

Istraživanje podataka

Seminarski rad

Analiza podataka iz serijala

"Igra Prestola"

Filip Kristić 335/2015

Ana Petrović 195/2015

I Uvod

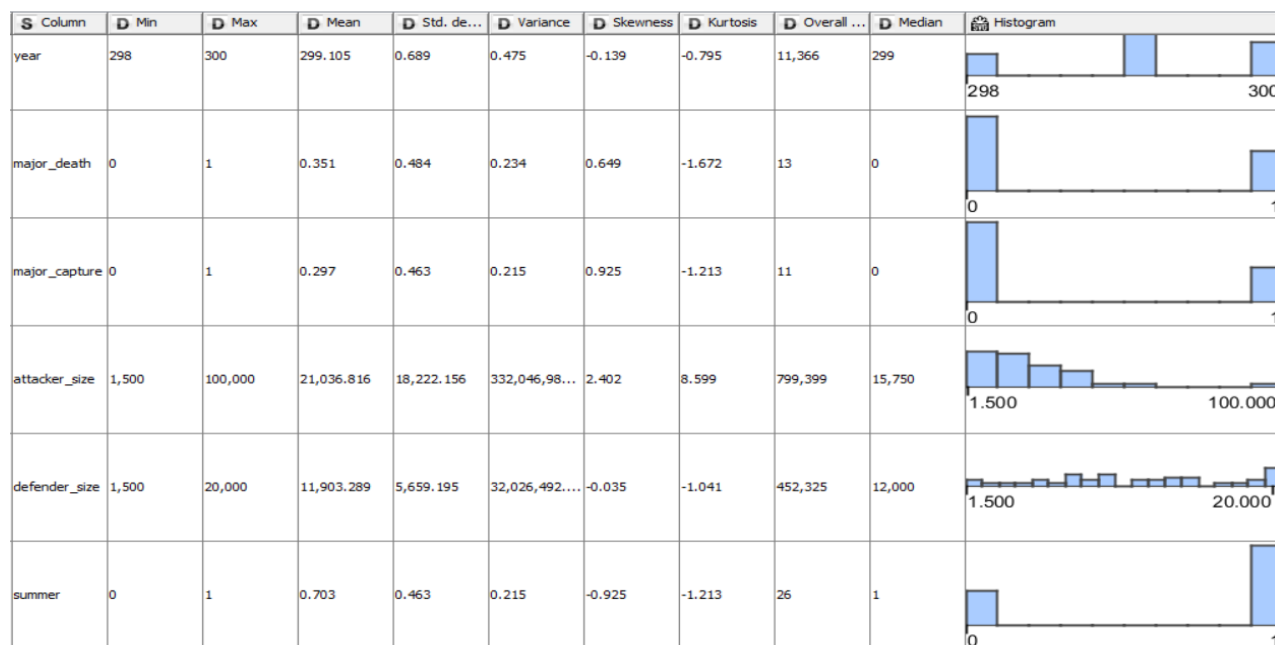
Skup podataka koji će biti analiziran se može pronaći na sledećoj [adresi](#) .

Za analizu je korišćen alat KNIME. Podaci se sastoje od tri tabele, koje sadrže informacije o bitkama i smrti likova iz serijala knjiga „Ples leda i vatre“.

U prvoj datoteci - **battles.csv**, nalaze se podaci o svim bitkama, kojih ukupno ima 38. Tabela sadrži 25 atributa. U tabeli su prikazani samo neki od njih, kao i njihovi opisi:

| | |
|------------------|--|
| Name | Naziv bitke |
| Year | Godina bitke |
| battle_number | Redni broj bitke – id |
| attacker_king | Kralj koji napada |
| defender_king | Kralj koji brani |
| attacker_outcome | Ishod bitke za napadača – može biti win ili loss |
| battle_type | Tip bitke – može biti pitched battle, ambush, razing ili seige |
| major_death | Važna smrt – može biti 1 ili 0 |
| major_capture | Važan zarobljenik – može biti 1 ili 0 |
| attacker_size | Veličina vojske napadača |
| deffender_size | Veličina vojske koja brani |
| Summer | Da li je leto - vrednosti 0 ili 1 |
| location | Lokacija bitke |

Ime bitke, kao i redni broj su kolone koje ničim ne doprinose, tako da njih nećemo uzimati u obzir pri daljoj analizi (brišemo ih iz tabele koristeći čvor Column Filter). Nakon što je datoteka učitana, upotrebićemo čvor Statistics iz kojeg ćemo dobiti informacije o nekim statističkim merama numeričkih podataka, kao što su medijana, srednja vrednost, kvartili, itd.








Iz ovoga vidimo da su godine bitki između 298. i 300. godine, kao i da su veličine napadačke vojske u proseku skoro duplo veće od vojske koja je napadnuta.

Druga datoteka – **character-deaths.csv**, sadrži podatke o smrti svakog lika iz knjiga. Ona ima 13 kolona, koje su izlistane u sledećoj tabeli:

| | |
|--------------------|---|
| Name | Ime lika |
| Allegiances | Kuće sa kojima je dati lik u savezu |
| Death year | Godina smrti |
| Book of death | U kojoj od pet knjiga je dati lik umro |
| Death chapter | Poglavlje u knjizi u kojem se dešava data smrt – od 0 do 80 |
| Book intro chapter | Poglavlje u kojem se lik prvi put pojavljuje |
| Gender | Pol lika – 1 označava muški pol, 0 ženski |
| Nobility | Da li je lik plemenitog porekla – 1 ili 0 |
| GoT | Pojavljivanje u 1. knjizi – 1 ili 0 |
| CoK | Pojavljivanje u 2. knjizi – 1 ili 0 |
| SoS | Pojavljivanje u 3. knjizi – 1 ili 0 |
| FfC | Pojavljivanje u 4. knjizi – 1 ili 0 |
| DwD | Pojavljivanje u 5. knjizi – 1 ili 0 |

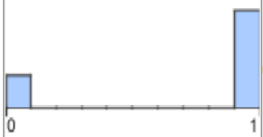

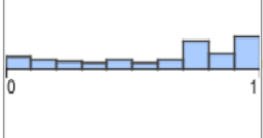
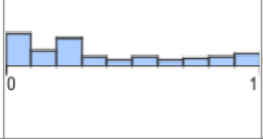
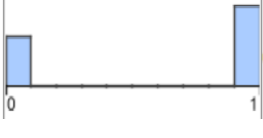
Na ovu datoteku primenjujemo čvor Statistics, kao i na prethodnu:

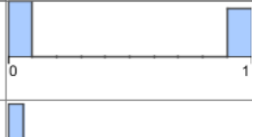
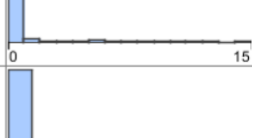

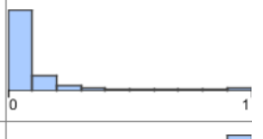
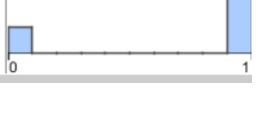

| Column | Min | Max | Mean | Std. deviation | Variance | Skewness | Kurtosis | Overall sum | No. missings | Median | Histogram |
|-------------------|-----|-----|---------|----------------|----------|----------|----------|-------------|--------------|--------|---|
| Death Year | 297 | 300 | 299.157 | 0.703 | 0.495 | -0.402 | -0.317 | 91,243 | 612 | 299 |  |
| Book of Death | 1 | 5 | 2.928 | 1.326 | 1.76 | 0.234 | -0.994 | 899 | 610 | 3 |  |
| Death Chapter | 0 | 80 | 40.07 | 20.47 | 419.032 | -0.151 | -0.955 | 11,981 | 618 | 39 |  |
| Book Intro Cha... | 0 | 80 | 28.862 | 20.166 | 406.639 | 0.353 | -0.863 | 26,120 | 12 | 27 |  |
| Gender | 0 | 1 | 0.829 | 0.377 | 0.142 | -1.749 | 1.06 | 760 | 0 | 1 |  |

Odavde možemo zaključiti da su sve godine smrti imeđu 297. i 300. godine, kao i to da postoji mnogo više muških likova nego ženskih. Ova tabela je slična trećoj, i ne poseduje neke značajne informacije, stoga je nećemo razmatrati u nastavku.

Treća tabela zove se **character-predictions.csv**, ima 33 atributa, od kojih nećemo koristiti sve (zbog velikog broja nedostajućih vrednosti) i 1946 redova, i u njoj se nalaze neka predviđanja o smrti likova. Prikazaćemo neke osobine podataka, kao i za prve dve tabele:

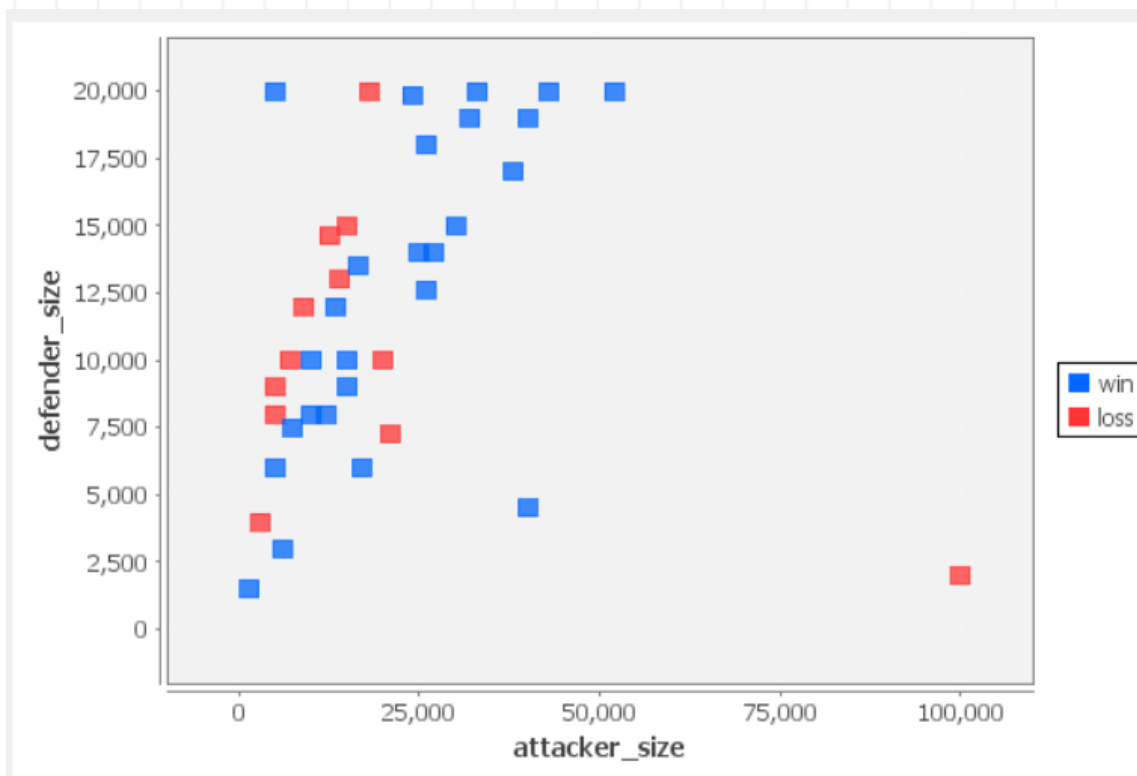
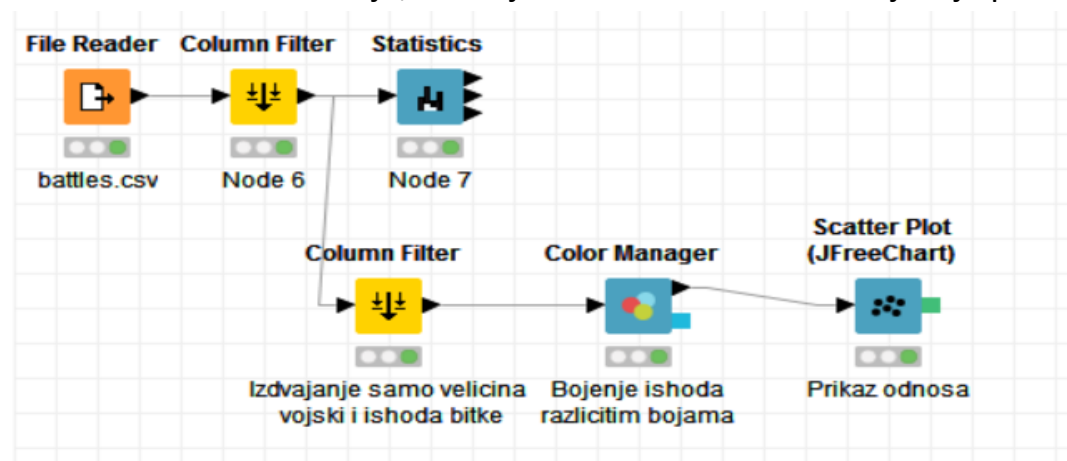
| | |
|-----------------------------------|--|
| actual | 1 ako živ, 0 ako je mrtav |
| dateOfBirth | Datum rođenja lika |
| dateOfDeath | Datum smrti lika |
| culture | Kultura lika – Andals, Astapor, Braavosi, Dornish, Dothraki, Ironborn itd. |
| male | Ako je muško 1, inače 0 |
| popularity | Procenat popularnosti |
| house | Kojoj kući pripada |
| isNoble | Da li pripada plemenitoj porodici, 1 ili 0 |
| book1, book2, book3, book4, book5 | Pojavljivanje u knjigama, 1 ili 0 |
| numDeadRelations | Broj mrtvih sa kojima je lik povezan |

