

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	1
2	Analiza podataka	1
3	Pravila pridruživanja	3
4	Klasterovanje	5
4.1	SPSS	6
4.2	KNIME	9
5	Klasifikacija	11
5.1	SPSS	13

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

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

Tabela 2: Podaci o pobjednicima

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 lopti gubitnika
l_bpFaced	broj izgubljenih gemova posle brejka gubitnika

Tabela 3: Podaci o gubitnicima

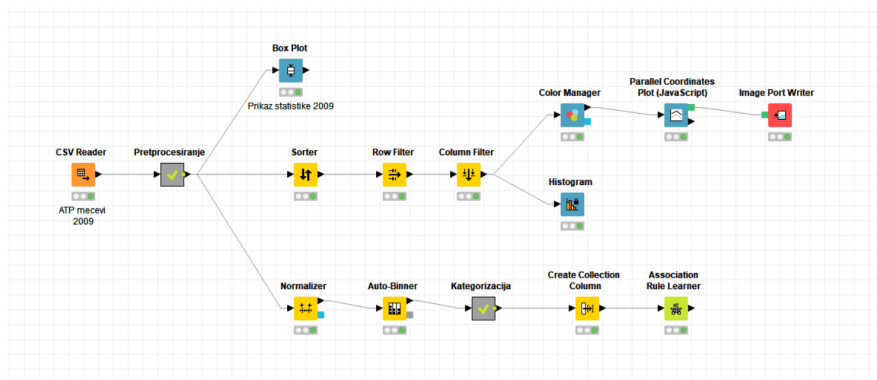
U tabeli 2 prikazani su podaci o pobjedniku meča.

U tabeli 3 prikazani su podaci o gubitniku meča.

S obzirom na veliki broj raspoloživih godina, prvo smo se detaljno upoznali sa podacima i 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.

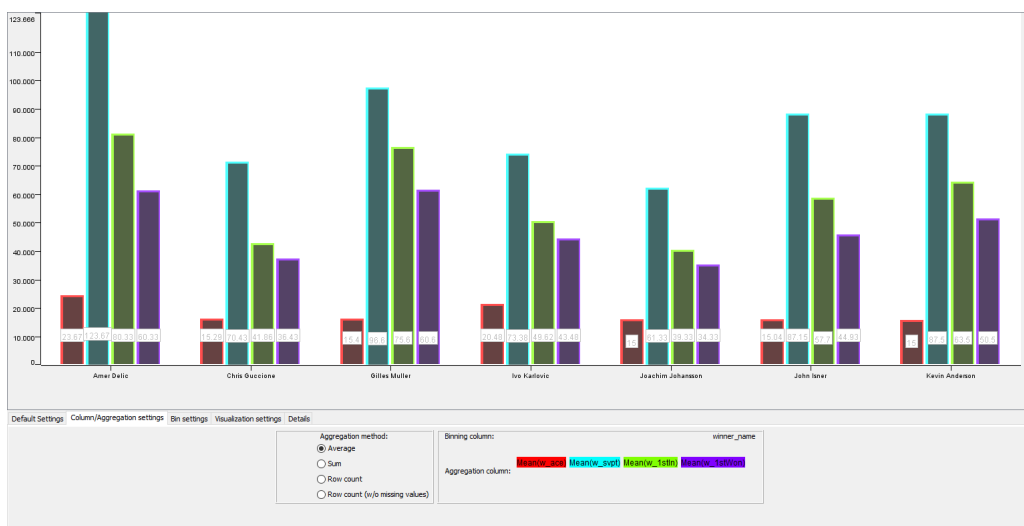
3 Pravila pridruživanja

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.

