

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 | KNIME | 12 |
| 5.1.1 | SVM | 12 |
| 5.2 | SPSS | 14 |

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.

| 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 šansi pobjednika |
| w_bpFaced | broj brejk šansi na servis pobjednika |

Tabela 2: Podaci o pobjednicima

U tabeli 1 prikazani su podaci o turniru.

U tabeli 2 prikazani su podaci o pobjedniku meča.

U tabeli 3 prikazani su podaci o gubitniku meča.

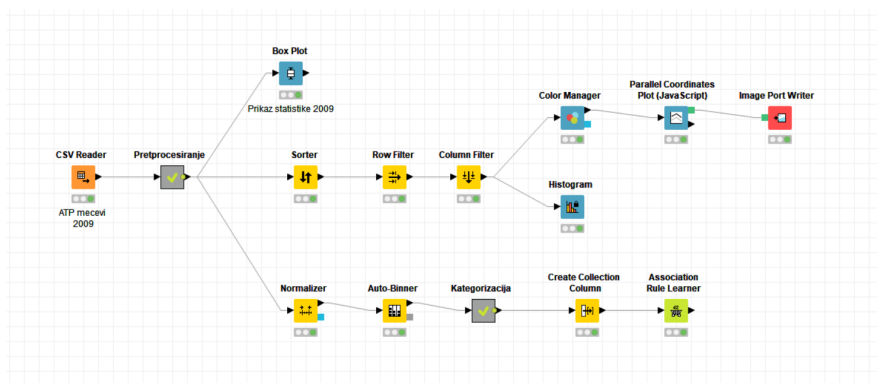
S obzirom na veliki broj raspoloživih godina, prvo smo se detaljno upoznali sa podacima, šta nam koja godina pruža i koji su najzanimljiviji atributi za svaku godinu. U zavisnosti od toga smo, po potrebama metoda, koristili različite godine, ali svuda smo se ograničili na četiri maksimalno.

| 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

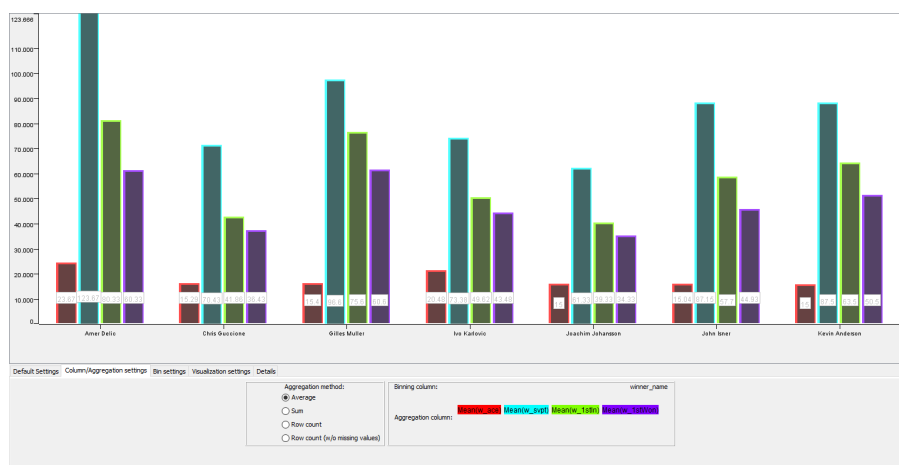
3 Pravila pridruživanja

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.

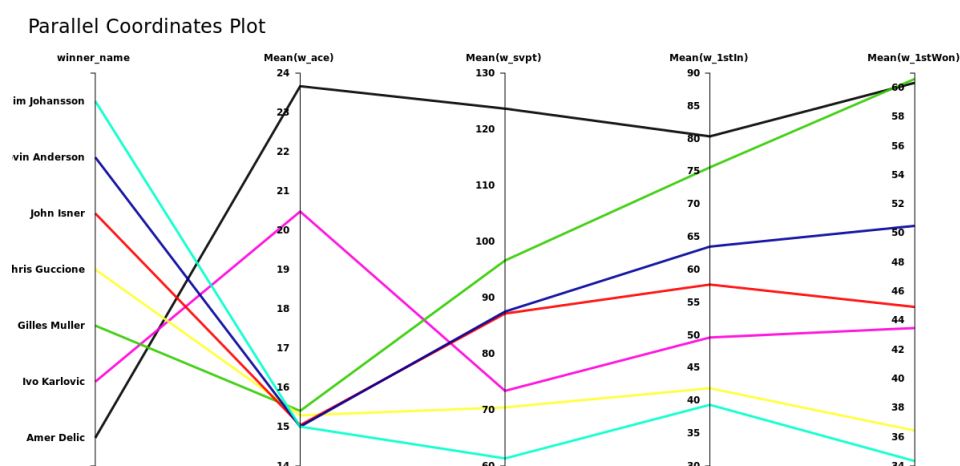


Slika 1: KNIME implementacija

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

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, 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