| S Column | D Min | D Max | D Mean | D Std. de... | D Variance | D Skewness | D Kurtosis | D Overall ... | D Median | Histogram |
|----------|-------|-------|--------|--------------|------------|------------|------------|---------------|----------|---|
| actual | 0 | 1 | 0.746 | 0.436 | 0.19 | -1.129 | -0.726 | 1,451 | 1 |  |
| pred | 0 | 1 | 0.687 | 0.464 | 0.215 | -0.807 | -1.349 | 1,337 | 1 |  |
| alive | 0 | 1 | 0.634 | 0.313 | 0.098 | -0.664 | -0.87 | 1,234.678 | 0.736 |  |
| plod | 0 | 1 | 0.366 | 0.313 | 0.098 | 0.664 | -0.87 | 711.322 | 0.265 |  |
| male | 0 | 1 | 0.619 | 0.486 | 0.236 | -0.491 | -1.76 | 1,205 | 1 |  |

| S Column | D Min | D Max | D Mean | D Std. deviation | D Variance | D Skewness | D Kurtosis | D Overall sum | D Median | Histogram |
|----------------|-------|-------|--------|------------------|------------|------------|------------|---------------|----------|---|
| isNoble | 0 | 1 | 0.461 | 0.499 | 0.249 | 0.157 | -1.977 | 897 | 0 |  |
| numDeadRel... | 0 | 15 | 0.306 | 1.384 | 1.915 | 5.715 | 37.393 | 595 | 0 |  |
| boolDeadRel... | 0 | 1 | 0.075 | 0.263 | 0.069 | 3.243 | 8.526 | 145 | 0 |  |
| isPopular | 0 | 1 | 0.059 | 0.236 | 0.056 | 3.742 | 12.018 | 115 | 0 |  |
| popularity | 0 | 1 | 0.09 | 0.161 | 0.026 | 3.699 | 15.229 | 174.331 | 0.033 |  |
| isAlive | 0 | 1 | 0.746 | 0.436 | 0.19 | -1.129 | -0.726 | 1,451 | 1 |  |

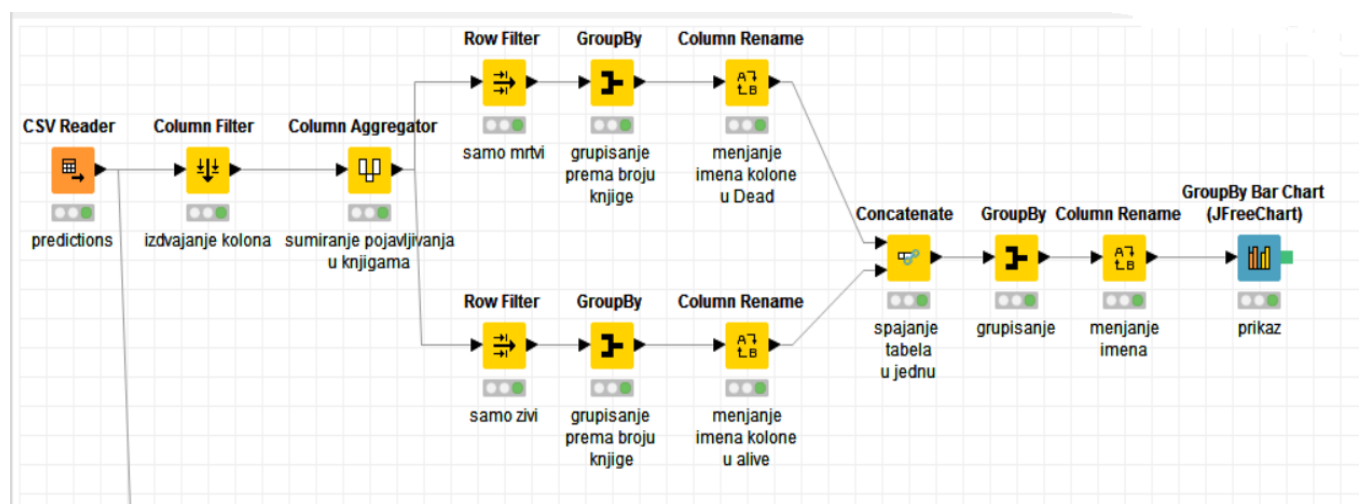
II Vizualizacija podataka

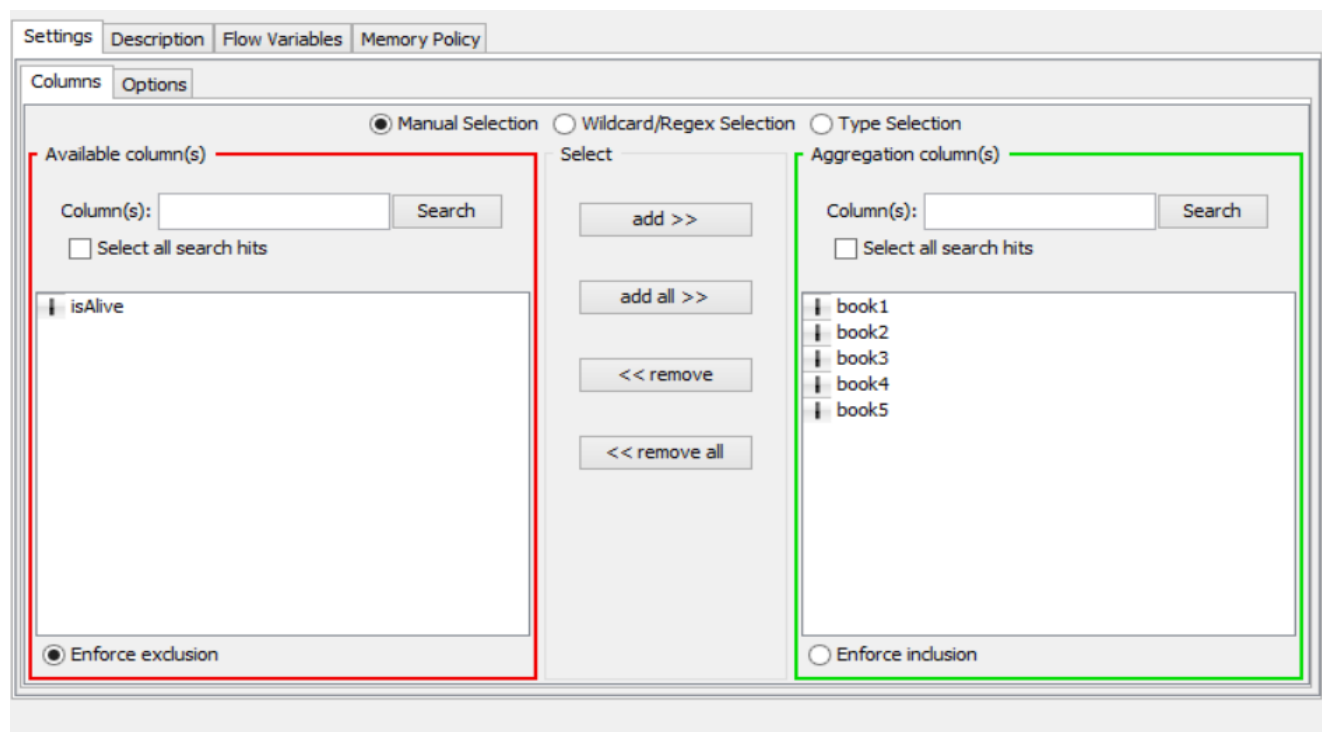
Vizuelizacijom podataka možemo grafički prikazati odnose između podataka, i tako odgovoriti na neka pitanja. Na primer, možemo videti kako veličina vojske utiče na ishod bitke za napadača. To u KNIME-u radimo tako što na učitane podatke najpre primenimo čvor Column Filter, kojim izdvajamo samo 3 kolone – attacker_size, deffender_size i attacker_outcome. Nakon toga, primenjujemo Color Manager, kojim ćemo obojiti svaku pobedu plavom bojom, a svaki poraz crvenom. Na samom kraju, ubacujemo čvor Scatter Plot koji daje prikaz zavisnosti.



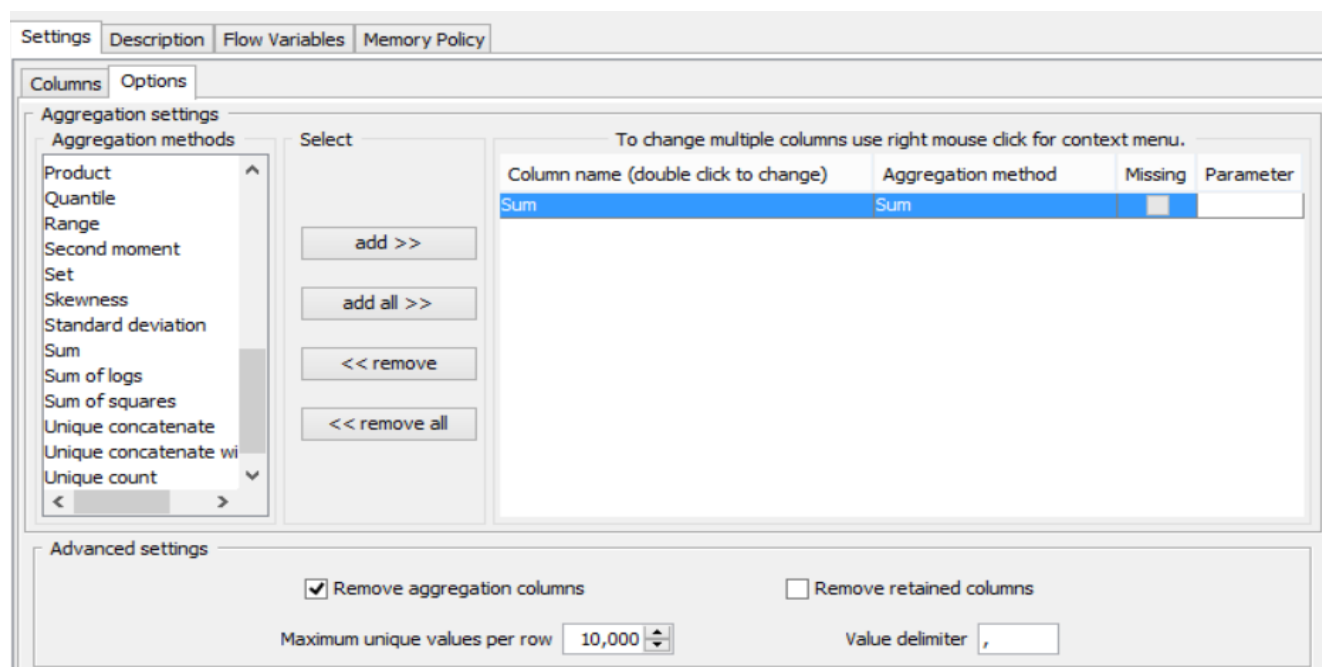
Odavde možemo da vidimo da u većem broju slučaja napadačka vojska pobeđuje, i uglavnom je veća od one koju napada. Naravno, postoje izuzeci gde vidimo da veličina vojske ne mora da presudi. Međutim, postoji jedna bitka koja se upadljivo izdvaja od ostalih, u kojoj je napadač imao neuporedivo veću vojsku, a ipak je ishod bio loss. Ta bitka predstavlja autlajer, ali s obzirom da su podaci izmišljeni, i da je ona ubačena radi razvoja događaja same priče, ne daje nam nikakve korisne informacije.