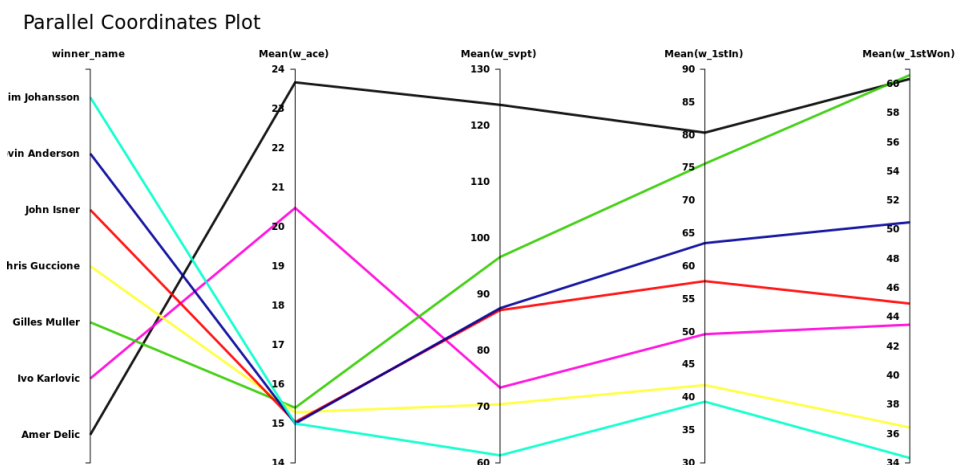


Slika 1: KNIME implementacija

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, procenat ubačenog prvog 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

Iznenadenje 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 (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

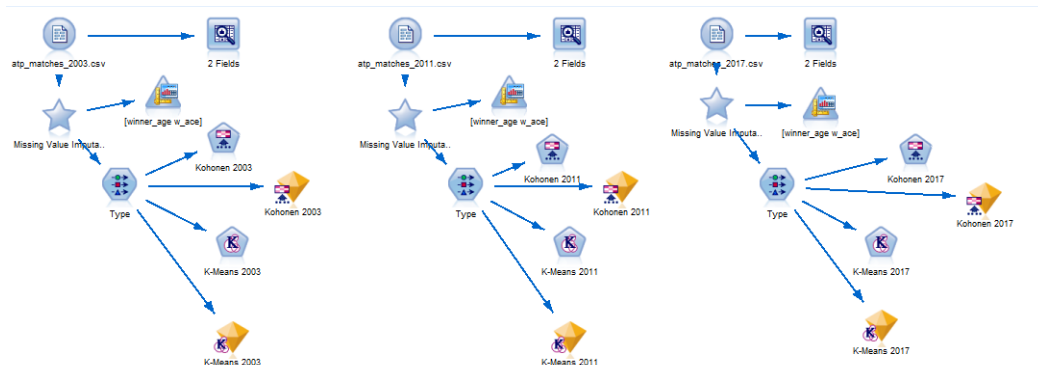
Row ID	D Support	D Confide...	D ▼ Lift	S Conseq...	S implies	(...) Items
rule2	0.177	0.837	1.36	BPS 1	<---	[ACE 2,WON 2,DF 1]
rule4	0.197	0.816	1.326	BPS 1	<---	[ACE 2,DF 1]
rule6	0.207	0.857	1.279	ACE 1	<---	[BPS 2,WON 2,DF 1]
rule10	0.217	0.83	1.239	ACE 1	<---	[BPS 2,DF 1]
rule7	0.207	0.955	1.174	WON 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

4 Klasterovanje

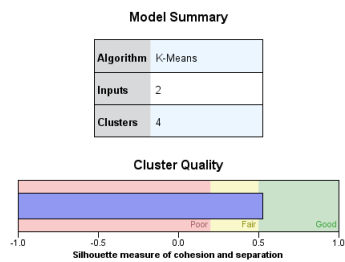
Sto se tiče klasterovanja, s obzirom da podaci po godinama dosta osciliraju, odlučili smo da klasterovanje izvršimo za više godina. Izabrali smo 2003, 2011 i 2017 godinu. Pre svega nas je zanimala zavisnost broja godina pobednika i broj asova pobednika. Prvo smo obradili nedostajuće vrednosti. Klasterovanje smo obradili u alatima SPSS i KNIME (slike 5 i 8).

4.1 SPSS

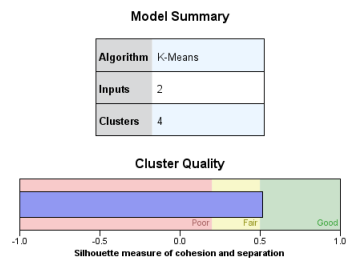


Slika 5: SPSS klasterovanje

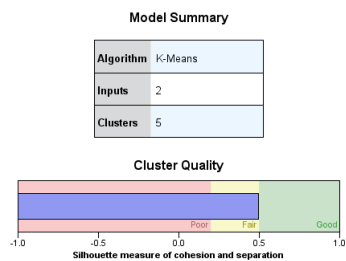
U SPSS-u pomoću siluete smo pratili kako nam se kvalitet klasterovanja razlikuje u zavisnosti od broja klastera. *Kohonen* algoritam nam je za broj klastera 3-5 davao "osrednji" kvalitet klasterovanja, pa ga nismo detaljno razmatrali. (Podseti me da probam sa nekim drugim ulaznim argumentima za *Kohonen* i da napisem ovde za koje smo probali.) S druge strane, *K-Means* algoritam nam je davao dosta šarenolike 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", promeniš broj klastera na 4 silueta je prešla malo 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", promeniš 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 smose dlucili da stavimo 5 klastera za 2017 godinu. Konačno, odlučili smo se za broj i kvalitet klastera koji su prikazani na slici 6.



