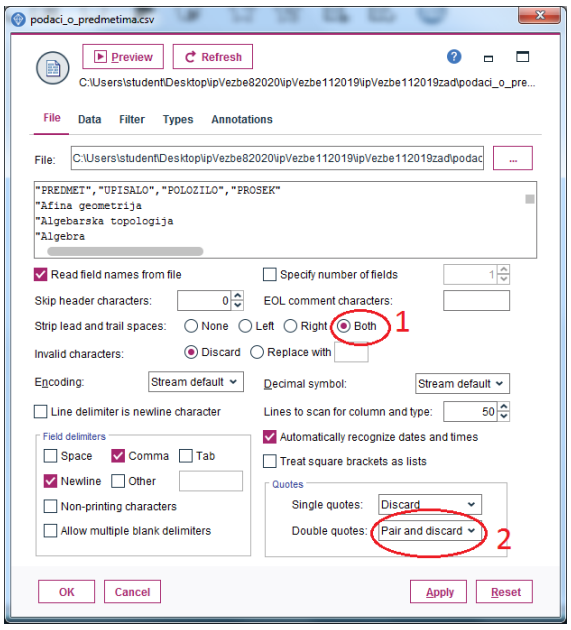
Istraživanje podataka 1 - vežbe 8, 2020.

Primer 2: Izvršiti klasterovanje predmeta primenom algoritma K-sredina u alatu IBM SPSS Modeler. Skup *podaci_ o_ predmetima.csv* ima atribute:

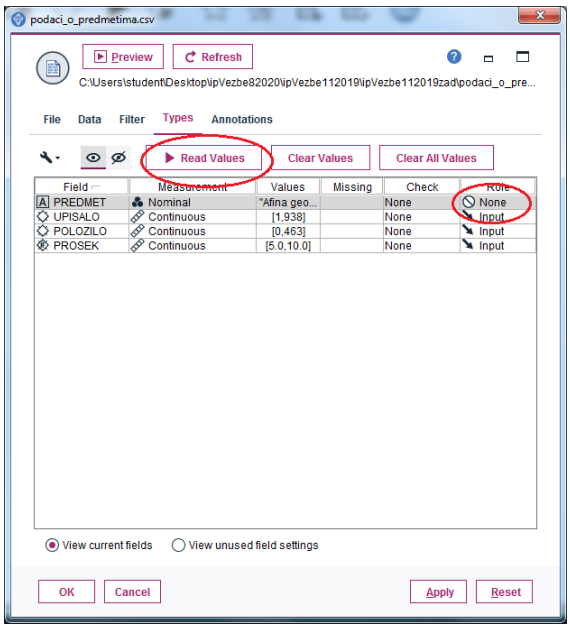
- *predmet* - naziv predmeta
- *upisalo* - broj studenata koji su upisali predmet
- *polozilo* - broj studenata koji su položili ispit iz predmeta
- *prosek* - prosečna ocena na položenim ispitima iz predmeta. Za predmete koje nijedan student nije položio, prosek je 5.

U radnom toku se prvo učitava skup pomoću čvora *Var. File*. Za pravilno učitavanje u odeljku *File* bitno je postaviti da se eliminišu beline sa početka i kraja u nazivu predmeta (označeno sa 1 na slici 1) i da se upare i eliminišu dvostruki navodnici u skupu. Tekst između dvostrukih navodnika je vrednost jedne ćelije (označeno sa 2 na slici 1).