Drugi primer vizualizacije predstavimo u tabeli character-predictions koristeći Bar Chart. Pokazaćemo kako pojavljivanje lika u više knjiga utiče na to da li će lik preživeti. Kao i u prethodnom primeru, povezujemo učitane podatke sa čvorom Column filter i u njemu izdvajamo atribut actual, book1, book2, book3, book4, book5. Sledeći je čvor Column Agregator u kome ćemo kolone za svih pet knjiga sumirati u jednu (slike 1.1, 1.2) i time ćemo dobiti tabelu sa dve kolone, 1. da li je lik živ i 2. broj knjiga u kojima se taj lik pojavljuje. Zatim ćemo razdvojiti podatke na dve tabele, za žive likove i mrtve likove. To radimo sa čvorovima row filter, pa za njim GroupBy, kojim grupišemo prema broju knjige. Te dve tabele spajamo čvorom Concatenate, a onda ponovo grupišemo, tako da se napravi tabela koja sadrži kolone BrojKnjigaPojavljivanje, BrojMrtvih, BrojŽivih. Čvorom Column Rename menjamo ime koloni. Na kraju povezujemo sa čvorom GroupBy Bar Chart. Bar Chart prikazuje na x-osi broj knjige (0 je ako se lik ne pojavljuje ni u jednoj knjizi, 5 ako se pojavljuje u svih 5), a na y-osi imamo broj živih/mrtvih likova.





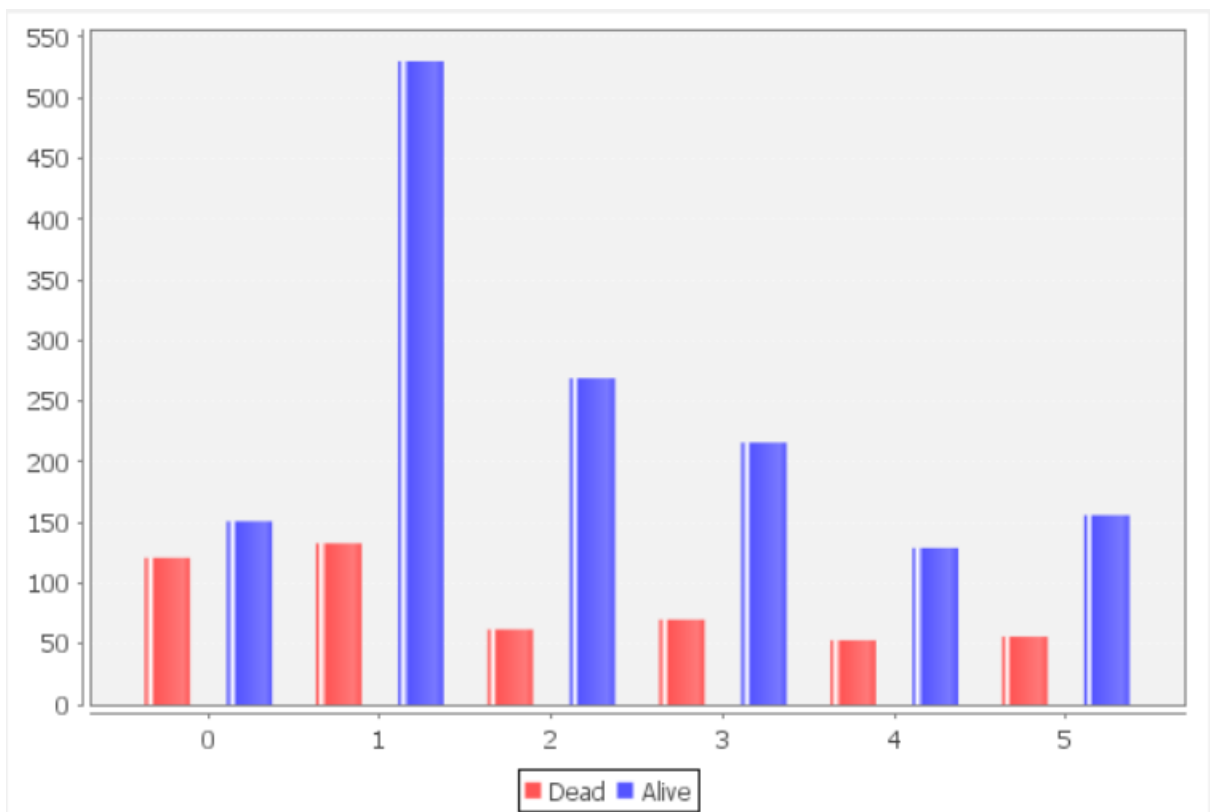
Slika 1.1.



Slika 1.2.

| Row ID | Pojavljivanje u knjizi | Dead | Alive |
|--------|------------------------|------|-------|
| Row0 | 0 | 121 | 151 |
| Row1 | 1 | 133 | 530 |
| Row2 | 2 | 62 | 269 |
| Row3 | 3 | 70 | 216 |
| Row4 | 4 | 53 | 129 |
| Row5 | 5 | 56 | 156 |

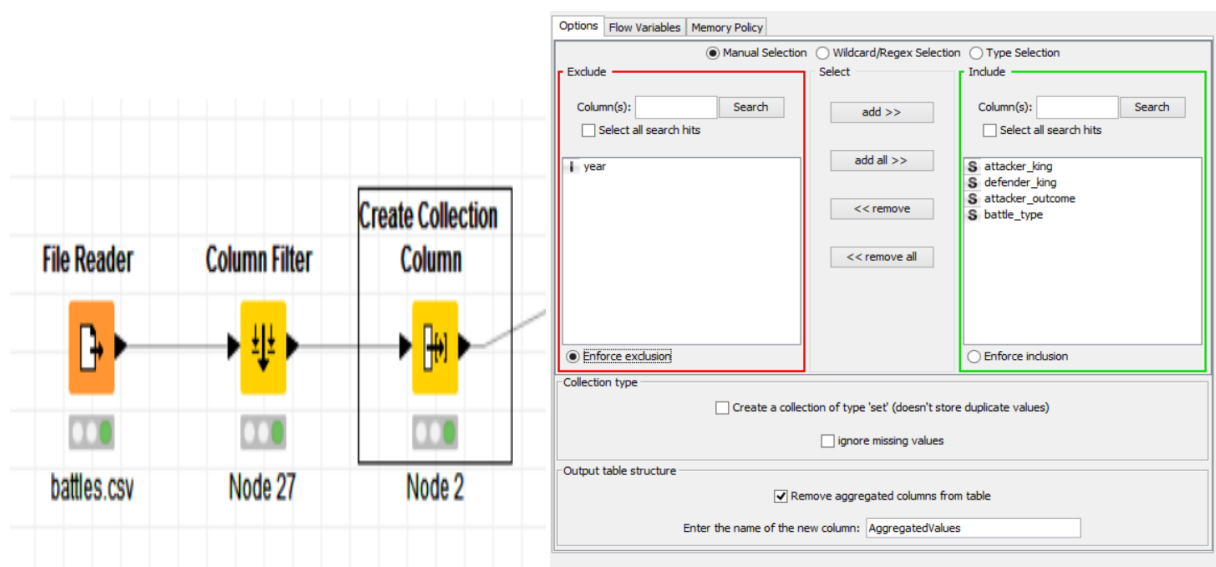
Naš Bar Chart izgleda ovako:



Vidimo da ako se likovi ne pojavljuju ni u jednoj od pet knjiga, ima više onih koji su živi nego mrtvi. U prvoj knjizi ima najviše živih, što je i logično, jer je većina likova upoznato u prvoj knjizi. Taj broj opada sa svakom sledećom knjigom. U prvoj knjizi i umre više likova nego u bilo kojoj drugoj. Ako se likovi pojavljuju u svih pet knjiga, vidi se da od ukupno 156 likova, 56 njih ne preživi.

III Pravila pridruživanja

Pravila pridruživanja na našem skupu ćemo primeniti u tabeli battles.csv da ustanovimo zavisnosti između kraljeva koji napadaju, kraljeva koji brane, tipa bitke i ishoda bitke. Najpre izdvajamo ta četiri atributa iz učitanih podataka. Koristimo Create Collection Column, kojim pravimo listu stavki :



Čvor kojim izdvajamo česte skupove i pravila pridruživanja je Association Rule Learner. U njemu navodimo minimalnu podršku (podrška meri koliko puta se dva skupa stavki zajedno pojavljuju u odnosu na ceo skup transakcija). Za podršku ćemo uneti 0.05, tj. 5%. Ako u konfiguraciji ovog čvora ne označimo Output association rules, dobićemo samo česte skupove.

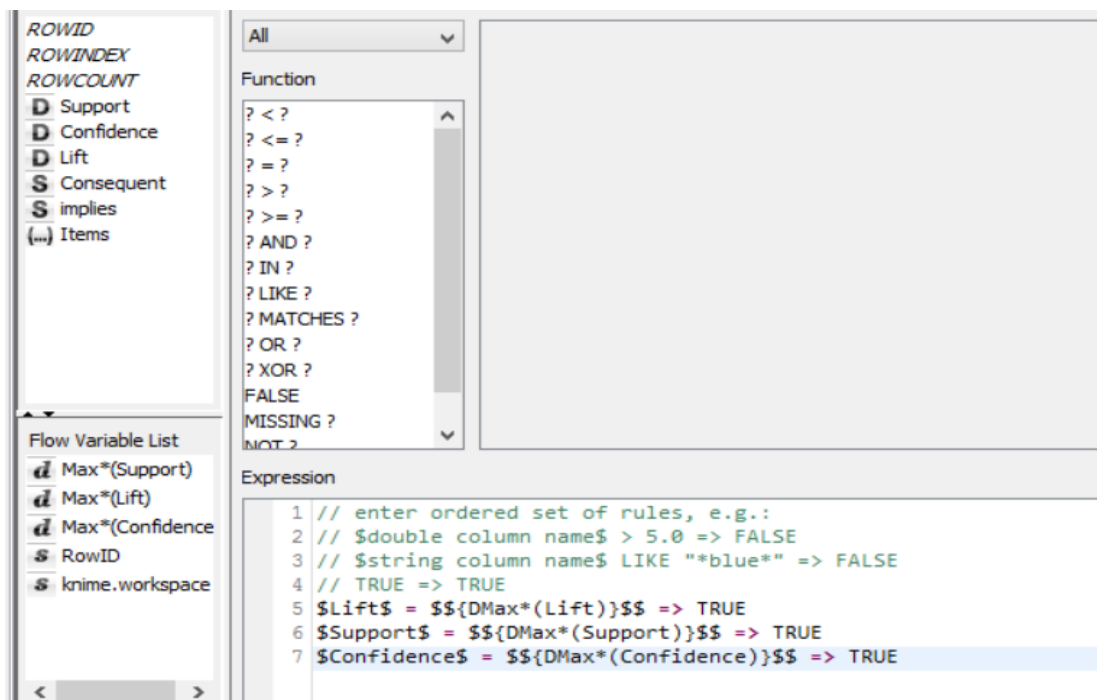
The screenshot shows the 'Association Rule Learner' configuration window. It has three tabs: 'Options', 'Flow Variables', and 'Memory Policy'. The 'Options' tab is active. It contains sections for 'Itemset Mining', 'Output', and 'Association Rules'. In 'Itemset Mining', 'Column containing transactions' is set to 'AggregatedValues', 'Minimum support (0-1)' is 0.05, and 'Underlying data structure' is 'ARRAY'. In 'Output', 'Itemset type' is 'FREE' and 'Maximal itemset length' is 10. In 'Association Rules', 'Output association rules' is checked and 'Minimum confidence' is 0.3.