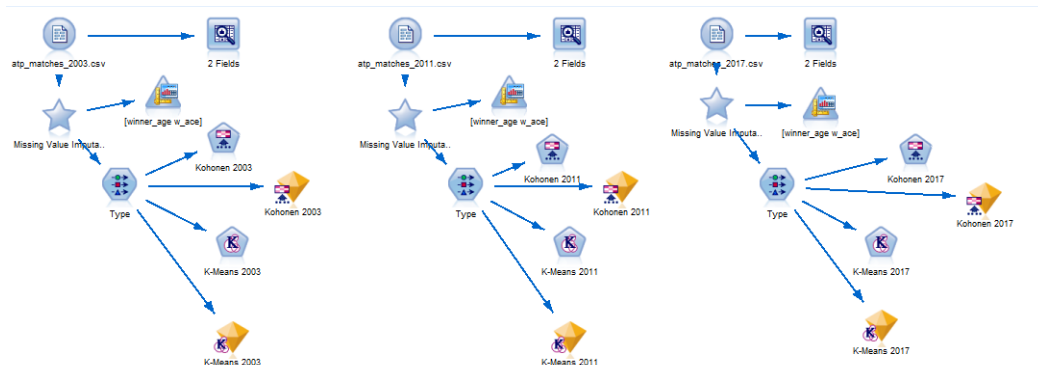
| Row ID | D Support | D Confide... | D ▼ Lift | S Conseq... | S implies | (...) Items |
|--------|-----------|--------------|----------|-------------|-----------|---------------------|
| rule2 | 0.177 | 0.837 | 1.36 | BPS 1 | <--- | [ACE 2,WON 2,DF 1] |
| rule4 | 0.197 | 0.816 | 1.326 | BPS 1 | <--- | [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

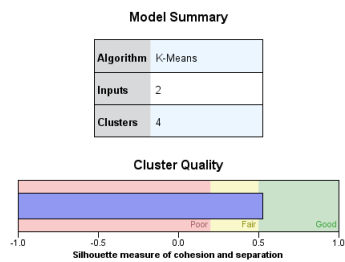
Što se tiče klasterovanja, s obzirom da podaci po godinama dosta osciliraju, odlučili smo da klasterovanje izvršimo za više godina. Izabrali smo 2003., 2011. i 2017. godinu. Pre svega, zanimala nas je zavisnost broja godina pobednika i broj asova pobednika. Prvo smo obradili nedostajuće vrednosti. Klasterovanje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

4.1 SPSS

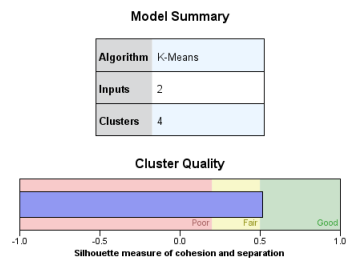


Slika 5: SPSS klasterovanje

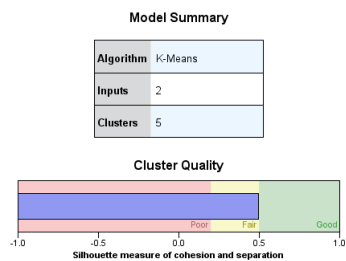
U alatu SPSS smo pomoću siluete pratili kako nam se kvalitet klasterovanja razlikuje u zavisnosti od broja klastera. *Kohonen* algoritam nam je za broj klastera između 3 i 5 davao "osrednji" kvalitet klasterovanja, pa ga nismo detaljno razmatrali. (Podseti me da probam sa nekim drugim ulaznim argumentima za Kohonena i da napisem ovde za koje smo probali.) 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", promeniš 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"; 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 smo odlučili da 2017. godinu odbradimo sa 5 klastera. Konačno, odlučili smo se za broj i kvalitet klastera koji su prikazani na slici 6.