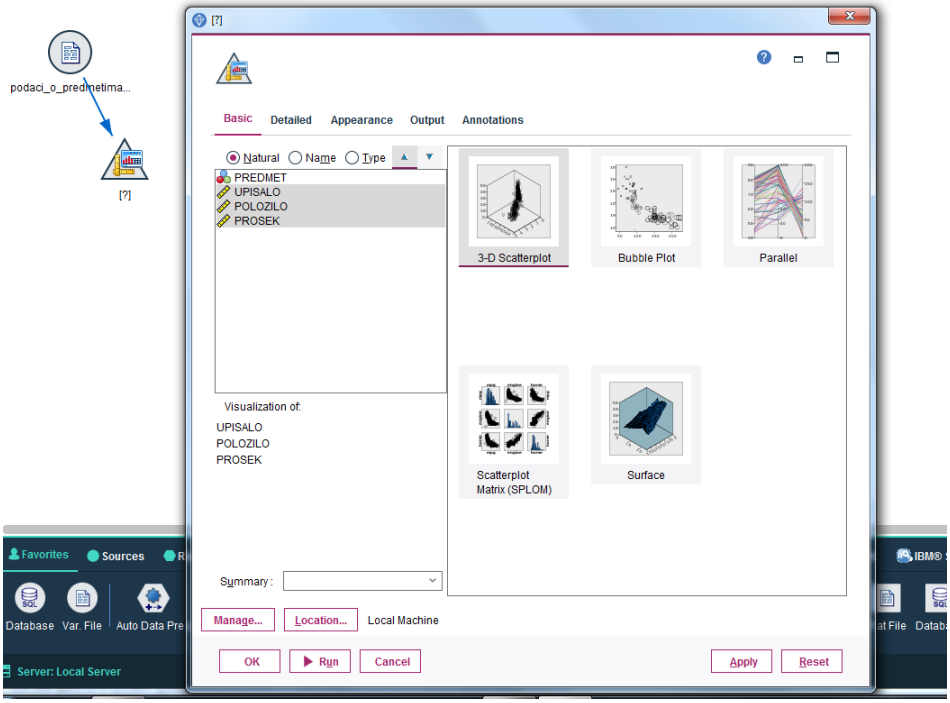
Slika 1: Postavke pri učitavanju skupa

U odeljku *Types* klikom na dugme *Read Values* učitavaju se podaci o vrednostima koje se javljaju u atributima skupa. Atributima koji učestvuju u klasterovanju uloga (*Role*) se postavlja na *Input*. Kako svaki predmet ima jedinstveno ime, naziv predmeta nema značaj u klasterovanju, njegova uloga se postavlja *None*, tj. neće biti korišćen pri klasterovanju (Slika 2).



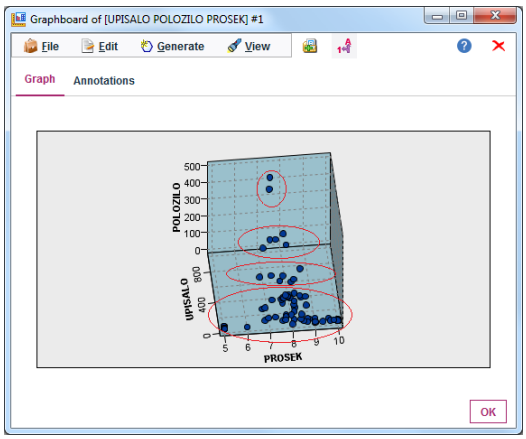
Slika 2: Učitavanju vrednosti u atributima skupa i dodela uloga atributim

Pošto se iz skupa za klasterovanje koriste tri numerička atributa, instance se mogu prikazati grafički pomoću 3D šeme sa raspršenim elementima, da bi se na osnovu vizuelnog upoznavanja sa skupom stekao utisak o broju klastera koji će biti zadat kao parametar u algoritmu K-sredina. Za odabir i prikaz grafika koristi se čvor *Graphboard* (Slika 3). Pomoću miša slika može da se rotira kako bi se šema sa raspršenim elementima videla iz različitih uglova.



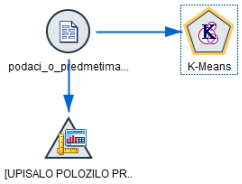
Slika 3: Izbor 3D šeme sa raspršenim elementima za grafički prikaz podataka

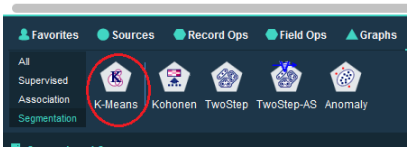
Na osnovu vizuelnog prikaza podataka, kao željeni broj klastera možemo zadati vrednosti od 2 do 4 (Slika 4). Primetiti: neke od označenih grupa na slici nisu globularnog oblika, te je za očekivati da ih algoritam K-sredina kao takve neće izdvojiti.