Ovim smo dobili 75 čestih skupova koji imaju podršku veću od 0.05. Ako označimo Output association rules i za minimum confidence podesimo 0.4 (pouzdanost meri koliko često se javlja stavka B u transakcijama koje sadrže stavku A), dobijamo sva pravila pridruživanja $A \Rightarrow B$ za koje je podrška stavke A U B veća ili jednaka od minsup (naša je 5%) i pouzdanost pravila je veća ili jednaka od minconf (naša je 40%). Odavde dobijamo 89 pravila pridruživanja, prikazaćemo samo prvih nekoliko :

| Table "default" - Rows: 89 | | | | | | |
|----------------------------|-----------|-------------------|--------|--------------------------|----------------|--|
| | | Spec - Columns: 6 | | Properties | Flow Variables | |
| Row ID | D Support | D Confidence | D Lift | S Consequent | S implies | (...) Items |
| rule0 | 0.053 | 0.667 | 0.938 | Joffrey/Tommen Baratheon | <--- | [razing] |
| rule1 | 0.053 | 0.667 | 0.905 | win | <--- | [razing] |
| rule2 | 0.053 | 0.667 | 0.938 | Joffrey/Tommen Baratheon | <--- | [loss,Stannis Baratheon] |
| rule3 | 0.053 | 0.5 | 1.9 | loss | <--- | [Joffrey/Tommen Baratheon,Stannis Baratheon] |
| rule4 | 0.053 | 0.5 | 0.792 | Robb Stark | <--- | [Balon/Euron Greyjoy,pitched battle] |
| rule5 | 0.053 | 0.4 | 1.086 | pitched battle | <--- | [Balon/Euron Greyjoy,Robb Stark] |
| rule6 | 0.053 | 1 | 1.583 | Robb Stark | <--- | [ambush,loss] |
| rule7 | 0.053 | 1 | 1.583 | Robb Stark | <--- | [ambush,Balon/Euron Greyjoy] |
| rule8 | 0.053 | 0.4 | 1.52 | ambush | <--- | [Balon/Euron Greyjoy,Robb Stark] |
| rule9 | 0.053 | 0.667 | 1.056 | Robb Stark | <--- | [loss,Balon/Euron Greyjoy] |
| rule10 | 0.053 | 0.4 | 1.52 | loss | <--- | [Balon/Euron Greyjoy,Robb Stark] |
| rule11 | 0.053 | 1 | 1.357 | win | <--- | [pitched battle,?] |
| rule12 | 0.053 | 0.667 | 1.81 | pitched battle | <--- | [?,win] |
| rule13 | 0.053 | 0.667 | 0.905 | win | <--- | [Balon/Euron Greyjoy,siege] |
| rule14 | 0.053 | 0.667 | 0.938 | Joffrey/Tommen Baratheon | <--- | [siege,Stannis Baratheon,win] |
| rule15 | 0.053 | 1 | 1.357 | win | <--- | [Joffrey/Tommen Baratheon,siege,Stannis Baratheon] |
| rule16 | 0.053 | 1 | 3.455 | siege | <--- | [Joffrey/Tommen Baratheon,Stannis Baratheon,win] |
| rule17 | 0.079 | 1 | 1.357 | win | <--- | [?] |
| rule18 | 0.079 | 0.429 | 1.629 | loss | <--- | [Stannis Baratheon] |
| rule19 | 0.079 | 0.429 | 0.603 | Joffrey/Tommen Baratheon | <--- | [Balon/Euron Greyjoy,win] |
| rule20 | 0.079 | 1 | 1.357 | win | <--- | [Joffrey/Tommen Baratheon,Balon/Euron Greyjoy] |
| rule21 | 0.079 | 0.429 | 0.679 | Robb Stark | <--- | [Balon/Euron Greyjoy,win] |
| rule22 | 0.079 | 0.6 | 0.814 | win | <--- | [Balon/Euron Greyjoy,Robb Stark] |
| rule23 | 0.079 | 0.75 | 1.018 | win | <--- | [Balon/Euron Greyjoy,pitched battle] |
| rule24 | 0.079 | 0.429 | 1.163 | pitched battle | <--- | [Balon/Euron Greyjoy,win] |
| rule25 | 0.079 | 0.75 | 1.018 | win | <--- | [siege,Stannis Baratheon] |
| rule26 | 0.079 | 0.75 | 2.591 | siege | <--- | [Stannis Baratheon,win] |
| rule27 | 0.079 | 0.75 | 1.056 | Joffrey/Tommen Baratheon | <--- | [loss,pitched battle,Robb Stark] |

Kolona Lift nam pokazuje lift meru pravila. Zanimljiva su nam samo pravila koja imaju Lift meru različitu od 1 – ona koja imaju veću od 1 su pravila kod kojih se stavke pojavljuju zajedno više nego što očekujemo, a ona koja imaju lift manje od 1 su pravila gde se stavke zajedno pojavljuju manje nego što je očekivano. Iz tabele vidimo da su Lift mere u rasponu od 0.603 do 3.445. Izvući ćemo najzanimljivija pravila korišćenjem čvora GroupBy. Biramo ona koja imaju najveću Lift meru, najveću podršku i najveću pouzdanost. Koristimo čvor Rule-based Row Filter i u njemu navodimo pravila da Lift treba da bude maksimalan, kao i Support

i Confidence (slika 3.1.) . Pre toga, povežemo i čvor Table Row to Variable kojim pretvaramo red tabele u promenljivu.



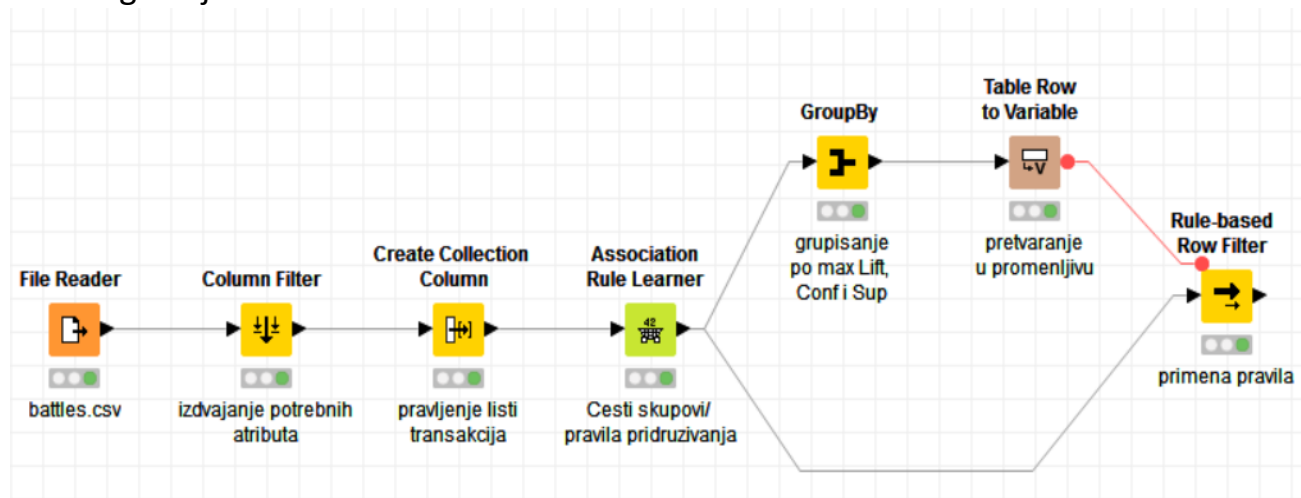
Slika 3.1.

| Row ID | D Support | D Confide... | D Lift | S Consequent | S implies | (...) Items |
|--------|-----------|--------------|--------|--------------------------|-----------|--|
| rule6 | 0.053 | 1 | 1.583 | Robb Stark | <--- | [ambush,loss] |
| rule7 | 0.053 | 1 | 1.583 | Robb Stark | <--- | [ambush,Balon/Euron Greyjoy] |
| rule11 | 0.053 | 1 | 1.357 | win | <--- | [pitched battle,?] |
| rule15 | 0.053 | 1 | 1.357 | win | <--- | [Joffrey/Tommen Baratheon,siege,Stannis Baratheon] |
| rule16 | 0.053 | 1 | 3.455 | siege | <--- | [Joffrey/Tommen Baratheon,Stannis Baratheon,win] |
| rule17 | 0.079 | 1 | 1.357 | win | <--- | [?] |
| rule20 | 0.079 | 1 | 1.357 | win | <--- | [Joffrey/Tommen Baratheon,Balon/Euron Greyjoy] |
| rule48 | 0.105 | 1 | 1.357 | win | <--- | [Joffrey/Tommen Baratheon,siege,Robb Stark] |
| rule54 | 0.132 | 1 | 1.357 | win | <--- | [siege,Robb Stark] |
| rule63 | 0.184 | 1 | 1.357 | win | <--- | [Joffrey/Tommen Baratheon,siege] |
| rule65 | 0.184 | 1 | 1.583 | Robb Stark | <--- | [Joffrey/Tommen Baratheon,ambush,win] |
| rule69 | 0.211 | 1 | 1.583 | Robb Stark | <--- | [Joffrey/Tommen Baratheon,ambush] |
| rule71 | 0.211 | 1 | 1.583 | Robb Stark | <--- | [ambush,win] |
| rule78 | 0.263 | 1 | 1.583 | Robb Stark | <--- | [ambush] |
| rule87 | 0.553 | 0.75 | 1.056 | Joffrey/Tommen Baratheon | <--- | [win] |
| rule88 | 0.553 | 0.778 | 1.056 | win | <--- | [Joffrey/Tommen Baratheon] |

Ovime dobijamo ovih 16 pravila. Pravilo iz tabele rule 16 ima najveći Lift. Može se zaključiti da su se neočekivano javile zajedno stavke iz pravila: kralj Joffrey Baratheon je napao kralja Stannisa i pritom pobedio => tip bitke je bio siege (opsada). Podrška ovog pravila je skoro na granici minimalne podrške, što nam govori da pravilo nije naročito korisno. Pouzdanost pravila je maksimalna (1), a to nam govori da je pravilo precizno, tj. da je visok nivo uzročnosti. Pravila sa najvećom podrškom su ona koja se javljaju često, što znači da su korisna i da je