(a) 2003. godina



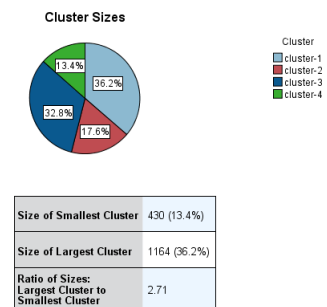
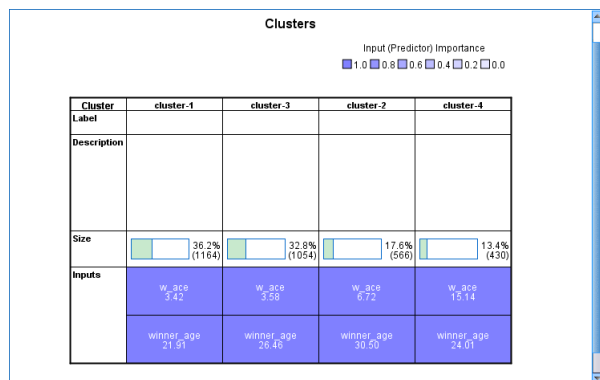
(b) 2011. godina



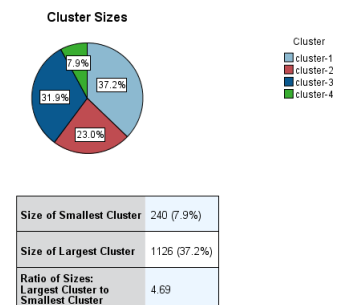
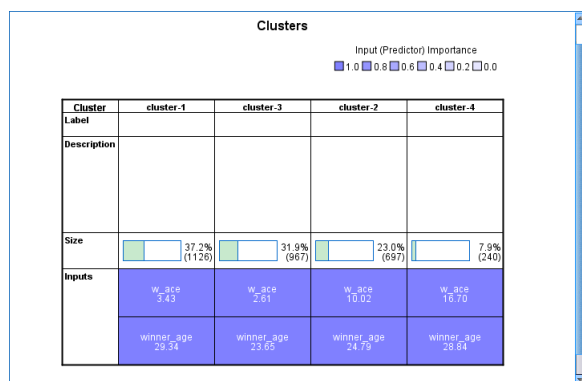
(c) 2017. godina

Slika 6: Kvalitet klasterovanja

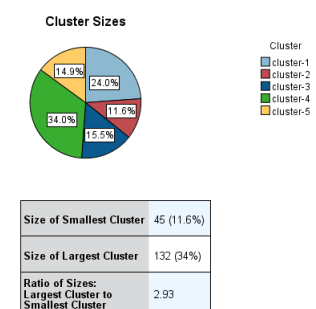
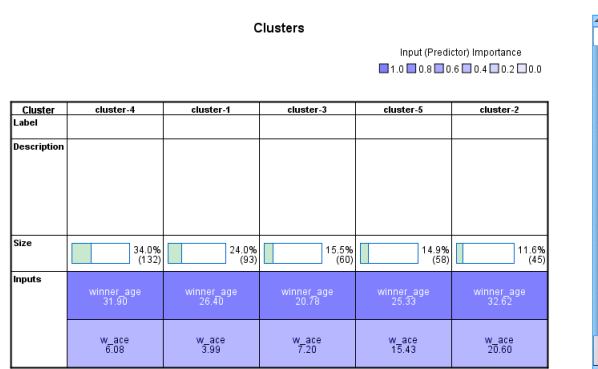
Kao što se vidi na slici 7, u sve tri godine su dobijeni interesantni podaci. Na primer, u 2003. godini najstariji igrači imaju slabiji prosek asova, dok u 2011. i 2017. godini imamo dva klastera sa prosekom godina oko 30; u jednom klasteru nam je broj asova mali, dok je u drugom najveći. Ovo nam je 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