Slika 4: Prikaz podataka pomoću 3D šeme sa raspršenim elementima

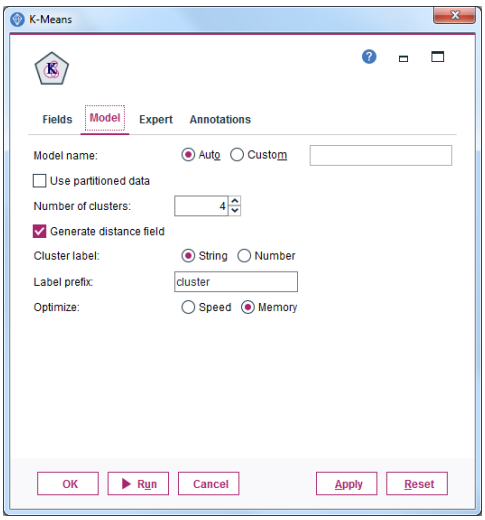
Da bi se primenio algoritm K-sredina na skup, čvor sa skupom podataka povezuje se sa čvorom *K-means* (podsećanje kako: klik na čvor sa skupom, taster F2, klik na čvor *K-means*) (Slika 5).





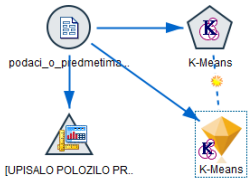
Slika 5: Izbor čvora *K-means*

Preko opcija dostupnih u čvoru *K-means*, u odeljku *Model* postavlja se broj željenih klastera na 4 i da se za svaku instancu u okviru modela klasterovanja prikaže rastojanje do najbližeg centroida, tj. da se skupu doda atribut *$KMD-K-means*. Za ostale opcije ostaju podrazumevane vrednosti (Slika 6).



Slika 6: Postavljanje vrednosti za opcije u čvoru *K-means*

Izborom opcije *Run* pravi se model klasterovanja koji je u radnom toku prikazan čvorom u obliku dijamanta (Slika 7).



Slika 7: Radni tok sa napravljenim modelom klasterovanja

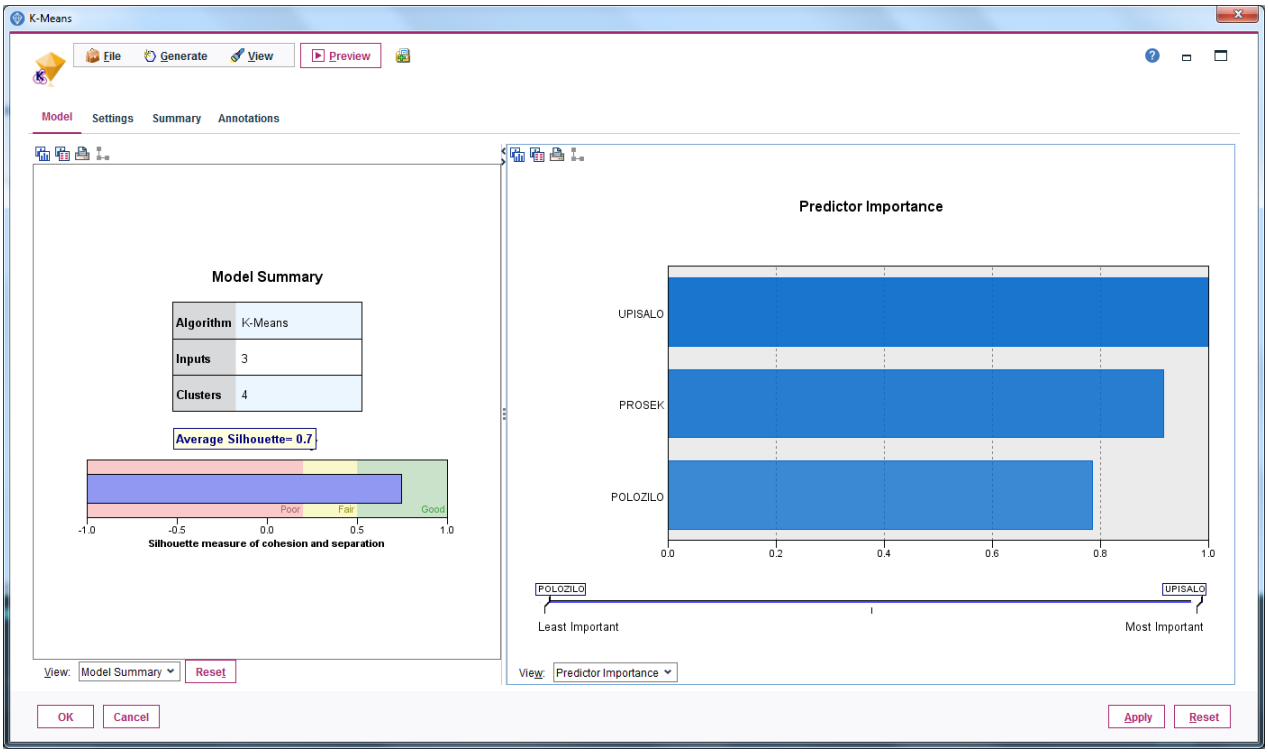
Duplim klikom na model klasterovanja može da se vidi rezultat klasterovanja i da se izvrši detaljnija analiza izdvojenih klastera.