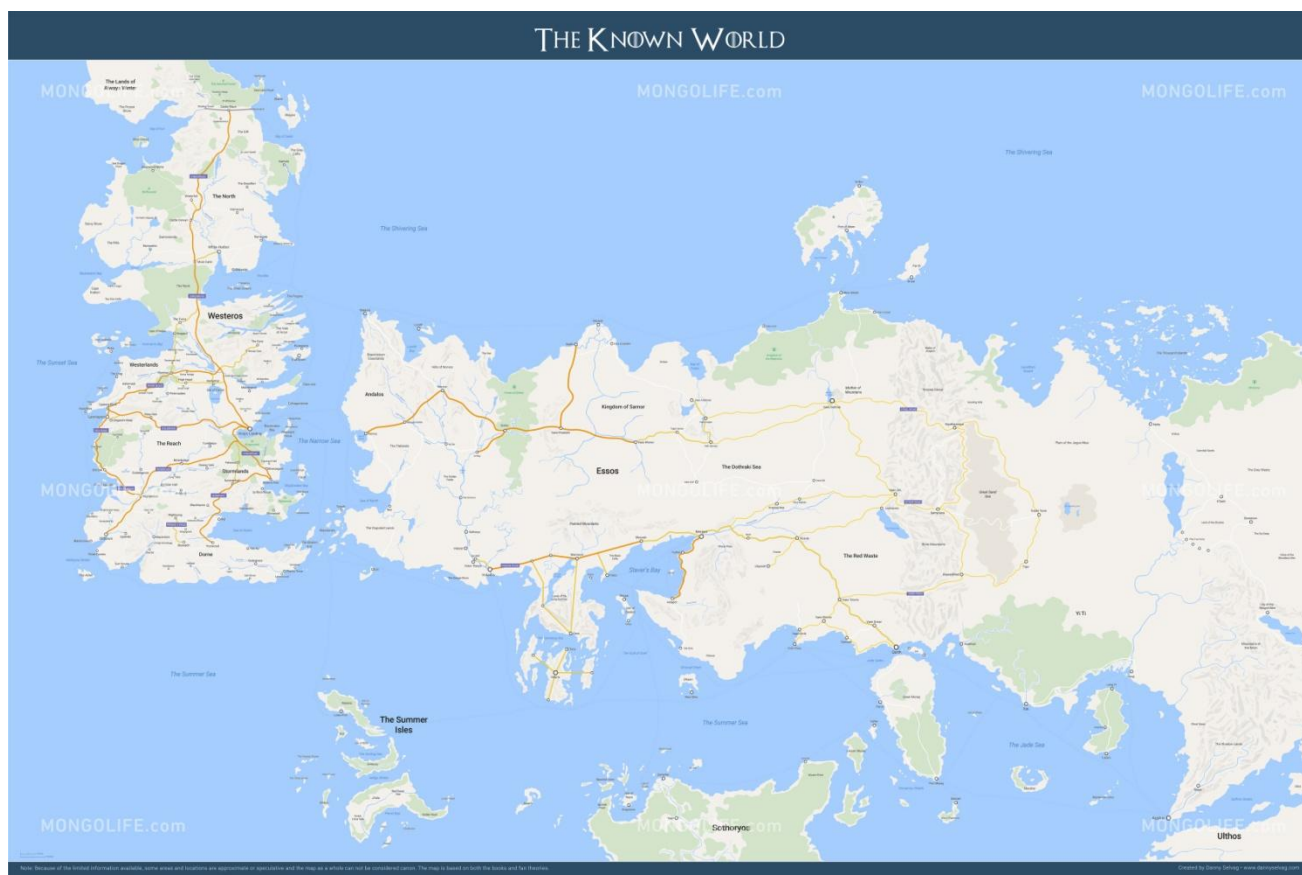
manja verovatnoća da su se pojavila slučajno. Ta pravila iz tabele su poslednja dva: Ako je ishod bitke pobjeda napadača, sledi da je napadač Joffrey Baratheon, i obrnuto, ako je napadač Joffrey, sledi da je on pobedio. Pravilo sa najvišom pouzdanošću je npr. drugo pravilo iz tabele: Ako je tip bitke zaseda, i napadač je bio GreyJoy, sledi da je napao Roba Starka.

Ovako izgledaju naši čvorovi u KNIME-u :



IV Klasterizacija podataka

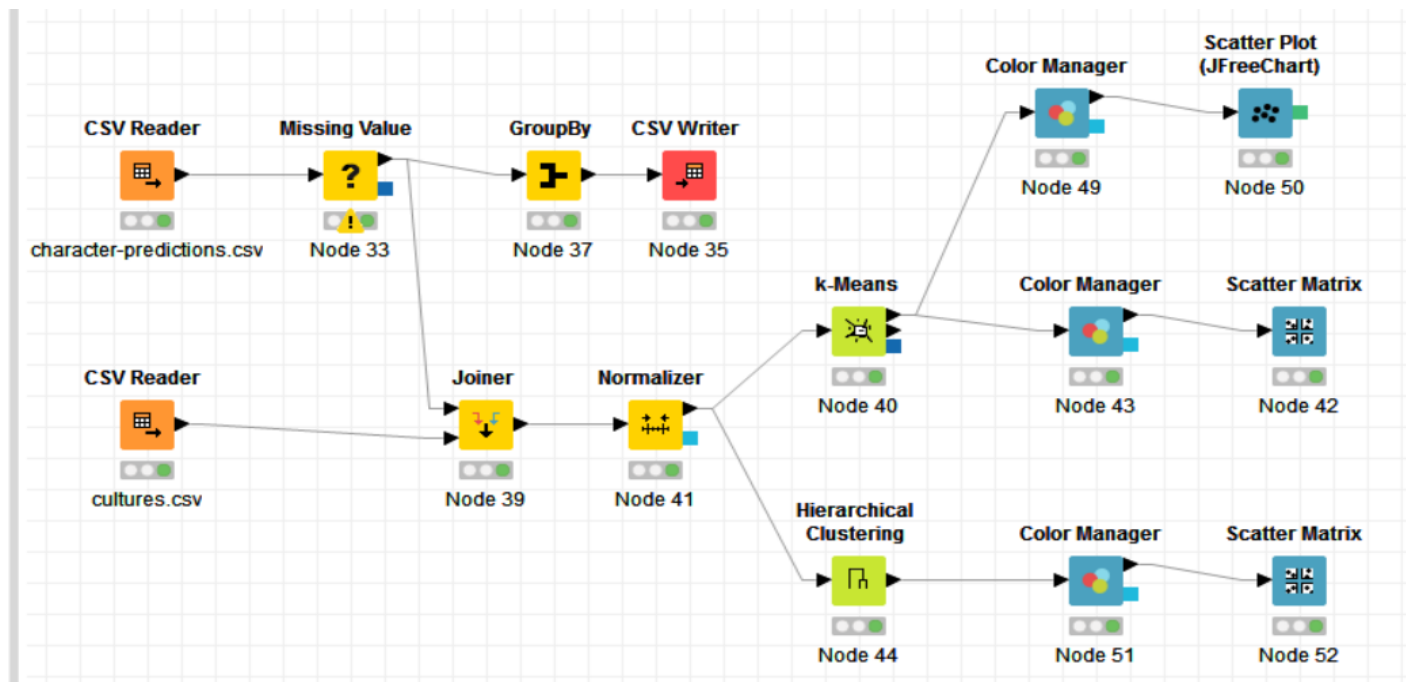
Ovaj skup podataka i nije baš pogodan za klasterovanje, jer ne poseduje neke značajne numeričke podatke. Klasterovanje ćemo izvršiti nad tabelom character-predictions. Grupisaćemo podatke u odnosu na kulturu lika. Pošto su nam kulture kategorički podaci, da bismo mogli da radimo klaster analizu, mi ćemo za svaku kulturu odrediti položaj tog naroda na mapi Vesterosa, na osnovu domenskog znanja iz serije. Položaji će biti određeni x i y koordinatom, a to su pikseli na slici mape (slika 4.1.). Napravićemo novu tabelu cultures.csv u kojoj će pored kolone sa kulturama za svaku kulturu postojati i kolona x i kolona y. Te dve kolone unećemo ručno u tabelu, nakon što pronađemo mesto gde narod živi, na osnovu koordinata piksela. Spojićemo tu tabelu sa početnom i onda ćemo da vršimo klasterizaciju u odnosu na x, y, i popularity iz prvobitne tabele, i iz toga videti da li će popularni likovi biti smešteni u istoj regiji na mapi.



Slika 4.1. – položaji mesta su koordinate piksela na slici

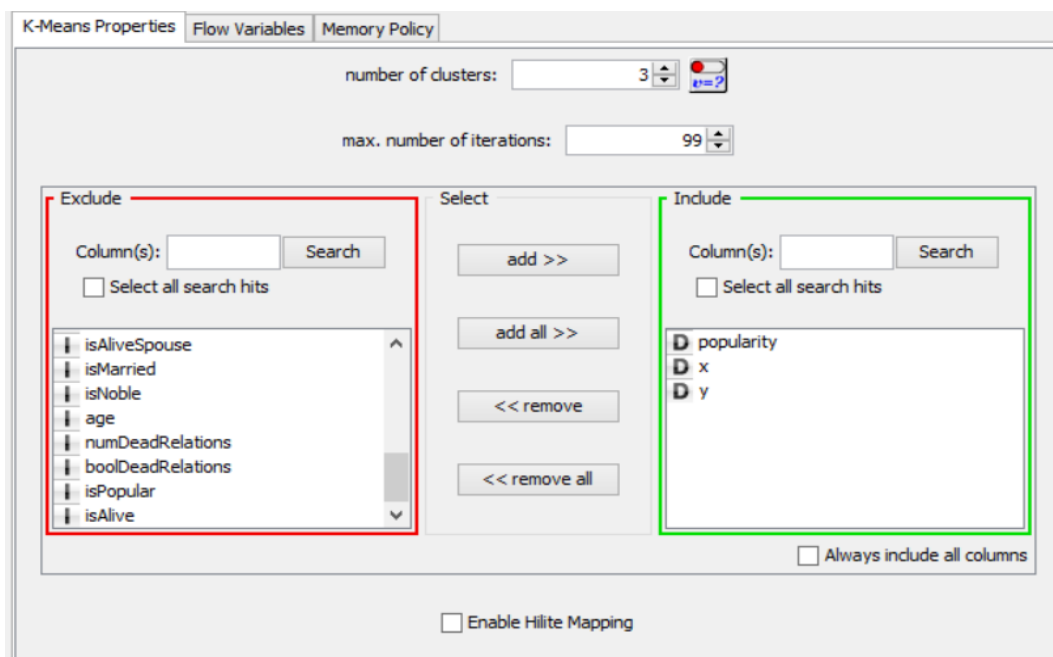
Prvo ćemo napraviti posebnu tabelu koja sadrži samo kolonu sa kulturama. Iz početne tabele izbacujemo sve redove u kojima je kultura nedostajuća vrednost pomoću čvora Missing Value. Zatim primenjujemo čvor GroupBy, kojim grupišemo po kulturama i tako dobijamo tabelu koja sadrži 64 različite kulture. Novodobijenu tabelu čuvamo u novom fajlu **cultures.csv** koristeći čvor CSV Writer (prikaz tabele na slici 4.2.). Tu tabelu učitaćemo sa novim CSV Reader čvorom, a zatim iskoristiti čvor Joiner kojim ćemo spojiti početnu tabelu sa našom novom tabelom. Spajanje vršimo po koloni cultures. Nakon toga, Joiner povezujemo sa čvorom Normalizer, jer se vrednosti pre klasterizacije moraju normalizovati. Njega spajamo sa čvorom k-means, koji vrši klasterovanje podataka algoritmom k-sredina (slika 4.3.). K-sredina radi tako što nasumično odabere k instanci koje predstavljaju polazne centroide, a klasteri se dobijaju tako što se te instance dodeljuju najbližim centroidima koristeći npr. Euklidsko rastojanje. To se obavlja u nekoliko iteracija, sve dok se centroidi menjaju. Joiner ćemo povezati i sa Hierarchical Clustering čvorom, koji vrši algoritam hijerarhijskog klasterovanja. Iz njega ćemo moći da vidimo dendogram (slika 4.4.). Izabraćemo da broj klastera bude 3. Klasterovanje vršimo u odnosu na atribut popularity, x i y. Za prikaz

klastera koristićemo Scatter Matrix (slika 4.5.) , kao i Scatter Plot da prikažemo raspored klastera u odnosu na x i y osu (slika 4.6.). Povezani čvorovi u KNIME-u izgledaju ovako:

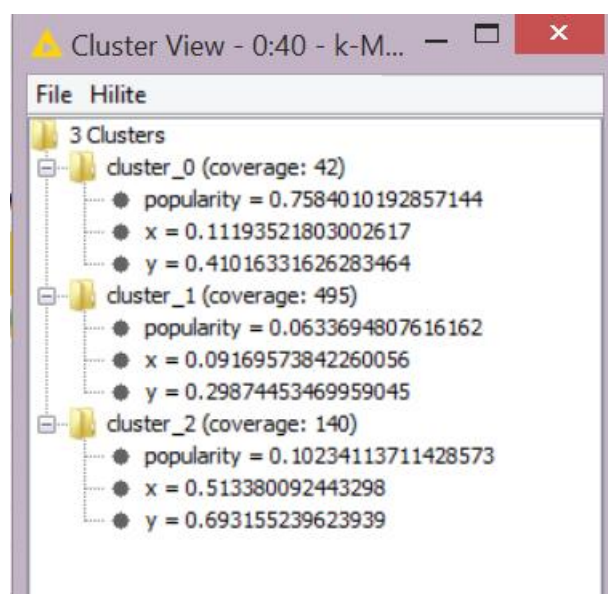


| Table "cultures%20(1).csv" - Rows: 64 Spec - Columns: 3 Prope | | | |
|---|--------------|------|------|
| Row ID | S cultures | x | y |
| Row0 | Andal | 963 | 999 |
| Row1 | Andals | 965 | 1001 |
| Row2 | Asshai | 2905 | 1745 |
| Row3 | Asshai'i | 2906 | 1748 |
| Row4 | Astapor | 1733 | 1533 |
| Row5 | Astapori | 1735 | 1533 |
| Row6 | Braavos | 932 | 790 |
| Row7 | Braavosi | 931 | 790 |
| Row8 | Crannogmen | 460 | 715 |
| Row9 | Dorne | 506 | 1417 |
| Row10 | Dornish | 506 | 1419 |
| Row11 | Dornishmen | 508 | 1417 |
| Row12 | Dothraki | 1887 | 1189 |
| Row13 | First Men | 963 | 999 |
| Row14 | Free Folk | 488 | 188 |
| Row15 | Free folk | 486 | 188 |
| Row16 | Ghiscari | 1863 | 1366 |
| Row17 | Ghiscaricari | 1863 | 1368 |
| Row18 | Ibbenese | 2179 | 539 |
| Row19 | Ironborn | 247 | 860 |
| Row20 | Ironmen | 247 | 862 |
| Row21 | Lhazareen | 2139 | 1169 |
| Row22 | Lhazarene | 2139 | 1170 |
| Row23 | Lysene | 951 | 1455 |

Slika 4.2. – izgled tabele cultures

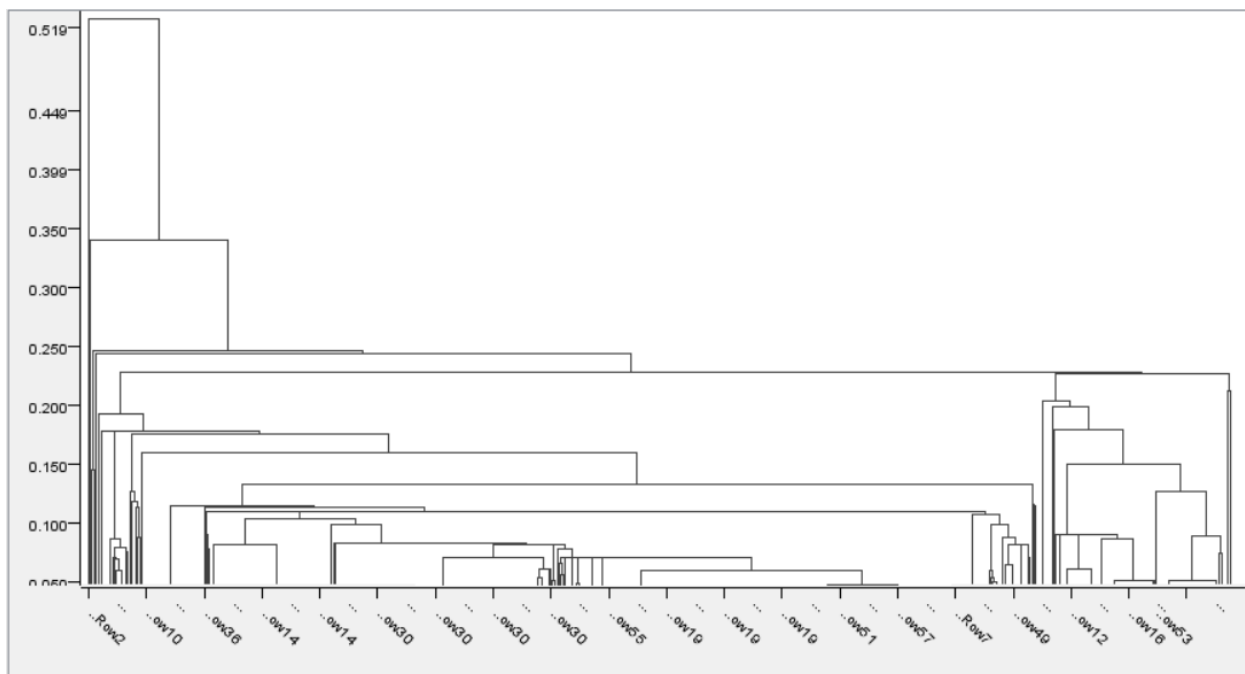


Slika 4.3. – konfiguracija čvora k-means

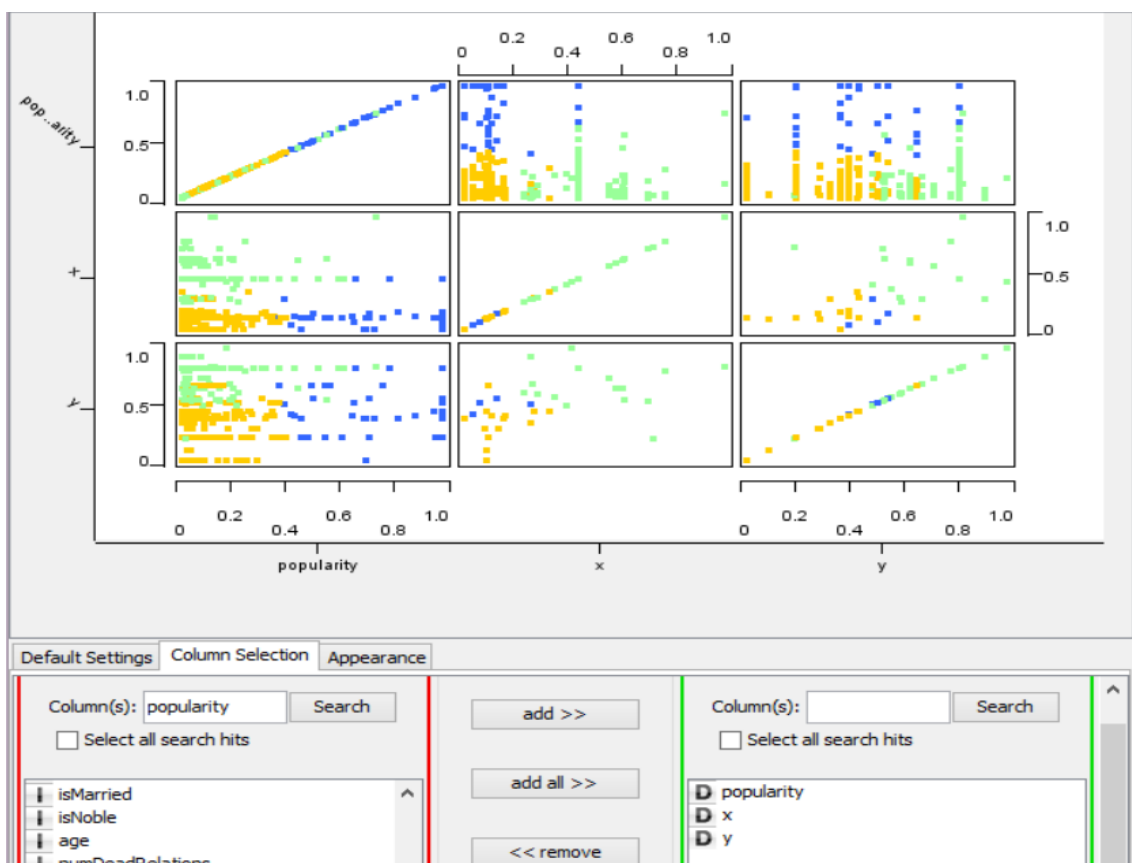


Na ovoj slici vidimo 3 klastera dobijena algoritmom k-sredina. Prvi klaster je svrstao popularnije likove i našao koordinate mesta. Druga dva klastera su manje popularni likovi i njihove koordinate. Koordinate $x = 0.11193$ i $y = 0.4101633$ iz prvog klastera vratili smo u nenormalizovani oblik i dobili koordinate lokacije na mapi.

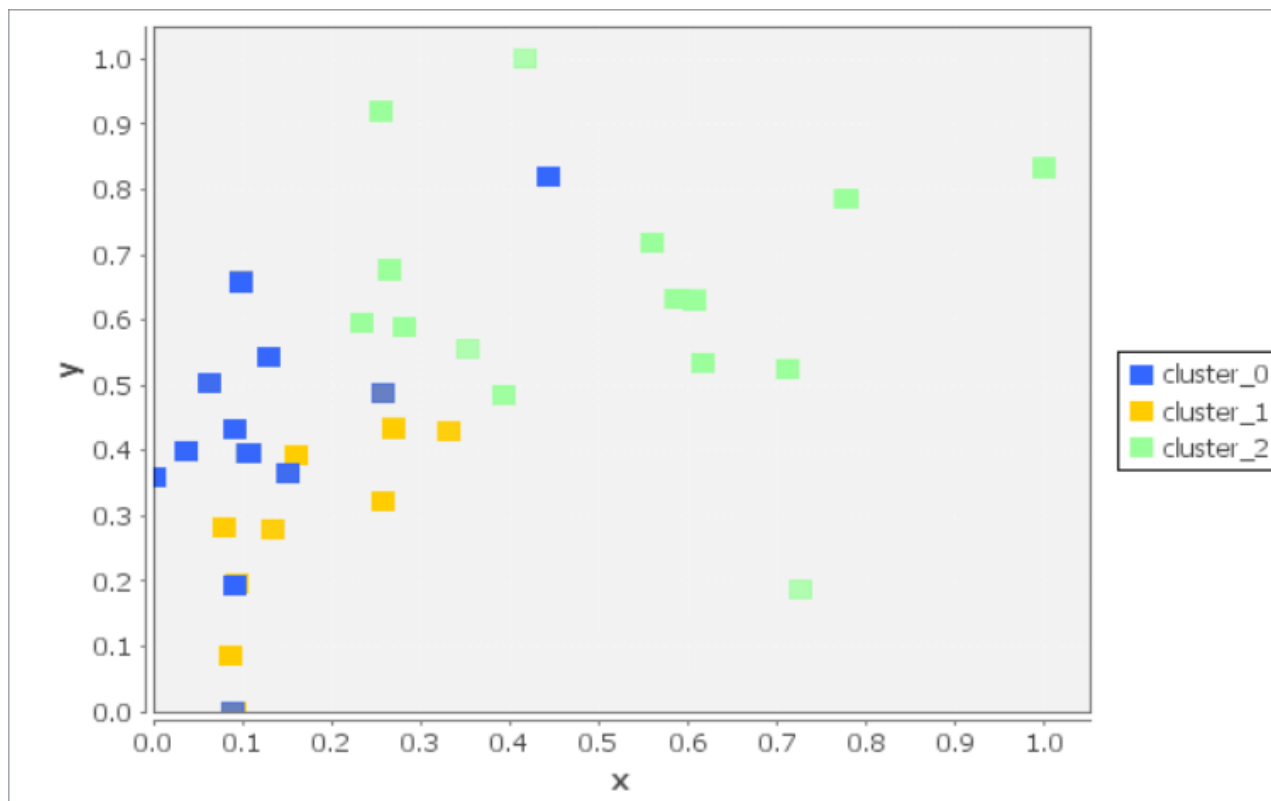
Ta lokacija na slici je negde na sredini između mesta na kojima se u samom serijalu najviše radnje odvija i bitna su za samu priču. Time vidimo da su popularniji likovi smešteni na tim bitnijim područjima.