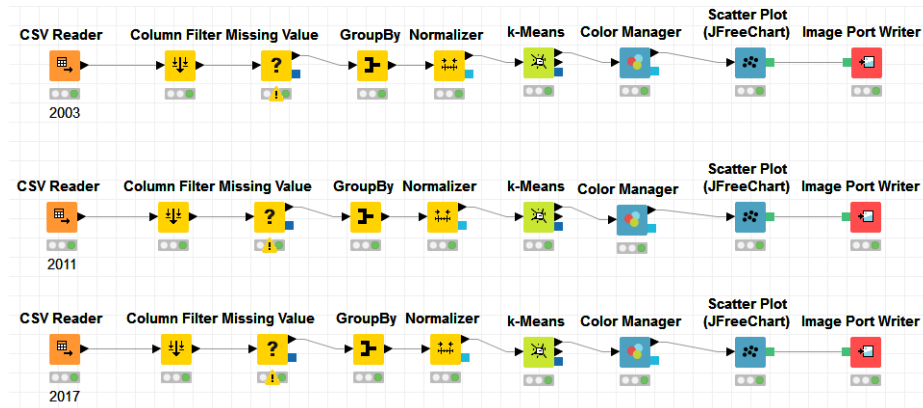
(b) 2011. godina



(c) 2017. godina

Slika 7: Modeli klastera

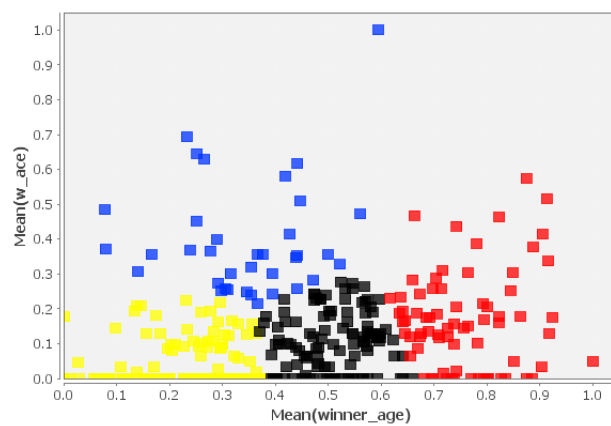
4.2 KNIME



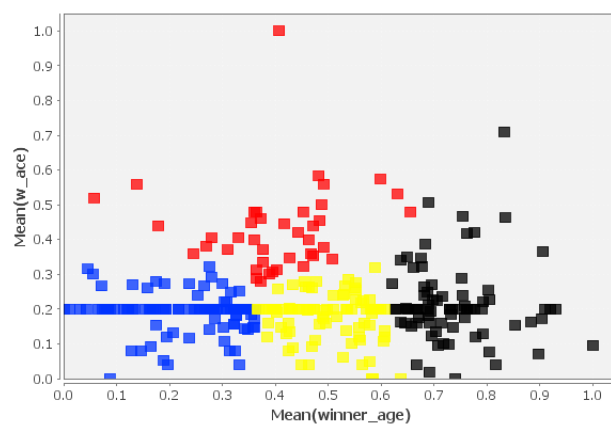
Slika 8: KNIME klasterovanje

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.

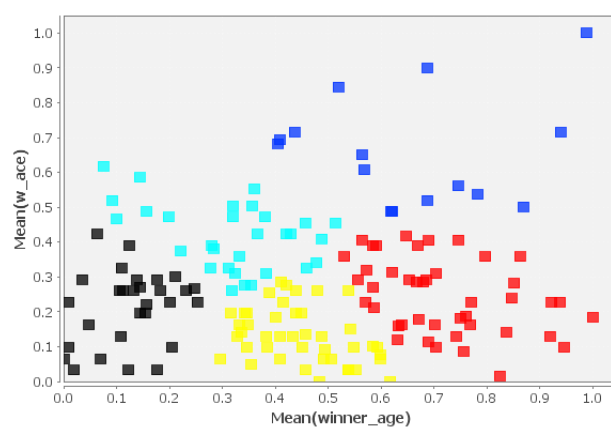
U KNIME-u smo se opredelili da klasterovanje vršimo sa istim brojem klastera kao što smo činili u SPSS-u. Dobijeni klasteri su prikazani na slici 9.



