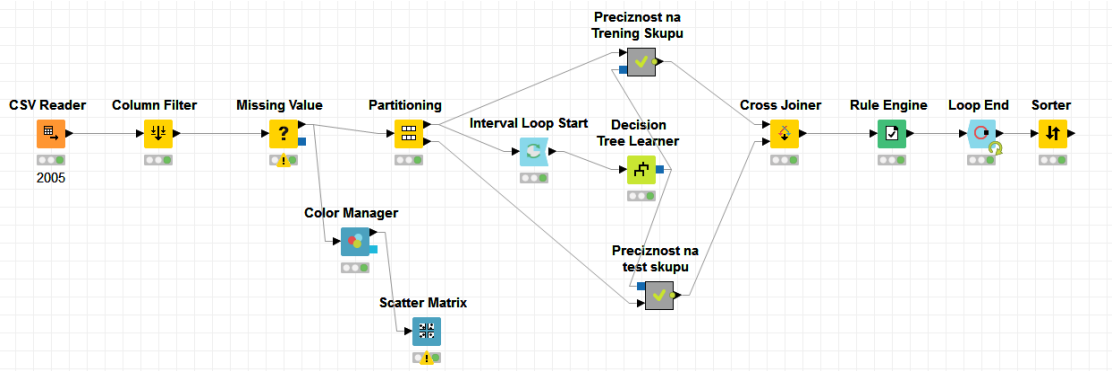
Sveobuhvatni prikaz rada algoritma *C5.0* dat je u fajlovima: 2005, 2008, 2015.

S obzirom da nismo bili zadovoljni rezultatima dobijenim algoritmom *C5.0*, kao i to da nam je dobijena preciznost na trening i test skupu ukazivala da je možda došlo do preprilagođavanja, odlučili smo da na 2005. i 2015. godinu izvršimo istu analizu u alatu KNIME. Još jedan od razloga za istraživanje podataka u drugom alatu, bio je i taj što smo želili moguće izvršiti odsecanje drveta ranije u toku grananja.

5.2 KNIME

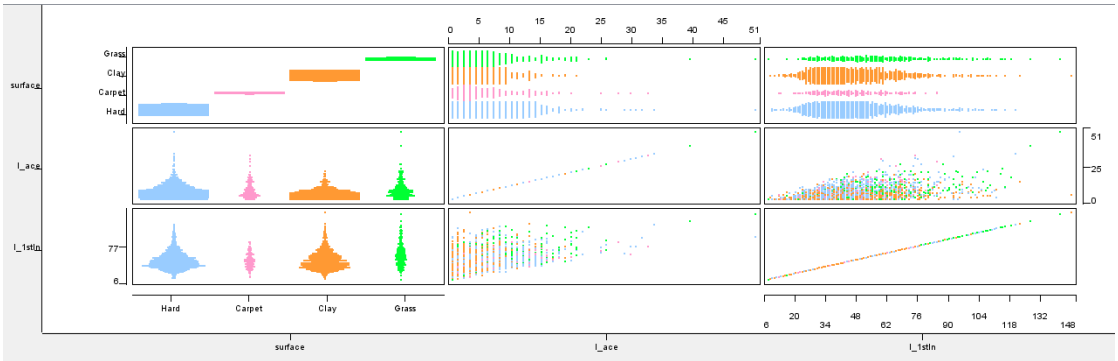
U alatu KNIME smo za klasifikaciju koristili Drveta odlučivanja, K najbližih suseda i metod potpornih vektora (SVM). Podatke smo takođe podelili na trening i test skup u odnosu 70-30.