(a) 2003.godina



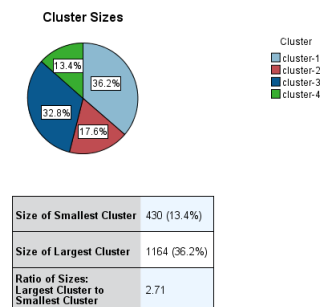
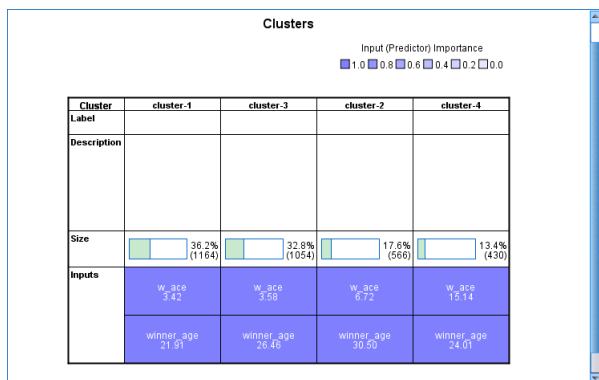
(b) 2011.godina



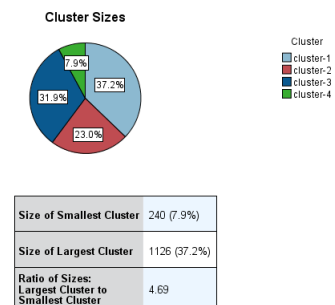
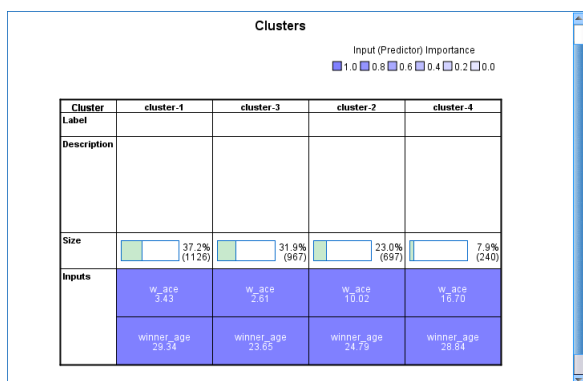
(c) 2017.godina

Slika 6: Kvalitet klasterovanja

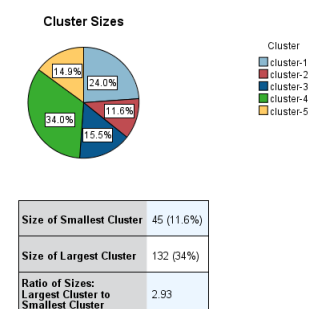
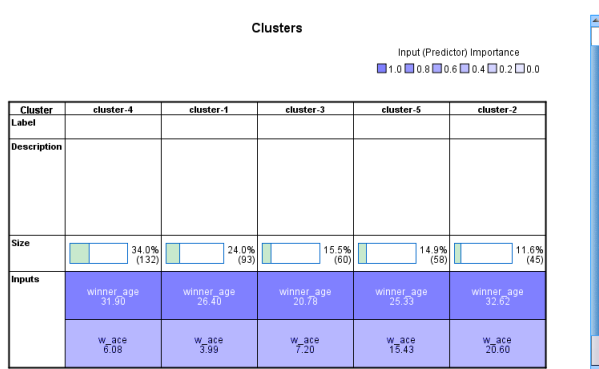
Kao što možemo da vidimo na slici 7. U sve tri godine su dobijeni interesantni podaci. Na primer, u 2003. godini najstariji igrači su nam sa slabijim prosekom asova, dok u 2011. i 2017. godini imamo dva klastara sa prosekom godina oko 30, u jednom klasteru nam je broj asova mali, dok je u drugom najveći. Ovo nam je govorilo da možda imamo neki autlajer koji je uticao na kreiranje dodatnog klastera. Odlučili smo da proverimo šta ćemo dobiti u KNIME-u.