(a) 2003. godina



(b) 2011. godina



(c) 2017. godina

Slika 9: Klasteri po godinama

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 10.

| Row ID | S winner_name | D ▼ Mean(winner_age) | D ▼ Mean(w_ace) |
|--------|-------------------|----------------------|-----------------|
| Row91 | Frederic Niemeyer | 27.162 | 31 |

(a) 2003. godina

| Row ID | S winner_name | D Mean(winner_age) | D ▼ Mean(w_ace) |
|--------|-----------------|--------------------|-----------------|
| Row97 | Fritz Wolmarans | 24.903 | 25 |

(b) 2011. godina

| Row ID | S winner_name | D Mean(winner_age) | D ▼ Mean(w_ace) |
|--------|---------------|--------------------|-----------------|
| Row68 | Ivo Karlovic | 37.864 | 30.75 |

(c) 2017. godina

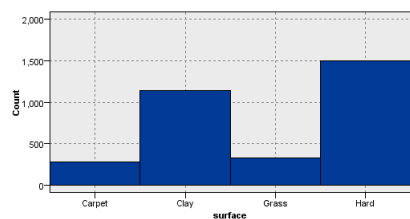
Slika 10: Elementi van granica

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. **prevereno!**

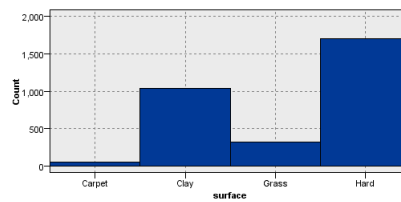
5 Klasifikacija

Za klasifikaciju smo odlučili 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 gađamo na osnovu proseka?** 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, godine za klasterovanje smo izabrali što ravnomernije. Nakon detaljne analize raspodele podloga po turnirima, odlučili smo se za 2005., 2008. i 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.

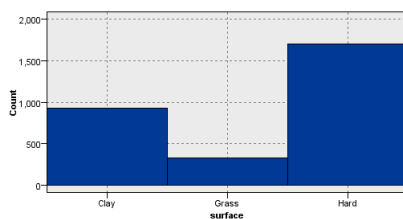
Klasifikaciju smo vršili na četiri normalizovana atributa: broj asova gubitnika, broj duplih servis grešaka gubitnika, broj ubačenih prvih servisa gubitnika, broj brejk šansi na servis gubitnika. Ispitivana je zavisnost ovih atributa u odnosu na podlogu na kojoj se igra meč. **Trebalo bi da uskladimo attribute koje koristimo za klasifikaciju, ja sam u SPSS sve gubitnikove karakteritike stavila.**



(a) 2005. godina



(b) 2008. godina



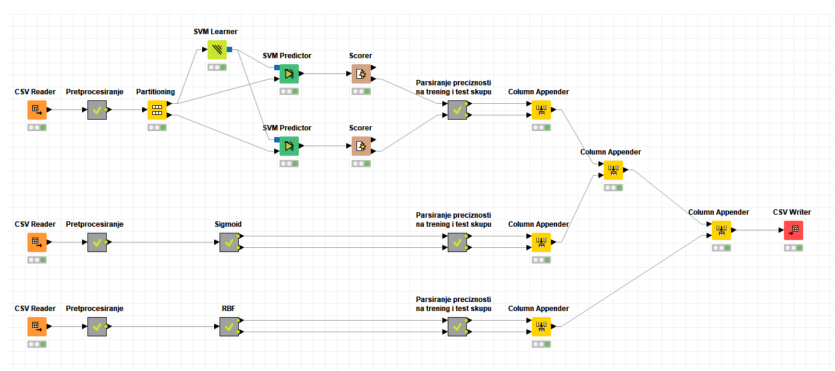
(c) 2015. godina

Slika 11: Podloge terena

Klasifikaciju smo, takođe, obradili u SPSS-u i u KNIME-u. (slike 20 i 12).

5.1 KNIME

5.1.1 SVM



Slika 12: KNIME implementacija SVM tehnike

Vršili smo klasifikaciju tehnikom *SVM*. Normalizovane podatke smo podelili na trening i test skup u odnosu 70-30. Primenili smo sva tri raspoloživa kernela (polinomijalni trećeg stepena, sigmoid, Gausov(RBF)). Na slici 13 se mogu videti preciznosti za sva tri kernela, i za trening i za test skup.

| | | | | | |
|----------------------------------|------------------------------|-------------------------------------|-------------------------|---------------------------------|-----------------------------|
| Acc_Training_Poly 0.355419043 | Acc_Test_Poly 0.351738241 | Acc_Training_Sigmoid 0.492321194 | Acc_Test_Sigmoid 0.5 | Acc_Training_RBF 0.482229048 | Acc_Test_RBF 0.460122699 |
|----------------------------------|------------------------------|-------------------------------------|-------------------------|---------------------------------|-----------------------------|