5.2.1 Drveta odlučivanja

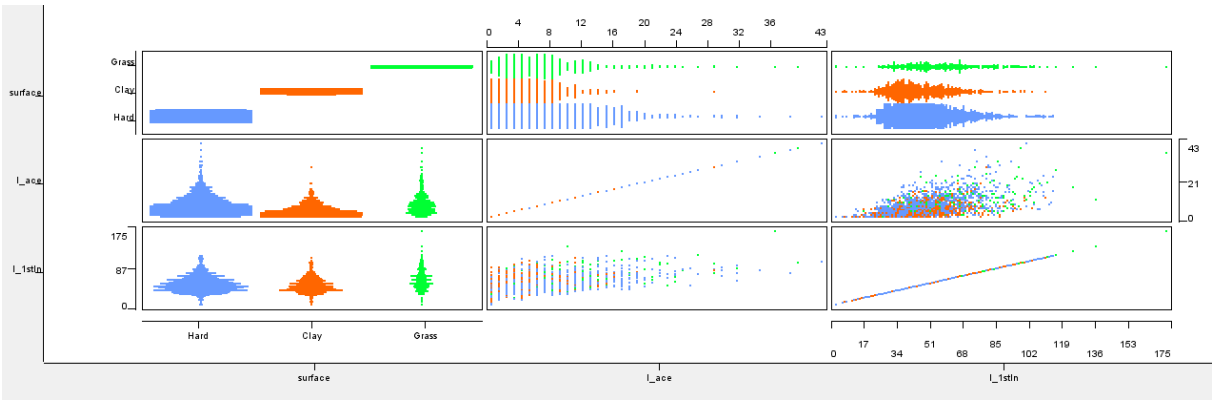


Slika 22: KNIME implementacija tehnike Drveta odlučivanja

Pre same klasifikacije, na slikama 23 i 24 je prikazan odnos između nekih atributa po kojima je vršena klasifikacija. Možemo primetiti tesnu povezanost između broja asova i broj ubačenih prvih servisa gubitnika.



Slika 23: Korelacija atributa - 2005. godina



Slika 24: Korelacija atributa - 2015. godina

Dobijene matrice konfuzije su prikazane na slikama 25 i 26. Za 2005. godinu, i na trening i na test skupu vidimo da se tepih i trava klasifikuju pre svega kao beton, a potom i kao šljaka. Za 2015. godinu, i na trening i na test skupu vidimo da se trava i dalje klasifikuje uglavnom kao beton.

Row ID	Hard	Clay	Carpet	Grass
Hard	640	307	0	0
Clay	288	417	0	0
Carpet	118	55	0	0
Grass	155	58	0	0

(a) Trening skup

Row ID	Hard	Clay	Carpet	Grass
Hard	266	140	0	0
Clay	124	178	0	0
Carpet	45	29	0	0
Grass	70	22	0	0

(b) Test skup

Slika 25: Matrice konfuzije - 2005

Row ID	Hard	Clay	Grass
Hard	971	79	0
Clay	444	118	0
Grass	214	11	0

(a) Trening skup

Row ID	Hard	Clay	Grass
Hard	412	39	0
Clay	198	43	0
Grass	89	7	0

(b) Test skup

Slika 26: Matrice konfuzije - 2015

Preciznost se može videti na slici 27. Da bismo dobili što bolju preciznost, primenili smo algoritam

za vrednosti od 5 do 100 sa korakom 5 za minimalni broj podloga po čvoru u drvetu odlučivanja. Primetimo da je najveća preciznost na trening i test skupu u 2005. godini dobijena u rasponu od 5 do 15 podloga po čvoru. Isti princip smo primenili i za 2015. godinu i dobili da je preciznost na test skupu najveća za vrednosti u rasponu od 45 do 60 podloga po čvoru.

Korišćenjem ovog alata, nismo dobili značajnu razliku u odnosu na preciznost dobijenu u alatu SPSS, ali je primetna razlika u dubini drveta odlučivanja, što se može videti na slikam 28 i 29. Kao meru nečistoće koristili smo Ginijev indeks i MDL metod za odsecanje stabla.