Na pogledu *Model Summary* vidi se da je silueta koeficijent 0,7, čime se smatra da je izvršeno dobro klasterovanje. Na pogledu *Predictor Importance* vidi se da su svi atributi značajni za klasterovanje, pri čemu je atribut sa najvećim značajem *upisalo*, a sa najmanjim *prosek* (Slika 8).

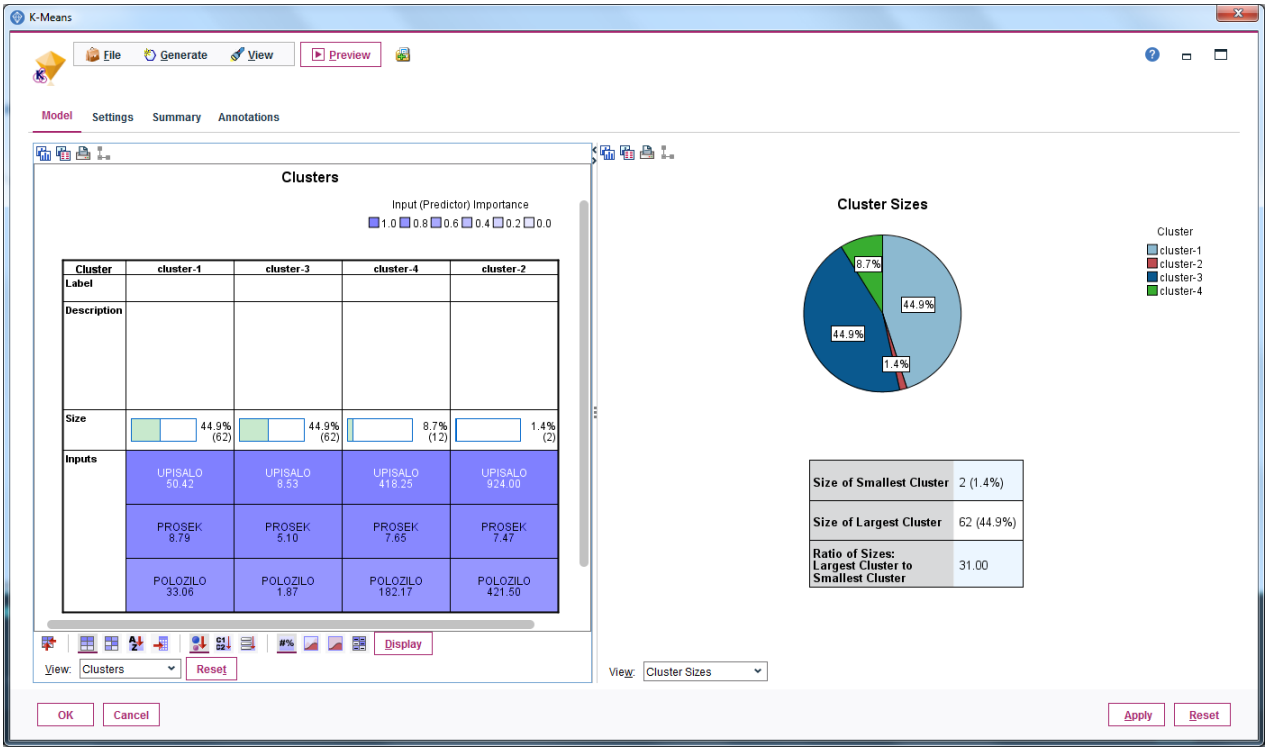
Na pogledu *Cluster Sizes* (Slika 9) vidi se da su izdvojena dva velika klastera (sa po 44,9% instanci) i dva mala (sa 8,7% i 1,4% instanci). Preko pogleda *Clusters* i *Cluster Comparison* (Slika 9 i Slika 10) se može uočiti šta je specifično za svaki klaster:

- *cluster 1* sadrži predmete sa visokom prosečnom ocenom (prosečna vrednost za atribut prosek je 8,79)
- *cluster 3* sadrži predmete koje je upisao mali broj studenata (prosečna vrednost za atribut upisalo je 8,53), a koje skoro nijedan ili mali broj studenata od upisanih je položio (prosečna vrednost za atribut prosek je 5,10, a za broj studenata koji su položili ispit 1,87). Tu spadaju predmeti sa doktorskih studija kojih ima puno, a u trenutku kada je baza pravljena mali broj studenata je iste i položio.
- *cluster 2* i *cluster 4* imaju slične prosečne vrednosti za prosek na položenim ispitima, ali je opseg vrednosti za prosek veći u klasteru 4 nego u klasteru 2. Razlikuju se značajno

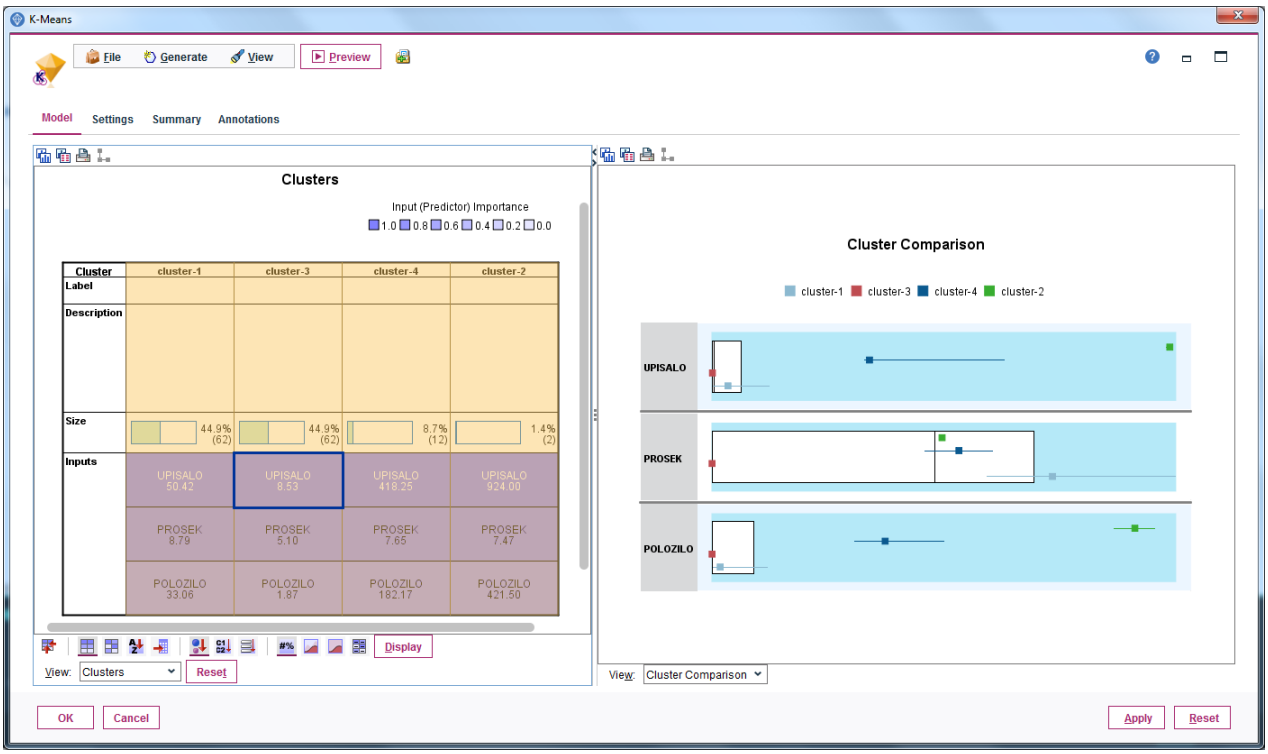
po broju studenata koji su upisali predmete, a zatim i po broju studenata koji su položili predmete.