Slika 13: 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 14 i 15 vidimo da su podaci pogrešno klasifikovani u mečeve koji su odigrani na šljaci.

| Row ID | Hard | Clay | Carpet | Grass |
|--------|------|------|--------|-------|
| Hard | 11 | 1039 | 0 | 0 |
| Clay | 1 | 799 | 0 | 0 |
| Carpet | 1 | 195 | 0 | 0 |
| Grass | 6 | 227 | 0 | 0 |

Slika 14: Trening podaci za polinomijalni kernel

| Row ID | Hard | Clay | Carpet | Grass |
|--------|------|------|--------|-------|
| Hard | 2 | 449 | 0 | 0 |
| Clay | 1 | 342 | 0 | 0 |
| Carpet | 0 | 84 | 0 | 0 |
| Grass | 3 | 97 | 0 | 0 |

Slika 15: 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 16 i 17. Primetimo da su podaci uglavnom raspoređeni u klase koje se odnose na beton i šljaku.

| Row ID | Hard | Carpet | Clay | Grass |
|--------|------|--------|------|-------|
| Hard | 974 | 2 | 74 | 0 |
| Carpet | 183 | 0 | 13 | 0 |
| Clay | 647 | 5 | 148 | 0 |
| Grass | 217 | 1 | 15 | 0 |

Slika 16: Trening podaci za sigmoid kernel

| Row ID | Hard | Carpet | Clay | Grass |
|--------|------|--------|------|-------|
| Hard | 425 | 0 | 26 | 0 |
| Carpet | 81 | 0 | 3 | 0 |
| Clay | 276 | 3 | 64 | 0 |
| Grass | 93 | 0 | 7 | 0 |

Slika 17: Test podaci za sigmoid kernel

Koristeći Gausov kernel, dobili smo lošiju klasifikaciju za tvrdu podlogu, dosta bolju klasifikaciju za šljaku i malo bolju klasifikaciju za travu (slike 18 i 19).

| Row ID | Hard | Clay | Grass | Carpet |
|--------|------|------|-------|--------|
| Hard | 743 | 210 | 97 | 0 |
| Clay | 376 | 335 | 89 | 0 |
| Grass | 182 | 30 | 21 | 0 |
| Carpet | 133 | 39 | 24 | 0 |

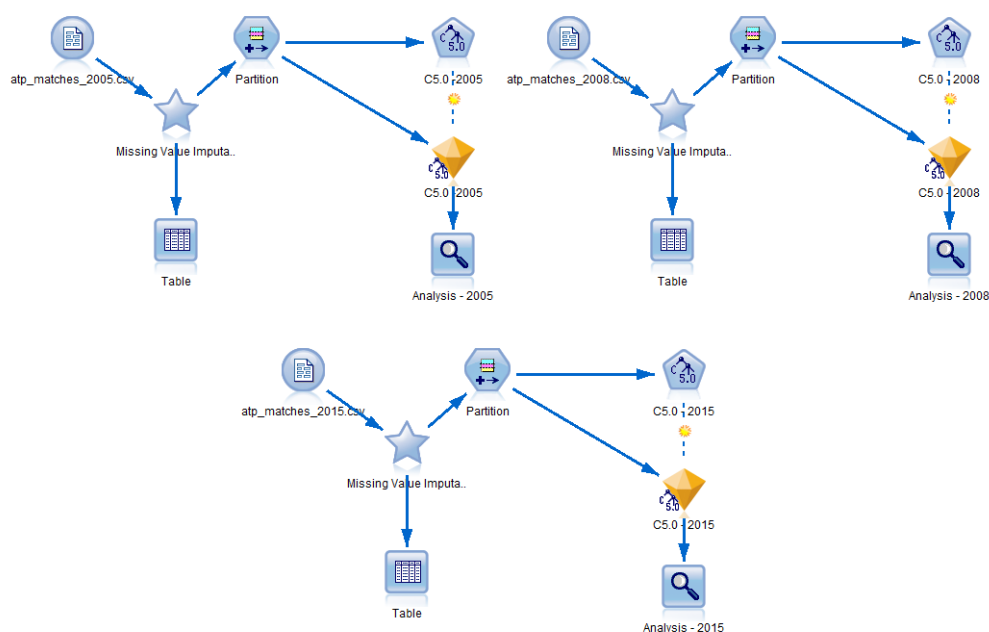
Slika 18: Trening podaci za Gausov kernel