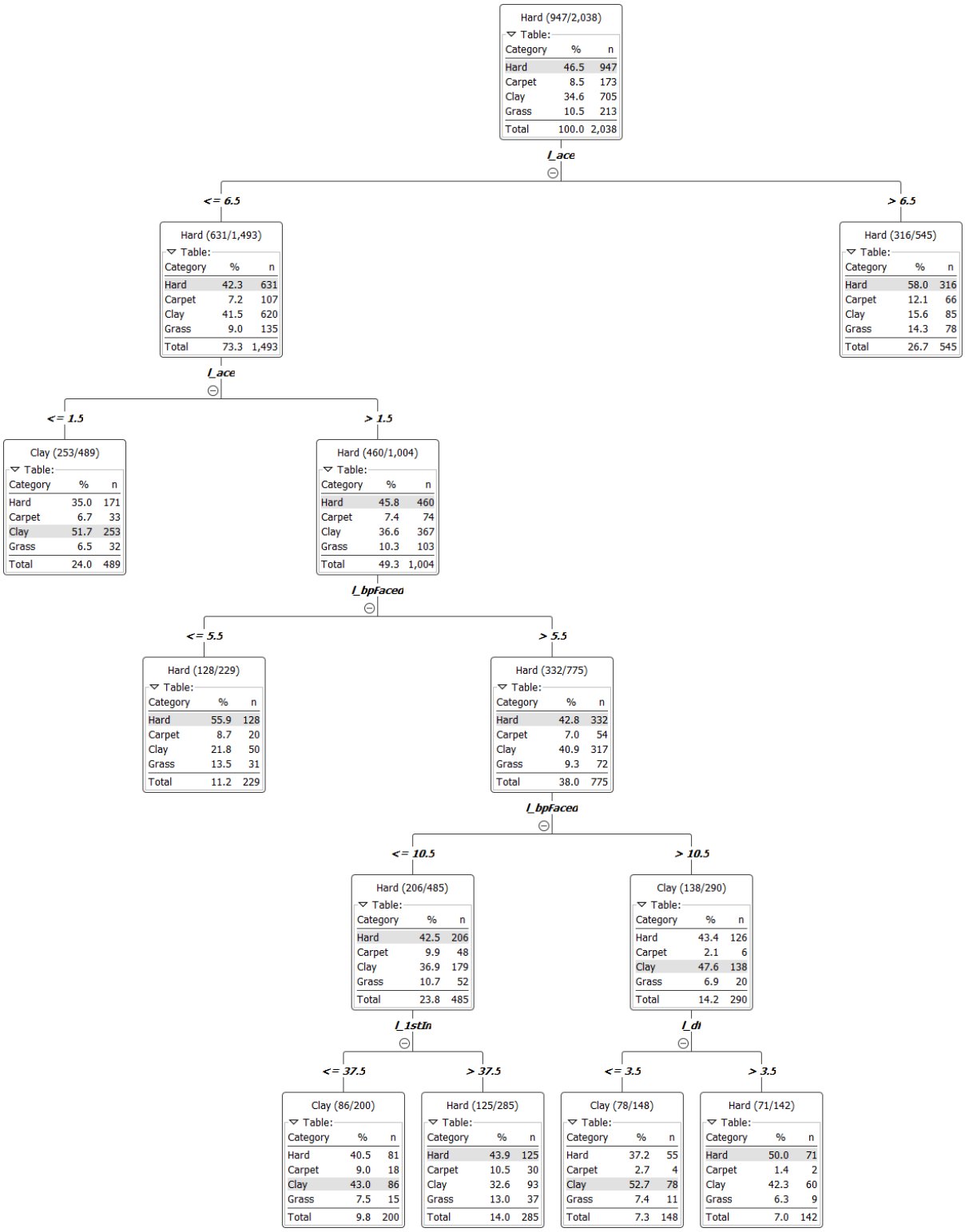
Row ID	D Accur...	D Accur...	I minLoop	I Iteration
Overall_Ov...	0.553	0.514	15	2
Overall_Ov...	0.553	0.515	5	0
Overall_Ov...	0.55	0.516	10	1
Overall_Ov...	0.548	0.51	20	3
Overall_Ov...	0.535	0.507	25	4
Overall_Ov...	0.535	0.507	30	5
Overall_Ov...	0.535	0.507	35	6
Overall_Ov...	0.535	0.507	40	7
Overall_Ov...	0.535	0.507	45	8
Overall_Ov...	0.535	0.507	50	9
Overall_Ov...	0.533	0.491	55	10
Overall_Ov...	0.531	0.493	60	11
Overall_Ov...	0.528	0.492	75	14
Overall_Ov...	0.528	0.492	80	15
Overall_Ov...	0.527	0.487	70	13
Overall_Ov...	0.527	0.491	65	12
Overall_Ov...	0.521	0.506	85	16
Overall_Ov...	0.521	0.506	90	17
Overall_Ov...	0.521	0.506	95	18
Overall_Ov...	0.519	0.508	100	19