Slika 8: Pogledi: *Model Summary* i *Predictor Importance*

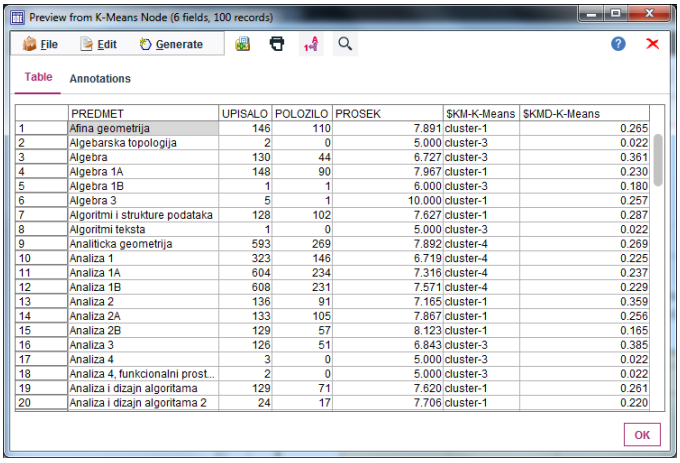


Slika 9: Pogledi: *Clusters* i *Cluster Sizes*



Slika 10: Pogledi *Clusters* i *Cluster Comparison*

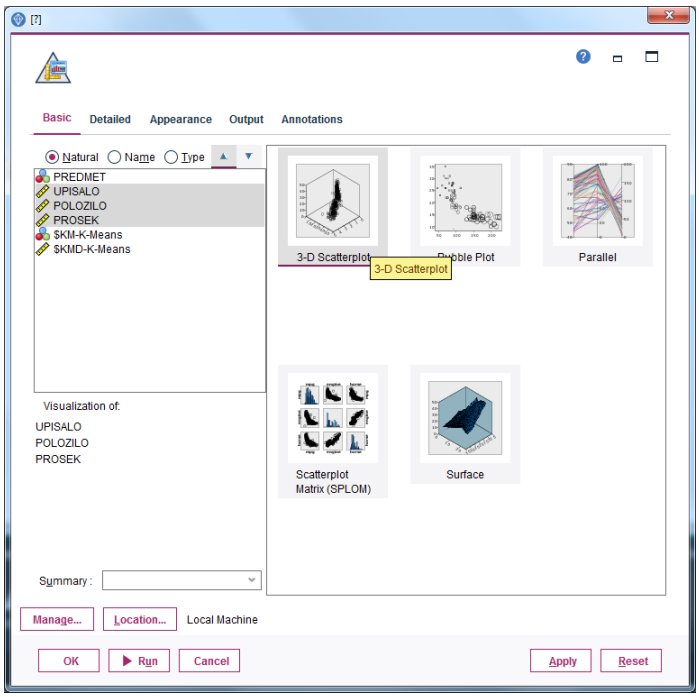
U modelu, klikom na dugme *Preview* može se videti za svaku instancu kom klasteru je dodeljena i koliko je udaljena od najbližeg centroida (Slika 11).