(a) 2003.godina



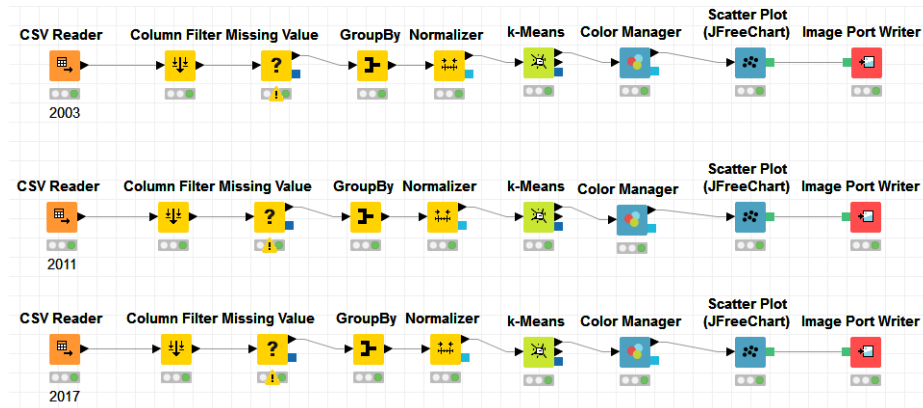
(b) 2011.godina



(c) 2017.godina

Slika 7: Modeli klastera

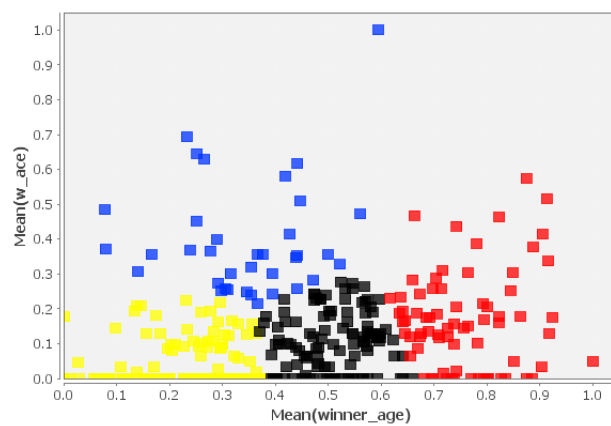
4.2 KNIME



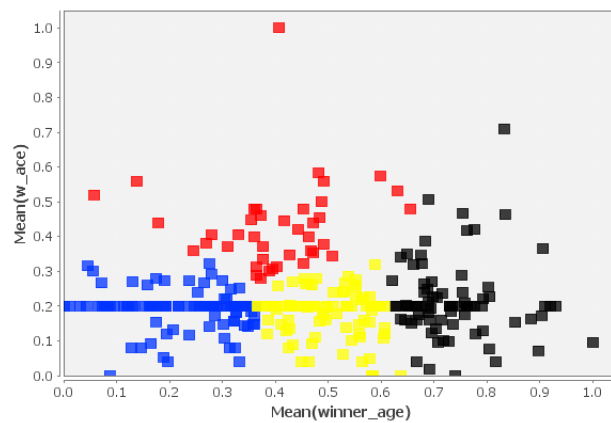
Slika 8: KNIME klasterovanje

Postojao je jedan igrač kome je bio nepoznat broj godina, njega smo obrisali. U situacija kada je broj asova bio nepoznat, stavljali smo vrednost na 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 morali smo još da normalizujemo podatke kako bi broj godina i broj asova imali isti uticaj na računanje rastojanja među instancama.

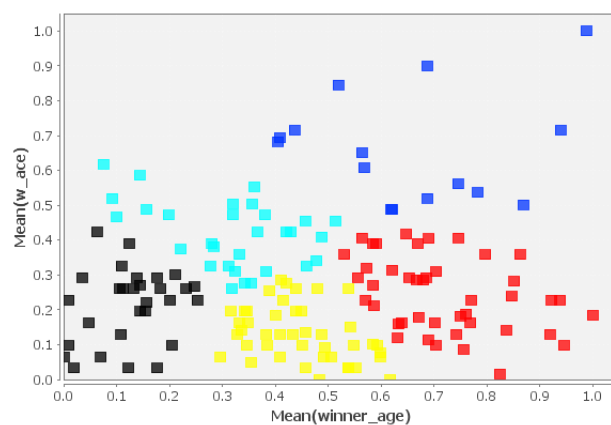
U KNIME-u smo se opredelili da klasterovanje vršimo sa istim broj klastera kao što smo činili u SPSS-u. Dobijeni klasteri su prikazani na slici 9.