(a) 2005. godina

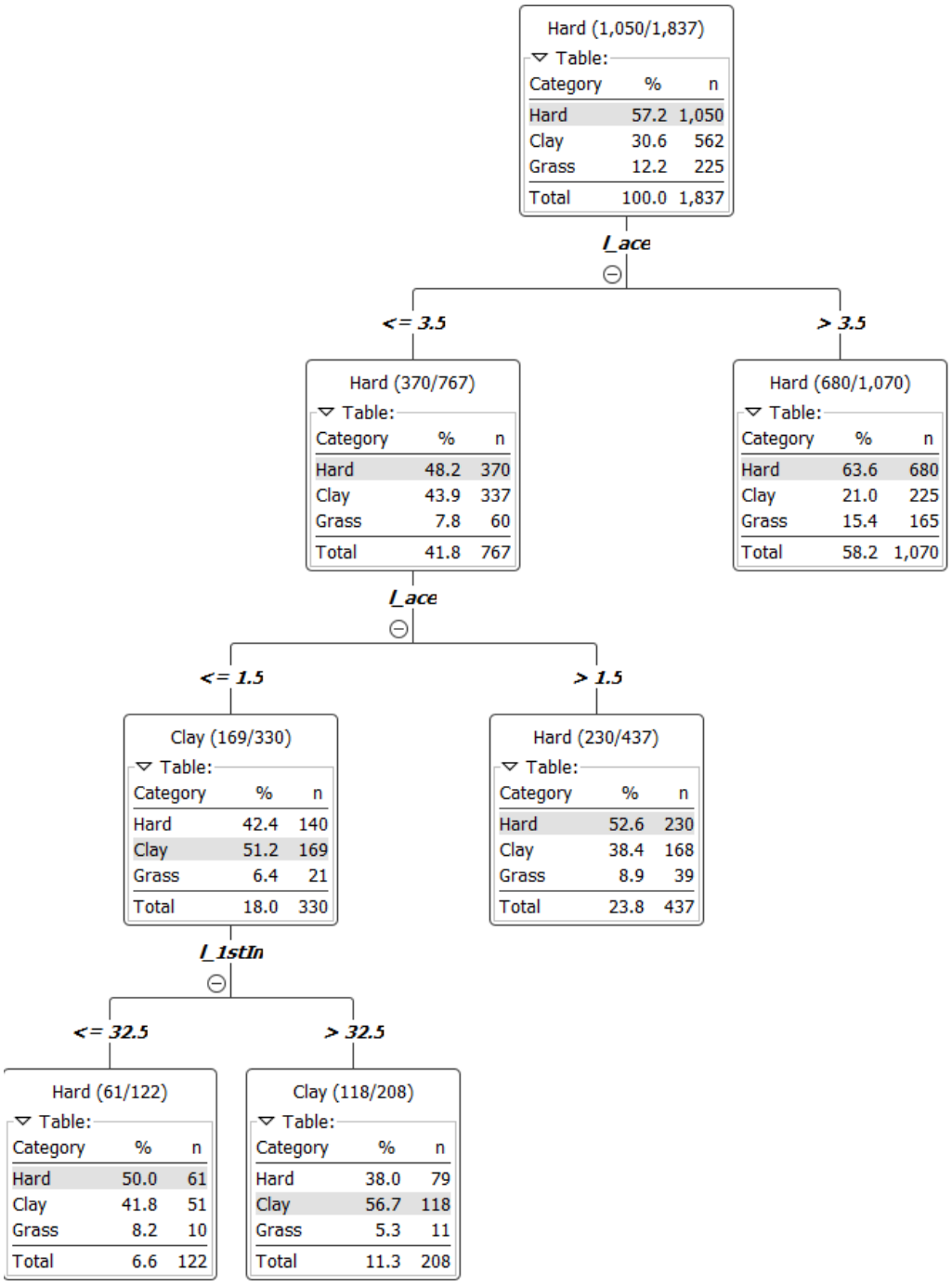
Row ID	D Accur...	D Accur...	I minLoop	I Iteration
Overall_Ov...	0.622	0.584	5	0
Overall_Ov...	0.618	0.585	10	1
Overall_Ov...	0.618	0.585	15	2
Overall_Ov...	0.618	0.585	20	3
Overall_Ov...	0.615	0.575	25	4
Overall_Ov...	0.615	0.575	30	5
Overall_Ov...	0.609	0.594	45	8
Overall_Ov...	0.609	0.594	50	9
Overall_Ov...	0.609	0.589	40	7
Overall_Ov...	0.609	0.586	35	6
Overall_Ov...	0.6	0.595	55	10
Overall_Ov...	0.6	0.595	60	11
Overall_Ov...	0.593	0.577	65	12
Overall_Ov...	0.593	0.577	70	13
Overall_Ov...	0.593	0.577	75	14
Overall_Ov...	0.593	0.577	80	15
Overall_Ov...	0.593	0.577	85	16
Overall_Ov...	0.593	0.577	90	17
Overall_Ov...	0.593	0.577	95	18
Overall_Ov...	0.593	0.577	100	19