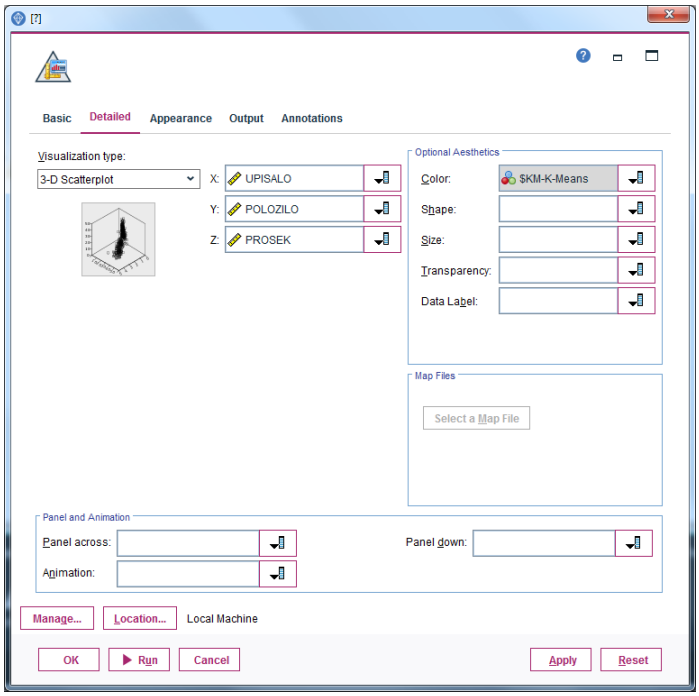
	PREDMET	UPISALO	POLOZILO	PROSEK	$KM-K-Means	$KMD-K-Means
1	Afina geometrija	146	110	7.891	cluster-1	0.265
2	Algebarska topologija	2	0	5.000	cluster-3	0.022
3	Algebra	130	44	6.727	cluster-3	0.361
4	Algebra 1A	148	90	7.967	cluster-1	0.230
5	Algebra 1B	1	1	6.000	cluster-3	0.180
6	Algebra 3	5	1	10.000	cluster-1	0.257
7	Algoritmi i strukture podataka	128	102	7.627	cluster-1	0.287
8	Algoritmi teksta	1	0	5.000	cluster-3	0.022
9	Analiticka geometrija	593	269	7.892	cluster-4	0.269
10	Analiza 1	323	146	6.719	cluster-4	0.225
11	Analiza 1A	604	234	7.316	cluster-4	0.237
12	Analiza 1B	608	231	7.571	cluster-4	0.229
13	Analiza 2	136	91	7.165	cluster-1	0.359
14	Analiza 2A	133	105	7.867	cluster-1	0.256
15	Analiza 2B	129	57	8.123	cluster-1	0.165
16	Analiza 3	126	51	6.843	cluster-3	0.385
17	Analiza 4	3	0	5.000	cluster-3	0.022
18	Analiza 4, funkcionalni prost...	2	0	5.000	cluster-3	0.022
19	Analiza i dizajn algoritama	129	71	7.620	cluster-1	0.261
20	Analiza i dizajn algoritama 2	24	17	7.706	cluster-1	0.220

Slika 11: Prikaz atributa koje je dodao model na originalan skup atributa

Pomoću 3D šeme sa raspršenim elementima može se i vizuelno prikazati rezultat klasterovanja (Slika 12). Svakom klasteru se dodeljuje jedinstvena boja, a svaka instanca se boji bojom koja je dodeljena klasteru kome pripada. (Slika 13).



Slika 12: Izbor 3D šeme sa raspršenim elementima za prikaz rezultata klasterovanja



Slika 13: Dodela jedinstvene boje svakom klasteru zbog prikaza rezultat klasterovanja

Za vežbu: izvršiti klasterovanje za drugačije vrednosti parametara i uporediti rezultate.