| Row ID | Hard | Clay | Grass | Carpet |
|--------|------|------|-------|--------|
| Hard | 304 | 96 | 51 | 0 |
| Clay | 157 | 139 | 47 | 0 |
| Grass | 74 | 19 | 7 | 0 |
| Carpet | 59 | 16 | 9 | 0 |

Slika 19: Test podaci za Gausov kernel

Prikazani rezultati su za 2005. godinu i, kako je u svim slučajevima klasifikacija koja se odnosila na tepih davala nulu, slični rezultati su očekivani i za 2008. godinu. Stoga nismo obrađivali ostale godine *SVM* metodom, već smo pokušali da obradimo podatke algoritmom *C5.0* u SPSS-u.

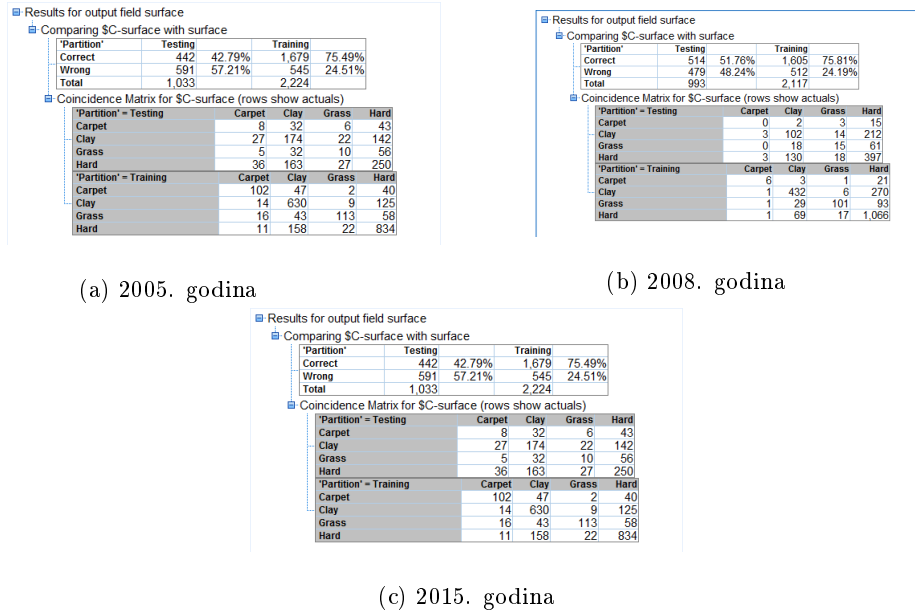
5.2 SPSS



Slika 20: SPSS klasifikacija

Primenili smo *C5.0* algoritam sa podelom na trening i test skup u odnosu 70-30 i dobili smo neočekivane matrice konfuzije koje su prikazane

na slici 21.



Slika 21: Matrice konfuzije - C5.0

U 2005. godini, za razliku od *SVM* metode, podloga "tepih" se na trening skupu klasifikovala dobro u 102 slučaja, dok u ostalim pretežno bira "beton" ili "šljaku". Situacija na test skupu je raznolika, uz dominaciju "betona" i "šljake". Takođe, možemo da primetimo da nam je preciznost klasifikacije test skupa jako loša, tj. 57.21% podataka se pogrešno klasifikovalo, dok je trening skup pokazao 24.51% loše klasifikovanih. U ostalim godinama možemo videti da, kako se "tepih" gubi to se i smanjuje procenat pogrešno klasifikovanih podloga, ali opet uz glavnu dominaciju "beton" podloge. U 2015. godini gotovo sve podloge na test skupu su pretežno klasifikovane u "beton", što je i očekivano s obzirom da se najviše turnira igra upravo na "betonu".

C5.0 nam je dao i prikaz najznačajnijih atributa, kao i drvo odlučivanja. Prikaz najznačajnijih atributa je dat na slici 22, a detaljniji prikaz drveta odlučivanja: 2005, 2008, 2015.

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