(b) 2015. godina

Slika 27: Preciznost

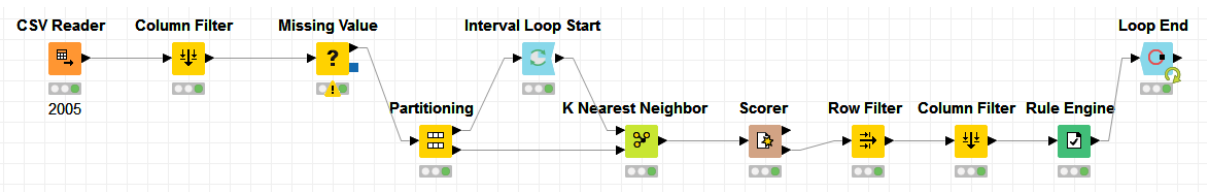


Slika 28: Ginijev indeks - 2005. godina



Slika 29: Ginijev indeks - 2015. godina

5.2.2 K najbližih suseda



Slika 30: KNIME implementacija tehnike k najbližih suseda

Koristeći metodu k najbližih suseda, dobili smo slične rezultate. Matrice konfuzije su date na slici 31.

Row ID	Hard	Carpet	Clay	Grass
Hard	318	0	88	0
Carpet	58	0	16	0
Clay	174	0	128	0
Grass	79	0	13	0

(a) 2005. godina

Row ID	Hard	Clay	Grass
Hard	445	6	0
Clay	230	11	0
Grass	95	0	1

(b) 2015. godina

Slika 31: Matrice konfuzije

Preciznost smo izračunali za vrednosti k u rasponu od 3 do 100 sa korakom 1. Na slikama 32 i 33 se može videti da je najbolja preciznost na test skupu dobijena za 2005. godinu, za k=35 i iznosi 0.527. U 2015. godini, najveća preciznost na test skupu je dobijena za vrednost k=47 i iznosi 0.589.

Row ID	D Ac...	k	Iteration
Overall#32	0.527	35	32
Overall#30	0.524	33	30
Overall#33	0.523	36	33
Overall#25	0.522	28	25
Overall#29	0.522	32	29
Overall#31	0.522	34	31
Overall#35	0.522	38	35
Overall#26	0.521	29	26
Overall#23	0.519	26	23
Overall#24	0.519	27	24
Overall#28	0.519	31	28
Overall#94	0.519	97	94
Overall#34	0.518	37	34
Overall#85	0.518	88	85
Overall#19	0.517	22	19
Overall#27	0.517	30	27
Overall#55	0.517	58	55
Overall#49	0.516	52	49
Overall#56	0.516	59	56
Overall#93	0.516	96	93
Overall#36	0.515	39	36
Overall#60	0.515	63	60