Slika 4.4. – Dendogram

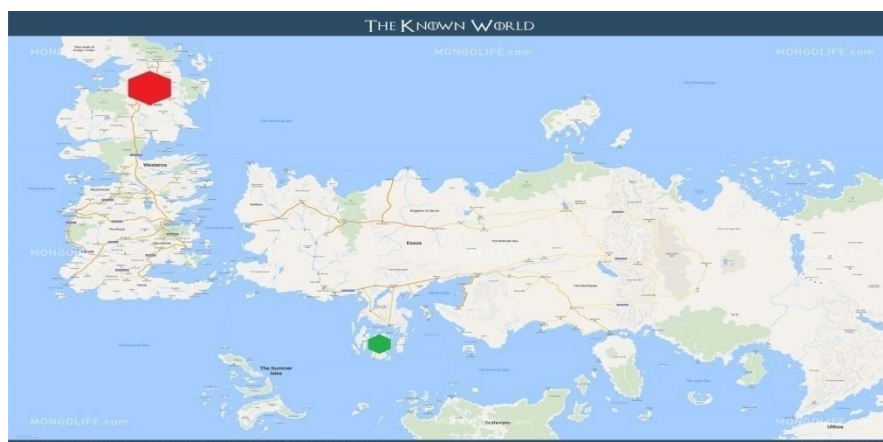


Slika 4.5. – Scatter Matrix



Slika 4.6. – Scatter Plot

Sa ove slike vidimo da su instance plavog klastera, koji predstavlja prvi klaster sa najpopularnijim likovima svi smešteni na bliskim koordinatama. Jedino tačka sa koordinatama $x = 0.444$ i $y = 0.82$ se ne nalazi blizu ostalih. Pronalaženjem tih koordinata u tabeli zaključujemo da ta tačka predstavlja kulturu Valyrian, za koju znamo da jeste jedna od poznatijih i bitnih za priču, a da se ne nalazi blizu ostalih bitnih regija.

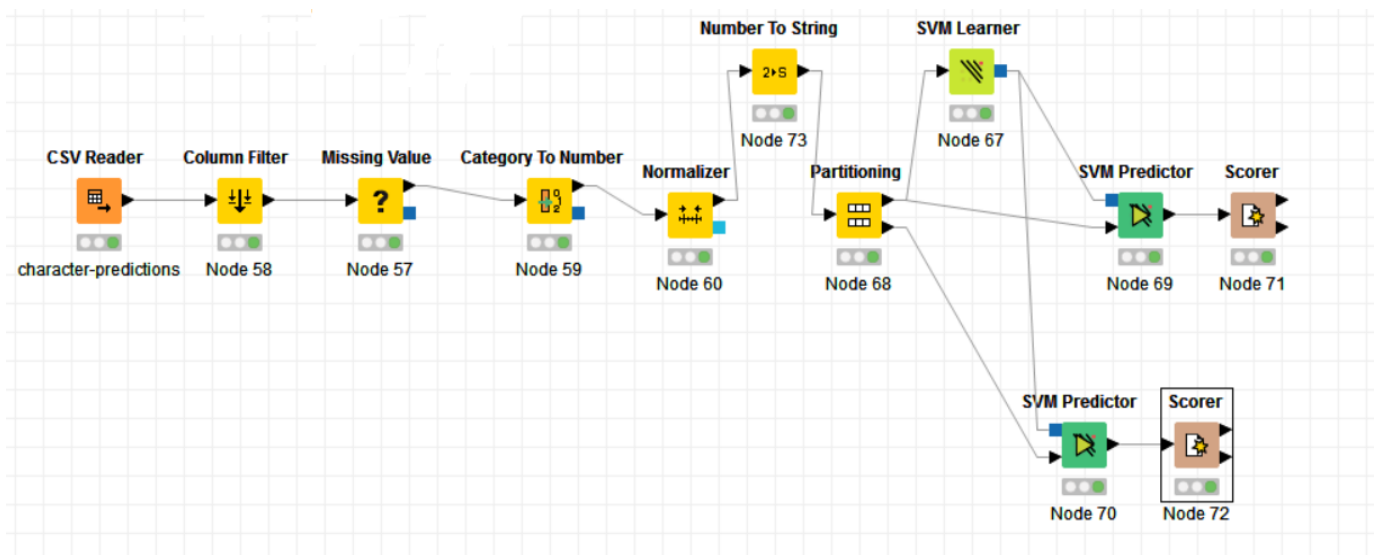


Na slici smo crvenom bojom označili prosečne koordinate kultura najpopularnijih likova koje smo dobili u prvom klasteru, a zelenom bojom je označen narod kulture Valyrian. Sa slike se vidi da su prilično udaljeni, što se i vidi na Scatter Plotu .

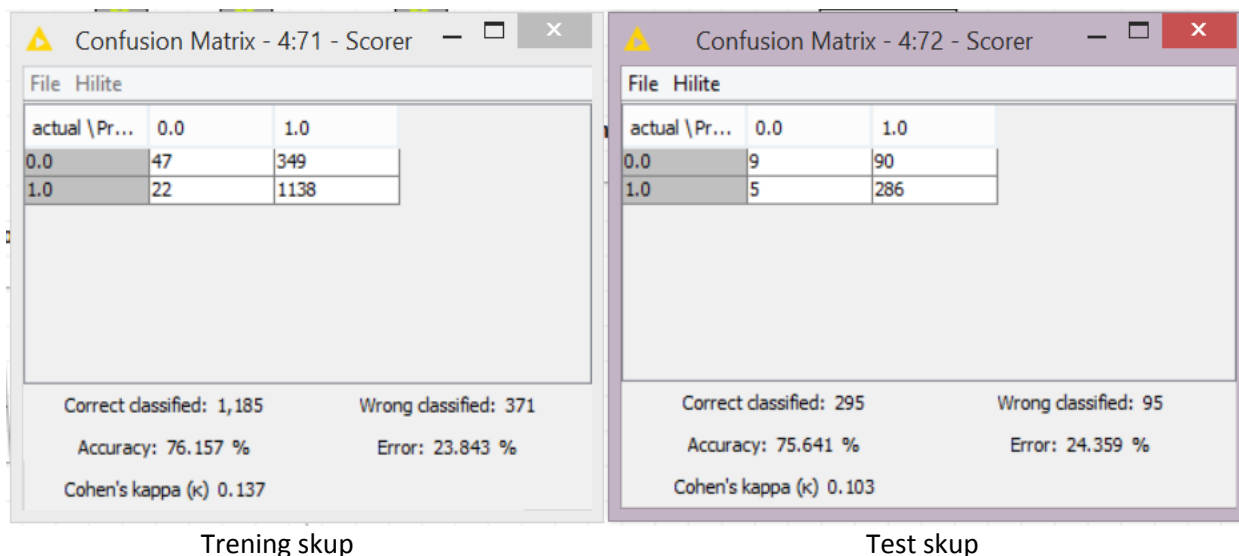
V Klasifikacija

Klasifikaciju ćemo izvršiti nad skupom character-predictions. Pokušaćemo da predvidimo ko od likova će preživeti, tj. svrstavamo u klase Dead i Alive. Nakon što učitamo podatke, prvo ćemo iskoristiti Column filter i izbaciti sve nepotrebne attribute, kao što su name, plod, pred, isAliveMother itd. Sačuvali smo kolone: actual (da li je živ), title, male, culture, mother, father, pojavljivanje u knjigama, age itd. Zatim smo obradili nedostajuće vrednosti (čvor Missing Value), i za svaki nedostajući string postavili Unknown, a za svaki integer smo uneli -1. Primenili smo dve tehnike klasifikacije : SVM (Support Vector Machine) i drvo odlučivanja.

Prva tehnika je zasnovana na ideji vektorskih prostora. Kod nje je model formula, a ne skup pravila. Potrebno je u vektorskom prostoru u kome su predstavljeni podaci odrediti razdvajajuću hiper-ravan, tako da su podaci iz jedne klase sa iste strane ravni. Nakon toga, izračunava se rastojanje tačaka od hiper-ravni i na osnovu toga se određuje klasa (iznad i ispod ravni). Da bi čvor SVM Learner prihvatio podatke, prebacili smo sve kategoričke podatke u numeričke (čvor Category To Number). Normalizujemo podatke (čvor Normalizer), kolonu actual pretvaramo u String (Number To String). Čvorom Partitioning razdvajamo podatke za trening i podatke za test. Izabrali smo 80% podataka za trening, a ostatak za test. Imaćemo jedan čvor SVM Learner u kome se obrađuju podaci za trening, i dva čvora SVM Predictor – za trening i za test skup. Čvorovima Scorer dobijamo rezultate klasifikacije – matrice konfuzije i tabele sa statistikama o preciznosti. Izvući ćemo kolonu Accuracy i uporediti za oba skupa.



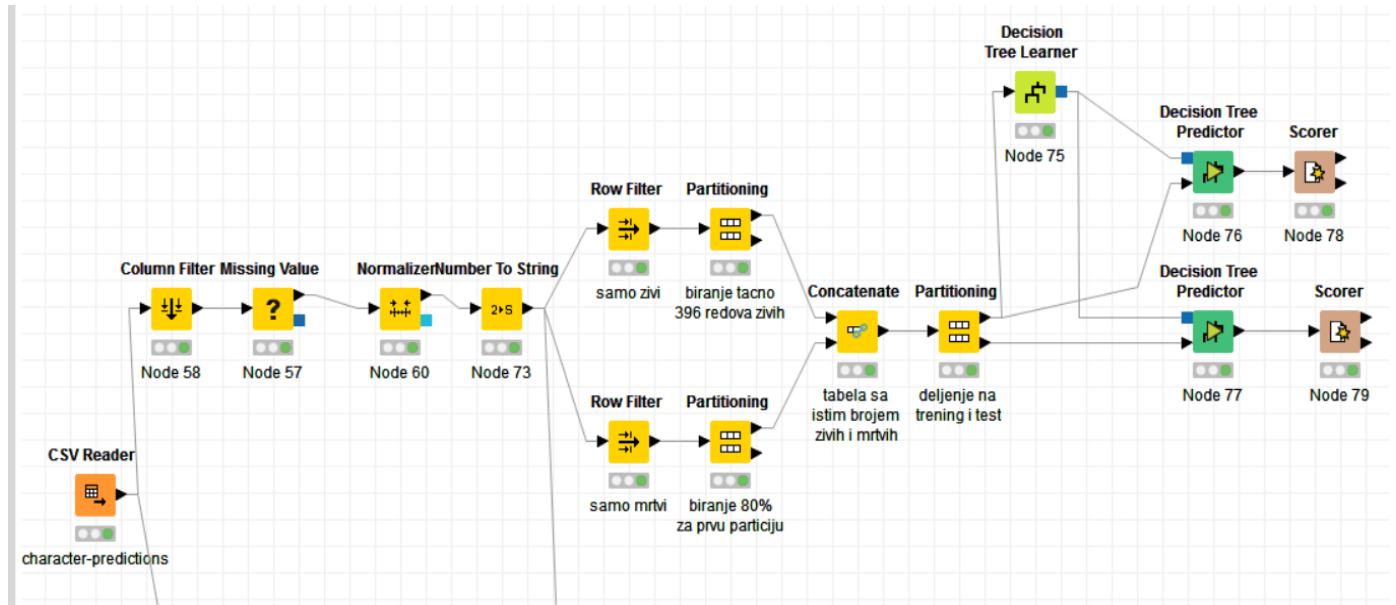
Matrice konfuzije dobijene iz ova dva čvora izgledaju ovako :