(a) 2003.godina



(b) 2011.godina



(c) 2017.godina

Slika 9: Klasteri

Možemo primetiti da stvarno postoje autlajeri koji su uticali na to da se naprave novi klasteri. Igrači koji predstavljaju autlajere su dati na slici 10.

Row ID	S winner_name	D Mean(winner_age)	D Mean(w_age)
Row91	Frederic Niemeyer	27.162	31

(a) 2003.godina

Row ID	S winner_name	D Mean(winner_age)	D Mean(w_age)
Row97	Fritz Wolmarans	24.903	25

(b) 2011.godina

Row ID	S winner_name	D Mean(winner_age)	D Mean(w_age)
Row68	Ivo Karlovic	37.864	30.75

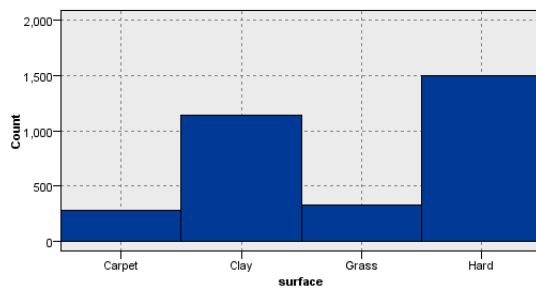
(c) 2017.godina

Slika 10: Klasteri

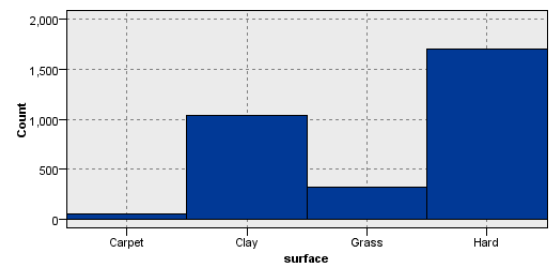
Najinteresantiji je definitivno Ivo Karlović, koji je na meču Autralian Open-a protiv Orasia Zebaljosa postigao 75 asova. Međutim treba imati na umu da je ukupan broj mečeva koje je odigrao na ovom turniru je 4. Još moramo napomenuti, kako je skup objavljen 2017., dosta turnira nije još uvek upisano, samim tim ni potpuni, trenutni skup ima samo podatke sa turnira do sredine jula. **proveriti do kad ima statistika (Ovo smo mi lupili :P)**

5 Klasifikacija

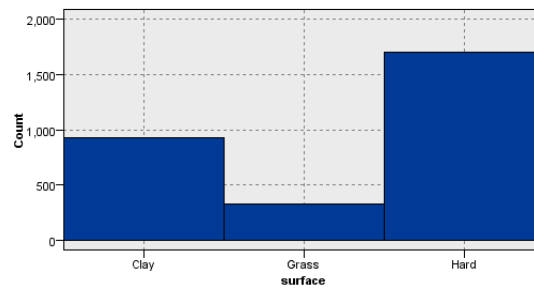
Odlučili smo da nam klase budu podloge terena, a pripadnost svakoj klasi se određuje na osnovu karakteristika gubitnika meča. **Možda i ovde da probamo da grupišemo prvo po podlogama, pa pripadnost da gadamo na osnovu proseka?** Kao i kod klasterovanja i ovde smo želeli da vidimo šta se dešava u više godina. S obzirom na to da se su neke podloge dominantnije u odnosu na ostale, godine za klasterovanje smo izabrali koliko je to moguće najpravnomernije. Nakon detaljne analize raspodele podloga po turnirima odlučili smo se za 2005, 2008, 2015. godinu. Razlog što smo odabrali baš ove godine jeste polako gubljenje ~tepiha~ kao podloge (slika 11), pa nas je zanimalo kako će ova činjenica uticati na sam proces klasifikacije.