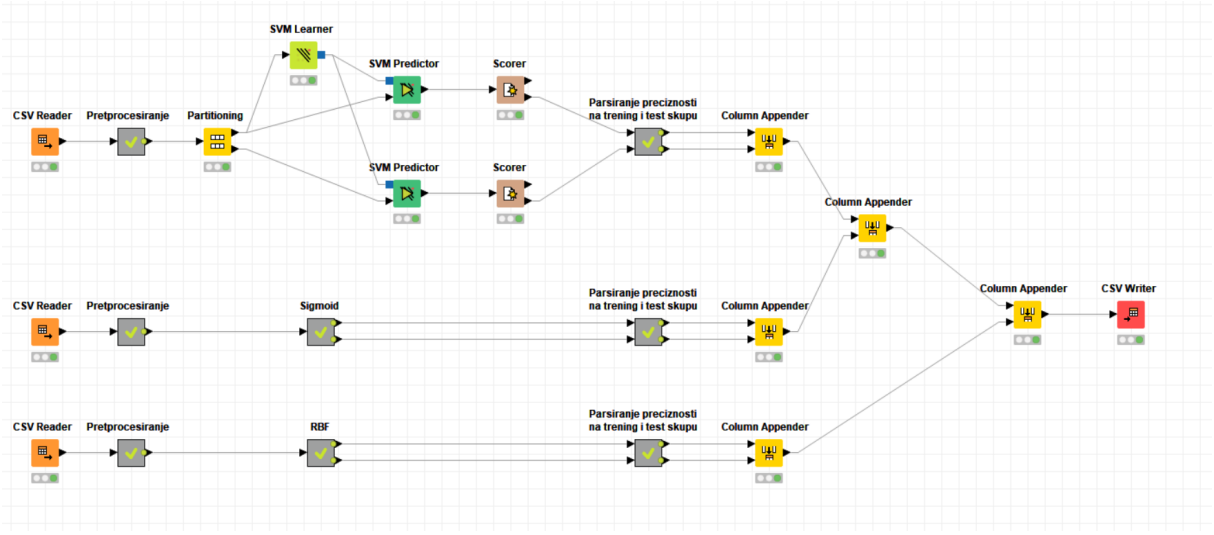
Slika 32: Preciznost kNN - 2005. godina

Row ID	Ac...	k	Iteration
Overall#44	0.589	47	44
Overall#29	0.588	32	29
Overall#30	0.585	33	30
Overall#31	0.585	34	31
Overall#32	0.585	35	32
Overall#33	0.585	36	33
Overall#34	0.585	37	34
Overall#70	0.585	73	70
Overall#5	0.584	8	5
Overall#19	0.584	22	19
Overall#26	0.584	29	26
Overall#28	0.584	31	28
Overall#37	0.584	40	37
Overall#40	0.584	43	40
Overall#41	0.584	44	41
Overall#47	0.584	50	47
Overall#21	0.582	24	21
Overall#27	0.582	30	27
Overall#25	0.581	28	25
Overall#38	0.581	41	38
Overall#49	0.581	52	49

Slika 33: Preciznost kNN - 2015. godina

Dobijeni rezultati su približno isti kao i preciznosti dobijene na test skupu metodom drveta odlučivanja. Prikazaćemo rezultate koje smo dobili još metodom SVM.

5.2.3 SVM



Slika 34: KNIME implementacija SVM tehnike

Na normalizovanim podacima, primenili smo sva tri raspoloživa kernela (polinomijalni trećeg stepena, sigmoid, Gausov(RBF)) za 2005. godinu. Na slici 35 se mogu videti preciznosti za sva tri kernela, i za trening i za test skup.

Acc_Training_Poly	Acc_Test_Poly	Acc_Training_Sigmoid	Acc_Test_Sigmoid	Acc_Training_RBF	Acc_Test_RBF
0.498903028	0.491820041	0.465555068	0.474437628	0.468626591	0.489775051

Slika 35: Preciznost za različite kernele

Koristeći polinomijalni kernel trećeg stepena, dobili smo izuzetno loše rezultate. Naime, skoro 50% redova (1501 od 3257) odgovaraju mečevima koji su odigrani na tvrdoj podlozi. Na slikama 36 i 37 vidimo da su podaci pogrešno klasifikovani u mečeve koji su odigrani na šljaci.

Row ID	Hard	Clay	Carpet	Grass
Hard	957	93	0	0
Clay	620	180	0	0
Carpet	179	17	0	0
Grass	218	15	0	0

Slika 36: Trening podaci za polinomijalni kernel

Row ID	Hard	Clay	Carpet	Grass
Hard	407	44	0	0
Clay	269	74	0	0
Carpet	76	8	0	0
Grass	90	10	0	0

Slika 37: Test podaci za polinomijalni kernel

Koristeći sigmoid kernel, situacija se promenila utoliko što su podaci vezani za tvrdu podlogu vrlo dobro klasifikovani, što se može videti na slikama 38 i 39. Primetimo da su podaci uglavnom raspoređeni u klase koje se odnose na beton, šljaku i tepih.

Row ID	Hard	Carpet	Clay	Grass
Hard	851	99	100	0
Carpet	158	24	14	0
Clay	519	95	186	0
Grass	198	23	12	0

Slika 38: Trening podaci za sigmoid kernel

Row ID	Hard	Carpet	Clay	Grass
Hard	363	49	39	0
Carpet	66	9	9	0
Clay	210	41	92	0
Grass	85	6	9	0

Slika 39: Test podaci za sigmoid kernel

Koristeći Gausov kernel, dobili smo lošiju klasifikaciju za tvrdu podlogu, dosta bolju klasifikaciju za šljaku, bolju klasifikaciju za travu, dok je tepih u potpunosti promašen (slike 40 i 41).

Row ID	Hard	Clay	Grass	Carpet
Hard	767	172	111	0
Clay	424	278	98	0
Grass	180	30	23	0
Carpet	142	33	21	0

Slika 40: Trening podaci za Gausov kernel

Row ID	Hard	Clay	Grass	Carpet
Hard	336	78	37	0
Clay	167	138	38	0
Grass	82	13	5	0
Carpet	62	10	12	0

Slika 41: Test podaci za Gausov kernel