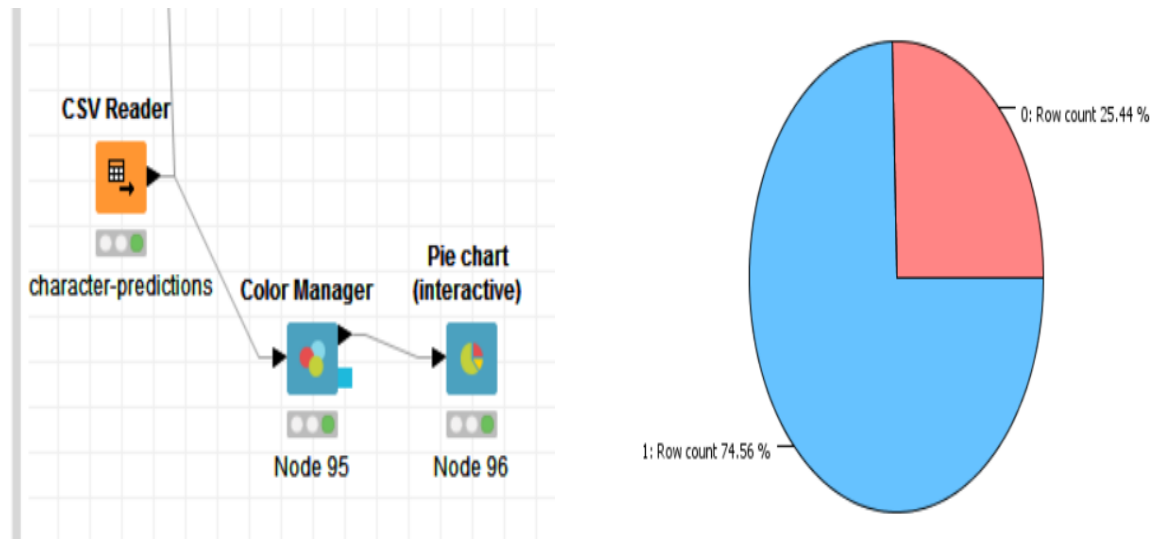
Vidimo da je Accuracy sličan u oba skupa, i iznosi oko 0.76, a time je stopa greške oko 24%. Ovde je klasifikacija u redu za žive likove jer vidimo da u matrici ćelija (1,1) ima mnogo više nego u (1,0), što znači da je za žive likove dobro predviđanje. Međutim, za mrtve ima više promašenih nego pogođenih, pa zato ovo nije dobar model. Dobar model bi izgledao tako da su na dijagonali veći brojevi nego brojevi van dijagonale. Ovi podaci su prilično nasumični i nemaju mnogo smisla, te nisu baš pogodni za klasifikaciju, i zato ne možemo dobiti klasifikaciju sa boljom preciznošću od ove.

Sada ćemo da pokušamo da dobijemo bolje matrice konfuzije, gde će na dijagonali biti više elemenata nego van nje. Pošto možemo videti da postoji mnogo više živih nego mrtvih likova (slika 5.1.), pokušaćemo da izbalansiramo taj odnos i da vidimo da li će to uticati na bolje predviđanje. Pretprocerisanje podataka je isto kao i za prethodni deo. Ovde ćemo pomoću čvorova Row Filter izvojiti samo žive i samo mrtve likove. Od mrtvih ćemo izabrati 80% pomoću čvora Partitioning (slika 5.2.) . Time dobijamo 396 redova u prvoj particiji. Onda ćemo od živih uzeti tačno 396 nasumičnih redova (slika 5.3.). Te dve tabele spojimo čvorom Concatenate. Ta konkatenirana tabela sada ulazi u čvor Partitioning kojim delimo na trening i test skup kao u prethodnom primeru. Povezujemo sa čvorom Decision Tree Learner, na kome podešavamo kolonu za koju vršimo predviđanje na actual, za meru kvaliteta smo izabrali Ginijev indeks. Ponovo imamo dva

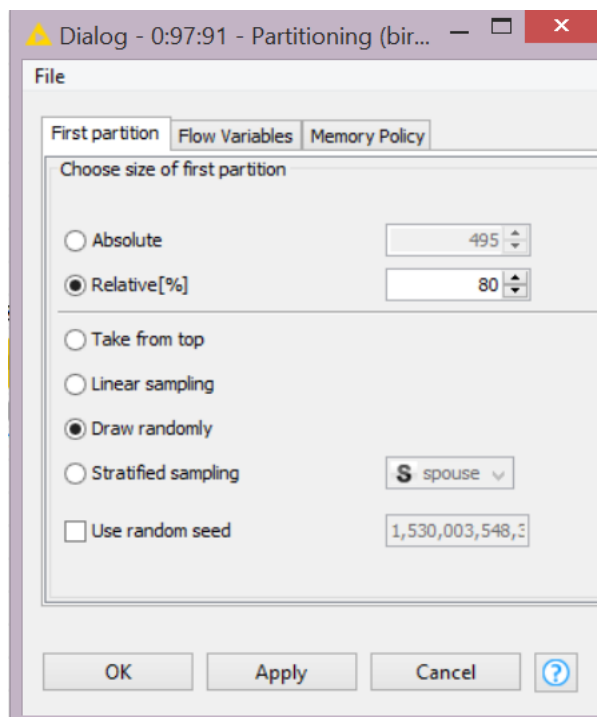
Predictora (Decision Tree Predictor) i dva Scorer čvora. Prikaz matrica konfuzije za trening i test skup je dat na slici 5.4., a za drvo odlučivanja na slici 5.5.



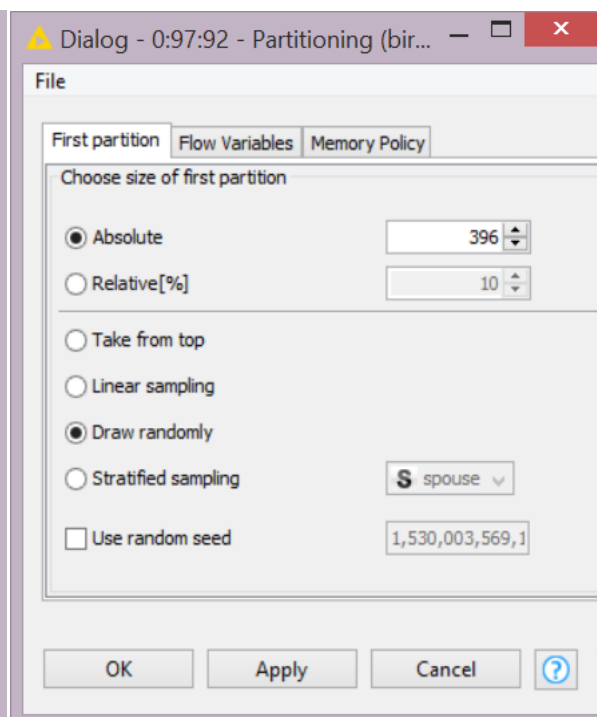
Prikaz čvorova u KNIME-u



Slika 5.1. – Živih ima skoro 75%



Slika 5.2



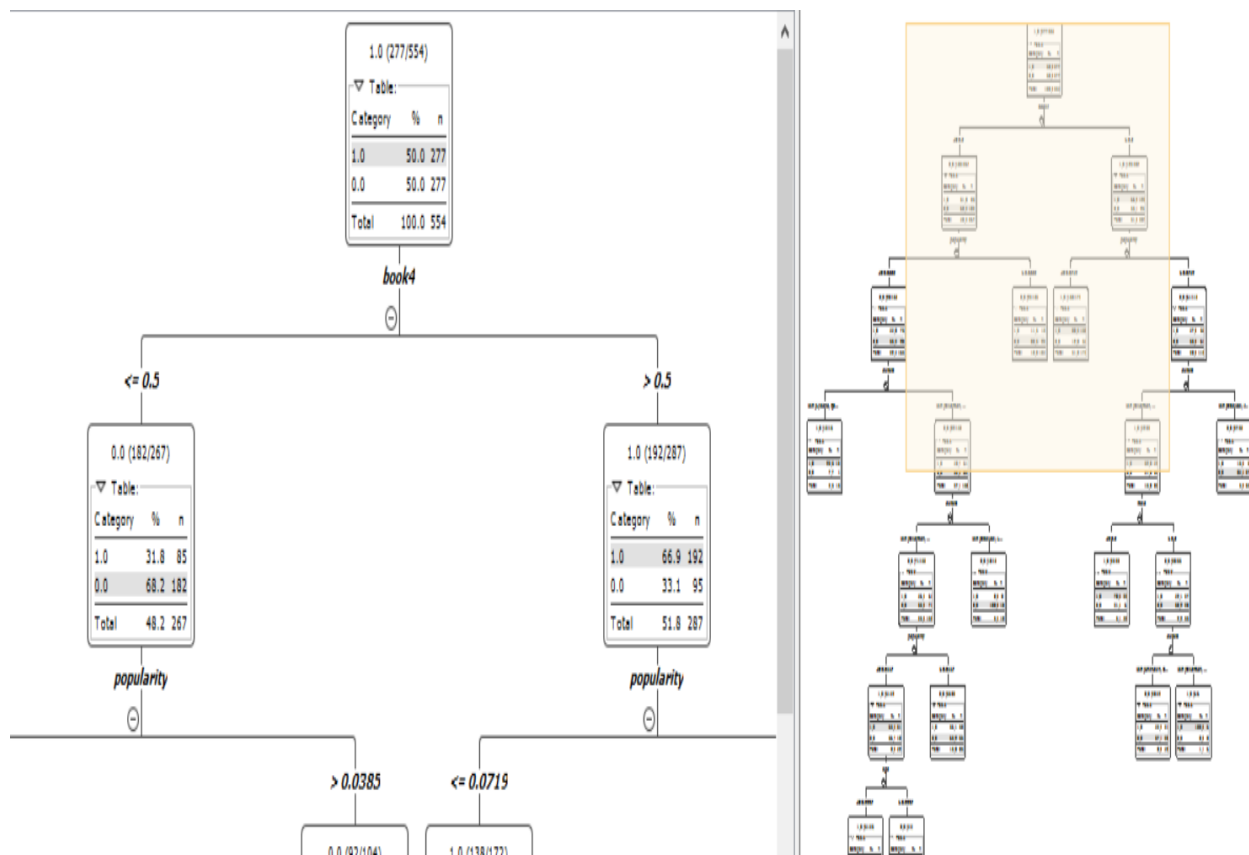
Slika 5.3.

| actual \ Prediction (actual) | 0.0 | 1.0 |
|---------------------------------|-----|-----------------------|
| 0.0 | 223 | 54 |
| 1.0 | 68 | 209 |
| Correct classified: 432 | | Wrong classified: 122 |
| Accuracy: 77.978 % | | Error: 22.022 % |
| Cohen's kappa (κ) 0.56 | | |

| actual \ Prediction (actual) | 0.0 | 1.0 |
|----------------------------------|-----|----------------------|
| 0.0 | 93 | 24 |
| 1.0 | 33 | 85 |
| Correct classified: 178 | | Wrong classified: 57 |
| Accuracy: 75.745 % | | Error: 24.255 % |
| Cohen's kappa (κ) 0.515 | | |

Slika 5.4.

Dobili smo da su veći brojevi na dijagonali, tj. ima više pogodjenih nego promašenih. Preciznost se malo poboljšala za trening skup, ali ne značajno. Ponovo naglašavamo da su podaci izmišljeni i vrlo nasumični, te se ne može se dobiti bolja klasifikacija od ovakve.



Slika 5.5.

Drvo prvo deli podatke na osnovu pojavljivanja lika u knjizi 4, što ima smisla jer ako su likovi preživeli do preposlednje knjige, mogu da umru samo još u poslednjoj. U sledećem nivou razdvaja podatke prema popularnosti, što je jedan od retkih atributa koji je zapravo korisna informacija u celom skupu.