(a) 2005.godina



(b) 2008.godina

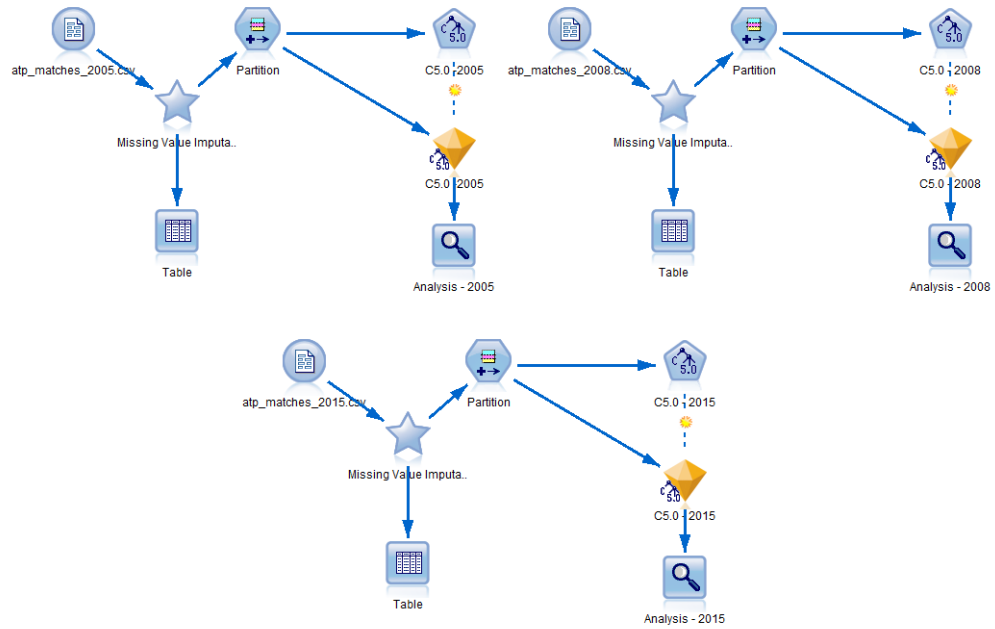


(c) 2015.godina

Slika 11: Podloga terena

Klasifikaciju smo, takođe, obradili u SPSS-u i u KNIME-u. (slike 12 i ??).

5.1 SPSS



Slika 12: SPSS klastifikacija

Pre svega nas je zanimalo ponašanje *C-50* algoritma. Dobijene su matrice konfuzije prikazane na 13.

Results for output field surface

Comparing SC-surface with surface

'Partition'	Testing	Training
Correct	501 53.13%	1,428 70.87%
Wrong	442 46.87%	587 29.13%
Total	943	2,015

Coincidence Matrix for SC-surface (rows show actuals)

'Partition' = Testing	Clay	Grass	Hard
Clay	67	19	222
Grass	16	12	86
Hard	68	31	422

'Partition' = Training	Clay	Grass	Hard
Clay	232	19	369
Grass	13	84	117
Hard	50	19	1,112

(a) 2005.godina

Results for output field surface

Comparing SC-surface with surface

'Partition'	Testing	Training
Correct	514 51.76%	1,605 75.81%
Wrong	479 48.24%	512 24.19%
Total	993	2,117

Coincidence Matrix for SC-surface (rows show actuals)

'Partition' = Testing	Carpet	Clay	Grass	Hard
Carpet	0	2	3	15
Clay	3	102	14	212
Grass	0	18	15	61
Hard	3	130	18	397

'Partition' = Training	Carpet	Clay	Grass	Hard
Carpet	6	3	1	21
Clay	1	432	6	270
Grass	1	29	101	93
Hard	1	69	17	1,066

(b) 2008.godina

Results for output field surface

Comparing SC-surface with surface

'Partition'	Testing	Training
Correct	442 42.79%	1,679 75.49%
Wrong	591 57.21%	545 24.51%
Total	1,033	2,224

Coincidence Matrix for SC-surface (rows show actuals)

'Partition' = Testing	Carpet	Clay	Grass	Hard
Carpet	8	32	6	43
Clay	27	174	22	142
Grass	5	32	10	56
Hard	36	163	27	250

'Partition' = Training	Carpet	Clay	Grass	Hard
Carpet	102	47	2	40
Clay	14	630	9	125
Grass	16	43	113	58
Hard	11	158	22	834

(c) 2015.godina

Slika 13: Matrica konfuzije C-50

Prikaz najznačajnijih atributa može se videti na 14, a detaljniji prikaz drveća odlučivanja: 2005, 2008, 2015.
Sveobuhvatna analiza je data: 2005, 2008, 2015.

